

MAC-CPTM Situations Project

Situation 04: Representing Standard Deviation*

Prompt

In prior lessons, students learned to compute mean, mode and median. The teacher presented the formula for standard deviation and had students work through an example of computing the standard deviation with data from a summer job context. The following written work developed during the example:

x	\bar{x}	$(x - \bar{x})$	$(x - \bar{x})^2$
140	200	-60	3600
190	200	-10	100
210	200	10	100
260	200	60	3600

$$\sum (x_i - \bar{x})^2 = 7400$$

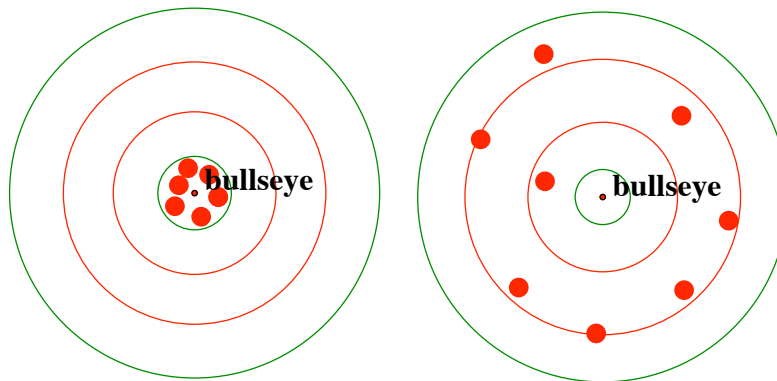
$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$$\sigma = \sqrt{\frac{7,400}{4}}$$

$$\sigma = \sqrt{1850}$$

$$\sigma \approx 43.01$$

The teacher then said, “Standard deviation is a measure of the consistency of our data set. Do you know what consistency means?” To explain “consistency” the teacher used the idea of throwing darts. One student pursued the analogy, “If you hit the bull’s eye your standard deviation would be lower. But if you’re all over the board, your standard deviation would be higher.” The student drew the following picture to illustrate his idea:



A student raised her hand and asked, “But what does this tell us about what we are trying to find?”

Commentary

This Prompt seems to deal with the analogy of the data points, mean, and the standard deviation of a data set as a dartboard target and the places on the target where the darts landed. The main issue raised by the student in the Prompt is the representation of standard deviation as either “hit[ting] the bull’s eye” or being “all over the board.” The bull’s eye in this analogy represents the mean. Standard deviation is a measure of the spread of a data set, so there are aspects of the target analogy that make it a viable representation because it addresses two key aspects of a visual representation: the objects and the relative positions of those objects with respect to the mean. However, the robustness of a representation depends upon how well these two aspects collectively represent the key mathematical objects being represented (e.g., data points, mean, standard deviation) and the relationships among those objects. The target representation contains severe limitations in this regard. An alternative representation (using a 1-dimensional number line rather than the 2-dimensional target) is suggested in Focus 5.

Note: Within the Prompt, there is not a consistent use of notation. To clarify, we must distinguish between parameters, which refer to the population, and statistics, which refer to a sample. This distinction is seen in the two formulas for standard deviation.

For the population, the standard deviation is defined to be $\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$, where μ is the population mean and N is the population size. For a sample, the standard deviation is defined to be $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$, where \bar{x} is the sample mean and n is the sample size.

The intent in the Prompt is to calculate a *statistic*, but a mixture of parameter and statistic notation is used, resulting in an incorrect formula for standard deviation. From this point forward, we will use parameter notation when referring to the population and statistic notation when referring to a sample.

Mathematical Foci

Mathematical Focus 1

Standard deviation is a measure of the spread of the data with respect to the mean.

Standard deviation is one way of indicating the spread of the data with respect to the mean, in its data distribution. The more spread out the data are from the mean, the larger the standard deviation for that particular data set. A smaller standard deviation indicates that the data are more tightly grouped about the mean. The standard deviation

for a data set is somewhat like the average distance the data points are from the mean. However, this is not a strict average, as is seen in the formula for standard deviation:

$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$. A truer average of the spread of the data with respect to the mean is

known as the Mean Absolute Deviation (*MAD*), and is calculated by the following

formula: $MAD = \frac{\sum|x - \bar{x}|}{n}$. Note that both of these formulas provide information about

the spread of the data with respect to the mean, and therefore they must yield non-negative values. In the formula for standard deviation, the values are non-negative because each difference, $(x - \bar{x})$, is squared. In the formula for the *MAD*, the values are non-negative because the absolute value is calculated for each difference, $(x - \bar{x})$. The *MAD* is a good introduction to the concept of measuring the spread of the data distribution, and therefore is a nice transitional way to help us understand the statistical concept of standard deviation. In theoretical statistics however, the formula for standard deviation is used rather than the *MAD*, because, among other things, $(x - \bar{x})^2$ is easier to work with mathematically than $|x - \bar{x}|$. Further rationale for the formula for standard deviation of a sample (in particular the denominator $n-1$) is elaborated upon in Situation 67.

Mathematical Focus 2

If elements of a visual representation of a formula represent symbolic elements of the formula, the relative positions of those elements should be consistent with the meaning of the symbols.

In the formula for the sample standard deviation, we have n , x , \bar{x} , and s . The number of data points, n , is represented by the number of darts on the target. Every data point (that is, each x -value) is represented by the place a dart lands. For the two targets shown in the Prompt, the bull's eye (center of the target) represents the sample mean, \bar{x} . Since distance is non-negative, each distance between a dart and the bull's eye is associated with a value of $|x - \bar{x}|$. The standard deviation, s , is indicated by how spread out the darts are. For example, if the darts are tightly clustered around the bull's eye, this represents a small standard deviation, and if the darts are "all over the board" the data set has a larger value for s .

Mathematical Focus 3

Data points may be compared either to the sample mean or the population mean—in either case, placement of these points in a visual representation may indicate something about the standard deviation of the data set.

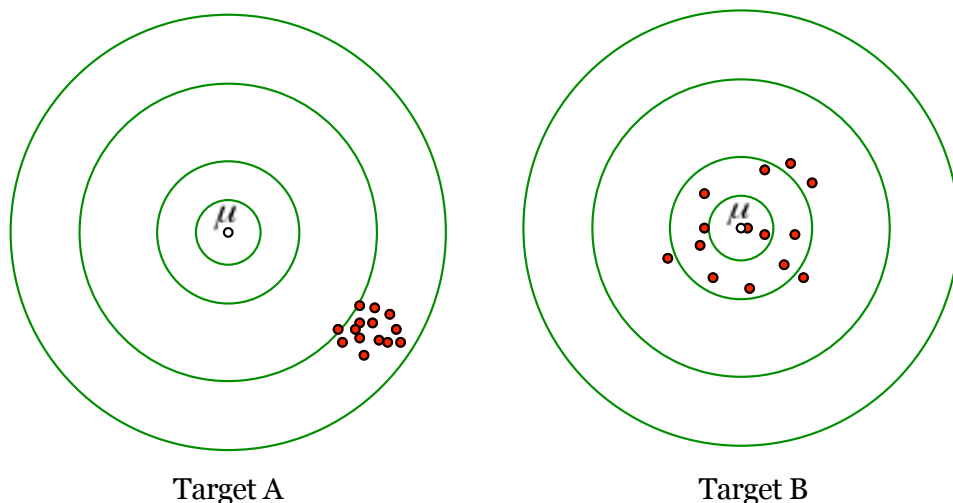


Figure 1

Targets A and B, shown in Figure 1, represent the distribution of two different sets of data. For these targets we will assume the center in each case represents the population mean, μ , indicated by an open circle on the target and the center of the data points represents the sample mean, \bar{x} . The center of Target A is not the same as the center of the data points, indicating that the mean calculated from this sample is not a desirable estimate of the population mean. We may say that the sample mean in Target A is a *biased* estimate (either an overestimate or underestimate) of the population mean. Most likely, the bias of the sample mean occurs because of the way the sample was selected. Target B, on the other hand, shows a sample whose mean is a more desirable estimate of the population mean, since the center of the data points is close to the center of the target. That is, in Target B, the sample mean, \bar{x} , is closer to an *unbiased* estimate of the population mean, μ .

Clearly, the spread of the distribution of the data points is different in each target representation. In Target A, the data points are more tightly clustered than the data points in Target B. This indicates a smaller sample standard deviation for the set of data points in Target A. That is, the data points in Target A have a greater consistency (i.e. smaller variability) than those in Target B.

Therefore, each target representation displays a characteristic of an ideal sample, which is typically achieved through random sampling and using large sample sizes. Random sampling, in the long run, would yield an unbiased estimate of the population mean (as in Target B). Using a large sample size yields more consistency in the distribution of

data points (as in Target A). Therefore random sampling reduces bias and large sample size reduces variability.

Mathematical Focus 4

The position of data points relative to the mean of the data set is best represented using a one-dimensional analogy because each data point is either greater than, equal to, or less than the mean.

From the two-dimensional target representation, we can create a one-dimensional representation of the data. In what follows, we will assume that the center of the target represents the sample mean, \bar{x} . To construct this representation, begin by placing the sample mean at the center of the target. Then place each data point so that its distance from the center of the target represents the absolute value of the difference between the data value and the sample mean, $|x - \bar{x}|$. The point may be anywhere on the target such that this criterion is met (i.e. the point may be anywhere on a circle with radius $|x - \bar{x}|$). For example, if the sample mean is 200, the data points 190 and 210 will lie on the same circle, because in each case the value of $|x - \bar{x}|$ is 10. However, it may seem counterintuitive to randomly place these two data points on this circle; rather, one may prefer to place them on opposite sides of the sample mean. In the figure below, note that the random placement of the data points may be confusing. In the following figure, the data set from the Prompt is being used: the x values from the sample are 140, 190, 210, and 260, and the sample mean is 200.

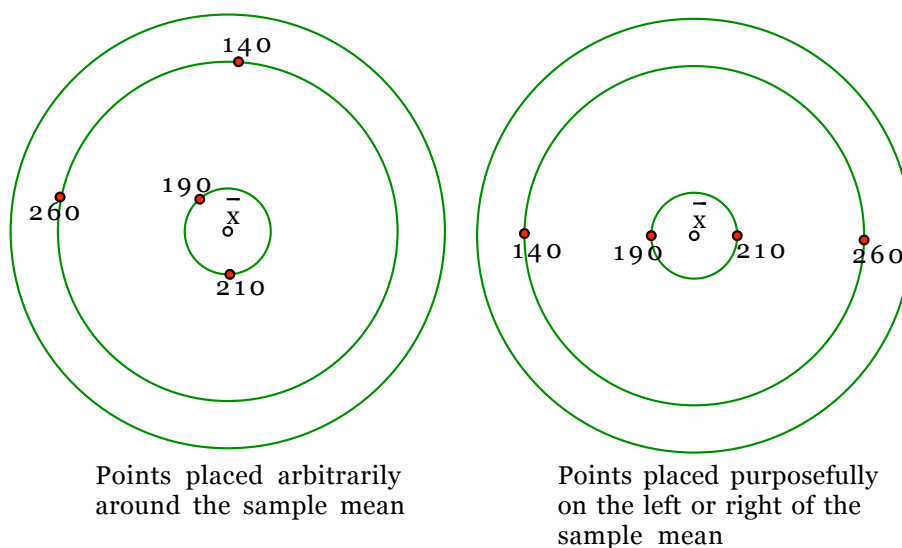


Figure 2

Several things are worth noting about the placement of data values in the two targets above. In the target on the left, the data values are placed arbitrarily on circles with

radii, $|x - \bar{x}|$; the only “rule” is that a point lie a certain distance from the center. In contrast, the target on the right displays an additional stipulation, where the data points lie on a diameter in increasing order from left to right. Note that this more purposeful placement of the data points implies a linear representation. The target on the right has the limitation that there is no way to indicate repetition of data points. That is, the distribution of the data remains unclear, and therefore information about the standard deviation is also unclear.

An alternative to the two-dimensional target representation involves a linear placement of the data points. Such a representation shows where data points lie in relation to the mean, and this relationship is one-dimensional—that is, the points are either less than or greater than the mean (or are equal to the mean). The following dot plot is a linear representation that shows the spread of the same data points used in the Prompt.

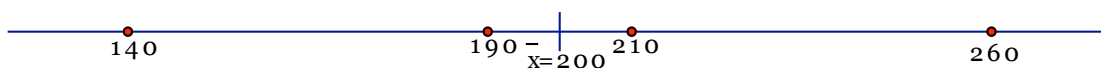


Figure 3

A dot plot has several advantages over either target representation. First, if there is more than one data point at a particular value, this can be indicated as follows:

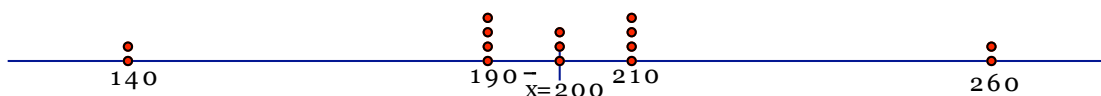


Figure 4

Second, the spread of the data about the sample mean provides a visual representation of the concept of standard deviation. We can compare the standard deviations of two data sets based on their dot plots. For example, in the dot plot in figure 4, data values are more tightly grouped about the mean than the data values in the dot plot in figure 3. This implies that the data distribution in figure 4 has a smaller standard deviation than the data distribution represented in figure 3. The larger sample size in figure 4 also plays a role in the standard deviation being smaller than the standard deviation for the data distribution in figure 3. Specifically, the standard deviation of the data set in figure 4 is approximately 32.95 which is less than the standard deviation of the data set in figure 3, which is approximately 49.67.

Mathematical Focus 5

Different representations may be more or less effective in representing the spread of a data set.

A box and whisker plot also shows us something about the spread of a data set. However, in this type of representation the measure of center tendency is the *median* of the data set rather than the mean. The box and whisker plot provides a visual representation of the five-number summary of the data set (minimum, Q_1 , median, Q_3 , and the maximum), so that the data set is separated into four intervals. Whereas the box and whisker plot does not indicate the shape of the distribution of data values within a particular interval, the box and whisker plot as a whole shows us something about the spread and possible skewness of the data set around the median.

There are limitations to using either the target representation or the box and whisker plot. Neither representation indicates the sample size, n . In the target representation, repeated data values are not necessarily apparent, and the box and whisker plot does not show particular data values other than the minimum and maximum of the data set. Because n is not necessarily known in either representation, there are limitations to the conclusions that can be made about the spread of the data. For example, the data sets represented in Figures 3 and 4 have the same five-number summary, so their box and whisker plots are exactly the same (See Figures 5 and 6). Also, as discussed earlier, the standard deviations of these two data sets are different. Lastly, because of the arbitrary placement of points on circles of radius $|x - \bar{x}|$ in the target representation, the possible skewness of the distribution of the data is not evident. The box and whisker plot, on the other hand, can illustrate possible skewness through the five-number summary. In this case, the box plots indicate that the data distribution appears to be symmetric.

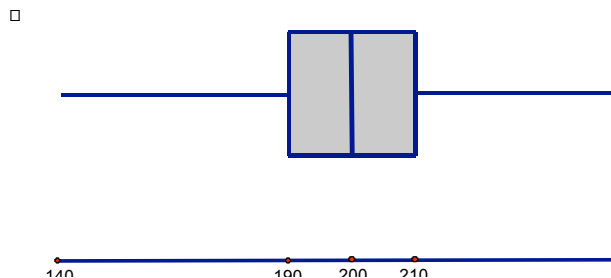


Figure 5: Box and whisker plot for the data values in Figure 3

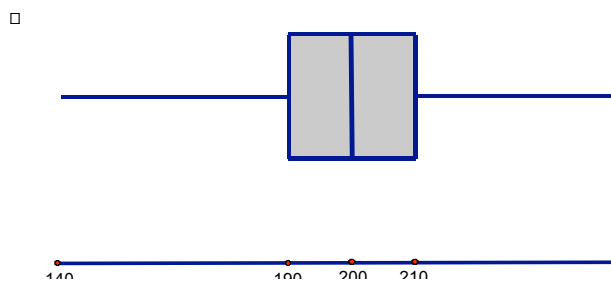


Figure 6: Box and whisker plot for the data values in Figure 4

In conclusion, there is value to each representation described above depending upon the intended purpose. The dot plot is useful in that it displays the actual data points and therefore illustrates the spread of these data values about the mean, possible bias, and consistency of the data distribution. The box and whisker plot is useful in that the five-number summary indicates the spread of the data in relation to measures of position (i.e. minimum, Q_1 , median, Q_3 , and the maximum). The target representation is useful for indicating possible bias and consistency of the data distribution.