

**Summarize and describe distributions.**

5. Summarize numerical data sets in relation to their context.

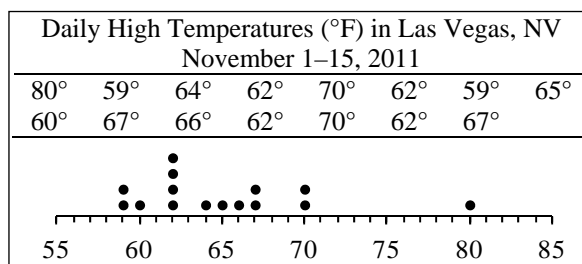
Under the Nevada State Standards, students summarized data sets as early as grade 4, computing mode, median, and range. Mean was introduced in grade 5 and interquartile range in grade 8. In the Common Core State Standards (CCSS), the only data analysis before grade 6 is rudimentary graphing of categorical and quantitative data. Thus, it is in sixth grade when students have their first real exposure to statistics and the approach is quite different than the Nevada State Standards.

There are multiple parts to this standard, covering multiple aspects of data distribution including the number of observations in a data set, the nature of the variable being studied, computing measures of center and variability, examining departures from distribution patterns, and connecting measures of center and variability to the distribution's shape. While each part can be viewed as a separate concept or skill, they should not be taught in isolation. It is important to see the standard as a whole greater than the sum of its parts.

There are two main ideas of which teachers must remain cognizant when teaching statistics: *variability* and *context*. Variability is inherent in our world. Whether measuring heights of bean plants, daily high temperatures, or opinions of likely voters, differences are inevitable. Sometimes the variability is natural, sometimes it is induced, and sometimes it is due to measurement. Regardless of its source, it is why we do statistics.

The other key word in the standard and each of its subparts is *context*. Data are not simply numbers; they are numbers with a context. Context provides meaning. In mathematics, we may temporarily remove context to look for patterns. When analyzing data, the meaning of patterns relies on context. We cannot draw conclusions from data without awareness of the who, what, when, where, why, and how. Without context, data are of little interest.

The data set at right is a list of the daily high temperatures in Las Vegas, Nevada in the first half of November 2011 as recorded by the National Weather Service. The variability of the high temperatures is apparent from the table, but even more visible when organized into a dot plot.



The remainder of this letter will focus on the third part of 6.SP.5.

- c. Giving quantitative measures of center (median and/or mean) and variability (interquartile range and/or mean absolute deviation), as well as describing any overall pattern and any striking deviations from the overall pattern with reference to the context in which the data were gathered.

The *measures of center* (sometimes called *measures of central tendency*), *mean* and *median*, are familiar. The *measures of variability* (sometimes called *measures of spread*), *interquartile range* and *mean absolute deviation*, may not be. But let's begin with a well-known measure of variability that is not mentioned in the CCSS: *range*.

$$\text{range} = \text{value of the largest observation} - \text{value of the smallest observation}$$

Since it's not in the CCSS, does that mean it should be excluded from our curriculum? No. Knowing how to compute range sets the stage for *interquartile range*. In the data set above, the range = 80°F – 59°F = 21°F.

$$\text{interquartile range (IQR)} = \text{value of the third quartile} - \text{value of the first quartile}$$

*Interquartile range* is the difference between the third and first quartiles in a data set. Quartiles are those points that divide a data set into roughly four equally-sized parts. To divide the data into fourths, find the median. There are 15 observations in the set, so the median (*M*) is the 8<sup>th</sup> value (64°F). The first quartile (*Q*<sub>1</sub>) is median of all values below the overall median (62°F).

59°	59°	60°	62°	62°	62°	62°	64°	65°	66°	67°	67°	70°	70°	80°
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

The third quartile (*Q*<sub>3</sub>) is the median of all values above the overall median (67°F). So, the interquartile range is  $\text{IQR} = 67^\circ\text{F} - 62^\circ\text{F} = 5^\circ\text{F}$ . This tells us that the spread across the middle half of the data is 5°F. We may conclude the temperatures in the first half of November were fairly consistent. (Note the quartiles divide the data into four roughly

equally-sized groups, but do not divide the range into four equal distances.)

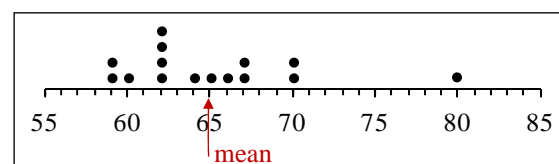
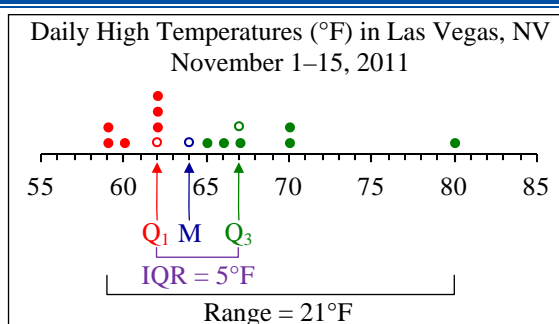
While the interquartile range tells us something about how the data are spread out, it is based on just two locations within the data set. It does not take into account all values of the data. One measure of variability that does is the *mean absolute deviation*. We compute this in four steps.

1. Compute the mean of the data set.
2. Determine each observation's *deviation*. It is the difference between the observation and the mean.
3. Find the *absolute value* of each deviation.
4. Compute the *mean* of the absolute deviations.

The mean of the temperatures is **65°F**.

Observations	59°	59°	60°	62°	62°	62°	62°	64°	65°	66°	67°	67°	70°	70°	80°
Deviations	-6°	-6°	-5°	-3°	-3°	-3°	-3°	-1°	0°	1°	2°	2°	5°	5°	15°
Absolute Deviations	6°	6°	5°	3°	3°	3°	3°	1°	0°	1°	2°	2°	5°	5°	15°

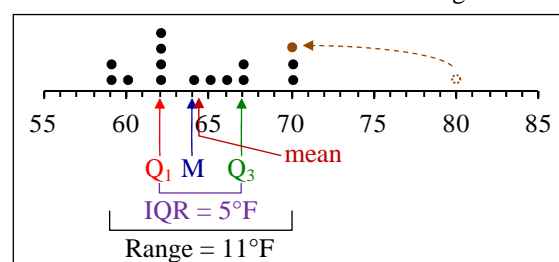
The mean absolute deviation (MAD) is 4°F. We interpret this by saying that the temperature on any given day in the first 15 days of November was about 4°F from the mean of 65°F. Some days were less than 4°F from the mean, some were more, but on average the difference was 4°F. The temperatures appear to be fairly consistent.



So, why do we have multiple measures of center (mean, median) and multiple measures of spread (interquartile range, mean absolute deviation), and why is range not emphasized in the standards? It has to do with how extreme values (i.e. outliers) affect those measures. (It also has to do with the shape of the data's distribution, but we will touch on that at a later date.)

Because range depends on only two observations, it is of dubious value as a way to describe the spread of a data set. It is greatly influenced by extreme observations (i.e. outliers) and can sometimes give an incorrect picture of overall variability. Notice the observation of **80°F** appears to be an outlier. Let's pretend for a moment that it was the same as the next highest temperature of 70°F. The range would then be 11°F, a little more than half as much as 21°F. The outlier has a big influence on the range. Statisticians would say that range is not *resistant* to outliers. While range is not completely useless, it should be used only when it gives meaningful information and should receive less emphasis than the other measures of variability.

What happens to the other two measures of variability if we were to change that 80°F observation to 70°F? The first and third quartiles do not change, thus the interquartile range is still 5°F. The IQR is resistant to outliers, as is the median which remains 64°F.



In the case of the mean absolute deviation, changing 80° to 70°F would first change the mean to 64.3°F, a drop of nearly a degree. (This shows that the mean is not resistant to outliers, especially in small data sets.) Recalculating the MAD with our new mean, we would get about 3.3°F, again a drop of nearly a degree. While that reduction of 0.7°F may not seem like much, it is change of almost 18% from the MAD of 4°F. That's a substantial difference. The mean absolute deviation is not resistant to outliers, especially in small data sets.

The median and the IQR are usually reported together, as are the mean and MAD. Which pair we choose to report depends on a combination of the shape of the data, the presence of outliers, and the size of the data set.

Computing measures of center and spread cannot be done in isolation. They must be done along with graphs of the data. Students have been making line plots since grade 3 and will go on to compare plots of multiple data sets in later grades, along with making scatter plots. The operations involved in computing MAD connect to other standards in grade 6, particularly working with signed numbers and absolute value.

Finally, what is the deal with this "new-fangled" mean absolute deviation? Well, it's actually been around for a while, but has not been part of the middle school curriculum. It gives us a companion measure of spread for the mean and sets the stage for a better companion students will learn in high school: *standard deviation*.

### Connections

- 3.MD.4
- 4.MD.4
- 5.MD.1-2
- 6.RP.3
- 6.NS.5-7
- 6.EE.2
- 6.SP.1-4
- 7.SP.1-4
- 8.SP.1
- N.Q.1-3
- S.ID.1-4
- S.ID.6-7
- S.IC.6