Homework 3: ECO220Y – SOLUTIONS

Reminder: The Readings page in Quercus has complete solutions to textbook exercises.

Required Problems:

(1) (a) It is positively skewed and unimodal.

(b) Y = 10 + X. The shape remains positively skewed and unimodal. The mean increases by 10 and the s.d. is unchanged. This gives everyone the same boost regardless of their performance: high performers may feel that this is unfair. This adjustment does not change the variability across employees.

(c) Y = 1.5*X. The shape remains positively skewed and unimodal. The mean increases by 50 percent and the s.d. increases by 50 percent. This gives the highest performers the biggest boost: low performers may feel that this is unfair. This adjustment greatly increases the variability across employees.

(d) Y = 1.25*(X + 5). The shape remains positively skewed and unimodal. The mean increases overall by 11.8 points (combination of unit increase and percentage increase) and the s.d. increases by 25 percent. This is a hybrid of the previous methods, but fairness is a subjective concept.

(e) All of the adjustments considered are examples of linear transformations. Yes, changes in units of measurement are examples of linear transformations. This is good news: it means you do not have to worry that your inferences about the shape of the distribution will be influenced by the units of measurement you choose when drawing the histogram.

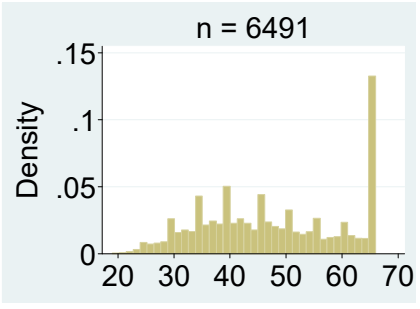
(2) We can approximate:

frac_read	Freq.	Percent	Cum.
0	19	13.10	13.10
.25	6	4.14	17.24
.5	19	13.10	30.34
.75	45	31.03	61.38
1	56	38.62	100.00
Total	145	100.00	

This gives a mean of about 0.69 and a standard deviation of 0.34. (Review Section 5.7 of your textbook if you are having trouble with this.) This means that on average the class completed 69 percent of the reading with a s.d. of 34 percentage points (a lot of variation). However, this distribution is clearly not symmetric so we may consider the median and IQR as summary statistics. The median is about 0.75 and the IQR is about 0.5 (the 75th percentile is 1 and the 25th percentile is 0.5). This means that that median student completed about 75 percent of the reading and a student at the 75th percentile completed 50 percentage points more of the reading than a student at the 25th percentile.

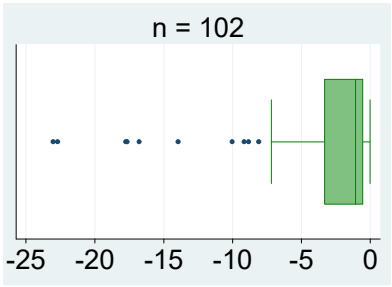
(3) Mean is 0.6572, median is 0.7, mode is 1, 25th percentile is 0.4, 75th percentile is 1, IQR is 0.6, range is 1. Show work.

(4) These data are unusual in that there are a large number of people reporting an age of 66: at least 10% of the sample. Between 5% and 10% of the sample are 30 years old or less: hence, somewhere between 325 and 649 people are 30 years old or less. One person is 19 years old. We know that there are at least 3 people and less than 65 people who are 20 years old. If it were normal then the Empirical Rule should hold: roughly 68% of the data should be between 35 and 60 and 95% should be between 23 and 72. This is a large sample size, 6,491, which means that large deviations from expectation cannot be explained by sampling noise. Hence we



have sufficient evidence to infer that the population that this sample is taken from is not normal. Here is the histogram of these data, which also supports this inference:

(5) It is extremely negatively skewed and unimodal. Also, all values are negative and range from just below zero to nearly 25 below zero.



(6) (a) Mean is 5.332, s.d. is 2.119, and the IQR is 3. (This is an application of the concepts in Chapter 5 (and Lecture 3). If you are having trouble, review those.)

(b) Not quite. This distribution has some features in common with the Normal distribution but it is *not* Normal (Bell shaped) and it is discrete (integers only). Notice that the figure is based on an extremely large sample size so we *cannot* say that it is probably just sampling error that is causing the shape to diverge from the Normal shape.

(c) One potential non-sampling error is that the target population is all adults but the sampled population will be only those that can be contacted (e.g. by phone): it would likely exclude institutionalized people (e.g. prison, old age, etc.) and the homeless. (And the happiness of these people likely differs substantially from the rest of the population.) Another non-sampling error would be non-response bias: it is very hard to get people to respond to surveys and those that choose to respond are not a random subset of those invited. There are other possible non-sampling errors as well.

(7) We can see that 14.04th percentile is 25. Hence we can approximate that the 10th percentile is somewhere below a debt-to-GDP of 25% (but above 0%). Similarly the 15th percentile will be above a debt-to-GDP of 25% and likely just above that. The median should mark the point where half the area is above and half below. The first two bars include 52.64% of our sample of 57 countries. Hence the median is somewhere between a debt-to-GDP of 25% to 50% and likely closest to 50%. The 75th percentile will be somewhere between a debt-to-GDP of 75% to 100% (and likely closest to 75%). Finally we can find the 96.5th percentile exactly: it is a debt-to-GDP of 125%.

(8) (a) These are time series data. The unit of observation is an hour.

(b) No, we'd need to look at a time series plot. Review Section 5.13 of the textbook (pp. 115 – 118).

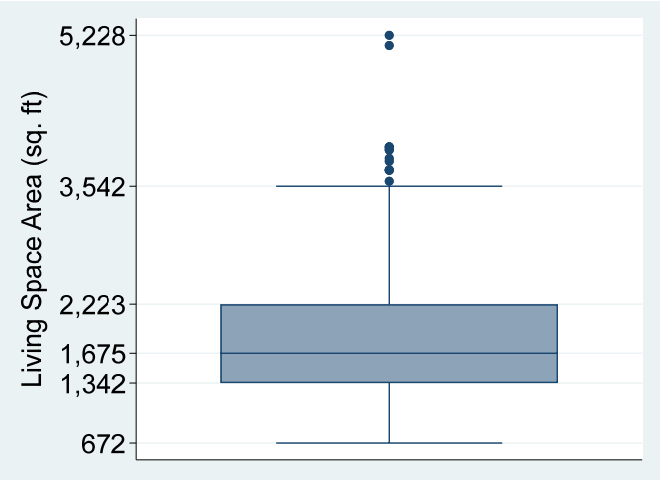
(c) Statement **(A)** is FALSE.

(9) A z-score of 0 means that you earn exactly the class average mark average; in other words, your score is zero standard deviations from the mean. Your friend Wei earned a mark that is 2.15 standard deviations below the class average (not good). Your friend Tina earned a mark is 1.86 standard deviations above the class average (good). Your percentage mark is 69% (= 69 + 0*12), Wei's is 43% (= 69 - 2.15*12), and Tina's is 91% (= 69 + 1.86*12).

(10) Activity.

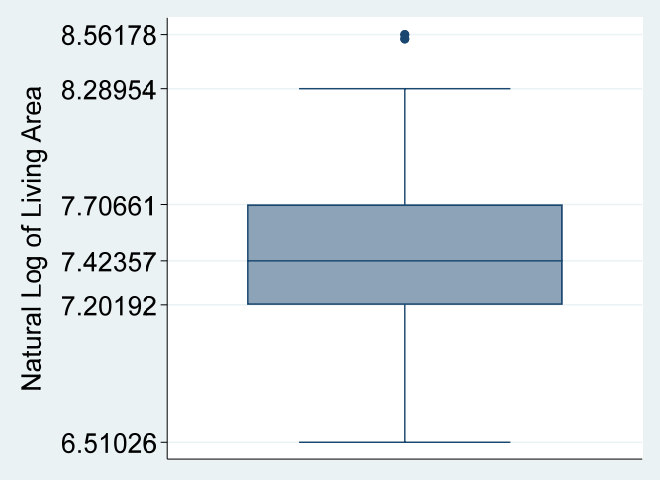
(11) (a) Yes, they match up. Notice that the textbook gives a frequency histogram whereas the STATA histogram is a density histogram. Also, there is a slight difference in the number of bins. However, both histograms give a good summary of the distribution of the living space variable.

(b)



(c) You can perfectly predict all of the percentiles and the smallest and largest values. These are all measures of relative standing and are unchanged by a monotonic transformation (like the natural log). For example, if the 10th percentile of living space area is 1,056 square feet then we know that after the natural log transformation the 10th percentile will still be the same house. If it has a size of 1,056 square feet then the natural log is 6.96 $(= ln(1{,}056))$. However, the mean and standard deviation cannot be predicted. This comes down to basic math: $\ln(X + Y) \neq \ln(X) + \ln(Y)$. (Both the mean and s.d. involve sums.) (Note that part (d) provides the STATA output for the natural log of living area.)

(d)



(e) As expected, the histogram and box plot are both much more symmetric after the natural log transformation because the natural log reins in the big numbers. In other words, notice how much closer the biggest house of 5,228 square feet is to the median after the natural log transformation compared to before it: it was more than three times as big as the median house in the original variable, but only 15% bigger than the median house after the transformation.

(12) Figure 3 illustrates how the rates of ownership of consumer durables, like cars, widescreen TVs, and refrigerators, have changed between 1985 and 2012 for the highest income Americans versus the lowest income Americans. For example, refrigerator ownership among people in the top income decile – incomes above the 90th percentile – has been high throughout the period, increasing from 99% ownership to 100% ownership. In contrast, those in the lowest income decile – incomes below the 10th percentile – have made considerable progress and almost closed the gap by income: in 1985 about 92% owned a refrigerator but by 2012 nearly 99% percent did. A similar trend exists for cooking durables. However, a continuing gap exists for dishwasher, washers/dryers, and vehicles. For example, over 90% of the richest Americans own a car and that has been steady between 1985 and 2012. The poorest Americans have made only modest

gains going from about 66% vehicle ownership in 1985 to about 72% in 2012. One special category of durables is entertainment – televisions, computers, etc. – where there is very little difference in ownership rates between the richest and poorest: both have sharply increased such that by 2012 nearly 100% of households own some entertainment durable, even among those below the 10th percentile in the income distribution. [This is my summary. However, you may also view the original paper for the authors' summary: http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.30.2.3.]