ECN 102: Analysis of Economics Data

Homework 1 Answer Key

Remy Beauregard

Question 1: Data types

Give an example of the following types of data:

a) Observational, categorical, time-series data **What is my most-visited coffee shop each month for 12 months?**

b) Observational, numerical, panel data **How many times does each student in 102 visit the Canvas page each month throughout the quarter?**

c) Experimental, discrete numerical, cross-sectional data **How many cases of COVID do treated individuals in a vaccine trial develop relative to a placebo group?**

d) Experimental, categorical, repeated cross-sectional data **Which of 5 new flavors of a drink do multiple waves of subjects prefer in a brand experiment?**

What type of data are the following:

e) Midterm 1 scores for one class of ECN 102 **observational, continuous numerical, cross-sectional**

f) Final grades (%) for ECN 102 from the last 5 years **observational, continuous numerical, repeated cross-section**

g) All homework grades (graded pass/fail) for one student in ECN 102 this quarter **observational, categorical, time-series data**

h) All homework grades (graded pass/fail) for all students in ECN 102 this quarter **observational, categorical, panel**

Question 2: Univariate data

For the given sample 0, 4, 5, 2, 3, 2, 11, 17, 6:

a) Calculate the mean, median, mode, variance, and standard deviation from first principles $\bar{x} = 5.\bar{5}, Median = 4, Mode = 2, s^2 = 28.3, s = 5.3$

b) Without using the formula, is this data symmetric or skewed? How do you know? Is it normally distributed? **Since $\bar{x} > Median$, the data is right-skewed and not symmetric. This means it cannot be normal.**

c) How would your above calculations change if we subtracted 7 from each observation? If we multiplied by 2 and *then* subtracted 7? How do these differ? **Addition and subtraction does not affect our measures of spread and would shift our measures of central tendency down by 7. Multiplying by two and subtracting 7 would shift our measures of spread by 2^2 (s^2) or 2 (s) and our measures of central tendency by $2x - 7$.**

d) Compute z-scores for each observation. What is the mean of these z-scores? What is the standard deviation? Does this surprise you? $z_i = \frac{x_i - 5.\bar{5}}{5.3}$; $z \sim (0, 1)$ **as we expect.**

e) Do you expect these z-scores to be normally distributed? Why or why not? **No. As our underlying sample is not normally distributed, standardizing our data will not make it become normal.**

Question 3: Manual data entry

Use the following code to input *some* data into your Stata browser (it does not have to match my data) and compute summary statistics.

```
input myvar // this can be any name you wish
0
3
5
1
4
0
9
3
1
4
5
end
summarize myvar, detail
```

```
                             myvar
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              1       Obs                  11
25%            1              1       Sum of wgt.          11

50%            3                      Mean           3.181818
                        Largest       Std. dev.      2.676497
75%            5              4
90%            5              5       Variance       7.163636
```

```
95%            9              5        Skewness       .6829952
99%            9              9        Kurtosis       3.041582
```

a) What is the IQR of your data? What is the skewness and kurtosis? What do these values mean? **IQR = 75%ile-25%ile. Skewness > 0 is right-skewed, < 0 is left-skewed. Kurtosis > 3 means fatter tails than the normal; < 3 means thinner tails.**

b) Obtain a table of frequencies for your data using `tabulate [varname]`.

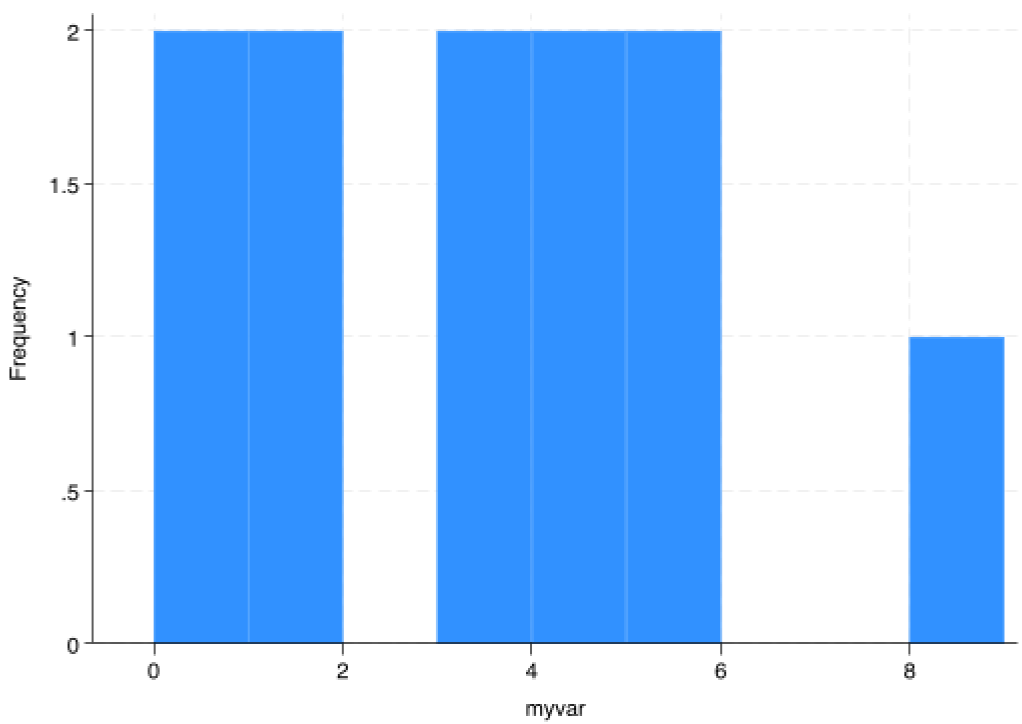
```
tab myvar
```

```
      myvar |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          2       18.18       18.18
          1 |          2       18.18       36.36
          3 |          2       18.18       54.55
          4 |          2       18.18       72.73
          5 |          2       18.18       90.91
          9 |          1        9.09      100.00
------------+-----------------------------------
      Total |         11      100.00
```

c) Give a histogram of the data with bin width one using `histogram [varname], width(1) frequency`.

```
histogram myvar, w(1) freq
```

```
(bin=9, start=0, width=1)
```



Question 4: Data in Stata

Download AED_CALELECTRICITY.DTA from the website above and bring it into Stata.

a) Describe the data using `describe` - what variables are here? What do they correspond to?

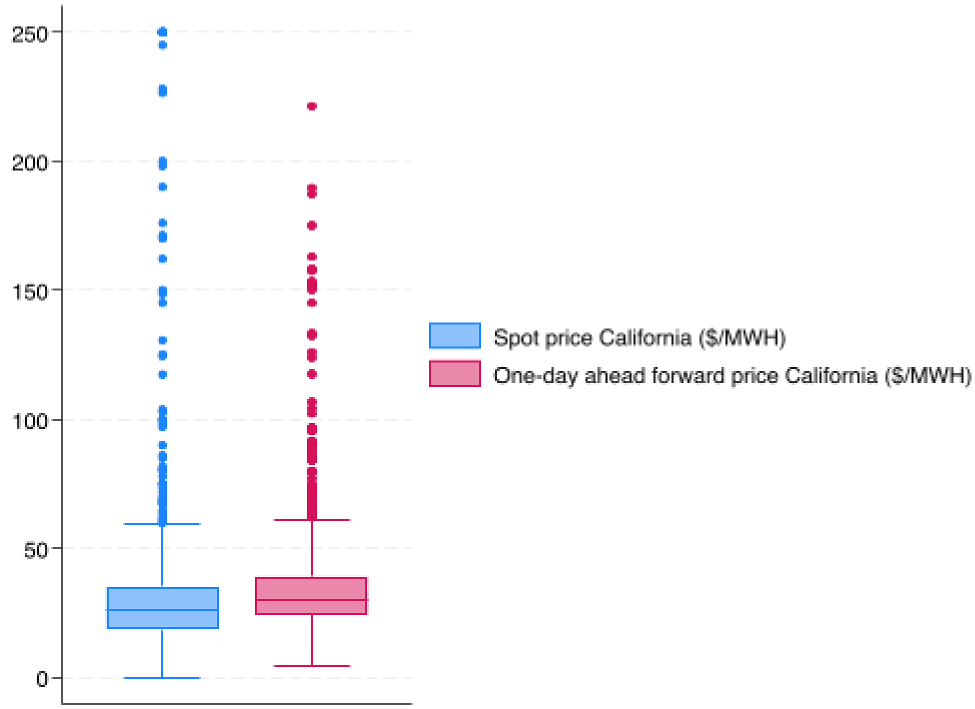
```
qui use AED_CALELECTRICITY, clear
describe
```

```
Contains data from AED_CALELECTRICITY.dta
 Observations:           682                  Data for A. Colin Cameron
                                                (2015): Analysis of Economics
                                                Data, W.W. Norton
    Variables:             7                  2 Mar 2015 20:17
--------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
--------------------------------------------------------------------------------
month           byte    %8.0g                 Month of year number
year            int     %8.0g                 Year
day             byte    %8.0g                 Day of month
hour            byte    %8.0g                 Hour of day (24 hour clock)
npx             float   %9.0g                 One-day ahead forward price
                                                California ($/MWH)
niso            float   %9.0g                 Spot price California ($/MWH)
diff            float   %9.0g                 niso - npx
--------------------------------------------------------------------------------
Sorted by:
```

b) Obtain a box plot for both the spot price and one-day ahead forward price of electricity using *graph box [varnames]* and save it with `graph export myboxplot.png, replace`.

```
graph box niso npx
graph export myboxplot.png, replace
```

```
file myboxplot.png written in PNG format
```



c) Obtain summary statistics for both series. Which has a higher degree of dispersion? Are they symmetric or skewed?

```
summarize niso npx, detail
```

```
                 Spot price California ($/MWH)
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%      4.99753              0
10%      9.37237              0       Obs                 682
25%        18.67              0       Sum of wgt.         682

50%     26.03347                      Mean           37.46019
                        Largest       Std. dev.      45.81533
75%     35.00999            250
90%     60.00001            250       Variance       2099.044
95%     124.3951            250       Skewness       3.542202
99%          250            250       Kurtosis       15.79749

       One-day ahead forward price California ($/MWH)
-------------------------------------------------------------
      Percentiles      Smallest
 1%        6.507         4.0046
 5%      15.4918         4.4966
10%        19.99         4.9313       Obs                 681
```

```
25%      24.1723         4.9316       Sum of wgt.         681

50%       29.788                      Mean           37.22629
                        Largest       Std. dev.      27.09195
75%        38.99         174.99
90%        60.01       187.1325       Variance       733.9737
95%        87.93       189.6182       Skewness       3.171462
99%       157.99         221.27       Kurtosis       15.13346
```

To compare dispersion, we should compare the coefficient of variation, $CV = \frac{s}{\bar{x}}$.

$CV_{niso} = \frac{45.8}{37.5} > CV_{npx} = \frac{27.1}{37.2}$, **so the spot price has higher dispersion.**

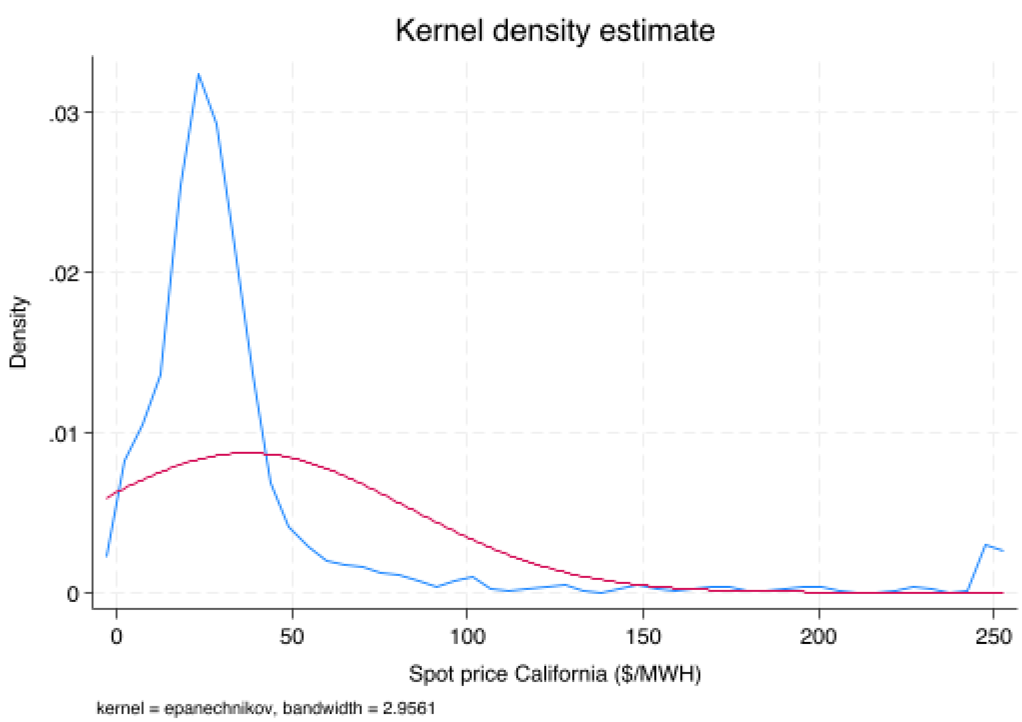
d) Log transform the spot price variable and create a new variable *ln_niso* using `generate ln_niso = ln(niso)`. Label this variable "Log of spot price California" using `label variable ln_niso [VARLABEL]`

```
gen ln_niso = ln(niso)
la var ln_niso "Log of spot price California"
```

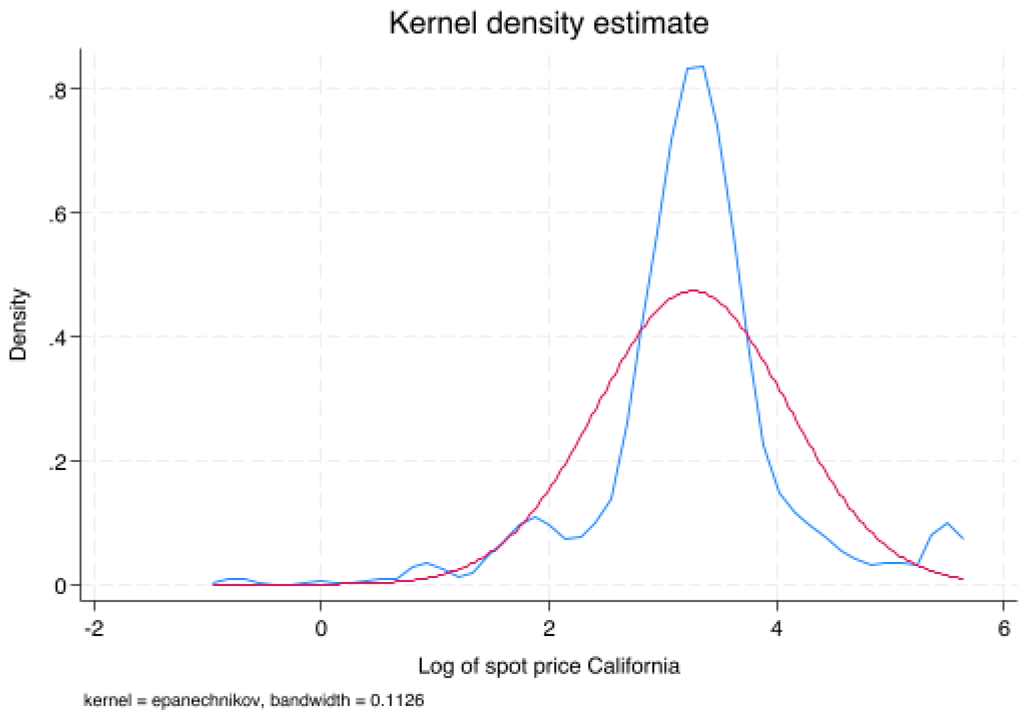
```
(9 missing values generated)
```

e) Plot the kernel density functions of both *niso* and *ln_niso* (in separate graphs) against a normal density function with `kdensity [varname], normal`. How does the shape of the distribution change? Is this surprising? Which appears more normal?

```
kdensity niso, normal legend(off)
```



```
kdensity ln_niso, normal legend(off)
```

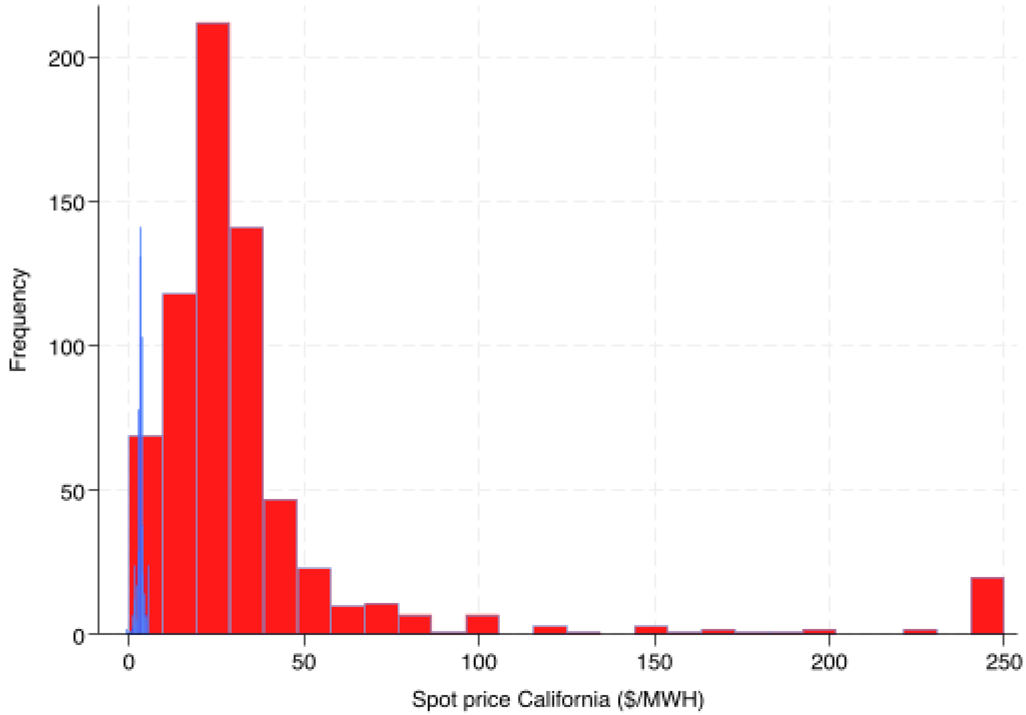


The data becomes less right-skewed and more normal with the log transformation, as we expect. Log transformations can be used to somewhat normalize right-skewed data.

f) Plot a histogram of *niso* and *ln_niso* together on the same graph with the code below. Does this look strange? Why might this be?

```
histogram niso, freq legend(off) fcolor(red) addplot(hist ln_niso, freq fcolor(blue))
```

```
(bin=26, start=0, width=9.6153858)
```



This looks strange, as the data are on different scales. One is in levels, one is in logs. This is not a good histogram for us to show!