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

Due September 16, 2021
56 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., *Cash Problem Set 3.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. (**50 points**) For this problem use the file *TexasEM2007-08.dta* on Github. These data represent test performance and other characteristics of Texas elementary and middle schools during the 2007-08 academic year. Each observation is a school (N=6,354).

   (a) (**5 points**) The variables called *ca311tmr, ca311tcr, ca311tsr,* and *ca311trr* provide the percent of students in a school testing at the "proficient" level or higher in math, science, social studies, and reading, respectively. Provide a "five number summary" (min, Q1, median, Q3, max) for these four variables and include the interquartile range. Do this once for the whole population of schools, and then a second time restricting the sample to schools in Houston. (There is an indicator variable called *houston* than equals one for schools in Houston). How do the distributions of scores compare? Which subject has the lowest median, and which has the greatest variability based on the IQR?

   (b) (**5 points**) Create a boxplot that shows the distribution of student attendance rates (*ca0atr*), restricting the analysis to schools in Houston. What do the whiskers (tails) represent in this graph? Are there any outlier values of attenance rates?

   (c) (**5 points**) Now create a boxplot that shows the distribution of student attendance rates specifically for Black, Hipsanic, and white students, restricting the analysis to schools in Houston. These subgroup-specific attendance rates are reported as separate variables (*cb0atr, ch0atr, cw0atr*). How do these distributions compare?

   (d) (**5 points**) How would you describe the skewness of the variables you have examined thus far (proficiency and attendance rates)? Use any summary statistics or graphical summary that is appropriate.

   (e) (**5 points**) Consider the variable called *cpemallp*, which represents the school's percentage of students who attended that school less than 83% percent of the school year. (They refer to this as the "mobility" rate). Use the skewness statistic to assess the skewness of this variable. In your do file, calculate the standard error of the skewness (see the lecture notes for the formula) and determine whether this distribution is "significantly" skewed or not.

   (f) (**5 points**) Generate a new variable that contains the *natural log* of *cpemallp.* Find its skewness statistic and standard error of the skewness. Has this log transformation reduced the severity of skewness in this variable? Are all of the values of *cpemallp* valid for the log transformation?

   (g) (**5 points**) As an alternative to the log transformation, generate a new variable that contains the inverse hyperbolic sine of *cpemallp.* The IHS function for a variable x is defined as:

$$\text{IHS} = ln(x + sqrt(x^2 + 1))$$

How does the skewness of this variable compare the original *cpemallp* variable?

(h) (**5 points**) The variable *cpetecop* contains the percent of students in the school who are considered to be economically disadvantaged. Use this variable to create a z-score for *cpetecop* 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.

(i) (**5 points**) Using the information from part (h), what level of economic disadvantage corresponds to a z-score of 1.2? Of -1.2? Interpret these values in words.

(j) (**5 points**) What proportion of schools have a level of economic disadvantage between a z-score of -1 and +1? Why isn't this value 68% (or at least closer to it), as the Empirical Rule would suggest?