Chapter 10: Statistics

Index:

Name: _____   Date: _____ Period: _____
Algebra   Graphical Displays of Data   10A

Quantitative data on a single variable is often collected in order to understand how a characteristic of a group differs amongst the group members or between groups. When we ask a question like "How old is a typical fast food worker?" it is helpful to take a **survey** and then see graphically how the ages differ amongst the group.
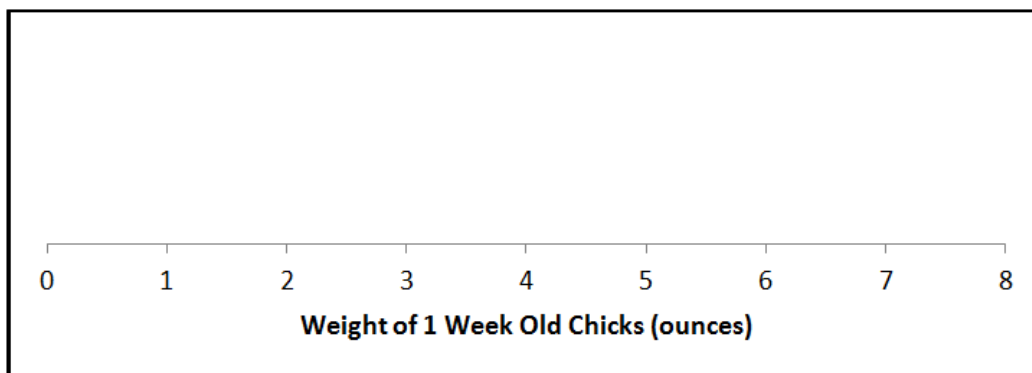
Exercise #1: Charlie's Food Factory currently employs 28 workers whose ages are shown below on a **dot plot**. Answer the following questions based on this plot.

Ages of Workers at Charlie's Food Factory
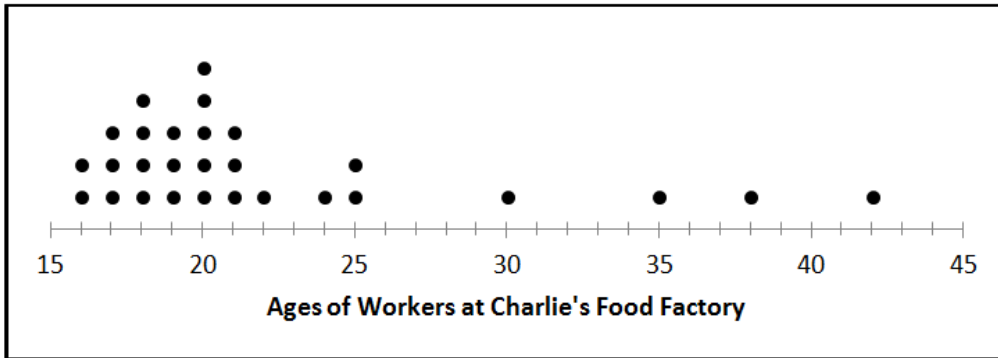
(a) How many of the workers are 18 years old?

(b) What is the **range** of the ages of the workers?

(c) Would you consider this distribution symmetric?

(d) The mean (average) age for a worker is 22 years old. Why is this average not representative of a typical worker?

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

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

Weight of 1 Week Old Chicks (ounces)

2

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

(a) How many workers have ages between 19 and 21 years?

Ages of Workers at Charlie's Food Factory

Number of Workers

16-18 19-21 22-24 26-28 29-31 32-34 35-37 38-40 42-44

Age Ranges of Workers

(b) What is the disadvantage of a histogram compared to a dot plot?

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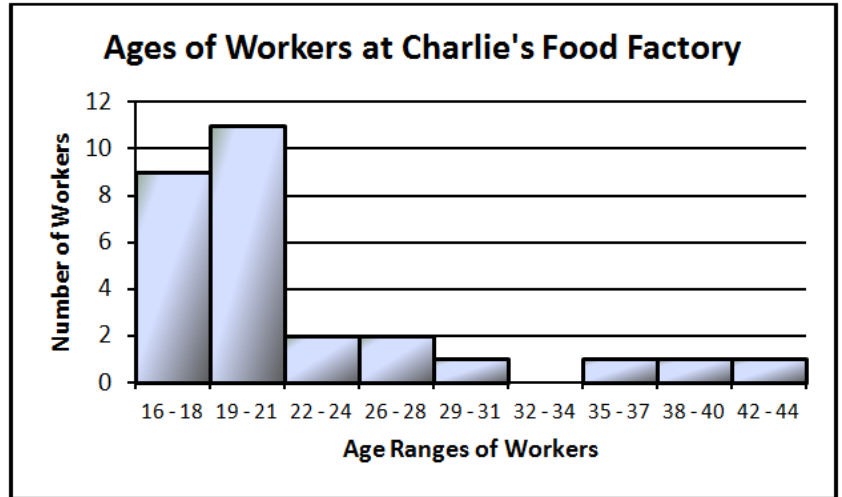
Exercise #4: The 2006 – 2007 Arlington High School Varsity Boy's basketball team had an excellent season, compiling a record of 15 – 5 (15 wins and 5 losses). The total points scored by the team for each of the 20 games are listed below in the order in which the games were played:

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

(a) Complete the frequency table below.

(b) Construct the histogram below.

POINTS SCORED	TALLY	FREQUENCY
40 - 49		
50 - 59		
60 - 69		
70 - 79		
80 - 89		
90 - 99		
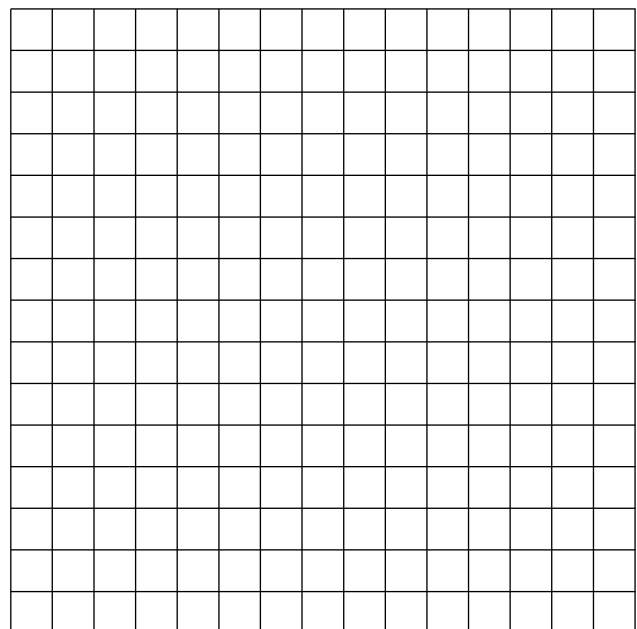
A local marketing company did a survey of 30 households to determine how many devices the household contained that family members watched video on (i.e. TV's, tablets, smart phones, etcetera). The dot plot of the responses is shown below.

Number of Screens Capable of Watching Video
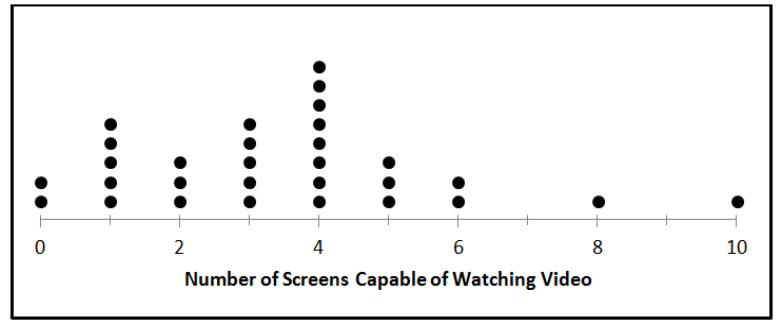
_____1. How many households have three devices capable of showing video on them?
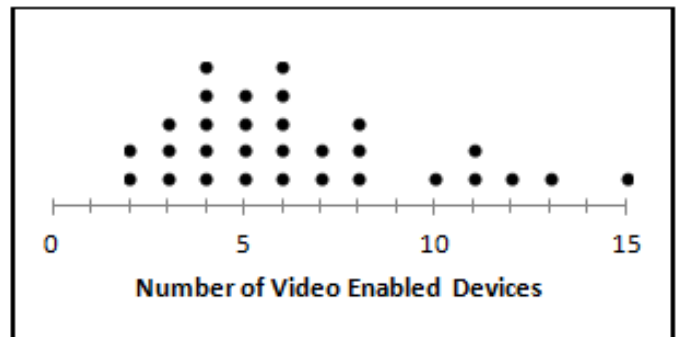
(1) 1   (3) 7
(2) 2   (4) 5

_____2. More households had 4 devices to watch video on than any other number. Which of the following
    is closest to the percent of households that have 4 devices?
(1) 22%   (3) 27%
(2) 34%   (4) 45%

3. The marketing company would like to claim that the majority of households have either 3 or 4 screens capable of watching video on. Does the information displayed on the dot plot support this claim? Explain your reasoning.

4. The same marketing company then surveyed 30 households that contained at least one teenager. The dot plot for the video enabled devices is shown below. The mean number of screens for the first survey was 3.4. Based on the second dot plot, do you think its mean will be higher or lower? Explain.


Number of Video Enabled Devices

On a recent Precalculus quiz, Mr. Weiler found the following distribution of scores, which are arranged in 5 point intervals (with the exception of the last interval).

___5. How many students scored in the 75 to 79 point range?

(1) 8   (3) 25

(2) 10   (4) 5

___6. Students do not pass the quiz if they receive lower than a 70. How many students did not pass?

(1) 8   (3) 7

(2) 5   (4) 15

___7. How many total student took the quiz?

(1) 25   (3) 56

(2) 104   (4) 91

8. Twenty-two students scored in the 80 to 84 range on this test. Does the histogram provide us with enough information to conclude that a student must have scored on 82 on this test? Explain your thinking.

9. A random survey of 100 cars found the following frequency distribution for the fuel efficiency of the car, as measured in miles per gallon. Construct a histogram below that effectively shows the distribution of this data set.

Fuel Efficiency (miles per gallon)	Number of Cars
10 to 14	4
15 to 19	17
20 to 24	36
25 to 29	24
30 to 34	10
35 to 39	6
40 to 44	3

_____ 10.) The local deli charges a fee for delivery. On Monday, they delivered two dozen bagels to an office at a total cost of $8. On Tuesday, three dozen bagels were delivered at a total cost of $11. Which system of equations could be used to find the cost of a dozen bagels, b, if the delivery fee is f?

(1) $b + 2f = 8$
$\quad b + 3f = 11$

(3) $b + 2f = 8$
$\quad 3b + f = 11$

(2) $2b + f = 8$
$\quad b + 3f = 11$

(4) $2b + f = 8$
$\quad 3b + f = 11$

_____ 11.) Which inequality is shown in the graph below?
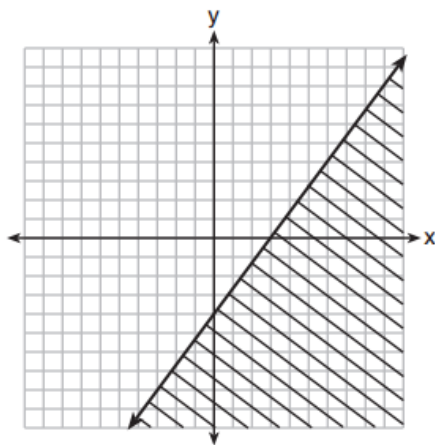
(1) $y \le \frac{4}{3}x + 3$

(3) $y \le \frac{4}{3}x - 4$

(2) $y \ge \frac{4}{3}x + 3$

(4) $y \ge \frac{4}{3}x - 4$

12.) Write a quadratic equation in standard form that has roots of -12 and 2.

Name: _____

Homework Answers

Date: _____ Period: _____
Algebra   Graphical Displays of Data   10A HW

1.) 4

2.) 3

3.) Information does not support the claim

4.) The mean for this distribution will be higher.
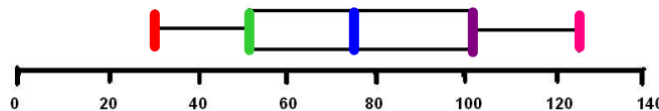
5.) 2

6.) 3

7.) 4

8.) No

9.) GRAPH

10.) 4

11.) 3

12.) $x^2 + 10x - 24 = 0$

Box and Whisker Plots are graphs created to display data (very similar to a histogram) at a quick glance. Box and Whisker Plots can also be called Box Plots. There are 5 key pieces.

Annual snow depth at Mathsville Ski Resort

Lowest Value:
The smallest number that is in the set of data.

Highest Value:
The largest number that is in the set of data.

Lower (First) Quartile:
The median of the lower half of the given

Median (Second Quartile):
The middle of the given data Represents 50%

Upper (Third) Quartile:
The median of the upper half of the given data.
Represents 75%

Range:
The amount of numbers represented in the set of data. To identify the range, you will subtract the lowest value from the highest value.

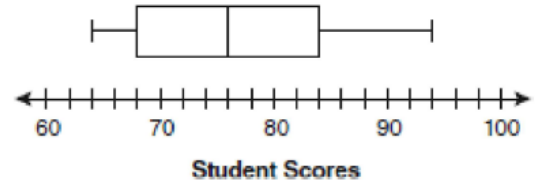*Exercise #1*: Given the following box and whisker plot what is:

a) the lowest value:____**55**_____

b) the highest value:_____**100**____

c) the lower quartile:____**60**____

d) the upper quartile:_____**85**____

e) the median value:____**75**_____

f) What is the range? ___**100 - 55 = 45**____

8
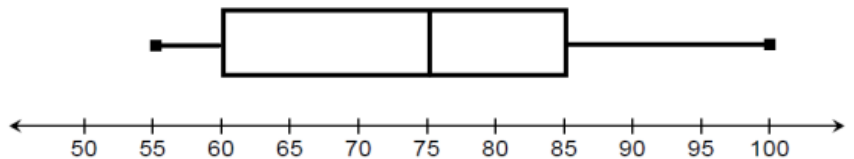
Exercise #2: The box-and-whisker plot below represents students scores on a recent English test. What is the value of the lower quartile?

    (1) 68   (2) 76

    (3) 84   (4) 94

Student Scores

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

Exercise #3: Shown below are the scores 16 students received on a math quiz.

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

(a) What is the median of this data set?

(b) Find the range of the data set.

(c) What is the median of the lower half of this data set? What is another name for this term?

(d) What is the median of the upper half of this data? What is another name for this term?

Exercise #4: Using the same data set construct a box plot on the number line given below.

Exercise #5: The ages of the 15 employees of the Red Hook Curry House are given below.

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

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

Lowest Value = _____   Highest Value = _____
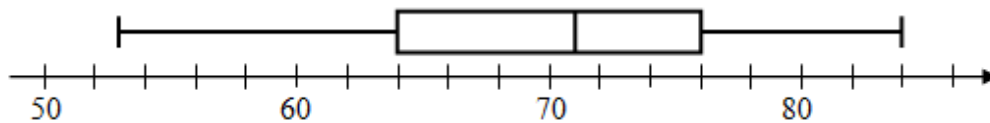
Lower Quartile = _____   Median = _____   Upper Quartile = ____

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

Exercise #6: Twenty of Mr. Ouimet's math students recently took a quiz. The results of this quiz are shown in the following box-and-whiskers diagram. Assume that all scores are whole numbers.

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

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

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

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

____ 1. Which of the following data sets, given in ascending order, has the greatest range?
     (1) $\{3, 4, 7, 10, 18\}$ (3) $\{-2, 5, 8, 11, 26\}$

     (2) $\{65, 66, 70, 72\}$ (4) $\{-5, -2, 4, 7, 10\}$

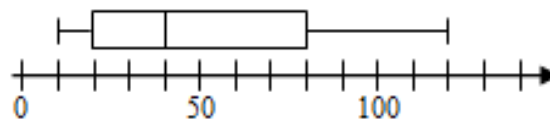____ 2. Given the box plot shown below, which of the following represents the third quartile value for this data set?

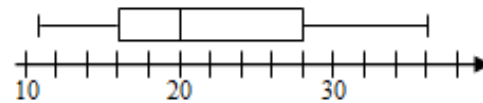   (1) 12   (3) 6

     (2) 18   (4) 19
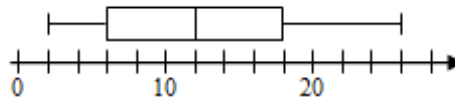
____ 3. Given the box plot shown below, which of the following represents the range of this data set?

     (1) 110   (3) 60

     (2) 40   (4) 75

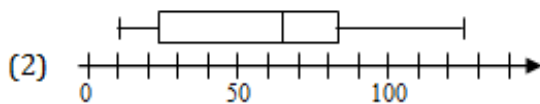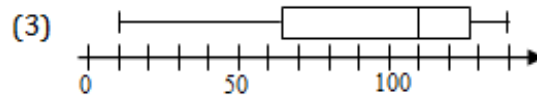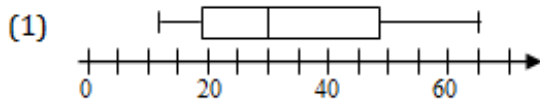____ 4. According to the following box-and-whiskers diagram, which of the following values represents the lower
     quartile of this data set?

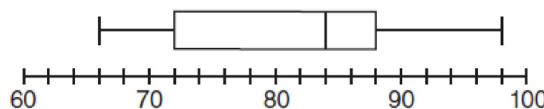   (1) 20   (3) 28

     (2) 13   (4) 16

____ 5. Which of the following box-and-whiskers diagram represents a data set whose median value is equal to 65?

     (1)   (3)

     (2)   (4)
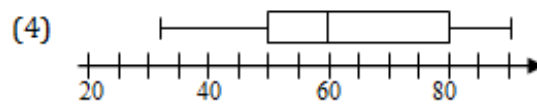
____ 6. The box-and-whisker plot below represents the math scores of 20 students. What percentage
       of the test scores are less than 72?
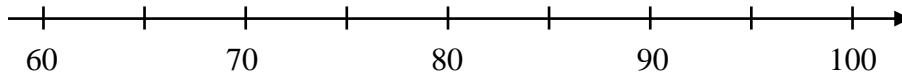   (1) 25   (2) 50
       (3) 75   (4) 100

7. Mr. Ramirez gives a math test and records the grades of his 17 students as follows:

67, 72, 74, 74, 78, 80, 80, 82, 85, 85, 86, 87, 90, 92, 92, 95, 98

Create a box-and-whisker diagram of this data set below.

```
  +----+----+----+----+----+----+----+----+----→
  60        70        80        90        100
```

8. The speeds, in miles per hour, of 24 cars on a particular road are recorded and represented on the box-and-whiskers diagram shown below. Answer each of the following questions based on this diagram.

(a) What is the range of this data set?

```
      |----[----|----]---|
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-→
      20        30        40        50
```
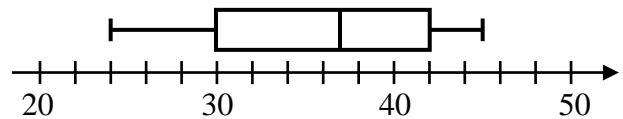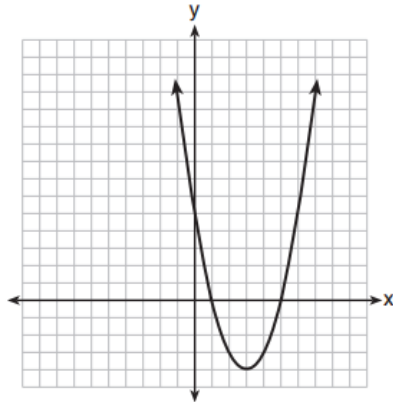
(b) What is the maximum speed of the 24 drives?

(c) How many drivers drove between 30 and 42 miles per hour?

(d) If the speed limit on this part of the road is 35 miles per hour, are more people speeding or are more people going below the speed limit? Justify your answer.

Review Section:

____ 9.) The equation $y = ax^2 + bx + c$ is graphed on the set of axes below.

Based on the graph, what are the roots of the equation $ax^2 + bx + c = 0$?

(1) 0 and 5   (3) 1 and 5

(2) 1 and 0   (4) 3 and −4

____ 10.) What is the solution of $4x - 30 \geq -3x + 12$?

(1) $x \geq 6$ (3) $x \geq -6$

(2) $x \leq 6$ (4) $x \leq -6$

11.) Solve the following system of equations algebraically for all values of x and y.

$$y = x^2 + 2x - 8$$
$$y = 2x + 1$$

13

1.) 3

2.) 2

3.) 1

4.) 4

5.) 2

6.) 1

7.) Box Plot

8.) a.) range = 21 mph
    b.) 45 mph
    c.) 50% of the drives fall into this range
    d.) More people are speeding.

9.) 3

10.) 1

11.) $(-3, -5)$ and $(3, 7)$

Name: _____   Date: _____ Period: _____
Algebra   Measures of Central Tendency   10C

In our day to day activities, we deal with many problems that involve related items of numerical information called *data*. *Statistics* is the study of sets of such numerical data. When we gather numerical data, besides displaying it, we often want to know a single number that is representative of the data as a whole. We call these types of numbers **measures of central tendency**. The two most common measures of central tendency are the **mean** and the **median**.

Exercise #1: A survey was taken amongst 12 people on the number of passwords they currently have to remember. The results in ascending order are shown below. State the median number of passwords and the mean number of passwords (to the nearest tenth).

<p align="center">0, 1, 1, 1, 2, 2, 3, 3, 3, 3, 4, 6</p>

Exercise #2: Students in Mr. Tobin's algebra class were trying to determine if people speed along a certain section of roadway. They collected speeds of 10 vehicles, as displayed in the table below.
(a) Find the mean and median for this data set.

Speed (mph)	Number of Cars
29	1
33	4
34	2
35	3

(b) The speed limit along this part of the highway is 34 mph. Based on your results from part (a), is it fair to make the conclusion that the average drives foes speed on this roadway?

When conducting a statistical study, it is not always possible to obtain information about every person or situation to which the study applies. Unlike a census, in which every person is counted, some studies use only a **sample** or **portion** of the items being investigated. Whenever a sample is taken, it is vital that it be **fair**; in other words, the sample reflects the overall population.

Exercise #3: To determine which television programs are the most popular in a large city, a poll is conducted by selecting a sample of people at random and interviewing them. Outside which of the following locations would the interviewer be most likely to find a fair sample? Explain your choice and why the others are inappropriate.

    (1) A baseball stadium   (3) A grocery store

    (2) A concert hall   (4) A comedy club

Exercise #4: Truong is trying to determine the average height of high school male students. Because he is on the basketball team, he uses the heights of the 14 players on the team, which are given below in inches.

$$69, 70, 72, 72, 74, 74, 74, 75, 76, 76, 76, 77, 77, 82$$

(a) Calculate the mean and median for this data set. Round any non-integer answers to the nearest tenth.

(b) Is the data set above a **fair sample** to use to determine the average height of high school male students? Explain your answer.

Data sets can have members that are far away from all of the rest of the data set. These elements are called **outliers,** which can result in a mean that does not represent the true "average" of a data set.

Exercise #5: In Mr. Petrovic's Advanced Calculus, eight students recently took a test. Their grades were as follows:
45, 78, 82, 85, 87, 89, 93, 95

(a) Calculate the mean and median of this data set.   (b) What score is an **outlier** in this data set?

(c) Which value, the mean or the median, is a better measure of how well the average student did on Mr. Petrovic's quiz?

Name: _____   Date: _____ Period: _____
Algebra   Measures of Central Tendency   10C HW

____ 1. The Student Government at Arlington High School decided to conduct a survey to determine where to go on
a senior field trip. They asked students the following question: "Would you rather go to a sports event or to an
IMAX movie?" At which of the following locations would they most likely get a fair sample?

(1) The gym, after a game   (3) A randomly chosen study hall

(2) The auditorium after a play   (4) At the Nature Club meeting.

2. For the following data set, calculate the mean and median. Any non-integer answers should be rounded to the
nearest tenth.

$$3, 5, 8, 8, 12, 16, 17, 20, 24$$

3. For the following data set, calculate the mean and median. Any non-integer answers should be rounded to the
nearest tenth.

$$5, 5, 9, 10, 13, 16, 18, 20, 22, 22$$

____ 4. Which of the following is true about the data set {3, 5, 5, 7, 9}?

(1) median > range   (3) mean > median

(2) median = mean   (4) median > mean

____ 5. Which of the following data sets has a median of 7.5?

(1) {6, 7, 8, 9, 10}   (3) {1, 3, 7, 10, 14}

(2) {3, 5, 7, 8, 10, 14}   (4) {2, 7, 9, 11, 14, 17}

6. A survey is taken by an insurance company to determine how many car accidents the average New York City resident has gotten into in the past 10 years. The company surveyed 20 people who are getting off a train at a subway station. The following table gives the results of the survey.

(a) Calculate the mean and median number of accidents of this data set. Remember, there are 6-zeros in this data set, 8-1's, etc.

Number of Accidents	Number of People
0	6
1	8
2	4
3	1
11	1

(b) Are there any outliers in this data set? If so, what data value?

(c) Which number, the mean or the median, better represents the number of accidents an average person in this survey had over this 10 year period? Explain your answer.

(d) Does this sample fairly represent the average number of accidents a typical New York City resident would get into over a 10 year period? Why or why not?

(e) Construct a dot plot that represents this data on the set of axes provided.

Is this a symmetric plot? Explain your thinking.

Number of Traffic Accidents

19
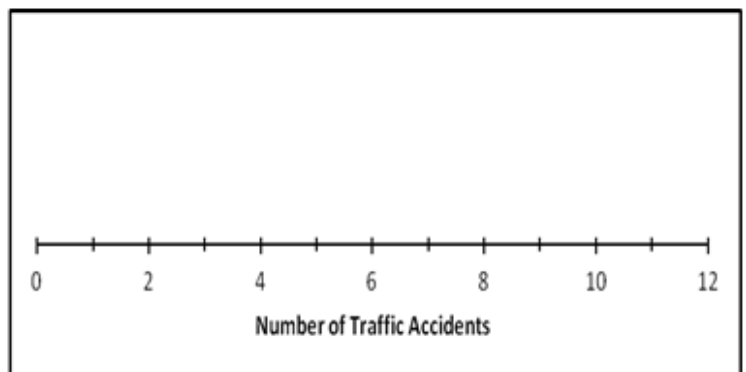
Review Section:

_____ 7.) What is an equation of the line that passes through the points $(2,1)$
and $(6,-5)$?

(1) $y = -\frac{3}{2}x - 2$ (3) $y = -\frac{2}{3}x - 1$

(2) $y = -\frac{3}{2}x + 4$ (4) $y = -\frac{2}{3}x + \frac{7}{3}$

8.) Carla bought a dress at a sale for 20% off the original price. The sale price of the dress was $28.80.
Find the original price of the dress, in dollars.

9.) A DVD costs twice as much as a music CD. Jack buys 2 DVDs and 2 CDs and spends $45.
Determine how much one CD costs, in dollars. [Only an algebraic solution can receive full credit.]

Name: ___*Homework Answers*_____   Date: _____ Period: _____
Algebra   Measures of Central Tendency   10C HW

1.) 3

2.) Mean = 12.6   Median = 12

3.) Mean = 14   Median = 14.5

4.) 3

5.) 2

6.) a.) Mean = 1.5   Median = 1 accident
    b.) Yes, the one person who had 11 accidents
    c.) The median best represents the average number of accidents a person has over this 10 year period.
    d.) Probably not.
    e.) It is NOT symmetric.

7.) 2

8.) The original price of the dress is $36.00

9.) One CD costs $7.50

Name: _____   Date: _____ Period: _____
Algebra   Variation within a Data Set   10D

Measures of central tendency give us numbers that describe the typical data value in a given data set. But, they do not let us know how much **variation** there is in the data set. Two data sets can have the same mean but look radically different depending on how varied the numbers are in the set.

Exercise #1: The scores of a recent math quiz are represented below. What are the values of the first quartile and the third quartile? What is the value of the interquartile range?

First Quartile = 6   Third Quartile = 10

Interquartile Range = 10 – 6 = 4
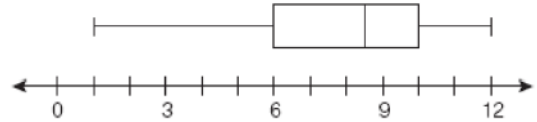
Exercise #2: The two data sets below each have equal means but differ in the variation within the data set. Use your calculator to determine the **Interquartile Range (IQR)** of each data set. The IQR is defined as the difference between the third quartile value and the first quartile value.

Data Set #1: 3, 3, 4, 4, 5, 5, 6, 6, 7, 8, 8, 9, 9, 10, 10, 11, 11   **Data Set #2:** 5, 5, 6, 6, 7, 7, 8, 8, 9, 9

The interquartile range gives a good measure of how spread out the data set is. But, the best measure of variation within a data set is the **standard deviation**. The actual calculation of standard deviation is complex and we will not go into it here. We will rely on our ***calculators*** for its calculation.

How to use the calculator to find the Standard Deviation:

1.) Press Stat

2.) Press Edit (if you have data in the list already, use the up arrow to highlight L1, then press clear and enter).

3.) Type in your data set. Be very careful and take your time!

4.) Press Stat

5.) Scroll to the right to (Calc)

6.) Choose option #1 (1-Var Stats)

7.) Press Enter three times

8.) $\bar{x} = Mean$

$n = Sample\ Size$

$Sx = Sample\ Size\ Standard\ Deviation$

$\sigma x = Population\ Standard\ Deviation$

Exercise #3: Using the same data sets, use your calculator to produce the mean, sample size and standard deviation (shown as σ_x on the calculator) of the two data sets. Round your answers to the nearest tenth.

Data Set #1: 3, 3, 4, 4, 5, 5, 6, 6, 7, 8, 8, 9, 9, 10, 10, 11, 11   **Data Set #2:** 5, 5, 6, 6, 7, 7, 8, 8, 9, 9

   Mean = _____   Mean = _____

   Sample Size = _____   Sample Size = _____

   σx = _____ σx = _____

Standard Deviation

The **standard deviation** of a data set tells us, on **average**, how **far** a **data point** is **away** from the **mean** of the data set. The **larger** the **standard deviation**, the **greater** the **variation** within the data set.

Exercise #4: A farm is studying the weight of baby chickens (chicks) after 1 week of growth. They find the weight, in ounces, of 20 chicks. The weights are shown below. Find the mean, the interquartile range and the standard deviation for this data set. Round any non-integer values to the nearest tenth. Include appropriate units in your answers. Give an interpretation of the standard deviation.

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

Mean = _____   Interquartile Range = _____ σx = _____

Exercise #5: A marketing company is trying to determine how much diversity there is in the age of people who drink different soft drinks. They take a **sample** of people and ask them which soda they prefer. For the two sodas, the age of those people who preferred them is given below.

Soda A: **Soda B:**

18, 16, 22, 16, 28, 18, 21, 38, 22, 29, 25, 44, 36, 27, 40   25, 22, 18, 30, 27, 19, 22, 28, 25, 19, 23, 29, 26, 18, 20

(a) Explain why standard deviation is a better measure of the diversity in age than the mean.

(b) Which soda appears to have a greater diversity in the age of people who prefer it? How did you decide on this?

(c) Use your calculator to determine the **sample standard deviation**, normally given as s_x, for both data sets. Round your answers to the nearest tenth. Did this answer reinforce your pick from (b)? How?

Population Versus Sample Standard Deviation

When we are working with **every possible data point** of interest, we call this a **population** and use the population standard deviation, σ. When we have only a **sample** of all possible values we use the sample standard deviation, s. The formulas for these two differ very slightly, so their values tend to be slightly different.

Exercise #6: Which of the following data sets would have a standard deviation (population) closest to zero? Do this without your calculator. Explain how you arrived at your answer.

(1) $\{-5, -2, -1, 0, 1, 2, 5\}$

(2) $\{5, 8, 10, 16, 20\}$

(3) $\{11, 11, 12, 13, 13\}$

(4) $\{3, 7, 11, 11, 11, 18\}$

1. For each of the following data sets, use your calculator to help find the interquartile range and the population standard deviation. Show your calculation for the IQR. Round all non-integer values to the nearest *tenth*.

    (a) 4, 6, 8, 10, 15, 19, 22, 25

    (b) 3, 3, 4, 5, 5, 6, 6, 7, 7, 8

____ 2. For the data set shown in the dot plot below, which of the following is closest to its population standard deviation?

(1) 2.7   (3) 3.3

(2) 4.2   (4) 5.8
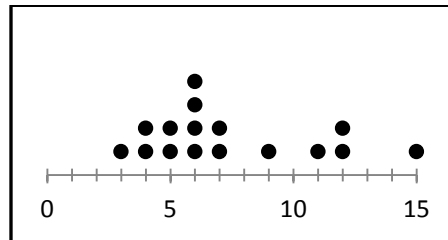
____ 3. What is the interquartile range of the data set represented in the box plot shown below?

(1) 24   (3) 8

(2) 14   (4) 12
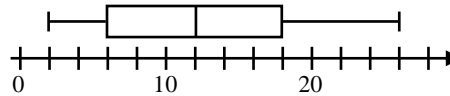
____ 4. Which of the following best measures the average distance that a data value lies away from the mean?

(1) mean   (3) median

(2) standard deviation   (4) range

____ 5. Which of the following data sets would have the largest standard deviation?

(1) $\{3, 3, 4, 5, 5\}$ (3) $\{2, 8, 18, 26, 35\}$

(2) $\{72, 73, 74, 75, 76\}$ (4) $\{8, 10, 12, 14, 16\}$

6. We are going to revisit our survey of households that have video enabled devices (televisions, smart phones, tablets, etcetera). Recall that two surveys were done, each with 30 participants. In the first case (Survey A), the survey was random, in the second case (Survey B), the survey only included families with at least one teenager. The dot plots of the results are shown below.
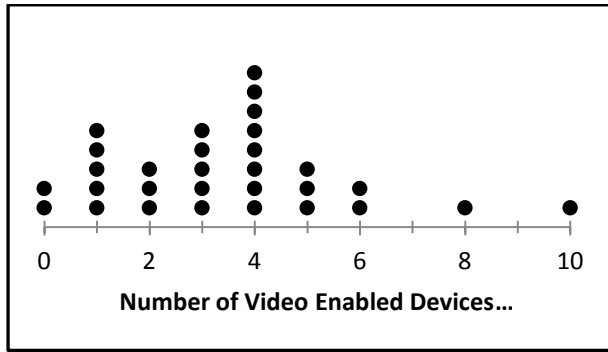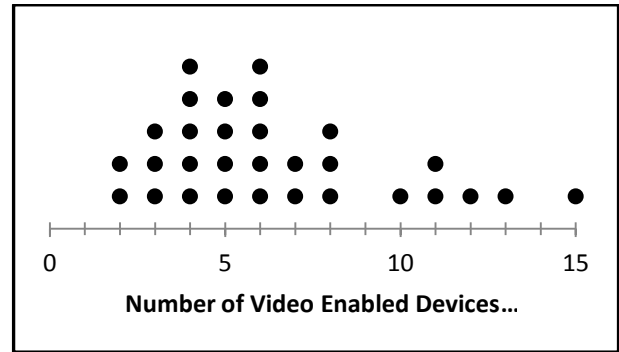
Survey A   Survey B

(a) Enter the data into your calculator and use it to calculate the mean number of devices, the interquartile range, and the standard deviation of both data sets. Round all non-integers to the nearest tenth. Remember, you will have to enter a given data point more than once. For example, in Survey A, you will need to enter 2-0's, 5-1's, 3-2's, etcetera. Use the **sample standard deviation**.

Survey A Statistics: **Survey B Statistics:**

(b) Which of these two survey data sets had the greatest variation in the data? Explain based on the statistics you found in part (a).

(c) How many of the 30 values in Survey B fall within one standard deviation of the mean? To do this calculation, add the standard deviation and subtract the standard deviation from the mean and then count the number of values between the results of this addition and subtraction.

_____ 7.) The expression $\dfrac{24x^6y^3}{-6x^3y}$ is equivalent to

(1) $-4x^2y^3$ (3) $-4x^9y^4$

(2) $-4x^3y^3$ (4) $-4x^3y^2$

_____ 8.) In the box-and-whisker plot below, what is the 2nd quartile?
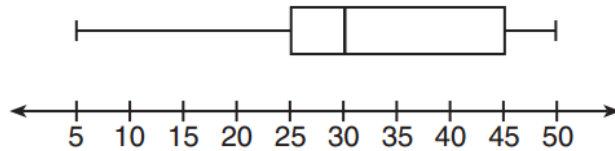
(1) 25   (3) 45

(2) 30   (4) 50

9.) The equations $6x + 5y = 300$ and $3x + 7y = 285$ represent the money collected from selling gift baskets in a school fundraising event. If x represents the cost for each snack gift basket and y represents the cost for each chocolate gift basket, what is the cost for each chocolate gift basket?

Name: ___*Homework Answers*_____   Date: _____ Period: _____
Algebra   Variation within a Data Set   10D HW

1.) a.) $Q_1 = 7$, $Q_3 = 20.5$, Interquartile Range = 13.5 σx = 7.3

   b.) $Q_1 = 4$, $Q_3 = 7$, Interquartile Range = 3 σx = 1.6

2.) 3

3.) 4

4.) 2

5.) 3

6.) a.) Survey A: Mean = 3.4, $Q_1 = 2$, $Q_3 = 4$, Interquartile Range = 2 σx = 2.2
        Survey B: Mean = 6.4, $Q_1 = 4$, $Q_3 = 8$, Interquartile Range = 4 σx = 3.3

   b.) Survey B

   c.) Low data point = 3.1   High data point = 9.7   19 values

7.) 4

8.) 2

9.) Each chocolate gift basket costs $30.

So far we have worked with **quantitative data** for a **single variable**, for example weight of baby chicks or number of video enabled devices. We can also work with **categorical data** or data that shows how many things **surveyed** fall into a **given category**.

Exercise #1: Let's do a quick categorical survey in this class. By a show of hands, determine how many students fall into each of the following categories for eye color.

Brown Eyes	Blue Eyes	Green Eyes	Other Color

Although surveys of data that contain only one category are interesting, statisticians are often interested in how responses to two categories relate to one another. For example, we may want to know how a person's gender (one category) affects what profession (a second category) they would prefer when they grow up. We may want to know if a person's hair color (one category) has any relationship to their eye color (a second category). This type of data is summarized in a **two-way frequency table**.

Exercise #2: A class of 20 students recorded their hair color and eye color which are shown in the **two-way frequency table** below.

		Hair Color			
		Black	Blond	Red	Total
Eye Color	Blue	3	4	1	8
	Brown	5	2	0	7
	Green	1	1	3	5
	Total	9	7	4	20

(a) How many students had blond hair and blue eyes?   (b) How many students had black hair?

   4 students have blond hair and blue eyes   **9 students have black hair**

(c) Construct a table that shows the **joint relative frequencies** and the **marginal relative frequencies (percentages)** for the data above.

		Hair Color			
		Black	Blond	Red	Total
Eye Color	Blue	$\frac{3}{20} = .15 = 15\%$	$\frac{4}{20} = .2 = 20\%$		
	Brown	$\frac{5}{20} = .25 = 25\%$	$\frac{2}{20} = .1 = 10\%$		
	Green	$\frac{1}{20} = .05 = 5\%$	$\frac{1}{20} = .05 = 5\%$		
	Total	$\frac{9}{20} = .45 = 45\%$	$\frac{7}{20} = .35 = 35\%$		

29

We would like to understand **associations** or **trends** within the data set, i.e. would a response to one category tell us something about the response to the other category?

Exercise #3: Let's see if there is a connection between eye color and hair color by using **conditional relative frequencies**.

		Hair Color		
		Black	Blond	Red	Total
Eye Color	Blue	3	4	1	8
	Brown	5	2	0	7
	Green	1	1	3	5
	Total	9	7	4	20

(a) What is the conditional relative frequency of having green eyes if you have red hair? (This is equivalent to asking what the percent of people with red hair have green eyes.)

(b) What is the conditional relative frequency of having green eyes if you have black hair?

(c) Does it appear that having green eyes has a dependency or at least an association with having red hair? Explain.

(d) Is it more likely that a person with black hair has blue eyes or that a person with blond hair has brown eyes? Use conditional marginal frequencies to support your answer.

Exercise #4: A survey of 52 graduating seniors was conducted to determine if there was a connection between the gender of the student and whether they were going on to college. Based on this data, what is more likely: that someone going to college is female or that someone who is female is going to college? These may seem like the same thing, but are quite different.

	Gender		
	Male	Female	Total
Going to College	16	13	29
Not Going to College	14	9	23
Total	30	22	52

A survey was done to determine the relationship between gender and subject preference. A total of 56 students were surveyed to determine if they liked math, English, social studies, or science as their favorite subject. The results were then broken down based on whether the respondent was male or female.

	Math	English	Social Studies	Science	Total
Female	8	6	11	5	30
Male	10	4	8	4	26
Total	18	10	19	9	56

____ 1. Which of the following is closest to the joint relative frequency of being a male who likes social studies?

(1) 0.42   (3) 0.31

(2) 0.14   (4) 0.56

____ 2. Which of the following is the marginal relative frequency of liking math?

(1) $\dfrac{18}{36}$ (3) $\dfrac{10}{18}$

(2) $\dfrac{8}{10}$ (4) $\dfrac{18}{56}$

____ 3. What percent of female students liked English as their favorite subject?

(1) 20%   (3) 11%

(2) 16%   (4) 60%

4. A person looking at this table concludes that it is more likely that a female student will like social studies than a male student will like math. Is this correct? Justify your answer.

5. Is it more likely that a person who likes social studies will be female or that a person who is female will like social studies? Justify.

Demographers are trying to understand the association between where a person lives and how they commute to work. They survey 100 people in three cities with the results shown below.

	Car	Train	Walk	Total
New York	5	25	10	40
Los Angeles	18	12	5	35
Chicago	8	14	3	25
Total	31	51	18	100

6. Fill in the table below with the relative frequencies.

	Car	Train	Walk	Total
New York				
Los Angeles				
Chicago				
Total				

_____ 7. Given that a person rides a train to work, what is the conditional relative frequency that they live in New York?

(1) 0.25   (3) 0.49

(2) 0.63   (4) 0.82

_____ 8. If a person lives in Los Angeles, what is the conditional relative frequency that they drive a car?

(1) 0.42   (3) 0.68

(2) 0.16   (4) 0.51

_____ 9. Which of the following is the marginal frequency of walking to work?

(1) 18%   (3) 25%

(2) 60%   (4) 44%

10. Is a person more likely to ride a train if they live in New York or if they live in Chicago? Justify your answer.

Review Section:

____ 11.) When $9x^2 - 100$ is factored, it is equivalent to $(3x - b)(3x + b)$.
What is a value for b?

(1) 50   (3) 3

(2) 10   (4) 100

____ 12.) What is the slope of a line passing through points $(-7,5)$ and $(5,-3)$?

(1) $-\frac{3}{2}$ (3) $\frac{2}{3}$

(2) $-\frac{2}{3}$ (4) $\frac{3}{2}$

13.) On the set of axes below, solve the following system of inequalities graphically.

$$y + 3 < 2x$$
$$-2y \le 6x - 10$$

State the coordinates of a point in the solution set.
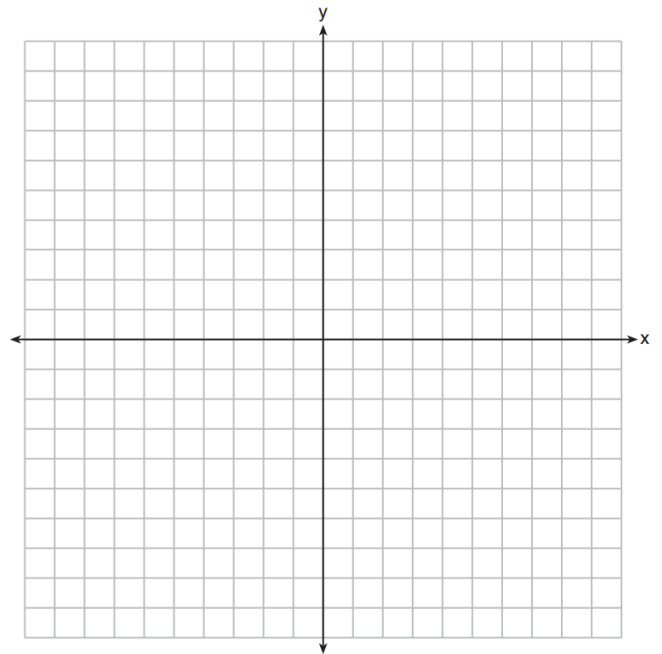
Name: _____ *Homework Answers*   Date: _____ Period: _____
Algebra   Two Way Frequency Tables   10E HW

1.) 2

2.) 4

3.) 1

4.) Its more likely that a male student will like math more than a female student will like social students.

5.) It is more likely that a person who likes social studies will be female that a person who is female will like social studies.

6.) Table

7.) 3

8.) 4

9.) 1

10.) A Person is more likely to be riding the train if they live in New York than in Chicago.

11.) 2

12.) 2

13.) GRAPH

Exercise #1: A survey was taken of 8 low and high temperatures, in Fahrenheit, in the month of April to try to establish a relationship between a day's low temperature and high temperatures.

Low Temperature, x	22	24	26	28	30	32	34	36
High Temperature, y	30	38	34	42	53	46	58	56

(a) Given the following data points, plot on the following graph grid.
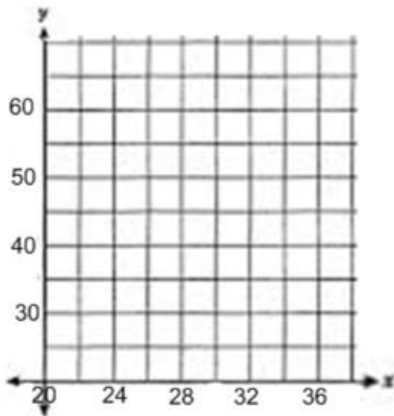
This is called a **Scatter Plot**.
Scatter plots are similar to line graphs in that they use horizontal and vertical axes to plot data points. However, they have a very specific purpose. Scatter plots show how much one variable is affected by another. The relationship between two variables is called their **correlation** .

The **line of best fit** , also called a trendline or a **linear regression**, is a straight line that best illustrates the overall picture of what the collected data is showing. It helps us to see if there is a relationship or correlation between the two factors being studied. This trendline helps us to predict future events relating to the data being studied.
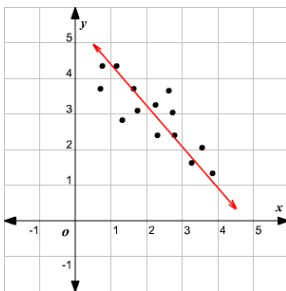
***Exercise* #2:** Which scatter plot has the most accurate line of best fit?

(1)

Hours of Sleep vs. Amount of Coffee Drank

(2)

Hours of Sleep vs. Amount of Coffee Drank

(3)

Hours of Sleep vs. Amount of Coffee Drank

(4)

Hours of Sleep vs. Amount of Coffee Drank
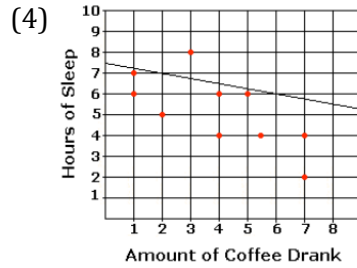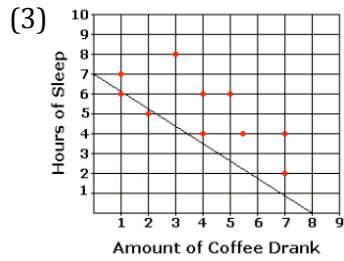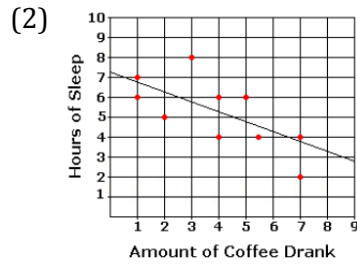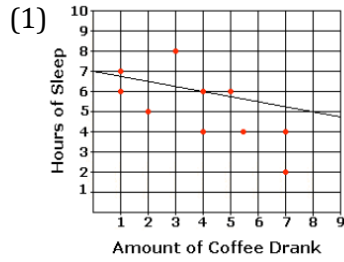
We have now discovered how to sketch the line of best fit. However, this isn't 100% accurate. Now we are going to learn how to get the line of best fit using our calculator.

Let's take a closer look at ***Exercise* #1:**

Exercise #3: A survey was taken of 8 low and high temperatures, in Fahrenheit, in the month of April to try to establish a relationship between a day's low temperature and high temperatures.

Low Temperature, x	22	24	26	28	30	32	34	36
High Temperature, y	30	38	34	42	53	46	58	56

(a) Use your calculation to find the equation for the line of best fit. Round the slope of the line to the nearest hundredth and the y-intercept to the nearest integer.

How to find the Equation of the Line of Best Fit (Linear Regression):

MODE
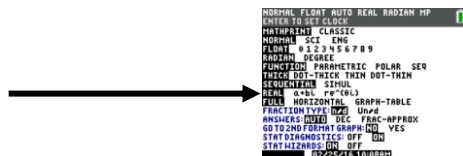
Step 1: STAT ➔ 1: Edit....

Step 2: Insert information into two lists:
$L_1 = x \ values$ **Make sure both lists have the same amount**
$L_2 = y \ values$

Step 3: STAT ➔ CALC ➔ 4: LinReg (ax+b)

Step 4: Make sure your Xlist and Ylist are L_1 and L_2. Then, scroll down to **Calculate** and hit **Enter**

Step 5: Use given information to create the Line of Best Fit.

Equation for the Line of Best Fit:

Exercise #4: Generally, the fuel efficiency of a car changes with the weight of the car. A survey of some cars with their weights and gas mileages is shown below.

Weight (1000's of lbs)	3.7	4.5	3.2	5.1	6.8	4.9	4.8	5.5
Mileage (miles per gallon)	38	26	48	24	18	30	28	21

(a) Find the equation for the line of best fit using your calculator. Round both coefficients to the nearest tenth. List what the variables x and y represent in this problem.

(b) Given the graph of this scatter plot, would you consider the **correlation** between weight and mileage to be **positive** or **negative?** Explain.
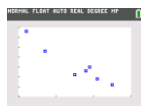
(c) Which **parameter** of the linear model predicts whether the **correlation** is positive or negative? Use this model to help explain your answer.

(d) If a car had a weight of 4,300 pounds, what would this model predict as its fuel efficiency? Round to the nearest integer. Use appropriate units and make sense of your answer.

(e) If we wanted to purchase a car that got 40 miles to a gallon, what weight of car, to the nearest 100 pounds, should we purchase? Solve algebraically.

Name: _____   Date: _____ Period: _____
Algebra   Linear Regression on Calculator   10F HW

1) The table below gives the number of hours spent studying for a science exam and the final exam grade.

   (a) Draw a scatter plot of the data and sketch in the line of best fit.

   (b) What is the equation for the line of best fit? (Use calc)
        Round all values to nearest tenth.

Study Hours	3	2	5	1	0	4	3
Grade	84	77	92	70	60	90	75
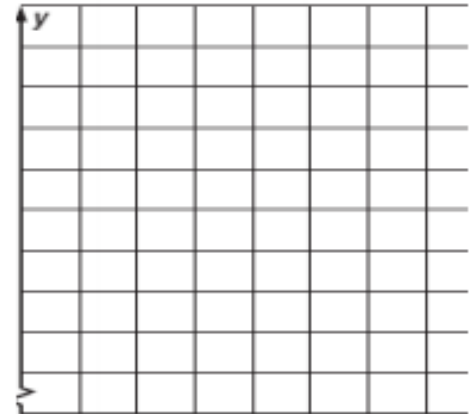
   (c) Could this line go on forever? Why or why not?

2) The table shows the average and maximum longevity of various animals in captivity.

   (a) Draw a scatter plot and determine, what relationship, if any, exists.

   (b) Sketch in the line of best fit.

Longevity (years)							
Avg.	12	25	15	8	35	40	41	20
Max.	47	50	40	20	70	77	61	54

   (c) What is the equation for the line of best fit? (Use calc)
        Round all values to nearest tenth.

Animal Longevity (Years)
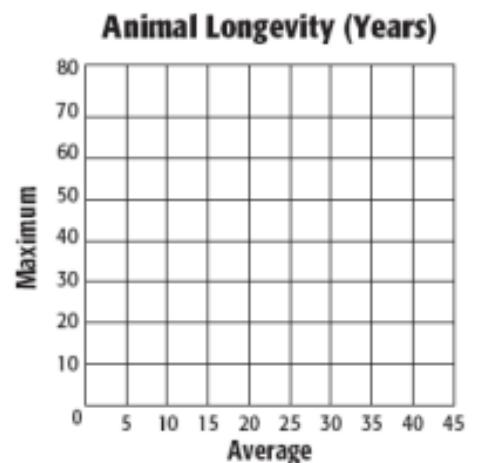
3) A survey was done at Ketcham High School to determine the effect of time spent on studying and grade point average. The table below shows the results for 10 students randomly selected.

Study time (Hours per week)	2	4	5	7	10	12	14	17	19	20
GPA (out of 100)	64	71	69	74	81	86	84	94	91	96

(a) Enter the data in your calculator and use it to generate the equation for the line of best fit. Round your slope to the nearest tenth and round your *y*-intercept to the nearest integer.

(b) According to the linear regression model from part (a), what GPA, to the nearest integer, would result from studying for 15 hours in a given week? Justify your answer.

(c) A passing average is defined as a 65% or above. Does the model predict a passing average if the student spends no time studying in a given week? Justify your answer.

(d) For each additional hour that a student studies per week, how many points does the model predict a GPA will rise? Explain how you arrived at your answer.

4) The mean annual temperature of a location generally depends on its elevation above sea level. A collection of nine locations in Nevada were chosen and had their elevation and mean annual temperature recorded. The data is shown below.

Elevation (feet)	1200	4125	6230	2378	5625	6328	4375	1864	3160
Mean Temperature (°F)	62	45	36	51	48	32	40	58	49

(a) Use your calculator to determine the equation for the line of best fit. Round your slope to the nearest *thousandth*. Note that it will be a small number. Round your *y*-intercept to the nearest integer.

(b) What does the y-intercept tell you about the temperature in Nevada?

(c) Using correct units, give an interpretation of the slope of this line.

(d) Using your model from part (a), what would be the predicted mean temperature at an elevation of 3000 feet above sea level?

(e) Would you characterize this correlation as being positive or negative? How can you tell this from the equation itself?

Review Section:

5) In the function $f(x) = (x + 1)^2 + 4$, identify the vertex.

6) When solving the equation $x^2 - 8x - 7 = 0$ by completing the square, which equation is a step in the process?

(1) $(x - 4)^2 = 9$ (3) $(x - 4)^2 = 23$

(2) $(x - 8)^2 = 9$ (4) $(x - 8)^2 = 23$

If you don't have work you will not receive full credit on your homework

Name: _____*Homework Answers*_____   Date: _____ Period: _____
Algebra   Linear Regression on Calculator   10F HW

1.) (a) Sketch a line of best fit   (b) $y = 6.3x + 62.0$
     (c) No; with appropriate explanation

2.) (a) Positive Correlation   (b) Sketch a line of best fit
     (c) $y = 1.2x + 22.4$

3.) (a) $y = 1.7x + 62$ (b) 88
     (c) No; with appropriate explanation   (d) GPA will rise 1.7 points for each additional hour of studying.

4.) (a) $y = -.005x + 65$ (b) The mean temperature is 65 degrees Fahrenheit.
     (c) For every foot of increase above sea level, the mean temperature decreases by .005 degrees Fahrenheit.
     (d) 50 degrees Fahrenheit   (e) Negative correlation; the slope is negative.

5.) (-1,4)

6.) (3)

In the last two lessons we fit **bivariate data sets** with **lines of best fit**. Sometimes, though, linear models are not the best choice. We can fit data with all sorts of curves, the most common of which are **linear**, **exponential**, and **quadratic**. But, there are many other types. Before we look at exponential and quadratic regression, recall the general shapes of these two types of functions.

EXPONENTIAL AND QUADRATIC GRAPHS

EXPONENTIAL GRAPHS   **QUADRATIC GRAPHS**
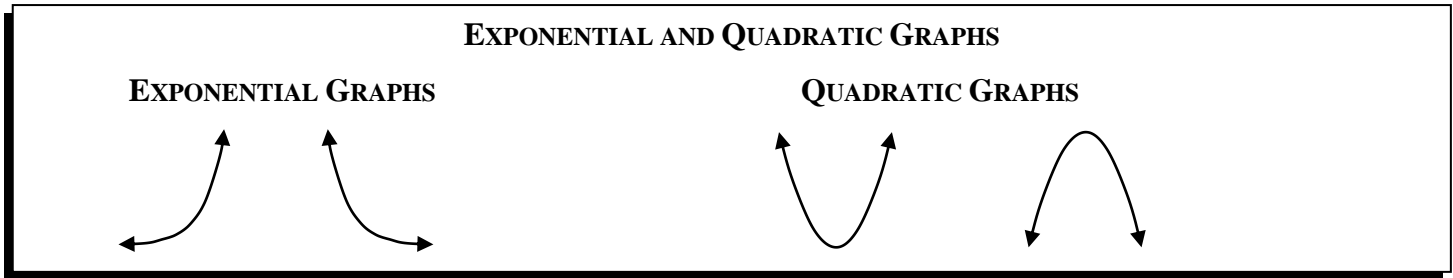
Exercise #1: For each scatterplot shown below, determine if it is best fit with a linear, exponential, or quadratic function. Draw a curve of best fit depending on your choice.

(a)

Type: _____

(b)

Type: _____

(c)

Type: _____

(d)

Type: _____

(e)

Type: _____

(f)
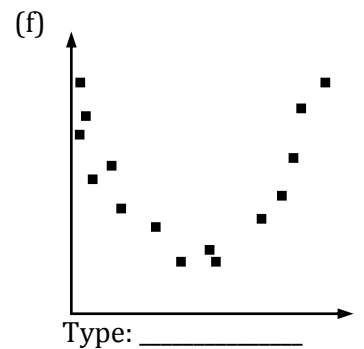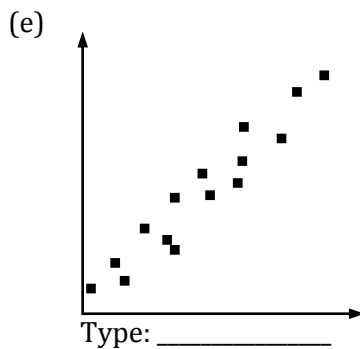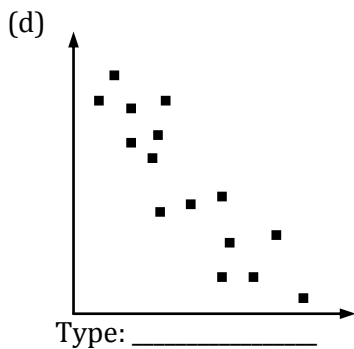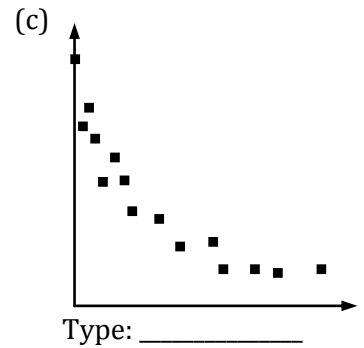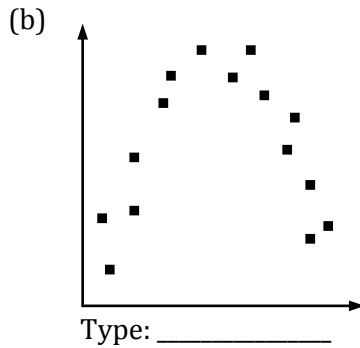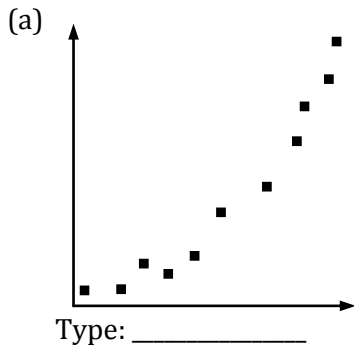
Type: _____

Our calculators can produce equations for **exponentials of best fit** and **quadratics of best fit** (along with a lot of other types of curves).

43

Exercise #2: Biologists are modeling the number of flu cases as it spreads around a particular city. The total number of cases, y, was recorded each day, x, after the total first reached 16. The data for the first week is shown in the table below.

x, days	0	1	3	4	6	7
y, cases	16	18	22	25	33	35

(a) Use your calculator to find the **exponential regression equation** for this data set in the form $y = a(b)^x$ Round all parameters to the nearest *hundredth*.

==**To find the exponential regression equation, insert lists into L_1 and L_2. Click on STAT, CALC. Instead of using 4:LinReg, use 0: ExpReg****==

(b) Based on the regression equation, how many total cases of flu will there be after two weeks?

(c) According to your model, by what percent are the flu cases increasing on a daily basis?

Exercise #3: An application developer released a new app to be downloaded. The table below gives the number of downloads for the first four weeks after the launch of the app.

Number of Weeks	1	2	3	4
Number of Downloads	120	180	270	405

(a) Write an exponential equation that models this data.

(b) Use this model to predict how many downloads the developer would expect in the 26th week if this trend continues. Round your answer to the *nearest download*.

Name: _____ Date: _____ Period: _____
Algebra   Other Types of Regression   10G HW

1. For each scatterplot below, determine the best type of regression from: linear, exponential, or quadratic. Draw
    a representative curve (line, exponential, or parabola) through the data.

(a)   (b)   (c)

    Type: _____   Type:_____   Type: _____

(d)   (e)   (f)

    Type: _____   Type: _____   Type: _____
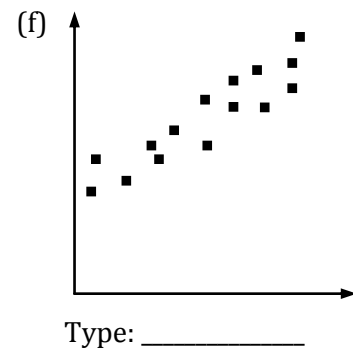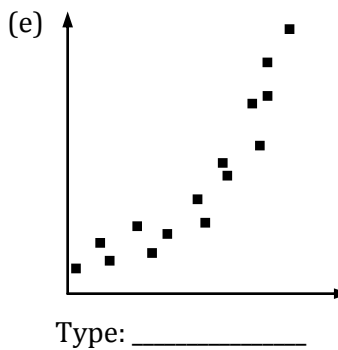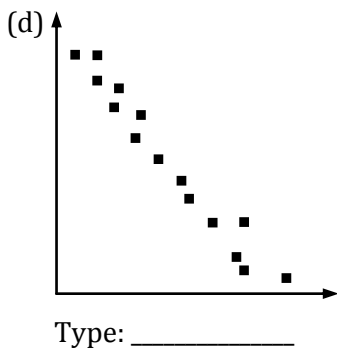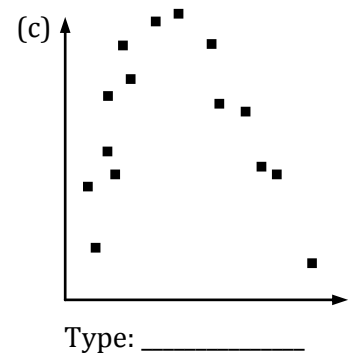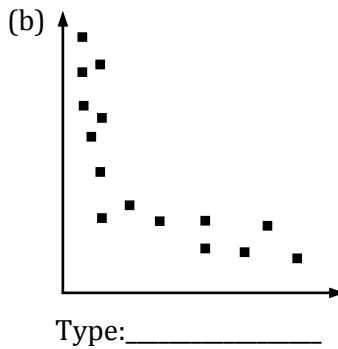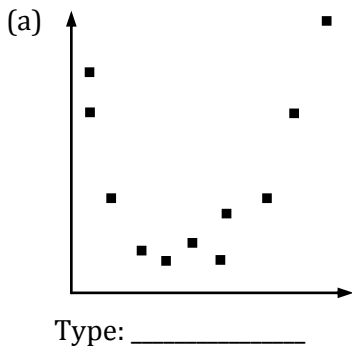
2. Given the scatterplot below, which of the following equations would best model the data? Explain your choice.

(1) $y = -3x + 6$ (3) $y = -4x^2 + 20x + 3$

(2) $y = 6(2)^x$ (4) $y = 2x^2 - 6x + 4$
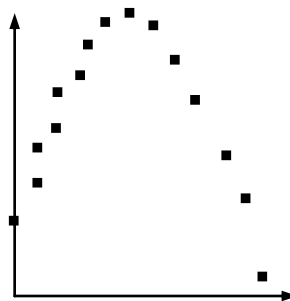
3. The cost per widget produced by a factory generally drops as more are produced but then starts to rise again
due to overtime costs and wear on the equipment. Quality control engineers recorded data on the cost per widget
compared to the number of widgets produced. Their data is shown below.

Number of widgets, x	35	88	110	135	154	190
Cost per widget, y	9.32	2.63	1.42	1.32	2.12	5.50

Why should a quadratic model be considered for this data set as opposed to linear or exponential?

4. A marketing company is keeping track of the number of hits that a website receives on a daily basis. Their data for the first two weeks is shown below. A scatterplot of the data is also shown.

Days	Hits
0	120
3	145
5	162
10	220
14	270

Daily Hit Count for Site

Days After the Website Launched
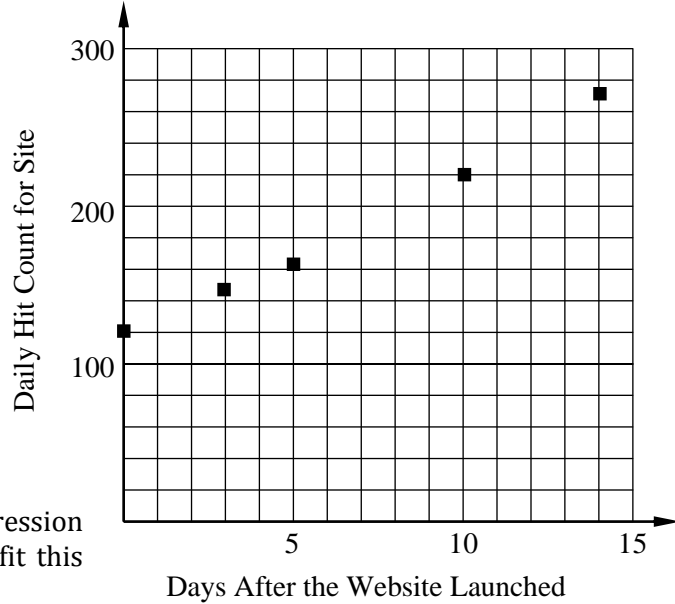
(a) Of the two (Linear/Exponential) types of regression we have studied which seems most likely to fit this data? Explain your choice.

(b) Find a linear equation, in the form $y = ax + b$, that best models this data and an exponential equation, in the form $y = a(b)^x$ that best models this data. Round all parameters to the nearest *hundredth*.

Linear Model   **Exponential Model**

(c) How close are the two model's outputs when $x = 10$? Show the values you find.

(d) How close are the two model's outputs when $x = 30$? Show the values that you find.

(e) Which model will predict faster growth of website hits over time? Explain your answer. You may want to experiment by graphing both models.

5. The cost of belonging to a gym can be modeled by $C(m) = 50m + 79.50$, where $C(m)$ is the total cost for m months of membership.

State the meaning of the slope and y-intercept of this function with respect to costs associated with the gym membership.

6. When $(2x - 3)^2$ is subtracted from $5x^2$, the result is _____?

Name: _____   Date: _____ Period: _____

Homework Answers

Algebra   Other Types of Regression   10G HW

1.) (a) Quadratic   (b) Exponential   (c) Quadratic
      (d) Linear   (e) Exponential   (f) Linear

2.) (3)

3.) The outputs decrease and the increase. Linear/Exponential only increase or only decrease.

4.) (a) Linear; with appropriate explanation
      (b) Linear: $y = 10.83x + 114.07$ Exponential: $y = 121.09(1.06)^x$
      (c) $y = 22.37 \; and \; y = 216.85$
      (d) $y = 438.97 \; and \; y = 695.48$
      (e) Exponential; grows faster over time

5.) Proper explanations

6.) $x^2 + 12x - 9$

Name: _____   Date: _____ Period: _____
Algebra   Quantifying Predictability   10H

In the last few lessons we have worked with generating **lines and curves of best** fit for **bivariate** data sets. In every circumstance, though, the data did not fall on a straight line or on a perfect curve. We have never answered the question of how well specifically a **linear model** does in predicting the **correlation** between the two variables.

Correlation (r) is a relationship between two or more things which change (variables) that can be described mathematically. Correlation refers to how closely two sets of information or data are related. The linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The value of r is such that $-1 < r < +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

   - ***Positive correlation***: If x and y have a strong positive linear correlation, r is close to +1.
          An r value of exactly +1 indicates a perfect positive fit.   Positive values indicate a relationship
          between x and y variables such that as values for x increases, values for y also increase.
   - ***Negative correlation:***   If x and y have a strong negative linear correlation, r is close to -1.
          An r value of exactly -1 indicates a perfect negative fit.   Negative values indicate a
          relationship between x and y such that as values for x increase, values for y decrease.
   - ***No correlation:*** If there is no linear correlation or a weak linear correlation, r is close to 0. A
          value near 0 means that there is a random, nonlinear relationship between the two variables
   - ***Perfect correlation:*** of ± 1 occurs only when the data points all lie exactly on a straight line. If
          r = +1, the slope of this line is positive. If r = -1, the slope of this line is negative.
   - ***Strong & Weak correlations***: A correlation greater than 0.8 is generally described as strong,
          whereas a correlation less than 0.5 is generally described as weak. These values can vary
          based upon the "type" of data being examined. A study utilizing scientific data may require a
          stronger correlation than a study using social science data.
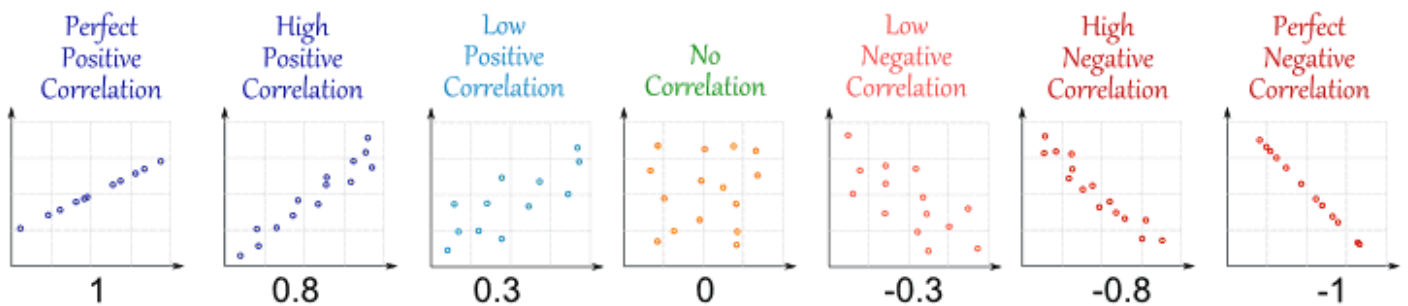
Your teacher will explain how to ensure that your calculator has its "r-value" on. Since this varies by graphing calculator, write down the procedure below if necessary.

MODE →

Exercise #1: In the following exercises four data sets with equal *x*-values are given to illustrate different types of **positive correlations**. For each, enter the data, observe the scatter plot, and record the *r*-value, known as the **correlation coefficient**, for a **linear fit** to the nearest *thousandth*.

(a)

x	2	5	8	11	15	18
y	4	13	22	29	43	52

(b)

x	2	5	8	11	15	18
y	16	14	22	41	37	51

NORMAL FLOAT AUTO REAL RADIAN MP

$r = .999$
Very Good fit (Close to +1)

(c)

x	2	5	8	11	15	18
y	18	8	41	28	62	44

(d)

x	2	5	8	11	15	18
y	44	51	30	55	45	47

Exercise #2: The following data set is that of two variables that have a **negative correlation**. Enter the data, produce the scatter plot, and record the *r*-value. How is the negative correlation reflected in the *r*-value?

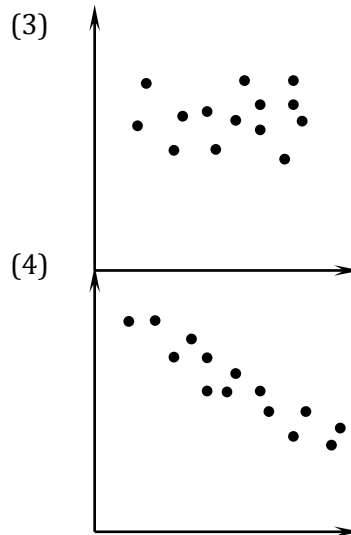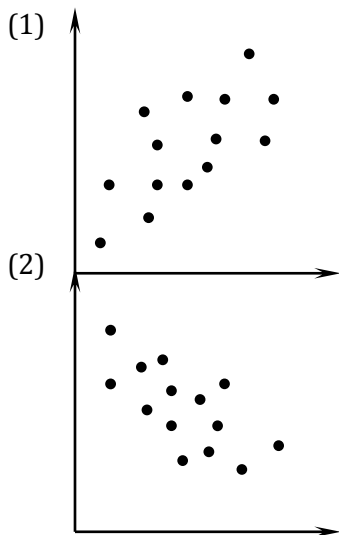| x	2	5	8	11	15	18
y	52	47	28	32	25	10

Exercise #3: Given the scatter plot shown below, which of the r-values would most likely represent the correlation between the two variables? Explain your choice.

(1) $r = 0.88$ (3) $r = 1$

(2) $r = 0.28$ (4) $r = -0.94$
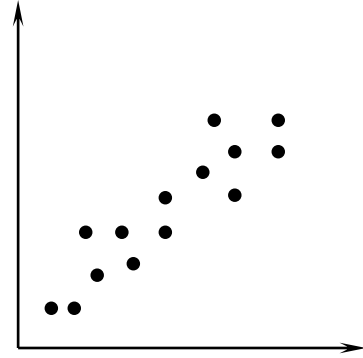
***Exercise #4:*** Which of the following scatter plots would have a correlation coefficient closest to -1?

(1)

(3)

(2)

(4)

Exercise #5: There are two primary types of crude oil sold in the world, West Texas Intermediate (WTI) and Brent Crude. Each is priced differently on a daily basis and each has a correlation with the average price per gallon for unleaded gasoline. The two linear regression models, along with their r-values, are shown below. Give a prediction for the price per gallon of unleaded gasoline, y, on a day when the price for WTI is $103 and the price for Brent is $109, x. Which model did you choose and why?

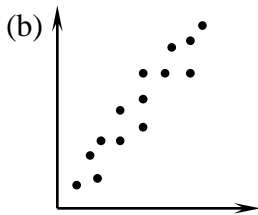Brent Crude: $y = 0.028x + 0.71$, $r = 0.973$ WTI Crude: $y = 0.031x + 0.67$, $r = 0.924$

1. Below there are six scatter plots, six correlation coefficients, and six terms. Match the appropriate *r*-value with the scatter plot it most likely corresponds to. Then match the term you think is most appropriate to the *r*-value as well (not to the graph).

(a) is done for you already

(a)
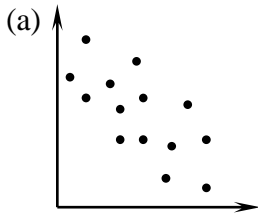
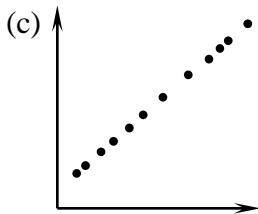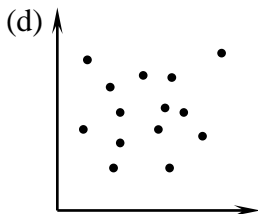$r = 1.0$
(c)

Weak Negative
(a)

(b)

$r = 0.35$
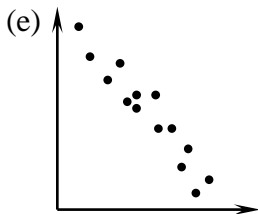
Perfect Positive

(c)

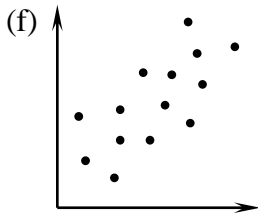$r = -0.82$

Strong Positive

(d)

$r = 0$

Weak Positive

(e)

$r = -0.56$

Moderate Negative

(f)

$r = 0.93$

No Correlation

2. A solar power company is trying to correlate the total possible hours of daylight (simply the time from sunrise to sunset) on a given day to the production from solar panels on a residential unit. They created a scatter plot for one such unit over the span of five months. The scatter plot is shown below.

The equation line of best fit for this bivariate data set was:

$$y = 2.26x + 20.01$$
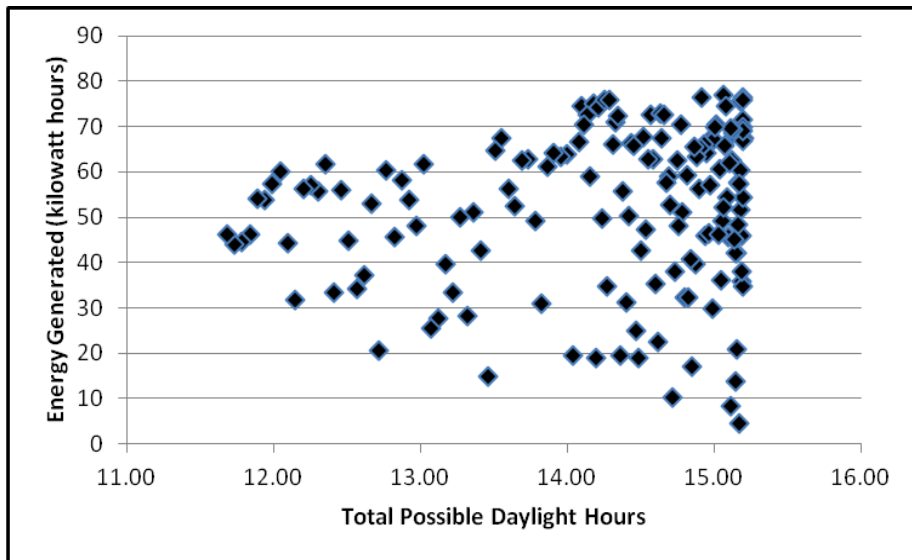
(a) How many kilowatt hours would the model predict on a day that has 14 hours of possible daylight?

(b) To the nearest tenth of an hour, how many hours of possible daylight would be needed to produce 50 kilowatt hours of energy?

(c) The correlation coefficient for this regression was $r = 0.134$. Would you characterize this as strongly positive, moderately positive, or a weakly positive correlation? Explain.

(d) Based on (c), do you have confidence in the model to accurately predict the energy production based on the total possible daylight hours? Explain.

(e) What environmental factors might contribute to the "noise" in the data? Noise are factors that prevent the correlation from being perfect.

53

Review Section:

3. Evaluate the following as an equivalent trinomial.
$$(x + 3)^2$$

4. Find the roots of the following:
$$8x^2 - 18x + 9 = 0$$

Name: _____ **Homework Answers**   Date: _____ Period: _____
Algebra   Quantifying Predictability   10H HW

1.) (b) f and c   (c) e and b   (d) d and f   (e) a and e   (f) b and d

2.) (a) 51.65 kwh   (b) 13.3 hours
   (c) Weak positive correlation; with appropriate explanation
   (d) No; with appropriate explanation
   (e) Clouds, shade, snow, etc…

3.) $x^2 + 6x + 9$

4.) $x = \left\{\frac{3}{4}, \frac{3}{2}\right\}$

Do Now: **Exercise #1:** A skydiver jumps from an airplane and an attached micro-computer records the time and speed of the diver for the first 12 seconds of the diver's freefall. The data is shown in the table below.

| Time (sec)	0	2	4	6	8	10	12
Speed (ft/sec)	0	25	46	60	68	72	74

(a) Find the equation for the *line of best fit* for this data set. Round both coefficients to the nearest *tenth*. As well, determine the correlation coefficient and round it to the nearest *hundredth*. Based on the correlation coefficient, characterize the fit as positive or negative and how strong of a fit it is.

In the last lesson, we saw how the **correlation coefficient** (or *r*-value) measures the **predictability** of the model (or how well it will do in its predictions). Although the *r*-value is an excellent measure, it does not tell us whether the model is **appropriate** only whether it does a good job at **predicting**. Today we will examine what are known as **residuals and residual graphs** to determine if a linear model makes sense.

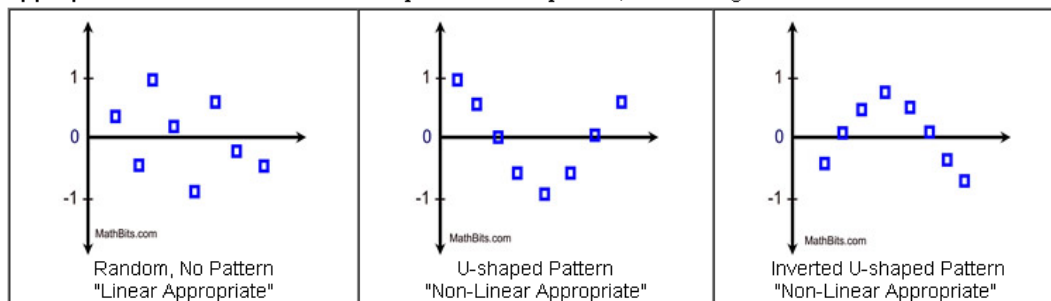- Residuals help to determine if a curve (shape) is appropriate for the data. (linear vs. non-linear)

Definition:	Residual = Observed y-value - Predicted y-value

A residual is the difference between the observed *y*-value (from scatter plot) and the predicted *y*-value (from regression equation line).
It is the vertical **distance** from the actual plotted point to the point on the regression line.
You can think of a residual as how far the data "fall" from the regression line (sometimes referred to as "*observed error*").
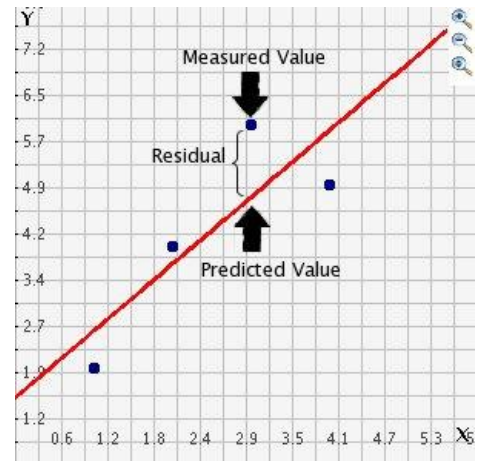
Appropriate linear model: when plots are randomly placed, above and below *x*-axis ($y = 0$).
**Appropriate non-linear model:** when plots follow a pattern, resembling a curve.

Random, No Pattern "Linear Appropriate"	U-shaped Pattern "Non-Linear Appropriate"	Inverted U-shaped Pattern "Non-Linear Appropriate"

When a **pattern** is observed in a residual plot, a **linear regression model** is probably **not appropriate** for your data.

56
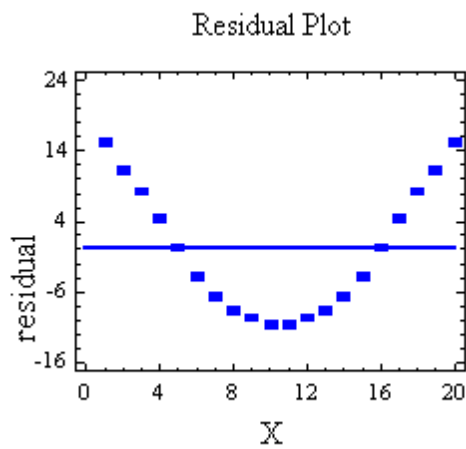
Exercise #2: Tell what the residual plot indicated about the appropriateness of the linear model that was fit to the data.
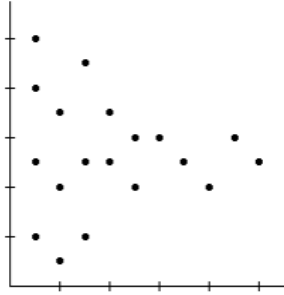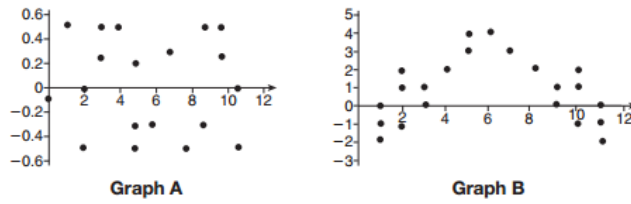

(1) Model is not appropriate. The relationship is nonlinear.
(2) Model is appropriate.
(3) Model may not be appropriate. The spread is changing.

**Exercise #3:** Would you consider this model to be a "good fit"? Explain your answer.

Residual Plot

1) The residual plots from two different sets of bivariate date are graphed below.

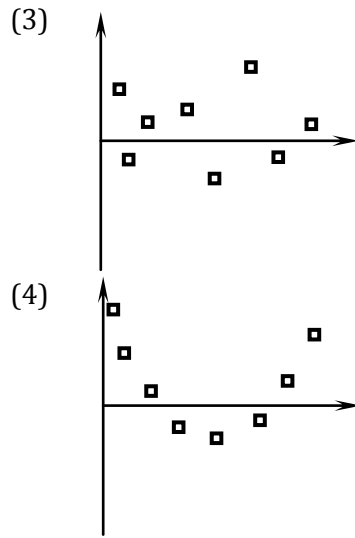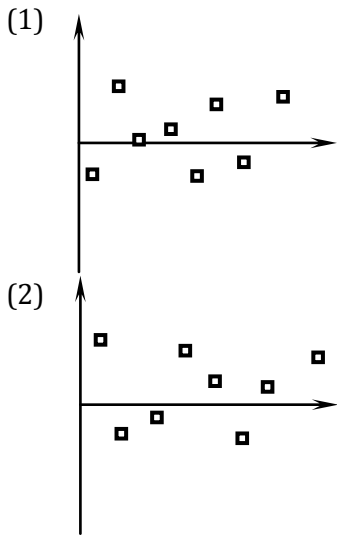Graph A   Graph B

Explain using evidence from graph A and graph B, which graph indicates that the model for the data would be a good fit.

2) Which of the following residual plots indicates a model that is most appropriate?

(1)

(3)

(2)

(4)
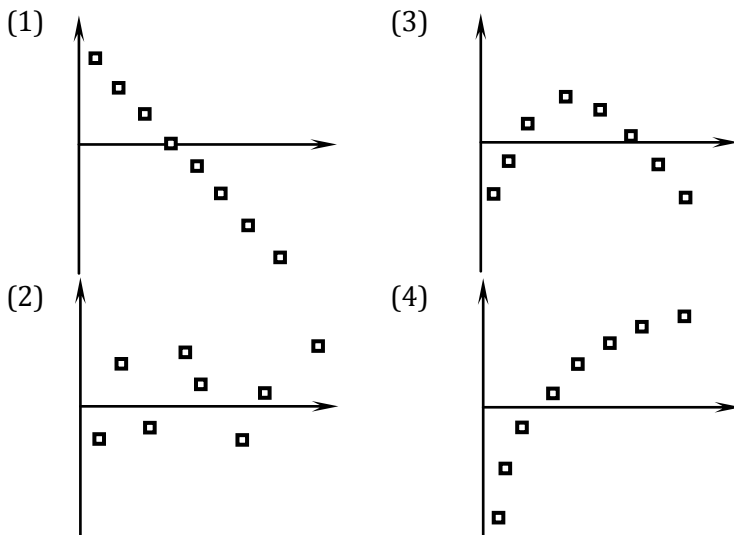
3) Which of the following residual plots would indicate the linear model used to produce it was an inappropriate choice?

(1)

(3)

(2)

(4)

4) Physics students are performing a lab where they allow a ball to roll down a ramp and record the distance that it has rolled versus the time it has been rolling. The data for one such experiment are shown below.

Time (sec)	0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Distance (ft)	0	0.4	1.5	3.2	5.6	8.5	12.6	17.2	22.8

(a) Determine the equation for the line of best fit. Round your coefficients to the nearest *tenth*. Also, determine the correlation coefficient. Round it to the nearest *hundredth*.
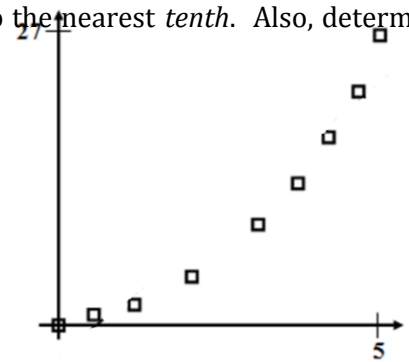
(b) To the right is a scatter plot of the data. Sketch a line of best fit.
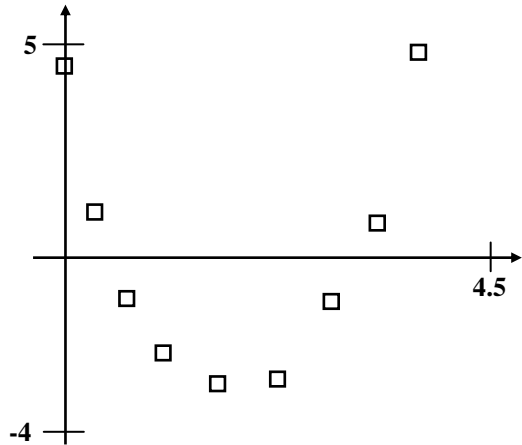
(c) To the right is a graph of the residuals. Is the linear model appropriate for this data set given the residual plot? Explain below.

5) Beverly did a study this past spring using data she collected from a cafeteria. She recorded the data weekly for ice creams sales and soda sales. Beverly found the line of best fit and the correlation coefficient, as shown in the diagram below.

Beverly's Cafeteria Study
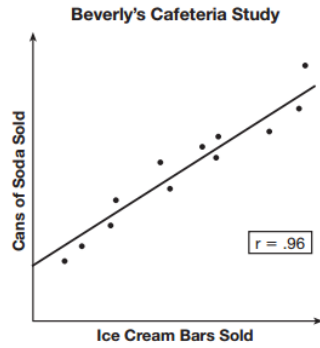
Given the information, which statement(s) can correctly be concluded?

I: Eating more ice cream causes a person to become thirsty.

II: Drinking more soda causes a person to become hungry.

III: There is a strong correlation between ice cream sales and soda sales.

   (1) I, only   (2) III, only   (3) I and III   (4) II and III

6) Solve the following by completing the square:

$$x^2 + 6x - 7 = 0$$

Name: _____ ***Homework Answers***   Date: _____ Period: _____
Algebra   Residuals   10I HW

1.) Graph A; with appropriate explanation

2.) (2)

3.) (4)

4.) (a) $y = 5.6x - 3.3$ $r = .96$ (b) Sketch a line of best fit
    (c) No; with appropriate explanation

5.) (2)

6.) $x = \{-7,1\}$