Chapter 4
Statistics

The first time that students are officially introduced to statistics is in grade 6; up until now the focus has been on the gathering and representation of data. They have learned about measurement and data, making bar graphs in grade 3 and line plots in grade 5. Students will build on this knowledge as they explore the concepts of center and variability, to be developed in later grades as a major theme of statistics.

What is a Statistical Question?

Recognize a statistical question as one that anticipates variability in the data related to the question and accounts for it in the answers. 6.SP.1

In grade 6 students begin to engage with statistical analysis. We start with the question: what is meant by *a statistical question*? A statistical question is one whose answer involves *variability*; a question that generates a range of answers, with some description of the likelihood of the correctness of any of the answers. The input may also be variable.

For example, my local lumber distributor may have a wide variety of kits to make garden sheds. If I want to build a garden shed, I will decide on the measures of the shed, and then go to the lumber yard to buy the kit that makes that shed. This is called a *deterministic* problem: there is only one set of input conditions, and only one kit that makes that particular shed. However, the lumber distributor has a different problem: there may be hundreds of homeowners who want to build garden sheds of a variety of sizes. The distributor wants to sell everyone the shed that meets their needs, and not have any left over in the stockroom. So, she may gather information on the gardens in the neighborhood in order to estimate the set of sizes that are in demand, and for each such size, the likelihood of the demand for that size. This is the nature of statistics: a *distribution* of inputs and an output of a collection viable answers, each associated with a number. In this case, the answer consists of the collection of available kits, and a recommended stocking number for each kit. Another characteristic of statistical analyses is that the result is not perfect - the distributor will not meet every demand with no kits left over, some customers will be dissatisfied, and some kits will have to be restocked. But the analysis, if done right, gives a guess that minimizes the number of dissatisfied customers and the number of kits to be restocked.

To make clear the distinction between *deterministic* and *statistical*, here is another example. We may ask "for how many days over the past year has the sun risen before 7:00 AM?" This is a deterministic question, for there is only one number (it can be found by consulting an almanac) that answers the question. But now, if we ask our students to estimate the number of days over the past year in which *they* arose before 7:00 AM, we will get a variety of responses which we will then analyze to find out what is the *typical* wake-up time for students, and how broad the range is from the typical. Now, we may ask: "How long did it take me to get from home to school today?" This is not a statistical question because there is a single, definite answer. "How long did it take the students at my school to arrive from home?" is a statistical question because the answer will change (vary) from student to student. This variation (difference in answer from one student to another) arises because of the several modes of travel (walking, biking, riding the bus, etc.) and the different distances travelled.

Statistical questions generate data of two types. The first type is called *numerical data* (also called *quantitative data*) because the responses to the question are measurements or counts (quantities). If you posed the statistical question to the students at a particular school "How long did it take to arrive from home today?" you would get a variety of numerical answers, most likely measured in minutes. However, if you asked the same children the statistical question "What is your favorite color?" you would get several different answers such as "blue," "green," or "pink" that are not measurements or counts. This illustrates the second type of data, called *categorical* (or *qualitative*) because it involves a non-numeric description or quality about the data.

A word of caution is in order. While all numerical data involves numbers (because they are measurements or counts), some categorical data might be numerical, for example, phone numbers, athlete's jersey numbers and PINs, One cannot say anything meaningful about two different people because the phone number of one person is larger than the other or that one player is better or worse based solely on the number on their jersey.

Design of Experiments

Summarize numerical data sets in relation to their context, such as by:

a. reporting the number of observations;

b. describing the nature of the attribute under investigation, including how it was measured and its units of measurement. 6.SP.5ab

An **experiment** is a procedure to be used to better understand an attribute of a given population. To design an experiment, we

a) identify the population;

b) identify the attribute to be studied and its measure, including listing the range of possible outcomes, in terms of that measure.

c) select a sample size: the *sample* is the subset of the population for which you will determine the measures of the attribute being studied;

d) determine a bias free way of selecting the sample, and measure the attribute for each member of the sample;

e) analyze the data using statistical tools so as to assess the likelihood of each of the possible outcomes.

Basically, this is what this chapter is about. Indeed it is what Statistics is about, and in subsequent grades, the tools of analysis become more and more incisive. Let us illustrate:

Example 1. Suppose the district nutritionist wants to obtain information on the breakfast-eating habits of the elementary school students in the district (of which there are 2300) in order to create the most appropriate choices for snack time. She knows that students who have had little or no breakfast will need a substantial snack, and students who have had a robust breakfast will make do with a glass of milk. She also wants to come close to meeting everyone's need with minimal waste of food. How does she most efficiently make the best guess as to what data should be collected?

SOLUTION. Let's follow the above outline.

a) the population is the set of all 2300 elementary school students.

b) the attribute is: *breakfast content*. There are many ways to measure the content of a meal: caloric content, protein content, balance (among the major food groups), satisfaction of the participant. The researcher is just starting this study, so focuses on one measure: caloric content, and assumes that the range will be between 0 and 750 calories.

She makes the following table relating the caloric content, as reported by the student, to the recommended snack.

Breakfast	**Snack**
0-200 calories	milk, fruit, crackers, vegetables
201-450 calories	milk, fruit, crackers
451-650 calories	milk and fruit
651 or more calories	fruit

c) She might, on one particular day, ask all of the students "what did you have for breakfast?" for she knows how to convert that information into caloric content. But she is looking for the typical response from the students, not just the data for one particular day. But questioning 2300 students on a sequence of days does not fit within her financial or time budget. So she decides to ask this question of a subset of the students on three randomly selected days. The size of the subset she chooses is 5%, or 115 of the students: she will ask the question 3 times, and then record the data.

d) Our researcher has made an attempt to discount bias, by choosing at random 5% of the students, and asking them the question on three different occasions. Does this overcome the possibility of bias? Probably not, since *students who come to school regularly* and *students who come sporadically* are different collections of students, and the method selected preselects the first population. However, she could feel that the conclusion of the study should respond to the needs of those students who come to school regularly.

e) When the data are collected, she will record the number of responses in each of the categories of the table above. Using that she will be able to estimate the total caloric content of the snack that will meet the needs (generally) of the students.

Clearly, in the above example, or any other example you may conjure up, there are choices to be made. The population and the attribute to be studied are specified by the goal of the study. After that, the measure of the attribute and the size of the sample chosen are up to the researcher. The question then comes up: how valid are the conclusions? The subject of statistics provides a large variety of measures of validity of conclusions that are dependent upon sample size, number of categories of the attribute and so forth. In 6th grade the intent is to introduce students to statistical methods, saving for later the validity of those measures.

Section 1. Measures of Shape

Summarize numerical data sets in relation to their context, such as by:

a. reporting the number of observations;

b. describing the nature of the attribute under investigation, including how it was measured and its units of measurement. 6.SP.5ab

Understand that a set of data collected to answer a statistical question has a distribution which can be described by its center, spread, and **overall shape**. *6.SP.2*

Display numerical data in plots on a number line, including **dot plots, histograms***, and box plots. 6.SP.4*

When we have numerical responses to a statistical question, we often want to "see" the data. An easy graph to make for such numerical data is called a *dot plot*. Some students may recall *line plots* from earlier grades used to represent data, but with an x instead of a dot. Dot plots and line plots are essentially the same, but in sixth grade we start to analyze the illustrated data more deeply. To create a dot plot, first draw a number line and then place a dot above the number line at the location of each data value. If a value is repeated, this is represented by placing another dot above the previous instance(s) of that value. This type of graph allows us to identify clusters (data points together in a group), gaps (intervals without any reported values), peaks (data where there are more responses than for nearby values), and outliers (values that are significantly different from the rest of the data). This is illustrated in the next example.

Example 2. Jayden and several friends are having a fitness contest. They count the number of jumping jacks that each one of them can perform in one minute. The results are: 36, 35, 34, 28, 40, 35, 30, 35, 33, and 29. Create a dot plot for the data and identify any clusters, gaps, peaks, and outliers.

Here these terms are to be understood by common usage. Later they will have mathematical definitions, but often intuition is a better gauge of validity.

SOLUTION. This is provided in the dot plots below. The top figure is the dot plot and the second figure identifies the features of the dot plot.

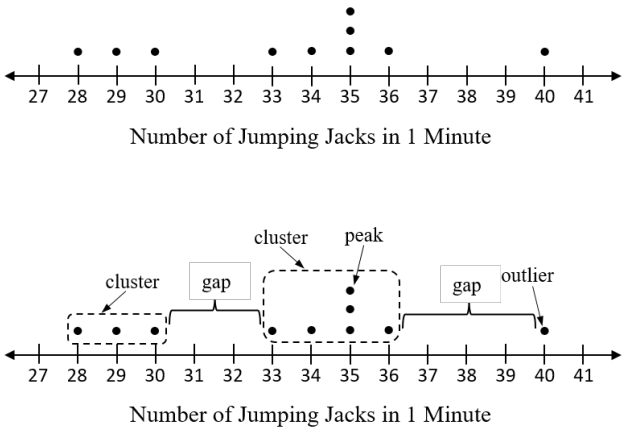


Figure 1. Dot Plot and Dot Plot Showing Features

A word of caution about *outliers*. We have identified the value 40 as an outlier, because it is far removed from the rest of the data. Note that *far removed* is pretty vague: if the value were 38 or 39, would it still be an outlier? Shouldn't the three points on the left also be considered as outliers? To resolve this vagueness, there is a statistical test (the 1.5 IQR test, which will be introduced later on). If we apply this test here, 40 would be identified as an outlier. The three points on the left are not part of the main cluster, and their existence pulls the data to the left, and that may be why 40 shows up as an outlier. So where are the outliers? Are they the three values to the left or the one value to the right, or both? Only the context can help decide on the response to this question. Maybe the outlier to the right is indicating a need for further research, and the trio of points on the left are all outliers.

The point is that the issue of outliers cannot be resolved by looking at the data abstractly, either by eyeballing, or by the 1.5 IQR test. These strategies identify *potential* outliers. We can identify true outliers only by reference to the context. Do we know that Jayden and his friends are representative of the school population? If so, then the 40 is an important outlier, and that person should be put on the jumping jack team. But maybe Jayden and his friends are especially non-athletic: in this case the 40 could be a less than average score. The message here is: Given a specific problem, we may use a statistical test to identify potential outliers, but to justify eliminating an outlier as

an aberrant data value we must do so in the context of our experiment and its goals.

Example 3. In another fitness challenge, Jayden and friends see how many pushups they can do in one minute. The results are: 17, 15, 16, 16, 13, 18, 15, 14, 16, and 15. Create a dot plot for this data and comment on the shape of the graph.

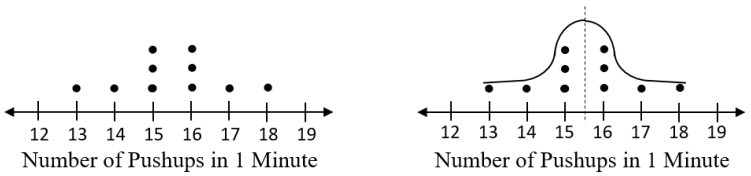


Figure 2. Dot Plot and Dot Plot Shape

SOLUTION. The image on the left of Figure 2 is the dot plot of the data. This plot is symmetrical about a vertical line half way the values 15 and 16. Also, those values have the largest values. We say that the data are *bell-shaped*: the image on the right illustrates that shape.

For larger data sets, a dot plot can be cumbersome, and not very revealing. Another way to visualize data is a *histogram*, drawn off of a *frequency table* summarizing the results. Histograms are commonly used in the media to illustrate data sets. To make the frequency table, first, arrange the numbers in groups (called *bins* or *classes*) that are all the same width. This can be accomplished by finding the range (the interval between the highest and lowest values), and then dividing by the desired number of groups to get the width of each bin (referred to as the *intervals*, but sometimes called *class width*). Then count the number of data values from the set that fall into each bin. This is called the *frequency* or *tally*. This completes the frequency table. From this table, make the histogram by drawing rectangles above each interval so that the height of the rectangle represents the frequency (number of data values) in that interval. The right side of one rectangle touches the left side of the next rectangle to demonstrate that where one bin or class ends, the next one begins.

The word *desired* is deliberate, for it is the researcher who will decide on the format that provides the most information.

Example 4. Recall the jumping jack data of Example 1: 36, 35, 34, 28, 40, 35, 30, 35, 33, 29. Create a frequency table and histogram representing these data.

SOLUTION. Figure 3 shows the Frequency Table and Histogram.

Figure 4 below shows how the heights of the bars in a histogram correspond to the number of dots in each interval of the dot plot.

Note that, although the histogram gives a quick snapshot of the shape of the data, it misses the within-bar details. For example, for the third and fourth bars, the data are to the left of the bar. Noticing this, the researcher might redefine the bins as 28-32, 33-37, 38-42; the corresponding histogram gives a better representation of the data: a cluster to the left, a central cluster, and a single data point to the right.

Histograms and dot plots give a visual summary of the data. The dot plot retains each of the original data values. In a histogram, without knowing the original data, we can only say how many data points are in the interval but we cannot say precisely what they are. The selection of representation, and the size of the bins is left to the researcher to best show the data.

With both the dot plot and histogram, we can tell how many total data values there are. For the dot plot, we simply count the number of dots since each dot represents one data point. For the histogram, we add up the frequencies (heights) of each bar to find how many data points are in the set.

Jumping Jacks Intervals	Frequency / Tally	Total
28 – 31	|||	3
32 – 35	|||||	5
36 – 39	|	1
40 – 43	|	1

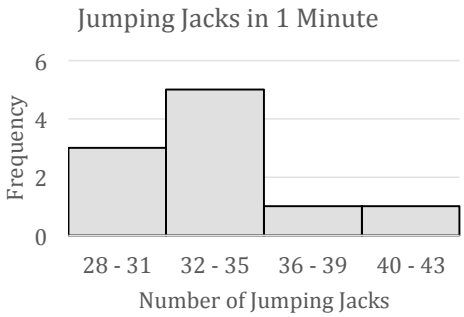


Figure 3. Frequency Table and Histogram

Figure 4

The shape of the histogram tells a lot about the data distribution; thus there are terms to describe the principal characteristic of a shape.

For the number of pushups (Example 3), the graph has a symmetrical bell shape (often referred to as the *bell-curve* or *normal distribution*). If the histogram looks somewhat like a bell shape but one tail is heavier / longer than the other, we say that the graph is skewed in the direction of the longer tail (*skewed left* or *skewed right*). These possibilities are illustrated in Figure 5. It is important to be able to interpret the shape of a histogram in terms

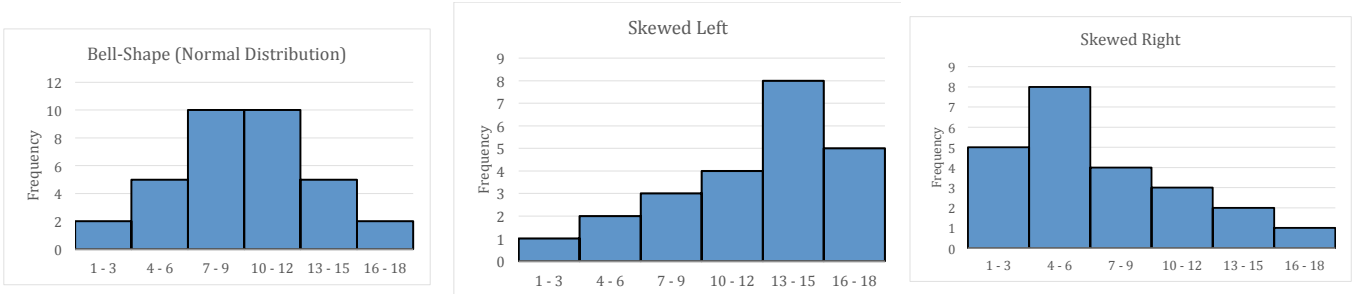


Figure 5. Shapes of Histograms

of the context. For example, suppose that the histograms of Figure 5 are showing us the distribution of grades in a particular exam in three different classes. For the Figure to the left, about half the students will get letter

grades of C, one-quarter will get grades of D or F, and one-quarter will get grades of A or B. However, for the middle histogram, about half the class will get grades A or B, and the rest will earn a C or lower. In the histogram to the right, half the class has failed the exam, about one quarter will receive a C, and one-quarter a B or an A. In all three cases, the average grade might be the same. However, teacher B will have to pay more attention to the weaker students, and teacher C will have to work to move all students upwards, with special attention to the stronger students.

Section 2. Measures of Center

Understand that a set of data collected to answer a statistical question has a distribution which can be described by its **center, spread, and overall shape**. *6.SP.2*

Recognize that a measure of center for a numerical data set summarizes all of its values with a single number while a measure of variation describes with a single number how its values vary. 6.SP.3

Summarize numerical data sets in relation to their context, such as by c: Giving quantitative **measures of center (median and/or mean)**, *and variability (interquartile range and/or mean absolute deviation), as well as 6.SP.5c*

When dealing with numerical data sets, we often want to know what is the center (or middle) of the data set. This is meant to be the single number that best represents the data set. There are three such measures in common use, to be chosen according to the goal of the researcher. These are: a) the arithmetic average of the data, b) the number that divides the data, when put in numerical order, into two pieces of the same size, c) the number that occurs the most often. In statistics these are respectively called: a) *mean*, b) *median*, c) *mode*. They each serve different purposes, which we now describe.

The Mean

The most common measure of center is the *mean*. The mean is the arithmetic average, often referred to simply as "average." Many people are familiar with the procedure of computing the mean: add up all the data values and then divide by the number of data values. The significance of the mean or average is derived from this procedure. When you sum up all the numbers, you are collecting the total of all the values. Then, when you divide by the number of data values, you are giving each person a fair share of the total. This may be seen in Examples 5 and 6 below. In later grades students will learn of a measure of distance between data sets. In terms of this distance, the mean is the single-value data set that comes closest to the given data.

Example 5. After a class party, four friends find that they each have won some candies: Aiden has 4, Ben has 5, Chloe has 9, and Diego has 2. They decide to share their candies equally, so they put them together for a total of 20. Then, in turn, each child takes one candy until they are gone (a fair share). Since there are 4 participants, everyone will have 20/4 = 5 candies. So, the average (or *mean*) of the numbers 4, 5, 9, and 2 is 5. Another way to view the average is to note that since the average is 5, if Chloe gives one of her candies to Aiden and three candies to Diego, then each child will have exactly 5 candies.

Given two numbers a and b, the *average* of the two numbers is $(1/2)(a + b)$: it is the midpoint of the line segment on the number line joining a to b. The mean is an extension of this concept to any set of numbers.The extension is best explained in terms of *fair share*. When Mary and Maria go trick-or-treating, they decide that they will pool their acquisitions and split them evenly. Suppose Mary receives a pieces of candy, and Maria receives b. Then a fair, or even, split of the spoils gives each $(a + b)/2$ candies. Suppose next year, Juan and Jason join the team and accept the rule of division of spoils. Once again Mary gets a pieces of candy, Maria gets b pieces of candy, and the haul for Juan is c, and for Jason, d. Then, in all, the team has collected $a + b + c + d$ pieces of candy, and a fair share is one-fourth of that: $(a + b + c + d)/4$.

Based on these examples, if N kids join this group and accept the fair-share agreement, and the jth kid collects a_j

pieces of candy, then a fair share is

$$\text{Mean} = \frac{a_1 + a_2 + a_3 + \cdots + a_N}{N}.$$

Another way to view the mean is as a balance point: the sum of the distances of the data points from the mean for those points below the mean, is equal to the same sum for all the points above the mean.

Example 6. Let's revisit the example of the four children sharing candies. If Aiden puts his 4 candies on the table and then Ben puts his 5, there are 4 + 5 = 9 candies. If they were to split them evenly between the two children, they would each get 4.5 candies. This can be seen by putting a dot on a number line at 4 and another dot at 5. These two dots would balance each other if we put a fulcrum (triangle) at 4.5 as illustrated in Figure 6:

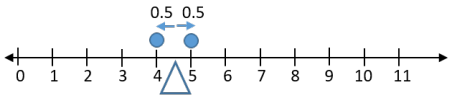


Figure 6

Note that the two dots balance each other out as they are each 0.5 units away from, but in opposite directions from, the mean of 4.5. We call the distance between a data point and the mean an absolute deviation, which is a positive number since it is a distance. This illustrates that the mean of 4 and 5 is 4.5.

When Chloe adds her 9 candies to be shared, this new addition to the sum of candies is higher than the previous average (4.5) so it will make the new mean larger. The average of 4, 5, and 9 is 6. This is illustrated below. Note how the two arrows (absolute deviations) pointing to the left of distances 1 and 2 balance out the one arrow (absolute deviation) pointing to the right of length 3. If Chloe were to take three of her candies and give two candies to Aiden and one candy to Ben, then all three children would each have 6 candies, 6 being the mean.

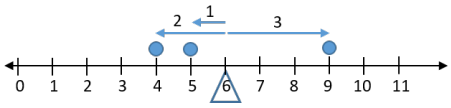


Figure 7

Finally, when Diego adds his 2 candies, for the diagram to balance, the mean or balance point for all four children changes again. Since the new number added (2) is less than the mean at the previous step (6), the mean will have to decrease in order to balance: the mean of 4, 5, 9, and 2 is 5. To balance out the dots on the number line, the triangle needs to be placed at 5 (see Figure 8). Note how the two absolute deviations to the left (of length 1 and 3) and the absolute deviation to the right (of length 4) from the mean of 5 balance each other out. The data value at the mean of 5 does not cause the number line to tip in either direction since it is exactly over the mean and has an absolute deviation of 0. If Chloe were to take 4 of her candies and give 3 to Diego and 1 to Aiden, then each child would have an equal share of 5 candies.

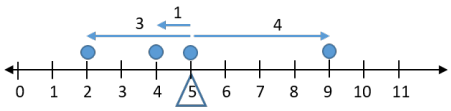


Figure 8

Example 7. Find the mean for the timed jumping jack data (exercises per minute) from Examples 1 and 3 as seen in the dotplot of Figure 1.

SOLUTION. The actual data are: $36, 35, 34, 28, 40, 35, 30, 35, 33, 29$. We have indicated in Figure 9 the deviation from the mean (indicated by the triangle). The mean is the sum of the values divided by the number of data values:

$$\text{Mean} = \frac{36 + 35 + \cdots + 29}{10} = 33.5 .$$

Revisiting the dot plot, we can see that the mean or balance point is 33.5 and that the sum of the absolute deviations on the left of the mean (5.5 + 4.5 + 3.5 + 0.5 = 14) and the sum of the absolute deviations on the right of the mean (0.5 + 1.5 + 1.5 + 1.5 + 2.5 + 6.5 = 14) are the same.

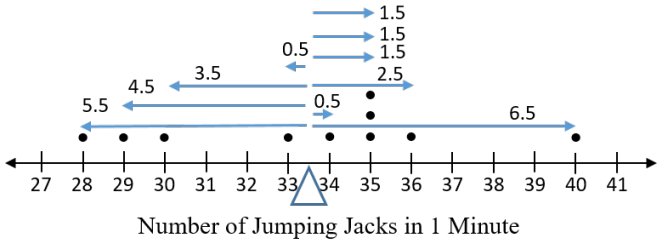


Figure 9

It may be worthwhile to observe that what is happening here is an application of Archimedes' *Principle of the Lever* (see Figure 9). If the classroom has the apparatus available, one might illustrate this with a ruler above a fulcrum. Select 8 whole numbers (say between 1 and 20) at random. Place tokens of equal weight along the ruler at the numerical value of each number chosen. Now move the fulcrum to the position at which balance is achieve. Calculate that this position is the mean of the selected numbers.

The Median

In addition to the mean, there are another two measures of center; here we discuss the *median*. This is a number that divides an *ordered* data set into two parts with an equal number of values in each part. If there are an odd number of data points, the median is the number right in the middle. If there are an even number of data points, the median is the number halfway between the two middle values (their mean).

Example 8. Earlier, we looked at an example studying the number of jumping jacks that Jayden and friends could perform in a minute. The data set as given is the first line of Figure 10 below. To find the median of the data, we first put the values in order, generally from lowest to highest, as in the second line of Figure 10. Since there are 10 data values, the median will be halfway between the two middle values of 34 and 35, namely 34.5.

Original List	36	35	34	28	40	35	30	35	33	29
Sorted List	28	29	30	33	34	35	35	35	36	40

Figure 10

Note the importance of ordering the data: if we were to simply find the halfway point between the two data values in the list as originally given, 40 and 35, we would get 37.5. This can not be the actual median because only one data value was larger than 37.5 while nine were smaller so 37.5 is not truly the middle value.

Finding the median for this data set can be visualized by creating a strip of paper with equal sized portions for each data value. Next, write the data values in increasing order in each of the spaces. Now fold the strip of paper in half; where the crease appears is the median. Since there are an even number of values in our data set, the crease will be on the border between two data points (34 and 35) so we take the number halfway between them, 34.5. In this way, 34.5 will have half (the lower 5) of the data values below it and half (the remaining upper 5) above it.

If there are an odd number of data values, then the median will be the middle value once the data has been put in order.

Recall that the mean for this data set is 33.5. Since the median is larger, if tells us that there are more data points to the right of of the mean than to the left. Go back to Figure 3 to see this.

Example 9. Riley's basketball team played 13 games this season. The number of points Riley scored in each game is 2, 4, 4, 5, 3, 4, 8, 3, 7, 6, 3, 6, and 5. Find the median number of points Riley scored for the season.

SOLUTION. When put in increasing order, the data are 2, 3, 3, 3, 4, 4, 4, 5, 5, 6, 6, 7, 8. Since there are an odd number (13) of data points, the median will be the number that cuts the data into two equal halves. This will be the 7th data value so that there will be 6 numbers higher and 6 numbers lower. The 7th data value is the third instance of the number 4, so the median is 4.

Again, the median can be visualized by making a strip of paper with equal spaces for each data point. Finding the median is as easy as folding the paper in half and locating the crease, on the number 4 in this case.

The mean for these data is 4.6. Since the median is 4, we can conclude that when Riley is good, he is very good, but he's good only half the time.

The Mode

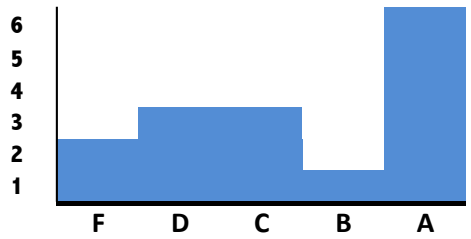
The third measure of center is called the *mode*. This is the number that appears more often than any other number(s). For the example of Jayden and the jumping jacks, the mode is 35 because it occurred 3 times, which is more than any other number of jumping jacks in the data. The mode is easy to pick out from the dot plot because it is the value that has the highest peak.

Now, the mode may not be unique: there may be two data values that have the maximum number of appearances. In the case of Riley's basketball season (Example 9), the values 3 and 4 both occur at the maximum number (3) of times. In this case, we say that the data are *bimodal* with modes 3 and 4.

Example 10. The midterm grades in my class with 15 students are as in the first line of the table below. In the second line, these are ordered from lowest to highest.

Grades	71	62	23	91	93	71	87	54	62	100	95	91	68	91	62
Ordered	23	54	62	62	62	68	71	71	87	91	91	91	93	95	100

In the second line above we see that the largest number of repetitions of a grade is 3, and occurs twice: at 62 and 91. In this case, taking the average of the these two values (76.5) is useless, because it is nowhere near any specific value. After some thought, we might conclude that these data are *bimodal* with modes of 62 and 91. Looking at te historgram (below) we see this hypothesis supported



The fact that a distribution is bimodal is often very useful: in this case, I could conclude that, for some reason, the class divides into two groups: those who get it and those who don't. This is valuable information, leading to an examination of what is not working for the group clustered around the grade of 62.

In Example 5, the values were 4, 5, 9, and 2. Since no value occurs more often than any other, we say that there is no mode. The median is 7, and the mean is 5.

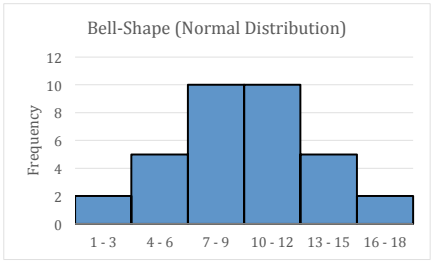
An important feature of the mode is that it can be used on both numerical (quantitative) and categorical (qualitative) data. The mean and median make sense only for numerical data.

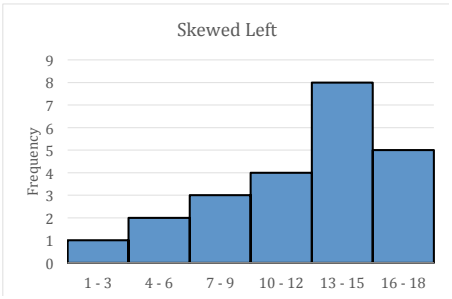
Because the mean and median both measure a kind of middle, students may wonder which is the best measure to use. When data are close to a symmetrical bell-shaped curve, the mean and median (and mode) are very close (if not the same). For the pushup data (Example 3) both the mean and median are 15.5 (and there are two modes: 15 and 16). In such cases, either of these measures is a good measure of center and accurately represents the typical values.

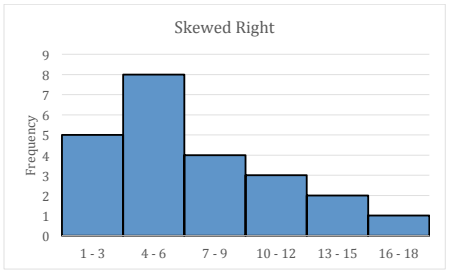
However, in cases where the mean and median are significantly far apart, we may need to look more deeply to decide which is meaningful.

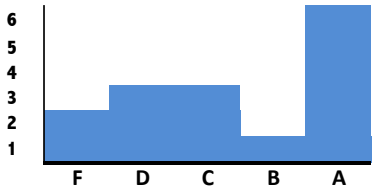
Example 11. Consider the histograms of Figure 5 and the one directly above. Compare the means, medians, modes.

SOLUTION. For convenience, we have copied Figure 5 below. The top left figure is symmetrical about the central line, so that the median and the mean coincide at the value 9.5. For the figure to the right, the median is at 12, but the mean is somewhat higher, and the reverse is true for the left bottom figure. The right bottom figure is also skewed to the right, but the meaningful attribute of these data is that they are bimodal.









Keep in mind that median is that value on the base axis such that the area of the left part of the histogram is the same as area to the right. When the data are skewed, the median is usually to the side of the mean in the direction of the skew. In other words, if the data are skewed left, the mean will generally be to the left of the median. This is illustrated in the next example.

Example 12. Emilia worked for her neighbors on the weekends for the past three months. The amounts, in dollars, earned each week were 6, 9, 10, 12, 8, 12, 9, 9, 15, 10, 20, and 24. Find the mean, median, and mode for this data set. Which measure best represents the center?

The sum of these values is 144 and there are 12 of values, so the mean is 144/12 = 12. Next, put the data in order: 6, 8, 9, 9, 9, 10, 10, 12, 12, 15, 20, and 24. The median is 10 and the mode is 9, since it occurs more often than any other data value.

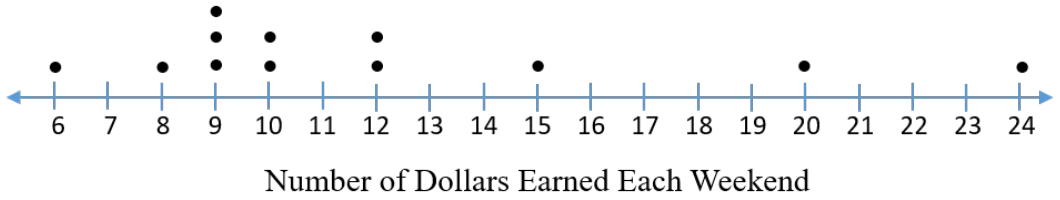


Figure 11

A dot plot or histogram of the data shows that these data are skewed to the right. When dealing with skewed data, the median value is often closer to the typical values, the "center", than the mean. In this case, two-thirds of the data is within 2 units of the median, 10, while only one third of the data is within 2 units of the mean, 12. Figure 11 shows three data points that are not at all typical of her weekend work schedule: 15, 20 and 24. These qualify for consideration as "outliers," and perhaps Emilia can justify that, saying that those were weekends on which her neighbors had parties, and so Emilia way paid time and a half. In the next section we will describe a quantitative test for data points to be outliers, and you should check that according to this test, only 20 and 24 qualify. Although it is called a test for outliers, we should call it a test for *potential* outliers, for we cannot dismiss any data point without understanding its in-context significance.

For example, suppose this same data set represented the scores of 12 candidates, on an exam of 25 questions, for status as astronaut. In this context we are looking for the outliers, and possibly only the grade of 24 is what we are looking for.

To check the relationship of these measures of centrality, let's consider extreme results in a particular example. Suppose that we select 99 responses to the request: give me an integer between 0 and 100.

- If the data set is the set of numbers {1, 2, 3, . . . , 98,99} then the distribution is uniform, the mean and the median are both 50 and there is no mode.
- If the data set consists of ninety-eight 1s and one 99, then the data are skewed left, the mean is 1.99 and median and the mode are both 1
- If the data set consists of one 1 and ninety-eight 99's, then the data are skewed left, the mean is 98.1 and median and the mode are both 99.

Section 3. Measures of Variability

Understand that a set of data collected to answer a statistical question has a distribution which can be described by its center, **spread** *and overall shape. 6.SP.2*

Recognize that a measure of center for a numerical data set summarizes all of its values with a single number while a **measure of variation** *describes with a single number how its values vary. 6.SP.3*

Display numerical data in plots on a number line, including dot plots, histograms, and box plots. 6.SP.4

Summarize numerical data sets in relation to their context, such as by
c) Giving quantitative measures of center (median and/or mean) , and variability (interquartile range and/or mean absolute deviation), as well as 6.SP.5c

Summarize numerical data sets in relation to their context, such as by
d) Relating the choice of measures and variability to the shape of the data distribution and the context in which the data were gathered. 6.SP.5d

After learning how to exhibit the center of the data by using the mean, median and mode, the next step is for students to learn about the *spread* (or *variability*) of the data. The first, and easiest, measure is the *range*. This is given by the highest and lowest data values. While the difference between these two values is easy to compute, it only tells us how far apart the extreme values are. It does not give any indication about the spread of the data values between the two extremes.

Mean Absolute Deviation

The *mean absolute deviation* (MAD) is a measure of variability (or spread) of the data that uses each data value. Now the range (from lowest value to highest) gives a measure of variability, but it is not descriptive of the overall spread of the data. For example, the lowest and highest values may be far from the mean, while *all* other values are very close. In a way, the MAD is an expression of the overall deviation from the mean. Since the MAD automatically accords the mean as the significant measure of center, in data sets where this is not true, the MAD has little value.

Students learn the mean absolute deviation in 6th grade in preparation to learn about the standard deviation of data and its uses in later grades, once square roots have been mastered. The standard deviation is the most frequently used measure of variability for data.

To compute the MAD, we recall the work done with the mean as a balance point. There we introduced absolute deviations, the distance of each data point to the mean. This can be found by taking the absolute value of the difference between the data point and the mean: $|x - \overline{x}|$. The mean absolute deviation is simply the mean of these absolute deviations for all the data points.

Example 13. Recall the jumping jacks Examples 1 and 3. The dot plot with absolute deviations is repeated below in Figure 12. The sum of the absolute deviations to the left of the mean and to the right of the mean were 14 and 14, confirming the mean as the balance point. These absolute deviations sum to 28, and when divided by the number of data values (10), gives a mean absolute deviation of 2.8. Note that no single data value is 2.8 units away from the mean (which is usually the case) but that this is an average of all the distances from the mean.

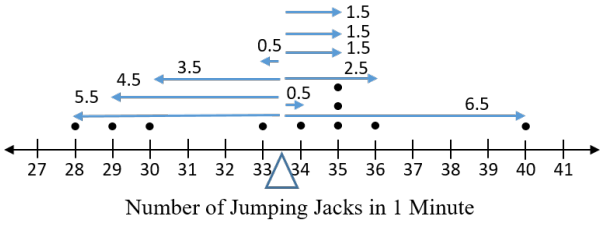


Figure 12

Example 14. Jayden's friends also counted how many sit-ups they could do in 1 minute. The data has been put in order: 31, 32, 32, 32, 33, 33, 34, 35, 36, and 37. Make a dot plot of these data and then compute the mean absolute deviation. Use the MADs to compare the variability between the number of jumping jacks and sit-ups

the children can do in 1 minute.

SOLUTION. See Figure 13. The mean of the data is 33.5 and is represented by the triangle below the number line. The absolute deviation of each data value has been written by the point.

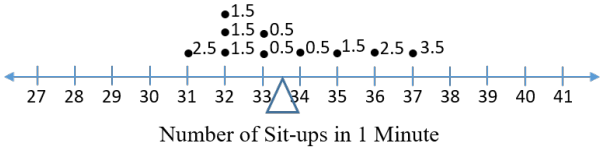


Figure 13

The sum of the absolute deviations is 2.5 + 1.5 + 1.5 + 1.5 + 0.5 + 0.5 + 0.5 + 1.5 + 2.5 + 3.5 = 16. When divided by the number of data values, 10, we arrive at the MAD of 1.6. Note that the MAD for the number of sit-ups is smaller than the MAD of 2.8 for jumping jacks. Since the jumping jacks data has more variability, the MAD is higher than for the sit-ups data.

IQR and 5 Number Summary

Just as there were multiple ways to measure center (mean and median), there are other ways to measure variability. In contexts where the median seems to be a better choice of data center, there is a measure of variability (called the *interquartile range*, IQR), using the median. Recall that, once the data are put in increasing order, the median splits the data into two parts, each with the same number of data points. We can also divide the data into four equal parts, each of which is a *quartile*.

The quartile numbers (the endpoints of these quartiles) separate the data set into four parts with equal numbers of data values in each part. The division points, starting at the lower end, are denoted by Q1, Q2, Q3 in increasing order. Note that Q2 is just the median of the data set. The *5 number summary* of a data set are these 5 numbers:

Minimum Q1 Q2 (median) Q3 Maximum

The values of the 5 number summary are found as follows:

- Order the data from smallest to largest. The number furthest left is the **Minimum**, and the number furthest right is the **Maximum**.
- Find the median of the whole data set. If there are an odd number of data values, remove the median so that there are two equal pieces remaining. If there are an even number of data values, the median already cuts the set into two equal halves. This is **Q2**.
- Find the median of the lower half, called quartile 1 (denoted Q1) .
- Find the median of the upper half, called quartile 3 (denoted Q3).

Example 15. Find the quartiles for the jumping jacks data: 36, 35, 34, 28, 40, 35, 30, 35, 33, 29.

SOLUTION. First put the values in order from lowest to highest: 28, 29, 30, 33, 34, 35, 35, 35, 36, 40. There are an even number of data points so the median is 34.5. This is Q2. The lower half of the data consists of 28, 29, 30, 33, and 34 and the median (Q1) is 30. The upper half of the data consists of 35, 35, 35, 36, and 40 and has a median of 35 (Q3). Therefore, the five number summary is

Minimum: 28 Q1: 30 Q2 (median): 34.5 Q3: 35 Maximum: 40.

Let's return to our visualization of the median as the halfway point on a strip of paper with the data values. If we were to fold the strip of paper into fourths, we would find that there is a crease at Q1 of 30 and another crease at Q3 of 35. This helps show that the data was cut into four equal pieces by the quartiles.

The distance from Q1 to Q3 is called the *interquartile range, IQR*, and represents the interval for the middle 50% of the data: from 25% below the median to 25% above. The IQR for the jumping jacks data is $35 - 30 = 5$, meaning that the middle 50% of the data are all within 5 units of each other.

Example 16. Find the 5 number summary, the range and the interquartile range for the sit-up data:

31, 32, 32, 32, 33, 33, 34, 35, 36, and 37.

SOLUTION. Note that we have already ordered the data. The median is 33 (halfway between the two 33s) and divides the data into a lower half of 31, 32, 32, 32, 33 and an upper half of 33, 34, 35, 36, 37. The medians of these two halves give the quartiles, namely Q1 = 32 and Q3 = 35. The interquartile range is $35 - 32 = 3$. The minimum is 31, the maximum is 37, and so the range is 6. The 5 number summary is

Minimum	Q1	Q2 (median)	Q3	Maximum
31	32	33	35	37

and the range and IQR are:

Range = 6, IQR = 3.

Box Plots

Box plots provide a visual image of the 5 number summary, plotting a box to represent the IQR, a vertical bar for the median, and horizontal bars to the extreme values (maximum and minimum). To create a box plot, use a number line and draw a rectangle from Q1 to Q3. Then draw a (vertical) line segment inside the rectangle at the median. Finally, draw a horizontal line segment from the lower edge of the box to the minimum and another (horizontal) line segment from the upper edge of the box to the maximum (these are called the whiskers). The dot plot of the jumping jacks data and its corresponding box plot are shown below.

Example 17. Draw a box plot for the jumping jacks.

SOLUTION. The five number summary (see Example 15) consists of: 28 (min), 30 (Q1), 34.5 (median), 35 (Q3), and 40 (max). Figure 14 is the box plot for the jumping jacks data.

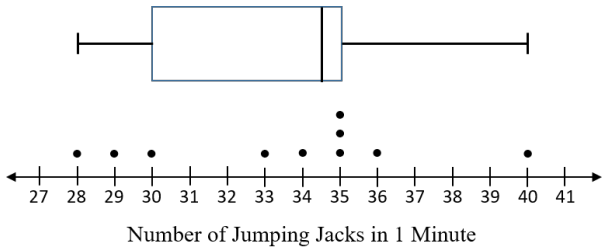


Figure 14

To read a box plot, it is important to keep in mind how the data are represented: a) the same number of data values fall within each quartile; b) the breadth of the quartile box tells us the breadth of variability in the quartile. This can be confusing to students: each box has the same number of data points, what we are seeing is the spread of

those data points. Here, the spread from the minimum to the left edge of the box represents the lowest 25% of the data (the left whisker) and it has a modest spread. The second quarter of the data (the lower portion of the box) is wider than the first part of the graph because the data points are more spread out there. The next quartile (the upper portion of the box) is very narrow because the data is clustered around the point 35. Finally, the highest 25% of the data, represented by the right whisker, is also very spread out as the data stretches up to the maximum of 40. Note that the vertical bar representing the median is not in the center of the box because the data are skewed the left. Also note that the middle 50% of the data, the interquartile range represented by the box, is from 30 to 35.

Example 18. Draw a box plot for the sit-up data and then compare with the box plot for jumping jacks.

SOLUTION. Earlier we found that Q1, median, and Q3 were 32, 33, and 35 respectively. With the minimum of 31 and maximum of 37, the five-number summary becomes 31, 32, 33, 35, and 37.

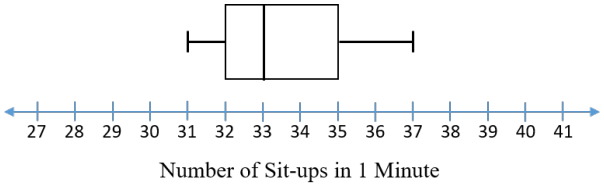


Figure 15

Graphing the two box plots on the same axes, and labeling which graph is which, we are better able to compare the two data sets (see Figure 16).

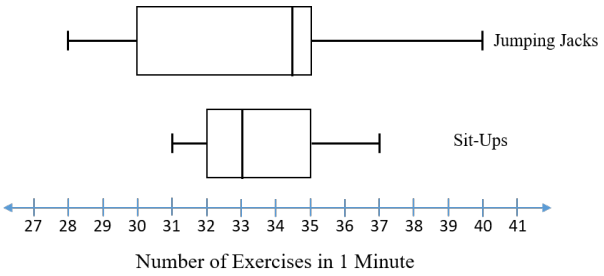


Figure 16

There are many comparisons we can make about the data based on the graphs:

- The overall spread of the jumping jacks data is much greater than the spread of the sit-ups data. This was quantified in the MAD for jumping jacks of 2.8 while for sit-ups the MAD was 1.6.
- The IQR for jumping jacks is 5 while for sit-ups it was 3, showing that the middle 50% of the sit-ups data was less spread out than the middle 50% of the jumping jacks.
- The median for the number of sit-ups was lower than the median for jumping jacks.
- The upper quartile for both types of exercise was 35.
- The lowest number of sit-ups was higher than at least 25% of the jumping jacks data.

Section 4. Interpret and Draw Conclusions about Data

Summarize numerical data sets in relation to their context, such as by

a. Reporting the number of observations.

b. Describing the nature of the attribute under investigation, including how it was measured and its units of measurement.

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.

d. Relating the choice of measures of center and variability to the shape of the data distribution and the context in which the data were gathered. *6.SP.5.*

In one way or another all of these objectives have been covered in the preceding text. Here is a problem that summarizes those discussions.

Example 19: Hunter and Maria play basketball in the same league. Dot plots for the number of points each girl scored per game are shown in Figure 17.

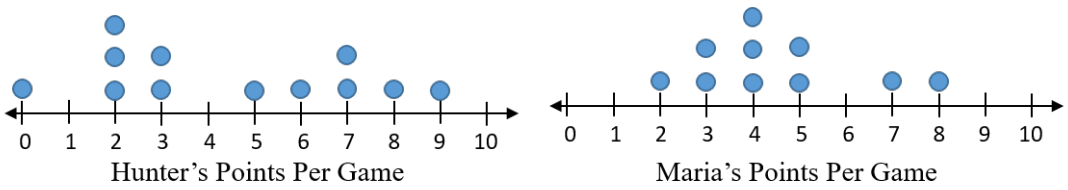


Figure 17

Using these data, answer the following:

a) Which player played in more games?

b) Which player has the higher average points per game?

c) Calculate the mean, median and mode for each player. What do these show?

d) Find the IQR for each player.

e) Find the MAD for each player.

f) Which player is more consistent (less variability) in points scored per game? Explain your answer.

SOLUTION.

a) By counting the number of dots, we see that Hunter played in 12 games and Maria played in 10. Therefore Hunter played in more games than Maria.

b) The average for Hunter is 54/12 = 4.5. For Maria, the average is 45/10 = 4.5. Both players had the same average points per game of 4.5.

c) Here are the data:

	Mean	Median	Mode
Hunter	4.5	4	2
Maria	4.5	4	4

There is not much to be learned from these computations, since the measures of central tendency are the same except in the case of the mode, and we can't yet determine what that means.

d) The IQR for Hunter is 5 and for Maria it is 2. This also lets us know that the middle 50% of Maria's data has less variability than the middle 50% of Hunter's, suggesting that Maria is likely to be the more consistent player.

e) For Hunter, the MAD is 2.5 and for Maria it is 1.4. This suggests that there is much more variance in Hunter's play than in that of Maria. Putting that suggestion together with another look at the dotplots, we see more clearly that we should consider Hunter's play as almost bimodal: one peak at 2 and the other at 7.

The conclusion to be made is that, although both players contribute, on average, the same number of points, Hunter has bad and good days, while Maria is much more consistent in her play. With this knowledge, the coach will play Hunter or Maria, depending upon the needs of a game. If a superior performance is needed to win a game, the coach will play Hunter, but if consistent play is key to a victory, Maria will be selected.