MONT 105Q – Mathematical Journeys
Problem Set 1 – Basic Descriptive Statistics
Due: Friday, February 19

Directions

Do the following problems and prepare individual writeups to hand in at the start of class on February 19.

I. Consider the following data set (list of numbers):

0.7, 1.6, 9.8, 3.2, 5.4, 0.8, 7.7, 6.3, 2.2, 4.1
8.1, 6.5, 3.7, 0.6, 6.9, 9.9, 8.8, 3.1, 5.7, 9.1

(A) What are the mean (μ) and median of this data set?

(B) Suppose the last number 9.1 was altered to 15.3. *Without doing any computations*, say what would happen to the mean and the median. Does each of them increase, decrease, or stay the same?

(C) What is the SD (standard deviation, or σ) of this data set?

(D) To what extent is the pattern from the diagram on page 26 of *Naked Statistics* true for this data set? Count how many of the numbers in the list are in each of the intervals $[\mu-2\sigma, \mu-\sigma)$, $[\mu-\sigma, \mu)$, $[\mu, \mu+\sigma)$, $[\mu+\sigma, \mu+2\sigma]$ and compute the percent of the numbers in the whole list that lie in each interval. We never expect exact agreement with something like the percentages from page 26, but are the actual percentages even in the right ballpark for this data set? Does this look like a random sample from a population described by a normal distribution? Explain.

(E) Another way to understand the distribution of a data set like this is to draw a diagram called a *histogram*. There are several possible choices statisticians might make to do this. Follow these directions to produce first a *frequency histogram*, and then a *standardized histogram* with total area 1:

(1) The numbers in the data set are all contained in the range $[0, 10]$. Suppose we divide this interval into 5 "bins" $[0, 2), [2, 4), [4, 6), [6, 8), [8, 10]$. How many numbers from the data set fall into each of the bins? (By the way, there's nothing magic about using 5 bins. Any number is possible, but larger numbers of bins tend to spread out the data too much and produce less informative histograms, so 5 or 6 is a common choice! Also, to avoid having to make awkward choices, it's usually good to choose bin boundaries so that no numbers from the data set cooincide with them. Note that that worked out fine for us here.)

(2) Draw perpendicular x- and y-axes to make a coordinate system and plot the 5 "bins" along the x-axis. Then make rectangular boxes over each of those bins where the height

gives the number of values from the data set in the corresponding bin. This is the *frequency histogram*.

(3) Then determine a vertical scaling factor λ (a number) with the property that multiplying each height from the *frequency histogram* by λ will yield a collection of boxes whose *areas add up to 1*. Redraw the histogram with the new scale on the y-axis showing this standardized version of the histogram. (The shape should be exactly the same, only the vertical scale changes!)

II. A study on college students showed that the 150 men studied had an average mass of 66kg, with an SD of 9kg. The 200 women studied had an average mass of 55kg, with an SD of 9kg.

(A) About how many of the 150 men would we expect to have masses between 57kg and 75kg?

(B) If you took the men and women together would the SD for the whole group of 350 students be exactly 9kg, larger than 9kg, or smaller than 9kg? Why?