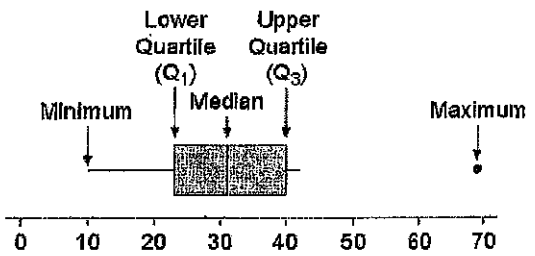
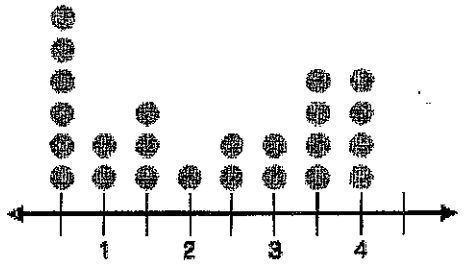
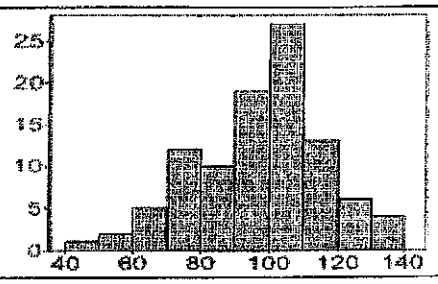
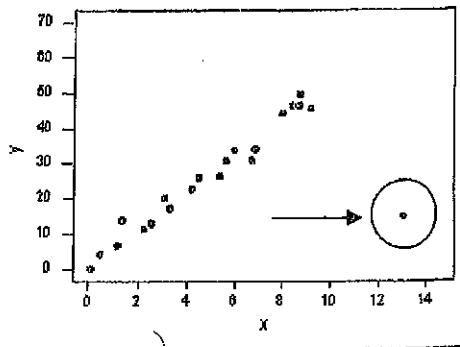
Accelerated CCGPS Algebra/Geometry
Unit 6: Describing Data

Date	Topic/Classwork	Assignment
Feb. 25	Dot Plots, Histograms, Box Plots Notes: pages 1-2	Pages 3-6
Feb. 26	Measures of Central Tendency -mean -median -mode -Interquartile Range (IQR) -Mean Absolute Deviation (MAD)	Pages 7-10b
March 1	Interpret shapes, centers and spread of data Pages 11-15	Pages 16-17
March 2	Two-Way Frequency Tables Page 18	Pages 19-21
March 3	Help Session 10:30 -11:30	Make sure homework is all caught up!
March 4	QUIZ	Make sure homework is all caught up!
March 5	Correlation VS Causation Correlation Coefficient Pages 22-25	Pages finish up 22-25
March 8	Line of Best Fit Pages 26-27	Pages 28-29
March 9	Line of Best Fit (exponential and linear) Pages 30-31	Page 32
March 10	Help Session 10:30-11:30	
March 11	Residuals Page 33A	Pages 33-35
March 12	Review	
March 15	TEST	

Day 1 and 2 Introduction to Statistics - Vocabulary

Term	Describe	Example
Box Plot		
Dot Plot		
Histogram		
Median		median of all data, second quartile 65, 65, 70, 75, 80, 80, 85, 90, 95, 100 median of lower part, first quartile median of upper part, third quartile
First and Third Quartiles		median of all data, second quartile 65, 65, 70, 75, 80, 80, 85, 90, 95, 100 median of lower part, first quartile median of upper part, third quartile

Interquartile Range		**Subtract** **Third Quartile (Q_3) – First Quartile (Q_1) = IQR**
Outlier		
Mean		**5 + 4 + 2 + 6 + 3 = 20** $\frac{20}{5} = 4$ **The Mean is 4.**
Mean Absolute Deviation (MAD)		Steps: 1. Find the Mean 2. Calculate the absolute value of the difference between each data value and the mean 3. Determine the average of the differences in step 2. This average is the mean absolute deviation
Measures of Center		Find the Mean and Median for the following data. ***Hint:*** (Must order the numbers first before finding the Median) 2 1 5 4 3 **Mean**: $\frac{15}{5} = 3$ **Median** = 3
Measures of spread		***Examples of Measures of Spread:*** 1. Range 2. Interquartile Range (IQR) 3. Mean Absolute Deviation -MAD

Name: _____ Date: _____

Graphical Displays for Data

Example 1: A pharmacy records the number of customers each hour that the pharmacy is open. The staff is using the information to determine how many people need to be working at the pharmacy at each time of the day. The number of customers is in the table below. Use the table to create a histogram to help the pharmacy staff understand how many customers are in the pharmacy at each time of day.

Time Frame	Number of customers
8:00 A.M. – 9:00 A.M.	2
9:00 A.M. – 10:00 A.M.	0
10:00 A.M. – 11:00 A.M.	8
11:00 A.M. – 12:00 P.M.	14
12:00 P.M. – 1:00 P.M.	23
1:00 P.M. – 2:00 P.M.	12
2:00 P.M. – 3:00 P.M.	7
3:00 P.M. – 4:00 P.M.	3
4:00 P.M. – 5:00 P.M.	5

Example 2: Anna and Ethan watch 10 thirty-minute shows during the month of June. They record the number of food commercials that air during each show in the table below. Create a dot plot to display the number of food commercials that aired during the 10 shows.

Shows	# of Commercials
A	7
B	7
C	5
D	7
E	4
F	7
G	5
H	9
I	5
J	6

Example 3: Ray's scores on his mathematics tests were 70, 85, 78, 90, 84, 82, and 83. Draw a box plot to represent Ray's Data.

Find the IQR.

Are there any outliers?

Adapted from: Walch Education Resources: CCGPS Coordinate Algebra Teacher Resource Binder

Example 4: A company keeps track of the age at which employees retire. It is considered an early retirement if the employee retires before turning 65. The age of the 11 employees who took early retirement this year are listed in the table below. Draw a box plot for the data. Are there any striking deviations in the data?

Employee	Age at early retirement
1	56
2	55
3	60
4	51
5	53
6	58
7	56
8	64
9	59
10	42
11	48

Example 5: Elizabeth records her scores each time she goes bowling. The scores from her last 13 games are in the table below.

Game	Score
1	206
2	210
3	198
4	209
5	194
6	200
7	216
8	212
9	196
10	224
11	228
12	231
13	207

Construct a box plot of her data.

Find the IQR.

Are there any outliers?

Adapted from: Walch Education Resources: CCGPS Coordinate Algebra Teacher Resource Binder

Name: _____ Date: _____

Graphical Displays for Data Homework

Kirsten plays softball in the spring. Each game, she records the number of times she reaches first base without being called out. Use the data in the table to solve problems 1 -5.

Game	Number of times at first	Game	Number of times at first
1	5	10	0
2	1	11	1
3	2	12	1
4	0	13	0
5	2	14	5
6	2	15	5
7	4	16	4
8	4	17	0
9	0	18	4

1. Create a dot plot showing the number of times Kirsten reached first base in each game.

2. Find the minimum, maximum, first quartile, and third quartile of the data set.

   a. Minimum:

   b. Maximum:

   c. First Quartile:

   d. Third Quartile:

3. Create a box plot showing the number of times Kirsten reached first base.

4. Find the interquartile range of the data. Are there any outliers?

5. Kirsten wants to analyze her performance using this data. She wants to understand the range of her data and the frequency of different results. Which graph, the dot plot or the box plot, will be most useful to Kirsten? Explain.

Adapted from: Walch Education Resources: CCGPS Coordinate Algebra Teacher Resource Binder

Dr. Singh is a veterinarian. He records the weights of each pet. The weights of 10 German shepherds, all 4-year-old males, are in the table below, rounded to the nearest pound. Use this information to solve problems 6-10.

Weight in pounds
80
78
82
84
81
89
83
81
81
82

6. Create a histogram showing the weights of Dr. Singh's German shepherds.

7. Find the minimum, maximum, first quartile, and third quartile of the data set.

a. Minimum:

b. Maximum:

c. First Quartile:

d. Third Quartile:

8. Create a box plot showing the weights of the German shepherds.

9. Find the interquartile range of the data. Are there any outliers?

10. Dr. Singh wants to analyze the weights of the German shepherds. He wants to understand the center and spread of his data, so that he has a better idea of an expected weight for a 4-year-old male German shepherd. Which graph would be most useful to Dr. Singh? Explain.

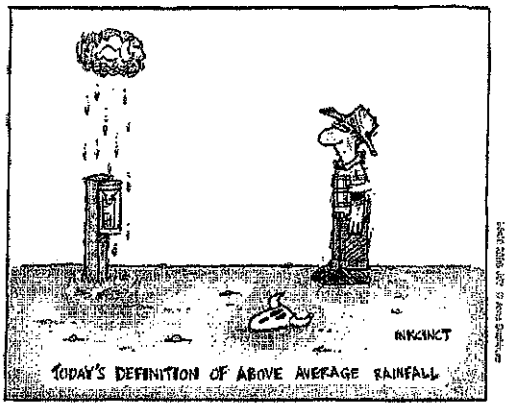
Name: _____ Date: _____

Measures of Central Tendency

1. Some people use "average" interchangeably for both mean and median. Consider this statement:

   "Just think of how stupid the average person is, and then realize half of them are even stupider?" George Carlin

   What type of "average" is George Carlin referring to, mean or median? Is it possible to have more than half of a population above this kind of average?



2. What is the difference between mean and median?

3. Give an example of data when the mean and median might have the same value.

   Give an example when the mean and the median do NOT have the same value.

4. Can the following statement be true? Why or why not?
   "Welcome to Lake Wobegon, where all the women are strong, all the men are good-looking, and all the children are above average." Garrison Keillor

5. Is it possible to have more than half of data values above (or below) the mean?

6. Find the mean, median, and mode for this set of data.
   5, 11, 16, 8, 4, 7, 15, 6, 11, 7

7. Kara had 85, 83, 92, 88, and 69 on her first five math tests. She knows that she needs an average of 85 to get a B. What score must she get on her last test to get a B?

Adapted from: Mathematics Vision Project

Measures of Spread – Range, IQR, and Mean Absolute Deviation

Man Absolute Deviation of a numerical data set is the average positive deviations of the data from the mean.

$$Mean\ Absolute\ Deviation = \frac{\left|x_1 - \bar{x}\right| + \left|x_2 - \bar{x}\right| + \ldots + \left|x_n - \bar{x}\right|}{n}$$

A measure of distribution is a measure of how spread out data is, or how the data is distributed from its smallest values to its largest values. Suppose, for instance, that Joe has test scores of 60, 68, 69, 78, 90, 95, and 100. Sammy scores 78, 78, 79, 79, 82, 82, and 82.

8. Calculate Joe's mean test score. Then calculate Sam's mean test score. What do you notice about Joe's scores compared to Sammy's?

Measuring the mean will not tell you much about the characteristics of the test takers. A measure of distribution, or spread, will help you see that Sam consistently scores near 80, while Joe's scores are spread out, or distributed, over a much larger range.

9. To examine the distribution of test scores, find the mean absolute deviation. Follow the steps below to find the mean absolute deviation of Sam's test scores (Joe's example is given).

Steps	Joe	Sam
a) Calculate the mean, symbolically, $\bar{x}$, of the data.	Mean: $\bar{x} = \frac{60+68+69+78+90+95+100}{7}$, $\bar{x} = 80$	
b) Find the deviation, or distance from the mean, for each piece of data.	Deviation: $60-80=-20$ $68-80=-12$ $69-80=-11$ $78-80=-2$ $90-80=10$ $95-80=15$ $100-80=20$	
c) Find the absolute value of the mean deviations.	20 12 11 2 10 15 20	
d) Find the average of the positive deviations found in part c.	MAD: $\frac{20+12+11+2+10+15+20}{7} = 12.86$	

10. Why do you think the MAD of Joe's test scores is higher than the MAD of Sammy's test scores?

Name: _____ Date: _____

Central Tendency and Spread Homework

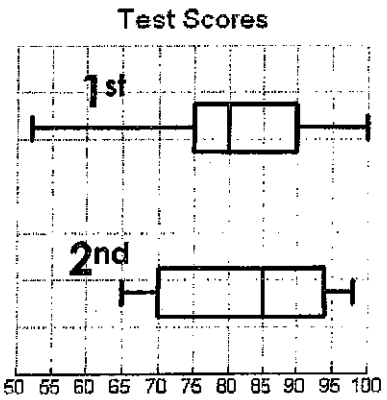
1. The table shows the scores from the top 10 players of our Homecoming basketball game.

Which player scored more than the upper quartile of the data?

A. Matt
B. Michael
C. Jim
D. Bobby

Player	Points	Player	Points
Michael	12	Dave	9
Brendan	6	Heath	15
Andrew	21	Jack	3
Jim	14	Bobby	10
Andre	5	Matt	18

For #2-3, use the graph to the right.



2. Fill in the blanks:
   - The median for 1st period is _____
   - The median for 2nd period is _____
   - The lowest score for 1st period is _____
   - The lower quartile for 2nd period is _____
   - The spread of the middle 50% for 2nd period is _____

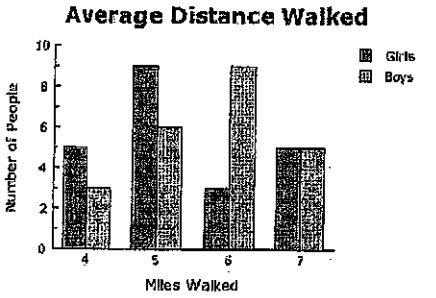
3. Which statement below is NOT true?
   A. 2nd period had the highest score on the test
   B. The median for 2nd period is 5 less than the median for 3rd
   C. The LQ for 2nd period is 5 less than LQ for 3rd period
   D. The UQ for 2nd period is 94

Sample A: 2, 4, 4, 4, 8, 8, 10, 12, 12, 14 **Sample B: 0, 1, 4, 7, 9, 9, 10, 12, 12, 15**

4. Which statement accurately compares the two samples?
   A. The mean for Sample A is 1 greater than the mean of Sample B.
   B. The mean for Sample B is 1 greater than the mean of Sample A.
   C. The mean for Sample A is 0.1 greater than the mean of Sample B.
   D. The mean for Sample B is 0.1 greater than the mean of Sample A.

5. Your scores on the first 4 tests in Algebra were 85, 80, 90, and 93. What do you need to make on the 5th test to have a 90 average in the class?

6. Which measure of central tendency is MOST EASILY affected by outliers?
7. Forty-five people were asked about how many miles they walked in one week. The results are shown in the graph. How does the median number of miles walked for boys compare with the median number of miles walked for girls?



8. The table below shows the running times for science-fiction movies. Find the Mean Absolute Deviation of the data.

Running Times for Movies (min)					
98	87	93	88	126	108

9. The summary statistics for all of the workers at a steel factory are shown. Three sample groups were taken from each of the three shifts. For which sample group is the mean deviation greater than that of the population?

Steel Factory Workers Ages

Mean Deviation: 11.23

Shift 1	Shift 2	Shift 3
23	19	21
19	22	23
50	24	25
49	40	40
67	45	35
34	29	19
30	33	70
59	29	40
40	39	22
33	59	23

Adapted from: Mathematics Vision Project

50 | Mean Absolute Deviation

The mean absolute deviation or M.A.D. measures the spread of a set of data just like the interquartile range (IQR). Unlike IQR, however, M.A.D. uses every data point. Because they both use every data point, the mean (center) and M.A.D. (spread) tend to be used together to describe a set of data. These two measures of center and spread are appropriate for symmetric distributions.

Finding the mean absolute deviation

1. Find the mean of the data
2. Find the difference of every data point from the mean (called deviation)
3. Make every difference positive (absolute value)
4. Find the mean of the absolute differences

Find the M.A.D. for the following data sets. Round the answers to the nearest tenth.

1. 68, 70, 72, 73, 74, 75
2. 72, 75, 73, 99, 68, 79, 48, 60, 52, 59
3. 250, 300, 200, 400, 650, 225, 760, 1215
4. 22, 31, 57, 29, 62, 24
5. 1, 1, 3, 3, 6, 6, 5, 5, 10, 12

Name: _____ Date: _____

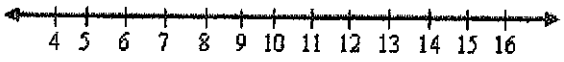
Performance Task: The Basketball Star
Is Bob or Alan a Basketball Star?

MCC9-12.S.ID.1 Represent data with plots on the real number line (dot plots, histograms, and box plots). Choose appropriate graphs to be consistent with numerical data: dot plots, histograms, and box plots.
MCC9-12.S.ID.2 Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation-Advanced Algebra) of two or more different data sets. Include review of Mean Absolute Deviation as a measure of variation.
MCC9-12.S.ID.3 Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers). Students will examine graphical representations to determine if data are symmetric, skewed left, or skewed right and how the shape of the data affects descriptive statistics.

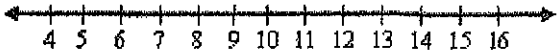
Bob's Points per Game

8, 15, 10, 10, 10, 15, 7, 8, 10, 9, 12, 11, 11, 13, 7, 8, 9, 9, 8, 10, 11, 14, 11, 10, 9, 12, 14, 14, 12, 13, 5, 13, 9, 11, 12, 13, 10, 8, 7, 8

1. Bob believes he is a basketball star and so does his friend Alan. Create a dot plot and box plot of Bob's points for the last 40 games.



Bob's Points



Bob's Points

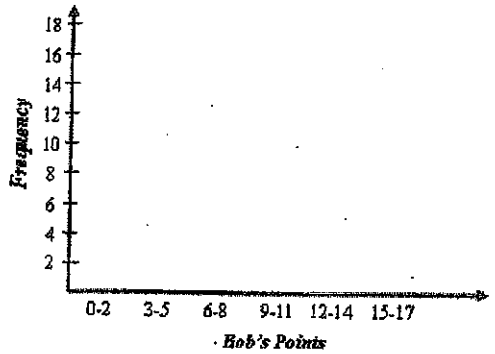
2. Describe Bob's data in terms of center, spread, and shape.

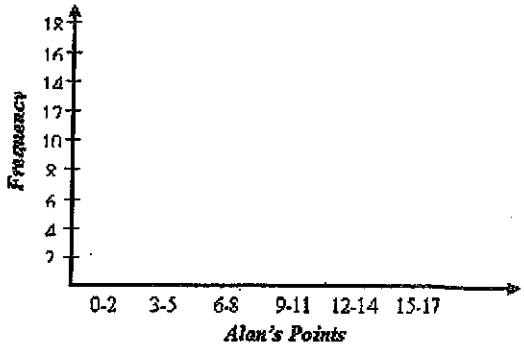
Bob's friend Alan has the following points:

Alan's Points per Game

1, 3, 0, 2, 4, 5, 7, 7, 8, 10, 4, 4, 3, 2, 5, 6, 6, 6, 8, 8, 10, 11, 11, 10, 12, 12, 5, 6, 8, 9, 10, 15, 10, 12, 11, 11, 6, 7, 7, 8

3. Create a histogram of both Bob's and Alan's data.





4. Describe the shape of the two histograms from problem #3.

5. Use summary statistics to compare Bob and Alan's points per game.

	Min	Quartile 1 (Q1)	Median (Q2)	Quartile 3 (Q3)	Max	Mean	Range	IQR	MAD
Bob									
Alan									

6. Which graphical representation best displayed Bob's and Alan's data?

7. Based on the summary statistics is either friend a basketball star? Justify your answer.

Name: _____ Date: _____

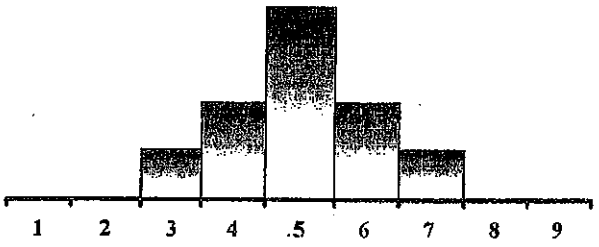
How to Compare Distributions

When you compare two or more data sets, focus on four features:

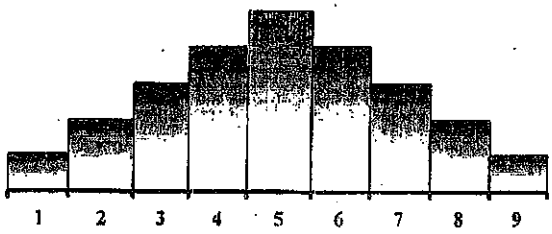
- **Center.** Graphically, the center of a distribution is the point where about half of the observations are on either side.
- **Spread.** The spread of a distribution refers to the variability of the data. If the observations cover a wide range, the spread is larger. If the observations are clustered around a single value, the spread is smaller.
- **Shape.** The shape of a distribution is described by symmetry, skewness, number of peaks, etc.
- **Unusual features.** Unusual features refer to gaps (areas of the distribution where there are no observations) and outliers.

SPREAD

The spread of a distribution refers to the variability of the data. If the data cluster around a single central value, the spread is smaller. The further the observations fall from the center, the greater the spread or variability of the set.



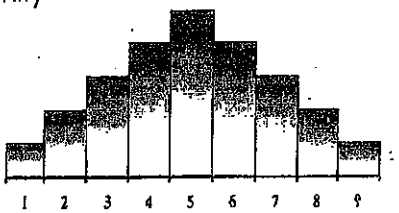
Less Spread



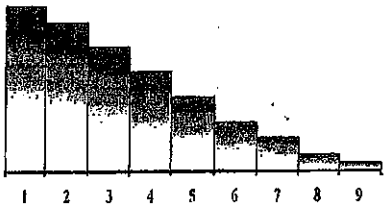
More Spread

SHAPE

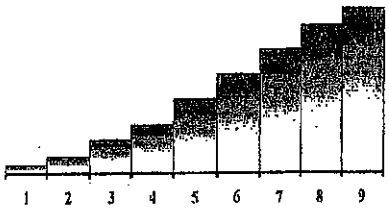
The shape of a distribution is described by symmetry, number of peaks, direction of skew, or uniformity



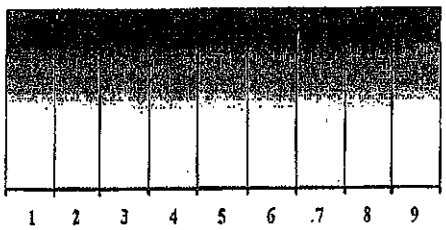
Symmetric, Unimodal, Bell-shaped



Skewed Right

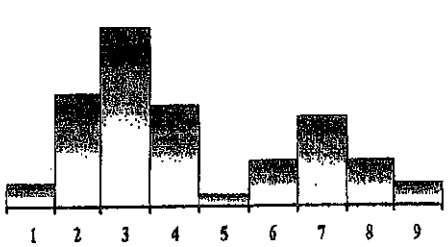


Skewed Left



Uniform

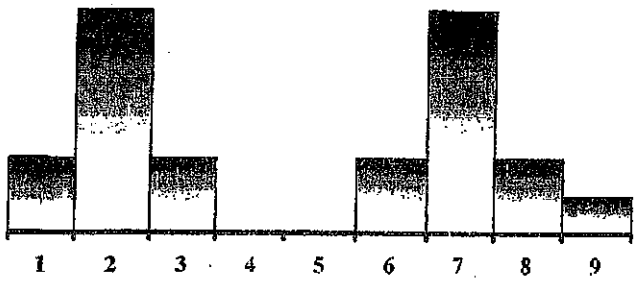
Symmetric, Bimodal



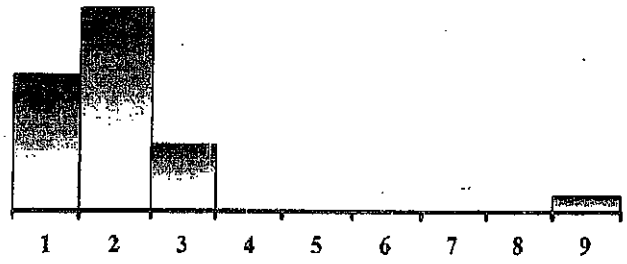
Non-Symmetric, bimodal

UNUSUAL FEATURES

Sometimes, statisticians refer to unusual features in a set of data. The two most common unusual features are gaps and outliers.



Gap



Outlier

Name: _____ Date: _____

Learning Task: If the Shoe Fits!

MCC9-12.S.ID. 1 Represent data with plots on the real number line (dot plots, histograms, and box plots).
MCC9-12.S.ID. 2 Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, mean absolute deviation) of two or more different data sets.
MCC9-12.S.ID. 3 Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

Welcome to CSI at School! Over the weekend, a student entered the school grounds without permission. Even though it appears that the culprit was just looking for a quiet place to study undisturbed by friends, school administrators are anxious to identify the offender and have asked for your help. The only available evidence is a suspicious footprint outside the library door.

After the incident, school administrators arranged for the data in the table below to be obtained from a random sample of this high school's students. The table shows the shoe print length (in cm), height (in inches), and gender for each individual in the sample.

Shoe Print Length	Height	Gender	Shoe Print Length	Height	Gender
24	71	F	24.5	68.5	F
32	74	M	22.5	59	F
27	65	F	29	74	M
26	64	F	24.5	61	F
25.5	64	F	25	66	F
30	65	M	37	72	M
31	71	M	27	67	F
29.5	67	M	32.5	70	M
29	72	F	27	66	F
25	63	F	27.5	65	F
27.5	72	F	25	62	F
25.5	64	F	31	69	M
27	67	F	32	72	M
31	69	M	27.4	67	F
26	64	F	30	71	M
27	67	F	25	67	F
28	67	F	26.5	65.5	F
26.5	64	F	30	70	F
22.5	61	F	31	66	F
			27.25	67	F

1. Explain why this study was an observational study and not an experiment.

2. Why do you think the school's administrators chose to collect data on a random sample of students from the school? What benefit might a random sample offer?

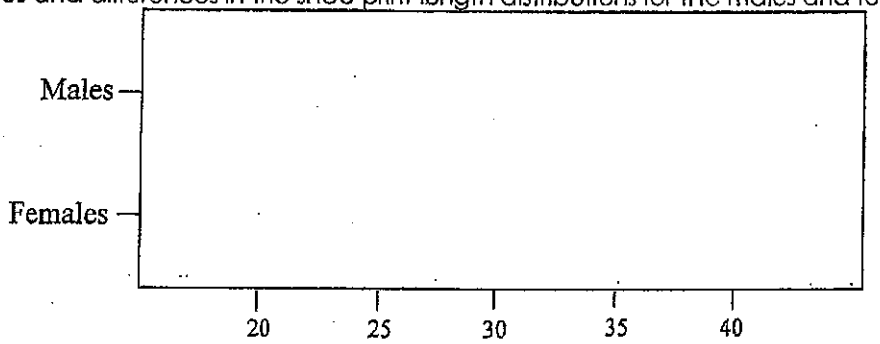
3. Suggest a graph that might be used to use to compare the shoe print length data distributions for females and males.

4. Describe one advantage of using comparative box plots instead of comparative dot plots to display these data.

5. For each gender calculate the five-number summary for the shoe print lengths. Additionally, for each gender, determine if there are any outlying shoe print length values.

	Minimum	Quartile 1 (Q1)	Median (Q2)	Quartile 3 (Q3)	Maximum
Male					
Female					

6. Construct comparative box plots for the shoe print lengths of males and females. Discuss the similarities and differences in the shoe print length distributions for the males and females in this sample.



7. For each gender calculate the mean shoe print length. What information does the mean shoe print length provide?

8. The mean will give us an indication of a typical shoe print length. In addition to knowing a typical length we would also like to know how much variability to expect around this length. For each gender calculate the **Range**; **Interquartile Range**; and **Mean Absolute Deviation** of the shoe print lengths. Interpret each of the calculated values.

	Range	IQR	M.A.D.
Male			
Female			

9. If the length of a student's shoe print was 32 cm...
   A. Would you think that the print was made by a male or a female?

   B. How sure are you that you are correct? Explain your reasoning. Use results from Questions 5 through 8 in your explanation.

10. How would you answer Question 9 if the suspect's shoe print length was 27 cm?

Name: _____ Date: _____

Two-way Frequency Charts

MCC9-12.S.ID.5 Summarize categorical data for two categories in two-way frequency tables. Interpret relative frequencies in the context of the data (including joint, marginal, and conditional relative frequencies). Recognize possible associations and trends in the data.

There are essentially two types of data: **quantitative** and **categorical**.

- Examples of categorical data: color, type of pet, gender, ethnic group, religious affiliation, etc.
- Examples of quantitative data: age, years of schooling, height, weight, test score, etc.

Researchers use both types of data but in different ways. Bar graphs and pie charts are frequently associated with categorical data. Box plots, dot plots, and histograms are used with quantitative data. The measures of central tendency (mean, median, and mode) apply to quantitative data. Frequencies can apply to both categorical and quantitative.

Bivariate data consists of pairs of linked numerical observations, or frequencies of things in categories. Numerical bivariate data can be presented as ordered pairs and in any way that ordered pairs can be presented: as a set of ordered pairs, as a table of values, or as a graph on the coordinate plane.

- An example would be the number of people that play certain sports or are in certain clubs at your school broken down by gender.

A bivariate or **two-way frequency chart** is often used with data from two categories. Each category is considered a variable, and the categories serve as labels in the chart. Two-way frequency charts are made of cells. The number in each cell is the frequency of things that fit both the row and column categories for the cell. From the two-way chart below, we see that there are 12 males in the band and 3 females in the chess club.

Participation in School Activities			
School Club	Gender		
	Male	Female	Totals
Band	12	21	33
Chorus	15	17	32
Chess	16	3	19
Latin	7	9	16
Yearbook	28	7	35
Totals	78	57	135

If no person or thing can be in more than one category per scale, the entries in each cell are called **joint frequencies**. The frequencies in the cells and the totals tell us about the percentages of students engaged in different activities based on gender. For example, we can determine that if we picked at random from the students, we are least likely to find a female in the chess club because only 3 of 135 students are females in the chess club. These frequencies are converted to percents in the chart below.

Participation in School Activities			
School Club	Gender		
	Male	Female	Totals
Band	8.9%	15.6%	24.5%
Chorus	11.1%	12.6%	23.7%
Chess	11.9%	2.2%	14.1%
Latin	5.2%	6.7%	11.9%
Yearbook	20.7%	5.2%	25.9%
Totals	57.8%	42.3%	100%

There is also what we call **marginal frequencies** in the bottom and right margins (grayed cells). These frequencies lack one of the categories. For our example, the frequencies at the bottom represent percents of males and females in the school population. The marginal frequencies on the right represent percents of club membership.

Lastly, associated with two-way frequency charts are **conditional frequencies.** These are not usually in the body of the chart, but can be readily calculated from the cell contents. One conditional frequency would be the percent of females that are in the chorus out of the total number of females in some type of club. 17 of the 57 females are in the chorus, so 29.8%. This could also be stated as "Given that a female in a club is selected, what is the probability that she is in the chorus?"

Practice #1:
Elizabeth surveys 9th graders, 10th graders, and 11th graders in her school. She asks each student how many hours they spend doing homework each night. She records the responses in the table below.

Grade	Hours spent on homework		
	0–2	2–4	More than 4
9	38	12	2
10	21	25	9
11	14	18	20

a) How many 9th graders spend 0–2 hours on homework each night? What frequency is that?

b) How many 10th graders spend 2–4 hours on homework each night? What frequency is that?

c) Which response was the most popular among 11th graders?
0–2 hours, 2–4 hours, or more than 4 hours?

Practice #2:
Cameron surveys students in his school who play sports, and asks them which sport they prefer. He records the responses in the table below.

Gender	Preferred sport		
	Baseball	Soccer	Basketball
Male	49	52	16
Female	23	64	33

a) What is the joint frequency of male students who prefer soccer? (How many male students prefer soccer?)

b) What is the marginal frequency of each type of sport? (Total the number of males and females who played baseball. Then give total for the other two sports.)

Practice #3:
Abigail surveys students in different grades; and asks each student which pet they prefer. The responses are in th table below.

Grade	Preferred pet			
	Bird	Cat	Dog	Fish
9	3	49	53	22
10	7	36	64	10

a) What is the joint frequency of 10th graders who prefer having fish or a cat as a pet?

b) What is the marginal frequency of each type of preferred pet?

Name: _____ Date: _____

Task – Public Opinions

MCC9-12.S.ID.5 Summarize categorical data for two categories in two-way frequency tables. Interpret relative frequencies in the context of the data (including joint, marginal, and conditional relative frequencies). Recognize possible associations and trends in the data.

A public opinion survey explored the relationship between age and support for increasing the minimum wage. The results are found in the following two-way frequency table.

	For	Against	No Opinion	TOTAL
Ages 21-40	25	20	5	50
Ages 41-60	30	30	15	75
Over 60	50	20	5	75
TOTAL	105	70	25	200

Frequency Count

1. In the 41 to 60 age group, what percentage supports increasing the minimum wage? Explain how you arrived at your percentage. What type of probability is this? Joint, marginal, or conditional?

2. Out of the people that have no opinion, what percentage is over 60 years old?

3. What are the marginal frequencies?

4. What are the joint frequencies?

5. Why are joint and marginal frequencies important when describing trends or associations in data? Do you see any significant trends when looking at the frequencies?

Adapted from: Jordan-Granite Consortium (2012). http://secmathccss.wordpress.com/

Task – Leisure Time

1. Using the table below, construct a table displaying the joint and marginal frequencies.

	Dance	Sports	Movies	TOTAL
Women	16	6	8	30
Men	2	10	8	20
TOTAL	18	16	16	**50**

	Dance	Sports	Movies	TOTAL
Women				
Men				
TOTAL				

2. After the basketball game, the statistician did not have time to compute Jana's relative frequency. Complete the table determining the relative frequency for Jana. Discuss any trends or associations from the table below concerning points scored by two basketball players.

Point Value	*Frequency for Jana*	*Relative Frequency for Jana*	*Frequency for Jill*	*Relative Frequency for Jill*
0	0		1	0.025
1	0		1	0.025
2	0		2	0.05
3	0		2	0.05
4	0		3	0.075
5	1		3	0.075
6	0		5	0.125
7	3		4	0.1
8	6		5	0.125
9	5		1	0.025
10	7		4	0.1
11	5		5	0.125
12	4		3	0.075
13	4		0	0
14	3		0	0
15	2		1	0.025
TOTALS	**40**	**1**	**40**	**1**

Adapted from: Jordan-Granite Consortium (2012). http://secmathccss.wordpress.com/

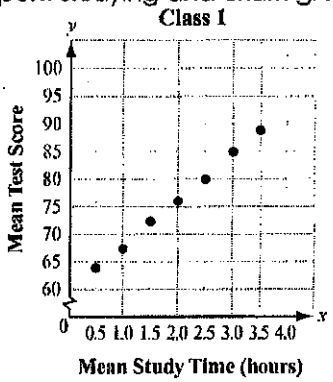
Name: _____ Date: _____

Correlation

MCC9-12.S.ID.6 Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.
MCC9-12.S.ID.9 Distinguish between correlation and causation.

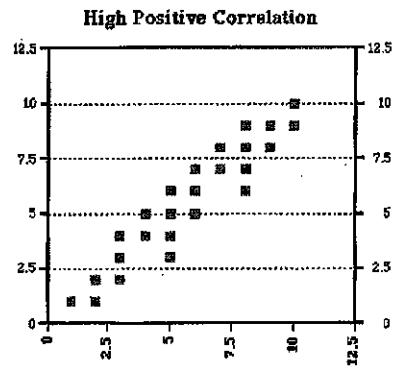
A **scatter plot** is often used to present bivariate **quantitative** data. Each variable is represented on an axis and the axes are labeled accordingly.

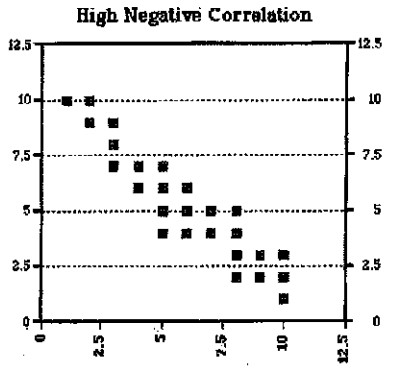
A scatter plot displays data as points on a grid using the associated numbers as coordinates or ordered pairs (x, y). The way the points are arranged by themselves in a scatter plot may or may not suggest a relationship between the two variables. For instance, by reading the graph below, do you think there is a relationship between the hours spent studying and exam grades?

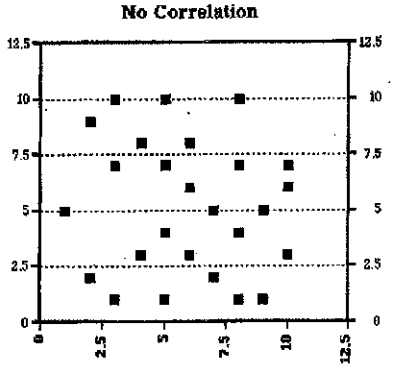


If y tends to increase as x increases, then the data have **positive** correlation.

If y tends to decrease as x increases, then the data have **negative** correlation.





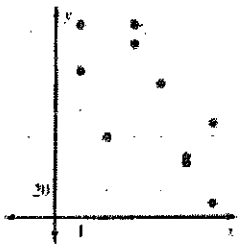


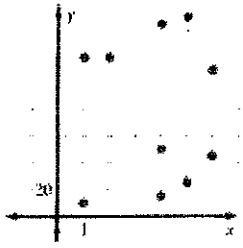
A correlation coefficient, denoted by r, is a number from -1 to 1 that measures how well a line fits a set of data pairs (x, y). If r is near 1, the points lie close to a line with a positive slope. If r is near -1, the points lie close to a line with a negative slope. If r is near 0, the points to not lie close to any line.

Give an example of negative correlation: _____

Practice Problems:

For each scatter plot, tell whether the data have a positive correlation, a negative correlation, or no correlation. Then, tell whether the correlation is closest to -1, -0.5, 0, 0.5, or 1.

1. 

2. 

3. Positive, negative, or no correlation?

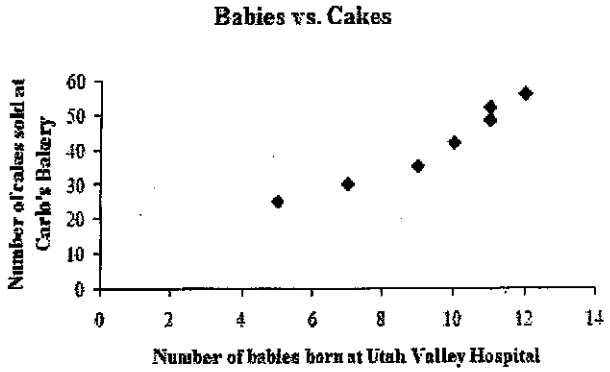
a. Amount of exercise and percent of body fat _____
b. A person's age and the number of medical conditions they have _____
c. Temperature and number of ice cream cones sold _____
d. The number of students at Hillgrove and the number of dogs in Atlanta _____
e. Age of a tadpole and the length of its tail _____

Correlation vs. Causation

When a scatter plot shows a correlation between two variables, even if it's a strong one, there is *not necessarily a cause-and-effect relationship*. Both variables could be related to some third variable that actually causes the apparent correlation. Also, an apparent correlation simply could be the result of chance.

Example 1: During the month of June the number of new babies born at the Utah Valley Hospital was recorded for a week. Over the same time period, the number of cakes sold at Carlo's Bakery in Hoboken, New Jersey was also recorded. What can be said about the correlation? Is there causation? Why or why not?

Number of babies born	Number of cakes sold
5	25
7	30
9	35
10	42
11	48
11	52
12	56



Example 2: Mr. Jones gave a math test to all the students in his school. He made the startling discovery that the taller students did better than the short ones. His Causation Statement: *As your height increases, so does your math ability.*
What can be said about the correlation? Is there causation? Why or why not?

Example 3: In this present economy families are trying to find ways to save money Families might be thinking about not eating out to spend less money. Causation Statement: *The more you eat out, the more money you spend at restaurants.*
What can be said about the correlation? Is there causation? Why or why not?

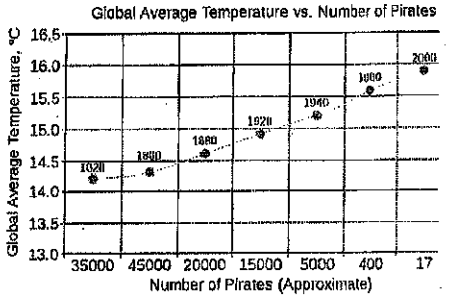
Name: _____ Date: _____

Correlation and Causation Homework

1. From the information given,
   a. Determine if the correlation is positive, negative or none.

   b. Estimate the correlation coefficient.

   c. Is there causation? Why or why not?

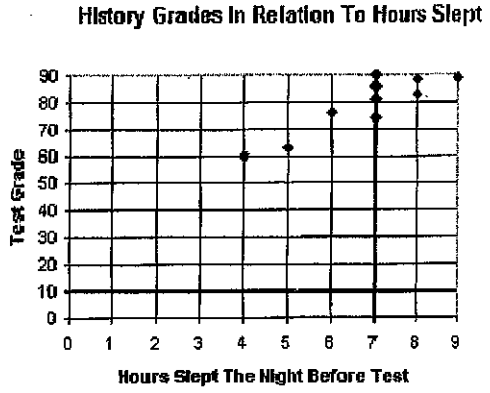


2. A history teacher asked her students how many hours of sleep they had the night before a test. The data above shows the number of hours the student slept and their score on the exam. The graph is a scatter plot from the given data.

   a. Determine if the correlation is positive, negative, or none.

   b. Estimate the correlation coefficient.

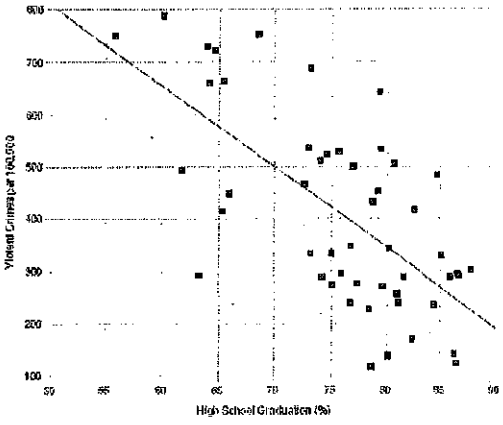
   c. Is there causation? Would this information affect your behavior the night before a test?



3. The following chart shows violent crime rates compared to high school graduation for all fifty states.
   a. Determine if the correlation is positive, negative, or none.

   b. Estimate the correlation coefficient.

   c. Is this an illustration of cause and effect, or are these two variables simply correlated?



Adapted from: Mathematics Vision Project

For the given situations below,

a. Is the association positive, negative or none?

b. Is the causation statement is true or false?

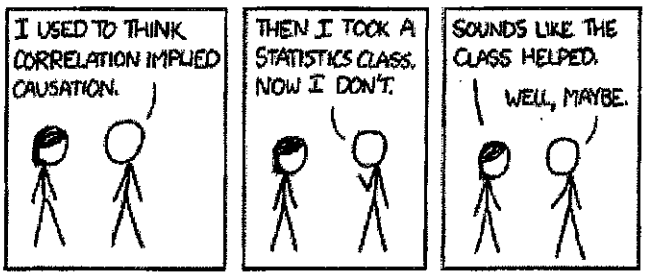
4. When you are on a diet, the less calories you eat daily vs. the more weight you lose. Causation statement: *Therefore, eating less calories makes you lose weight.*

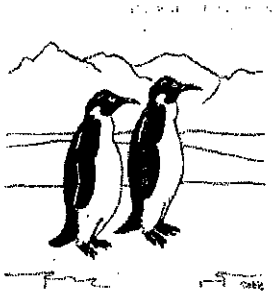
5. The more ice cream consumed on a beach vs. the increased number of people who go in the water. Causation statement: *Therefore, eating more ice cream on the beach makes people go in the water.*

6. The more people in a family vs. the increased number of cars the family owns. Causation Statement: Therefore, the more people there are in a family determines how many cars a family owns.

7. The average speed cars travel from Philadelphia to New York on the turnpike vs. the average amount of times it takes. Causation Statement: *Therefore, the speed cars travel from Philadelphia to New York determines the time it takes to go between them.*

8. How much you pay for a house vs. how much you pay for a car. Causation statement: *Therefore the more you pay for a house makes you spend more for a car.*





"Do you think all these film crews brought on global warming or did global warming bring on all these film crews?"

Adapted from: Mathematics Vision Project

Name: _____ Date: _____

Scatter Plots and Line of Best Fit – TV Task

MCC9-12.S.ID.6 Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.
MCC9-12.S.ID.6a Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use the given functions or choose a function suggested by the context. Emphasize linear and exponential models.
MCC9-12.S.ID.6c Fit a linear function for a scatter plot that suggests a linear association.

1. Students in Ms. Garth's Algebra II class wanted to see if there are correlations between test scores and height and between test scores and time spent watching television. Before the students began collecting data, Ms. Garth asked them to predict what the data would reveal. Answer the following questions that Ms. Garth asked her class.

   a. Do you think students' heights will be correlated to their test grades? If you think a correlation will be found, will it be a positive or negative correlation? Will it be a strong or weak correlation?

   b. Do you think the average number of hours students watch television per week will be correlated to their test grades? If you think a correlation will be found, will it be a positive or negative correlation? Will it be a strong or weak correlation? Do watching TV and low test grades have a cause and effect relationship?

2. The students then created a table in which they recorded each student's height, average number of hours per week spent watching television (measured over a four-week period), and scores on two tests. Use the actual data collected by the students in Ms. Garth's class, as shown in the table below, to answer the following questions.

Student	1	2	3	4	5	6	7	8	9	10	11	12	13
Height (in inches)	60	65	51	76	66	72	59	58	70	67	65	71	58
TV hrs/week (average)	30	12	30	20	10	20	15	12	15	11	16	20	19
Test 1	60	80	65	85	100	78	75	95	75	90	90	80	75
Test 2	70	85	75	85	100	88	85	90	90	90	95	85	85

   a. Which pairs of variables seem to have a positive correlation? Explain.

   b. Which pairs of variables seem to have a negative correlation? Explain.

   c. Which pairs of variables seem to have no correlation? Explain.

3. Using the statistical functions of your graphing calculator, determine a line of good fit for each of the following categories.

   a. Score on Test 1 versus hours watching television:

   b. Score on test 1 versus score on test 2:

   c. Hours watching television versus score on test 2:

4. Use your answer to 3a to predict the test 1 score of someone who watches tv for 40 hours per week.

5. Use your answer to 3c to predict the test 2 score of someone who watches tv for 5 hours per week.

TI 30X LINE OF BEST FIT STEPS

1. 2nd DATA choose 2-VAR
2. DATA (enter data and use down arrow)
3. STAT VAR
4. Arrow over to find

   a =

   b =

   r =
5. The equation of the line is y = **a**x + **b.**
6. Correlation Coefficient is r.
7. To predict use **a(predict #) + b**. *Estimated method*

TI 30 MULTIVIEW LINE OF BEST FIT STEPS

1. DATA (type in data)
2. 2nd DATA
3. 2 VAR L1 L2 CALC (enter)

   TI-36 Pro 2 VAR L1 L2 Frequency of 1 Calc
4. a =

   b =

   r =

   ★ You can use the x variable button to find a, b, and r.
5. The equation of the line is y = **a**x + **b.**
6. Correlation Coefficient is r.
7. To predict use **a(predict #) + b**. *Estimated method*

TI 83 or 84 of Best Fit Steps

1. DATA, then EDIT (type in data)
2. DATA, then CALC
3. 4: LinReg(ax+b)
4. a =

   b =

   r =

   ★ You can use the x variable button to find a, b, and r.
5. The equation of the line is y = **a**x + **b.**
6. Correlation Coefficient is r.
8. To predict use **a(predict #) + b**. *Estimated method*

Name: _____ Date: _____

Scatter Plots and Line of Best Fit

MCC9-12.S.ID.6 Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.
MCC9-12.S.ID.6a Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use the given functions or choose a function suggested by the context. Emphasize linear and exponential models.
MCC9-12.S.ID.6c Fit a linear function for a scatter plot that suggests a linear association.

The **best fitting line or curve** is the line that lies as close as possible to all the data points.

Regression is a method used to find the equation of the best fitting line or curve.

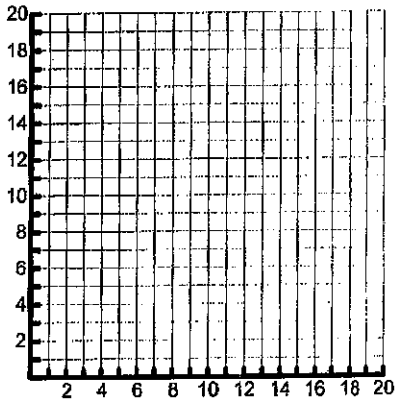
Extrapolation – the use of the regression curve to make predictions outside the domain of values of the independent variable.

Interpolation – Interpolation is used to make predictions within the domain of values of the independent variable.

Line of Best Fit by Hand:

1) The environment club is interested in the relationship between the number of canned beverages sold in the cafeteria and the number of cans that are recycled. The data they collected are listed in this chart.

Beverage Can Recycling								
Number of Canned Beverages Sold	18	15	19	8	10	13	9	14
Number of Cans Recycled	8	6	10	6	3	7	5	4



a) Plot the points to make a scatter plot.

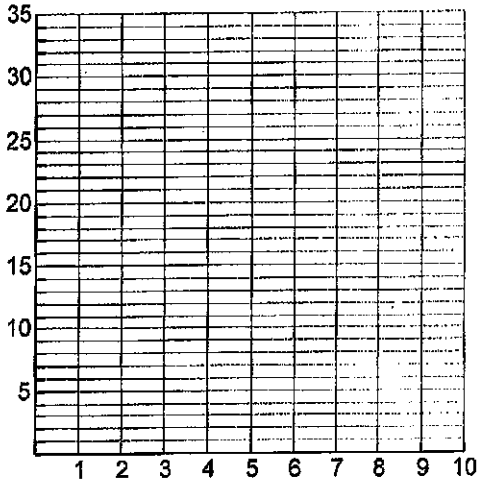
b) Use a straightedge to approximate the line of best fit by hand.

c) Find an equation of the line of best fit for the data.

---

2. Mike is riding his bike home from his grandmother's house. In the table below, x represents the number of hours Mike has been biking and y represents the number of miles Mike is away from home. Make a scatter plot for this data on the grid below.

Hours (x)	1	2	3	4	5	6	7	8
Miles (y)	35	29	26	20	16	9	6	0



a) Describe the association between the data points on the scatter plot.

b) Use a straightedge to approximate the line of best fit.

c) Find an equation of the line of best fit for the data.

d) What does the slope represent in the context of the problem? What does the y-intercept represent in the context of the problem?

e) Could you use your equation to predict how far Mike would be after 10 hours? Use mathematics to justify your answer.

Line of Best Fit using the calculator

3) Use the table below to answer the questions about the population p (in millions) in Florida.

Year, t	2002	2003	2004	2005
Population (millions)	16.4	17.0	17.4	17.8

a) Find the best-fitting line for the data and the correlation coefficient.

b) Using this model, what will be the population in 2020?

---

4) Use the table below to answer the questions about the U.S. residential carbon dioxide emissions from 1993 to 2002. Emissions are measured in million metric tons.

Year, t	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Emissions	1027.6	1020.9	1026.5	1086.1	1077.5	1083.3	1107.1	1170.4	1163.3	1193.9

a) Find the best-fitting line for the data and the correlation coefficient.

b) Using this model, how many residential tons were emitted in 1990? In 2010?

---

5) Use the table below to answer the questions about the operating costs in thousands of a small business from 2000 to 2007.

Year, t	2000	2001	2002	2003	2004	2005	2006	2007
Operating Costs	2.3	2.6	3.1	3.3	4.0	5.2	5.9	7.0

a) Find the best-fitting line for the data and the correlation coefficient.

b) Using this model, what will be the operating costs in 2015?

Name: _____ Date: _____

Exponential Regression

MCC9-12.S.ID.6 Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.
MCC9-12.S.ID.6a Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use the given functions or choose a function suggested by the context. Emphasize linear and exponential models.
MCC9-12.S.ID.8 Compute (using technology) and interpret the correlation coefficient of a linear fit.

Hot Coffee

The data at the right shows the cooling temperatures of a freshly brewed cup of coffee after it is poured from the brewing pot into a serving cup. The brewing pot temperature is approximately 180° F.

Time (mins)	Temp (F°)
0	179.5
5	168.7
8	158.1
11	149.2
15	141.7
18	134.6
22	125.4
25	123.5
30	116.3
34	113.2
38	109.1
42	105.7
45	102.2
50	100.5

a) Determine an exponential regression model equation to represent this data.

b) Decide whether the new equation is a "good fit" to represent this data.

c) Based upon the new equation, what was the initial temperature of the coffee? What is the decay rate?

d) Interpolate data: When is the coffee at a temperature of 106 degrees?

e) Extrapolate data: What is the predicted temperature of the coffee after 1 hour?

f) In 1992, a woman sued McDonald's for serving coffee at a temperature of 180° that caused her to be severely burned when the coffee spilled. An expert witness at the trial testified that liquids at 180° will cause a full thickness burn to human skin in two to seven seconds. It was stated that had the coffee been served at 155°, the liquid would have cooled and avoided the serious burns. The woman was awarded over 2.7 million dollars. As a result of this famous case, many restaurants now serve coffee at a temperature around 155°. How long should restaurants wait (after pouring the coffee from the pot) before serving coffee, to ensure that the coffee is not hotter than 155°?

g) If the temperature in the room is 76° F, what will happen to the temperature of the coffee, after being poured from the pot, over an extended period of time?

Practice Problems:

1. Estimates for world population vary, but the data in the accompanying table are reasonable estimates of the world population from 1800 to 2000.

Year	Total Population (millions)
1800	980
1850	1260
1900	1650
1950	2520
1970	3700
1980	4440
1990	5270
2000	6080

a. Identify your independent and dependent variables.

b. Generate a best fit exponential function using your variables. Round to 3 decimals.

c. What does your model give for the growth rate? Describe this in the context of the problem.

d. Using the function, estimate the world population in 1750 and 2050 to 3 decimal places.

---

2. **Town Planning:** The town planners designed their town for an optimal growth of 8% per year. The present school construction will serve a population of 200,000. Below is a table representing the growth from 1997 to 2003.

Year	Population
1997	50,000
1998	54,000
1999	58,000
2000	62,986
2001	68,024
2002	73,466
2003	79,344

a) Find and write the model of a linear regression. Use the model to determine what the population was in 1977. Round to 2 decimals.

b) Find and write the model of an exponential regression. Use the model to determine what the population was in 1977. Round to 2 decimals.

c) Determine which model is better to use. Explain why you selected your model.

d) Using the better model, predict what the population will be in the year 2017.

e) In what year will the population double for the better model?

Name: _____ Date: _____

Exponential Regression Homework

1) **Baseball Salaries.** Ball players have been signing ever-larger contracts. The highest salaries (in millions of dollars per season) for some notable players are given in the following table.

Player	Year	Salary (millions $)
Nolan Ryan	1980	1.0
George Foster	1982	2.04
Kirby Puckett	1990	3.0
Jose Canseco	1990	4.7
Roger Clemens	1991	5.3
Ken Griffey, Jr.	1996	8.5
Albert Belle	1997	11.0
Pedro Martinez	1998	12.5
Mike Piazza	1999	12.5
Mo Vaughn	1999	13.3
Kevin Brown	1999	15.0
Carlos Delgado	2001	17.0
Alex Rodriguez	2001	25.2

a) Find an exponential model for this data. Use years since 1980 for the x variable.

b) What is the growth rate of salaries? State the growth rate in the context of the problem?

c) Predict a salary for the year 2012.

d) Using the Internet, see if you can find out the largest baseball salary this year? How does it compare to your prediction?

2) **Internet Users.** Reliable data about Internet use are hard to come by. But Nau Internet Surveys cites estimates of 18 million Internet users in the U.S. in 1995, 76 million users in 1998, and 119.2 million in 1999.

a) Using regression, determine whether a linear or an exponential function would be a better model? Explain. Write down both equations, rounding to 3 decimals.

b) What is the slope from the best-fit linear function? Interpret this in the context of the problem.

c) What was the annual growth factor and growth rate from the best-fit exponential function? Interpret the rate in the context of the problem.

d) Using the model that you felt was the best in part a, how many users do you predict for 2020?

3) **Coins.** A box containing 1,000 coins is shaken and the coins are emptied onto a table. Only the coins that land heads up are returned to the box, and then the process is repeated. The accompanying table shows the number of trials and the number of coins returned to the box after each trial.

Trial	0	1	3	4	6
Coins Returned	1000	610	220	132	45

a) Write an exponential regression equation, rounding the calculated values to the nearest ten-thousandth.

b) Use the equation to predict how many coins would be returned to the box after the eighth trial.

AC CCGPS Alg/Geo
Unit 4: Describing Data
Class notes on **RESIDUALS**

Name:_____

My notes from power point:

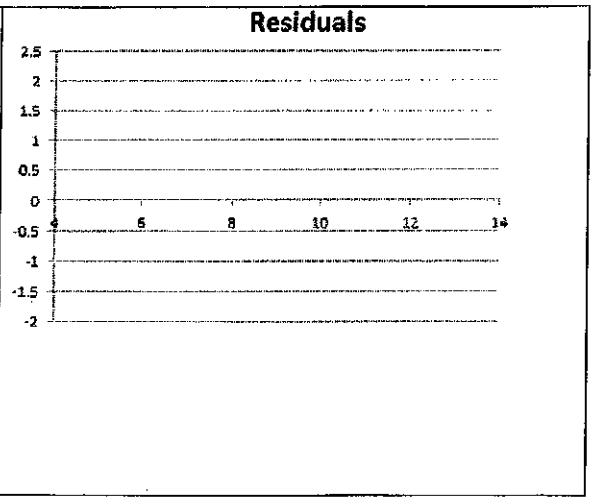
Main idea of residuals:

Example:

Height in inches	Hand size in inches (from tip of middle finger to wrist)

Line of best fit:_____

Height in inches	Predicted length of hand (use the line of best fit)	Residual (actual-predicted)



33A

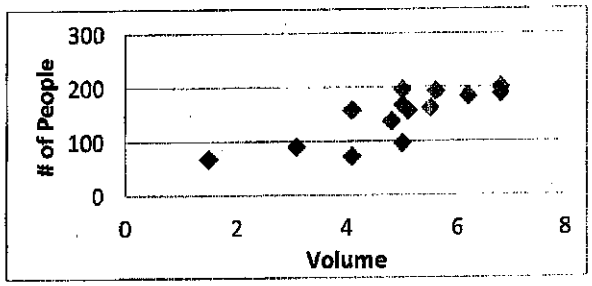
Name: _____ Date: _____

Residuals Task - Carnival

You work for a traveling carnival, and your exhibit has been struggling with a lack of customers. You bought a new megaphone and decide to run an experiment. Each day, you randomly choose a volume setting (from 1 to 10) on the megaphone. You use the megaphone throughout the day and record the number of customers that visit your exhibit. The data is given in the table below.

Volume Setting	5.	5.	3.1	6.8	6.2	5.	6.8	5.5	4.1	4.1	4.8	1.5	5.6	5.1
# People	195	96	90	188	183	166	200	161	72	157	137	68	192	156

Consider the scatter plot of the data.



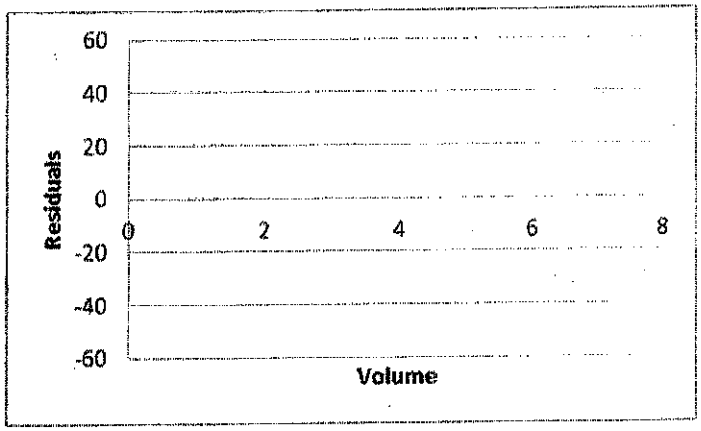
1. Describe the form, direction, and strength of the data.

2. Find an equation of the best-fit line and the correlation coefficient. Label each variable.

3. What does the slope and y-intercept represent in this scenario?

4. **Residual Plots**: Is a line a good fit of the data?

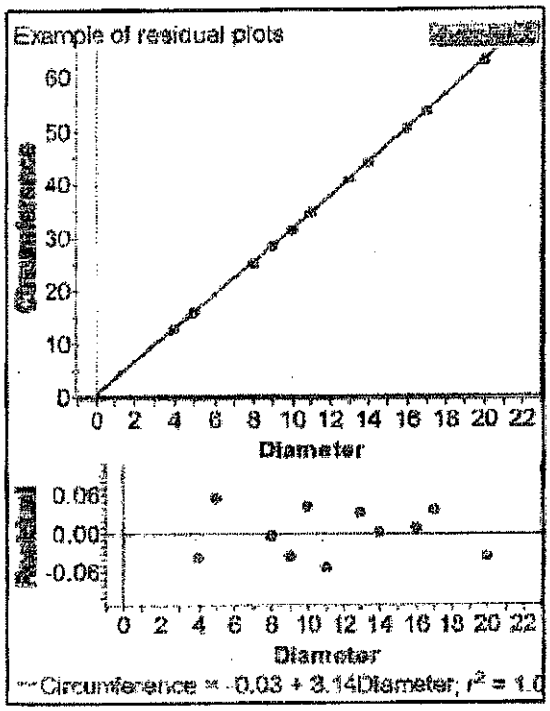
Volume	# of People	Predicted	Residual
5	195		
5	96		
3.1	90		
6.8	188		
6.2	183		
5	166		
6.8	200		
5.5	161		
4.1	72		
4.1	157		
4.8	137		
1.5	68		
5.6	192		
5.1	156		

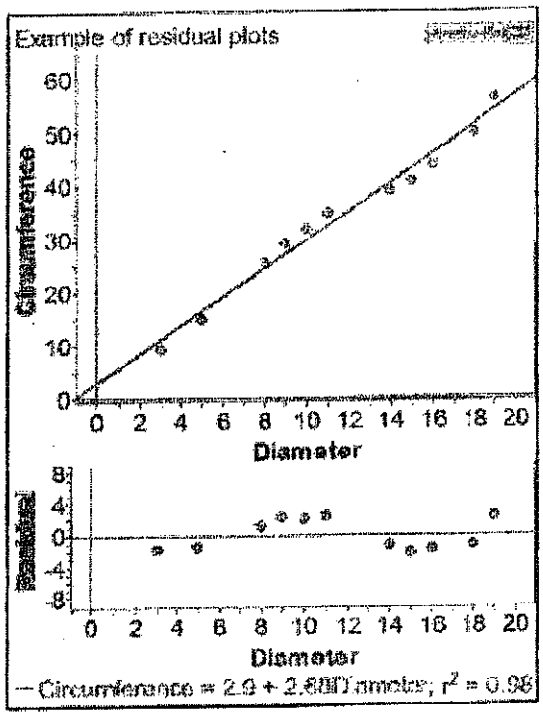


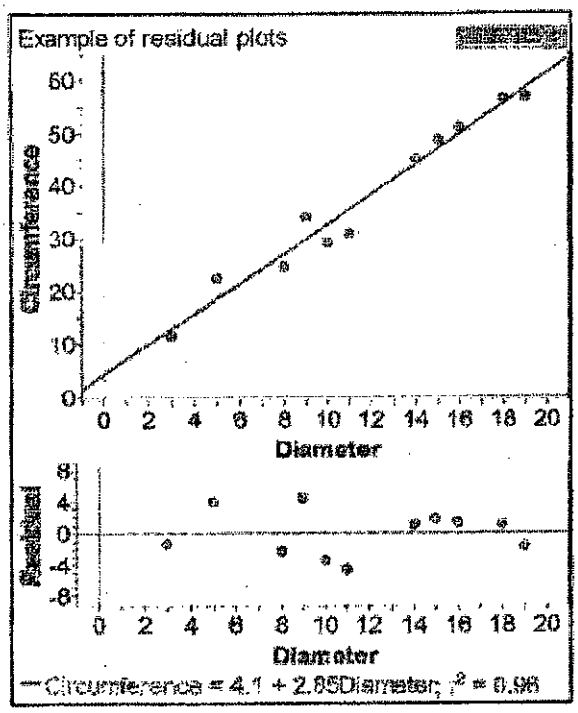
Name: _____ Date: _____

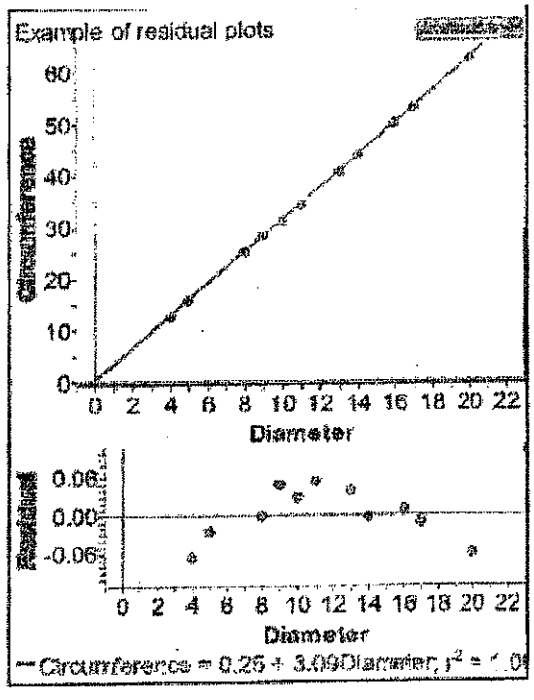
Residuals Homework

Consider the following plots made for the circumference of a ball vs. the diameter of the ball. Based on the corresponding residual plots, which graphs would you think that the linear model is a good fit for?









Consider the following data:

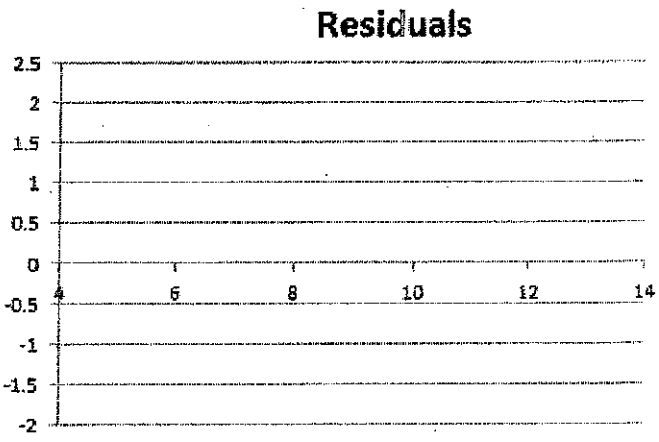
The shoe sizes and heights (in inches) for men.

Shoe Size (x)	Height (y)
8.5	66.0
9.0	68.5
9.0	67.5
9.5	70.0
10	70.0
10	72.0
10.5	71.5
10.5	69.5
11.0	71.5
11.0	72
11.0	73
12.0	73.5
12.0	74
12.5	74

1. **Find the equation for the line of best fit, as well as the correlation coefficient.**

2. **The equation of the line you just wrote is called a prediction equation. For each shoe size (x), calculate the predicted value of y and the residual.**

Shoe Size (x)	Predicted Height	Residual (Actual-Predicted)
8.5		
9.0		
9.0		
9.5		
10		
10		
10.5		
10.5		
11.0		
11.0		
11.0		
12.0		
12.0		
12.5		

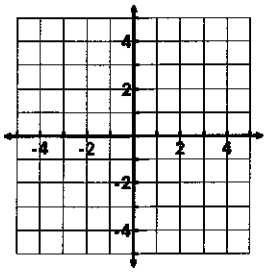


3. **Is there a pattern? Is the prediction line the best model for the data? How can you tell?**

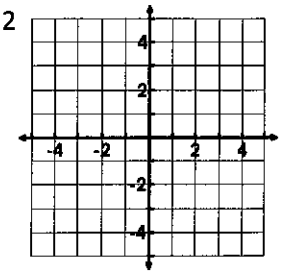
AC CCGPS ALG/GEO
Unit 4: Describing Data CFA

Name:_____

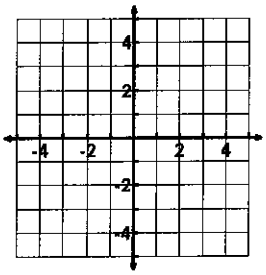
Sketch each of the following:

1. 

Positive, weak

2 

negative, strong

3. 

no correlation

4. If the correlation coefficient is r = -.76, what can you conclude about the data?

5. What is the line of best fit (linear) for the data provided: The data is the money a waitress made one night

hours worked	1	2	3	4	5	6
Money made	12	18	23	30	36	41

Line of best fit:_____

6. Make a box and whisker plot with the following data:
5, 6, -7, 0, 7, 8, 5, 10, 5, 7

Are there any outliers? If so, what are the limitations?_____

What is the interquartile range?_____

7. What is the Mean Absolute Value of the data above?

8. If the exponential regression equation is $y = 121.3(.89)^x$ ....

a) What is the initial value?_____

b) What is the rate of decay?_____ c) what is the value of y when x = 3?_____

---

9. A study of graduates' average grades and degrees showed the following results.

Degree	C	B	A	Total
B.S.	5	8	15	
B.A.	7	12	8	
Total				

If a graduate student is selected at random, find these conditional frequencies.

a. The graduate has a B.S. degree, given that he or she has an A average.

b. Given that the graduate has a B.A. Degree, the graduate has a C Average.

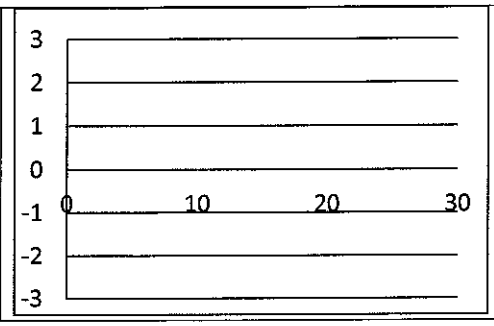
---

10. For the set of data, perform a residual analysis. Fill in the table with the predicted values and residual values (use 3 decimal places).

$y = .786x - 3.45$

Construct the residual plot:

x	y	predicted	residual
5	9.12		
12	7.89		
22	12.87		
18	8.1		
3	9		
14	5.6		



Based on the residuals, is this a good fit?_____ Why or why not?