Investigation 4

Variability and Associations in Numerical Data

When playing basketball, it helps to be tall and to have long arms. The average player in the National Basketball Association is more than 6 feet 7 inches tall.

- How rare do you think it is for a man to be as tall as those average NBA players?
- Do you think height and arm span are closely related variables for NBA players?
- Do you think height and arm span are closely related variables for students in your class?

Working on the Problems in this Investigation will help you understand how to measure variability and associations of data values.

Common Core State Standards

8.SP.A.1 Construct and interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association.

8.SP.A.2 . . . For scatter plots that suggest a linear association, informally fit a straight line, and informally assess the model fit by judging the closeness of the data points to the line.

8.SP.A.3 Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept.

Also 8.EE.B.5, 8.EE.C.7, 8.F.A.1, 8.F.A.3, 8.F.B.4, 8.F.B.5, S.ID.A.2, S.ID.B.6, S.ID.B.6b, and S.ID.C.7

4.1 Vitruvian Man

Relating Body Measurements

More than 2,000 years ago, a Roman architect and writer named Vitruvius found patterns by relating two body measurements. He claimed a person's arm span is equal to his or her height.

- Do you think the relationship between arm span and height applies to the students in your class?
- How would you display and analyze data collected to test the claim made by Vitruvius?

Problem 4.1

The table shows the height and arm span of students in a CMP class.

Height (in.)	56	57	57	58	59	60	60	60	62	64	64	66	67	67	67	68
Arm span (in.)	54	57	54	61	56	58	59	60	62	63	62	62	65	67	69	67

Do you think the data support the claim that arm span and height are about equal?

A Analyze the data to test your ideas.

1. Plot the (height, arm span) data on a coordinate graph. The resulting graph is called a **scatter plot**.
2. Do you think the scatter plot supports the claim that arm span and height are about equal for most people?

continued on the next page >

Problem 4.1 continued

3. If each student in the class had arm span s equal to height h, what equation would relate the two variables?

   a. Graph the equation on your scatter plot.

   b. Which data points (if any) does your line pass through? Explain how arm span and height are related in those points.

   c. Choose several data points that are not on your line. Explain how arm span and height are related in each case. How do you describe the relationship shown on the graph?

B The tallest person in recorded history was Robert Pershing Wadlow. At age 22, he was 8 feet 11.1 inches (272 cm) tall. His arm span was 9 feet 5.75 inches (289 cm).

1. Where would you plot the point (height, arm span) for Robert Wadlow? Would the point be *on*, *above*, or *below* the line you drew in Question A, part (3)?

2. Does the data point for Robert Wadlow support the claim that arm span and height are roughly equal?

C The accuracy of fit for a linear model is measured by calculating errors from the model. These errors, called **residuals,** are the differences between the actual data and what the model predicts.

1. Find the arm span residuals (actual arm span − predicted arm span) using the model $s = h$ for the CMP class data.

Height (inches)	56	57	57	58	59	60	60	60	62	64	64	66	67	67	67	68
Arm span (inches)																
Actual	54	57	54	61	56	58	59	60	62	63	62	62	65	67	69	67
Predicted by Model	56	57	57	58	59	60	60	60	62	64	64	66	67	67	67	68
Residual	−2	0	−3	■	■	■	■	■	■	■	■	■	■	■	■	■

2. Describe the pattern of residuals. Do you think the equation $s = h$ is an accurate model for predicting arm span from height?

continued on the next page >

Problem 4.1 continued

D The dinosaur *Tyrannosaurus rex* grew to 20 feet in height with an arm span of about 10 feet.

1. Do you think the *T. rex* data point fits the pattern that arm span and height are roughly equal? Explain.
2. If you plot the data point, would it be *on*, *above*, or *below* the line you drew in Question A, part (4)?

ACE Homework starts on page 96.

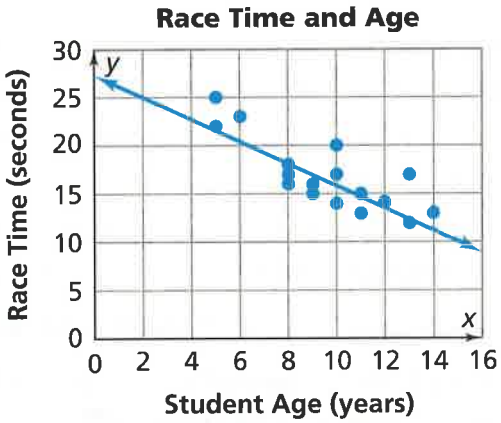
4.2 Older and Faster

Negative Correlation

Magnolia Elementary is a school with students who are 5 to 14 years old. One field day, all students were timed in a 100-meter race. The table shows data for some of the students.

Student Age (years)	5	5	6	8	8	8	9	9	10	10	10	11	11	12	13	13	14
Race Time (seconds)	25	22	23	18	16	17	15	16	17	20	14	15	13	14	17	12	13

The graph below shows the data from the table and a line that models the data.



- How would you describe the relationship between age and race time?
- Would you say the relationship is *strong* or *weak*?
- Are the data points close to the line or spread out?

Problem 4.2

Use the Race Time and Age graph.

A The line drawn on the graph models the relationship between age and race time.

1. What is the approximate slope of the line?
2. How does the slope help you understand the relationship between age and race time?
3. Do you think it makes sense to predict a race time for a 7-year-old student using the line? If so, what do you predict for a 7-year-old? How confident are you in your prediction?
4. Do you think it makes sense to predict a race time for a 21-year-old person using the line? If so, what do you predict for a 21-year-old? How confident are you in your prediction?

continued on the next page >

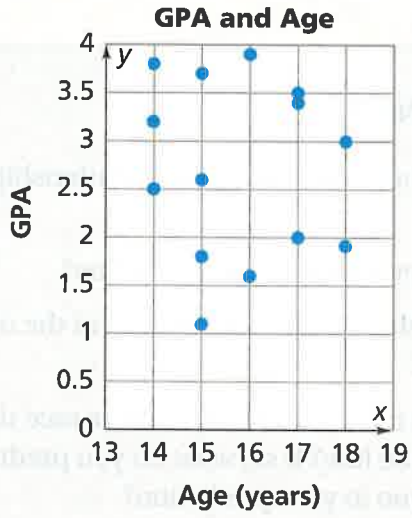
Problem 4.2 continued

B Some data points are very close to the line while others are far from it. The points far from the line don't seem to fit the model.

1. Find two points that don't seem to fit the model. What are their coordinates (age, race time)?
2. Why do you think the points don't match the overall pattern? Explain. Think about the relationship between race time and age.
3. In Problem 4.1, you used a line to model (height, arm span).
   a. If a 6-foot-9-inch NBA basketball player has a 7-foot-5-inch arm span, would that data point fit the model?
   b. Would you plot the data point, *on*, *above*, or *below* the $s = h$ line? Explain.

C The table and graph show age and grade point average (GPA) for 14 students at Magnolia High School.

Student Age (years)	14	14	14	15	15	15	15	16	16	17	17	17	18	18
GPA	2.5	3.2	3.8	1.8	2.6	3.7	1.2	1.6	3.9	2.0	3.4	3.5	1.9	3.0



1. Are age and GPA strongly related for these students? Explain.
2. How is your answer to part (1) supported by the table?
3. How is your answer to part (1) supported by the scatter plot?

ACE Homework starts on page 96.

4.3 Correlation Coefficients and Outliers

Roller coasters are popular rides at amusement parks. A recent survey counted 1,797 roller coaster rides in the world. 734 of them are in North America. Roller coasters differ in maximum drop, maximum height, track length, ride time, and coaster type (wood or steel).

Which roller coaster variables do you think are strongly related to the top speed on the ride?

Problem 4.3

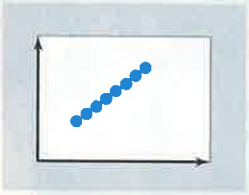
Statisticians measure the strength of a linear relationship between two variables using a number called the **correlation coefficient.** This number is a decimal between -1 and 1. When the points lie close to a straight line, the correlation coefficient is close to -1 or 1.

continued on the next page >

Problem 4.3 continued

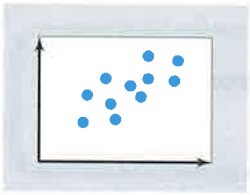
- When points cluster close to a line with positive slope, the correlation coefficient is almost 1, and with negative slope, the correlation coefficient is almost −1.
- Points that do not cluster close to any line have a correlation coefficient of almost 0.
- Positive association has correlation coefficients greater than 0 while negative association has correlation coefficients less than 0.

A **1.** The graph below has a correlation coefficient of 1.0. What do you think a correlation coefficient of 1.0 means?

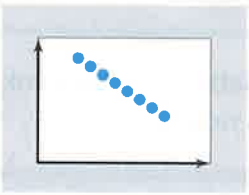


2. Which of the six scatter plots below (a)-(f) has a correlation coefficient of −1.0? What do you think a correlation coefficient of −1.0 means?

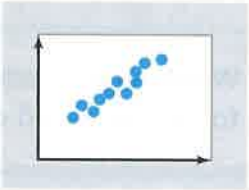
a.



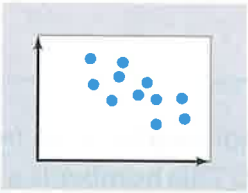
b.



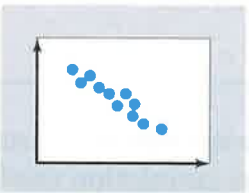
c.



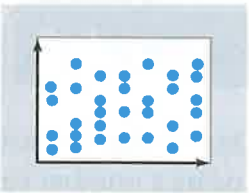
d.



e.



f.



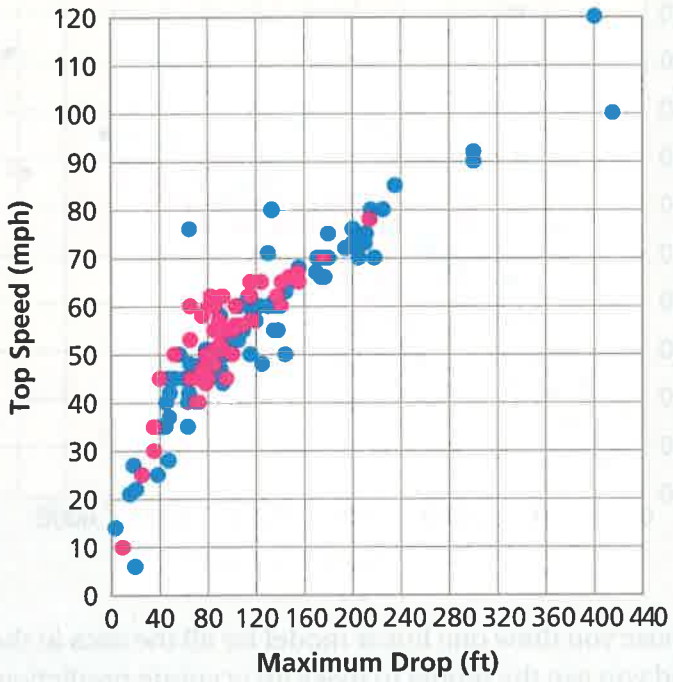
3. Match correlation coefficients −0.8, −0.4, 0.0, 0.4, and 0.8 with the other five scatter plots. Explain your reasoning.

continued on the next page >

Problem 4.3 continued

When you inspect a scatter plot, often you are looking for a strong association between the variables.

B The scatter plot below shows the relationship between the top speed of a roller coaster and its maximum drop. The pink dots represent wood-frame roller coasters. The blue dots represent steel-frame coasters.

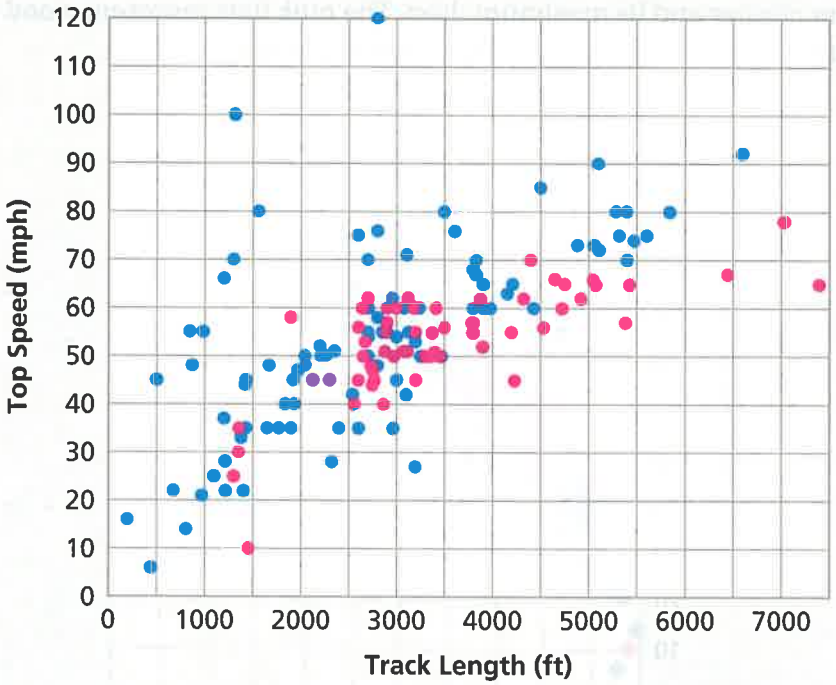


1. Suppose you drew one linear model for all the data in the graph. Could you use the model to make an accurate prediction about the top speed of the roller coaster with a given maximum drop? Explain.
2. Estimate the correlation coefficient for the top speed and the maximum drop. Is the correlation coefficient closest to -1, -0.5, 0, 0.5, or 1?

continued on the next page >

Problem 4.3 continued

C The scatter plot below shows the relationship between the top speed of a roller coaster and its track length. The pink dots represent wood-frame roller coasters. The blue dots represent steel-frame coasters.

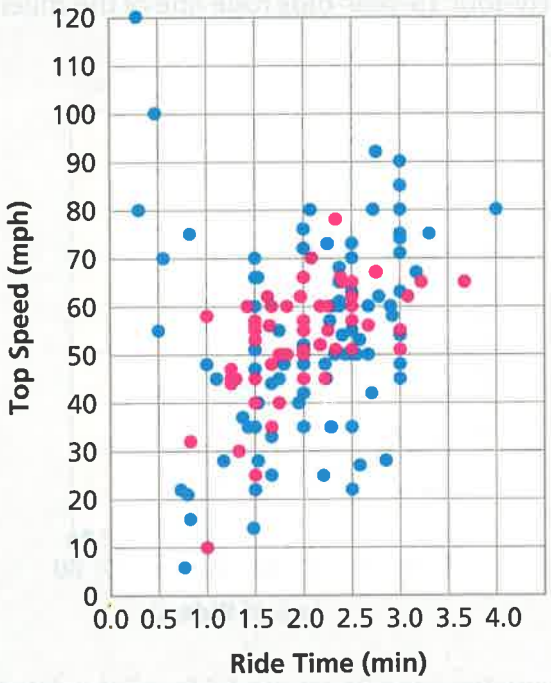


1. Suppose you drew one linear model for all the data in the graph. Could you use the model to make an accurate prediction about the top speed of the roller coaster with a given track length? Explain.
2. Estimate the correlation coefficient for the top speed and track length. Is the correlation coefficient closest to -1, -0.5, 0, 0.5, or 1?

continued on the next page >

Problem 4.3 continued

D The scatter plot below shows the relationship between the top speed of a roller coaster and the ride time. The pink dots represent wood-frame roller coasters. The blue dots represent steel-frame coasters.



1. Suppose you drew one linear model for all the data in the graph. Could you use the model to make an accurate prediction about the top speed of the roller coaster with a given ride time? Explain.
2. Estimate the correlation coefficient for the top speed and ride time. Is the correlation coefficient closest to -1, -0.5, 0, 0.5, or 1?
3. Suppose most of the points on a scatter plot cluster near a line, with only a few that don't fit the pattern. The points that lie outside a cluster are called **outliers.** Use the graph above. Find each point. Then decide whether the point is an outlier. If it is, explain why you think it is an outlier.

   a. (1.75, 50) **b.** (0.30, 80) **c.** (3.35, 75)

   d. (0.28, 120) **e.** (0.80, 21) **f.** (1.0, 10)

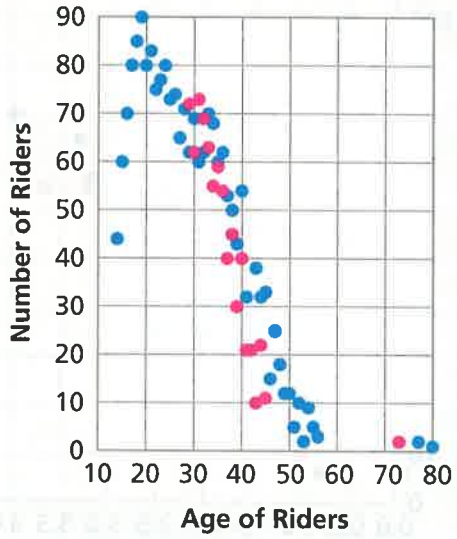
   g. Use the scatter plot in Question C. Find two outliers on that graph and estimate their coordinates (track length, top speed).

continued on the next page >

Problem 4.3 continued

E The scatter plot shows the number of roller coaster riders and their ages on a given day. The pink dots represent wood-frame roller coasters. The blue dots represent steel-frame coasters.

On that day, forty-four 15-year-olds rode one of the roller coasters. The data point is (15, 44).



1. Suppose you drew one linear model for all the data in the graph. Could you use the model to make an accurate prediction about the number of riders on the roller coaster with a given age? Explain.
2. Estimate the correlation coefficient for the number of riders and age of riders. Is the correlation coefficient closest to -1, -0.5, 0, 0.5, or 1?
3. Are any of the data points outliers? If so, estimate the coordinates of those points.

F Is it possible to have a correlation coefficient close to -1 or 1 with only a few outliers? Explain your thinking.

ACE Homework starts on page 96.

4.4 Measuring Variability
Standard Deviation

A height of 6 feet 7 inches is unusual for an adult man.

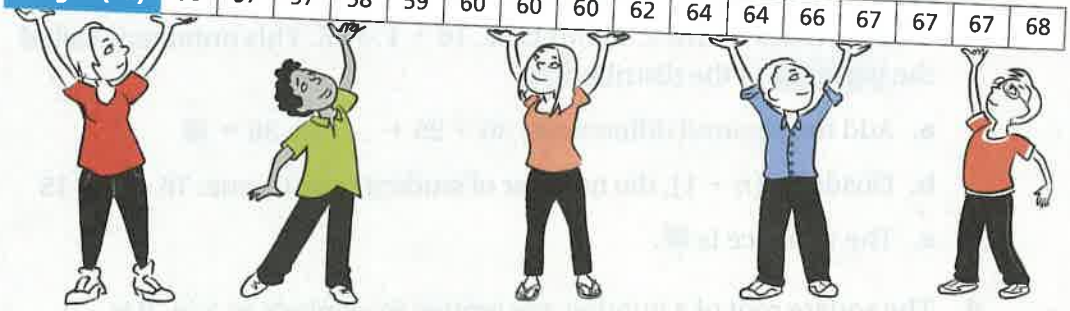
What height would make an eighth-grade boy or girl above average?

You can use range and interquartile range to describe how data values in a sample vary. You can also use the mean absolute deviation (MAD) to measure the spread of data values. This problem reviews those measures and introduces a measure of spread called *standard deviation*.

Problem 4.4

The table shows the heights of several CMP students. You used this information in Problem 4.1.

Height (in.)	56	57	57	58	59	60	60	60	62	64	64	66	67	67	67	68



A Make a line plot to show the distribution of the data.

B Calculate the summary statistics below, and explain what each number says about the distribution of heights.

1. Range
2. Mean
3. Mean Absolute Deviation (MAD)

continued on the next page >

Problem 4.4 continued

C Like the MAD, you calculate the **standard deviation** of a data set from the differences between data values and the mean. To calculate the standard deviation for the height data, complete each part below.

1. Find the difference of each data value and the mean. In the table below, for example, Jayne's height is 56 inches and the mean is 62 inches. The difference is $(56 - 62) = -6$. Copy the table and complete the middle row with the differences.

2. Square each difference. For example, for Jayne's height, $(-6)^2 = 36$. Complete the third row with the squares of the differences.

Jayne's height

Height (inches)	56	57	57	58	59	60	60	60	62	64	64	66	67	67	67	68
Height − Mean	−6	−5	−5	−4	■	■	■	■	■	■	■	■	■	■	■	■
Squares of differences	36	25	■	■	■	■	■	■	■	■	■	■	■	■	■	■

Square the difference.

3. Next, sum the squared differences and divide by $(n - 1)$, the number of data values minus 1. In this case, $16 - 1 = 15$. This number is called the **variance** of the distribution.

   a. Add the squared differences: $36 + 25 + \ldots + 36 =$ ■

   b. Divide by $(n - 1)$, the number of students minus one: $16 - 1 = 15$

   c. The variance is ■.

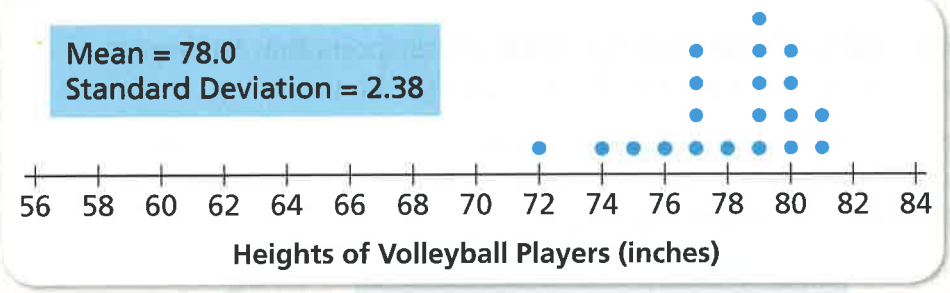
4. The square root of a number n is written in symbols as $\sqrt{n}$. It is the positive number you multiply by itself to equal n. For example, $\sqrt{25} = 5$ and $\sqrt{6.25} = 2.5$.

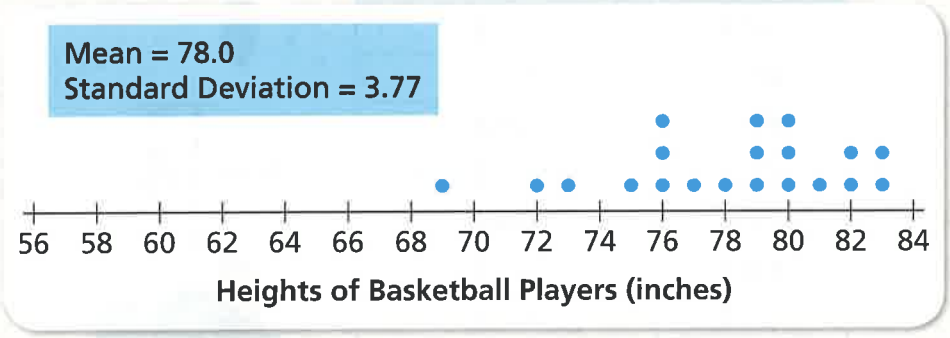
   Take the square root of the variance. This number is the standard deviation of the distribution of heights.

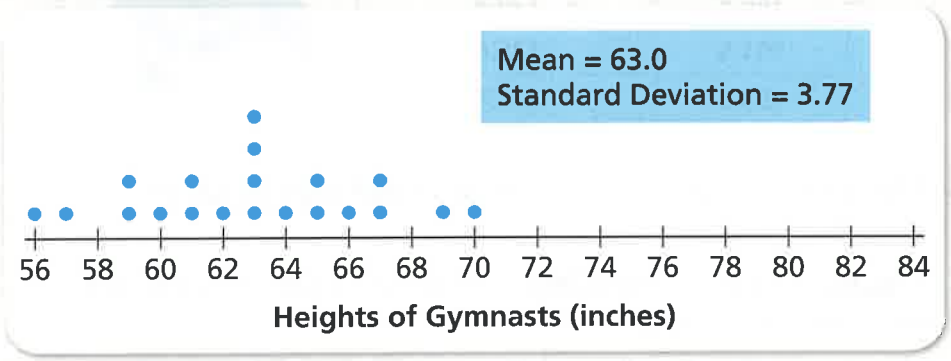
continued on the next page >

Problem 4.4 continued

D Each dot plot shows the distribution, mean, and standard deviation of heights of 20 athletes. The 20 athletes are a random sample.







1. Compare the heights of volleyball players with the heights of basketball players. What can you say about the similarities and differences using the dot plots?
2. Compare the gymnasts with the basketball players. What can you say about the similarities and differences using the dot plots?

ACE Homework starts on page 96.

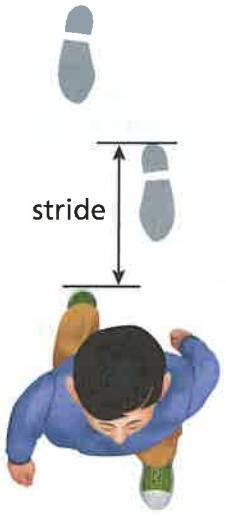
Applications

For Exercises 1 and 2, use the table below. It shows the height and stride distance for 10 students.

For humans, walking is the most basic form of transportation. An average person is able to walk at a pace of about 3 miles per hour.

The distance a person covers in one step depends on his or her stride. To measure stride distance, measure from the heel of the first foot to the heel of that same foot on the next step.

Height (cm)	Stride Distance (cm)
150.8	125.2
149.5	124.2
151.2	125.2
153.1	126.8
150.6	124.4
149.9	123.8
146.5	121.8
146.5	120.8
151.5	125.6
153.5	126.8



1. a. What is the median height of these students? Explain how you found the median.

b. What is the median stride distance of these students? Explain how you found the median.

c. What is the ratio of median height to median stride distance? Explain.