Lesson 1: Distributions and Their Shapes

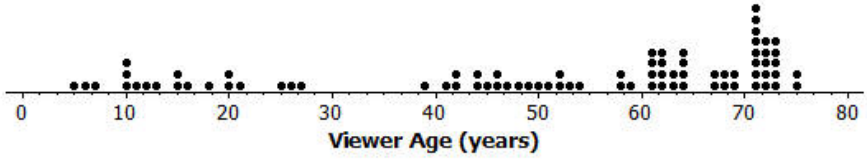
Classwork

Statistics is all about data. Without data to talk about or to analyze or to question, statistics would not exist. There is a story to be uncovered behind all data—a story that has characters, plots, and problems. The questions or problems addressed by the data and their story can be disappointing, exciting, or just plain ordinary. This module is about stories that begin with data.

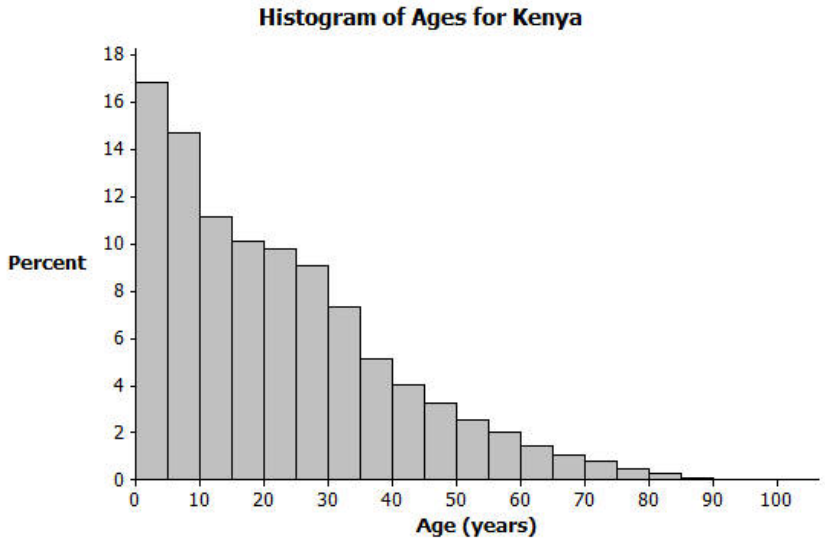
Example 1: Graphs

Data are often summarized by graphs; the graphs are the first indicator of variability in the data.

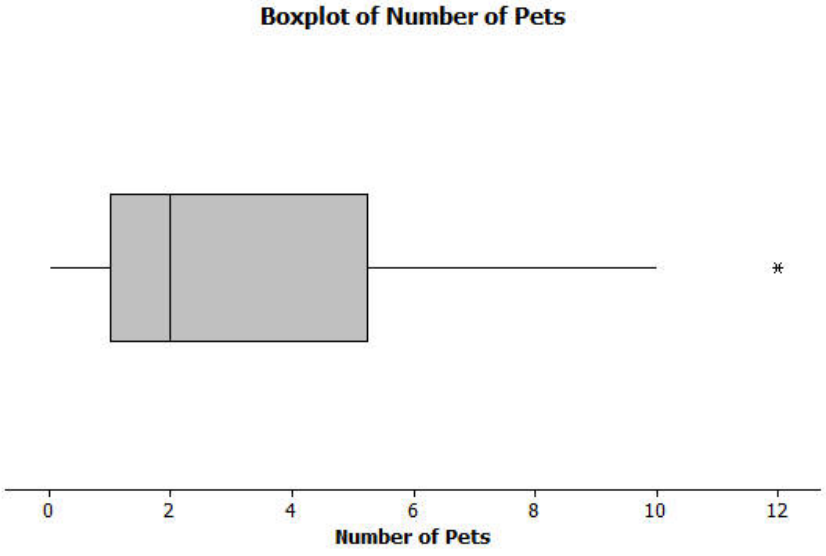
- **Dot plots**: A plot of each data value on a scale or number line.



- **Histograms**: A graph of data that groups the data based on intervals and represents the data in each interval by a bar.



- **Box plots**: A graph that provides a picture of the data ordered and divided into four intervals that each contains approximately 25% of the data.

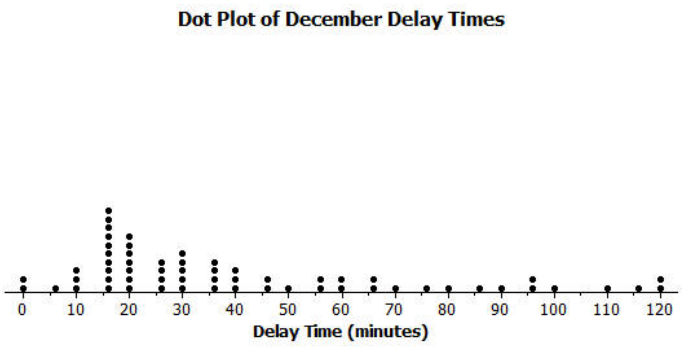


Exercises 1–15

Answer the questions that accompany each graph to begin your understanding of the story behind the data.

Transportation officials collect data on flight delays (the number of minutes past the scheduled departure time that a flight takes off).

Consider the dot plot of the delay times for sixty BigAir flights during December 2012.

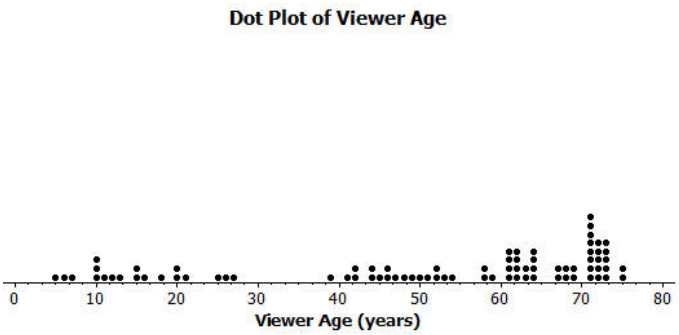


1. What do you think this graph is telling us about the flight delays for these sixty flights?

2. Can you think of a reason why the data presented by this graph provides important information? Who might be interested in this data distribution?

3. Based on your previous work with dot plots, would you describe this dot plot as representing a symmetric or a skewed data distribution? (Recall that a skewed data distribution is not mound shaped.) Explain your answer.

A random sample of eighty viewers of a television show was selected. The dot plot below shows the distribution of the ages (in years) of these eighty viewers.

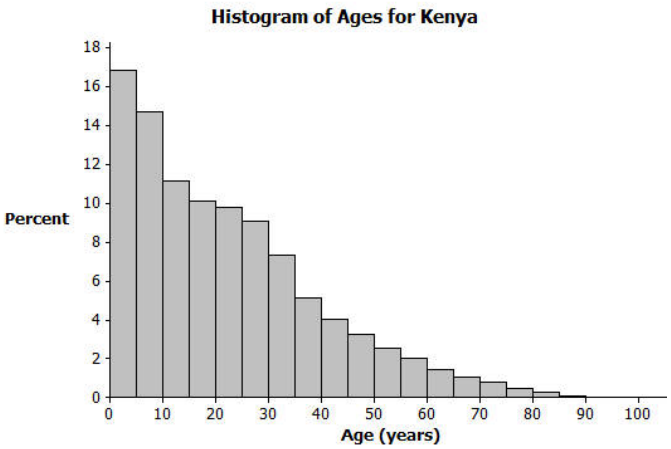


4. What do you think this graph is telling us about the ages of the eighty viewers in this sample?

5. Can you think of a reason why the data presented by this graph provides important information? Who might be interested in this data distribution?

6. Based on your previous work with dot plots, would you describe this dot plot as representing a symmetric or a skewed data distribution? Explain your answer.

The following histogram represents the age distribution of the population of Kenya in 2010.

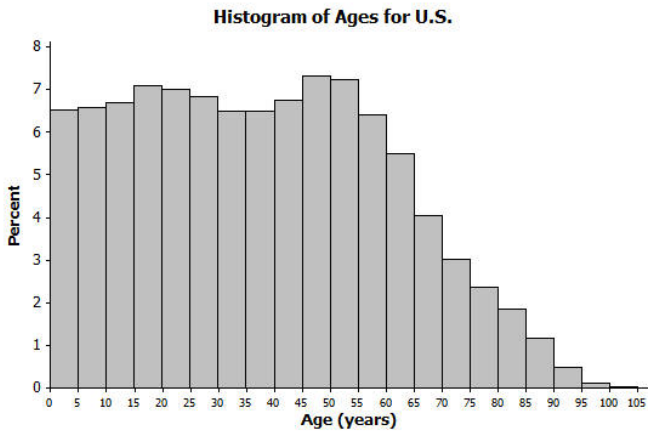


7. What do you think this graph is telling us about the population of Kenya?

8. Why might we want to study the data represented by this graph?

9. Based on your previous work with histograms, would you describe this histogram as representing a symmetrical or a skewed distribution? Explain your answer.

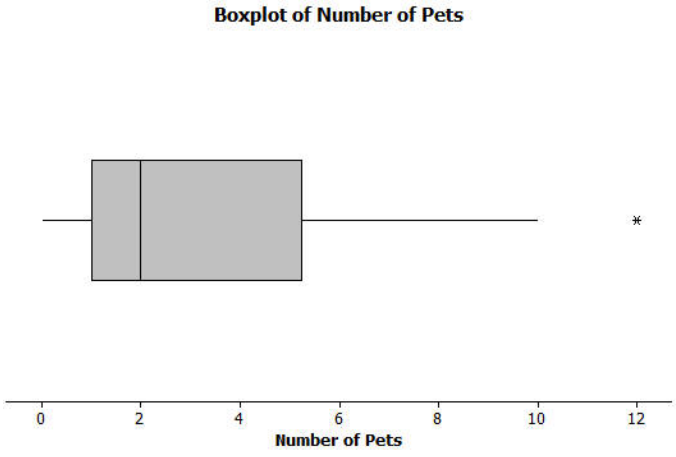
The following histogram represents the age distribution of the population of the United States in 2010.



10. What do you think this graph is telling us about the population of the United States?

11. Why might we want to study the data represented by this graph?

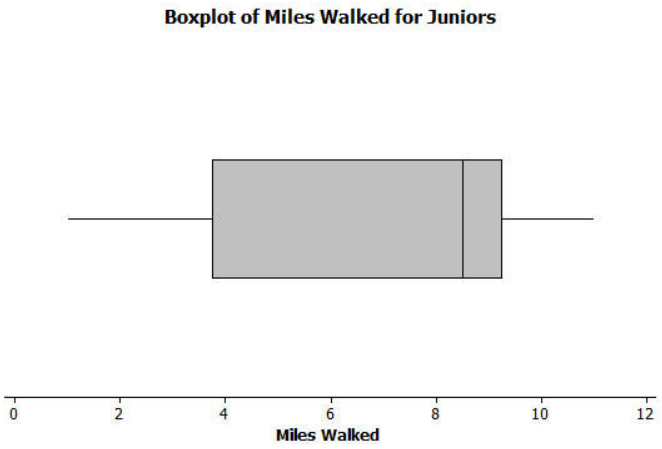
Thirty students from River City High School were asked how many pets they owned. The following box plot was prepared from their answers.



12. What does the box plot tell us about the number of pets owned by the thirty students at River City High School?

13. Why might understanding the data behind this graph be important?

Twenty-two juniors from River City High School participated in a walkathon to raise money for the school band. The following box plot was constructed using the number of miles walked by each of the twenty-two juniors.



14. What do you think the box plot tells us about the number of miles walked by the twenty-two juniors?

15. Why might understanding the data behind this graph be important?

Lesson Summary

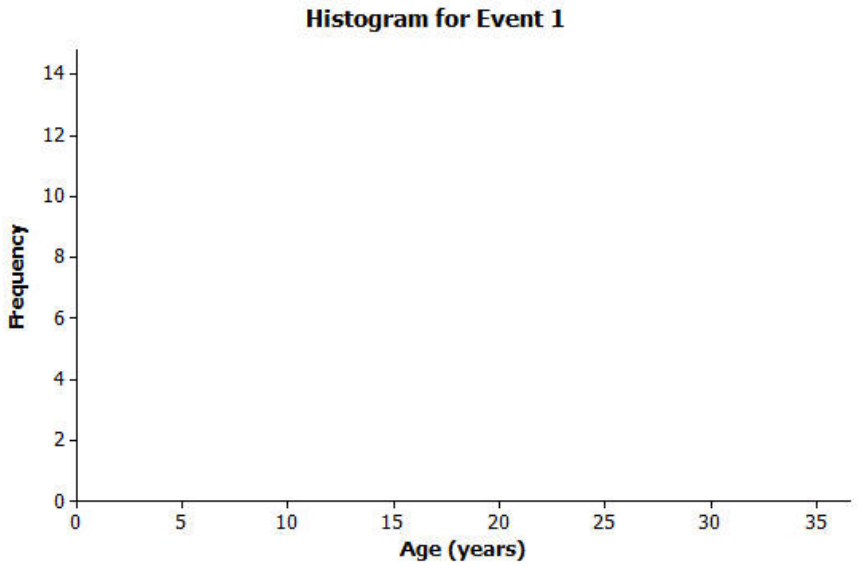
Statistics is about data. Graphs provide a representation of the data distribution and are used to understand the data and to answer questions about the distribution.

Problem Set

1. Twenty-five people were attending an event. The ages of the people are as follows:

$$3, 3, 4, 4, 4, 4, 5, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 16, 17, 22, 22, 25.$$

   a. Create a histogram of the ages using the provided axes.



   b. Would you describe your graph as symmetrical or skewed? Explain your choice.

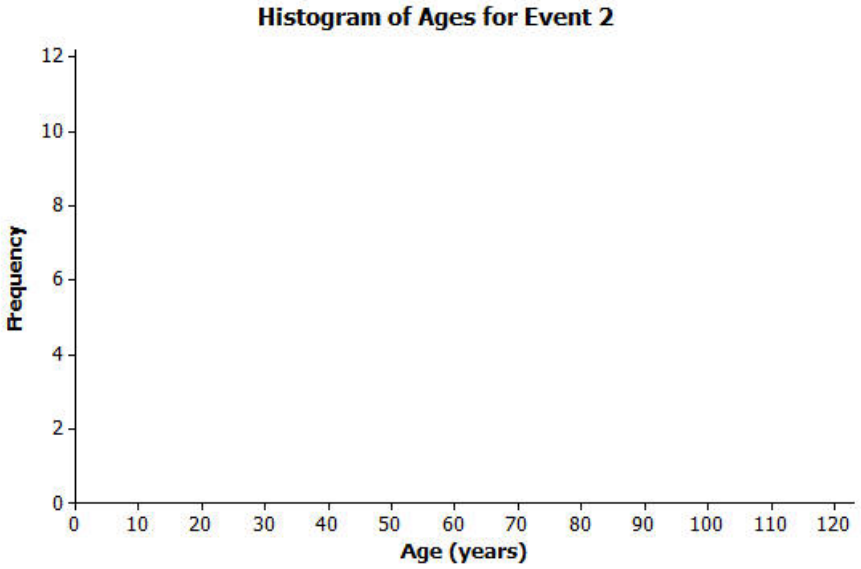
   c. Identify a typical age of the twenty-five people.

   d. What event do you think the twenty-five people were attending? Use your histogram to justify your conjecture.

2. A different forty people were also attending an event. The ages of the people are as follows:

6, 13, 24, 27, 28, 32, 32, 34, 38, 42, 42, 43, 48, 49, 49, 49, 51, 52, 52, 53,
53, 53, 54, 55, 56, 57, 57, 60, 61, 61, 62, 66, 66, 66, 68, 70, 72, 78, 83, 97.

a. Create a histogram of the ages using the provided axes.



b. Would you describe your graph of ages as symmetrical or skewed? Explain your choice.

c. Identify a typical age of the forty people.

d. What event do you think the forty people were attending? Use your histogram to justify your conjecture.

e. How would you describe the differences in the two histograms?

Lesson 2: Describing the Center of a Distribution

Classwork

In previous work with data distributions, you learned how to derive the mean and the median of a data distribution. This lesson builds on your previous work with a center.

Exploratory Challenge

Consider the following three sets of data.

Data Set 1: Pet owners

Students from River City High School were randomly selected and asked, "How many pets do you currently own?" The results are recorded below.

0	0	0	0	1	1	1	1	1	1	1	1	1	1	2
2	2	2	3	3	4	5	5	6	6	7	8	9	10	12

Data Set 2: Length of the east hallway at River City High School

Twenty students were selected to measure the length of the east hallway. Two marks were made on the hallway's floor, one at the front of the hallway, and one at the end of the hallway. Each student was given a meter stick and asked to use the meter stick to determine the length between the marks to the nearest tenth of a meter. The results are recorded below.

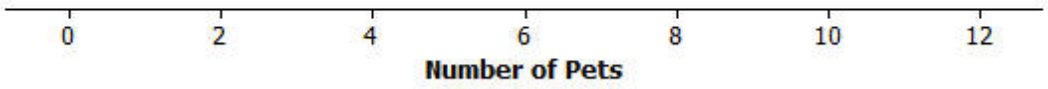
8.2	8.3	8.3	8.4	8.4	8.5	8.5	8.5	8.5	8.5
8.6	8.6	8.6	8.6	8.7	8.7	8.8	8.8	8.9	8.9

Data Set 3: Age of cars

Twenty-five car owners were asked the age of their cars in years. The results are recorded below.

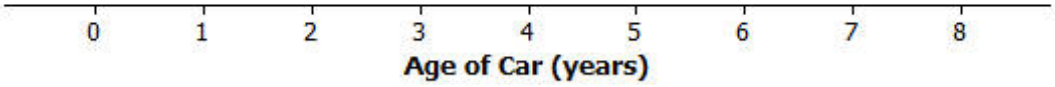
0	1	2	2	3	4	5	5	6	6	6	7	7
7	7	7	7	8	8	8	8	8	8	8	8	

1. Make dot plot of each of the data sets. Use the following scales.



Number of Pets

East Hallway Length Measurement (meters)



Age of Car (years)

2. Calculate the mean number of pets owned by the thirty students from River City High School. Calculate the median number of pets owned by the thirty students.

3. What do you think is a typical number of pets for students from River City High School? Explain how you made your estimate.

4. Why do you think that different students got different results when they measured the same distance of the east hallway?

5. What is the mean length of the east hallway data set? What is the median length?

6. A construction company will be installing a handrail along a wall from the beginning point to the ending point of the east hallway. The company asks you how long the handrail should be. What would you tell the company? Explain your answer.

7. Describe the distribution of the age of cars.

8. What is the mean age of the twenty-five cars? What is the median age? Why are the mean and the median different?

9. What number would you use as an estimate of the typical age of a car for the twenty-five car owners? Explain your answer.

Lesson Summary

- A dot plot provides a graphical representation of a data distribution, helping us to visualize the distribution.
- The mean and the median of the distribution are numerical summaries of the center of a data distribution.
- When the distribution is nearly symmetrical, the mean and the median of the distribution are approximately equal. When the distribution is not symmetrical (often described as skewed), the mean and the median are not the same.
- For symmetrical distributions, the mean is an appropriate choice for describing a typical value for the distribution. For skewed data distributions, the median is a better description of a typical value.

Problem Set

Consider the following scenario. The company that created a popular video game, "Leaders," plans to release a significant upgrade of the game. Users earn or lose points for making decisions as the leader of an imaginary country. In most cases, repeated playing of the game improves a user's ability to make decisions. The company will launch an online advertising campaign, but at the moment, they are not sure how to focus the advertising. Your goal is to help the company decide how the advertising campaign should be focused. Five videos have been proposed for the following target audiences:

Video 1: Target females with beginning level scores

Video 2: Target males with advanced level scores

Video 3: Target all users with middle range level scores

Video 4: Target males with beginning level scores

Video 5: Target females with advanced level scores

1. Why might the company be interested in the developing different videos based on user score?

2. Thirty female users and twenty-five male users were selected at random from a database of people who play the game regularly. Each of them agreed to be part of a research study and report their scores. A leadership score is based on a player's answers to leadership questions. A score of 1 to 40 is considered a beginning level leadership score, a score of 41 to 60 is considered a middle level leadership score, and a score of greater than 60 is considered an advanced level leadership score.

   Use the following data to make a dot plot of the female scores, a dot plot of the male scores, and a dot plot of the scores for the combined group of males and females.

Female scores:

10	20	20	20	30	30	30	40	40	40
50	50	55	65	65	65	65	65	70	70
70	70	76	76	76	76	76	76	76	76

Male scores:

15	20	20	25	25	25	25	30	30	30
30	30	30	35	35	35	35	35	40	40
40	45	45	45	50					

3. What do you think is a typical score for a female user? What do you think is a typical score for a male user? Explain how you determined these typical scores.

4. Why is it more difficult to report a typical score for the overall group that includes both the males and females?

5. Production costs will only allow for two video advertisements to be developed. Which two videos would you recommend for development? Explain your recommendations.

Lesson 3: Estimating Centers and Interpreting the Mean as a Balance Point

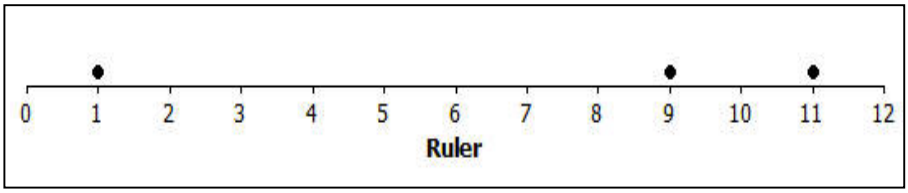
Classwork

Example 1

Your previous work in mathematics involved estimating a balance point of a data distribution. Let's review what we learned about the balance point of a distribution. A 12-inch ruler has several quarters taped to positions along the ruler. The broad side of a pencil is placed underneath the ruler to determine an approximate balance point of the ruler with the quarters.

Exercises 1–7

Consider the following example of quarters taped to a lightweight ruler.



1. Sam taped 3 quarters to his ruler. The quarters were taped to the positions 1 inch, 9 inches, and 11 inches. If the pencil was placed under the position 5 inches, do you think the ruler would balance? Why or why not?

2. If the ruler did not balance, would you move the pencil to the left or to the right of 5 inches to balance the ruler? Explain your answer.

3. Estimate a balance point for the ruler. Complete the following based on the position you selected.

Position of Quarter	Distance from Quarter to your Estimate of the Balance Point
1	
9	
11	

4. What is the sum of the distances to the right of your estimate of the balance point?

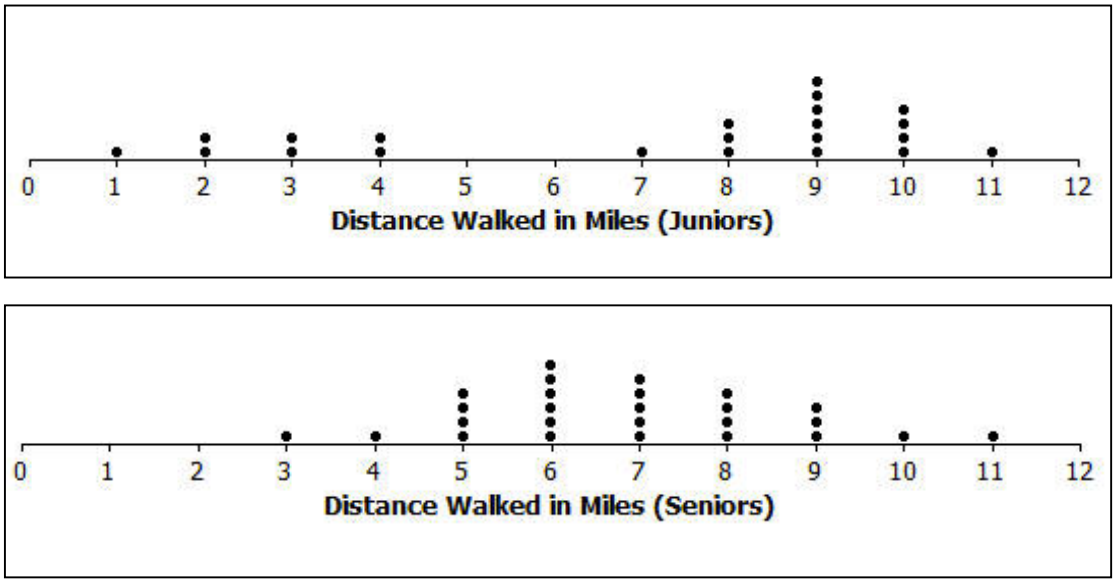
5. What is the sum of the distances to the left of your estimate of the balance point?

6. Do you need to adjust the position of your balance point? If yes, explain how.

7. Calculate the mean and the median of the position of the quarters. Does the mean or the median of the positions provide a better estimate of the balance point for the position of the 3 quarters taped to this ruler? Explain why you made this selection.

Exercises 8–20

Twenty-two students from the junior class and twenty-six students from the senior class at River City High School participated in a walkathon to raise money for the school's band. Dot plots indicating the distances in miles students from each class walked are as follows.



8. Estimate the mean number of miles walked by a junior, and mark it with an "X" on the junior class dot plot. How did you estimate this position?

9. What is the median of the junior data distribution?

10. Is the mean number of miles walked by a junior less than, approximately equal to, or greater than the median number of miles? If they are different, explain why. If they are approximately the same, explain why.

11. How would you describe the typical number of miles walked by a junior in this walkathon?

12. Estimate the mean number of miles walked by a senior, and mark it with an "X" on the senior class dot plot. How did you estimate this position?

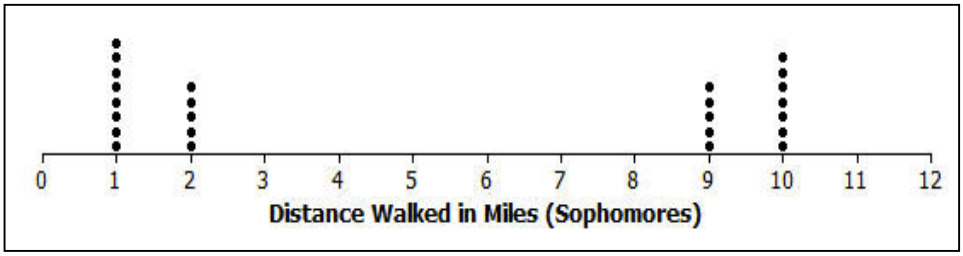
13. What is the median of the senior data distribution?

14. Estimate the mean and the median of the miles walked by the seniors. Is your estimate of the mean number of miles less than, approximately equal to, or greater than the median number of miles walked by a senior? If they are different, explain why. If they are approximately the same, explain why.

15. How would you describe the typical number of miles walked by a senior in this walkathon?

16. A junior from River City High School indicated that the number of miles walked by a typical junior was better than the number of miles walked by a typical senior. Do you agree? Explain your answer.

Finally, the twenty-five sophomores who participated in the walkathon reported their results. A dot plot is shown below.



17. What is different about the sophomore data distribution compared to the data distributions for juniors and seniors?

18. Estimate the balance point of the sophomore data distribution.

19. What is the median number of miles walked by a sophomore?

20. How would you describe the sophomore data distribution?

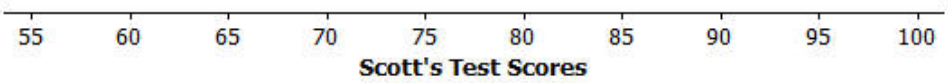
Lesson Summary

The mean of a data distribution represents a balance point for the distribution. The sum of the distances to the right of the mean is equal to the sum of the distances to the left of the mean.

Problem Set

Consider another example of balance. Mr. Jackson is a mathematics teacher at Waldo High School. Students in his class are frequently given quizzes or exams. He indicated to his students that an exam is worth 4 quizzes when calculating an overall weighted average to determine their final grade. During one grading period, Scott got an 80% on one exam, a 90% on a second exam, a 60% on one quiz, and a 70% on another quiz.

How could we represent Scott's test scores? Consider the following number line.



Scott's Test Scores

1. What values are represented by the number line?

2. If one "•" symbol is used to represent a quiz score, how might you represent an exam score?

3. Represent Scott's exams and quizzes on this number line using "•" symbols.

4. Mr. Jackson indicated that students should set an 85% overall weighted average as a goal. Do you think Scott met that goal? Explain your answer.

5. Place an X on the number line at a position that you think locates the balance point of all of the "•" symbols. Determine the sum of the distances from the X to each "•" on the right side of the X.

6. Determine the sum of the distances from the X to each "•" on the left side of the X.

7. Do the total distances to the right of the X equal the total distances to the left of the X?

8. Based on your answer to Problem 7, would you change your estimate of the balance point? If yes, where would you place your adjusted balance point? How does using this adjusted estimate change the total distances to the right of your estimate and the total distances to the left?

9. Scott's weighted average is 81. Recall that each exam score is equal to 4 times a quiz score. Show the calculations that lead to this weighted average.

10. How does the calculated mean score compare with your estimated balance point?

11. Compute the total distances to the right of the mean and the total distances to the left of the mean. What do you observe?

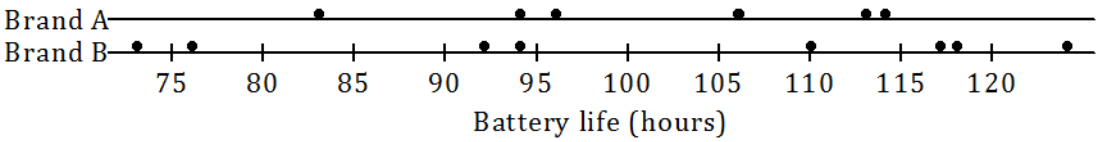
12. Did Scott achieve the goal set by Mr. Jackson of an 85% average? Explain your answer.

Lesson 4: Summarizing Deviations from the Mean

Classwork

Exercises 1–4

A consumers' organization is planning a study of the various brands of batteries that are available. As part of its planning, it measures lifetime (i.e., how long a battery can be used before it must be replaced) for each of six batteries of Brand A and eight batteries of Brand B. Dot plots showing the battery lives for each brand are shown below.



1. Does one brand of battery tend to last longer, or are they roughly the same? What calculations could you do in order to compare the battery lives of the two brands?

2. Do the battery lives tend to differ more from battery to battery for Brand A or for Brand B?

3. Would you prefer a battery brand that has battery lives that do not vary much from battery to battery? Why or why not?

The table below shows the lives (in hours) of the Brand A batteries.

Life (Hours)	83	94	96	106	113	114
Deviation from the Mean						

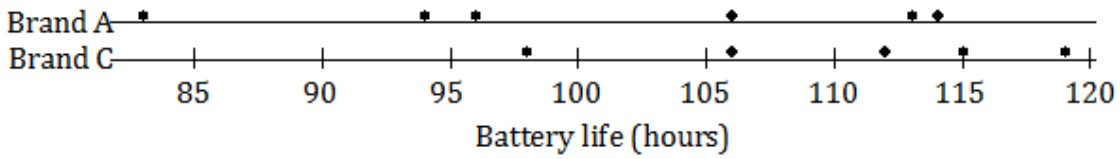
4. Calculate the deviations from the mean for the remaining values, and write your answers in the appropriate places in the table.

The table below shows the battery lives and the deviations from the mean for Brand B.

Life (Hours)	73	76	92	94	110	117	118	124
Deviation from the Mean	-27.5	-24.5	-8.5	-6.5	9.5	16.5	17.5	23.5

Exercises 5–10

The lives of five batteries of a third brand, Brand C, were determined. The dot plot below shows the lives of the Brand A and Brand C batteries.



5. Which brand has the greater mean battery life? (You should be able to answer this question without doing any calculations.)

6. Which brand shows greater variability?

7. Which brand would you expect to have the greater deviations from the mean (ignoring the signs of the deviations)?

The table below shows the lives for the Brand C batteries.

Life (Hours)	115	119	112	98	106
Deviation from the Mean					

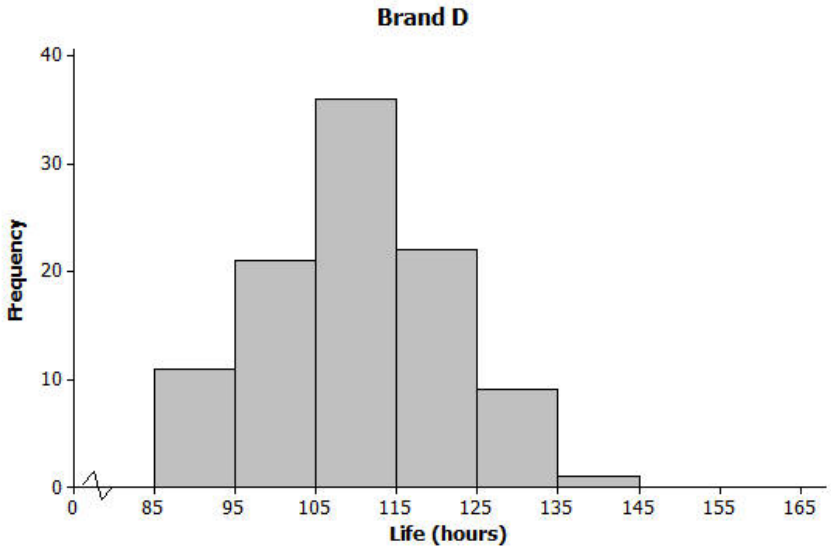
8. Calculate the mean battery life for Brand C. (Be sure to include a unit in your answer.)

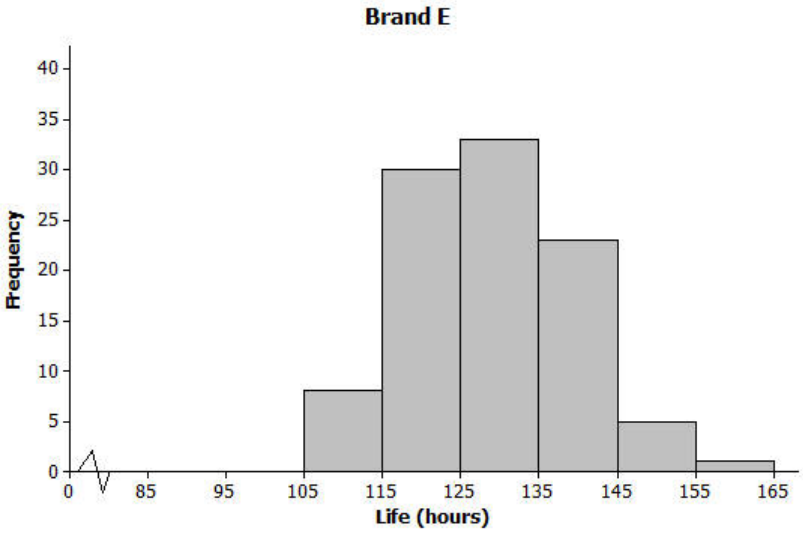
9. Write the deviations from the mean in the empty cells of the table for Brand C.

10. Ignoring the signs, are the deviations from the mean generally larger for Brand A or for Brand C? Does your answer agree with your answer to Exercise 7?

Exercises 11–15

The lives of 100 batteries of Brand D and 100 batteries of Brand E were determined. The results are summarized in the histograms below.





11. Estimate the mean battery life for Brand D. (Do not do any calculations.)

12. Estimate the mean battery life for Brand E. (Do not do any calculations.)

13. Which of Brands D and E shows the greater variability in battery lives? Do you think the two brands are roughly the same in this regard?

14. Estimate the largest deviation from the mean for Brand D.

15. What would you consider a typical deviation from the mean for Brand D?

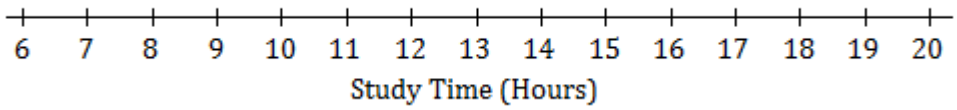
Lesson Summary

- For any given value in a data set, the deviation from the mean is the value minus the mean. Written algebraically, this is $x - \bar{x}$.
- The greater the variability (spread) of the distribution, the greater the deviations from the mean (ignoring the signs of the deviations).

Problem Set

1. Ten members of a high school girls' basketball team were asked how many hours they studied in a typical week. Their responses (in hours) were 20, 13, 10, 6, 13, 10, 13, 11, 11, 10.

   a. Using the axis given below, draw a dot plot of these values. (Remember, when there are repeated values, stack the dots with one above the other.)

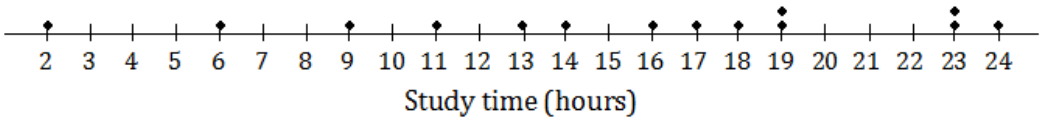


   b. Calculate the mean study time for these students.

   c. Calculate the deviations from the mean for these study times, and write your answers in the appropriate places in the table below.

Number of Hours Studied	20	13	10	6	13	10	13	11	11	10
Deviation from the Mean										

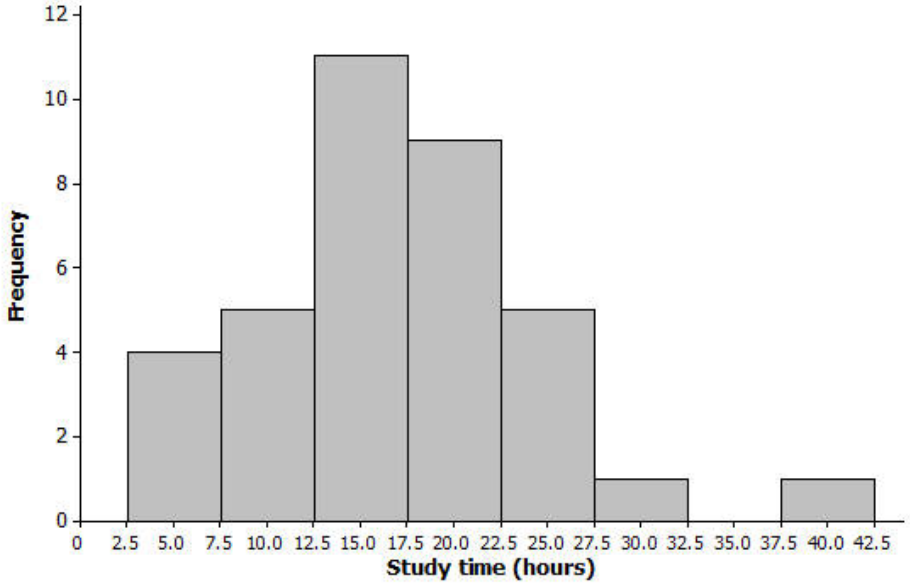
   d. The study times for fourteen girls from the soccer team at the same school as the one above are shown in the dot plot below.



   Based on the data, would the deviations from the mean (ignoring the sign of the deviations) be greater or less for the soccer players than for the basketball players?

2. All the members of a high school softball team were asked how many hours they studied in a typical week. The results are shown in the histogram below.

   (The data set in this question comes from Core Math Tools, www.nctm.org.)



   a. We can see from the histogram that four students studied around 5 hours per week. How many students studied around 15 hours per week?

   b. How many students were there in total?

   c. Suppose that the four students represented by the histogram bar centered at 5 had all studied exactly 5 hours, the five students represented by the next histogram bar had all studied exactly 10 hours, and so on. If you were to add up the study times for all of the students, what result would you get?

   d. What is the mean study time for these students?

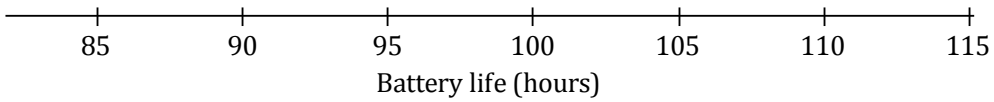
   e. What would you consider to be a typical deviation from the mean for this data set?

Lesson 5: Measuring Variability for Symmetrical Distributions

Classwork

Example 1: Calculating the Standard Deviation

Here is a dot plot of the lives of the Brand A batteries from Lesson 4.



How do you measure variability of this data set? One way is by calculating **standard deviation**.

- First, find each deviation from the mean.
- Then, square the deviations from the mean. For example, when the deviation from the mean is -18 the squared deviation from the mean is $(-18)^2 = 324$.

Life (Hours)	83	94	96	106	113	114
Deviation from the Mean	-18	-7	-5	5	12	13
Squared Deviations from the Mean	324	49	25	25	144	169

- Add up the squared deviations:

  $$324 + 49 + 25 + 25 + 144 + 169 = 736.$$

  This result is the *sum* of the squared deviations.

The number of values in the data set is denoted by n. In this example, n is 6.

- You divide the sum of the squared deviations by $n - 1$, which here is $6 - 1 = 5$.

$$\frac{736}{5} = 147.2$$

- Finally, you take the square root of 147.2, which to the nearest hundredth is 12.13.

That is the standard deviation! It seems like a very complicated process at first, but you will soon get used to it.

We conclude that a typical deviation of a Brand A battery lifetime from the mean battery lifetime for Brand A is 12.13 hours. The unit of standard deviation is always the same as the unit of the original data set. So, the standard deviation to the nearest hundredth, with the unit, is 12.13 hours. How close is the answer to the typical deviation that you estimated at the beginning of the lesson?

Exercises 1–5

Now you can calculate the standard deviation of the lifetimes for the eight Brand B batteries. The mean was 100.5. We already have the deviations from the mean.

Life (Hours)	73	76	92	94	110	117	118	124
Deviation from the Mean	-27.5	-24.5	-8.5	-6.5	9.5	16.5	17.5	23.5
Squared Deviation from the Mean								

1. Write the squared deviations in the table.

2. Add up the squared deviations. What result do you get?

3. What is the value of n for this data set? Divide the sum of the squared deviations by $n - 1$, and write your answer below. Round your answer to the nearest thousandth.

4. Take the square root to find the standard deviation. Record your answer to the nearest hundredth.

5. How would you interpret the standard deviation that you found in Exercise 4? (Remember to give your answer in the context of this question. Interpret your answer to the nearest hundredth.)

Exercises 6–7

Jenna has bought a new hybrid car. Each week for a period of seven weeks, she has noted the fuel efficiency (in miles per gallon) of her car. The results are shown below.

45 44 43 44 45 44 43

6. Calculate the standard deviation of these results to the nearest hundredth. Be sure to show your work.

7. What is the meaning of the standard deviation you found in Exercise 6?

Lesson Summary

- The standard deviation measures a typical deviation from the mean.
- To calculate the standard deviation,
  1. Find the mean of the data set;
  2. Calculate the deviations from the mean;
  3. Square the deviations from the mean;
  4. Add up the squared deviations;
  5. Divide by $n-1$ (if you are working with a data from a sample, which is the most common case);
  6. Take the square root.
- The unit of the standard deviation is always the same as the unit of the original data set.
- The larger the standard deviation, the greater the spread (variability) of the data set.

Problem Set

1. A small car dealership tests the fuel efficiency of sedans on its lot. It chooses 12 sedans for the test. The fuel efficiency (mpg) values of the cars are given in the table below. Complete the table as directed below.

Fuel Efficiency (miles per gallon)	29	35	24	25	21	21	18	28	31	26	26	22
Deviation from the Mean												
Squared Deviation from the Mean												

   a. Calculate the mean fuel efficiency for these cars. Calculate the mean fuel efficiency for these cars.
   b. Calculate the deviations from the mean, and write your answers in the second row of the table.
   c. Square the deviations from the mean, and write the squared deviations in the third row of the table.
   d. Find the sum of the squared deviations.
   e. What is the value of n for this data set? Divide the sum of the squared deviations by $n-1$.
   f. Take the square root of your answer to (e) to find the standard deviation of the fuel efficiencies of these cars. Round your answer to the nearest hundredth.

2. The same dealership decides to test fuel efficiency of SUVs. It selects six SUVs on its lot for the test. The fuel efficiencies (in miles per gallon) of these cars are shown below.

   21 21 21 30 28 24

   Calculate the mean and the standard deviation of these values. Be sure to show your work, and include a unit in your answer.

3. Consider the following questions regarding the cars described in Problems 1 and 2.

   a. What is the standard deviation of the fuel efficiencies of the cars in Problem 1? Explain what this value tells you.

   b. You also calculated the standard deviation of the fuel efficiencies for the cars in Problem 2. Which of the two data sets (Problem 1 or Problem 2) has the larger standard deviation? What does this tell you about the two types of cars (sedans and SUVs)?

Lesson 6: Interpreting the Standard Deviation

Classwork

Example 1

Your teacher will show you how to use a calculator to find the mean and standard deviation for the following set of data.

A set of eight men have heights (in inches) as shown below.

67.0 70.9 67.6 69.8 69.7 70.9 68.7 67.2

Indicate the mean and standard deviation you obtained from your calculator to the nearest hundredth.

Mean: _____

Standard Deviation: _____

Exercise 1

1. The heights (in inches) of nine women are as shown below.

   68.4 70.9 67.4 67.7 67.1 69.2 66.0 70.3 67.6

   Use the statistical features of your calculator or computer software to find the mean and the standard deviation of these heights to the nearest hundredth.

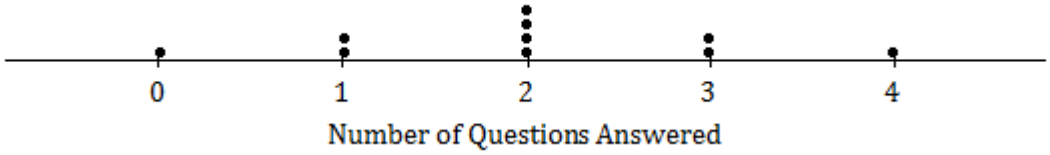
   Mean: _____

   Standard Deviation: _____

Exploratory Challenge/Exercises 2–5

2. A group of people attended a talk at a conference. At the end of the talk, ten of the attendees were given a questionnaire that consisted of four questions. The questions were optional, so it was possible that some attendees might answer none of the questions, while others might answer 1, 2, 3, or all 4 of the questions (so, the possible numbers of questions answered are 0, 1, 2, 3, and 4).

   Suppose that the numbers of questions answered by each of the ten people were as shown in the dot plot below.

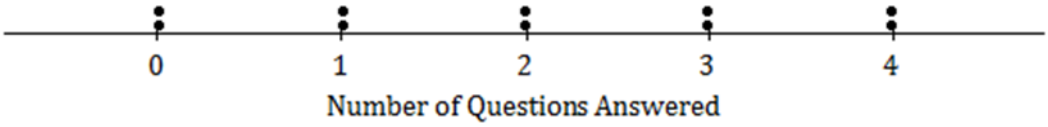


   Use the statistical features of your calculator to find the mean and the standard deviation of the data set.

   Mean: _____

   Standard Deviation: _____

3. Suppose the dot plot looked like this:

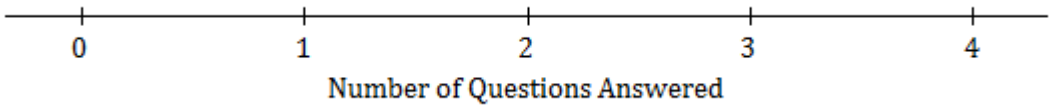


   a. Use your calculator to find the mean and the standard deviation of this distribution.

   b. Remember that the size of the standard deviation is related to the size of the deviations from the mean. Explain why the standard deviation of this distribution is greater than the standard deviation in Exercise 2.

4. Suppose that all ten people questioned answered all four questions on the questionnaire.

   a. What would the dot plot look like?

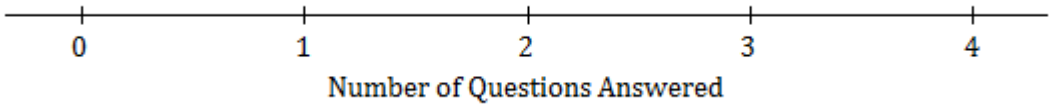


   b. What is the mean number of questions answered? (You should be able to answer without doing any calculations!)

   c. What is the standard deviation? (Again, don't do any calculations!)

5. Continue to think about the situation previously described where the numbers of questions answered by each of ten people was recorded.

   a. Draw the dot plot of the distribution of possible data values that has the largest possible standard deviation. (There were ten people at the talk, so there should be ten dots in your dot plot.) Use the scale given below.



   b. Explain why the distribution you have drawn has a larger standard deviation than the distribution in Exercise 4.

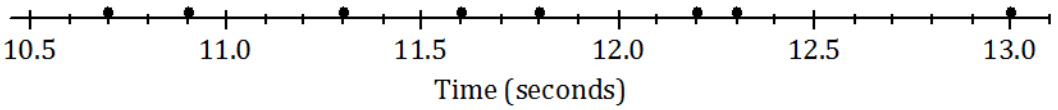
Lesson Summary

- The mean and the standard deviation of a data set can be found directly using the statistical features of a calculator.
- The size of the standard deviation is related to the sizes of the deviations from the mean. Therefore, the standard deviation is minimized when all the numbers in the data set are the same and is maximized when the deviations from the mean are made as large as possible.

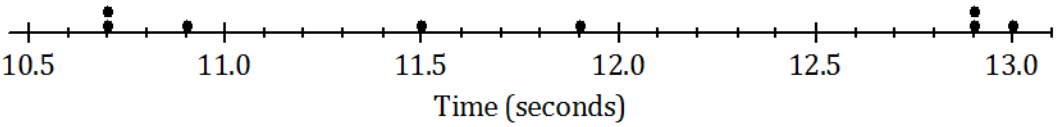
Problem Set

1. At a track meet, there are three men's 100 m races. The times for eight of the sprinters are recorded to the nearest $\frac{1}{10}$ of a second. The results of the three races for these eight sprinters are shown in the dot plots below.

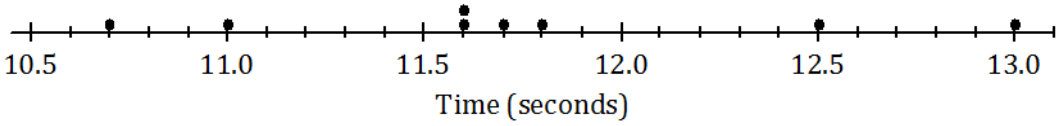
Race 1



Race 2



Race 3



a. Remember that the size of the standard deviation is related to the sizes of the deviations from the mean. Without doing any calculations, indicate which of the three races has the smallest standard deviation of times. Justify your answer.

b. Which race had the largest standard deviation of times? (Again, don't do any calculations!) Justify your answer.

c. Roughly what would be the standard deviation in Race 1? (Remember that the standard deviation is a typical deviation from the mean. So, here you are looking for a typical deviation from the mean, in seconds, for Race 1.)

d. Use your calculator to find the mean and the standard deviation for each of the three races. Write your answers in the table below to the nearest thousandth.

	Mean	Standard Deviation
Race 1		
Race 2		
Race 3		

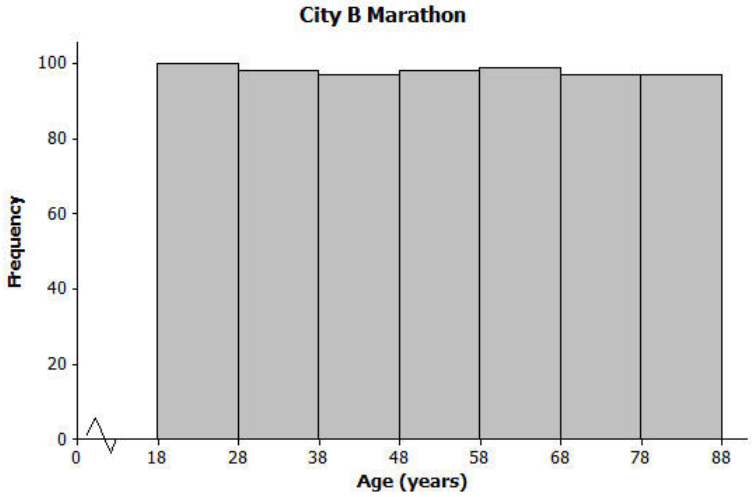
e. How close were your answers (a)–(c) to the actual values?

2. A large city, which we will call City A, holds a marathon. Suppose that the ages of the participants in the marathon that took place in City A were summarized in the histogram below.

a. Make an estimate of the mean age of the participants in the City A marathon.

b. Make an *estimate* of the standard deviation of the ages of the participants in the City A marathon.

A smaller city, City B, also held a marathon. However, City B restricts the number of people of each age category who can take part to 100. The ages of the participants for one race are summarized in the histogram below. The ages of the participants are summarized in the histogram below.



c. Approximately what was the mean age of the participants in the City B marathon? Approximately what was the standard deviation of the ages?

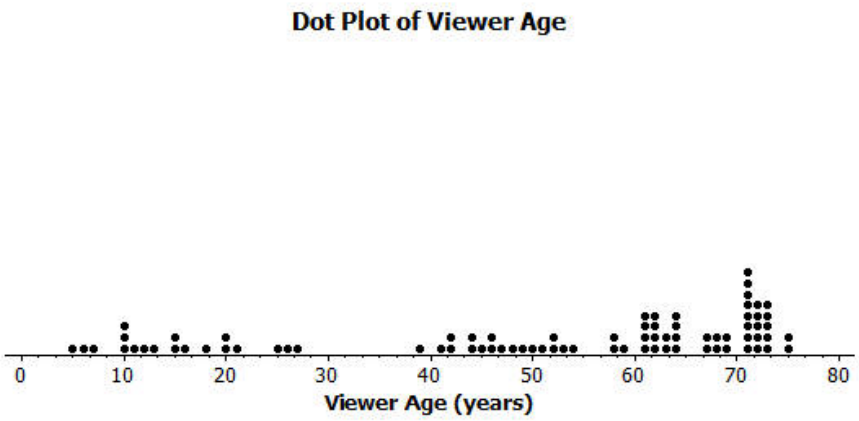
d. Explain why the standard deviation of the ages in the City B marathon is greater than the standard deviation of the ages for the City A marathon.

Lesson 7: Measuring Variability for Skewed Distributions (Interquartile Range)

Classwork

Exploratory Challenge 1/Exercises 1–3: Skewed Data and its Measure of Center

Consider the following scenario. A television game show, *Fact or Fiction*, was canceled after nine shows. Many people watched the nine shows and were rather upset when it was taken off the air. A random sample of eighty viewers of the show was selected. Viewers in the sample responded to several questions. The dot plot below shows the distribution of ages of these eighty viewers.



1. Approximately where would you locate the mean (balance point) in the above distribution?

2. How does the direction of the tail affect the location of the mean age compared to the median age?

3. The mean age of the above sample is approximately 50. Do you think this age describes the typical viewer of this show? Explain your answer.

Exploratory Challenge 2/Exercises 4–8: Constructing and Interpreting the Box Plot

4. Using the above dot plot, construct a box plot over the dot plot by completing the following steps:
   i. Locate the middle 40 observations, and draw a box around these values.
   ii. Calculate the median, and then draw a line in the box at the location of the median.
   iii. Draw a line that extends from the upper end of the box to the largest observation in the data set.
   iv. Draw a line that extends from the lower edge of the box to the minimum value in the data set.

5. Recall that the 5 values used to construct the dot plot make up the 5-number summary. What is the 5-number summary for this data set of ages?

   Minimum age: _____

   Lower quartile or Q1: _____

   Median Age: _____

   Upper quartile or Q3: _____

   Maximum age: _____

6. What percent of the data does the box part of the box plot capture?

7. What percent of the data falls between the minimum value and Q1?

8. What percent of the data falls between Q3 and the maximum value?

Exercises 9–14

An advertising agency researched the ages of viewers most interested in various types of television ads. Consider the following summaries:

Ages	Target Products or Services
30–45	Electronics, home goods, cars
46–55	Financial services, appliances, furniture
56–72	Retirement planning, cruises, health care services

9. The mean age of the people surveyed is approximately 50 years old. As a result, the producers of the show decided to obtain advertisers for a typical viewer of 50 years old. According to the table, what products or services do you think the producers will target? Based on the sample, what percent of the people surveyed about the *Fact or Fiction* show would have been interested in these commercials if the advertising table is accurate?

10. The show failed to generate the interest the advertisers hoped. As a result, they stopped advertising on the show, and the show was cancelled. Kristin made the argument that a better age to describe the typical viewer is the median age. What is the median age of the sample? What products or services does the advertising table suggest for viewers if the median age is considered as a description of the typical viewer?

11. What percent of the people surveyed would be interested in the products or services suggested by the advertising table if the median age were used to describe a typical viewer?

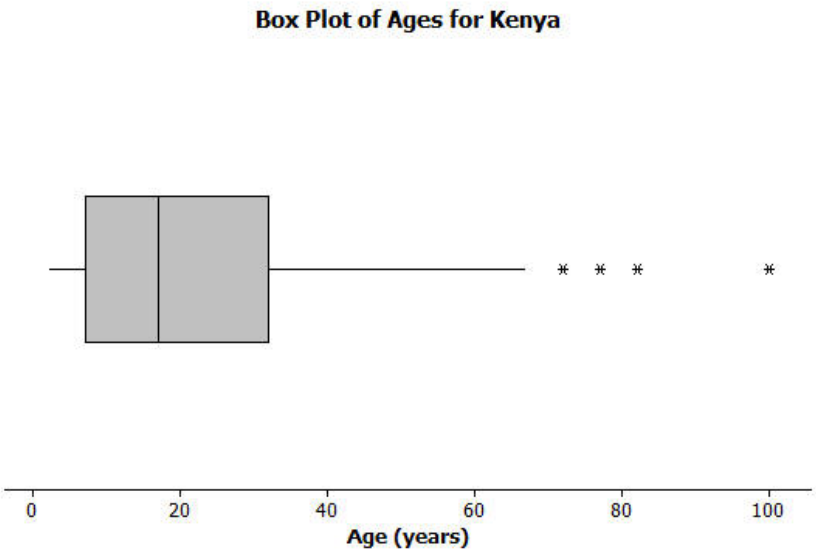
12. What percent of the viewers have ages between Q1 and Q3? The difference between Q3 and Q1, or Q3 – Q1, is called the **interquartile range**, or **IQR**. What is the interquartile range (IQR) for this data distribution?

13. The IQR provides a summary of the variability for a skewed data distribution. The IQR is a number that specifies the length of the interval that contains the middle half of the ages of viewers. Do you think producers of the show would prefer a show that has a small or large interquartile range? Explain your answer.

14. Do you agree with Kristin's argument that the median age provides a better description of a typical viewer? Explain your answer.

Exploratory Challenge 3/Exercises 15–20: Outliers

Students at Waldo High School are involved in a special project that involves communicating with people in Kenya. Consider a box plot of the ages of 200 randomly selected people from Kenya.



A data distribution may contain extreme data (specific data values that are unusually large or unusually small relative to the median and the interquartile range). A box plot can be used to display extreme data values that are identified as **outliers**.

Each "*" in the box plot represents the ages of four people from this sample. Based on the sample, these four ages were considered outliers.

15. Estimate the values of the four ages represented by an *.

An outlier is defined to be any data value that is more than $1.5 \times (IQR)$ away from the nearest quartile.

16. What is the median age of the sample of ages from Kenya? What are the approximate values of Q1 and Q3? What is the approximate IQR of this sample?

17. Multiply the IQR by 1.5. What value do you get?

18. Add $1.5 \times (IQR)$ to the 3rd quartile age (Q3). What do you notice about the four ages identified by an *?

19. Are there any age values that are less than $Q1 - 1.5 \times (IQR)$? If so, these ages would also be considered outliers.

20. Explain why there is no * on the low side of the box plot for ages of the people in the sample from Kenya.

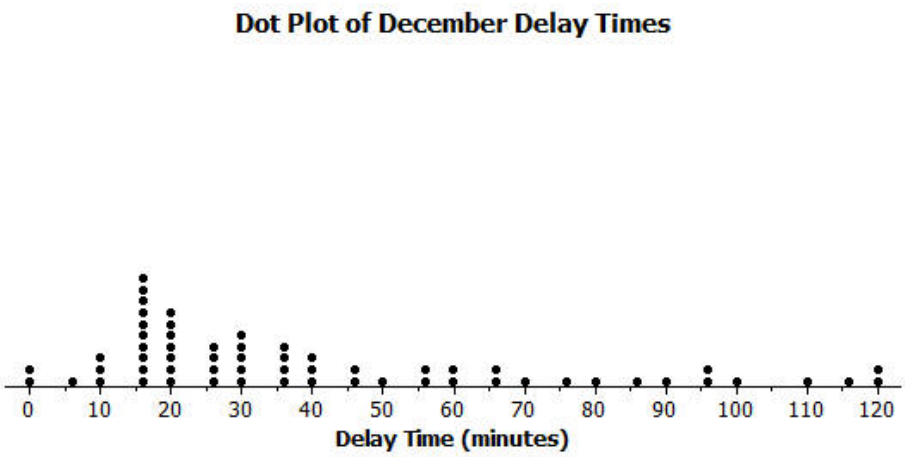
Lesson Summary

- Non-symmetrical data distributions are referred to as skewed.
- Left-skewed or skewed to the left means the data spreads out longer (like a tail) on the left side.
- Right-skewed or skewed to the right means the data spreads out longer (like a tail) on the right side.
- The center of a skewed data distribution is described by the median.
- Variability of a skewed data distribution is described by the interquartile range (IQR).
- The IQR describes variability by specifying the length of the interval that contains the middle 50% of the data values.
- Outliers in a data set are defined as those values more than 1.5(IQR) from the nearest quartile. Outliers are usually identified by an "*" or a "•" in a box plot.

Problem Set

Consider the following scenario. Transportation officials collect data on flight delays (the number of minutes a flight takes off after its scheduled time).

Consider the dot plot of the delay times in minutes for 60 BigAir flights during December 2012:



1. How many flights left more than 60 minutes late?

2. Why is this data distribution considered skewed?

3. Is the tail of this data distribution to the right or to the left? How would you describe several of the delay times in the tail?

4. Draw a box plot over the dot plot of the flights for December.

5. What is the interquartile range, or IQR, of this data set?

6. The mean of the 60 flight delays is approximately 42 minutes. Do you think that 42 minutes is typical of the number of minutes a BigAir flight was delayed? Why or why not?

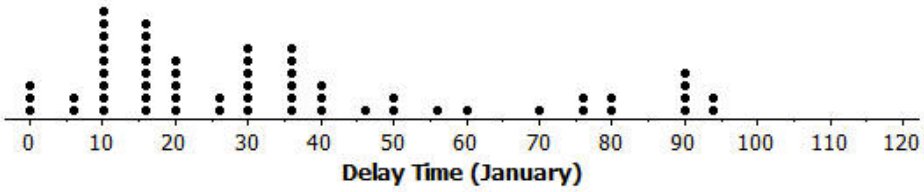
7. Based on the December data, write a brief description of the BigAir flight distribution for December.

8. Calculate the percentage of flights with delays of more than 1 hour. Were there many flight delays of more than 1 hour?

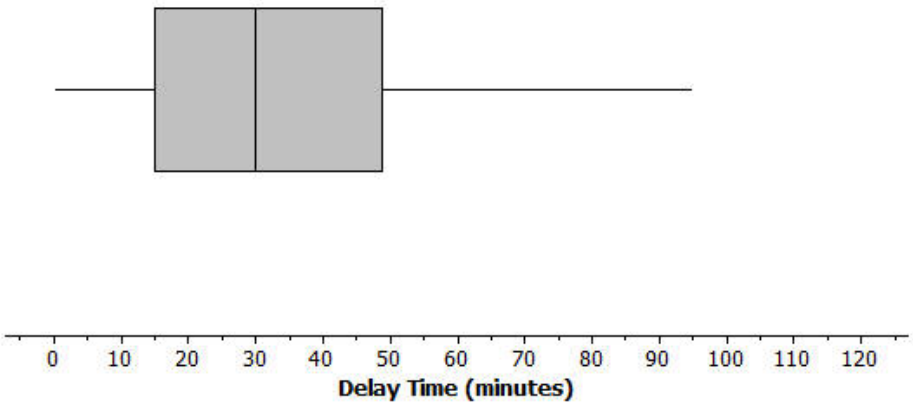
9. BigAir later indicated that there was a flight delay that was not included in the data. The flight not reported was delayed for 48 hours. If you had included that flight delay in the box plot, how would you have represented it? Explain your answer.

10. Consider a dot plot and the box plot of the delay times in minutes for 60 BigAir flights during January 2013.

    How is the January flight delay distribution different from the one summarizing the December flight delays? In terms of flight delays in January, did BigAir improve, stay the same, or do worse compared to December? Explain your answer.



Box Plot of January Delay Times



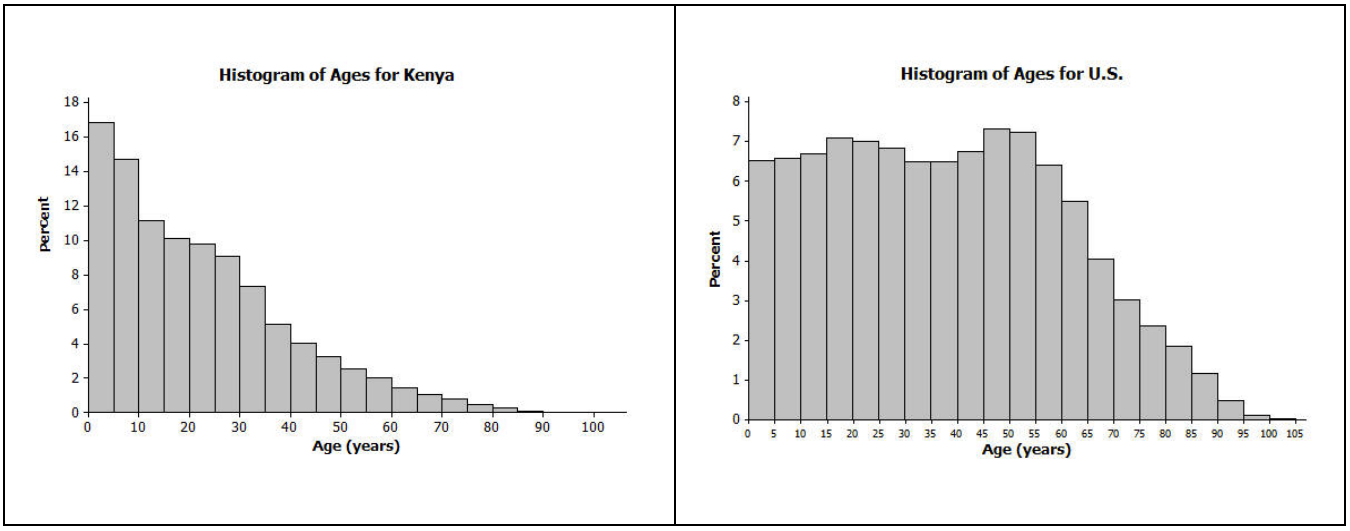
Lesson 8: Comparing Distributions

Classwork

Exploratory Challenge 1: Country Data

A science museum has a "Traveling Around the World" exhibit. Using 3D technology, participants can make a virtual tour of cities and towns around the world. Students at Waldo High School registered with the museum to participate in a virtual tour of Kenya, visiting the capital city of Nairobi and several small towns. Before they take the tour, however, their mathematics class decided to study Kenya using demographic data from 2010 provided by the United States Census Bureau. They also obtained data for the United States from 2010 to compare to data for Kenya.

The following histograms represent the age distributions of the two countries.



Exercises 1–8

1. How do the shapes of the two histograms differ?

2. Approximately what percent of people in Kenya are between the ages of 0 and 10 years?

3. Approximately what percent of people in the United States are between the ages of 0 and 10 years?

4. Approximately what percent of people in Kenya are 60 years or older?

5. Approximately what percent of people in the United States are 60 years or older?

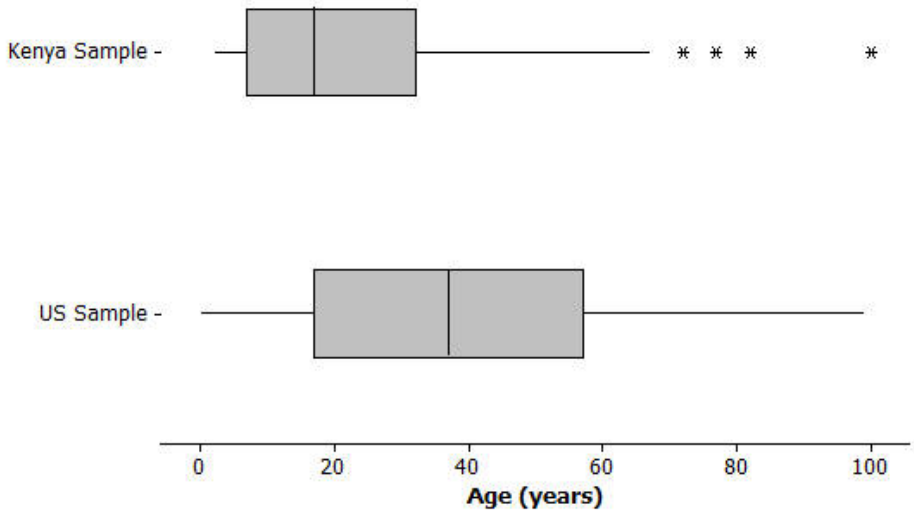
6. The population of Kenya in 2010 was approximately 41 million people. What is the approximate number of people in Kenya between the ages of 0 and 10 years?

7. The population of the United States in 2010 was approximately 309 million people. What is the approximate number of people in the United States between the ages of 0 and 10 years?

8. The Waldo High School students started planning for their virtual visit of the neighborhoods in Nairobi and several towns in Kenya. Do you think they will see many teenagers? Will they see many senior citizens who are 70 or older? Explain your answer based on the histogram.

Exploratory Challenge 2: Learning More about the Countries using Box Plots and Histograms

A random sample of 200 people from Kenya in 2010 was discussed in previous lessons. A random sample of 200 people from the United States is also available for study. Box plots constructed using the ages of the people in these two samples are shown below.



Exercises 9–16

9. Adrian, a senior at Waldo High School, stated that the box plots indicate that the United States has a lot of older people compared to Kenya. Would you agree? How would you describe the difference in the ages of people in these two countries based on the above box plots?

10. Estimate the median age of a person in Kenya and the median age of a person in the United States using the box plots.

11. Using the box plot, 25% of the people in the United States are younger than what age? How did you determine that age?

12. Using the box plots, approximately what percent of people in Kenya are younger than 18 years old?

13. Could you have estimated the mean age of a person from Kenya using the box plot? Explain your answer.

14. The mean age of people the United States is approximately 38 years. Using the histogram, estimate the percentage of people in the United States who are younger than the mean age in the United States.

15. If the median age is used to describe a "typical" person in Kenya, what percent of people in Kenya are younger than the median age? Is the mean or median age a better description of a "typical" person in Kenya? Explain your answer.

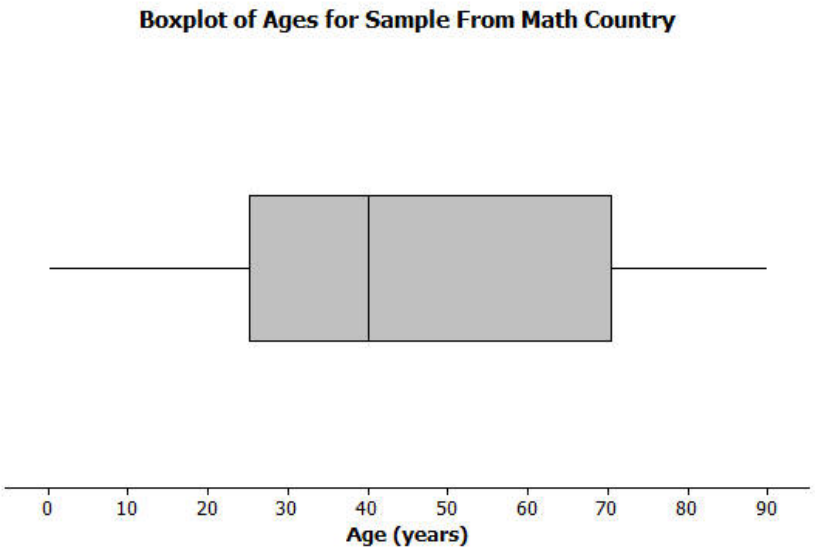
16. What is the IQR of the ages in the sample from the United States? What is the IQR of the ages in the sample from Kenya? If the IQRs are used to compare countries, what does a smaller IQR indicate about a country? Use Kenya and the United States to explain your answer.

Lesson Summary

- Histograms show the general shape of a distribution.
- Box plots are created from the 5-number summary of a data set.
- A box plot identifies the median, minimum, and maximum values, and the upper and lower quartiles.
- The interquartile range (IQR) describes how the data is spread around the median; it is the length of the interval that contains 50% of the data values.
- The median is used as a measure of the center when a distribution is skewed or contains outliers.

Problem Set

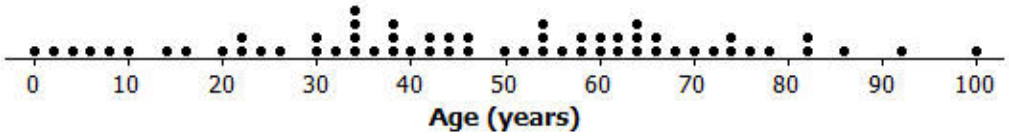
The following box plot summarizes ages for a random sample from a made up country named Math Country.



1. Make up your own sample of forty ages that could be represented by the box plot for Math Country. Use a dot plot to represent the ages of the forty people in Math Country.

2. Is the sample of forty ages represented in your dot plot of Math Country the only sample that could be represented by the box plot? Explain your answer.

3. The following is a dot plot of sixty ages from a random sample of people from Japan in 2010.
   Draw a box plot over this dot plot.



4. Based on your box plot, would the median age of people in Japan be closer to the median age of people in Kenya or the United States? Justify your answer.

5. What does the box plot of this sample from Japan indicate about the possible differences in the age distributions of people from Japan and Kenya?

Lesson 9: Summarizing Bivariate Categorical Data

Classwork

Recall from your work in Grade 6 and Grade 8 that categorical data are data that are not numbers. **Bivariate categorical data** results from collecting data on two categorical variables. In this lesson, you will see examples involving categorical data collected from two survey questions.

Exploratory Challenge 1: Superhero Powers

Superheroes have been popular characters in movies, television, books, and comics for many generations. Superman was one of the most popular series in the 1950s while Batman was a top rated series in the 1960s. Each of these characters was also popular in movies released from 1990 to 2013. Other notable characters portrayed in movies over the last several decades include Captain America, She-Ra, and the Fantastic Four. What is special about a superhero? Is there a special superhero power that makes these characters particularly popular?

High school students in the United States were invited to complete an online survey in 2010. Part of the survey included questions about superhero powers. More than $1{,}000$ students responded to this survey that included a question about a favorite superhero power. 450 of the completed surveys were randomly selected. A rather confusing breakdown of the data by gender was compiled from the 450 surveys:

- 100 students indicated their favorite power was "to fly." 49 of those students were females.
- 131 students selected the power to "freeze time" as their favorite power. 71 of those students were males.
- 75 students selected "invisibility" as their favorite power. 48 of those students were females.
- 26 students indicated "super strength" as their favorite power. 25 of those students were males.
- And finally, 118 students indicated "telepathy" as their favorite power. 70 of those students were females.

Exercises 1–4

Several superheroes portrayed in movies and television series had at least one extraordinary power. Some superheroes had more than one special power. Was Superman's power "to fly" the favorite power of his fans, or was it his "super strength"? Would females view the power "to fly" differently than males, or in the same way? Use the survey information given in Example 1 to answer the following questions.

1. How many more females than males indicated their favorite power is "telepathy"?

2. How many more males than females indicated their favorite power was "to fly"?

3. Write survey questions that you think might have been used to collect this data.

4. How do you think the 450 surveys used in Example 1 might have been selected? You can assume that there were $1{,}000$ surveys to select from.

Exploratory Challenge 2: A Statistical Study Involving a Two-Way Frequency Table

The data in Example 1 prompted students in a mathematics class to pose the statistical question, "Do high school males have different preferences for superhero powers than high school females?" Answering this statistical question involves collecting data as well as anticipating variability in the data collected.

The data consist of two responses from each student completing a survey. The first response indicates a student's gender, and the second response indicates the student's favorite superpower. For example, data collected from one student was "male" and "to fly." The data are bivariate categorical data.

The first step in analyzing the statistical question posed by the students in their mathematics class is to organize this data in a two-way frequency table.

A two-way frequency table that can be used to organize the categorical data is shown below. The letters below represent the frequency counts of the cells of the table.

	To Fly	**Freeze time**	**Invisibility**	**Super Strength**	**Telepathy**	**Total**
Females	(a)	(b)	(c)	(d)	(e)	(f)
Males	(g)	(h)	(i)	(j)	(k)	(l)
Total	(m)	(n)	(o)	(p)	(q)	(r)

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

Exercises 5–12

5. Describe the data that would be counted in cell (a).

6. Describe the data that would be counted in cell (j).

7. Describe the data that would be counted in cell (l).

8. Describe the data that would be counted in cell (n).

9. Describe the data that would be counted in cell (r).

10. Cell (i) is the number of male students who selected "invisibility" as their favorite superpower. Using the information given in Example 1, what is the value of this number?

11. Cell (d) is the number of females whose favorite superpower is "super strength." Using the information given in Example 1, what is the value of this number?

12. Complete the table below by determining a frequency count for each cell based on the summarized data.

	To Fly	Freeze Time	Invisibility	Super Strength	Telepathy	Total
Females						
Males						
Total						

Lesson Summary

- *Categorical data* are data that take on values that are categories rather than numbers. Examples include male or female for the categorical variable of gender or the five superpower categories for the categorical variable of superpower qualities.
- A *two-way frequency table* is used to summarize bivariate categorical data.
- The number in a two-way frequency table at the intersection of a row and column of the response to two categorical variables represents a *joint frequency*.
- The total number of responses for each value of a categorical variable in the table represents the *marginal frequency* for that value.

Problem Set

Several students at Rufus King High School were debating whether males or females were more involved in after-school activities. There are three organized activities in the after-school program—intramural basketball, chess club, and jazz band. Due to budget constraints, a student can only select one of these activities. The students were not able to ask every student in the school whether they participated in the after-school program or what activity they selected if they were involved.

1. Write questions that could be included in the survey to investigate the question the students are debating. Questions that could be used for this study include the following:

2. Rufus King High School has approximately $1{,}500$ students. Sam suggested that the first 100 students entering the cafeteria for lunch would provide a random sample to analyze. Janet suggested that they pick 100 students based on a school identification number. Who has a better strategy for selecting a random sample? How do you think 100 students could be randomly selected to complete the survey?

3. Consider the following results from 100 randomly selected students:
   - Of the 60 female students selected, 20 of them played intramural basketball, 10 played chess, and 10 were in the jazz bland. The rest of them did not participate in the after-school program.
   - Of the male students, 10 did not participate in the after-school program, 20 played intramural basketball, 8 played in the jazz band, and the rest played chess.

   A two-way frequency table to summarize the survey data was started. What label is needed in the table cell identified with a "???."

	Intramural Basketball	Chess Club	Jazz Band	???	Total
Female					
Male					
Total					

4. Complete the above table for the 100 students who were surveyed.

5. The table shows the responses to the after-school activity question for males and females. Do you think there is a difference in the responses of males and females? Explain your answer.

Lesson 10: Summarizing Bivariate Categorical Data with Relative Frequencies

Classwork

This lesson expands on your work with two-way frequency tables from Lesson 9.

Exploratory Challenge 1: Extending the Frequency Table to a Relative Frequency Table

Determining the number of students in each cell presents the first step in organizing bivariate categorical data. Another way of analyzing the data in the table is to calculate the *relative frequency* for each cell. Relative frequencies relate each frequency count to the total number of observations. For each cell in this table, the *relative frequency* of a cell is found by dividing the frequency of that cell by the total number of responses.

Consider the two-way frequency table from the previous lesson.

Two-Way Frequency Table:

	To Fly	Freeze Time	Invisibility	Super Strength	Telepathy	Total
Females	49	60	48	1	70	228
Males	51	71	27	25	48	222
Total	100	131	75	26	118	450

The relative frequency table would be found by dividing each of the above cell values by 450. For example, the relative frequency of females selecting "To Fly" is $\frac{49}{450}$, or approximately 0.109, to the nearest thousandth. A few of the other relative frequencies to the nearest thousandth are shown in the following relative frequency table:

	To Fly	Freeze Time	Invisibility	Super Strength	Telepathy	Total
Females	$\frac{49}{450} \approx 0.109$					$\frac{228}{450} \approx 0.507$
Males			$\frac{27}{450} \approx 0.060$			
Total		$\frac{131}{450} \approx 0.291$			$\frac{118}{450} \approx 0.262$	

Exercises 1–7

1. Calculate the remaining relative frequencies in the table below. Write the value in the table as a decimal rounded to the nearest thousandth or as a percent.

   Two-Way Frequency Table:

	To Fly	Freeze Time	Invisibility	Super Strength	Telepathy	Total
Females						
Males						
Total						

2. Based on previous work with frequency tables, which cells in this table would represent the joint relative frequencies?

3. Which cells in the relative frequency table would represent the marginal relative frequencies*?*

4. What is the joint relative frequency for females who selected “invisibility” as their favorite superpower?

5. What is the marginal relative frequency for “freeze time”? Interpret the meaning of this value.

6. What is the difference in the joint relative frequencies for males and for females who selected "to fly" as their favorite superpower?

7. Is there a noticeable difference between the genders and their favorite superpowers?

Exploratory Challenge 2: Interpreting Data

Interest in superheroes continues at Rufus King High School. The students who analyzed the data in the previous lesson decided to create a comic strip for the school website that involves a superhero. They thought the summaries developed from the data would be helpful in designing the comic strip.

Only one power will be given to the superhero. A debate arose as to what power the school's superhero would possess. Students used the two-way frequency table and the relative frequency table to continue the discussion. Take another look at those tables.

Scott initially indicated that the character created should have "super strength" as the special power. This suggestion was not well received by the other students planning this project. In particular, Jill argued, "Well, if you don't want to ignore more than half of the readers, then I suggest 'telepathy' is the better power for our character."

Exercises 8–10

Scott acknowledged that "super strength" was probably not the best choice based on the data. "The data indicate that "freeze time" is the most popular power for a superhero," continued Scott. Jill, however, still did not agree with Scott that this was a good choice. She argued that "telepathy" was a better choice.

8. How do the data support Scott's claim? Why do you think he selected "freeze time" as the special power for the comic strip superhero?

9. How do the data support Jill's claim? Why do you think she selected "telepathy" as the special power for the comic strip superhero?

10. Of the two special powers "freeze time" and "telepathy," select one and justify why you think it is a better choice based on the data.

Lesson Summary

- *Categorical data* are data that take on values that are categories rather than numbers. Examples include male or female for the categorical variable of gender or the five superpower categories for the categorical variable of superpower qualities.
- A *two-way frequency table* is used to summarize bivariate categorical data.
- A *relative frequency* compares a frequency count to the total number of observations. It can be written as a decimal or percent. A two-way table summarizing the relative frequencies of each cell is called a *relative frequency table*.
- The marginal cells in a two-way relative frequency table are called the *marginal relative frequencies*, while the joint cells are called the *joint relative frequencies*.

Problem Set

1. Consider the Rufus King High School data from the previous lesson regarding after-school activities:

	Intramural Basketball	Chess Club	Jazz Band	Not Involved	Total
Males	20	2	8	10	40
Females	20	10	10	20	60
Total	40	12	18	30	100

Calculate the relative frequencies for each of the cells to the nearest thousandth. Place the relative frequencies in the cells of the following table. (The first cell has been completed as an example.)

	Intramural Basketball	Chess Club	Jazz Band	Not Involved	Total
Males	$\frac{20}{100} = 0.200$				
Females					
Total					

2. Based on your relative frequency table, what is the relative frequency of students who indicated they play basketball?

3. Based on your table, what is the relative frequency of males who play basketball?

4. If a student were randomly selected from the students at the school, do you think the student selected would be a male or a female?

5. If a student were selected at random from school, do you think this student would be involved in an after-school program? Explain your answer.

6. Why might someone question whether or not the students who completed the survey were randomly selected? If the students completing the survey were randomly selected, what do the marginal relative frequencies possibly tell you about the school? Explain your answer.

7. Why might females think they are more involved in after-school activities than males? Explain your answer.

Lesson 11: Conditional Relative Frequencies and Association

Classwork

After further discussion, the students involved in designing the superhero comic strip decided that before any decision is made, a more careful look at the data on the special powers a superhero character could possess was needed. There is an association between gender and superpower response if the superpower responses of males are not the same as the superpower responses of females. Examining each row of the table can help determine whether or not there is an association.

Exploratory Challenge 1: Conditional Relative Frequencies

Recall the two-way table from the previous lesson.

	To Fly	Freeze Time	Invisibility	Super Strength	Telepathy	Total
Females	49	60	48	1	70	228
Males	51	71	27	25	48	222
Total	100	131	75	26	118	450

A *conditional relative frequency* compares a frequency count to the marginal total that represents the condition of interest. For example, the condition of interest in the first row is females. The row conditional relative frequency of females responding "invisibility" as the favorite superpower is $\frac{48}{228}$, or approximately 0.211. This conditional relative frequency indicates that approximately 21.1% of females prefer "invisibility" as their favorite superpower. Similarly, $\frac{27}{222}$, or approximately 0.122 or 12.2%, of males prefer "invisibility" as their favorite superpower.

Exercises 1–5

1. Use the frequency counts from the table in Exploratory Challenge 1 to calculate the missing row of conditional relative frequencies. Round the answers to the nearest thousandth.

	To Fly	Freeze Time	Invisibility	Super Strength	Telepathy	Total
Females			$\frac{48}{228} \approx 0.211$			
Males	$\frac{51}{222} \approx 0.230$					$\frac{222}{222} = 1.000$
Total						

2. Suppose that a student is selected at random from those who completed the survey. What do you think is the gender of the student selected? What would you predict for this student's response to the superpower question?

3. Suppose that a student is selected at random from those who completed the survey. If the selected student is male, what do you think was his response to the selection of a favorite superpower? Explain your answer.

4. Suppose that a student is selected at random from those who completed the survey. If the selected student is female, what do you think was her response to the selection of a favorite superpower? Explain your answer.

5. What superpower was selected by approximately one-third of the females? What superpower was selected by approximately one-third of the males? How did you determine each answer from the conditional relative frequency table?

Exploratory Challenge 2: Possible Association Based on Conditional Relative Frequencies

Two categorical variables are associated if the row conditional relative frequencies (or column relative frequencies) are different for the rows (or columns) of the table. For example, if the selection of superpowers selected for females is different than the selection of superpowers for males, then gender and superpower favorites are associated. This difference indicates that knowing the gender of a person in the sample indicates something about their superpower preference.

The evidence of an association is strongest when the conditional relative frequencies are quite different. If the conditional relative frequencies are nearly equal for all categories, then there is probably not an association between variables.

Exercises 6–10

Examine the conditional relative frequencies in the two-way table of conditional relative frequencies you created in Exercise 1. Note that for each superpower, the conditional relative frequencies are different for females and males.

6. For what superpowers would you say that the conditional relative frequencies for females and males are very different?

7. For what superpowers are the conditional relative frequencies nearly equal for males and females?

8. Suppose a student is selected at random from the students who completed the survey. If you had to predict which superpower this student selected, would it be helpful to know the student's gender? Explain your answer.

9. Is there evidence of an association between gender and a favorite superpower? Explain why or why not.

10. What superpower would you recommend the students at Rufus King High School select for their superhero character? Justify your choice.

Exploratory Challenge 3: Association and Cause-and-Effect

Students were given the opportunity to prepare for a college placement test in mathematics by taking a review course. Not all students took advantage of this opportunity. The following results were obtained from a random sample of students who took the placement test.

	Placed in Math 200	Placed in Math 100	Placed in Math 50	Total
Took Review Course	40	13	7	60
Did Not Take Review Course	10	15	15	40
Total	50	28	22	100

Exercises 11–16

11. Construct a row conditional relative frequency table of the above data.

	Placed in Math 200	Placed in Math 100	Placed in Math 50	Total
Took Review Course				
Did Not Take Review Course				
Total				

12. Based on the conditional relative frequencies, is there evidence of an association between whether a student takes the review course and the math course in which the student was placed? Explain your answer.

13. Looking at the conditional relative frequencies, the proportion of students who placed into Math 200 is much higher for those who took the review course than for those who did not. One possible explanation is that taking the review course caused improvement in placement test scores. What is another possible explanation?

Now consider the following statistical study:

Fifty students were selected at random from students at a large middle school. Each of these students was classified according to sugar consumption (high or low) and exercise level (high or low). The resulting data are summarized in the following frequency table.

		Exercise Level		
		High	**Low**	**Total**
Sugar Consumption	**High**	14	18	32
	Low	14	4	18
	Total	28	22	50

14. Calculate the row conditional relative frequencies, and display them in a row conditional relative frequency table.

		Exercise Level		
		High	**Low**	**Total**
Sugar Consumption	**High**			
	Low			
	Total			

15. Is there evidence of an association between sugar consumption category and exercise level? Support your answer using conditional relative frequencies.

16. Do you think it is reasonable to conclude that high sugar consumption is the cause of the observed differences in the conditional relative frequencies? What other explanations could explain a difference in the conditional relative frequencies? Explain your answer.

Lesson Summary

- A conditional relative frequency compares a frequency count to the marginal total that represents the *condition* of interest.
- The differences in conditional relative frequencies are used to assess whether or not there is an association between two categorical variables.
- The greater the differences in the conditional relative frequencies, the stronger the evidence that an association exits.
- An observed association between two variables does not necessarily mean that there is a cause-and-effect relationship between the two variables.

Problem Set

Consider again the summary of data from the 100 randomly selected students in the Rufus King High School investigation of after-school activities and gender.

	Intramural Basketball	Chess Club	Jazz Band	Not Involved	Total
Females	20	10	10	20	60
Males	20	2	8	10	40
Total	40	12	18	30	100

1. Construct a row conditional relative frequency table for this data. Decimal values are given to the nearest thousandth.

	Intramural Basketball	Chess Club	Jazz Band	Not Involved	Total
Females					60
Males					40
Total					

2. For what after-school activities do you think the row conditional relative frequencies for females and males are very different? What might explain why males or females select different activities?

3. If John, a male student at Rufus King High School, completed the after-school survey, what would you predict was his response? Explain your answer.

4. If Beth, a female student at Rufus King High School, completed the after-school survey, what would you predict was her response? Explain your answer.

5. Notice that 20 female students participate in intramural basketball and that 20 male students participate in intramural basketball. Is it accurate to say that females and males are equally involved in intramural basketball? Explain your answer.

6. Do you think there is an association between gender and choice of after-school program? Explain.

Column conditional relative frequencies can also be computed by dividing each frequency in a frequency table by the corresponding column total to create a column conditional relative frequency table. Column conditional relative frequencies indicate the proportions, or relative frequencies, based on the column totals.

7. If you wanted to know the relative frequency of females surveyed who participated in chess club, would you use a row conditional relative frequency or a column conditional relative frequency?

8. If you wanted to know the relative frequency of band members surveyed who were female, would you use a row conditional relative frequency or a column conditional relative frequency?

9. For the superpower survey data, write a question that would be answered using a row conditional relative frequency.

10. For the superpower survey data, write a question that would be answered using a column conditional relative frequency.

Lesson 12: Relationships Between Two Numerical Variables

Classwork

A scatter plot is an informative way to display numerical data with two variables. In your previous work in Grade 8, you saw how to construct and interpret scatter plots. Recall that if the two numerical variables are denoted by x and y, the scatter plot of the data is a plot of the (x, y) data pairs.

Example 1: Looking for Patterns in a Scatter Plot

The National Climate Data Center collects data on weather conditions at various locations. They classify each day as clear, partly cloudy, or cloudy. Using data taken over a number of years, they provide data on the following variables.

$x =$ elevation above sea level (in feet)

$y =$ mean number of clear days per year

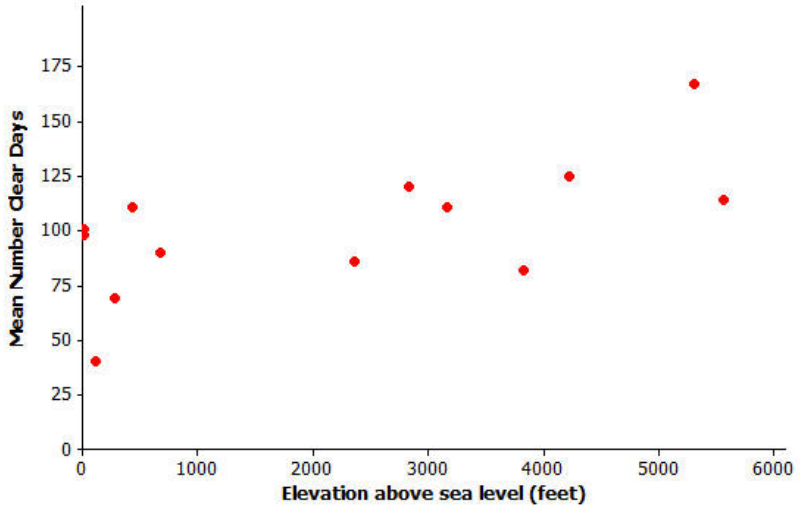
$w =$ mean number of partly cloudy days per year

$z =$ mean number of cloudy days per year

The table below shows data for 14 U.S. cities.

City	$x =$ Elevation Above Sea Level (ft.)	$y =$ Mean Number of Clear Days per Year	$w =$ Mean Number of Partly Cloudy Days per Year	$z =$ Mean Number of Cloudy Days per Year
Albany, NY	275	69	111	185
Albuquerque, NM	5,311	167	111	87
Anchorage, AK	114	40	60	265
Boise, ID	2,838	120	90	155
Boston, MA	15	98	103	164
Helena, MT	3,828	82	104	179
Lander, WY	5,557	114	122	129
Milwaukee, WI	672	90	100	175
New Orleans, LA	4	101	118	146
Raleigh, NC	434	111	106	149
Rapid City, SD	3,162	111	115	139
Salt Lake City, UT	4,221	125	101	139
Spokane, WA	2,356	86	88	191
Tampa, FL	19	101	143	121

Here is a scatter plot of the data on elevation and mean number of clear days.



Data Source: http://www.ncdc.noaa.gov/oa/climate/online/ccd/cldy.html

Exercises 1–3

1. Do you see a pattern in the scatter plot, or does it look like the data points are scattered?

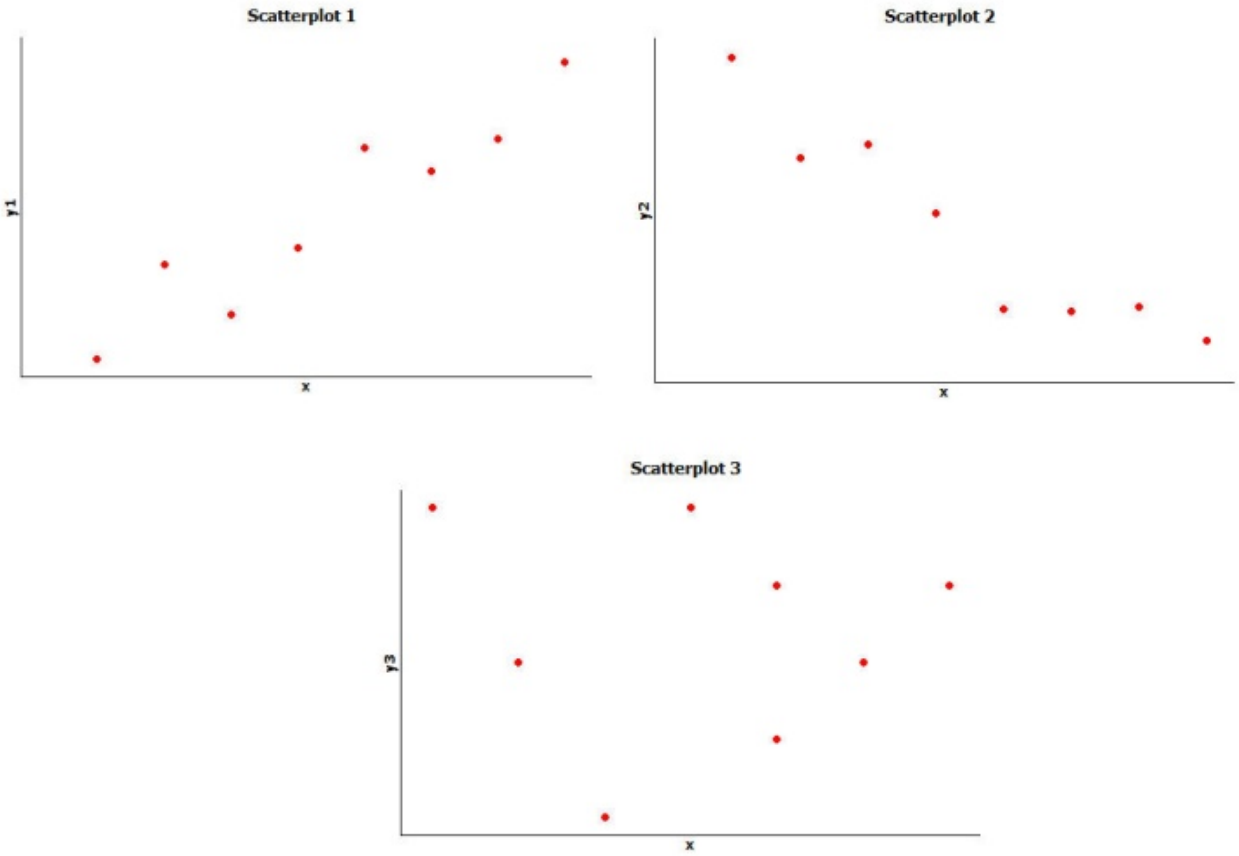
2. How would you describe the relationship between elevation and mean number of clear days for these 14 cities? That is, does the mean number of clear days tend to increase as elevation increases, or does the mean number of clear days tend to decrease as elevation increases?

3. Do you think that a straight line would be a good way to describe the relationship between the mean number of clear days and elevation? Why do you think this?

Exercises 4–7: Thinking about Linear Relationships

Below are three scatter plots. Each one represents a data set with eight observations.

The scales on the x- and y-axes have been left off these plots on purpose so you will have to think carefully about the relationships.



4. If one of these scatter plots represents the relationship between height and weight for eight adults, which scatter plot do you think it is and why?

5. If one of these scatter plots represents the relationship between height and SAT math score for eight high-school seniors, which scatter plot do you think it is and why?

6. If one of these scatter plots represents the relationship between the weight of a car and fuel efficiency for eight cars, which scatter plot do you think it is and why?

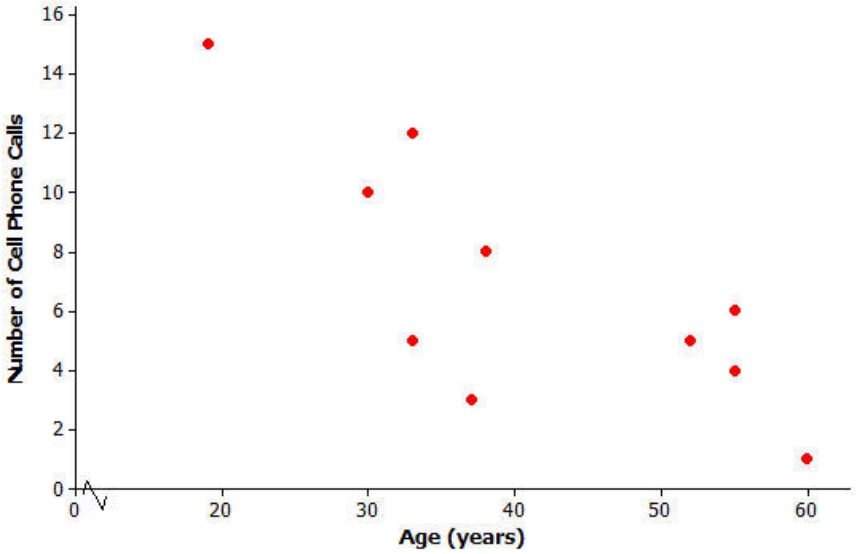
7. Which of these three scatter plots does *not* appear to represent a linear relationship? Explain the reasoning behind your choice.

Exercises 8–13: Not Every Relationship Is Linear

When a straight line provides a reasonable summary of the relationship between two numerical variables, we say that the two variables are *linearly related* or that there is a *linear relationship* between the two variables.

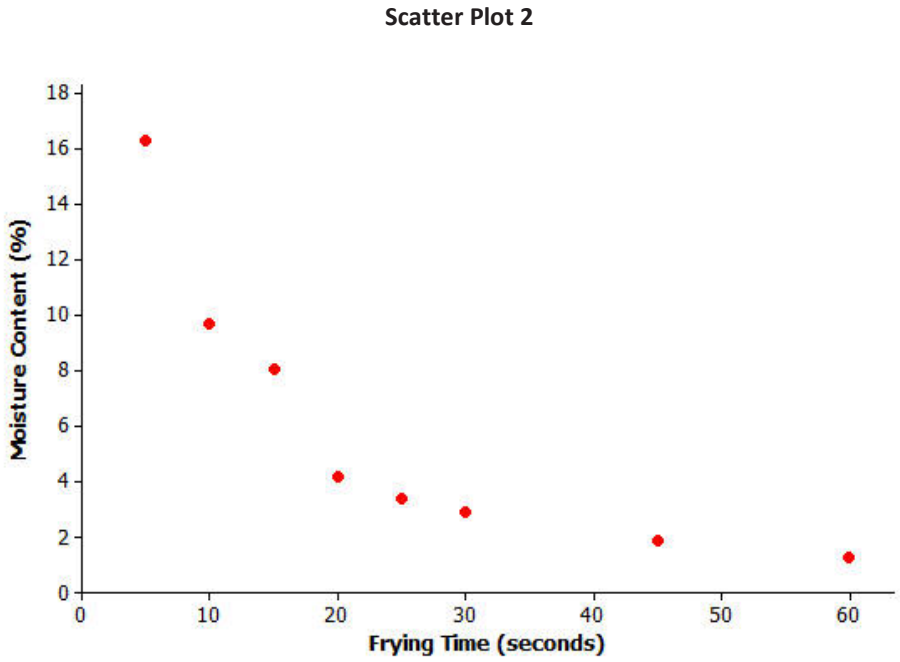
Take a look at the scatter plots below and answer the questions that follow.

Scatter Plot 1



8. Is there a relationship between number of cell phone calls and age, or does it look like the data points are scattered?

9. If there is a relationship between number of cell phone calls and age, does the relationship appear to be linear?

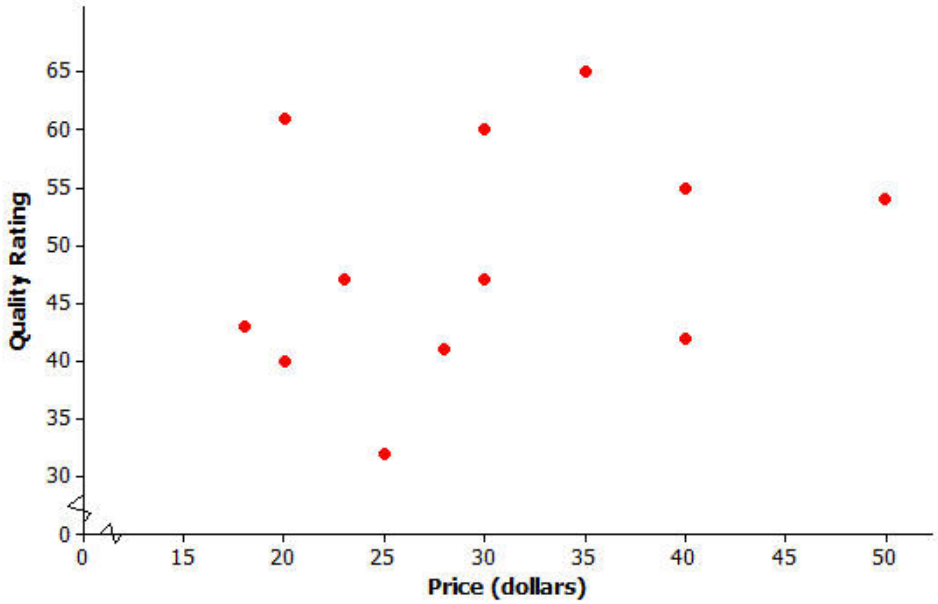


Data Source: R.G. Moreira, J. Palau, V.E. Sweat, and X. Sun, "Thermal and Physical Properties of Tortilla Chips as a Function of Frying Time," *Journal of Food Processing and Preservation,* 19 (1995): 175.

10. Is there a relationship between moisture content and frying time, or do the data points look scattered?

11. If there is a relationship between moisture content and frying time, does the relationship look linear?

Scatter Plot 3



Data Source: www.consumerreports.org/health

12. Scatter plot 3 shows data for the prices of bike helmets and the quality ratings of the helmets (based on a scale that estimates helmet quality). Is there a relationship between quality rating and price, or are the data points scattered?

13. If there is a relationship between quality rating and price for bike helmets, does the relationship appear to be linear?

Lesson Summary

- A scatter plot can be used to investigate whether or not there is a relationship between two numerical variables.
- A relationship between two numerical variables can be described as a linear or nonlinear relationship.

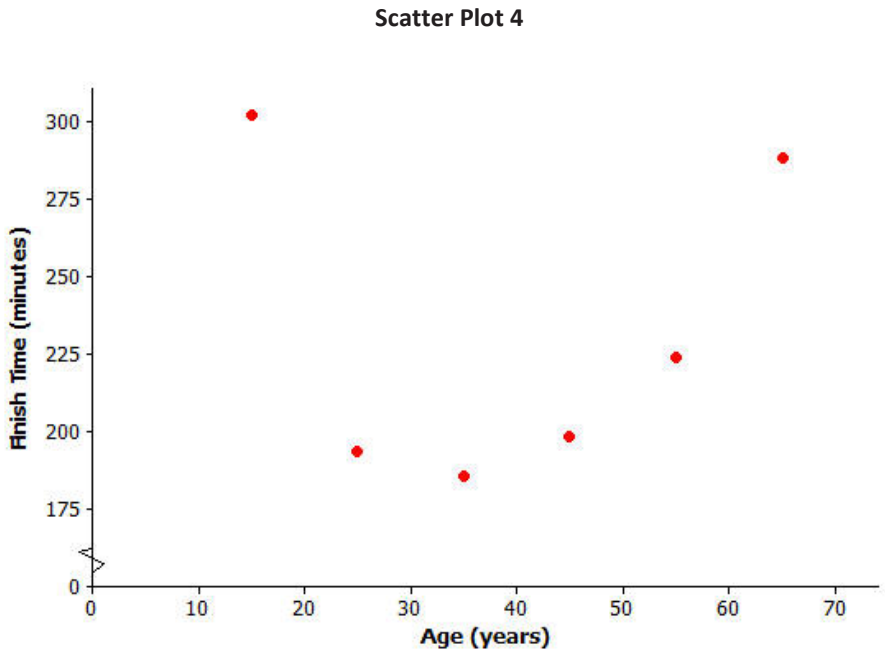
Problem Set

1. Construct a scatter plot that displays the data for $x =$ elevation above sea level (in feet) and $w =$ mean number of *partly cloudy days per year*.

City	$x =$ Elevation Above Sea Level (ft.)	$y =$ Mean Number of Clear Days per Year	$w =$ Mean Number of Partly Cloudy Days per Year	$z =$ Mean Number of Cloudy Days per Year
Albany, NY	275	69	111	185
Albuquerque, NM	5,311	167	111	87
Anchorage, AK	114	40	60	265
Boise, ID	2,838	120	90	155
Boston, MA	15	98	103	164
Helena, MT	3,828	82	104	179
Lander, WY	5,557	114	122	129
Milwaukee, WI	672	90	100	175
New Orleans, LA	4	101	118	146
Raleigh, NC	434	111	106	149
Rapid City, SD	3,162	111	115	139
Salt Lake City, UT	4,221	125	101	139
Spokane, WA	2,356	86	88	191
Tampa, FL	19	101	143	121

2. Based on the scatter plot you constructed in Question 1, is there a relationship between elevation and the mean number of partly cloudy days per year? If so, how would you describe the relationship? Explain your reasoning.

Consider the following scatter plot for Questions 3 and 4.

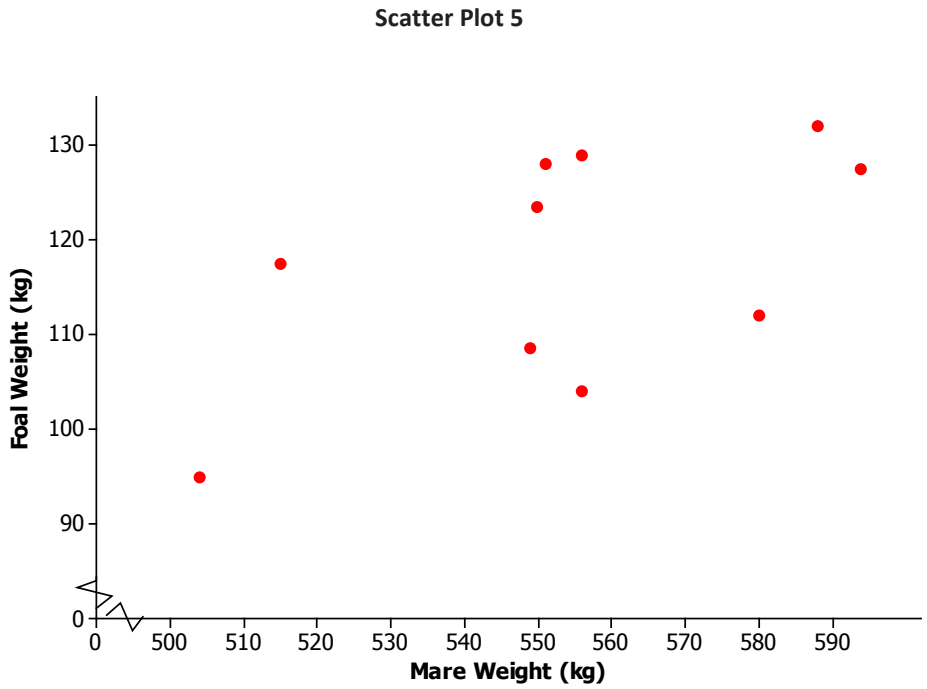


Data Source: Sample of six women who ran the 2003 NYC marathon

3. Is there a relationship between finish time and age, or are the data points scattered?

4. Do you think there is a relationship between finish time and age? If so, does it look linear?

Consider the following scatter plot for Questions 5 and 6.



Data Source: Elissa Z. Cameron, Kevin J. Stafford, Wayne L. Linklater, and Clare J. Veltman, "Suckling behaviour does not measure milk intake in horses, equus caballus," *Animal Behaviour,* 57 (1999): 673.

5. A mare is a female horse and a foal is a baby horse. Is there a relationship between a foal's birth weight and a mare's weight, or are the data points scattered?

6. If there is a relationship between baby birth weight and mother's age, does the relationship look linear?

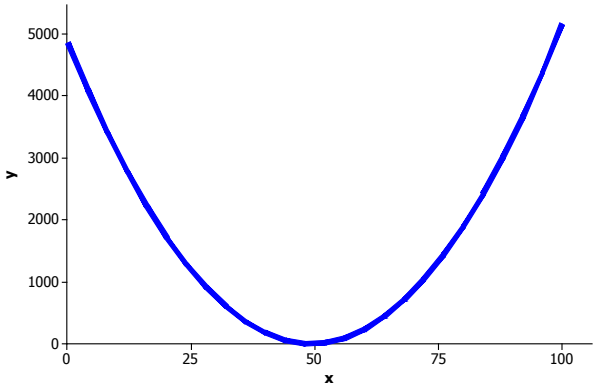
Lesson 13: Relationships Between Two Numerical Variables

Classwork

Not all relationships between two numerical variables are *linear*. There are many situations where the pattern in the scatter plot would best be described by a curve. Two types of functions often used in modeling nonlinear relationships are *quadratic* and *exponential* functions.

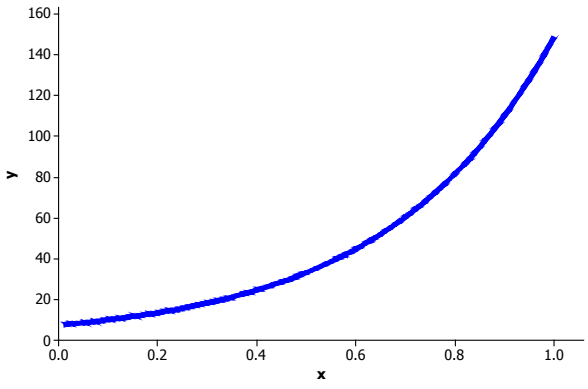
Example 1: Modeling Relationships

Sometimes the pattern in a scatter plot will look like the graph of a quadratic function (with the points falling roughly in the shape of a U that opens up or down), as in the graph below.

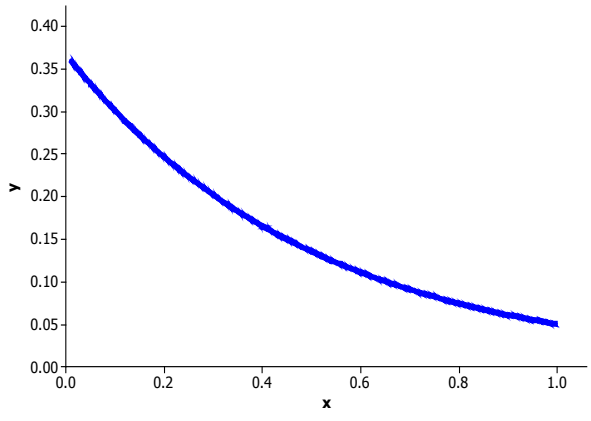


In other situations, the pattern in the scatter plot might look like the graphs of exponential functions that either are upward sloping (Graph 1) or downward sloping (Graph 2).

Graph 1: Exponential – upward sloping



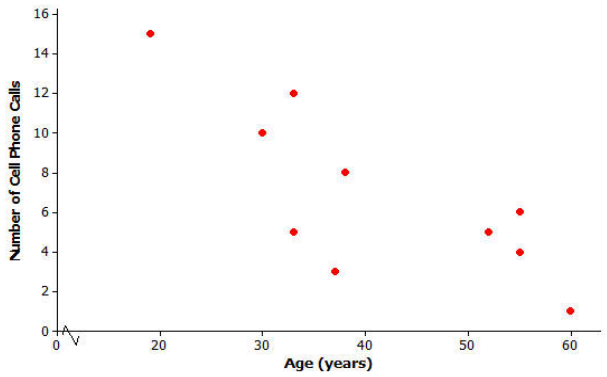
Graph 2: Exponential – downward sloping



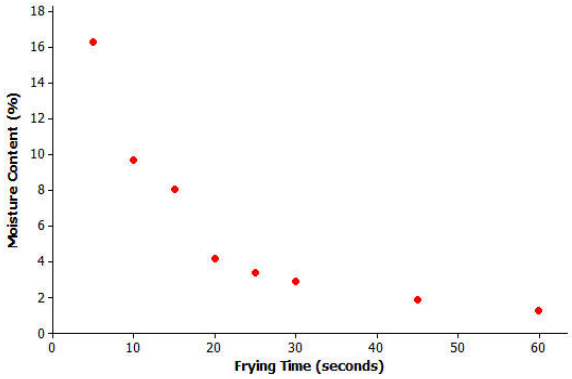
Exercises 1–6

Consider again the five scatter plots discussed in the previous lesson.

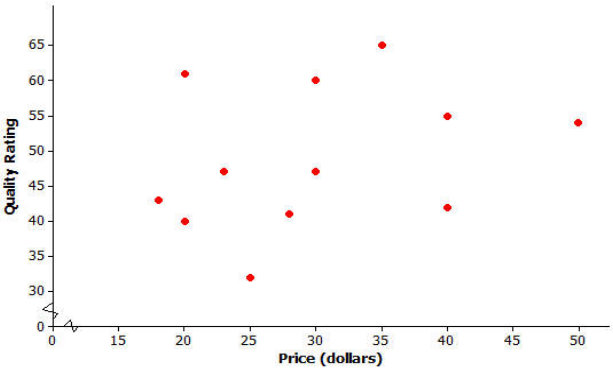
Scatter Plot 1



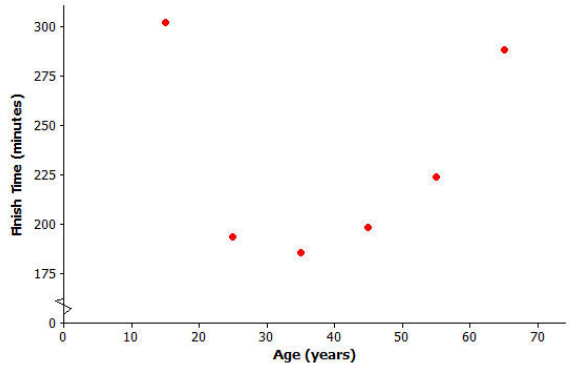
Scatter Plot 2



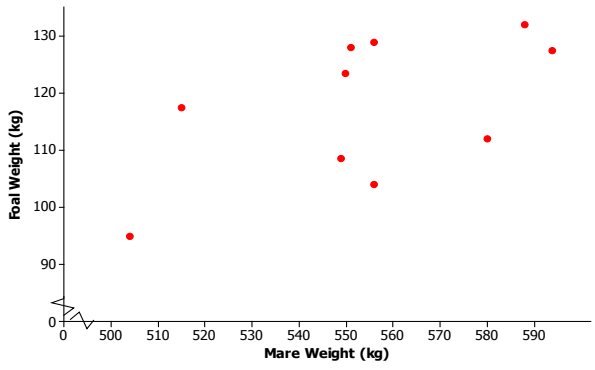
Scatter Plot 3



Scatter Plot 4



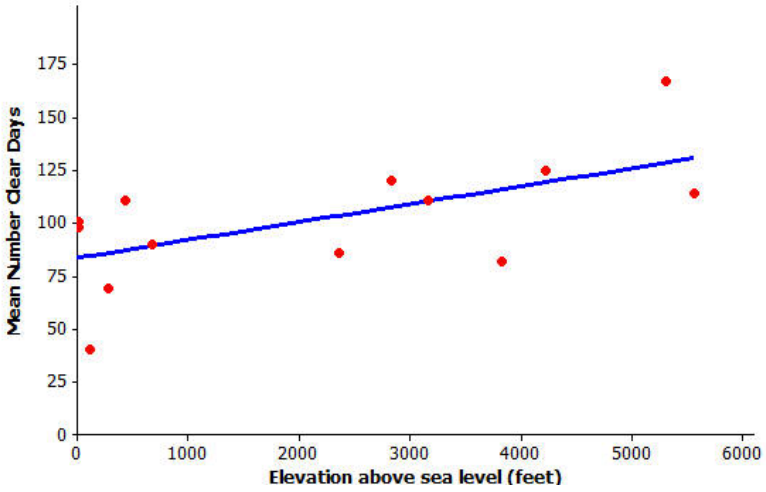
Scatter Plot 5



1. Which of the five scatter plots from Lesson 12 show a pattern that could be reasonably described by a quadratic curve?

2. Which of the five scatter plots show a pattern that could be reasonably described by an exponential curve?

Let's revisit the data on elevation (in feet above sea level) and mean number of clear days per year. The scatter plot of this data is shown below. The plot also shows a straight line that can be used to model the relationship between elevation and mean number of clear days. (In Grade 8, you informally fit a straight line to model the relationship between two variables. The next lesson shows a more formal way to fit a straight line.) The equation of this line is $y = 83.6 + 0.008x$.



3. Assuming that the 14 cities used in this scatter plot are representative of cities across the United States, should you see more clear days per year in Los Angeles, which is near sea level, or in Denver, which is known as the mile-high city? Justify your choice with a line showing the relationship between elevation and mean number of clear days.

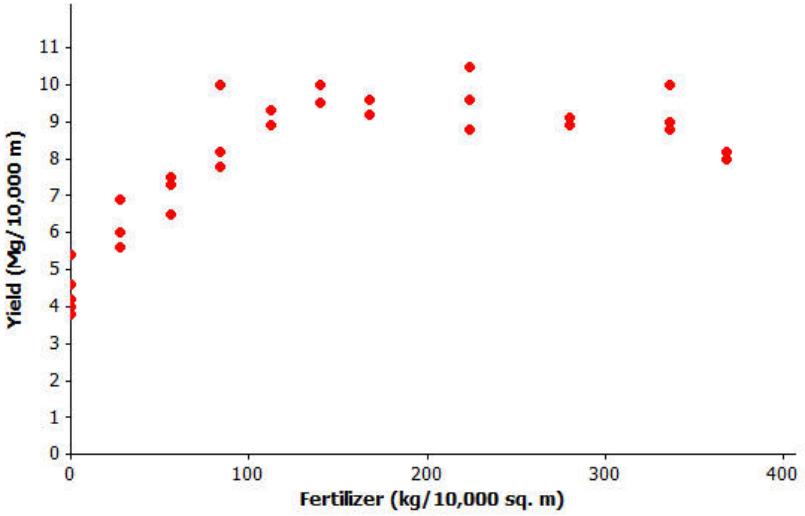
4. One of the cities in the data set was Albany, New York, which has an elevation of 275 feet. If you did not know the mean number of clear days for Albany, what would you predict this number to be based on the line that describes the relationship between elevation and mean number of clear days?

5. Another city in the data set was Albuquerque, New Mexico. Albuquerque has an elevation of 5,311 feet. If you did not know the mean number of clear days for Albuquerque, what would you predict this number to be based on the line that describes the relationship between elevation and mean number of clear days?

6. Was the prediction of the mean number of clear days based on the line closer to the actual value for Albany with 69 clear days or for Albuquerque with 167 clear days? How could you tell this from looking at the scatter plot with the line shown above?

Example 2: A Quadratic Model

Farmers sometimes use fertilizers to increase crop yield, but often wonder just how much fertilizer they should use. The data shown in the scatter plot below are from a study of the effect of fertilizer on the yield of corn.



Data Source: M.E. Cerrato and A.M. Blackmer, "Comparison of Models for Describing Corn Yield Response to Nitrogen Fertilizer" *Agronomy Journal,* 82 (1990): 138.

Exercises 7–9

7. The researchers who conducted this study decided to use a quadratic curve to describe the relationship between yield and amount of fertilizer. Explain why they made this choice.

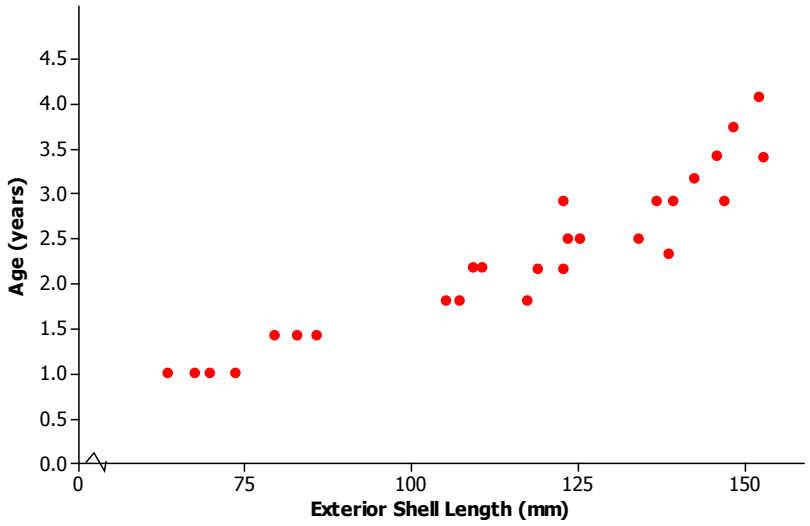
8. The model that the researchers used to describe the relationship was $y = 4.7 + 0.05x - 0.0001x^2$, where x represents the amount of fertilizer (kg per $10{,}000$ sq. m) and y represents corn yield (Mg per $10{,}000$ sq. m). Use this quadratic model to complete the following table. Then sketch the graph of this quadratic equation on the scatter plot.

x	y
0	
100	
200	
300	
400	

9. Based on this quadratic model, how much fertilizer per $10{,}000$ square meters would you recommend that a farmer use on his cornfields in order to maximize crop yield? Justify your choice.

Example 3: An Exponential Model

How do you tell how old a lobster is? This question is important to biologists and to those who regulate lobster trapping. To answer this question, researchers recorded data on the shell length of 27 lobsters that were raised in a laboratory and whose ages were known.

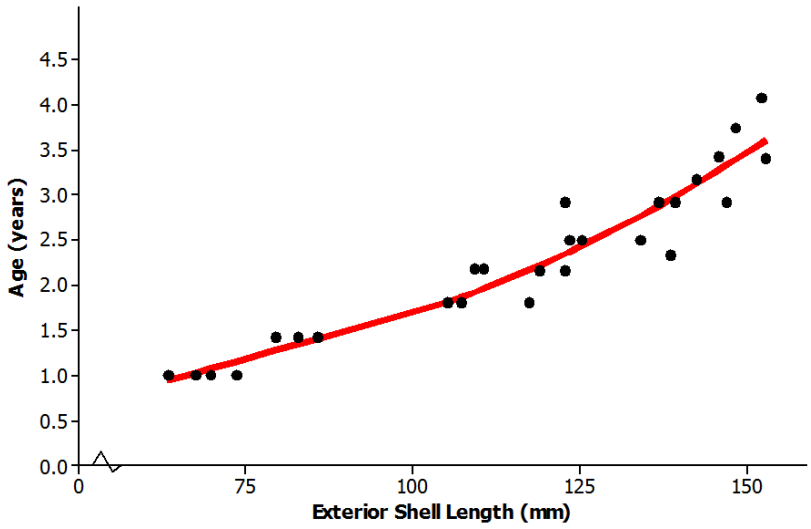


Data Source: Kerry E. Maxwell, Thomas R. Matthews, Matt R.J. Sheehy, Rodney D. Bertelsen, and Charles D. Derby, "Neurolipofuscin is a Measure of Age in *Panulirus argus*, the Caribbean Spiny Lobser, in Florida" *Biological Bulletin,* 213 (2007): 55.

Exercises 10–13

10. The researchers who conducted this study decided to use an exponential curve to describe the relationship between age and exterior shell length. Explain why they made this choice.

11. The model that the researchers used to describe the relationship is $y = 10^{-0.403 + 0.0063x}$, where x represents the exterior shell length (mm) and y represents the age of the lobster (years). The exponential curve is shown on the scatter plot below. Does this model provide a good description of the relationship between age and exterior shell length? Explain why or why not.



12. Based on this exponential model, what age is a lobster with an exterior shell length of 100 mm?

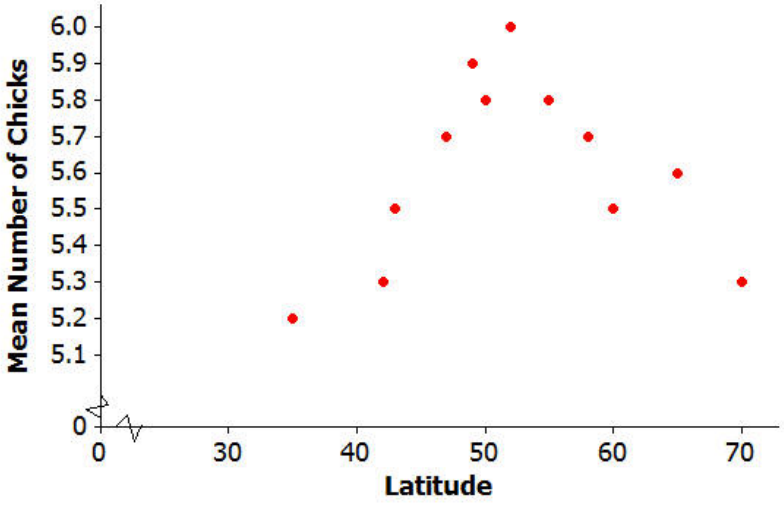
13. Suppose that trapping regulations require that any lobster with an exterior shell length less than 75 mm or more than 150 mm must be released. Based on the exponential model, what are the ages of lobsters with exterior shell lengths less than 75 mm? What are the ages of lobsters with exterior shell lengths greater than 150 mm? Explain how you arrived at your answer.

Lesson Summary

- A scatter plot can be used to investigate whether or not there is a relationship between two numerical variables.
- Linear, quadratic, and exponential functions are common models that can be used to describe the relationship between variables.
- Models can be used to answer questions about how two variables are related.

Problem Set

Biologists conducted a study of the nesting behavior of a type of bird called a flycatcher. They examined a large number of nests and recorded the latitude for the location of the nest and the number of chicks in the nest.



Data Source: Juan José Sanz, "Geographic variation in breeding parameters of the pied flycatcher *Ficedula hypoleuca*" *Ibis,* 139 (1997): 107.

1. What type of model (linear, quadratic or exponential) would best describe the relationship between latitude and mean number of chicks?

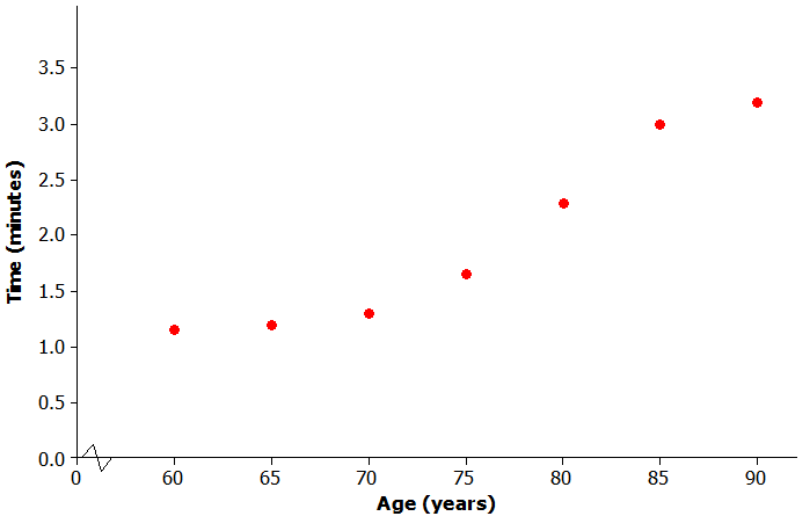
2. One model that could be used to describe the relationship between mean number of chicks and latitude is $y = 0.175 + 0.21x - 0.002x^2$, where x represents the latitude of the location of the nest and y represents the number of chicks in the nest. Use the quadratic model to complete the following table. Then sketch a graph of the quadratic curve on the scatter plot above.

x (degrees)	y
30	
40	
50	
60	
70	

3. Based on this quadratic model, what is the best latitude for hatching the most flycatcher chicks? Justify your choice.

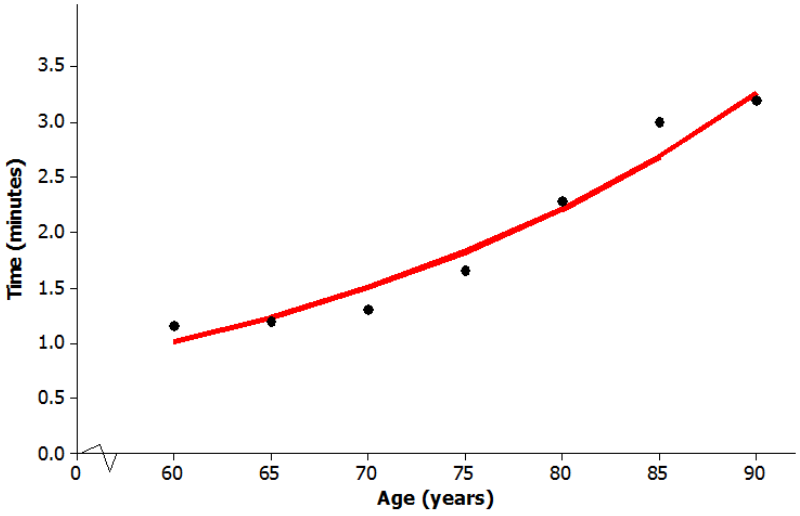
Suppose that social scientists conducted a study of senior citizens to see how the time (in minutes) required to solve a word puzzle changes with age. The scatter plot below displays data from this study.

Let x equal the age of the citizen and y equal the time (in minutes) required to solve a word puzzle for the seven study participants.



4. What type of model (linear, quadratic, or exponential) would you use to describe the relationship between age and time required to complete the word puzzle?

5. One model that could describe the relationship between age and time to complete the word puzzle is $y = 10^{-1.01 + 0.017x}$. This exponential curve is shown on the scatter plot below. Does this model do a good job of describing the relationship between age and time to complete the word puzzle? Explain why or why not.



6. Based on this exponential model, what time would you predict for a person who is 78 years old?

Lesson 14: Modeling Relationships with a Line

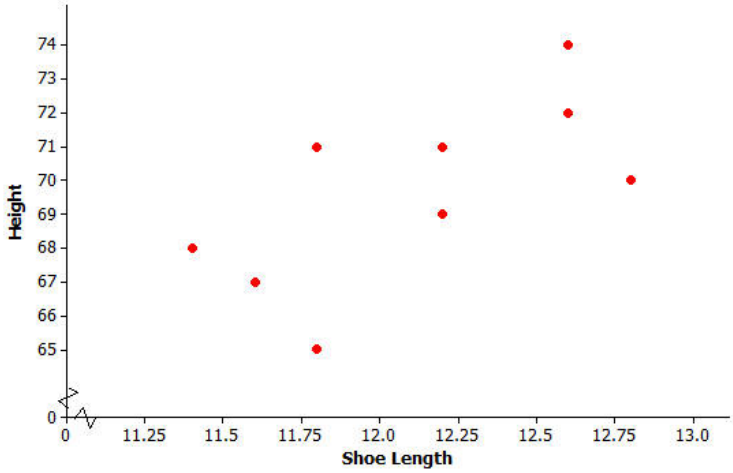
Classwork

Example 1: Using a Line to Describe a Relationship

Kendra likes to watch crime scene investigation shows on television. She watched a show where investigators used a shoe print to help identify a suspect in a case. She questioned how possible it is to predict someone's height is from his shoe print.

To investigate, she collected data on shoe length (in inches) and height (in inches) from 10 adult men. Her data appear in the table and scatter plot below.

x = Shoe Length	y = Height
12.6	74
11.8	65
12.2	71
11.6	67
12.2	69
11.4	68
12.8	70
12.2	69
12.6	72
11.8	71



Exercises 1–2

1. Is there a relationship between shoe length and height?

2. How would you describe the relationship? Do the men with longer shoe lengths tend be taller?

Example 2: Using Models to Make Predictions

When two variables x and y are linearly related, you can use a line to describe their relationship. You can also use the equation of the line to predict the value of the y variable based on the value of the x variable.

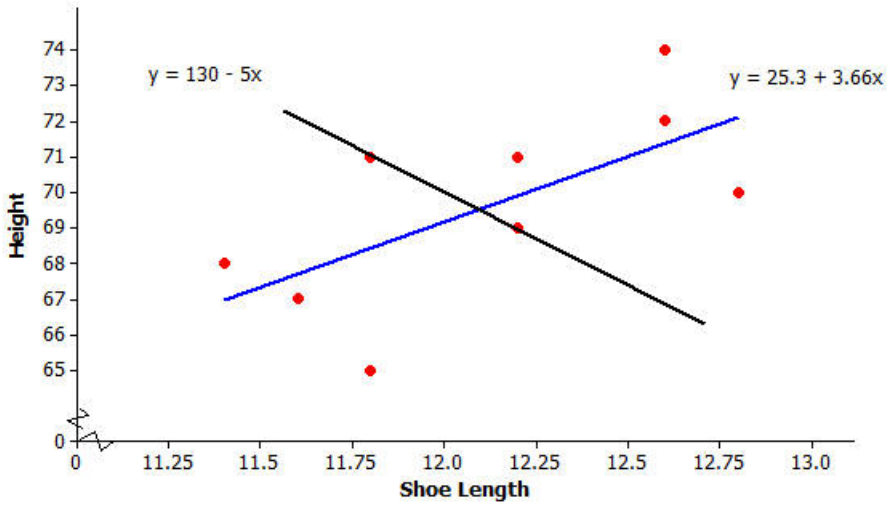
For example, the line $y = 25.3 + 3.66x$ might be used to describe the relationship between shoe length and height, where x represents shoe length and y represents height. To predict the height of a man with a shoe length of 12, you would substitute 12 in for x in the equation of the line and then calculate the value of y.

$$y = 25.3 + 3.66x = 25.3 + 3.66(12) = 69.22$$

You would predict a height of 69.22 inches for a man with a shoe length of 12 inches.

Exercises 3–7

3. Below is a scatter plot of the data with two linear models, $y = 130 - 5x$ and $y = 25.3 + 3.66x$. Which of these two models does a better job of describing how shoe length (x) and height (y) are related? Explain your choice.



4. One of the men in the sample has a shoe length of 11.8 inches and a height of 71 inches. Circle the point in the scatter plot in Question 3 that represents this man.

5. Suppose that you do not know this man's height, but do know that his shoe length is 11.8 inches. If you use the model $y = 25.3 + 3.66x$, what would you predict his height to be? If you use the model $y = 130 - 5x$, what would you predict his height to be?

6. Which model was closer to the actual height of 71 inches? Is that model a better fit to the data? Explain your answer.

7. Is there a better way to decide which of two lines provides a better description of a relationship (rather than just comparing the predicted value to the actual value for one data point in the sample)?

Example 3: Residuals

One way to think about how useful a line is for describing a relationship between two variables is to use the line to predict the y-values for the points in the scatter plot. These predicted values could then be compared to the actual y-values.

For example, the first data point in the table represents a man with a shoe length of 12.6 inches and height of 74 inches. If you use the line $y = 25.3 + 3.66x$ to predict this man's height, you would get:

$$\begin{aligned} y &= 25.3 + 3.66x \\ &= 25.3 + 3.66(12.6) \\ &= 71.42 \text{ in.} \end{aligned}$$

Because his actual height was 74 inches, you can calculate the prediction error by subtracting the predicted value from the actual value. This prediction error is called a *residual*. For the first data point, the residual is calculated as follows:

$$\begin{aligned} \text{Residual} &= \text{actual } y\text{-value} - \text{predicted } y\text{-value} \\ &= 74 - 71.42 \\ &= 2.58 \text{ in.} \end{aligned}$$

Exercises 8–10

8. For the line $y = 25.3 + 3.66x$, calculate the missing values and add them to complete the table.

$x =$ Shoe Length	$y =$ Height	Predicted y-value	Residual
12.6	74	71.42	2.58
11.8	65		−3.49
12.2	71		
11.6	67	67.76	−0.76
12.2	69	69.95	−0.95
11.4	68	67.02	
12.8	70	72.15	−2.15
12.2	69		−0.95
12.6	72	71.42	0.58
11.8	71	68.49	2.51

9. Why is the residual in the table's first row positive, and the residual in the second row negative?

10. What is the sum of the residuals? Why did you get a number close to zero for this sum? Does this mean that all of the residuals were close to 0?

Exercises 11–13

When you use a line to describe the relationship between two numerical variables, the *best* line is the line that makes the residuals as small as possible overall.

11. If the residuals tend to be small, what does that say about the fit of the line to the data?

The most common choice for the *best* line is the line that makes the sum of the *squared* residuals as small as possible. Add a column on the right of the table in Exercise 8. Calculate the square of each residual and place the answer in the column.

12. Why do we use the sum of the squared residuals instead of just the sum of the residuals (without squaring)? Hint: Think about whether the sum of the residuals for a line can be small even if the prediction errors are large. Can this happen for squared residuals?

13. What is the sum of the squared residuals for the line $y = 25.3 + 3.66x$ and the data of Exercise 11?

Example 4: The Least squares Line (Best-Fit Line)

The line that has a smaller sum of squared residuals for this data set than any other line is called the *least squares line*. This line can also be called the *best-fit line* or the *line of best fit* (or regression line).

For the shoe-length and height data for the sample of 10 men, the line $y = 25.3 + 3.66x$ is the least squares line. No other line would have a smaller sum of squared residuals for this data set than this line.

There are equations that can be used to calculate the value for the slope and the intercept of the least squares line, but these formulas require a lot of tedious calculations. Fortunately, a graphing calculator can be used to find the equation of the least squares line.

Your teacher will show you how to enter data and obtain the equation of the least squares line using your graphing calculator or other statistics program.

Exercises 14–17

14. Enter the shoe-length and height data and then use your calculator to find the equation of the least squares line. Did you get $y = 25.3 + 3.66x$? (The slope and y-intercept here have been rounded to the nearest hundredth.)

15. Assuming that the 10 men in the sample are representative of adult men in general, what height would you predict for a man whose shoe length is 12.5 inches? What height would you predict for a man whose shoe length is 11.9 inches?

Once you have found the equation of the least squares line, the values of the slope and y-intercept of the line often reveals something interesting about the relationship you are modeling.

The slope of the least squares line is the change in the predicted value of the y variable associated with an increase of one in the value of the x variable.

16. Give an interpretation of the slope of the least squares line $y = 25.3 + 3.66x$ for predicting height from shoe size for adult men.

The y-intercept of a line is the predicted value of y when x equals zero. When using a line as a model for the relationship between two numerical variables, it often does not make sense to interpret the y-intercept because an x-value of zero may not make any sense.

17. Explain why it does not make sense to interpret the y-intercept of 25.3 as the predicted height for an adult male whose shoe length is zero.

Lesson Summary

When the relationship between two numerical variables x and y is linear, a straight line can be used to describe the relationship. Such a line can then be used to predict the value of y based on the value of x. When a prediction is made, the prediction error is the difference between the actual y-value and the predicted y-value. The prediction error is called a residual, and the residual is calculated as residual = actual y-value − predicted y-value. The least squares line is the line that is used to model a linear relationship. The least squares line is the "best" line in that it has a smaller sum of squared residuals than any other line.

Problem Set

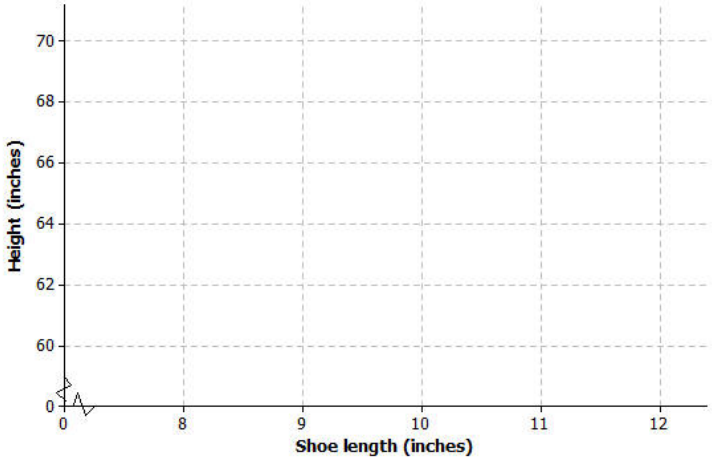
Kendra wondered if the relationship between shoe length and height might be different for men and women. To investigate, she also collected data on shoe length (in inches) and height (in inches) for 12 women.

x = Shoe Length (Women)	y = Height (Women)
8.9	61
9.6	61
9.8	66
10.0	64
10.2	64
10.4	65
10.6	65
10.6	67
10.5	66
10.8	67
11.0	67
11.8	70

1. Construct a scatter plot of these data.

2. Is there a relationship between shoe length and height for these 12 women?

3. Find the equation of the least squares line. (Round values to the nearest hundredth.)



4. Suppose that these 12 women are representative of adult women in general. Based on the least squares line, what would you predict for the height of a woman whose shoe length is 10.5 inches? What would you predict for the height of a woman whose shoe length is 11.5 inches?

5. One of the women in the sample had a shoe length of 9.8 inches. Based on the regression line, what would you predict for her height?

6. What is the value of the residual associated with the observation for the woman with the shoe length of 9.8?

7. Add the predicted value and the residual you just calculated to the table below. Then, calculate the sum of the squared residuals.

$x =$ Shoe Length (Women)	$y =$ Height (Women)	Predicted Height	Residual
8.9	61	60.72	0.28
9.6	61	62.92	-1.92
9.8	66		
10.0	64	64.18	-0.18
10.2	64	64.81	-0.81
10.4	65	65.44	-0.44
10.6	65	66.07	-1.07
10.6	67	66.07	0.93
10.5	66	65.76	0.24
10.8	67	66.7	0.3
11.0	67	67.33	-0.33
11.8	70	69.85	0.15

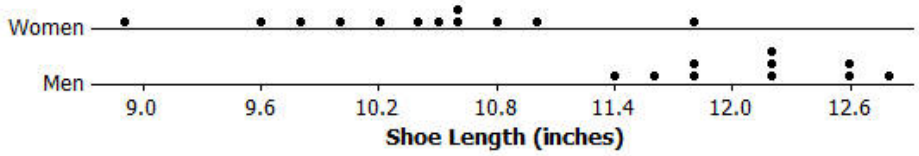
8. Provide an interpretation of the slope of the least squares line.

9. Does it make sense to interpret the y-intercept of the least squares line in this context? Explain why or why not.

10. Would the sum of the squared residuals for the line $y = 25 + 2.8x$ be greater than, about the same as, or less than the sum you computed in Question 7? Explain how you know this. You should be able to answer this question without calculating the sum of squared residuals for this new line.

11. For the men, the least squares line that describes the relationship between $x =$ shoe length (in inches) and $y =$ height (in inches) was $y = 25.3 + 3.66x$. How does this compare to the equation of the least squares line for women? Would you use $y = 25.3 + 3.66x$ to predict the height of a woman based on her shoe length? Explain why or why not.

12. Below are dot plots of the shoe lengths for women and the shoe lengths for men. Suppose that you found a shoe print and that when you measured the shoe length, you got 10.8 inches. Do you think that a man or a woman left this shoe print? Explain your choice.



13. Suppose that you find a shoe print and the shoe length for this print is 12 inches. What would you predict for the height of the person who left this print? Explain how you arrived at this answer.

Lesson 15: Interpreting Residuals from a Line

Classwork

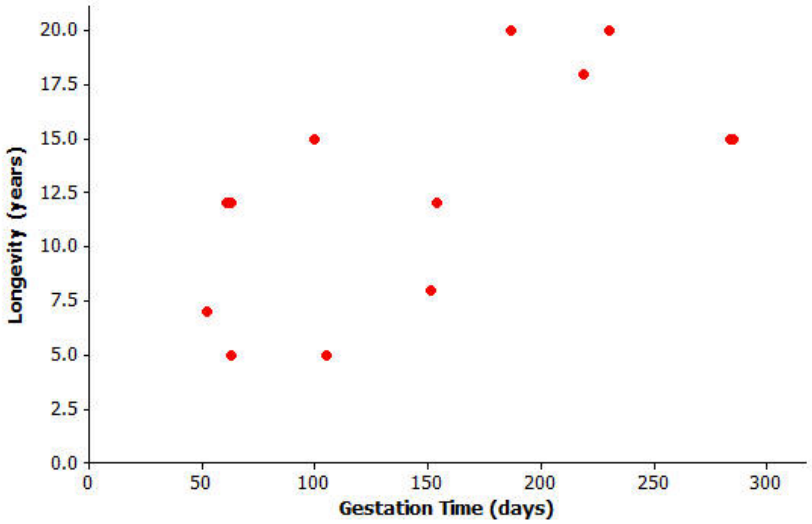
Example 1: Calculating Prediction Errors

The gestation time for an animal is the typical duration between conception and birth. The longevity of an animal is the typical lifespan for that animal. The gestation times (in days) and longevities (in years) for 13 types of animals are shown in the table below.

Animal	Gestation Time (days)	Longevity (years)
Baboon	187	20
Black Bear	219	18
Beaver	105	5
Bison	285	15
Cat	63	12
Chimpanzee	230	20
Cow	284	15
Dog	61	12
Fox (Red)	52	7
Goat	151	8
Lion	100	15
Sheep	154	12
Wolf	63	5

Data Source: *Core Math Tools,* www.nctm.org

Here is the scatter plot for this data set:

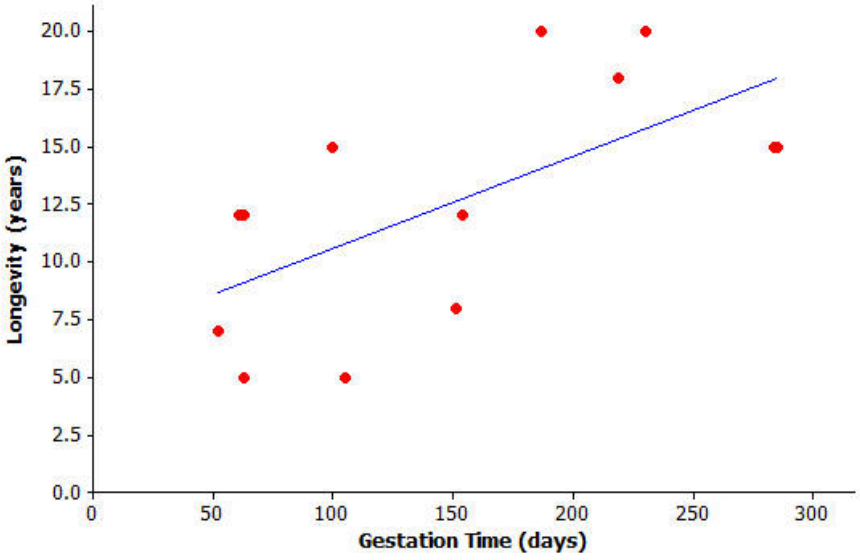


Exercises 1–4

Finding the equation of the least squares line relating longevity to gestation time for these types of animals provides the equation to predict longevity. How good is the line? In other words, if you were given the gestation time for another type of animal not included in the original list, how accurate would the least squares line be at predicting the longevity of that type of animal?

1. Using a graphing calculator, verify that the equation of the least squares line is $y = 6.642 + 0.03974x$, where x represents the gestation time (in days) and y represents longevity (in years).

   The least squares line has been added to the scatter plot below.



2. Suppose a particular type of animal has a gestation time of 200 days. Approximately what value does the line predict for the longevity of that type of animal?

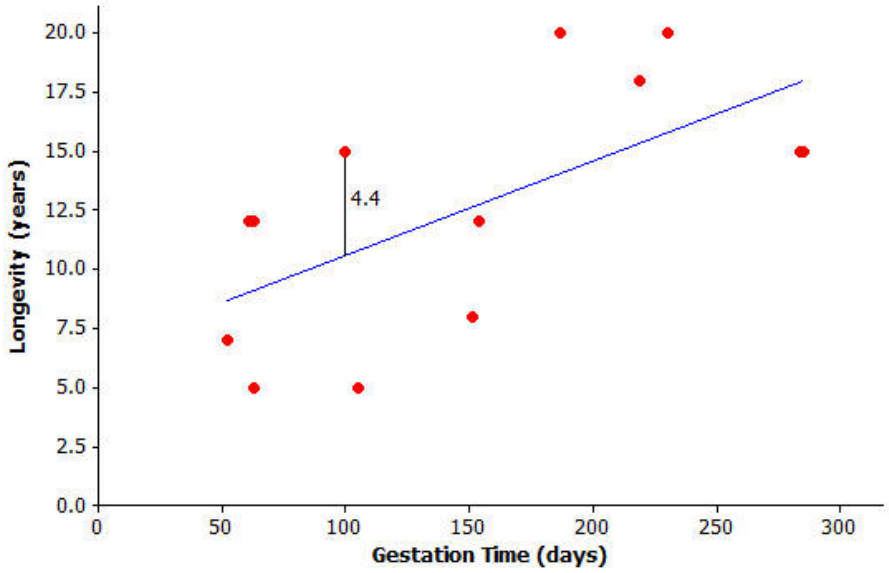
3. Would the value you predicted in Exercise 2 necessarily be the exact value for the longevity of that type of animal? Could the actual longevity of that type of animal be longer than predicted? Could it be shorter?

You can investigate further by looking at the types of animals included in the original data set. Take the lion, for example. Its gestation time is 100 days. You also know that its longevity is 15 years, but what does the least squares line *predict* for the lion's longevity?

Substituting $x = 100$ days into the equation, you get $y = 6.642 + 0.03974(100)$ or approximately 10.6. The least squares line predicts the lion's longevity to be approximately 10.6 years.

4. How close is this to being correct? More precisely, how much do you have to add to 10.6 to get the lion's true longevity of 15?

You can show the prediction error of 4.4 years on the graph like this:



Exercises 5–6

5. Let's continue to think about the gestation times and longevities of animals. Let's specifically investigate how accurately the least squares line predicted the longevity of the black bear.
   a. What is the gestation time for the black bear?

b. Look at the graph. Roughly what does the least squares line predict for the longevity of the black bear?

c. Use the gestation time from part (a) and the least squares line $y = 6.642 + 0.03974x$ to predict the black bear's longevity. Round your answer to the nearest tenth.

d. What is the actual longevity of the black bear?

e. How much do you have to add to the predicted value to get the actual longevity of the black bear?

f. Show your answer to part (e) on the graph as a vertical line segment.

6. Repeat this activity for the sheep.

a. Substitute the sheep's gestation time for x into the equation to find the predicted value for the sheep's longevity. Round your answer to the nearest tenth.

b. What do you have to add to the predicted value in order to get the actual value of the sheep's longevity? (Hint: Your answer should be negative.)

c. Show your answer to part (b) on the graph as a vertical line segment. Write a sentence describing points in the graph for which a negative number would need to be added to the predicted value in order to get the actual value.

Example 2: Residuals as Prediction Errors

In each exercise above, you found out how much needs to be added to the predicted value to find the true value of an animal's longevity. In order to find this you have been calculating

$$\text{actual value} - \text{predicted value.}$$

This quantity is referred to as a residual. It is summarized as

$$\text{residual} = \text{actual } y\text{-value} - \text{predicted } y\text{-value.}$$

You can now work out the residuals for all of the points in our animal longevity example. The values of the residuals are shown in the table below.

Animal	Gestation Time (days)	Longevity (years)	Residual
Baboon	187	20	5.9
Black Bear	219	18	2.7
Beaver	105	5	−5.8
Bison	285	15	−3.0
Cat	63	12	2.9
Chimpanzee	230	20	4.2
Cow	284	15	−2.9
Dog	61	12	2.9
Fox (Red)	52	7	−1.7
Goat	151	8	−4.6
Lion	100	15	4.4
Sheep	154	12	−0.8
Wolf	63	5	−4.1

These residuals show that the actual longevity of an animal should be within six years of the longevity predicted by the least squares line.

Suppose you selected a type of animal that is not included in the original data set, and the gestation time for this type of animal is 270 days. Substituting $x = 270$ into the equation of the least squares line you get

$$\begin{aligned} y &= 6.642 + 0.03974(270) \\ &= 17.4 \text{ years.} \end{aligned}$$

Exercises 7–8

Think about what the *actual* longevity of this type of animal might be.

7. Could it be 30 years? How about 5 years?

8. Judging by the size of the residuals in our table, what kind of values do you think would be reasonable for the longevity of this type of animal?

Exercises 9–10

Continue to think about the gestation times and longevities of animals. The gestation time for the type of animal called the ocelot is known to be 85 days.

The least squares line predicts the longevity of the ocelot to be

$$y = 6.642 + 0.03974(85) = 10.0 \text{ years.}$$

9. Based on the residuals in Example 3, would you be surprised to find that the longevity of the ocelot was 2 years? Why, or why not? What might be a sensible range of values for the actual longevity of the ocelot?

10. We know that the actual longevity of the ocelot is 9 years. What is the residual for the ocelot?

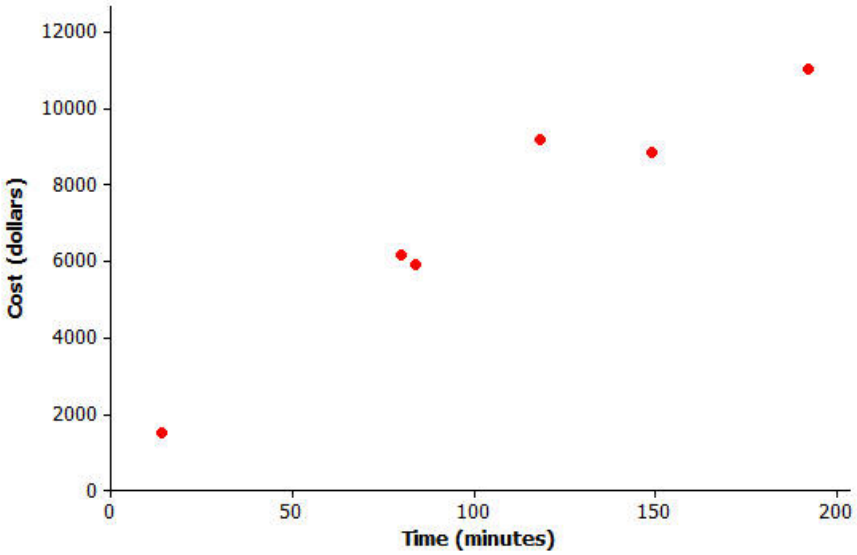
Lesson Summary

- When a least squares line is used to calculate a predicted value, the prediction error can be measured by
$$\text{residual} = \text{actual } y\text{-value} - \text{predicted } y\text{-value}.$$
- On the graph, the residuals are the vertical distances of the points from the least squares line.
- The residuals give us an idea how close a prediction might be when the least squares line is used to make a prediction for a value that is not included in the data set.

Problem Set

The time spent in surgery and the cost of surgery was recorded for six patients. The results and scatter plot are shown below.

Time (minutes)	Cost ($)
14	1,510
80	6,178
84	5,912
118	9,184
149	8,855
192	11,023



1. Calculate the equation of the least squares line relating cost to time. (Indicate slope to the nearest tenth and y-intercept to the nearest whole number.)

2. Draw the least squares line on the graph above. (Hint: Substitute $x = 30$ into your equation to find the predicted y-value. Plot the point $(30,$ your answer$)$ on the graph. Then substitute $x = 180$ into the equation and plot the point. Join the two points with a straightedge.)

3. What does the least squares line predict for the cost of a surgery that lasts 118 minutes? (Calculate the cost to the nearest cent.)

4. How much do you have to add to your answer to question 3 to get the actual cost of surgery for a surgery lasting 118 minutes? (This is the residual.)

5. Show your answer to question 4 as a vertical line between the point for that person in the scatter plot and the least squares line.

6. Remember that the residual is the actual y-value minus the predicted y-value. Calculate the residual for the surgery that took 149 minutes and cost $\$8{,}855$.

7. Calculate the other residuals, and write all the residuals in the table below.

Time (minutes)	Cost ($)	Predicted Value ($)	Residual
14	1,510		
80	6,178		
84	5,912		
118	9,184		
149	8,855		
192	11,023		

8. Suppose that a surgery took 100 minutes.
   a. What does the least squares line predict for the cost of this surgery?
   b. Would you be surprised if the actual cost of this surgery were $\$9{,}000$? Why, or why not?
   c. Interpret the slope of the least squares line.

Lesson 16: More on Modeling Relationships with a Line

Classwork

Example 1: Calculating Residuals

The curb weight of a car is the weight of the car without luggage or passengers. The table below shows the curb weights (in hundreds of pounds) and fuel efficiencies (in miles per gallon) of five compact cars.

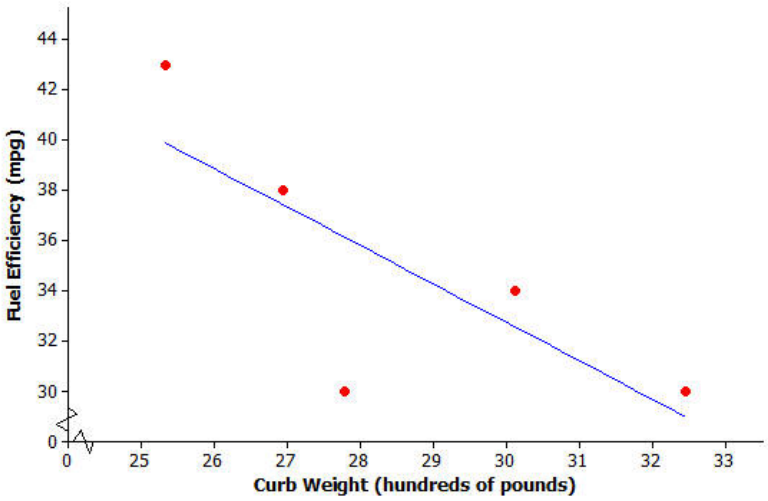
Curb Weight (100 lb.)	Fuel Efficiency (miles per gallon)
25.33	43
26.94	38
27.79	30
30.12	34
32.47	30

Using a calculator, the least squares line for this data set was found to have the equation:

$$y = 78.62 - 1.5290x,$$

where x is the curb weight (in hundreds of pounds) and y is the predicted fuel efficiency (in miles per gallon).

The scatter plot of this data set is shown below, and the least squares line is shown on the graph.



You will calculate the residuals for the five points in the scatter plot. Before calculating the residual, look at the scatter plot.

Exercises 1–2

1. Will the residual for the car whose curb weight is 25.33 be positive or negative? Roughly what is the value of the residual for this point?

2. Will the residual for the car whose curb weight is 27.79 be positive or negative? Roughly what is the value of the residual for this point?

The residuals for both of these curb weights are calculated as follows:

Substitute $x = 25.33$ into the equation of the least squares line to find the predicted fuel efficiency. $y = 78.62 - 1.5290(25.33)$ $= 39.9$ Now calculate the residual. residual $=$ actual y-value $-$ predicted y-value $= 43 - 39.9$ $= 3.1$ mpg	Substitute $x = 27.79$ into the equation of the least squares line to find the predicted fuel efficiency. $y = 78.62 - 1.5290(27.79)$ $= 36.1$ Now calculate the residual. residual $=$ actual y-value $-$ predicted y-value $= 30 - 36.1$ $= -6.1$ mpg

These two residuals have been written in the table below.

Curb Weight (100 lb)	Fuel Efficiency (miles per gallon)	Residual
25.33	43	3.1
26.94	38	
27.79	30	-6.1
30.12	34	
32.47	30	

Exercises 3–4

Continue to think about the car weights and fuel efficiencies from Example 1.

3. Calculate the remaining three residuals and write them in the table.

4. Suppose that a car has a curb weight (in hundreds of pounds) of 31.
   a. What does the least squares line predict for the fuel efficiency of this car?

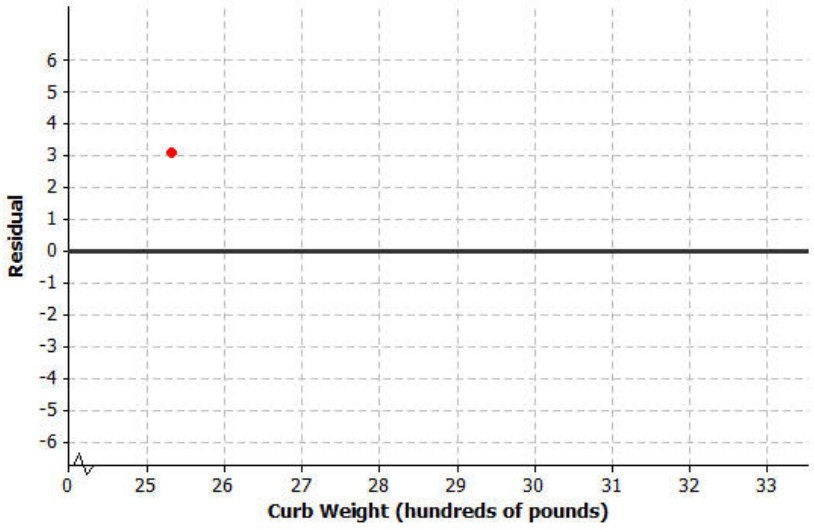
   b. Would you be surprised if the actual fuel efficiency of this car was 29 miles per gallon? Explain your answer.

Example 2: Making a Residual Plot to Evaluate a Line

It is often useful to make a graph of the residuals, called a residual plot. You will make the residual plot for the compact car data set.

Plot the original x variable (curb weight in this case) on the horizontal axis and the residuals on the vertical axis. For this example, you need to draw a horizontal axis that goes from 25 to 32 and a vertical axis with a scale that includes the values of the residuals that you calculated. Next, plot the point for the first car. The curb weight of the first car is 25.33 and the residual is 3.1. Plot the point $(25.33, 3.1)$.

The axes and this first point are shown below.



Exercise 5–6

5. Plot the other four residuals in the residual plot started in Example 3.

6. How does the pattern of the points in the residual plot relate to pattern in the original scatter plot? Looking at the original scatter plot, could you have known what the pattern in the residual plot would be?

Lesson Summary

- The predicted y-value is calculated using the equation of the least squares line.
- The residual is calculated using

$$\text{residual} = \text{actual } y\text{-value} - \text{predicted } y\text{-value}.$$

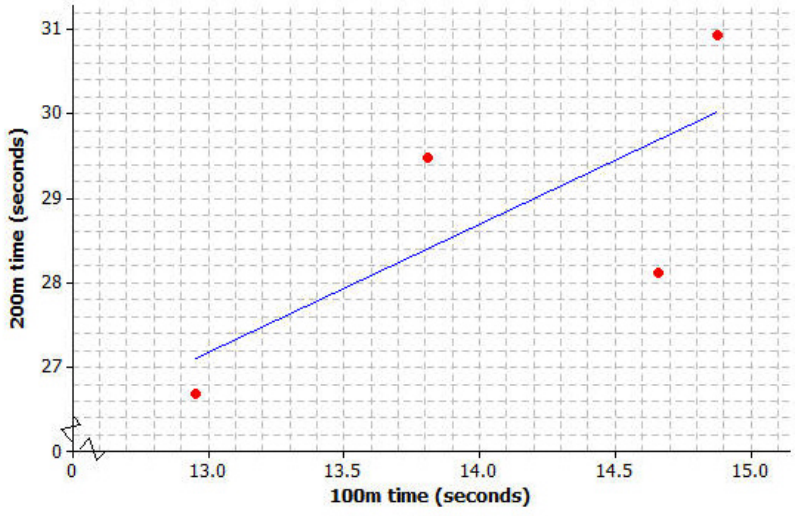
- The sum of the residuals provides an idea of the degree of accuracy when using the least squares line to make predictions.
- To make a residual plot, plot the x-values on the horizontal axis and the residuals on the vertical axis.

Problem Set

Four athletes on a track team are comparing their personal bests in the 100- and 200-meter events. A table of their best times is shown below.

Athlete	100 m time (seconds)	200 m time (seconds)
1	12.95	26.68
2	13.81	29.48
3	14.66	28.11
4	14.88	30.93

A scatter plot of these results (including the least squares line) is shown below.



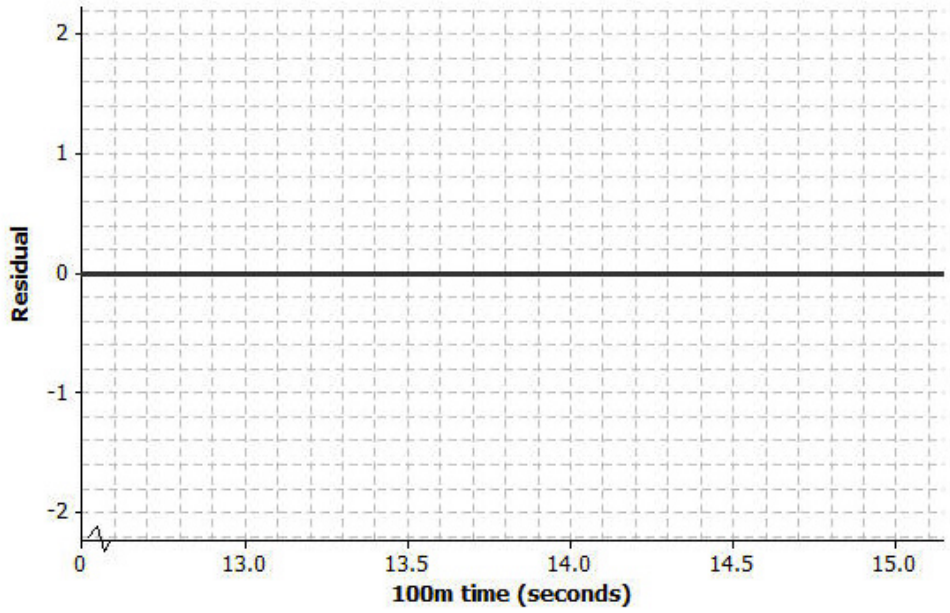
1. Use your calculator or computer to find the equation of the least squares line.

2. Use your equation to find the predicted 200-meter time for the runner whose 100-meter time is 12.95. What is the residual for this athlete?

3. Calculate the residuals for the other three athletes. Write all the residuals in the table given below.

Athlete	100 m time (seconds)	200 m time (seconds)	Residual
1	12.95	26.68	
2	13.81	29.48	
3	14.66	28.11	
4	14.88	30.93	

4. Using the axes provided below, construct a residual plot for this data set.

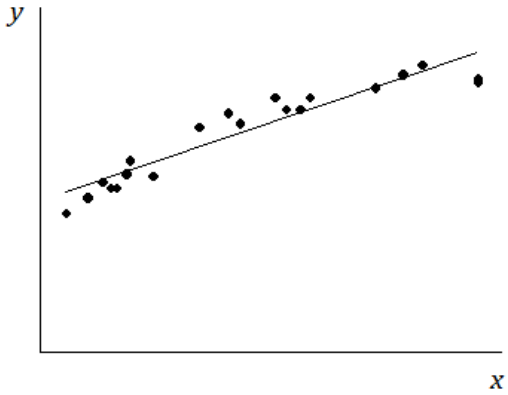


Lesson 17: Analyzing Residuals

Classwork

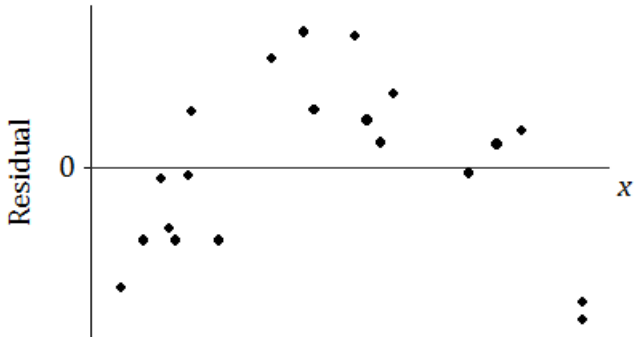
Example 1: Predicting the Pattern in the Residual Plot

Suppose you are given a scatter plot and least squares line that looks like this:



Describe what you think the residual plot would look like.

The residual plot has an arch shape, like this:

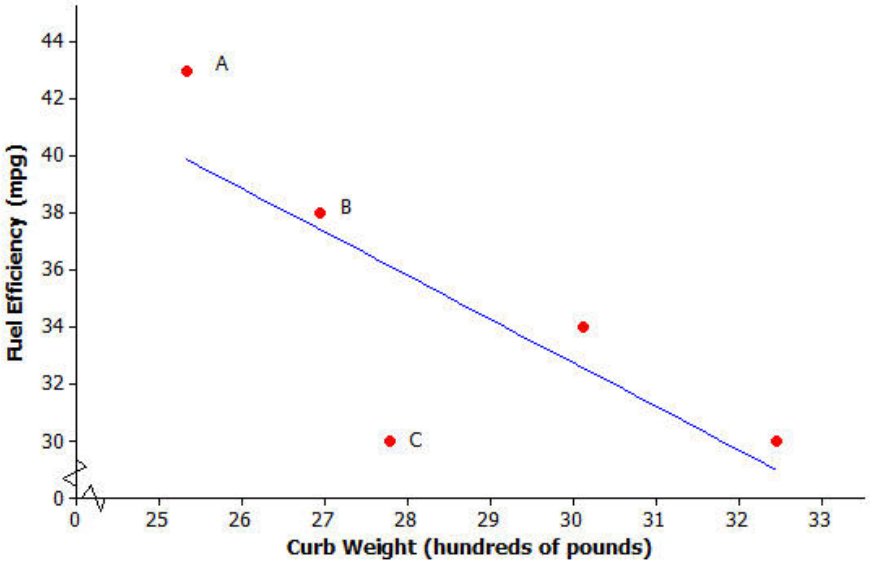


Why is looking at the pattern in the residual plot important?

Example 2: The Meaning of Residuals

Suppose that you have a scatter plot and that you have drawn the least squares line on your plot. Remember that the residual for a point in the scatter plot is the vertical distance of that point from the least squares line.

In the previous lesson, you looked at a scatter plot showing how fuel efficiency was related to curb weight for five compact cars. The scatter plot and least squares line are shown below.



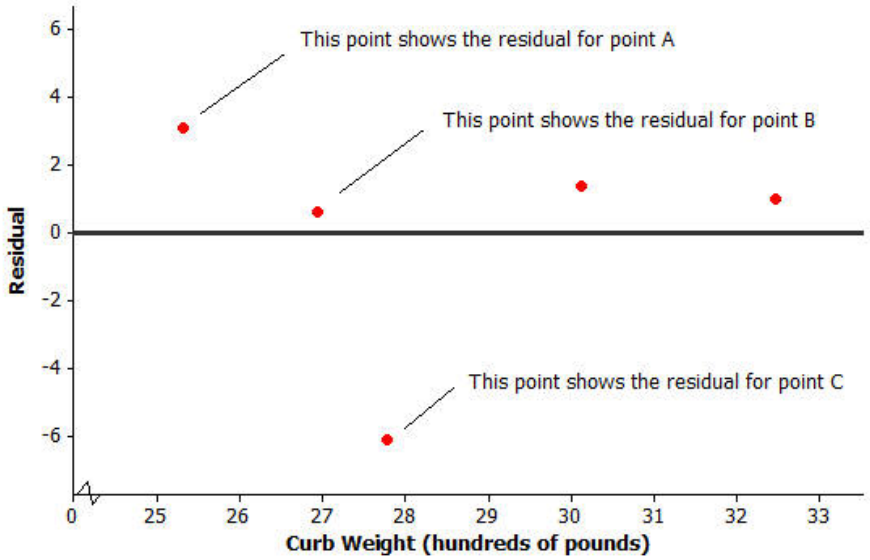
Consider the following questions:

- What kind of residual will Point A have?

- What kind of residual will Point B have?

- What kind of residual will Point C have?

You also looked at the residual plot for this data set:



Your teacher will now show how to use a graphing calculator or graphing program to construct a scatter plot and a residual plot. Consider the following exercise.

Example 3: Using a Graphing Calculator to Construct a Residual Plot

In an earlier lesson you looked at a data set giving the shoe lengths and heights of 12 adult women. This data set is shown in the table below.

Shoe Length (x)	Height (y)
inches	inches
8.9	61
9.6	61
9.8	66
10.0	64
10.2	64
10.4	65
10.6	65
10.6	67
10.5	66
10.8	67
11.0	67
11.8	70

Use a calculator to construct the scatter plot (with least squares line) and the residual plot for this data set.

Lesson Summary

- After fitting a line, the residual plot can be constructed using a graphing calculator.
- A pattern in the residual plot indicates that the relationship in the original data set is not linear.

Problem Set

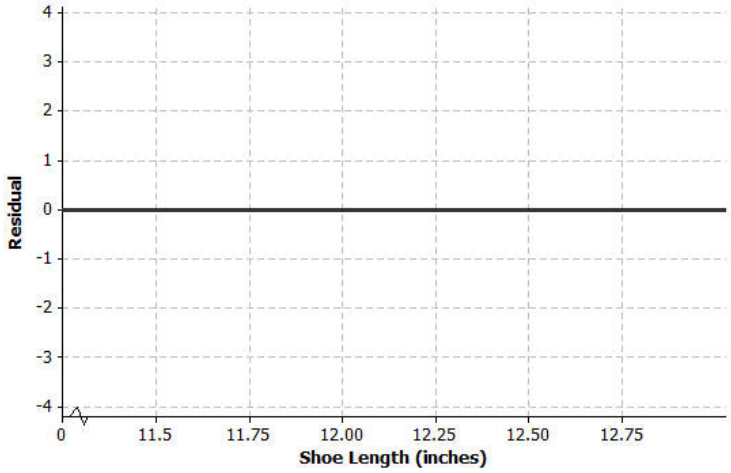
Consider again a data set giving the shoe lengths and heights of 10 adult men. This data set is shown in the table below.

Shoe Length (x)	Height (y)
inches	inches
12.6	74
11.8	65
12.2	71
11.6	67
12.2	69
11.4	68
12.8	70
12.2	69
12.6	72
11.8	71

1. Use your calculator or graphing program to construct the scatter plot of this data set. Include the least squares line on your graph. Explain what the slope of the least squares line indicates about shoe length and height.

2. Use your calculator to construct the residual plot for this data set.

3. Make a sketch of the residual plot on the axes given below. Does the scatter of points in the residual plot indicate a linear relationship in the original data set? Explain your answer.



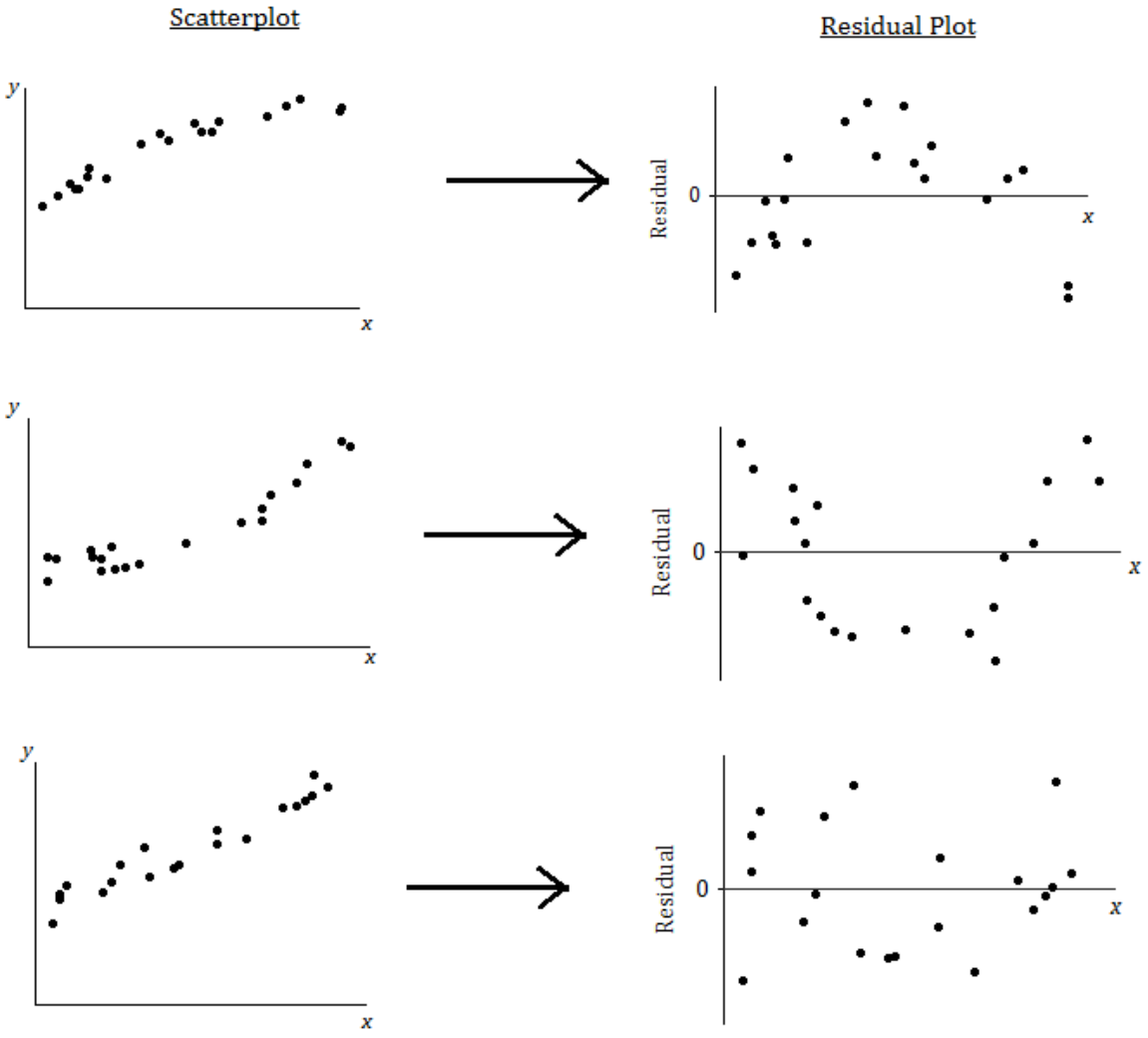
Lesson 18: Analyzing Residuals

Classwork

The previous lesson shows that when data is fitted to a line, a scatter plot with a curved pattern produces a residual plot that shows a clear pattern. You also saw that when a line is fit, a scatter plot where the points show a straight-line pattern results in a residual plot where the points are randomly scattered.

Example 1: The Relevance of the Pattern in the Residual Plot

Our previous findings are summarized in the plots below:



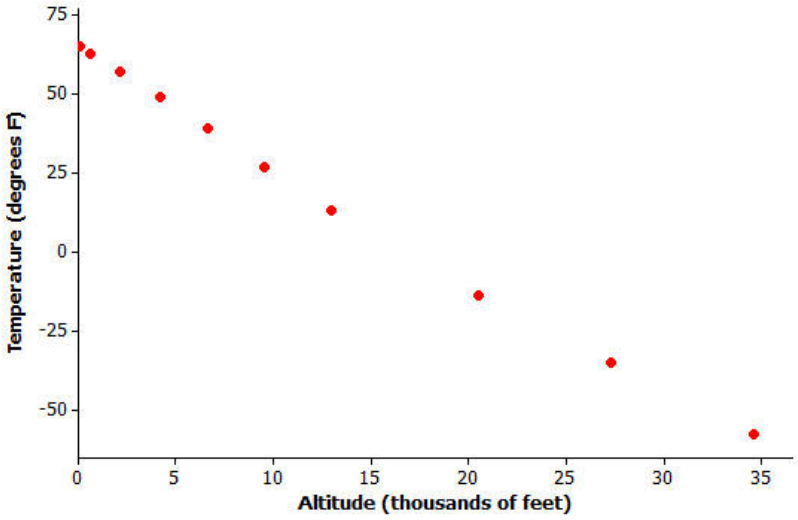
What does it mean when there is a curved pattern in the residual plot?

What does it mean when the points in the residual plot appear to be scattered at random with no visible pattern?

Why not just look at the scatter plot of the original data set? Why was the residual plot necessary? The next example answers these questions.

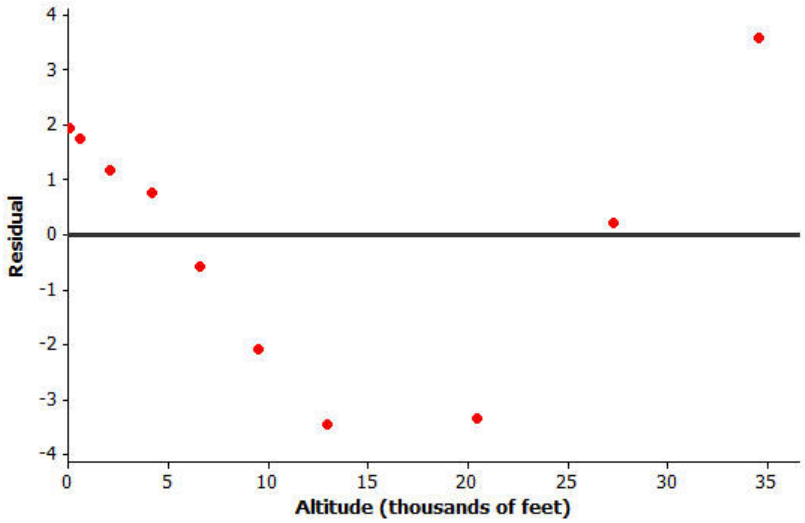
Example 2: Why Do You Need the Residual Plot?

The temperature (in degrees Fahrenheit) was measured at various altitudes (in thousands of feet) above Los Angeles. The scatter plot (below) seems to show a linear (straight line) relationship between these two quantities.



Data source: *Core Math Tools,* www.nctm.org

However, look at the residual plot:



There is a clear curve in the residual plot. So what appeared to be a linear relationship in the original scatter plot was, in fact, a nonlinear relationship.

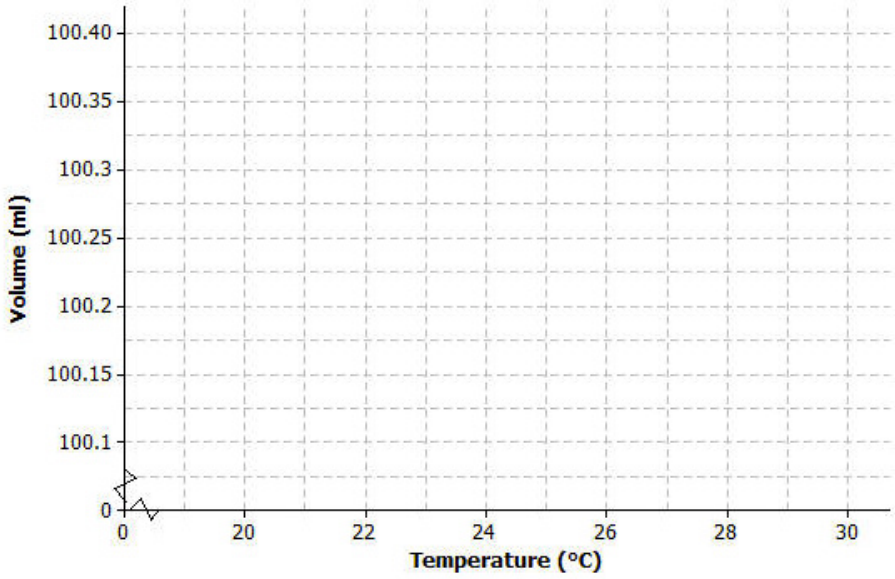
How did this residual plot result from the original scatter plot?

Exercises 1–3: Volume and Temperature

Water expands as it heats. Researchers measured the volume (in milliliters) of water at various temperatures. The results are shown below.

Temperature (°C)	Volume (ml)
20	100.125
21	100.145
22	100.170
23	100.191
24	100.215
25	100.239
26	100.266
27	100.290
28	100.319
29	100.345
30	100.374

1. Using a graphing calculator, construct the scatter plot of this data set. Include the least squares line on your graph. Make a sketch of the scatter plot including the least-squares line on the axes below.



2. Using the calculator, construct a residual plot for this data set. Make a sketch of the residual plot on the axes given below.

3. Do you see a clear curve in the residual plot? What does this say about the original data set?

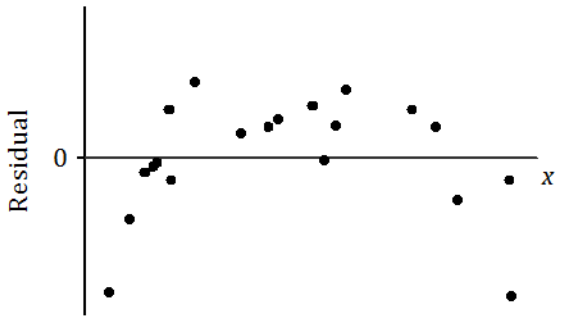
Lesson Summary

- After fitting a line, the residual plot can be constructed using a graphing calculator.
- A curve or pattern in the residual plot indicates a nonlinear relationship in the original data set.
- A random scatter of points in the residual plot indicates a linear relationship in the original data set.

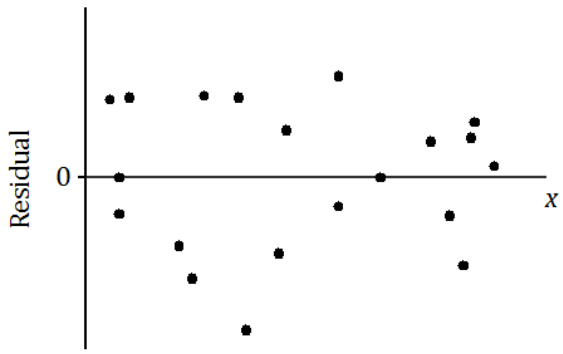
Problem Set

1. For each of the following residual plots, what conclusion would you reach about the relationship between the variables in the original data set? Indicate whether the values would be better represented by a linear or a nonlinear relationship.

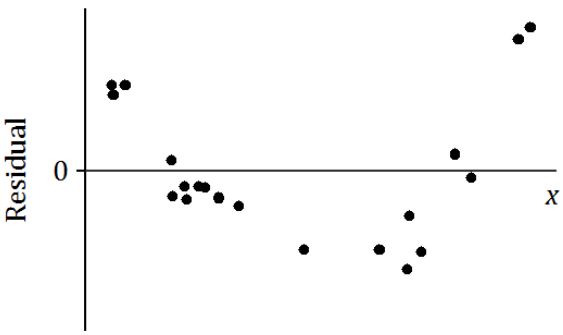
a.



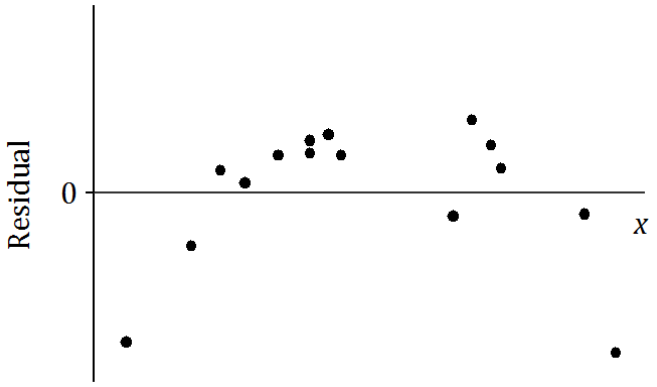
b.



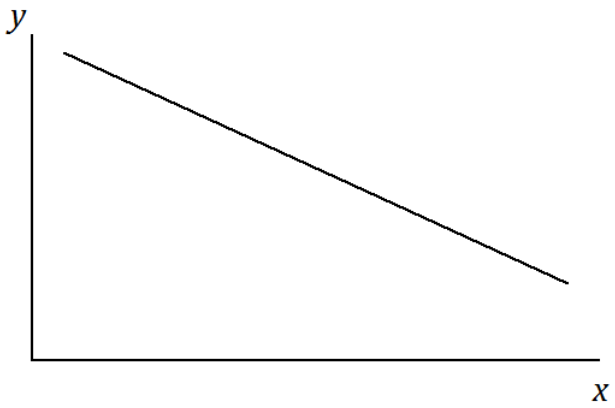
c.



2. Suppose that after fitting a line, a data set produces the residual plot shown below.



An incomplete scatter plot of the original data set is shown below. The least squares line is shown, but the points in the scatter plot have been erased. Estimate the locations of the original points and create an approximation of the scatter plot below.

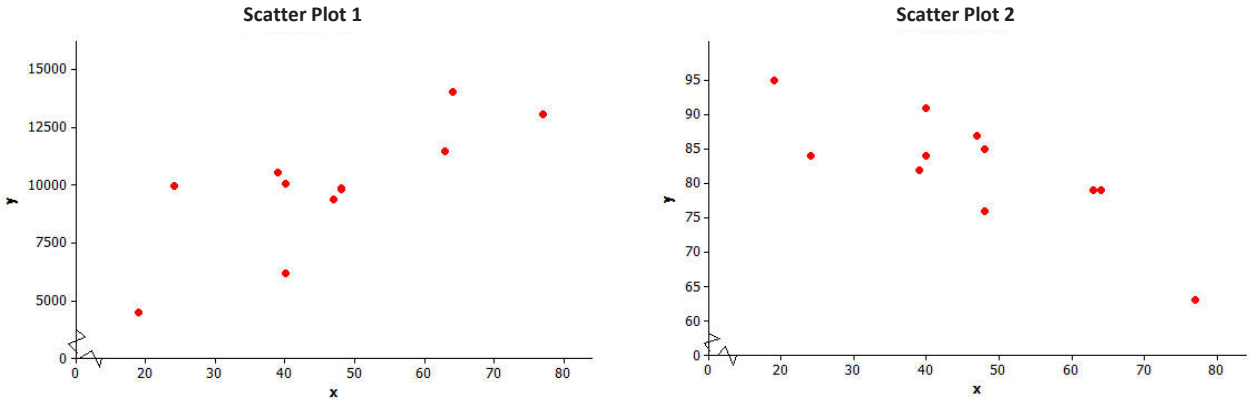


Lesson 19: Interpreting Correlation

Classwork

Example 1: Positive and Negative Linear Relationships

Linear relationships can be described as either positive or negative. Below are two scatter plots that display a linear relationship between two numerical variables x and y.



Exercises 1–4

1. The relationship displayed in scatter plot 1 is a positive linear relationship. Does the value of the y-variable tend to increase or decrease as the value of x increases? If you were to describe this relationship using a line, would the line have a positive or negative slope?

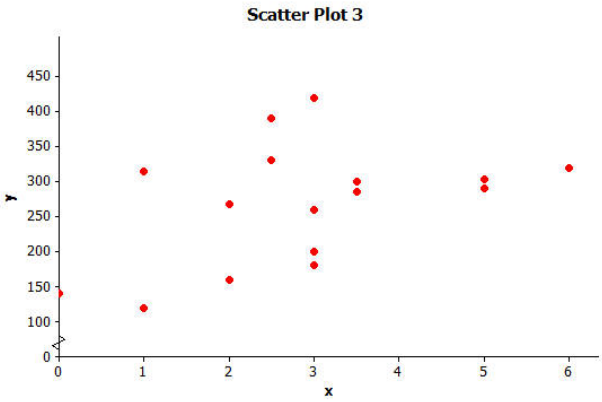
2. The relationship displayed in scatter plot 2 is a negative linear relationship. As the value of one of the variables increases, what happens to the value of the other variable? If you were to describe this relationship using a line, would the line have a positive or negative slope?

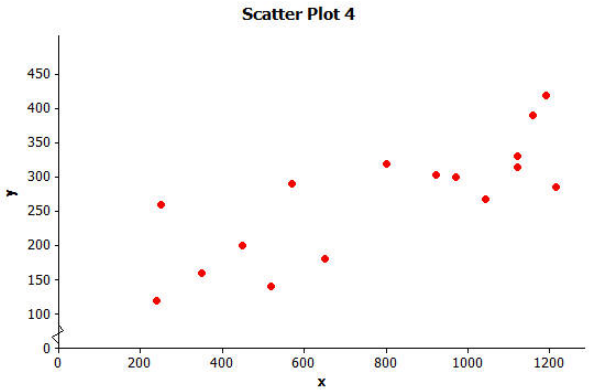
3. What does it mean to say that there is a positive linear relationship between two variables?

4. What does it mean to say that there is a negative linear relationship between two variables?

Example 2: Some Linear Relationships Are Stronger than Others

Below are two scatter plots that show a linear relationship between two numerical variables x and y.





Exercises 5–9

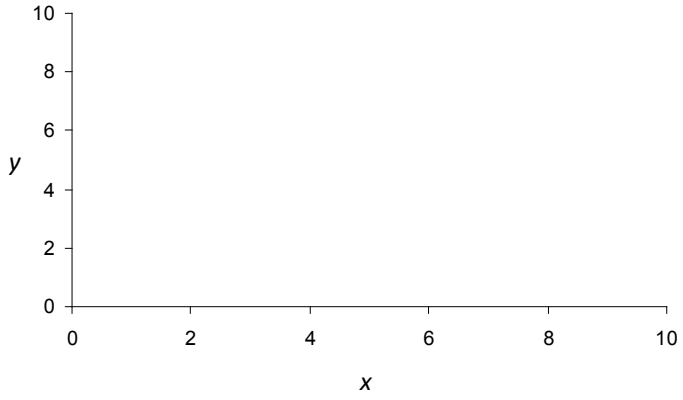
5. Is the linear relationship in scatter plot 3 positive or negative?

6. Is the linear relationship in scatter plot 4 positive or negative?

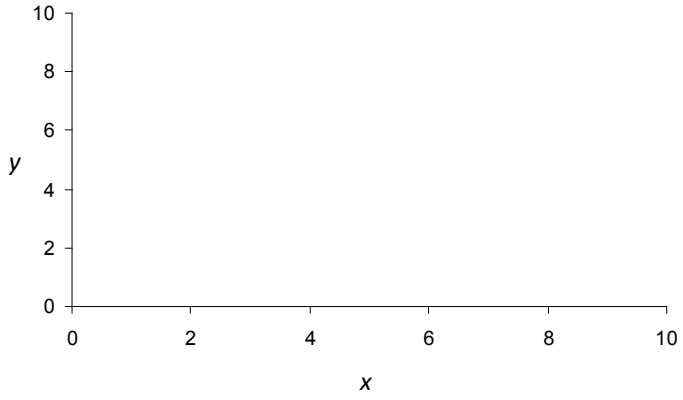
It is also common to describe the strength of a linear relationship. We would say that the linear relationship in scatter plot 3 is weaker than the linear relationship in scatter plot 4.

7. Why do you think the linear relationship in scatter plot 3 is considered weaker than the linear relationship in scatter plot 4?

8. What do you think a scatter plot with the strongest possible linear relationship might look like if it is a positive relationship? Draw a scatter plot with five points that illustrates this.



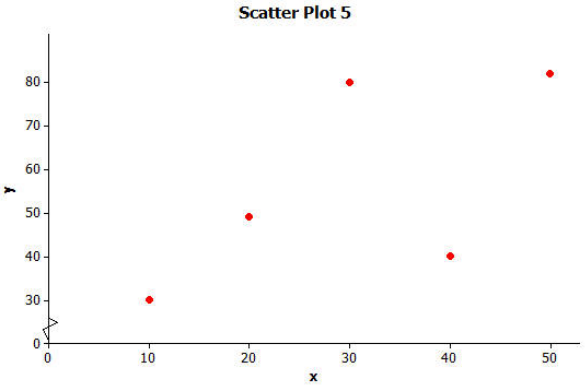
9. How would a scatter plot that shows the strongest possible linear relationship that is negative look different from the scatter plot that you drew in the previous question?

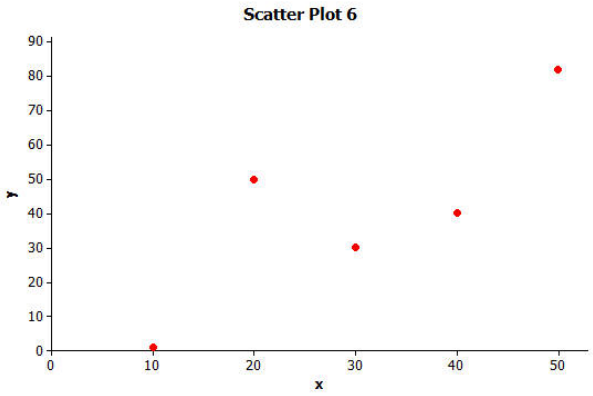


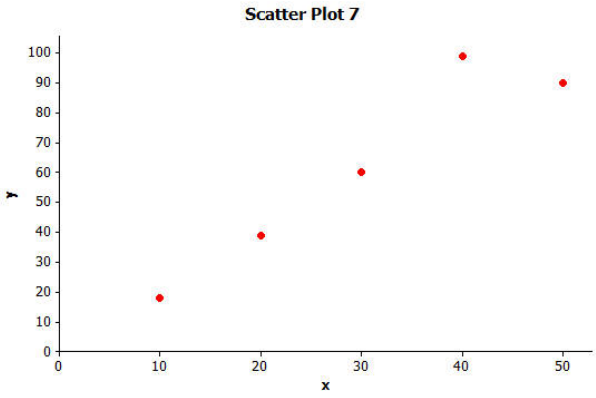
Exercises 10–12: Strength of Linear Relationships

10. Consider the three scatter plots below. Place them in order from the one that shows the strongest linear relationship to the one that shows the weakest linear relationship.

Strongest		Weakest

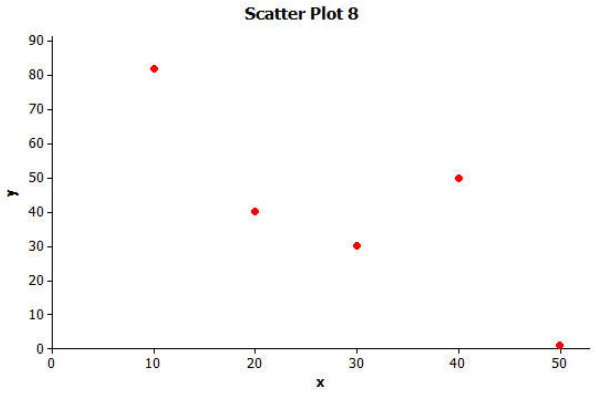


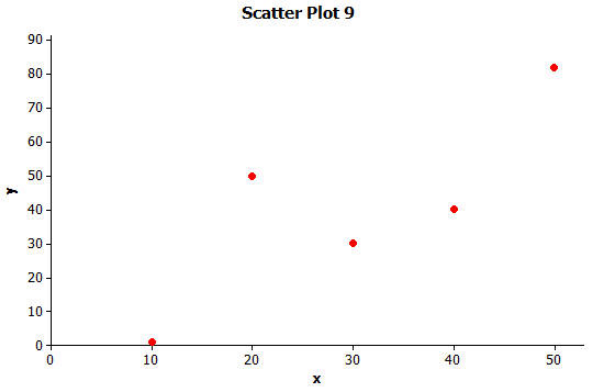




11. Explain your reasoning for choosing the order in Exercise 10.

12. Which of the following two scatter plots shows the stronger linear relationship? (Think carefully about this one!)

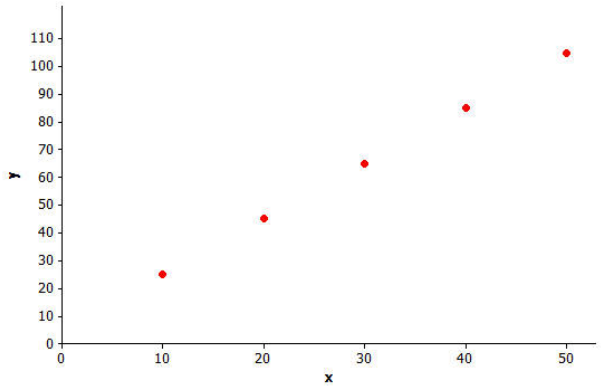


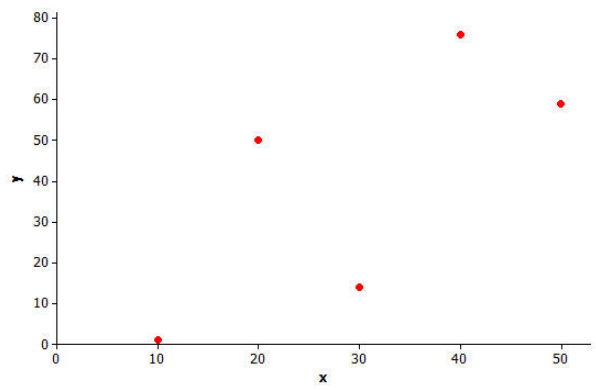
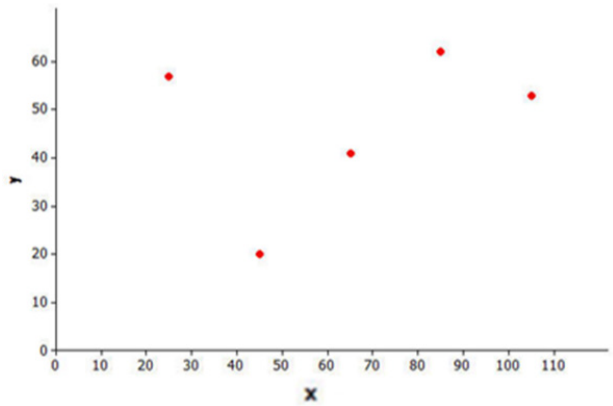
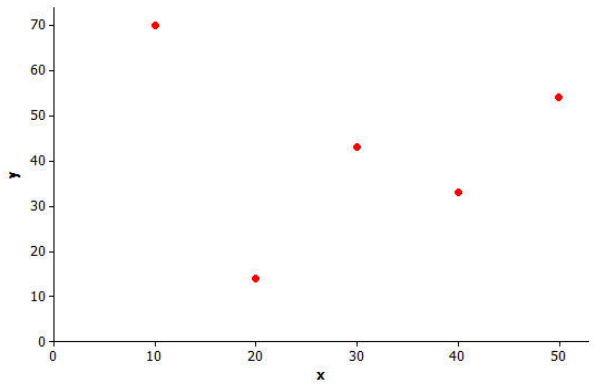


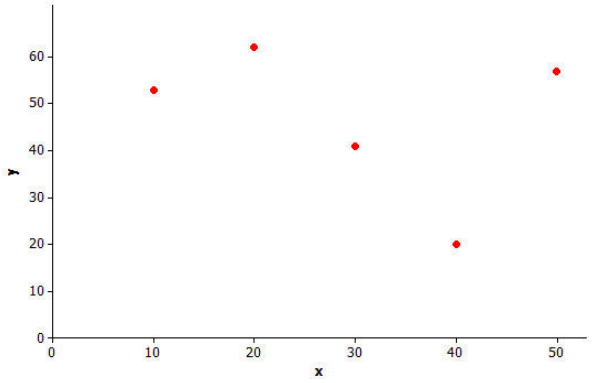
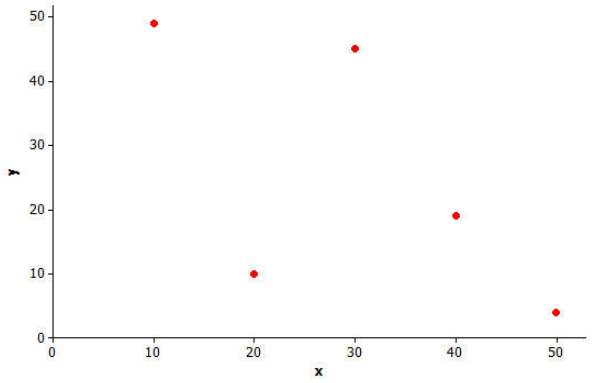
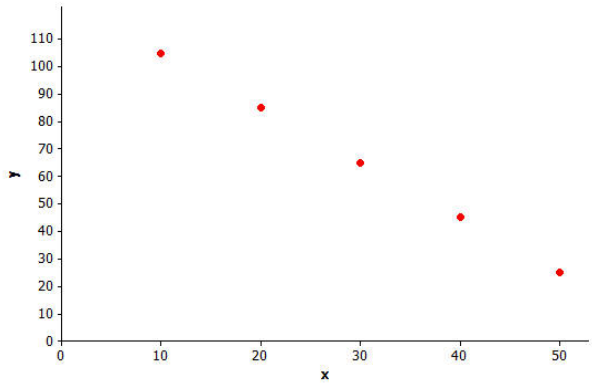
Example 3: The Correlation Coefficient

The **correlation coefficient** is a number between -1 and $+1$ (including -1 and $+1$) that measures the strength and direction of a linear relationship. The correlation coefficient is denoted by the letter r.

Several scatter plots are shown below. The value of the correlation coefficient for the data displayed in each plot is also given.

	$r = 1.00$

	$r = 0.71$
	$r = 0.32$
	$r = -0.10$

	$r = -0.32$
	$r = -0.63$
	$r = -1.00$

Exercises 13–15

13. When is the value of the correlation coefficient positive?

14. When is the value of the correlation coefficient negative?

15. Is the linear relationship stronger when the correlation coefficient is closer to 0 or to 1 (or -1)?

Looking at the scatter plots in Example 4, you should have discovered the following properties of the correlation coefficient:

Property 1: The sign of r (positive or negative) corresponds to the direction of the linear relationship.

Property 2: A value of $r = +1$ indicates a perfect positive linear relationship, with all points in the scatter plot falling exactly on a straight line.

Property 3: A value of $r = -1$ indicates a perfect negative linear relationship, with all points in the scatter plot falling exactly on a straight line.

Property 4: The closer the value of r is to $+1$ or -1, the stronger the linear relationship.

Example 4: Calculating the Value of the Correlation Coefficient

There is an equation that can be used to calculate the value of the correlation coefficient given data on two numerical variables. Using this formula requires a lot of tedious calculations that will be discussed in later grades. Fortunately, a graphing calculator can be used to find the value of the correlation coefficient once you have entered the data.

Your teacher will show you how to enter data and how to use a graphing calculator to obtain the value of the correlation coefficient.

Here is the data from a previous lesson on shoe length in inches and height in inches for 10 men.

Shoe Length (x) (inches)	**Height (y)** (inches)
12.6	74
11.8	65
12.2	71
11.6	67
12.2	69
11.4	68
12.8	70
12.2	69
12.6	72
11.8	71

Exercises 16–17

16. Enter the shoe length and height data in your calculator. Find the value of the correlation coefficient between shoe length and height. Round to the nearest tenth.

The table below shows how you can informally interpret the value of a correlation coefficient.

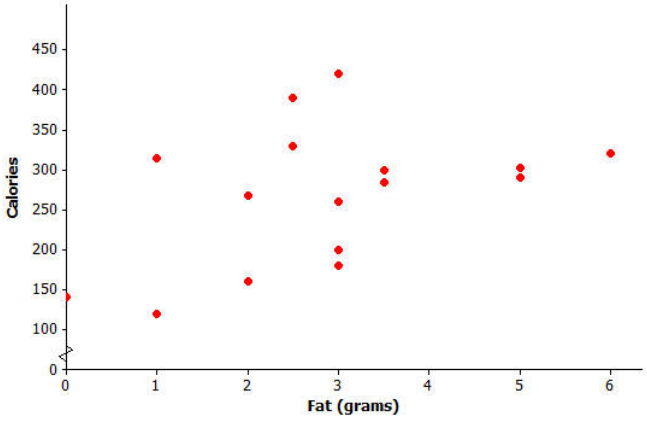
If the value of the correlation coefficient is between …	You can say that …
$r = 1.0$	There is a perfect positive linear relationship.
$0.7 \leq r < 1.0$	There is a strong positive linear relationship.
$0.3 \leq r < 0.7$	There is a moderate positive linear relationship.
$0 < r < 0.3$	There is a weak positive linear relationship.
$r = 0$	There is no linear relationship.
$-0.3 < r < 0$	There is a weak negative linear relationship.
$-0.7 < r \leq -0.3$	There is a moderate negative linear relationship.
$-1.0 < r \leq -0.7$	There is a strong negative linear relationship.
$r = -1.0$	There is a perfect negative linear relationship.

17. Interpret the value of the correlation coefficient between shoe length and height for the data given above.

Exercises 18–24: Practice Calculating and Interpreting Correlation Coefficients

Consumer Reports published a study of fast-food items. The table and scatter plot below display the fat content (in grams) and number of calories per serving for sixteen fast-food items.

Fat (g)	Calories (kcal)
2	268
5	303
3	260
3.5	300
1	315
2	160
3	200
6	320
3	420
5	290
3.5	285
2.5	390
0	140
2.5	330
1	120
3	180



Data Source: *Consumer Reports*

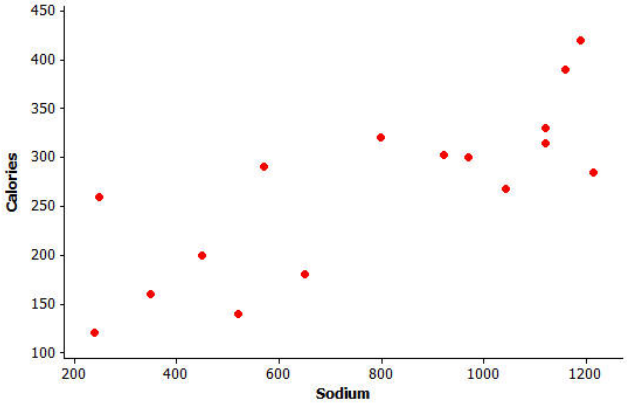
18. Based on the scatter plot, do you think that the value of the correlation coefficient between fat content and calories per serving will be positive or negative? Explain why you made this choice.

19. Based on the scatter plot, estimate the value of the correlation coefficient between fat content and calories.

20. Calculate the value of the correlation coefficient between fat content and calories per serving. Round to the nearest hundredth. Interpret this value.

The *Consumer Reports* study also collected data on sodium content (in mg) and number of calories per serving for the same sixteen fast food items. The data is represented in the table and scatter plot below.

Sodium (mg)	Calories (kcal)
1042	268
921	303
250	260
970	300
1120	315
350	160
450	200
800	320
1190	420
570	290
1215	285
1160	390
520	140
1120	330
240	120
650	180



21. Based on the scatter plot, do you think that the value of the correlation coefficient between sodium content and calories per serving will be positive or negative? Explain why you made this choice.

22. Based on the scatter plot, estimate the value of the correlation coefficient between sodium content and calories per serving.

23. Calculate the value of the correlation coefficient between sodium content and calories per serving. Round to the nearest hundredth. Interpret this value.

24. For these sixteen fast-food items, is the linear relationship between fat content and number of calories stronger or weaker than the linear relationship between sodium content and number of calories? Does this surprise you? Explain why or why not.

Example 5: Correlation Does Not Mean There is a Cause-and-Effect Relationship Between Variables

It is sometimes tempting to conclude that if there is a strong linear relationship between two variables that one variable is causing the value of the other variable to increase or decrease. But you should avoid making this mistake. When there is a strong linear relationship, it means that the two variables tend to vary together in a predictable way, which might be due to something other than a cause-and-effect relationship.

For example, the value of the correlation coefficient between sodium content and number of calories for the fast food items in the previous example was $r = 0.79$, indicating a strong positive relationship. This means that the items with higher sodium content tend to have a higher number of calories. But the high number of calories is not caused by the high sodium content. In fact sodium does not have any calories. What may be happening is that food items with high sodium content also may be the items that are high in sugar or fat, and this is the reason for the higher number of calories in these items.

Similarly, there is a strong positive correlation between shoe size and reading ability in children. But it would be silly to think that having big feet causes children to read better. It just means that the two variables vary together in a predictable way. Can you think of a reason that might explain why children with larger feet also tend to score higher on reading tests?

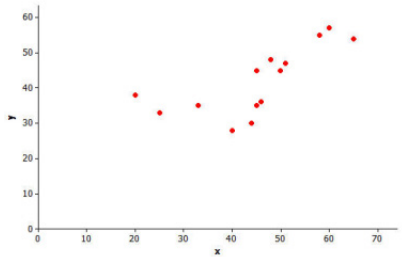
Lesson Summary

- Linear relationships are often described in terms of strength and direction.
- The correlation coefficient is a measure of the strength and direction of a linear relationship.
- The closer the value of the correlation coefficient is to $+1$ or -1, the stronger the linear relationship.
- Just because there is a strong correlation between the two variables does not mean there is a cause-and-effect relationship.

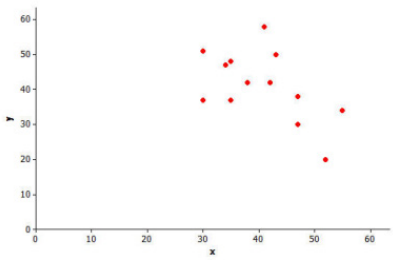
Problem Set

1. Which of the three scatter plots below shows the strongest linear relationship? Which shows the weakest linear relationship?

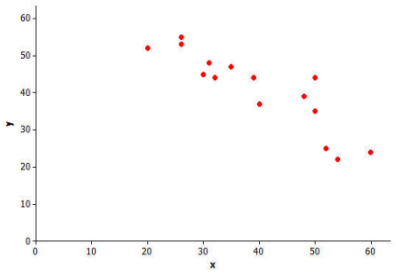
Scatter Plot 1



Scatter Plot 2

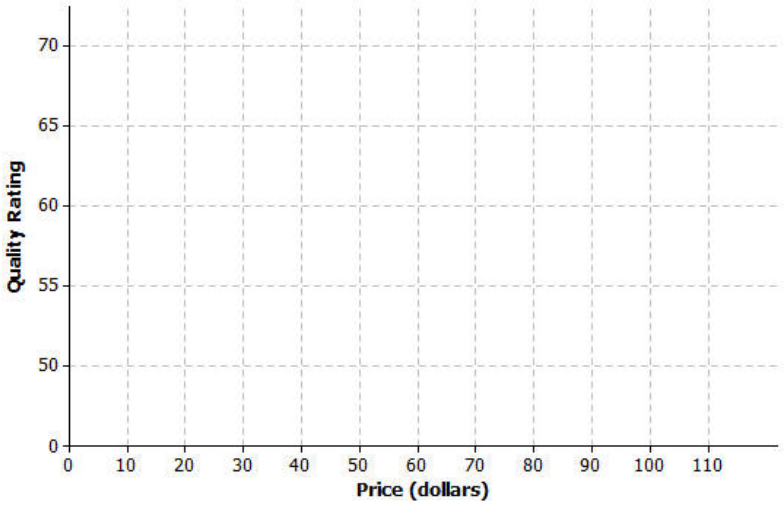


Scatter Plot 3



2. *Consumer Reports* published data on the price (in dollars) and quality rating (on a scale of 0 to 100) for 10 different brands of men's athletic shoes.

Price ($)	Quality Rating
65	71
45	70
45	62
80	59
110	58
110	57
30	56
80	52
110	51
70	51



a. Construct a scatter plot of these data using the grid provided.

b. Calculate the value of the correlation coefficient between price and quality rating and interpret this value. Round to the nearest hundredth.

c. Does it surprise you that the value of the correlation coefficient is negative? Explain why or why not.

d. Is it reasonable to conclude that higher priced shoes are higher quality? Explain.

e. The correlation between price and quality rating is negative. Is it reasonable to conclude that increasing the price causes a decrease in quality rating? Explain.

3. *The Princeton Review* publishes information about colleges and universities. The data below are for six public 4-year colleges in New York. Graduation rate is the percentage of students who graduate within six years. Student-to-faculty ratio is the number of students per full-time faculty member.

School	Number of Full-Time Students	Student-to-Faculty Ratio	Graduation Rate
CUNY Bernard M Baruch College	11,477	17	63
CUNY Brooklyn College	9,876	15.3	48
CUNY City College	10,047	13.1	40
SUNY at Albany	14,013	19.5	64
SUNY at Binghamton	13,031	20	77
SUNY College at Buffalo	9,398	14.1	47

a. Calculate the value of the correlation coefficient between the number of full-time students and graduation rate. Round to the nearest hundredth.

b. Is the linear relationship between graduation rate and number of full-time students weak, moderate or strong? On what did you base your decision?

c. True or False? Based on the value of the correlation coefficient, it is reasonable to conclude that having a larger number of students at a school is the cause of a higher graduation rate.

d. Calculate the value of the correlation coefficient between the student-to-faculty ratio and graduation rate. Round to the nearest hundredth.

e. Which linear relationship is stronger: graduation rate and number of full-time students or graduation rate and student-to-faculty ratio? Justify your choice.

Lesson 20: Analyzing Data Collected on Two Variables

Classwork

Lessons 12–19 included several data sets that were used to learn about how two numerical variables might be related. Recall the data on elevation above sea level and the number of clear days per year for fourteen cities in the United States. Could a city's elevation above sea level by used to predict the number of clear days per year a city experiences? After observing a scatter plot of the data, a linear model (or the least-squares linear model obtained from a calculator or computer software) provided a reasonable description of the relationship between these two variables. This linear model was evaluated by considering how close the data points were to the corresponding graph of the line. The equation of the linear model was used to answer the statistical question about elevation and the number of clear days.

Several data sets were also provided to illustrate other possible models, specifically quadratic and exponential models. Finding a model that describes the relationship between two variables allows you to answer statistical questions about how the two numerical variables vary. For example, the following statistical question was posed in Lesson 13 regarding latitude and the mean number of flycatcher chicks in a nest: What latitude is best for hatching flycatcher chicks? A quadratic model of the latitude and the mean number of flycatcher chicks in a nest was used to answer this statistical question.

In Lessons 12–19, you worked with several data sets and models used to answer statistical questions. Select one of the data sets presented in Lessons 12–19, and develop a poster that summarizes how a statistical question is answered that involves two numerical variables. Your poster should include the following: a brief summary of the data, the statistical question asking the relationship between two numerical variables, a scatter plot of the data, and a brief summary to indicate how well the data fit a specific model.

After you identify one of the problems to summarize with a poster, consider the following questions to plan your poster:

1. What two variables were involved in this problem?

2. What was the statistical question? Remember, a statistical question involves data. The question also anticipates that the data will vary.

3. What model was used to describe the relationship between the two variables?

4. Based on the scatter plot of the data, was the model a good one?

5. Was the residual plot used to evaluate the model? What did the residual plot indicate about the model?

6. How would you use the model to predict values not included in the data set?

7. Does the model answer the statistical question posed?

Examples of posters involving two numerical variables can be found at the website of the American Statistical Association (www.amstat.org/education/posterprojects/index.cfm).