Bouma, Soc 110: Social Problems

Handout #2- Data Analysis
Investigating the Effect of Race and Gender on Earnings in the US[1]

Now that we have walked through the national level data in class, you are going to replicate these analyses yourself and then examine data for Kentucky and another state of your choosing. To do so, you are asked to work in pairs.

Please choose the state you would like to examine: ________________________________

GETTING STARTED:

1) To get started, go to the website:
http://ssdan.net/datacounts/webchip/

2) On the left, you will see the title Choose Dataset. Underneath, in the upper left box that says **Collection**, hit the down arrow and scroll down to choose the dataset "geo2016." In the box below that says **Dataset**, click on the down arrow and you will see a list of datasets, each containing **state level data sets**, with the suffix indicating the state. For example, Earn_KY contains the data for Kentucky. Click on the data set you want to analyze. All of you will be analyzing KY data, and then pairs of you will each examine a different state. Go ahead and explore the state you have chosen. (Please note, to find **national data**, choose Earn_States.)

A) FREQUENCIES: A frequency table gives an overall sense of the distribution of a particular variable or set of variables. **To examine the frequencies of each variable:**

1) **Click on "Compute Marginals" and scroll down to see the output.**

You will see a table for each variable. For example, for the racial composition of the fulltime, year-round workforce in **Kentucky**, you will see the following table:

Racial Composition of Fulltime Year-Round Workforce in Kentucky, 2016

NHWhite	Black	AmIndian	Asian	NHMulti	NHOther	Hispanic
86.7%	**7.39%**	**0.18%**	**1.42%**	**0.97%**	**0.14%**	**3.2%**
1,208,400	**103,028**	**2,506**	**19,755**	**13,506**	**1,897**	**44,604**

Note that the first row of data gives percentages, and the second column gives the total number. Thus, in Kentucky, 1,208,400 fulltime workers are white, which represents 87% of fulltime workers. Generally, for comparison purposes, we talk about percentages, rather than raw numbers.

B) After examining the frequency tables for age, gender and income, answer the following questions for **your STATE:**

- What percentage of all full-time workers are Black? ________ Hispanic? _________ Asian? ____
- What percentage of all full-time workers make less than $25,000? ____________
  Less than $35,000: _____________ More than $69,000? _______
- What percentage of all full-time workers are 16-24? _________; older than 65? _________

[1] This exercise is adapted from a module originally developed by Tim Thornton, SUNY-Brockport.

3) Now describe these frequencies in easily understood English. In other words, how would you describe all full-time workers for the state you chose in the year 2016?

4) Once again, please make hypotheses about what you expect to find about differences in earnings in your chose state:

HYPOTHESES FOR State ____________

1) SEX: _________ will have higher incomes than __________.
2) RACE ___________ will have the highest incomes and _________ will have the lowest.
3) AGE: People in the age group ________ will earn the most money, and people in the ________ age group will earn the least.

5) **CROSS-TABS** (Bivariate Analysis): To run cross-tabs, you will need to tell the program which variable is your dependent variable, which is your independent, and how to percentage your tables (across or down). Remember the following:
INDEPENDENT VARIABLE (X) - the variable that influences or affects another variable
DEPENDENT VARIABLE (Y) the variable that is influenced by, or depends upon, another variable
Do the following:

A) Scroll down the side menu to the section labeled "Choose Variables" on the left.
B) Click the down arrow next to the box labeled "Row" and select a variable. For the purposes of this assignment, always **put your dependent variable in the rows**. So, for this example, choose earnings for your row variable by highlighting "Earn3".
C) Now you will need to choose your **column variable**. For this assignment, always **put your independent variables in the columns**. So, choose sex as your independent variable by highlighting "Sex" as the column variable.
D) Now scroll down to the section labeled "Generate Table" and click **"percentage down."**
E) This should bring up your bivariate, cross-tab table.
F) To look at the effect of race on earnings, follow the steps above, this time choosing "RaceEth" as the column variable

Making a chart:
Scroll down further on the left, and you'll see "Generate Chart" with four options. Play around with the four options to find what makes most sense for graphically depicting your data. (Hint, not all of these charts will make sense.)
Once you have examined the frequencies and cross-tabs and graphs for your state, go back and examine the data for Kentucky, and repeat the steps above. You will also need to go back to the data set for the nation as a whole and repeat the steps above.

Overall, you need the following information for (a) the nation as a whole (in geo2016 Earn_States) (b) Kentucky (in geo2016, Earn_KY) and (c) a state of your own choosing:
1) frequencies (marginals) of age, gender, race, and income distributions
2) the cross-tab of earnings by sex
3) the cross-tab of earnings by race

After you have analyzed these data, please write a 5-8 page essay that compares your state to either the U.S. as a whole or to Kentucky. The essay is described below.

DATA ANALYSIS PAPER

For this assignment, you are asked to write a paper which analyzes the effects of sex and race on income. You are asked to do this for a state of your own choosing and to compare the findings to the US as a whole or to Kentucky (or another state). Your paper should follow the outline of a general research paper:

Introduction: You should start with a general introduction that lays out the main questions you will explore. Please do NOT start by saying "This paper is a data analysis about….".

Brief Literature review: You should refer briefly to studies, helping us understand the issue. What have you learned about race and sex differences in income from readings for this class and other sources? You can refer to studies referenced in our book or other sources. (Wright and Rogers are especially good sources for this.) Make sure that you properly cite your sources – ask me if you do not know how to do this properly. I know that you are not an expert on these issues, so I fully expect that you will summarizing and paraphrasing (and therefore citing) others who are.

Sample: For this section, briefly describe who is in your sample – describe the full-time working populations of your state, as well as either Kentucky or the U.S. To do this, describe the MARGINAL FREQUENCIES (not the cross-tabs). What is the racial composition, gender composition, and earnings distribution for your chosen state and either KY or the U.S. (You do not need to do the age distribution). You'll need to include the frequency tables/marginal tables for this section. Your description of the racial, sex and earnings distribution can be relatively brief. Please **do not** describe every single statistic (you'll put me to sleep). Instead, try to succinctly summarize the information that you think is most important. (Note, this was question #3 on page 2.) This will take you a few tries. Try reading it aloud to a friend and having them tell you what they think you mean. The point of this part of the exercise is to become comfortable using and discussing **basic** statistics. This part should be comprehensible and easily understood. Do not just list statistics; try to keep the reader engaged.

Results: This is the heart of your paper, where you are describing your bivariate tables (cross-tabs). Start by giving your hypotheses: which groups did you expect to make the most and least? Then you need to describe how earnings differ by race and sex, both in your state and either in KY or the US. Do men or women make more money? How does your state compare to the nation as a whole or to Kentucky? Where do you see the greatest evidence of a wage gap in earnings? You also need to describe racial differences in earnings. Which racial groups have the highest and lowest earnings? Again, report specific findings both for your state and either Kentucky or the nation. Where do you see the largest wage gap in terms of race? Make sure you include the appropriate tables with your paper.

Tables: Tables should always be numbered, titled, and clearly labeled (and please number your tables in the order in which you refer to them). You should have the marginal/frequency tables for the Sample Section and cross-tab tables for the Results Section. **Each table should have the totals at the bottom**. I usually put 100% = [total #]. In order to get **the totals** for your cross-tabs tables, first follow the directions for generating a cross-tab; then after you have clicked on "percent down" to get the table with percentages, click on "frequency" to get the actual numbers. You'll find the total number there. You also need to have the source at the bottom of your tables. An example of a properly formatted table follows:

Table 1: 2016 Earnings by Sex for U.S. Full-time Year-Round Workers, PUMS

Earnings	Male	Female	Total
<25K	21.1%	26.2%	23.2%
25-34K	13.8%	18.3%	15.7%
35-49K	18.2%	21.1%	19.4%
50-69K	17.7%	16.6%	17.2%
70-99K	13.7%	10.4%	12.3%
100K+	15.5%	7.4%	12%
Total	100%=	100%=	100%
	60,913,292	45,349,937	106,263,229

source: Earn.dat: 2012-16, wgtd PUMS, Frey/U-Mich for SSDAN

Conclusion: Sum up your main findings, and discuss reasons you think we find these differences. Why do we find the patterns we see? One reason is probably discrimination, but there are many, many other reasons for these differences. You should NOT conclude that the differences you see are simply due to discrimination (though this is indeed a part) – there are many other forces at work. Again, consult our text and other sources.

Bibliography: Make sure to include the full citation for any of the literature you refer to in you brief literature review. To cite the data sets, you can cite each state using the following:

Frey, William H. "2012-16 Full-time, year-round workers, age 16+, [*name of state*], wgtd PUMS, Frey/U-Mich for SSDAN. 2016. < http://ssdan.net/datacounts/webchip> [date you got your information from the website].

If you decide to compare your state to the national statistics, then cite the national stats as follows:

Frey, William H. "2012-16, Earnings by race, and sex, age 16+, wgtd PUMS, Frey/U-Mich for SSDAN < http://ssdan.net/datacounts/webchip > [date you got your information from the website].

Context Statement: This project was designed to introduce you to and give you practice with the following skills: read and report basic frequencies in a table, interpret numbers in a bivariate table, convert raw data into formatted tables, identify independent and dependent variables, analyze the relationship between two variables, tell a story using numbers. For your paper, please also include a "Context Statement." In a couple paragraphs, please describe your experience writing a paper based on data analysis. What did you learn from this experience? What did you find most challenging? Has this helped you feel more confident about working with numbers? In what ways would you like more help or practice? You may comment directly on the skills listed above or discuss any other issues relevant to this assignment.

FORMAT:

- Make sure to **double-space your paper**.
- Use a heading for each section (Introduction, Literature Review, Sample, etc.).
- Put tables in the order in which you refer to them in the paper.
- Number your pages.
- Do not use "you," and do not split tables across pages. If you choose, you may write this paper with a partner.

DUE DATES:
Monday, 2/26
1) Finish handout 2
2) **Type up and turn in: A) draft section of literature review and**
B) one table and a description of this table
3) Be ready to present your findings to the class:

- Sex: My hypothesis was that ____ will earn more than ____ .
- Race: My hypothesis was that ___ would earn the most and ___ would earn the least.

I found:

- 1) start with general statement
- 2) use data to support your general statement
- 3) offer brief conclusion

4) start drafting your paper, to identify where you need help (and practice presenting in class);

Wed, 2/28: Bring early draft to work on in class.
Fri, 3/2: Peer Review: Bring **typed draft** to class – this counts for 10% of the grade of the paper. **Bring a printed copy** so that a classmate can work directly on your paper (I will check all of these in class.);
Wed, 3/14: Final Draft due (I will not accept emailed copies).

Rubric for Data Analysis Papers

	Grade	Strengths – what to focus on	Areas to improve:/typical problems
Peer Review 10			
Intro 5		Clear, lays out main points	Needs to be broader
Literature review 10		Relevant to paper, discusses gender and race income inequality; brings in outside sources; properly cites sources	Info not relevant Didn't bring in sources Didn't cite properly (if improper citing or plagiarism, **paper fails**)
Sample (describe freqs) 15		Clearly describes frequencies (marginals) of own state and other state or U.S.	Didn't adequately describe frequencies
Results Section:			
Income by gender 15		Uses language properly Tells a story	
Income by race 15		Uses language properly Tells a story	
Tables 15		Properly numbered, labeled, titled have totals and source at bottom, etc.	
Conclusion 5 Sums up findings, refers back to relevant lit, offers reasons why we see patterns			
Works Cited 5			If improper citing/plagiarism – **paper fails**
Context Statement 5			

Soc 110 – reading and writing about statistics: TELL A STORY WITH NUMBERS.

Rules:

- start with general statement without numbers (e.g. “Men make more than women.”)
- back up claim with stats – and follow same direction of statement. For example, if you say “Men make more than women” then show how more men fall into the high-income categories, not how more women fall into the low-income categories. If you say “Women make less than men,” then show how more women fall into the low-income categories.)
- when using stats, always say ___% of ___ (e.g. “10% of Hispanics”). Do NOT say “Hispanics are 10%.”
- when showing similarities across groups, use ‘and”
- when showing differences, use “compared to” “while only” etc.
- have a summary statement at the end.
- Write as if you care about what you found – these numbers mean something!

Useful phrases:

- A higher proportion of … fall into the (lower income/higher income) category
- a disproportionately high/low number of … fall into the …
- … are overrepresented/underrepresented among low-income/high-income groups…

Bouma’s attempt at describing race differences in income for the US from earlier data: (ACS 2008)

When examining race differences in earnings in 2008 for the U.S., we see that Non-Hispanic Whites and Asians have the highest earnings, and African Americans, Native Americans and Hispanics have the lowest. For example, we see that more than one-quarter of all Asians and one-fifth of whites earn above $70,000. This compares to only about one in ten African Americans or Native Americans, and less than one in twelve Hispanics earning this much. When we examine the low-income categories, we now see that Blacks, Native Americans, and especially Hispanics are over-represented. Over 30% of both African Americans and Native Americans earn less than $25,000 every year, and a full 43% of Hispanics earn this little. This means that about two out of every five Hispanics earned less than $25,000 in 2008. This compares to just 19% of Whites and 21% of Asians. Overall, then, we see that Asians and Whites fall disproportionately into the high-income categories, and Blacks, Native Americans, and especially Hispanics are fall disproportionately into the low-income categories.

2008 Earnings by Race for U.S. Full-time Civilian Workers, ACS

	NHWhite	Black	Asian	Hispanic	AmIndian	NHOther	NHMulti	TOTAL
<15K	5.5%	9.5%	6.2%	13.5%	11.6%	10.1%	7.5%	7.1%
15-24K	13.5%	21.8%	14.8%	29.4%	24.2%	20.9%	17.3%	16.8%
25-34K	17.5%	22.6%	15.2%	21.0%	21.5%	21.0%	20.0%	18.4%
35-49K	21.8%	22.1%	18.4%	17.7%	20.1%	18.8%	21.9%	21.1%
50-69K	18.5%	13.9%	16.9%	10.3%	12.6%	13.4%	16.4%	16.7%
70-99K	12.1%	6.8%	14.7%	5.0%	6.3%	9.2%	9.7%	10.6%
100K+	11.2%	3.3%	13.8%	3.1%	3.6%	6.6%	7.1%	9.3%
TOTAL	100% 66,678,276	100% 10,610,592	100% 4,694,340	100% 13,309,425	100% 611,753	100% 216,348	100% 962,917	97,083,651

Source: wgtd 2006-08 ACS, SSDAN/U-Michigan