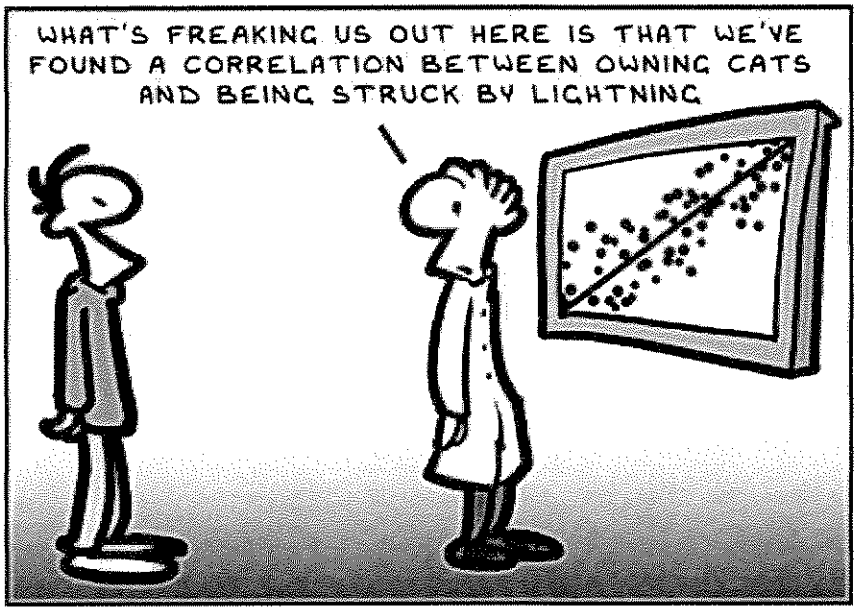
 Name: _____

Two Variable Statistics

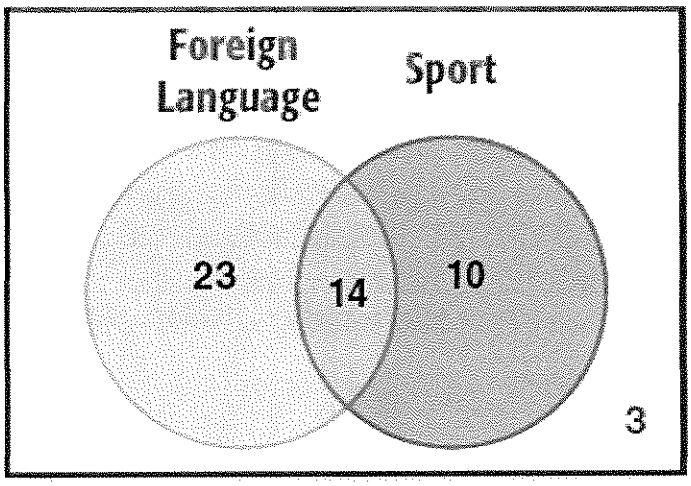
Date	Day	Lesson	Assignment
Mon. 5/6	1	Two Way Tables	
Tues. 5/7	2	Two Way Tables and Conditional Probability	
Wed. 5/8	3	**Project Day: Due @ end of class**	
Thurs 5/9	4	Linear Regression Models	
Fri. 5/10	5	Choosing an appropriate model	
Mon. 5/13	6	Residual Plots	
Tues. 5/14	7	**Review/Quiz**	
Wed. 5/15	7	Correlation & Causation	
Thur. 2/16	7	Review Day	
Fri. 2/17	7	**Unit 7 Test**	

Homework Grade:



Investigation: Two-Way Tables

1 The data from a survey of 50 students is shown in the Venn diagram below. The students were asked whether or not they were taking a foreign language and whether or not they played a sport.



1. How many students are taking a foreign language?

2. How many students play a sport?

3. How many students do both?

4. How many students do not play a sport and do not take a foreign language?

5. How many students play a sport but do not take a foreign language?

A ***two-way table*** is similar to a Venn diagram. A two-way table shows data that pertain to two different categories, which requires us to only use categorical variables. The data from one sample group is shown as it relates to two different categories. One variable is represented by rows, and the other is represented by columns.

Use the data from above to fill in the two-way table.

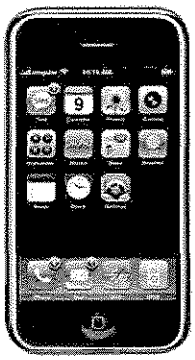
	Play a Sport	Do Not Play a Sport	Total
Take a Foreign Language			
Do Not Take a Foreign Language			
Total			

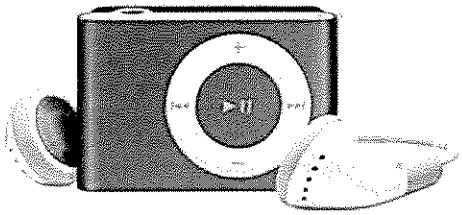
Check in with your teacher before moving on!

2 Felipe surveyed students at his school. He found that 78 students own a cell phone and 57 of those students own an MP3 player. There are 13 students that do not own a cell phone, but own an MP3 player. Nine students do not own either device.

Construct a two-way frequency table summarizing the data.

	MP3 Player	*No MP3 Player*	*Total*
Cell Phone			
No Cell Phone			
Total			





Marginal distributions are the totals of each individual category. These are located in the margins of the table. Use your table above to fill in the following:

_____ students have MP3 players. _____ students do not have MP3 players.

_____ students have cell phones. _____ students do not have cell phones.

Joint distributions are the values that "join" the two variables together. Use your table to fill in the following:

_____ students have cell phones and MP3 players

_____ students have cell phones, but not a MP3 players

_____ students have an MP3 player, but not a cell phone

_____ students have neither a cell phone nor an MP3 player

The tables you have created so far use frequencies. Some people better understand data if displayed as a percent. We can do this by created a two-way relative frequency table. Remember from Unit 1 that relative frequency is found by dividing the frequency and the overall total.

Create a two-way relative frequency table below. Round to the nearest hundredth.

	MP3 Player	No MP3 Player	Total
Cell Phone			
No Cell Phone			
Total			

Because we are working with percents, what should your overall total be? Why? Use complete sentences.

Use the table above to answer the following:

1. What percent of students have a cell phone, but not an MP3 player? _____
2. What percent of students have neither a cell phone nor an MP3 player? _____
3. What percent of students have an MP3 player, but not cell phone? _____
4. What percent of students have a cell phone and an MP3 player? _____

By converting the table to percents, we have also given ourselves probabilities!

Another way to state your answer to #1 is "The probability a person will have a cell phone and not have an MP3 player is....."

Restate questions 2—4 from above as probabilities, including the values you found.

2.

3.

4.

Two-Way Tables

A two-way table has a _____ and a _____.

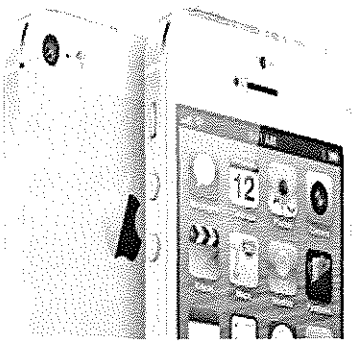
	Facebook	Twitter	YouTube	Total
Male	Number of males who prefer Facebook	Number of males who prefer Twitter	Number of males who prefer YouTube	Total number of males
Female	Number of females who prefer Facebook	Number of females who prefer Twitter	Number of females who prefer YouTube	Total number of females
Total	Total number of people who prefer Facebook	Total number of people who prefer Twitter	Total number of people who prefer YouTube	Total people surveyed

Job Opportunity

Apple needs you! They are planning to promote the iPhone and iPad in a new series of commercials. Based on our data, should the add focus on the use of Facebook, Twitter, or YouTube apps?

Why?

What other factors would impact your decision?



SURVEY SAYS....

	Facebook	Twitter	YouTube	Total
Male				
Female				
Total				

What does this all mean?

Interpreting Joint Frequencies

_____ males prefer Facebook

_____ females prefer Facebook

_____ males prefer Twitter

_____ females prefer Twitter

_____ males prefer YouTube

_____ females prefer YouTube

Marginal Frequencies

_____ people prefer Facebook.

_____ people prefer Twitter.

_____ people prefer YouTube.

Some people are picky...

In general, unless a total sample size is referenced repeatedly, people relate better to data that is given as a percent. These are called _____.

To find the joint probability for each element of the table, divide the frequency by the total.

Round to the nearest hundredth!

Original Two-Way Table

	Facebook	Twitter	YouTube	Total
Male				
Female				
Total				

Joint Frequencies

	Facebook	Twitter	YouTube	Total
Male				
Female				
Total				

In Other Words...

_____ % of males prefer Facebook.

_____ % of females prefer Facebook.

_____ % of males prefer Twitter.

_____ % of females prefer Twitter.

_____ % of males prefer YouTube.

_____ % of females prefer YouTube.

What would Steve Jobs do?

Use the information collected to find the best advertising site. Use information collected to defend your answer.

Unit 7 Day 1

Homework

Cali surveyed the students in the cafeteria about the number of times they bring their lunch to school per month. The table shows her findings.

Number of Times per Month	Males	Females
0– 5	35	25
6–10	23	16
11–15	22	13
16–20	18	8

1. Create a two-way frequency table.

2. Answer the following using your two-way table.

   a. How many students did Cali survey? _____

   b. How many males brought their lunch more than 11 times per month? _____

   c. How many females brought their lunch fewer than 6 times per month? _____

3. Create a two-way relative frequency table.

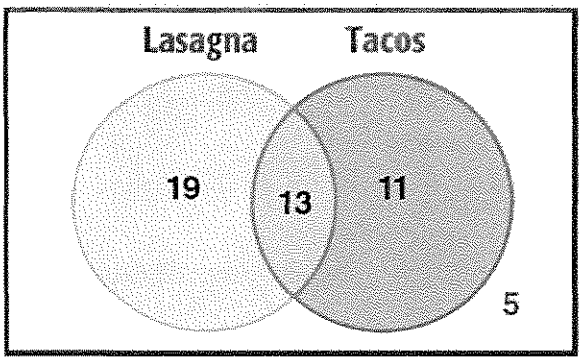
4. Answer the following using your two-way relative frequency table.

   a. What percent of females bring their lunch more than 15 times per month? _____

   b. What percent of students bring their lunch more than 15 times per month? _____

   c. Which joint frequency is highest? Use a complete sentence to describe your answer.

Students were asked to vote what they would prefer to eat at the upcoming awards banquet. The Venn Diagram displays the results.



5. Create a two-way frequency table.

6. Using complete sentences, describe all four marginal distributions. (Use a separate sheet of paper.)

7. Using complete sentences, describe all four joint distributions. (Use a separate sheet of paper.)

8. Create a two-way relative frequency table.

9. Rewrite all of your sentences from 6—7 using relative frequencies. (Use a separate sheet of paper.)

Unit 7 Day 2 — Homework

On the Wakefield Track team:

Of the people with black hair, 7 have brown eyes, 2 have blue eyes, 2 have hazel eyes, and 1 has green eyes.

Of the people with brown hair, 12 have brown eyes, 5 have hazel eyes, 3 have green eyes, and 8 have blue eyes.

Of the people with blonde hair, 9 have blue eyes, 1 has brown eyes, 1 has hazel eyes, and 2 have green eyes.

Of the people with red hair, 1 has hazel eyes, 1 has green eyes, 2 have blue eyes, and 3 have brown eyes.

Use the information from above to fill in the two-way frequency table below.

		Hair Color				
		Black	Brown	Red	Blonde	Total
Eye Color	Brown					
	Blue					
	Hazel					
	Green					
	Total					

Use the information to find the following conditional probabilities:

1. Given that a member of the team has blue eyes, what is the probability that he/she has:

   Brown hair? Blonde hair?

2. Given that a member of the team has brown hair, what is the probability that he/she has:

   Hazel eyes? Doesn't have green eyes?

3. Given that a member of the team has black hair, what is the probability that he/she has:

   Blue eyes? Brown eyes?

4. JOINT/MARGINAL: What is the probability a member of the team has brown hair? Blue eyes?

Best Fit Line

You have learned how to find and write equations for lines of fit by hand. Many calculators use complex algorithms that find a more precise line of fit called the best-fit line.

One algorithm is called linear regression. We can find the linear regression.

To enter the data: [STAT] EDIT L_1 is independent variable; L_2 is dependent variable

Calculator Steps: [STAT] CALC 4: LinReg [VARS] Y-VARS FUNCTION 1: Y_1 [Enter]

Your calculator may also compute a number called the correlation coefficient. This number will tell you if your correlation is positive or negative and how closely the equation is modeling the data. The closer the correlation coefficient is to 1 or -1, the more closely the equation models the data.

To turn the correlation coefficient on: [2nd] [0] DiagnosticOn [Enter]

- If the correlation coefficient is close to 1 or -1, the fit is _____.
- The farther away from 1 or -1, the _____ the fit.
- If the scatterplot appears random, there is _____.
- If the correlation coefficient is positive, the slope will be _____.
- If the correlation coefficient is negative, the slope will be _____.

We will often need to interpret the slope and y-intercept in the context of the problem.

Slope can often follow this pattern:

(Topic of data) (increases/decreases) (slope) (y-units) per (x-units).

The y-intercept is the starting value, or what the dependent variable is when the independent variable is 0.

EXAMPLE: The average lifespan of American women has been tracked, and the model for the data is $y = 0.2t + 73$, where $t = 0$ corresponds to 1960.

INTERPRETATION:

Real World Example 1: Box Office

The table shows the amount of money made by movies in the United States. Use a graphing calculator to write an equation for the best-fit line for that data.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Income ($ billion)	7.48	8.13	9.19	9.35	9.27	8.95	9.25	9.65	9.85	10.21

1. Enter the data into a list using the graphing calculator.
   Let x = the number of years after 2000.

2. Find the best fit line using the graphing calculator. _____

r = _____ Describe the fit. _____

Interpret the slope. _____

_____

Interpret the y-intercept. _____

_____

3. EXTRAPOLATION: Use the equation and the table in the graphing calculator to predict what the box office income will be in 2013. State your answer as a complete sentence.

_____

_____

_____

4. INTERPOLATION: Use the equation and the table in the graphing calculator to predict what the expected box office income was in 2008. How does the compare to the actual box office income given in the table? What is the difference?

_____

_____

_____

_____

REAL WORLD EXAMPLE 2: HOCKEY

The table below shows the number of goals scored by the Mustang Girls Hockey Team per year. Let x represent the number of years after 2003.

Year	2003	2004	2005	2006	2007	2008	2009	2010
Goals	63	44	55	63	81	85	93	84

1. Find the best fit line using the graphing calculator.

_____

r = _____ Describe the fit. _____

Interpret the slope. _____

_____

Interpret the y-intercept. _____

_____

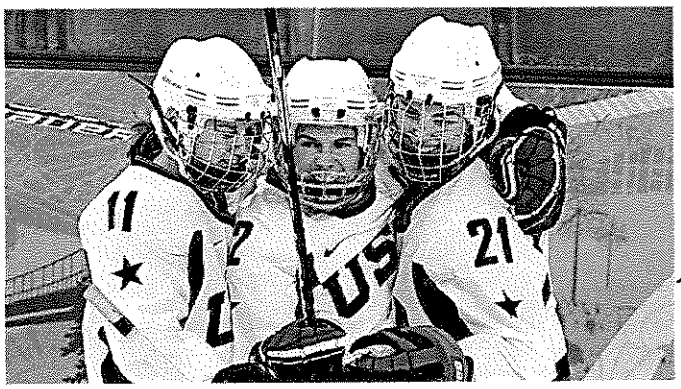
2. ANALYSIS: If this were your hockey team, would you want to use this model to predict the number of goals expected in 2013? Why or why not?

_____

_____

_____

_____



Did you know that Minnesota was the first state to establish women's hockey as a varsity high school sport in 1994?

Women's hockey first appeared in the Winter Olympics in 1998.

A local university is keeping track of the number of students who use the pottery studio each day.

Day	1	2	3	4	5	6	7
Students	10	15	18	15	13	19	20

1. Find the best-fit equation.

2. Interpret the slope and y-intercept.

3. State the correlation coefficient and describe the goodness-of-fit.

The table gives the number of young adults who auditioned for the Toledo Youth Orchestra by year.

Year	2004	2005	2006	2007	2008	2009	2010
Auditions	22	19	25	37	32	35	42

4. Find the best-fit line. Let x represent the number of years after 2004.

5. Interpret the slope and y-intercept.

6. Using your model, find the number of young adults predicted to audition in 2008. How does this compare to the actual number of auditions? Using this, is the best-fit line a good fit?

The table below shows the number of people participating in high school athletics.

Year Since 1970	1	10	20	30	35
Athletes	3,960,932	5,356,913	5,298,671	6,705,223	7,159,904

7. Find the best-fit line.

8. Interpret the slope and y-intercept.

9. Predict the number of participants in 1988. State your answer as a complete sentence.

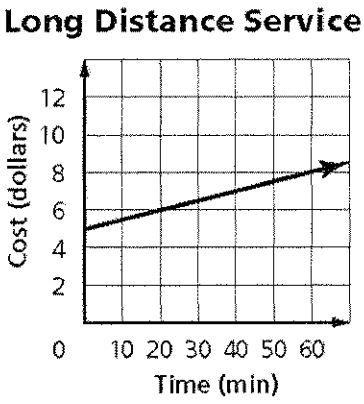
10. Predict the number of athletes in 2013. State your answer as a complete sentence.

Interpret the slope and y-intercept for each real-world situation.

11. The function, $h(x) = 2.90 + 0.79x$, models the cost of a hamburger with varying numbers of toppings.

12. The height of a candle, in inches, as a function of time, in hours, when burning is modeled by $h(t) = -0.2t + 12$.

13.



14.

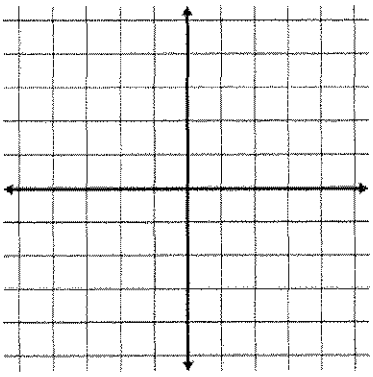
Nan's Weekly Salary					
Sales Made	1	2	3	4	5
Salary	$250	$300	$350	$400	$450

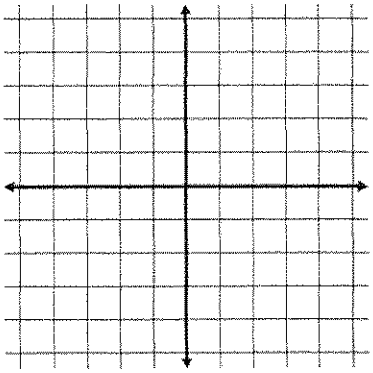
Enduring Question: What kind of equation best models a set of data?

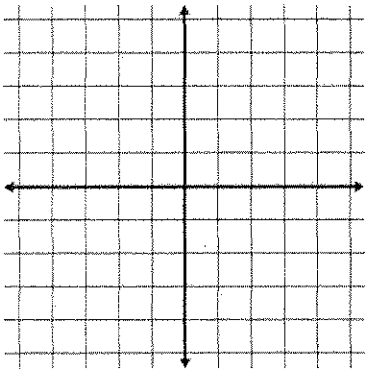
The three most common types of models are:

_____ _____ _____

General Form _____ General Form _____ General Form _____







x	*y*
-3	7
-1	13
1	19
3	25
5	31

x	*y*
-2	-2
-1	-8
0	-32
1	-128
2	-512

x	*y*
-1	20
0	8
1	3.2
2	1.28
3	0.512

CALCULATOR TIPS

Let's Try Some!

1.

X	Y
-1	20
0	8
1	3.2
2	1.28
3	0.512

2.

X	Y
0	0
1	1.5
2	6
3	13.5
4	24

3. **The table below shows the population of a small town. Let $t = 0$ correspond to the year 1990.**

Years	Pop
0	5100
5	5700
10	6300
15	6900
20	7500

a. Graph the data. Does the graph suggest a linear, exponential, or quadratic model? _____

b. What is the difference in years? _____

c. Find the differences of consecutive terms. Divide by the difference in years to find possible common differences. _____

d. Write a linear equation to model the data based on your answer to part (c). _____

e. Predict the population in 2020. _____

Practice 10-9

Choosing a Linear, Quadratic, or Exponential Model

Which kind of function best models the data? Write an equation to model the data.

1. $(-1, 3), (1, 3), (3, 27), (5, 75), (7, 147)$
2. $(-2, 4), (-1, 2), (0, 0), (1, -2), (2, -4)$
3. $\left(-2, \frac{1}{16}\right), \left(-1, \frac{1}{4}\right), (0, 1), (1, 4), (2, 16)$
4. $(-6, -1), (-3, 0), (0, 1), (3, 2), (6, 3)$
5. $\left(-2, \frac{1}{3}\right), (-1, 1), (0, 3), (1, 9), (2, 27)$
6. $(-4, -32), (-2, -8), (0, 0), (2, -8), (4, -32)$
7.

x	y
−3	$\frac{9}{2}$
−2	2
−1	$\frac{1}{2}$
0	0

8.

x	y
−1	−2
0	−4
1	−6
2	−8

9.

x	y
−4	−4
−2	−1
0	0
2	−1

10.

x	y
0	−2
1	−8
2	−32
3	−128

11.

x	y
−7	−245
−5	−125
−3	−45
−1	−5

12.

x	y
−2	$\frac{3}{2}$
0	$\frac{1}{2}$
2	$-\frac{1}{2}$
4	$-\frac{3}{2}$

13. $\left(-2, \frac{1}{3}\right), \left(-1, \frac{1}{3}\right), \left(0, \frac{1}{3}\right), \left(1, \frac{1}{3}\right), \left(2, \frac{1}{3}\right)$
14. $\left(-1, -\frac{1}{4}\right), \left(0, -\frac{1}{2}\right), (1, -1), (2, -2), (3, -4)$
15. The cost of shipping computers from a warehouse is given in the table below.

Number of Computers	50	75	100	125
Cost (dollars)	1700	2500	3300	4100

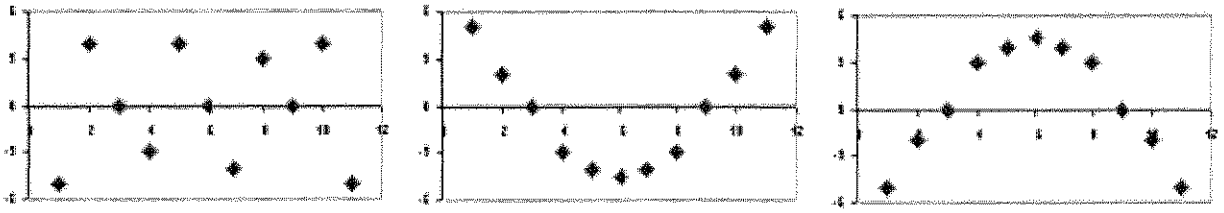
   a. Determine which kind of function best models the data.
   b. Write an equation to model the data.
   c. On the basis of your equation, what is the cost of shipping 27 computers?
   d. On the basis of your equation, how many computers could be shipped for $5500?

16. During a scientific experiment, the bacteria count was taken at 5-min intervals. The data shows the count at several time periods during the experiment.

Time Interval	0	1	2	3
Count	110	132	159	190

   a. Determine which kind of function best models the data.
   b. Write an equation to model the data.
   c. On the basis of your equation, what is the count 1 hr, 45 min after the start of the experiment?

- A residual plot is a graph that shows the difference between the actual data (what is provided through a table or graph) and the predicted data (what the model says should happen).
- The independent variable is graphed on the horizontal axis and the residual value (actual – predicted) is graphed on the vertical axis.
- If the residual plots are randomly scattered around the horizontal axis, a linear model is the best choice to model the data.
- If the residual plot shows a pattern, and does not appear to random or scattered, a non-linear model would most likely be a better fit.



Random pattern

Non-random: U-shaped curve

Non-random: Inverted U

Best fit is a linear model.

Best fit is a non-linear model.

Best fit is a non-linear model.

Let's start with some data!

The data below shows the number of active woodpecker clusters in the DeSoto National Forest.

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000
Active Clusters	22	24	27	27	34	40	42	45	51

1. Enter the data into your calculator and find the line of best fit. Let x represent the number of years after 1992.

2. What is the correlation coefficient? Describe the goodness-of-fit.

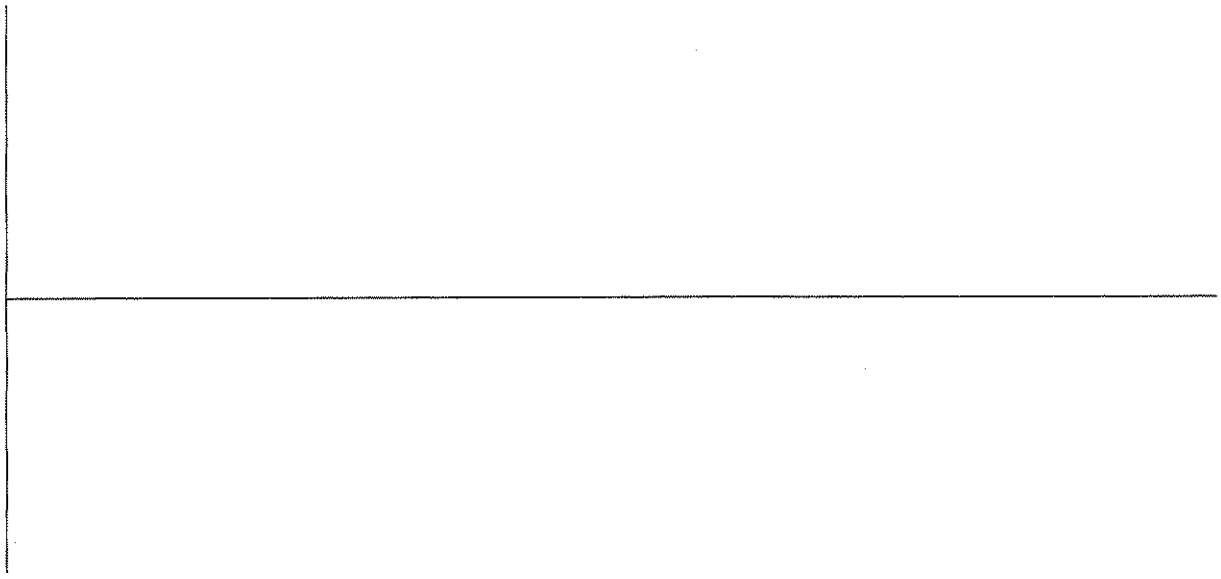
3. Using your equation from question 1, find the predicted number of active woodpecker clusters. You can do this using the TABLE feature in your graphing calculator

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000
Predicted Active Clusters									

4. To find the residual plots, we need to find the difference between what actually happened (original table) and what is predicted to happen (table from #3).

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000
Residual Value									

5. Now let's construct a residual plot. On the horizontal axis will be our independent variable. On the vertical axis will the residual value.



6. Would you describe the residual plot as scattered and random or do you see a pattern? Do you think a linear model is best?

More data!

Below is population data for Jamestown, Virginia.

Year	2002	2004	2005	2007	2009
Population	5564	6121	6300	6812	7422

1. Find the line of best fit. Let x represent the number of years after 2000.

2. Recreate the table using the model from question 1 to find the predicted population.

3. Find the difference between actual and predicted and make a residual plot.

4. Do you think a linear model is best? Why or why not?

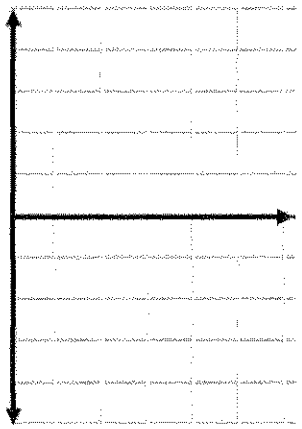
5. If you think the data would best be modeled by a non-linear model, find this model. To help you do this, look at the scatterplot in your calculator!

Unit 7 Day 6 — Homework

Complete each table using the given linear regression (Round answers to one decimal place). Construct a residual plot.

1. Linear regression equation: $y = 0.5x$

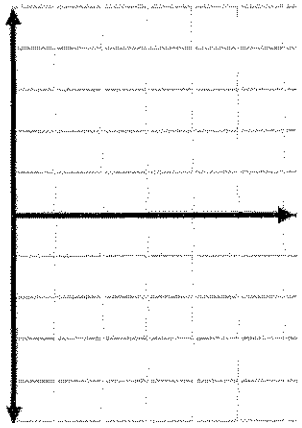
x	y	Predicted Value	Residual Value
5	3		
10	4		
15	9		
20	7		
25	13		
20	15		



Does the residual plot suggest a linear relationship? Explain.

2. Linear regression equation: $y = -0.4x + 16.3$

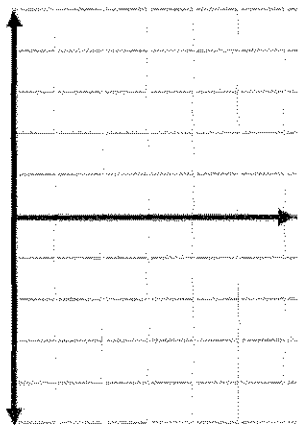
x	y	Predicted Value	Residual Value
2	5		
4	15		
6	26		
8	23		
10	11		
12	3		



Does the residual plot suggest a linear relationship? Explain.

3. Linear regression equation: $y = 4.9x + 16.4$

x	y	Predicted Value	Residual Value
100	505		
90	406		
80	415		
70	360		
60	305		
50	265		



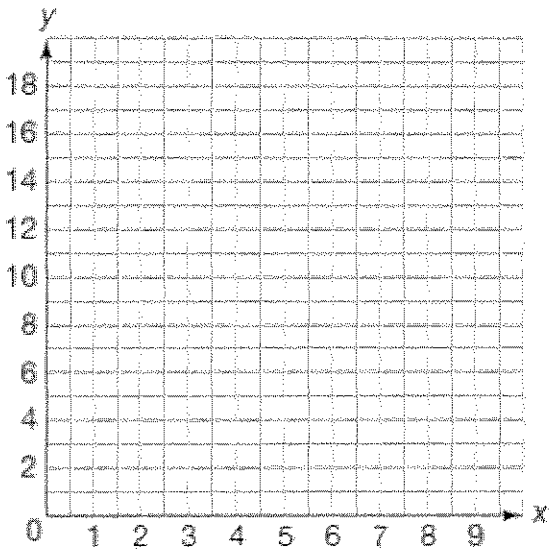
Does the residual plot suggest a linear relationship? Explain.

4. The table below shows the percent of the United States population who did not receive needed dental care services due to cost. Let $x = 0$ represent the number of years after 1999.

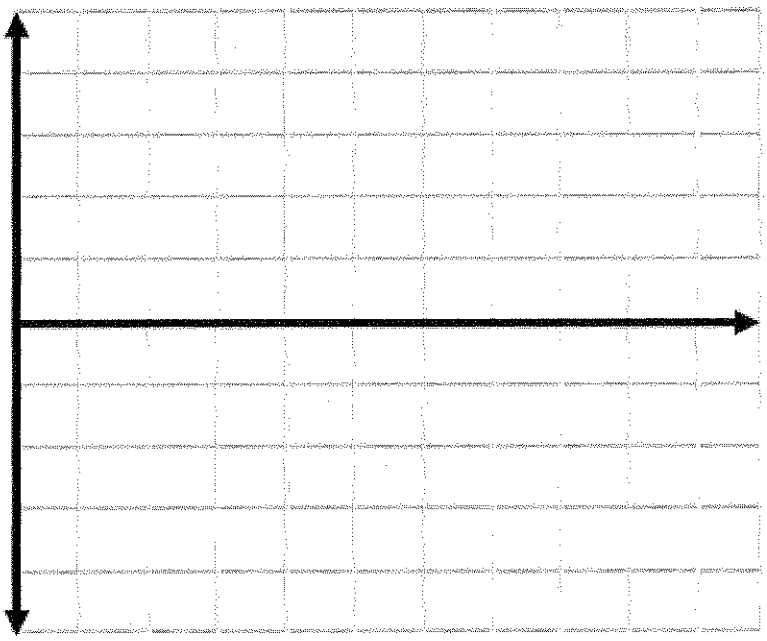
Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Percent	7.9	8.1	8.7	8.6	9.2	10.7	10.7	10.8	10.5	12.6	13.3

a. Sketch the scatter plot.

b. Choosing two points, estimate the line of best fit.



c. Using the estimated line of best fit equation, calculate the residuals for the set of data (rounding to the nearest tenth. Construct a residual plot for the data.



Unit 7 Day 8 — Notes: Correlation/Causation

A _____ is a measure or degree of relationship between two variables. A set of data can be positively correlated, negatively correlated or not correlated at all. As one set of values increases the other set tends to increase then it is called a positive correlation. As one set of values increases the other set tends to decrease then it is called a negative correlation. If the change in values of one set does not effect the values of the other, then the variables are said to have "no correlation" or zero correlation".

A _____ _____ between two events exists if the occurrence of the first causes the other. The first event is called the cause and the second event is called the effect. A correlation between two variables does not imply causation. On the other hand, if there is a causal relationship between two variables, they must be correlated.

Example 1: A study shows that there is a negative correlation between a student's anxiety before a test and the student's score on the test.

In other words, the higher the student's anxiety, the _____ the test score.

Causation: Does anxiety cause a student to earn a low test score?

Example 2: A study shows that there is a positive correlation between the number of hours a student spends studying and the student's score on a test.

In other words, the more hours spent studying, the _____ the test score.

Causation: Does more studying result in a higher grade?

Example 3: Using the graphing calculator, enter the data below.

Weekly Data Collection	
The weight of a growing puppy in New York	**The retail price of snowshoes in Alaska**
8 pounds	$32.45
8.5	$32.95
9	$33.45
9.6	$34.00
10.1	$34.50
10.7	$35.10
11.5	$35.63

a. Find the equation of the best-fit line.

b. What is the correlation coefficient?

c. Looking at the graph of the scatter plot and the line, describe the correlation. Does this correspond with the correlation coefficient?

d. Correlation vs. Causation: Is this an example of correlation, causation, or both? Justify your answer.

Unit 7 Day 8 — Homework

Identify the relationship between the two quantities in the given question as *causation* or *correlation*.

1. The number of cold, snowy days and the amount of hot chocolate sold at a ski resort.

2. The number of miles driven and the amount of gas used.

3. The number of additional calories consumed and the amount of weight gained.

4. The age of a child and his/her shoe size.

5. The amount of cars a salesperson sells and how much commission he makes.

6. The number of cars traveling over a busy holiday weekend and the number of accidents reported.

7. The number of homework assignments turned in and how well an individual does in class.

8. The annual salary and blood pressure for men ages 20-60

9. Which of the following statements shows a relationship that is correlated but *not* causal?

A) The amount of rainfall received and level of water in the lake.
B) The number of lights left on each day and the amount of the electric bill.
C) The increase of warm, sunny days and the number of ice cream vendors visible.
D) The number of hours worked and how much money is made.

10. Which of the following statements shows a relationship that is correlated but *not* causal?

A) The number of tardies to class and the number of detentions received.
B) The season of the year and the number of water related injuries/deaths.
C) As the temperature rises, more the mercury in a thermometer will expand and rise.
D) The larger the dimensions of a rectangular patio, the more square footage there will be.

11. Which of the following statements shows a causal relationship and *not* just a correlated one?

A) An individual's decision to work in construction and his diagnosis of skin cancer.
B) A decrease in temperature and the increase in attendance at an ice skating rink.
C) As a child's weight increases so does her vocabulary.
D) The number of minutes spent exercising and the amount of calories burned.

12. Which statement below might be caused by the statement, "The more the furnace runs...."?

A) the less time individuals will spend outside
B) the longer you will have to let your car warm up
C) the colder it is outside
D) the warmer the house becomes

13. Consider a large number of countries around the world. There is a positive correlation between the number of Nintendo games per person x and the average life expectancy y. Does this mean that we could increase the life expectancy in Rwanda by shipping Nintendo games to that country?

A) Yes: the correlation says that as Nintendos go up, so does life expectancy.
B) No: if the correlation were negative we could accept that conclusion, but this correlation is positive.
C) Yes: positive correlation means that if we increase x, then y will also increase.
D) No: the positive correlation just shows that richer countries have both more Nintendos and higher life expectancies.
E) It makes no sense to calculate correlation between these variables.