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

Due September 13, 2022
45 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., *Swift 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. (**33 points - 3 each**) On Github, locate the Stata dataset called *TNDOE schools 2018-19.* This dataset is a compilation of selected demographic and school performance measures for 1,756 schools in Tennessee in 2018-19. 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) How many variables are in this dataset? What is an example of a *string*-type variable, and what is an example of a *numeric* variable?

(b) Create a tabular relative frequency distribution for the "grades served" variable. What is the most common grade span in Tennessee in 2018-19?

(c) In Tennessee's school accountability system, schools receive 0-4 points for various indicators (achievement, growth, chronic absenteeism, etc.) Create a bar graph showing the relative frequency distribution of the indicator score for school achievement. How many schools were included in this graph? What is the modal number of points earned on this metric? What percentage of schools received a score of 2 *or lower* on this metric?

(d) Create a pie graph showing the relative share of schools receiving each indicator score for school achievement (i.e., using the same variable you used in part c).

(e) Create a histogram for the high school graduation rate. How many schools were included in this graph? How would you describe the *shape* of this distribution?

(f) Create a histogram for the school per pupil expenditure. How many schools were included in this graph? How would you describe the *shape* of this distribution?

(g) Repeat part (f), but *exclude* schools with a per-pupil expenditure above $18,000. How many schools were excluded that had non-missing expenditure above $18,000? How does this change the shape of the distribution, if at all? (Hint: if needed, refer to the Stata basics handout on Github to see how to execute a command for a subset of cases where a condition is true or not true).

(h) Find the mean and median high school graduation rate for the schools in this dataset. How do they compare?

(i) Now find the mean and median high school graduation rate for schools in the Metro Nashville Public Schools (*district_id* equal to 190). How do they compare, and how do they compare to the state as a whole?

(j) The variable *izone* is a dichotomous variable that equals one if the school is part of a district Innovation Zone (an approach to turning around low-performing schools). What is the mean of the *izone* variable and how should it be interpreted?

(k) Finally, explain why the mode is not a useful measure of central tendency for the school per pupil expenditure variable.