# Unit 4: 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. 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.

## KEY STANDARDS

### 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, ~~standard deviation~~) of two or more different data sets.★ (*Standard deviation is left for Advanced Algebra, use MAD as a measure of spread*.)

**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).★

### 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** Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear, ~~quadratic~~, and exponential models.★

**MCC9-12.S.ID.6b** Informally assess the fit of a function by plotting and analyzing residuals.★

**MCC9-12.S.ID.6c** Fit a linear function for a scatter plot that suggests a linear association.★

### 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 of a linear fit.★

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

# SUMMARIZE, REPRESENT, AND INTERPRET DATA ON A SINGLE COUNT OR MEASURABLE VARIABLE

## 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 numerical order. 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 of lowest to highest.

   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 $Q_3 - Q_1$.

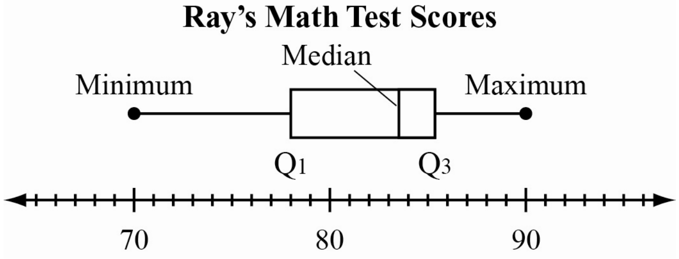
**Example:**

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.

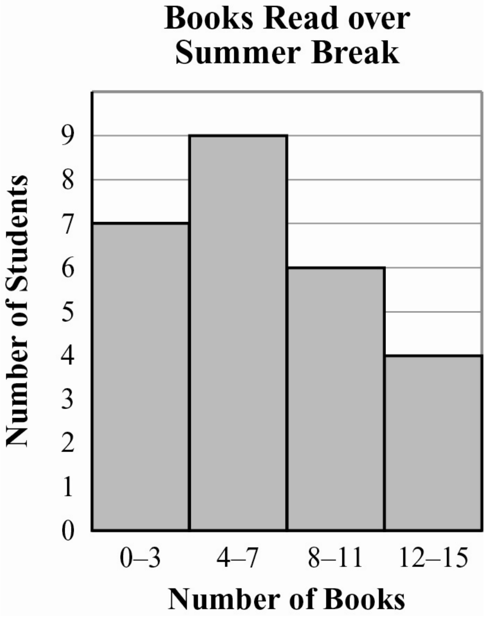
5. The most common displays for quantitative data are dot plots, histograms, box plots, and frequency distributions. A ***box plot*** is a diagram used to display a data set that uses quartiles to form the center box and the minimum and maximum to form the whiskers.

**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 $Q_1 - 1.5 \cdot \text{IQR}$ or above $Q_3 + 1.5 \cdot \text{IQR}$**.**

**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. For example, 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. A ***normal distribution*** shows quantitative data that vary randomly from a mean. The pattern follows a symmetrical, bell-shaped curve called a normal curve. As the distance from the mean increases on both sides of the mean, the number of data points decreases. ***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 positve 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.

8. Another way to describe the variability of a set of data is to use its ***mean absolute deviation***. The mean absolute deviation is the average distance between each data value and the mean.

**Example**:

The table shown below displays the running times for science-fiction movies.

| Running Times for Movies (min) | | | | | |
|---|---|---|---|---|---|
| 98 | 87 | 93 | 88 | 126 | 108 |

Find the mean of the movie running times.

$$\frac{98+87+93+88+126+108}{6}=100$$

Find the absolute value of the differences between each data value and the mean.

$|98 - 100| = 2$

$|87 - 100| = 13$

$|93 - 100| = 7$

$|88 - 100| = 12$

$|126 - 100| = 26$

$|108 - 100| = 8$

Then, average the differences.

$$\frac{2+13+7+12+26+8}{6}=11.33$$

The mean absolute deviation for the movie running times is about 11.33.

### *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.

## REVIEW EXAMPLES

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

**Total Tips by Day**

| Day | Josh's Tips | Richard's Tips |
|---|---|---|
| Monday | $40 | $40 |
| Tuesday | $20 | $45 |
| Wednesday | $36 | $53 |
| Thursday | $28 | $41 |
| Friday | $31 | $28 |

a. Who had the greatest 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 Richard's earnings from tips?

**Solution:**

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

$20, $28, **$31**, $36, $40

Josh's median earnings from tips is $31.

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

$28, $40, **$41**, $45, $53

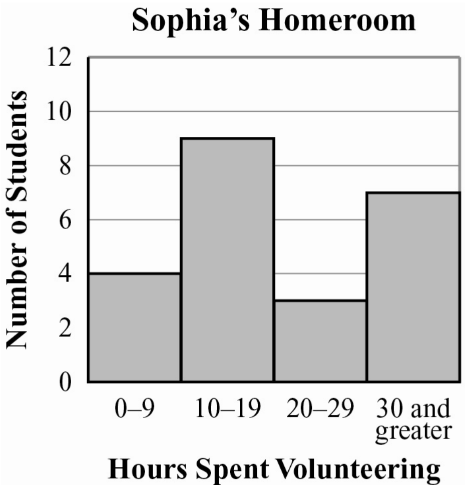
Richard had the greatest median earnings from tips. The difference in the median of the earnings from tips is $41 – $31 = $10.

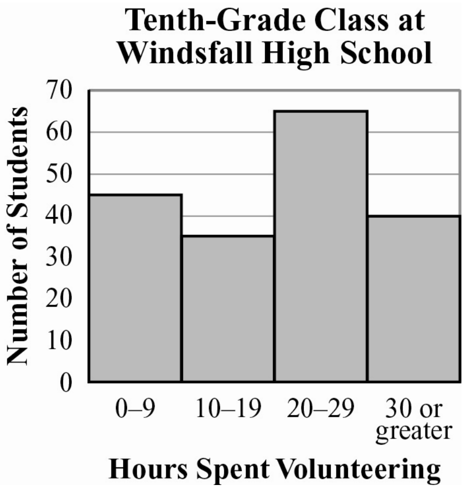
b. For Josh's earnings from tips, the lower quartile is $24 and the upper quartile is $38. The interquartile range $38 – $24, or $14.

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

d. Does either of the histograms reflect a normal distribution? Explain your answer.

**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 0–9 hours interval.

You can add the numbers 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 lower quartile of the number of hours spent volunteering by each student in Sophia's is less than the interval as 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 either within 30 and 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 probably more 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 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 number of hours spent volunteering by each student in the tenth-grade class.

d. Neither histogram appears to reflect a normal distribution. If the distribution were normal, most of the number of hours spent volunteering would be represented by the middle bars. The heights of the bars in the histogram for Sophia's homeroom vary without showing a definite pattern of skewness. The histogram for the tenth-grade class is slightly skewed to the left.

3) Mr. Storer, the physical education teacher, measured the height of the students in his first period class. He organized his data in this chart.

| Height (in inches) | Frequency |
|---|---|
| 42 | 1 |
| 43 | 2 |
| 44 | 4 |
| 45 | 5 |
| 46 | 4 |
| 47 | 2 |
| 48 | 1 |

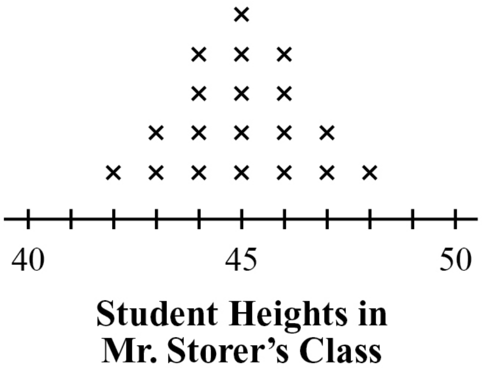
a. Make a dot plot for the data.

b. Make a histogram for the data.

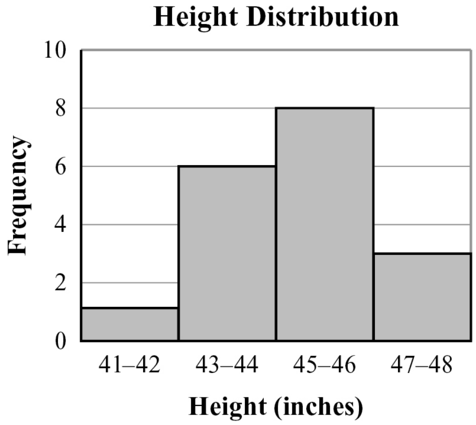
c. Make a box plot for the data.

d. Does the distribution of heights appear normal/bell shaped?

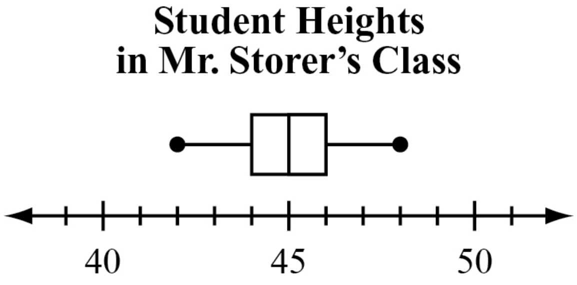
**Solution:**



a.



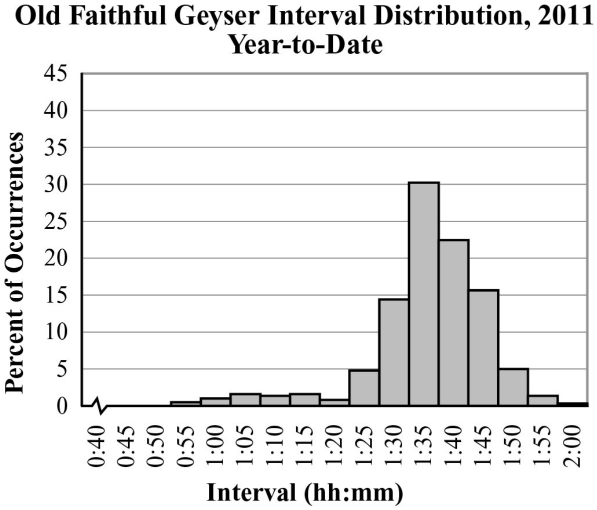
b.

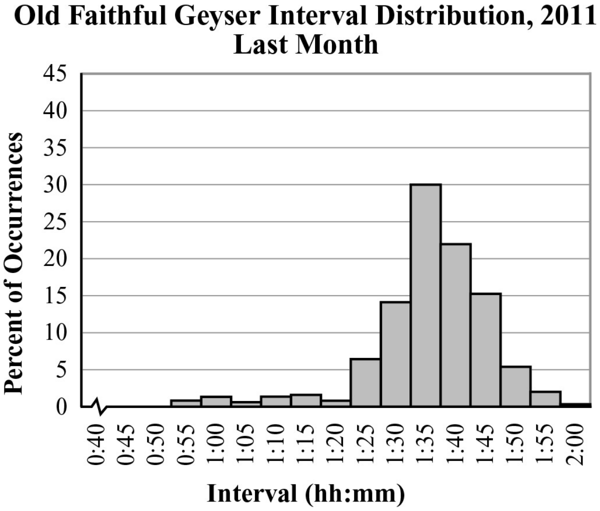


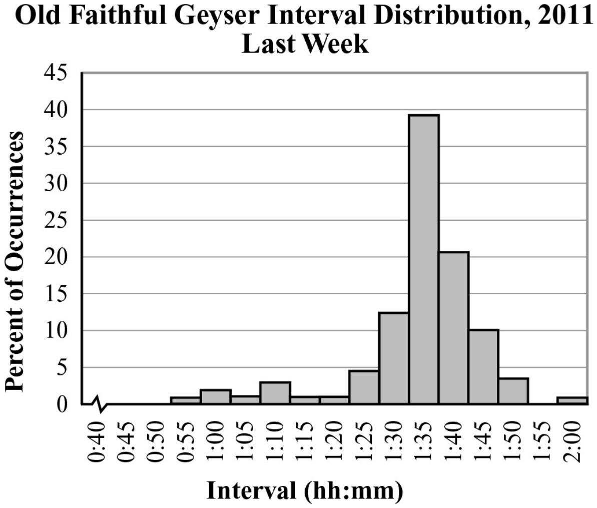
c.

d. Yes, the distribution of student heights appears fairly normal with a concentration in the middle and lesser frequencies in the tails.

4) Old Faithful, a geyser in Yellowstone National park, is renowned for erupting fairly regularly. In more recent times, it has become less predictable. It was observed that the time interval between eruptions was related to the duration of the most recent eruption. The distribution of its interval times for 2011 is shown below.







a. Does the Year-to-Date distribution seem normal, skewed, or uniform?

b. Compare Last Week's distribution to Last Month's.

c. What does the Year-to-Date distribution tell you about the interval of time between Old Faithful's eruptions?

**Solution**:

a. The Year-to-Date distribution appears to be skewed to the left (negative). Most of the intervals approach 90 minutes. In a normal distribution, the peak would be in the middle.

b. Last Week's distribution seems more skewed to the left than Last Month's. It is also more asymmetric. Last Month's distribution appears to have the highest percent of intervals longer than 1 hour 30 minutes between eruptions.

c. The Year-to-Date distribution shows Old Faithful rarely erupts an hour after its previous eruption. Most visitors will have to wait more than 90 minutes to see two eruptions.

5) The top five salaries and bottom five salaries at Technology Incorporated are shown in the table below.

| **Salaries at Technology Incorporated (in thousands)** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Top Five Salaries** | | | | | **Bottom Five Salaries** | | | | |
| \$90.4 | \$73 | \$69 | \$68.5 | \$61 | \$29 | \$28.5 | \$25.6 | \$24.5 | \$22.8 |

a. Find the mean absolute deviation for each set of data. Round to the nearest hundredth.

b. Compare the variations of the data sets.

**Solution:**

a. Find the mean absolute deviation of the top five salaries.

Find the mean of the top five salaries.

$$\frac{90.4+73+69+68.5+61}{5}=72.38$$

Find the absolute value of the differences between each data value and the mean. Then, average the differences.

$|90.4 - 72.38| = 18.02$

$|73 - 72.38| = 0.62$

$|69 - 72.38| = 3.38$

$|68.5 - 72.38| = 3.88$

$|61 - 72.38| = 11.38$

$$\frac{18.02+0.62+3.38+3.88+11.38}{5}=7.46$$

The mean absolute deviation for the top five salaries is about $7460.

Find the mean absolute deviation of the bottom five salaries.

Find the mean of the bottom five salaries.

$$\frac{29+28.5+25.6+24.5+22.8}{5}=26.08$$

Find the absolute value of the differences between each data value and the mean. Then, average the differences.

$|29 – 26.08| = 2.92$

$|28.5 – 26.08| = 2.42$

$|25.6 – 26.08| = 0.48$

$|24.5 – 26.08| =1.58$

$|22.8 – 26.08| = 3.28$

$$\frac{2.92+2.42+0.48+1.58+3.28}{5}=2.14$$

The mean absolute deviation for the bottom five salaries is about $2140.

b. The mean absolute deviation for the bottom five salaries is less than the mean absolute value for the top five salaries. There is less variability in the data values for the bottom five salaries than there is with the data values for the top five salaries.

## *EOCT Practice Items*

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

| **Day** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|
| **Temperature, in °F, in Macon, GA** | **71** | **72** | **66** | **69** | **71** | **73** |
| **Temperature, in °F, in Charlotte, NC** | **69** | **64** | **68** | **74** | **71** | **75** |

**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 lower than the lower quartile for the temperatures in Charlotte.

**C.** The mean and median temperatures of Macon were higher than the mean and median temperatures of Charlotte.

**D.** The upper quartile for the temperatures in Charlotte was lower than the upper quartile for the temperatures in Macon.

[Key: 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 ranges of the number of coats collected per class for the freshman and for the sophomores.**

- **The freshman 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 classes?**

**A.** 4 to 14

**B.** 6 to 14

**C.** 8 to 15

**D.** 12 to 15

[Key: D]

**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.

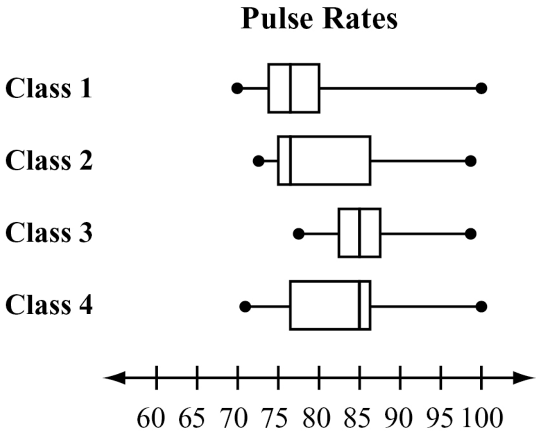
**B.** The lower quartile is 44.

**C.** The median is 45.

**D.** The maximum is 51.

[Key: B]

**4) A science teacher recorded the pulse rates for 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 had the highest pulse rates after climbing the stairs?**

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

[Key: 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 of 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.

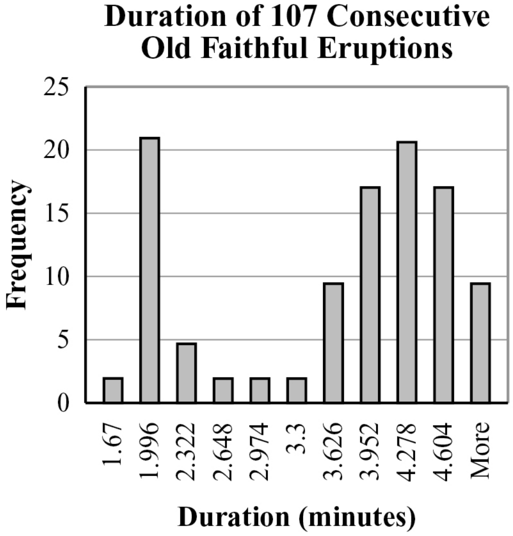
**B.** Peter had one very bad score in Week 2.

**C.** Peter did not improve as he did the first week.

**D.** Peter had one very good score in Week 1.

[Key: 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.** multi-outliers

**D.** skewed to the right

[Key: A]

**7) A teacher determined the median scores and interquartile ranges of scores for a test she gave to two classes.**

- **In Class 1, the median score was 70 points, and the interquartile range was 15 points.**
- **In Class 2, the median score was 75 points, and the interquartile range was 12 points.**

**Which range of numbers includes only third quartile of scores for both classes?**

**A.** 70 to 87 points
**B.** 70 to 85 points
**C.** 75 to 87 points
**D.** 75 to 85 points

[Key: D]

**8) This table shows admission price for various museums in the same city.**

| Museum Prices | | | | |
|---|---|---|---|---|
| $9.00 | $12.00 | $9.75 | $8.25 | $11.25 |

**Which is the mean absolute deviation for this set of data?**

**A.** $1.26
**B.** $6.30
**C.** $10.05
**D.** $10.13

[Key: A]

# SUMMARIZE, REPRESENT, AND INTERPRET DATA ON TWO CATEGORICAL AND QUANTITATIVE VARIABLES

## KEY IDEAS

1. There are essentially two types of data: ***quantitative*** and ***categorical***. 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.

2. ***Bivariate data*** consists 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 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 chart below, we see that there are 12 males in the band and 3 females in the chess club.

**Participation in School Activities**

| **School Club** | **Gender** | | |
|---|---|---|---|
| | Male | Female | **Totals** |
| Band | 12 | 21 | 33 |
| Chorus | 15 | 17 | 32 |
| Chess | 16 | 3 | 19 |
| Latin | 7 | 9 | 16 |
| Yearbook | 28 | 7 | 35 |
| **Totals** | 78 | 57 | **135** |

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 percents 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 boys as girls.

**Participation in School Activities**

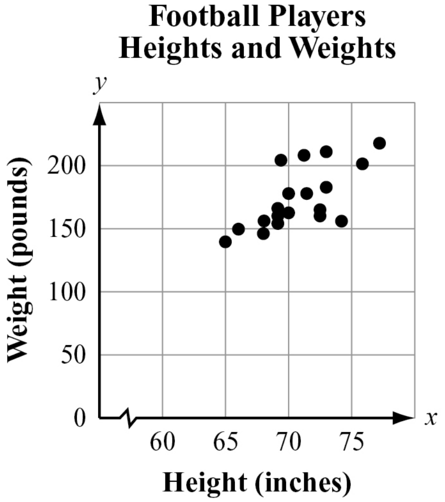
| **School Club** | **Gender** | | |
|---|---|---|---|
| | **Male** | **Female** | **Totals** |
| Band | 8.9% | 15.6% | 24.5% |
| Chorus | 11.1% | 12.6% | 23.7% |
| Chess | 11.9% | 2.2% | 14.1% |
| Latin | 5.2% | 6.7% | 11.9% |
| Yearbook | 20.7% | 5.2% | 25.9% |
| **Totals** | 57.8% | 42.3% | **100%** |

Note that the total, 100%, is a rounded value.

There are also what we call ***marginal frequencies*** in the bottom and right margins (grayed cells). These frequencies lack one of the categories. For our example, the frequencies at the bottom represent percents of males and females in the school population. The marginal frequencies on the right represent percents 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 percent 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 42.3% of the student body is female, then 12.6% / 42.3%, or 29.8%, of the females in the school are in the chorus (also 17 of 57 girls).

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



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 above 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 earlier, 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. For the purposes of this unit we will consider linear models.

**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.

| **Class 1 Test Score Analysis** | |
|---|---|
| **Mean Study Time (hours)** | **Mean Test Score** |
| 0.5 | 63 |
| 1 | 67 |
| 1.5 | 72 |
| 2 | 76 |
| 2.5 | 80 |
| 3 | 85 |
| 3.5 | 89 |

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.

| **Class 2 Test Score Analysis** | |
|---|---|
| **Mean Study Time (hours)** | **Mean Test Score** |
| 0.5 | 60 |
| 1 | 61 |
| 1.5 | 63 |
| 2 | 68 |
| 2.5 | 74 |
| 3 | 82 |
| 3.5 | 93 |

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. The ***second differences*** are 1, 3, 1, 2, and 3. Since the second differences are fairly close to constant, it is likely that a different model known as an exponential function would be employed for the Class 2 data.

This table shows Melissa’s data for Class 3.

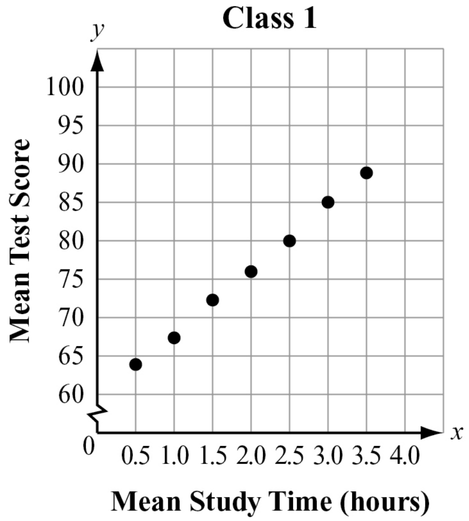
| **Class 3 Test Score Analysis** | |
|---|---|
| **Mean Study Time (hours)** | **Mean Test Score** |
| 0.5 | 71 |
| 1 | 94 |
| 1.5 | 87 |
| 2 | 98 |
| 2.5 | 69 |
| 3 | 78 |
| 3.5 | 91 |

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.

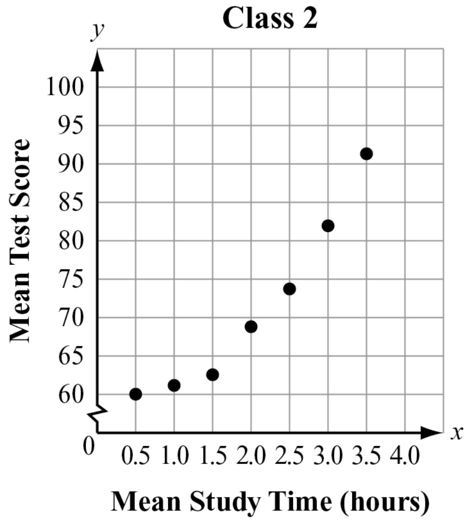
**Example:**

This graph shows Melissa's data for Class 1.



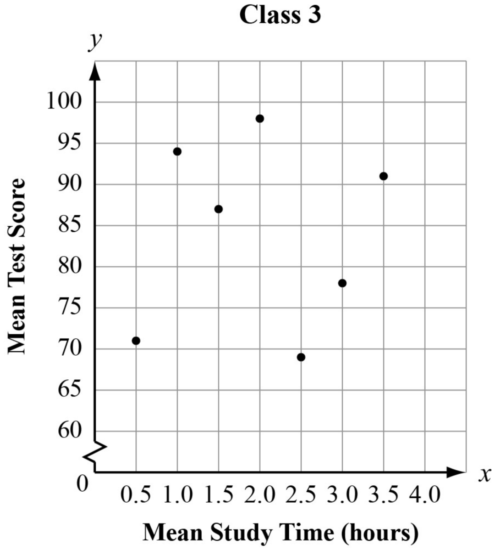
In this graph, the data points are all very close to being on the same line. This is further confirmation that a linear model is appropriate for this class.

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 an 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. Neither a linear model nor an exponential model is 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 an exponential model would produce a better fitting curved.

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.

**Items on the EOCT ask students to determine the equation of a line of best fit when given a graph. Students may also be asked to estimate a line of best fit for a given scatter plot. Items may require data interpretation. **

~~The median-median line is a way to estimate a line of best fit that involves relatively simple calculations. Since it involves using medians, it is also somewhat resistant to the effect of outliers in the data.~~

~~To calculate a median-median line, order the data from the least value to the greatest value of the $x$-coordinate. Order the data points, which have the same value of $x$, from the least to the greatest value of the $y$-coordinate. Next, use this ordering to divide the data into three equal groups. If the number of data values is not divisible by 3, split them up so that the first and last groups are the same size. Find the median $x$-coordinate value for~~

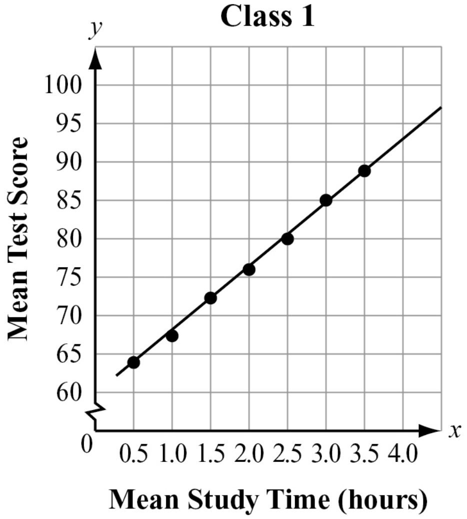
~~each of the three groups (low values, middle values, high values). Then find the median of the *y*-values associated with each of these data points. The median values of *x* and *y* may not be associated with the same data point. Find the equation of the line containing the points from the low-*x* value set and the high-*x* value set. Next adjust the position of the line by moving it $\frac{1}{3}$ of the way toward the middle point.~~

~~***Note*: Each item on the EOCT that asks students to find the median-median line requires this method of calculation. Graphing calculators are currently not permitted for use during the CCGPS Coordinate Algebra EOCT. Students should become familiar with this method as preparation for the assessment.**~~

**Note: Sample items have been adjusted to reflect the changes in the text on pages 147–148. Students do not have to solve line of best fit problems using the median-median approach. (10/03/2012)**

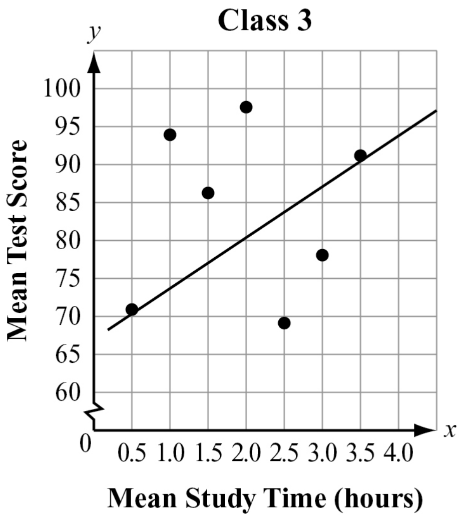
**Example:**

This graph shows Melissa's data for Class 1 with the line of best fit added. The equation of the line is $y = 8.8x + 58.4$.



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.

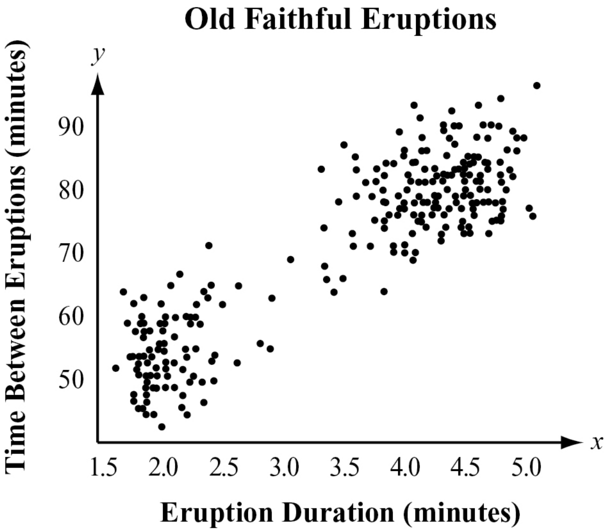
This next graph shows Melissa's data for Class 3 with the line of best fit added. The equation of the line is $y = 0.8\,x + 83.1$.



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.

## 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.



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. Do you think the scatter plot could be modeled with a linear function?

**Solution**:

a. After a 4-minute eruption, it would be between 75 to 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. There seem to be two major regions in the scatter plot, so a single line may not be a good predictor of both regions.

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.

| **Beverage Can Recycling** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Number of Canned Beverages Sold** | 18 | 15 | 19 | 8 | 10 | 13 | 9 | 14 |
| **Number of Cans Recycled** | 8 | 6 | 10 | 6 | 3 | 7 | 5 | 4 |

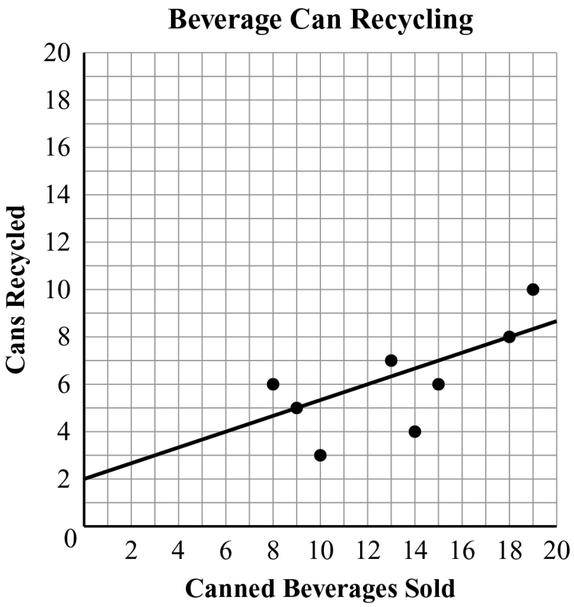
Find an equation of a line of best fit for the data.

**Solution:**

Write the data as ordered pairs.

{(8, 6) (9, 5) (10, 3) (13, 7) (14, 4) (15, 6) (18, 8) (19, 10)}

Plot the ordered pairs on a coordinate grid and draw a line that approximates the trend of the data. Draw the line so that it has about the same number of points above and below the line. It does not always have to cross directly through data points.



To find the slope of the line, choose two points on the line, such as (9, 5) and (18, 8). Then calculate the slope.

$$\text{Slope} = \frac{8-5}{18-9} = \frac{3}{9} = \frac{1}{3}$$

The line appears to cross the $y$-axis at (0, 2), so estimate the $y$-intercept of this line as 2.

So, the equation of a line of best fit for this data could be $y = \frac{1}{3}x + 2$.

2) A fast food restaurant wants to determine if the season of the year affects the choice of soft-drink size purchased. They surveyed 278 customers and the table below shows their 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 that fit both row and column titles. On the bottom and in the right margin are the totals.

| | **Spring** | **Summer** | **Fall** | **TOTALS** |
|---|---|---|---|---|
| **Small** | 24 | 22 | 18 | 64 |
| **Medium** | 23 | 28 | 19 | 70 |
| **Large** | 18 | 27 | 29 | 74 |
| **Jumbo** | 16 | 21 | 33 | 70 |
| **TOTALS** | 81 | 98 | 99 | 278 |

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

b. What percent of those surveyed purchased the small drinks?

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

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

**Solution:**

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

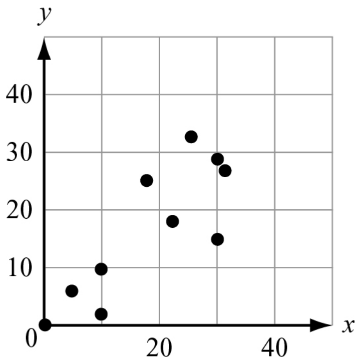
b. Twenty-three percent (64/278 = 23%) of the 278 surveyed purchased the small drinks.

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

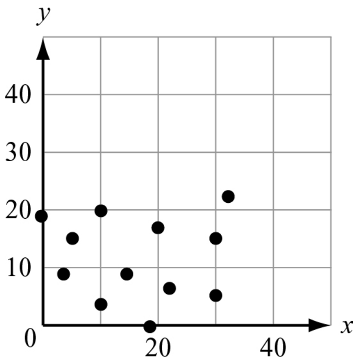
d. The fast food restaurant probably learned that customers tend to purchase the larger drinks in the spring and fall, the smaller drinks in the summer.

## *EOCT Practice Items*

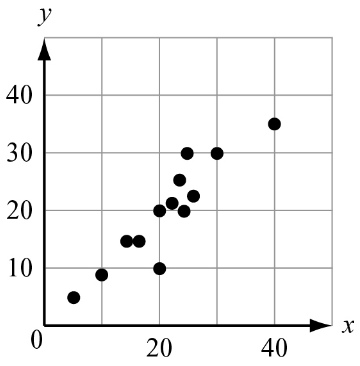
**1) Which graph displays a set of data for which a linear function is the model of best fit?**



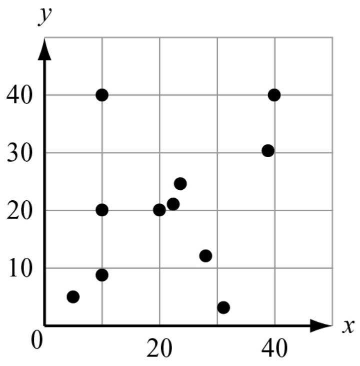
**A**



**B**



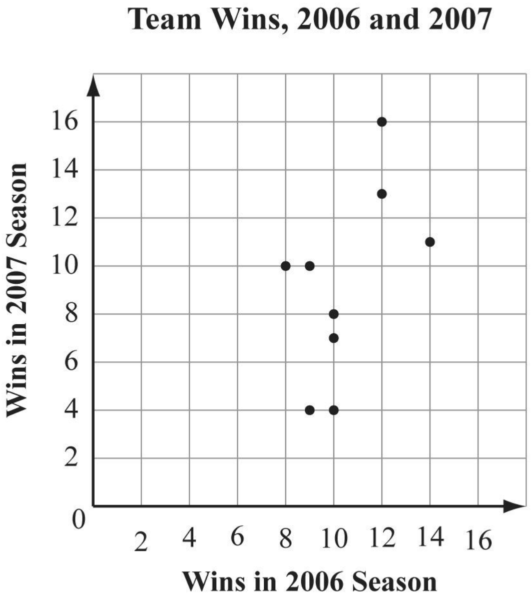
**C**



**D**

[Key: C]

**2) This graph plots the number of wins in the 2006 and 2007 seasons for a sample of professional football teams.**



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

**A.** $y = \frac{1}{2}x$

**B.** $y = \frac{1}{2}x + 8$

**C.** $y = 2x - 6$

**D.** $y = 2x - 12$

[Key: D]

# INTERPRET LINEAR MODELS

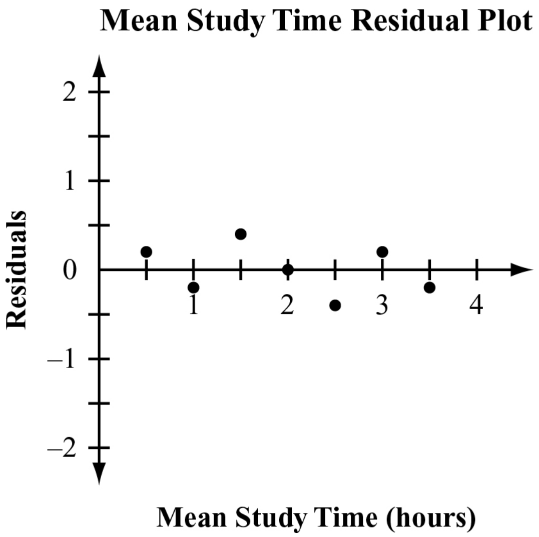
## KEY IDEAS

1. Once a model for the scatter plot is determined, its goodness of fit is very important. The goodness of fit depends on the model's accuracy in predicting values. ***Residuals***, or error distances, are used to measure the goodness of fit. A residual is the difference between the observed value and the model's predicted value. For a regression model, a residual = observed value – predicted value. A residual plot is a graph that shows the residual values on the vertical axis and the independent variable (*x*) on the horizontal axis. A residual plot shows where the model fits best, and where the fit is worst. A good regression fit has very short residuals.

**Example**:

Take the data from the test scores for Class 1 used in the last section. The observed mean test scores were 63, 67, 72, etc. The best fit model was a linear model with the equation $y = 8.8x + 58.4$. We can calculate the residuals for this data and consider the fit of the regression line.

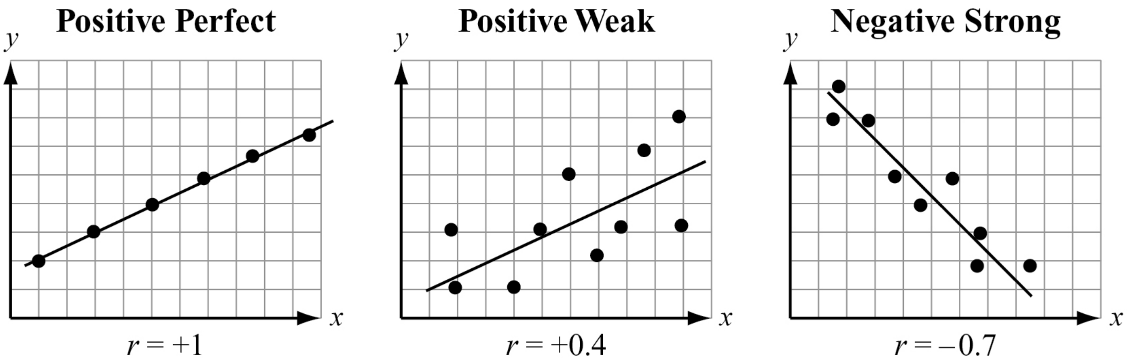
| Mean Study Time (hours) | Mean Test Score | Predicted Score $y = 8.8x + 58.4$ | Residual |
|---|---|---|---|
| 0.5 | 63 | 62.8 | 0.2 |
| 1.0 | 67 | 67.2 | -0.2 |
| 1.5 | 72 | 71.6 | 0.4 |
| 2.0 | 76 | 76 | 0 |
| 2.5 | 80 | 80.4 | -0.4 |
| 3.0 | 85 | 84.8 | 0.2 |
| 3.5 | 89 | 89.2 | -0.2 |



Notice the numbers in the residual column tell us how far the predicted mean test score was from the observed, as seen in the regression scatter plot for Class 1. The regression passes through one of the actual points in the plot of the points where the residual is 0. Notice also, the residuals add up to 0. Residuals add up to 0 for a properly calculated regression line. The goal is to minimize all of the residuals.

2. 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* get 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 positive slope and both variables are growing. However, if the correlation coefficient is negative, the line of best fit has negative slope and the dependent variable is decreasing. The numerical value is an indicator of how closely the data points come to the line.

**Example**:



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.

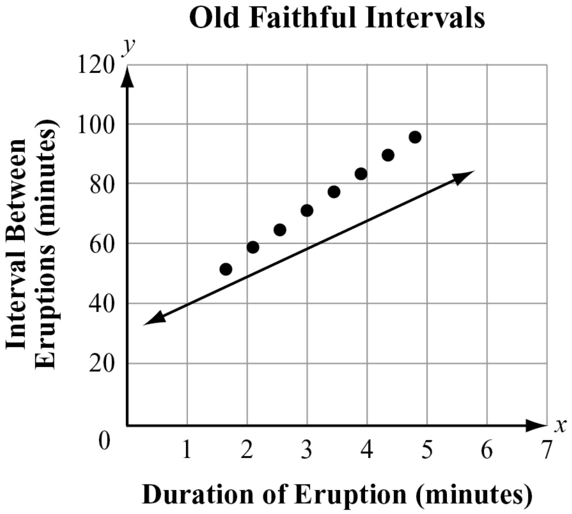
**Example:**

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×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 today. The data points are from a histogram with intervals of 0.5 minutes for *x*-values. The *y*-values are the average interval time for an eruption in that duration interval.

## Old Faithful Eruptions

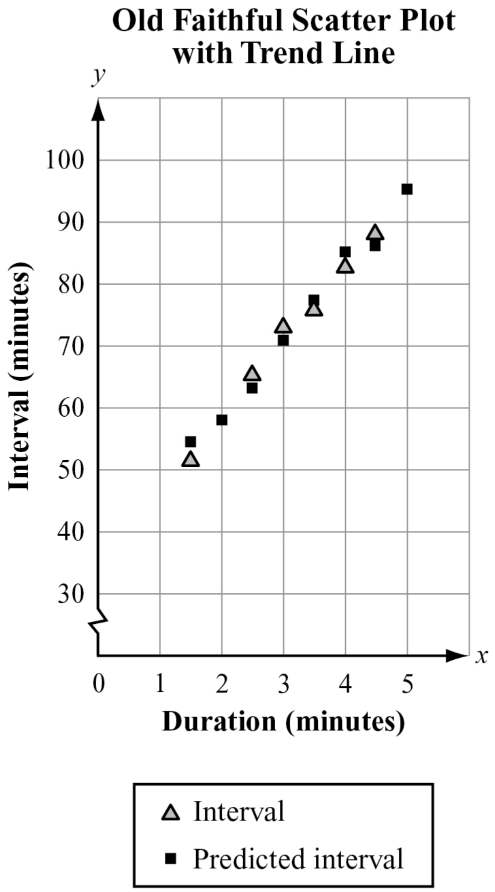
| Duration (*x*) | Interval (*y*) | Prediction | Residual |
|---|---|---|---|
| 1.50 | 51.00 | 45 | 6 |
| 2.00 | 58.00 | 50 | 8 |
| 2.50 | 65.00 | 55 | 10 |
| 3.00 | 71.00 | 60 | 11 |
| 3.50 | 76.00 | 65 | 11 |
| 4.00 | 82.00 | 70 | 12 |
| 4.50 | 89.00 | 75 | 14 |
| 5.00 | 95.00 | 80 | 15 |

The residuals display a pretty clear pattern. The Park Service's regression line on the scatter plot shows the same reality. They keep increasing by small increments. The formula, $I = 10 \times D + 30$ 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 above 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. This time we'll use statistical functions of the graphing calculator to find the linear regression equation.



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 $y = 12.36\,x + 33.2$. While the new regression line appears to come much closer to the observed data points, there are still residuals, 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 these days, and farther apart.

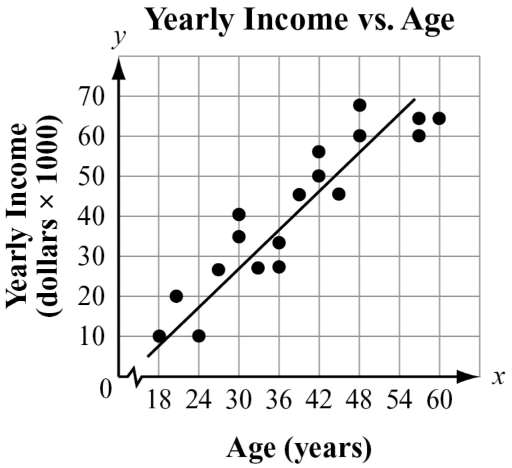
The technology also provides a correlation coefficient. From the picture of the regression points above, it looks like the number should be positive and fairly close to 1. The technology provided a correlation coefficient of 0.9992. Indeed, the length of the interval between Old Faithful's eruptions 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 that high, the two are definitely related to one another. You should never confuse correlation with causation. 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.

## REVIEW EXAMPLES

1) This scatter plot suggests a relationship between the variables age and income. Answer the questions below based on the pictured scatter plot.

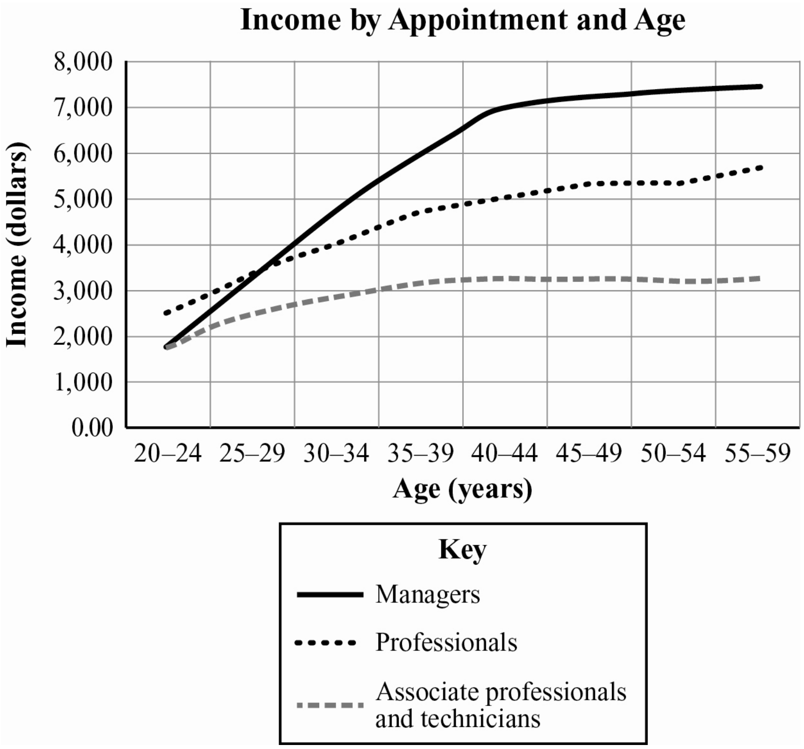


a. What type of a relationship is suggested by the scatter plot (positive/negative, weak/strong)?
b. What is the domain of ages considered by the researchers?
c. What is the range of incomes?
d. Do you think age causes income level to increase? Why or why not?

**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) Another 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.

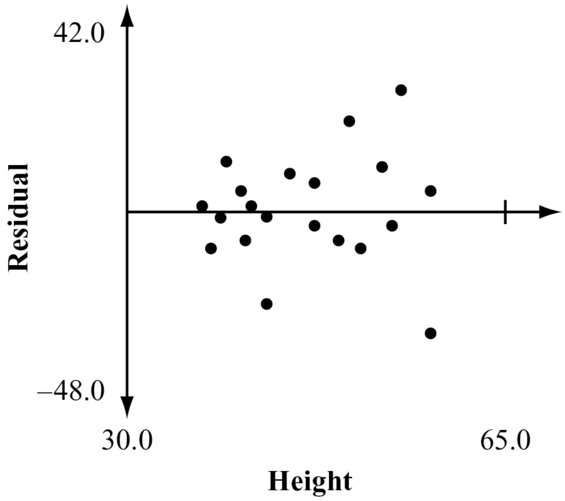


a. Does there appear to be a relationship between age and income?
b. Do all three types of employees appear to share the same benefit of aging when it comes to income?
c. Does a linear model appear to fit the data for any of the employee types?
d. Does the effect of age vary over a person's lifetime?

**Solution:**

a. Yes, as people get older their income tends to increase.
b. The managers appear to benefit more from increased age than either the professionals or the associates.
c. The rate of growth appears to vary for all three categories, making a linear model unsuitable for modeling this relationship over a longer domain.
d. The effect of increasing income appears to diminish with age, showing that income level does not always increase with age.

3) Consider the residual plot below. Each vertical segment represents the difference between an observed weight and a predicted weight of a person, based on height.



a. Do you think the regression line is a good predictor of weight?
b. Why do the residuals appear to be getting longer for greater heights?

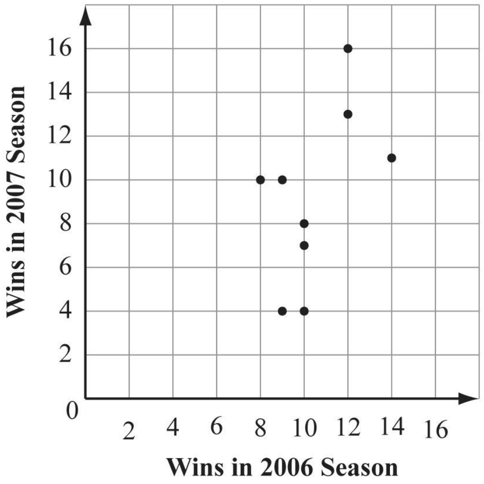
**Solution:**

a. The regression line does not appear to be a good predictor. Some of the predicted weights were off. The points on the regression plot art dispersed.
b. Many other factors not shown in the data can affect weight for taller people.

## *EOCT Practice Items*

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

**Team Wins, 2006 and 2007**

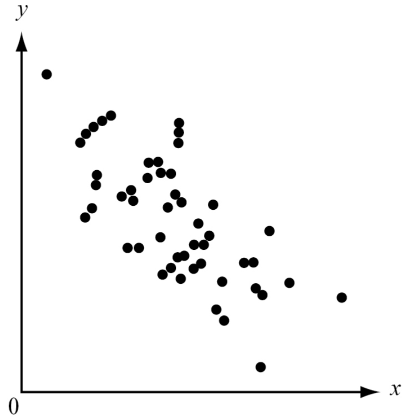


**Based on the regression model, what is the predicted number of 2007 wins for a team that won 5 games in 2006?**

**A.** 3
**B.** 4
**C.** 5
**D.** 6

[Key: A]

**2) How would you describe the correlation of the two variables based on the scatter plot?**



**A.** positive, strong linear
**B.** negative, weak linear
**C.** negative, fairly strong linear
**D.** little or no correlation

[Key: C]