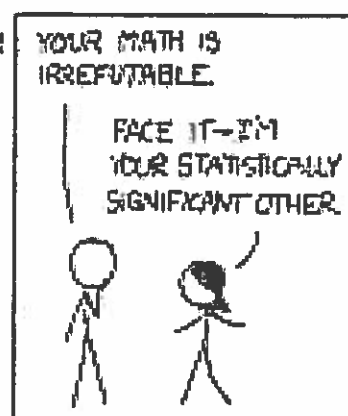
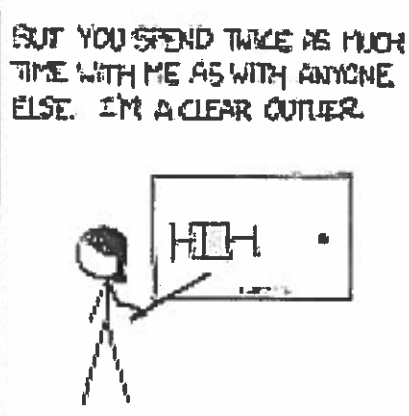
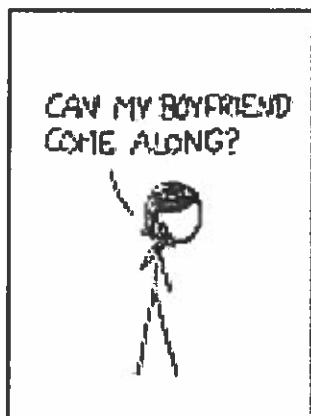
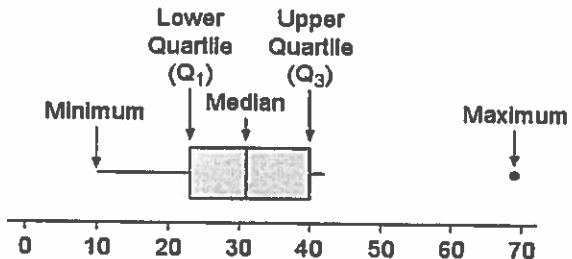
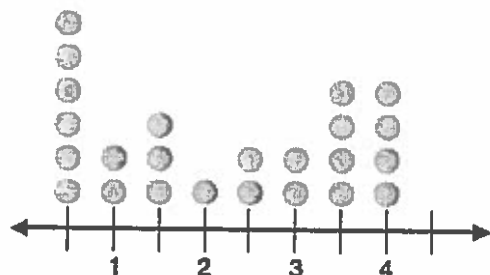
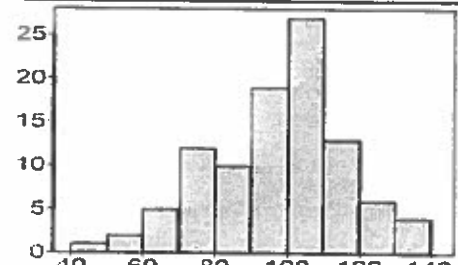
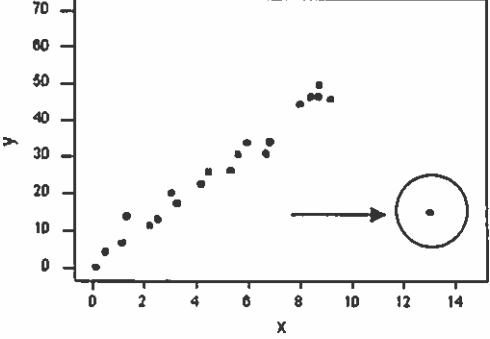


# Unit 6: Statistics





Term	Describe	Example
<b>Box Plot</b>		 <p>Lower Quartile (Q<sub>1</sub>)    Upper Quartile (Q<sub>3</sub>) Minimum    Median    Maximum</p>
<b>Dot Plot</b>		
<b>Histogram</b>		
<b>Median</b>		<p>median of all data, second quartile</p> <p>65, 65, 70, 75, 80, 80, 85, 90, 95, 100</p> <p>median of lower part, first quartile      median of upper part, third quartile</p>
<b>First and Third Quartiles</b>		<p>median of all data, second quartile</p> <p>65, 65, 70, 75, 80, 80, 85, 90, 95, 100</p> <p>median of lower part, first quartile      median of upper part, third quartile</p>

<p><b>Interquartile Range</b></p>		<p><b>Subtract</b></p> <p>Third Quartile (<math>Q_3</math>) – First Quartile (<math>Q_1</math>) = IQR</p>																																
<p><b>Outlier</b></p>		 <p>more than 1.5 times IQR from <math>Q_1</math> or <math>Q_3</math></p>																																
<p><b>Mean</b></p>		<p><math>5 + 4 + 2 + 6 + 3 = 20</math></p> <p><math>\frac{20}{5} = 4</math>      <b>The Mean is <u>4</u>.</b></p>																																
<p><b>Mean Absolute Deviation (MAD)</b></p>		<p>Steps:</p> <ol style="list-style-type: none"> <li>1. Find the Mean</li> <li>2. Calculate the absolute value of the difference between each data value and the mean</li> <li>3. Determine the average of the differences in step 2. This average is the mean absolute deviation</li> </ol>																																
<p><b>Two-Way Frequency Table</b></p>		<table border="1" data-bbox="901 1333 1274 1522"> <thead> <tr> <th></th> <th>Play Daily</th> <th>Play Occasionally</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>Boys</th> <td>16</td> <td>8</td> <td>24</td> </tr> <tr> <th>Girls</th> <td>4</td> <td>12</td> <td>16</td> </tr> <tr> <th>Total</th> <td>20</td> <td>20</td> <td>40</td> </tr> </tbody> </table> <p>Joint frequencies are in the body of the table.</p> <p>Marginal frequencies are in the "Total" row and "Total" column.</p> <table border="1" data-bbox="901 1596 1274 1785"> <thead> <tr> <th></th> <th>Play Daily</th> <th>Play Occasionally</th> <th>Total</th> </tr> </thead> <tbody> <tr> <th>Boys</th> <td>40%</td> <td>20%</td> <td>60%</td> </tr> <tr> <th>Girls</th> <td>10%</td> <td>30%</td> <td>40%</td> </tr> <tr> <th>Total</th> <td>50%</td> <td>50%</td> <td>100%</td> </tr> </tbody> </table> <p>Conditional frequencies are in the body of the table.</p> <p>Marginal frequencies are in the "Total" row and "Total" column.</p>		Play Daily	Play Occasionally	Total	Boys	16	8	24	Girls	4	12	16	Total	20	20	40		Play Daily	Play Occasionally	Total	Boys	40%	20%	60%	Girls	10%	30%	40%	Total	50%	50%	100%
	Play Daily	Play Occasionally	Total																															
Boys	16	8	24																															
Girls	4	12	16																															
Total	20	20	40																															
	Play Daily	Play Occasionally	Total																															
Boys	40%	20%	60%																															
Girls	10%	30%	40%																															
Total	50%	50%	100%																															

# Constructing and Analyzing Two-Way Frequency Tables

**UNDERSTAND** Data can be classified as being either quantitative data or categorical data. **Quantitative data** involve numbers that usually result from measurement. Temperature, height, cost, and population are examples of quantitative data. **Categorical data** take on values that are names or labels. Gender, profession, and nationality are examples of categorical data.

When researchers collect data, they often ask more than one question. Comparing the results of those questions can reveal relationships among the data. To compare two categorical variables, you can enter the frequencies for each category into a **two-way frequency table**.

The two-way frequency table below displays the results of a survey that examined the relationship between gender and video game play. The table shows **joint frequencies** and **marginal frequencies**.

	Play Daily	Play Occasionally	Total
Boys	16	8	24
Girls	4	12	16
Total	20	20	40

Joint frequencies are in the body of the table.

Marginal frequencies are in the "Total" row and "Total" column.

Sometimes you are less interested in the actual frequency count than in the percentage of data values that fall into each category. These percentages are the **relative frequencies**. When displayed in a table, they form a **two-way relative frequency table**. The percentages in the middle of a relative frequency table are called **conditional frequencies**.

	Play Daily	Play Occasionally	Total
Boys	40%	20%	60%
Girls	10%	30%	40%
Total	50%	50%	100%

Conditional frequencies are in the body of the table.

Marginal frequencies are in the "Total" row and "Total" column.

Two-way tables help us see associations between two variables. For example, the above table shows that 40% of the students surveyed are boys who play video games daily and that 10% of the students surveyed are girls who play video games daily, so 50%, or half, of the students surveyed play video games daily. Based on this survey, it seems that boys are more likely to play video games daily than girls.

## Connect

Kyra asked students and parents of students at her high school whether they are in favor of or against a proposal to remove the juice machine from the school cafeteria. The two-way frequency table on the right displays the results of the survey.

	For	Against	Total
Students	5	37	42
Parents	20	18	38
Total	25	55	80

Identify and interpret the marginal and joint frequencies in the table.

- 1 Identify and interpret the marginal frequencies.

Marginal frequencies are in the "Total" column and in the "Total" row.

The marginal frequencies in the "Total" column show that 42 students and 38 parents were surveyed. Roughly equal numbers of parents and students were surveyed.

The marginal frequencies in the "Total" row show that 25 people surveyed supported the proposal and 55 were against it. More than twice as many people surveyed were against the proposal as were in favor of it.

Both sets of marginal frequencies show that Kyra surveyed a total of 80 people.

- 2 Identify and interpret the joint frequencies by row.

Joint frequencies are in the body of the table, not the "Total" column or row.

The first row shows that 5 students support the proposal, while 37 oppose it.

A large majority of students do not support removing the juice machine.

The second row shows that 20 parents support the proposal, while 18 oppose it.

Parents are about evenly split on the proposal.

- 3 Identify and interpret the joint frequencies by column.

The first column shows that 5 students and 20 parents are for the proposal.

Many more parents than students support removing the juice machine.

The second column shows that 37 students and 18 parents are against the proposal.

Many more students than parents are against the proposal.

### DISCUSS

Would your understanding of the situation be different if you only had the marginal frequencies? What do you learn from the joint frequencies that is not shown in the marginal frequencies?

**EXAMPLE A** The P.E. teachers at a high school are organizing an intramural league. They asked ninth-grade students which sport they would most like to play. The results are shown in the frequency table below.

	Basketball	Kickball	Volleyball	Total
Boys	50	30	12	92
Girls	18	32	58	108
Total	68	62	70	200

Create a two-way relative frequency table for the entire table. Based on the data, should the P.E. teachers create a basketball league? Explain.

1

Calculate each relative frequency.

Find the relative frequencies for the entire table. Each relative frequency will be the quotient of the corresponding frequency divided by the total frequencies

	Basketball	Kickball	Volleyball	Total
Boys	$\frac{50}{200} = 0.25$	$\frac{30}{200} = 0.15$	$\frac{12}{200} = 0.06$	$\frac{92}{200} = 0.46$
Girls	$\frac{18}{200} = 0.09$	$\frac{32}{200} = 0.16$	$\frac{58}{200} = 0.29$	$\frac{108}{200} = 0.54$
Total	$\frac{68}{200} = 0.34$	$\frac{62}{200} = 0.31$	$\frac{70}{200} = 0.35$	$\frac{200}{200} = 1.00$

2

Determine whether basketball is the most popular choice.

The two-way relative frequency table shows 34% of students surveyed prefer basketball. That is not significantly more than the percent who prefer kickball and is less than the percent who prefer volleyball.

- ▶ The data show that there is some support, but not overwhelming support, for a basketball league.

DISCUSS

Based on the data, is there an obvious choice for which sport the teachers should select? Explain your thinking.

**EXAMPLE** Carter surveyed 20 ninth-grade students and 30 twelfth-grade students at random. He asked the students whether they were involved in school clubs. After creating a two-way frequency table of his results, he calculated the relative frequencies for each row of his table. The relative frequencies are shown on the right.

	One or More Clubs	No Clubs	Total
9th Grade	30%	70%	100%
12th Grade	80%	20%	100%
Total	60%	40%	100%

Create a frequency table for Carter's data. Then create a two-way relative frequency table for the columns in the frequency table.

1

Use the relative frequencies and the given information to create a frequency table.

You know that Carter surveyed 20 ninth-graders and 30 twelfth-graders. Use those numbers to fill in the Total column. Then use the relative frequencies to calculate the frequencies.

	One or More Clubs	No Clubs	Total
9th Grade	$0.3 \cdot 20 = 6$	$0.7 \cdot 20 = 14$	20
12th Grade	$0.8 \cdot 30 = 24$	$0.2 \cdot 30 = 6$	30
Total	$0.6 \cdot 50 = 30$	$0.4 \cdot 50 = 20$	50

2

Use the frequencies in the table you created to find the relative frequencies by column.

To create a two-way relative frequency table based on columns, divide each value in a column by the total frequency for that column.

	One or More Clubs	No Clubs	Total
9th Grade	$\frac{6}{30} = 20\%$	$\frac{14}{20} = 70\%$	$\frac{20}{50} = 40\%$
12th Grade	$\frac{24}{30} = 80\%$	$\frac{6}{20} = 30\%$	$\frac{30}{50} = 60\%$
Total	$\frac{30}{30} = 100\%$	$\frac{20}{20} = 100\%$	$\frac{50}{50} = 100\%$

**DISCUSS**

What associations do you find in the data in the relative frequency table? Do you see different associations in the data when you look at the relative frequency table by rows compared to the relative frequency table by columns?



# Practice

7

Circle and label marginal frequencies and either joint frequencies or conditional frequencies.

1.

	Smart Phone	No Smart Phone	Total
Boys	15	10	25
Girls	19	6	25
Total	34	16	50

**HINT** A two-way frequency table has joint frequencies and marginal frequencies.

2.

	Smart Phone	No Smart Phone	Total
Boys	30%	20%	50%
Girls	38%	12%	50%
Total	68%	32%	100%

Fill in each blank with an appropriate word or phrase.

3. A two-way frequency table allows you to organize \_\_\_\_\_ data.
4. \_\_\_\_\_ frequencies are entries in the "Total" row and "Total" column of a frequency table.
5. \_\_\_\_\_ frequencies are entries in the body of a two-way relative frequency table.
6. Given a two-way frequency table, you can find relative frequencies for each \_\_\_\_\_, for each \_\_\_\_\_, or for the entire table.

Use the information and the two-way frequency table for questions 7 and 8.

A group of U.S. history teachers asked students where they would most like to go for an overnight field trip. The table shows the results.

	Washington, D.C.	Williamsburg, VA	Total
Boys	77	28	105
Girls	20	75	95
Total	97	103	200

7. Interpret the marginal frequencies.

---

---

8. Interpret the joint frequencies.

---

---

Use the information and the two-way relative frequency table for questions 9 and 10.

Byron asked fellow high school students and their parents if they support a proposal to replace the current school food vendor with a new food vendor.

	New Vendor	Current Vendor	Total
Parents	0.36	0.14	0.50
Students	0.24	0.26	0.50
Total	0.60	0.40	1.00

9. Interpret the marginal frequencies.

---



---

10. Interpret the conditional frequencies.

---



---

Use this information for question 11.

Twenty students were asked which type of music they like best. Three boys said hip-hop, four boys said jazz, and two boys said rock. Six girls said hip-hop, one girl said jazz, and four girls said rock.

11. Use the grid below to create a two-way frequency table for the data.


Use the information below for questions 12-15.

Erika asked ten high school seniors if they owned a car and if they had an after-school job. Her results are shown in the table.

Car	yes	yes	no	no	yes	no	no	yes	no	yes
Job	yes	no	yes	yes	yes	no	no	yes	no	yes

12. Use Erika's results to complete the two-way frequency table below.

	Car	No Car	Total
Job			
No Job			
Total			

13. Complete the table below to show relative frequencies for each column in the table you created for question 12. Express the frequencies as percentages.

	Car	No Car	Total
Job			
No Job			
Total			

14. Does the two-way relative frequency table show a possible association between owning a car and having an after-school job? Explain.

---



---



---

15. Does the two-way relative frequency table show a possible association between **not** owning a car and having an after-school job? Explain.

---



---



---

Use the information and table for questions 16–20.

The two-way frequency table shows the results of a survey in which ninth-grade students were asked which world language elective they most wanted to take next semester.

	Spanish	French	German	Total
Boys	80	30	10	120
Girls	30	20	30	80
Total	110	50	40	200

Use the grids below to create three different two-way relative frequency tables for the data. Express frequencies as decimals. Round to the nearest thousandth.

16. Show relative frequencies for the entire table.

	Spanish	French	German	Total
Boys				
Girls				
Total				

17. Show relative frequencies for each row.

	Spanish	French	German	Total
Boys				
Girls				
Total				

18. Show relative frequencies for each column.

	Spanish	French	German	Total
Boys				
Girls				
Total				

19. **EXAMINE** Examine the two-way relative frequency tables you created above. Describe two or more associations you see in the data.

---



---



---

20. **CONJECTURE** The school will offer a total of 8 sections of world language classes for ninth-grade students next semester. How many sections should be Spanish? French? German? Explain your answers.

---



---



---

Name: \_\_\_\_\_

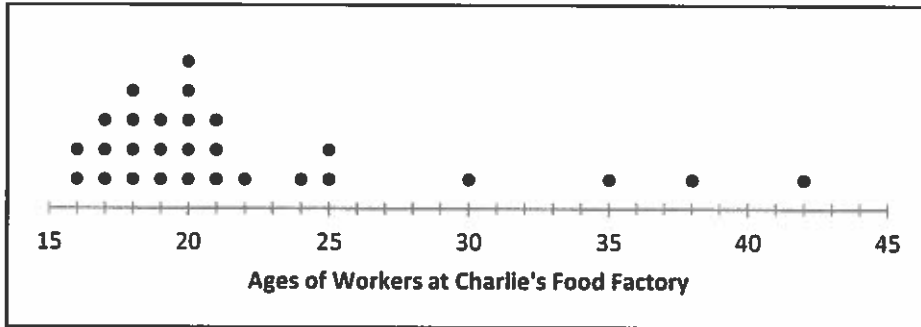
Date: \_\_\_\_\_

### GRAPHICALLY REPRESENTING DATA



**Quantitative** data on a single variable is often collected in order to understand how a characteristic of a group differs amongst the group members or between groups. When we ask a question like “How old is a typical fast food worker?” it is helpful to take a **survey** and then see graphically how the ages differ amongst the group.

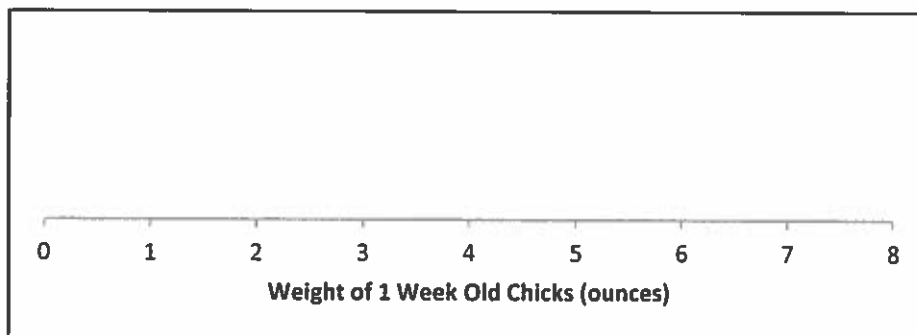
**Exercise #1:** Charlie’s Food Factory currently employs 28 workers whose ages are shown below on a **dot plot**. Answer the following questions based on this plot.



- (a) How many of the workers are 18 years old?
- (b) What is the range of the ages of the workers?
- (c) Would you consider this distribution symmetric?
- (d) The mean (average) age for a worker is 22 years old. Why is this average not representative of a typical worker?

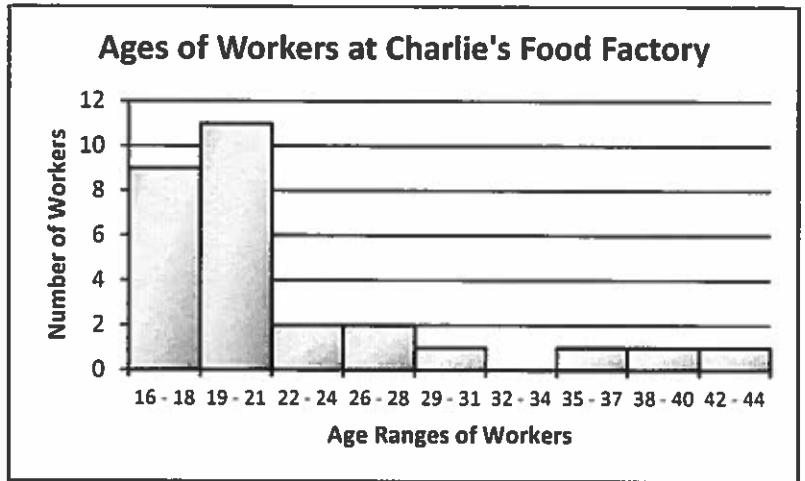
**Exercise #2:** A farm is studying the weight of baby chickens (chicks) after 1 week of growth. They find the weight, in ounces, of 20 chicks. The weights are shown below. Construct a dot plot on the axes given.

2, 1, 3, 4, 2, 2, 3, 1, 5, 3, 4, 4, 5, 6, 3, 8, 5, 4, 6, 3



**Exercise #3:** The following histogram shows the ages of the workers at Charlie's Food Factory (from Exercise #1) but in a different format.

- (a) How many workers have ages between 19 and 21 years?
- (b) What is the disadvantage of a histogram compared to a dot plot?
- (c) Does the histogram have any advantages over the dot plot?



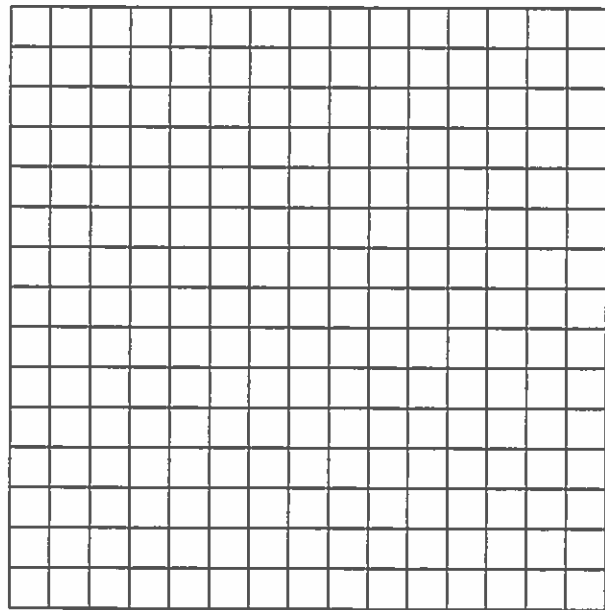
**Exercise #4** The 2006 – 2007 Arlington High School Varsity Boy's basketball team had an excellent season, compiling a record of 15 – 5 (15 wins and 5 losses). The total points scored by the team for each of the 20 games are listed below in the order in which the games were played:

76, 55, 76, 64, 46, 91, 65, 46, 45, 53, 56, 53, 57, 67, 58, 64, 67, 52, 58, 62

(a) Complete the frequency table below.

POINTS SCORED	TALLY	FREQUENCY
40 - 49		
50 - 59		
60 - 69		
70 - 79		
80 - 89		
90 - 99		

(b) Construct the histogram below.



Name: \_\_\_\_\_

Date: \_\_\_\_\_

## QUARTILES AND BOX PLOTS



Another visual representation of how a data set is **distributed** comes in the form of a box plot. We create box plots by dividing the data up roughly into quarters by finding the **quartiles** of the data set.

**Exercise #1:** Shown below are the scores 16 students received on a math quiz.

52, 60, 66, 66, 68, 72, 72, 73, 74, 75, 80, 82, 84, 91, 92, 98

- (a) What is the median of this data set?
- (b) Find the **range** of the data set (defined as the difference between the largest data value and the smallest data value).
- (c) What is the median of the lower half of this data set (known as the **first quartile,  $Q_1$** )?
- (d) What is the median of the upper half of this data set (known as the **third quartile,  $Q_3$** )?

The first and third quartiles are sometimes known as the lower and upper quartiles, respectively. The quartiles, the median, and the lowest and highest values in a data set comprise what is known as the **five number summary** and can be graphically represented on a **box plot**. This type of plot is also sometimes known as a **box and whiskers plot**.

**Exercise #3:** Using the same data set construct a box plot on the number line given below.



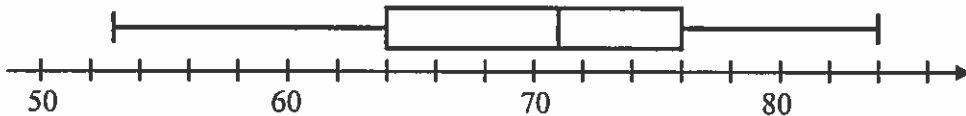
**Exercise #4:** The ages of the 15 employees of the Red Hook Curry House are given below.

16, 17, 17, 18, 19, 22, 25, 26, 29, 33, 33, 37, 40, 42, 44

(a) Determine the median and quartile values for this data set.

(b) Create a box-and-whiskers diagram below.

**Exercise #5:** Twenty of Mr. Ouimet's physics students recently took a quiz. The results of this quiz are shown in the following box-and-whiskers diagram. Assume that all scores are whole numbers.



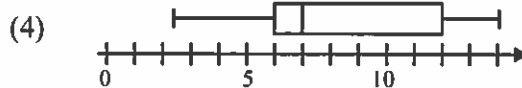
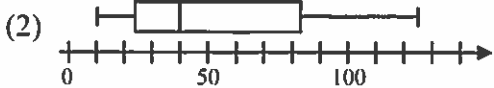
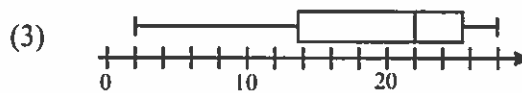
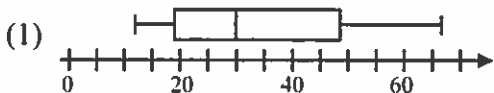
(a) What was the median score on Mr. Ouimet math quiz?

(b) What was the range of the scores on Mr. Ouimet's physics quiz?

(c) What score was greater than or equal to 75% of all other scores on this quiz?

(d) Mr. Ouimet regularly sets the passing grade on his quizzes to be the score of the lower quartile. What is the passing grade on this quiz?

**Exercise #6:** Which of the following box plots shows a data set with the greatest median?





Name: \_\_\_\_\_ Date: \_\_\_\_\_

### Graphical Displays for Data Homework

Kirsten plays softball in the spring. Each game, she records the number of times she reaches first base without being called out. Use the data in the table to solve problems 1 -5.

Game	Number of times at first	Game	Number of times at first
1	5	10	0
2	1	11	1
3	2	12	1
4	0	13	0
5	2	14	5
6	2	15	5
7	4	16	4
8	4	17	0
9	0	18	4

1. Create a dot plot showing the number of times Kirsten reached first base in each game.

2. Find the minimum, maximum, first quartile, and third quartile of the data set.

- a. Minimum:
- b. Maximum:
- c. First Quartile:
- d. Third Quartile:

3. Create a box plot showing the number of times Kirsten reached first base.

4. Find the interquartile range of the data. Are there any outliers?

5. Kirsten wants to analyze her performance using this data. She wants to understand the range of her data and the frequency of different results. Which graph, the dot plot or the box plot, will be most useful to Kirsten? Explain.

16

Dr. Singh is a veterinarian. He records the weights of each pet. The weights of 10 German shepherds, all 4-year-old males, are in the table below, rounded to the nearest pound. Use this information to solve problems 6-10.

Weight in pounds
80
78
82
84
81
89
83
81
81
82

6. Create a histogram showing the weights of Dr. Singh's German shepherds.

7. Find the minimum, maximum, first quartile, and third quartile of the data set.

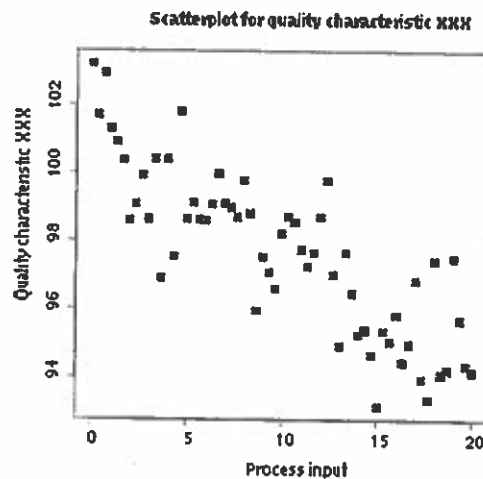
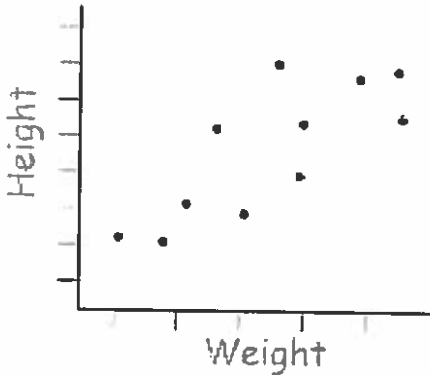
- Minimum:
- Maximum:
- First Quartile:
- Third Quartile:

8. Create a box plot showing the weights of the German shepherds.

9. Find the interquartile range of the data. Are there any outliers?

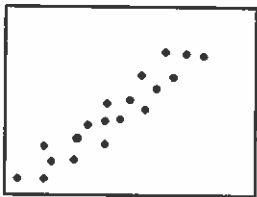
10. Dr. Singh wants to analyze the weights of the German shepherds. He wants to understand the center and spread of his data, so that he has a better idea of an expected weight for a 4-year-old male German shepherd. Which graph would be most useful to Dr. Singh? Explain.

Displaying data visually can help you see relationships. A **scatter plot** is a graph with points plotted to show a possible relationship between two sets of data. A scatter plot is an effective way to display some types of data.



Is a scatter plot discrete or continuous?

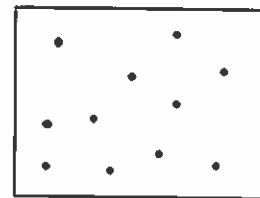
A *scatter plot* is helpful in understanding the form, direction, and strength of the relationship between two variables. **Correlation** is the strength and direction of the linear relationship between the two variables.



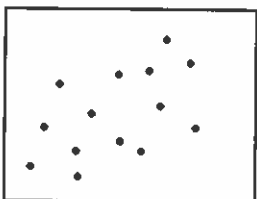
Strong Positive



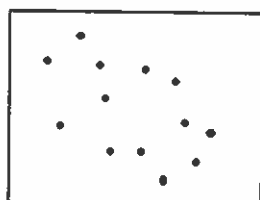
Strong Negative



None



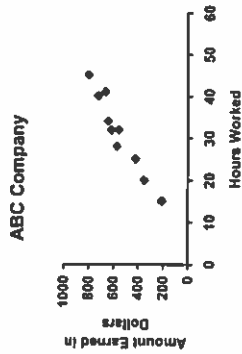
Weak Positive



Weak Negative

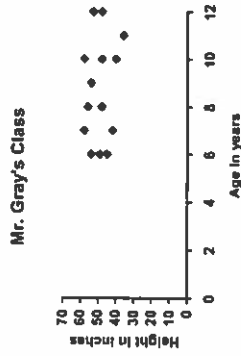
## Scatter Plots

- 1) The scatter plot below shows a relationship between hours worked and money earned. Which best describes the relationship between the variables?



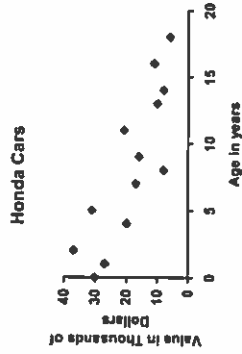
- A) Strong positive correlation
- B) Weak positive correlation
- C) Strong negative correlation
- D) Weak negative correlation

- 2) This scatter plot shows a relationship between age and height. Which best describes the relationship between the variables?



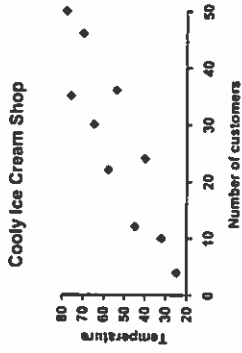
- A) Strong positive correlation
- B) Weak positive correlation
- C) Strong negative correlation
- D) No correlation

- 3) This scatter plot shows the relationship between the age of a car and its value. Which best describes the relationship between the variables?



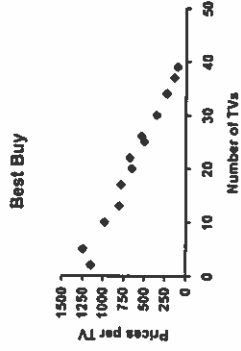
- A) Strong positive correlation
- B) Weak negative correlation
- C) Strong negative correlation
- D) No correlation

- 4) This scatter plot shows a relationship between the outdoor temperature and number of customers in an ice cream store. Which best describes the relationship between the variables?



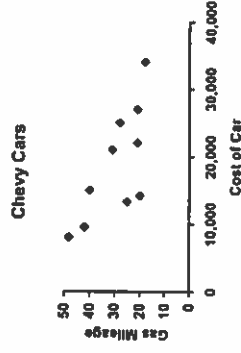
- A) Strong positive correlation
- B) Weak positive correlation
- C) Weak negative correlation
- D) No correlation

- 5) This scatter plot shows a relationship between the TVs purchased and prices. Which best describes the relationship between the variables?



- A) Strong positive correlation
- B) Weak positive correlation
- C) Strong negative correlation
- D) Weak negative correlation

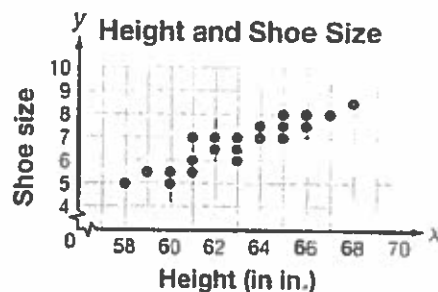
- 6) This scatter plot shows a relationship between the cost of Chevy cars and their gas mileage. Which best describes the relationship between the variables?



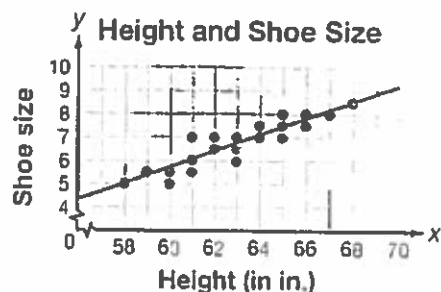
- A) Strong positive correlation
- B) Weak positive correlation
- C) Weak negative correlation
- D) No correlation

# Constructing and Analyzing Scatter Plots

**UNDERSTAND** When you study the relationship between two variables—such as the heights and shoe sizes of a group of students—you are working with **bivariate data**. Bivariate data can be written as a set of  $(x, y)$  ordered pairs and graphed on a coordinate plane. This kind of graph is called a **scatter plot**. A scatter plot can help you interpret bivariate data. The scatter plot below shows a set of ordered pairs in which the  $x$ -values represent heights and the  $y$ -values represent shoe sizes.



Look at the shape formed by the plotted points. The shape resembles a straight line. This suggests a linear relationship between the variables. You can draw a line to fit, or model, the data. The line you draw represents a linear function. If the line is a good fit, you can use the graph and the equation of the line to interpret and make predictions about the data.

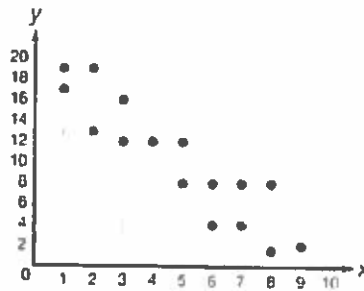
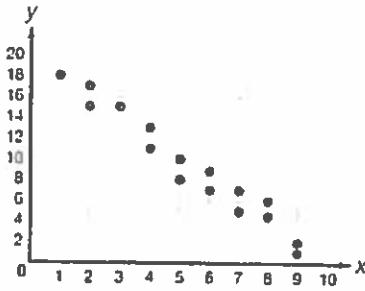


The line appears to be a good fit. The data points slant up from left to right, indicating a positive linear relationship. The line has a positive slope and is close to most data points.

You could also show that the line is a good fit for the data by calculating **residuals**. For each point  $(x, y)$  on the scatter plot, there is a corresponding point  $(x, \hat{y})$  on the line of fit. A residual is equal to the difference  $y - \hat{y}$ . Residuals measure the difference of each actual  $y$ -value and the expected  $y$ -value ( $\hat{y}$ ), which is based on the equation of the line of fit.

Residuals help you determine how accurately the linear function could predict actual points on the scatter plot. That is, if the values of the residuals are relatively small, the linear function is a good fit for the data. So, for any value of  $x$ , you could use the equation of the line to make a good prediction about what the value of  $y$  would be, and vice versa.

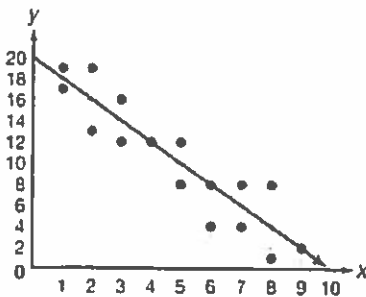
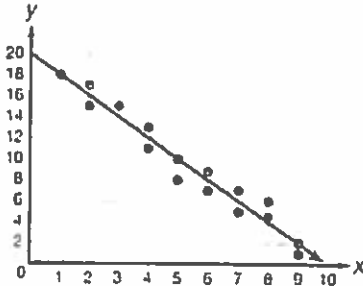
Draw a line of fit for each of the scatter plots. Determine how well each fits the data.



1

Draw a line to model the data for each scatter plot.

For each plot, draw a line that has about as many points above it as below it.



2

Use residuals to determine how well the lines fit the data in the first plot.

Pick several data points, such as (1, 18), (4, 11), (6, 7), and (8, 6). Find the corresponding points,  $(x, \hat{y})$ , on the line for those  $x$ -values: (1, 18), (4, 12), (6, 8), and (8, 4). Calculate the residuals.

(1, 18):  $y - \hat{y} = 18 - 18 = 0$

(4, 11):  $y - \hat{y} = 11 - 12 = -1$

(6, 7):  $y - \hat{y} = 7 - 8 = -1$

(8, 6):  $y - \hat{y} = 6 - 4 = 2$

None of the residuals have large values. The line fits the first data set well.

3

Use residuals to determine how well the line fits the data in the second plot.

Pick several data points: (1, 19), (4, 12), (6, 4), and (8, 8). Find the corresponding points on the line: (1, 18), (4, 12), (6, 8), and (8, 4). Calculate the residuals.

(1, 19):  $y - \hat{y} = 19 - 18 = 1$

(4, 12):  $y - \hat{y} = 12 - 12 = 0$

(6, 4):  $y - \hat{y} = 4 - 8 = -4$

(8, 8):  $y - \hat{y} = 8 - 4 = 4$

Some of the residuals have large values. The line does not fit the second data set well.

DISCUSS

Are the lines drawn the only possible lines of fit that could have been drawn for these scatter plots? Why or why not?

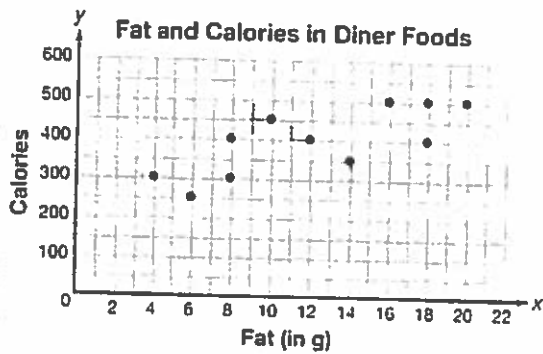
**EXAMPLE** For a health project, Dylan recorded the number of grams of fat and the number of calories in lunch entrees sold at his favorite diner.

Fat (in grams)	4	6	8	8	10	12	14	16	18	18	20
Calories	300	250	300	400	450	400	350	500	400	500	500

Create a scatter plot for the data. Draw a line to fit the data. Find the equation of the line.

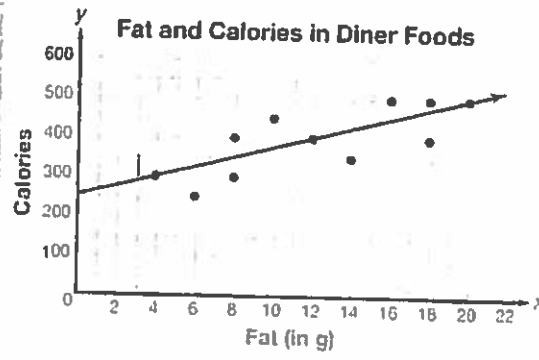
1

Use the ordered pairs of data items to create a scatter plot.



2

Draw a line to fit the data.



3

Write an equation for the line of fit.

The points (4, 300) and (12, 400) are on the line. Use those points to find the slope.

$$m = \frac{400 - 300}{12 - 4} = \frac{100}{8} = 12.5$$

The y-intercept is at (0, 250), so  $b = 250$ .

▶ The equation of the line is  $y = 12.5x + 250$ .

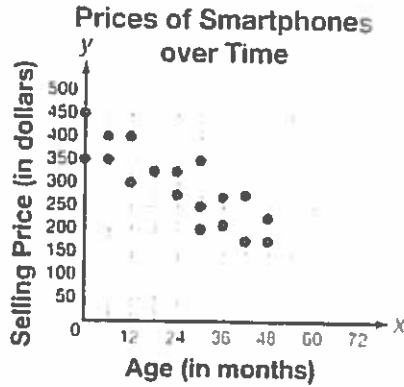
**DISCUSS**

Explain what the slope of the line tells you in this context. Do the data show a positive linear relationship or a negative linear relationship?

# Problem Solving

## READ

The scatter plot shows the ages of various Model Z smartphones, in months, and the prices for which they sold. Predict how much Trent will pay if he buys a Model Z smartphone that is 5 years old.



## PLAN

Draw a \_\_\_\_\_ to fit the data. Write the equation of the \_\_\_\_\_, and use it to predict the price for a phone that is 5 years, or \_\_\_\_\_ months, old.

## SOLVE

On the scatter plot, draw a line that fits the data.

Choose two points on the line, (\_\_\_\_\_, \_\_\_\_\_) and (\_\_\_\_\_, \_\_\_\_\_).

Use the points to find the slope of the line.  $m =$  \_\_\_\_\_

In this context, the slope represents \_\_\_\_\_

Find the y-intercept of the line. Extend the line to the y-axis if necessary.  $b =$  \_\_\_\_\_

The equation for the line is  $y =$  \_\_\_\_\_

In this context, the y-intercept represents \_\_\_\_\_

To predict the cost of a 5-year-old smartphone, substitute 60 for  $x$  in the equation. \$ \_\_\_\_\_

## CHECK

Pick three data points from the scatter plot: (\_\_\_\_\_, \_\_\_\_\_), (\_\_\_\_\_, \_\_\_\_\_), (\_\_\_\_\_, \_\_\_\_\_).

Find the points with corresponding  $x$ -values on the line of fit:

(\_\_\_\_\_, \_\_\_\_\_), (\_\_\_\_\_, \_\_\_\_\_), (\_\_\_\_\_, \_\_\_\_\_).

Calculate the residual for each point. Each residual is relatively \_\_\_\_\_

Does the line fit the data well? \_\_\_\_\_ Is your answer a reasonable prediction? \_\_\_\_\_

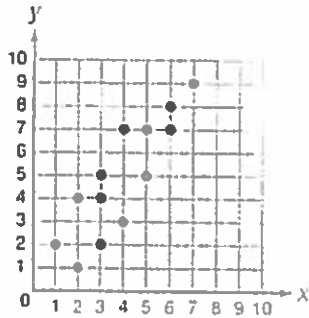
A good prediction is that Trent will pay about \_\_\_\_\_ for a Model Z smartphone that is 5 years old.



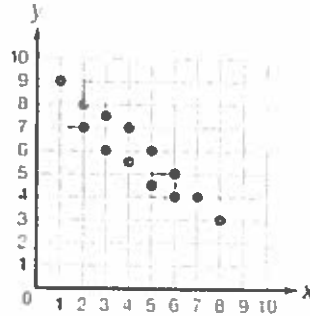
# Practice

Describe the relationship shown in each scatter plot as either *positive* or *negative*.

1.



2.



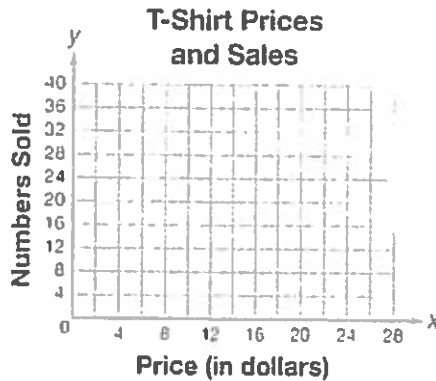
**HINT** A line that slants from lower left to upper right has a positive slope.

Use the information and table below for questions 3 and 4.

The table below shows T-shirt sales data for a store one weekend.

Price, $x$ (in dollars)	4	8	8	12	12	16	20	20	24	24
Number Sold, $y$	32	26	30	22	26	20	12	20	14	10

3. Create a scatter plot for the data. Then draw a line of fit for the data.



4. Find the slope of the line of fit. What does it represent in the context of this problem?

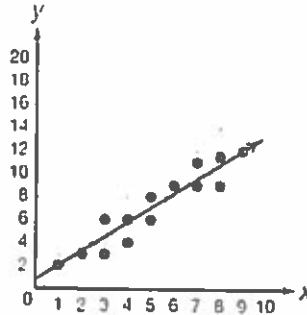
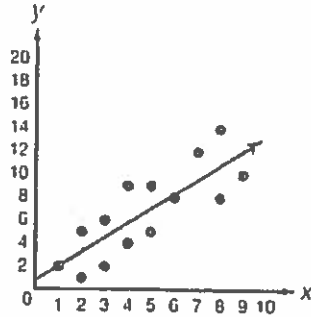
---



---

Assess the fit of the lines to the data.

5. The lines of fit in the scatter plots below are identical.



Which line better fits the data in its scatter plot? How did you determine your answer?

---

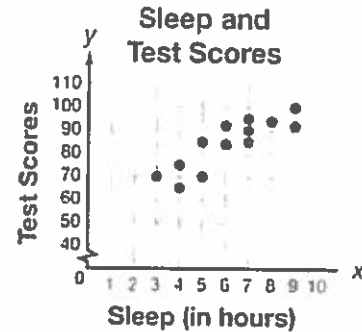


---

Use the information and scatter plot below for questions 6 and 7.

The scatter plot shows the number of hours of sleep that students got the night before a test and their scores on the test.

6. **INTERPRET** Draw a line of fit for the scatter plot. Identify the slope and y-intercept of the line. What does each represent in the context of this problem?




---



---



---

7. **PREDICT** Write the equation of the line. Then use the equation to predict a student's test score if she gets only 2 hours of sleep before the next test.

---



---

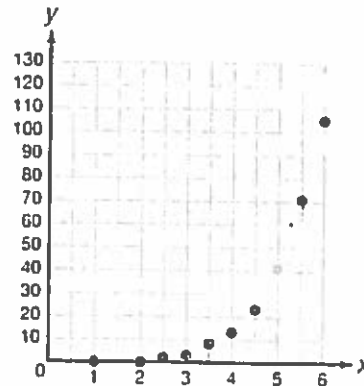


---

# Best Fit and Correlation

**UNDERSTAND** You can draw a line of fit for a scatter plot by analyzing the data visually. Someone else, however, could look at the same data and draw a slightly different line. To find the line that best fits the data, you need to use a process called regression analysis. Regression analysis helps you find the function that minimizes residuals.

When there seems to be a linear relationship in the data, regression analysis can find the equation of a **line of best fit**. But not all bivariate data show a linear association. In some cases, the relationship between the variables is better modeled by a curve, as in the scatter plot shown. For data that do not have a linear association, you will need to find a **curve of best fit**. To find the equation of either a line of best fit or a curve of best fit, you can use a graphing calculator to perform a regression analysis.



**UNDERSTAND** Once you have determined the line of best fit for bivariate data, you can use the **correlation coefficient**,  $r$ , to describe the strength and direction of the relationship between the two variables.

These statements will help you interpret a correlation coefficient.

- The value of  $r$  is always in the range  $-1 \leq r \leq 1$ .
- If  $r$  is close to 1, the data show a strong positive correlation
- If  $r$  is close to  $-1$ , the data show a strong negative correlation.
- If  $r = 0$ , the data do not show a linear correlation.

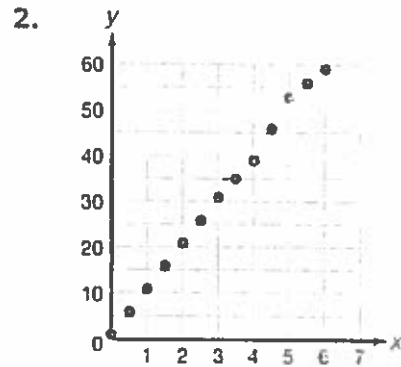
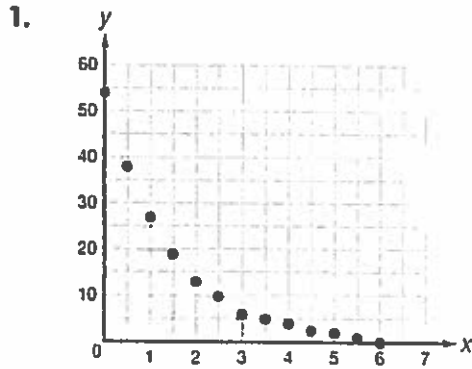
When bivariate data have a strong correlation, the predictions you make by using the line of best fit are likely to be very accurate. When there is a weak correlation, these predictions will tend to be less accurate. A positive correlation means that as one variable increases, the other variable tends to increase also. A negative correlation means that as one variable increases, the other tends to decrease.

The correlation coefficient,  $r$ , is calculated using a rather complex formula involving residuals. Fortunately, you can use a calculator to do that work for you!

Keep in mind that there is a crucial difference between correlation and causation. A strong correlation does not tell you that  $x$  is the cause of  $y$ . For example, buying lemonade and going to the beach might be strongly correlated, but one does not cause the other.

# Practice

Use *line* or *curve* to tell which kind of model best fits each data set.



Use *strong*, *weak*, *positive*, *negative*, or *no linear correlation* to describe what each correlation coefficient,  $r$ , tells you about a bivariate data set.

3.  $r = 0$

4.  $r = 0.250$

5.  $r = -0.895$

REMEMBER The closer  $r$  is to 1 or  $-1$ , the stronger the correlation.

Write *true* or *false* for each statement. If false, rewrite the statement so it is true.

6. A line of best fit will help you predict values for variables with complete accuracy.

---



---

7. Not all bivariate data show a linear correlation, so sometimes data are better modeled by a curve than a line.

---



---

8. If regression analysis shows that there is a strong correlation between two variables,  $x$  and  $y$ , then  $x$  must cause  $y$ .

---

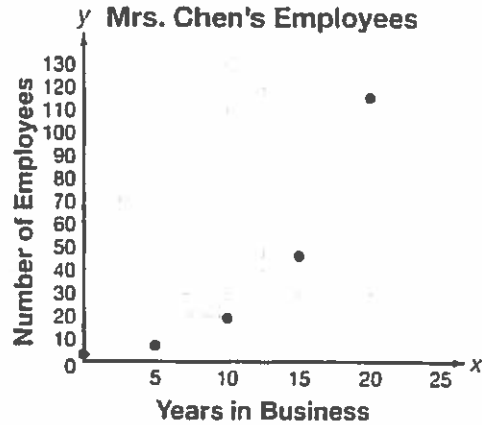


---

Use the information, table, and scatter plot below for question 14.

Mrs. Chen started a business 20 years ago. The table and scatter plot show the number of employees her growing business has had over a period of 20 years.

Years in Business	Number of Employees
0	3
5	7
10	19
15	46
20	115

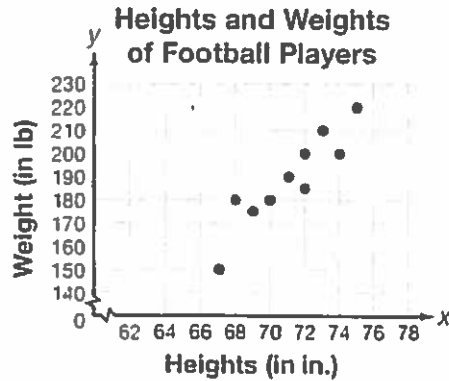


14. Use your calculator to perform an exponential regression for the data. What is the equation of the curve of best fit? Graph that curve on the scatter plot above.
- 

Use the information, table, and scatter plot below for questions 15 and 16.

The table and scatter plot both show the heights and weights of a randomly selected sample of football players from an all-star team.

Height (in inches)	Weight (in pounds)
67	150
68	180
69	175
70	180
71	190
72	185
72	200
73	210
74	200
75	220



15. *Draw the line of best fit.* What is the equation of the line of best fit? What is the correlation coefficient?
- 

16. Graph that line on the scatter plot above. How good a fit is the line?
-

200

Name: \_\_\_\_\_ Date: \_\_\_\_\_

## Scatter Plots and Line of Best Fit

The **best fitting line or curve** is the line that lies as close as possible to all the data points.

**Regression** is a method used to find the equation of the best fitting line or curve.

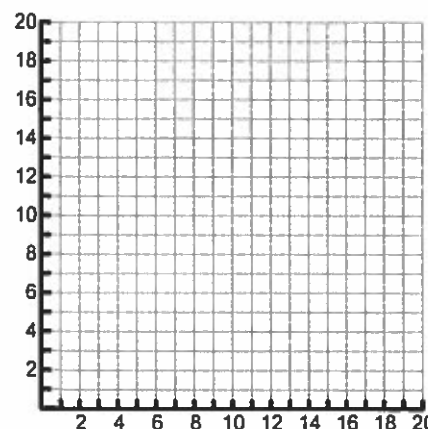
**Extrapolation** – the use of the regression curve to make predictions outside the domain of values of the independent variable.

**Interpolation** – Interpolation is used to make predictions within the domain of values of the independent variable.

### Line of Best Fit by Hand:

1) The environment club is interested in the relationship between the number of canned beverages sold in the cafeteria and the number of cans that are recycled. The data they collected are listed in this chart.

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

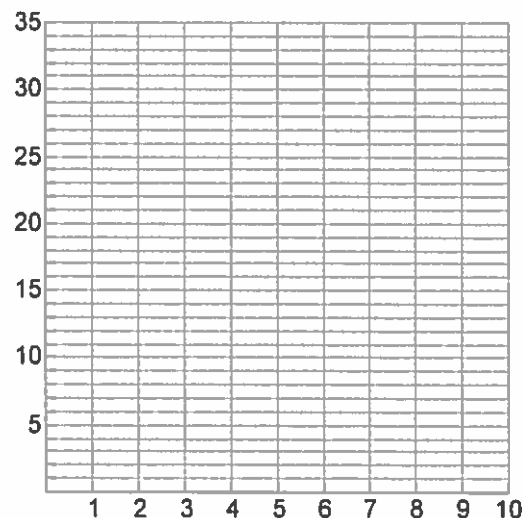


- Plot the points to make a scatter plot.
- Use a straightedge to approximate the line of best fit by hand.
- Find an equation of the line of best fit for the data.

2. Mike is riding his bike home from his grandmother's house. In the table below,  $x$  represents the number of hours Mike has been biking and  $y$  represents the number of miles Mike is away from home. Make a scatter plot for this data on the grid below.

<b>Hours (x)</b>	1	2	3	4	5	6	7	8
<b>Miles (y)</b>	35	29	26	20	16	9	6	0

- Describe the association between the data points on the scatter plot.
- Use a straightedge to approximate the line of best fit.
- Find an equation of the line of best fit for the data.
- What does the slope represent in the context of the problem? What does the  $y$ -intercept represent in the context of the problem?
- Could you use your equation to predict how far Mike would be after 10 hours? Use mathematics to justify your answer.

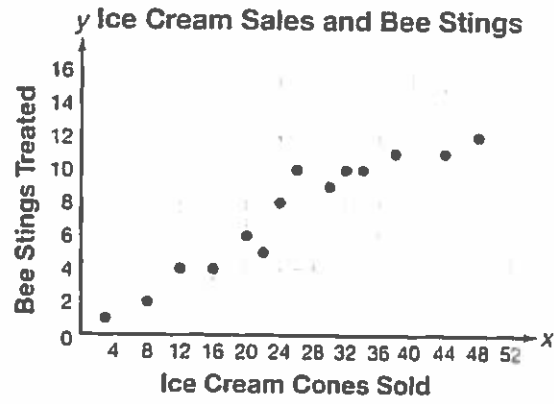


### Line of Best Fit using the calculator

- Use the table below to answer the questions about the

28

The scatter plot below shows data for the number of ice cream cones sold and the number of bee stings treated at a lake resort. Based on the data, can you conclude that eating ice cream causes bee stings? If not, what can you conclude?



---

---



# Mean Absolute Deviation (MAD)

Consider the data set: 100, 106, 180, 41, 161, 292, 116, 213

Step 1: Calculate the mean \_\_\_\_\_

Step 2: Subtract the mean from each data point and take the absolute value (all values should be positive)

\_\_\_\_\_

Step 3: Calculate the mean of the new data set \_\_\_\_\_  
(THIS IS THE M.A.D.)

\*\*The M.A.D. determines the variability of a data set. The higher the M.A.D., the higher variability of the data.

## Examples

1. 64, 68, 71, 77, 81, 82, 86, 88, 93, 93, 95, 97

Mean:

MAD:

2. 5, 9, 11, 12, 13, 15, 15, 22, 60

Mean:

MAD:

3. These are the golf scores for a tournament. Find the mean and MAD for each set.

Men	67	69	69	71	74	76
Women	68	70	72	73	74	75

---

**Mean, Median, Mode, MAD Practice WS**

---

**Find the mean, median, and modes of the following data:**

1. 6, 1, 3, 8, 5, 11, 1, 5

Mean:

Median:

Mode:

2. 15, 27, 10, 25, 9, 22, 25

Mean:

Median:

Mode:

3. 23, 6, 8, 14, 28, 8, 13, 28

Mean:

Median:

Mode:

4. 4.2, 2.2, 3.7, 2.8, 1.1

Mean:

Median:

Mode:

5. 89, 86, 96, 87, 100, 86

Mean:

Median:

Mode:

**Find the range and Mean Absolute Deviation:**

6. 10, 7, 13, 10, 8

Range:

MAD:

7. 110, 114, 104, 108, 106

Range:

MAD:

8. 87, 75, 85, 77, 74, 82

Range:

MAD:

9. 15, 17, 15, 17, 21, 17, 15, 23

Range:

MAD:

10. 58.8, 51.6, 51.9, 52, 52.5, 52.8, 53.1

Range:

MAD: