Unit 7 Notes

Statistics

Name: _____

Tentative Schedule

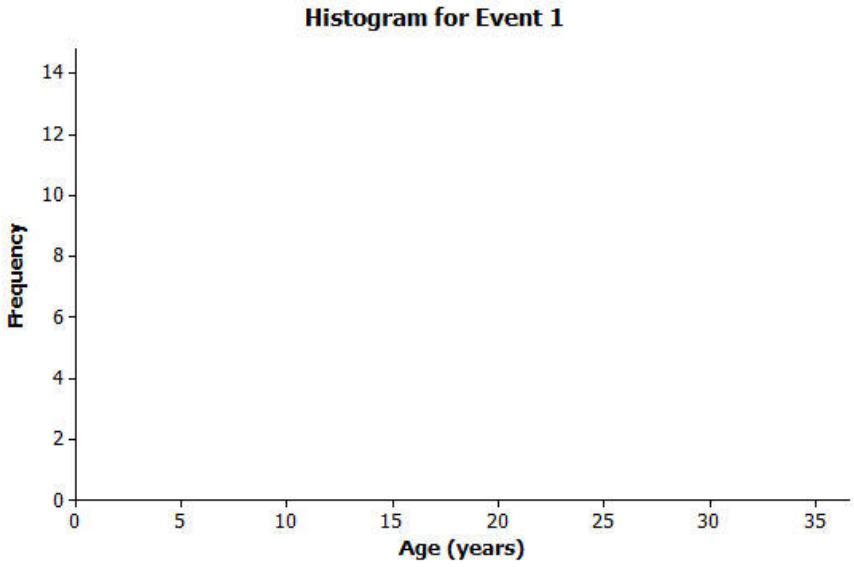
Day	*Topic*	*Assignment*
Mon. 12/15	Distributions, Shapes, and Centers **(p. 3)**	1 - 6
Tues. 12/16 Wed. 12/17	Estimating Centers and Interpreting Mean as a Balance Point **(p. 7)**	7 - 18
Thurs. 12/18	Summarizing Deviations from the Mean Measuring Variability for Symmetrical Distributions Interpreting Standard Deviation **(p. 11)**	19 - 24
Fri. 12/19 Mon. 1/5	Interpreting Standard Deviation Measuring Variability for Skewed Distributions Comparing Distributions **(p. 17)**	25 - 32
Tues. 1/6	Summarizing Bivariate Categorical Data with Relative Frequencies **(p. 23)**	33 - 42
Wed. 1/7 Thurs. 1/8	Conditional Frequencies and Association **(p. 28)** **Take-home Mid-Module Assessment due**	43 - 52
Fri. 1/9	Relationships Between Two Numerical Variables and Modeling Relationships with a Line **(p. 32)**	53 - 66
Mon. 1/12 Tues. 1/13	Interpreting Residuals from a Line More Modeling with a Line **(p. 36)**	67 - 75
Wed. 1/14	Analyzing Residuals **(p. 41)**	76 - 80
Thurs. 1/15 Fri. 1/16	Interpreting Correlation and Analyzing Data **(p. 45)**	81 - 86
Tues. 1/20	Review	Review for Test #7
Wed. 1/21 Thurs. 1/22	**Test #7**	TBA

Notes 7.1 - Distributions, Shapes, and Centers

1.) Twenty-five people were attending an event. The ages of the people are indicated below:

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

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



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

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

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

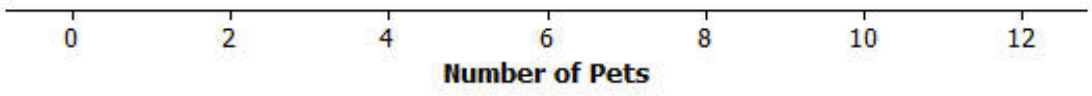
Consider the following three sets of data.

2.) Make dot plot of each of the data sets. Use the following scales:

Data Set I: Pet owners

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

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



Number of Pets

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

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

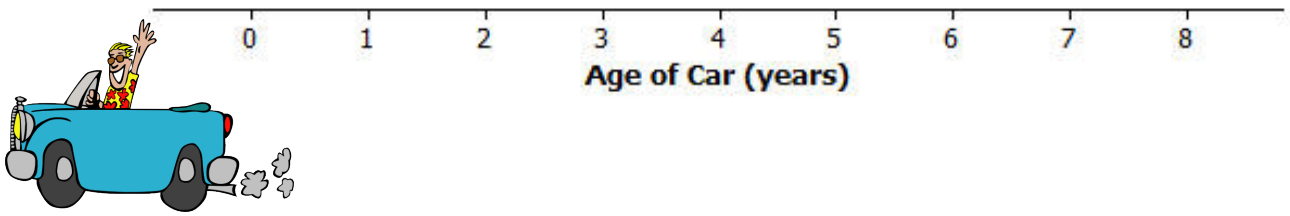
8.2	8.3	8.3	8.4	8.4	8.5	8.5	8.5	8.5	8.5
8.6	8.6	8.6	8.6	8.7	8.7	8.8	8.8	8.9	8.9

East Hallway Length Measurement (meters)

Data Set 3: Age of cars

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

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



3.) Calculate the mean number of pets owned by the 30 students from River City High School. Calculate the median number of pets owned by the thirty students.

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

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

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

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

8.) Describe the distribution of the age of cars.

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

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

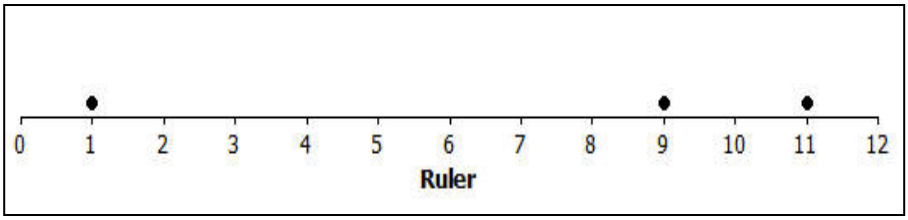
Lesson Summary:

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

Notes 7-2 - Estimating Centers and Interpreting Mean as a Balance Point

Exercises 1-7

Consider the follow example of quarters taped to lightweight ruler:



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

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

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

Position of quarter	Distance from quarter to your estimate of the balance point
1	
9	
11	

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

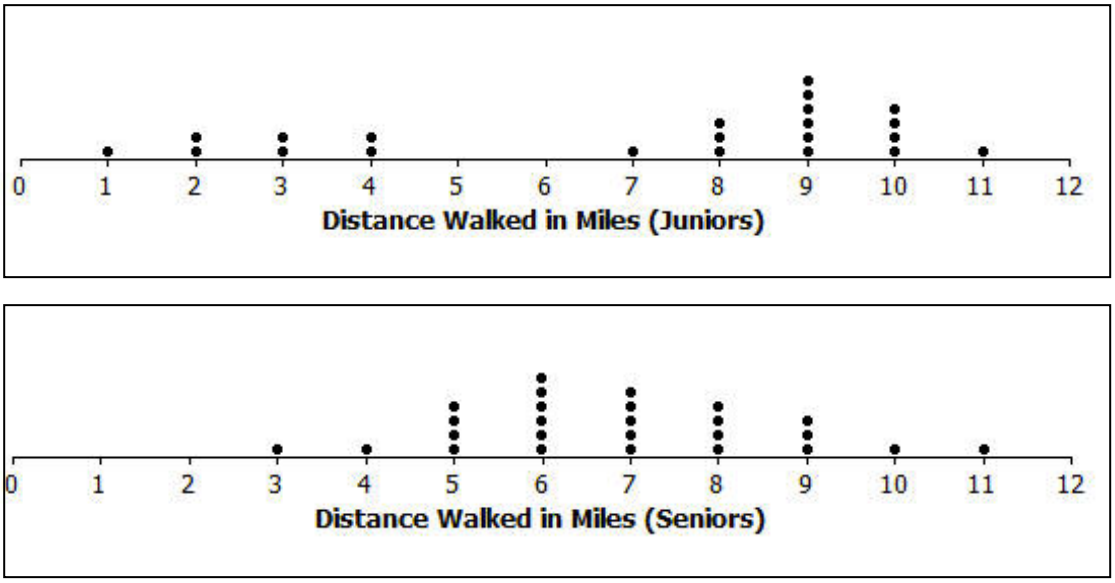
5.) What is the sum of the distances to the left of your estimate of the balance point?

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

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

Exercises 8-20

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



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

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

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

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

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

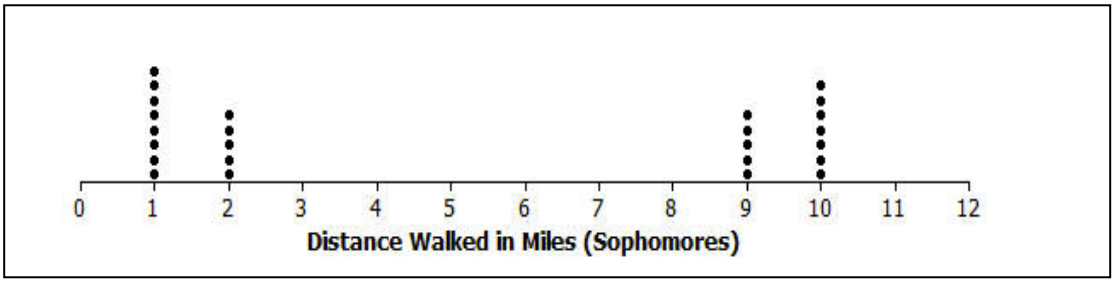
13.) What is the median of the senior data distribution?

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

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

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

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



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

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

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

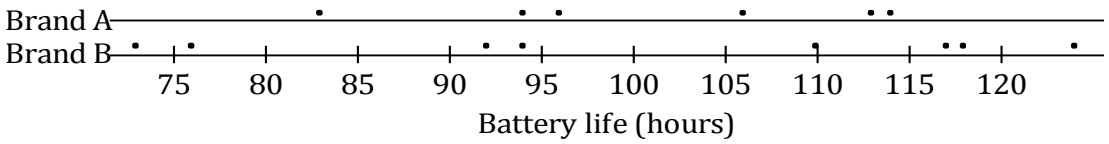
20.) How would you describe the sophomore data distribution?

Lesson Summary:

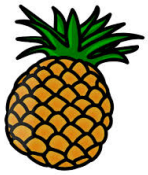
The mean of a data distribution represents a balance point for the distribution. The sum of the distances to the right of the mean is equal to the sum of the distances to the left of the mean.

Notes 7-3 - Summarizing Deviations from the Mean
Measuring Variability for Symmetrical Distributions
Interpreting Standard Deviation

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



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



To find the deviations from the mean of any value, you need to:

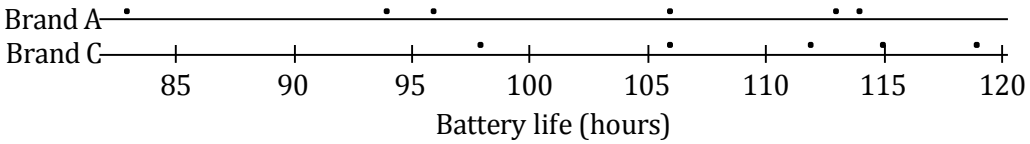
- Calculate the mean of the values. The mean is represented by $\bar{x}$.
- Find the difference between the mean and the data value.

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

Life (Hours)	83	94	96	106	113	114
Deviation from the Mean		-7				+13

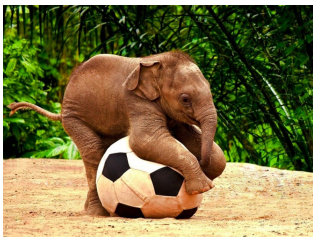
2.) Calculate the deviations from the mean for the remaining values, and write your answers in the appropriate places in the table.

The lives of 5 batteries of a third brand, Brand C, were determined. The dot plot below shows the lives of the Brand A and Brand C batteries.



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

4.) Which brand shows the greater variability?



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

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

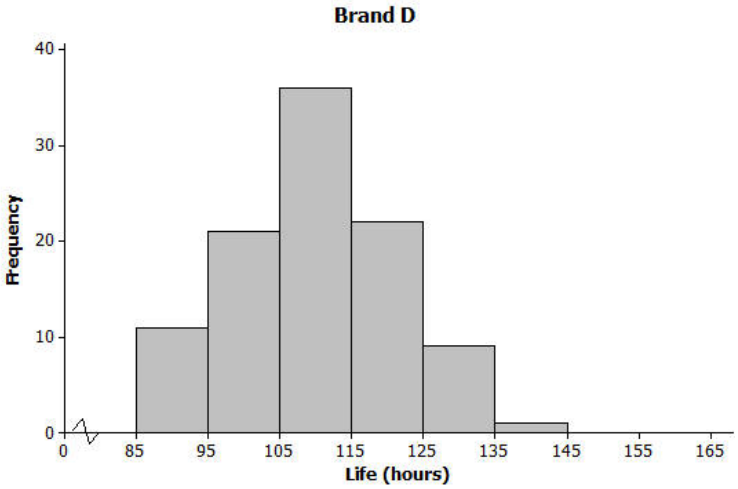
Life (Hours)	115	119	112	98	106
Deviation from the Mean					

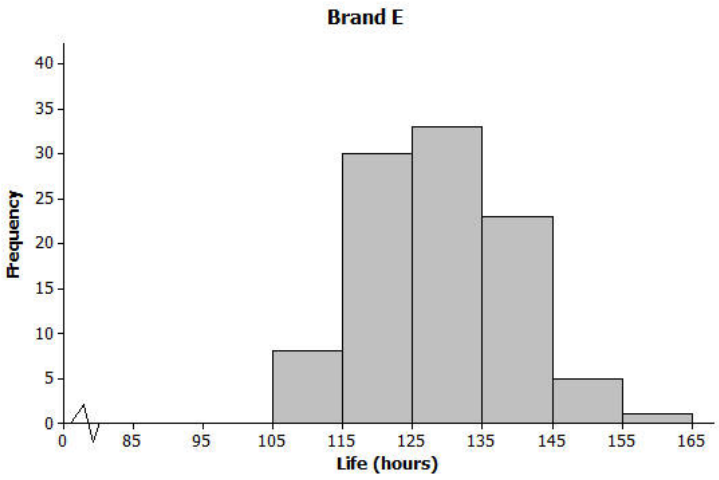
6.) Calculate the mean for Brand C. (Be sure to include a unit in your answer.)

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

8.) Ignoring the signs, are the deviations from the mean generally larger for Brand A or for Brand C? Does your answer agree with your answer to Exercise 5?

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



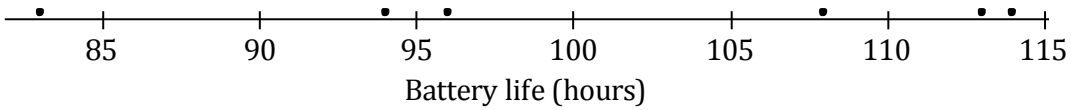


9.) Estimate the mean life for Brand D. (Do not do any calculations!)

10.) Estimate the mean life for Brand E. (Again, no calculations!)

11.) Which of Brands D and E shows the greater variability in lives? Or do you think the two brands are roughly the same in this regard?

Here's a dot plot of the lives of the Brand A batteries from the previous lesson.



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

Steps to find standard deviation by hand:

- Find each deviation from the mean.
- Square the deviations from the mean.

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

- Add up the squared deviations:
  $324 + 49 + 25 + 25 + 144 + 169 = 736$

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

- The number of values in the data set is denoted by n. In this example n is 6. You divide the sum of the squared deviations by $n - 1$, which here is $6 - 1 = 5$:

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

Take the square root of 147.2, or to the nearest hundredth is 12.13

We conclude that a typical deviation of a Brand A lifetime from the mean lifetime for Brand A is 12.13 hours. The unit of standard deviation is always the same as the unit of the original data set. So, here the standard deviation to the nearest hundredth, with the unit, is 12.13 hours.

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

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

Examples:

12.) The heights (in inches) of 9 women were as shown below.

68.4 70.9 67.4 67.7 67.1 69.2 66.0 70.3 67.6

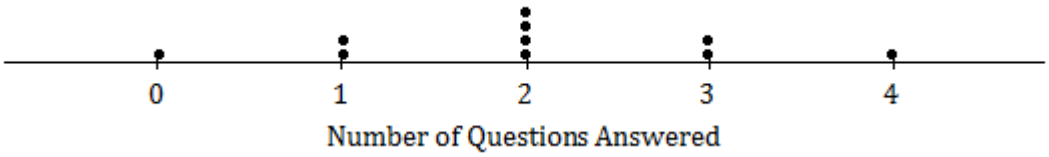
Use the statistical features of your calculator or computer software to find the mean and the standard deviation of these heights to the nearest hundredth.

Mean: _____ Standard Deviation: _____

Use the following situation for questions 13 - 16.

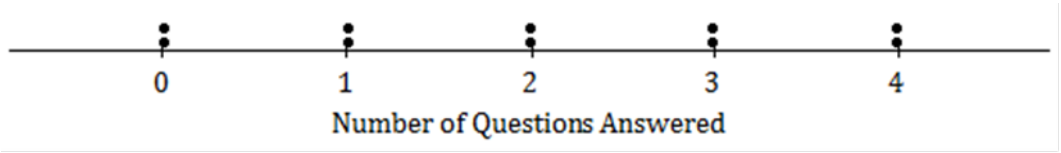
Ten people attended a talk at a conference. At the end of the talk, the attendees were given a questionnaire that consisted of four questions. The questions were optional, so it was possible that some attendees might answer none of the questions while others might answer 1, 2, 3, or all 4 of the questions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

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



Mean: _____ Standard Deviation: _____

14.) Suppose the dot plot looked like this:

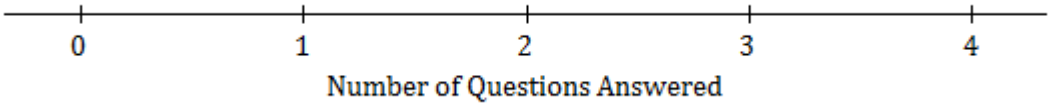


Mean: _____ Standard Deviation: _____

Remember that the size of the standard deviation is related to the size of the deviations from the mean. Explain why the standard deviation of this distribution is greater than the standard deviation in Example 13.

15.) Suppose that every person answers all four questions on the questionnaire.

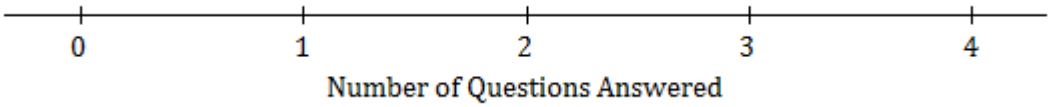
a. What would the dot plot look like?



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

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

16.) Continue to think about the situation previously described where the numbers of questions answered by each of ten people was recorded. Draw the dot plot of the distribution of possible data values that has the largest possible standard deviation. (There were ten people at the talk, so there should be ten dots in your dot plot.) Use the scale given below.



Number of Questions Answered

Explain why the distribution you have drawn would have the largest standard deviation.

Lesson Summary:

- For any given value in a data set, the deviation from the mean is the value minus the mean. Written algebraically, this is $x - \bar{x}$.
- The greater the variability (spread) of the distribution, the greater the deviations from the mean (ignoring the signs of the deviations).
- The standard deviation measures a typical deviation from the mean.
- The unit of the standard deviation is always the same as the unit of the original data set.
- The larger the standard deviation, the greater the spread (variability) of the data set.
- The size of the standard deviation is related to the sizes of the deviations from the mean. Therefore the standard deviation is minimized when all the numbers in the data set are the same, and is maximized when the deviations from the mean are made as large as possible.

Notes 7-4 - Measuring Variability in Skewed Distributions
Comparing Distributions

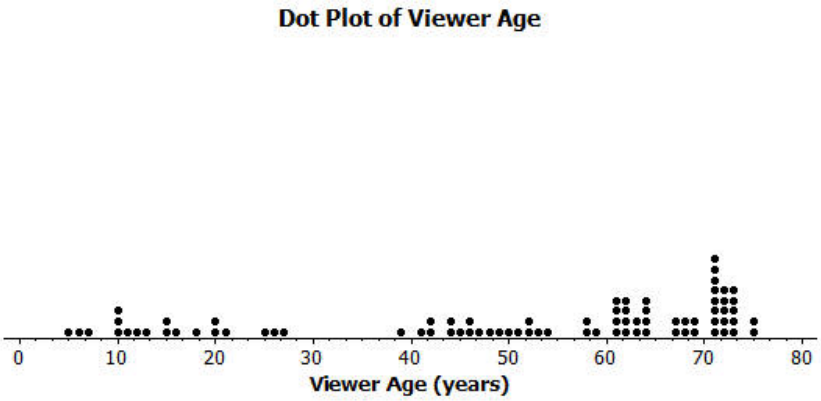
Symmetrical distribution - a distribution of data that has the same shape on both sides of the median

Skewed distribution - a distribution of data that leans either towards one end of the data values of the other; the stretched side is called a tail.

Consider the following scenario for questions 1-14.

A television game show, "Fact or Fiction", was canceled after nine shows. Many people watched the nine shows and were rather upset when it was taken off the air. A reviewer of the show called it a "cross generational show". A random sample of eighty viewers of the show was selected. Viewers in the sample responded to several questions.

The dot plot below shows the distribution of ages of these eighty viewers:



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

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

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

4.) Using the above dot plot, construct a box plot over the dot plot by completing the following steps:

a. Locate the middle 40 observations and draw a box around these values.

b. Calculate the median and then draw a line in the box at the location of the median.

c. Draw a line that extends from the upper end of the box to the largest observation in the data set.

d. Draw a line that extends from the lower edge of the box to the minimum value in the data set.

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

Minimum Age	Lower Quartile (Q1)	Median Age	Upper Quartile (Q3)	Maximum Age

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

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

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

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

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

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

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

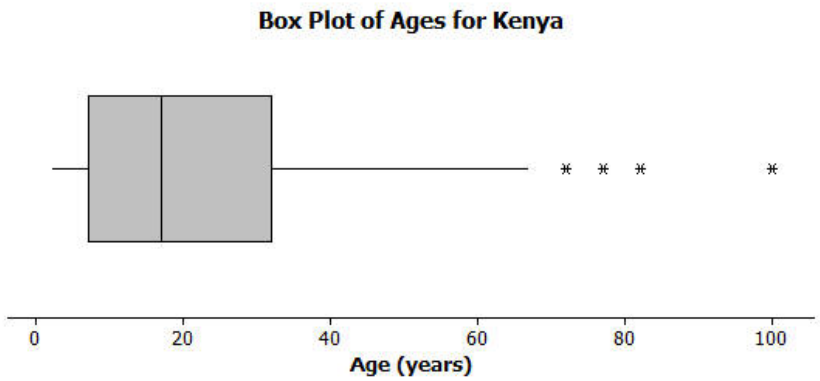
11.) What percentage of the people surveyed would be interested in the products or services suggested by the advertising table if the median age were used to describe a typical viewer?

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

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

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

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



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

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

15.) Estimate the values of the 4 ages represented by an "*".

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

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

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

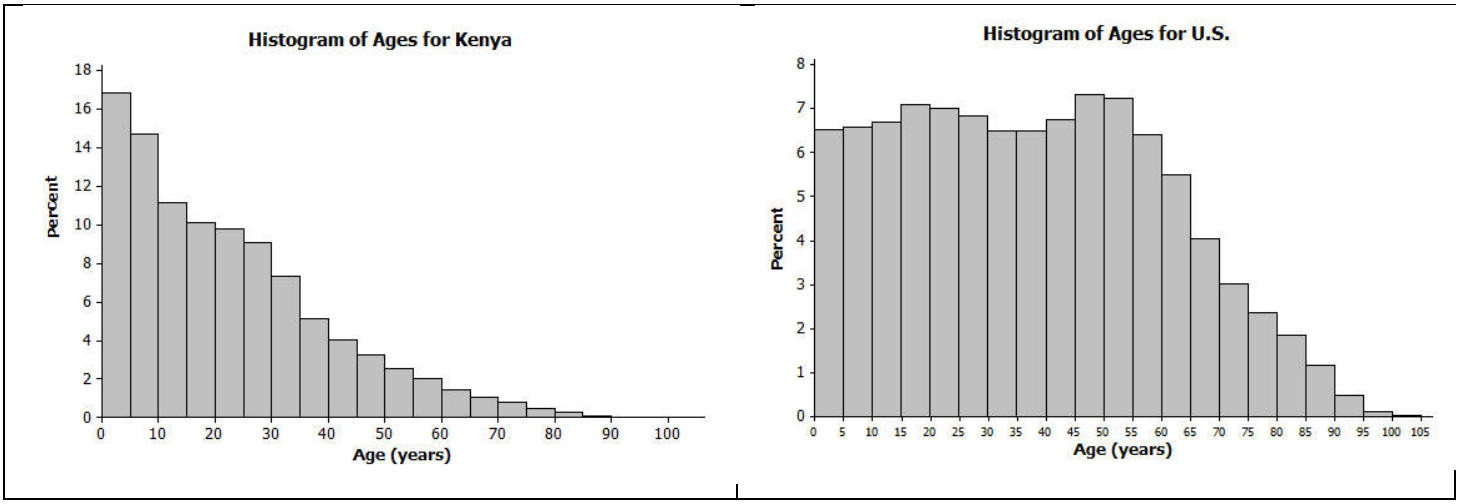
18.) Add $1.5 \times IQR$ to the 3rd quartile age (Q3). What do you notice about the four ages identified by an *?

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

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

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

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



21.) How do the shapes of the two histograms differ?

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

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

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

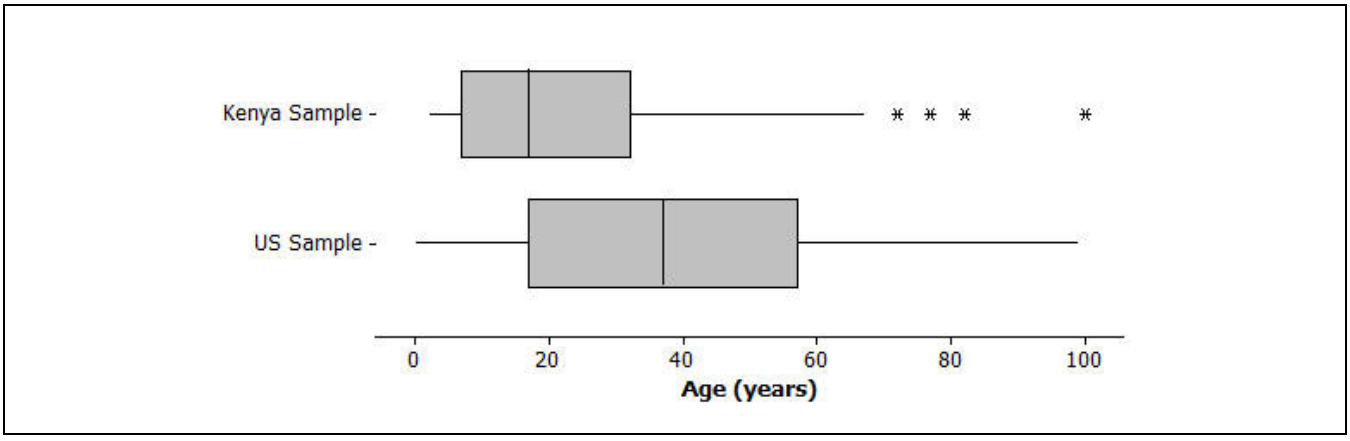
25.) Approximately what percent of people in the United States are 60 years or older?

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

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

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

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



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

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

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

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

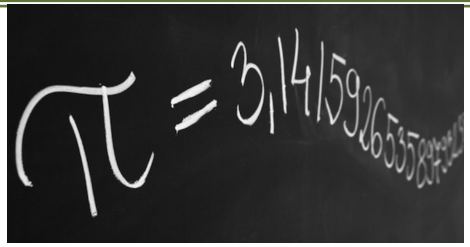
33.) Could you have estimated the mean age of a person from Kenya using the box plot? Explain your answer.

34.) The mean age of people the United States is approximately 38 years. Using the box plot, estimate the percentage of people in the United States who are younger than the mean age in the United States.

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

Lesson Summary:

- Non-symmetrical data distributions are referred to as skewed.
- Left-skewed or skewed to the left means the data spreads out longer (like a tail) on the left side.
- Right-skewed or skewed to the right means the data spreads out longer (like a tail) on the right side.
- The center of a skewed data distribution is described by the median.
- Variability of a skewed data distribution is described by the inter-quartile range (IQR).
- The IQR describes variability by specifying the length of the interval that contains the middle 50% of the data values.
- Outliers in a data set are defined as those values more than 1.5(IQR) from the nearest quartile. Outliers are usually identified by an "*" or a "·" in a box plot.
- Histograms show the general shape of a distribution.
- Box plots are created from the 5-number summary of a data set.
- A box plot identifies the median, minimum and maximum values, and the upper and lower quartiles.
- The interquartile range (IQR) describes how the data is spread around the median; it is the



Notes 7-5 - Summarizing Bivariate Categorical Data with Relative Frequencies

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

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

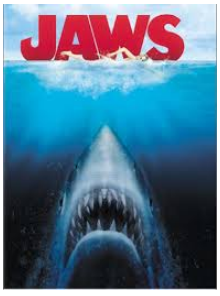
- 100 students indicated their favorite power was "to fly." 49 of those students were females.
- 131 students selected the power to "freeze time" as their favorite power. 71 of those students were males.
- 75 students selected "invisibility" as their favorite power. 48 of those students were females.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

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

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

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

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



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

The first step in analyzing the statistical question posed by the students in their mathematics class is to organize this data in a two-way frequency table. A two-way frequency table that can be used to organize the categorical data is shown below. The letters below represent the frequency counts of the cells of the table.

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

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

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

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

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

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

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

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

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

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

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

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

The relative frequency table would be found by dividing each of the above cell values by 450. For example, the relative frequency of females selecting "To Fly" is 49/ 450, or approximately 0.109 to the nearest thousandth.

A few of the other relative frequencies to the nearest thousandth are shown in the following relative frequency table:

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

13.) Calculate the remaining relative frequencies in the table above. Write the value in the table as a decimal rounded to the nearest thousandth.

14.) Which cells in this table would represent the joint relative frequencies?

15.) Which cells in the relative frequency table would represent the marginal relative frequencies?

16.) What is the joint relative frequency for females and "invisibility"? Interpret the meaning of this value.

17.) What is the marginal relative frequency for "freeze time"? Interpret the meaning of this value.

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

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

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

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

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

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

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

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

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

Lesson Summary:

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

Notes 7-6 - Conditional Frequencies and Associations

Recall the two-way table from the previous lesson:

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

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

1.) Use the frequency counts from the table to calculate the missing row conditional relative frequencies. Round the answers to the nearest thousandth.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

	Placed in Math 200	Placed in Math 100	Placed in Math 50	Total
Took Review Course				
Did not take Review Course				

12.) Based on the conditional relative frequencies, is there evidence of an association between whether or not a student takes the review course and the math course in which the student was placed? Explain your answer.

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

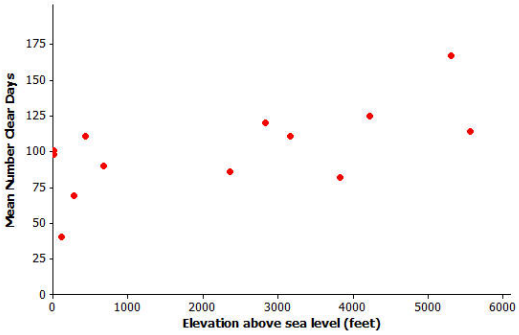
Lesson Summary

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

Notes 7-7 - Relationships Between Two Numerical Variables
Modeling Relationships with a Line

The National Climate Data Center collects data on weather conditions at various locations. They classify each day as clear, partly cloudy, or cloudy.

Here is a scatter plot of the data from 14 US cities, taken over several years, on elevation and mean number of clear days.

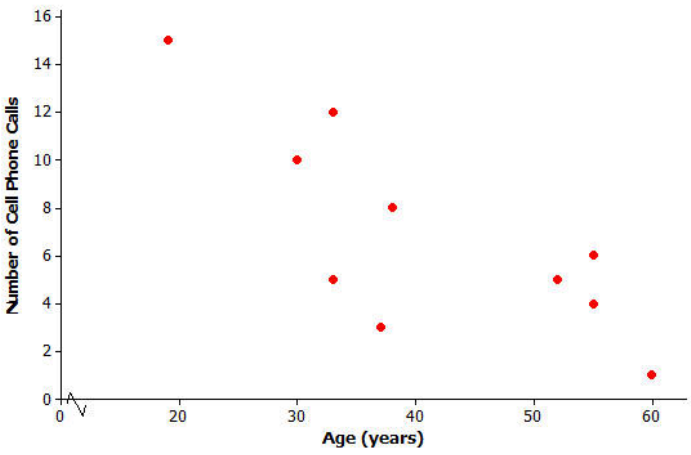


1.) Do you see a pattern in the scatter plot, or does it look like the data points are scattered?

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

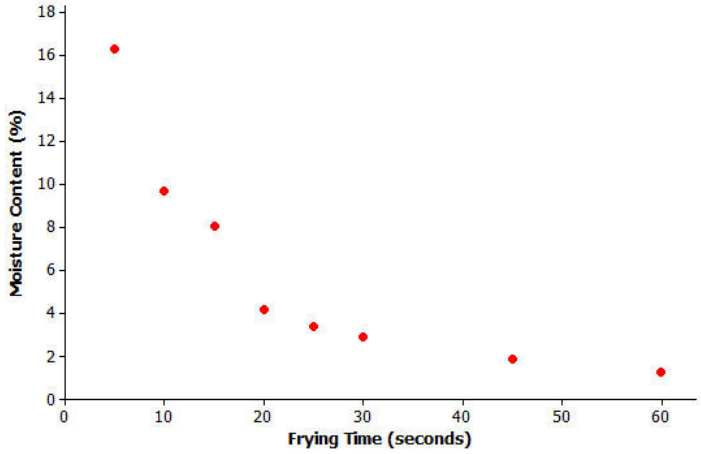
3.) Do you think that a straight line would be a good way to describe the relationship between the mean number of clear days and elevation? Why do you think this?

When a straight line provides a reasonable summary of the relationship between two numerical variables, we say that the two variables are *linearly related* or that there is a *linear relationship* between the two variables. Take a look at the scatter plots below and answer the questions that follow.



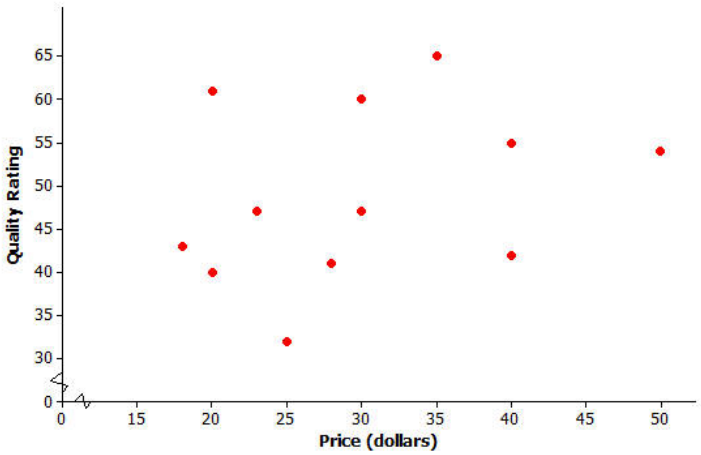
4.) Is there a relationship between number of cell phone calls and age, or does it look like the data points are scattered?

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



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

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



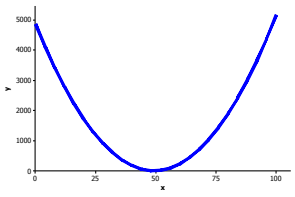
8.) Scatter plot 3 shows data for the prices of bike helmets and the quality ratings of the helmets (based on a scale that estimates helmet quality). Is there a relationship between quality rating and price, or are the data points scattered?

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

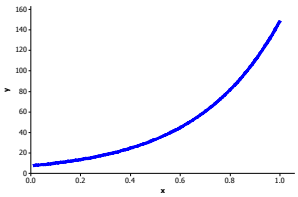
Sometimes the pattern in a scatter plot will look like the graph of a quadratic function (with the points falling roughly in the shape of a U that opens up or down), as in Graph 1 below.

In other situations, the pattern in the scatter plot might look like the graphs of exponential functions that either are upward sloping (Graph 2) or downward sloping (Graph 3):

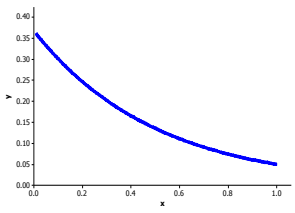
Graph 1: Quadratic



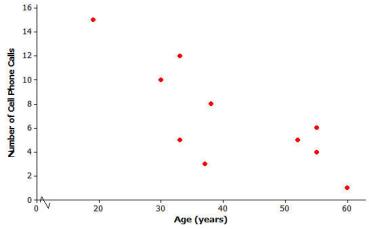
Graph 2: Exponential - upward sloping

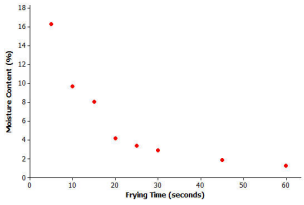


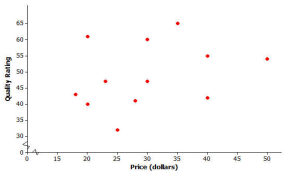
Graph 3: Exponential - downward sloping

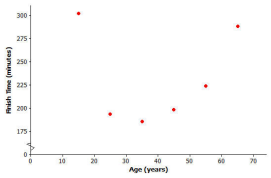


From the scatter plots below, identify which most closely shows a quadratic relationship and an exponential relationship.

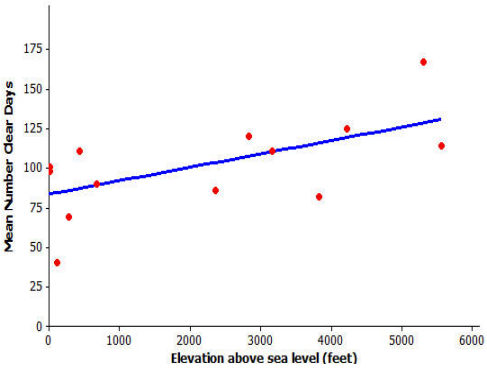








Let's revisit the data on elevation (in feet above sea level) and mean number of clear days per year. The scatter plot of this data is shown below. The plot also shows a straight line that can be used to model the relationship between elevation and mean number of clear days. The equation of this line is $y = 83.6 + 0.008x$.



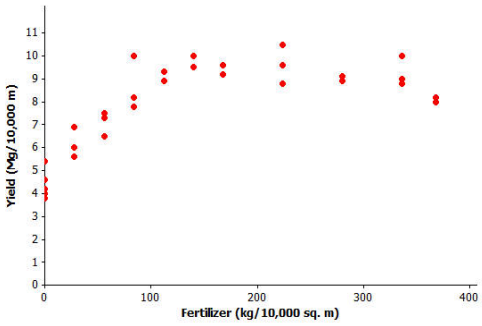
10.) Assuming that the 14 cities used in this scatter plot are representative of cities across the United States, should you see more clear days per year in Los Angeles, which is near sea level, or in Denver?

11.) One of the cities in the data set was Albany, New York, which has an elevation of 275 feet. If you did not know the mean number of clear days for Albany, what would you predict this number to be?

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

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

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

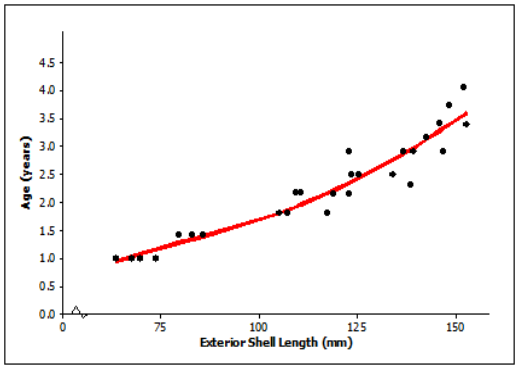


14.) The researchers who conducted this study decided to use a quadratic curve to describe the relationship between yield and amount of fertilizer. Explain why they made this choice.

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

x	Y
0	
100	
200	
300	
400	

How do you tell how old a lobster is? This question is important to biologists and to those who regulate lobster trapping. To answer this question, researchers recorded data on the shell length of 27 lobsters that were raised in a laboratory and whose ages were known. The model that the researchers used to describe the relationship is: $y = 10^{-0.403+0.0063x}$, where x represents the exterior shell length (mm) and y represents the age of the lobster (years). The exponential curve is shown on the scatter plot below.



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

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

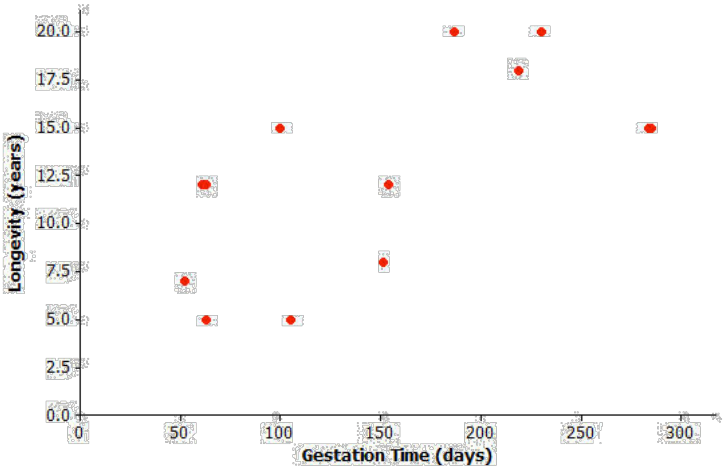
Lesson Summary

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

Notes 7-8 - Interpreting Residuals from a Line
More Modeling from a Line

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

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

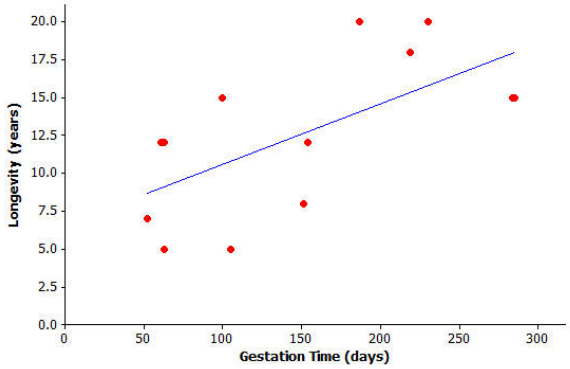


Finding the equation of the least-squares line (line of best fit) relating longevity to gestation time for these types of animal provides the equation to predict longevity. How good is the line? In other words, if you were given the gestation time for another type of animal not included in the original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

1.) Using a graphing calculator, verify that the equation of the least-squares line is $y = 6.642 + 0.03974x$

where x represents the gestation time (in days) and y represents longevity in years.

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



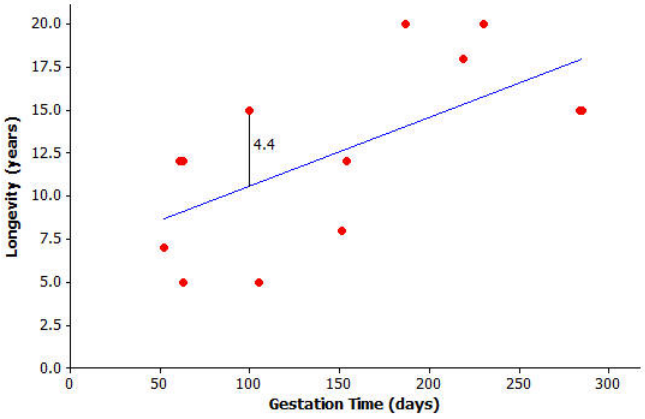
2.) Suppose a particular type of animal has a gestation time of 200 days. Approximately what value does the line predict for the longevity of that type of animal?

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

You can investigate further by looking at the types of animal included in the original data set. Take the lion, for example. Its gestation time is 100 days. You also know that its longevity is 15 years, but what does the least-squares line *predict* for the lion's longevity? Substituting x = 100 days into the equation, you get:
$y = 6.642 + 0.03974(100)$ or approximately 10.6. The least-squares line predicts the lion's longevity to be approximately 10.6 years.

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

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



5.) Let's continue to think about the gestation times and longevities of animals. Let's specifically investigate how accurately the least-squares line predicted the longevity of the black bear.

a.) What is the gestation time for the black bear?

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

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

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

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

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

6.) Repeat this activity for the sheep.

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

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

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

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

actual value - predicted value

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

residual = actual y value - predicted y value

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

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

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

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

$$y = 6.642 + 0.03974(270)$$

$$= 17.4 \text{ years.}$$

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

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

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

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

9.) Predict the longevity of the ocelot.

10.) Based on the residuals, would you be surprised to find that the longevity of the ocelot was 2 years? Why, or why not? What do you think might be a sensible range of values of the actual longevity of the ocelot?

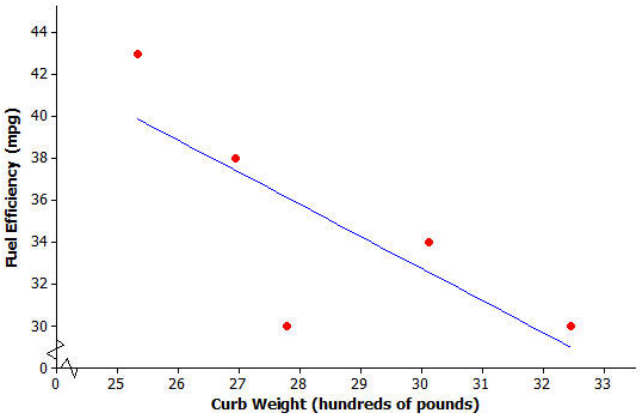
11.) You are now told that the actual longevity of the ocelot is 9 years. What is the residual for the ocelot?

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

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

Using a calculator, the least-squares line for this data set was found to have the equation $y = 78.62 - 1.5290x$ where x is the curb weight (in hundreds of pounds) and y is the predicted fuel efficiency (in miles per gallon).

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



12.) Will the residual for the car whose curb weight is 25.33 be positive or negative?

13.) Will the residual for the car whose curb weight is 27.79 be positive or negative?

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

Substitute $x = 25.33$ into the equation of the least-squares line to find the predicted fuel efficiency.

$y = 78.62 - 1.5290(25.33)$
$= 39.9$

Now calculate the residual.

residual = actual y value - predicted y value
= 43 - 39.9
= 3.1 mpg

Substitute $x = 27.79$ into the equation of the least-squares line to find the predicted fuel efficiency.

$y = 78.62 - 1.5290(27.79)$
$= 36.1$

Now calculate the residual.

residual = actual y value - predicted y value
= 30 - 36.1
= -6.1 mpg

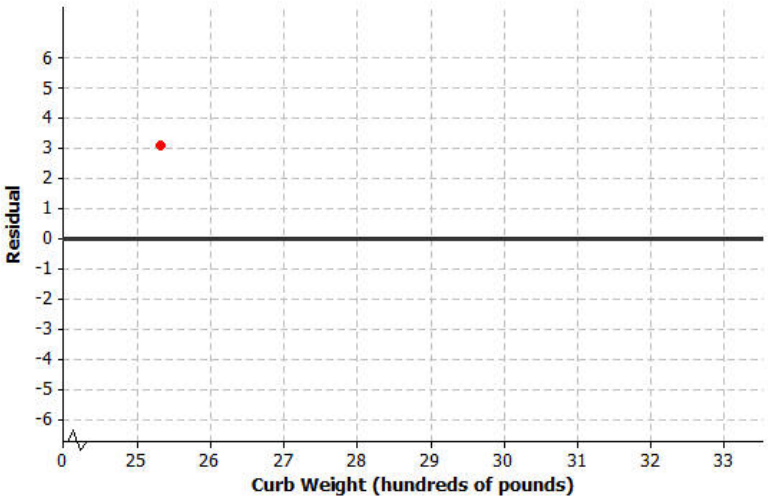
The residuals for all 5 cars have been written in the table below.

Curb weight (100 lb)	Fuel Efficiency (miles per gallon)	Residual
25.33	43	3.1
26.94	38	0.6
27.79	30	-6.1
30.12	34	1.4
32.47	30	1.0

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

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

The axes and this first point are shown below.



Plot the other four residuals in the residual plot above.

Lesson Summary

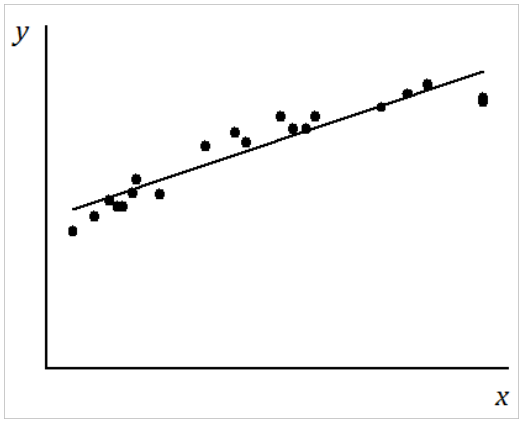
- The least-squares line is the line that is used to best model a linear relationship between the data.
- On the graph, the residuals are the vertical distances of the points from the least-squares line.
- The residuals give us an idea how close a prediction might be when the least-squares line is used to make a prediction for a value that is not included in the data set.
- The predicted y-value is calculated using the equation of the least-squares line.
- The residual is calculated using:

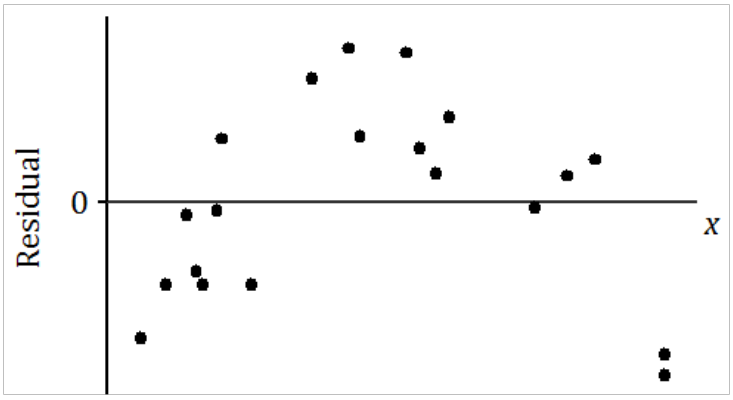
$$\text{residual} = \text{actual } y \text{ value} - \text{predicted } y \text{ value}$$

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

Notes 7-9 - Analyzing Residuals

Suppose you are given the scatter plot and least-squares line below. Also, the residual plot is shown.

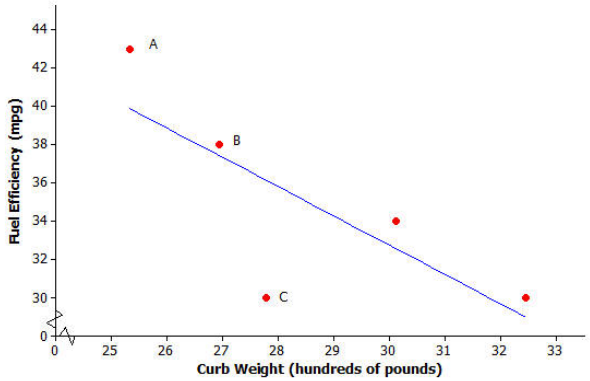




Why is looking at the pattern in the residual plot important?

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

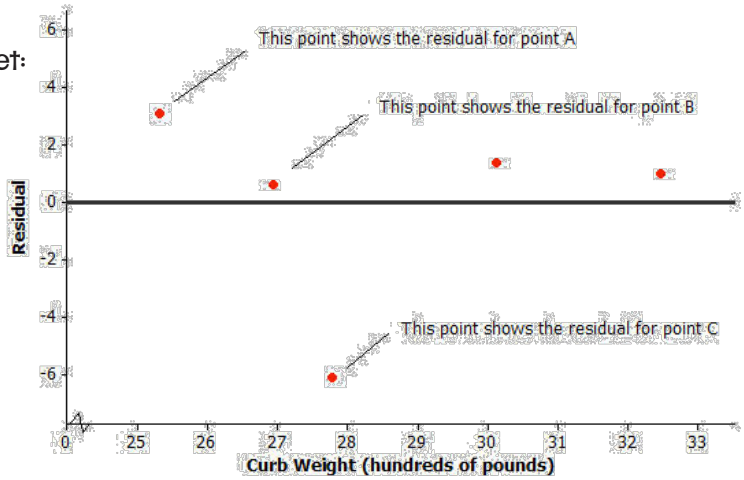
In the previous lesson, you looked at a scatter plot showing how fuel efficiency was related to curb weight for five compact cars. The scatter plot and least-squares line are shown below.



Consider the following questions:

- What kind of residual will Point A have?
- What kind of residual will Point B have?
- What kind of residual will Point C have?

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



You can use a graphing calculator or graphing program to construct a scatter plot and a residual plot.

Example:

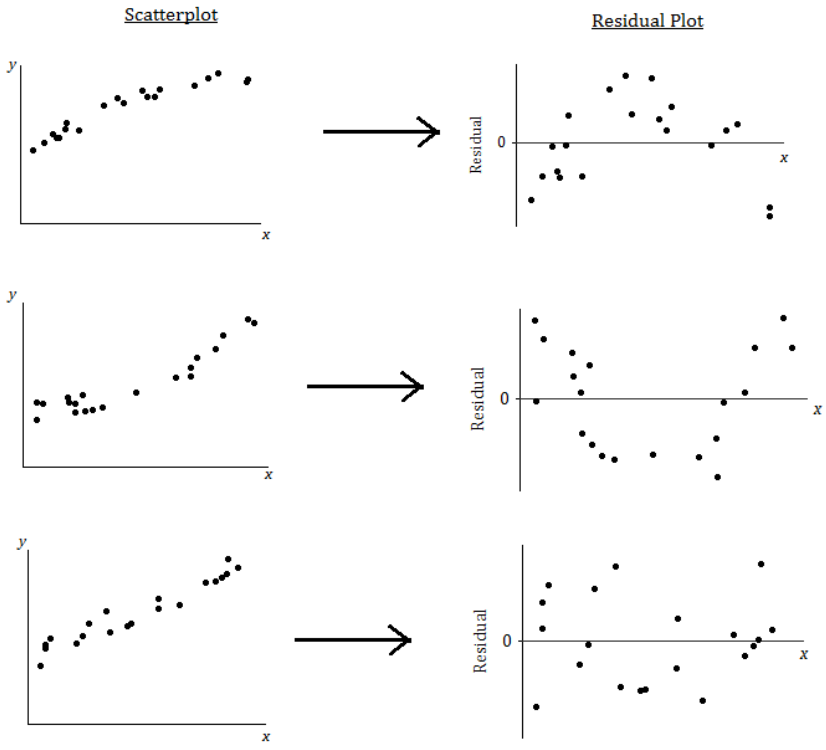
Kendra likes to watch crime scene investigation shows on television. She watched a show where investigators used a shoe print to help identify a suspect in a case. She questioned how possible it is to predict someone's height is from his or her shoe print.

To investigate, she collected data on shoe length (in inches) and height (in inches) from 12 adult women. Her data appear in the table.

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

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

The previous lesson shows that when a line is fit, a scatter plot with a curved pattern produces a residual plot that shows a clear curve. You also saw that when a line is fit, a scatter plot where the points show a straight-line pattern results in a residual plot where the points are randomly scattered. Our previous findings are summarized in the plots below:



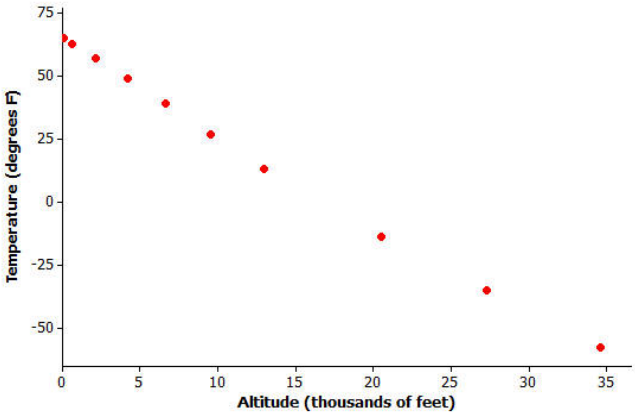
What does it mean when there is a curved pattern in the residual plot?

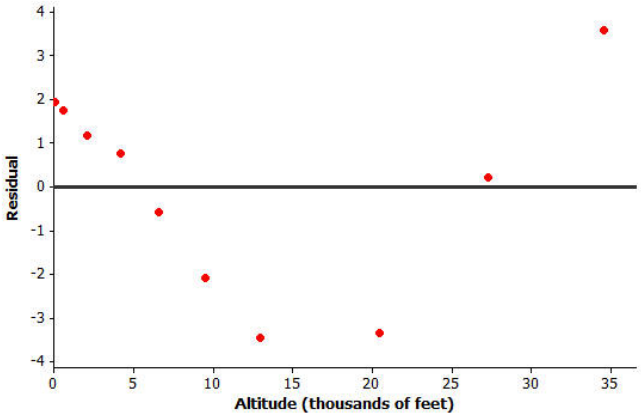
What does it mean when the points in the residual plot appear to be scattered at random with no visible pattern?

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

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

However, look at the residual plot:





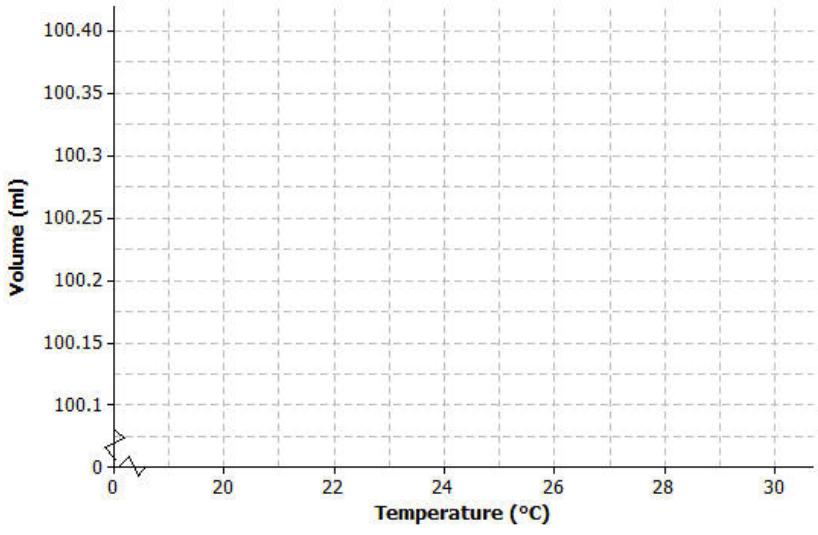
There is a clear curve in the residual plot. So what appeared to be a linear relationship in the original scatter plot was, in fact, a nonlinear (curved) relationship.

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

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

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

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



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

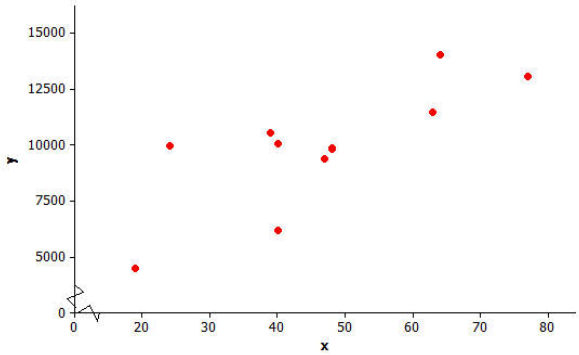
3.) Do you see a clear curve in the residual plot? What does this say about the original data set?

Lesson Summary

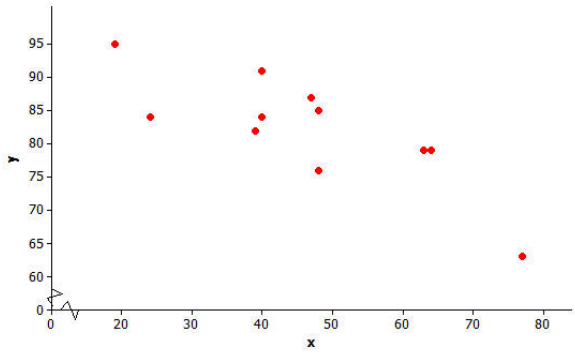
- After fitting a line, the residual plot can be constructed using a graphing calculator.
- A curve or pattern in the residual plot indicates a curved (nonlinear) relationship in the original data set.
- A random scatter of points in the residual plot indicates a linear relationship in the original data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

Notes 7-10 - Interpreting Correlation and Analyzing Data

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

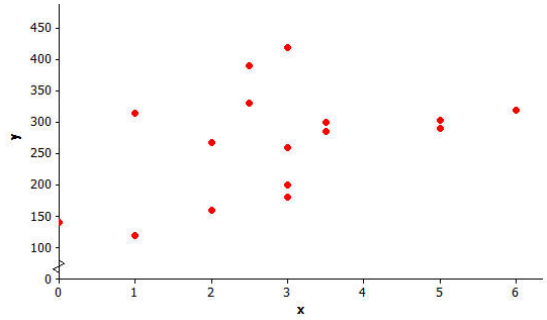


- Positive Relationship
- As x-values increase, y-values increase
- Positive Slope

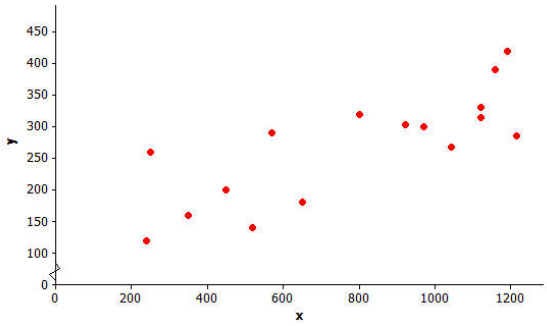


- Negative Relationship
- As x-values increase, y-values decrease
- Negative Slope

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

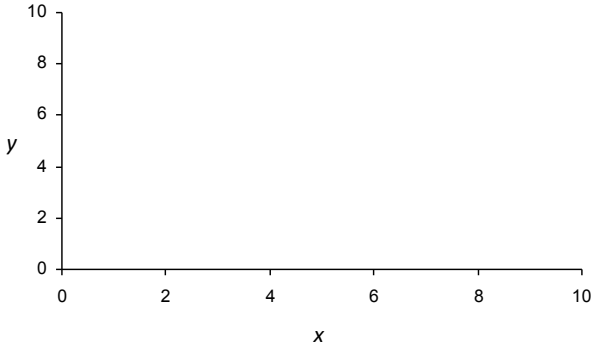


- Positive Relationship
- Weak Relationship (points are widely scattered)
- Positive Slope



- Positive Relationship
- Stronger Relationship (points are closer together)
- Positive Slope

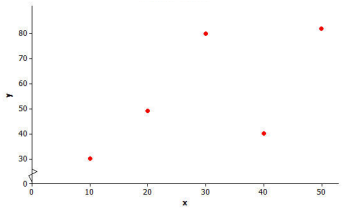
What do you think a scatter plot that shows the strongest possible positive linear relationship would look like? Draw a scatter plot with 5 points that illustrates this.



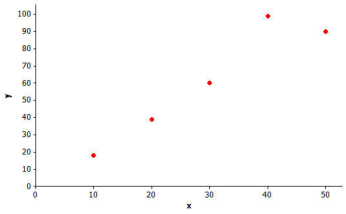
Consider the three scatter plots below. Place them in order from the one that shows the strongest linear relationship to the one that shows the weakest linear relationship.

Strongest	⇒	Weakest

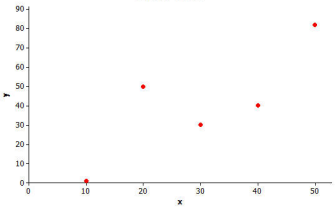
Scatter Plot 1



Scatter Plot 2

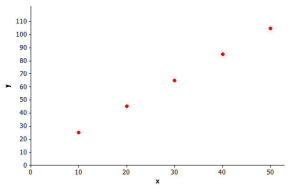


Scatter Plot 3

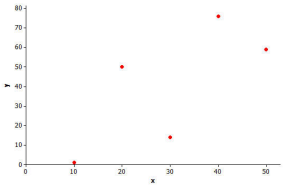


The correlation coefficient is a number between -1 and +1 (including -1 and +1) that measures the strength and direction of a linear relationship. The correlation coefficient is denoted by the letter r.

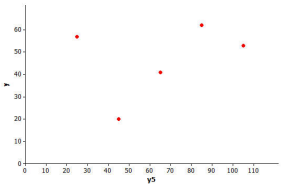
Several scatter plots are shown below. The value of the correlation coefficient for the data displayed in each plot is also given.



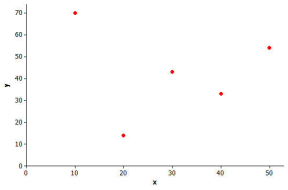
r = 1.00



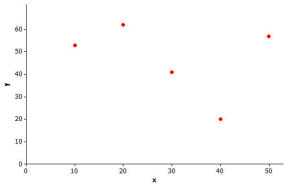
r = 0.71



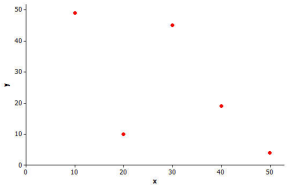
r = 0.32



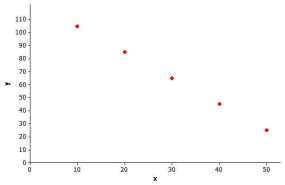
r = −0.10



r = −0.32



r = −0.63



r = −1.00

1.) When is the value of the correlation coefficient positive?

2.) When is the value of the correlation coefficient negative?

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

The properties of the correlation coefficient are as follows:

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

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

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

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

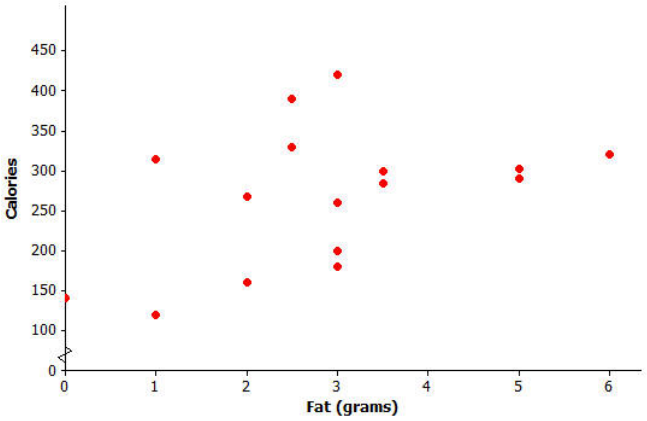
The table below shows how you can informally interpret the value of a correlation coefficient.

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

Examples:

Consumer Reports published a study of fast-food items. The table and scatter plot below display the fat content (in grams) and number of calories per serving for 16 fast-food items.

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

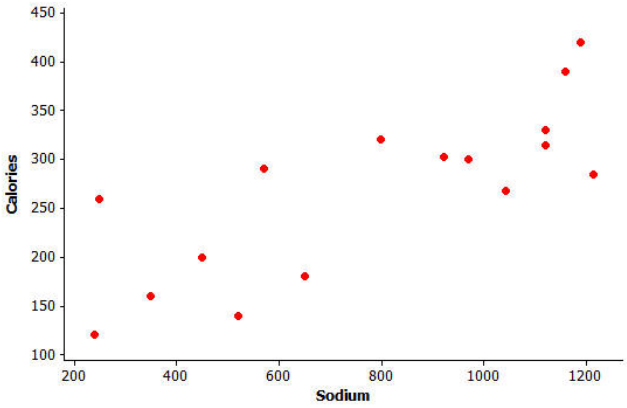


4.) Based on the scatter plot, do you think that the value of the correlation coefficient between fat content and calories per serving will be positive or negative? Explain why you made this choice.

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

The *Consumer Reports* study also collected data on sodium content (in mg) and number of calories per serving for the same 16 fast food items. The data is represented in the table and scatter plot below.

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



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

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

8.) For these 16 fast-food items, is the linear relationship between fat content and number of calories stronger or weaker than the linear relationship between sodium content and number of calories?

Correlation Does Not Mean There Is a Cause-and-Effect Relationship Between Variables!

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

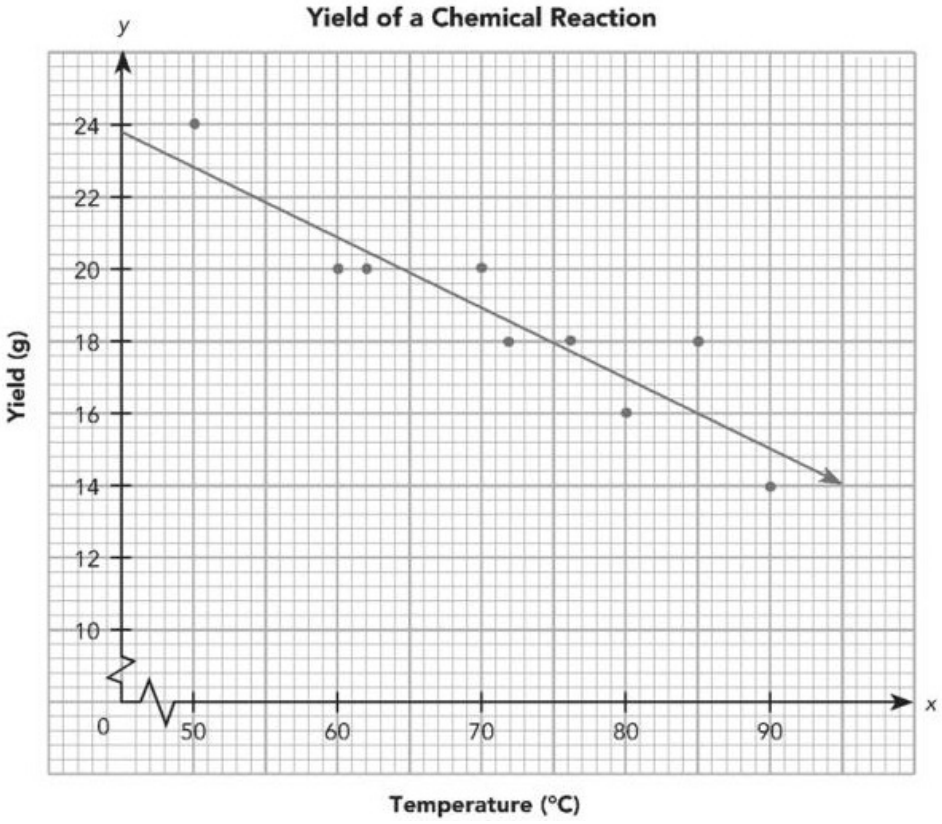
For example, the value of the correlation coefficient between sodium content and number of calories for the fast food items in the previous example was $r = 0.79$, indicating a strong positive relationship. This means that the items with higher sodium content tend to have a higher number of calories.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

Lesson Summary

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

Lab Notes - Lines of Best Fit

Line of Best Fit:



1.) Write an equation for a line of best fit.

2.) Use your equation to estimate the yield, in grams, of a chemical reaction when the temperature is 90°.

3.) Use your equation to estimate the temperature when the yield of a chemical reaction is 20 g.

To understand the relationship between area of living space within a home, x square feet, and cost of electricity, y dollars, data are collected for a particular month and recorded.

Area of Living Space (x square feet)	1,400	1,000	1,300	1,100	1,500	1,300
Electricity Cost (y dollars)	210	200	226	206	258	228

Area of Living Space (x square feet)	1,200	1,100	1,400	1,200	1,400
Electricity Cost (y dollars)	223	212	242	215	246

8. Use the graph paper on the next page. Construct the scatter plot. Use 1 centimeter on the horizontal axis to represent 100 square feet for the x interval from 1,000 to 1,500. Use 2 centimeters on the vertical axis to represent \$10 for the y interval from 200 to 260.

9. Sketch a line of best fit.

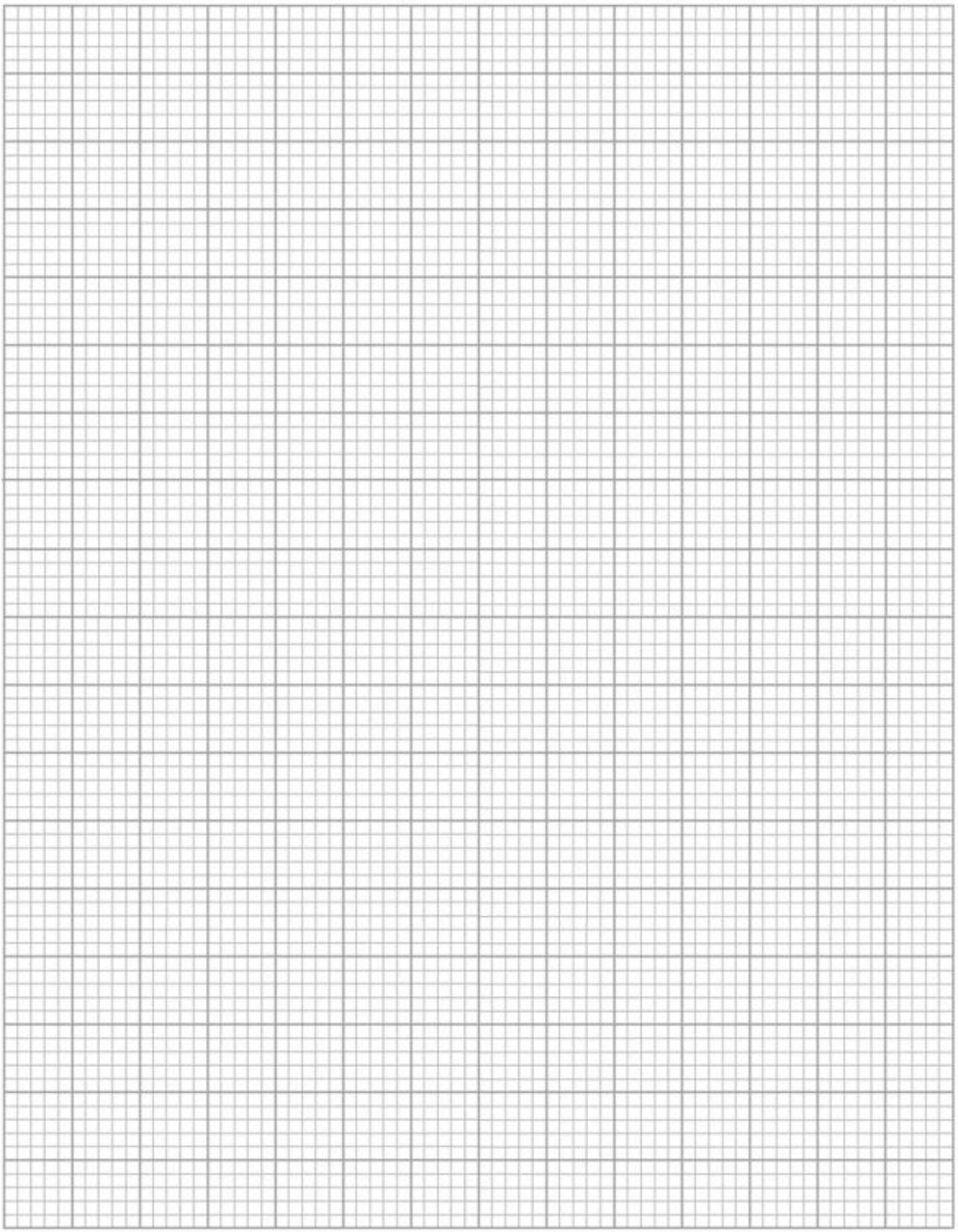
10. Find an equation for the line of best fit.

11. Describe the association between area of living space within a home and cost of electricity.

12. Identify the outlier. Give a likely explanation of the occurrence of the outlier.

13. Predict the cost for electricity for a 1,350 square-foot home.

14. Predict the area of living space within a home given an electricity cost of \$230.



The table shows the average rainfall, in millimeters per day, and the average number of hours of sunshine per day for a small town recorded over a period of 6 months.

Rainfall (x mm)	0.1	0.2	0.4	0.5	0.6	0.8
Sunshine (y hours)	2.5	3.0	3.5	3.0	2.0	1.5

Construct a scatter plot for this data. Use 1 centimeter on the horizontal axis to represent 0.1 millimeter of rainfall and 1 centimeter on the vertical axis to represent 0.5 hour.

The scatter plot displays bivariate data on length of leg, *x*, in centimeters, of each students and the distance of their single step, *y*, in centimeters.