

Instructor: Dr. Vahé Heboyan

Problem Set 2

Due: Sunday, November 13, 2022 by 11:59pm Eastern Time

100 points

This assignment focuses on topics related to continuous and categorical data analysis.

General rules for assignment submission:

- You need to submit 1 (one) document, which must be typed and submitted as a PDF (preferred) or Word file.
- Hand-written assignments are not accepted.
- Your answers must contain sufficient enough explanation for me to know that you comprehend the material. I will deduct points if you provide general or vague answers or explanations.
- When providing your answers/solutions, make sure question numbers below are referenced in your solution.
- Name your submission file as “PS2_FirstLastname”.
- Each assignment must be submitted through Assignment Submission system on course D2L website; Assessments > Assignments.
- Do not forget to provide your name on the document. 5 points will be deducted for missing student names.
- Please make sure that your submission is written and formatted neatly and professionally. Being able to present your results in a professional manner is an important factor in any discipline. Therefore, part of your grade for this assignment will also depend on how visually pleasing your submission is. Do not overdo. Just make sure it is clean and nicely formatted. Up to 10 points may be deducted for unprofessional submission.
- For all exercises in this assignment, you must use Stata to analyze your data.
- For all analyses, assume the level of significance (α) is set at 5%.
- Finally, since you will be using Stata for this assignment, the last page(s) of your assignment submission should include your Stata code. Your Stata code should include brief comments that explain to which question is a particular code line related to and what is it trying to accomplish. Keep it brief! Make sure when you paste the code from the Do-file into your assignment file, it comes out neat. The easiest way to do this is to ‘print’ your Do-file as a PDF (this saves your file as a PDF) and add to your PDF assignment file as the last pages (in Adobe Acrobat, you use Insert pages feature). Or you can copy and paste Do-file text into your assignment Word file. If you do this, it may require some editing to make it look nice (use Courier font after you paste the Do-file text). If you are having any difficulty with this step, please let me know.

1. Descriptive Statistics

- a) Use NHANS2 data to create descriptive summary for the following variables. For the continuous data, create a table that reports variable name, variable description, number of observations, mean, standard deviation, minimum, and maximum. For the categorical data, create a table that reports variable name, variable description, and category frequency and percentage of the total.

To receive full credit, make sure that your tables are professionally and aesthetically formatted and presented. You may consult journal article publications or research reports to see how they are done.

age, agegrp, bmi, bpdiastr, bpsystol, diabetes, heartatk, height, hlthstat, houssiz, race, rural, sex, tresult, tgresult, weight

- b) For the variables *bmi*, *bpdiastr*, and *bpsystol* create histograms that also have the normal curve overlayed. Graphically (histogram) and numerically test (e.g. Shapiro-Wilk Test) if these variables are normally distributed. Discuss what you learn from these histograms and tests.

2. Continuous Data Analysis

Using NHANS2 data, conduct hypothesis testing to answer the following research questions. In order to receive full credit, for each question below, you must explain which test you will use and why and provide a detailed explanation of your conclusions.

In these examples, given the large sample size, you should assume that the variables are approximately normally distributed.

- a) Is the average *bmi* in the sample 25?
- b) Is the average *bmi* for men and women equal?
- c) Is the average *bmi* for individuals of various races (black, white, other) equal?
- d) Is the average *bpsystol* in the sample 131?
- e) Is the average *bpsystol* for men and women equal?
- f) Is the average *bpsystol* for individuals of various races (black, white, other) equal?
- g) Test if the mean serum cholesterol levels in the sample are different by gender, race, and their interaction.

3. Categorical Data Analysis

- a) In our sample, *race* is represented by three categories (1=white, 2=black, 3=other). The US Census Bureau reports¹ that in 2019, 76.3% of the US population were White, 13.4% were African American, and the rest were of other races. Using data from NHANS2, test whether the observed percentages for the *race* categorical variable are significantly different from expected percentages reported by the Census Bureau. Explain in detail which test you would use and what your conclusions are.
- b) Use body mass index (BMI) variable in the NHANS2 data to create a categorical BMI variable with 4 categories that correspond to the standard BMI category ranges (underweight, normal, overweight, obese), which can be found here: https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html (look under “How is BMI interpreted for adults?”). Name the new categorical variable *bmi2*.

Note: in creating this new variable, make sure you code the ranges in a way that includes all values within the cutoff points. In our data, the variable *bmi* has 4 decimal digits but BMI cutoff points have a single digit. Therefore, if your first category is, say, “less than 7.5” and the 2nd category is 7.5-10.5, then if you only use single decimal digit value in coding in Stata (e.g. coding values between 0 and 7.4 as category 1), then 7.4 will be within the range of the 1st category. Since the 2nd category starts with 7.5, the value of 7.41 or 7.4999 will not be in the 1st category where it belongs. Therefore, if you code ranges using less than 4 decimal values, it will omit many values that have 4 decimals. To avoid this, make sure your value ranges are specified with 4 decimal points (e.g. 7.4999). Also, you should always check what your new variable looks like. When tabulating the new variable, if done correctly, it must only have 4 categories. You can use various methods to accomplish this in Stata. The easiest is to use `recode` command.

Finally, make sure all categories have meaningful labels, such as *underweight*, *normal*, *overweight*, and *obese*. Using `recode` command will help with this too. Please refer to the class example I demonstrated and/or Stata manual. Once done, display your new variable in a well formatted and professionally looking frequency table and a vertical bar graph. *Hint:* for the latter, you can create a histogram which displays percentages on the vertical axis.

Clarification on the Fisher’s Exact test. This test is used when the expected (not observed) cell value is less than 5. To calculate the expected values, you can use “, expected” option with `tabulate` twoway command in Stata. For example: `tab sex race, expected`

Using the NHANS2 dataset, answer research questions (c) to (f) below. Explain in detail which test you would use and what your conclusions are.

- c) Do men and women have the same risk of diabetes?
- d) Is the risk of a heart attack the same for individuals of all races?
- e) Is the risk of a heart attack the same for individuals of different BMI levels?
- f) Is there a relationship between being diabetic and suffering from a heart attack?

¹ Source: <https://www.census.gov/quickfacts/fact/table/US/PST045219>