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

September 16, 2021
56 total points

Problem Set 3 *Solutions*

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? **15 (the mean also triples)**

   (b) If each value in a distribution with standard deviation equal to 5 has been tripled, what is the new standard deviation? **15 (the standard deviation also triples). In general if one multiplies a variable by b the standard deviation of the transformed variable is |b| times the old standard deviation.**

   (c) If each value in a distribution with skewness equal to 1.14 has been tripled, what is the new skewness? **1.14 (the skewness is unchanged unless multiplying by a negative number)**

   (d) If each value in a distribution with mean equal to 5 has the constant 6 added to it, what is the new mean? **11 (the original mean +6)**

   (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? **Adding a constant to a variable has no effect on the standard deviation (5).**

   (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? **1.14 (the skewness is unchanged unless multiplying by a negative number)**

   (g) If each value in a distribution with mean equal to 5 has been multiplied by -2, what is the new mean? **-10. In general if one multiplies a variable by b the mean of the transformed variable is b times the old 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? **10. In general if one multiplies a variable by b the standard deviation of the transformed variable is |b| times the old 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? **-1.14. When multiplying a variable by a negative number, the skewness of the transformed variable is -1 times the old 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? **-1 (the original mean minus 6)**

(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? **Adding/subtracting a constant to a variable has no effect on the standard deviation (5).**

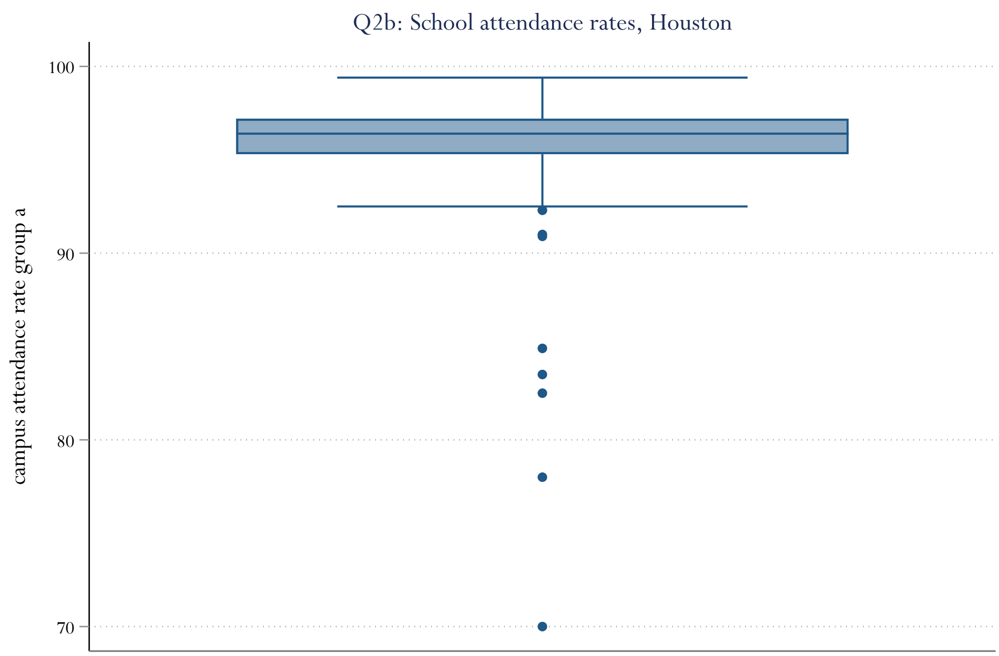
(l) If each value in a distribution with skewness equal to 1.14 has had a constant equal to 6 subtracted from it, what is the new skewness? **1.14. The skewness is unaffected unless the original variable has been multiplied by a negative value.**

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).

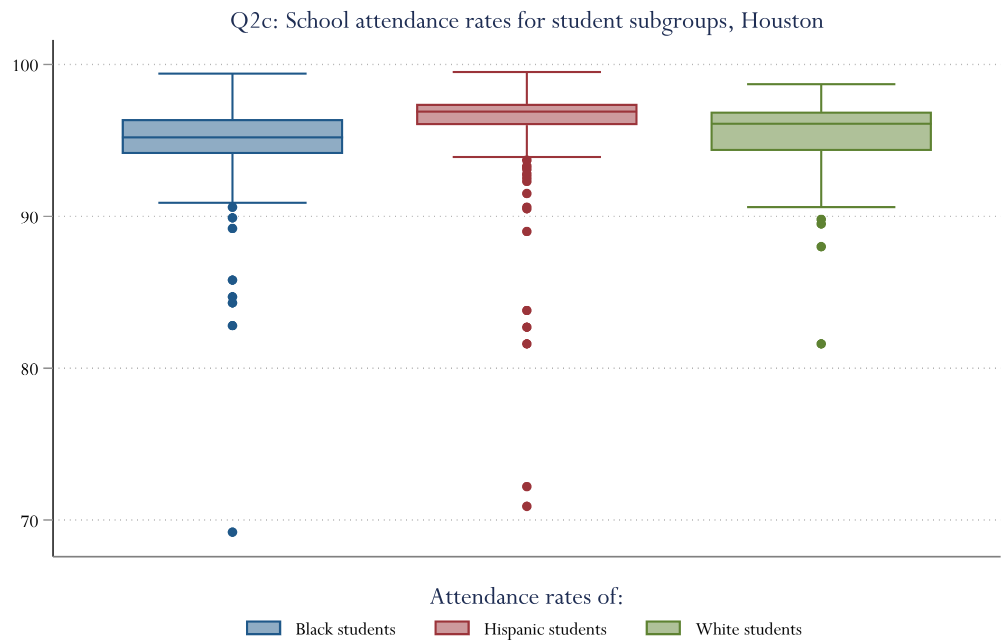
   See the attached log file. Note I used the user-written `graphlog` to integrate my output and graphs in to a PDF. The formatting is not great. `graphlog` requires a LaTeX installation (like MiKTeX). Other than that, using the command is quite easy: you create a log as a txt file, save your graphics as PDF files along the way using `graph export`, and then include a `graphlog` command at the end.

```
.
. // ***********************************************************************
. // LPO.8800 Problem Set 3 - Solution to Question 2
. // Last updated: September 15, 2021
. // ***********************************************************************
.
. /* QUESTION #2: Texas elementary and middle 2007-08.dta, represents test performance and
> other characteristics of Texas elementary and middle schools during the 2007-08 AY. Each
> observation is a school (N=6,354). */
.
. use https://github.com/spcorcor18/LPO-8800/raw/main/data/TexasEM2007-08.dta, ///   clear
.
. // ********
. // Part a
. // ********
. // 5 POINTS
.
. /* The variables called ca311tmr, ca311tcr, ca311tsr, and ca311trr provide the percent o
> f students in a school testing at the proficient level or higher in math, science, socia
> l 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 th
> e whole population of schools, and then a second time restricting the sample to schools
> in Houston. (There is an indicator variable called houston that equals one for schools i
> n Houston). How do the distributions of scores compare? Which subject has the lowest med
> ian, and which has the greatest variability based on the IQR? */
.
. tabstat ca311tmr ca311tcr ca311tsr ca311trr, stat(min p25 p50 p75 max iqr)
   stats |  ca311tmr  ca311tcr  ca311tsr  ca311trr
---------+----------------------------------------
     min |         4         5        26        11
     p25 |        79        67        86        87
     p50 |        87        79        91        92
     p75 |        93        88        95        96
     max |        99        99        99        99
     iqr |        14        21         9         9
--------------------------------------------------
.
. // Houston only
. tabstat ca311tmr ca311tcr ca311tsr ca311trr if houston==1, stat(min p25 p50 p75 max iqr)
   stats |  ca311tmr  ca311tcr  ca311tsr  ca311trr
---------+----------------------------------------
     min |        22         6        29        52
     p25 |        75        68        81        82
     p50 |        84        81        89        87
     p75 |        90        89        94        92
     max |        99        99        99        99
     iqr |        15        21        13        10
--------------------------------------------------
.
. // ***********************************************************************
. // ***********************************************************************
. /* The five-number summaries are shown above. The subject with the lowest median profici
> ency (79) and highest variability (21) is science. The median in the other three subject
> s is quite high. The minimum values and Q1 are lowest in math and science.*/
. // ***********************************************************************
. // ***********************************************************************
.
. // ********
```

```
. // Part b
. // ********
. /* Create a boxplot that shows the distribution of student attendance rates (ca0atr), re
> stricting the analysis to schools in Houston. What do the whiskers (tails) represent in
> this graph? Are there any outlier values of attenance rates? */
.
. graph box ca0atr if houston==1, title("Q2b: School attendance rates, Houston") ///
> name(attboxhouston, replace)
. graph export ca0atr.pdf, name(attboxhouston) as(pdf) replace
(file ca0atr.pdf written in PDF format)
```




```
.
. /* The whiskers extend to the maximum and minimum, or the adjacent values if there are o
> utliers. There are outliers at the bottom of the distribution. These represent attendanc
> e rates that are more than 1.5 IQR below the 25th percentile */
.
.
. // ********
. // Part c
. // ********
. /* Now create a boxplot that shows the distribution of student attendance rates specific
> ally for Black, Hispanic, and white students, restricting the analysis to schools in Hou
> ston. These subgroup-specific attendance rates arereported as separate variables (cb0atr
> , ch0atr, cw0atr). How do these distributions compare? */
.
. # delimit ;
delimiter now ;
. graph box cb0atr ch0atr cw0atr if houston==1, legend(position(6) row(1)         label(1
> "Black students") label(2 "Hispanic students")         label(3 "White students") title("
> Attendance rates of:"))         title("Q2c: School attendance rates for student subgroup
> s, Houston")         name(attboxhoustonr, replace);
. graph export attrace.pdf, name(attboxhoustonr) as(pdf) replace;
(file attrace.pdf written in PDF format)
```

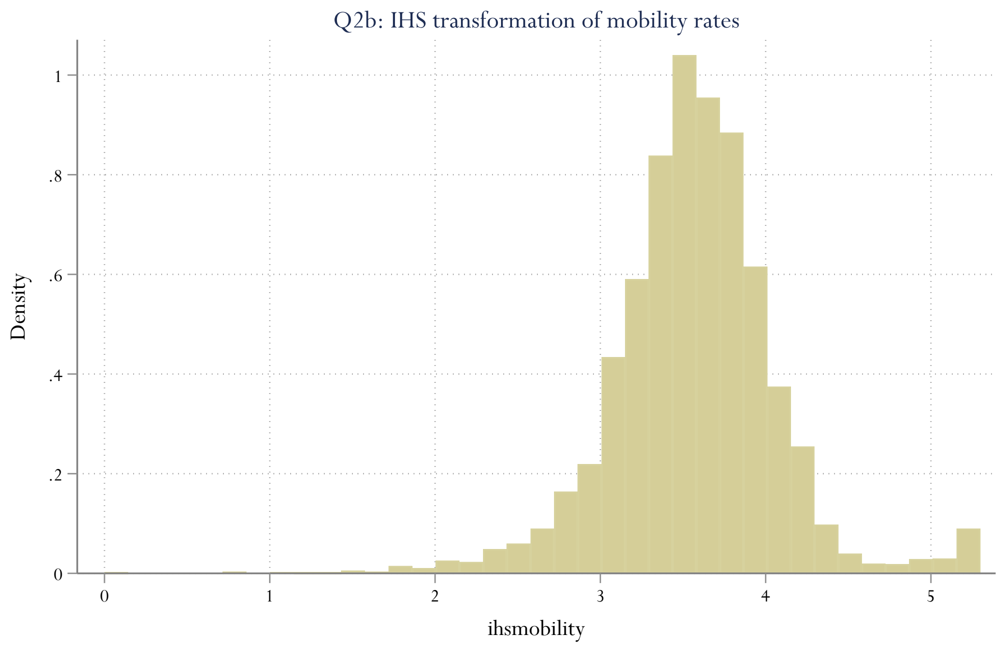



```
. # delimit cr
delimiter now cr
.
. /* The median attendance rate appears to be highest for Hispanic students, followed by w
> hite students and then Black. The attendance rates for Black students appears to be most
> variable, and for Hispanic students the least variable (although in both cases there are
> a lot of outliers on the low end). In all three cases, the vast majority of schools have
> an attendance rate of 90% or greater */
.
.
. // ********
. // Part d
. // ********
. /* How would you describe the skewness of the variables you have examined thusfar (profi
> ciency and attendance rates)? Use any summary statistics or graphicalsummary that is app
> ropriate. */
.
. /* Based on a visual inspection of the boxplots thus far, school proficiency and attenda
> nce rates appear to have a negative (left) skew. This is confirmed by a look at the skew
> ness statistic for each variable (again limiting the analysis to Houston). */
.
. tabstat ca0atr cb0atr ch0atr cw0atr if houston==1, stat(skew)
   stats |    ca0atr    cb0atr    ch0atr    cw0atr
---------+----------------------------------------
skewness | -5.070888 -4.100763  -5.20807 -2.307668
--------------------------------------------------
.
.
. // ********
. // Part e
. // ********
.
```

```
. /* Consider the variable called cpemallp, which represents the school's percentage of st
> udents 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.*/
.
. summ cpemallp, detail
                     mobility (percent)
-------------------------------------------------------------
      Percentiles      Smallest
 1%          4.3              0
 5%            8             .8
10%         10.2             .8       Obs               6,057
25%         13.6             .9       Sum of Wgt.       6,057
50%         17.8                      Mean           20.19678
                        Largest       Std. Dev.      12.91701
75%         23.1            100
90%           30            100       Variance       166.8493
95%         36.1            100       Skewness       3.630154
99%         96.8            100       Kurtosis        20.8207
.
. // Divide the skewness statistic (saved as "a") by the standard error of
. // the skewness (calculated as "b"). r(N) is the count of observations used
. // in the above command. The rule of thumb is that if this absolute value of
. // the ratio is >2, the distribution is significantly skewed. It is.
.
. scalar a=r(skewness)
. scalar b=sqrt((6*r(N)*(r(N)-1))/((r(N)-2)*(r(N)+1)*(r(N)+3)))
. display a
3.6301541
. display b
.03146584
. display a/b
115.3681
.
.
. // ********
. // Part f
. // ********
.
. /* 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 se
> verity of skewness in this variable? Are all of the values of cpemallp valid for the log
> transformation? */
.
. /* Results are below. The ratio of skewness to the standard error of the skewness is now
> less than 2 in absolute value. Logs are only valid for values >0, and we prefer that the
> y be >1. There is one case of cpemallp less than 0, and 3 values less than 1. */
.
. gen lnmobility=ln(cpemallp)
(298 missing values generated)
```

```
. summm lnmobility, detail
                         lnmobility
-------------------------------------------------------------
      Percentiles      Smallest
 1%     1.458615      -.2231435
 5%     2.079442      -.2231435
10%     2.322388      -.1053605       Obs               6,056
25%      2.61007       .3364722       Sum of Wgt.       6,056
50%     2.879198                      Mean            2.87257
                        Largest       Std. Dev.      .5026023
75%     3.139833        4.60517
90%     3.401197        4.60517       Variance       .2526091
95%     3.586293        4.60517       Skewness      -.0343937
99%     4.572647        4.60517       Kurtosis       6.081142
.
. scalar a=r(skewness)
. scalar b=sqrt((6*r(N)*(r(N)-1))/((r(N)-2)*(r(N)+1)*(r(N)+3)))
. display a
-.03439369
. display b
.03146844
. display a/b
-1.0929582
.
. count if cpemallp<1
  4
.
.
. // ********
. // Part g
. // ********
.
. /* As an alternative to the log transformation, generate a new variable that contains th
> e inverse hyperbolic sine of cpemallp. The IHS function for a variable x is defined as:
> IHS = ln(x + sqrt(x^2 + 1)). How does the skewness of this variable compare the original
> cpemallp variable? */
.
. /* Results are below. The transformed variable is less skewed than theoriginal cpemallp
> variable. */
.
. gen ihsmobility=ln(cpemallp + sqrt(cpemallp^2 + 1))
(297 missing values generated)
. summ ihsmobility, detail
                         ihsmobility
-------------------------------------------------------------
      Percentiles      Smallest
 1%     2.165017              0
 5%     2.776472       .7326683
10%     3.017929       .7326683       Obs               6,057
25%     3.304566       .8088669       Sum of Wgt.       6,057
50%     3.573134                      Mean            3.56668
                        Largest       Std. Dev.      .5019767
75%     3.833448       5.298342
90%     4.094622       5.298342       Variance       .2519806
95%     4.279632       5.298342       Skewness      -.0396856
99%     5.265821       5.298342       Kurtosis       6.179481
.
. scalar a=r(skewness)
. scalar b=sqrt((6*r(N)*(r(N)-1))/((r(N)-2)*(r(N)+1)*(r(N)+3)))
```

```
. display a
-.03968558
. display b
.03146584
. display a/b
-1.2612276
.
. histogram ihsmobility, title("Q2b: IHS transformation of mobility rates") ///        nam
> e(ihsmob, replace)
(bin=37, start=0, width=.14319844)
. graph export ihsmob.pdf, name(ihsmob) as(pdf) replace
(file ihsmob.pdf written in PDF format)
```




```
.
.
. // ********
. // Part h
. // ********
.
. /* The variable cpetecop contains the percent of students in the school who are consider
> ed 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 vari
> able has a mean of 0 and standard deviation of 1. */
.
. /* Results are below. The mean is ~0 and standard deviation ~1 */
.
. egen zecondis=std(cpetecop)
. summ zecondis
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    zecondis |      6,354    3.01e-10           1  -2.252633   1.489719
.
.
. // ********
. // Part i
. // ********
.
```

```
. /* 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. */
.
. /* Results are below, calculated using the mean and sd from the original cpetecop variab
> le. A school that is 1.2 standard deviations above the mean (z=1.2) in economic disadvan
> tage has a 92.3% share of disadvantaged students. A school that is 1.2 standard deviatio
> ns below the mean (z=-1.2) has a 28.1% share of disadvantaged students */
.
. summ cpetecop
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    cpetecop |      6,354    60.19298    26.72116          0        100
. display r(mean) + 1.2*r(sd)
92.258375
. display r(mean) - 1.2*r(sd)
28.127587
.
.
. // ********
. // Part j
. // ********
.
. /* 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 wo
> uld suggest? */
.
. /* Results are below. First I count the observations with a z-score between -1 and 1 and
> store it as "a". Then I count the number of non-missing z-scores and store this as "b".
> The proportion is a/b, or 60.6%. The Empirical Rule applies to **normal** distributions,
> which this is not */
.
. count if zecondis>-1 & zecondis<=1
  3,851
. scalar a = r(N)
. count if zecondis~=.
  6,354
. scalar b = r(N)
. display a/b
.60607491
.
. capture log close
```