Name: *Caleb McWhorter — Solutions*
MATH 100
Fall 2023
HW 13: Due 12/06

"When a theory is generated by ransacking data, we can't use these pillaged data to test the theory."
–Gary Smith

Problem 1. (10pt) Consider the following dataset:

$$318.8,\quad 179.7,\quad 2111.3,\quad 395.7,\quad 412.1,\quad 371,\quad 78.8,\quad 146.3,\quad 231.3,\quad 412,\quad 220.4$$

Showing all your work, complete the following:

(a) Find the median of this dataset.

(b) Find the 40th percentile for this dataset.

(c) Find the 5-number summary for this dataset.

(d) Sketch a box-plot for this dataset.

Solution. We first order the data:

$$78.8,\quad 146.3,\quad 179.7,\quad 220.4,\quad 231.3,\quad 318.8,\quad 371,\quad 395.7,\quad 412,\quad 412.1,\quad 2111.3$$

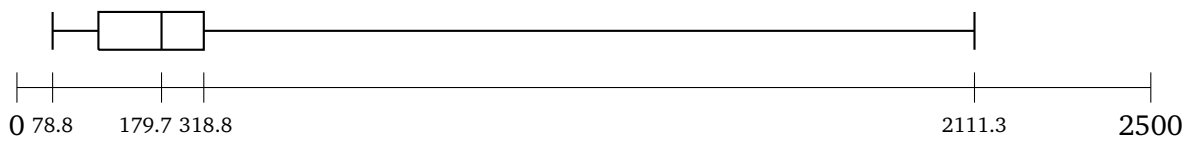
(a) There are 11 numbers in this dataset. Therefore, the median is the $\frac{11}{2} = 5.5 \rightsquigarrow 6$th number in the dataset. Therefore, the median is 318.8.

(b) The 40th percentile will be given by the $L = \frac{40}{100} \cdot 11 = 4.4 \rightsquigarrow 5$th number in the dataset. Therefore, the 40th percentile is 231.3.

(c) Clearly, the minimum of the dataset is 78.8 and the maximum is 2111.3. From (a), we know that the median is 318.8. We need only find the first and third quartile, i.e. the 25th and 75th percentile, respectively. The 25th percentile is given by the $L = \frac{25}{100} \cdot 11 = 2.75 \rightsquigarrow 3$rd value in the dataset and the 75th percentile is given by the $L = \frac{75}{100} \cdot 11 = 8.25 \rightsquigarrow 9$th value in the dataset. But then the first quartile is 179.7 and the third quartile is 412. Therefore, the 5-number summary is...

Min	Q_1	Median	Q_3	Max
78.8	179.7	318.8	412	2111.3

(d)



Problem 2. (10pt) Consider the following dataset:

$$23, \qquad 38, \qquad 17, \qquad 20, \qquad 26$$

Showing all your work, complete the following:

(a) Find the mean of for this dataset.

(b) Find the standard deviation for this dataset.

(c) If you increased each of the numbers above by 3, what happens to the mean and standard deviation? Explain.

(d) Would the median or the mean be a more robust measure of center for this dataset? Explain.

Solution.

(a) The mean of this dataset is...

$$\overline{x} = \frac{\sum x_i}{n} = \frac{23 + 38 + 17 + 20 + 26}{5} = \frac{124}{5} = 24.8$$

(b) Recall that the variance is $\sigma^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2$ and the standard deviation is $\sigma = \sqrt{\sigma^2}$.

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
23	-1.8	3.24
38	13.2	174.24
17	-7.8	60.84
20	-4.8	23.04
26	1.2	1.44
	Total:	262.8

Therefore, the variance is...

$$\sigma^2 = \frac{1}{n-1}\sum(x_i - \overline{x})^2 = \frac{1}{5-1} \cdot 262.8 = \frac{1}{4} \cdot 262.8 = 65.7$$

But then the standard deviation is $\sigma = \sqrt{\sigma^2} = \sqrt{65.7} = 8.10555$.

(c) Because each value would go up by 3, the mean would go up by 3. However, because all the values 'stay in the same position' relative to all other numbers in the dataset, the total 'spread' of the dataset must stay the same. Therefore, the standard deviation remains unchanged. Alternatively, if we introduce a linear change of variables $mx + b$, the mean changes from $\overline{x}$ to $m\overline{x} + b$ and the standard deviation is multiplied by a factor of $|m|$.

(d) The median would be a more robust measure of center than the mean. The median is more resistant to the presence of outliers compared to the mean. If the value 10^{10} were added to the dataset, the mean would change to nearly $\frac{1}{5} \cdot 10^{10}$ whereas the median would merely change from 23 to 26.