Bouma, SOC 110

***Handout #1,* Data Analysis Project**

**Investigating the Effect of Race and Gender on Earnings in the US[[1]](#footnote-1)**

For the past couple weeks, we have been discussing inequality in America. Often, we think of power and inequality in terms of income. Who has more money, and thus, in general, more access to opportunities and resources? Who makes the most money, and who makes the least? Does income differ for men and women, and for whites and people of color? In this exercise, we will examine earnings data for all full-time workers in the US. The data come from the 2021 American Community Survey (ACS) data collected by the Census Bureau. You will be able to examine data for the nation as a whole, for Kentucky, and for a state of your choosing.

**LEARNING GOALS**:

A*. Substantive: Students will be able to:*

1. Discuss and write about the influence of race and gender on earnings, by examining 2021 ACS data. You will be able to use data to explain whether men make more than women and whether whites earn more than people of color.

2. Use data to make national, regional and state comparisons in terms of earnings. You will be able to compare Kentucky to the nation as a whole, in terms of income inequality, as well as comparing Kentucky to another state of your choosing. The data shown in the tables for this worksheet come from

[**http://ssdan.net/webchip/webchip4/**](http://ssdan.net/webchip/webchip4/) which access the acs21\_yr data collection and then the data set Earn.

B. *Methodological/Quantitative Skills: Students will acquire the following skills, related to data analysis*:

1. Interpretation: Students will be able to:

* Read and report basic frequencies from a large data set
* Describe bivariate tables, both orally and in writing

2. Representation: Students will be able to:

* Take raw data in tabular form and create properly formatted and labeled tables

3. Application/Analysis: Students will be able to:

Manipulate variables in a large data set using a basic statistical package

* Identify independent and dependent variables
* Form hypotheses about the relationship between two variables
* Analyze relationship between two variables as presented in bivariate table

4. Communication: Students will be able to:

* Present hypotheses and findings for relationship between two variables in informal presentation and in a paper that tells a story about the effects of race and gender on income, using numbers as evidence

5. Confidence: Students will feel more comfortable

* reading and discussing data from a table
* writing about and using percentages to make an argument

**VARIABLES:** For this exercise, we will be examining the following variables: earnings, race, sex, and age. Although there are many ways in which each could be conceptualized, the following ways are those used by the U.S. Census Bureau and the Bureau of Labor Statistics:

***EARNINGS***: Money a person makes from working, as wages, salary, or a form of self-employment, expressed as an annual amount

***SEX*** - individual’s self-identification as either male or female.

# AGE: - age divided into the following categories: 16-24, 25-34, 35-44, 45-54, 55-64, 65+

***RACE (RaceLat***) – individual’s self-identification as:

* **Non-Hispanic White (NHwhite)** – all persons who indicated their race as white and not of Hispanic origin.
* **Black** – all persons who indicated their race as black.
* **Asian (or Pacific Islander)** – includes all persons who indicated their race or ethnicity as Chinese, Filipino, Japanese, Asian Indian, Korean, Vietnamese, Cambodian, Hmong, Laotian, Thai, or other Asian as well as Hawaiian, Samoan, Guamanian or other Pacific Islander.
* **Hispanic**– persons of white or “other” races who identified themselves as Mexican, Puerto Rican, Cuban, or Other Spanish/Hispanic. This category can refer to ancestry, nationality group, lineage, or country of birth of the person’s parents or ancestors before their arrival in the U.S.
* **American Indian (AmIndian)** – all persons who classified themselves as American Indian, Eskimo or Aleut.
* **Non-Hispanic Multiracial (NHMulti)** – any Non-Hispanic persons who identified as more than 1 race
* **Non**-**Hispanic Other (NHOther)** – any Non-Hispanic persons of a single race that was not white, black, American Indian/Alaskan Native (AIAN), Asian or Pacific Islander

## FREQUENCIES: A frequency table gives an overall sense of the distribution of a particular variable or set of variables. Here are the frequencies for the variable RaceEth (race) for the PUMS sample of full-time, year-round workers in 2021.

## RaceEth:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **NHWhite** | **Black** | **AmIndian** | **Asian** | **NHMulti** | **NHOther** | **Hispanic** |
| 61.49% | 11.42% | 0.46% | 6.24% | 2.38% | 0.48% | 17.53% |
| 69,845,203 | 12,968,881 | 527,236 | 7,087,716 | 2,706,784 | 550,683 | 19,907,556 |

Note that the first row of data gives percentages, and the second row gives the total number. Thus, in the U.S., 69,845,203 fulltime, year-round workers in 2021 were Non-Hispanic White, which represents 61% of fulltime workers. For comparison purposes, we talk about percentages, rather than raw numbers. We see that about two-thirds of US full-time workers in 2021 were White, about 11% of all workers were Black, about 6% Asian, about 18% Hispanic, less than 1% were Native American, and very few identified as Non-Hispanic Other or Non-Hispanic Multi-racial. Now examine the frequencies for age, sex, and earnings below:

**Sex:**

|  |  |
| --- | --- |
| Male | Female |
| 56.83% | 43.17% |
| 64,554,439 | 49,039,620 |

Earnings

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **<25K** | **25-34K** | **35-49K** | **50-69K** | **70-99K** | **100K+** |
| 17.81% | 14.14% | 18.93% | 18.48% | 14.47% | 16.17% |
| 20,228,579 | 16,056,884 | 21,506,646 | 20,997,854 | 16,435,776 | 18,368,320 |

 Age

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **16-24** | **25-34** | **35-44** | **45-54** | **55-64** | **65+** |
| 6.79% | 23.73% | 23.66% | 23.05% | 18.39% | 4.38% |
| 7,713,060 | 26,957,372 | 26,875,640 | 26,189,029 | 20,886,686 | 4,972,272 |

Answer the following:

* What percentage of all full-time workers are men? \_\_\_\_\_\_\_\_\_\_
* What percentage of all full-time workers make less than $25,000? \_\_\_\_\_\_\_\_\_\_\_\_

Less than $35,000: \_\_\_\_\_\_\_\_\_\_\_\_\_ $50,000 or more? \_\_\_\_\_\_\_

* What percentage of all full-time workers are 16-24? \_\_\_\_\_\_\_\_\_ older than 65? \_\_\_\_\_\_\_\_\_

Now describe these frequencies in easily understood English. In other words, how would you describe all full-time workers in the year 2021?

Now you get to be the sociologists. You can see that about 18% of all full-time workers make less than $25,000. Who are they? About 31% of all full-time workers make more than $70,000. Who are they? Are they men or women? Are they White, African American, Latinx or Asian workers? Which age group makes the most? The least? Begin by making hypotheses about what you expect to find:

HYPOTHESES:

1. SEX: \_\_\_\_\_\_\_\_\_\_ will have higher incomes than \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
2. RACE/ETHNICITY: \_\_\_\_\_\_\_\_\_\_\_\_\_ 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.

In order to investigate your hypotheses, you will need to do cross-tabulations, also called bivariate tables. In a cross-tabulation, you are simply exploring the association between two variables. Before running these, think about what you really want to know. It makes sense to say that a person’s sex affects his or her earnings. It does not make sense to say that earnings affect a person’s sex.

The following two definitions are important:

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

You can write the relationship between the two as X 🡪 Y . In this case, we are interested in how sex influences earnings. Sex would be the independent variable (X), and earnings would be the dependent variable (Y). In other words, earnings to some extent, depend on sex:

 Independent 🡪 Dependent

 Variable variable

 X 🡪 Y

 Sex 🡪 Earnings

Below is a bivariate table (cross-tabulation) of sex and earnings:

# Table 1: 2021 Earnings by Sex for U.S. Full-time Year-Round Workers, ACS

|  |  |  |  |
| --- | --- | --- | --- |
|  **Earnings** | **Male** | **Female** | **Total** |
| **<25K** | 16.2% | 20% | 17.8% |
| **25-34K** | 12.3% | 16.6% | 14.1% |
| **35-49K** | 17.5% | 20.8% | 18.9% |
| **50-69K** | 18.5% | 18.5% | 18.5% |
| **70-99K** | 15.5% | 13.1% | 14.5% |
| **100K+** | 20% | 11.1% | 16.2% |
| **Total** | 100% | 100% | 100% |
|  |

|  |  |  |
| --- | --- | --- |
| 64,554,439 |  |  |

 | 49,039,620 | 113,594,059 |

 source: Earn.dat wgtd ACS21, Frey/U-Mich for SSDAN

Do incomes differ for men and women in the US, and if so, how? Below, write a description of this table:

Here’s my description: From Table 1, we see that female full-time workers typically make less than male full-time workers. For example, 36.6% of women make $34,000 or less, compared to only 28.5% of men. By contrast, 35.5% of full-time male workers make $70,000 or more, compared to only 24.2% of women.

Note my description: I started with a broad, generalized statement: “female full-time workers typically make less than male full-time workers.” Then I used specific statistics from the table to make my case. Note that since I **percentaged down the rows**, and I **compared across the columns**. In other words, I compared percentages **across men and women**. Now let’s examine the influence of race on earnings:

# Table 2: 2021 Earnings by Race for U.S. Full-time Year-Round Workers, ACS

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Earnings | **NHWhite** | **Black** | **AmIndian** | **Asian** | **NHMulti** | **NHOther** | **Hispanic** | **Total** |
| **<25K** | 14.7% | 22.7% | 24.5% | 14.6% | 17.6% | 19.5% | 26.4% | 17.8% |
| **25-34K** | 12% | 19% | 19% | 10.1% | 14.2% | 15.4% | 19.9% | 14.1% |
| **35-49K** | 18.3% | 21.8% | 21.8% | 13.7% | 18.7% | 19.8% | 21.1% | 18.9% |
| **50-69K** | 19.9% | 17.2% | 17.2% | 15.3% | 18.2% | 18% | 15.6% | 18.5% |
| **70-99K** | 16.2% | 10.3% | 10.3% | 16.9% | 14.9% | 13.2% | 9.6% | 14.5% |
| **100K+** | 19% | 7.2% | 7.2% | 29.4% | 16.4% | 14% | 7.3% | 16.2% |
| **Total** | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| **Total** | 69,845,203 | 12,968,881 | 527,236 | 7,087,716 | 2,706,784 | 550,683 | 19,907,556 | 113,594,059 |

 Source: Source: Earn.dat: wgtd ACS21, Frey/U-Mich for SSDAN

***For Wednesday, write a description of the above table, following suggested rules above. This will likely take you several tries***.

1. This exercise is based on a module originally developed by Tim Thornton, SUNY-Brockport. [↑](#footnote-ref-1)