Describing Data **Name: _____**

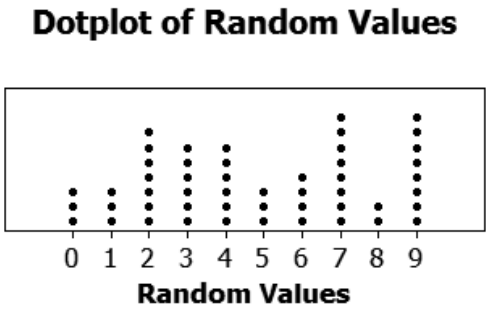
Date: _____

Summarize, represent, and interpret data on a single count or measurement variable
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, standard deviation) of two or more different data sets.★ *(Standard deviation is left for Advanced Algebra, use MAD as a measure of spread.)*
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).
Summarize, represent, and interpret data on two categorical and quantitative variables
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.
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 given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models.
MCC9-12.S.ID.6b Informally assess the fit of a function by plotting and analyzing residuals.
MCC9-12.S.ID.6c Fit a linear function for a scatter plot that suggests a linear association.
Interpret linear models
MCC9-12.S.ID.7 Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.
MCC9-12.S.ID.8 Compute (using technology) and interpret the correlation coefficient of a linear fit.
MCC9-12.S.ID.9 Distinguish between correlation and causation.

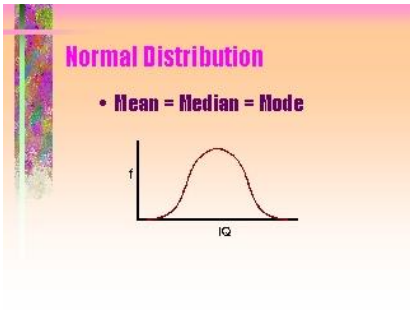
Lesson 4.1 Representing Data Graphically

Two **measures of central tendency** that help describe a data set are **mean** and **median**. The **mean** is the sum of the data values divided by the total number of data values. The **median** is the middle value when the data values are written in numerical order. If a data set has an even number of data values, the median is the mean of the two middle values.

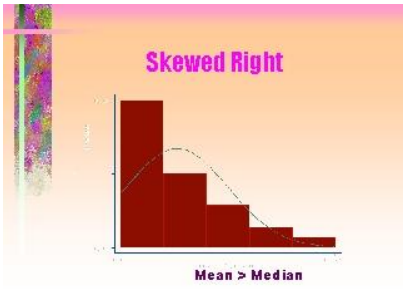
The **dot plot** as a representation of a distribution consists of group of data points plotted on a simple scale:



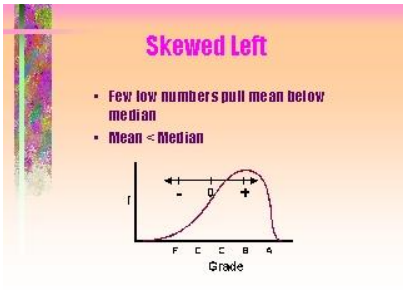
In a normal distribution the peak is at the center (median = average = mode):



A data distribution is skewed right if the peak of the data is to the left of the center:

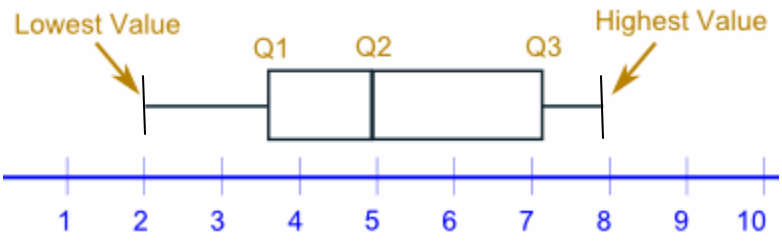


A data distribution is skewed left if the peak of the data is to the right of the center:



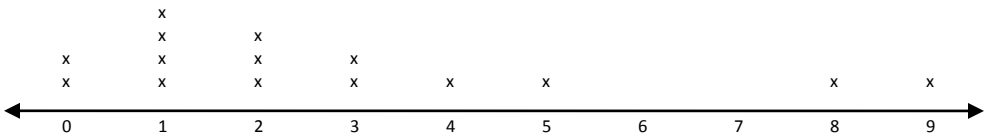
The first quartile or lower quartile, Q1, is the median of the lower half of the data set. The third quartile or upper quartile, Q3, is the median of the upper half of the data set.

Box and whisker plots show a box whose left side and right side are the 25th (Q1) and 75th (Q2) percentile (the lower and upper quartiles), respectively. The band near the middle of the box is always the 50th percentile (the median). The ends of the whiskers represent the minimum and maximum of all the data.



Example: Construct a dot plot and box-and-whisker plot for the data: $2, 0, 5, 1, 2, 1, 0, 8, 4, 3, 9, 1, 2, 3, 1.$

Dot Plot:



Box-and-Whisker Plot: First you have to find the lowest number, highest number, median, 25th percentile, and 75th percentile.

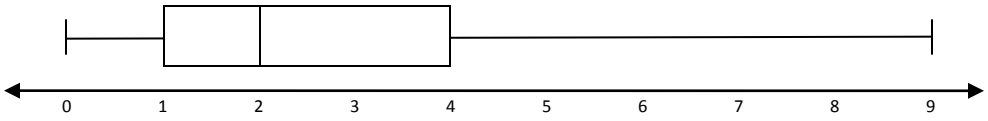
Lowest Number: 0 Highest Number: 9

To find the median list the numbers in ascending order and pick the middle number: $0, 0, 1, 1, 1, 1, 2, \textcircled{2}, 2, 3, 3, 4, 5, 8, 9$

The 25th percentile is the median of the lower half: $0, 0, 1, \textcircled{1}, 1, 1, 2$

The 75th percentile is the median of the upper half: $2, 3, 3, \textcircled{4}, 5, 8, 9$

Now we have all the info to construct a box and whisker plot.



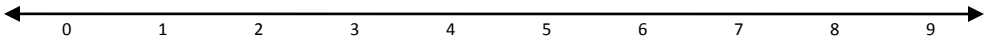
Based on the plot you can conclude the following:

1. 25% of the numbers are 1 or less.
2. 50% of the numbers are 2 or less.
3. 75% of the numbers are 4 or less.
4. The lowest number is 0.
5. The highest number is 9.

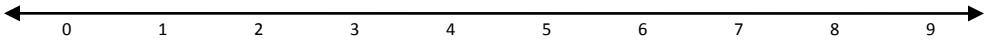
Problems:

1. Construct a dot plot and a box-and-whisker plot for following data set: $0,1,4,3,2,5,8,6,1,2,0,9,4,4,6,7,7,3,2$

Dot Plot:



Box and Whisker Plot:



Based on the plot fill in the information:
1. 25% of the numbers are _____ or less.
2. 50% of the numbers are _____ or less.
3. 75% of the numbers are _____ or less.
4. The lowest number is _____.
5. The highest number is _____.

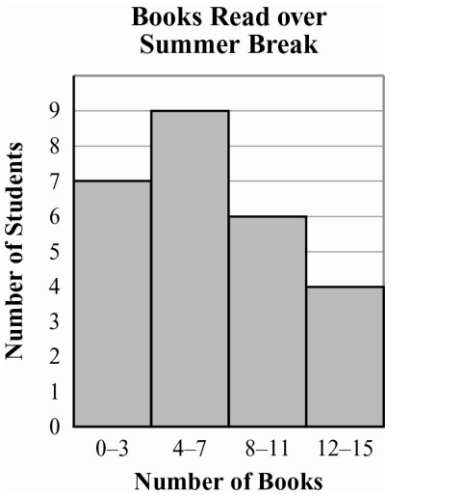
2. Mr. Corson's recent science test had the following scores:
$90,95,100,70,85,65,90,80,65,70,75,80,85,80,60,80,75,85.$

a. Construct a box-and-whisker plot.

b. Do the data represent a normal distribution or are they skewed to the left or right?

Lesson 4.2 Comparing Distributions

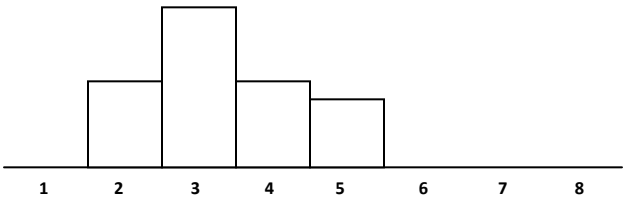
A **histogram** is a graphical display that subdivides the data into class intervals, called bins, and uses a rectangle to show the frequency of observations in those intervals—for example, you might use intervals of 0–3, 4–7, 8–11, and 12–15 for the number of books students read over summer break.



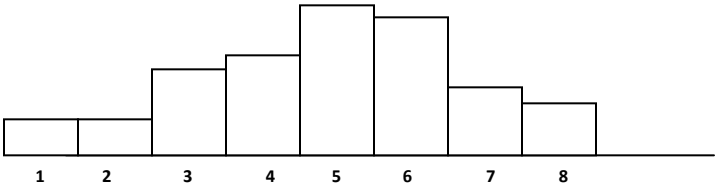
When you compare two or more data sets, you need to be able to talk about four features:

1. Center Graphically, the center of a distribution is the point where about half of the observations are on either side.

2. 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.

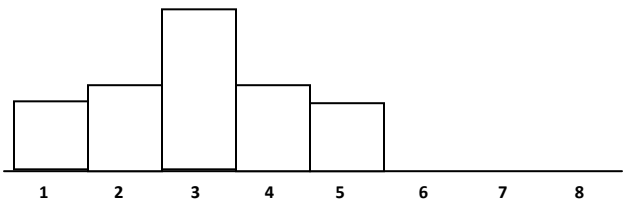


Less Spread

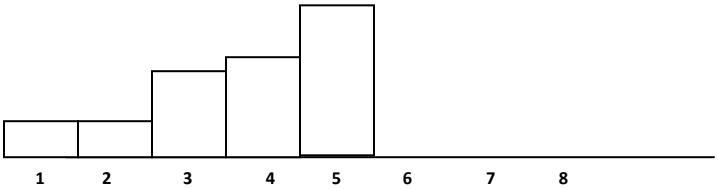


More Spread

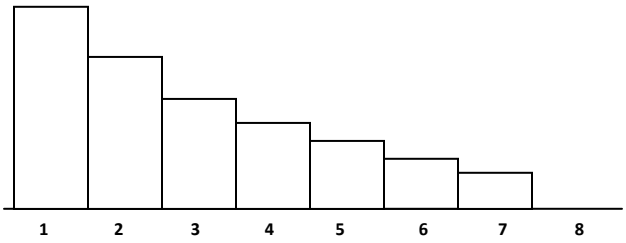
3. Shape The shape of a distribution is described by symmetry, skewness, number of peaks, etc.



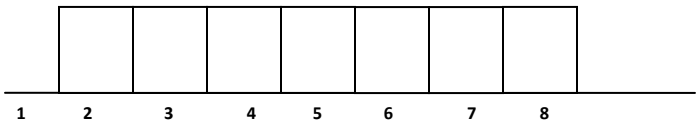
symmetric, bell-shaped, uni-modal



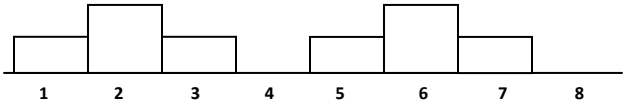
skewed left



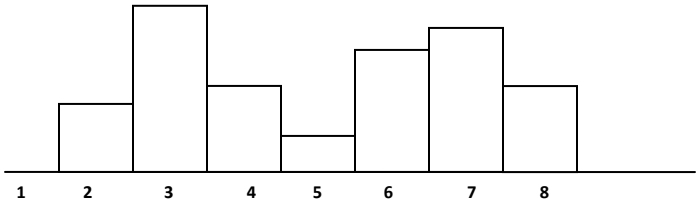
skewed right



uniform

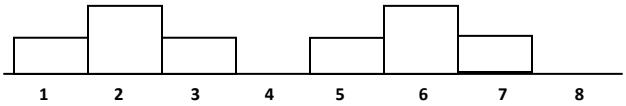


symmetric, bi-modal

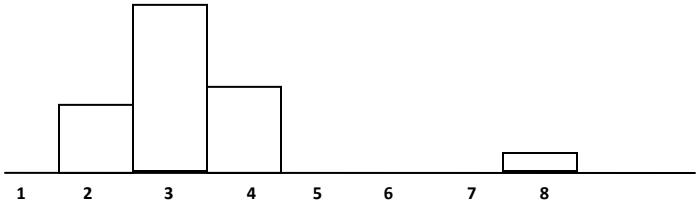


non-symmetric, bi-modal

4. Unusual features Unusual features refer to gaps (areas of the distribution where there are no observations) and outliers.



gap



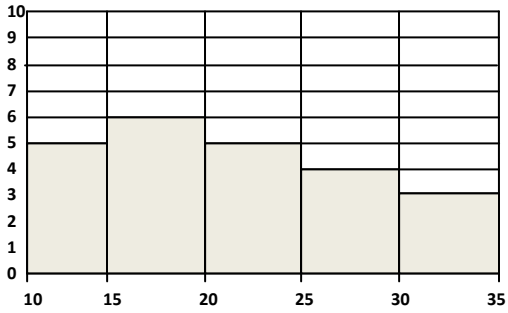
outlier

Sometimes, distributions are characterized by extreme values that differ greatly from the other observations. These extreme values are called **outliers**. A data value is an **outlier** if it is less than Q1 – 1.5 • IQR or above Q3 + 1.5 • IQR.

For example, consider these two sets of quiz scores:
Student P: {8, 9, 9, 9, 10} Mean = 9, Median = 9, Mode = 9
Student Q: {3, 9, 9, 9, 10} Mean = 8, Median = 9, Mode = 9
Both students consistently performed well on quizzes and both have the same median and mode score, 9. Student Q, however, has a mean quiz score of 8, while Student P has a mean quiz score of 9. Although many instructors accept the use of a mean as being fair and representative of a student's overall performance in the context of test or quiz scores, it can be misleading because it fails to describe the variation in a student's scores, and the effect of a single score on the mean can be disproportionately large, especially when the number of scores is small.

PROBLEMS

1. Analyze the given histogram which displays the ACT composite score of several randomly chosen students.



a. How many students are represented by the histogram?

b. How many student scores fall between 15 and 25?

c. How many students have scores less than 30?

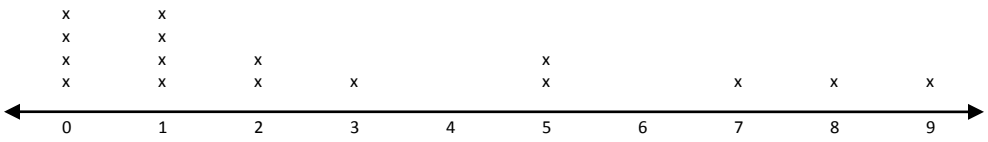
d. How many students have scores between 10 and 15 and between 20 and 25?

e. Can you determine how many students scored a 20? Why or why not?

f. Describe the distribution of the data and explain what it means in terms of the problem situation.

2. In your own words describe an outlier. Show an example in a data set of your own choice.

3. Analyze the given dot plot which displays the number of home runs by each of the members of the Atlanta Braves team this season so far and answer the questions accordingly.



a. Describe the distribution of the data. Are the skewed left or right? What do the data mean in terms of the problem situation?

b. How many players are on the team?

c. How many players hit more than 2 home runs?

d. How many players hit at least 1 home run?

e. How many players hit more than 1 and fewer than 9 home runs?

f. How many players scored more than 9 home runs?

g. How many players hit more than 1 and fewer than 5 home runs?

h. How many players scored less than 3 home runs?

i. Compute the mean of the data set (average homeruns per player).

Lesson 4.3 Measures of Central Tendency and Dispersion

Goal: Describe data using statistical measures.

Statistics are numerical values used to summarize and compare sets of data.

The mean (or average) of n numbers is the sum of the numbers divided by n. The mean is denoted by $\bar{x}$.

The median of n numbers is the middle number when arranged in order. If n is even, then the median is the average of the two middle numbers.

The mode of n numbers is the number which occurs most often.

The standard deviation is given by: $\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ...(x_n - \bar{x})^2}{n}}$.

The standard deviation is a measure of dispersion. A larger standard deviation means a larger dispersion. Another measure of dispersion is the range, which is the difference between the largest and smallest value.

Example 1: Find the mean, median, mode, range, and standard deviation of the data set.

Data Set: 65, 86, 90, 100, 55, 69, 82, 77, 81

$$Mean = \frac{65+86+90+100+55+69+82+77+81}{9} \approx 78.3$$

$Median$: $55, 65, 69, 77, 81, 82, 86, 90, 100$ $The\ middle\ number\ (median)\ is\ 81.$

$Mode$: There is no mode since every numbers appears only once.

$Range$: $100 - 55 = 45$

$$\sigma = \sqrt{\frac{(65-78.3)^2 + (86-78.3)^2 + (90-78.3)^2 + (100-78.3)^2 + (55-78.3)^2 + (69-78.3)^2 + (82-78.3) + (77-78.3) + (81-78.3)^2}{9}} = 12.9$$

Problems:

14. Find the mean, median, mode, range, and standard deviation of the data set: 49, 111, 91, 53, 55, 64, 62, 64.

Another way to describe the variability of a set of data is to use its **mean absolute deviation**. The **mean absolute deviation** is the average distance between each data value and the mean.

Find the mean absolute deviation for the data set 5,7,7,5,9,3.

First find the mean: $Mean = \frac{5+7+7+5+9+3}{6} = 6$

$$Mean\ Absolute\ Deviation = \frac{(x_1 - \bar{x}) + (x_2 - \bar{x}) + ...(x_n - \bar{x})}{n}$$

Plug in the corresponding values:

$$\frac{|5-6|+|7-6|+|7-6|+|5-6|+|9-6|+|3-6|}{6} = \frac{|-1|+|1|+|1|+|-1|+|3|+|-3|}{6} = \frac{1+1+1+1+3+3}{6} = \frac{10}{6} \approx 1.7$$

15. Find the absolute deviation of the data set: 49, 111, 91, 53, 55, 64, 62, 64.

16. The data shown below displays the running times in minutes for science-fiction movies.

Find the absolute deviation of the data set. 98, 87, 93, 88, 126, 108

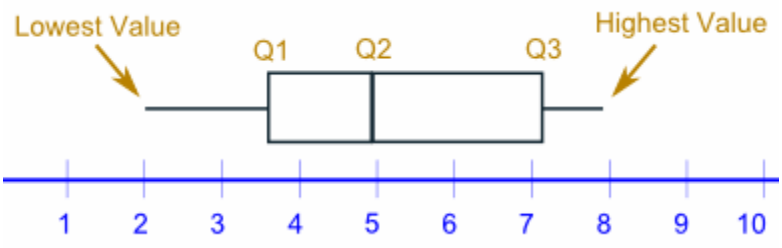
Interquartile Range – Box and Whisker Plots

In descriptive statistics, the interquartile range (IQR), also called the midspread or middle fifty, is a measure of statistical dispersion, being equal to the difference between the upper (Q3)and lower (Q1) quartiles (IQR = Q3 − Q1). (*Upton, Graham; Cook, Ian (1996). Understanding Statistics. Oxford University Press. p. 55. ISBN 0-19-914391-9.)*

To find the interquartile range you have to follow these steps:

1. Put the numbers in ascending order
2. Find the median.
3. Split the data into the upper half (all numbers above the median) and the lower half (all numbers below the median).
4. Find the median of the upper half and the median of the lower half.
5. The interquartile range (IQR) is the number between the median of the upper half and the median of the lower half.

The corresponding **box and whisker plot** is constructed using the lowest value of the data, Q1 (lower quartile or median of the lower half), Q2 (the median), Q3 (upper quartile or median of upper half), and the highest value of the data.



Example 1

The data are 5, 8, 4, 4, 6, 3, 10
1. Put them in order: 3, 4, 4, 5, 6, 8, 8
2. The median is 5.
3. The data for the upper half are 6,8,10 and the data for the lower half are 3,4,4
4. The median of the upper half is 8, the median of the lower half is 4.
5. The interquartile range is 8 - 4 = 4

The corresponding box and whisker plot looks like:

Example 2

The data are 1, 3, 3, 4, 5, 6, 6, 7, 8, 8

1. The numbers are already in order
2. The median is (5+6)/2 = 5.5
3. The data for the upper half are 6,6,7,8,8 and the data for the lower half are 1,3,3,4,5.
4. The median of the upper half is 7 and the median of the lower half is 3.
5. The interquartile range is 7 - 3 = 4.

The corresponding box and whisker plot looks like:

Problems: Find Q1, Q2, and Q3 of the following data sets; then construct a box and whisker plot.

1. 4,6,5,1,9,7,8,12,15,15,12,2,4,

2. 34,55,23,12,44,34,67,34,23,36

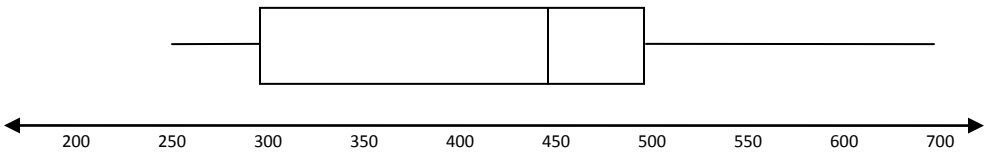
3. 500, 600, 420, 120, 670, 550, 900

4. 3,4,5,6,5,1,3,4,2

5. Find the lower quartile, upper quartile, median, the lowest and highest number based on the box and whisker plot:

Lower Quartile (Q1):
Upper Quartile (Q2):
Median:
Lowest Number:
Highest Number:

6. Find the lower quartile, upper quartile, median, the lowest and highest number based on the box and whisker plot:



Lower Quartile (Q1):
Upper Quartile (Q2):
Median:
Lowest Number:
Highest Number:

Lesson 4.4 Comparing Data Sets and Review Problems

Statistics is the study of the collection, organization, analysis, interpretation, and presentation of data. It deals with all aspects of this, including the planning of data collection in terms of the design of surveys and experiments (Dodge, Y. (2003) The Oxford Dictionary of Statistical Terms, OUP. ISBN 0-19-920613-9).

In statistics, the term **central tendency** relates to the way in which quantitative data tend to cluster around some value. A measure of central tendency is any of a number of ways of specifying this "central value". In practical statistical analysis, the terms are often used before one has chosen even a preliminary form of analysis: thus an initial objective might be to "choose an appropriate measure of central tendency" (Dodge, Y. (2003) The Oxford Dictionary of Statistical Terms, OUP. ISBN 0-19-920613-9).

In the simplest cases, the measure of central tendency is an average of a set of measurements, the word average being variously construed as mean, median, or other measure of location, depending on the context.

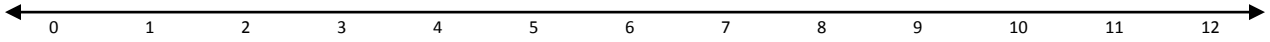
Problems:

1. Create a dot plot of the given data set. Calculate the mean and the median (see lesson 4.3). Determine which measure of center (the mean or the median) best describes each data set.

1,3,2,0,7,2,1,10,1,12,1,2,0,3,4

Mean:

Median:



Is the mean or the median the better measure of center of the data set?

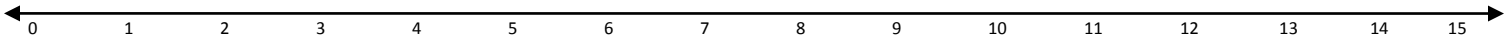
Why? (Talk about which way the data are skewed)

2. Create a dot plot of the given data set. Calculate the mean and the median (see lesson 4.3). Determine which measure of center (the mean or the median) best describes each data set.

7,2,9,9,10,12,17,10,6,11,9,10,8,11,8

Mean:

Median:



Is the mean or the median the better measure of center of the data set?

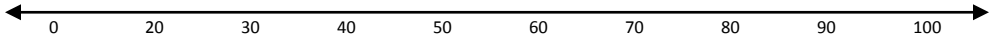
Why? (Talk about which way the data are skewed)

3. Create a dot plot of the given data set. Calculate the mean and the median (see lesson 4.3). Determine which measure of center (the mean or the median) best describes each data set.

50,50,40,70,60,50,20,50,80,40,60,40,50

Mean:

Median:



Is the mean or the median the better measure of center of the data set?

Why? (Talk about which way the data are skewed)

4. Josh and Richard each earn tips at their part-time job. This table shows their earnings from tips for five days.

TOTAL TIPS BY DAY		
Day	Josh's Tips	Richard's Tips
Monday	$40	$40
Tuesday	$20	$45
Wednesday	$36	$53
Thursday	$28	$41
Friday	$31	$28

a. Who had the greatest median earnings from tips? What is the difference in the median of Josh's earnings from tips and the median of Richard's earnings from tips?

b. What is the difference in the interquartile range for Josh's earnings from tips and Richard's earnings from tips?

5. Mr. Schroder, the physical education teacher, measured the height of the students in his first period class. He organized his data in this chart.

Height (in inches)	Frequency
42	1
43	2
44	4
45	5
46	4
47	2
48	1

a. Make a dot plot for the data.

b. Make a histogram for the data.

c. Make a box plot for the data.

d. Does the distribution of heights appear normal/bell shaped?

EOCT Practice Item

6. This table shows the average low temperature, in ˚F, recorded in Macon, GA, and Charlotte, NC, over a six-day period.

Day	1	2	3	4	5	6
Temperature in ˚F in Macon	71	72	66	69	71	73
Temperature in ˚F in Charlotte	69	64	68	74	71	75

Which conclusion can be drawn from the data?

A. The interquartile range of the temperatures is the same for both cities.

B. The lower quartile for the temperatures in Macon is lower than the lower quartile for the temperatures in Charlotte.

C. The mean and median temperatures of Macon were higher than the mean and median temperatures of Charlotte.

D. The upper quartile for the temperatures in Charlotte was lower than the upper quartile for the temperatures in Macon.

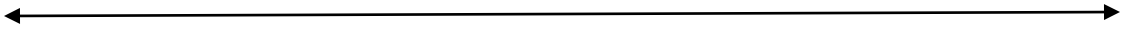
7. Mr. Anzalone finished grading quizzes for one of his Coordinate Algebra classes. The table below shown is the recorded grades of the class.

Student	Grade	Student	Grade	Student	Grade
A	85	J	52	S	98
B	89	K	81	T	53
C	66	L	61	U	62
D	74	M	71	V	55
E	77	N	53	W	64
F	72	O	71	X	62
G	64	P	90	Y	56
H	55	Q	65	Z	87
I	61	R	55		

a. Mr. Anzalone is worried that students may not have understood the material covered on the quiz. He would like to get a better idea of how the class did as a whole. Would you recommend that he make a dot plot, a box-and – whisker plot, or a histogram to display this data? Explain your reasoning.

b. Construct a dot plot and a histogram of the data in the table.

Dot Plot:



Histogram:

c. Describe the distribution of the graphs. What do you notice?

d. What information does the dot plot provide that the histogram does not?

e. The students argue that more than half the students failed the quiz. So they think Mr. Anzalone should let them retake it. A grade of 56 is failing.
Construct a box- and – whisker plot of the data.

f. Describe the distribution of the box-and-whisker plot. Explain what it means in terms of this problem situation.

g. Are the students correct? Explain your reasoning.

REVIEW

8. Math each definition to its corresponding term.

1. Interquartile range (IQR)	**a.** A data value significantly greater than or less than the other values in the data set.
2. Outlier	**b.** A value created by subtracting Q1 from Q3.
3. Median	**c.** The sum of a data set divided by the number of the data.
4. Mean	**d.** The middle number of a data set in ascending order.

9. Describe "Standard Deviation" in your own words.

10. Describe "Range" in your own words.

11. Calculate the mean and standard deviation of the data set. The data are 0, 3, 6, 7, and 9.

Lesson 4.5 Summarize, Represent Data on Two Categorical and Quantitative Variables

KEY IDEAS

There are essentially two types of data: **quantitative** and **categorical**.
Examples of **categorical** data are: color, type of pet, gender, ethnic group, religious affiliation, etc.
Examples of **quantitative** data are: 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.

PROBLEMS:

1. Fill in the correct associations based on the paragraph above:

Examples of categorical data:	**Examples of quantitative data:**
Graphs associated with categorical data :	**Graphs associated with quantitative data:**

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.

Categorial Example: frequencies of gender and club membership for 9th graders

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 from the chart 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 CCGPS Coordinate Algebra chess club. The most popular club is yearbook, with 35 of 135 students in that club.

PROBLEMS:

2. How many female students are in the latin club?

3. How many students are in the chess club?

4. How many male students are in the chorus and the band?

The values in the table can be converted to percents which will give us an idea of the composition of each club by gender. We see that close to 14% of the students are in the chess club, and there are more than five times as many boys as girls.

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 are 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.

PROBLEMS:

5. What is the percentage of female students are in the Chess club?

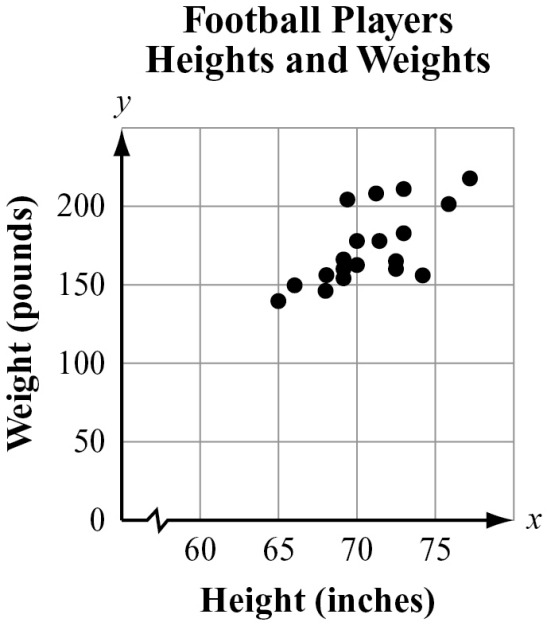
6. What is the percentage of male students in the band?

7. What is the percentage of male and female students in the Latin club?

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 chorus members that are female. The working condition is that the person is female. If 12.6% of the entire school population is females in the chorus, and 42.3% of the student body is female, then 12.6% / 42.3%, or 29.8%, of the females in the school are in the chorus (also 17 of 57 girls).

<u>Quantitative example:</u> <u>Chart of heights and weights of players on a football team</u>

A **scatter plot** is often used to present bivariate quantitative data. Each variable is represented on an axis and the axes are labeled accordingly. Each point represents a player's height and weight. For example, one of the points represents a height of 66 inches and weight of 150 pounds. The scatter plot shows two players standing 70 inches tall because there are two dots above that height.



A **scatter plot** displays data as points on a grid using the associated numbers as coordinates. The way the points are arranged by themselves in a scatter plot may or may not suggest a relationship between the two variables. In the scatter plot about the football players shown earlier, it appears there may be a relationship between height and weight because, as the players get taller, they seem to generally increase in weight; that is, the points are positioned higher as you move to the right. Bivariate data may have an underlying relationship that can be modeled by a mathematical function. For the purposes of this unit we will consider linear models.

Example:
Melissa would like to determine whether there is a relationship between study time and mean test scores. She recorded the mean study time per test and the mean test score for students in three different classes.

This is the data for Class 1.

Class 1 Test Scores	
Mean Study Time (Hours)	**Mean Test Score**
0.5	63
1	67
1.5	72
2	76
2.5	80
3	85
3.5	89

Notice that, for these data, as the mean study time increases, the mean test score increases. It is important to consider the rate of increase when deciding which algebraic model to use. In this case, the mean test score increases by approximately 4 points for each 0.5-hour increase in mean study time. When the rate of increase is close to constant, as it is here, the best model is most likely a linear function.

This next table shows Melissa's data for Class 2.

Class 2 Test Scores	
Mean Study Time (Hours)	**Mean Test Score**
0.5	60
1	61
1.5	63
2	68
2.5	74
3	82
3.5	93

In these data as well, the mean test score increases as the mean study time increases. However, the rate of increase is not constant. The differences between each successive mean test score are 1, 2, 5, 6, 8, and 11. The second differences are 1, 3, 1, 2, and 3. Since the second differences are fairly close to constant, it is likely that a different model known as an exponential function would be employed for the Class 2 data.

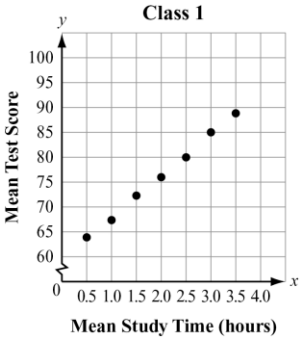
This table shows Melissa's data for Class 3.

Class 3 Test Scores	
Mean Study Time (Hours)	**Mean Test Score**
0.5	71
1	94
1.5	87
2	98
2.5	69
3	78
3.5	91

In these data, as the mean study time increases, there is no consistent pattern in the mean test score. As a result, there does not appear to be any clear relationship between the mean study time and mean test score for this particular class. Often, patterns in bivariate data are more easily seen when the data is plotted on a coordinate grid.

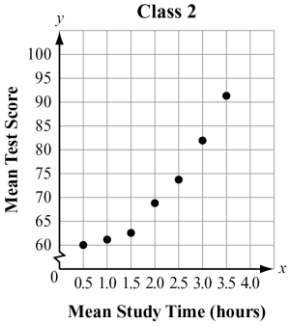
Example:

This graph shows Melissa's data for Class 1:



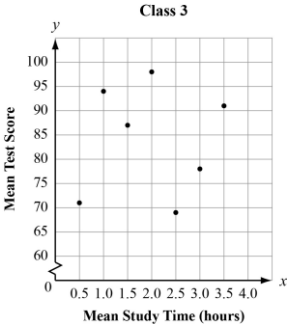
In this graph, the data points are all very close to being on the same line. This is further confirmation that a linear model is appropriate for this class.

This graph shows Melissa's data for Class 2:



In this graph, the data points appear to lie on a curve, rather than on a line, with a rate of increase that increases as the value of x increases. It appears that an exponential model may be more appropriate than a linear model for these data.

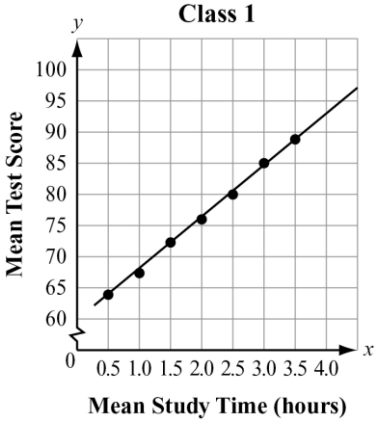
This graph shows Melissa's data for Class 3:



In this graph, the data points do not appear to lie on a line or on a curve. Neither a linear model nor an exponential model is appropriate to represent the data.

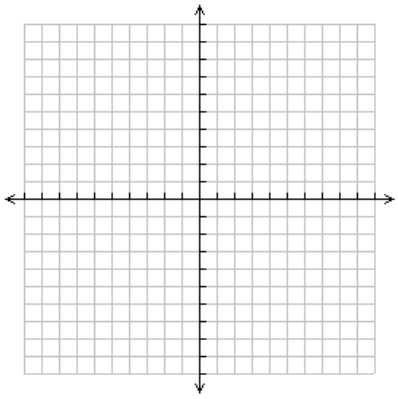
A **line of best fit (trend or regression line)** is a straight line that best represents the data on a scatter plot. This line may pass through some of the points, none of the points, or all of the points. In the previous examples, only the Class 1 scatter plot looks like a linear model would be a good fit for the points. In the other classes, a curved graph would seem to pass through more of the points. For Class 2, perhaps an exponential model would produce a better fitting curved. When a linear model is indicated there are several ways to find a function that approximates the y-value for any given x-value. A method called regression is the best way to find a line of best fit, but it requires extensive computations and is generally done on a computer or graphing calculator.

Example:
This graph shows Melissa's data for Class 1 with the line of best fit added. The equation of the line is y = 8.8x + 58.4.

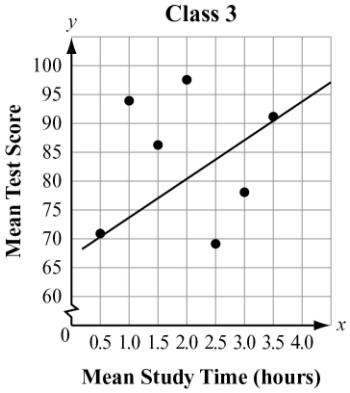


Notice that five of the seven data points are on the line. This represents a very strong positive relationship for study time and test scores, since the line of best fit is positive and a very tight fit to the data points.

PROBLEMS:
8. Draw the line of best fit for the following data: (0,1), (2,3), (3,3), (3,4), (5,5), (6,8), (7,5), (8,8), (9,7).



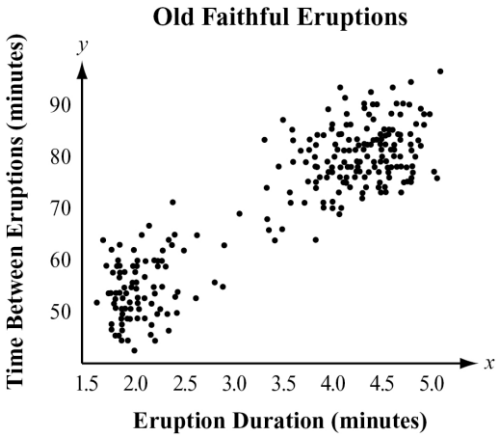
This next graph shows Melissa's data for Class 3 with the line of best fit added. The equation of the line is y = 0.8 x+ 83.1.



Although a line of best fit can be calculated for this set of data, notice that most of the data points are not very close to the line. In this case, although there is some correlation between study time and test scores, the amount of correlation is very small.

PROBLEMS:

9. Barbara is considering visiting Yellowstone National Park. She has heard about Old Faithful,the geyser, and she wants to make sure she sees it erupt. At one time it erupted just about every hour. That is not the case today. The time between eruptions varies. Barbara went on the Web and found a scatter plot of how long an eruption lasted compared to the wait time between eruptions. She learned that, in general, the longer the wait time, the longer the eruption lasts. The eruptions take place anywhere from 45 minutes to 125 minutes apart. They currently average 90 minutes apart.



a. For an eruption that lasts 4 minutes, about how long would the wait time be for the next eruption?

b. What is the shortest duration time for an eruption?

c. Do you think the scatter plot could be modeled with a linear function? Why or why not?

Example:

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.

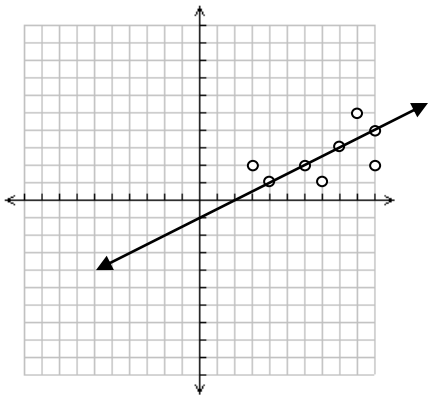
Beverage Can Recycling								
Number of Canned Beverages Sold	10	8	7	3	6	10	9	4
Number of Cans Recycled	4	3	1	2	2	2	5	1

Find an equation of a line of best fit for the data.

Solution:

1. First lot the ordered pairs:

2. Then draw a straight line which Represent the data points.



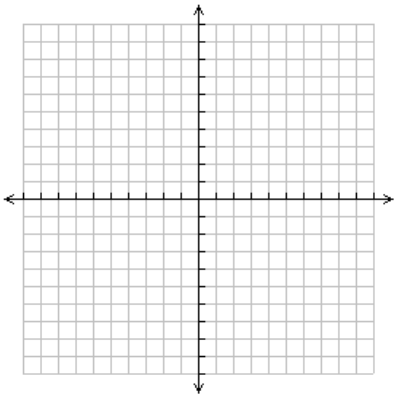
3. Find the equation of the line:
   The y-intercept of the line is -1.
   The slope of the line is ½. (Slope is rise/run; you rise 1 unit and run 2 units).
   Therefore, the equation of the line is $y = \frac{1}{2}x - 1$.

PROBLEMS:

10. Mr. Landolt likes to go fishing. He is interested in the relationship of hours spent fishing and numbers of fish caught. He collected the following data in 2011.

Fish Caught During the 2011 Season								
Number of Hours Spent Fishing	1	5	4	3	1	2	3	2
Number of Fish Caught	4	5	6	4	3	2	4	1

Find an equation of a line of best fit for the data.



11. A fast food restaurant wants to determine if the season of the year affects the choice of softdrink size purchased. They surveyed 278 customers and the table below shows their results. The drink sizes were small, medium, large, and jumbo. The seasons of the year were spring, summer, and fall. In the body of the table, the cells list the number of customers that fit both row and column titles. On the bottom and in the right margin are the totals.

	Spring	**Summer**	**Fall**	**Totals**
Small	24	22	18	64
Medium	23	28	19	70
Large	18	27	29	74
Jumbo	16	21	33	70
TOTALS	81	98	99	278

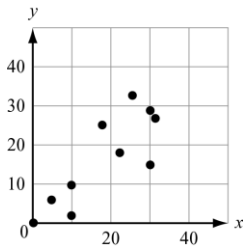
a. In which season did the most customers prefer jumbo drinks?

b. What percent of those surveyed purchased the small drinks?

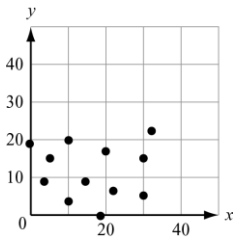
c. What percent of those surveyed purchased medium drinks in the summer?

d. What do you think the fast-food restaurant learned from their survey?

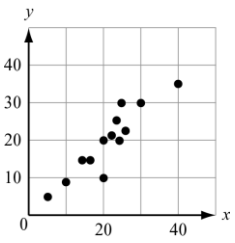
12. Which graph displays a set of data for which a linear function is the model of best fit?



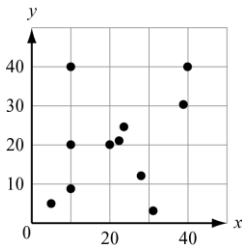
A



B

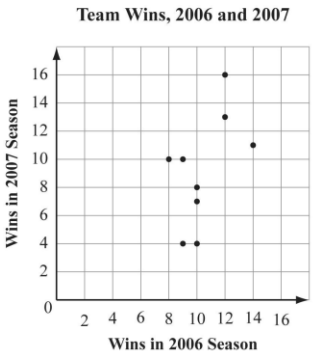


C



D

13. This graph plots the number of wins in the 2006 and 2007 seasons for a sample of professional football teams.



Which equation BEST represents a line that matches the trend of this data?

A. $y = \frac{1}{2}x$ B. $y = \frac{1}{2}x + 8$ C. $y = 2x - 6$ D. $y = 2x - 12$

Lesson 4.6 Interpret Linear Models

Once a model for the scatter plot is determined, its goodness of fit is very important. The goodness of fit depends on the model's accuracy in predicting values. **Residuals**, or error distances, are used to measure the goodness of fit. A residual is the difference between the observed value and the model's predicted value.
For a regression model, a **residual = observed value – predicted value**.
A residual plot is a graph that shows the residual values on the vertical axis and the independent variable (x) on the horizontal axis. A residual plot shows where the model fits best, and where the fit is worst. A good regression fit has very short residuals.

Example:
Take the data from the test scores for Class 1 used in the last section. The observed mean test scores were 63, 67, 72, etc. The best fit model was a linear model with the equation $y = 8.8x + 58.4$. We can calculate the residuals for this data and consider the fit of the regression line.

Mean Study Time (hours)	Mean Test Score	Predicted Score $y = 8.8x + 58.4$	Residual
0.5	63	62.8	0.2
1.0	67	67.2	-0.2
1.5	72	71.6	0.4
2.0	76	76	0
2.5	80	80.4	-0.4
3.0	85	84.8	0.2
3.5	89	89.2	-0.2

Notice the numbers in the residual column tell us how far the predicted mean test score was from the observed, as seen in the regression scatter plot for Class 1. The regression passes through one of the actual points in the plot of the points where the residual is 0. Notice also, the residuals add up to 0. Residuals add up to 0 for a properly calculated regression line. The goal is to minimize all of the residuals.

PROBLEMS:

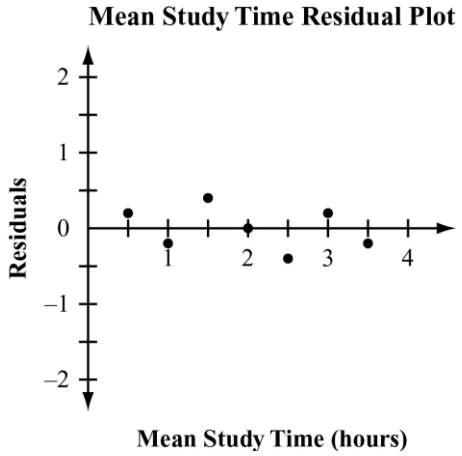
1. Based on the table above and the regression equation $y = 8.8x + 58.4$, what score would you expect if a student studied

a. 0 hours

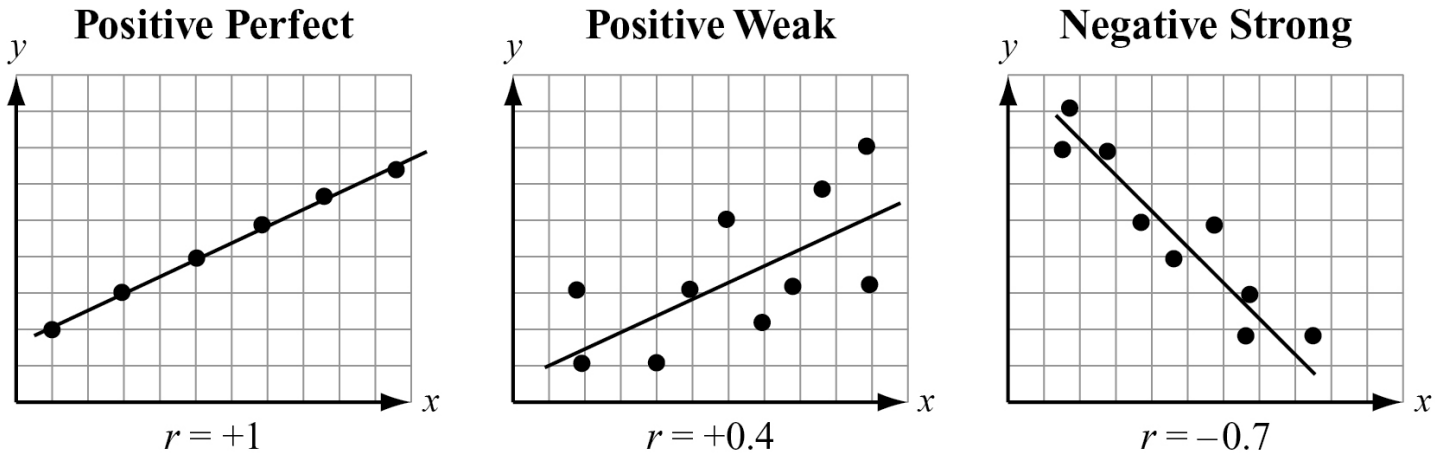
b. 4 hours

c. 2.75 hours

The table of the previous page (first column vs. last column) is plotted below. The regression passes through one of the actual points in the plot of the points where the residual is 0.



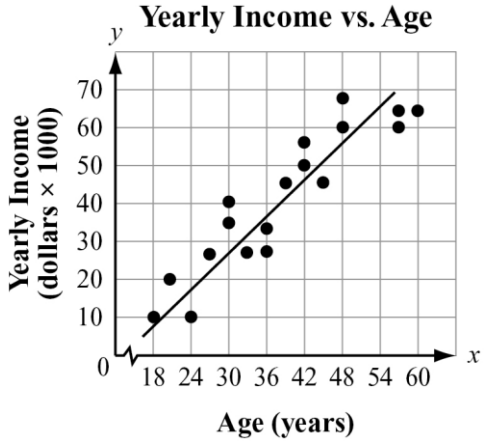
A **correlation coefficient** is a measure of the strength of the linear relationship between two variables. It also indicates whether the dependent variable, y, grows along with x, or y get smaller as x increases. The correlation coefficient is a number between –1 and + 1 including –1 and +1. The letter r is usually used for the correlation coefficient. When the correlation is positive, the line of best fit will have positive slope and both variables are growing. However, if the correlation coefficient is negative, the line of best fit has negative slope and the dependent variable is decreasing. The numerical value is an indicator of how closely the data points come to the line.



The correlation between two variables is related to the slope and the goodness of the fit of a regression line. However, data in scatter plots can have the same regression lines and very different correlations. The correlation's sign will be the same as the slope of the regression line. The correlation's value depends on the dispersion of the data points and their proximity to the line of best fit.

PROBLEMS:

2. This scatter plot suggests a relationship between the variables age and income. Answer the questions below based on the pictured scatter plot.



a. What type of a relationship is suggested by the scatter plot (positive/negative, weak/strong)?

b. What is the domain of ages considered by the researchers?

c. What is the range of incomes?

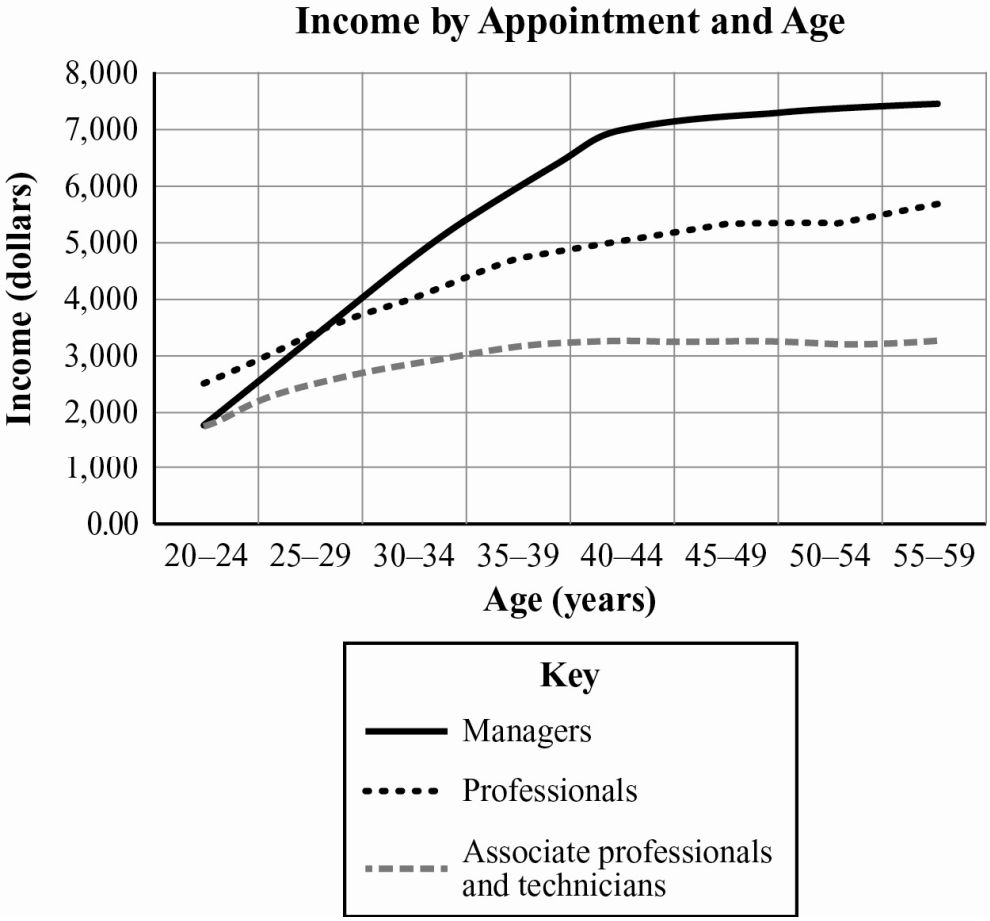
d. Do you think age causes income level to increase? Why or why not?

e. Based on the graph, what is the yearly income when you are 30 years old?

f. Based on the graph, what is the yearly income when you are 54 years old?

g. Can you assume that the line can be extrapolated (extended so that you can estimate the incomes for ages 60 and older)? Why or why not?

3. A group of researchers looked at income and age in Singapore. Their results are shown below. They used line graphs instead of scatter plots so they could consider the type of occupation of the wage earner.



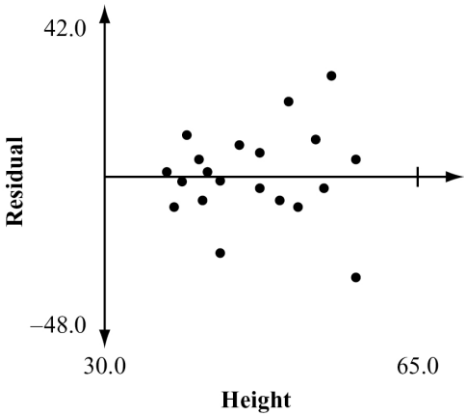
a. Does there appear to be a relationship between age and income?

b. Do all three types of employees appear to share the same benefit of aging when it comes to income?

c. Does a linear model appear to fit the data for any of the employee types?

d. Does the effect of age vary over a person's lifetime?

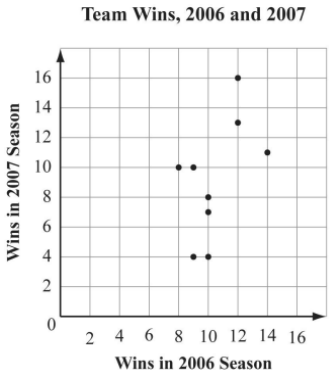
4. Consider the residual plot below. Each vertical segment represents the difference between an observed weight and a predicted weight of a person, based on height.



a. Do you think the regression line is a good predictor of weight?

b. Why do the residuals appear to be getting longer for greater heights?

5. This graph plots the number of wins in the 2006 and 2007 seasons for a sample of professional football teams.

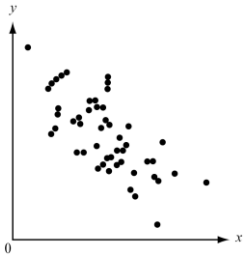


Based on the regression model, what is the predicted number of 2007 wins for a team that won 5 games in 2006?

A. 3
B. 5
C. 6
D. 7

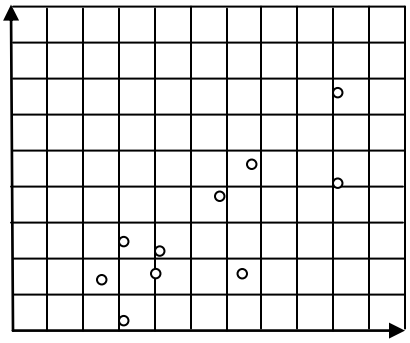
State your reason why you picked the answer:

6. How would you describe the correlation of the two variables based on the scatter plot?

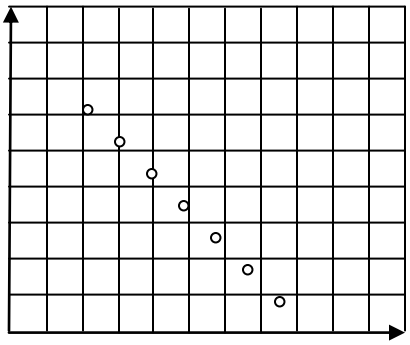


A. positive, strong linear
B. negative, weak linear
C. negative, fairly strong linear
D. little or no correlation

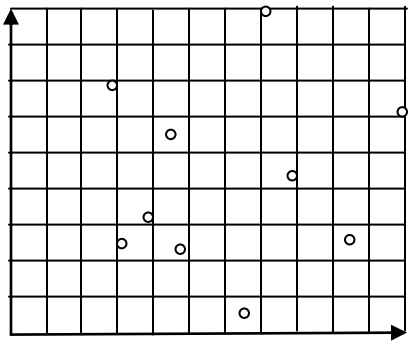
7. Decide whether the there is weak, strong or no correlation to the data. Then circle whether the correlation is positive or negative.



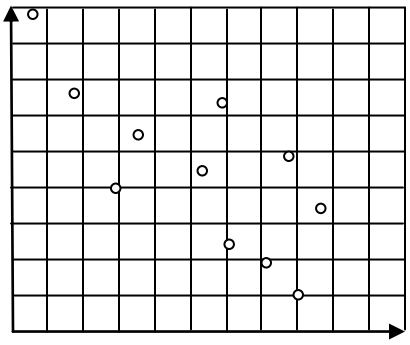
Correlation:
Pos Neg



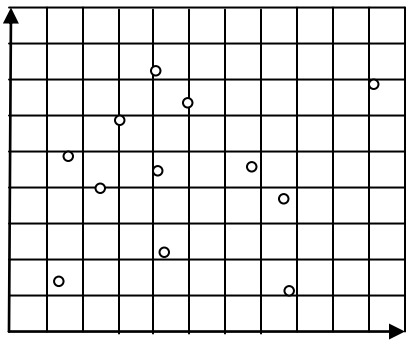
Correlation:
Pos Neg



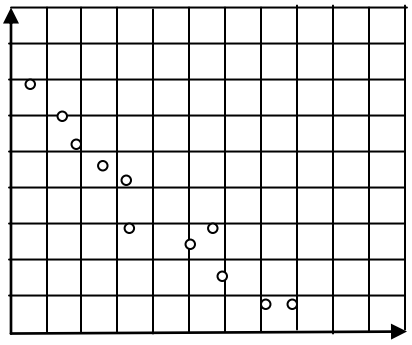
Correlation:
Pos Neg



Correlation:
Pos Neg



Correlation:
Pos Neg



Correlation:
Pos Neg