**Biostatistics 203A**

**Mid-Quarter Project**

Due: November 2, 2023 at 5:00PM

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 primarily in employment-related characteristics and concerns before and after the 2008 economic recession. 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 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.
2. Subset the **W2007** and **W2011** datasetssuch 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

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 brief summary of the results (< 1 page), followed by a set of approximately 4 tables containing the requested results (as many pages as needed, see Submission Template for an example format). Even if you do not plan to use the Submission Template, you may want to review it to better understand expectations with regards to content and format.

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,” even though 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 uploaded to BruinLean in a PDF format no later than 5:00 PM on Thursday, November 2nd**.

Results that should be included in the submitted project:

**Demographic and Employment Characteristics in 2007**

Calculate and tabulate frequencies and percentages for each of the following variables found in your 2007 dataset. For each of these variables, calculate frequencies and percentages separately for two samples: (1) those respondents who appeared in the 2007 file but not the 2011 file and (2) those respondents who appeared in both the 2007 and 2011 files. Individuals are uniquely identified using the FAMID variable.

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 (unless otherwise noted) missing values should 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 available categories.

1. Gender (Use the Corrected 7/25/94 variable)
2. 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)
3. Currently employed (either part-time or full-time)
4. Currently married or cohabitating in an intimate relationship
5. 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.

1. How secure is your primary job?
2. How satisfied with your job as a whole?
   * Extremely or very dissatisfied
   * Somewhat dissatisfied
   * Somewhat satisfied
   * Extremely or very satisfied

**Demographic Descriptive Statistics in 2007**

Tabulate means and standard deviations for the following variables. For each of these variables, calculate means and standard deviations separately for two samples: (1) those respondents who appeared in the 2007 file but not the 2011 file and (2) those respondents who appeared in both the 2007 and 2011 files. You should also 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).

1. 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)
2. Annual household income in dollars (Use variable E5HI17)
3. *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.

**Financial Stress in 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. 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.

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

**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.

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

In summarizing your results in the “Results” section, you may want to consider the following questions:

* Among the sample of respondents who completed the 2007 assessment, were there demographic or employment differences between those that did and did not go on to complete the 2011 assessment?
* Among the subsample with assessments completed at both time points, does anything appear to have changed between 2007 and 2011?