

Eureka Math™ Homework Helper

2015–2016

Algebra I Module 2 *Lessons 1–19*

Eureka Math, A Story of Functions®

Published by the non-profit Great Minds.

Copyright © 2015 Great Minds. No part of this work may be reproduced, distributed, modified, sold, or commercialized, in whole or in part, without consent of the copyright holder. Please see our [User Agreement](#) for more information. “Great Minds” and “Eureka Math” are registered trademarks of Great Minds.

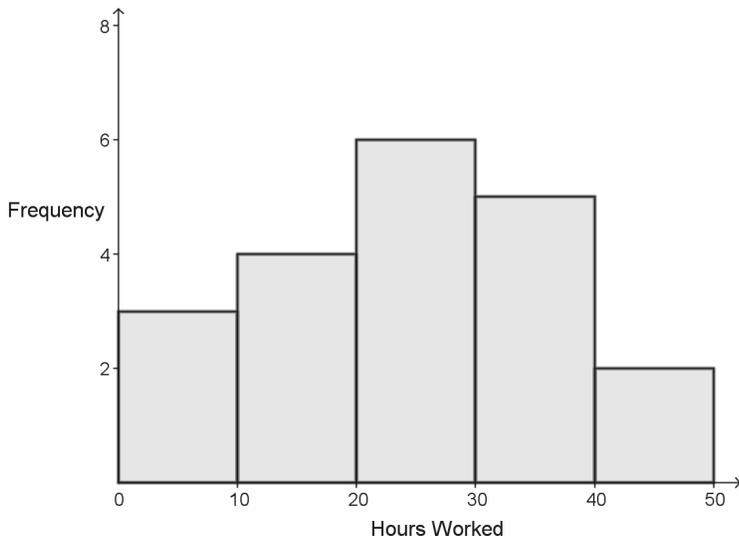
Lesson 1: Distributions and Their Shapes

Create a Histogram to Represent a Distribution

The hours worked for 20 employees are listed below:

5, 5, 8, 10, 12, 12, 15, 20, 20, 20, 22, 24, 25, 30, 30, 31, 35, 39, 40, 41

1. Create a histogram of the hours.



I need to group the data into intervals. The hours range from 5 to 41 hours. I think I will use 10 hour intervals with the first interval being 0 hours up to but not including 10 hours. Then, I need to count the number of data points in each interval to create the height of each bar.

Understand and Answer Questions about the Data

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

The graph is symmetrical. Most people work between 20 and 30 hours, and there are roughly equal numbers of people who work more or less than that amount.

Symmetrical graphs have the most frequent data points in the middle of the distribution. Skewed graphs have the most frequent data points on the left or right end of the distribution.

7. Identify the typical hours worked by employees of this business.

Just over half of the employees work 40 or more hours per week.

I need to think of a type of business where most people would work a typical 40 hour work week. When comparing the histograms, I need to think about what is the same and what is different. Since I used the same hours intervals, it is easier to compare the distribution of the two data sets.

8. What type of business might employ people who work these types of hours? Use the histogram to justify your answer.

This is a larger business where more workers appear to work a typical 40 hour work week. Perhaps this is a small manufacturing company, an insurance agency, or a small bank where employees keep full-time, regular hours.

Compare Two Distributions and Their Graphs

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

The major differences are in the center and distribution of the data. One is symmetric with the hours distributed evenly around the center of the data, and the other is skewed with most employees working 40 or more hours per week. Each hours interval had an entry for the first company but no one at the second company worked from 30 up to 40 hours per week. The second set of data was a much larger set, so the frequencies in each interval are larger since both graphs use the same hours intervals.

Lesson 2: Describing the Center of a Distribution

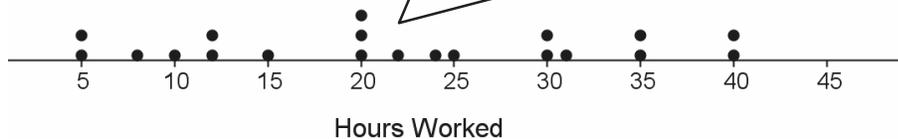
Create a Dot Plot to Represent a Distribution

The hours worked for 20 employees are listed below:

5, 5, 8, 10, 12, 12, 15, 20, 20, 20, 22, 24, 25, 30, 30, 31, 35, 35, 40, 40

1. Create a dot plot of the hours worked by the 20 employees.

I need to make a number line that includes the highest and lowest values in the data set. I need to scale the number line so I can easily plot the data values, so I will scale it by 5's. Repeated data get stacked on top of one another. Since there are 3 people that worked 20 hours, I need 3 dots at 20. I need to label the graph.



Calculate the Mean and Median of a Data Set

2. What is the mean of this data set?

Add the values in the data set, and divide this sum by the number of values. Since there were 20 employees, I need to divide the sum total of hours by 20.

$$\frac{5 + 5 + 8 + 10 + 12 + 12 + 15 + 20 + 20 + 20 + 22 + 24 + 25 + 30 + 30 + 31 + 35 + 35 + 40 + 40}{20} = 21.95$$

The mean hours worked is 21.95.

I need to decide whether to use the mean or the median. This distribution is skewed to the left, so the median would be a better choice. This data set is already listed in order from least to greatest, so I do not need to worry about rewriting it in numerical order to find the median. Out of 53 elements, the middle one is the 27th.

6. How many hours per week is typical for employees of this company? Explain how you determined your answer.

Since this distribution is skewed to the left, the median is a better choice to describe the center of the data set. The median hours is 40.

7. Why would it be difficult to report a typical number of hours if we combined the hours employees worked at both companies?

The distribution would have two points that appeared to be the center of the distribution, one around 20 to 24 hours and another around 40 hours. The two companies have very different patterns of hours worked, so combining the two may not accurately represent the trends at either company.

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

Balance Point

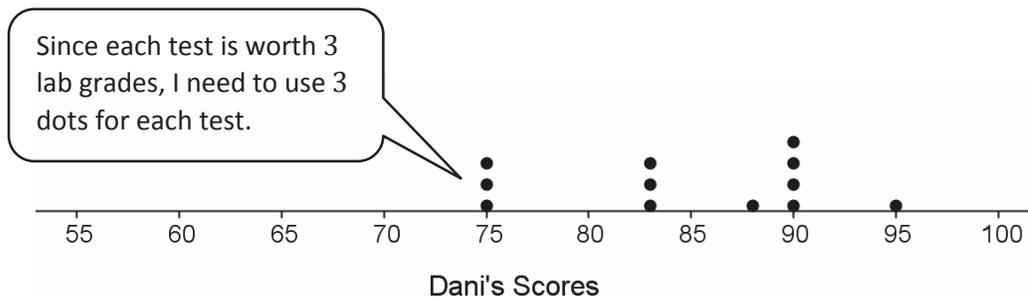
Create a Dot Plot to Represent a Distribution and Estimate a Balance Point

In Mr. Moreno's science class, each test is worth the same amount as three lab reports. Dani earned 75%, 83%, and 90% on her tests, and she earned 90%, 88%, and 95% on her lab reports.

Here is a number line that can be used to plot Dani's grades in science class.



- On the number line, create a dot plot of Dani's science grades. Let one "•" symbol represent one lab report score.



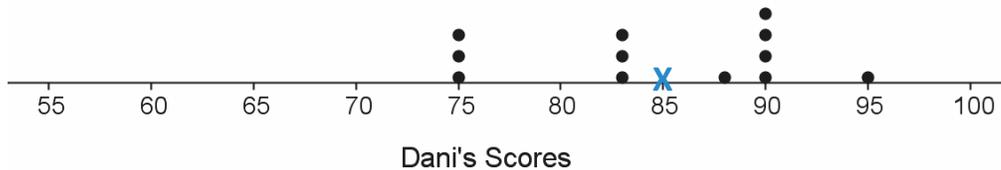
I already counted the tests as three labs, so they are weighted appropriately.

- To be eligible for the Honors Award, Dani's weighted average in the class must be 85% or higher. Do you think Dani will get the award? Explain your answer.

She has the same number of scores above 85% as she does below 85%. She might be just underneath that average because three of her scores were 75%.

3. Place an X on the number line at a position that you think indicates the balance point of all the “●” symbols.

The X should be located so that the sums of the distances above and below the X and each data value are equal.



4. Determine the sum of the distances from the X to each “●” to the left. Determine the sum of the distance from the X to each “●” to the right. Then, explain whether or not you need to adjust your balance point based on these sums.

The distance from 85 to 75 is 10.

$$85 - 75 = 10$$

The distance from 85 to 83 is 2.

$$85 - 83 = 2$$

To find the distance, I can subtract the smaller number from the larger one or use absolute value to get a positive number.

The sum of the distances to the left of 85 is $3 \cdot 10 + 3 \cdot 2$ which is equal to 36.

The distance from 88 to 85 is 3.

$$|85 - 88| = 3$$

The distance from 90 to 85 is 5.

$$|85 - 90| = 5$$

The distance from 95 to 85 is 10.

$$|85 - 95| = 10$$

Since there were four 5's, I can multiply each 5 by 4 and then add to the other distances.

The sum of the distances to the right of 85 is $3 + 4 \cdot 5 + 10$ which is equal to 33.

The balance point should be slightly lower than 85 since the sum of the distances to the left was greater than the sum of the distances to the right.

The balance point is where the sums are equal. I have too large a sum to the left, so the point needs to be lower.

Calculate a Weighted Average and Compare Means and Balance Points

If some data points in a distribution are worth more than others, then a weighted average can be used to describe the center of the distribution. Each data point is counted according to its weight. In this case, tests are worth three lab report grades, so we can multiply each test score by three. When calculating a weighted average, you need to account for the different “weights” of each score, which is why the example below divides the total by 12 instead of 6.

5. Based on these test and lab report grades, what is Dani’s weighted average?

Each test is worth 3 lab reports, so multiply those scores by 3. Then, count a total of 12 grades when calculating the mean.

$$\frac{(3 \cdot 75) + (3 \cdot 83) + (3 \cdot 90) + 88 + 90 + 95}{12} = 84.75$$

6. How does the calculated mean compare with your estimated balance point.

They are very close with the mean slightly below the estimated balance point.

Lesson 4: Summarizing Deviations from the Mean

Calculate Deviations from the Mean

The vertical jump in inches of 26 players in an NBA draft is given in the table below. (Data set from Core Math Tools, www.nctm.org)

32	33	33	34	36	36	37	37	37	38	38	38	38
38	38	38	38	38	38	39	39	39	39	40	41	43

It's quicker to write the sum using multiplication when values are repeated.

- Calculate the mean vertical jump for these players in the NBA draft.

$$\frac{32 + 2 \cdot 33 + 34 + 2 \cdot 36 + 3 \cdot 37 + 10 \cdot 38 + 4 \cdot 39 + 40 + 41 + 43}{26} = 37.5$$

- Calculate the deviations from the mean for these vertical jumps, and write your answers in the table below.

Vertical Jump	32	33	33	34	36	36	37	37	37	38	38	38	38
Deviation from the Mean	-5.5	-4.5	-4.5	-3.5	-1.5	-1.5	-0.5	-0.5	-0.5	0.5	0.5	0.5	0.5
Vertical Jump	38	38	38	38	38	38	39	39	39	39	40	41	43
Deviation from the Mean	0.5	0.5	0.5	0.5	0.5	0.5	1.5	1.5	1.5	1.5	2.5	3.5	5.5

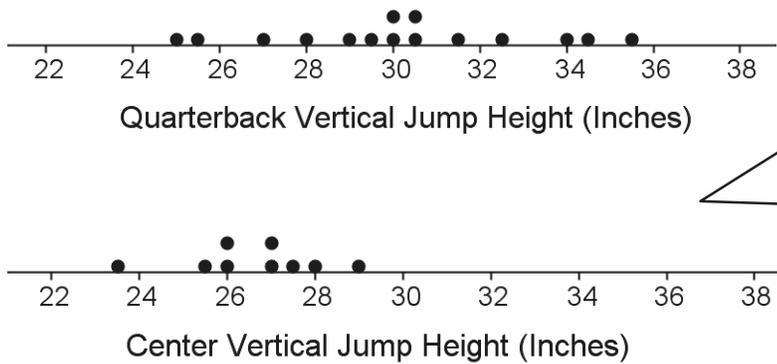
- Write an expression for the deviation from the mean for a jump height of 32 inches.

$$32 - 37.5$$

We don't use absolute value for deviation from the mean, so the deviations will be either positive, negative, or zero depending on whether the value is above, below, or at the mean.

Compare the Variability of Two Distributions Considering Deviations from the Mean

The vertical jumps of the quarterbacks and the centers from a recent NFL combine are shown on the dot plots below.



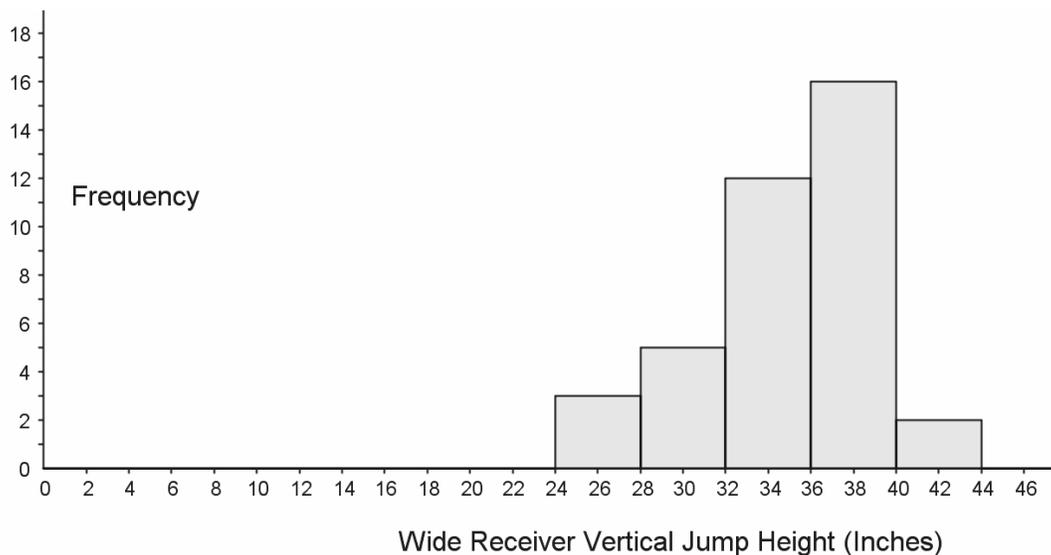
I need to look at how far away each data point would be from the mean. Since the distributions are fairly symmetrical, the mean will be near the middle.

4. Based on the data, which position, quarterback or center, has the greatest deviation from the mean?

The quarterback position has the greater deviation from the mean because the jump heights range from 25 to nearly 36 inches. For the centers, the jump heights range from around 23 to 29 inches.

Estimate Mean and Deviation from the Mean Given a Histogram

The vertical jumps of the wide receivers from a recent NFL combine are shown on the histogram below.



5. How many wide receivers jumped around 34 inches?

Twelve wide receivers jumped around 34 inches.

In a histogram, the height of the bar represents the number of values in each interval.

6. How many wide receivers participated in the combine in total?

The height of the first bar is 3, which represents 3 players in that interval. The sum of the frequencies in each interval is $3 + 5 + 12 + 16 + 2$ which is equal to 38. There were 38 receivers in the combine.

7. Suppose the three players represented by the bar centered at 26 inches each jumped exactly 26 inches and the players in the next bar each jumped exactly 30 inches, and so on. If you were to add up all the jump heights, what result would you get?

$$3 \cdot 26 + 5 \cdot 30 + 12 \cdot 34 + 16 \cdot 38 + 2 \cdot 42 = 1328$$

I can multiply the middle value in each bar by the frequency and add.

8. What is the mean jump height for the wide receivers?

$$\frac{1328}{38} \approx 34.9$$

The mean jump height is approximately 34.9 inches.

9. What is a typical deviation from the mean for this data set? Explain your reasoning.

A typical deviation from the mean would be 4 to 6 inches. Each bar represents a 4-inch interval of jumps. Most jumps are within 4 inches of the mean, but a few are outside of that range, so the typical deviation from the mean should be a little bit greater than 4 inches.

I need to think about how far each value of the data set would be from 34.9 inches. Based on the bar heights (frequency), I can see that almost all the heights were within 4 to 6 inches of the mean.

Lesson 5: Measuring Variability for Symmetrical Distributions

Calculate the Standard Deviation of a Data Set

1. Ten of the members of a high school boys' basketball team were asked how many hours they studied in a typical week. Their responses (in hours) are shown in the table below.

Number of Hours Studied	20	12	9	6	13	10	14	11	11	12
Deviation from the Mean	8.2	0.2	-2.8	-5.8	1.2	-1.8	2.2	-0.8	-0.8	0.2
Squared Deviation from the Mean	67.24	0.04	7.84	33.64	1.44	3.24	4.84	0.64	0.64	0.04

a. Calculate the mean study time for data set.

$$\frac{20 + 12 + 9 + 6 + 13 + 10 + 14 + 11 + 11 + 12}{10} = 11.8$$

Find the sum of the data values, and divide by the total number of players, which is 10.

b. Calculate the deviations from the mean, and write your answers in the second row of the table.

First entry: $20 - 11.8 = 8.2$
Second entry: $12 - 11.8 = 0.2$
Third entry: $9 - 11.8 = -2.8$

I need to subtract the mean from each value in the data set.

Continue these calculations; the remaining values are in the second row of the table above.

c. Square the deviations from the mean, and write them in the third row of the table.

First entry: $8.2^2 = 8.2 \cdot 8.2 = 67.24$
Second entry: $0.2^2 = 0.2 \cdot 0.2 = 0.04$

Squaring a number means to multiply that number by itself.

Continue these calculations; the remaining values are in the third row of the table above.

d. Find the sum of the square deviations.

The sum means to add all of the numbers.

$$67.24 + 0.04 + 7.84 + 33.64 + 1.44 + 3.24 + 4.84 + 0.64 + 0.64 + 0.04 = 119.6$$

- e. What is the value of n for this data set?

Divide the sum of the squared deviations by $n - 1$.

$$n = 10$$

$$\frac{119.6}{9} = 13.2\bar{8}$$

n stands for the number of values in the data set. We divide by one less than that number.

- f. Take the square root of your answer to part (e).

Round your answer to the nearest hundredth.

$$\sqrt{13.2\bar{8}} \approx 3.65$$

The nearest hundredth means two places after the decimal. Use the " \approx " symbol when rounding. I will need to use a calculator to compute the square root.

2. Find the standard deviation of the following data set: 3, 5, 10, 23, 23, 30, 34, 40.

Find the mean.

$$\text{Mean} = \frac{3+5+10+23+23+30+34+40}{8} = 21$$

Find the deviations from the mean.

$$3 - 21 = -18$$

$$5 - 21 = -16$$

$$10 - 21 = -11$$

$$23 - 21 = 2$$

$$23 - 21 = 2$$

$$30 - 21 = 9$$

$$34 - 21 = 13$$

$$40 - 21 = 19$$

I can follow the steps below to find standard deviation:

- Find the mean.
- Find the deviations from the mean.
- Square the deviations.
- Sum the squares of the deviations.
- Divide by $n - 1$.
- Take the square root.

Sum the square of the deviations.

$$(-18)^2 + (-16)^2 + (-11)^2 + 2^2 + 2^2 + 9^2 + 13^2 + 19^2 = 1320$$

The sum of the squares is divided by $n - 1$.

$$\frac{1320}{7} \approx 188.57$$

The standard deviation is the square root of the number.

$$\sqrt{188.57} \approx 13.73$$

Standard deviation is greater when a data set has more variability. It is a measure of the spread of the data.

3. Which data set, the one from Exercise 1 or the one from Exercise 2, has the greatest spread (variability)?

The one from Exercise 2 because it had the larger standard deviation.

Lesson 6: Interpreting Standard Deviation

Calculate the Mean and Standard Deviation Using a TI-83 or TI-84 Calculator

Instructions may vary based on the type of calculator or software used. The instructions below are based on a Texas Instruments TI-83 or TI-84 calculator using data stored in L1.

- From the home screen, press STAT, ENTER to access the stat editor.

L1	L2	L3	1
████████	-----	-----	
L1(1)=			

After pressing STAT and ENTER, I see this screen, and I can start typing the data values in the L1 column.

- If there are already numbers in L1, clear the data from L1 by moving the cursor to "L1" and pressing CLEAR, ENTER.
- Move the cursor to the first entry of L1, type the first data value, and press ENTER. Continue entering the remaining data values to L1 in the same way.

L1	L2	L3	1
20	-----	-----	
12			
9			
6			
████████			
L1(5)=			

This is what the data set {20,12,9,6} would look like after I enter it in L1.

- Press 2ND, QUIT to return to the home screen.
- Press STAT, select CALC, select 1-Var Stats, and press ENTER.

This step is optional.

EDIT	CALC	TESTS
1:1-Var Stats		
2:2-Var Stats		
3:Med-Med		
4:LinReg(ax+b)		
5:QuadReg		
6:CubicReg		
7:QuartReg		

This step assumes I have already entered data specifically into L1.

6. The screen should now show summary statistics for your data set. The mean is the \bar{x} value, and the standard deviation for a sample is the s_x value.

This is the mean.

```

1-Var Stats
x̄=11.75
Σx=47
Σx²=661
Sx=6.020797289
σx=5.214163404
↓n=4
            
```

This is the standard deviation we use for a sample.

If data is stored in another list, it will need to be referred to after selecting 1-Var Stats in Step 5. For example, if data was entered in L2:

L1	L2	L3	2
-----	20	-----	
	12		
	9		
	6		
	13		
	10		
L2(7) =			

```

1-Var Stats L2
            
```

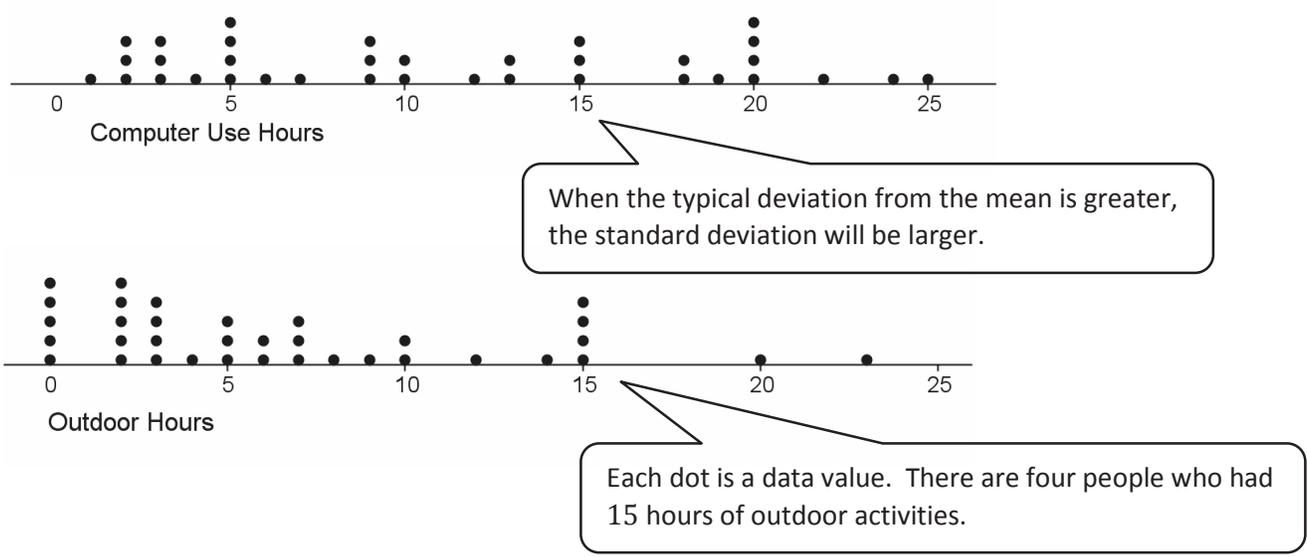
```

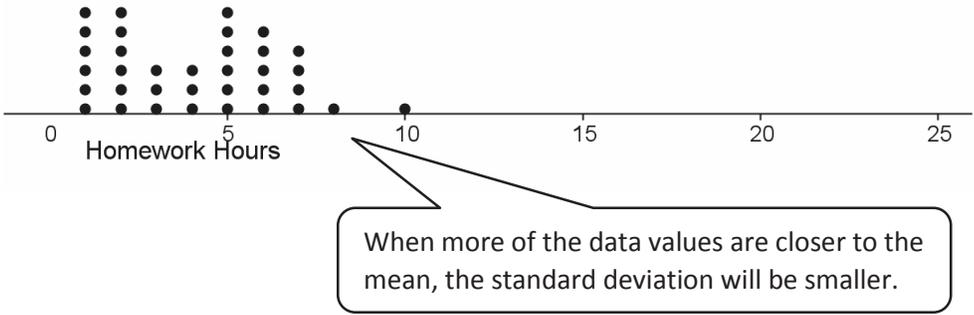
1-Var Stats
x̄=11.66666667
Σx=70
Σx²=930
Sx=4.760952286
σx=4.346134937
↓n=6
            
```

Press STAT, select CALC, select 1-Var Stats, and then refer to L2. This is done by pressing 2ND, L2 (i.e., "2ND" and then the "2" key). The screen will display 1-Var Stats L2. Then, press ENTER.

Calculating the Standard Deviation Given a Box Plot

A random sample of 35 ninth graders reported they spent the following number of hours each week on the following activities as shown in the dot plots below.





1. Which of the three activities has the smallest standard deviation? Which had the largest? Justify your answer.

Homework has the smallest standard deviation because the times are clustered around the mean (center) of the data. Either Computer Use or Outdoor Activities will have the largest standard deviation because the times are more spread out from the mean (center) of the data.

2. Estimate the mean and standard deviation of computer use hours.

The data is pretty evenly distributed, so the mean will be near the center. The mean is approximately 11. The standard deviation is a measure of typical deviation from the mean, which appears to be around 6 or 7.

3. Use a calculator to determine the mean and standard deviation of each variable, and record the information in the table below. Round answers to the nearest hundredth.

First, I need to enter each data set into a list on the calculator.

	Homework	Outdoor Activities	Computer Use
Mean	4.14	6.86	11.11
Standard Deviation	2.39	6.03	7.30

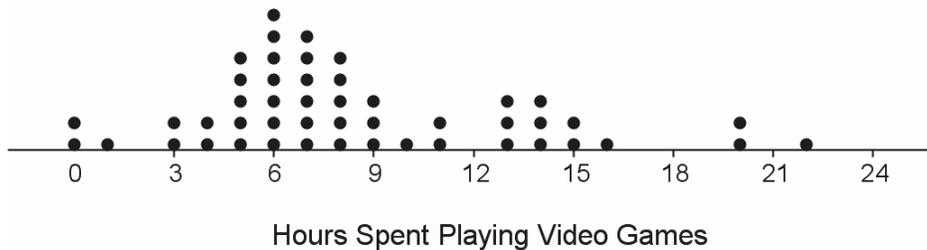
On the calculator, \bar{x} is the mean, and s_x is the standard deviation.

Then, I need to press STAT, CALC, 1-Var Stats, and ENTER for data stored in L1.

For data entered in another list, I need to type that list name after 1-Var Stats before pressing ENTER.

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

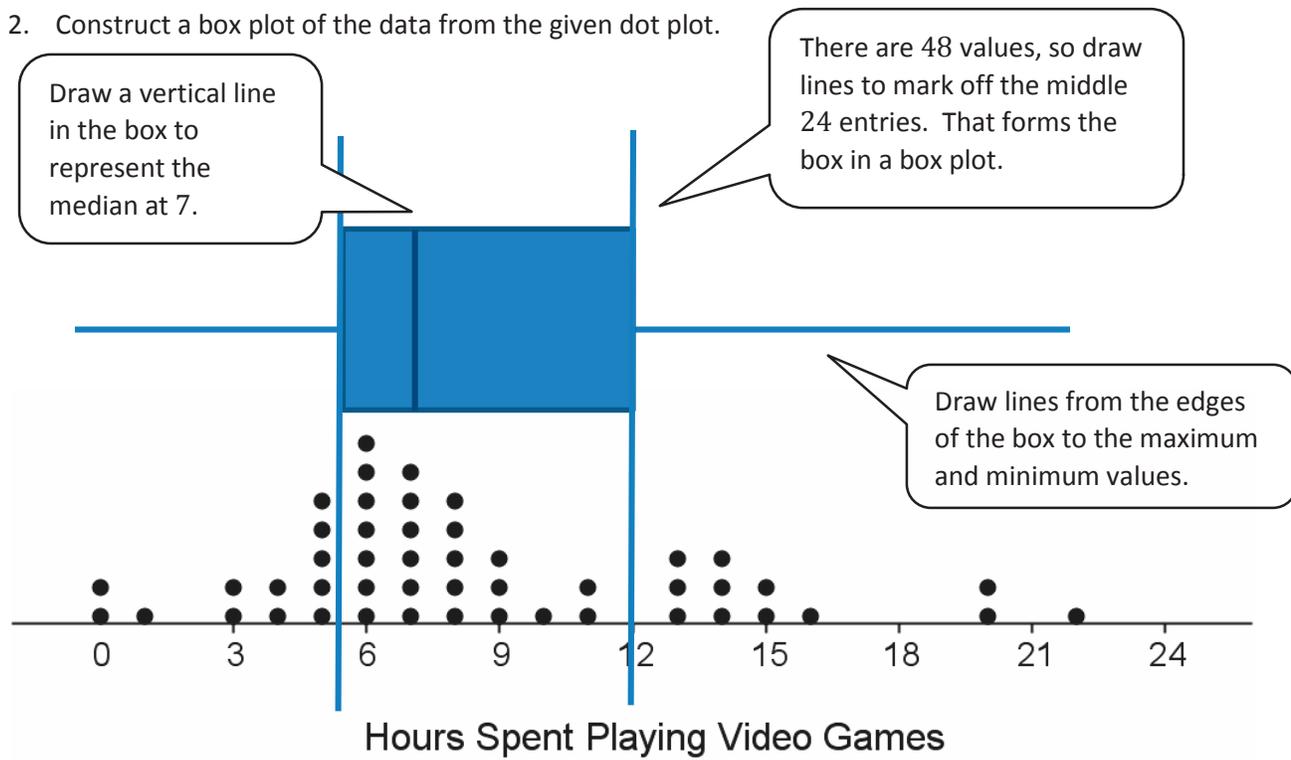
The dot plot displays the number of hours a sample of 48 ninth graders spend playing video games during one week.



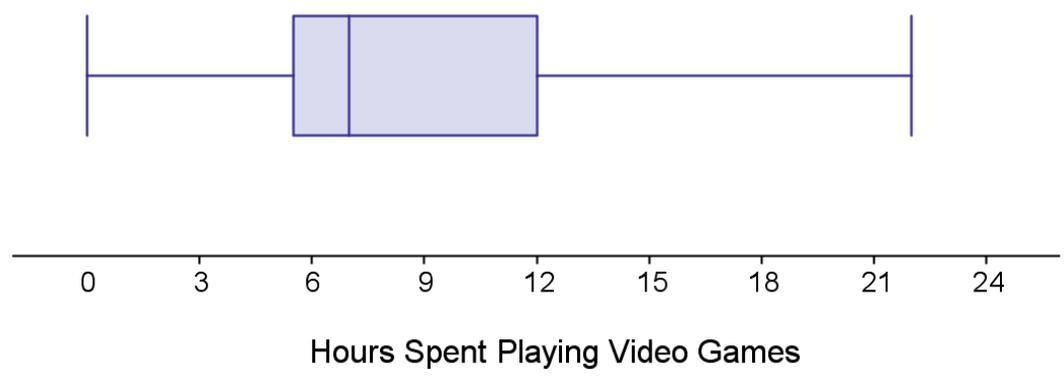
- Identify the data set shown in the dot plot as skewed to the left or skewed to the right.

The data set is skewed right because it spreads out longer on the right side.

- Construct a box plot of the data from the given dot plot.

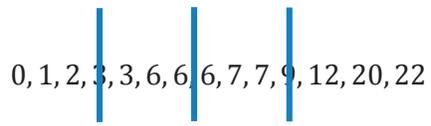


The final box plot should look like this.



Calculate the Five-Number Summary for a Data Set and the Interquartile Range

3. What is the five-number summary for the data set?



- Minimum Value: 0**
- Lower Quartile or Q1: 3**
- Median: 6**
- Upper Quartile or Q3: 9**
- Maximum Value: 22**

There are 14 data values. Once the set is in order, I need to cut the data set into quarters. The middle is between the 7th and 8th values, and the quartiles are the 4th and 11th values.

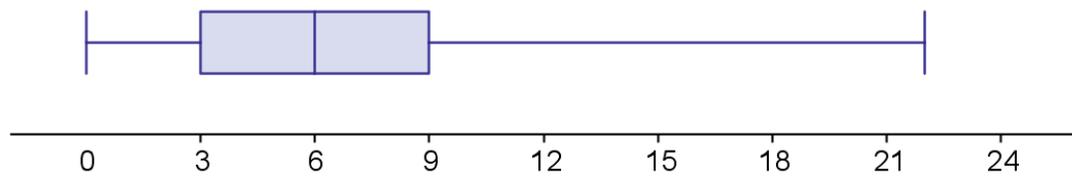
The interquartile range (IQR) is the difference between the third and first quartiles.

4. What is the interquartile range?

The lower quartile is 3. The upper quartile is 9. The difference is $9 - 3 = 6$. The interquartile range is 6.

Construct a Box Plot for a Data Set Using the Five-Number Summary

5. Construct a box plot of the data set from Problem 3.



The data ranges from 0 to 22, so I will make the number line go slightly above and slightly below these numbers.

The lower quartile is 3, and the upper quartile is 9. These values mark the sides of the box. Then, draw a line at the median, 6.

Finally, extend a horizontal line from the sides of the box to the minimum value 0 and the maximum value 22.

6. Identify any outliers in the data set from Problem 3.

The interquartile range is 6. Next, multiply this number by 1.5.

$$1.5 \cdot 6 = 9$$

Any data value that is more than 9 away from the upper or lower quartile is considered an outlier. Thus, both 20 and 22 are outliers in this data set since $9 + 9 = 18$.

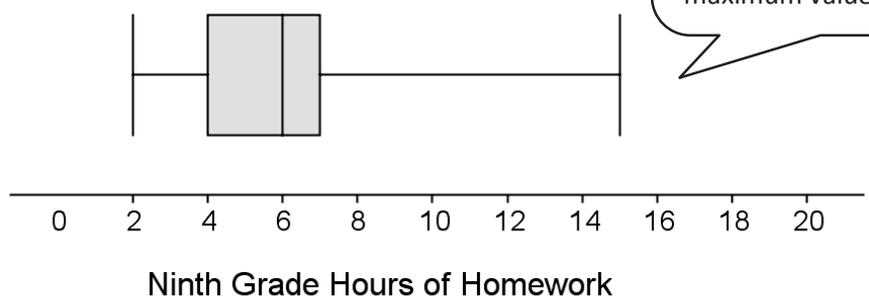
Lesson 8: Comparing Distributions

Construct a Possible Dot Plot from a Box Plot

The box plot displays the number of hours a random sample of 24 ninth graders at a high school spent doing homework during one week.

The edges of the box are the upper and lower quartiles, and the line inside the box is the median.

A box plot gives the values of the five-number summary. The ends of the lines are the minimum and maximum values.



1. Construct a possible dot plot of the sample of 24 ninth graders.

a. What is the five-number summary for this data set?

Minimum value: 2

Lower Quartile (Q1): 4

Median: 6

Upper Quartile (Q3): 7

Maximum Value: 15

b. How many data values are located between the upper and lower quartiles?

Half of the data set lies between the upper and lower quartiles. There must be 12 data values.

c. How many data values will be below the lower quartile, and how many will be above the upper quartile?

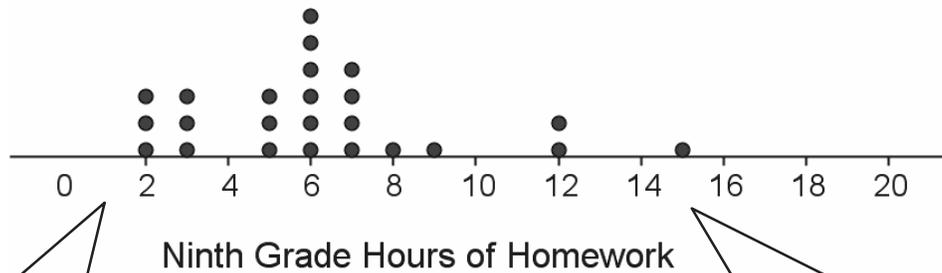
There must be 6 values below the lower quartile and 6 values above the upper quartile.

d. List your possible data set below.

2, 2, 2, 3, 3, 3, 5, 5, 5, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 8, 9, 12, 12, 15

The median must be 6. I need to have the minimum value, 2, and the maximum value, 15, in my data set.

e. Create a dot plot using the sample you created.

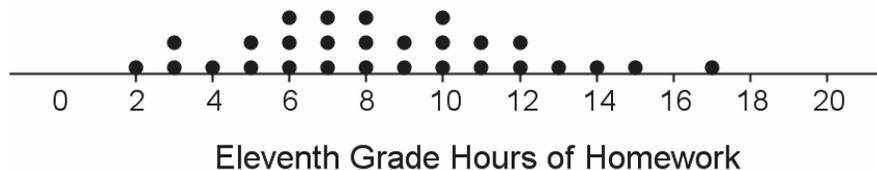


My possible data set needs to have $Q1 = 3$ and $Q3 = 9$.

This is not the only possible data set that could be represented by the given box plot.

Construct a Box Plot from a Dot Plot

The dot plot below shows the hours of homework for a random sample of 28 juniors from the same high school during one week.



2. Construct a box plot from this dot plot.

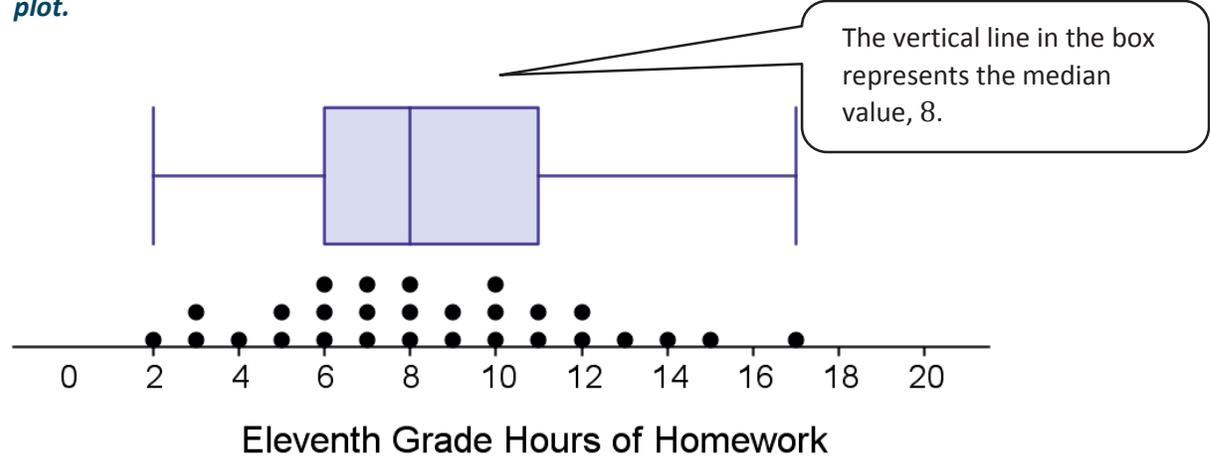
a. What is the five-number summary for this data set?

- Minimum value: 2**
- Lower Quartile: 6**
- Median: 8**
- Upper Quartile: 11**
- Maximum Value: 17**

The lower quartile is between the 7th and 8th data values. Both of these values are 6, so the lower quartile

- b. Create a box plot using the information in the five-number summary.

The box plot is superimposed on the dot plot to illustrate how to construct the box plot from a dot plot.



Compare Two Distributions

The median describes a typical value for a data set. The interquartile range (IQR) describes the variability of a data set.

3. What is a typical amount of homework for ninth graders and eleventh graders?

The median for ninth graders is 6 hours and for eleventh graders it is 8 hours.

4. Do the two data sets have a similar variability? Use the inner quartile range to support your answer.

The IQR is the difference between the upper and lower quartiles. The ninth grade IQR is 3 hours, and the eleventh grade IQR is 5 hours. The data sets have different variability. Eleventh graders have more variety in the number of hours spent doing homework than ninth graders do because they have the greater IQR.

5. Why might eleventh graders typically spend more time on homework than ninth graders?

Perhaps they are taking more challenging courses or have more academic courses that require homework.

6. Why might the hours that eleventh graders spend on homework have more variability than the hours spent by ninth graders?

Most ninth graders could be taking similar classes. By the time you are in eleventh grade, you have more options for courses or career paths such as advanced placement, vocational, or regular courses, so homework hours could vary more.

Lesson 9: Summarizing Bivariate Categorical Data

Survey Design

A random sample of 50 ninth graders were surveyed regarding their favorite type of music. Twenty-nine of the students in the sample were females.

- Eight students liked country, and 6 of those were females.
- Eight students liked rock, and 5 of those were males.
- Only 4 females and no males liked pop music.
- Ten females and 6 males liked rap/hip hop.
- Five students liked techno/electronica, and 2 of those were females
- Four females and 5 males preferred other types of music.

Some of the numbers here are represented with symbols and some with words.

1. What questions might have been asked to gather this information?

What is your gender? What is your favorite type of music?

There are two types of data in the bulleted list: gender and types of music. So my questions need to ask about that.

How could you best randomly survey ninth graders at your high school about their favorite type of music?

Answers will vary. Sample response: Randomly select five students from each ninth grade English class since all ninth graders take an English class.

I need to think of a way that doesn't bias the results but provides a fairly easy way to get answers to my questions.

2. Would the results of a random survey of ninth graders at your high school be representative of all ninth graders in the United States? Explain your reasoning.

No because this would not account for differences in geographic locations.

This answer should be no but student reasons will vary. Reasons should point out differences due to demographics and geographic location not necessarily representing the total U.S. population of ninth graders.

Summarize Bivariate Categorical Data in a Two-Way Frequency Table

3. Complete a two-way frequency table using the survey results from the 50 ninth graders.

The values of the favorite type of music categorical variable are the different types of music in the top row including the other value.

Since there were 29 females out of 50 total, there must be 21 males. The entry in the lower right cell is always the total number surveyed.

	Country	Rock	Pop	Rap/ Hip hop	Techno/ Electronica	Other	Total
Female	6	3	4	10	2	4	29
Male	2	5	0	6	3	5	21
Total	8	8	4	16	5	9	50

If 6 out of the 8 people that liked country were girls, then the other two have to be boys. The 6 and 2 are called joint frequencies.

The marginal frequencies in the bottom row and right-most column are the total number of responses for each value of the categorical variable.

The marginal frequencies should always add up to the total surveyed.

4. Do you think there is a difference in the responses of males and females? Explain your answer.

Answers will vary.

Making comparisons of the joint frequencies is tricky because the numbers of males and females are not equal.

Lesson 10: Summarizing Bivariate Categorical Data with Relative Frequencies

Construct a Relative Frequency Table and Interpret the Results

Consider the two-way frequency table for a random sample of 50 ninth graders surveyed regarding their favorite type of music.

	Country	Rock	Pop	Rap/ Hip hop	Techno/ Electronica	Other	Total
Females	6	3	4	10	2	4	29
Males	2	5	0	6	3	5	21
Total	8	8	4	16	5	9	50

Three places after the decimal point

- Calculate the relative frequencies for each of the cells to the nearest thousandth.

	Country	Rock	Pop	Rap/ Hip hop	Techno/ Electronica	Other	Total
Females	$\frac{6}{50} = 0.120$	$\frac{3}{50} = 0.060$	$\frac{4}{50} = 0.080$	$\frac{10}{50} = 0.200$	$\frac{2}{50} = 0.040$	$\frac{4}{50} = 0.080$	$\frac{29}{50} = 0.580$
Males	$\frac{2}{50} = 0.040$	$\frac{5}{50} = 0.100$	$\frac{0}{50} = 0.000$	$\frac{6}{50} = 0.120$	$\frac{3}{50} = 0.060$	$\frac{5}{50} = 0.100$	$\frac{21}{50} = 0.420$
Total	$\frac{8}{50} = 0.160$	$\frac{8}{50} = 0.160$	$\frac{4}{50} = 0.080$	$\frac{16}{50} = 0.320$	$\frac{5}{50} = 0.100$	$\frac{9}{50} = 0.180$	$\frac{50}{50} = 1.000$

I need to divide each count by the total surveyed and write my answer as a decimal.

This means that 12% of the people surveyed were boys that liked rap. It doesn't mean that 12% of boys liked rap. To convert a decimal to a percent, I need to think of it using hundredths.

$$0.120 = \frac{120}{1000} = \frac{12}{100} = 12\%$$

2. What is the relative frequency of students whose favorite music is rap/hip hop?

The relative frequency is 0.320 or 32%. This is the relative frequency for the cell that corresponds to the total number of students whose favorite music was rap/hip hop.

3. What is the relative frequency of males whose favorite music is rap/hip hop?

The relative frequency is 0.120 or 12%. This is the relative frequency for the cell that corresponds to the total number of males whose favorite music was rap/hip hop.

4. Why might someone question whether or not the students who completed the survey were selected at random? Explain your answer.

You would expect to see equal numbers of males and females. Nearly 60% of those surveyed were females.

If the survey values are very different from the population, then the survey might not be random.

5. If another student was selected at random from this school, do you think their favorite type of music would be pop? Explain your answer.

No, looking at the relative frequencies in the last row, we can see that only 8% of students reported pop as their favorite type of music.

Survey results can be used to make predictions about a population.

Lesson 11: Conditional Relative Frequencies and Association

Construct a Row Conditional Relative Frequency Table and Interpret the Results

Consider the two-way frequency table for a random sample of 50 ninth graders surveyed regarding their favorite type of music.

	Country	Rock	Pop	Rap/ Hip hop	Techno/ Electronica	Other	Total
Females	6	3	4	10	2	4	29
Males	2	5	0	6	3	5	21
Total	8	8	4	16	5	9	50

The word *row* indicates that I need to divide each frequency count in a given row by the row total.

- Construct a row conditional relative frequency table for this data. Give answers to the nearest thousandth.

The first row total is 29. The frequency count in the first cell is 6. The row relative frequency for females whose favorite music is country rounded to the nearest thousandth would be

$$\frac{6}{29} \approx 0.207.$$

The row relative frequency for males whose favorite music is country would be $\frac{2}{21} \approx 0.095$.

	Country	Rock	Pop	Rap/ Hip hop	Techno/ Electronica	Other	Total
Females	$\frac{6}{29} \approx 0.207$	$\frac{3}{29} \approx 0.103$	$\frac{4}{29} \approx 0.138$	$\frac{10}{29} \approx 0.345$	$\frac{2}{29} \approx 0.069$	$\frac{4}{29} \approx 0.138$	$\frac{29}{29} = 1.000$
Males	$\frac{2}{21} \approx 0.095$	$\frac{5}{21} \approx 0.238$	$\frac{0}{21} = 0.000$	$\frac{6}{21} \approx 0.286$	$\frac{3}{21} \approx 0.143$	$\frac{5}{21} \approx 0.238$	$\frac{21}{21} = 1.000$
Total	$\frac{8}{50} = 0.160$	$\frac{8}{50} = 0.160$	$\frac{4}{50} = 0.080$	$\frac{16}{50} = 0.320$	$\frac{5}{50} = 0.100$	$\frac{9}{50} = 0.180$	$\frac{50}{50} = 1.000$

This means that approximately 28.6% of the boys surveyed indicated that their favorite music was rap/hip hop. It does not mean that approximately 28.6% of those that liked rap/hip hop were boys.

2. For what types of music are the row conditional relative frequencies for females and males very different?

They were fairly different for all categories. The most similar was rap/hip hop, which was the most popular type of music for males and females.

3. If Pedro, a ninth grade male at this school, completed the favorite type of music survey, what would you predict was his response?

He would probably like either rap/hip hop, rock, or other.

I need to think about which entries were greatest in each row.

4. If Ali, a ninth grade female at this school, completed the favorite type of music survey, what would you predict was her response?

She would most likely indicate her favorite type of music was rap/hip hop or maybe country.

5. Is it fair to say that males and females equally prefer *other* types of music since they had nearly equal frequency counts?

No. The row conditional relative frequencies are different.

6. Do you think there is an association between gender and favorite type of music for ninth graders at this school? Explain.

While the survey revealed differences between the genders, the differences were not that large, and the survey only included 50 students. We cannot say there is strong evidence for an association.

The word *column* indicates that I need to divide each frequency count in a given column by the column total.

Construct a Column Conditional Relative Frequency Table and Interpret the Results

7. Construct a column conditional relative frequency table for this data. Give answers to the nearest thousandth.

The first column total is 8. The frequency count in the first cell is 6. The column relative frequency for females whose favorite music is country rounded to the nearest thousandth is

$$\frac{6}{8} = 0.750.$$

The column relative frequency for males whose favorite music is country would be $\frac{2}{8} = 0.250$.

	Country	Rock	Pop	Rap/ Hip hop	Techno/ Electronica	Other	Total
Females	$\frac{6}{8} = 0.750$	$\frac{3}{8} = 0.375$	$\frac{4}{4} = 1.000$	$\frac{10}{16} = 0.625$	$\frac{2}{5} = 0.400$	$\frac{4}{9} \approx 0.444$	$\frac{29}{50} = 0.580$
Males	$\frac{2}{8} = 0.250$	$\frac{5}{8} = 0.625$	$\frac{0}{4} = 0.000$	$\frac{6}{16} = 0.375$	$\frac{3}{5} = 0.600$	$\frac{5}{9} \approx 0.556$	$\frac{21}{50} = 0.420$
Total	$\frac{8}{8} = 1.000$	$\frac{8}{8} = 1.000$	$\frac{4}{4} = 1.000$	$\frac{16}{16} = 1.000$	$\frac{5}{5} = 1.000$	$\frac{9}{9} = 1.000$	$\frac{50}{50} = 1.000$

This means that 37.5% of those surveyed that liked rap/hip hop were boys. It does not mean that 37.5% of boys liked rap/hip hop.

8. If you wanted to know the relative frequency of females surveyed whose favorite music was country, would you use a row conditional relative frequency or a column conditional relative frequency?

I would use a row conditional relative frequency.

The category is females, and the condition is country music. That category is in a row.

9. If you wanted to know the relative frequency of students who liked rap/hip hop that were males, would you use a row conditional relative frequency or a column conditional relative frequency?

I would use a column conditional relative frequency.

The category is rap/hip hop and the condition is male. That category is in a column.

Lesson 12: Relationships Between Two Numerical Values

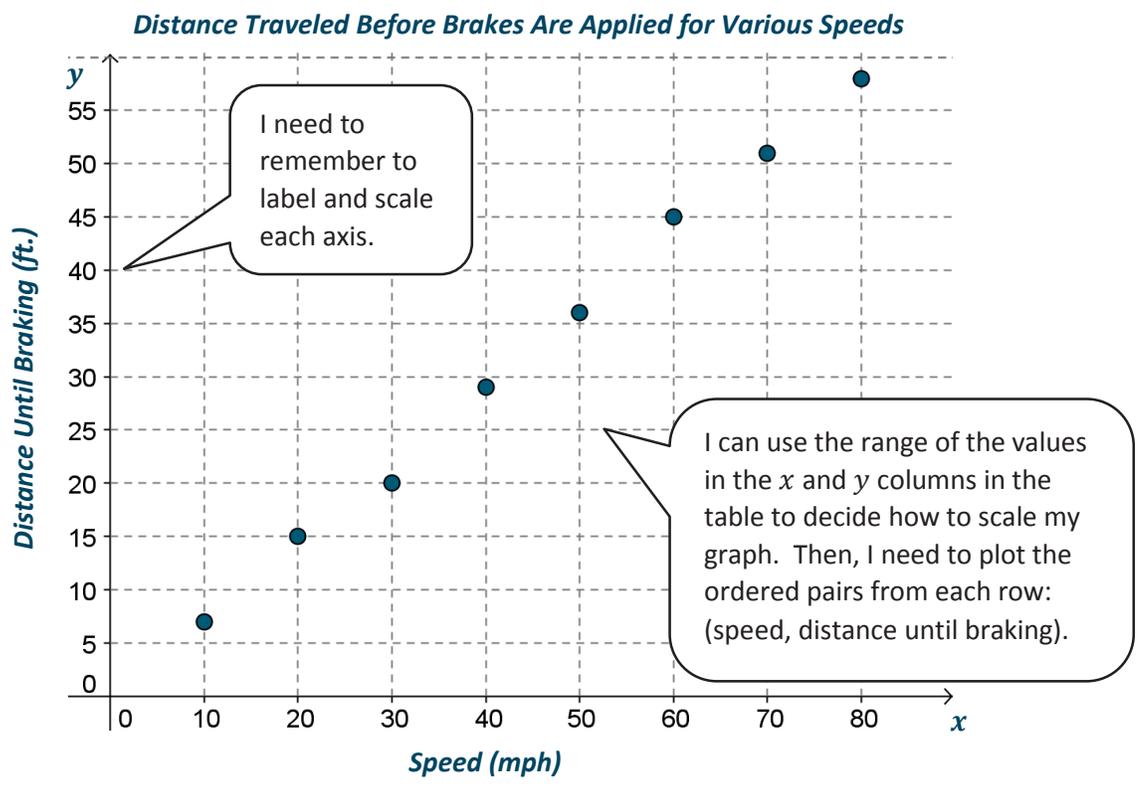
Construct a Scatter Plot and Analyze Relationships

The table below gives typical automobile braking road test results (distance traveled before brakes are applied and distance traveled until a complete stop after brakes are applied) for various speeds.

x (Speed in mph)	y (Distance Until Braking in ft.)	z (Distance Until Stopped in ft.)
10	7	5
20	15	17
30	20	37
40	29	65
50	36	105
60	45	150
70	51	205
80	58	265

(Data set from Core Math Tools, www.nctm.org)

- Construct a scatter plot that displays the data where x represents speed (in mph) and y represents distance (in feet) traveled before braking.

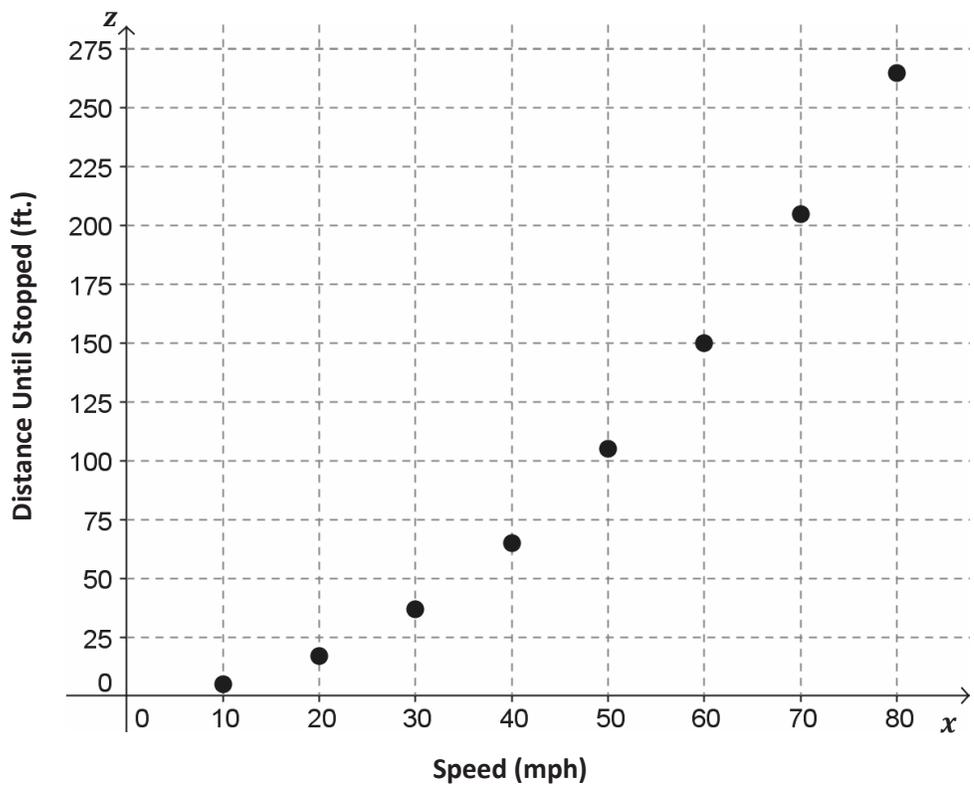


2. Based on the scatter plot, is there a relationship between the speed and the distance until braking? If so, how would you describe the relationship? Explain your reasoning.

There appears to be a relationship. As the speed increases, the distance until braking increases. The pattern suggests a linear model since the distance until braking increases at a nearly constant rate as the speed increases every 10 mph.

Consider the scatter plot where x represents the speed in mph and z represents the distance until stopped in feet shown below.

Distance Until Stopped After Brakes Are Applied for Various Speeds



3. Is there a relationship between the speed of an automobile and the distance until stopped after breaking or are the points scattered?

As the speed increases, the distance until stopped increases by larger and larger intervals. There is a pattern, so the points are not scattered.

4. Do you think there is a relationship between the speed and the distance until stopped? If so, does it look linear?

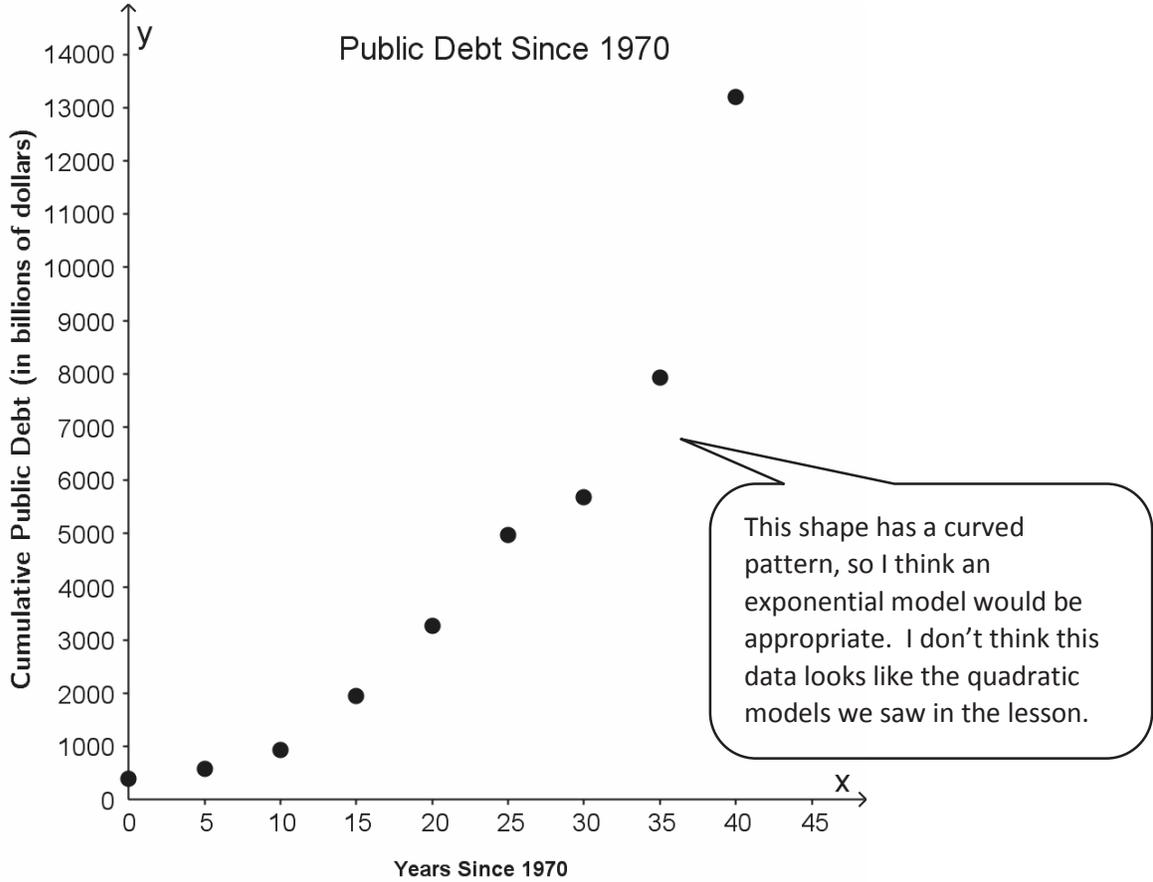
The relationship does not appear to be linear.

Lesson 13: Relationships Between Two Numerical Values

Select a Model and Make Predictions

The scatter plot below shows the cumulative public debt (in billions of dollars) of the United States government at five-year intervals since 1970. (Source: U.S. Department of the Treasury, The Public Debt Online.)

Let x represent the years since 1970, and let y represent the cumulative public debt in billions of dollars.



1. What type of model (linear, quadratic, or exponential) would you use to describe the relationship between years since 1970 and the cumulative public debt in billions of dollars?

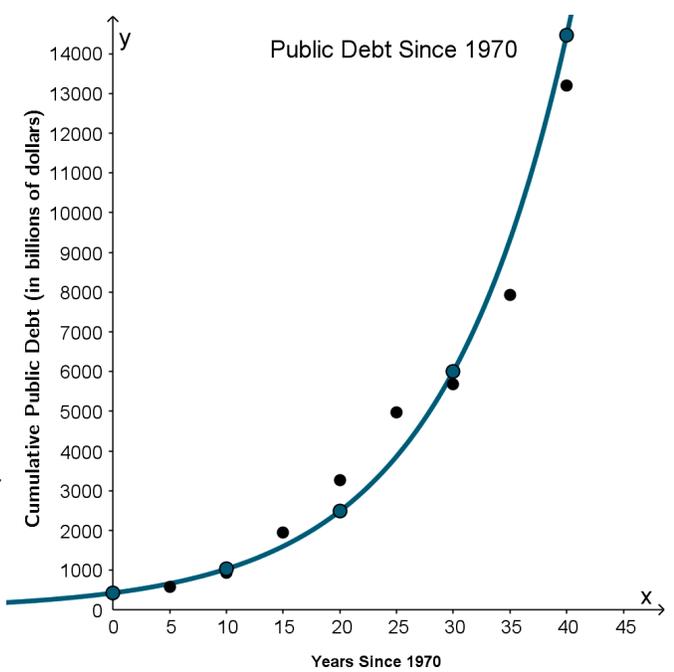
An exponential model could be used to describe this relationship.

2. One model that could describe the relationship between the years since 1970 and the cumulative public debt is $y = 428.093(1.092)^x$. Use the exponential model to complete the table. Then sketch a graph of the exponential curve on the scatter plot above.

I need to substitute the years into the equation for x and use a calculator to evaluate and find y . I can round to the nearest whole number.

Years since 1970	Public Debt in Billions of \$
0	429
10	1032
20	2489
30	6001
40	14,469

To graph the model, I can plot the ordered pairs represented in the table and connect them with a smooth curve.



3. Based on this model, how much cumulative public debt would you predict in the year 2015?

The model predicts cumulative public debt of 22,468 billion dollars in 2015.

I know that 2015 is 45 years since 1970. I will substitute 45 into the model and find y .

Lesson 14: Modeling Relationships with a Line

Lesson Notes

Finding the Regression Line (TI-84 Plus)

- From your home screen, press STAT, and then from the STAT menu, select the EDIT option. (EDIT, ENTER)
- Enter the x -values of the data set in L1, and enter the y -values of the data set in L2.

```

EDIT  CALC TESTS
1:Edit...
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor
  
```

L1	L2	L3	2
10	7		
20	15		
30	20		
40	29		
50	36		
60	45		
70	51		

L2 = {7, 15, 20, 29, ...}

- Select STAT. Move the cursor to the menu item CALC, and then move the cursor to option 4: LinReg($ax + b$) or option 8: LinReg($a + bx$). Press ENTER. (Note: Both options 4 and 8 are representations of a linear equation.)
- With option 4 or option 8 on the screen, enter L1, L2, and Y1 as described in the following notes.
LinReg($a + bx$) L1, L2, Y1 and select ENTER to see results.

```

EDIT  CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
  
```

```

LinReg(ax+b) L1,
L2, Y1
  
```

To obtain Y1, go to VARS, and then move the cursor to Y-VARS and then to Functions (ENTER). You are now at the screen highlighting the y -variables. Move the cursor to Y1, and hit ENTER. Y1 is the least squares regression line and will be stored in Y1. To see the scatter plot, move the cursor to Plot1, and press ENTER.

```

LinReg
y=ax+b
a=.7369047619
b= -.5357142857
  
```

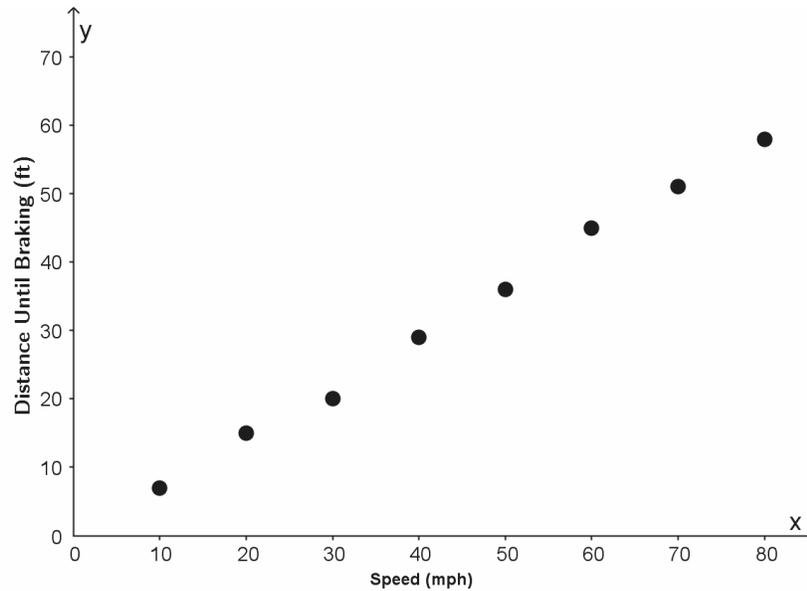
```

Plot1 Plot2 Plot3
\Y1=.73690476190
476X+ -.535714285
714
\Y2=
\Y3=
\Y4=
\Y5=
  
```

Find a Least Squares Regression Line

The table and scatter plot below give the typical distance an automobile will travel once a driver decides to hit the brakes before the driver actually engages the brakes. (Data set from Core Math Tools, www.nctm.org)

x is Speed (mph)	10	20	30	40	50	60	70	80
y is Distance Until Braking (ft.)	7	15	20	29	36	45	51	58



I can follow the steps on the previous page to create the least squares line on a graphing calculator.

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

The least squares line is $y = 0.74x - 0.54$.

Desmos.com is a free online graphing calculator that can also be used to create the least squares line. There is a tutorial on their website that I can use to guide me through the steps.

I need to substitute 55 and 100 into the least squares line.

- Predict the distance until braking 70 ft. for a car traveling 55 mph? What would you predict for a car traveling 100 mph?

When $x = 55$, $y = 0.74(55) - 0.54 = 40.16$. When $x = 100$, $y = 0.74(100) - 0.54 = 73.46$. The car travels approximately 40 feet before the brakes are applied when traveling 55 mph and approximately 73 feet before the brakes are applied when traveling 100 mph.

3. Calculate the predicted value and the residual value for a car traveling 50 mph, and add it to the table below. Then calculate the sum of the squared residuals.

$y = 0.74(50) - 0.54 = 36.46$

Residual: $36 - 36.46 = -0.46$

I need to subtract the predicted value from the actual value.

Sum of squared residuals:

$0.14^2 + 0.74^2 + (-1.66)^2 + (0.06)^2 + (-0.46)^2 + 1.14^2 + (-0.26)^2 + (-0.66)^2 = 5.3408$

Speed (mph)	Actual Distance Until Braking (ft.)	Predicted Distance Until Braking (ft.)	Residual
10	7	6.86	0.14
20	15	14.26	0.74
30	20	21.66	-1.66
40	29	29.06	-0.06
50	36	36.46	-0.46
60	45	43.86	1.14
70	51	51.26	-0.26
80	58	58.66	-0.66

I remember that slope is a rate of change. It is the coefficient of x in the least squares line.

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

We would predict an additional 0.74 feet before braking for a speed increase of 1 mph.

5. Does it make sense to interpret the y -intercept of the least squares line in this context?

No. The distance cannot be negative. The y -intercept is close to 0, which is the value that would make sense. When the speed is 0 mph, the car is not moving so the distance would also be 0 ft.

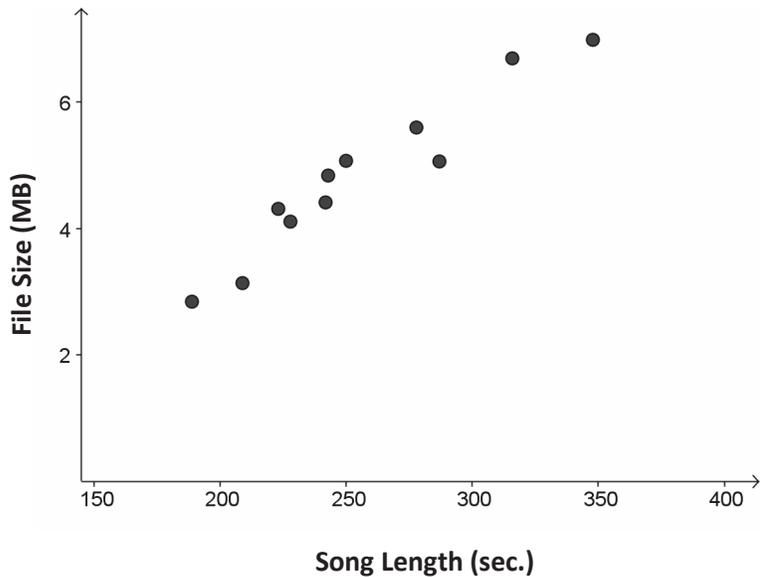
6. Would the sum of the residuals for the line $y = 0.9x - 1$ be greater than, about the same as, or less than the sum you computed in Problem 3?

It would be greater because the least squares line has the smallest sum of the residuals of any linear model.

Lesson 15: Interpreting Residuals from a Line

The song length in seconds and file size in MB for several songs is shown in the table and scatter plot.

Song Length (sec.)	File Size (MB)
228	4.103
223	4.306
250	5.071
243	4.835
278	5.595
242	4.410
348	6.975
316	6.683
287	5.062
189	2.846
209	3.134



The equation for the least squares line is $y = 0.026x - 1.859$ where x is the time in seconds and y is the file size in megabytes (MB). (Data set from Core Math Tools, www.nctm.org)

1. Draw the least squares line on the graph.

The line is drawn on the graph in the solution for Exercise 5.

Let $x = 200$; then $y = 0.026(200) - 1.859 = 3.341$.

Let $x = 300$; then $y = 0.026(300) - 1.859 = 5.941$.

I need to determine the coordinates of two points on the line and then plot the points.

2. Interpret the slope of the least squares line.

The slope is 0.026 MB per second. For each additional second, the file size increases by 0.026 MB.

The slope is the coefficient of x .

3. What does the least squares line predict for the file size of a song that is 250 seconds long?

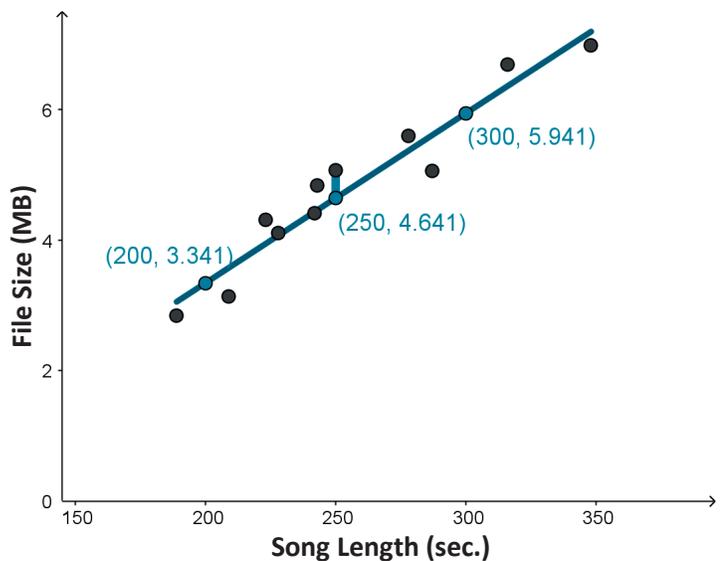
When $x = 250$, $y = 0.026(250) - 1.859 = 4.641$. The predicted size is 4.641 MB.

4. What is the difference between the actual file size of a 250 second song and the predicted file size? (This is the residual.)

$$5.071 - 4.641 = 0.43$$

5. Show your answer to Exercise 4 as a vertical line between the point on the scatter plot and the least squares line.

I need to plot the point (250, 4.641) and connect that point to the point (250, 5.071). The length of the segment is 0.43.



6. Calculate all the residuals, and write them in the table below.

Song Length (sec.)	Actual File Size (MB)	Predicted File Size (MB)	Residual
228	4.103	4.069	0.034
223	4.306	3.939	0.367
250	5.071	4.641	0.430
243	4.835	4.459	0.376
278	5.595	5.369	0.226
242	4.410	4.433	-0.023
348	6.975	7.189	-0.214
316	6.683	6.357	0.326
287	5.062	5.603	-0.541
189	2.846	3.055	-0.209
209	3.134	3.575	-0.441

Using the graphing calculator, the residuals are stored in a list named RESID. I can insert this list where I enter other lists. Or, I can just calculate the residuals one at a time like I did in Exercises 3 and 4.

7. What does the least squares line predict for the file size of a 350 second song?

When $x = 350$, $y = 0.026(350) - 1.859 = 7.241$. The predicted size is 7.241 MB.

I need to compare this residual to the others in the table. I can see it is quite a bit larger.

8. Would you be surprised if the actual file size of a 350 second song was 8 MB? Why or why not?

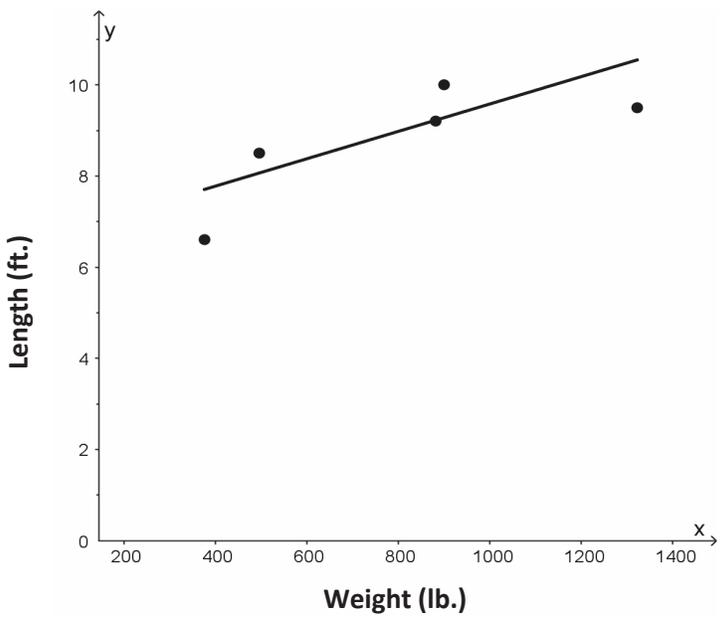
The residual would be $8 - 7.241 = 0.759$. This number is larger than the other residuals, so a file this large would be surprising for a 350 second song.

Lesson 16: More on Modeling Relationships with a Line

Create a Residual Plot

The lengths (in feet) and weights (in pounds) of five types of seals are shown in the table and scatter plot. (Source: "Grzimek's Encyclopedia, Mammals" V4. New York: McGraw-Hill, 1990, accessed via Core Math Tools, www.nctm.org)

Seal	Weight (in lb.)	Length (in ft.)
Crabeater	496	8.5
Harbor	375	6.6
Hooded	900	10.0
Monk	881	9.2
Weddell	1,323	9.5



I learned how to create the least squares line using a graphing calculator in Lesson 14.

The equation for the least squares line is $y = 0.003x + 6.58$, where x is the weight in pounds and y is the length in feet, and is included on the scatter plot.

- Use your equation to find the predicted length of the hooded seal. What is the residual?
When $x = 900$, $y = 0.003(900) + 6.58 = 9.280$. The predicted length is 9.280 ft.
Residual = Actual – Predicted = 10.0 – 9.28 = 0.82

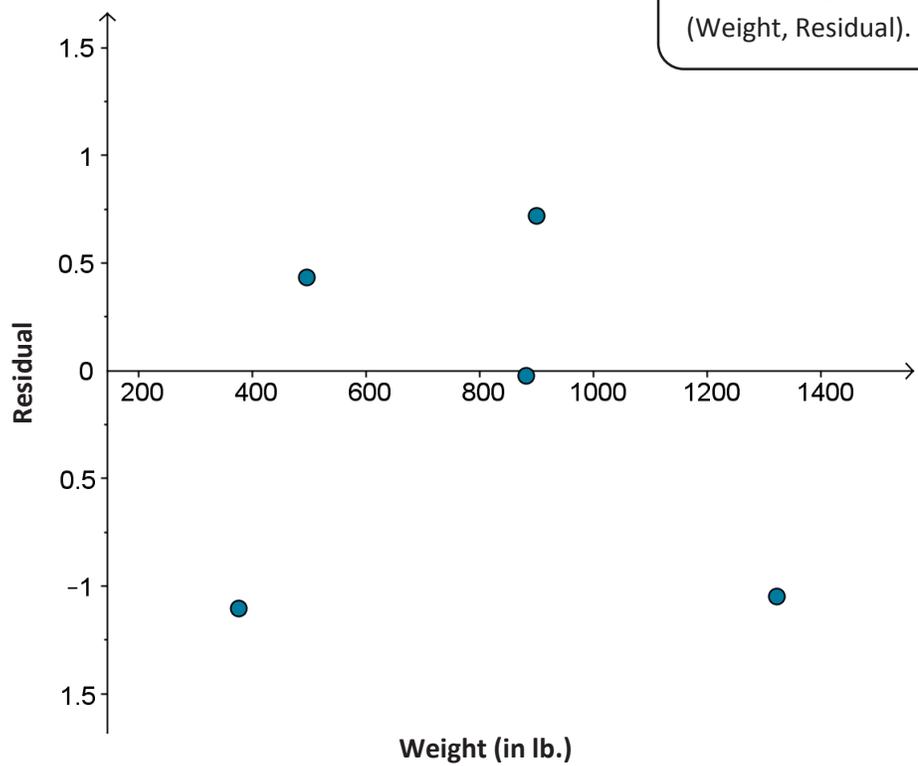
2. Calculate the residuals for the other seals. Write the residuals in the table below.

Seal	Weight (in lb.)	Actual Length (in ft.)	Predicted Length (in ft.)	Residual
Crabeater	496	8.5	8.068	0.432
Harbor	375	6.6	7.705	-1.105
Hooded	900	10.0	9.280	0.720
Monk	881	9.2	9.223	-0.023
Weddell	1323	9.5	10.549	-1.049

I can use the least squares line to get the predicted length.

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

The residuals are shown below.



I need to plot the ordered pairs from the table in Problem 2: (Weight, Residual).

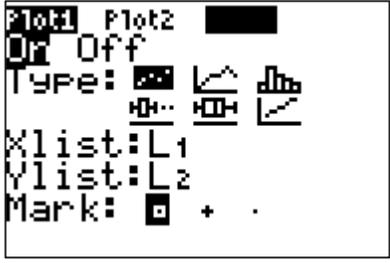
Lesson 17: Analyzing Residuals

Lesson Notes

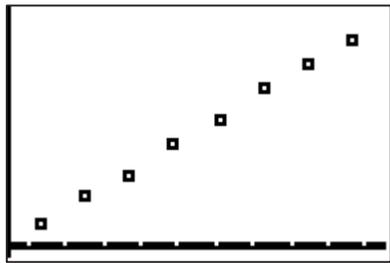
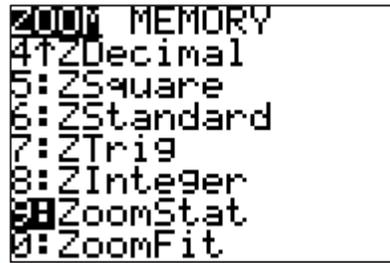
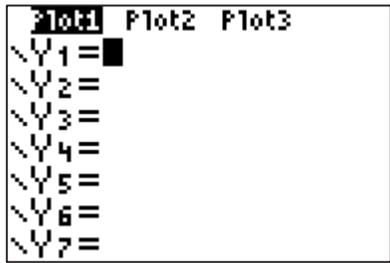
Students need access to a graphing calculator or other technology to complete this problem set. Directions for using a TI-84 Plus graphing calculator are listed below. Other types of calculators or tools may give slightly different values when calculating regression equations. Students can also use a spreadsheet program, free mathematics graphing software such as Geogebra, or a web-based application such as Desmos.com.

Construction of Scatter Plot:

- From the home screen press 2nd, STAT PLOT, and then select Plot1, and press ENTER.
- Select *On*. Under *Type*, choose the first (scatter plot) icon, for Xlist enter L₁, for Ylist enter L₂, and under *Mark*, choose the first (square) symbol.



- Press 2nd, QUIT to return to the home screen.
- Press Y = and go to any unwanted graph equations and press CLEAR. Make sure that only Plot1 is selected (not Plot2 or Plot3).
- Press Zoom, select ZoomStat (option 9), and press ENTER.



*these are general directions. The data is not the same as that given in the exercises on next pages.

Create a Scatter Plot and Least Squares Line

The table below shows the hippopotamus population sizes for various years. (Data accessed from Core Math Tools, www.nctm.org)

I will let the x represent the year and y represent the population.

To view the scatter plot on my calculator, I need to set up an appropriate viewing window by pressing **WINDOW**. The table values can be used to select appropriate **xmin**, **xmax**, **ymin**, and **ymax** values.

Year	Hippopotamus Population
1970	2,815
1972	2,919
1975	2,342
1976	4,501
1977	5,147
1978	4,765
1979	5,151
1981	4,884
1982	6,293
1983	6,544

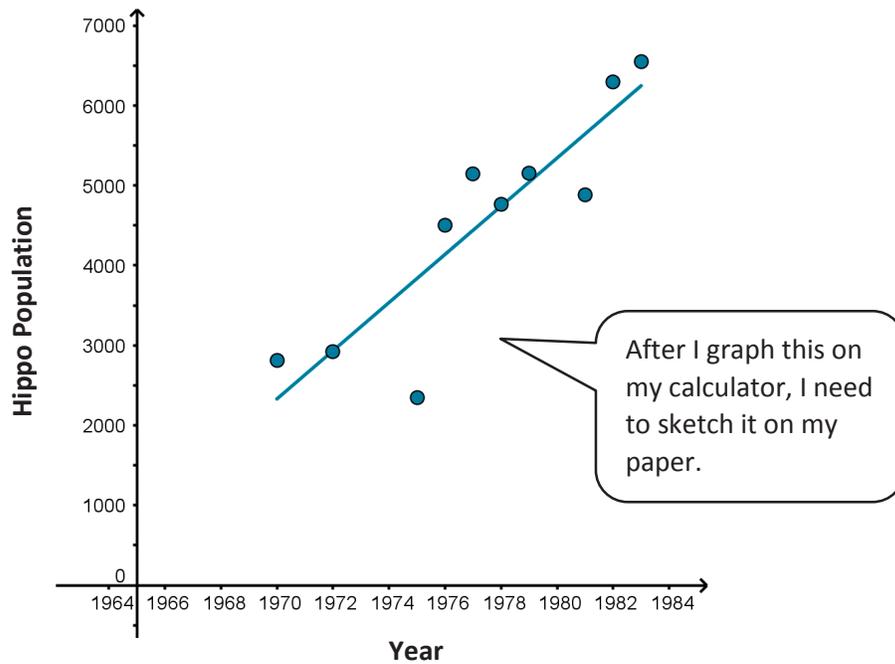
After I enter the data into two lists by pressing **STAT** and **Edit** and then typing in the data, I need to set up a scatter plot on my calculator in **STATPLOT** (**2nd** **Y=**).

I learned the steps to create the least squares line in Lesson 14.

1. Use a calculator or computer to construct the scatter plot of this data set. Include the least squares line on your graph. Explain what the slope of the least squares line indicates about the hippopotamus population.

The least squares equation is $y = 301.3x - 591231.4$, where x is the year and y is the hippopotamus population for that year.

The slope means that the population of hippos is increasing by about 301 additional hippos each year.



Construction of Residual Plot:

1. From the home screen, press 2nd, STAT PLOT, and then select Plot2, and press ENTER.
2. Select On. Under Type, choose the first (scatter plot) icon, for Xlist enter L₁, for Ylist enter RESID, and under Mark choose the first (square) symbol. (RESID is accessed by pressing 2nd, LIST, selecting NAMES, scrolling down to RESID and pressing ENTER.)

```

STAT PLOTS
1:Plot1...On
  [ ] L1 L2 [ ]
2:Plot2...Off
  [ ] L1 L2 [ ]
3:Plot3...Off
  [ ] L1 L2 [ ]
4:PlotsOff
    
```

```

NAMES OPS MATH
1:L1
2:L2
3:L3
4:L4
5:L5
6:L6
7:RESID
    
```

```

Plot1 Plot2 Plot3
[ ] Off
Type: [ ] [ ] [ ]
      [ ] [ ] [ ]
Xlist:L1
Ylist:RESID
Mark: [ ] + .
    
```

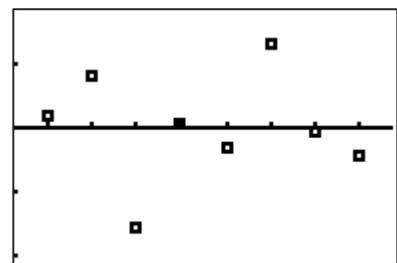
3. Press Y=. First, deselect the equation of the least squares line in Y₁ by going to the = sign for Y₁ and pressing ENTER. Then deselect Plot1, and make sure that Plot2 is selected.
4. Press Zoom, select ZoomStat (option 9), and press ENTER. The graph will be displayed.

```

Plot1 Plot2 Plot3
\Y1=.73690476190
476X+-.535714285
714
\Y2=[ ]
\Y3=[ ]
\Y4=[ ]
\Y5=[ ]
    
```

```

ZOOM MEMORY
4:ZDecimal
5:ZSquare
6:ZStandard
7:ZTrig
8:ZInteger
9:ZoomStat
0:ZoomFit
    
```



Construct and Analyze a Residual Plot

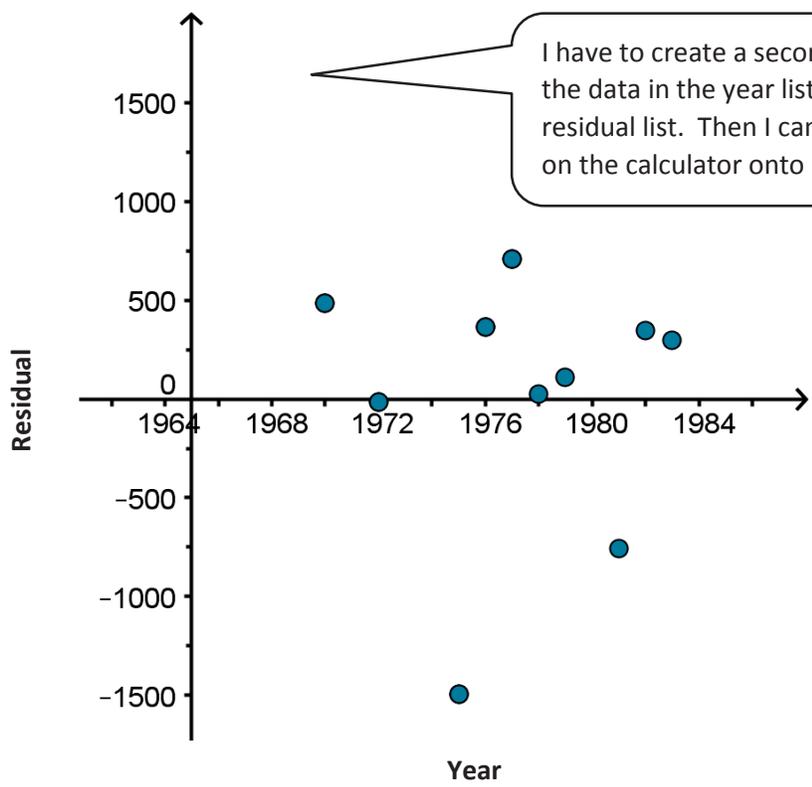
2. Use your calculator to construct a residual plot for this data set, and make a sketch on the axes given below. Does the scatter points on the residual plot indicate a linear relationship in the original data set? Explain your answer.

Year	Actual Population	Predicted Population	Residual
1970	2,815	2330	485
1972	2,919	2932	-13
1975	2,342	3836	-1494
1976	4,501	4137	364
1977	5,147	4439	708
1978	4,765	4740	25
1979	5,151	5041	110
1981	4,884	5644	-760
1982	6,293	5945	348
1983	6,544	6247	298

The graphing calculator stores residuals in a list named RESID. I will make a list for the predicted population and one for the residuals. I learned in Lessons 15 and 16 that: $\text{residual} = \text{actual} - \text{predicted}$.

I got these values from my calculator and rounded them to the nearest whole number.

There is not a clear pattern in the data, so the residual plot would indicate a linear relationship.

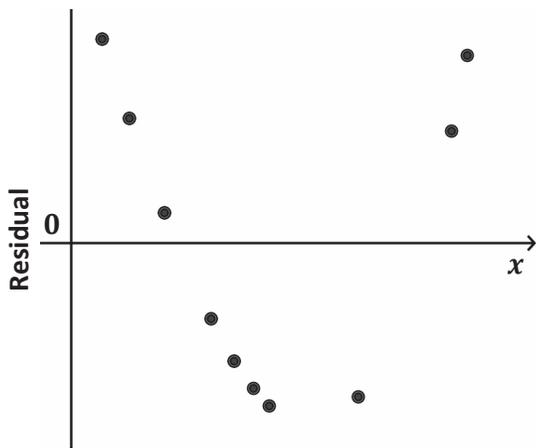


Lesson 18: Analyzing Residuals

Analyzing a Residual Plot

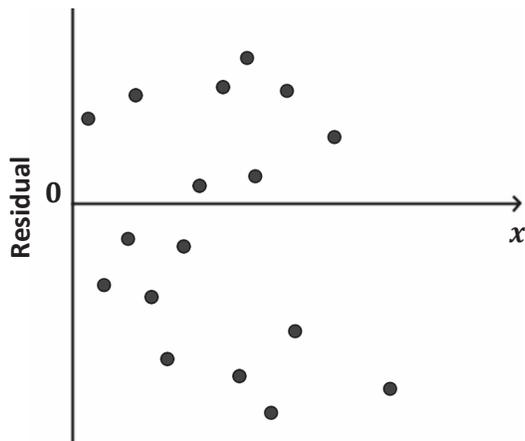
For each residual plot, what conclusion would you reach about the relationship between the variables in the original data set? Indicate whether the values would be better represented by a linear or nonlinear relationship.

1.



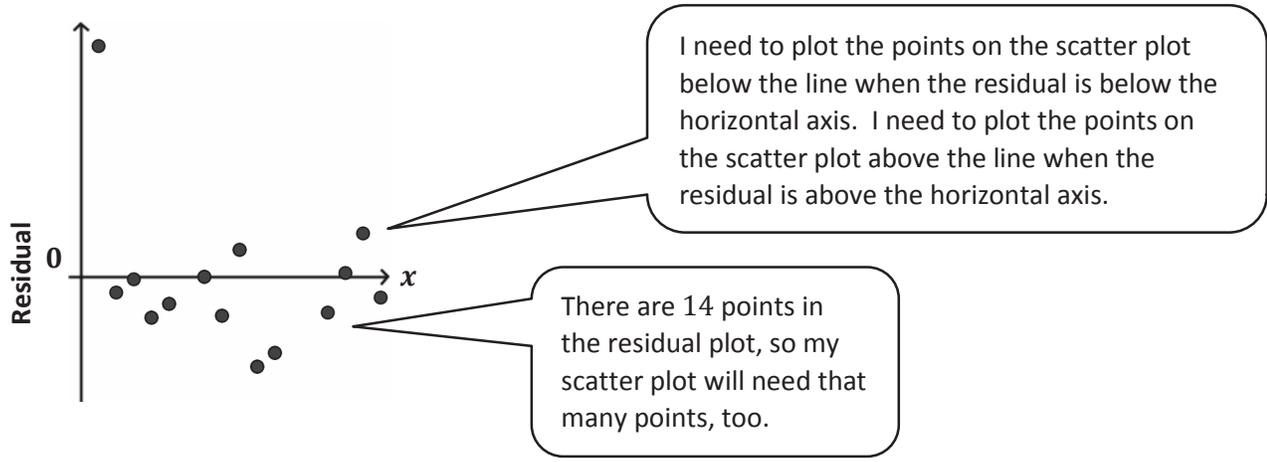
There is a pattern in the residuals for these data. The values would be better represented by a nonlinear relationship.

2.

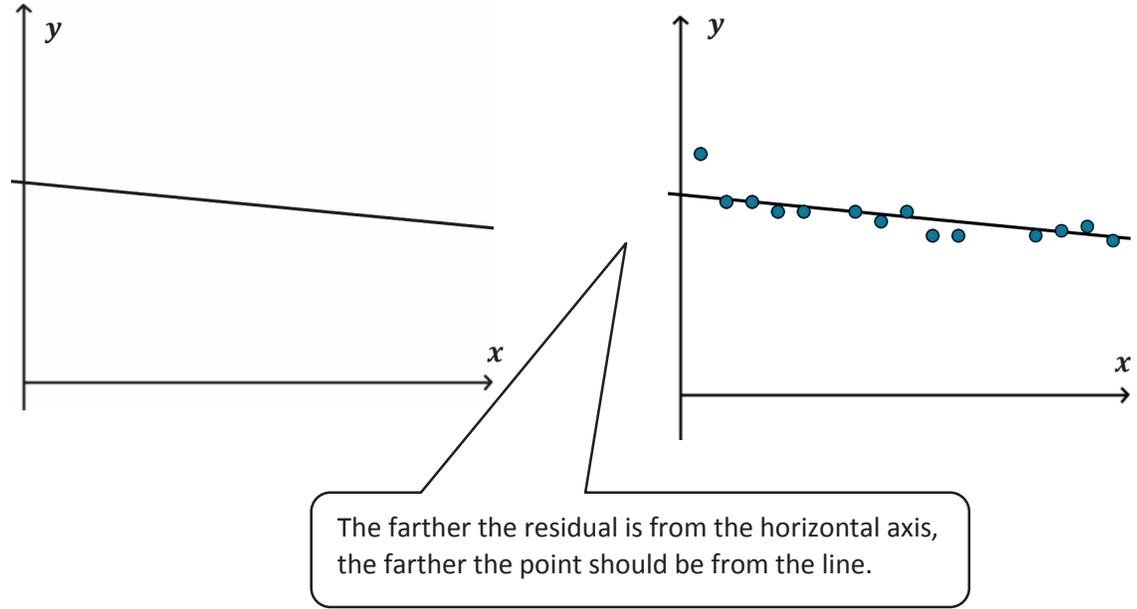


There is no pattern in the residuals. The values would be better represented by a linear relationship.

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



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



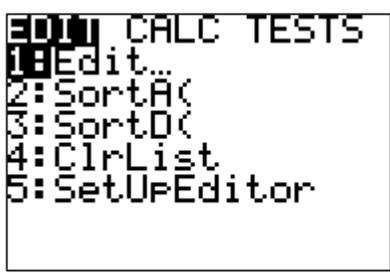
Lesson 19: Interpreting Correlation

Lesson Notes

The steps below describe how to calculate a correlation coefficient using a TI-84 Plus graphing calculator. Different types of calculators or tools may return slightly different values in regression equations. If using a different graphing calculator, graphing software, or other graphing applications, students will need to consult the user guides for their technology.

Steps for Calculating the Correlation Coefficient Using a TI-84 Plus:

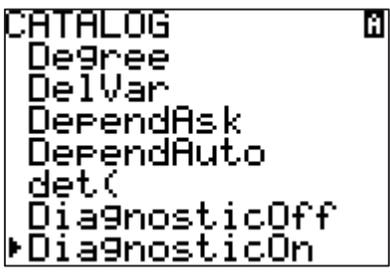
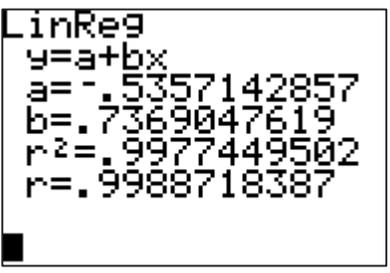
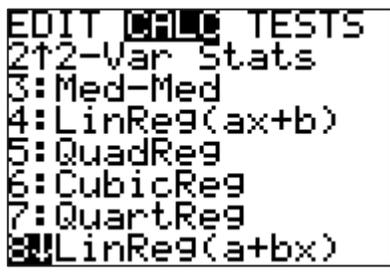
- Determine which variable represents x and which variable represents y based on x - and y -variable designations.
- From the home screen, select STAT and then Edit by pressing ENTER.
- Enter the values of x in L1 and the values of y in L2. When complete, enter 2ND QUIT.



L1	L2	L3	3
10	7		
20	15		
30	20		
40	29		
50	36		
60	45		
70	51		
L3(1)=			

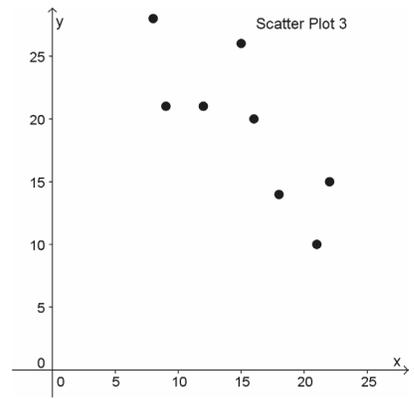
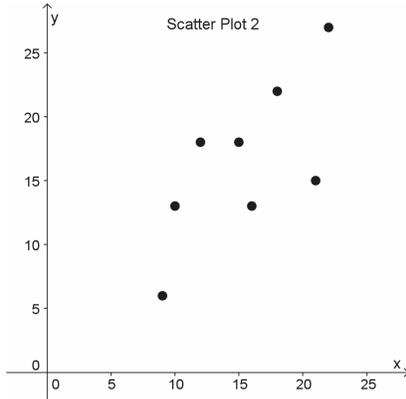
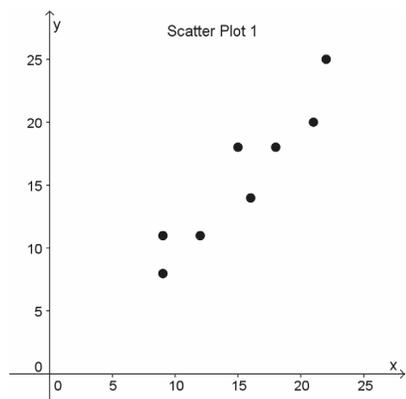
- Select STAT. With the arrows, move the top cursor over to the option CALC, and move the cursor down to 8: LinReg($a + bx$) and then ENTER.
- With LinReg($a + bx$) on the screen, enter L1, L2, Y1, and then ENTER. The value of r , the correlation coefficient, should appear on the screen. The least squares equation will be stored in Y1.

NOTE: If the r value does not appear, select 2ND CATALOG, move the cursor down to DiagnosticOn, and then ENTER. Press ENTER one more time. Repeat steps 4 and 5 above.



Identifying the Strength of a Linear Relationship

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



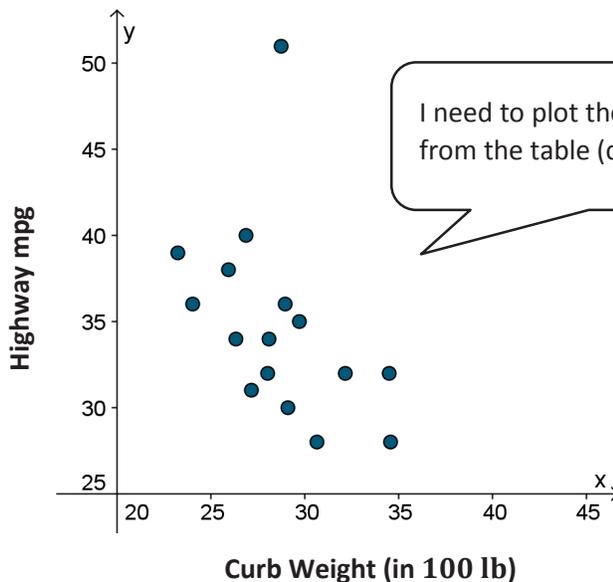
Scatter plot 1 has the strongest correlation. The weakest linear relationship is shown in scatter plot 2.

Calculating and Interpreting Correlation Coefficient

Edmunds.com published data on curb weight (in 100 lb) and highway miles per gallon for several compact car models.

Car Make and Model	Curb Weight (in 100 lb)	Highway mpg
Audi A4	34.50	32
Chevrolet Cobalt	32.16	32
Ford Focus	26.36	34
Honda Civic	26.90	40
Honda Civic Hybrid	28.75	51
Hyundai Accent	24.03	36
Kia Spectra	29.72	35
Mazda3	28.11	34
Mercedes-Benz C280	34.60	28
Nissan Sentra	28.97	36
Saturn ION	28.05	32
Subaru Impreza	30.67	28
Suzuki Aerio	27.16	31
Toyota Corolla	25.95	38
Toyota Yaris	23.26	39
VW Rabbit	29.11	30

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



After entering the data into lists, I calculate the least squares line. The correlation coefficient is given by the r value shown on my calculator screen.

3. Calculate the value of the correlation coefficient between curb weight and highway miles per gallon and interpret this value. Round to the nearest hundredth.

The correlation coefficient is $r = -0.43$. This indicates a moderately negative relationship between weight in hundreds of pounds and highway mpg.

The least squares line would have a negative slope. I know that means the correlation coefficient will be negative as well.

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

No. Heavier cars should need more gas to move them forward provided other variables are equal such as tire quality, engine efficiency and size, type of fuel, etc.

5. Is it reasonable to conclude heavier cars have a higher highway miles per gallon?

No. It appears that as the cars get heavier, the highway miles per gallon decreases.

6. Is it reasonable to conclude that increasing the weight of a car decreases the highway miles per gallon?

No. Just because there is a correlation between these two variables does not mean there is a cause-and-effect relationship between the two.

I need to remember that correlation does not imply that one variable causes the changes in the other.