NAME:_____ Math 7.2, Period _____

Mr. Rogove Date:_____

Statistics Study Guide

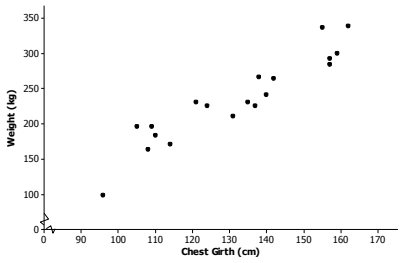
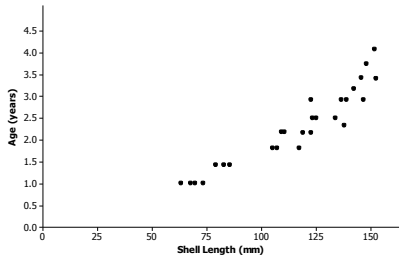
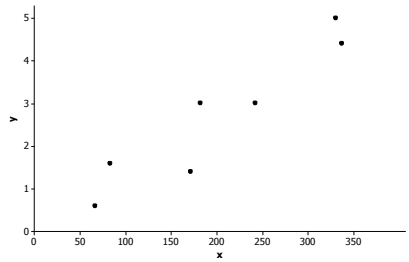
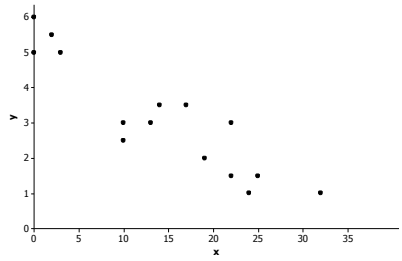
Instructions: *This study guide covers the material that will be on our next test covering mostly statistics. Do your best and turn in the completed study guide when you take your test. Thanks!*

SCATTER PLOT

A **scatter plot** is a graph of bivariate numerical data.

Patterns in Scatter Plots:
If you can see the value of one variable tend to vary in a predictable way as the values of the other variable changes, there is a statistical relationship.

Linear Relationship: If the data looks like it is varying along a straight line, we can say there is a linear relationship.	**Non-Linear Relationship**: If the data is varying along a curve, or a pattern other than a straight line, it is said to be non-linear.
Positive Trend: if the data move in a pattern up and to the right, there is a positive trend.	**Negative Trend**: If the data move in a pattern down and to the left, there is a negative trend.

Linear Relationship	Non Linear Relationship
	
Positive trend in the data	Negative trend in the data
	

NAME:_____ Math 7.2, Period _____

Mr. Rogove Date:_____

Independent Variable: this is the **explanatory variable** or the **predictor variable**. This is the variable that is not changed by the action of the other variables. This is the *x*-value, on the horizontal axis.

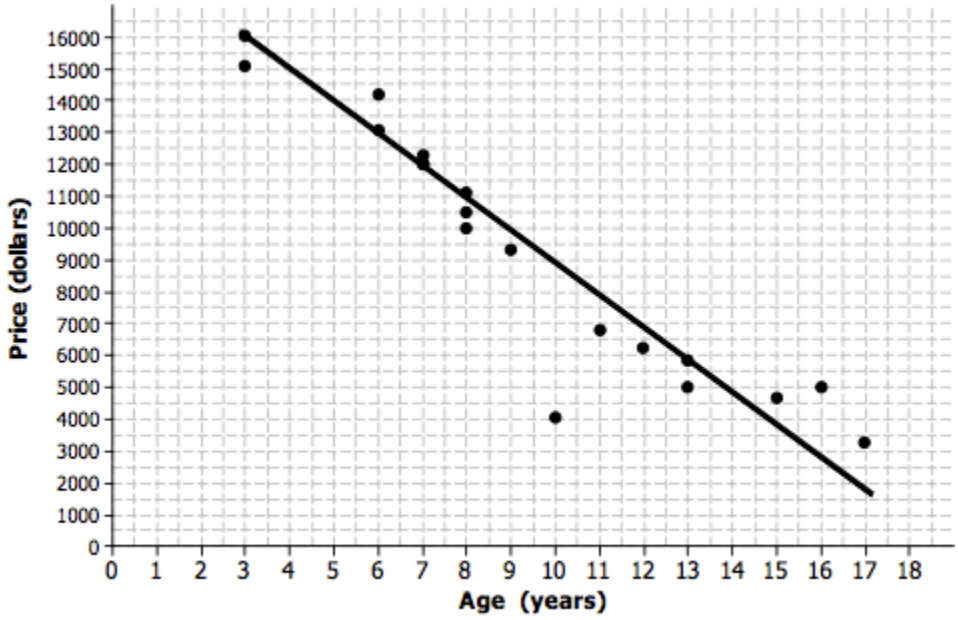
Dependent Variable: This is **response** variable or the **predicted** variable. This is the variable that you are trying to make predictions about. This is the *y*-value on the vertical axis.

Independent v. Dependent Variable: We can use the information about the independent variable to make predictions about the values of the dependent variable (*y*- axis).

Line of Best Fit: This is a straight line that represents the trend in the data. The line of best fit should be drawn as close to as many points on the graph as possible. We can write an equation for this line by identifying two points on the line, finding a slope and a *y*-intercept.

The **slope** of the line of best fit measures the impact that the explanatory variable has on the response variable.

The **y-intercept** is the value of the response variable when the explanatory variable has no effect. In linear models, the y-intercept might not make sense in the context of the real world situation.



NAME:_____ Math 7.2, Period _____

Mr. Rogove Date:_____

BIVARIATE CATEGORICAL DATA AND TWO WAY TABLES

Categorical Variables: Variables that represent data evaluated using specific categories or descriptions.

Bivariate Categorical Data is organized and summarized in a **two-way frequency table**.

		Favorite Snack					
		Candy Bar	**Baked Goods**	**Salty**	**Spicy**	**Healthy**	**Total**
Gender	**Male**	9	10	15	5	8	**47**
	Female	2	13	14	1	10	**40**
	Total	**11**	**23**	**29**	**6**	**18**	**87**

Relative Frequency: A description of the frequency of the occurrences of each of the pieces of categorical data in relation to the whole. This is a **proportion** measured by the following fraction: $\frac{frequency}{total}$.

Example: The proportion of all students who are male AND preferred salty snacks is $\frac{15}{87}$ or 0.17

Row Relative Frequency: A description of the frequency of the occurrences of pieces of categorical data in relation to the total of a row. This is a proportion measured by the following fraction: $\frac{frequency}{row\ total}$.

Example: The proportion of female students who like healthy food is $\frac{10}{40}$ or 0.25.

Column Relative Frequency: A description of the frequency of the occurrences of pieces of categorical data in relation to the total of a column. This is a proportion measured by the following fraction: $\frac{frequency}{column\ total}$.

Example: Of the students who like candy bars the proportion of them who are boys is $\frac{9}{11}$ or 0.82.

NAME:_____ Math 7.2, Period _____

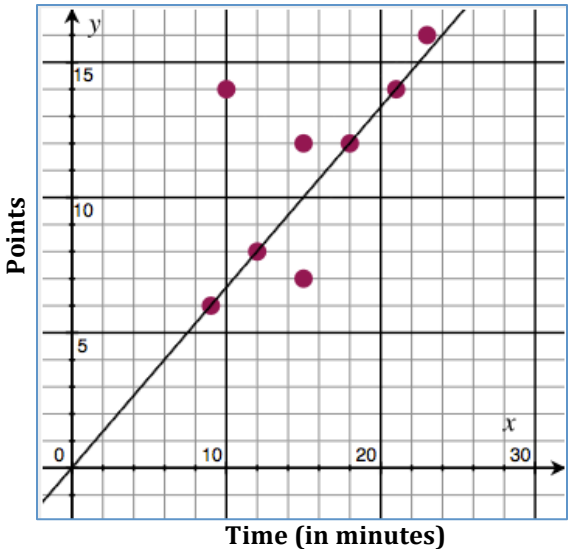
Mr. Rogove Date:_____

Problem Set.

2. Below is data that measures minutes Nicole played in basketball games and the number of points she scored.

Minutes Played	Points Scored		Minutes Played	Points Scored
15	7		12	8
18	12		23	16
9	6		15	12
21	14		10	14

a. Draw a scatterplot of the data above in the space provided below. Clearly label your graph.



b. What pattern(s) do you notice in the data?

Linear positive relationship. It's also proportional since it goes through the origin.

c. Draw a line of best fit on the graph above. Write the equation for the line below. Show how you determined the equation using calculations.

I think a reasonable line of best fit would go through (12,8) and (18, 12), and so the slope of that line would be $\frac{12-8}{18-12} = \frac{4}{6} = \frac{2}{3}$, which would make the equation $y = \frac{2}{3}x + b$…substitute in $(12, 8) for\ x$ and y and you get $8 = \frac{2}{3}(12) + b, b = 0$

Equation is $y = \frac{2}{3}x$

d. Verbally describe the relationship between the number of minutes Nicole plays and the number of points she scores. What does the slope mean in the context of the situation?

The slope is 2/3 which means that will score about 2 points for every 3 minutes she plays, or 2/3 or a point per minute.

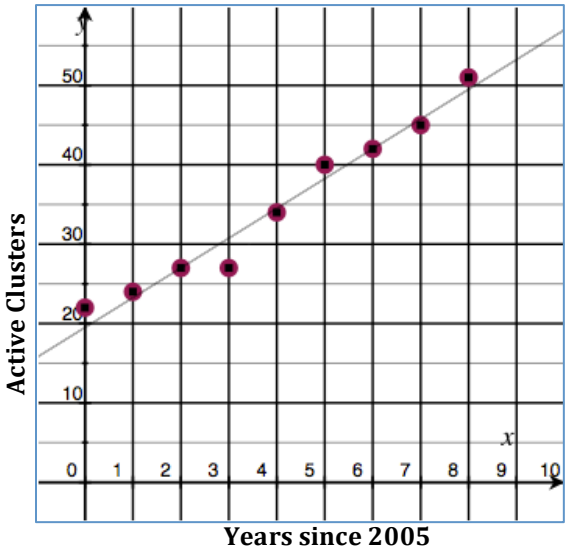
NAME:_____ Math 7.2, Period _____

Mr. Rogove Date:_____

3. The table shows the number of active woodpecker clusters in a part of the De Soto National Forest in Mississippi.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013
Active Clusters	22	24	27	27	34	40	42	45	51

a. Create a scatterplot of the data.
Represent the x-axis as the number of years since 2005.



b. One reasonable line of best fit goes through the 2007 and 2011 data. Find the equation of that line.

The 2007 data is $(2, 27)$ and the 2011 data is $(6, 42)$…so we find slope: $\frac{42-27}{6-2} = \frac{15}{4}$.

Now, to find y-intercept, plug in a coordinate and the slope:

$y = mx + b$ $\qquad 27 = 2\left(\frac{15}{4}\right) + b$

$27 = \frac{15}{2} + b$ → $\qquad b = \frac{39}{2}$

$$\boldsymbol{y = \frac{15}{4}x + \frac{39}{2}}$$

c. Predict the number of active clusters in 2020.

2020 is 15 since 2005, so the x-value would be 15…we need to find y when $x = 15$.

$$y = \frac{15}{4}(15) + \frac{39}{2}$$
$$y = \frac{225}{4} + \frac{78}{4}\ y$$
$$= \frac{303}{4} = 75\frac{3}{4}$$

There should be about 76 active clusters in 2020.

NAME:_____ Math 7.2, Period _____

Mr. Rogove Date:_____

4. A survey was conducted of 400 people that asked them questions about their gender and their preferred footwear. Some of the results are as follows:

- 240 people surveyed were female.
- 80 people surveyed preferred heels
- 60 females preferred sneakers
- 32 males preferred sandals
- 160 people surveyed preferred sneakers.
- 40 people surveyed preferred sandals
- 78 females preferred heels

a. Complete the two-way frequency table that summarizes the data on footwear and gender.

	Footwear Preference				
	Sneakers	**Heels**	**Sandals**	**Flats/Dress Shoes**	**Total**
Female	**60**	**78**	**8**	**94**	**240**
Male	**100**	**2**	**32**	**26**	**160**
Total	**160**	**80**	**40**	**120**	**400**

b. What proportion of the participants is female?
240/400 or .60

c. If there were no association between gender and footwear preference, should you expect more females than males to prefer sneakers or fewer females than males to prefer sneakers?
I would expect more females to prefer sneakers because there are more females surveyed.

d. Make a table of the row relative frequencies of each footwear type for the male and female row.

	Footwear Preference				
	Sneakers	**Heels**	**Sandals**	**Flats/Dress Shoes**	**Total**
Female	**.25**	**.33**	**.03**	**.39**	**1.00**
Male	**.63**	**.01**	**.20**	**.16**	**1.00**
Total	**.40**	**.20**	**.10**	**.30**	**1.00**

e. If you chose a survey participant at random, what kind of footwear would you expect them to prefer? Explain.
Sneakers because they are the most popular overall. 160 students prefer sneakers.

f. If you know that the randomly selected participant was a female, would this change the prediction from part (e)? Why or why not? What associations can you can make between the variables? Yes, it would probably be flats or dress shoes. More than you'd think, women prefer flat/heels.

NAME:_____ Math 7.2, Period _____

Mr. Rogove Date:_____

5. A survey of 58 7th grade students was conducted that asked many interesting questions about gender and salsa preference. Some results are as follows:

- 12 students total liked hot salsa
- 11 students liked medium salsa
- 13 boys like mild salsa
- 6 boys don't like salsa at all
- 15 students don't like salsa
- Only 2 girls like medium salsa
- 9 boys like hot salsa

a. Complete the two-way frequency table that summarizes the data on salsa preference and gender.

	Salsa Preference				
	No salsa	**Mild**	**Medium**	**Hot**	**Total**
Male	**6**	**13**	**9**	**9**	**37**
Female	**9**	**7**	**2**	**3**	**21**
Total	**15**	**20**	**11**	**12**	**58**

b. What proportion of the participants are females who like do not like salsa at all?
9/58 or .16

c. If there were no association between gender and salsa preference, would you expect to find that more girls do not like salsa or more boys do not like salsa? Explain your answer.
I would expect that more boys wouldn't like salsa because there are more boys surveyed overall.

d. Create a ROW relative frequency table of values for salsa preference for each gender.

	Salsa Preference				
	No salsa	**Mild**	**Medium**	**Hot**	**Total**
Male	**.16**	**.35**	**.24**	**.24**	**1.00**
Female	**.43**	**.33**	**.10**	**.14**	**1.00**
Total	**.26**	**.35**	**.19**	**.21**	**1.00**

e. Are there any associations you can make between salsa preference and gender? What are they?
It looks like there's an association between gender and those do not like salsa—if you don't like salsa you're more likely to be a girl. While not as strong, there looks to be an association between gender and medium and hot in that boys are more likely than girls to like medium and hot salsa.

NAME:_____ Math 7.2, Period _____

Mr. Rogove Date:_____

6. In the same survey we asked students about the amount of sleep they got and the time they went to bed. Below are the results in a two way table.

		BEDTIME			
		Between 8PM - 9PM	**Between 9 - 10PM**	**After 10PM**	**Total**
Sleep each night	**Less than 6**	0	0	1	1
	Between 6 and 8	2	18	7	27
	More than 8	8	21	1	30
	TOTAL	10	39	9	58

a. Make a conclusion (based on math) about the association between going to bed after 10PM and getting less than 8 hours of sleep. Write a few sentences explaining your thoughts.

If you go to bed after 10PM, you're more likely to get less than 8 hours of sleep. While less than half of all respondents (28/58) reported getting less than 8 hours of sleep, those who went to bed after 10PM overwhelmingly (8/9) got less than 8 hours of sleep.

b. Create a COLUMN relative frequency table below

		BEDTIME			
		Between 8PM - 9PM	**Between 9 - 10PM**	**After 10PM**	**Total**
Sleep each night	**Less than 6**	**0**	**0**	**.11**	**.02**
	Between 6 and 8	**.20**	**.46**	**.78**	**.46**
	More than 8	**.80**	**.54**	**.11**	**.52**
	TOTAL	**1**	**1**	**1**	**1**

c. Based on your column relative frequency table, for which bedtime category is there the least association with the amount of sleep? Explain this.

Because the 9-10PM bedtime relative frequencies most closely resemble the total relative frequency, I would say that column has the least association…If you go to bed between 9-10PM, it provides little information on how many hours of sleep you might get.