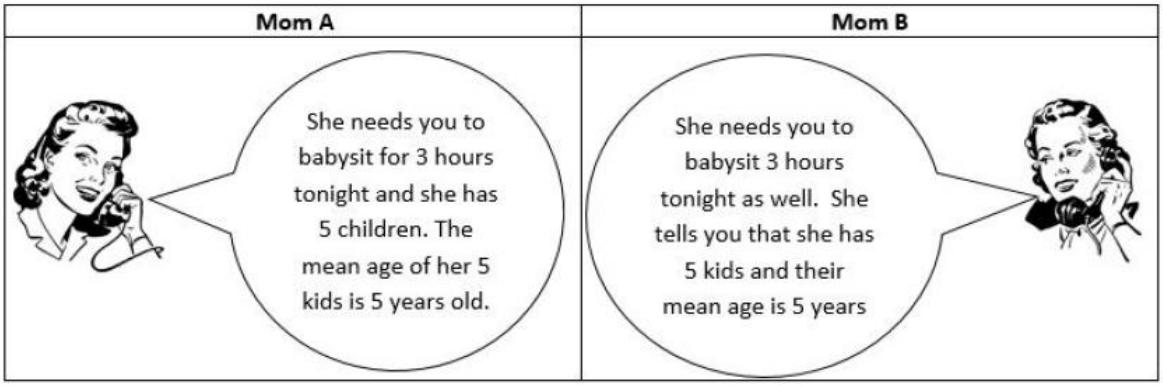
Section 6.2: Measures of Variation

Measures of variation (or spread) refers to a set of numerical summaries that describe the degree to which the data are spread out.

Why do we need them? Why is using measures of center not sufficient in describing data sets? To answer these kind of questions consider the following example.

Example 1. Babysitting

Suppose you are starting a new babysitting business. You have advertised in your neighborhood and it does not take long for calls to come in. You get calls from two moms, "Mom A" and "Mom B". The information given by both moms is illustrated in the picture below.



As you can see the information provided is not enough to distinguish between the two families. It is highly unlikely that these two moms have kids that are the exact same ages. You politely ask the moms for the actual ages of their children.

Mom A	Mom B
1, 2, 4, 8, 10	3, 4, 5, 6, 7

We will use these two groups of data to find a new set of measures called *measures of variation: range and standard deviation*. These measures will help us quantify the age variation for each family and better understand our data.

Objective: Compute a range

Definition. *The **range** of a set of data is the difference between the largest value and the smallest value in the data set.*

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

The larger the range value, the more scattered the data.

Example 2. Babysitting

Find the range for our data:

Mom A	Mom B
1, 2, 4, 8, 10	3, 4, 5, 6, 7

Range of children's ages of Mom A $=10-1=9$

Range of children's ages of Mom B $=7-3=4$

Since the range value 9 is the higher of the two range values, we conclude that children's ages of Mom A have greater age variation than the children's ages of Mom B.

A disadvantage of using the range is that it only considers the maximum and the minimum values and ignores the rest of data.

Objective: Compute a standard deviation

We would prefer a measure of variation to quantify spread with respect to the center as well as to use all the data values in the set. This measure is the standard deviation.

Definition 3. *The **standard deviation** is the average distance the data values are from the mean.*

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

Example 3. Babysitting

Consider the "Babysitting example" again. Find the standard deviation for Mom A children's ages.

Step 1: Start by setting up a table like the one below.

Step 2: Write each data value in the first column of the table. In column 2 find the difference between each age and the mean. Next square the values from column 2 and enter the results in column 3 (see the computations in the table below).

Mom A

Children's ages mean is $\mu = 5$

Age x	$Age - Mean$ $x - \mu$	$(Age - Mean)^2$ $(x-\mu)^2$
1	$1-5=-4$	$(-4)^2=16$
2	$2-5=-3$	$(-3)^2=9$
4	$4-5=-1$	$(-1)^2=1$
8	$8-5=3$	$(3)^2=9$
10	$10-5=5$	$(5)^2=25$

Step 3: Substitute the values in the formula for standard deviation and perform the computations.

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$ Standard deviation formula.

$$\sigma = \sqrt{\frac{16+9+1+9+25}{5}}$$ Add the values from column 3 of the table in the numerator. In the denominator put the number of children of Mom A.

$$\sigma = \sqrt{\frac{60}{5}}$$ Divide 60 by 5

$$\sigma = \sqrt{12}$$ Take square root of 12

$$\sigma \approx \mathbf{3.46}$$ This is the standard deviation of children's ages for Mom A.

Example 4. Babysitting

Find the standard deviation for Mom B children's ages.

We repeat the same steps as in previous example to find the standard deviation of children ages for Mom B.

Step 1: Set up the table.

Step 2: We write each data value in the first column of the table. In column 2 we find the difference between each age and the mean. Square the values from column 2 and enter the results in column 3 (see the computations in the table below).

Mom B

Children's ages mean is $\mu = 5$

Age x	$Age - Mean$ $x - \mu$	$(Age - Mean)^2$ $(x-\mu)^2$
3	$3-5=-2$	$(-2)^2=4$
4	$4-5=-1$	$(-1)^2=1$
5	$5-5=0$	$(0)^2=0$
6	$6-5=1$	$(1)^2=1$
7	$7-5=2$	$(2)^2=4$

Step 3: Substitute the values in the formula for standard deviation and perform the computations.

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$ Standard deviation formula.

$$\sigma = \sqrt{\frac{4+1+0+1+4}{5}}$$ Add the values from column 3 of the table in the numerator. In the denominator put the number of children of Mom B.

$$\sigma = \sqrt{\frac{10}{5}}$$ Divide 10 by 5.

$$\sigma = \sqrt{2}$$ Take square root of 2.

$$\sigma \approx \mathbf{1.41}$$ This is the standard deviation of children's ages for Mom B.

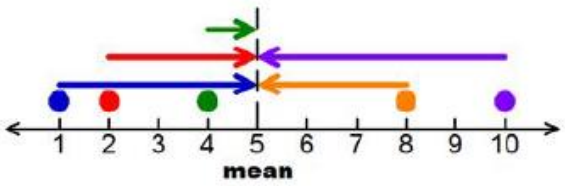
Objective: Compare distributions with different standard deviations

As mentioned previously, the standard deviation can be used to determine the variation (spread) of data. That is, the larger the standard deviation, the more the data values are spread. This also permits us to compare data sets with different standard deviations. Let us look at a few examples.

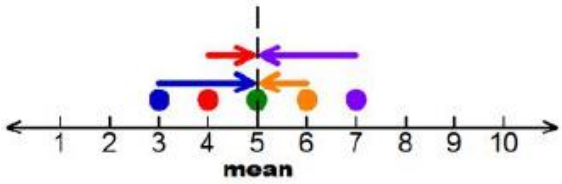
Example 5. Babysitting

We would like to compare the age variation (spread) between the two families. To visualize the idea of spread consider the dot plots below corresponding to our data sets.

Mom A



Mom B



Clearly, the data in the dot plot A is spread out more from the mean, while the data in the dot plot B is closer to the mean. Thus, the graphs show that Mom A children's ages are more spread out from the mean than the Mom B children's ages.

We can also compare the standard deviations calculated previously for each family. We have a standard deviation value of 3.46 which is larger than the standard deviation value of 1.41. Thus, we reach the same conclusion that Mom A children's ages have a greater age variation than Mom B children's ages.

Example 6. Emergency Room Waiting Time

Consider the following scenario:

The mean of waiting times in an emergency room is 90 minutes with a standard deviation of 12 minutes for people who are admitted for additional treatment. The mean waiting time for patients who are discharged after receiving treatment is 130 minutes with a standard deviation of 19 minutes.

In this example, by comparing the standard deviation values (13 min. and 19 min.), we can conclude that the waiting time for patients who are discharged after receiving treatment are more spread out than the waiting time for people who are admitted for additional treatment.

Example 7. Route to Work

You have two different routes to work each morning. The summary statistics are below.

Route 1: Mean = 14 minutes, Standard deviation = 4.3 minutes

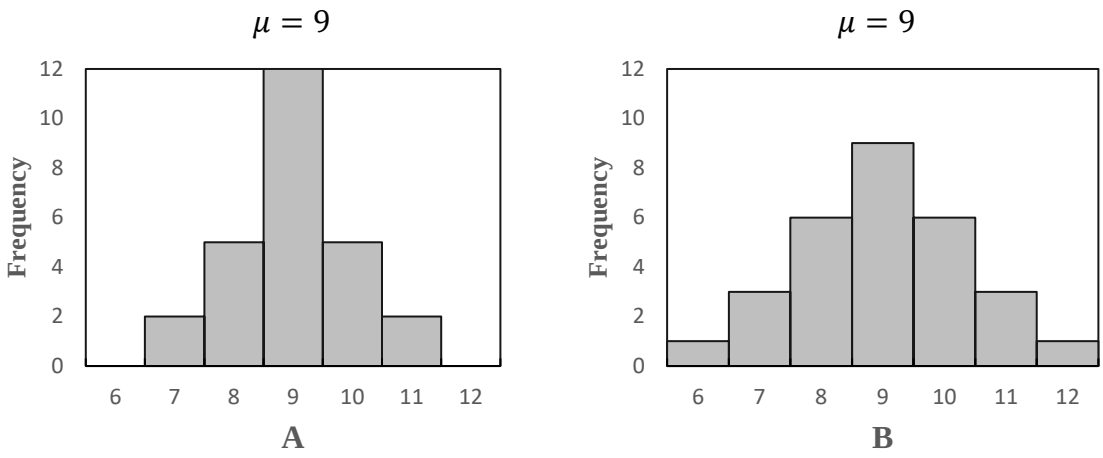
Route 2: Mean = 16 minutes, Standard deviation = 1.1 minutes

Which route offers you are more consistent commute time?

Route 2 has a standard deviation that is much smaller then Route 1. This will tells us that the route times they used to compute the summary statistics are more similar to one another. This ensures a more consistent commute.

Example 8. Bank Waiting Time

We can also determine which data set has a greater variation by looking at histograms. Consider the two histograms below which illustrate the waiting time (in minutes) of customers at two banks.



The data in histogram B is spread out more from the mean, while the data in histogram A is closer to the mean. This means that histogram B has a larger standard deviation than histogram A. Therefore, Bank A has more consistency in time.

6.2 Practice

1. The number of hours a student went to work after school last week are:

   1, 4, 3, 2, 6

   a. Find the range.
   b. Calculate the standard deviation.

2. The daily vehicle pass charge in dollars for six National parks are:

   15, 10, 6, 15, 20, 12

   a. Find the range.
   b. Calculate the standard deviation.

3. The range of scores on a statistics test was 40. The highest score was 99. What was the lowest score?

4. The weather station announced that the temperature fluctuates between a low of 80 F and a high of 93 F. Which measure of spread could be calculated using just this information? What is its value?

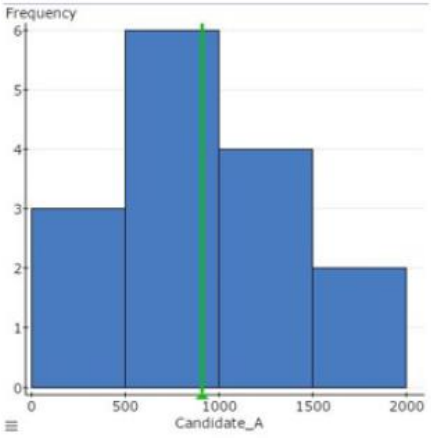
5. The table below shows the preparation tax return times (in hours) of two certified public accountants.

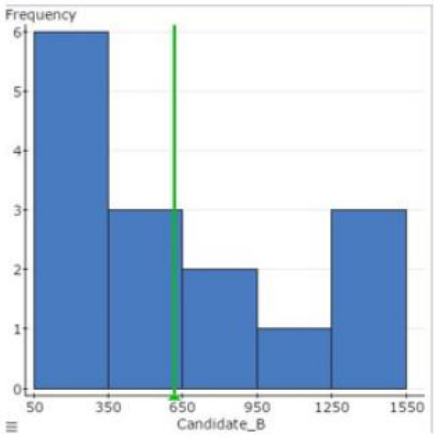
Accountant X	7	9	5	11	8
Accountant Z	5	11	14	3	7

   a. Find the range preparation time for each accountant.
   b. Calculate the standard deviations for each accountant.
   c. Who has the more consistent tax preparation time?

6. The average number of days construction workers miss per year is 11 with a standard deviation of 2.3. The average number of days factory workers miss per year is 8 with a standard deviation of 1.8. Which class of workers is more variable in terms of days missed?

7. The histograms below illustrate the contribution (in dollars) made to two presidential candidates in a recent election. Which histogram has a larger standard deviation?





8. The following table summarizes the wait times at a call center. Which center had a greater variation of wait times?

Call Center A	**Call Center B**
$n = 200$ calls	$n = 200$ calls
$\mu = 4.5$ minutes	$\mu = 3.2$ minutes
$\sigma = 1.2$ minutes	$\sigma = 1.6$ minutes

9. A movie director is trying to choose between 2 production companies. Company H spends an average of \$88 million with a standard deviation of \$15 million, while Company T spends an average of \$95 million with a standard deviation of \$9 million. Which company is more consistent in spending?

6.2 Answers

1.
   a. Range = 5
   b. $\sigma = 1.72$
2.
   a. Range = 14
   b. $\sigma = 4.40$

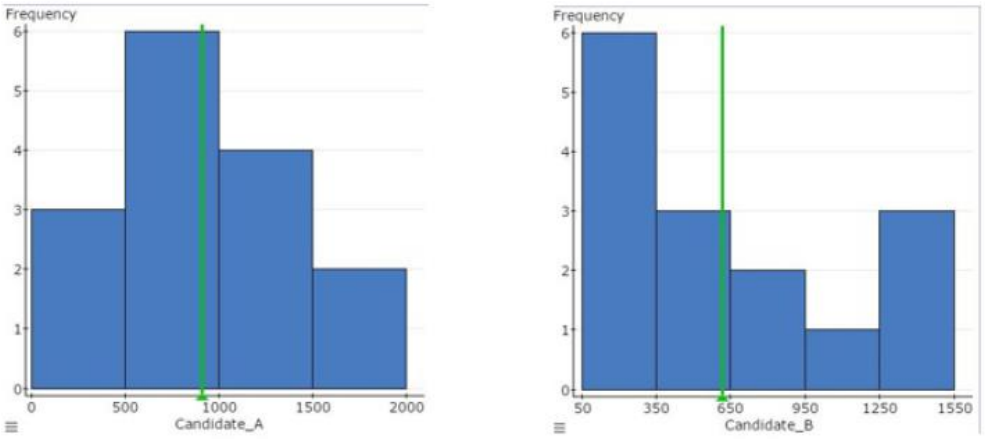
3. Lowest score is 59.

4. Range. The range is 13°F.

5.
   a. Range (Accountant X) = 6 hours Range (Accountant Z) = 11 hours
   b. σ (Accountant X) $= 2$ σ (Accountant Z) $= 4$
   c. Accountant X because their standard deviation is smaller.

6. The construction workers are more variable with their days missed because their standard deviation is larger. This means that their data is more varied.

7. The histograms below illustrate the contribution (in dollars) made to two presidential candidates in a recent election. Which histogram has a larger standard deviation?



Candidate B will have a larger standard deviation. You can tell because of the higher bars in the histogram are farther away from the mean (the green line). This means that more data are in the \$50 to \$349 and the \$1250 to \$1549 classifications.

8. Call center B does because it has a larger standard deviation.

9. Company T because it has a smaller standard deviation.