This module is updated based on one created by Jeffrey Lashbrook of SUNY-Brockport to include more recent data and for ease of use with WebCHIP 4.0.

SOC 100

Introduction to Sociology Gender Inequality in the United States Data Analysis Assignment

By Jeffrey Lashbrook

For this assignment we will explore the impact of gender on the earnings of full-time workers in general and within a specific occupation. The purpose of this assignment is to give you additional practice with a data analysis software program (StudentCHIP), to develop some familiarity with working with Census data, to apply what you have learned in the course concerning cross-tabulation analysis to this data, and to link course material on gender to explain differences in earnings.

I. OBJECTIVES

Social science research is often categorized in terms of two objectives—to describe and to analyze the social world. The first presents data to build a portrait of the way things are (or were) while the latter brings conceptual frameworks to bear on the data to explain why things are the way they are. For this assignment, our objective is to both describe and analyze gender differences in earnings. We will answer the following questions:

- What is the nature of gender inequality in earnings among U.S. full-time workers, ages 25 and over, in the most recent year for which full census data are available?
- How do selected state results compare to the national picture regarding gender inequality?
- What might be some possible explanations for observed gender inequality in earnings?

Other learning outcomes for this assignment include honing analytical skills associated with generating, reading, and interpreting rudimentary forms of data analysis. Our goal is to interpret the information for a lay audience (e.g., write a summary that any newspaper reader could understand).

II. CENSUS DATA

(This description draws from W. Frey's Investigating Change in American Society.) The U.S. government can probably be accused of a lot of things, but when it comes to data-gathering efforts, they're not too shabby. The reason, of course, is that they have much more money to spend on it than do most researchers. As you're probably aware, one of the government's biggest efforts is the decennial census

(mandated by our Constitution). The Census' original purpose was for apportioning our congressional representatives every ten years. The first one was done in 1790; the most recent, of course, was the 2000 census. Although originally meant to count every single person, it does a lot more than that since it collects a lot of other information. Recently, for instance, the census bureau sent out two different surveys, a "short" and a "long" form. The "short" form asked just for basic sociodemographic information (e.g., age, gender, race/ethnicity etc.), while the "long" form added many more items on a variety of social characteristics (e.g., occupation, education, language proficiency). These characteristics represent some variables. The information on these characteristics is of immense importance for both planners at all governmental levels (federal, state, and local) and social science researchers. For this class, we can examine stratification-related phenomena like people's earnings, their occupational category, educational level, and much more.

Before we get to the actual assignment, one thing to remember is that no research effort is perfect, including the census's. They are never able to count everyone (e.g., all the homeless), and some people may be more likely to answer the different forms than others (we refer to this potential problem as response bias). The sample sizes are so large, however, that it still represents a pretty accurate picture of American people. Now, we turn our attention briefly to our software program to see how we will access our census data.

III. A PRIMER ON THE LOGIC OF SOME BASIC FORMS OF DATA ANALYSIS

(Note: Much of this repeats points made in class. Forgive the repetition, but a fundamental principle of learning theory is that multiple exposures to material help us learn.) Social scientists are interested in whether or not there are patterns or relationships among the social phenomena on which we collect information. A major conduit for such information comes in the form of surveys. Survey research translates our phenomena of interest into variables. A variable is anything that can vary, i.e., take different values, and it represents our way of measuring the concepts in which we are interested. To illustrate, for sociology, socioeconomic status, which we often refer to as social class, is one of our important concepts, and it can vary, right? People can be located higher or lower in the socioeconomic structure. In order to try to measure that in the real world, I have to find a way to put the concept into operation. For instance, maybe I could construct a survey in which I ask respondents to identify their social class location and I gave them choices of "upper class," "middle class," "working class," or "poor" (in methodological terms, I call this the "response set" for this question) from which to select. That's one way of measuring our concept—social class. Now, to look for patterns, I might examine the variation on just this one variable by looking at the distribution of cases [respondents] across the variable's response set (we call this doing "univariate analysis" through generating a "frequency distribution"). Doing that would tell me how many of my survey respondents classified themselves as

"middle class," "working class" and so on. While it doesn't take me real far, this type of analysis is still useful in its own right.

But since my hypothetical survey also asked about a lot of other stuff, I could expand my analysis to see what other kinds of things are associated with my respondent's self-reported social class status. Examining the relationship between two variables is important since social scientists often are trying to make sense of how variables are related to one another. To do so, we work with conceptual frameworks (theories) that lead us to expect certain relationships between our variables of interest. Here's an example continuing with social class. We know that social class is connected to many things including how parents might raise their children. One conceptual framework suggests that middle class parents, because they tend to work in less regimented environments might actually be less regimented in their child rearing practices. A concrete variable that represents such child rearing practices could be whether or not someone favors physical forms of punishment. Let's say that the actual measures for these two concepts are my earlier measure on social class and the degree to which a parent believes in spanking as a form of discipline (measured with a response set of "strongly agree," "agree," etc.).

One simple, yet useful, way to investigate relationships between variables like these is through cross-tabulation. A "cross-tab" is a table that presents the distribution (in frequencies and/or percents) of one variable across the categories of another variable(s) (e.g., what percentage of parents from of middle class background agree with spanking). Since it let's us look at two variables, we label this a bivariate analysis. Typically, in crosstabs and many other statistical techniques, we conceptualize the relationship between the two variables in terms of one influencing the other. The language we use to capture such relationships is to call one variable an independent variable (IV--it's doing the influencing) and the second variable a dependent variable (DV--it's the one that is being influenced). To run a crosstab you decide on two variables that you think might be related to one another. Typically, drawing from a conceptual framework, one next states a hypothesis for the relationship between your chosen variables. A hypothesis is an educated guess about what you think you will find. Hypotheses should always state a specific relationship and specify the comparison. Extending my example, the conceptual framework I noted above would lead to the following expectation: respondents from a lower social class background are more likely to favor physical punishment of children compared to those from higher social class locations. In this example, "social class" is the IV and "attitude towards spanking" is the DV. Then I'd go on and examine the data to see if my expectation is borne out.

To illustrate, consider the results below. These come from the General Social Survey and are analyzed by the SPSS program which I demonstrated in class. I've chosen educational level as a type of "proxy" for social class status. You may recall from that demo that the output looked different. I took what SPSS produced and

cleaned it up to make it more presentable to a lay audience.

Table 1: The Relationship Between Educational Level and Attitudes Toward Spanking for Discipline

	Educational Level	
Favor Spanking to Discipline A Child	High School Grad or Less	College Degree or Higher
Agree	76.1% (542)	66.3% (189)
Disagree	23.9% (170)	33.7% (96)
	100%	100%

Interpreting the results: StudentChip (or SPSS) does the number-crunching for you, but you still have to interpret what the tables are "saying." Remember what you're after: you want to see if the IV (education) has any association with the DV (favor spanking). To do that, you want to compare percentage differences between the categories of the IV (e.g., compare those with lower levels of education against those with higher). [Note: to do this properly, you want to make sure each category of the IV adds up to 100%. Analytically, you're asking what % of people with a high school education favor spanking versus what % of people with a college education? That's why, in building the tables in StudentChip, you'll see below that we request that the program "percent down" through the education variable which constitutes the columns.]

In this example, if you compare DV categories (those who "favor spanking" versus those who do not), you see that a higher percentage of people approve of spanking than don't. I'm not saying that's not interesting, but it doesn't address our conceptual question of whether or not education makes a difference in such attitudes. Besides, you could have already noted the differences in those who approve or disapprove spanking in a univariate analysis (frequency distribution).

Remember the three basic questions from which we frame our interpretation of crosstabs output:

1. Is there a relationship between the IV and DV? (You're looking to see if there is enough of a difference in percentages between the columns to matter. Some guidelines are found in the third bullet here. Your answer at this point is a simple yes or no.)

2. What is the relationship? (Here you want to put in sentence form what it is that the table is telling you. For the sample output: "Those with a lower level of education are more likely to favor spanking compared to those with higher levels." Then you could go on to flesh that out with some specific percentages or some such comparison. Pay attention to how you phrase things, because if you put it any differently you may end up saying something that the results don't show. Also, please be aware that I've given a simple example here. Many of our variables of interest have more than two values they take. If your IV has 3 or more categories, the logic is still the same. You want to just compare across the IV categories. If the categories on each end seem to represent extreme differences, then you can look at those.
3. How strong is the relationship? (For this one, you want to look at the % differential. Or, if the numbers are really small in absolute terms, you can calculate the proportional difference. In the above example, it's roughly 10%, which, in the social sciences, is a moderate relationship. Here's my rule of thumb for gauging strength: If it's 1- 3%, then I'm hesitant to say there's much of a relationship. If it's 4-7%, then it's a slight relationship. An 8-15% difference is a moderate relationship, while anything aproaching 20% is a substantial relationship. This is only meant as a rough guideline. These aren't technical rules.)

IV. ACCESSING DATA SETS

You will access 2 different data sets. One file contains data on a selected state from the 2000 census (state is specified below), while the second is a data file for one specific occupation—doctors—from the 1990 census. You can access these files and run StudentChip in a couple ways. First, you can use the program on our campus terminals in Dailey. If you do it this way, you will need to download the datasets you are using onto a floppy disk from Angel.

Instructions:

1. Log on to our course in Angel and click on the "Lessons" tab.
2. Under the "Lessons" tab in Angel you will find a folder called "StudentCHIP Data Sets" which contains the necessary data sets. You'll find one for each state and they are labeled by a suffix of their postal abbreviations. For your assignment, obtain the data set for Washington D.C. The other data set you need is "doctor9.dat."
3. To download the data, click on one of the datasets you need. Angel will take you to a new page stating: "The requested file can be accessed using the link below."
4. Right click on the file (it's listed at the top of the page with the extension .dat). Highlight the "Save Target As..." option.
5. In the "Save In" window, select "3 1/2 Floppy (A:)" or the appropriate drive

or folder where you want to save the dataset and click on the "Save" button.
6. You probably saved it on the A: drive so now you have it on your floppy. You are now able to get into StudentChip and call up the data to analyze.

The other way to access StudentChip is through a website. Use these instructions:

1. You can access WebCHIP through the following link: http://ssdan.net/datacounts/webchip
2. Next, select the dataset earn2kc.
3. This will bring up the data set in the CHIP program and it is ready for analysis. Guidelines for running the analysis are found further below.

Now, on to our assignment! I have given you this hard copy so you could look at it in detail in class. Also, you should use this as a worksheet. Please follow the instructions and complete the necessary information for the two tasks found below. Answer all questions associated with each task. The actual assignment you turn in will be done on the computer. Each of you will get a version of the assignment through Angel. You will call up the file in Word, type in your data into the charts and your answers to the questions on the computer. Be sure to put your name in the space provided and then use the drop-box facility found on Angel under "Lessons" tab. Look for the drop box named "StudentChip Assignment." The due date for this assignment is 3 PM on Monday, April 7th.

V. THE ASSIGNMENT NAME

Before doing your analyses, I present univariate and bivariate results from the national data set. This data set is a little different than the one we used in class. To get a listing of the variables and their frequencies, I go to the "Command" menu and click on "Marginals." This gives the following output:

RaceLat

NHwhite	Black	Hispanic	Asian	AmIndian	Total
73.6	11.7	9.9	4.0	0.7	= 100.0%

Gender

Male	Female	Total
58.7	41.3	= 100.0%

Earning

<15K	15-25K	25-35K	35-50K	50-75K	75K+	Total
12.3	22.3	20.9	20.4	14.7	9.4	= 100%

WkAge

16-24	25-34	35-44	45-54	55-64	65+	Total
7.8	24.4	30.2	25.0	10.7	1.9	= 100%

These results are frequency distributions for each of the 4 variables in this CPS sample of full-time, year-round workers in 2000. According to the table, 58.7% of all full-time workers are male as opposed to 41.3% female. 12.3% of full-time workers earn less than $15,000 a year while 9.4% earn more than $75,000. 73.6% of full- time workers are non-Hispanic white, 11.7% are black, 9.9% are Hispanic, 4.0% are Asian, and only 0.7% are American Indian. Workers between the ages of 35-54 account for 55.2% of all full-time workers. Only 1.9% of full-time workers are over the age of 65.

Next, producing a crosstabulation of gender and earnings gives us the following nation-wide results:

	Male	Female	All
75K+	13.3%	3.9%	9.4%
50-75K	18.6%	9.3%	14.7%
35-50K	21.9%	18.3%	20.4%
25-35K	19.1%	23.4%	20.9%
15-25K	17.9%	28.5%	22.3%
<15K	9.2%	16.7%	12.3%
Total	100%	100%	100%
	(N=56,212,417)	(N=39,610,636)	(N=95,823,053)

As I mentioned above, these results are a bit different than those presented in class because this is actually from a slightly different data set. Overall though, the pattern is the same: men are much more likely to be represented in the higher income brackets. We see that 31.9% of male full-time workers make $50,000 or more a year as opposed to only 13.2% of females. You will use these results to compare with data from a selected state.

TASK #1: *Produce a crosstabulation of gender and earnings for Washington D.C.*

Directions:

1. If you are using the Chip program which is on our servers, you need to have first downloaded the dataset from Angel (instructions were given above) to a floppy. From the "Start" button go to Programs>StudentChip. Click on File>Open. Here, you'll probably have to change the path to find your file since from the earlier instructions, you probably saved it to a floppy disk. In the "Look In" window change the path to a:/ and click on the state data file you've been assigned. With either StudentChip on our computers or Webchip, once you've accessed the dataset, you'll get a one- or two-line description of what's there. For this data set, it will say "2000 Full-time, Year-round civilian workers in [particular state]." It will also give you the total number of individuals these data are drawn from (N=.....).
2. Next, produce a crosstabulation of earnings by gender. Click on Command>Crosstab. Pick Earnings as your first variable, then click select. Next, pick Gender and click select. Next we have to tell Chip how to percentage the table. We're interested in the percentage of people of a particular gender who fall into the various earnings bracket so choose Table>Percent down. Your table should now appear. With WebChip, use Earnings as the row variable and Gender as the column variable and also select % Down to help construct the table. After those choices, click on "CROSSTAB" and you will have your table.
3. Based on your results, neatly fill in the table below. Include the Ns below plus the name of the state in the title.

Table 1: Gender Differences in Earnings among Full-time Workers, Aged 25-64, from Census 2000 for the State of

_____

	Male	Female	All
75K+			
50-75K			
35-50K			
25-35K			
15-25K			
<15K			
Total	100%	100%	100%
	(N=)	(N=)	(N=)

Questions:

1. Do male or female full-time workers typically earn more money in your selected state? Summarize the results shown in Table 1 (keep in mind guidelines I mentioned in class).
2. How do these results compare with national data? Is the overall trend the same or different? Is the gap larger or smaller for this state compared to the national data?
3. Why is there a gap? Draw from the text to briefly summarize some possible explanations for the existing gender gap. While your discussion must draw from the text, you may also include other things you've thought of.

Task #2: *Assess some of the proposed explanations using other data.*

Directions:

1. To answer Question #2, you gave some informed speculation on potential factors playing a role in the earnings gap. From the table above, however, we cannot support any of the explanations with any certainty. We need to "zoom in" on some more specific information, like we did in class, to help assess some of these explanations. To help us do that, I want you to produce another crosstab. Retrieve the data set—doctors9.dat. This is found on Angel

under the "Lessons" tab in the StudentChip data files folder. Or, if you're using WebChip, go to the website mentioned in the beginning of this handout and click on the "WebCHIP" launcher link. Using the "Menu-Enabled Launcher" highlight "cen1990" and click "Select Target."

2. Next highlight "doctors9.dat" and then click "Select Target." This will open the data set and you're ready for generating the table. Depending on which version of StudentChip you're using, follow the appropriate instructions and produce a table of earnings by gender again (caution: be sure you percentage the table the correct way for this analysis). Fill in the appropriate information below.

Table 2: Gender Differences in Earnings among Full-time Doctors, Aged 25-64, from Census 1990

	Male	Female	All
150K+			
125-150K			
100-125K			
85-100K			
70-85K			
55-70K			
40-55K			
<40K			

Questions:

1. Do male or female full-time doctors typically earn more money? Summarize the results as with the other crosstabs.
2. While we cannot say with absolute certainty, briefly summarize one or two of the explanations mentioned in the text and/or class that become more plausible given these results. Include your reasoning. How about one or two that become less plausible given these results? Why?