EOC Review- Unit 4: Statistics Book Pages

20.3 Finding Data Missing From the Mean (DOK 2)

Example 3: Mara knew she had an 88 average in her biology class, but she lost one of her papers. The three papers she could find had scores of 98%, 84%, and 90%. What was the score on her fourth paper?

Step 1: Calculate the total score on four papers with an 88% average. $.88 \times 4 = 3.52$

Step 2: Add together the scores from the three papers you have. $.98 + .84 + .9 = 2.72$

Step 3: Subtract the scores you know from the total score. $3.52 - 2.72 = .80$. She had 80% on her fourth paper.

Find the data missing from the following problems. (DOK 2)

1. Gabriel earns 87% on his first geography test. He wants to keep a 92% average. What does he need to get on his next test to bring his average up?

2. Rian earned $68.00 on Monday. How much money must she earn on Tuesday to have an average of $80 earned for the two days?

3. Haley, Chuck, Dana, and Chris enter a contest to see who could bake the most chocolate chip cookies in an hour. They bake an average of 75 cookies. Haley bakes 55, Chuck bakes 70, and Dana bakes 90. How many does Chris bake?

4. Four wrestlers make a pact to lose some weight before the competition. They lose an average of 7 pounds each over the course of 3 weeks. Carlos loses 6 pounds, Steve loses 5 pounds, and Greg loses 9 pounds. How many pounds does Wes lose?

5. Three boxes are ready for shipment. The boxes average 26 pounds each. The first box weighs 30 pounds; the second box weighs 25 pounds. How much does the third box weigh?

6. The five jockeys running in the next race average 92 pounds each. Nicole weighs 89 pounds. Jon weighs 95 pounds. Jenny and Kasey weigh 90 pounds each. How much does Jordan weigh?

7. Jessica makes three loaves of bread that weigh a total of 45 ounces. What is the average weight of each loaf?

8. Celeste makes scented candles to give away to friends. She has 2 pounds of candle wax which she melted, scented, and poured into 8 molds. What is the average weight of each candle?

9. Each basketball player has to average a minimum of 5 points a game for the next three games to stay on the team. Ben is feeling the pressure. He scored 3 points the first game and 2 points the second game. How many points does he need to score in the third game to stay on the team?

20.5 Standard Deviation (DOK 2)

Standard deviation is the measure of the variability (or spread) of the values in a data set. It can be applied to a population or a smaller, sample data set. The formulas for standard deviation for a set $\{x_1, x_2, ..., x_n\}$ is

Sample Standard Deviation

$$S = \sqrt{\frac{1}{N-1}\left(\sum_{i=1}^{N}(x_i - \overline{x})^2\right)}$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N}(x_i - \overline{x})^2\right)}$$

where x_i are the elements in the set $\{x_1, x_2, ..., x_n\}$ and $\overline{x}$ is the mean of the data set.

Example 6: A randomly-selected group of 6 automobiles was formed, and the odometer reading was taken for each car. The number of miles on each was as follows:

Car 1: 23,000 miles Car 3: 120,000 miles Car 5: 27,000 miles
Car 2: 44,000 miles Car 4: 31,000 miles Car 6: 19,000 miles

What is the standard deviation of this data sample?

Step 1: Calculate the mean $\overline{x}$ of the data set.

$$\overline{x} = \frac{23,000 + 44,000 + 120,000 + 31,000 + 27,000 + 19,000}{6} = 44,000$$

miles

Step 2: Subtract the mean from each data value. The difference is called a **deviation**.

$x_1 - \overline{x} = 23,000$ miles $- 44,000$ miles $= -21,000$ miles
$x_2 - \overline{x} = 44,000$ miles $- 44,000$ miles $= 0$ miles
$x_3 - \overline{x} = 120,000$ miles $- 44,000$ miles $= 76,000$ miles
$x_4 - \overline{x} = 31,000$ miles $- 44,000$ miles $= -13,000$ miles
$x_5 - \overline{x} = 27,000$ miles $- 44,000$ miles $= -17,000$ miles
$x_6 - \overline{x} = 19,000$ miles $- 44,000$ miles $= -25,000$ miles

Step 3: Now, plug the values into the equation $S = \sqrt{\frac{1}{N-1}\left(\sum_{i=1}^{N}(x_i - \overline{x})^2\right)}$.

$$S = \sqrt{\frac{(x_1-\overline{x})^2 + (x_2-\overline{x})^2 + (x_3-\overline{x})^2 + (x_4-\overline{x})^2 + (x_5-\overline{x})^2 + (x_6-\overline{x})^2}{N-1}}$$

$$= \sqrt{\frac{(-21,000)^2 + (0)^2 + (76,000)^2 + (-13,000)^2 + (-17,000)^2 + (-25,000)^2}{5}}$$

$$= \sqrt{\frac{7,300,000,000}{5}} = \sqrt{1,460,000,000} \approx 38,209.95 \text{ miles}$$

The standard deviation of the data set is 38,209.95 miles.

Example 7: During his short football career, a running back had 7 rushes of 6 yards, 14 yards, 11 yards, 20 yards, 22 yards, 1 yard, and 3 yards, respectively. What is the standard deviation of this set of data?
Hint: This set of data represents the entire population of rushes that the running back had during his career.

Step 1: Calculate the mean of the set of data.

$$\overline{x} = \frac{6+14+11+20+22+1+3}{7} = \frac{77}{7} = 11 \text{ yards}$$

Step 2: Subtract the mean from each data value.

$x_1 - \overline{x} = 6 - 11 = -5$ $\quad x_5 - \overline{x} = 22 - 11 = 11$
$x_2 - \overline{x} = 14 - 11 = 3$ $\quad x_6 - \overline{x} = 1 - 11 = -10$
$x_3 - \overline{x} = 11 - 11 = 0$ $\quad x_7 - \overline{x} = 3 - 11 = -8$
$x_4 - \overline{x} = 20 - 11 = 9$

Step 3: Now, plug the values into the population equation $\sigma = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N}(x_i - \overline{x})^2\right)}$.

$$\sigma = \sqrt{\frac{(-5)^2 + (3)^2 + (0)^2 + (9)^2 + (11)^2 + (-10)^2 + (-8)^2}{7}}$$

$$= \sqrt{\frac{400}{7}} \approx 7.56$$

The standard deviation of the population set is 7.56 yards.

** Remember to use the correct formula for standard deviation. The population standard deviation and sample standard deviation are different from each other. In this case, the population standard deviation is 7.56 yards, but the sample standard deviation is 8.16 yards.

Calculate the standard deviation of each of the following data samples. Use the appropriate unit of measurement in each answer. (DOK 2)

1. 200 pounds, 60 pounds, 150 pounds, 160 pounds, 110 pounds, 90 pounds
2. 7 days, 12 days, 13 days, 2 days, 22 days, 20 days, 15 days, 21 days
3. 33 millimeters, 6 millimeters, 4 millimeters, 5 millimeters, 7 millimeters
4. $32,000, $19,000, $17,000, $29,000, $31,000, $27,000, $25,000
5. 4.5 radians, 1.8 radians, 2.6 radians, 0.9 radians, 3.9 radians, 4.1 radians
6. 6 pints; 9 pints, 7 pints, 8 pints, 6 pints, 5 pints, 8 pints, 4 pints, 6 pints, 5 pints
7. 1900 people, 1400 people, 2100 people, 2300 people, 3300 people, 3000 people
8. 3 goals, 2 goals, 4 goals, 3 goals, 5 goals, 4 goals, 5 goals, 1 goal, 2 goals
9. 44 watts, 29 watts, 51 watts, 49 watts, 33 watts, 22 watts, 53 watts, 57 watts

Calculate the standard deviation of each of the following population sets. (DOK 2)

10. 314, 306, 299, 351, 301, 308, 349
11. 7, 6, 8, 7, 8, 5, 6, 9
12. 89, 76, 82, 83, 93, 74, 79
13. 45, 48, 30, 39, 58, 103
14. 17, 3, 5, 21, 16, 18, 1, 2
15. 43, 34, 54, 49, 66

20.6 Quartiles and Extremes (DOK 1)

In statistics, large sets of data are separated into four equal parts. These parts are called **quartiles.** The **median** separates the date into two halves. Then, the median of the upper half is the **upper quartile**, and the median of the lower half is the **lower quartile**. The distance between the upper quartile and the lower quartile is the **interquartile range**. The interquartile range is sometimes used in the place of range, especially when there are outliers in the data set. Interquartile range is also another type of variability of the data.

The **extremes** are the highest and lowest values in a set of data. The lowest value is called the **lower extreme**, and the highest value is called the **upper extreme.**

Example 8: The following set of data shows the high temperatures (in degrees Fahrenheit) in cities across the United States on a particular autumn day. Find the median, the upper quartile, the lower quartile, the upper extreme, and the lower extreme of the data.

lower half | upper half

15 21 23 25 32 37 40 48 51 53 55 56 57 57 59 60 61 61 65 66 67 69 79

- 15: **lower extreme** (lowest value)
- 37: **lower quartile** (middle data point of lower half)
- 56: **median** (middle data point)
- 61: **upper quartile** (middle data point of upper half)
- 79: **upper extreme** (highest value)

Example 9: The following set of data shows the fastest race car qualifying speeds in miles per hour. Find the medium, the upper quartile, the lower quartile, the upper extreme, and the lower extreme of the data.

lower half | upper half

152 153 153 154 156 158 160 161 163 163 164 164 164 165 165 166

- 152: **lower extreme**
- 155: **lower quartile**
- 162: **median**
- 164: **upper quartile**
- 166: **upper extreme**

Note: When you have an even number of data points, the median is the average of the two middle points. The lower middle number is then included in the lower half of the data, and the upper middle number is included in the upper half.

Find the median, the upper quartile, the lower quartile, the upper extreme, and the lower extreme of each set of data given below. (DOK 1)

1. 0 0 1 1 1 2 2 3 3 4 5
2. 15 16 18 20 22 22 23
3. 62 75 77 80 81 85 87 91 94
4. 74 74 76 76 77 78
5. 3 3 3 5 5 6 6 7 7 7 8 8
6. 190 191 192 192 194 195 196
7. 6 7 9 9 10 10 11 13 15
8. 21 22 24 25 27 28 32 35

20.7 Box-and-Whisker Plots (DOK 2)

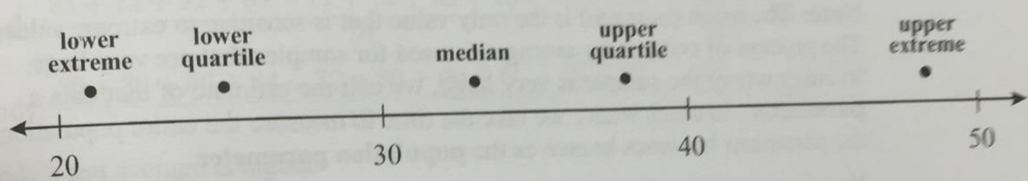
Box-and-whisker plots are used to summarize data as well as to display data. A box-and-whisker plot summarizes data using the median, upper and lower quartiles, and the lower and upper extreme values. Consider the data below: a list of employees' ages at the Acme Lumber Company:

[21] 21 22 23 24 24 24 [25] 26 27 28 29 30 32 32 [33] 33 33 34 35 36 37 37 [38] 38 39 40 40 41 44 [48]

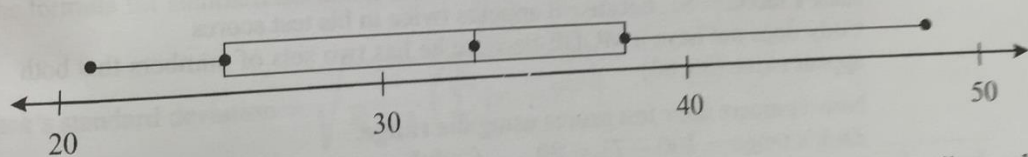
↑ lower extreme — ↑ lower quartile — ↑ median — ↑ upper quartile — ↑ upper extreme

Step 1: Find the median, upper quartile, lower quartile, upper extreme, and lower extreme just like you did in the previous section.

Step 2: Plot the 5 data points found in step 1 above on a number line as shown below.



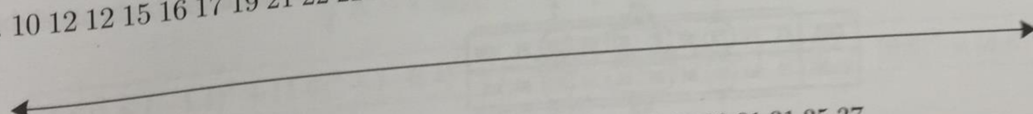
Step 3: Draw a box around the quartile values, and draw a vertical line through the median value. Draw whiskers from each quartile to the extreme value data points.



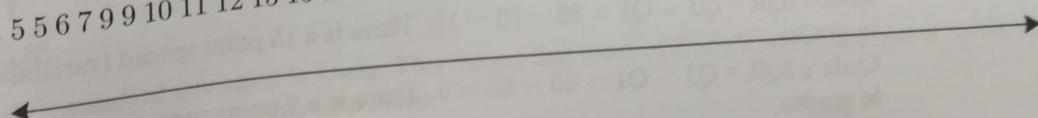
This box-and-whisker displays five types of information: lower extreme, lower quartile, median, upper quartile, and upper extreme.

Draw a box-and-whisker plot for the following sets of data. (DOK 2)

1. 10 12 12 15 16 17 19 21 22 22 25 27 31 35 36 37 38 38 41 43 45 50 51 56 57 58 59



2. 5 5 6 7 9 9 10 11 12 15 15 16 17 18 19 19 20 22 24 26 27 27 30 31 31 35 37



20.8 Comparing Data Sets (DOK 2, 3)

The best way to compare statistics is to use summary statistics. To compare, you could use the mean, median, mode, range, interquartile range, or quartiles.

Example 10: Compare Zack and Cody's test scores using the mean, median, mode, range, interquartile range, and quartiles.

Zack	81	72	91	88	71	73	82	100	81	86
Cody	86	80	86	84	72	80	90	85	89	92

Compare the **mean** of Zack and Cody's test scores. Whose average is higher?

Zack: $\frac{81+72+91+88+71+73+82+100+81+86}{10} = \frac{825}{10} = 82.5$

Cody: $\frac{86+80+86+84+72+80+90+85+89+92}{10} = \frac{844}{10} = 84.4$

Cody's test average is higher.

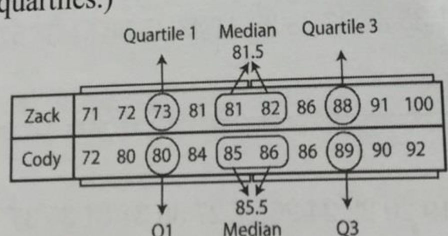
Note: The mean (average) is the only value that is sensitive to extreme values. The process of comparing averages is used for samples that are very large. In cases where the sample is very large, we call the estimate of that data a **parameter**. In cases where we take the time to measure the entire population, the parameter becomes known as the **population parameter**.

Now compare their test scores using the **median**.
Zack's median = 81.5 Cody's median = 85.5
Cody's average is still higher!

Now compare their test scores using the **mode**.
Zack's mode = 81, because it appears twice in his test scores
Cody does not have a MODE because he has two sets of numbers that both appear twice (80, 86)

Now compare their test scores using the **range**.
Zack's range = 100 − 71 = 29 Cody's range = 92 − 72 = 20
Cody's range is smaller meaning his grades are a little more localized around one point, or grade.

Now compare their test scores using the **interquartile range**. (Remember: The interquartile range (IQR) is the measure of variance between the 3rd and 1st quartiles.)



Zack's IQR = Q3 − Q1 = 88 − 73 = 15. There is a 15 point spread (variability) in grades.
Cody's IQR = Q3 − Q1 = 89 − 80 = 9. There is a 9 point spread (variability) in grades.

It is important to know that the mean, median, and mode are measures of the **center**. The range and interquartile range are measures of the **spread (variability)**.

Statistics can also be used to compare a small group of data to a large group of data.

The next set of test scores include one more test for Zack and Cody.

The best way to compare statistics is to use summary statistics. To compare, you could use the mean or standard deviations.

Example 11: Compare Zack and Cody's test scores using the mean and standard deviation.

Zack	81	72	91	88	71	73	82	100	81	91
Cody	86	80	86	84	72	80	90	85	89	92

Compare the **mean** of Zack and Cody's test scores. Whose average is higher?

Zack: $\frac{81+72+91+88+71+73+82+100+81+91}{10} = \frac{825}{10} = 83$

Cody: $\frac{86+80+86+84+72+80+90+85+89+92}{10} = \frac{844}{10} = 84.4$

Cody's test average is higher.

Now compare their test scores using the **standard deviation**.

The formula for standard deviation is $S = \sqrt{\frac{1}{N-1}\left(\sum_{i=1}^{N}(x_i - \overline{x})^2\right)}$

Zack's standard deviation $= \sqrt{\frac{1}{N-1}\left(\sum_{i=1}^{N}(x_i - \overline{x})^2\right)}$

$$= \sqrt{\frac{(-2)^2+(-11)^2+(8)^2+(5)^2+(-12)^2+(-10)^2+(-1)^2+(17)^2+(-2)^2+(8)^2}{10-1}}$$

$= 9.52$

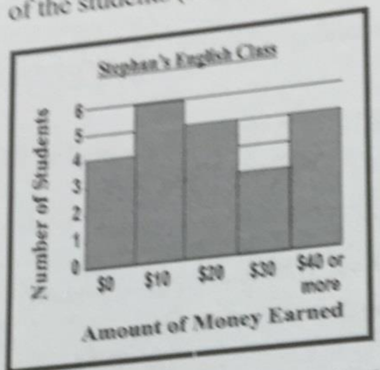
Cody's standard deviation $= \sqrt{\frac{1}{N-1}\left(\sum_{i=1}^{N}(x_i - \overline{x})^2\right)}$

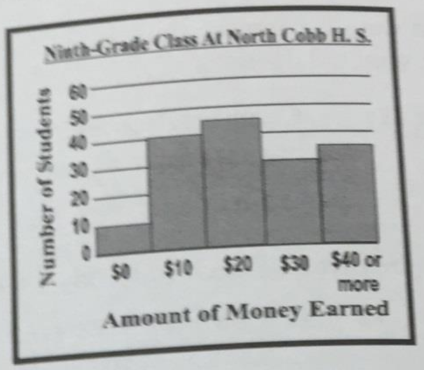
$$= \sqrt{\frac{(1.6)^2+(-4.4)^2+(1.6)^2+(-0.4)^2+(-12.4)^2+(-4.4)^2+(-5.6)^2+(0.6)^2+(4.6)^2+(7.6)^2}{10-1}}$$

$= 5.85$

Cody's standard deviation is lower. This means that most of Cody's scores are closer to his mean, then Zack's scores are to his own mean.

Example 12: Stephan is a student at North Cobb High School. Look at the following histograms. The first histogram displays the amount of money earned each week by each student (rounded to the nearest ten) in Stephan's English class. The second histogram displays the amount of money earned each week by each of the students (rounded to the nearest ten) in the ninth-grade class at his school.





A. Compare the medians of the data in the histograms.

B. Compare the lower quartiles of the data in the histograms.

C. Compare the modes of the data in the histograms.

Solution:

A. To find the median of Stephan's English class, first find the total number of students in his class. By adding up the number of students in each bar $(4 + 6 + 5 + 3 + 5)$, we find there are 23 students in Stephan's English class. The median is the middle number in the data set that is written from least to greatest. The middle number is Stephan's English class is 12, which is found within the third bar ($20).
The total number of students in ninth-grade at North Cobb High School is found by adding up the number of students in each bar $(10 + 40 + 45 + 40 + 35)$. There are 170 students in ninth-grade at North Cobb High School. The median is between the 85th and 86th student, which is found within the third bar ($20).
The median of the amount of money earned by each of the students in the ninth-grade class at North Cobb High School is the same as the median of the amount of money earned by each of the students in Stephan's English class.

B. The lower quartile is the median of the first half of data.
The first half of data in Stephan's English class includes the first 12 students, so the median of this is found within the second bar ($10).
The first half of data in ninth-grade at North Cobb High School includes the first 85 students, so the median of this is found with the second bar ($10).
The lower quartile of the amount of money earned by each of the students in the ninth-grade class at North Cobb High School is the same as the lower quartile of the amount of money earned by each of the students in Stephan's English class.

C. The mode is easy to find. The mode is the value that is listed the most.
In Stephan's English class, the mode is $10.
In the ninth-grade class at North Cobb High School, the mode is $20.
The mode of the amount of money earned by each of the students in the ninth-grade class at North Cobb High School is greater than the mode of the amount of money earned by each of the students in Stephan's English class.

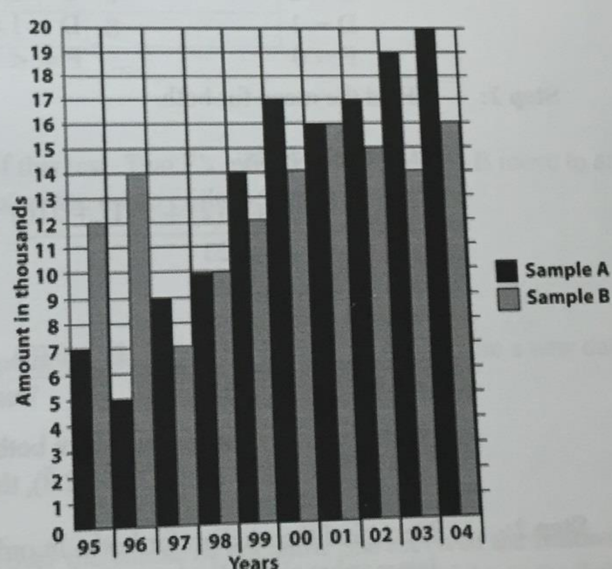
Use the chart below for problems 1–5. (DOK 2, 3)

Sample A	60	71	73	69	80	82
Sample B	71	74	73	79	81	80

1. Find the means.
2. Find the medians.
3. Find the ranges.
4. Find the interquartile range.
5. Find the standard deviations.

Use the bar graph below for problems 6–10. (DOK 2, 3)

6. Find the means.
7. Find the medians.
8. Find the ranges.
9. Find the interquartile range.
10. Find the standard deviations.

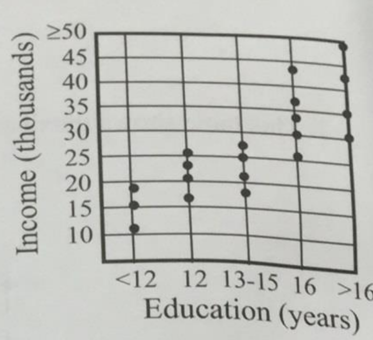


11. Kristen has decided to investigate whether coffee drinkers or tea drinkers consume more cups of their preferred beverage. To do so she interviewed a random group of 6 coffee drinkers and asked them how many cups of coffee they drank in the last week. She then interviewed a random group of 6 tea drinkers and asked them the same question. The mean amount of cups of coffee that a coffee drinker drank in the last week was 26, and the standard deviation of the coffee drinkers' responses was 4 cups. Also, the mean amount of cups of tea that a tea drinker drank in the last week was 22, and the standard deviation of the tea drinkers' responses was 3 cups. Which data set has the higher mean? Which data set has the higher standard deviation? What can you conclude?

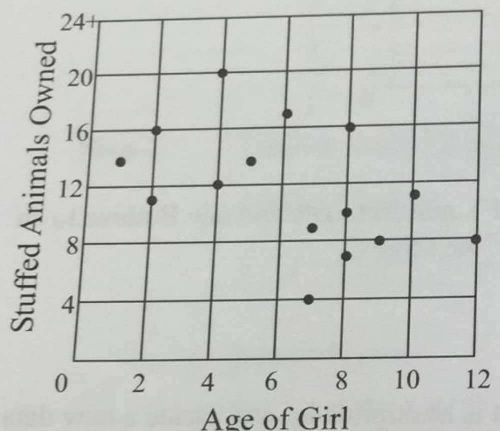
20.10 Scatter Plots (DOK 1)

A **scatter plot** is a graph of ordered pairs involving two sets of data. These plots are used to detect whether two sets of data, or variables, are truly related.

In the example to the right, two variables, income and education, are being compared to see if whether or not they are related. Twenty people were interviewed, ages 25 and older, and the results were recorded on the chart.

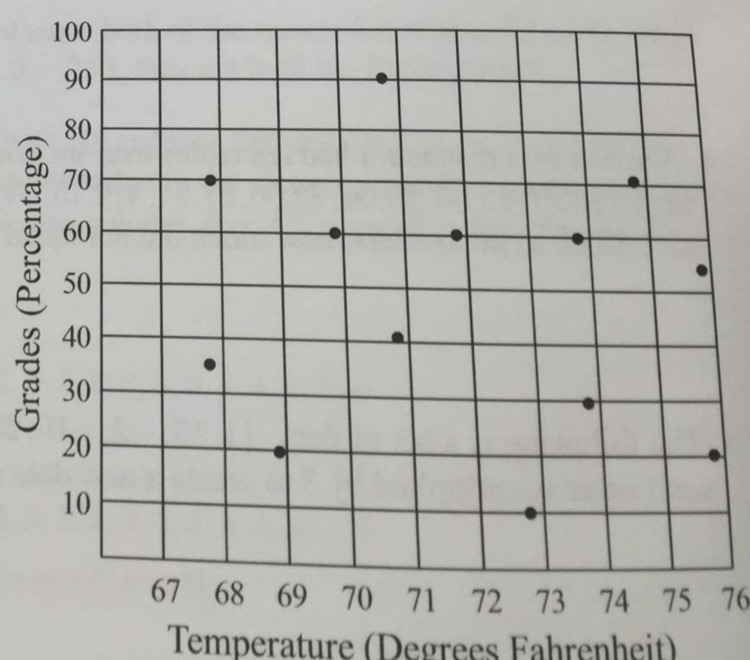


Imagine drawing a line on the scatter plot where half of the points are above the line and half the points are below it. In the plot on the right, you will notice that this line slants upward and to the right. This line direction means there is a **positive** relationship between education and income. In general, for every increase in education, there is a corresponding increase in income.



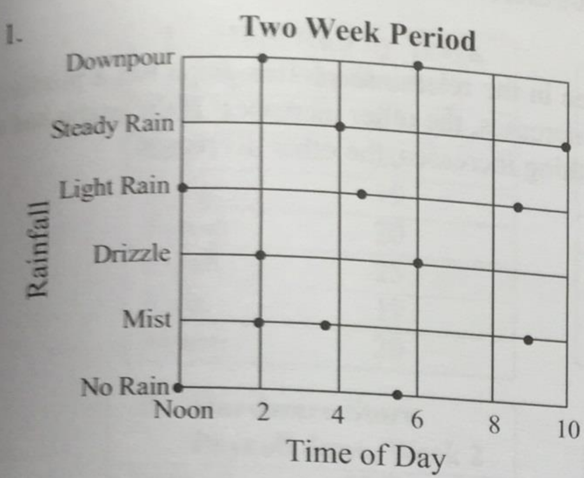
Now, examine the scatter plot on the left. In this case, 15 girls ages 2-12 were interviewed and asked, "How many stuffed animals do you currently have?" If you draw an imaginary line through the middle points, you will notice that the line slants downward and to the right. This plot demonstrates a **negative** relationship between the ages of girls and their stuffed animal ownership. In general, as the girls' ages increase, the number of stuffed animals owned decreases.

Finally, look at the scatter plot shown on the right. In this plot, Rita wanted to see the relationship between the temperature in the classroom and the grades she received on tests she took at that temperature. As you look to your right, you will notice that the points are distributed all over the graph. Because this plot is not in a pattern, there is no way to draw a line through the middle of the points. This type of point pattern indicates there is **no relationship** between Rita's grades on tests and the classroom temperature.

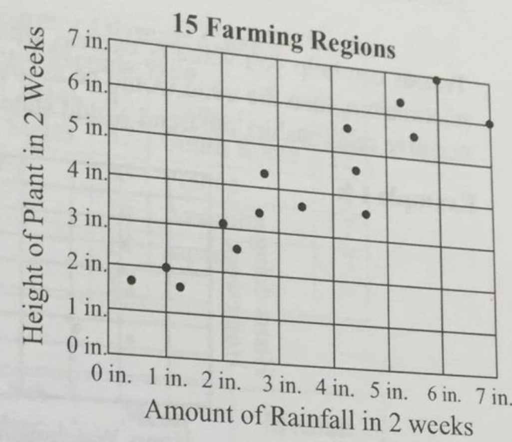


Examine each of the scatter plots below. Write whether the relationship shown between the two variables is "positive", "negative", or "no relationship". (DOK 1)

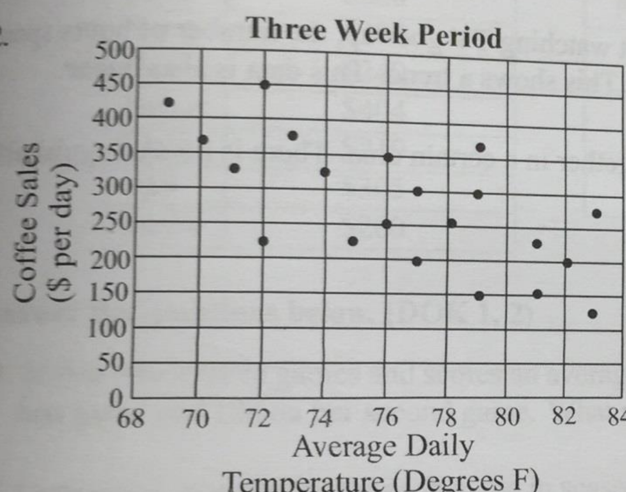
1.



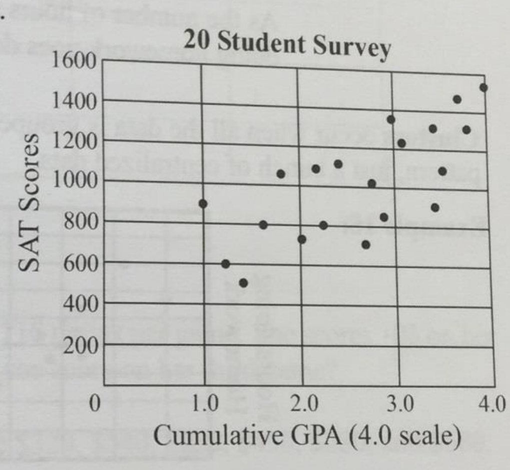
4.



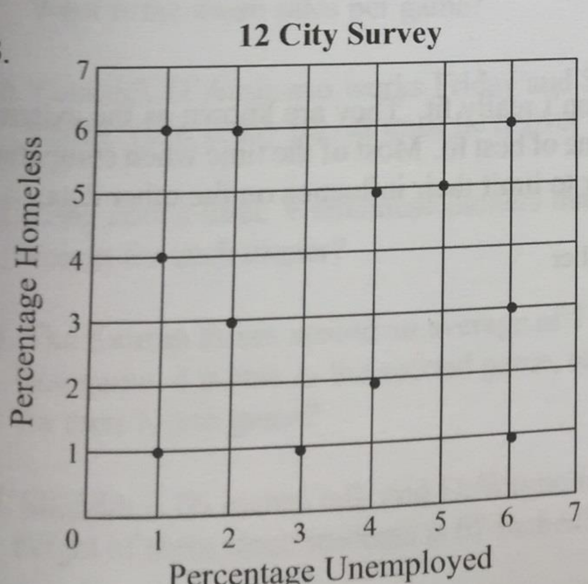
2.



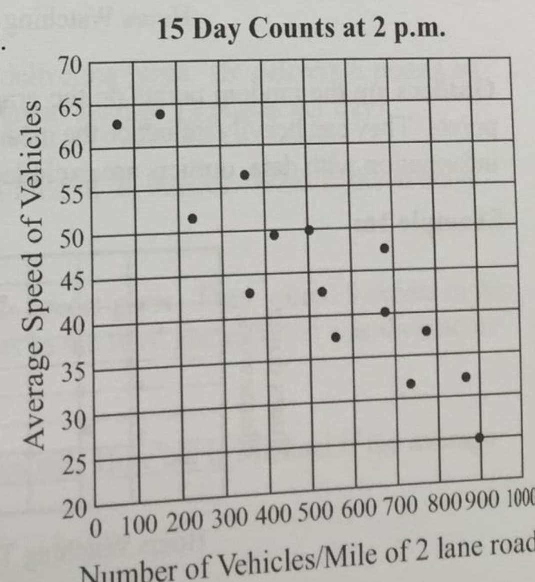
5.



3.



6.

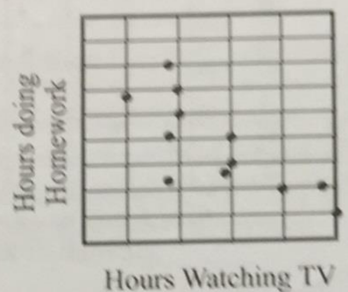


20.11 Making Conclusions About Data (DOK 1)

You've already learned how to recognize positive, negative, or no relationships from data. Now you will see and learn about **trends, clusters**, and **outliers**.

Trends can help you verbally express what you see in the relationship. If a graph has a positive relationship, then the trend would be as one thing increases, the other increases. If the graph has a negative relationship, the trend would state as one thing increases, the other decreases.

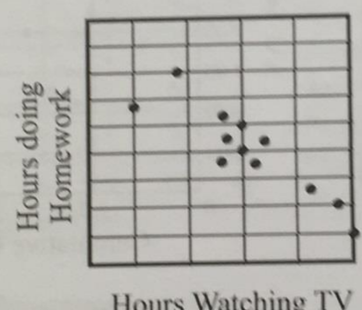
Example 14:



As the number of hours spent watching TV goes up, the number of hours spent doing homework goes down. This shows a trend. This data is also **linear**.

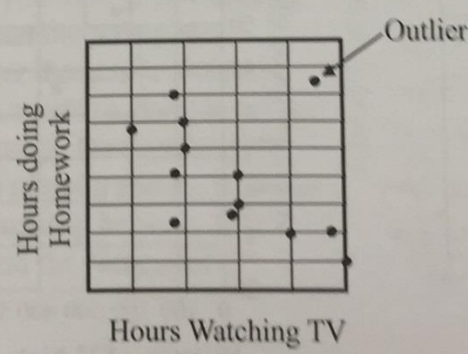
Clusters occur when all the data is grouped together in a certain area. There is no distinguishable pattern, just a bunch of centralized data.

Example 15:



Outliers are the random points on the graph that don't really fit. They are known as the extreme points. They can heavily influence the mean or the line of best fit. Most of the time when computing information with data, outliers are excluded in order to limit their influence on the other data.

Example 16:

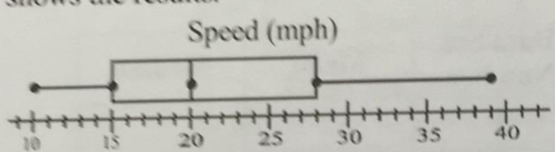


Calculate the standard deviation of each of the following data samples. Use the appropriate unit of measurement in each answer. (DOK 2)

10. 13 light years, 9 light years, 19 light years, 8 light years, 13 light years, 4 light years
11. 811 hertz, 799 hertz, 803 hertz, 804 hertz, 800 hertz, 802 hertz, 810 hertz, 809 hertz
12. 48 milligrams, 57 milligrams, 51 milligrams, 50 milligrams, 60 milligrams
13. 29 bushels, 39 bushels, 38 bushels, 31 bushels, 28 bushels, 37 bushels, 67 bushels

Read the problem below and answer the questions that follow. (DOK 3)

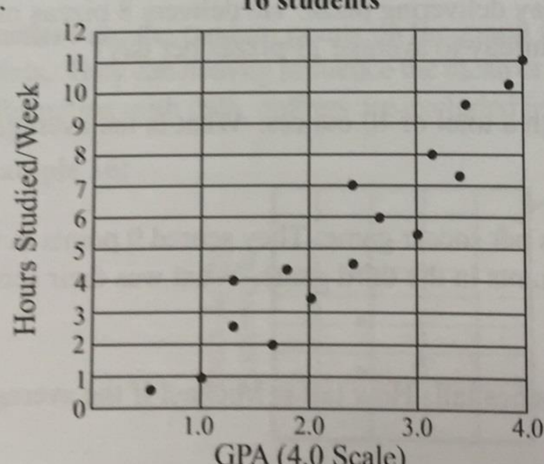
14. As part of a traffic safety campaign, a speed detector was used to record speed in the entrance driveway to the Jefferson High School student parking lot. The box-and-whisker plot below shows the results.



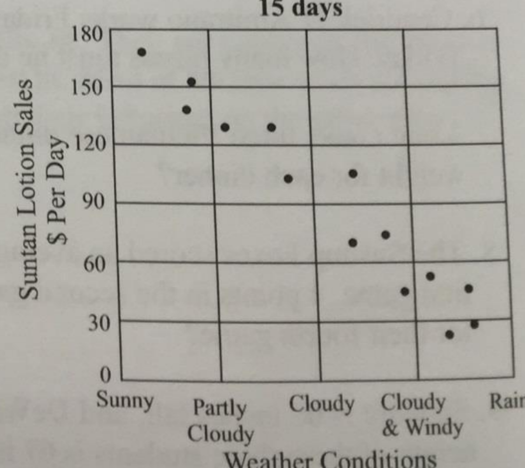
(A) Find the median recorded speed, the range of speeds, and the interquartile range.
(B) The speed limit for the entrance is 15 mph. The high school principal argues that speeds up to 20 mph are acceptable, so nothing needs to be done about students speeding into the entrance. The vice principal thinks that there is a problem that needs to be addressed, probably by issuing speeding tickets. How could the vice principal use this box-and-whisker to plot and the principal's statement that "speeds up to 20 mph are acceptable" to argue her position? Use mathematics in your explanation.

Determine whether the relationship shown between the two variables is "positive", "negative", or "no relationship". (DOK 1)

15.

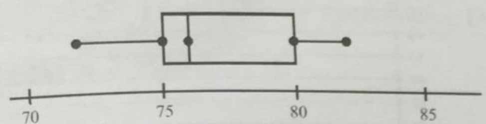


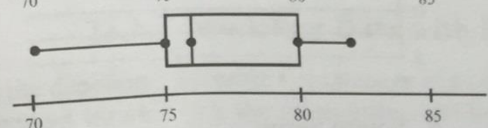
16.

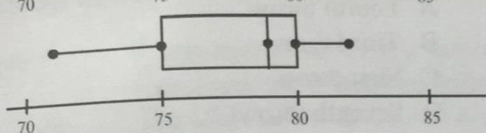


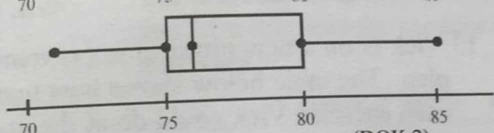
Chapter 20 Test

1 George measured the height of 11 football players in inches.
$73, 72, 75, 82, 75, 76, 77, 81, 76, 80, 77$
Which of the box-and-whisker plots below correctly represents these heights?

A


B


C


D


(DOK 2)

2 A random sample of 7 students had the following scores on a quiz: 8 points; 9 points; 8 points; 6 points; 7 points; 6 points; 2 points. Which of the following is the best approximation of the standard deviation of the sample of quiz scores?

A 1.30 points
B 2.13 points
C 2.30 points
D 3.13 points

(DOK 2)

3 Examine the following two data sets:
Set #1: $49, 55, 68, 72, 98$
Set #2: $20, 36, 47, 68, 75, 82, 89$
Which of the following statements is true?

A They have the same mode.
B They have the same median.
C They have the same mean.
D None of the above.

(DOK 2)

4 What is the median of the following set of data? $33, 31, 35, 24, 38, 30$

A 32
B 31
C 30
D 29

(DOK 1)

5 Which of the following sets of numbers has a range of 51?

A $\{29, 19, 72, 68, 39\}$
B $\{81, 85, 37, 41, 60\}$
C $\{17, 12, 9, 47, 82\}$
D $\{62, 86, 44, 78, 95\}$

(DOK 1)

6 The scores on the math quiz were 94, 73, 87, 81, 82, 62, 55, 60. What is the upper quartile value of these scores?

A 77
B 80
C 84.1
D 84.5

(DOK 1)

7 What is the mean of 36, 54, 66, 45, 36, 36, and 63?

A 36
B 45
C 48
D 63

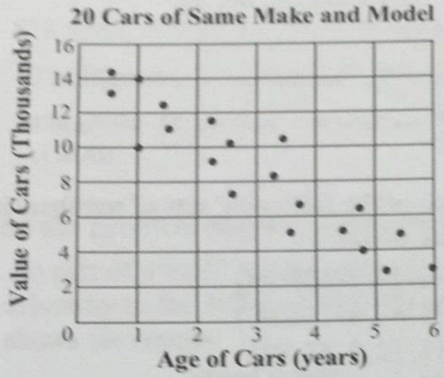
(DOK 1)

8 The mean of a data sample was calculated, and it was then subtracted from each data value in the sample. Each difference was then squared, and the squared differences were then added together. Finally, the sum of the squared differences was divided by 1 less than the number of data values. What is the term used to describe the result?

A Deviation
B Mean
C Sample Standard Deviation
D Population Standard Deviation

(DOK 1)

9 Which of the following accurately describes the relationship between age and value of cars in the graph below?



A Negative
B No relationship
C Positive
D Cannot be determined

(DOK 1)

10 Here is a data set: $\{4, 7, 9, 5, 6, 8, 9\}$. A new data set is created by adding 5 to every value in the old data set. What is the range of the new data set?

A 25
B 10
C 15
D 5

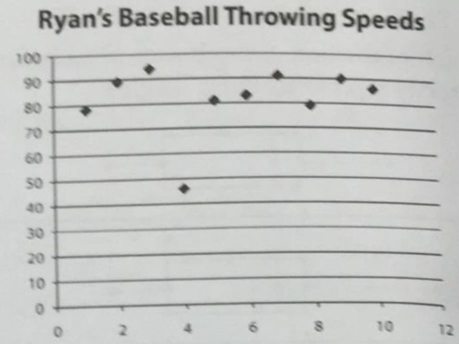
(DOK 2)

11 The following are Raelee's grades in math this semester: $\{93, 68, 83, 87, 79, 91\}$. If the teacher drops the lowest grade, which statement is true?

A The median increases by 2.
B The median increases by 5.
C The mean increases by 2.
D The mean increases by 5.

(DOK 2)

12 The scatter plot below shows the speed at which Ryan threw the baseball in 10 practice throws. Out of the 10 throws, which could be considered an outlier?



A Fourth throw
B Third throw
C First throw
D Seventh throw

(DOK 1)

13 Vick is on a new physical improvement plan. The table below shows how many of each exercise Vick could do at the end of week 1 and week 6.

Exercise	End of Week 1	End of Week 6
Sit-ups	22	56
Push-ups	14	45
Run	50 yd	200 yd
Jumping Jacks	17	75

Which statement below is true about the data in the table?

A The range of sit-ups Vick was able to do between the end of week 1 and the end of week 6 is 36.
B The mean number of push-ups Vick was able to do between the end of week 1 and the end of week 6 is 29.5.
C The range of runs Vick was able to do between the end of week 1 and the end of week 6 is 250 yards.
D The mean number of the jumping jacks Vick was able to do between the end of week 1 and the end of week 6 is 47.

(DOK 3)

Step 2: Determine the y-intercept of the equation of the line.

The equation of the line is $y = 9x + b$, and one of the points on the line is $(1, 16)$. Using this information, substitute the point into the equation to solve for b.

$16 = 9(1) + b$

$16 = 9 + b$

$16 - 9 = b$

$b = 7$

The equation of the line is $y = 9x + 7$.

*Note: Even though the y-intercept of the equation of the line is 7, the line segment that is the model for this situation does not actually cross the y-axis, since the first day the barbershop had any patrons was day 1.

Determine whether or not each of the following situations can be modeled with a linear function. (DOK 2)

1. For every $0.50 increase in the price of a loaf of bread, the number of loaves purchased per day is cut in half.
2. Every 5 minutes the number of cookies in a cookie jar decreases by 3.
3. Every time the number of fans in a stadium increases by 1,000, the noise level increases by 10 percent.
4. For every mile a runner runs, she burns 100 calories.
5. For every new employee a company hires, it spends $4,000 on training.
6. A car is travelling at 25 miles per hour.
7. The area of a square is equal to the length of a side of the square multiplied by itself.
8. The total number of visits to a website is quadrupling every month.
9. To produce a ton of paper, 24 trees must be cut down.

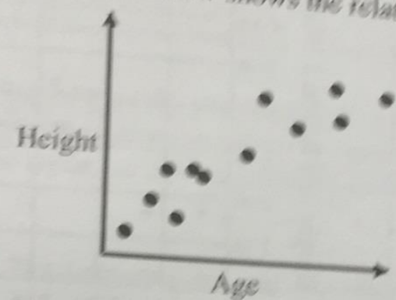
Find the equation of the linear function that can be used to model each situation. (DOK 3)

10. (1 hr, 5 meters), (4 hrs, 41 meters)
11. ($1.50, 78 gallons), ($1.90, 54 gallons)
12. (26 people, 6 cans), (38 people, 9 cans)
13. (7 boxes, 44 lb), (12 boxes, 79 lb)
14. (88°, 70 customers), (95°, 56 customers)
15. (3 days, 120 clicks), (9 days, 200 clicks)

21.3 Interpreting Data in Scatter Plots (DOK 3)

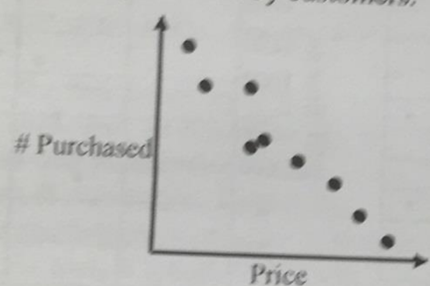
You have already learned that scatter plots show the relationship between two variables. Now, you will learn how to explain the relationship between variables.

Example 4: The graph below shows the relationship between height and age.



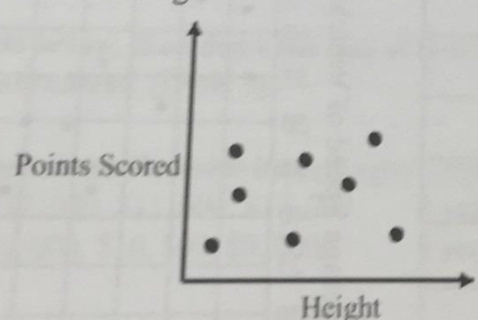
Although it isn't linear, there is clearly a positive relationship between age and height. This means that as age increases, height increases.

Example 5: The graph below shows the relationship between price of an object and the number purchased by customers.



This illustrates a negative relationship. This means as price of an object increases, the number purchased decreases. In other words, if the price of an object goes up, fewer people will buy that object.

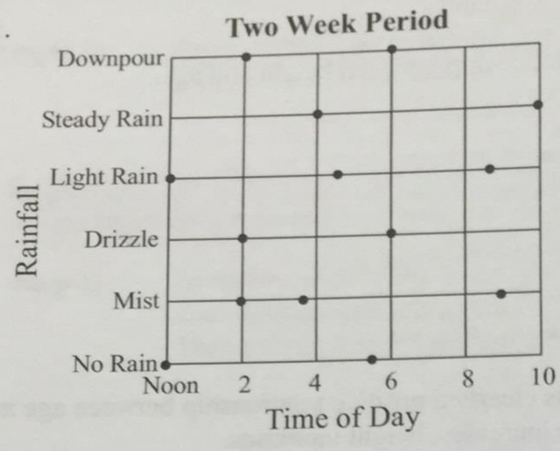
Example 6: The graph below shows the relationship between number of points scored on a test and height



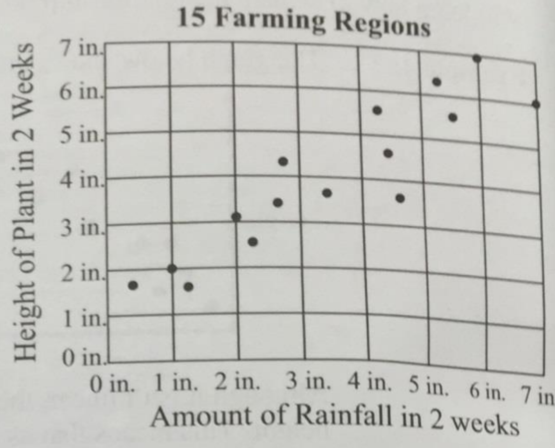
There is no relationship between height and points scored because there is no definable pattern. Therefore, height and number of points scored on a test are not correlated.

Explain the relationship between the variables in each of the following scatter plots. (DOK 3)

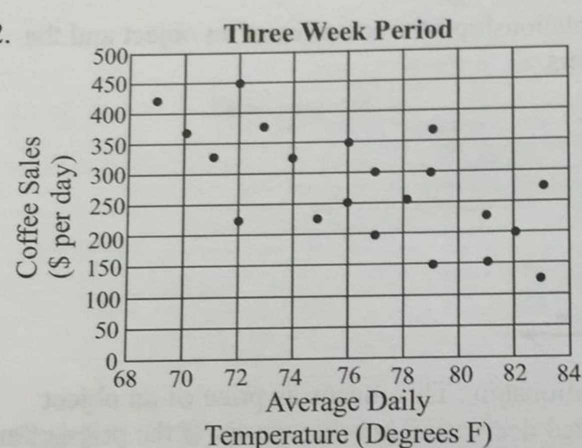
1.



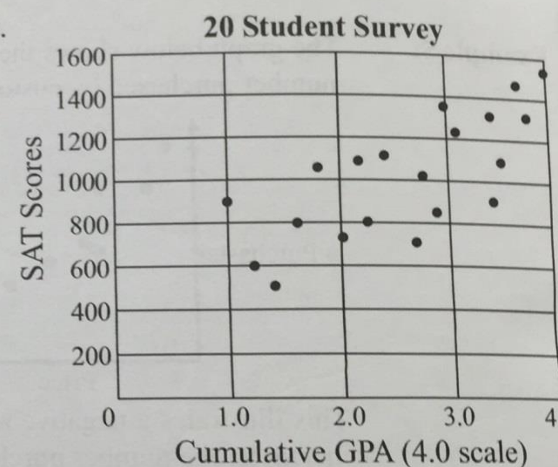
4.



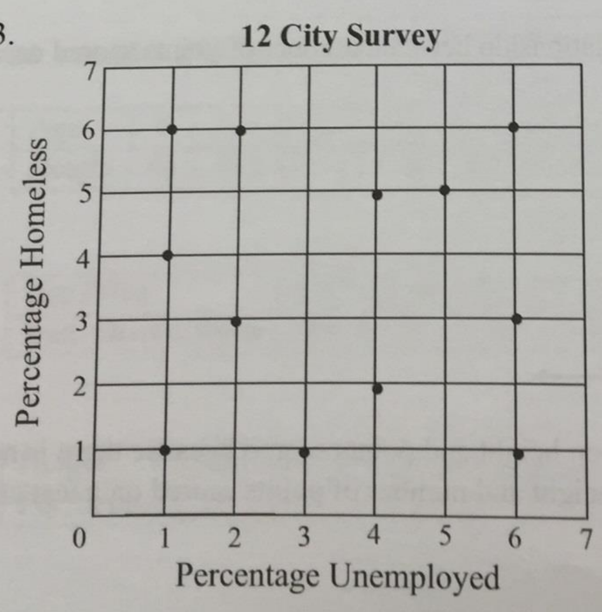
2.



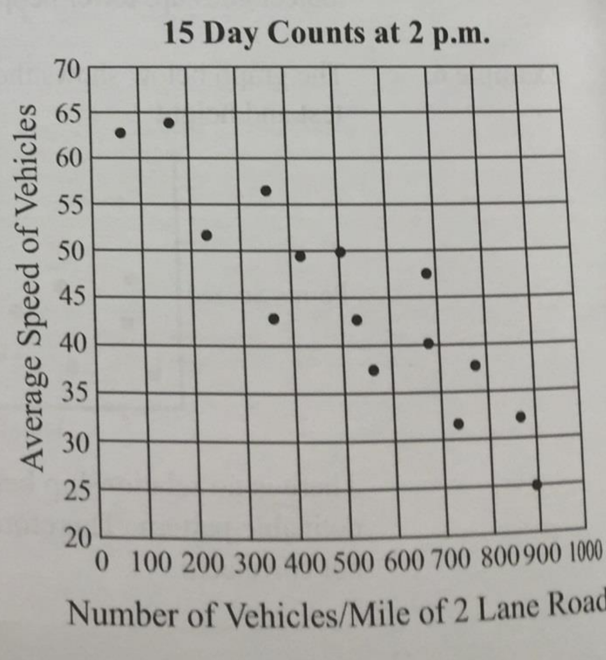
5.



3.



6.



21.4 The Line of Best Fit (DOK 3)

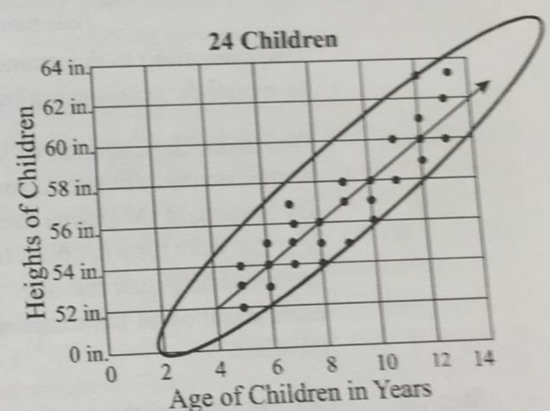
At this point, you now understand how to plot points on a Cartesian plane. You also understand how to find the data trend on a Cartesian plane. These skills are necessary to accomplish the next task, determining the line of best fit. The line of best fit is a straight line that demonstrates the relationship between two variables. The line does not necessarily divide the plotted points into two areas, but this sometimes is the best way to estimate the line of best fit.

To estimate the line of best fit, you must first draw a scatter plot of all data points. Once this is accomplished, draw an oval around all of the points plotted. Draw a line through the points in such a way that the line separates half the points from one another. You may now use this line to answer questions.

Example 7: The following data set contains the heights of children between 5 and 13 years old. Make a scatter plot and draw the line of best fit to represent the trend. Using the graph, determine the height for a 14-year old child.

Age 5: 4'6",4'4",4'5"
Age 6: 4'7", 4'5", 4'6"
Age 7: 4'9", 4'7", 4'6", 4'8"
Age 8: 4'8", 4'6", 4'7"
Age 9: 4'9", 4'7", 4'10"
Age 10: 4'9", 4'8", 4' 10"
Age 11: 5'0", 4'10"
Age 12: 5'1", 4'11", 5'0", 5'3"
Age 13: 5'3", 5'2", 5'0", 5'1"

In this example, the data points lay in a positive sloping direction. To determine the line of best fit, all data points were circled, then a line of best fit was drawn. Half of the points lay below and half above the line of best fit drawn bisecting the narrow length of the oval. The is called "eye-balling."
To find the height of a 14–year old, simply continue the line of best fit forward. In this case, the height is 62 inches.



Plot the data sets below, then draw the line of best fit. Next, use the line to estimate the value of the next measurement. (DOK 3)

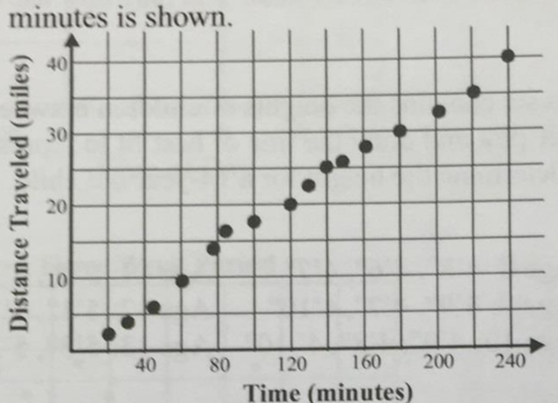
1. Selected values of the Sleekster Brand Light Compact Vehicles: New Vehicle: $13,000.
   1 year old: $12,000, $11,000, $12,500
   2 year old: $9,000, $10,500, $9,500
   3 year old: $8,500, $8,000, $9,000
   4 year old: $7,500, $6,500, $6,000
   5 year old: ?

2. The relationship between string length and kite height for the following kites:
   (L = 500 ft, H = 400 ft) (L = 250 ft, H = 150ft) (L = 100 ft, H = 75 ft)
   (L = 500ft, H = 350 ft) (L = 250 ft, H = 200 ft) (L = 100 ft, H = 50 ft)
   (L = 600 ft, H = ?)

21.5 More Lines of Best Fit (DOK 2)

Relationships that can be modeled with linear functions usually are not exactly linear. In other words, the linear model is only an approximation, and there are points that do not lie exactly on the line. When this is the case, methods such as eyeballing (which we studied in the previous section) and finding the median-median line can be used to determine the equation of the linear model.

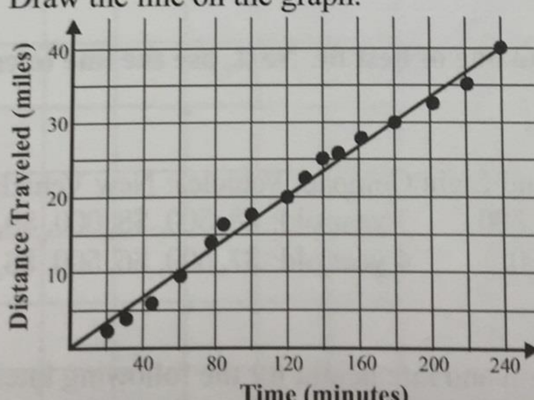
Example 8: Jake rode his bicycle for a total of 240 minutes, and he travelled a total of 40 miles. However, he did not travel at a constant speed, so the graph representing his distance travelled as a function of time is not exactly linear. A scatter plot representing Jake's distance travelled in miles as a function of time passed in minutes is shown.



Use eyeballing to determine the equation of the linear function that can be used to model the relationship between the time passed and the distance travelled.

Step 1: Determine the equation of a line that would be a good approximate representation of the relationship being modeled.
By looking at the points included in the scatter plot, it appears that if the line $y = \frac{1}{6}x$ were drawn, about the same number of points would be above the line as would be below the line. For this reason, the line $y = \frac{1}{6}x$ would be a good representation of the relationship between the time passed and the distance travelled.

Step 2: Draw the line on the graph.



Since about the same number of points are, in fact, above the line as are below the line, the line $y = \frac{1}{6}x$ is an appropriate model.

2. Dr. Silverstein charted 8 of his patients' candy consumption versus their number of cavities in a year.

Patient	Candies Per Day	Cavities Per Year
Kurt	1	1
Erin	4	2
Sarah	4	3
Brian	9	5

Patient	Candies Per Day	Cavities Per Year
Chris	2	0
Olivia	7	3
Rimas	5	2
Tonya	10	4

How many cavities should Jessica have if she ate 8 candies per day? Use the exact the line of best fit to determine your answer.

3. A group of 9 students in Mrs. Van Wyck's math class were given the assignment to determine if there was a strong relationship between the number of people in their household and the amount that their household spent on groceries (excluding pet food).
Their data points were as follows (Number of people in household, cost of groceries):

(3, \$143), (4, \$156) , (2, \$89), (2, \$127), (6, \$201), (5, \$180), (3, \$171), (3, \$152), (3, \$135)

Mrs. Van Wyck has a household of 7 people. How many dollars would you estimate her grocery bill to be? Use the line of best fit to determine your answer.

For questions 4–8, use the linear regression feature on a graphing calculator to determine the equation of the linear function that can be used to model the relationship in each question. (DOK 2)

4. (10 years, 3.5 mm), (11 years, 9.2 mm), (13 years, 14.5 mm), (14 years, 22.8 mm)

5. (2.2 min, 75 beats), (4 min, 149 beats), (6 min, 213 beats), (8.8 min, 298 beats)

6. (\$2.47, 11 pts), (\$2.55, 16 pts), (\$2.58, 19 pts), (\$2.61, 20 pts), (\$2.66, 27 pts)

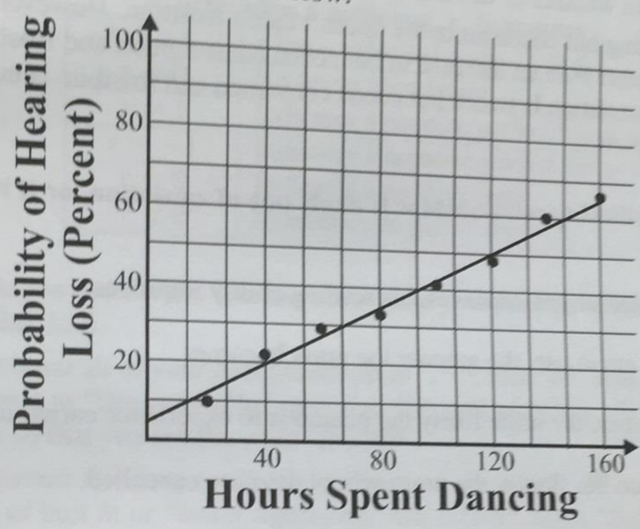
7. (4 laps, 42 calories), (10 laps, 113 calories), (19 laps, 176 calories), (23 laps, 258 calories)

8. (2 carries, 9 yards), (3 carries, 21 yards), (7 carries, 28 yards), (8 carries, 43 yards)

21.10 Correlation and Causation (DOK 3)

When two variables in a relationship are correlated, a change in one variable suggests a change in the other. However, correlation does not imply causation. In other words, just because a change in variable x suggests a change in variable y, the change in x does not necessary cause the change in y. Both the change in x and the change in y could be caused by a third factor.

Example 13: A study has shown that as the number of hours a person spends dancing increases, the probability that the person will experience hearing loss also increases. The relationship between hours spent dancing and probability of hearing loss is shown below:



Because there is an obvious correlation between hours spent dancing and probability of hearing loss, the study concluded that dancing causes hearing loss. Is it likely that the study's conclusion is correct?

Step 1: Determine if a third factor could cause a change in both variables.

What could cause both an increase in the number of hours spent dancing and an increase in the probability of hearing loss? How about loud music?

Step 2: Decide if the relationship between the variables is likely one of causation, or if it is only a correlation.

Common sense dictates that loud music is much more likely to cause hearing loss than dancing. Therefore, the relationship between hours spent dancing and probability of hearing loss is only a correlation, and there is probably no causation involved.

Example 14: Give an example of a variable that could be in correlation with the amount of hot cocoa a person consumes. Then explain whether or not this is an example of causation.

Step 1: Give an example.

One possible example of a variable that could be in correlation with the amount of hot cocoa a person consumes is the probability that the person will experience frostbite. The more hot cocoa a person consumes, the more likely that person is to experience frostbite.

Step 2: Explain whether or not this correlation is an example of causation.

Drinking hot cocoa probably doesn't cause frostbite. However, cold weather can cause both an increase in hot cocoa consumption and frostbite. Therefore, the correlation between hot cocoa consumed and frostbite is not an example of causation.

Decide if each of the following relationships is likely one of causation, or if it is only a correlation. (DOK 3)

1. As a person's corrected vision improves, his reading ability improves.
2. The more umbrellas are in use, the greener the grass becomes.
3. The more a person types, the more likely the person is to experience carpal tunnel syndrome.
4. The more skiers are on the slopes, the more school days are cancelled.
5. As a person sits down more, he becomes more intelligent.
6. The windier it is, the more people are flying kites.
7. As the number of bicycles sold increases, the number of hybrid cars sold increases.
8. The noisier it is in a basketball area, the more hot dogs are sold in the arena.
9. As a person's income increases, the amount she pays in taxes increases.

Give an example of a variable that could be in correlation with each of the following variables. Then explain whether or not this is an example of causation. (DOK 3)

10. a student's height
11. the number of bicycles on the road
12. the number of aspirin a person takes
13. the number of fires in a city
14. the number of children in a family
15. the number of teeth a toddler has

21.11 Correlation Coefficient (DOK 2)

In section 20.4 we learned that the line of best fit demonstrates the relationship, or correlation, between two variables. To measure how well the line fits the data, we use what is known as the **correlation coefficient**. The **correlation coefficient** is a number that is the measure of the strength and direction of the correlation between two variables, commonly represented by the letter, r. The correlation, r, has a value between -1 and 1. The closer r is to -1, the less scattered the points are and the stronger the relationship between two variables. Whenever $r = 1$ or $r = -1$, the data points on the scatterplot must be in a perfectly straight line. When $r < 0$, the data has a negative association, and when $r > 0$, the data has a positive association.

Correlation Coefficient, r, Summary

$-1 \leq r \leq 1$	r has a value between -1 and 1
$r = 0$	No linear correlation
$r = 1$	Indicates a positive perfect linear fit
$r = -1$	Indicates a negative perfect linear fit
$r > 0$	Indicates a positive slope
$r < 0$	Indicates a negative slope

To find the correlation coefficient, you must first enter your data points into the calculator. This is explained in section 15.6.

Once you have entered all of your data points, press "2nd", then "0". This will bring up the catalog screen. Scroll down to "DiagnosticOn" (pressing "D" will automatically scroll down to the D's). Press Enter twice so that you see the word "DONE".

Once you have turned the diagnostics on, go back to the STAT menu and find where you can calculate the line of best fit or "linear regression." On a TI-83, press "right" and you will be under the CALC heading at the top of the STAT menu. Find option 4, "LinReg(ax+b)", then press ENTER. Now you will be back on the main screen with "LinReg(ax+b)" displayed. Type in the following on the TI-83: L1, L2. The correlation coefficient, r, can be seen.

*Note: Correlation does not imply causation. In other words, just because a change in variable x suggests a change in variable y, the change in x does not necessarily <u>cause</u> the change in y. Both the change in x and the change in y could be caused by a third factor.

Example 15: The following data set contains heights in inches and weight in pounds of the students in Mr. Harmon's math class.

Height (in)	Weight (lb)	Height (in)	Weight (lb)
58 in	115 lb	71 in	183 lb
65 in	156 lb	72 in	167 lb
72 in	159 lb	63 in	139 lb
64 in	142 lb	65 in	145 lb
69 in	129 lb	68 in	154 lb
72 in	159 lb	66 in	165 lb

Enter the data: (STAT, Edit,...) Put the height in the L1 column and weight in the L2 column like the following:

L1 (x)	L2 (y)
58	115
65	156
72	159
64	142
69	129
72	159
71	183
72	167
63	139
65	145
68	154
66	165

Now make sure the diagnostics are on:
Press "2nd", then 0. Scroll down to "DiagnosticOn."
Hit ENTER twice.
Now calculate the correlation coefficient by pressing STAT and going to the CALC menu. Find LinReg(ax+b) and press enter. Now you will be back on the main screen with "LinReg(ax+b)" displayed. Type in the following on the TI-83: L1, L2.
Output looks like:
LinReg
$y = ax + b$
$a = 3.005531411$
$b = -50.53773212$
$r^2 = 0.517885189$
$r = 0.7196424036$

So, the correlation coefficient, r, is 0.72. Therefore, there is a strong positive correlation between the heights and weights of the students in Mr. Harmon's class.

Follow the directions in each problem below. (DOK 2)

1. Use the given data sets to find the correlation coefficient, r. Round to the nearest thousandths.

Cumulative GPA (4.0 scale)	SAT Scores
1.0	900
2.8	1050
3.5	1300
1.4	600
2.1	1100
1.8	1050
2.4	1125
4.0	1500
3.0	1380
3.5	1100

2. Which correlation coefficient indicates the stronger linear relationship between random variables for a fixed sample size?

A $r = 0.8$
B $r = 0.5$
C $r = -0.2$
D $r = -0.9$

3. Use the given data sets to find the correlation coefficient, r. Round to the nearest thousandths.

Age of Child	Height of Child (in)
5	53
11	58
6	54
7	54
7	57
8	56
9	55
13	63
10	58
12	63

4. Which correlation coefficient indicates the weaker negative relationship between random variables for a fixed sample size?

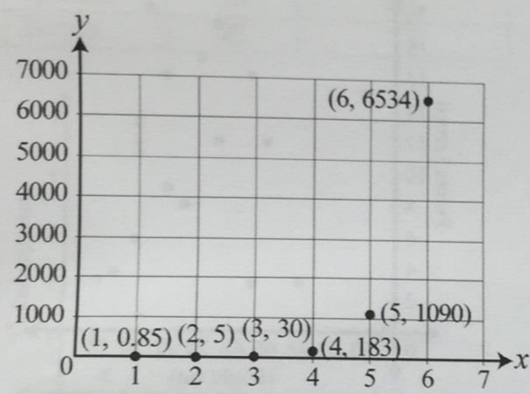
A $r = 0.1$
B $r = -0.2$
C $r = 0.9$
D $r = -0.08$

Determine whether or not each of the following situations can modeled with an exponential function. (DOK 3)

20. As kudzu grows, it doubles its ground coverage weekly.
21. Dad mows the lawn twice a week for a whole summer.
22. A $\$100,000$ mortgage has a 4.2% monthly interest rate.
23. The worth of a car decreases 1% for every $1,000$ miles it is driven.

For each of the following, identify which type of function best models the data: linear or exponential. (DOK 3)

24.



25.

x	-4	-2	-1	0	4	6	7
y	6	7	8	8	10	11	12

26. $y = 6.4$

27.

x	-10	-2	0	3	4	7	9
y	25	22	7	1	0.5	0.2	0.1

Decide if each of the following relationships is likely one of causation, or if it is only a correlation. (DOK 3)

28. the more electricity is used, the more ice cream is eaten

29. the more gifts are given, the more carols are sung

30. as a person eats fewer sugary foods, he gets fewer cavities

31. the more fireworks are detonated, the more bratwursts are eaten

32. the more hours a wage-earner spends working, the more money he/she makes

33. as the grass becomes greener, more sunglasses are worn