Appendix F

**Handy Dandy Guide to Analyzing the General Social Survey**

While this textbook is intended to help people understand the research process and to make sense of research results, some readers may want hands-on experience analyzing data. There are many software packages available that are often used in research methods and statistical analysis courses. Often these are costly and beyond the reach of casual users, but some universities make some analytical software available. Some software companies offer modified student versions that are affordable and work quite well.

But for those who want to just play with data analysis tools and have some fun testing ideas about why people think as they do, the University of California at Berkeley provides a free site. It has made available data from the General Social Survey (GSS) conducted by National Opinion Research Center (NORC) out of the University of Chicago. This site provides free data analysis tools that enable users to do simple and more complex analyses.

The site is called Survey Documentation and Analysis (SDA) and invites people to test the analysis tools for fast results. You can find it at: sda.berkeley.edu. The General Social Survey (GSS) cumulative files can be found in the archives or through this link: <http://sda.berkeley.edu/cgi-bin/hsda?harcsda+gss10>. The site provides documentation and helpful guides, but I have created this one to walk people through the basic steps in analyzing descriptive data using SDA software.

Before I start, please note that I have no connection to UC-Berkeley, NORC or SDA. I recommend this site because NORC have been asking question about public policy issues in surveys for a long time. In fact, I have used GSS data to illustrate various analytical concepts in the text. As a teacher, I appreciate that they provide a great learning opportunity to students by making data and easy to use analytical tools freely available.

My intention is to keep the analyses in this guide as simple as possible. I will not use weighted data, so the results will be simply the information reported by the respondents. People doing dissertation research will have to read all the details in the coding books and understand more about sample designs and weights used by NORC, and the various statistical tools provided by SDA.

I also will show how the SDA tables look in only the first few examples. After that, the tables will be consistent with format used in this textbook.

Any mistakes and errors in this guide are mine.

**Basic Concepts**

 Data File: Sometimes called a database or data set, this is a specific set of data that is in a file format which resembles a spread sheet with rows and columns. The rows contain the information for each person or case; this is called a “record.” Every record will have a unique identifier, usually called an ID, along with information about how each person answered all the survey questions.

 Variables: These are the specific questions that were asked in the survey. For example in the typical survey data set, age, gender, race, income, opinion about gun control, etc. are variables.

 Variable Name: This is the shorthand name given to the variable. Sometimes it is self-explanatory, like age or sex. But other times, it is a short abbreviation like RIncome and Gunlaw.

 Variable Labels: Variable labels are a more in-depth description (usually closer to the English language) of the variables. What does RIncome mean? It means the respondents income. Gunlaw asks about whether people favor or oppose gun permits. For the GSS, the variable names and labels are already programmed.

Values: Each variable contains information about how people answered the question. Values are short-hand code used in data entry. For example, gender might be coded as M or F, but more typically it is coded with numbers, such as 1 or 2. Just remember, that even though the data is entered into the data filet using numbers, variables like race, gender, and religion are still nominal-level data. They may look like numbers, but they should not be treated like numbers: you cannot average gender or religion, even though the computer will calculate it if you ask.

Similarly, if you ask people whether they agree or disagree with raising taxes, the data will be typically be coded with numbers: 1-5 on a 5 point scale. This is an ordinal measure—going from more to less agreement. It is possible to construct a mean and many researchers find means to be a quick way to make sense of the data. I prefer to use percentage distributions, but this is an unsettled area.

If you are working with a 5-point agree-disagree scale, the data might be coded as 1 for very strongly disagree, and 5 for strongly agree; alternatively, it might be coded as 1 for strongly agree and 5 for strongly disagree. Researcher can code variables however they want, so this it helps to check to see how the data are coded.

For some variables, the values are the actual data. For example, the GSS asks people their age, the number of hours they worked last week, and the number of children. Their responses are the actual numbers that are entered into the computer. These are interval/ratio level data. If you ask the computer to calculate the mean age or hours worked, you will get a valid result.

Just a note here: if people do not answer a question, it will be coded as “missing”—again using a number, often a 9 or 99. You will typically not include missing in your analysis. The good news: SDA already excludes missing data by default.

Value Labels: These labels show up on the output table. So, in our gender example, instead of having 1s and 2s, the computer will print out the value labels of Male and Female (or men and women, depending on what the programmer told the computer.

 On the plus side, the SDA is pretty much set up with variable names and labels, as well as values and value labels. The tables include them automatically.

Code Book: Sometimes called a data dictionary, this is the crosswalk between the survey instrument and the data file. It describes exactly how every variable is labeled and how all values are coded. If you click on the SDA’s codebook tab, you will be able to look at the alphabetical list of the variables.

**What Is On the SDA Site?**

On the top of the page you will see some pull-down menus: Analysis, Create Variables, Download, Codebooks, Search, and Get Started.

I will be using the GSS 1972-2010 Cumulative Data File (although they will continue to update this as new surveys are conducted).

**Crunching Data**

The SDA is pretty easy to use and pop-up windows provide additional information. Still, like learning any new skill, it is a process and I will walk through this step by step, keeping it simple. Variables used in analyses will be highlighted: GunLaw. However, feel free to choose other variables (I provide a short list at the end of this guide or select some from the extensive variable list).

Please note that I have approached the analyses for the guide outside the context of theory, literature review, or rigorous hypothesis testing. Research guided by theory helps maintain focus. This theory-less approach is sometimes called a fishing expedition, and while fun, it can get overwhelming without a roadmap.

**Let’s Begin: Frequencies (Frequency and Percent Distributions of Variables)**

* Go to the Analysis tab and it should give you some options.
* Click on Frequencies and Crosstabulations.

This page is divided roughly in half.

* On the left side is the Variable Selection table. The list of all the variables for the GSS 1972-2010 Cumulative data file is listed with blue highlighting. Scrolling through, you will see a whole lot of categories. When a category is selected by clicking, more options are then provided. It might take a few clicks to get to the desired variable. Once a variable is selected, it automatically goes into the Selected box. When the View button is clicked, a pop-up table will appear that shows the coding as well as the data for that variable.
* On the right side is the analysis table. This is where we tell the computer what we want: the variables and the desired analysis. Hang with me on this as I walk your through this.
1. **Required Variable names**

Row:

If we want to know how many respondents favored or opposed requiring gun permits, we would enter GUNLAW here in the Row variable.

1. **Table Options**
* The computer will automatically check column percent, but I made some other changes: To simply the analysis, I selected “unweighted” data rather than weighted data, on N cases to display.
	+ I unselected the color coding—I find it distracting, but it does provide some useful information. It is an aesthetics thing.
	+ I clicked the question text so it prints out: it helps to know exactly what was asked.
1. **Other options**
	* At the very bottom are options about decimals. I changed the Number of decimal places to display to zero. I prefer whole numbers unless having decimal places really matter, which they typically do not when looking at percent distributions..
	* To the right, you can choose to have the data presented as a chart. I chose “no chart,” but feel free to play this feature.

Once I am done with all my selections, I click on Run the Table

This is how the SDA table looks:

|  |
| --- |
| **Frequency Distribution** |
| Cells contain:-**Column percent**-N of cases | **Distribution** |
| **GUNLAW** | 1: FAVOR | **77.0**27,438 |
| 2: OPPOSE | **23.0**8,200 |
| ***COL TOTAL*** | ***100.0****35,638* |

Each cell contains the Column percent (the percent who favor and oppose in bold) and the N of cases (the number of respondents who favor and oppose). In technical terms, these are the percent and frequency distributions. So, of the 35,638 people who gave an opinion about whether they favor or oppose gun permits in the various surveys administered between 1972 and 2010, 77 percent favor a law that would require gun permits and 23 percent oppose such a law.

Let’s go through frequencies again. Suppose we want to find out how many men and women answered these surveys.

On the left side is the variable listing organized by main categories and sub-categories

* Click on Respondent Background and more options appear:
	+ Click on the sub-category Age, Gender, Race, and Ethnicity .
* Click on Sex

The variable Sex now appears in the Select box. Click on the View button to the right and it will give you a table that shows the frequency distribution, values and value labels. The View table looks like this:

|  |  |  |
| --- | --- | --- |
|

|  |
| --- |
| **Description of the Variable** |
| 23. Code respondent's sex |

 |
|

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Percent** |  | **N** |  | **Value** |  | **Label** |
| 44.0 |  | 24,260 |  | **1** |  | MALE |
| 56.0 |  | 30,827 |  | **2** |  | FEMALE |
| **100.0** |  | **55,087** |  |  |  | **Total** |

 |

From this table we see that 44 percent of the respondents were men and 56 percent were women. A total of 55,087 responded with valid data. Males were coded as 1 and females were coded as 2.

**Practice**: What are the frequencies of these variables? (variable name is highlighted)

* Race (Race of Respondent)
* Relig (R Religious preference)
* Degree (R’s Highest Degree)

**Measures of Central Tendency: Means, Medians, and Modes**

You can also use the Frequencies tool to summarize numeric, continuous/ interval/ratio level data: that is, data that are real numbers. Age and Childs (the number of children) are two real numerical variables.

What is the average age of the respondents? To get the frequency distribution and the measures of central tendency (means, medians and mode), put Age in the required row variable. I chose unweighted, no color coding, and no chart. I chose the text question.

* To get means, medians and modes (and some other stuff), click on the Summary Statistics.
* Then click on Run the Table.

As a general rule, you will probably not want to print out the table for Age because it is very long: it has the number of people who reported their exact age, ranging from 18 to 89. This is where the summary statistics like the measures of central tendency ( mean, median, and mode) become very handy in describing the data. SDA’s summary statistics table looks like:

|  |
| --- |
| **Summary Statistics** |
| Mean = | 44.32 |  | Std Dev = | 16.97 |  | Coef var = | .38 |  |
| Median = | 42.00 |  | Variance = | 288.11 |  | Min = | 18.00 |  |
| Mode = | 23.00 |  | Skewness = | .46 |  | Max = | 89.00 |  |
| Sum = | 2,365,240.40 |  | Kurtosis = | -.66 |  | Range = | 71.00 |  |

The mean (average) age is 44. The midpoint (the point where 50 percent are above and below) is age 42. The mode (the most commonly reported age) is 23. Note: the computer will often provide more information than needed to do a very simple analysis to describe the data. Advanced researchers will make use of this other data.

What about the number of children reported by the respondents? They were asked: “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).”

Childs goes into the Row, select summary statistics, and then run the table. The summary statistics tell us that the average number of children was 1.96 and the median was 2. The most frequent response (the mode) was none..

**Practice: Describe the amount of hours worked, hours spend watching TV and the years of education completed.**

* HRS1 (hours worked)
* HRSTV (hours watching TV)
* Educ (years of education completed)

**Crosstabulations**

Crosstabulations (crosstabs for short) is the workhorse of survey research. It is used to look at two or more variables simultaneously. The level of data matters: crosstab analysis is used with categorical data—both nominal data (such as gender and race) and ordinal data (such as opinion questions with scales: agree-disagree, favor-oppose, etc). It is not used with interval/ratio level data.

To do crosstabs, we will use the Frequency/Crosstab option in the Analysis menu. The programing table is on the right side of the screen: SDA-frequency/crosstabulation program.

Suppose I want to find out if men and women differ in terms of their attitudes about requiring gun permits. I would put GunLaw in the row variable. I put Sex in the column because I want to know if men and women have different opinions.

**A side story: Independent and Dependent Variables**

Once we get into crosstabs, we typically have to think about which is the independent variable and which one is the dependent variable. Basically, if we are trying to explain attitudes about gun permits, then GunLaw is the dependent variable. When we think that gender might explain differences in views about gun permits (testing the stereotype that women are kinder and gentler than men and therefore would be more likely to favor gun permits that would likely restrict access to guns), Sex becomes the independent variable.

Why does this matter? It matters because we need to tell the computer how to calculate the percent. What we want is a table that shows us the percent of men who favor and oppose gun permits, so we can compare it to the percent of women who favor and oppose gun permits. Therefore, we want the percentages to go down the columns. As it happens, SDA assumes that the dependent variable is in the row and its default is to percentage down the columns.

Go to the Table Options:

* Make sure Percentage the Column (the default) is checked

I did a few other things in Table Options to keep analysis simple:

* I checked unweighted cases to display
* I also checked to see the question text
* I also scrolled down to change the number of decimal places for percents to zero.
* Click on the Table button and you should see this table:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|

|  |
| --- |
| **Frequency Distribution** |
| Cells contain:-**Column percent**-N of cases | **SEX** |
| 1MALE | 2FEMALE | ***ROWTOTAL*** |
| **GUNLAW** | 1: FAVOR | **69**11,015 | **83**16,423 | ***77****27,438* |
| 2: OPPOSE | **31**4,894 | **17**3,306 | ***23****8,200* |
| ***COL TOTAL*** | ***100****15,909* | ***100****19,729* | ***100****35,638* |

 |
| **Text for 'GUNLAW'** 86. 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?**Text for 'SEX'** 23. Code respondent's sex |

Each cell contains the column percent for men, for women, and the total. It also shows the N of cases (the number of respondents). You can always tell the direction of the percent distribution because it will add to 100 percent. The Row Total is the same as the frequency distribution table for GunLaw alone: 77 percent favor and 23 percent oppose. The total number of respondents is 35,638.

**Interpreting the Data**: Now—to answer our question: do women view the issue differently? Looking at the table, both men and women favor laws requiring gun permits, but a somewhat higher percentage of women favor it: 69 percent of the men favor gun permits as compared to 83 percent of the women. It is, in my opinion, a small difference, but worth noting.

**Practice: Run these crosstabs and draw conclusions based on the results**

* Do men and women differ in terms of their views on Capital punishment for murder? Put CapPun in Row, Sex in Column
* Do views on capital punishment vary by race? Variables: CapPun and Race
* Are Democrats more likely than Republicans to support Social Security?

 Variables: NatSoc and PartyID

* Are people from the south more likely than people from the northeast to oppose gun permits? Variables: GunLaw and Region
* Who owns guns? OwnGun by demographic variables, such as Sex or Region

**Three-way Crosstabs: Adding a Control Variables**

The SDA programming gives us a chance to add more variables. One is a **Control variable**.This means we can add a third variable to the analysis to give more detail and complexity.

What happens when we disaggregate the data about gun permits by education? Using the Degree variable, which asks respondents the highest degree they have completed, we will now generate tables that looks at GunLaw and Sex for each category of Degree: less than high school, high school, junior college, college, and graduate degree.

* Put GunLaw in the row
* Put Sex in the column
* Put Degree in the control
* The table is set up as the first crosstab, with percentaging in the column.
* Run the Table

SDA produces five tables to show the crosstab data of GunLaw and Sex controlling for education (Degree). I have opted to simplify the presentation into a single table format used in this textbook. What a researcher needs to see in terms of technical information is not what a user of the results needs to see.

Exhibit Appendix XXX **Gendered Views on Gun Permits by Education Level**

Question: 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?

Percent of men and women **favoring** a law requiring gun permits

|  |  |  |  |
| --- | --- | --- | --- |
| Degree | Men | Women | Total |
| Less than high school | 68 | 81 | 75 |
| High school diploma | 67 | 83 | 76 |
| Associate’s  | 73 | 84 | 79 |
| Bachelor’s  | 73 | 87 | 80 |
| Graduate  | 80 | 90 | 84 |
| Total | 69 | 84 | 77 |

*Source:* National Opinion Research Center, General Social Survey, Combined Surveys 1972-2010

**Interpreting the data**: The majority (77 percent) of all respondents favor gun permits. While women do tend to favor gun permits slightly more than men across all education levels, we also see that those with more education tend to be more supportive of requiring gun permits. For example, 80 percent of men and 90 percent of women with graduate degrees favor gun permits as compared to 67 percent of the men and 83 percent of the women with high school diplomas. So, some nuance but basically the message is that the majority favor gun permits.

**Practice:**

* What about taxes? GSS asked: Do you consider the amount of federal income tax which you have to pay as too high, about right, or too low? Variable Name is Tax, and Label is Federal Income tax.
* Do views about taxes vary by Race? Education (Degree)? Sex? What about political party affiliation (PartyID)?
* What happens when you look at Tax and PartyID controlling for Sex?

**Selection Filters**

The SDA program offers another option: **Selection Filters**. Selection filters allows us to choose to focus on very specific segments of the respondents.

Here is how it works. Let’s go back to the crosstab of GunLaw and Sex. The cumulative data file covers almost 40 years but I am interested in the most recent survey in this file: 2010.

Look at the variable list on the left side of the page: find Case identification and Year and click on it. We really don’t care about the Identification variable, but we will find the year of the survey useful in our practice analyses.

 » YEAR - GSS YEAR FOR THIS RESPONDENT

This tells us what year the respondent participated in the GSS survey. We can use the Selection Filter to tell the computer that we want to look at a single value of the variable, or, in this case, a particular survey year.

Enter YEAR(2010) as the Selection Filter, then click on Run the Table.

For comparison, I want to see if views have changed much over time. I decide to look at 1973, the first time this question was asked. The enter Year(1973) as the Selection Filter, then click on Run the Table.

I then summarized the results of the two analyses into a single table:

**Exhibit Appendix xxx : Men and Women’s Views on Gun Permits in 1973 and 2010**

Question: 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?

Percent

|  |  |  |
| --- | --- | --- |
|  | 1973 | 2010 |
|  | Men | Women | Total | Men | Women | Total |
| Favor | 68 | 81 | 75 | 69 | 79 | 75 |
| Oppose | 32 | 19 | 25 | 31 | 21 | 25 |
| Total | 100% | 100% | 100%n=1,470 | 100% | 100% | 100%n=1,271 |

Source: Source: National Opinion Research Center, General Social Survey, 1973, 2010

Clearly, there is not much difference between these two points of time spanning 37 years.. It makes me wonder how in 2013 it seemed to come as a surprise that the majority of people wanted some limits on who owns guns. It appears from this data that a substantial majority have favored gun permits for almost four decades.

**Practice:** What other views might have changed over time?

* + What about the legalization of Marijuana (Grass)? Are views different in 2010 as compared to 1980?
		- To calculate, place Grass in row and Year(1980) in Selection Filter, and then run table.
		- Then place Grass in row, and Year(2010) in selection filter.
	+ You might want to look at views on Grass based on Sex, and then filter for Year(1980) and Year(2010). Alternatively, we could look at Grass and Sex, with Year as a control variable. This would give us the results for every year that an opinion question was asked about marijuana.
	+ Of course, we could use Year as the independent variable. This would be a very large table, but would be useful if we wanted to construct a trend line.

**Comparison of Means Analysis**

A comparison of means is used when the dependent variable is a real number. If I want to look at whether there are income differences between men and women, I need to use comparison of means. Looking the buttons at the top of the page, click on Analysis, and then select Comparison of Means.

We can use the GSS created income data called: ConRIinc. GSS created this variable that took the respondents’ income data and converted them to constant dollars (I believe to the year 2000). We can then select Sex as the independent variable.

The analysis table is set up a little differently than in crosstabs. The required fields are **Dependent Variable** and **Row Variable**.

If I want to compare the average salaries of men and women in the cumulative files, I will use ConRInc as the dependent variable. The row variable is Sex, which is the independent variable.

To keep the table simple, do not include any statistics or the chart.

Exhibit xxx: Comparison of Average Salaries of Men and Women

.

|  |  |
| --- | --- |
|  | Average Salary |
| Men | $39,619 |
| Women | $22,550 |
| Total | $31,453 |

Source: National Opinion Research Center, General Social Survey, Combined Surveys 1972-2010.

Looking at almost four decades of respondent information, with the dollars controlled for inflation, the average salary is just over $31,000, with men earning more than women, with a mean of $39,619 as compared to $22,550. Again, one survey is suggestive but for serious researcher would want to look at other data as well, including data from the U.S. Census or the Internal Revenue Service for comparison. .

**Practice:**

* Do men watch more TV than women? TVHours andSex.
* Is there a difference between men and women in terms of hours worked? HRS1 andSex
* Is there a difference in income (ConRInc) based on race or degree?
* Does the number of children (Childs) vary based on education (Degree)?

**Controlling for a Third Variable**

Sometimes researchers want more detail and complexity, so more variables are included in the analysis. This is similar to what we did with crosstabs. Does education make a difference in terms of the incomes of men and women? To test this, I will use Degree (a categorical variable) as my third variable and place it in the column field. (note: this is a bit confusing, since it gives the option of both the column and control—but trust me: put it in the column).

The data shows that education makes a big difference, but average salaries are still less for women than for men in all education categories.

Exhibit xxx: Comparison of Average Salaries for Men and Women with Different Educational Degrees

|  |  |  |  |
| --- | --- | --- | --- |
| Degree | Men | Women | Total |
| Less than high school | 24,679 | 12,539 | 19,889 |
| High school diploma | 34,013 | 18,674 | 26,341 |
| Associate’s  | 39,495 | 26,139 | 32,399 |
| Bachelor’s  | 54,949 | 31,307 | 43,619 |
| Graduate  | 74,354 | 46,206 | 62,038 |
| Total | 39,640 | 22,556 | 31,456 |

Source: National Opinion Research Center, General Social Survey, Combined Surveys 1972-2010

Looking at data prompts more questions. Maybe the income disparity between genders was greater earlier on and less so now (or maybe I have been watching too many episodes of Mad Men). What would the data look like if I compare average salaries by gender in 1975 to those in 2010?

To do this, I will use the Select Filter option. I will run table selecting for Year(1975), and then run another table selecting for Year(2010). I then combined those tables into one:

Exhibit xxx: Comparison of Average Salaries (Constant Dollars) for Men and Women with Different Educational Degrees: 1975 and 2010

|  |  |  |
| --- | --- | --- |
|  | 1975N=880 | 2010N=1,202 |
| Degree | Men | Women | Total | Men | Women | Total |
| Less than high school | $26,852 | $12,511 | $20,381 | $19,966 | $10,644 | $16,292 |
| High school diploma | 34,415 | 14,1556 | 25,803 | 29,780 | 17,419 | 23,361 |
| Associate’s  | 40,145 | 12,219 | 32,488 | 42,608 | 23,330 | 30,344 |
| Bachelor’s  | 45,536 | 15,361 | 32,825 | 59,021 | 37,003 | 44,688 |
| Graduate  | 49,438 | 30,037 | 42,341 | 59,021 | 57,342 | 58,124 |
| Total | 34,392 | 14,297 | 25,842 | 37,136 | 26,354 | 31,438 |

Source: National Opinion Research Center, General Social Survey, Combined 1975 and 2010

**Interpreting the Data**: Looking at the Total, average salaries increased for both men and women, but women earn less on average than men. Looking at education levels, women earn less on average than men at every education level. But what else? The average income of those with less than a high school degree is less in 2010 than it was in 1975 for both men and women. For men with a high school degree also showed a decline in average salary, although women showed an increase. For men with a graduate degree, income stayed about the same, but there was a substantial increase for women. Again, this data is controlled for inflation, so the actual dollar amount in 2010 will be higher, but by controlling for inflation, income is comparable over time.

**Practice:** Is there a difference in income based on Race or Degree?

**Another Side Story: Understanding Surveys**

While I have just been picking variables to look at, sometimes the data does not make sense. For example, I looked at data about the 2008 presidential election. Specifically, I wanted to see if there was a gender gap: that is, whether women were more likely than men to vote for the Democratic candidate (Obama) than for the Republican candidate (McCain). This is a simple crosstab using two variables: Pres08 (in the row, the dependent variable), and Sex (in the column, the independent variable).

Here is the data:

Exhibit new xxx 2008 Presidential Election: A Gender Gap? (in percent)

Did you vote for Obama or McCain?

|  |  |  |  |
| --- | --- | --- | --- |
|  | Men | Women | Total |
| Obama | 52 | 63 | 59 |
| McCain | 44 | 34 | 38 |
| Other Candidate | 4 | 2 | 3 |
| Didn’t Vote | 0 | 0 | 0 |
| Total | 100% | 100% |  *n* = 1,352 |

*Source:* National Opinion Research Center, General Social Survey, 2010.

Yes, women were more likely to vote for Obama than McCain (63% to 34%), and more likely than men to vote for Obama (63% to 52%). But notice, very few people said they did not vote. That struck me as odd because generally about 50 percent of eligible voters actually vote in presidential elections.

What is going on here? Maybe this was question was embedded in a skip pattern, meaning that they first asked people if they voted? I checked, and yes, that is what they did. The survey first asked if the respondent voted: Vote08 (Did R vote in 2008 election). If yes, the survey asked whether they voted for Obama or McCain: Pres08. If no, the surveyed skipped to a question about who they would have voted for: If08Who (who would you have voted for). This is how a skip question works.

**Practice:**

* How much of a factor was race on the 2008 election? Pres08 and Race.
* What about PartyID (party identification)? Or PolViews (liberal-conservative views)?

**Recoding Variables: A Useful Tool**

Sometimes researchers need to change the variables so they are easier to use. For example, using a continuous variable like Age is problematical because it will create a large table that is of little use in a crosstab. Try it. So, we need a way to convert continuous data or variables with a large number of categories into fewer categories. In other words, we will create a new variable by recoding. SDA provides a tool for recoding variables. It is found in the top menu under **Create Variables**.

**Recoding Age**

First it helps to know how a variable is coded. Type Age on the left side of the screen in the Select box, and click the View button. The pop-up table will show the frequency distribution of the variable—but more importantly, how it was coded. This is important. People were asked their age, so this is numeric data (real numbers)..

**How to recode?**

 Go the top menu and click on Create Variables. Then click Recode. A new menu will pop up on the right side of the screen. It will then ask for instructions:

* **Name of New Variable**: I call it AgeCat4 (but you can call it whatever you want. I use the original variable name, then add Cat, which stands for category, and then 4, meaning I am going to have 4 age categories.
* **Replace the Variable**: The default is set not to replace any existing variables—a wise choice.
* **The variable (or variables) that are to be recoded**. We will only recode a single variable: Age. For purposes of this recode, Age becomes Var 1.

Now I have to tell the computer specifically the **Recoding Rules**. I want to divide the people into 4 age categories: young adult (18-29), adult (30-49), 50-64) (older adults), and 65+ (retirement age). Clearly, these are arbitrary category choices. I am thinking life cycle here. You can make other choices. Warning: the age groups cannot overlap: if I had 18-30, and 30-49, those who said they were 30 would be counted in both categories.

**Recoding Rules: Creating the New Categories**: The values are the actual ages in what is now Var 1 and I am telling the computer to put everyone who said their age was between age 18 and 29 into a new output value “1”. Similarly, I tell the computer to put everyone who said their age was between age 30 and 49 into a new output value called “2”. Then everyone between age 50 and 64 is output value “3”. The last category is a bit different because I have to remember that missing values are coded 99, and I don’t want missing data included. If I said everyone between ages 65 and 100, I would have included the missing folks who were coded 99. Looking more closely at the variable, I see that SDA had already recoded people 89 years and above as 89. So, my retirement-age category is for those aged 65-89, and they are output value “4”.

**Recoding Rules Table:** Basically, Var 1 values are recoded into the output values and then get new output labels.

|  |  |  |
| --- | --- | --- |
| Output Value | Output Label | Var 1 |
| 1 | young adult (18-29) | 18-29 |
| 2 | adult (30-49), | 30-49 |
| 3 | Older adults (50-64)  | 50-64 |
| 4 | Retirement age (65+) | 65-89 |

Click on Start Recoding. A new table will appear for AgeCat4 and you want to check the frequency distribution just to make sure it looks fairly equitably divided.

|  |
| --- |
| **Description of the derived variable** |

**AGECAT4 Age 4 Categories**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Percent** |  | **N** |  | **Value** |  | **Label** |
| 21.5 |  | 11,800 |  | **1** |  | Young adult (18-29) |
| 40.0 |  | 21,957 |  | **2** |  | adult (30-49) |
| 21.1 |  | 11,559 |  | **3** |  | older adults (50-64) |
| 17.4 |  | 9,574 |  | **4** |  | retirement age (65+) |
|  |  | 197 |  | **.** |  | (No Data) |
| **100.0** |  | **55,087** |  |  |  | **Total** |

If I wanted to have a more even distribution, I could break adult into 30-39, and 40-49, since 40% are ages 30-49. I would then have five age categories. But for my purpose here, the 4 categories will work to answer this this question: For those who voted in the 2008 presidential election was there any difference based on age? I go back to the Frequency/Crosstabulation Analysis and tell the computer that Pres08 is in the row, and AgeCat4 is in the column.

 Exhibit xxx: Percent Voting for Obama or McCain in 2008 Presidential Election by Age Groups

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Candidates | Young Adult(18-29) | Adult(30-49) | Older adult(50-64) | Retirement Age (65+) | Total |
| Obama | 67% | 58% | 57% | 57% | 59% |
| McCain | 26 | 39 | 40 | 41 | 38 |
| Other | 6 | 3 | 3 | 2 | 3 |
| Total | 100% | 100% | 100% | 100% | 100%n=1,350 |

Source: National Opinion Research Center, General Social Survey 2010.

Yes, there appears to be some differences based on age, with the young adults somewhat more likely to have reported they voted for Obama than the retirement age group. I am surprised that it is not as great a difference as I expected, knowing that the CNN’s exit polls found that a majority (53 percent) of those 65 and older reported to have voted for McCain.[[1]](#footnote-1) So, there is a discrepancy and a reminder that survey results can differ. It is possible that exit polls might be a bit more accurate because the winners have not been determined, whereas a retrospective poll might have a bit of bias toward people reporting that they voted for the winner. However, this “winner bias” is just another theory to be tested.

**Practice**: Recode Age into 5 Categories and see if there are any differences in terms of age in who people voted for in various Presidential elections: Pres08, Pres04, Pres00?

**Recoding a Categorical Variable**

Respondents were asked: “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 to extremely conservative.” In this variable, labeled **Political Views** (PolViews), there were seven categories. While it is useful to see a gradation of strength of views, I want to simplify it to just three categories: Conservative, Moderate, and Liberal.

**Recoding PolViews**

* Go the top menu, click on Create Variables, and then click on Recode. A new menu will pop up on the right.
* **Name of New Variable**: I call it PolViews3 (but you can call it whatever you want). I use the original variable name and then 3, meaning this new variable will have only 3 categories.
* The default is set not to replace any existing variables—a wise choice.
* **Variable (or variables) that are to be recoded**. We will only recode a single variable: PolViews. This is, for purposes of this recode, Var 1.

It is important to see how PolViews is coded and the distribution. Put PolViews is the Select box and click the View button to see the pop-up table with the coding data. This actually is a good example of a bell-shaped distribution, with the greatest percentage being Moderates.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Percent** |  | **N** |  | **Value** |  | **Label** |
| 2.7 |  | 1,249 |  | **1** |  | EXTREMELY LIBERAL |
| 11.6 |  | 5,338 |  | **2** |  | LIBERAL |
| 13.0 |  | 5,973 |  | **3** |  | SLIGHTLY LIBERAL |
| 38.7 |  | 17,781 |  | **4** |  | MODERATE |
| 16.1 |  | 7,423 |  | **5** |  | SLGHTLY CONSERVATIVE |
| 14.8 |  | 6,800 |  | **6** |  | CONSERVATIVE |
| 3.1 |  | 1,438 |  | **7** |  | EXTRMLY CONSERVATIVE |
|  |  | 6,777 |  | **0** |  | IAP |
|  |  | 2,014 |  | **8** |  | DK |
|  |  | 294 |  | **9** |  | NA |
| **100.0** |  | **55,087** |  |  |  | **Total** |

**Recoding Rules:** Now I have to tell the computer specifically that I want to divide the respondents into 3 categories: Liberal, Moderates, and Conservatives. Again, this is arbitrary. The “extremely liberal” and “extremely conservative” might be quite different in their views than the other liberals and conservatives. It is also true that that some might have political views that are not captured on the liberal to conservative continuum. But we have what we have.

I recoded the **Output Variable** (my new Polviews3) with its values and value labels:

|  |  |  |
| --- | --- | --- |
| Output Value | Output Label | Var 1 |
| 1 | Liberal | 1-3 |
| 2 | Moderate | 4 |
| 3 | Conservative  | 5-7 |

Click on Recode the Variable. It will show this result:

|  |  |  |
| --- | --- | --- |
| Output Value | Output Label | Percent |
| 1 | Liberal | 27% |
| 2 | Moderate | 39% |
| 3 | Conservative  | 35% |
| Total |  | 100%n=46,002 |

Now let’s run a crosstab: Of those who did not vote in 2008, who would they have voted for given their political views? Variables: If08Who (dependent variable is in the row)and PolViews3 (independent variable is in the column), and the column is percentaged.

Exhibit xxx: Political Views and How Non-voters would have voted in 2008 Presidential Race

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Candidates | Liberal | Moderate | Conservative | Total |  |
| Obama | 72 | 63 | 51 | 61 |  |
| McCain | 11 | 16 | 30 | 19 |  |
| Other | 17 | 21 | 19 | 19 |  |
| Total | 100% | 100% | 100% | 100%n=554 | 100% |

Source: National Opinion Research Center, General Social Survey, 2010

Not a whole lot of insight other than only 30 percent who called themselves Conservative said they would have voted for McCain. It is important to remember, however, that they be be some “voting for a winner” bias here.

**Practice**: Recode PartyID and look at percent who voted in 2008 (Vote08) and 2000 (Vote00).

**Recode**: Relig (respondents religious preference), and look at who they voted for in the 2008 and 2000 Presidential elections? Did religion make a difference?

**Your Turn: Describing Data**

Frequencies, measures of central tendency, crosstabulations and comparison of means are essential analytical tools used to describe survey data. Given the number of variables in the GSS, researchers can explore for a long time. It is possible to look at the impact of birth signs on income if you want ( Zodiac). This is a fishing expedition and I encourage you to continue to play with this data. However, serious researchers are best served by having a very clear and defined analysis plan that answers very specific research questions or tests specific hypotheses.

**Variables to Play with**

>>Relig (religious preference)

>>NatWefare (Welfare)

>>NatPark (National Parks)

>>Gender Issues>>FEPol(men or women better suited to politics)

>>Abortion>>ABAny (abortion for any reason)

>>Family Planning, Sex, Contraception>>HomoSex (Homosexual sexual relations)

>> Family Planning>>TeenSex (Sex Before Marriage 14-16)

>>Family Planning>>EXmarSex (Sex with person other than spouse)

 >>Civic>>ObVote (Obligation to vote in elections)

 >>Government>>Helppoor (should government improve standard of living

>>Government>>HelpSick (should government help pay for medical care)

>>Happy (general happiness)

>>Life (Is life exciting or dull)

>>Trust (can people be trusted)

>>Anomie7 (public officials interest in average person)

>>Anomoie5 (lot of average man getting worse)

**Practice:** Telling a story about public confidence in various institutions: Most/least confidence? Changes over time? Variation based on demographic characteristics?

* + ConFed (confidence in executive branch federal government)
	+ ConLegis (Confidence in Congress)
	+ ConJudge (Confidence in US Supreme Court
	+ ConPress (Confidence in the press)
	+ ConTV (Confidence in TV)
	+ ConFinan (confidence in banks and financial institutions)
1. Exit poll reported on CNN: <http://www.cnn.com/ELECTION/2008/results/polls/#val=USP00p1> [↑](#footnote-ref-1)