UNIT 6: DESCRIBING DATA

In this unit, students will learn informative ways to display both categorical and quantitative data. They will learn ways of interpreting those displays and pitfalls to avoid when presented with data. Students will learn how to determine the mean absolute deviation. Among the methods they will study are two-way frequency charts for categorical data and lines of best fit for quantitative data. Measures of central tendency will be revisited along with measures of spread.

Summarize, Represent, and Interpret Data on a Single Count or Measurable Variable

MCC9-12.S.ID.1 Represent data with plots on the real number line (dot plots, histograms, and box plots).

MCC9-12.S.ID.2 Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, mean absolute deviation, standard deviation) of two or more different data sets.

MCC9-12.S.ID.3 Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

KEY IDEAS

1. Two measures of central tendency that help describe a data set are mean and median.
   - The mean is the sum of the data values divided by the total number of data values.
   - The median is the middle value when the data values are written in order from least to greatest. If a data set has an even number of data values, the median is the mean of the two middle values.

2. The first quartile, or the lower quartile, \( Q_{1} \), is the median of the lower half of a data set.

   Example:
   Ray’s scores on his mathematics tests were 70, 85, 78, 90, 84, 82, and 83. To find the first quartile of his scores, write them in order from least to greatest:
   
   \[ 70, 78, 82, 83, 84, 85, 90 \]

   The scores in the lower half of the data set are 70, 78, and 82. The median of the lower half of the scores is 78.

   So, the first quartile is 78.

3. The third quartile, or the upper quartile, \( Q_{3} \), is the median of the upper half of a data set.
Example:
Referring to the previous example, the upper half of Ray’s scores is 84, 85, and 90. The median of the upper half of the scores is 85.
So, the third quartile is 85.

4. The **interquartile range (IQR)** of a data set is the difference between the third and first quartiles, or \( Q_3 - Q_1 \).

Example:
Referring again to the example of Ray’s scores, to find the interquartile range, subtract the first quartile from the third quartile. The interquartile range of Ray’s scores is 85 – 78 = 7.

5. The most common displays for quantitative data are dot plots, histograms, box plots, and frequency distributions. A **box plot** is a diagram used to display a data set that uses quartiles to form the center box and the minimum and maximum to form the whiskers.

Example:
For the data in Key Idea 2, the box plot would look like the one shown below:
A histogram is a graphical display that subdivides the data into class intervals, called bins, and uses a rectangle to show the frequency of observations in those intervals—for example, you might use intervals of 0–3, 4–7, 8–11, and 12–15 for the number of books students read over summer break.

6. Sometimes, distributions are characterized by extreme values that differ greatly from the other observations. These extreme values are called outliers. A data value is an outlier if it is less than $Q_1 - 1.5 \cdot IQR$ or above $Q_3 + 1.5 \cdot IQR$.

**Example:**

This example shows the effect that an outlier can have on a measure of central tendency.

The mean is one of several measures of central tendency that can be used to describe a data set. The main limitation of the mean is that, because every data value directly affects the result, it can be affected greatly by outliers. Consider these two sets of quiz scores:

- **Student P:** {8, 9, 9, 9, 10}
- **Student Q:** {3, 9, 9, 9, 10}

Both students consistently performed well on quizzes, and both have the same median and mode score, 9. Student Q, however, has a mean quiz score of 8, while student P has a mean quiz score of 9. Although many instructors accept the use of a mean as being fair and representative of a student’s overall performance in the context of test or quiz scores, it can be misleading because it fails to describe the variation in a student’s scores, and the effect of a single score on the mean can be disproportionately large, especially when the number of scores is small.
7. **Mean absolute deviation** is the distance each data value is from the mean of the data set. This helps to get a sense of how spread out a data set is.

**Example:**

This example shows two sets of data that have the same mean but different mean absolute deviations. Consider the quiz scores of two students:

- **Student R:** {3, 6, 8, 8, 9, 10, 12}
- **Student S:** {1, 1, 3, 7, 14, 15, 15}

The mean score of student R is 8, and the mean score of student S is also 8. Determining the mean does not provide us with which student was more consistent. Which student is more consistent is what the mean absolute deviation will provide.

We can use this formula to find the mean absolute deviation. To apply the formula, we need to find the sum of the difference of the terms and the mean. So,

\[
\frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}
\]

Student R: \[|3 - 8| + |6 - 8| + |8 - 8| + |8 - 8| + |9 - 8| + |10 - 8| + |12 - 8| = 14\]

Student S: \[|1 - 8| + |1 - 8| + |3 - 8| + |7 - 8| + |14 - 8| + |15 - 8| + |15 - 8| = 40\]

The final step is to divide the sums by the number of data, \(n\).

Student R: \[\frac{14}{8} = 1.75\]

Student S: \[\frac{40}{8} = 5\]

Since the mean absolute deviation of student R is smaller than the mean absolute deviation of student S, this means the quiz scores of student R were more consistent.
8. **Skewness** refers to the type and degree of a distribution’s asymmetry. A distribution is skewed to the left if it has a longer tail on the left side and has a negative value for its skewness. If a distribution has a longer tail on the right, it has positive skewness. Generally, distributions have only one peak, but there are distributions called **bimodal** or **multimodal** where there are two or more peaks, respectively. A distribution can have symmetry but not be a normal distribution. It could be too flat (uniform) or too spindly. A box plot can present a fair representation of a data set’s distribution. For a normal distribution, the median should be very close to the middle of the box and the two whiskers should be about the same length.

![Skewed to the left and right](image)

**Bimodal representation**

**Important Tip**

The extent to which a data set is distributed normally can be determined by observing its skewness. Most of the data should lie in the middle near the median. The mean and the median should be fairly close. The left and right tails of the distribution curve should taper off. There should be only one peak, and it should neither be too high nor too flat.
Unit 6: Describing Data

REVIEW EXAMPLES

1. Josh and Richard each earn tips at their part-time jobs. This table shows their earnings from tips for five days.

<table>
<thead>
<tr>
<th>Day</th>
<th>Josh’s Tips</th>
<th>Richard’s Tips</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>$40</td>
<td>$40</td>
</tr>
<tr>
<td>Tuesday</td>
<td>$20</td>
<td>$45</td>
</tr>
<tr>
<td>Wednesday</td>
<td>$36</td>
<td>$53</td>
</tr>
<tr>
<td>Thursday</td>
<td>$28</td>
<td>$41</td>
</tr>
<tr>
<td>Friday</td>
<td>$31</td>
<td>$28</td>
</tr>
</tbody>
</table>

a. Who had the greater median earnings from tips? What is the difference in the median of Josh’s earnings from tips and the median of Richard’s earnings from tips?

b. What is the difference in the interquartile range for Josh’s earnings from tips and the interquartile range for Richard’s earnings from tips?

Solution:

a. Write Josh’s earnings from tips in order from the least to greatest. Then, identify the middle value.

$20, $28, $\textbf{31}$, $36, $40$

Josh’s median earnings from tips are $31.

Write Richard’s earnings from tips in order from the least to greatest. Then, identify the middle value.

$28, $40, $\textbf{41}$, $45, $53$

Richard had the greater median earnings from tips. The difference in the median of the earnings from tips is $41 – $31 = $10.$

b. For Josh’s earnings from tips, the lower quartile is $24$ and the upper quartile is $38$. The interquartile range is $38 –$24, or $14$.

For Richard’s earnings from tips, the lower quartile is $34$ and the upper quartile is $49$. The interquartile range is $49 –$34, or $15$.

The difference in Josh’s interquartile range and Richard’s interquartile range is $15 – $14, or $1$.

2. Sophia is a student at Windsfall High School. These histograms give information about the number of hours spent volunteering by each of the students in Sophia’s homeroom and by each of the students in the tenth-grade class at her school.
a. Compare the lower quartiles of the data in the histograms.
b. Compare the upper quartiles of the data in the histograms.
c. Compare the medians of the data in the histograms.

**Solution:**

a. You can add the number of students given by the height of each bar to find that there are 23 students in Sophia’s homeroom. The lower quartile is the median of the first half of the data. That would be found within the 10–19 hours interval.

You can add the number of students given by the height of each bar to find that there are 185 students in the tenth-grade class. The lower quartile for this group is found within the 10–19 hours interval.

The interval of the lower quartile of the number of hours spent volunteering by each student in Sophia’s homeroom is the same as the interval of the lower quartile of the number of hours spent volunteering by each student in the tenth-grade class.

b. The upper quartile is the median of the second half of the data. For Sophia’s homeroom, that would be found in the 30 or greater interval.

For the tenth-grade class, the upper quartile is found within the 20–29 hours interval.

The upper quartile of the number of hours spent volunteering by each student in Sophia’s homeroom is greater than the upper quartile of the number of hours spent volunteering by each student in the tenth-grade class.

c. The median is the middle data value in a data set when the data values are written in order from least to greatest. The median for Sophia’s homeroom is found within the 10–19 hours interval.

The median for the tenth-grade class is found within the 20–29 hours interval.

The median of the number of hours spent volunteering by each student in Sophia’s homeroom is less than the median of the number of hours spent volunteering by each student in the tenth-grade class.
3. Mr. Storer, the physical education teacher, measured and rounded, to the nearest whole inch, the height of each student in his first-period class. He organized his data in this chart.

<table>
<thead>
<tr>
<th>Height (inches)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>42</td>
<td>1</td>
</tr>
<tr>
<td>43</td>
<td>2</td>
</tr>
<tr>
<td>44</td>
<td>4</td>
</tr>
<tr>
<td>45</td>
<td>5</td>
</tr>
<tr>
<td>46</td>
<td>4</td>
</tr>
<tr>
<td>47</td>
<td>2</td>
</tr>
<tr>
<td>48</td>
<td>1</td>
</tr>
</tbody>
</table>

a. Make a dot plot for the data.
b. Make a histogram for the data.
c. Make a box plot for the data.

Solution:

a. 

```
  x
 x  x  x
 x  x  x
 x  x  x  x
 x  x  x  x  x

Student Heights in Mr. Storer’s Class
```
Unit 6: Describing Data

b.

![Height Distribution Chart]

<table>
<thead>
<tr>
<th>Frequency</th>
<th>41 or 42</th>
<th>43 or 44</th>
<th>45 or 46</th>
<th>47 or 48</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

Height (inches)

41 or 42 43 or 44 45 or 46 47 or 48


c.

![Box Plot for Student Heights]

Student Heights in Mr. Storer's Class

40 45 50

41 or 42 43 or 44 45 or 46 47 or 48
4. A geyser in a national park erupts fairly regularly. In more recent times, it has become less predictable. It was observed last year that the time interval between eruptions was related to the duration of the most recent eruption. The distribution of its interval times for last year is shown in the following graphs.
Unit 6: Describing Data

a. Does the Last Year distribution seem skewed or uniform?
b. Compare Last Week’s distribution to Last Month’s distribution.
c. What does the Last Year distribution tell you about the interval of time between the geyser’s eruptions?

Solution:

a. The Last Year distribution appears to be skewed to the left (negative). Most of the intervals approach 90 minutes.
b. Last Week’s distribution seems more skewed to the left than Last Month’s. It is also more asymmetric because of its high number of 1-hour-and-35-minute intervals between eruptions. Last Month’s distribution appears to have the highest percentage of intervals longer than 1 hour 30 minutes between eruptions.
c. The Last Year distribution shows that the geyser rarely erupts an hour after its previous eruption. Most visitors will have to wait more than 90 minutes to see two eruptions.
SAMPLE ITEMS

1. This table shows the average low temperature, in °F, recorded in Macon, GA, and Charlotte, NC, over a six-day period.

<table>
<thead>
<tr>
<th>Day</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature in Macon, GA (°F)</td>
<td>71</td>
<td>72</td>
<td>66</td>
<td>69</td>
<td>71</td>
<td>73</td>
</tr>
<tr>
<td>Temperature in Charlotte, NC (°F)</td>
<td>69</td>
<td>64</td>
<td>68</td>
<td>74</td>
<td>71</td>
<td>75</td>
</tr>
</tbody>
</table>

Which conclusion can be drawn from the data?

A. The interquartile range of the temperatures is the same for both cities.
B. The lower quartile for the temperatures in Macon is less than the lower quartile for the temperatures in Charlotte.
C. The mean and median temperatures in Macon were higher than the mean and median temperatures in Charlotte.
D. The upper quartile for the temperatures in Charlotte was less than the upper quartile for the temperatures in Macon.

Correct Answer: C

2. A school was having a coat drive for a local shelter. A teacher determined the median number of coats collected per class and the interquartile range of the number of coats collected per class for the freshmen and for the sophomores.

- The freshmen collected a median number of coats per class of 10, and the interquartile range was 6.
- The sophomores collected a median number of coats per class of 10, and the interquartile range was 4.

Which range of numbers includes the third quartile of coats collected for both freshmen and sophomore classes?

A. 4 to 14
B. 6 to 14
C. 10 to 16
D. 12 to 15

Correct Answer: C
3. A reading teacher recorded the number of pages read in an hour by each of her students. The numbers are shown below.

44, 49, 39, 43, 50, 44, 45, 49, 51

For this data, which summary statistic is NOT correct?

A. The minimum is 39.
B. The lower quartile is 44.
C. The median is 45.
D. The maximum is 51.

Correct Answer: B

4. A science teacher recorded the pulse of each of the students in her classes after the students had climbed a set of stairs. She displayed the results, by class, using the box plots shown.

Which class generally had the highest pulse after climbing the stairs?

A. Class 1
B. Class 2
C. Class 3
D. Class 4

Correct Answer: C
5. Peter went bowling, Monday to Friday, two weeks in a row. He only bowled one game each time he went. He kept track of his scores below.

Week 1: 70, 70, 70, 73, 75  
Week 2: 72, 64, 73, 73, 75

What is the BEST explanation for why Peter’s Week 2 mean score was lower than his Week 1 mean score?

A. Peter received the same score three times in Week 1.  
B. Peter had one very low score in Week 2.  
C. Peter did not beat his high score from Week 1 in Week 2.  
D. Peter had one very high score in Week 1.

Correct Answer: B

6. This histogram shows the frequency distribution of duration times for 107 consecutive eruptions of the Old Faithful geyser. The duration of an eruption is the length of time, in minutes, from the beginning of the spewing of water until it stops. What is the BEST description for the distribution?

A. bimodal  
B. uniform  
C. multiple outlier  
D. skewed to the right

Correct Answer: A
Summarize, Represent, and Interpret Data on Two Categorical and Quantitative Variables

MCC9-12.S.ID.5 Summarize categorical data for two categories in two-way frequency tables. Interpret relative frequencies in the context of the data (including joint, marginal, and conditional relative frequencies). Recognize possible associations and trends in the data.

MCC9-12.S.ID.6 Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

  MCC9-12.S.ID.6a Decide which type of function is most appropriate by observing graphed data, charted data, or by analysis of context to generate a viable (rough) function of best fit. Use this function to solve problems in context. Emphasize linear, quadratic, and exponential models.

  MCC9-12.S.ID.6c Using given or collected bivariate data, fit a linear function for a scatter plot that suggests a linear association.

KEY IDEAS

1. There are essentially two types of data: **categorical** and **quantitative**. Examples of categorical data are color, type of pet, gender, ethnic group, religious affiliation, etc. Examples of quantitative data are age, years of schooling, height, weight, test score, etc. Researchers use both types of data but in different ways. Bar graphs and pie charts are frequently associated with categorical data. Box plots, dot plots, and histograms are used with quantitative data. The measures of central tendency (mean, median, and mode) apply to quantitative data. Frequencies can apply to both categorical and quantitative data.

2. **Bivariate data** consist of pairs of linked numerical observations, or frequencies of things in categories. Numerical bivariate data can be presented as ordered pairs and in any way that ordered pairs can be presented: as a set of ordered pairs, as a table of values, or as a graph on the coordinate plane.

   Categorical example: frequencies of gender and club memberships for 9th graders.

A bivariate chart, or **two-way frequency chart**, is often used with data from two categories. Each category is considered a variable, and the categories serve as labels in the chart. Two-way frequency charts are made of cells. The number in each cell is the frequency of things that fit both the row and column categories for the cell. From the two-way frequency chart on the next page, we see that there are 12 males in the band and 3 females in the chess club.
# Unit 6: Describing Data

## Participation in School Activities

<table>
<thead>
<tr>
<th>School Club</th>
<th>Gender</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td>Totals</td>
<td></td>
</tr>
<tr>
<td>Band</td>
<td>12</td>
<td>21</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>Chorus</td>
<td>15</td>
<td>17</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Chess</td>
<td>16</td>
<td>3</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Latin</td>
<td>7</td>
<td>9</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Yearbook</td>
<td>28</td>
<td>7</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>78</td>
<td>57</td>
<td>135</td>
<td></td>
</tr>
</tbody>
</table>

If no person or thing can be in more than one category per scale, the entries in each cell are called **joint frequencies**. The frequencies in the cells and the totals tell us about the percentages of students engaged in different activities based on gender. For example, we can determine from the chart that if we picked at random from the students, we are least likely to find a female in the chess club because only 3 of 135 students are females in the chess club. The most popular club is yearbook, with 35 of 135 students in that club. The values in the table can be converted to percentages, which will give us an idea of the composition of each club by gender. We see that close to 14% of the students are in the chess club, and there are more than five times as many males as females.

- **Band**: 8.9% (Male), 15.5% (Female), 24.4% (Totals)
- **Chorus**: 11.1% (Male), 12.6% (Female), 23.7% (Totals)
- **Chess**: 11.9% (Male), 2.2% (Female), 14.1% (Totals)
- **Latin**: 5.2% (Male), 6.7% (Female), 11.9% (Totals)
- **Yearbook**: 20.7% (Male), 5.2% (Female), 25.9% (Totals)

There are also what we call **marginal frequencies** in the bottom and right margins (the shaded cells in the table above). These frequencies lack one of the categories. For our example, the frequencies at the bottom represent percentages of males and females in the school population. The marginal frequencies on the right represent percentages of club membership.
Lastly, associated with two-way frequency charts are **conditional frequencies**. These are not usually in the body of the chart but can be readily calculated from the cell contents. One conditional frequency would be the percentage of chorus members that are female. The working condition is that the person is female. If 12.6% of the entire school population is females in the chorus and 42.2% of the student body is female, then 12.6% / 42.2%, or 29.9%, of the females in the school are in the chorus (also 17 of 57 females).

Quantitative example: Consider this chart of heights and weights of players on a football team.

![Football Players' Heights and Weights](chart.png)

A scatter plot is often used to present bivariate quantitative data. Each variable is represented on an axis, and the axes are labeled accordingly. Each point represents a player’s height and weight. For example, one of the points represents a height of 66 inches and weight of 150 pounds. The scatter plot shows two players standing 70 inches tall because there are two dots on that height.

3. **A scatter plot** displays data as points on a grid using the associated numbers as coordinates. The way the points are arranged by themselves in a scatter plot may or may not suggest a relationship between the two variables. In the scatter plot about the football players shown previously, it appears there may be a relationship between height and weight because, as the players get taller, they seem to generally increase in weight; that is, the points are positioned higher as you move to the right. Bivariate data may have an underlying relationship that can be modeled by a mathematical function. Many of the examples in this review focus on linear models, but the models may take other forms, especially quadratic and exponential functions.
**Example:**
Melissa would like to determine whether there is a relationship between study time and mean test scores. She recorded the mean study time per test and the mean test score for students in three different classes.

This is the data for Class 1.

<table>
<thead>
<tr>
<th>Mean Study Time (hours)</th>
<th>Mean Test Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>63</td>
</tr>
<tr>
<td>1</td>
<td>67</td>
</tr>
<tr>
<td>1.5</td>
<td>72</td>
</tr>
<tr>
<td>2</td>
<td>76</td>
</tr>
<tr>
<td>2.5</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>85</td>
</tr>
<tr>
<td>3.5</td>
<td>89</td>
</tr>
</tbody>
</table>

Notice that, for these data, as the mean study time increases, the mean test score increases. It is important to consider the *rate of increase* when deciding which algebraic model to use. In this case, the mean test score increases by approximately 4 points for each 0.5-hour increase in mean study time. When the rate of increase is close to constant, as it is here, the best model is most likely a linear function.

This next table shows Melissa’s data for Class 2.

<table>
<thead>
<tr>
<th>Mean Study Time (hours)</th>
<th>Mean Test Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>60</td>
</tr>
<tr>
<td>1</td>
<td>61</td>
</tr>
<tr>
<td>1.5</td>
<td>63</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
</tr>
<tr>
<td>2.5</td>
<td>74</td>
</tr>
<tr>
<td>3</td>
<td>82</td>
</tr>
<tr>
<td>3.5</td>
<td>93</td>
</tr>
</tbody>
</table>

In these data as well, the mean test score increases as the mean study time increases. However, the rate of increase is not constant. The differences between each successive mean test score are 1, 2, 5, 6, 8, and 11.
This table shows Melissa’s data for Class 3.

<table>
<thead>
<tr>
<th>Mean Study Time (hours)</th>
<th>Mean Test Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>71</td>
</tr>
<tr>
<td>1</td>
<td>94</td>
</tr>
<tr>
<td>1.5</td>
<td>87</td>
</tr>
<tr>
<td>2</td>
<td>98</td>
</tr>
<tr>
<td>2.5</td>
<td>69</td>
</tr>
<tr>
<td>3</td>
<td>78</td>
</tr>
<tr>
<td>3.5</td>
<td>91</td>
</tr>
</tbody>
</table>

In these data, as the mean study time increases, there is no consistent pattern in the mean test score. As a result, there does not appear to be any clear relationship between the mean study time and mean test score for this particular class.

Often, patterns in bivariate data are more easily seen when the data is plotted on a coordinate grid.

**Example:**

This graph shows Melissa’s data for Class 1.

In this graph, the data points are all very close to being on the same line. This is further confirmation that a linear model is appropriate for this class.
This graph shows Melissa’s data for Class 2.

In this graph, the data points appear to lie on a curve, rather than on a line, with a rate of increase that increases as the value of $x$ increases. It appears that a quadratic or exponential model may be more appropriate than a linear model for these data.

This graph shows Melissa’s data for Class 3.

In this graph, the data points do not appear to lie on a line or on a curve. Linear, quadratic, and exponential models would not be appropriate to represent the data.
4. A **line of best fit** (trend or regression line) is a straight line that best represents the data on a scatter plot. This line may pass through some of the points, none of the points, or all of the points. In the previous examples, only the Class 1 scatter plot looks like a linear model would be a good fit for the points. In the other classes, a curved graph would seem to pass through more of the points. For Class 2, perhaps a quadratic model or an exponential model would produce a better-fitting curve. Since class 3 appears to have no correlation, creating a model may not produce the desired results.

When a linear model is indicated there are several ways to find a function that approximates the \( y \)-value for any given \( x \)-value. A method called **regression** is the best way to find a line of best fit, but it requires extensive computations and is generally done on a computer or graphing calculator.

**Example:**

This graph shows Melissa’s data for Class 1 with the line of best fit added. The equation of the line can be determined by using a calculator by entering the data into the STAT EDIT L1 and L2 in your calculator. L1 is for the \( x \)-values and L2 is for the \( y \)-values. Once all coordinates are entered, use the linear regression feature on the calculator. You will get values for \( a \) and \( b \) for the equation \( y = ax + b \). For the data below, the equation is \( y = 8.8x + 58.4 \).

![Class 1 Graph](image)

Notice that five of the seven data points are on the line. This represents a very strong positive relationship for study time and test scores, since the line of best fit is positive and a very tight fit to the data points.
This next graph shows Melissa’s data for Class 3 with the line of best fit added. The equation of the line is $y = 0.8x + 83.1$.

Although a line of best fit can be calculated for this set of data, notice that most of the data points are not very close to the line. In this case, although there is some correlation between study time and test scores, the amount of correlation is very small. This is called the correlation coefficient, which is discussed in more detail in the next section about linear models.
REVIEW EXAMPLES

1. Barbara is considering visiting Yellowstone National Park. She has heard about Old Faithful, the geyser, and she wants to make sure she sees it erupt. At one time, it erupted just about every hour. That is not the case today. The time between eruptions varies. Barbara went on the Web and found a scatter plot of how long an eruption lasted compared to the wait time between eruptions. She learned that, in general, the longer the wait time, the longer the eruption lasts. The eruptions take place anywhere from 45 minutes to 125 minutes apart. They currently average 90 minutes apart.

<scatter plot image>

a. For an eruption that lasts 4 minutes, about how long would the wait time be for the next eruption?

b. What is the shortest duration time for an eruption?

c. Determine whether the scatter plot has a positive or a negative correlation and explain how you know.

Solution:

a. After a 4-minute eruption, it would be between 75 and 80 minutes for the next eruption.

b. The shortest eruptions appear to be a little more than 1.5 minutes (1 minute and 35 seconds).

c. The scatter plot has a positive correlation because as the eruption duration increases, the time between eruptions increases.
2. The environment club is interested in the relationship between the number of canned beverages sold in the cafeteria and the number of cans that are recycled. The data they collected are listed in this chart.

<table>
<thead>
<tr>
<th>Beverage Can Recycling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of Canned Beverages Sold</strong></td>
</tr>
<tr>
<td><strong>Number of Cans Recycled</strong></td>
</tr>
</tbody>
</table>

Find an equation of a line of best fit for the data.

**Solution:**

Use a calculator by entering the data into the STAT EDIT L1 and L2 in your calculator. Enter the values of $x$ in the L1 column and the values of $y$ in the L2 column. Once all coordinates are entered, use the linear regression feature on the calculator. You will get values for $a$ and $b$ for the equation $y = ax + b$. For the data in the table, the equation is $y = 0.38x + 0.99$.

3. A fast-food restaurant wants to determine whether the season of the year affects the choice of soft-drink size purchased. It surveyed 278 customers, and the table below shows its results. The drink sizes were small, medium, large, and jumbo. The seasons of the year were spring, summer, and fall. In the body of the table, the cells list the number of customers who fit both row and column titles. On the bottom and in the right margin are the totals.

<table>
<thead>
<tr>
<th></th>
<th>Spring</th>
<th>Summer</th>
<th>Fall</th>
<th>TOTALS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>24</td>
<td>22</td>
<td>18</td>
<td>64</td>
</tr>
<tr>
<td>Medium</td>
<td>23</td>
<td>28</td>
<td>19</td>
<td>70</td>
</tr>
<tr>
<td>Large</td>
<td>18</td>
<td>27</td>
<td>29</td>
<td>74</td>
</tr>
<tr>
<td>Jumbo</td>
<td>16</td>
<td>21</td>
<td>33</td>
<td>70</td>
</tr>
<tr>
<td><strong>TOTALS</strong></td>
<td><strong>81</strong></td>
<td><strong>98</strong></td>
<td><strong>99</strong></td>
<td><strong>278</strong></td>
</tr>
</tbody>
</table>

a. In which season did the most customers prefer jumbo drinks?

b. What percentage of those surveyed purchased small drinks?

c. What percentage of those surveyed purchased medium drinks in the summer?

d. What do you think the fast-food restaurant learned from its survey?

**Solution:**

a. The most customers preferred jumbo drinks in the fall.

b. Twenty-three percent ($64/278 = 23\%$) of the 278 surveyed purchased small drinks.

c. Ten percent ($28/278 = 10\%$) of those customers surveyed purchased medium drinks in the summer.

d. The fast-food restaurant probably learned that customers tend to purchase the larger drinks in the fall and the smaller drinks in the spring and summer.
SAMPLE ITEMS

1. Which graph MOST clearly displays a set of data for which a quadratic function is the model of best fit?

A. 

B. 

C. 

D. 

Correct Answer: A
2. This graph plots the number of wins in the 2006 season and in the 2007 season for a sample of professional football teams.

Which equation BEST represents a line that matches the trend of the data?

A. $y = x + 2$

B. $y = x + 7$

C. $y = \frac{3}{5}x + 1$

D. $y = \frac{3}{5}x + 5$

Correct Answer: D
Interpret Linear Models

**MCC9-12.S.ID.7** Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

**MCC9-12.S.ID.8** Compute (using technology) and interpret the correlation coefficient “r” of a linear fit. (For instance, by looking at a scatter plot, students should be able to tell if the correlation coefficient is positive or negative and give a reasonable estimate of the “r” value.) After calculating the line of best fit using technology, students should be able to describe how strong the goodness of fit of the regression is, using “r.”

**MCC9-12.S.ID.9** Distinguish between correlation and causation.

**KEY IDEA**

Once a model for the scatter plot is determined, we can begin to analyze the correlation of the linear fit. We can also interpret the slope, or rate of change, and the constant term and distinguish between correlation and causation of the data.

1. A **correlation coefficient** is a measure of the strength of the linear relationship between two variables. It also indicates whether the dependent variable, $y$, grows along with $x$, or $y$ gets smaller as $x$ increases. The correlation coefficient is a number between –1 and +1 including –1 and +1. The letter $r$ is usually used for the correlation coefficient. When the correlation is positive, the line of best fit will have a positive slope and both variables are growing. However, if the correlation coefficient is negative, the line of best fit has a negative slope and the dependent variable is decreasing. The numerical value is an indicator of how closely the data points are modeled by a linear function.

When using a calculator, use the same steps as you did to find the line of best fit. Notice there is a value, $r$, below the values for $a$ and $b$. This is the correlation coefficient.

**Examples:**

![Positive Perfect](image1)

$r = +1$

![Positive Weak](image2)

$r = +0.4$

![Negative Strong](image3)

$r = -0.7$

The correlation between two variables is related to the slope and the goodness of the fit of a regression line. However, data in scatter plots can have the same regression lines and very different correlations. The correlation’s sign will be the same as the slope of the regression line. The correlation’s value depends on the dispersion of the data points and their proximity to the line of best fit.
Example:
Earlier we saw that the interval between eruptions of Old Faithful is related to the duration of the most recent eruption. Years ago, the National Park Service had a simple linear equation they used to help visitors determine when the next eruption would take place. Visitors were told to multiply the duration of the last eruption by 10 and add 30 minutes (I = 10 \cdot D + 30). We can look at a 2011 set of data for Old Faithful, with eight data points, and see how well that line fits the 2011 data. The data points are from a histogram with intervals of 0.5 minute for \( x \)-values. The \( y \)-values are the average interval time for an eruption in that duration interval.

<table>
<thead>
<tr>
<th>Duration (x)</th>
<th>Interval (y)</th>
<th>Prediction</th>
<th>Error Distances</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.50</td>
<td>51.00</td>
<td>45</td>
<td>6</td>
</tr>
<tr>
<td>2.00</td>
<td>58.00</td>
<td>50</td>
<td>8</td>
</tr>
<tr>
<td>2.50</td>
<td>65.00</td>
<td>55</td>
<td>10</td>
</tr>
<tr>
<td>3.00</td>
<td>71.00</td>
<td>60</td>
<td>11</td>
</tr>
<tr>
<td>3.50</td>
<td>76.00</td>
<td>65</td>
<td>11</td>
</tr>
<tr>
<td>4.00</td>
<td>82.00</td>
<td>70</td>
<td>12</td>
</tr>
<tr>
<td>4.50</td>
<td>89.00</td>
<td>75</td>
<td>14</td>
</tr>
<tr>
<td>5.00</td>
<td>95.00</td>
<td>80</td>
<td>15</td>
</tr>
</tbody>
</table>

The error distances display a clear pattern. The Park Service’s regression line on the scatter plot shows the same reality. They keep increasing by small increments. The formula \( I = 10 \cdot D + 30 \) no longer works as a good predictor. In fact, it is a worse predictor for longer eruptions.
Instead of using the old formula, the National Park Service has a chart like the one in this example for visitors when they want to gauge how long it will be until the next eruption. We can take the chart the National Park Service uses and see what the new regression line would be. But first, does the scatter plot on the previous page look like we should use a linear model? And, do the $y$-values of the data points in the chart have roughly a constant difference?

The answer to both questions is “yes.” The data points do look as though a linear model would fit. The differences in intervals are all 5s, 6s, and 7s. In cases like this, you can use technology to find a linear regression equation by entering the data points in the STAT feature of your calculator.

The technology determines data points for the new trend line that appear to fit the observed data points much better than the old line. The interval-predicting equation has new parameters for the model, $a = 12.36$ (up 2.36 minutes) and $b = 33.2$ (up 3.2 minutes). The new regression line would be $y = 12.36x + 33.2$. While the new regression line appears to come much closer to the observed data points, there are still error distances, especially for lesser duration times. The scientists at Yellowstone Park believe that there probably should be two regression lines now: one for use with shorter eruptions and another for longer eruptions. As we saw from the frequency distribution earlier, Old Faithful currently tends to have longer eruptions that are farther apart.
The technology also provides a correlation coefficient. From the picture of the regression points on the previous page, it looks like the number should be positive and fairly close to 1. Using the linear regression feature on the calculator, we get $r = 0.9992$. Indeed, the length of the interval between Old Faithful’s eruptions is very strongly related to its most recent eruption duration. The direction is positive, confirming the longer the eruption, the longer the interval between eruptions.

It is very important to point out that the length of Old Faithful’s eruptions does not directly cause the interval to be longer or shorter between eruptions. The reason it takes longer for Old Faithful to erupt again after a long eruption is not technically known. However, with a correlation coefficient so close to 1, the two variables are closely related to one another. However, you should never confuse correlation with causation. For example, research shows a correlation between income and age, but aging is not the reason for an increased income. Not all people earn more money the longer they live. Variables can be related to each other without one causing the other.

**Correlation** is when two or more things or events tend to occur at about the same time and might be associated with each other but are not necessarily connected by a cause/effect relationship. **Causation** is when one event occurs as a direct result of another event. For example, a runny nose and a sore throat may correlate to each other but that does not mean a sore throat causes a runny nose or a runny nose causes a sore throat. Another example is it is raining outside and the ground being wet. There is a correlation between how wet the ground gets and how much it rains. In this case, the rain is what caused the ground to get more wet, so there is causation.

**Example:**

Consider the correlation between the age, in years, of a person and the income, in dollars, each person earns in the table below.

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
<th>45</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Income (dollars)</td>
<td>30,000</td>
<td>37,000</td>
<td>43,000</td>
<td>39,000</td>
<td>53,000</td>
<td>54,000</td>
</tr>
</tbody>
</table>

It appears there is a correlation between age and income and that a person’s income increases as the person gets older. This does not mean age causes a person’s income.
REVIEW EXAMPLES

1. This scatter plot suggests a relationship between the variables age and income.

![Graph showing the relationship between age and yearly income.]

a. What type of a relationship is suggested by the scatter plot (positive/negative, weak/strong)?

b. What is the domain of ages considered by the researchers?

c. What is the range of incomes?

d. Do you think age causes income level to increase? Why or why not?

Solution:

a. The scatter plot suggests a fairly strong positive relationship between age and yearly income.

b. The domain of ages considered is 18 to 60 years.

c. The range of incomes appears to be $10,000 to $70,000.

d. No; the variables are related, but age does not cause income to increase.
2. A group of researchers looked at income and age in Singapore. Their results are shown below. They used line graphs instead of scatter plots so they could consider the type of occupation of the wage earner.

![Income by Occupation and Age](image)

**Key**
- Managers
- Professionals
- Associate professionals and technicians

a. Does there appear to be a relationship between age and income?
b. Do all three types of occupations appear to share the same benefit of aging when it comes to income?
c. Does a linear model appear to fit the data for any of the occupation types?
d. Does the relationship between age and income vary over a person’s lifetime?

**Solution:**
a. Yes, as people get older their income tends to increase.
b. No. The incomes grow at different rates until age 40. For example, the managers’ incomes grow faster than those of the other occupation types until age 40.
c. No. The rate of growth appears to vary for all three occupations, making a linear model unsuitable for modeling this relationship over a longer domain.
d. Yes, after about age 40, the income for each type of occupation grows slower than it did from age 22 to 40.
3. Consider the table below.

<table>
<thead>
<tr>
<th>Age (years)</th>
<th>Yearly Income (dollars × 1,000)</th>
<th>Prediction (dollars × 1,000)</th>
<th>Error Distances</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>10</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>24</td>
<td>10</td>
<td>17</td>
<td>7</td>
</tr>
<tr>
<td>27</td>
<td>27</td>
<td>23</td>
<td>4</td>
</tr>
<tr>
<td>30</td>
<td>35</td>
<td>29</td>
<td>6</td>
</tr>
<tr>
<td>32</td>
<td>28</td>
<td>32</td>
<td>6</td>
</tr>
<tr>
<td>36</td>
<td>28</td>
<td>37</td>
<td>9</td>
</tr>
<tr>
<td>39</td>
<td>45</td>
<td>40</td>
<td>5</td>
</tr>
<tr>
<td>42</td>
<td>50</td>
<td>49</td>
<td>1</td>
</tr>
<tr>
<td>45</td>
<td>45</td>
<td>50</td>
<td>5</td>
</tr>
</tbody>
</table>

Do you think the regression line for this table is a good predictor of yearly income?

**Solution:**
The regression line for this table would be a good predictor of yearly income because the sum of the error differences is zero.
SAMPLE ITEMS

1. This graph plots the number of wins in the 2006 season and in the 2007 season for a sample of professional football teams.

Based on the regression model, what is the predicted number of 2007 wins for a team that won 5 games in 2006?

A. 4  
B. 7  
C. 8  
D. 12

Correct Answer: C
2. Which BEST describes the correlation of the two variables shown in the scatter plot?

![Scatter plot](image)

- A. weak positive
- B. strong positive
- C. weak negative
- D. strong negative

**Correct Answer:** D

3. Which of these statements is an example of causation?

- A. When the weather becomes warmer, more meat is purchased at the supermarket.
- B. More people go to the mall when students go back to school.
- C. The greater the number of new television shows, the fewer the number of moviegoers.
- D. After operating costs are paid at a toy shop, as more toys are sold, more money is made.

**Correct Answer:** D
4. To rent a carpet cleaner at the hardware store, there is a set fee and an hourly rate. The rental cost, \( c \), can be determined using this equation when the carpet cleaner is rented for \( h \) hours.

\[
c = 25 + 3h
\]

Which of these is the hourly rate?

A. 3
B. 3\( h \)
C. 25
D. 25\( h \)

Correct Answer: A