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

Due September 20, 2022
66 total points

Problem Set 3

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., *Seinfeld PS3.pdf*). Working together is encouraged, but it is expected that all submitted work be that of the individual student.

---

1. (**6 points**) Answer each of the following questions about a variable that is the result of a linear transformation of another variable. (These do not require the use of Stata).

   (a) If each value in a distribution with mean equal to 5 has been tripled, what is the new mean?

   (b) If each value in a distribution with standard deviation equal to 5 has been tripled, what is the new standard deviation?

   (c) If each value in a distribution with skewness equal to 1.14 has been tripled, what is the new skewness?

   (d) If each value in a distribution with mean equal to 5 has the constant 6 added to it, what is the new mean?

   (e) 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?

   (f) If each value in a distribution with skewness equal to 1.14 has the constant 6 added to it, what is the new skewness?

   (g) If each value in a distribution with mean equal to 5 has been multiplied by -2, what is the new mean?

   (h) If each value in a distribution with standard deviation equal to 5 has been multiplied by -2, what is the new standard deviation?

   (i) If each value in a distribution with skewness equal to 1.14 has been multiplied by -2, what is the new skewness?

   (j) 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?

   (k) 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?

   (l) 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?

2. (**60 points**) For this problem use the file *mepssample.dta* on Github. 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.

   (a) (**4 points**) 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.

   (b) (**6 points**) 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 (a), but 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?

   (c) (**5 points**) 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?

   (d) (**5 points**) Now create a boxplot that shows the distribution of PCS separately by highest level of education completed. How do these distributions compare?

   (e) (**5 points**) Based on a visual inspection of the graphs above, how would you describe the skewness of the variables you have examined thus far (MCS, PCS, and doctor's office visits)?

   (f) (**5 points**) Use the skewness statistic to assess the skewness of these variables (MCS, PCS, and doctor's office visits). 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.

   (g) (**5 points**) 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).

   (h) (**5 points**) 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).

(i) (**5 points**) 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.

(j) (**5 points**) What level of medical expenditure corresponds to a z-score of 0.2 in this data? Of -0.2? Interpret these values in words.

(k) (**5 points**) 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?

(l) (**5 points**) What is the 43rd percentile for total medical expenses (*exp_tot*)? Explain/show how you got your answer.