6th Grade ~ Conceptual Foundations 9 – Statistics and Probability

Develop understanding of statistical variability.

1. Recognize a statistical question as one that anticipates variability in the data related to the question and accounts for it in the answers. *For example, "How old am I?" is not a statistical question, but "How old are the students in my school?" is a statistical question because one anticipates variability in students' ages.*

2. 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.

3. 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 how its values vary with a single number.

Summarize and describe distributions.

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

5. 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 (inter-quartile 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.

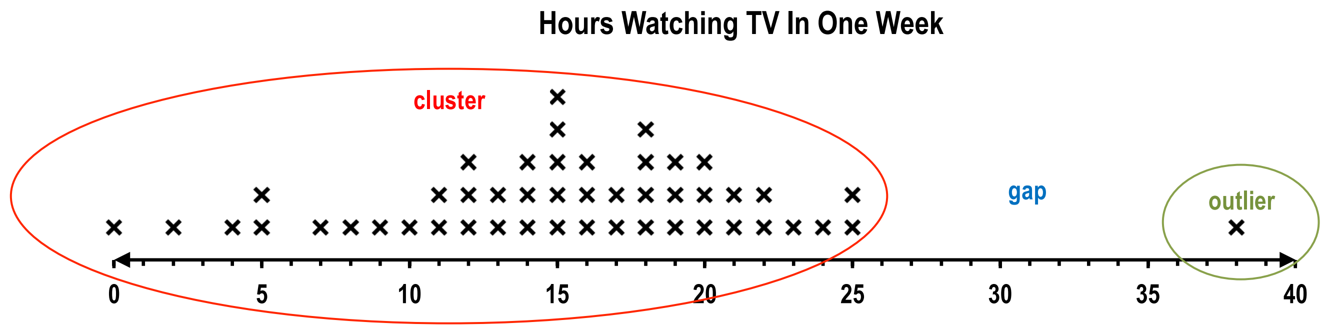
d. Relating the choice of measures of center and variability to the shape of the data distribution and the context in which the data were gathered.

Connections to Other Grades	Statistical Questions
In 1st grade students ask and answer questions about the data points in a graph. A study of line plots is provided in grades 2 through 5. Picture graphs and bar graphs are studied in grades 2 and 3. Therefore, 6th grade is a foundational year for histograms and box plots. Generating statistical questions and looking at the distribution of data to identify measures of center, spread, and overall shape is new in 6th grade. These skills lay the foundation for statistics and probability in 7th and 8th grades.	**What is a statistical question?** A question that generates a variety of answers is called a statistical question. Depending on the question, the type of data gathered can be either categorical or numerical. An example of a categorical question is "What is your favorite type of pizza?" The answers generated by this question will be categories of pizza types such as pepperoni, cheese, or sausage. An example of a numerical question is "How many pencils does each member of our class have in his or her desk?" A variety of numerical answers about the number of pencils would be given by a typical 6th grade class. In 6th grade, the focus should be on statistical questions that generate numerical data. Once the data is gathered, it can be organized in a table and/or displayed in a graph. The types of graphs to be focused on in 6th grade are dot plots (line plots), histograms, and box plots (box-and-whisker plots.) After organizing the data into tables and/or graphs, students will analyze the data finding measures of center and measures of variation to draw conclusions.

Examples of Statistical Questions	Non-Examples of Statistical Questions
• How old are the students in my school? • How many pets are owned by each student in my grade level? • What are the math test scores of the students in my class? • How many cupcakes of each type were made at the bakery in a week? • How many letters are in the names of each person in my class? • What is the height of each person in my class?	• How old am I? • How many pets do I own? • What is my math test score? • What is my favorite type of cupcake? • How many letters are in my name? • What is my height?

DATA DISTRIBUTION: The Shape of Data

The Shape of Data – A set of data can be distributed or placed on a graph in order to show characteristics of the data set. When placed on a graph, it is easier to see how the data is spread out or clustered together. To discuss the shape of the data set as a whole, students use the terms **cluster**, **gap**, and **outlier**. A **cluster** of data is a grouping of numbers that are close together in values. Looking at the line plot below, you can see the cluster of data is from 0 to 25. A **gap** is a place on the graph where no data values are present. On the graph below there is a **gap** between 26 and 39. Since the **gap** is very large between 26 and 37, the data value at 38 is called an **outlier**. An **outlier** is a number in a data set that is much larger or much smaller than the other numbers in the data set.



Why is it important to look at the shape of the data?

Observing the shape of the data on a graph gives a snapshot view of the overall characteristics. Generalizations can be made about the frequency and patterns of responses. For example, a teacher asks a class the question "How many hours of TV do you watch in a week?" By looking at a graph of the data, a general observation might be that a large number of students watch 10 to 20 hours of TV in a week. A second observation could be that one student watches almost double the amount of TV compared to other students.

DATA DISTRIBUTION: Center and Spread

There are two main ways that 6^{th} graders will summarize a data set. Students will examine and use measures of center and measures of variation.

- **Measures of center** (also called 'measures of central tendency') describe how data looks at the center. With a measure of center, we use a **single number** to summarize all of the values. The three most commonly used measures of center are mean, median, and mode. Students in 6^{th} grade are to focus on finding the mean and median.
- **Measures of variation** (also called 'measures of spread' or 'measures of dispersion') are ways to measure how much a collection of data is spread out. With a measure of variation, a single number describes how the values vary in a set. Students in 6^{th} grade should be able to find range and mean absolute deviation.

Measures of Center

Mean is the *sum of the values* in a set divided by the *number of values* in the set. In the dot plot below, each 'x' represents a data value for how many pets each student owns. Students need to add the values of each 'x' to find a total sum or number of pets owned by all class members. Then they divide that sum by the number of values, which is the same as the number of students. This gives the mean number of pets for each student.

(1+1+1+1+1+1+2+2+2+3+3+4+4+4+5+5+5+5+5) ÷19 = 2.9 pets for each student

When should the mean be utilized? Mean is useful when most of the data is tightly clustered as in the graph below. This means there are no extreme values or outliers.

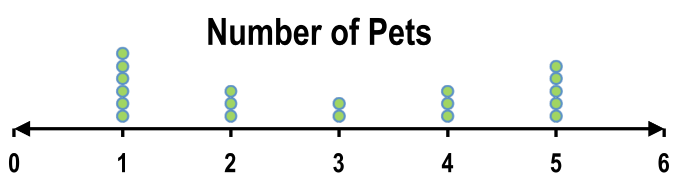
Median is the middle number of a set of values when the numbers are arranged in order from least to greatest. If there are two middle numbers, the median is the mean of those numbers. In the dot plot below, the median value is 3.

When should median be selected? Median is useful as a measure of center when there are extreme values or outliers and there are no big gaps in the middle of the data set. Median is also used in constructing box plots.

Mode is the number that appears most frequently in a set of numbers. There may be one mode, more than one mode, or no mode for a given data set. In the dot plot below, the mode is 1. The most frequent number of pets owned is 1.

When should the mode be selected? Mode can be a good choice when there are many identical data points because it describes what is typical about the set of data.

Dot Plot Example:



Mode = 1

Median = 3

Mean = 2.9

Measures of Variation (Spread)

Range is the difference between the maximum and the minimum in a set of data. In the box plot below, the highest math test score is 100% and the lowest math test score is 30%. The range would then be 100% minus 30% which is 70%.

100 – 30 = 70%

Why is range important? Range is valuable for knowing how far apart the minimum and maximum values are in a data set. It helps to know when the spread of data is close together or far apart.

Mean Absolute Deviation is an average of how far each data point in a set is from the mean of the set of data. A detailed description of how to find the mean absolute deviation for a set of data is included in this document.

Lower Quartile (Q1) is the median of the lower half of an ordered set of data. In the box plot below, the median of the lower half of the data is 50. This means that the middle test score in the lower half of the data was 50%.

Upper Quartile (Q3) is the median of the upper half of an ordered set of numbers. In the box plot below, the median of the upper half of the data is 93. The middle test score in the upper half of the data was 93%.

Why are the lower and upper quartiles important? Knowing the lower and upper quartiles helps to determine whether data points are outliers.

Interquartile Range is the difference between the upper quartile and the lower quartile. In the box plot below, the interquartile range is 93 – 50 = 43.

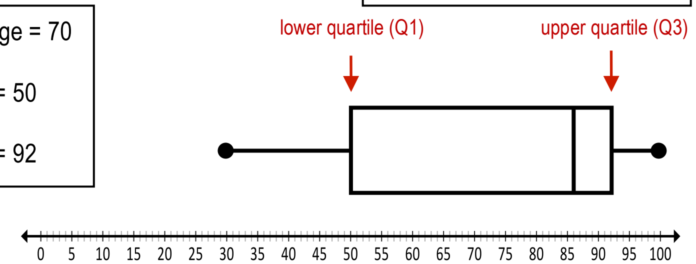
Box Plot (Box-and-Whisker Plot) Example:

Math Test Scores in Percents

Range = 70

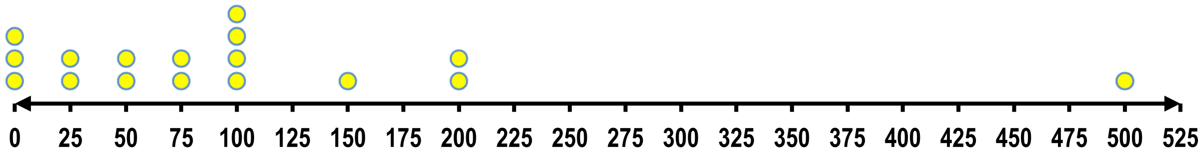
Q1 = 50

Q3 = 92



Matching a Statistical Question to a Graph

Sixth graders should be able to examine the shape of data on a graph and determine which statistical question best fits the shape of the data. For example, which question best fits the dot plot below? 1) How many glasses of milk does each member of our class drink a day? 2) How many letters are in the first name of each class member? 3) How many text messages does each class member send in a day? 4) How many library books are in each class member's desk?

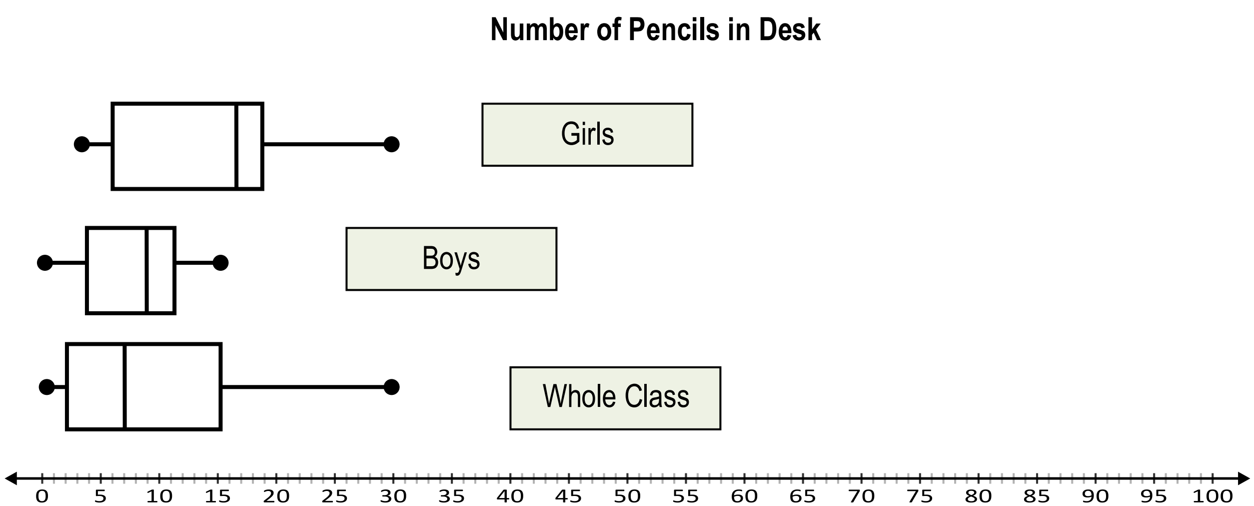


The best choice is "How many text messages does each class member send in a day?" The given numbers on the number line are not reasonable for glasses of milk, letters in a first name, or library books in a desk. This type of problem ties nicely into the Mathematical Practice Standard "Construct viable arguments and critique the reasoning of others." Students can justify their answers based on logical reasoning for each context in the possible answer choices.

Looking at Subgroups

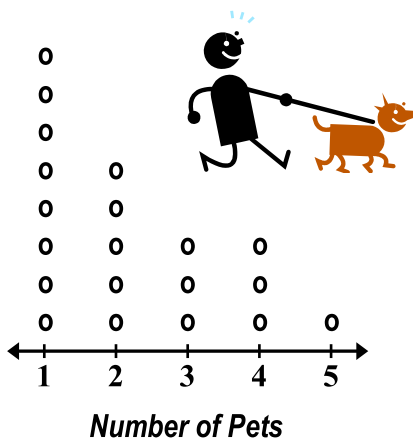
Depending on the question, an entire data set may be studied as a whole or as component parts. To look at it as component parts means to make comparisons between the responses of different subgroups that answered the question. For the question, "How many pencils does each member of our class have in his or her desk?" students might look at the class as a whole, or they might compare the numbers of pencils in the boys' desks compared to the girls' desks. Here we can see in the whole class box plot that the number of pencils in a desk range from 0 to 30, with the median being 7. Once this is broken down into subgroups, it can be seen that the lower extreme of 0 is found in the data about boys' desks, while the upper extreme of 30 pencils is found in the data about girls' desks. The median number of pencils found in girls' desks (17) is higher than the median number of pencils found in boys' desks (9). Additionally, students might generalize that boys have fewer pencils in their desks and closer to the same amount of pencils than found in girls' desks.

Example:



6th Grade ~ Conceptual Foundations 9 – Statistics and Probability

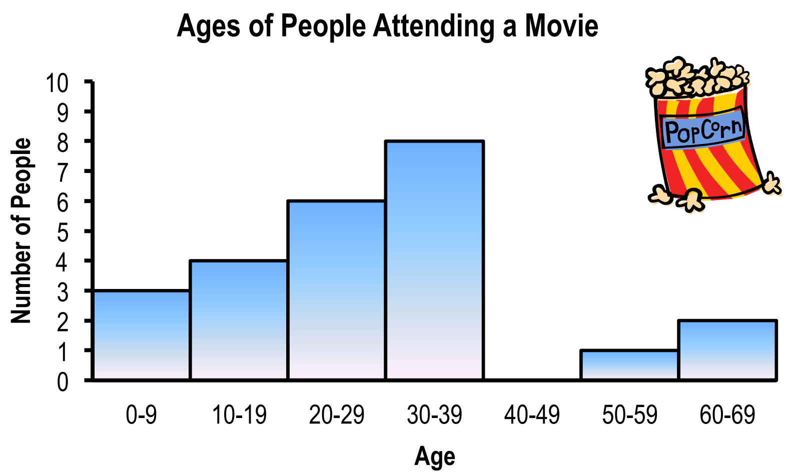
DATA DISPLAYS	
Importance of Data Displays	**6th Grade Focus**
Why are data displays important? We use data displays to organize information and make a visual representation for easy analysis. Imagine reading a newspaper article with numerous facts and figures that are too difficult to remember separately. Now imagine there is a graph attached to the article that takes all of those numbers and visually organizes the information. The information is then a valuable, informative tool.	There are three data displays 6th graders should be able to create and interpret. The first is a dot plot, which is a type of line plot. Although students have been making line plots since 2nd grade, the level of sophistication in the interpretation of the data is greatly increased by 6th grade. Students construct and interpret histograms, which require them to learn about intervals that reflect a range of numbers. Finally, 6th graders learn to make box plots (box-and-whisker plots). *Each type of graph is described below in detail with a different data set. Finally, one data set is modeled on all three graphs in order to compare and contrast the features of the displays.*

Dot Plots	Features
	**Definition:** A dot plot is a method of visually displaying a distribution of data values where each value is shown as a dot above a number line. (A dot plot is a type of line plot that uses dots instead of "x's" to show the frequency of values.) **Important Features:** A dot plot is formed from a number line. The mean, median, and mode are all measures of center, which can be calculated from a dot plot. Range and mean absolute deviation are also accessible in this data display. While there is no y-axis, it can be thought of as an imaginary frequency tally. **Advantages:** This is a simple graph that is easy to construct and interpret. It works best for a small set of data values, preferably 50 data values or less. Clusters, gaps, and outliers are easily identifiable when looking at the shape of the data. The exact data values are visually retained. **Common Misconceptions by Students:** Students may only make tick marks on the number line for values for which they have data. This will skew the shape of the data if equal intervals are not maintained for all consecutive numbers in the interval pattern. For example, if a student made equal tick marks on the number line and then labeled them as 1, 2, 4, 7, 10 because those were the only numerical responses given, the gaps in the responses would not be noticeable. Another common error is that students may count each mark as "1" rather than the value it represents. For example: An "x" above 4 has a value of 4, but a student may count it as a value of 1. Students must also take care to make each mark selected such as an 'x' equal in size.

Creating a Dot Plot

1) Draw a horizontal number line.
2) Determine and mark a scale of numbers below the line. Make sure to include the minimum and maximum values in the data set and all consecutive number values in between. Example: In the data set, there is a minimum value of 2 and a maximum value of 18. The number line must include tick marks for every number value from 2 through 18. A few numbers before the minimum and a few numbers after the maximum can be included.
3) A dot is tallied for each value above the corresponding number. Keep the imaginary y-axis as a frequency mark to ensure that dots are plotted correctly.
4) Put a title on the graph.

Histograms



Features

Definition: A histogram is a data display in which the labels for the bars are numerical intervals.

Important Features: A histogram has solid bars like a bar graph. However, there are no spaces between the bars unless no data is given for an interval. The intervals are listed below each bar on the x-axis. Exact values cannot be read because of the clustering of data. The y-axis is helpful for determining the value at the height of each bar.

Advantages: A histogram can be easy for students to read. Numerical data can be clustered into intervals and represented together on a graph. Additionally, large amount of data can be represented.

Common Misconceptions by Students: Students must pay attention to the intervals used on graphs in order to accurately understand the values being represented. Students may not pay attention to the intervals on the y-axis and think it is always labeled in multiples of one. They also may not understand that the intervals on the x-axis are a range of numbers that include the numbers seen and all numbers between. Many students may think a histogram is exactly the same as a bar graph (which displays categorical data rather than numerical data) and try to read or create it in the same manner.

Creating a Histogram

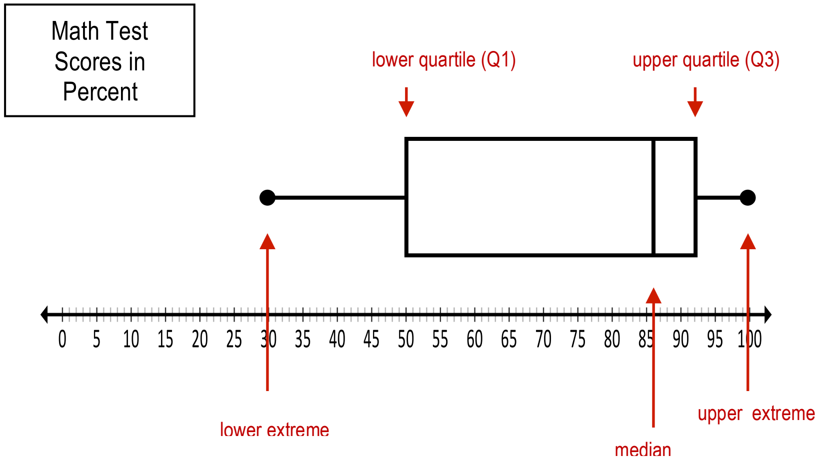
1) Make a frequency table of the data by selecting a range that will contain all of the data and then divide it into equal intervals. In the example above, the range of ages is from 0 to 69 so equal intervals of 10 years were selected.

<table>
<tr><th colspan="3">Age of People Attending a Movie</th></tr>
<tr><th>Age Ranges</th><th>Tally</th><th>Frequency</th></tr>
<tr><td>0 - 9</td><td>|||</td><td>3</td></tr>
<tr><td>10 - 19</td><td>||||</td><td>4</td></tr>
<tr><td>20 - 29</td><td>~~||||~~ |</td><td>6</td></tr>
<tr><td>30 - 39</td><td>~~||||~~ |||</td><td>8</td></tr>
<tr><td>40 - 49</td><td></td><td>0</td></tr>
<tr><td>50 - 59</td><td>|</td><td>1</td></tr>
<tr><td>60-69</td><td>||</td><td>2</td></tr>
</table>

2) Using graph paper, draw an x-axis where each box will represent an interval of numbers to represent the ranges.
3) Draw a y-axis with a scale of numbers appropriate for the data. Common scales are multiples of 1, 2, 5, 10 or 20.
4) Draw each bar on the histogram to correlate the intervals with the frequency of occurrence.
5) Title the graph and the x and y-axis.

6th Grade ~ Conceptual Foundations 9 – Statistics and Probability

Box Plots (Box-and-Whisker Plots)



Features

Definition: The CCSS Glossary defines a box plot as a method of visually displaying a distribution of data values by using the median, quartiles, and extremes of the data set. The box shows the middle 50% of the data, and the extended "whiskers" show the remaining 50% of the data.

Important Features: A box plot is formed from a number line. The graph can be thought of as a "5-Point Summary" of the data. It displays: 1) the median; 2) the lower quartile (Q1); 3) the upper quartile (Q3); 4) the lower extreme (minimum); and 5) the upper extreme (maximum).

Advantages: This graph can be used for very large data sets because it gives a general idea of how the data is clustered together. Exact values of each data point are not given in a box plot. A further advantage is that additional box plots can be drawn above the same number line to compare two or more data sets.

Common Misconceptions by Students: The most common misconception students have on box plots is they don't understand that each quartile represents 25% of the data. They also struggle with finding Q1 and Q3 and making the connection that Q1 is the median of the lower half of the data and Q3 is the median of the upper half of the data.

Creating a Box Plot (Box-and-Whisker Plot)

1) Write the data in order from least to greatest.
2) Draw a horizontal number line that can show the data in equal intervals.
3) Find the median of the data set and mark it on the number line.
4) Find the median of the upper half of the data. This is called the upper quartile (Q3). Mark it on the number line.
5) Find the median of the lower half of the data. This is called the lower quartile (Q1). Mark it on the number line.
6) Mark the lower extreme (minimum) on the number line.
7) Mark the upper extreme (maximum) on the number line.
8) Draw a box between the lower quartiles and the upper quartile. Draw a vertical line through the median to split the box.
9) Draw a "whisker" from the lower quartile to the lower extreme.
10) Draw a "whisker" from the upper quartile to the upper extreme.

6^{th} Grade ~ Conceptual Foundations 9 – Statistics and Probability

How to Calculate Mean Absolute Deviation

The **mean absolute deviation** is an average of how far each data point in a set is from the mean of the set of data. In other words, it is the "average distance from the average."

Example Problem:

> The weights of three students are 56 pounds, 78 pounds, and 91 pounds. What is the mean absolution deviation in weights?

Solution: Based upon the calculations in the four steps, the mean of the three weights is 75 pounds. By calculating the **mean absolute deviation**, it can be said that each person weighs an average of 12.67 pounds more or 12.67 pounds less than the mean of 75 pounds.

Step 1: Find the mean of the set of data. In the example problem, students add up the weights and then divide by 3 to find the mean.

$$(56+78+91) \div 3 = 75$$

Step 2: Determine the deviation or difference of each number in the data set from the mean. The mean of the example problem is 75. Subtract 75 from each number in the data set to find the difference.

$$56 – 75 = -19$$
$$78-75=3$$
$$91-75=16$$

Step 3: Find the absolute value of each deviation from Step 2.

$$| -19 | = 19$$
$$| 3 | = 3$$
$$| 16 | = 16$$

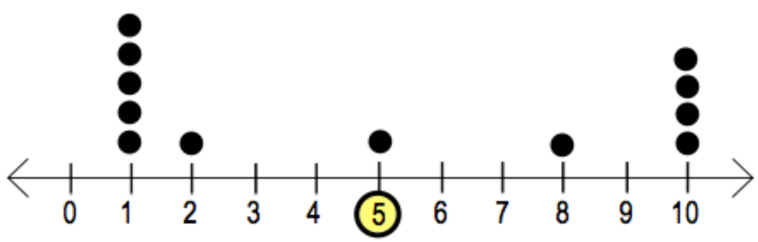
Step 4: Find the average of the absolute deviations from Step 3.

$$(19 + 3 + 16) \div 3 = 12.67$$

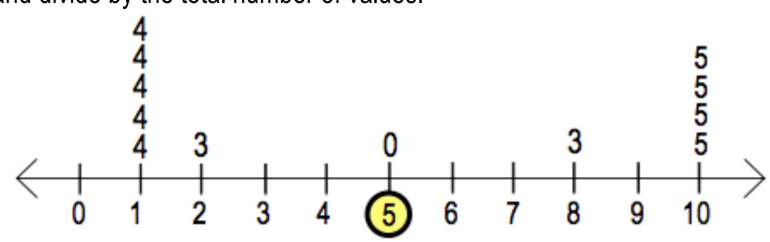
Answer: 12.67 is the mean absolute deviation.

Simplified Method for Determining the Mean Absolute Deviation

1) Place the values on a dot or line plot and determine the mean.



2) Mark the distances from the mean by each dot or in place of the dot. Add the distances from the mean and divide by the total number of values.



The mean absolute deviation is approximately 3.83, or each value is about 4 away from the mean.

Sample Questions Students Should Be Able to Answer About the Data in a Graph

What is the statistical question asked?
How many observations were made?
How many people were surveyed to gather the data?
What was measured or counted by the statistical question?
How was the attribute measured or counted?
What unit of measurement is being used to describe the data?
What is the shape of the data?

Where are the clusters or gaps in the data?
Are there any outliers in the data?
Is mean or median a better descriptor for the measure of center for a particular set of data?
What is the mean absolute deviation of the data points in the set?
What is the range of the data?
What is the lower quartile (Q1) or median of the lower half of the data?
What is the upper quartile (Q3) or median of the upper half of the data?