ALGEBRA UNIT ONE: Descriptive Data (2 weeks)

In this unit, students will learn to identify statistical and non-statistical questions. They will draw dot plots, histograms and other graphical representations to visually display data sets. Then, students will calculate measures of center and also analyze the spread of data sets by drawing box plots, finding the interquartile range, and finding standard deviation.

Students should first understand what a “statistical question” is and what it is not. Simply put, a statistical question is one that can be answered by collecting data and where there will be variability in that data. A statistical question anticipates an answer that varies from one individual to another; in doing so, the responses to a statistical question result in a set of data. Data are the variables produced in response to a statistical question. If answers to a statistical question do not predict variability, then the question is not statistical. For example, a student asking themselves, “How old am I?” is not a statistical question; the answer is predictable.

Example 1: Students should begin by simply “recognizing” statistical questions. Teachers should present a group of both statistical and non-statistical questions to students. Student teams or cooperative groups are allowed time to decide whether each question is statistical or non-statistical. Then team responses should be shared classroom wide.

Questions - Statistical or not?

1. How many pets do each of your teachers own?
2. How old is the oldest member of a household?
3. What is your classmate’s favorite flavor of ice cream?
4. How many times does a sixth grader eat (on average) each day?
5. What lunch item is served every day on “Pizza Day”?

Questions one through four are statistical; they can be responded to numerically or categorically and the questions will have varied responses. Question five is a non-statistical question. It is not statistical because the answer would be, predictably, pizza.

Example 2: Teachers should provide students with slides of questions (either via a power point or smartboard) to enhance student ability to identify “statistical” questions. Students will respond individually to whether the question is statistical or not with a thumbs up or thumbs down. If the question is statistical, students should decide whether the data is numerical (quantitative) or categorical (qualitative).

Possible questions:

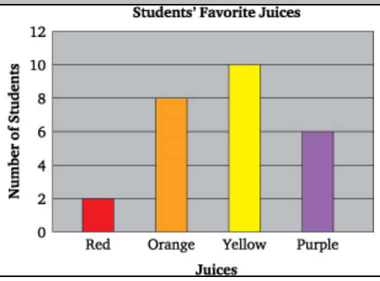
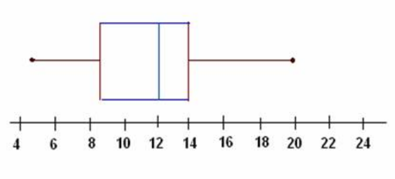
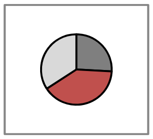
1. At what age do children learn to ride a 2-wheel bicycle?
2. What are the favorite colors of the students in the class?
3. Whom do you admire most?
4. What time do you wake up on school days?
5. What time does your school begin?
6. How many pairs of shoes do you currently own?
7. Which pair of shoes is your favorite?

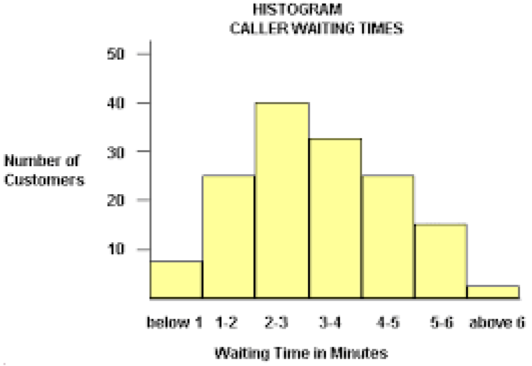
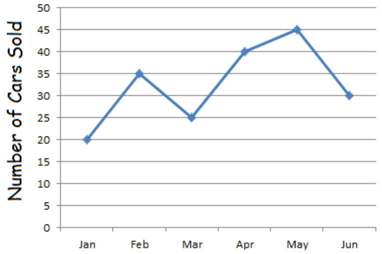
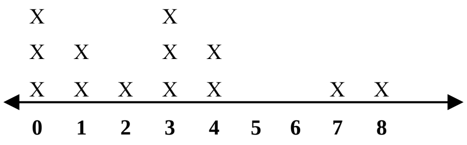
STUDENT REFLECTION: What makes a question numerically "statistical"? Use the word "variability" in your response. Provide an example (with possible responses) to support your view.

To begin working with data, we must first understand the two types of data we will encounter in this unit—categorical data and numerical data. **Categorical data (qualitative)** consists of names, labels or other non-numerical values such as movie preferences, types of animals, types of advertising, colors of medals given to winners of sporting events, etc. Categorical data is usually displayed in circle graphs, dot plots (pictographs) and bar graphs. **Numerical data (quantitative)** consists of numbers such as weights of animals, lengths of rivers, heights of waterfalls, rainfall in a particular city, number of siblings, etc. Numerical data can be broken down into two categories: Discrete and Continuous. Discrete data are values that are isolated points on the number line, think counting. A question that asks "How many pairs of shoes do you own?" will give us a discrete value. Continuous values can take on any number on the number line and refer to measurements like height, weight, temperature, etc. Numerical data is displayed in dot plots, bar graphs, histograms, line graphs, stem-and-leaf plots, and box-and-whisker plots.

<u>Unit 1.1: To display data</u>

As students begin to work with data displays, it is important that they choose an <u>appropriate</u> display for the data set. Below is a quick summary of each of the displays students will encounter in this unit.

Visual	Display	Usage/strength
	bar graph	To **compare categorical or discrete numerical** data. If categorical (like the one pictured), this graph type has no center and no spread.
	box-and-whisker plot or boxplot	To organize **numerical data** into four groups of approximately equal size. Used for large sets of data; appropriate for comparing two or more sets of data in terms of center, shape and spread.
	circle graph	Used to represent **categorical data** as parts of a whole.

<table>
<tr><td> </td><td>Histogram</td><td>Used to compare frequencies of <b>numerical data</b> that fall in equal intervals. Appropriate for displaying large sets of data or data sets with a large range. Strength is that it allows data to be manipulated by varying the intervals; gives an overall picture of the data by intervals without highlighting specific pieces of data within the set.</td></tr>
<tr><td> </td><td>line graph</td><td>Best used to display <b>numerical data</b> that change over time</td></tr>
<tr><td>Grades on a
Science Test
<table>
<tr><th>Stem</th><th>Leaf</th></tr>
<tr><td>7</td><td>2 2 4 5 6 9</td></tr>
<tr><td>8</td><td>1 4 5 7 7 9</td></tr>
<tr><td>9</td><td>0 1 3 5 8</td></tr>
<tr><td>10</td><td>0 0</td></tr>
</table>
Key: 7 / 2 means 72 percent</td><td>stem-and-leaf-plot or stemplot</td><td>Used to organize <b>numerical data</b> based on their digits. Best for small data sets and great for assessing shape. Can be used for comparison by doing back-to-back stemplots.</td></tr>
<tr><td><table>
<tr><th>Interval</th><th>Tally</th><th>Frequency</th></tr>
<tr><td>0 - 9</td><td>卌 卌 卌</td><td>15</td></tr>
<tr><td>10 - 19</td><td>卌 | |</td><td>7</td></tr>
<tr><td>20 - 29</td><td>|</td><td>1</td></tr>
<tr><td>30 - 39</td><td>卌 |</td><td>6</td></tr>
</table></td><td>frequency table</td><td>Used to <b>organize numerical data</b> according to the number of times the item occurs.</td></tr>
<tr><td><table>
<tr><th>Gender</th><th colspan="4">Preferred Program</th></tr>
<tr><th></th><th>Dance</th><th>Sports</th><th>Movies</th><th>Total</th></tr>
<tr><td><b>Women</b></td><td>16</td><td>6</td><td>8</td><td>30</td></tr>
<tr><td><b>Men</b></td><td>2</td><td>10</td><td>8</td><td>20</td></tr>
<tr><td><b>Total</b></td><td>18</td><td>16</td><td>16</td><td>50</td></tr>
</table></td><td>Two-way frequency table</td><td>Used to <b>organize data</b> between two categorical variables.</td></tr>
<tr><td> </td><td>Dot plot (line plot)</td><td>Used to display numerical or categorical data. Good to use for small to moderate sets of data with small to moderate range. Strength is that it is easy to create, highlights the distribution including clusters, gaps, and outliers</td></tr>
</table>

DOT PLOTS (LINE PLOTS)

Dot plots (line plots) is the most basic type of graph for representing data and can be used for categorical or discrete numerical data. It gives students a good visual representation of the shape of the data, where the center might be, and a visual of the data's variability or spread. However, it is not very useful when there are a lot of data pieces because it can be cumbersome to create.

FREQUENCY TABLE and HISTOGRAM

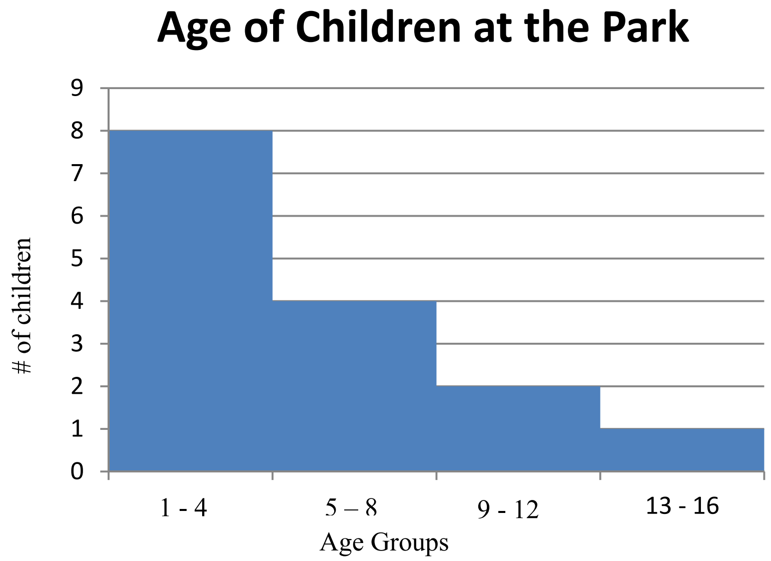
The *frequency* of a data value is the number of times it occurs in a data set. How can you tell the frequency of a data value by looking at a dot plot? It is sometimes more convenient to show data that has been divided into intervals than to display individual data values. A Histogram is a type of bar graph whose bars represent the frequencies of data within intervals. Unlike a bar chart, the bars "touch" and there is no space between the bars indicating the data is continuous.

Example: Make a histogram for the data given- Ages: 12, 3, 8, 1, 1, 6, 10, 14, 3, 6, 2, 1, 3, 2, 7

First, make a frequency table:

Interval	Tally	Frequency How many data values are in this interval?
1-4	𝍸 |||	8
5-8	||||	4
9-12	||	2
13-16	|	1

A histogram is made up of adjoining vertical rectangles or bars**.** In a histogram the horizontal axis typically identifies the topic of the graph and the vertical axis describes the frequency of those observations.



STEM-AND-LEAF PLOT

The following test scores are used to construct a stem-and-leaf plot:

82, 97, 70, 72, 83, 75, 76, 84, 76, 88, 80 81, 81, 82, 82

First determine how the stems will be defined. In this case, the stem will represent the tens column in the scores, the leaf will be represented by the ones column.

When the information is presented, it will be in two parts, the stem and leaf. For instance, 5 | 7 4 would be read as follows: The stem represents fifty, and the leaf has two scores, 7 and 4. Reading that information then gives 57 and a 54.

Since the lowest score is in the 70's and the highest is in the 90's, the stem will consist of 7, 8, and 9. Usually, the smaller stems are placed on top, but they can be arranged from largest to smallest. Leaves should be put in numerical order in the leaf portion. A key must be provided so there is no confusion as to what the stems and leaves stand for. For example, 8|1 could mean 81, but it could also be 810 or 8.1.

Student's Test Scores

Stems	Leaves
7	0 2 5 6 6
8	0 1 1 2 2 3 4 8 8
9	7

Key: 8|1 = 81

If the stem-and-leaf plot were to be rotated 90 degrees (a quarter turn), the graph would resemble a bar graph (which leads to the next type of graph to discuss).

Stems can also be "split" as they are in the example below. If there is a small amount of data it can be easier to see the shape by splitting the stems. The first "5" below refers to 50 – 54 and the second "5" below would have the stems 5, 6, 7, 8, and 9. This way the stems are evenly split. Stems can also NOT be skipped; if there is no value in that stem, leave it blank. Skipping the stem would give a false interpretation of shape and spread.

Comparative Stemplots:

Stemplots can also be used to compare two data sets on the same variable. Ask the students to place their heights on the board. Students could be given a sticky note and they would place the sticky note in the appropriate place on the back-to-back stem and leaf plot on the board. Divide the variable by gender to make comparisons between the two groups.

STUDENT HEIGHTS

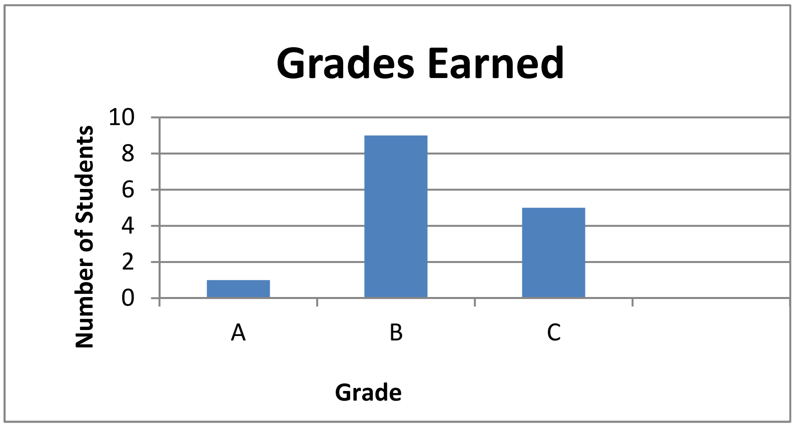
Male		Female
	5	4
9	5	6 8 8
4 4 2	6	0 2 2 4 4
9 8 8 7 6 6	6	5 6 7
2 1 0	7	
5	7	

KEY: 6|2 = 62 inches

BAR GRAPH

Using the same information above, let's construct a bar graph to show how many A, B, C, D, and F's there are. A's are defined as 90 and above, B's from 80 to 89, C's 70 to 79, etc.

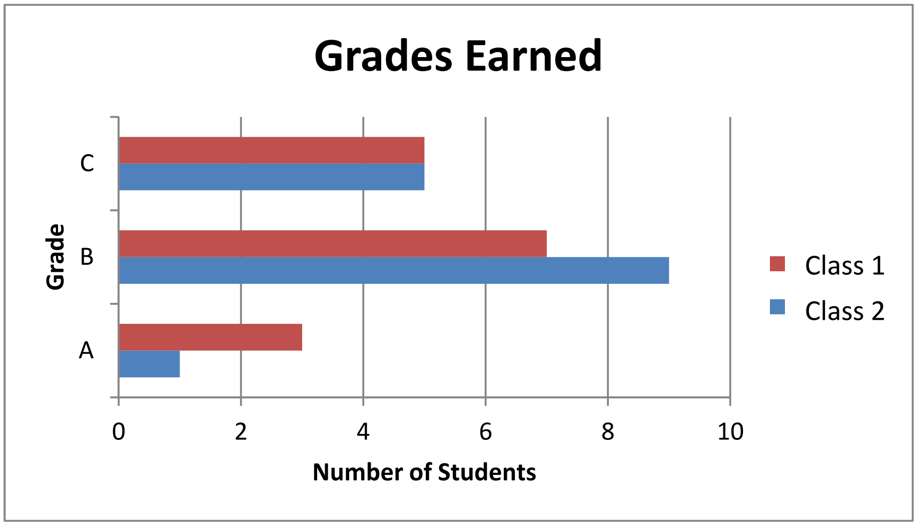
Sometimes, placing bars side-by-side in pairs makes it easier to display the kinds of comparisons you want to show. This is called a double-bar graph.



Let's compare the grades earned in the class discussed so far to another classroom of students. Both the old data and the new data is summed up in the table below:

	Number of A's	Number of B's	Number of C's
First classroom	1	9	5
Second classroom	3	7	5

The double bar graph could look like:



CIRCLE GRAPHS

Skill: interpret data from various formats including circle graphs** and scatter plots**.

A circle graph consists of a circular region partitioned into disjoint sections, each section representing a percentage of the whole. This type of graph shows how a whole is broken into parts. It does not have a center (a centerpoint ,yes, but no center of data) and does not show the spread of data.

Example: A family weekly income of $200 is budgeted in this manner; $60 food, $50 rent, $20 clothing, $20 books, $30 entertainment and $20 other. Construct a circle graph to illustrate that information.

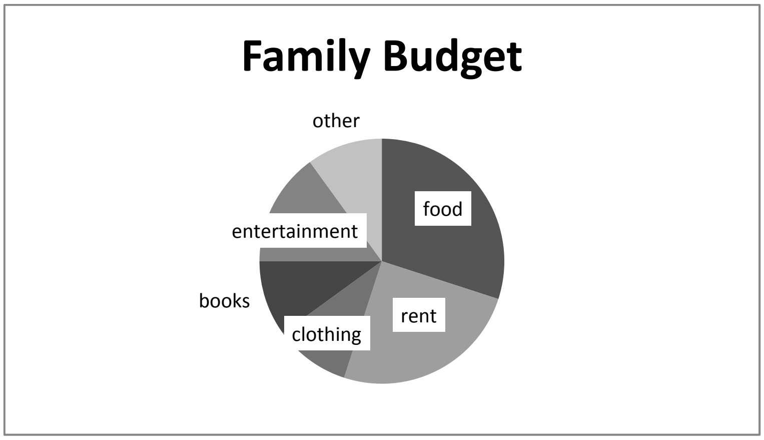
A circle has a total of 360º, therefore 360 represents the total amount of the budget or 100% of the expenses. To fill in the circle graph or pie chart, we have to determine what percent is spent for each expense.

To find that percent, I divide the expense by the total budgeted for the week. $60 out of $200 is budgeted for food. Converting that to a percent, we have 60/200 = 30%. So let's do percents.

Food	60/200 or 30%
Rent	50/200 or 25%
Clothing	20/200 or 10%
Books	20/200 or 10%
Entertainment	30/200 or 15%
Other	20/200 or 10%

The reason the dollar amounts are converted to percents is to determine what percentage of the circle will be dedicated to that category. Since food represents 30% of the pie, find 30% of 360º, which equals 108º $\left(.30 \times 360 = 108\right)$. Doing the same for clothing, 25% of 360º is 90º, 10% of 360º is 36º, and 15% of 360º is 54º.

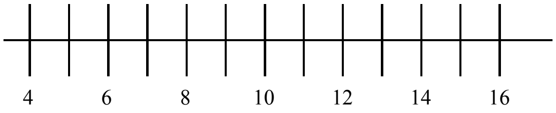
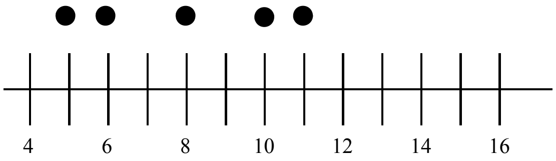
Here is what the pie chart looks like using those degree equivalents.

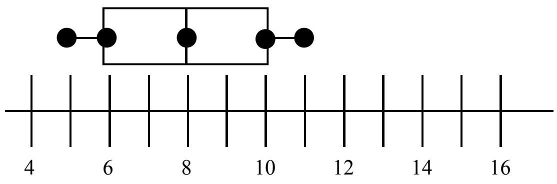


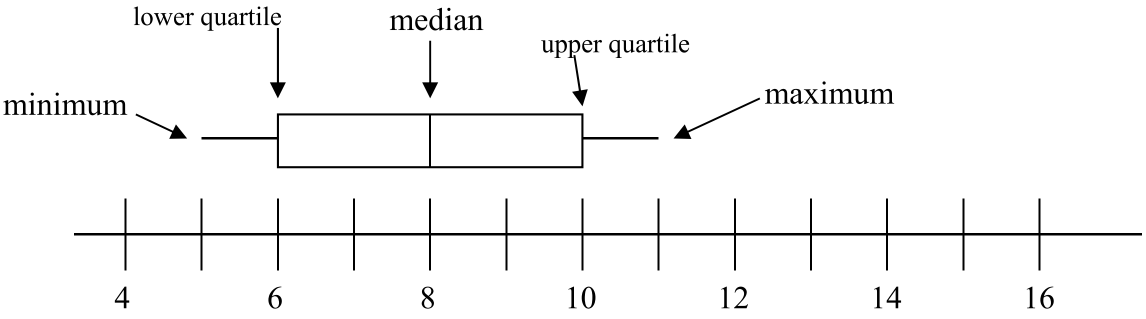
Box-and-Whisker Plot

The purpose of a box-and-whisker plot is to **organize numerical data** into four groups of approximately equal size. Let's take a look at what might be a way of giving notes to your students to help them learn the steps for creating a box-and-whisker plot. As you demonstrate the example in the third column, students are working with you taking the notes. Column 1 allows for guided, group or independent practice once they have an idea how to make a box-and-whisker plot.

Making a Box & Whisker Plot

Steps	Ex: 5 10 7 9 8 6 11
1. Arrange data in increasing order (minimum value & maximum value are the endpoints).	5 6 7 8 9 10 11
2. Find median of the entire list (median value).	5 6 7 8 9 10 11
a. If there is a number in the list that is the middle term, circle it and draw a line thru it. (median)	8 is the median 5 6 7 8 9 10 11
b. If there is not a number that is in the middle, draw a line between the two numbers. (Median is the number halfway between the two numbers.)	Does not apply to this problem
3. Look at the bottom half of the numbers. Find the median of the bottom half of numbers (lower quartile, same as #2 above). Do NOT include the median in either half.	5 6 7
4. Look at the top half of the numbers. Find the median of the upper half of numbers (upper quartile, same as #2 above).	9 10 11
5. Draw a number line that will cover the range of data (evenly spaced marks).	
6. Slightly above the number line place dots at the following points: minimum, lower quartile, median, upper quartile, and maximum.	

7. Draw a box with side borders being the lower and upper quartiles. Draw two lines, one from each side of the box, connecting the minimum point on one side and the maximum point on the other side. Draw a vertical line at the median point from the top to the bottom of the box.	



The data in the box represents the interquartile range – IQR, the average, the middle 50%.

The whisker on the left represents the bottom quartile, the bottom 25%; the whisker on the right represents the top 25%.

The difference between the upper and lower quartiles is called the "**interquartile range**" (IQR).

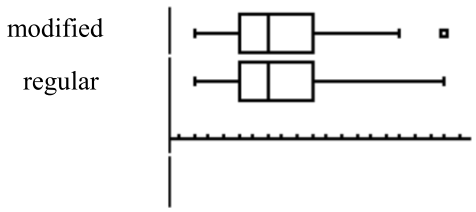
A statistic useful for identifying extremely large or small values of data is called an "outlier". An ***outlier*** is commonly defined as *any value of the data that lies more than 1.5 IQR units below the lower quartile or more than 1.5 IQR units above the upper quartile*.

In our example the lower quartile was at 6, the upper at 10.

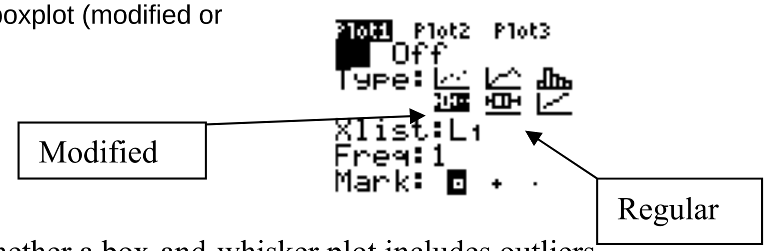
Using that the IQR $=10-6=4$. Multiplying that by 1.5, we have $(1.5)(4)=6$
Therefore, any score below 6 − 6 = 0 is an outlier, as is any score above $10+6=16$. There are no points below 0, so we are OK on the left. There are no points greater than 16, so we are OK on the right and there are no low outliers and no high outliers.

> This would be an ideal place to use technology. Next are the instructions for drawing a box-and-whisker plot on the TI-84. The examples will address outliers.

Entering Data and Drawing a Box-and-Whisker plot on the TI-84



1. STAT
2. EDIT
3. Enter the numbers in List 1 (L_1) or List 2 (L_2)
4. Return to the home screen
5. STAT PLOT
6. Turn Stat Plot 1 on and select the type of boxplot (modified or regular)
7. ZOOM
8. ZoomStat (9)
9. GRAPH



```
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Freq:1
Mark: □ + ·
```

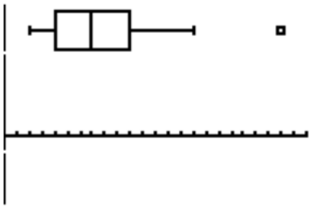
The TI-84 graphing calculator may indicate whether a box-and-whisker plot includes outliers. One setting on the graphing calculator gives the regular box-and-whisker plot which uses all numbers, so the furthest outliers are shown as being the endpoints of the whiskers.
Another calculator setting (modified) gives the box-and-whisker plot with the outliers specially marked (in this case, with a simulation of an open dot), and the whiskers going only as far as the highest and lowest values that aren't outliers.

Find the outliers and extreme values, if any, for the following data set, and draw the box-and-whisker plot. Mark any outliers with an asterisk and any extreme values with an open dot.

20, 21, 21, 23, 23, 24, 25, 25, 26, 27, 29, 33, 40

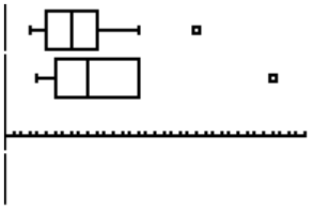
To find the outliers and extreme values, I first have to find the IQR. Since there are thirteen values in the list, the median is the seventh value, so $Q_2 = 25$. The first half of the list is **20, 21, 21, 23, 23, 24**, so $Q_1 = 22$; the second half is **25, 26, 27, 29, 33, 40** so $Q_3 = 28$. Then IQR = 28 – 22 = 6.

The outliers will be any values below 22 – 1.5×6 = 22 – 9 = 13 or above 28 + 1.5×6 = 28 + 9 = 37. The extreme values will be those below 22 – 3×6 = 22 – 18 = 4 or above 28 + 3×6 = 28 + 18 = 46



Another example: L_2 =21, 23, 24, 25, 29, 33, 49

So **I have an outlier at 49 but no extreme values**, so I won't have a top whisker because Q_3 is also the highest non-outlier, and my plot looks like this:



Unit 1.2 and 1.3 To describe and compare the shape of data distributions and the effect of outliers. To use measures of center and spread to describe and compare data sets.

3 Measures of Central Tendency:

1. Mean
2. Median
3. Mode

MEAN

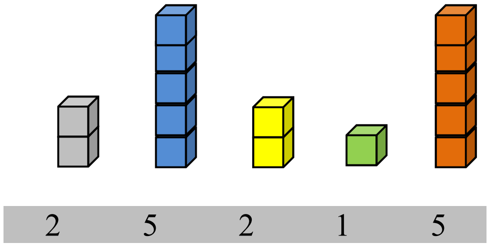
Students need to **conceptually** understand the concept of mean. Be sure to give them time to experience that mean is an "equal distribution" or everyone getting a "fair share".

1. One way to do this is to bring a large bag of treats (eg. M&M's, Jolly Ranchers, etc.) and distribute them to students at random in varying amounts. The students who get nothing will usually begin to comment or complain that they didn't get any and other will say they didn't get their share, others will say they got less than their neighbor . This sets the stage for a great discussion on "equal distribution" or "fair share" where everyone gets the same amount.
2. Another way to do this is give a problem and have students solve it using unifix cubes, blocks, chips, etc.

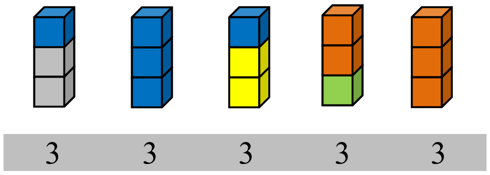
   Example: The table shows the number of students absent from 4th period last week. What was the mean number of students absent per day?

Day	# of students absent
Monday	2
Tuesday	5
Wednesday	2
Thursday	1
Friday	5

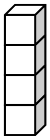
1. Have students model the data using their manipulatives.



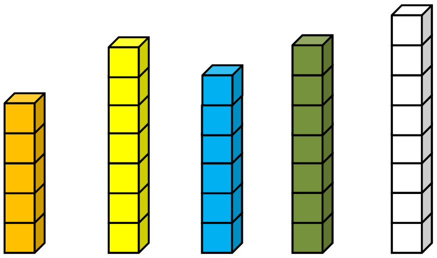
Have students "even out" or "equally distribute" the counters until each column has the same number of counters. (remind students they have 5 columns, Monday – Friday, and must have 5 columns in the end.)



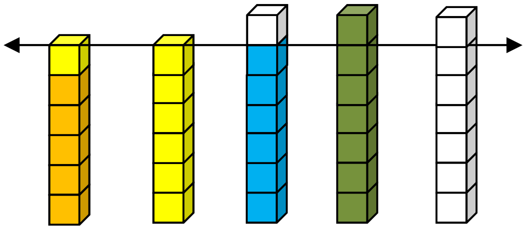
Example: With large groups or whole class, have each student create a stack (of unifix cubes) that represents the number of letters in their last name.
(eg. Long =)
Have students display their stacks together. Have students compute the mean of the letters in their classes last names using only the unifix cubes and redistributing them or evening them out.



Of course you must be ready to discuss remainders and what they represent. Let's say the last names you were working with looked like this:



Once distributed it might look like this.



Since each of the columns are 6 blocks high and a few extra, the mean is 6 and something.

What do we do with the 3 extra blocks? Since we have 3 extra blocks and 5 columns our mean is 6 3/5.

Procedurally, the mean is the one that is probably most familiar; it's the one often used in school for grades. To find the mean, you simply add all the scores and divide by the number of scores. For example, if a student scores 70, 80, and 90 on three tests, the mean is calculated as follows: add the three scores, 70 + 80 + 90 = 240, then divide the sum by the number of data pieces→ 240/3 = 80. The mean is 80. The average is 80.

Example: Find the mean of 72, 65, 93, 85, and 55.

First add those 5 scores together; 72 + 65 + 93 + 85 + 55 = 370
Second, divide that total by the number of scores, 370 ÷ 5 = 74
The mean is 74.

Example: Find the mean of 72, 65, 93, 85, and 25.
The new mean is: 72 + 65 + 93 + 85 + 25 = 340/5 = 68. Note that the low outlier greatly affected the mean by bringing our average down from 74 to 68. **We say that the mean is NOT RESISTANT**. That means that outliers greatly affect the mean.

Example: Five kids just finished bowling one game. The average score of the five kids is 82. What is the total of all 5 scores?

Having a mean of 82 does not mean each kid scored an 82—it means if the scores were distributed equally, they would each have 82. Their total score is $82 \cdot 5 = 410$.

When thinking of the mean, you need to think of the TOTAL units being distributed EQUALLY.

Problem: If the mean is 6, find the missing value for the set of numbers 3, 4, 5, □ , 9.

Method 1

Looking at the data given we know the mean is 6 and there are 5 data in the set.
So the total sum of the data must be 6 x 5 = 30.
Adding the data that we have 3 + 4 +5 + 9 = 21.
The missing value must be 30 – 21 = 9.

□ = 9

Method 2

Knowing the mean is 6, examine each piece of data given with reference to the mean being 6.

$$\begin{array}{cccc} 3 & 4 & 5 & 9 \\ \underline{-6} & \underline{-6} & \underline{-6} & \underline{-6} \\ -3 & -2 & -1 & +3 \end{array}$$

Summing the differences you get -3. Since you are 3 short negative) add 3 to the mean, 6 + 3 = 9. □ = 9

Method 3

$$\frac{3 + 4 + 5 + \boldsymbol{x} + 9}{6} = 30$$

$$\boldsymbol{x} + 21 = 30$$

$$\boldsymbol{x} = 9$$

OR

$$3 + 4 + 5 + \boldsymbol{x} + 9 = 6 + 6 + 6 + 6 + 6$$

$$\boldsymbol{x} + 21 = 30$$

$$\boldsymbol{x} = 9$$

MEDIAN

The median, often used in finance, is the middle score when the data is listed in either ascending or descending order. If there is no middle score, take the two middle scores, add them and divide by 2. It's also referred to as the average of the two middle scores. The median is a resistant measure. This means that the measure is NOT greatly influenced by outliers. Notice that the mean takes every value into account so that one value can greatly affect the average, however, the median ONLY looks at the middle number so it is not affected by very low or very high scores.

Example: Find the median of 72, 65, 93, 85, and 55.

Step 1: Rewrite the data in ascending order: 55, 65, 72, 85 and 93.

Step 2: The middle score is 72; therefore, the median is 72.

*Note: This example was used in the "mean" section and the mean equaled 74, not 72, like the median.

Example*:* Find the median of 72, 65, 93, 85, 74, and 55.

Step 1: Rewrite the data in ascending order: 55, 65, 72, 74, 85, 93
Step 2: Notice there is no middle score.
Step 3: Add the two middle scores together and divide by 2.

72 + 74 = 146, → 146 ÷ 2 = 73. The median is 73.

When thinking about the median, you need to think of the MIDDLE score if they are listed in ORDER.

MODE

The mode is the value point that appears most frequently.

Example: The following are test scores: 55, 64, 64, 76, 78, 81, 81, 81, and 92.
Find the mode.

The score that appears most often is 81.

When thinking about the mode, you need to think of the score that APPEARS MOST FREQUENTLY.

Note: a distribution may have **no mode, one mode or more than one mode.**

Example 1: Find the mode of 55, 64, 64, 76, 78, 81, 81, 81, and 92.
What scores appears most often? The mode is 81.

Example 2: Find the mode of 8, 9, 11, 14, 15, and 17.
What scores appears most often? None of these, so there is no mode.

Example 3: Find the mode of 17, 15, 15, 14, 14, 11.
What scores appears most often? Both 15 and 14 appear twice, so the modes are 15 and 14.

An **outlier** – an extreme value that is much smaller than or much larger than the rest of the data in a data set – can greatly affect the mean. It is important to note "which" measure of central tendency is "most" useful to use.

Measure	Most useful when
mean	the data are spread fairly symmetrically without outliers
median	the data set has an outlier or is greatly skewed
mode	the data involve a subject in which many data points of one value are important, such as election results

These three measures of central tendency, most often referred to as averages, describe a set of data using a single number. By condensing the information like that, the whole picture may not be seen.

The following example shows how using an average or mean to describe a performance may be misleading. Let's say Abe, Ben and Carl each bowl 3 games. Three games later, they all found they had a mean of 80. Here are the scores for each person:

Abe's scores:	80, 80, 80
Ben's scores:	70, 80, 90
Carl's scores:	65, 75, 100

In this case, the mean may not be a good indicator or each person's performance. Looking at Carl's scores, it appears he's a little erratic. It might be difficult to predict what he might score on the next game using the average. Abe, on the other hand, looks pretty stable as he'll probably score an 80 on the next game.

Abe and Ben both have the same average—this example shows that one mean is a pretty good descriptor, which would allow you to predict more comfortably what might happen next. In other words, the mean is doing a pretty good job of describing what's happening.

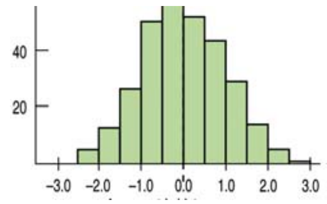
Carl's mean is not as good of a descriptor as Abe's. Although his average is 80 (like Abe's), the mean does not do a good job of describing what is happening. Abe's mean better describes what is occurring than Carl's mean. But, if the scores were not shown, it would not be evident how consistent Abe is and how erratic Carl is because their averages are both 80.

Sometimes, more information is needed. One way to do this is to look at all the scores and try to determine consistency. In math, rather than looking at the whole set of data, only the high and low scores might be examined. It might be someone just had one super high or low score that really affected the mean. In other words, determine the spread of the scores. In statistics, that's referred to as variability. There are three ways to measure this spread or variability.

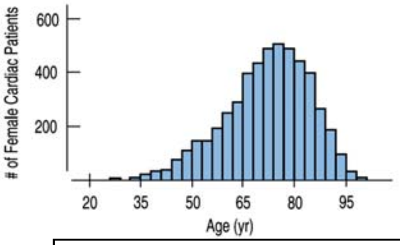
Describing the distribution

1. Shape
2. Outliers
3. Center
4. Spread

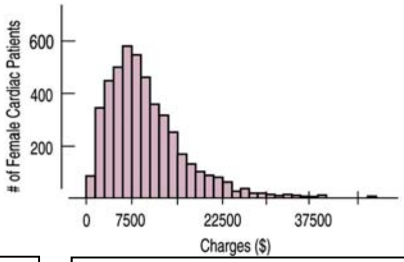
When we are describing the distribution of a data set or comparing two data sets, the four items above should be addressed. Center and outliers were discussed earlier and spread is discussed last. The only other thing to address is shape. When discussing shape, we should state whether the data is symmetric, meaning the data is approximately equally balanced on each side or whether the data is skewed (left or right).



Symmetric – Mirror image. Mean is approximately equal to the median.



Skewed to the Left – Tail is longer on the left. Mean is smaller than the median as the lower values pull down the average.



Skewed to the Right – Tail is longer on the right. Mean is larger than the median as the higher values increase the average.

Four Measures of Spread/Variability

1. Range
2. Interquartile Range (IQR)
3. Mean Absolute Deviation (MAD)
4. Standard Deviation (SD or *s*)

RANGE

The range is just the difference between the top score and the bottom score. The larger the range, the less likely the mean can be depended upon as a good descriptor or predictor. Range is rarely used to describe spread as it is NOT resistant and greatly affected by outliers.

In the last example, the range of Abe's scores was zero. The range of Carl's scores was 35 and Ben's range was 20.

INTERQUARTILE RANGE (IQR)

Students can describe measures of variability in a data set by finding the interquartile range, the mean absolute deviation (MAD) and the standard deviation (SD or *s*).

The interquartile range is found by subtracting the lower quartile from the upper quartile.

Example: Student's Heights (in inches)

60	58	54	56	63	61
65	61	62	59	56	58

When the student data is ordered it appears like this:

54	56	56	58	58	59	60	61	61	62	63	65

The median of the data is **59.5** (59 + 60/2). The significance of this number is that half the data is less than 59.5 and half the data is greater than 59.5.

The Lower Quartile (Q_1) is the median of the lower half of the data. In this example, the lower half of the data is 54, 56, 56, 58, 58, and 59. The median of these six data is **57** (56 + 58/2).

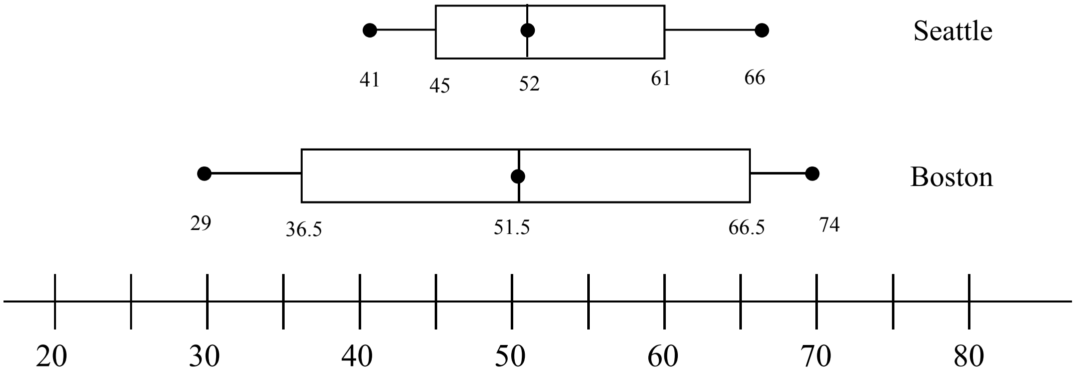
The Upper Quartile (Q_3) is the median of the upper half of the data. In this example, the upper half of the data is 60, 61, 61, 62, 63, and 65. The median of these six data is **61.5** (61 + 62/2).

The Interquartile Range (IQR) is the difference between Q_1 and Q_3. 61.5 – 57 = **4.5**.

The smaller the IQR, the less variability there is in the data. The greater the IQR, the greater the variability in the data.

An interesting reason to study box-and-whisker plots is when we begin to compare two or more plots and examine the data. Following are several examples to highlight this concept.

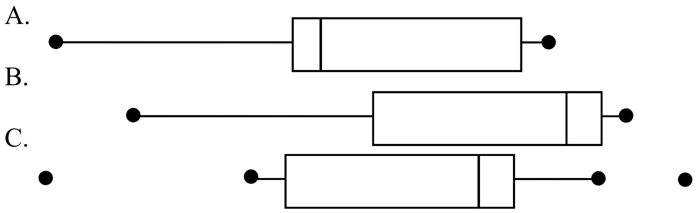
Example: Given below are boxplots displaying the annual temperatures for two cities, Seattle and Boston. What information and generalizations can you see in the plots?



Some information you should note during your discussion with students, could include, but not be limited to:

- The median temperature for Seattle and Boston is essentially the same.
- Boston's data is more spread out.
- Boston's range is greater so its temperature fluctuates more than Seattle's.
- Boston's temperature range is wider than Seattle's.
- Boston has greater high temperatures and lower low temperatures.
- More than 25% of the days in Boston the low temperatures are below Seattle's lowest temperature.
- 1/4 of the days with high temperatures in Boston are higher than Seattle's highest temperature.

Example: Look at the three boxplots below. Even without a scale, what can you say about the 3 temperature plots in general?



General discussions might include:

Plot A
- This city has a lot of cold days (in comparison to the other two).
- ¼ of the days the temperatures are very close (in the lower quartile).
- The number of days of lowest temperatures appear to have a great range.
- The number of days of temperatures in the upper quartile have a great range.

Plot B
- This city has a lot of hot days (in comparison to the other two).
- The variability of cold temperatures is great.
- The variability of high temperatures for ½ the year is small.

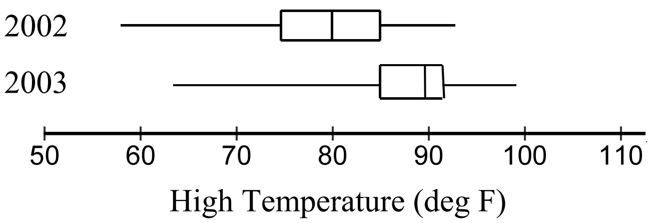
Plot C
- This city has real extremes - outliers.
- Without the outliers, this city has the smallest range of temperatures.

Looking back at the boxplots, if you were told one of the graphs represents temperatures in Las Vegas and another represents temperatures in Hawaii which graphs might represent them? Let students explain why they would choose one plot over another.

The impression of Las Vegas residents is that October 2003 was unusually warm. Half of the days had high temperatures of 89º F or more and fully three-fourths of days were at or above 84º. The coolest high temperature was 63º, but after looking at the raw data that seems like an anomaly when compared to the rest.

Was it really that warm in October 2003? How did it compare with the year before? One of the powers of boxplots is to answer questions about comparing distributions. In this case, we want to compare the highs in October 2003 with those from October 2002. The five-number summary for October 2002 is {59, 74, 80, 84, 92}.

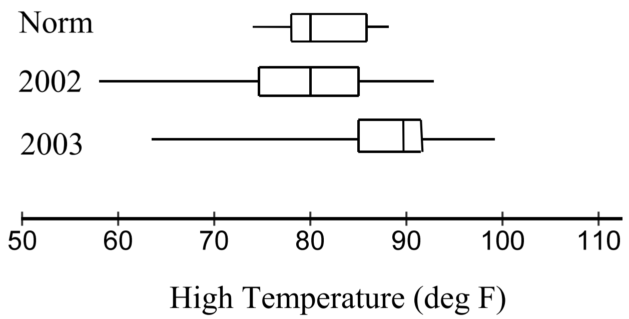
Parallel boxplots for both Octobers 2002 and 2003 are shown at right.



The boxplots clearly show that October 2003 was warmer, on the whole, than October 2002. The median high temperature in October 2003 is 9 degrees warmer than in October 2002. The coolest three-fourths of high temperatures in 2002 were below the first quartile in 2003. The middle halves of each data set don't even overlap!

We know that October 2003 was much warmer than October 2002, but how does it compare with what is considered "normal?" (Normal is the average daily high temperature since 1937.)

Consider the parallel boxplots at right and draw your own conclusions.



Useful suggestions for ways to connect these vocabulary words and concepts with your students way include:

Box and Whiskers Song

Sung to the tune of "Oh My Darling, Clementine"

Put in order
Find the median
Find the median of the top
Find the median of the bottom
Then you draw the whisker plot

Chorus:
Box and whiskers
Box and whiskers
Put the data to the test
Boxes show the middle 50
And the whiskers show the rest.

MEAN ABSOLUTE DEVIATION (MAD)

Students should additionally be introduced to mean absolute deviation (MAD) as a way to gauge variability. MAD is the average of how far away each piece of data is from the mean. Simply put, the greater the relative MAD, the more variability in the data set. The smaller the MAD, the less the variability of the data set as a whole. This topic is covered in middle school and could be reviewed at this point in anticipation of standard deviation.

To compute mean absolute deviation:

- Step 1. Find the mean of the data
- Step 2. For each piece of data, find the distance that data is from the mean and take it's absolute value.
- Step 3. Find the average of the distances from the mean

Example: Period 1 Test Scores

45	45	45	55	55	55	60	60	60	60	65	65
65	70	70	70	70	70	75	75	75	80	85	100

Step 1. Mean (1575/24 = 65.625= 66%)

Step 2. Find the distance each score is from the mean and take it's absolute value.

score	45	45	45	55	55	55	60	60	60	60	65	65
absolute distance from mean	21	21	21	11	11	11	6	6	6	6	1	1
score	65	70	70	70	70	70	75	75	75	80	85	100
absolute distance from mean	1	4	4	4	4	4	9	9	9	14	19	34

Step 3. Find the average of the distances: $234 \div 24 = 9.875$

The mean absolute variation is 9.875

Example: Period 2 Test Scores

score	45	70	70	75	75	80	80	80	80	85	85	85
absolute distance from mean	20	15	15	10	10	5	5	5	5	0	0	0
score	85	90	90	90	95	95	95	95	100	100	100	100
absolute distance from mean	0	5	5	5	10	10	10	10	15	15	15	15

Mean (2045/24 = 85.21= 85%)

225/24 = 9.375 Absolute Mean Deviation

Period 1 has an absolute mean deviation of 9.875; period 2 has an absolute mean deviation of 9.375 (similar, but slightly lower). What does this tell us about the data sets? First of all it tells us that according to this measure of variance, these two data sets are relatively similar. Overall, the data in each set is clustered reasonably close together. Outliers, students should understand, can have a great effect on the mean absolute deviation.

STANDARD DEVIATION (SD or s)

What is standard deviation?

Standard deviation is the average of the deviations from the mean. It shows how far the data spreads out from the mean. <u>It is the average amount that the data varies from the mean.</u>

Formula: $s = \sqrt{\dfrac{\sum(x_i - \bar{x})^2}{n-1}}$

Shortcut formula: $s = \sqrt{\dfrac{\sum x^2 - \dfrac{(\sum x)^2}{n}}{n-1}}$

SD Properties:

1. s = 0 only when there is no spread. Ex:: data: 5, 5, 5 – no spread!
2. s, like $\bar{x}$ (the mean), is NOT resistant. It is affected by outliers.
3. s is small if the observations are close to the mean, large if they are far from the mean. The more spread out the data, the larger they are.
4. The sum of the deviations from the mean are always equal to zero. This is why we square the deviations in the formula.

Example: Data: 2, 3, 5, 6, 9, 11 $\bar{x}$ = 36 / 6 = 6

X	x - $\bar{x}$	(x-$\bar{x}$)^2
2	2 – 6	16
3	3 – 6	9
5	5 – 6	1
6	6 – 6	0
9	9 – 6	9
11	11– 6	25

$\sum(x - \bar{x})^2 =$ 60

$s = \sqrt{\dfrac{60}{5}} = \sqrt{12} \approx \boxed{3.46}$

Students will need an understanding of the difference between measures of center (mean, median, and mode) versus a measure of variability (range, IQR, SD). These measures are typically taught as a unit (often in one day) however, it would be advantageous, in terms of student understanding, to separate them from the beginning of the lesson.

Measures of center describe the data with a single numerical value; however, a measure of variability, describes how the data differs (or varies) from other data values in the set.

Simple examples can help your students to see the difference between measures of center and measures of variability.

Example 1: The following data set represents the ages of people who attended the 70th birthday party of your grandmother at the senior center:

70, 75, 80, 72, 82, 81, 73, 78, 82

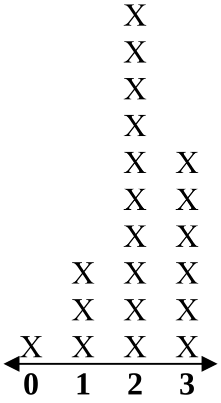
Measures of Center	***Measure of Variability***
Mean = 693/9 = 77	Range = 82 – 70 = 12
Median = 78	IQR = 81.5 – 72.5 = 9
Mode = 82	Standard Deviation = 4.61

The measures of center reasonably describe the data. If someone was to ask you, "How old were the people who attended your grandmother's birthday party?", any one of the measures of central tendency would be a "reasonable" representation of the data although in this case the mean and median are better than the mode. However, the range does not reasonably describe the data in one number; were the people who attended grandmother's birthday party around 12 years old? No, all the range tells us here is that the people who attended are elderly. The standard deviation tell us that on average, the ages deviated on average about 4.6 years from the mean of 78.

Other data sets where the range is significantly different from the measures of center provide students with an understanding that sets "measures of center" apart from "measures of variability".

Example 2: Provide students with a dot plot of data.

Student Scores on Math Constructed Response



Ask students to consider the data shown in the dot plot of constructed response scores among students.

The following questions are appropriate discussion questions:

- How many students are represented in the data set for students constructed response scores? How do you know the sample size?
- What are the measures of center (mean, median, and mode)? What do these values mean? How do they compare with one another?
- What is the range (measure of variability) of the data set? What does its value mean?
- If you have to choose one number to DESCRIBE the data set, which measure would you use? Defend your choice.
- In this example, why do you think the measures of center and the measure of variability are so similar? Explain your reasoning.

Student Reflection: Identify the measures of center. Identify a measure of variability. In what ways are they alike? Different? Which measure of center best describes a set of data?

Example C: A physical representation of each piece of data in the set is helpful when students are beginning to create box plots. One simple way to achieve this is by using sticky notes to represent each individual piece of data in the set.

The data set below was achieved by the teacher asking each of 32 students for their age in months. Each student wrote their age (in months) on a sticky note provided by the teacher. The sticky notes were then placed in order by the students in the front of the classroom.

130	130	131	131	131	132	132	132
133	134	136	136	137	137	138	139
139	139	140	140	141	141	142	142
142	143	144	145	147	148	149	150

Five Number Summary

Minimum: 130

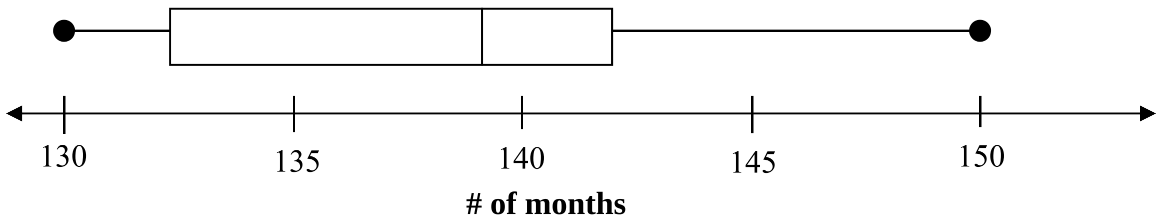
Quartile 1 (Q1): 132 + 133/2 = 132.5

Median: 139

Quartile 3 (Q3): 142

Maximum: 150

Ages in Months of a Class of 32 Sixth Grade Students

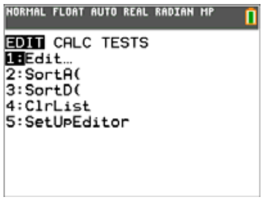


Have students complete the following to demonstrate their understanding of the data:

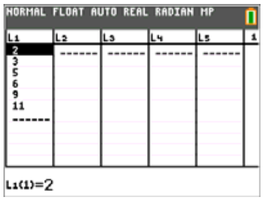
1. Each 1/4th of the data is equal to ______ students. This is because _____ divided by 4 is ______.
2. The youngest 1/4th of students is between _____ and _____ months old.
3. The second youngest one quarter of the students are between ____ and ____ months old.
4. The oldest one quarter of students is between _____ and ____ months old.
5. The _____ quartile has students that are between 139 and 142 months old.
6. The most "spread out" 1/4th of the data is the ____ quartile. I know this because ____.
7. The data is most "concentrated" in the _____quartile. I know this because ____.
8. The median class age in months is _____ months.
9. One half of the class is from ______ to ______ months old (remember this can be answered with three different, yet correct, answers).

Student Reflection: Identify one way to display data. Describe the strengths of your data display and what kind of data it is best suited for. Describe the limitations of your data display. Share your thoughts with a teammate.

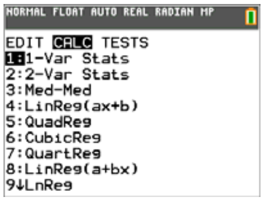
Technology: Make sure to introduce finding summary statistics on the graphing calculator. Place the data in a list and use the 1-VAR Stats feature. Press STAT to enter the data in a list. Then press STAT again and arrow right to CALC to use the 1-Var Stats feature. Note that the standard deviation is listed as S_x. Ignore the σ_x as that uses the formula divided by n, not n-1.



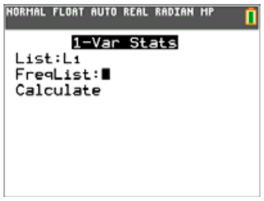
```
NORMAL FLOAT AUTO REAL RADIAN MP
EDIT CALC TESTS
1:Edit…
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor
```



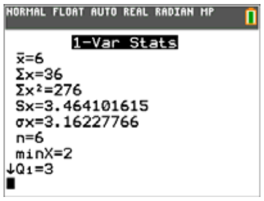
```
NORMAL FLOAT AUTO REAL RADIAN MP
L1     L2     L3     L4     L5     1
2      ------ ------ ------ ------
3
5
6
9
11
------
L1(1)=2
```



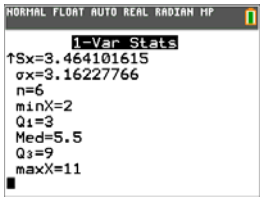
```
NORMAL FLOAT AUTO REAL RADIAN MP
EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
8:LinReg(a+bx)
9↓LnReg
```



```
NORMAL FLOAT AUTO REAL RADIAN MP
      1-Var Stats
List:L1
FreqList:█
Calculate
```



```
NORMAL FLOAT AUTO REAL RADIAN MP
      1-Var Stats
 x̄=6
 Σx=36
 Σx²=276
 Sx=3.464101615
 σx=3.16227766
 n=6
 minX=2
↓Q1=3
█
```



```
NORMAL FLOAT AUTO REAL RADIAN MP
      1-Var Stats
↑Sx=3.464101615
 σx=3.16227766
 n=6
 minX=2
 Q1=3
 Med=5.5
 Q3=9
 maxX=11
█
```

BIVARIATE CATEGORICAL DATA

Unit 1.4 and Unit 1.5 To read and interpret two way frequency tables. To interpret relative frequencies in the context of the data (joint, marginal, and conditional) and recognize possible associations and trends in the data.

Categorical data are data that take on values that are categories rather than numbers. Examples include male or female for the categorical variable of gender and football, basketball or baseball for the categorical variable of favorite sport. When the data consists of two responses from each variable, the data is called **bivariate categorical** data (bi means two and there were two different categories of data). A **two-way frequency table** is used to summarize bivariate categorical data. The number in a two-way table at the intersection of a row and column of the response of two categorical variables represents a **joint frequency.** The total number of responses for each value of a categorical variable in the table represents the **marginal frequency** for that value.

Example:
Juniors and Seniors at your high school were asked if they plan to attend college immediately after graduation, seek full-time employment, enter the military or choose some other option. A random sample of 100 students was selected from those who completed the survey.

	Attend College	Full-Time Employment	Military	Other Options	Totals
Juniors	27	5	4	9	45
Seniors	25	13	2	15	55
Totals	52	18	6	24	100

- The shaded cells are the **marginal frequencies.** They are located around the "margins" of the table and represent the totals of the rows or columns of the table.
- The non-shaded cells *within* the table are called **joint frequencies**. Each joint cell is the frequency count of responses from the two categorical variables located by the intersection of a row and column.

1. Find the relative frequencies for each of the cells.
   To do this, divide each value by the total amount: 100

	Attend College	Full-Time Employment	Military	Other Options	Totals
Juniors	27/100 = 0.27	5/100 = 0.05	4/100 = 0.04	9/100 = 0.09	45/100 = 0.45
Seniors	25/100 = 0.25	13/100 = 0.13	2/100 = 0.02	15/100 = 0.15	55/100 = 0.55
Totals	52/100 = 0.52	18/100 = 0.18	6/100 = 0.06	24/100 = 0.24	100/100 = 1.00

2. Based on the relative frequency table, what is the relative frequency of students who indicated they would seek out full-time employment?
   0.18 or 18% (this is the marginal frequency of the full-time column)

3. What is the relative frequency of Seniors looking at Other Options?
   0.15 or 15% (this is the joint frequency of Seniors and Other Options)
4. A school website article indicated that, "A Vast Majority of Students from our School Plan to Attend College." Do you agree or disagree with this statement? Explain.
   A majority is over 50% and since the "Attend College" column has a total of 52%, I would agree with statement. A clear majority (52%) of the students plan to attend college after high school.
5. Do you think juniors and seniors differ regarding after graduation options? Explain.
   There are some interesting differences. More Juniors stated they would be attending college than Seniors. Seniors had more students choosing Full-time Employment and Other Options than Juniors.

Conditional Relative Frequencies: A conditional relative frequency is the ratio of a joint relative frequency and the related marginal frequency. With conditional frequencies, there is a limitation or condition which is usually preceded by the word, given. We are no longer interested in the whole sample and only interested in a portion of it. This is why we divide by the related marginal frequency.

Example: Use the relative frequency table above.

6. What is the relative frequency of the student attending college, given they are a junior?
   Notice the word, "given" tells us it is a conditional frequency question. We are only interested in the Juniors, so we take the joint frequency of 0.27 and divide it by the marginal frequency of 0.45 and get 0.27 / 0.45 = 0.60 or 60%.
7. Given they are a senior, what is the relative frequency of the student attending college?
   Now, we are interested in the seniors only so take 0.25 / 0.55 = 0.45 or 45%.
8. What is the relative frequency they are a senior, given they are looking at Full-Time Employment?
   This time our subgroup is Full-time Employment. Take 0.13 / 0.18 = 0.72 = 72%.