Biostatistics 203A
Mid-Quarter Project
Due: November 15, 2019

The mid-quarter project will draw upon the skills you have learned so far in Biostatistics 203A, including reading in, manipulating, and summarizing data. You will be analyzing data from the Youth Development Study (https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/24881). Our interest will be primary in employment-related characteristics and concerns before and after the Great Recession of 2008. Thus, we will be working with data collected during the 2007 wave and the 2011 wave. This data will be contained within two text SAS datasets:

- **W2007.sas7bdat**
- **W2011.sas7bdat**

You will need to download the two files listed above from the course website and will also need to download a third data set called **FAMIDSubset.csv** which consists of two variables: SubsetNumber and FAMID. You will be given a SubsetNumber (please see the SubsetNumber.csv file) and you must subset the **FAMIDSubset.csv** file to include only those FAMIDs that correspond to your SubsetNumber. This subset of FAMIDs will correspond to the set of respondents that you will be analyzing throughout the entire project (each student will analyze a slightly different subset of all the respondents provided in the W2007 and W2011 datasets). Thus, it is a good idea to complete the following two steps first, before continuing on with the project:

(1) Read the **W2007.sas7bdat** and **W2011.sas7bdat** datasets into SAS and/or R (if using R, remember to use the read.sas7bdat() function we learned about in class).

(2) Subset the **W2007** and **W2011** datasets such that each dataset includes only FAMIDs within the **FAMIDSubset.csv** file that correspond to your SubsetNumber.

For this project you will need to rely heavily on the following supporting materials which were generated using SAS and can be found on the course website:

- YDS_2007_Contents.pdf
- YDS_2011_Contents.pdf
- YDS_Formats.pdf

To complete this project, you can choose to use SAS or R and you will not be asked to submit your code or output. Instead, you will be asked to submit a word document containing a description of the methods you used and a summary of the results (approximately 1.5-2 pages), followed by a set of approximately 4 tables and 1 figure containing the requested results (as many pages as needed).

Here is an example outline:

Methods
[text]
Results
[text]
Table 1. Demographic and Employment Characteristics of the Sample in 2007
Table 2. Mental and Physical Health by Employment Characteristics for the 2007 Sample
Table 3. Comparing Employment Characteristics between 2007 and 2011
Table 4. Employment-Related Concerns about the Future in 2007 and 2011
Figure 1. Bar Chart Depicting Average Number of Hours Worked by Day of Week in 2007 and 2011

Do not feel that you need to adhere to the above outline, it is simply provided as a guideline. You are required to ensure that your project contain the results described in detail below. The project should be prepared as though being disseminated to a non-statistical audience of public health professionals. You will be graded on the presentation of your report, including visual appeal and clarity of results presented. Since you will not be conducting statistical tests, the text summary of results should simply highlight key findings. It is acceptable to include statements such as "average income was higher among individuals with a college degree, relative to individuals with less education" despite that fact that no statistical tests were performed. You do not need to state every number that appears in the tables in the text, but select a handful to describe and interpret. The methods section should be relatively brief, non-technical, and should include descriptions of any new variables or subsets of data you created in completing this project.

Projects are to be emailed to haralis@mednet.ucla.edu in an editable word document format no later than 5:00 PM on Friday, November 15th.

Results that should be included in the submitted project:

Demographic and Employment Characteristics of the Sample in 2007

Calculate and tabulate frequencies and percentages for each of the following variables found in your 2007 dataset. You will need to use the information found in the files listed above to find the correct variable names in the data. Whenever missing values are present, the frequency and percentage of responses they represent should be presented and missing values should always be included in the denominator. When sub-bullets are listed below, you will be expected to present frequencies for each of the sub-bullets (this may involve some combining of categories). If sub-bullets are not listed, please include frequencies and percentages for all categories.

- Gender (Use the Corrected 7/25/94 variable)
- Highest Level of Education
    - High school or less (include High school or GED and Elementary or junior high)
    - Technical or vocational
    - Some college (include Some College and Associate degree)
    - Bachelor's degree
    - Graduate degree (include Master's degree and PhD or professional)
- Currently employed (either part-time or full-time)
- Currently married or cohabitating in an intimate relationship
- Do you have any children?

In calculating frequencies and percentages for the two variables below, include only those individuals who indicated being currently employed (either part-time or full-time). This will impact the number of missing values.

- How secure is your primary job?
- How satisfied with your job as a whole?
    - Extremely or very dissatisfied
    - Somewhat dissatisfied
    - Somewhat satisfied
    - Extremely or very satisfied

Tabulate means and standard deviations for the following variables. You should note somewhere on your table what number of records were used to calculate each of these means (use all non-missing values for each variable).

- Age in years (calculate using the difference between date of birth and October 1, 2007. To calculate date of birth, assume each participant was born on the first day of their birth month)
- Annual household income in dollars (Use variable E5HI17)

Mental and Physical Health by Employment Characteristics for the 2007 Sample

Once again, using variables from the 2007 wave, for each individual create two new variables to represent mental health and physical health:

1. *Body Mass Index (BMI)*
   BMI can be calculated by dividing weight in pounds by height in inches squared and multiplying the result by 703. Weight and height can both be found within the 2007 dataset.

2. *Mental Health Total Score*
   A total score can be calculated using items H13A17 thru H13O17. First, reverse items H13A17, H13D17, H13F17, H13I17, H13N17, and H13O17 so that all 15 items take values ranging from 1 to 5 with higher values indicating worse mental health symptoms. Second, sum across the reversed items and the other 9 items to obtain the total score.

For each of the two variables above, tabulate within-group means, medians, and standard deviations letting groups be defined by each of the following categorical variables. You should note somewhere on your table what number of records were used to calculate each of these descriptive statistics.

- Highest Level of Education
- Currently employed (either part-time or full-time)
- How secure is your primary job?
- How satisfied with your job as a whole?

For each of the above categorical variables, use the same categorizations as used in the previous section. Also, be sure to only calculate descriptive statistics among individuals who indicated being currently employed (either part-time or full-time) for the last two variables listed above.

Comparing Employment Characteristics between 2007 and 2011

To produce this table, you will first need to combine the 2007 and the 2011 datasets and retain only individuals who are present in BOTH datasets (individuals are uniquely identified using the FAMID variable). Somewhere in your description of the methods, note the following two numbers:

- How many individuals completed a survey in 2007 but not 2011?
- How many individuals completed a survey in 2011 but not 2007?

After having successfully combined the two datasets, use the resulting dataset to tabulate the mean and standard deviation in 2007 and the mean and standard deviation in 2011 for the following variables. For each variable, only include responses from individuals if the individual had a non-missing response in BOTH years. You should note somewhere on your table what number of records were used to calculate each of these means.

- How much stress have you felt in meeting financial obligations?
- How difficult is it for you to pay your bills on time?

Additionally, include somewhere in your table the frequency and percentage of individuals who were "Currently employed (either part-time or full-time)" in 2007 and in 2011. Once again, include only those

individuals who provided a non-missing response in both years in the denominator for both percentages.

Lastly, we would like to include information about household income in 2007 and 2011 and how incomes compared to poverty thresholds published by the US Census Bureau. For each individual in our dataset, we will convert his/her household income into new units:

$$Income\ as\ a\ Ratio\ of\ Poverty\ Threshold = \frac{Household\ Income}{Household{-}and\ Year{-}Specific\ Poverty\ Threshold}$$

The household- and year-specific poverty thresholds will be a function of year and size of household. We use thresholds for the 2006 and 2010 years for our 2007 and 2011 years, respectively, because it is assumed that survey respondents provided their previous year's income in response to the income question. You will use the following variables (variable names are in parentheses):

- 2007: Household income (E5HI17) and household size (F1417)
- 2011: Household income (INCYRH19) and household size (HOME19)

You should use the variables above, along with the poverty threshold table below to calculate "Income as a Ratio of Poverty Threshold" for each individual in 2007 and 2011. For this new variable, include the mean and standard deviation in 2007 and the mean and standard deviation in 2011. Once again, only include responses from individuals if the individual had a non-missing response in both years and note the number of records that were used to calculate each of these means.

	Poverty Thresholds in US Dollars	
Household Size	**2006**	**2010**
1	10,294	11,139
2	13,167	14,218
3	16,079	17,374
4	20,614	22,314
5	24,382	26,439
6	27,560	29,897
7	31,205	34,009
8	34,774	37,934
9+	41,499	45,220

Employment-Related Concerns about the Future in 2007 and 2011

To produce this table, you will need to use the combined 2007 and 2011 dataset you created for the previous table. For the following variables, calculate the frequency and percentage endorsing each item. Calculate these percentages separately for 2007 and 2011. For these percentages, you may consider the denominator to include all individuals present in your combined dataset.

Use the “Concerned About” version of the variable for each of the following items. Note the wording may have varied slightly between the 2007 and 2011 waves but it should be similar enough that you can correctly locate the items in both years.

- Lack of ability to get training or degree
- Lack of money to complete education or get started in my chosen career field
- I am considered "overqualified"
- Lack of openings in my field
- Relocation is difficult or impossible
- Illness, accident, or disability
- Caring for a sick parent or relative
- Transportation problems - difficulty in getting to or from work

Bar Chart Depicting Average Number of Hours Worked by Day of Week in 2007 and 2011

Using the combined dataset created previously, create a grouped bar chart with one bar per day of week per year (for instance, Monday-2007). Group bars such that the two bars for a given day of the week are adjacent to one another and space exists between bars representing difference days of the week. Bar height should represent the mean number of hours worked in the past week as reported by the entire sample (in your combined dataset) with non-missing values for the given variable.

Include labels, titles, and legends to make your chart as interpretable and visually appealing as possible. You should also include the number of records used to calculate each mean somewhere on the chart.

Variables you will need to create this chart are labeled similar to the following:

- Work hours: Sunday high
- Hours worked in past week (high) – Sunday

Include all days of the week (Monday through Sunday) and always use the “high” versions of the variables.