ECO220Y: Homework 2 -- SOLUTIONS

(See Appendix A of the textbook for answers to the exercises.)

Required Problems:

(1) It would sum to the number of observations (i.e. the sample size, n). It would sum to one. It would sum to one.

(2) (a) 100*(5 + 1)/78 = 7.7%

(b) ≈100*(0.06 + 0.015) = 7.5% (approximate because we roughly estimate the height)

(c) ≈ 100*(2*0.035 + 2*0.005) = 8.0% (approximate because we roughly estimate the height)

(d) As we discussed in lecture, the underlying data in this example are the same and the only difference is the type of histogram that is used. Of course the percent that work 12 hours or more does *not* change depending on the type of histogram. Any differences in our answers above are simply because we had to visually approximate using the graphs. (In this case because employees must be integers we can be confident in the exactness of our calculation in Part **(a)**.)

(3) Write up your own explanation using the idea of sampling error. (Recall the sampling experiment in class where a student drew little slips of paper out of a box: the same concept applies here. Instead of pieces of paper, we have employees. Instead of shapes, we have hours of work.)

(4) The high precision with which hours have been measured means that there are a lot of unique values: virtually every value except zero is unique. This makes the tabulation a poor summary of these data. By grouping data into bins the histogram, in a loose sense, "rounds" the data and lets us easily see the overall distribution of hours of work. In contrast the tabulation is listing each unique value separately and obscuring the larger picture. The only advantage of the tabulation over the histogram is that with the tabulation we can see exactly how many of the workers in our sample of 78 worked zero hours (7 workers).

(5) A histogram is used to summarize the distribution of interval data. Because of this, the horizontal axis in a histogram has a clear meaning: it is the real number line. A bar chart is used to summarize the relative frequency of values in ordinal or nominal data. Because of this, the horizontal axis in a bar chart is NOT the real number line: for example, you could put the categories in any order and the distance between categories has no meaning. Given this, it does not make sense to talk about the shape of a bar chart whereas it does make sense to talk about the shape of histogram (in fact, we spent a lot of time on this talking about symmetry, skewness, modality, etc). The vertical axis could be the same for a histogram and bar chart if we were talking about a frequency or relative frequency histogram. However, while a density histogram makes sense, a "density bar chart" does not.

(6) Thinking of the quantitative (interval) variable that records the number of defaults and that takes three unique values (0, 1, and 2), we need to use $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}}$. First, find $\bar{X}$ using $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{202*0+97*1+38*2}{337} = 0.513353$. Next, find $s = \sqrt{\frac{202*(0-0.513353)^2+97*(1-0.513353)^2+38*(2-0.513353)^2}{337-1}} = \sqrt{\frac{160.1899}{336}} = 0.69$.

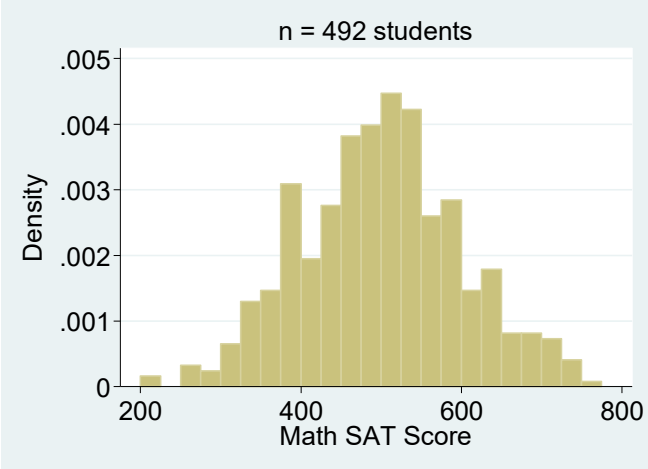
(7) Given that all three have the same sample mean and sample range, we can focus on the shape of the distribution to answer. Sample 1 has the smallest sample standard deviation and Sample 3 has the largest sample standard deviation. You should explain the intuition.

(8) If the variance were 5 then the standard deviation would be 2.24. We cannot use the Empirical Rule because this sample clearly comes from a non-Normal population. However, we can always use Chebysheff's theorem no matter how the population is distributed. According to Chebysheff's theorem at least 75% of the observations should be within

2 standard deviations of the mean. The mean in this case appears to be about 40. $40 \pm 2*2.24 \approx (35.5, 44.5)$. It is clear from the histogram that far fewer than 75% of the observations are in the interval (35.5, 44.5). However, <u>at least</u> 75% should be within 2 standard deviations of the mean. Hence, we can conclude that the variance of these data cannot possibly be 5. The variance must be greater than 5.

(9) (a) Recognize (roughly) Normal, use Empirical Rule. [In fact, I took this sample from a perfectly Normal population with mean 500 and s.d. 100.] Should get mean around 500 points and s.d. around 100 points by applying Empirical Rule. State at least part of that rule in support of your work.

(b) See that the bin width is 25 (e.g. there are 8 bins between 400 and 600 so width is 25). Can see that the fraction that fall in the bin just below 400 is about 0.075. Hence, the *area* of the bar in a density histogram should be about 0.075. Given that the bin width is 25 the height should be about 0.003.



(c) We would expect it to be negatively skewed because a top university will have lots of high scoring students (but max. possible bounded at 800). However, some will likely have relatively low scores (a left tail). It would certainly have a higher mean and likely a lower s.d.

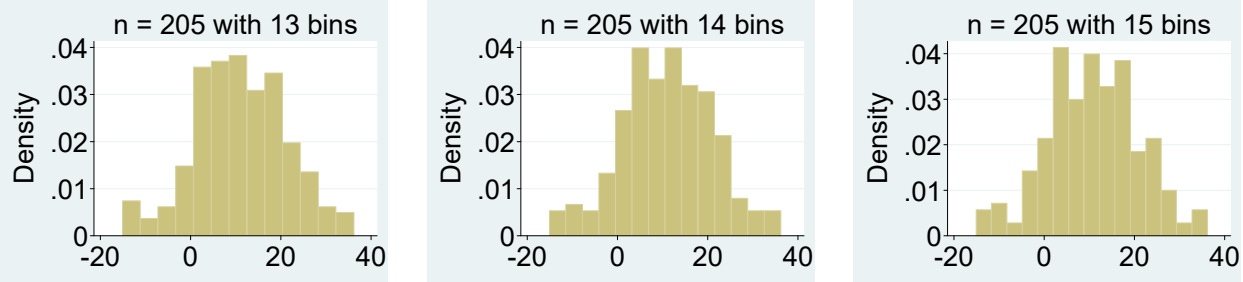
(10) (a) Activity.

(b) All four histograms showing the endowment distribution across universities (i.e. the bottom four histograms) are extremely positively skewed, which includes the last two graphs with attempt to cut off the long right tail but still leave very positively skewed distributions. The median endowments will be much smaller than the mean endowments (which are pulled up by the small number of very rich universities). While some of the % change in endowment graphs (i.e. the top four histograms) may appear to have something in common with the Normal distribution: they are NOT Normal. The tails are too thick. The first graph shows the problem in the extreme: if you go plus or minus one standard deviation about the mean you will get WAY more than about 68.3% of the universities. If it were Normal then the Empirical Rule should (at least approximately) hold.

Extra Problems:

(11) Symmetric, bell shaped, unimodal. Resist the temptation to call this bimodal or multi-modal: it is not. When we are describing the shape we see in a histogram we are trying to make inferences about the distribution of the population. However, the histogram is based on a SAMPLE: samples are subject to sampling noise. We do NOT want our conclusions to be driven by sampling noise. This means that we should look for major trends in the histogram and not get worked up about small deviations from the ideal bell shape: there will often be small deviations due to sampling noise. We will never see a perfect bell shape in a sample even if the sample is taken from a perfect bell shaped population. Modality refers to MAJOR PEAKS: the reason for the word "major" is to avoid counting up every small peak that can happen with sampling noise. In fact if we changed the number of bins slightly to 13 or to 15 instead of 14 (as in the original graph) the

exact same data would yield the histograms below. Notice how these tiny and subjective changes create minor peaks in arbitrary locations. In contrast if it had been a major peak then these small changes in the how we draw the histogram would not have affected it much at all.



Exactly 24 (11.7%) of the observations are less than 0, but you can't figure that exact number out with the given information. Just eye-balling the picture, you should get an answer between 10 – 37 observations (roughly 5% – 18%). If you are outside that range, you do not understand the meaning of the histogram.

(12) Positively skewed, unimodal. According to this graph, there are no negative observations in the sample (the first bin starts at zero). We cannot tell how many zeros there are in the data: there could be anywhere from no zeros in these data to about half zeros (up to about 0.41*1.25*62 = 32 observations).

(13) (a) The population is perfectly symmetric. Hence exactly half the data is above the mid-point (1) and exactly half is below the mid-point (1). The mean will also be at the mid-point because in averaging there is always an observation above average to exactly cancel out each observation below average. This is true for all symmetric populations but not all populations in general.

(b) The sample median is not exactly equal to the sample mean because of sampling error (pure chance). Even though this sample is taken from a population where the mean and median are equal the sample will differ from the population because of sampling error. In this relatively large sample of 500 observations we see that there is not too much sampling error (as expected) and the sample mean and median are in fact very close to each other.

(c) It is symmetric and unimodal. If you are thinking Bell shaped (Normal) that is not too far off but given the large sample of 500 we can see this is not quite Bell shaped (Bell shaped has longer and thinner tails and more of a hump in the center). This looks like a triangle shape. However, if we had a smaller sample size of say 50 there is no way we could tell the difference between a triangle shape and a Bell shape. There would be too much sampling error in a smaller sample size to make such a fine distinction.

(14) The sample mean is about 150. It is reasonable to infer that this sample has been taken from a Normal (bell shaped) population, which means that we can use the Empirical Rule as a guide. Hence, the sample standard deviation is about 15.