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

September 13, 2022
**45 total points**

---

**Problem Set 2 Solutions**

---

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 $x = 7$?

> See Stata syntax and results below (you can also obtain these easily by hand). The mean is $\bar{x} = \sum x_i/n = 70/13 = 5.4$. The median is 3 (the 7th value when ordered from smallest to largest). The mode is 1, a value that occurs 3 times. 33 is clearly an outlier, and increases the mean relative to the median (the median is 3 while the mean is 5.4). This would suggest the median is preferable to the mean, although it often pays to report both, and note the difference. Adding an observation of $x = 7$ will increase the mean (since it is greater than 5.4) and the median rises to 3.5 (the midpoint of 3 and 4).

```
clear
set obs 13
gen x=5 in 1
replace x=0 in 2
replace x=33 in 3
replace x=2 in 4
replace x=1 in 5
replace x=6 in 6
replace x=6 in 7
replace x=8 in 8
replace x=0 in 9
replace x=1 in 10
replace x=4 in 11
replace x=3 in 12
replace x=1 in 13
summ x, detail
```

```
                            x
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0             0
 5%            0             0
10%            0             1       Obs                  13
25%            1             1       Sum of Wgt.          13

50%            3                     Mean           5.384615
                        Largest       Std. Dev.      8.684646
75%            6             6
90%            8             6       Variance       75.42308
95%           33             8       Skewness       2.708977
99%           33            33       Kurtosis        9.27715

. tabstat x, stat(sum)

    variable |       sum
-------------+----------
           x |        70
------------------------

. egen modex=mode(x)
. table modex

----------------------
    modex |      Freq.
----------+-----------
        1 |         13
----------------------

. set obs 14
number of observations (_N) was 13, now 14

. replace x=7 in 14
(1 real change made)

. summ x, detail
                            x
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0             0
 5%            0             0
10%            0             1       Obs                  14
25%            1             1       Sum of Wgt.          14

50%          3.5                     Mean                5.5
                        Largest       Std. Dev.        8.3551
75%            6             6
90%            8             7       Variance       69.80769
95%           33             8       Skewness       2.757567
99%           33            33       Kurtosis         9.7783
```

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.

$$\begin{aligned}
\sum(x_i - \bar{x}) &= 0 \\
\sum x_i - n\bar{x} &= 0 \\
\frac{n\sum x_i}{n} - n\bar{x} &= 0 \\
n\bar{x} - n\bar{x} &= 0
\end{aligned}$$

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

$$\begin{aligned}
\sum(x_i - \bar{x})^2 &= \sum x_i^2 - n\bar{x}^2 \\
\sum(x_i^2 - 2x_i\bar{x} + \bar{x}^2) &= \\
\sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2 &= \\
\sum x_i^2 - 2\bar{x}\frac{n\sum x_i}{n} + n\bar{x}^2 &= \\
\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 &= \\
\sum x_i^2 - n\bar{x}^2 &= \sum x_i^2 - n\bar{x}^2
\end{aligned}$$

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?

   The `describe` command in Stata will provide answers to these questions. There are 1,756 observations (schools) and 38 variables. There are several string variables, including *grades_served* and *district_name*. There are many numeric variables, including *t_experienced_p*, the percentage of experienced teachers at a school. The variable type is visible in the "storage type" column. You can also find this information in the Properties window, and

in the Variables Manager.

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

The relative frequency distribution is shown below (the Percent column). I used `tabulate` with the `sort` option to make it clear that Grades PK-5 is the most common grade span. 19.4% of schools had this grade span in 2018-19.

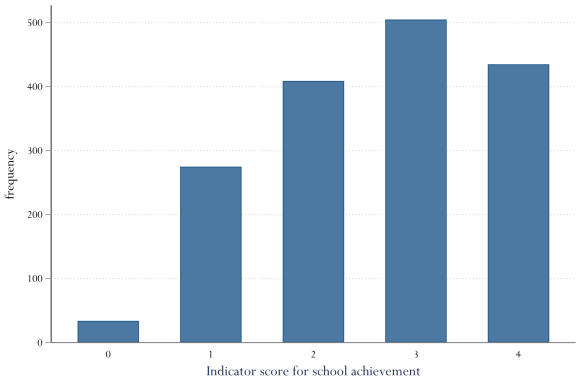
```
. tabulate grades_served, sort

GRADES_SERVE |
           D |      Freq.     Percent        Cum.
-------------+-----------------------------------
 Grades PK-5 |        341       19.43       19.43
 Grades 9-12 |        292       16.64       36.07
  Grades 6-8 |        217       12.36       48.43
  Grades K-5 |        199       11.34       59.77
 Grades PK-8 |        137        7.81       67.58
 Grades PK-4 |        111        6.32       73.90
  Grades 5-8 |         92        5.24       79.15
  Grades K-8 |         52        2.96       82.11
  Grades K-4 |         44        2.51       84.62
 Grades 6-12 |         28        1.60       86.21
 Grades PK-6 |         24        1.37       87.58
 Grades PK-2 |         23        1.31       88.89
  Grades K-6 |         21        1.20       90.09
  Grades 3-5 |         20        1.14       91.23
Grades PK-12 |         20        1.14       92.36
 Grades 7-12 |         16        0.91       93.28
 Grades K-12 |         13        0.74       94.02
 Grades PK-3 |         11        0.63       94.64
Grades 10-12 |          9        0.51       95.16
  Grades 7-8 |          9        0.51       95.67
  Grades K-3 |          7        0.40       96.07
  Grades 4-8 |          6        0.34       96.41
  Grades 5-6 |          5        0.28       96.70
...
-------------+-----------------------------------
       Total |      1,755      100.00
```

(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?

The bar graph is shown below. Using `tabulate`, it is clear that 1,658 schools have non-missing values of this indicator score (0-4). 43.3% of schools receive a score of 2 or lower on this metric.

The modal score is a 3 (30.5% of schools scored a 3).

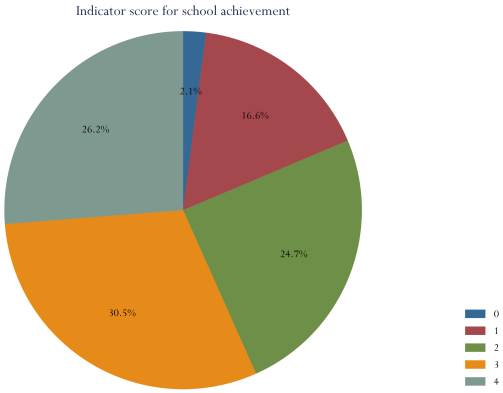
```
tabulate indicator_ach_all
graph bar (count), over(indicator_ach_all) ///
  title(Indicator score for school achievement, position(6))
```




(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).

The pie graph is shown below. Note the option for labeling the slices with the relative frequencies (percents).

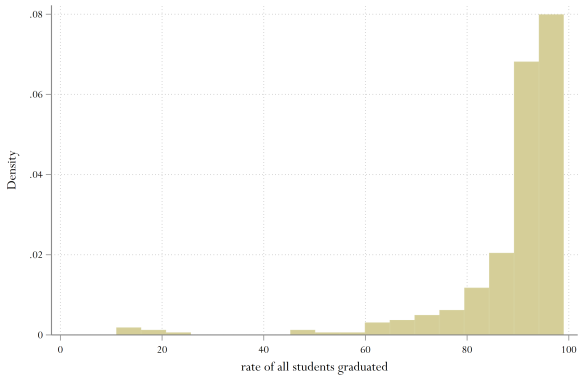
```
graph pie, over(indicator_ach_all) plabel(_all percent, format(%3.1f)) ///
   title(Indicator score for school achievement) scheme(modern)
```




(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?

The histogram is shown below. Using the `summarize` command we learn that 330 schools have non-missing graduation rates. (Those missing this data are probably not high schools, or may be new schools without a graduating class). The histogram indicates a very strong negative (left) skew.

```
summ grad_rate_all
histogram grad_rate_all
```

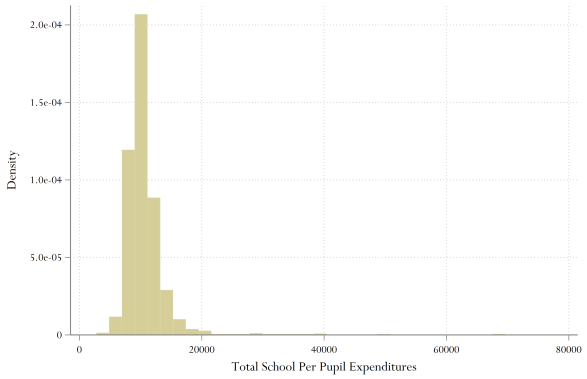



(f) Create a histogram for the school per pupil expenditure. How many schools were included in this graph? How would you describe the *shape* of this distribution?

The histogram is shown on the next page. Using the `summarize` command we learn that 1,751 schools have non-missing per-pupil spending. The histogram indicates a very strong positive (right) skew. This is due to some very large outliers. While median spending in the state is about $10,000 per student, there are a small number of schools with spending of $25,000 or more per student.

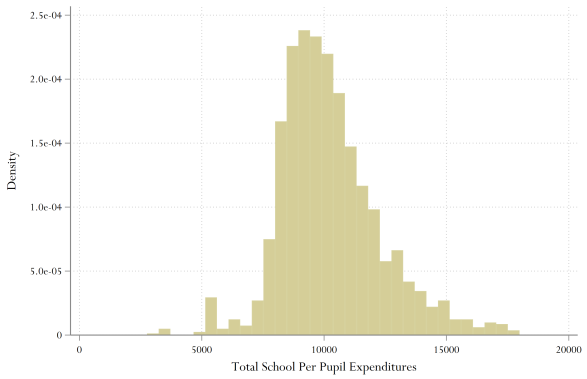
```
summ totalschoolperpupilexpenditu, detail
histogram totalschoolperpupilexpenditu
```

(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).




The histogram is shown below. Using the `summarize` command we learn that 1,712 schools have per-pupil spending at or below \$18,000. The histogram is still positively skewed, but it is more symmetric that the original histogram that included outliers.

```
summ totalschoolperpupilexpenditu if totalschoolperpupilexpenditu<=18000 , detail
histogram totalschoolperpupilexpenditu if totalschoolperpupilexpenditu<=18000
```




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

The mean graduation rate is 89% while the median is 93.2%. As the mean < median, this indicates a negatively skewed distribution. (This is confirmed by the skewness statistic).

```
. summ grad_rate_all,det

              rate of all students graduated
-------------------------------------------------------------
```

```
      Percentiles      Smallest
 1%         18.8            11
 5%         66.1          11.8
10%           77          14.8       Obs                 330
25%         88.3          18.8       Sum of Wgt.         330

50%         93.2                     Mean           89.02273
                       Largest       Std. Dev.      13.18537
75%         95.7          98.8
90%         97.3          98.9       Variance        173.854
95%           98            99       Skewness       -3.60364
99%         98.8            99       Kurtosis       18.67739
```

(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?

The mean graduation rate in MNPS is 79.9% while the median is 83.6.2%. As the mean < median, this indicates a negatively skewed distribution. Both the mean and median in MNPS are less than the state as a whole.

```
. summ grad_rate_all if district_id==190,det

              rate of all students graduated
-------------------------------------------------------------
      Percentiles      Smallest
 1%           20            20
 5%           64            64
10%         66.9          66.9       Obs                  24
25%        75.75          69.3       Sum of Wgt.          24

50%        83.55                     Mean            79.94583
                       Largest       Std. Dev.       15.53445
75%        89.35          92.1
90%         93.1          93.1       Variance        241.3191
95%         95.8          95.8       Skewness       -2.449619
99%         96.4          96.4       Kurtosis        10.32382
```

(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?

The mean of a dichotomous variable is interpreted as the proportion equal to one. The mean of *izone* is 0.023 (see below), meaning that 2.3% of the schools in the state are part of an Innovation Zone.

```
. summ izone

    Variable |        Obs        Mean    Std. Dev.       Min        Max
```

```
-------------+---------------------------------------------------------
       izone |      1,756     .022779    .1492409          0          1
```

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

Per-pupil spending is a variable with a large number of unique values and not many repeats. (It is technically a discrete variable, but treated as continuous for practical purposes). The mode is the most frequently occurring value in the distribution. Since few values of per-pupil spending recur, the mode is not that informative.