Statistical Methods in Education Research
Vanderbilt University
Prof. Sean P. Corcoran

Due September 12, 2023
68 total points

Problem Set 2

Instructions: Answer the following questions in their entirety in a separate document. Submit your completed problem set as a PDF document via email to `sean.corcoran@vanderbilt.edu`. Use your last name and problem set number as the filename (e.g., *Knowles Problem Set 2.pdf*). Working together is encouraged, but it is expected that all submitted work be that of the individual student.

---

1. (**6 points**) The following 13 values (x) are the reported number of doctor's visits in the past year for a small subsample of respondents to the National Health Interview Survey in 2020:

$$5, 0, 33, 2, 1, 6, 6, 8, 0, 1, 4, 3, 1$$

   (a) Find the mean, median, and mode for this sample data. Which would you say is "best" for characterizing the central tendency of this distribution, and why?

   (b) Does any observation (or observations) appear to be an outlier? Discuss its impact on how the mean compares to the median.

   (c) What would happen to the mean and median if another observation were added to the sample with $x = 7$?

2. (**6 points**) Use the definition of the sample mean (and the properties of summation) to show that:

   (a) $\sum(x_i - \bar{x}) = 0$, where $\bar{x}$ is the sample mean.

   (b) $\sum(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$

3. (**56 points - 4 each**) On Github, locate the Stata dataset called *mepssample.dta*. These data are an extract from the Medical Expenditures Panel Survey, a large-scale survey of households about their health and health expenditures. (See `https://www.meps.ahrq.gov/mepsweb/`). Each observation is a person (N=19,386). In some cases there are multiple persons within the same household; you can ignore this aspect of the data.

   Answer the questions below in a .do file that includes a copy of each question followed by Stata output (where applicable) and your response to the question. Graphs can be saved and submitted separately, or combined into a .pdf file with the Stata log.

(a) Create a tabular relative frequency distribution and (percent) bar graph for the family size variable. What is the most common (modal) family size in this sample, and how often was it reported?

(b) The variables *mcs12* and *pcs12* are summary scores of well-being. MCS is the Mental Component Summary, and PCS is the Physical Component Summary. What are the mean and standard deviation of these variables in the data? Provide a "five number summary" (min, Q1, median, Q3, max) for these two variables and include the IQR.

(c) Create a new ordinal variable called *highested* that contains the highest education completed by the individual. Use the four variables beginning in *ed_* to do this. For example, *highested=0* if *ed_hs=0* (no high school completed), *highested=1* if *ed_hs=1* (high school completed but no more), etc. Repeat part (b), but do this separately by highest level of education completed. How do the MCS and PCS distributions compare across levels of educational attainment? For example, how do their measures of central tendency compare? Their variation?

(d) Create a boxplot that shows the distribution of number of doctor's office visits (*use_off*). What do the whiskers (tails) represent in this graph? Are there any outlier values of doctor's office visits?

(e) Now create a boxplot that shows the distribution of PCS separately by highest level of education completed. How do these distributions compare? Hint: you can use the `over( )` option.

(f) Based on a visual inspection of the graphs above, how would you describe the skewness of the variables you have examined thus far (family size, MCS, PCS, and doctor's office visits)?

(g) Use the skewness statistic to assess the skewness of the MCS, PCS, and doctor's office visits variables. In your do file, calculate the standard error of the skewness (see the lecture notes for the formula) and determine whether these distributions are "significantly" skewed or not.

(h) You are considering doing a log transformation of the doctor's office visits variable to reduce the skewness. Would this help? Why or why not? (Try it and see what happens).

(i) You are considering doing a log transformation of the PCS variable to reduce the skewness. Would this help? Why or why not? (Try it and see what happens).

(j) The variable *exp_tot* reports the total amount of medical expenses incurred during

the year. Use this variable to create a z-score for *exp_tot* as shown in class. Run a full set of descriptive statistics to demonstrate this new variable has a mean of 0 and standard deviation of 1.

(k) What level of medical expenditure corresponds to a z-score of 0.2 in this data? Of -0.2? Interpret these values in words.

(l) What proportion of individuals have a z-score of medical expenditures between -1 and +1? Why isn't this value 68% (or at least closer to it), as the Empirical Rule would suggest?

(m) What is the 43rd percentile for total medical expenses (*exp_tot*)? Explain/show how you got your answer.

(n) The variable *ins_unins* is a dichotomous variable that equals 1 if the individual lacks health insurance (and 0 otherwise). What is the mean of this variable and how should it be interpreted?

BELOW: NOT REQUIRED

Test your knowledge about variables that have been created using a linear transformation of another variable:

- If each value in a distribution with mean equal to 5 has been tripled, what is the new mean?
- If each value in a distribution with standard deviation equal to 5 has been tripled, what is the new standard deviation?
- If each value in a distribution with skewness equal to 1.14 has been tripled, what is the new skewness?
- If each value in a distribution with mean equal to 5 has the constant 6 added to it, what is the new mean?
- If each value in a distribution with standard deviation equal to 5 has the constant 6 added to it, what is the new standard deviation?
- If each value in a distribution with skewness equal to 1.14 has the constant 6 added to it, what is the new skewness?
- If each value in a distribution with mean equal to 5 has been multiplied by -2, what is the new mean?

- If each value in a distribution with standard deviation equal to 5 has been multiplied by -2, what is the new standard deviation?
- If each value in a distribution with skewness equal to 1.14 has been multiplied by -2, what is the new skewness?
- If each value in a distribution with mean equal to 5 has had a constant equal to 6 subtracted from it, what is the new mean?
- If each value in a distribution with standard deviation equal to 5 has had a constant equal to 6 subtracted from it, what is the new standard deviation?
- If each value in a distribution with skewness equal to 1.14 has had a constant equal to 6 subtracted from it, what is the new skewness?