Numerical Summaries of Quantitative Data

Today, we will discuss ways in which to describe quantitative data sets numerically.

MEASURES OF CENTER:
The two most common measures of the center of a data set are the mean (or average) and median. The *mean*, denoted $\bar{x}$ (pronounced "x-bar"), is defined to be the sum of all of the data entries divided by the total number of entries; in equation form, it looks like

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$

where n is the total number of data points and x_i (for $i = 1, ..., n$) represent the n specific data points. The *median* is the middle entry of the data set, when the entries are all ordered from smallest to largest. If there are an even number of entries in the data set, the median is calculated by taking the average of the middle two entries.

Question 1: Find the mean and median of the following set of test scores:

91 84 75 78 100 86

Mean:_____ Median:_____

Question 2: Construct a data set that has a mean of 11 and a median of 12. Your data set must have at least $n = 3$ values.

Data set:_____

Question 3: Construct a data set of $n = 5$ values in which the mean is larger than the median. Then, try to force the median to be larger than the mean by changing a single value.

Original data set:_____

Altered data set:_____

Which value did you change? Why?

MEASURES OF SPREAD:

Two common measures of spread are the range and interquartile range. We will also discuss the standard deviation at a later point in the class. The *range* is defined to be the difference between the largest and smallest values of the data set. The *interquartile range* (IQR) is defined to be the difference between the upper and lower quartiles of the data set. Just as the median separates the data into halves, the quartiles separate it further into quarters. The lower quartile, Q_1 (also called the first quartile), is found by taking the median of the lower half of the data set (everything below the median). The upper quartile, Q_3 (also called the third quartile), is the median of the upper half of the data.

Question 4: Find the median, range, and IQR of the following set of $n = 20$ cholesterol levels:

196	212	200	242	206	178	184	198	160	182
198	182	222	198	188	166	204	178	164	230

Median:_____ Range:_____ IQR:_____

Question 5: The set of (Minimum, Q_1, Median, Q_3, Maximum) is called the *five-number summary* of a data set. Give the five-number summary for the cholesterol data:

Five-number summary:_____

OUTLIERS:

Outliers, or values that are not consistent with the rest of the data, are defined to be those values that lie less than 1.5× IQR to the left of Q_1 or more than 1.5× IQR to the right of Q_3.

Question 6: Are there any outliers in the cholesterol data set? If so, what are they?

Question 7: Consider the following weights (in pounds) for nine men on the Cambridge crew team:

188.5	183.0	194.5	185.0	214.0	203.5	186.0	178.5	109.0

Identify any outliers in this data set: _____

Does your answer make sense in the context of this data set? Why or why not?

Question 8: What effect do you think outliers have on our previously defined measures? For each of the following, decide whether you think that measure is *resistant* to outliers. (Note: if a measure is resistant to outliers, that means it is relatively unaffected by the addition of an outlier to the data).

Mean: _____

Median: _____

Range: _____

IQR: _____

You will read about the fourth feature of quantitative variables, shape, before our next meeting. As you read, think about what effects our first three features (center, spread, and outliers) may have upon the shape of a data set.

CHALLENGE:
Construct a data set that has all of the following characteristics.

- The mean is larger than the median.
- There are at least two outliers.
- The IQR is equal to 10.
- The data set contains at least $n = 6$ values.