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

September 20, 2022
66 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. (**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.

   See the attached log file

```
.
. // ********************************************************************
. // LPO.8800 Problem Set 3 - Solution to Question 2
. // Last updated: September 16, 2021
. // ********************************************************************
.
. /* QUESTION #2: 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
> .*/
.
.
. use https://github.com/spcorcor18/LPO-8800/raw/main/data/mepssample.dta, ///
>    clear
(Sample of MEPS 2004 data)

.
. // ********
. // Part 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.*/
.
. summ mcs12 pcs12

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       mcs12 |     19,386    50.22171    10.19464       1.35      75.06
       pcs12 |     19,386    49.01453    11.01185       4.56      72.17

. tabstat mcs12 pcs12, stat(min p25 p50 p75 max iqr)

   stats |     mcs12     pcs12
---------+--------------------
     min |      1.35      4.56
     p25 |      44.3     43.77
     p50 |     52.65     52.99
     p75 |     57.33     56.71
     max |     75.06     72.17
     iqr |     13.03     12.94
------------------------------

.
.
. // ********
. // Part 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?*/
```

```
.
. gen highested=0 if ed_hs==0
(5,764 missing values generated)

. replace highested=1 if ed_hs==1
(5,764 real changes made)

. replace highested=2 if ed_hsplus==1
(5,017 real changes made)

. replace highested=3 if ed_col==1
(3,307 real changes made)

. replace highested=4 if ed_colplus==1
(2,467 real changes made)

. label define hed 0 "no HS" 1 "HS" 2 "some college" 3 "college" 4 "college+",re
> place

. label values highested hed

. fre highested

highested
--------------------------------------------------------------------
                       |      Freq.    Percent      Valid       Cum.
-----------------------+--------------------------------------------
Valid   0 no HS        |       2831      14.60      14.60      14.60
        1 HS           |       5764      29.73      29.73      44.34
        2 some college |       5017      25.88      25.88      70.22
        3 college      |       3307      17.06      17.06      87.27
        4 college+     |       2467      12.73      12.73     100.00
        Total          |      19386     100.00     100.00           
--------------------------------------------------------------------

.
. /* when creating variables like highested above it pays to verify how the
> component ed_* variables are coded. As a check to see whether individuals are
> coded a "1" more than once in the ed_* variables, I used the code below.
> There are no such cases--variables seem mutually exclusive).*/
.
. egen check=rowtotal(ed_*)

. tabulate check

      check |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,831       14.60       14.60
          1 |     16,555       85.40      100.00
------------+-----------------------------------
      Total |     19,386      100.00

. drop check

.
. tabstat mcs12, by(highested) stat(mean sd min p25 p50 p75 max iqr)

Summary for variables: mcs12
     by categories of: highested

   highested |      mean        sd       min       p25       p50       p75
-------------+------------------------------------------------------------
       no HS |  47.52533  11.34479       1.9     39.96     48.77     56.94
          HS |   49.6437  10.61426      4.73    43.215        52     57.33
some college |  50.57164  10.11016      1.35     45.15     53.35     57.33
     college |    51.403  9.061986     12.45     46.98      54.1     57.16
    college+ |  52.37128  8.492424     11.71     48.84     54.37     57.63
-------------+------------------------------------------------------------
       Total |  50.22171  10.19464      1.35      44.3     52.65     57.33
--------------------------------------------------------------------------
```

```
   highested |       max       iqr
-------------+--------------------
       no HS |     74.15     16.98
          HS |     75.06    14.115
some college |     74.84     12.18
     college |     72.43     10.18
    college+ |     70.48  8.790001
-------------+--------------------
       Total |     75.06     13.03
----------------------------------

. tabstat pcs12, by(highested) stat(mean sd min p25 p50 p75 max iqr)

Summary for variables: pcs12
     by categories of: highested

   highested |      mean        sd       min       p25       p50       p75
-------------+------------------------------------------------------------
       no HS |  45.16709  12.34802      6.08     36.68     48.75     55.09
          HS |  47.97236  11.20774      7.57     41.98     51.93     56.15
some college |  49.26236  10.84113      4.56     44.27     53.18     56.71
     college |  51.44491  9.687731      7.33     48.18      54.8     57.57
    college+ |   52.1027  9.092791     11.29     49.46     55.13     57.76
-------------+------------------------------------------------------------
       Total |  49.01453  11.01185      4.56     43.77     52.99     56.71
--------------------------------------------------------------------------

   highested |       max       iqr
-------------+--------------------
       no HS |     65.73     18.41
          HS |     72.17     14.17
some college |     70.87     12.44
     college |      71.7  9.389999
    college+ |     69.86  8.299999
-------------+--------------------
       Total |     72.17     12.94
----------------------------------

.
. // ********************************************************************
. // ********************************************************************
. // Mental Component Survey (MCS) and Physical Component Survey (PCS): the
. // central tendency of these measures (both mean and median) rise as
. // educational attainment increases. Variability in these measures (both
. // the sd and the IQR) fall as educational attainment increases.
. // ********************************************************************
. // ********************************************************************
.
.
. // ********
. // Part 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?*/
.
. graph box use_off , title("Number of office-based provided visits") ///
>         name(boxoffice, replace)
```

```
. graph export boxoffice.pdf, name(boxoffice) as(pdf) replace
(file boxoffice.pdf written in PDF format)

.
. // *********************************************************************
. // *********************************************************************
. // The lower whisker extends to the minimum value in this case (0). Because
. // there are outlier values at the top of the distribution, the upper whisker
. // extends to the upper adjacent value--the last value observed in the data
. // before the the threshold used to determine outliers (1.5 IQR above the
. // 75th percentile).
. // *********************************************************************
. // *********************************************************************
.
. // note, to see without outliers:
. graph box use_off, nooutsides

.
.
. // ********
. // Part 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? */
.
. graph box pcs12, over(highested) name(boxpcs, replace)

. graph export boxpcs.pdf, name(boxpcs) as(pdf) replace
(file boxpcs.pdf written in PDF format)

.
. // *********************************************************************
. // *********************************************************************
. // The figures show visually what was found in part (a)--the distribution
. // of PCS increases with educational attainment and becomes less variable.
. // *********************************************************************
. // *********************************************************************
.
.
. // ********
. // Part 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)?  */
.
. // *********************************************************************
. // *********************************************************************
. // The doctor's office visit distribution was clearly very positively
. // skewed. Most respondents had zero or very few visits, while a small
. // share of respondents had comparably very large numbers of office visits.
. // The PCS and MCS distributions appear negatively skewed. There is a long
. // tail (and some outlying values) toward the bottom of the distribution.
. // *********************************************************************
. // *********************************************************************
.
```

```
.
. // ********
. // Part 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.  */
.
. summ mcs12, detail

               Mental health component of SF12
-------------------------------------------------------------
      Percentiles      Smallest
 1%        19.57           1.35
 5%        30.31            1.9
10%        35.69           4.73       Obs              19,386
25%         44.3            6.4       Sum of Wgt.      19,386

50%        52.65                      Mean           50.22171
                        Largest       Std. Dev.      10.19464
75%        57.33          74.15
90%        61.12          74.81       Variance       103.9306
95%        62.49          74.84       Skewness      -.9985878
99%        65.56          75.06       Kurtosis       3.792512

. scalar a=r(skewness)

. scalar b=sqrt((6*r(N)*(r(N)-1))/((r(N)-2)*(r(N)+1)*(r(N)+3)))

. display a
-.99858779

. display b
.0175913

. display a/b
-56.766005

.
. summ pcs12, detail

              Physical health component of SF12
-------------------------------------------------------------
      Percentiles      Smallest
 1%           17           4.56
 5%        24.69           6.08
10%         31.5           6.08       Obs              19,386
25%        43.77           6.26       Sum of Wgt.      19,386

50%        52.99                      Mean           49.01453
                        Largest       Std. Dev.      11.01185
75%        56.71          70.87
90%        58.96          70.89       Variance       121.2607
95%        60.45           71.7       Skewness      -1.237738
99%        63.43          72.17       Kurtosis       3.862569

. scalar a=r(skewness)
```

```
. scalar b=sqrt((6*r(N)*(r(N)-1))/((r(N)-2)*(r(N)+1)*(r(N)+3)))

. display a
-1.2377382

. display b
.0175913

. display a/b
-70.360816

.
. summ use_off, detail

                # office-based provider visits
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs              19,386
25%            0              0       Sum of Wgt.      19,386

50%            2                      Mean           5.802383
                        Largest       Std. Dev.      10.86976
75%            7            164
90%           15            166       Variance       118.1518
95%           23            167       Skewness       5.549091
99%           51            187       Kurtosis       54.47084

. scalar a=r(skewness)

. scalar b=sqrt((6*r(N)*(r(N)-1))/((r(N)-2)*(r(N)+1)*(r(N)+3)))

. display a
5.5490914

. display b
.0175913

. display a/b
315.44522

.
. // ***********************************************************************
. // ***********************************************************************
. // In all three cases above, I divided 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 previous command. The rule of thumb
. // is that if this absolute value of the ratio is >2, the distribution is
. // significantly skewed. All three ratios exceed 2.
. // ***********************************************************************
. // ***********************************************************************
.
.
. // ********
. // Part 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). */
```

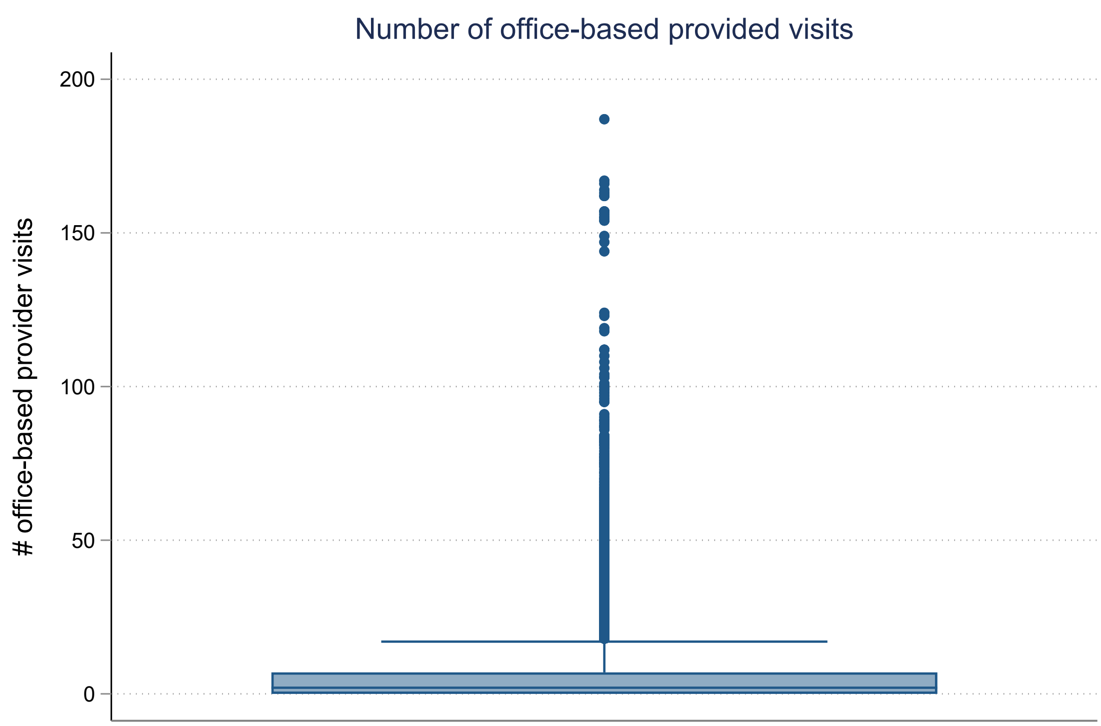
```
.
. gen lnoff=ln(use_off)
(5,673 missing values generated)

. histogram lnoff
(bin=41, start=0, width=.12758802)

. summ lnoff

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       lnoff |     13,713    1.490513    1.078619          0   5.231109

. count if use_off==0
  5,673

.
. summ lnoff, detail

                            lnoff
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs              13,713
25%     .6931472              0       Sum of Wgt.      13,713

50%     1.386294                      Mean           1.490513
                        Largest       Std. Dev.      1.078619
75%     2.302585       5.099866
90%     2.944439       5.111988       Variance       1.163419
95%     3.332205       5.117994       Skewness       .3161302
99%     4.060443       5.231109       Kurtosis        2.42947

. scalar a=r(skewness)

. scalar b=sqrt((6*r(N)*(r(N)-1))/((r(N)-2)*(r(N)+1)*(r(N)+3)))

. display a
.31613018

. display b
.02091519

. display a/b
15.114858

.
. // *******************************************************************
. // *******************************************************************
. // The distribution of the logged doctor's visits appears less skewed,
. // however, there are lots of zero values in the orginal variable and the
. // log transformation is not defined at zero.
. // *******************************************************************
. // *******************************************************************
.
.
. // ********
. // Part h
. // ********
```

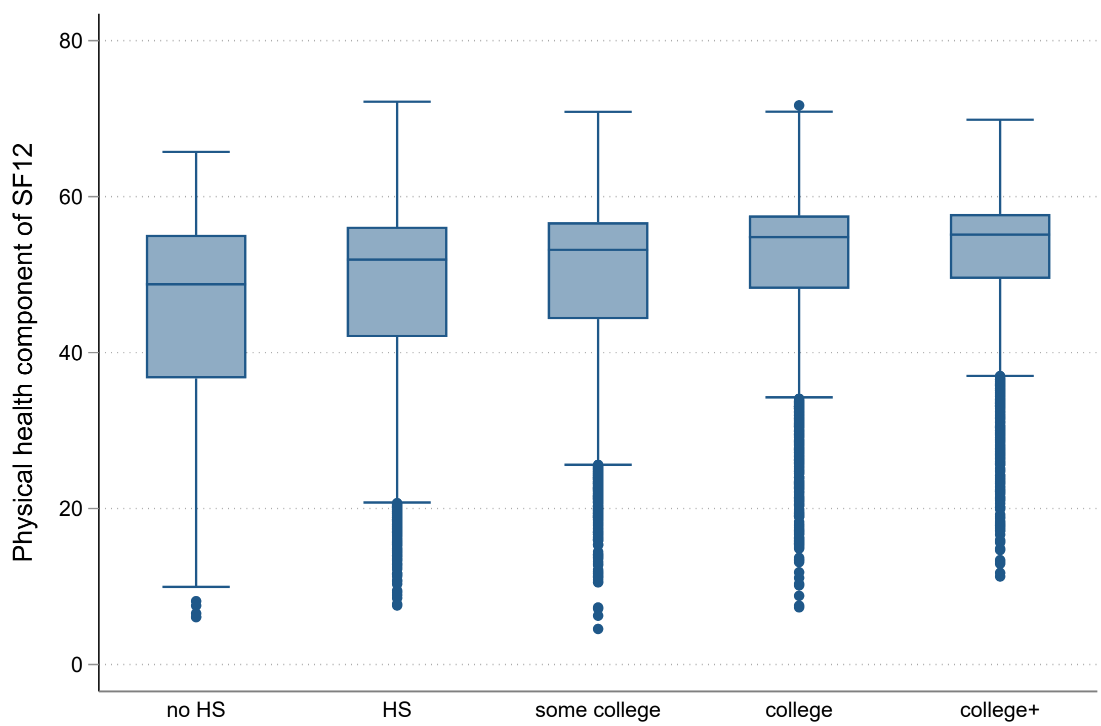
```
. // 5 POINTS
.
. /* You are considering doing a log transformation of the PCS variable to
> reduce the skewness. Would this help? Why or why not? (Try it and see what
> happens). */
.
. gen lnpcs=ln(pcs12)

. histogram pcs12, nodraw name(orig, replace)
(bin=42, start=4.5599999, width=1.6097619)

. histogram lnpcs, nodraw name(logged, replace)
(bin=42, start=1.5173227, width=.06575481)

. graph combine orig logged, row(1) ysize(4) xsize(6)

. graph export lnpcs.pdf, as(pdf) replace
(file lnpcs.pdf written in PDF format)

.
. summ lnpcs, detail

                            lnpcs
-------------------------------------------------------------
      Percentiles      Smallest
 1%     2.833213       1.517323
 5%     3.206398       1.805005
10%     3.449988       1.805005       Obs              19,386
25%     3.778949        1.83418       Sum of Wgt.      19,386

50%     3.970103                      Mean           3.857338
                        Largest       Std. Dev.      .2890377
75%     4.037951       4.260847
90%     4.076859       4.261129       Variance       .0835428
95%     4.101817       4.272491       Skewness      -2.048169
99%     4.149937       4.279025       Kurtosis       7.933831

. scalar a=r(skewness)

. scalar b=sqrt((6*r(N)*(r(N)-1))/((r(N)-2)*(r(N)+1)*(r(N)+3)))

. display a
-2.0481687

. display b
.0175913

. display a/b
-116.43078

.
. // *********************************************************************
. // *********************************************************************
. // The distribution of the logged PCS is more skewed than before!
. // We typically do log transformations to make a distributions less right-
. // skewed. This distribution was left skewed. Translating the original
. // variable (*-1) and adding a constant to get all values above 1 helps
. // (see below).
. // *********************************************************************
. // *********************************************************************
```

```
.
. qui summ pcs12

. gen lnpcs2=ln((-1*pcs12)+r(max)+1)

. histogram lnpcs2, nodraw name(logged2, replace)
(bin=42, start=0, width=.1006771)

. graph combine orig logged2, row(1) ysize(4) xsize(6)

. graph export lnpcs2.pdf, as(pdf) replace
(file lnpcs2.pdf written in PDF format)

.
.
. // ********
. // Part 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.*/
.
. egen zexp_tot=std(exp_tot)

. summ zexp_tot

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    zexp_tot |     19,386   -1.55e-09           1  -.3772595   44.71924

.
. // ********************************************************************
. // ********************************************************************
. // The mean of the z-score is indeed zero (or very close to it--there is a
. // small rounding difference) and the sd is 1.
. // ********************************************************************
. // ********************************************************************
.
.
. // ********
. // Part 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.*/
.
. summ exp_tot

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     exp_tot |     19,386     3685.25    9768.475          0     440524

. display r(mean) + 0.2*r(sd)
5638.9447

. display r(mean) - 0.2*r(sd)
1731.5548
```

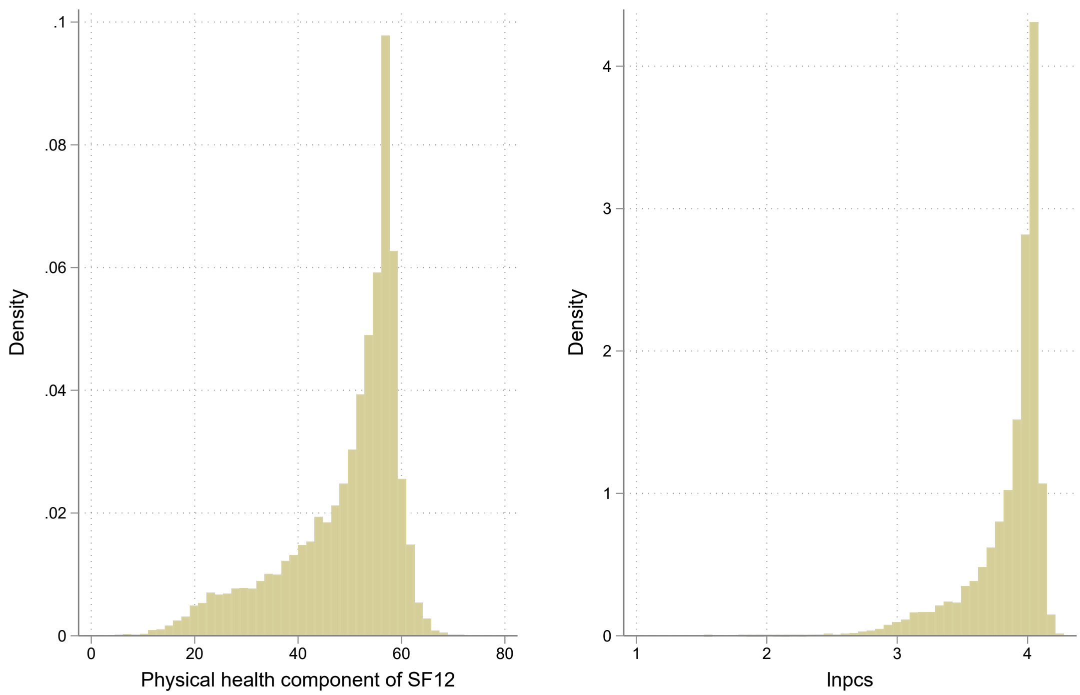
```
.
. // ********************************************************************
. // ********************************************************************
. // Results are above. $5,638 is 0.2 standard deviations above the mean.
. // $1,732 is 0.2 standard deviations below the mean.
. // ********************************************************************
. // ********************************************************************
.
.
. // ********
. // Part 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?*/
.
. count if zexp_tot>=-1 & zexp_tot<=1
  18,237

. scalar a = r(N)

. count if zexp_tot~=.
  19,386

. scalar b = r(N)

. display a/b
.94073042

.
. // ********************************************************************
. // ********************************************************************
. // Results are above. 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 94.7%.
. // The Empirical Rule applies to **normal** distributions, which this is not.
. // (This distribution is in fact very skewed).
. // ********************************************************************
. // ********************************************************************
.
.
. // ********
. // Part l
. // ********
. // 5 POINTS
.
. /* What is the 43rd percentile for total medical expenses (exp_tot)? Explain/
> show how you got your answer.*/
.
. centile exp_tot, centile(43)

                                                       -- Binom. Interp. --
    Variable |       Obs  Percentile    Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
     exp_tot |    19,386         43      622.41             597         651

.
. // ********************************************************************
```

```
. // ***********************************************************************
. // Results are above using the centile command. As an alternative, we can
. // use the index values of the observations as shown below.
. // ***********************************************************************
. // ***********************************************************************
.
. // we want the observation indexed as follows. It's a fractional value so we
. // take the next highest observation.
. display (43*r(N))/100
8335.98

.
. sort exp_tot

. list exp_tot if _n==8336

       +---------+
       | exp_tot |
       |---------|
 8336. |     622 |
       +---------+

.
.
. // Close log and convert to PDF
. log close
```

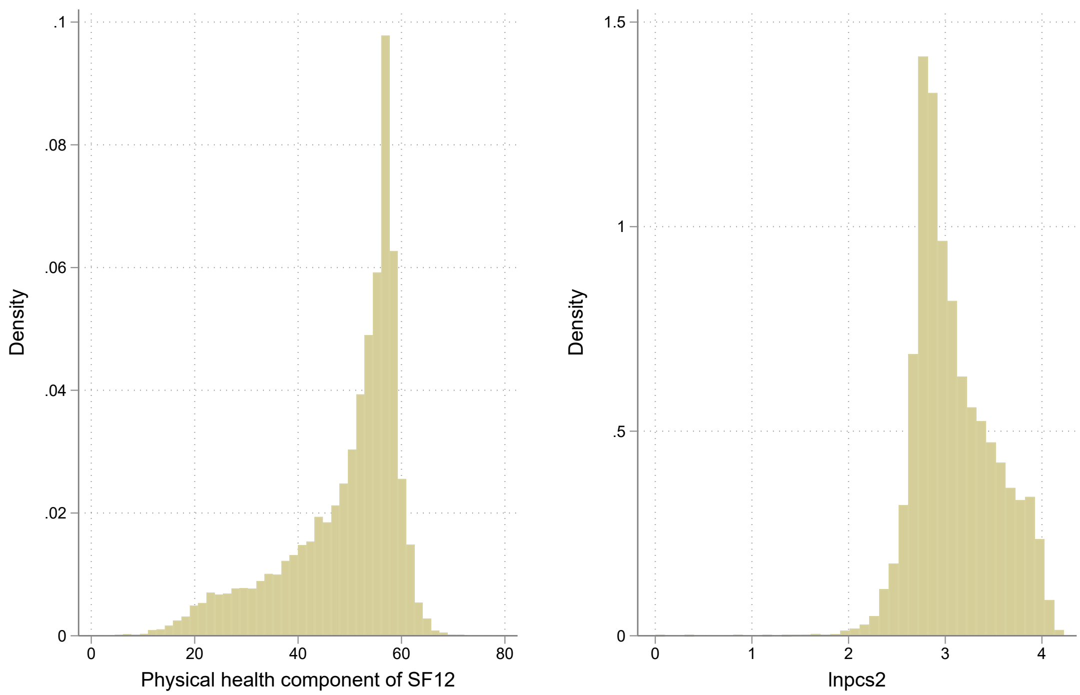

Number of office-based provided visits
office-based provider visits
200
150
100
50
0



Physical health component of SF12
80
60
40
20
0
no HS
HS
some college
college
college+



Density
.1
.08
.06
.04
.02
0
0
20
40
60
80
Physical health component of SF12
Density
4
3
2
1
0
1
2
3
4
lnpcs



Density
.1
.08
.06
.04
.02
0
0
20
40
60
80
Physical health component of SF12
Density
1.5
1
.5
0
0
1
2
3
4
lnpcs2