Section 1.5 Measures of Spread

Get Started – How do you solve a formula for a variable?

- How do you compute the range of data?
- What does the standard deviation tell you about data?
- How can we use the coefficient of variation to compare the standard deviations of different sets of data?
- How do you interpret quartiles and percentiles of data?
- How do you assemble a five-number summary of data?
- How are stem and leaf plots used to compare data?

Get Started – How do you solve a formula for a variable?

Key Terms

Equivalent form

Summary

In an earlier section, we examined how to solve an equation for a solution. The solution to an equation is a number or several numbers that when substituted for the variables results in a true statement.

$x=4$	solves	$2(x+2)+5=17$	since	$2(4+2)+5 \overset{TRUE}{=} 17$
$y=0.2$	solves	$5y+1=2$	since	$5(0.2)+1 \overset{TRUE}{=} 2$
$P=1500$ and $r=0.05$	solves	$1500.5=P+10r$	since	$1500.5 \overset{TRUE}{=} 1500+10(0.05)$

In each case above, replacing the variable with its corresponding value leads to a true statement.

Equations such as $1500.5=P+10r$ are often encountered in math. These equations may have 2 or more variables in them. If we want to solve them for a particular variable, we need to get that variable by itself on one side of the equation. It should also only appear on that side of the equation and nowhere else.

To solve for a variable, we need to rewrite the equation in an **equivalent form**. This means that the solutions to the equation and its equivalent form are the same. There are two rules that help us to get the equivalent form.

1. Adding an expression to both sides of an equation or or subtracting an expression from both sides of an equation gives an equivalent equation.
2. Multiplying both sides of an equation or dividing both sides of an equation by an expression gives an equivalent equation. Take care that what you multiply or divide by is not equal to zero.

Notes

Guided Example 1

Solve each of the equations for the indicated variable.

a. $2x+4y=20$ for y

Solution Start by isolating the term that contains the variable you are solving for. Subtracting $2x$ from both sides leads to

$$2x+4y=20$$

$$2x-2x+4y=20-2x \quad \text{Subtract } 2x \text{ from both sides}$$

$$4y=20-2x$$

To get y by itself, divide both sides of the equation by 4:

$$\frac{4y}{4}=\frac{20-2x}{4} \quad \text{Divide both sides by 4}$$

$$y=\frac{20-2x}{4} \quad \text{Simplify the left side}$$

$$y=\frac{20}{4}-\frac{2x}{4} \quad \text{Divide both terms on the right by 4 and simplify}$$

$$y=5-\frac{1}{2}x$$

The equation is solved for y in the second step. However, simplifying the right side often yields a simpler equation.

b. $A=P+Prt$ for t

Solution To solve for t, Subtract P from both sides of the equation and then divide to get t by itself:

Practice 1

Solve each of the equations for the indicated variable.

a. $10x+3y=30$ for x

b. $A=2hb+b$ for h

$A = P + Prt$ $A - P = P - P + Prt$ Subtract *P* from both sides and simplify $A - P = Prt$ $\frac{A-P}{Pr} = \frac{Prt}{Pr}$ Divide both sides by *Pr* and simplify $\frac{A-P}{Pr} = t$ c. $A = P + Prt$ for *P* **Solution** In solving for *P*, we need to notice that *P* appears in two places in the equation. Before we start isolating *P*, we need to factor *P* from the right side so that *P* appears in one place: $A = P + Prt$ $A = P(1 + rt)$ Factor *P* from both terms on the right You can check this step by distributing the P inside the parentheses. To get P by itself, divide both sides of the equation by $1 + rt$: $\frac{A}{1+rt} = \frac{P(1+rt)}{1+rt}$ Divide both sides by 1 + *rt* and simplify $\frac{A}{1+rt} = P$ d. $z = \frac{x-\mu}{\sigma}$ for μ **Solution** It is easier to work with this equation if we eliminate the fraction on the right side. Multiply both sides by σ: $z = \frac{x-\mu}{\sigma}$ $z\sigma = \frac{x-\mu}{\cancel{\sigma}} \cdot \cancel{\sigma}$ Multiply both sides by σ and simplify $z\sigma = x - \mu$	c. $A = 2hb + b$ for *b* d. $z = \frac{x-\mu}{\sigma}$ for *x*

To get μ by itself, subtract *x* from both sides and then multiply both sides by -1: $z\sigma - x = x - x - \mu$ $z\sigma - x = -\mu$ Subtract *x* from both sides and simplify $-1(z\sigma - x) = -1(-\mu)$ $-z\sigma + x = \mu$ Multiply both sides by -1	

How do you compute the range of data?

<u>Key Terms</u>

Range

<u>Summary</u>

Consider these three sets of quiz scores:

Section A: 5 5 5 5 5 5 5 5 5 5

Section B: 0 0 0 0 0 10 10 10 10 10

Section C: 4 4 4 5 5 5 5 6 6 6

All three of these sets of data have a mean of 5 and median of 5, yet the sets of scores are clearly quite different. In section A, everyone had the same score; in section B half the class got no points and the other half got a perfect score, assuming this was a 10-point quiz. Section C was not as consistent as section A, but not as widely varied as section B.

In addition to the mean and median, which are measures of the "typical" or "middle" value, we also need a measure of how "spread out" or varied each data set is.

There are several ways to measure this "spread" of the data. The first is the simplest and is called the range.

The **range** is the difference between the maximum value and the minimum value of the data set.

Notes

Guided Example 2	Practice 2
In a recent class, the overall percentages at the end of the semester were recorded. 82 86 80 77 83 92 81 61 68 73 86 83 84 94 86 78 65 90 Find the range of the data. **Solution** Examining the data, notice that the smallest value is 61 and the highest value is 94. Since the range is the difference between the maximum and minimum values, $\text{Range} = 94 - 61 = 33$	In a recent class, the overall percentages at the end of the semester were recorded. 73 71 72 73 64 71 77 38 58 75 85 36 86 39 61 Find the range of the data.

What does the standard deviation tell you about data?

Key Terms

Deviation from the mean

Squared deviation from the mean

Sample variance

Sample standard deviation

Population variance

Population standard deviation

Summary

Instead of looking at the difference between highest and lowest to find the range, let's look at the difference between each data value and the center. The center we will use is the mean. The difference between the data value and the mean is called the **deviation from the mean**.

Data Value Mean

$$\text{Deviation from the mean} = x - \bar{x}$$

To see how this works, let's use the data set from the data below describing temperatures in Flagstaff, Arizona.

71	59	69	68	63	57
57	57	57	65	67	

The mean was about 62.7°F.

To find the deviations from the mean, subtract 62.7 from each data value:

x	$x-\overline{x}$
71	8.3
59	-3.7
69	6.3
68	5.3
63	0.3
57	-5.7
57	-5.7
57	-5.7
57	-5.7
65	2.3
67	4.3
Sum	**0.3**

Notice that the sum of the deviations is around zero. If there is no rounding of the mean, then this should add up to exactly zero. So, what does that mean? Does this imply that on average the data values are zero distance from the mean? No. It just means that some of the data values are above the mean and some are below the mean. The negative deviations are for data values that are below the mean and the positive deviations are for data values that are above the mean. We need to get rid of the sign. How do we get rid of a negative sign? Squaring a number is a widely accepted way to make all the numbers positive.
Let's square all of the deviations. To find these values, **square the deviations from the mean**.

$$\text{Squared Deviation from the Mean} = \left(x-\overline{x}\right)^2$$

Also, you can think of this as being the squared distance from the mean.

x	$x-\overline{x}$	$(x-\overline{x})^2$
71	8.3	68.89
59	-3.7	13.69
69	6.3	39.69
68	5.3	28.09
63	0.3	0.09
57	-5.7	32.49
57	-5.7	32.49
57	-5.7	32.49
57	-5.7	32.49
65	2.3	5.29
67	4.3	18.49
Sum	**0.3**	**304.19**

Now that we have the sum of the squared deviations, we should find the mean of these values. However, since this is a sample, the normal way to find the mean, summing and dividing by *n*, does not estimate the true population value correctly. It would underestimate the true value. So, to calculate a better estimate, we will divide by a slightly smaller number, *n* - 1. This strange average is known as the sample variance.

The **sample variance** is the sum of the squared deviations from the mean divided by $n-1$. The symbol for sample variance is s^2 and the formula for the sample variance is

$$s^2 = \frac{\sum(x-\overline{x})^2}{n-1}$$

For this data set, the sample variance is $s^2 = \frac{304.19}{11-1} = 30.419$

The variance measures the average squared distance from the mean. Since we want to know the average distance from the mean, we will need to take the square root at this point.

The **sample standard deviation** is the square root of the variance. The standard deviation is a measure of the average distance the data values are from the mean. The symbol for sample standard deviation is s and the formula for the sample standard deviation is

$$s = \sqrt{s^2} = \sqrt{\frac{\sum\left(x - \bar{x}\right)^2}{n-1}}$$

For this data set, the sample standard deviation is $s = \sqrt{30.419} \approx 5.52\ ^{\circ}F$. Note that the units are the same as the original data.

Since the sample variance and the sample standard deviation are used to estimate the population variance and population standard deviation, we should define the symbols and formulas for those as well.

Population Variance: $\sigma^2 = \dfrac{\sum\left(x - \mu\right)^2}{N}$

Population Standard Deviation: $\sigma = \sqrt{\sigma^2} = \sqrt{\dfrac{\sum\left(x - \mu\right)^2}{N}}$

Notes

Guided Example 3

An instructor is comparing the performance of students in to classes. He records the final percentages students have earned.

Class A: 82 86 80 77 83 92 81 61 68 73 86 83 84 94 86 78 65 90

Class B: 73 71 72 73 64 71 77 38 58 75 85 36 86 39 61

a. Without computing the mean, explain which class has the higher mean and why you think this is so.

Solution Start by ordering each set of data from smallest to largest:

Class A: 61 65 68 73 77 78 80 81 82 83 83 84 86 86 86 90 92 94
Class B: 36 38 39 58 61 64 71 71 72 73 73 75 77 85 86

Examining Class A, the center of the data is in the low 80's. For class B, the middle of the data is in the low 70's. However, Class B has a number of very small data values which could draw the mean down. So, class A has the higher mean.

b. Compute the mean for each class and verify your answer to part a.

Solution For each set of data, add the data values and divide by the total number of data,

$$\text{Class A}: \bar{x} = \frac{\sum x}{n} = \frac{1449}{18} = 80.5$$

$$\text{Class B}: \bar{x} = \frac{\sum x}{n} = \frac{979}{15} \approx 65.3$$

These numbers are consistent with the estimates in part a.

c. Without computing the standard deviation for each class, explain which class has the higher standard deviation and why you think this is so.

Solution The standard deviation measures the spread in the data. Class A has scores that vary from a low of 61 to a high of 94 and Class B varies from a low of 36 to a high of 86. The range for Class A is 33 and for Class B has a range of 50. Since range is also a measure of spread and Class B has a larger range, the standard deviation for Class B is larger.

d. Compute the standard deviation for each class and verify your answer to part c.

Solution To calculate the standard deviation for each class, fill out the table below:

Class A x	$x - \bar{x}$	$\left(x-\bar{x}\right)^2$
82	1.5	2.25
86	5.5	30.25
80	-0.5	0.25
77	-3.5	12.25
83	2.5	6.25
92	11.5	132.25
81	0.5	0.25
61	-19.5	380.25
68	-12.5	156.25
73	-7.5	56.25
86	5.5	30.25
83	2.5	6.25
84	3.5	12.25
94	13.5	182.25
86	5.5	30.25
78	-2.5	6.25
65	-15.5	240.25
90	9.5	90.25
SUM		**1374.5**

Class B x	$x - \bar{x}$	$\left(x-\bar{x}\right)^2$
73	7.7	59.29
71	5.7	32.49
72	6.7	44.89
73	7.7	59.29
64	-1.3	1.69
71	5.7	32.49
77	11.7	136.89
38	-27.3	745.29
58	-7.3	53.29
75	9.7	94.09
85	19.7	388.09
36	-29.3	858.49
86	20.7	428.49
39	-26.3	691.69
61	-4.3	18.49
SUM		**3644.95**

Use the formula for the standard deviation:

$$s=\sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{n-1}}=\sqrt{\frac{1374.5}{18-1}}\approx 8.99 \qquad s=\sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{n-1}}=\sqrt{\frac{3644.95}{15-1}}\approx 16.14$$

Based on these standard deviations, Class A has the larger standard deviation and is more spread out. This is consistent with the answer from Part c.

Practice 3

1. A tennis promoter is comparing two brands of tennis ball to determine which one gives a faster serve. The following data represents the top speeds (in mph) clocked by comparable players using each ball.

   Ball A: 68, 76, 94, 74, 93, 70, 71, 86, 66, 92, 64, 62, 76, 85, 83, 69

   Ball B: 98, 86, 84, 92, 96, 89, 73, 71, 91, 58, 87, 93, 73, 89, 93, 69

a. Without computing the mean, explain which ball has the higher mean and why you think this is so.

b. Compute the mean for each ball and verify your answer to part a.

c. Without computing the standard deviation for each ball, explain which class has the higher mean and why you think this is so.

d. Compute the standard deviation for each ball and verify your answer to part c.

How do you interpret quartiles and percentiles of data?

Key Terms

Five number summary	Quartile	Percentile
Box and whisker plot	Interquartile range	

Summary

There are other calculations that we can do to look at spread. One of those is called **percentile**. This looks at what data value has a certain percent of the data at or below it.

A percentile is a value with that value of percent of the data at or below this value.

For example, if a data value is in the 80th percentile, then 80% of the data values fall at or below this value.

We see percentiles in many places in our lives. If you take any standardized tests, your score is given as a percentile. If you take your child to the doctor, their height and weight are given as percentiles. If your child is tested for gifted or behavior problems, the score is given as a percentile. If your child has a score on a gifted test that is in the 92^{nd} percentile, then that means that 92% of all children who took the same gifted test scored the same or lower than your child. That also means that 8% scored the same or higher than your child. This may mean that your child is gifted.

Guided Example 4	Practice 4
Suppose you took the SAT mathematics test and received your score as a percentile. a. What does a score in the 90^{th} percentile mean? **Solution** This means that 90 percent of the scores were at or below your score (You did the same as or better than 90% of the test takers.) b. What does a score in the 70^{th} percentile mean? **Solution** This tells you that 70% of the scores were at or below your score.	Suppose you took the GRE general test and received your score as a percentile. a. What does a score in the 50^{th} percentile mean? b. What does a score in the 99^{th} percentile mean?

c. If the test was out of 800 points and you scored in the 80th percentile, what was your score on the test? **Solution** You do not know! All you know is that you scored the same as or better than 80% of the people who took the test. If all the scores were low, you could have still failed the test. On the other hand, if many of the scores were high you could have gotten a 95% on the test.	c. If your score was in the 95th percentile, does that mean you passed the test?

There are three percentiles that are commonly used. They are the first, second, and third quartiles, where the quartiles divide the data into 25% sections.

First Quartile (Q_1): 25th percentile (25% of the data falls at or below this value.)
Second Quartile (Q_2 or M): 50th percentile, also known as the median (50% of the data falls at or below this value.).
Third Quartile (Q_3): 75th percentile (75% of the data falls at or below this value.)

To find the quartiles of a data set:

1. Sort the data set from the smallest value to the largest value.
2. Find the median (M or Q_2).
3. Find the median of the lower 50% of the data values. This is the first quartile (Q_1).
4. Find the median of the upper 50% of the data values. This is the third quartile (Q_3).

If we put the three quartiles together with the maximum and minimum values, then we have five numbers that describe the data set. This is called the **five-number summary**.

Five-Number Summary: Lowest data value known as the minimum (Min), the first quartile (Q_1), the median (M or Q_2), the third quartile (Q_3), and the highest data value known as the maximum (Max).

Also, since we have the quartiles, we can talk about how much spread there is between the 1st and 3rd quartiles. This is known as the **interquartile range** (IQR).

$$IQR = Q_3 - Q_1$$

There are times when we want to look at the five-number summary in a graphical representation. This is known as a **box-and-whiskers plot** or a **box plot**.

A box plot is created by first setting a scale (number line) as a guideline for the box plot. Then, draw a rectangle that spans from Q_1 to Q_3 above the number line. Mark the median with a vertical line through the rectangle. Next, draw dots for the minimum and maximum points to the sides of the rectangle. Finally, draw lines from the sides of the rectangle out to the dots.

Notes

Guided Example 5

The first 11 days of May 2013 in Flagstaff, AZ, had the following high temperatures (in °F):

71	59	69	68	63	57
57	57	57	65	67	

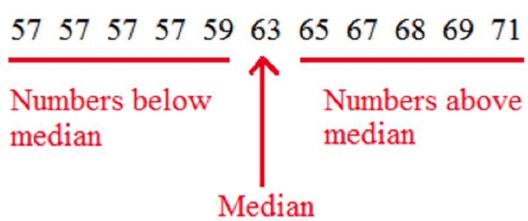
(Weather Underground, n.d.)

a. Find the five-number summary for this data.

Solution Notice that there is an odd number of data we will need to consider as we calculate medians.To find the five-number summary, you must first put the numbers in order from smallest to largest.

57, 57, 57, 57, 59, 63, 65, 67, 68, 69, 71

Now find the median. The number 63 is in the middle of the data set, so the median is 63°F.



To find Q_1, look at the numbers below the median. Since 63 is the median, you do not include that in the listing of the numbers below the median. To find Q_3, look at the numbers above the median. Since 63 is the median, you do not include that in the listing of the numbers above the median.

Looking at the numbers below the median, the median of those is 57so Q_1 = 57°F. Looking at the numbers above the median, the median of those is 68 so Q_3 = 68°F.

Now find the minimum and maximum. The minimum is 57°F and the maximum is 71°F.

Thus, the five-number summary is:

Practice 5

In a recent class, the overall percentages at the end of the semester were recorded.

73	71	72	73	64	71	77	39
38	58	75	85	36	86	61	

a. Find the five-number summary for this data.

Min = 57°F
Q_1 = 57°F
Med = Q_2 = 63°F
Q_3 = 68°F
Max = 71°F.

b. Find the interquartile range.

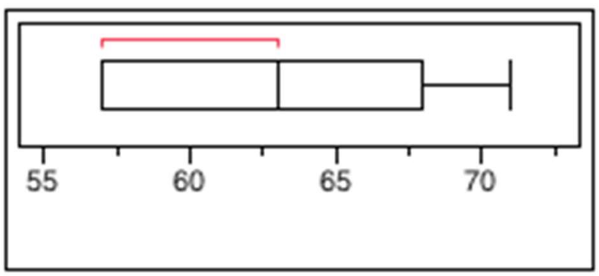
Solution The interquartile range is the difference between the third quartile and the first quartile,

$$IQR = Q_3 - Q_1 = 68 - 57 = 11°F$$

c. Draw a box plot for this data.

Solution Draw a box plot for this data set as follows:

Temperatures in °F in Flagstaff, AZ, in early May 2013



Notice that the median is basically in the center of the box, which implies that the data is not skewed. However, the minimum value is the same as Q_1, so that implies there might be a little skewing, though not much.

b. Find the interquartile range.

c. Draw a box plot for this data.

Guided Example 6

The first 12 days of May 2013 in Flagstaff, AZ, had the following high temperatures (in °F):

71	59	69	68	63	57
57	57	57	65	67	73

(Weather Underground, n.d.)

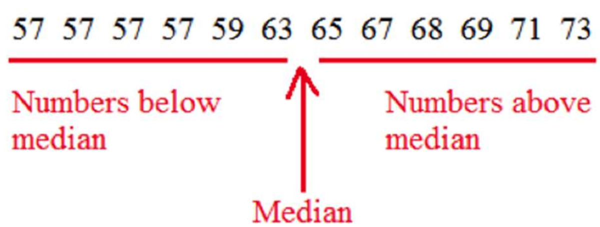
a. Find the five-number summary for this data.

Solution This set of data contains 12 values. To find the five-number summary, you must first put the data values in order from smallest to largest.

57, 57, 57, 57, 59, 63, 65, 67, 68, 69, 71, 73

Then find the median. The numbers 63 and 65 are in the middle of the data set, so the median is

$$\frac{63+65}{2} = 64^{\circ}F$$



To find Q_1, look at the numbers below the median. Since the number 64 is the median, you include all the numbers below 64, including the 63 that you used to find the median.
Looking at the numbers below the median 57, 57, 57, 57, 59, 63, the median of those is

$$\frac{57+57}{2} = 57^{\circ}F$$ so $Q_1 = 57$°F.

To find Q_3, look at the numbers above the median. Since the number 64 is the median, you include all the numbers above 64, including the 65 that you used to find the median.

Practice 6

In a recent class, the overall percentages at the end of the semester were recorded.

82	86	80	77	83	92	81	61	68
73	86	83	84	94	86	78	65	90

a. Find the five-number summary for this data.

<table>
<tr>
<td>
Looking at the numbers above the median 65, 67, 68, 69, 71, 73, the median of those is
$\frac{68+69}{2} = 68.5^{\circ} F$ so $Q_3 = 68.5$°F.
Now find the minimum and maximum. The minimum is 57°F and the maximum is 73°F.Thus, the five-number summary is:

Min = 57°F
$Q_1 = 57$°F
Med = $Q_2 = 64$°F
$Q_3 = 68.5$°F
Max = 73°F.

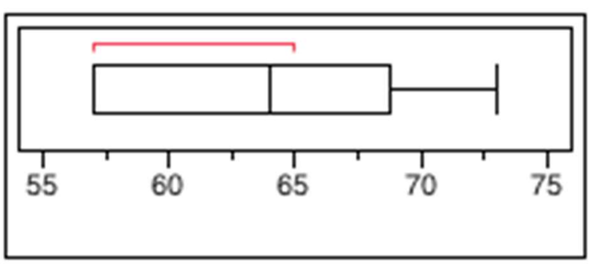
b. Find the interquartile range.

Solution The interquartile range is

$IQR = Q_3 - Q_1 = 68.5 - 57 = 11.5$°F

c. Draw a box plot for this data.

Solution Draw a box plot for this data set as follows:

Temperatures in °F in Flagstaff, AZ, in early May 2013


Notice that the median is basically in the center of the box, so that implies that the data is not skewed. However, the minimum value is the same as Q_1, so that implies there might be a little skewing, though not much.
</td>
<td>
b. Find the interquartile range.

c. Draw a box plot for this data.
</td>
</tr>
</table>