Integrated Math 1 Honors
Unit 9: Statistics
9.0

Name: ______________________

Date: __________ Period: _____

Part I Objective: To classify categorical vs. quantitative data.

Types of Data

Categorical Data: Represents types of data that can be divided into groups.

(Examples: hair color, make of car, state of birth, grade level classified by freshmen, sophomore, etc.)

Quantitative Data: Represents a certain quantity, amount, or range.

Discrete: Data based on counts - the values cannot be subdivided meaningfuly

(Examples: number of oranges on an orange tree, number of cars entering a college campus, grade level classified by 9, 10, 11, 12, distinctive points on a number line)

Continuous: Data that can be measured - it has an infinite number of possible values within a selected range.

(Examples: time until a light bulb burns out, height and weight, an interval on a number line)

__Example 1:__ Determine whether the following are categorical or quantitative? If they are quantitative, label them as discrete or continuous.

a.) Blood Type Categorical

b.) Household size (the number of people that live in a house) quantitative - discrete.

c.) Height of a waterfall quantitative - continuous

d.) Population of America quantitative - discrete

e.) Length of a movie quantitative - continuous

f.) Number of correct answers on a 100 question exam quantitative - discrete g.) Numbers from 0-10 quantitative - continuous

h.) Distance from Antioch to Skokie quantitative - continuous

i.) Number of dollar bills you have in your wallet quantitative - discrete

j.) Grade you are receiving in a class Categorical (A, B, C, ...) OR quantitative - continuous (89%, 67.3%, ...)

Part I Objective: To create dot plots.

Graphical Displays

For quantitative data, there are four ways that we will be focusing on to present data. They are:

- Dot Plots
- Histograms
- Box Plots
- Stem and Leaf Displays

- evenly spaced
- Same size dots
- one dot for each data value

Example 2: Create a dot plot that represents our class data.

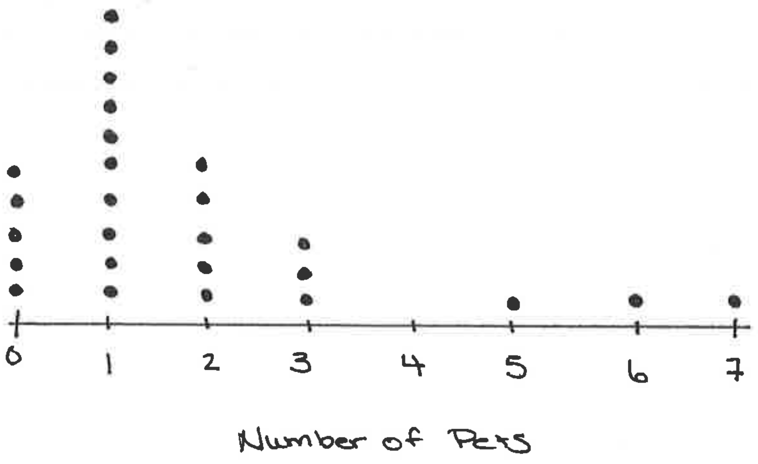
a. Number of Pets

0, 1, 1, 1, 2, 3, 7, 0, 1, 1, 5, 3, 2, 0, 0, 1, 1, 2, 2, 2, 1, 3, 1, 6, 0, 1

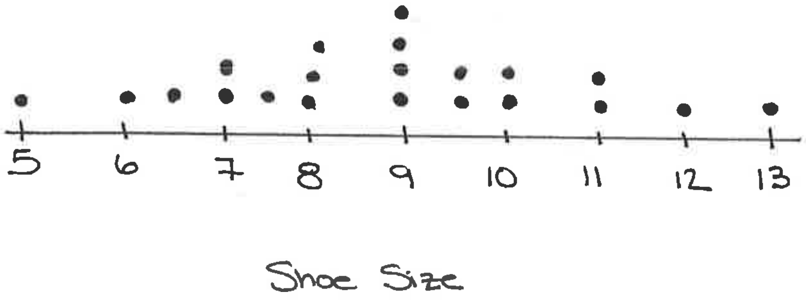
b. Shoe Size

7, 6.5, 8, 9, 9.5, 7.5, 12, 10, 11, 7, 9, 13, 6, 9.5, 5, 10, 11, 9, 8, 8, 9

Number of Pets Students in Mrs. Gusler's 3rd Hour Class Have



Shoe Sizes of Students in Mrs. Gusler's 5th Hour Class



Integrated Math 1 Honors
Unit 9: Statistics
9.1

Name: ____________________

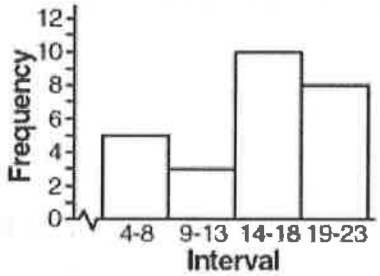
Date: __________ Period: ______

Objective: To use histograms to display data.

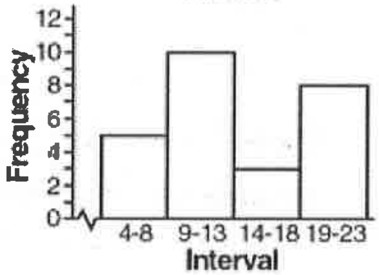
Predict: Which of the following histograms represents the data shown in the table below?

Interval	Frequency
4-8	8
9-13	3
14-18	10
19-23	5

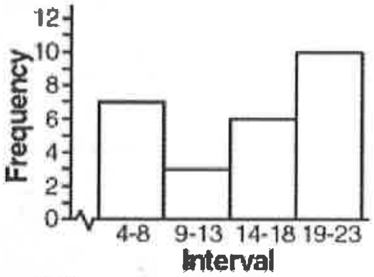
A)



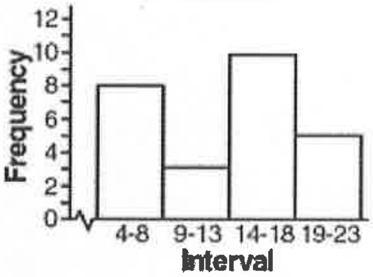
B)



C)



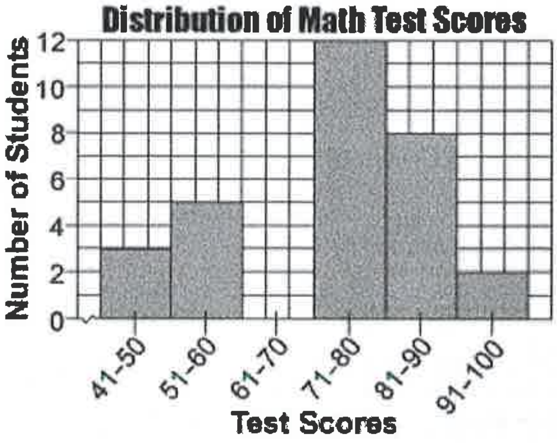
[✓ D)]



How is a histogram different from a bar graph?

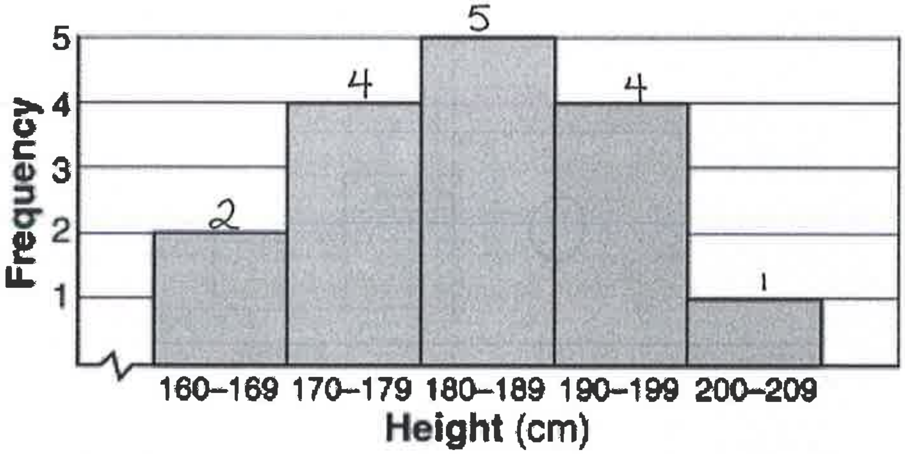
- Histograms are for quantitative data, bar graphs display categorical data!
- No space between bars on a histogram, bar graphs have spaces / gaps between bars
- Histograms plot intervals or ranges of data grouped together along the x-axis
- Histograms can have breaks (∿) so that they don't have to start at zero

Example 1: The graph below shows the distribution of scores of 30 students on a mathematics test. Complete the frequency table above using the data in the frequency histogram shown.



Test Scores	Frequency
91–100	2
81-90	8
71-80	12
61-70	0
51-60	5
41-50	3

Example 2: The accompanying histogram shows the heights of the students in Kyra's health class. What is the total number of students that are in her class?



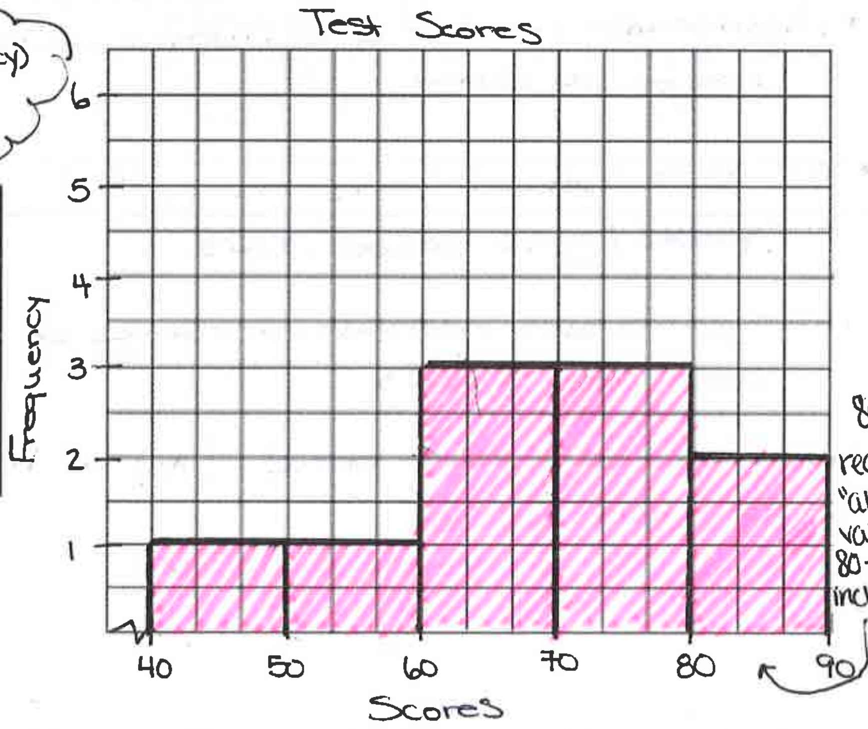
2+4+ 5+ 4 +1 = 16

There are 16 total students in Kyra's health class.

Example 3: The scores on a test were 70, 55, 61, 80, 72, 65, 40, 74, 68, and 84. Complete the accompanying table and use the table to construct a frequency histogram for these scores.

- must have a table and labeled axes (y-axis will always be frequency)
- evenly spaced
- bars touching
- intervals along x-axis

Score	Tally	Frequency
40-49	|	
50-59	|	
60-69	|||	
70-79	|||	
80-89	||	



80≤x<90 read as "all data values between 80-90, not including 90"

Integrated Math 1 Honors
Unit 9: Statistics
9.2

Name: ____________________

Date: __________ Period: _____

Objective: To create boxplots that represent a data set by hand and using a graphing calculator.

Warm up: Calculate the mean, median, and mode for the following data set: 5, 7, 12, 13, 11, 16, 4, 9, 10, 7, 8, 14

Mean: $\frac{5+7+12+13+11+16+4+9+10+7+8+14}{12} = 9.67$

Median: $\cancel{4}, \cancel{5}, \cancel{7}, \cancel{7}, \cancel{8}, \boxed{9, 10}, \cancel{11}, \cancel{12}, \cancel{13}, \cancel{14}, \cancel{16} = 9.5$

Mode: 7

Vocabulary

Measures of Central Tendency (Measures of Center):

Mean: the "average" – calculated by adding up all the data scores and then dividing by the total number of values

Median: the "middle" – calculated when data values are arranged least to greatest
if two values fall in the middle, take their average

Mode: the "most" – the score(s) ~~that~~ occur the most frequently.
* can have multiple modes or no mode.*

Measures of Spread:

Range: maximum – minimum

Interquartile Range (IQR): $Q_3 - Q_1$

Standard Deviation: *will visit later*

Example 1: For each of the following data sets, calculate the mean, median, range, and IQR by hand.

a. 4, 7, 7, 7, 10

Mean: 7

Median: 7

Range: 10-4=6

IQR: 8.5 - 5.5 = 3

$Q_1 = 5.5$

$Q_3 = 8.5$

b. 48, 56, 58, 60, 62, 70

Mean: 59

Median: 59

Range: 70 - 48 = 22

IQR: 62 - 56 = 6

$Q_1 = 56$

$Q_3 = 62$

Example 2: Use a graphing calculator to determine measures of center and spread.

Calculator Steps:

STAT → Edit

Enter the data into L1

STAT → Calc → 1-Var Stats

Mean: $\overline{X}$ Median: Med

Range: max-min IQR: $Q_3 - Q_1$

70	65	52	58
42	49	32	32
33	39	39	22
29	3	9	9

Creating Boxplots!

A boxplot is created using the five number summary.

The box is made up of the Q_1 (first quartile), median, and Q_3 (third quartile)

The whiskers are located at the maximum and minimum.

Let's create a boxplot out of the data below.

75	86	100
62	73	88
89	91	54
92	87	86
81	80	94
77	75	64
66	70	81

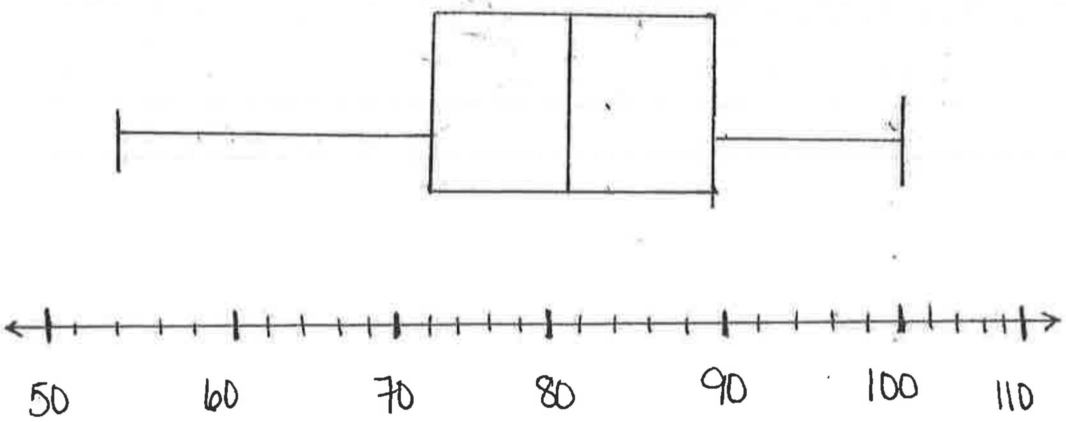
min: 54

Q_1: 71.5

med: 81

Q_3: 88.5

max: 100



We learned how to use the calculator to find mean, median, range, and IQR. We can also use it to create histograms and boxplots. First, we will need to enter the data into the list in the calculator exactly as we did before. Here are the steps to create these graphs:

Creating Graphs on the Calculator:

2ND → Y= (STAT PLOT)

Press enter on Plot #1 and then be sure to select "ON"

Type: You should see an icon that looks like a histogram (last graph in top row) and two icons that look like boxplots (first and second in bottom row)

The first boxplot icon is called a modified boxplot and will be the one that we will use

Xlist: Enter the list that you entered the data into (usually L1)

Freq: 1

Zoom → ZoomStat (#9) will make it so the entire graph is visible on the screen

Try it for the data set above!!!

Other calculator information:

Histogram:

Window → xscl (changes the bin width)

Box Plot (Box and Whisker):

Trace → use right and left arrows to see 5 number summary
(Min, Q_1, Med, Q_3, Max)

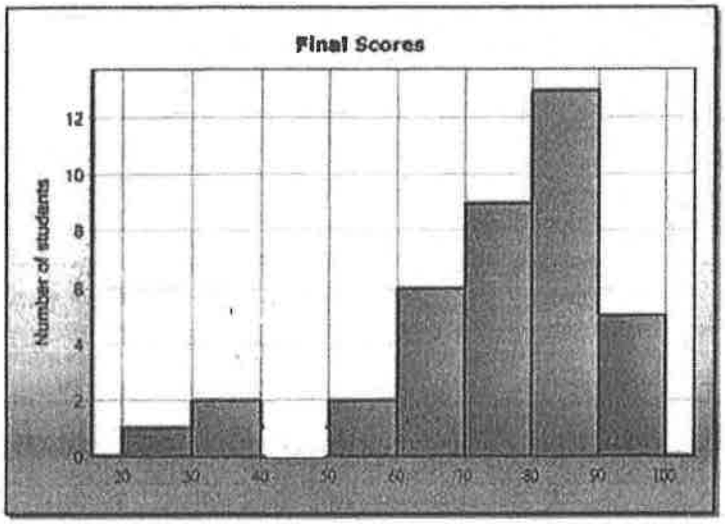
Integrated Math 1 Honors
Unit 9: Statistics
9.3

Name: ______________________

Date: ________ Period: _____

Part I Objective: To compare two data sets using double box plots.

Warm up: Use the histogram to answer the following questions about the data.



a. What is the range of the bin with the largest frequency?

80-90 or $80 \leq x < 90$

b. Which bin(s) has the smallest frequency?

0-20 $0 \leq x < 20$
40-50 or $40 \leq x < 50$

c. How many students scored between 70 and 90?

9+13 = 22 students

d. In what range did five students score in?

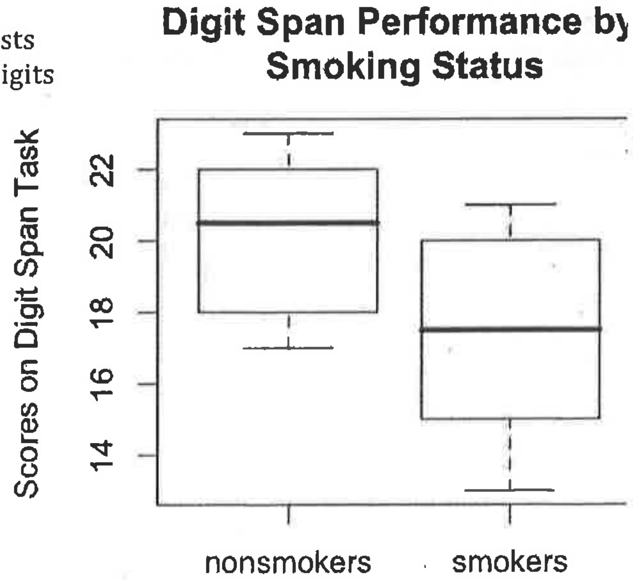
90-100 or $90 < x \leq 100$

e. How many students received an "A" on their final exam?

5 students

When comparing multiple data sets, boxplots are the easiest graphs to use. They can be drawn stacked using the same number line and scale. Histograms and dot plots would be difficult to draw using this stacked method, so we will stick with boxplots.

Example 1: On the Wechsler Adult Intelligence test, one of the subtests is called the digit span task. The score represents the longest list of digits that a person can repeat back in correct order immediately after presentation. Use the double boxplot below to answer the questions that follow.



a. Which group appears to score better on the digit span task? Explain.

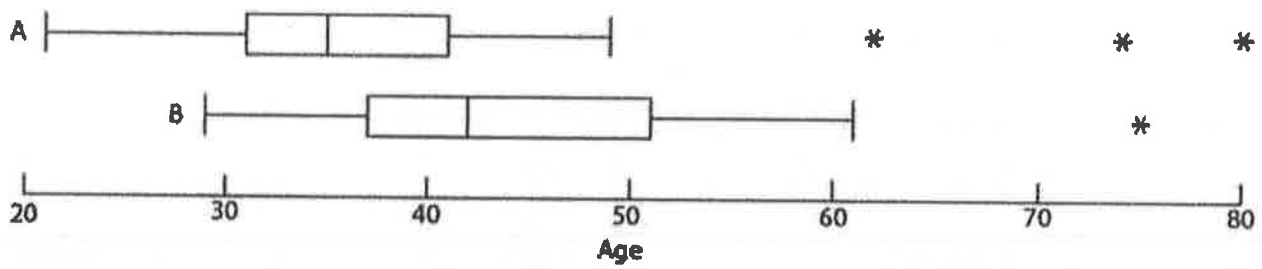
Nonsmokers - every score in the five number summary was higher, indicating that they can recall more digits.

b. Which data set appears to have a larger spread? Explain.

Smokers- range is bigger and IQR is bigger

Example 2: Use the double boxplot below to answer the following questions.

Ages of Oscar Winning Actors from 1975 to 2004



"A" denotes female actors.
"B" denotes male actors.

a. What do you think the stars represent?

Data that doesn't fit the rest – we call these "outliers"

(We'll discuss this more later)

b. What does the double boxplot suggest about the ages of Oscar winning male actors?

Male actors tend to win more Oscars as they get further into their careers

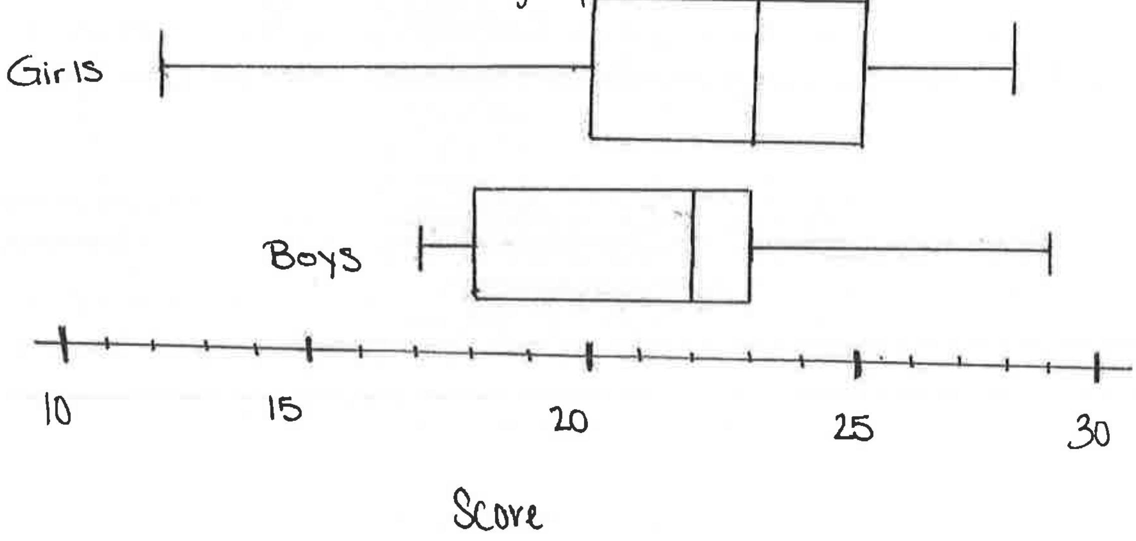
Example 3: The following data represents results from an agility test performed by fourth graders separated by gender. Draw a double box plot to compare the data.

Boys	22	17	18	29	22	22	23	24	23	17	21			
Girls	25	20	12	19	28	24	22	21	25	26	25	16	27	22

Boys
min: 17
Q₁: 18
med: 22
Q₃: 23
max: 29

Girls
min: 12
Q₁: 20
med: 23
Q₃: 25
max: 28

Fourth Grade Agility Test Scores



How do these fourth graders compare in terms of agility? Consider center, shape, and spread in your analysis.

Girls have a larger range of agility levels.

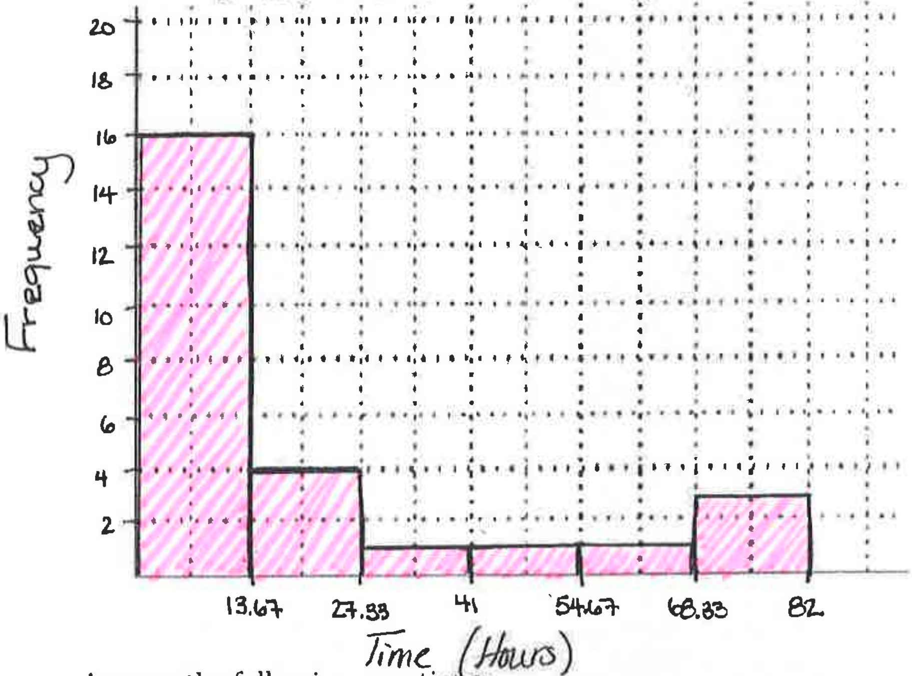
Boys had a higher maximum score.

Girls had a higher median score.

Part II Objective: To analyze and create graphical displays from a set of data using a graphing calculator.

Example 4: Use your graphing calculator to make a histogram regarding how many hours a week our class listens to music.

Number of Hours Spent Listening to Music per Week
for Mrs. Gusler's 3rd Hour Class



Class Results:

4 0 6 8 10
9 2 50 1 2
3 65 72 32
6 15 18 10
20 5 10 8
82 70 23 3

Answer the following questions:

a. Which bin has the largest frequency?

0-13.67

b. Draw a conclusion about the data you found.

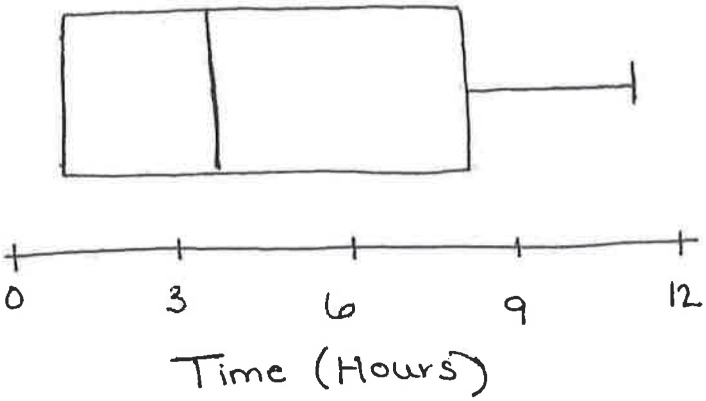
A lot of students listen to music between
0-13.67 hours per week.
Nobody listens to more than 82 hours of music per week.

Example 5: Use your graphing calculator to make a boxplot regarding how many hours a week our class watches television.

Class Results:

7, 1, 1.5, 8, 10, 1, 1, 2, 8,
5, 11, 1, 4, 3, 4, 1, 10

Amount of Hours Spent Watching TV
for Mrs. Guslar's 5th Hour Class



Integrated Math 1 Honors
Unit 9: Statistics
9.4

Name: ____________________________

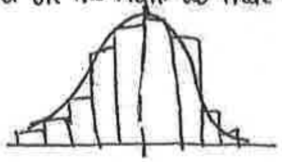
Date: __________ Period: _____

Objective: To classify the shape of a data set and choose an appropriate model.

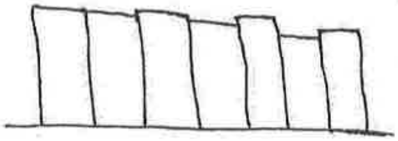
Vocabulary:

Classifying the Shape of a Distribution:

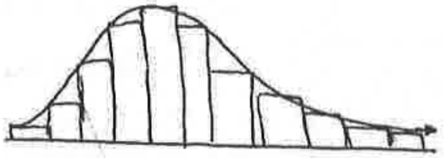
Symmetric: A graph that has a vertical line of symmetry. Approximately the Same amount of data on the right as there is on the left.



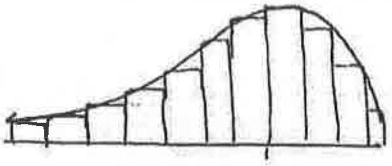
Uniform: Symmetric and all values are roughly the same.



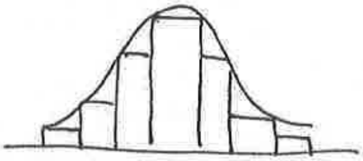
Skewed Right: Tail is to the right of the graph.



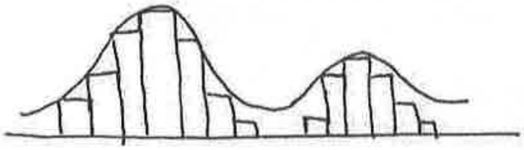
Skewed Left: Tail is to the left of the graph.



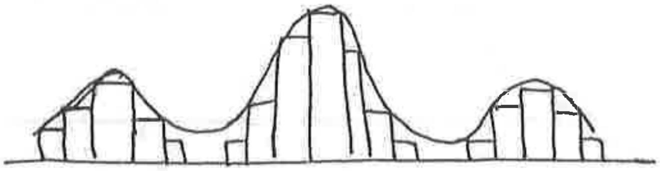
Unimodal: One Peak



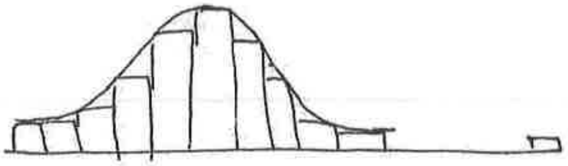
Bimodal: Two peaks.



Multimodal: Three or more peaks.

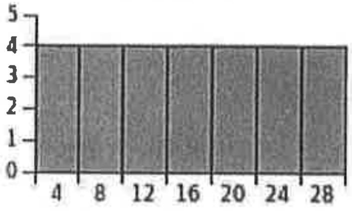


Outliers: Extreme data that does not match or fit with the rest of the data.



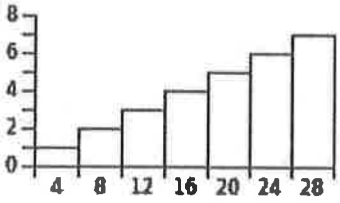
Example 1: Describe the shape of each distribution. Then, match the histogram to its corresponding boxplot.

1.



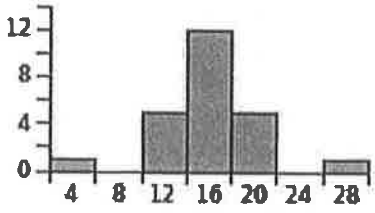
B- Uniform

2.



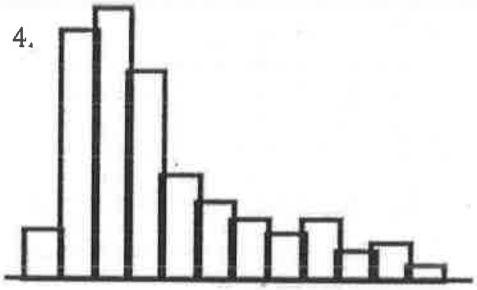
D- Skewed left

3.



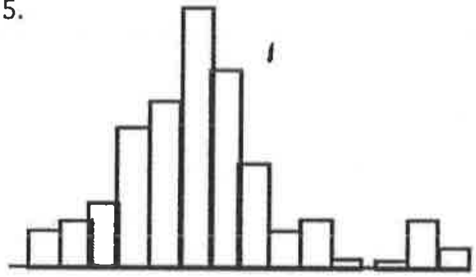
F - Symmetric, unimodal outliers

4.



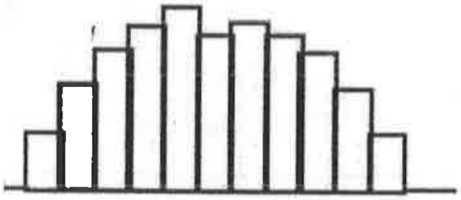
C - Skewed right

5.



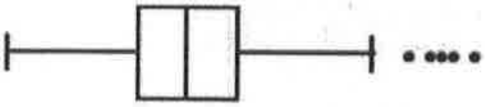
A - Unimodal w/ outliers OR bimodal

6.

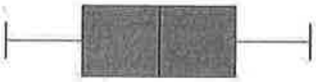


E - Symmetric, unimodal

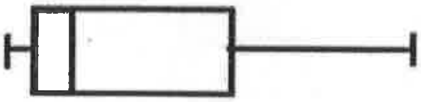
A)



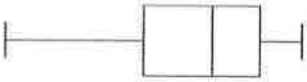
B)



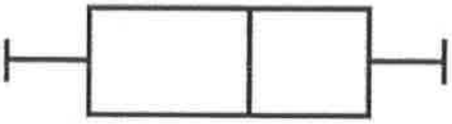
C)



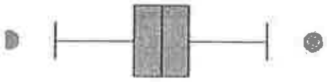
D)



E)



F)



Example 2: The salaries for 15 NFL players from the 2013-2014 Superbowl champs, the Seattle Seahawks, are listed in the table.

a) Do you think the data is symmetric, [✓ skewed right] or skewed left? Explain.

5 players have salaries in the millions (two are close to 10 million!). These values create a tail on the higher end.

b) Based on your answer in part a), what measures of center and spread would be appropriate to report?

Median and IQR

Sidney Rice	$9.7 million
Marshawn Lynch	$8.5 million
Percy Harvin	$4.9 million
Michael Bennett	$4.8 million
Cliff Avril	$3.75 million
Golden Tate	$880,000
Russell Wilson	$681,085
Steven Hauschka	$620,000
Richard Sherman	$600,606
Doug Baldwin	$560,833
Jeron Johnson	$560,000
Jermaine Kearse	$480,000
Spencer Ware	$360,535
Michael Robinson	$326,470
Korey Toomer	$298,059

c) Calculate the measures of center and spread that you reported in part b).

Median = $620,000

IQR = $4,320,000

Example 3: The average temperature for 14 US states are listed in the table.

a) Do you think the data is [✓ symmetric] skewed right, or skewed left? Explain.

Alaska and Florida are extreme values so they offset each other.

The rest of the data is fairly evenly distributed.

b) Based on your answer in part a), what measures of center and spread would be appropriate to report?

Mean and Standard Deviation

Alaska	26.6
North Dakota	40.4
Minnesota	41.2
Wyoming	42
Utah	48.4
Ohio	50.7
Indiana	51.7
New Mexico	53.4
Maryland	54.2
Kansas	54.3
Georgia	63.5
Texas	64.8
Louisiana	66.4
Florida	70.7

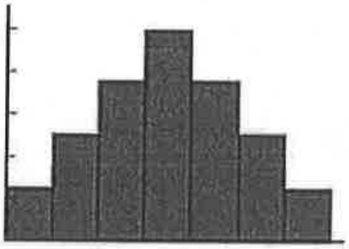
c) Calculate the measures of center and spread that you reported in part b).

Mean = 52° F

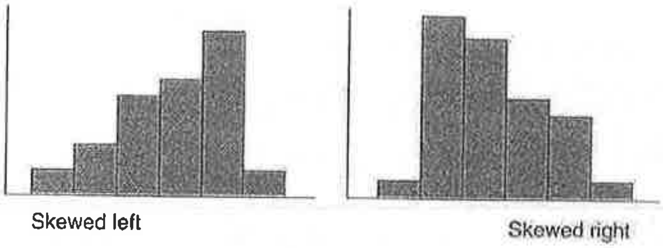
Standard Deviation = 11.6° F

Determining which data representation is the most appropriate

Symmetric



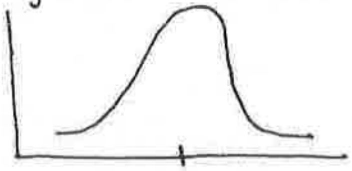
Skewed



Center:

What do you know about the mean and median of a symmetric graph?

they should be the same!



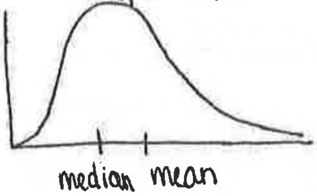
mean=median

What do you know about the mean and median of a skewed graph?

*mean gets "pulled" towards the extreme values

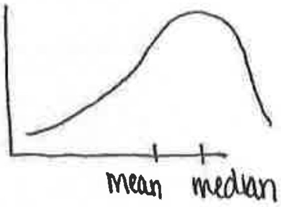
they're different!

Right Skew



median < mean

Left Skew



median > mean

Spread:

Range: The usefulness is limited because it is affected by outliers and skewed data making it a measure that is not typically reported in statistics.

IQR: Much less affected by outliers and skewed data making it an appropriate measure of spread for these types of data

Standard Deviation: Affected by extreme scores, which can be caused by outliers and/or skewed data, so it is only appropriate to use with symmetric data

In summary:

If data is **symmetric**, report:	If data is **skewed** or has **outliers**, report:
Center: mean	**Center:** median
Spread: Standard deviation	**Spread:** IQR

Integrated Math 1 Honors
Unit 9: Statistics
9.5

Name: ______________________________

Date: __________ Period: _____

Objective: To identify outliers of a data set.

Calculating Outliers:

To find outliers, we use a simple formula based on the components of the five-number summary.

$$\textit{Upper Fence (or Upper Limit)} = Q_3 + 1.5(IQR)$$

$$\textit{Lower Fence (or Lower Limit)} = Q_1 - 1.5(IQR)$$

Any data point(s) that fall outside of these values is considered an outlier.

Examples: Check for outliers in the following data sets. Check your answers by creating a modified boxplot in your calculator.

a. Temperatures in Chicago for 10 days in February

6	4
8	12
15	15
15	20
20	45
22	19

Q_1: 10

Q_3: 20

IQR= 20-10 = 10

Upper Fence: 20+1.5(10) = 35

Lower Fence: 10-1.5(10) = -5

Outliers? Yes, 45

b. Final Exam Scores

68	70	73	80	80
81	82	85	84	86
88	83	82	80	81
90	91	94	99	

Q_1: 80

Q_3: 88

IQR: 88-80 = 8

Upper Fence: 88+1.5(8) = 100

Lower Fence: 80-1.5(8) = 68

Outliers? None.

Integrated Math 1 Honors
Unit 9: Statistics
9.6

Name: ______________________

Date: ____________ Period: ______

Objective: To compare and contrast what the different graphical representations display.

Warm up: The following data represents the numbers of points scored by the winning team in the Super Bowl for the last 10 years: 43, 34, 21, 31, 31, 27, 17, 29, 21, 24. Determine the indicated statistical measures.

a. Mean: 27.8

b. Median: 28

c. Mode: 21 and 31

d. Range: 43-17=26

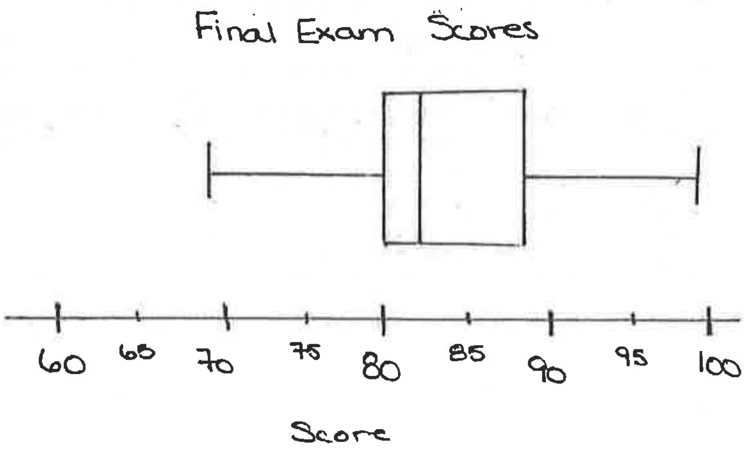
Example 1: Given the following data; find the five number summary and create a box plot.

Final Exam Scores:

68	70	73	80	80
81	82	85	84	86
88	83	82	80	81
90	91	94	99	

Five Number Summary

min: 68
Q_1: 80
med: 82
Q_3: 88
max: 99



Measures of Center:
Find the mean, median, OR mode – your choice! (Hint: which can you easily spot from the box plot?)

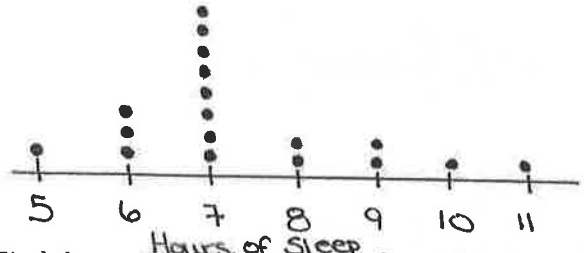
Median = 82

* already calculated to create the box plot *

Example 2: Given the following data; create a dot-plot.

Below are the results of a survey asking students how many hours of sleep they got the previous night:

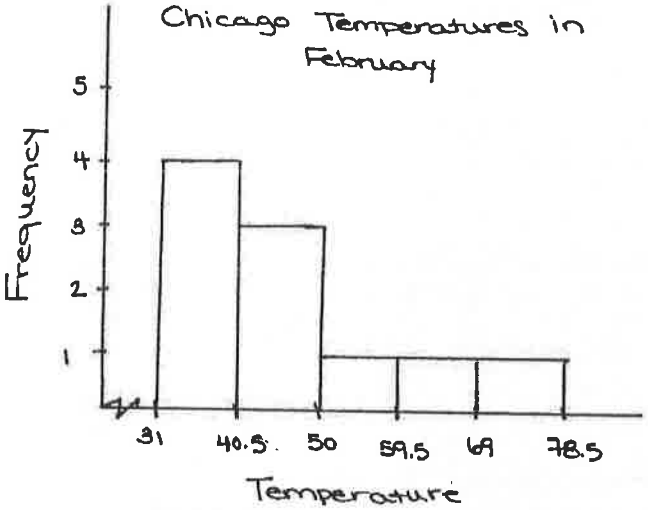
9	6	10	8	6	7	11	7	6	9	7	7	5	7	8	7	7	7



Find the mean, median, OR mode – your choice! Mode = 7

* easy to read off the graph just look for the tallest one *

Example 3: I gathered data on the temperature in Chicago during 10 days in February: 31, 31, 61, 37, 39, 46, 45, 69, 41, 53. Given this data; create a histogram.

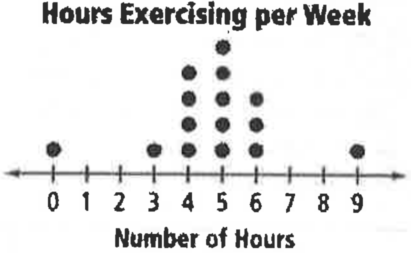


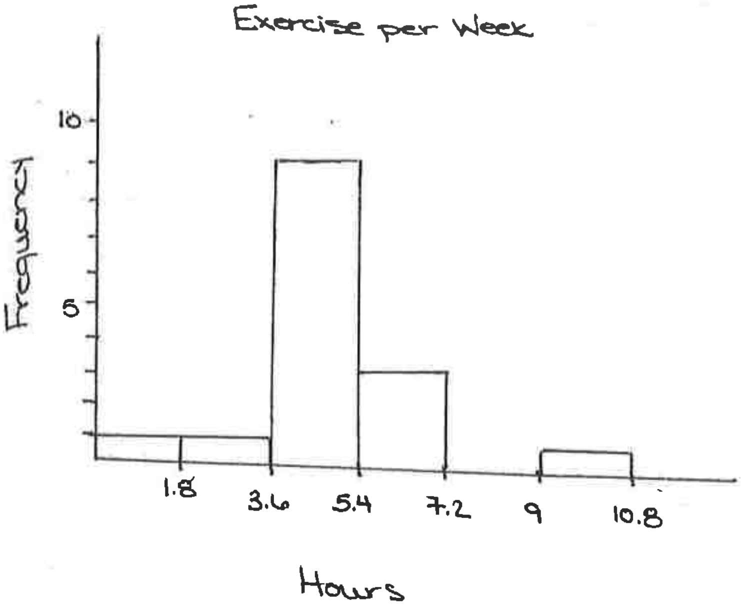
Find the mean, median, OR mode – your choice! * neither are easy to read off the graph *

Mean = 45.3 Median = 43 Mode = 31

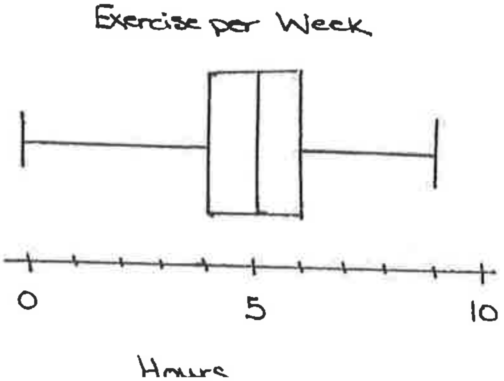
Example 4: Using the dot plot given, create a box-plot AND histogram.

0, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 9





min: 0
Q1: 4
med: 5
Q3: 6
max: 9



Integrated Math 1 Honors
Unit 9: Statistics
9.7

Name: ____________________

Date: __________ Period: _____

Objective: To organize quantitative data with scatterplots.

Warm up: What are some important aspects to include in a graph?

- title
- label axes (units and numbers)
- Counting by a consistent value
- breaks (if not starting at zero)

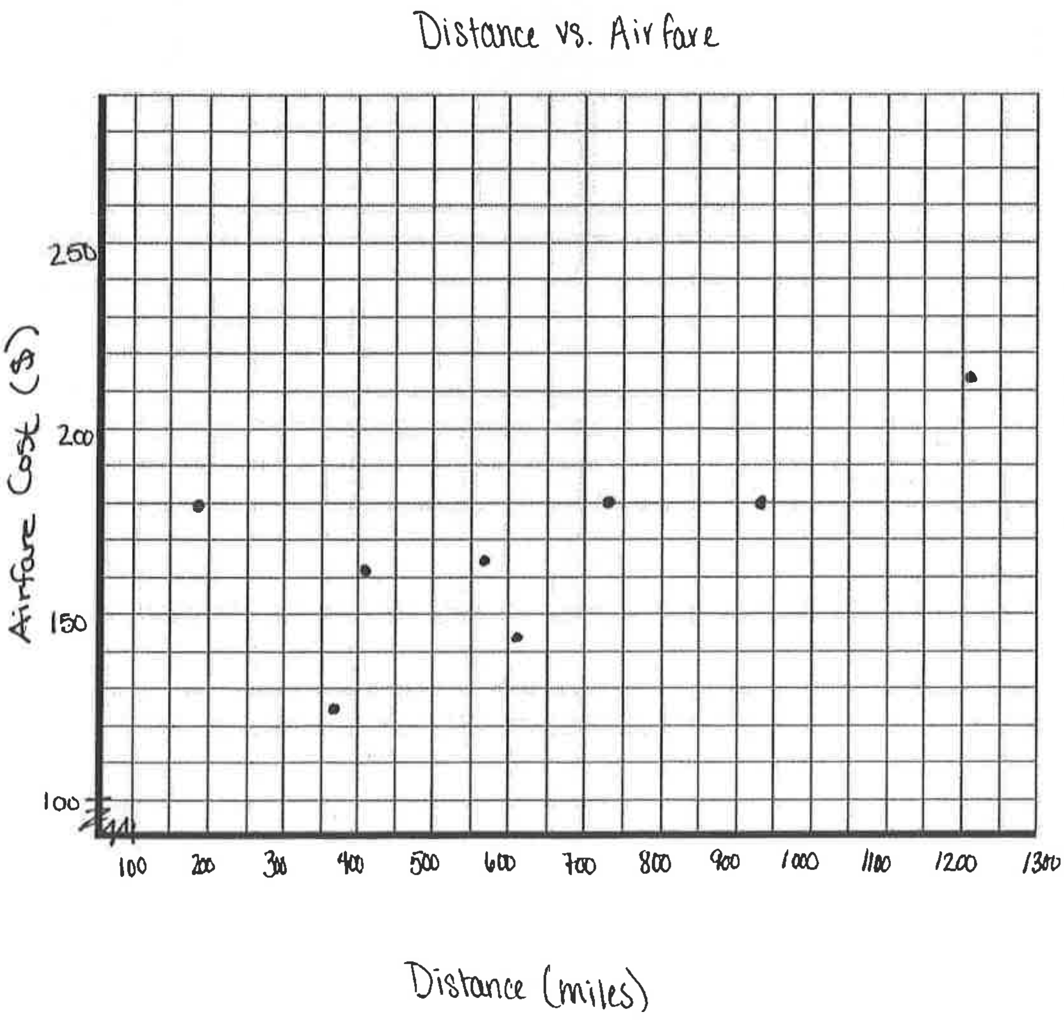
Scatter plots:

A scatter plot is used to show the relationship between two quantitative variables.

One of the variables (the independent variable) goes along the x-axis

and the other variable (the dependent variable) goes along the y-axis.

Example 1: Create a scatterplot for distance vs. airfare.



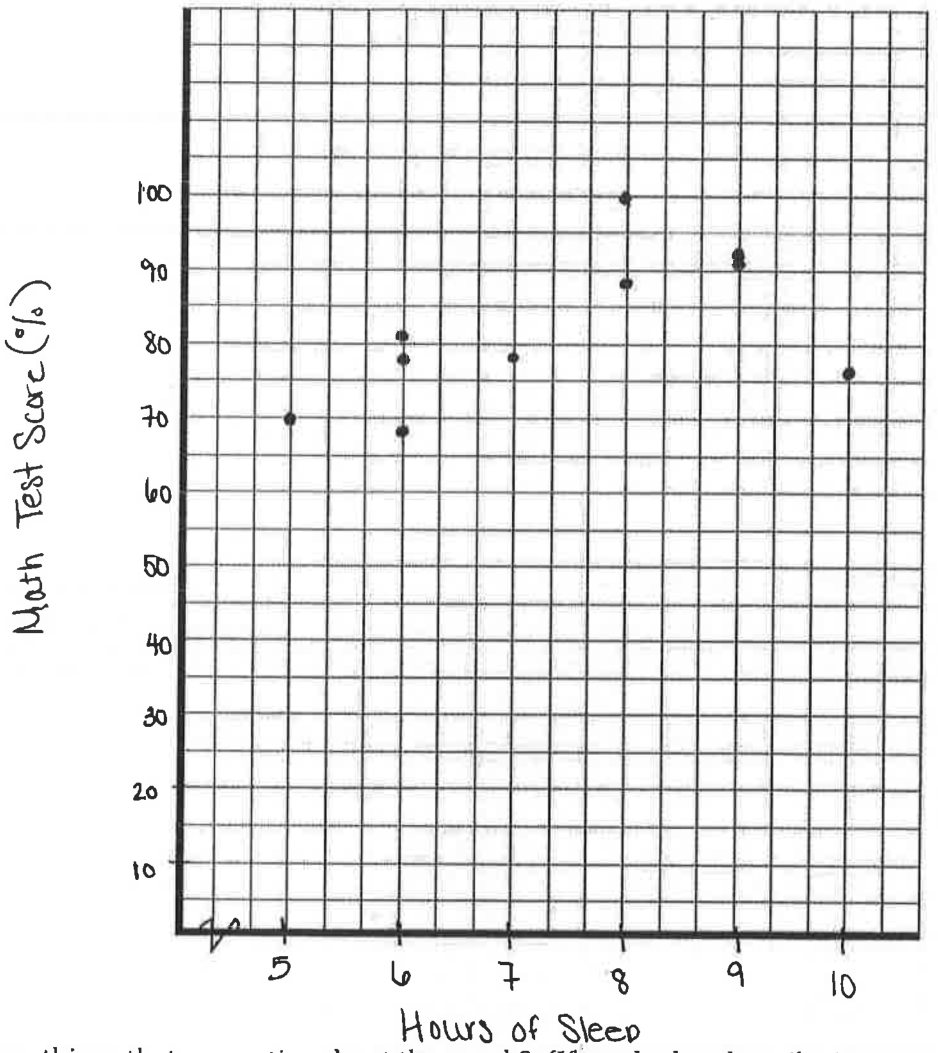
LOWEST-PRICED AIRFARES FROM BALTIMORE

Destination	Distance (in miles)	Airfare
Atlanta	576	$164
Boston	370	$124
Chicago	612	$143
Dallas	1,216	$260
Detroit	409	$161
Denver	1,502	$216
Miami	946	$180
New York	189	$108
St. Louis	737	$180

Example 2: Create a scatterplot for hours of sleep vs. math test score.

Hours of Sleep	9	5	6	6	8	9	10	7	6	8
Math Test Score	93	70	77	81	88	91	76	78	68	100

Hours of Sleep vs. Math Test Score



a. What are some things that you notice about the graph? (If you had to describe it to someone, what would you say?)

Points fall in the same general area

As the number of hours of sleep increase, test scores tend to increase.

b. Does it seem like there could be a relationship between hours of sleep and grade on a math test? Explain from what you see on the graph.

Yes- as sleep increased so did test scores.

Logically, this can't happen because if you sleep too much, you'll have less time to study so your scores won't be in the high percentages.

Creating Scatter Plots on the Calculator:

1. Enter the x-values (independent variable) into a list (usually L1)
2. Enter the y-values (dependent variable) into a list (usually L2)
3. 2nd → y= → Turn the stat plot on
4. The first option under "Type" is a scatter plot
5. Be sure that Xlist is set to the list used in step 1 and Ylist is set to the list used in step 2
6. Zoom → Stat (shows the graph in a perfect window)
7. Trace allows you to pinpoint an exact ordered pair

Predict:

a. Do you think that waist size and body fat percentage have a relationship? Explain.

b. Do you think that weight and body fat percentage have a relationship? Explain.

c. Which do you think is more closely related: [Waist size & Body Fat %] or [Weight & Body Fat %]? Explain.

Answers vary

Let's check it out!

Enter the following data into your calculator.

Waist → L1
Weight → L2
Body Fat % → L3

Hint: There should be 20 data points in each list!

1. Graph the scatter plot for waist $(L_1 \text{ vs. } L_3)$ size vs. body fat % in your calculator and describe the graph.

The points seem to indicate that as waist size increases, body fat % increases as well.

2. Graph the scatter plot for weight $(L_2 \text{ vs. } L_3)$ vs. body fat % in your calculator and describe the graph.

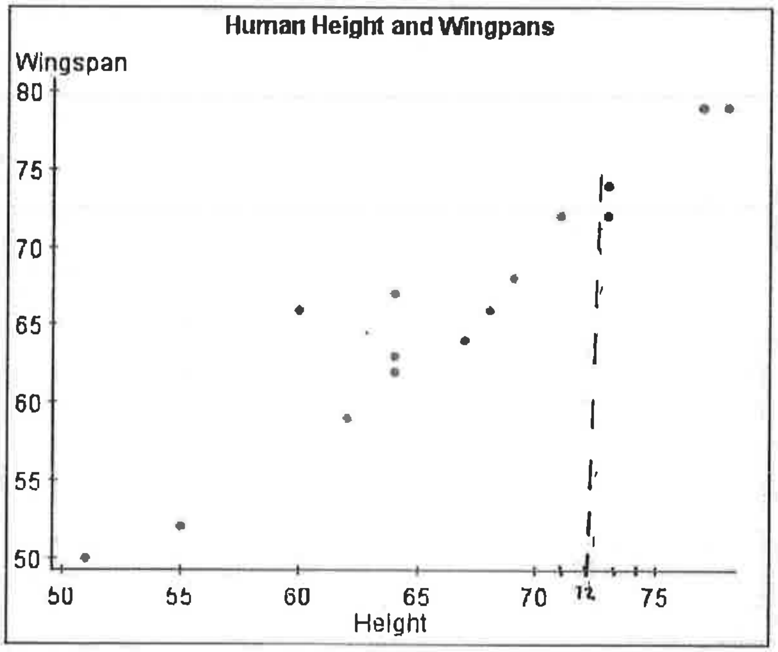
The points are more spread out, but still seem to follow the same pattern.

Waist (in) L1	Weight (lb) L2	Body Fat (%) L3	**Waist (in) L1**	Weight (lb) L2	Body Fat (%) L3
32	175	6	**33**	188	10
36	181	21	**40**	240	20
38	200	15	**36**	175	22
33	159	6	**32**	168	9
39	196	22	**44**	246	38
40	192	31	**33**	160	10
41	205	32	**41**	215	27
35	173	21	**34**	159	12
38	187	25	**34**	146	10
38	188	30	**44**	219	28

3. Now that you see the graphs, which do you think is more closely related and explain: [Waist size & Body Fat %] or [Weight & Body Fat %]

Waist size and body fat % - the points are less spread out (more consistent)

Example 3: The following graph shows human height (in inches) vs. wingspan (in inches).



a. Describe the graph.

The points seem to show that an increase in height matches with an increase in wingspan.

b. Do you think there is a relationship between height and wingspan? Explain.

There appears to be a relationship because the data almost forms a line and isn't very spread out.

c. Using the graph, what is your best prediction for the wingspan of someone who is 6' tall? Explain. (72 in.)

Somewhere between 67-75"

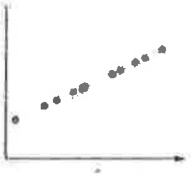
This would create a point that fits in nicely with the others.

Describing a Scatterplot:

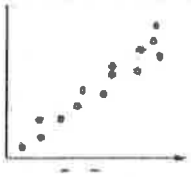
Direction	Form	Strength
Positive - Increase As x increases, y increases	Linear: points appear to be forming a line *no curves, a clear pattern*	Strong: points have very little deviation from the pattern and stay consistent.
Negative - Decrease As x increases, y decreases	Non-Linear: points will either curve or have no obvious pattern.	Moderate: points have some deviation from the pattern and/or are spread out a little bit.
		Weak: points have deviation from the pattern and/or are very inconsistent *may have no pattern to the line*

What would this look like????

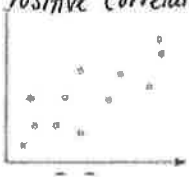
Perfect Positive Correlation



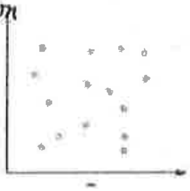
Strong Positive Correlation



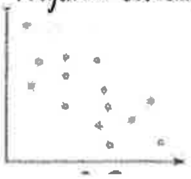
Weak Positive Correlation



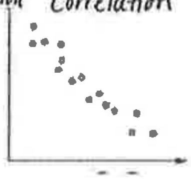
No Correlation



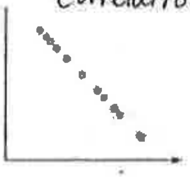
Weak Negative Correlation



Strong Negative Correlation

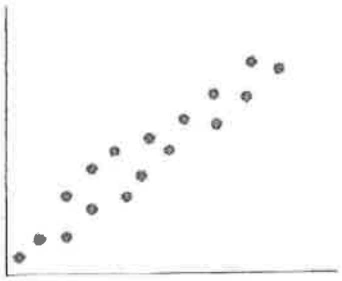


Perfect Negative Correlation



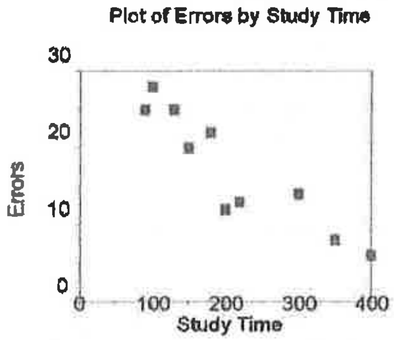
Example 4: Describe each graph. Be sure to comment on direction, form, and strength.

A.)



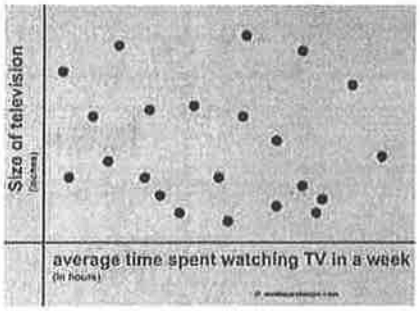
Positive, Linear, Strong

B.)



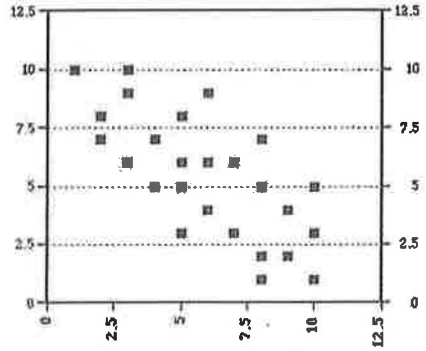
Negative, Linear, Moderate

C.)



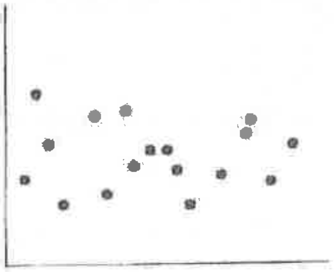
No Association, Weak Non-linear

D.)



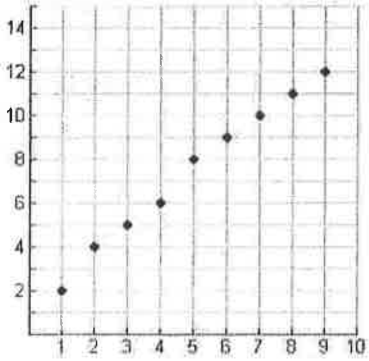
Negative, Linear, Weak

E.)



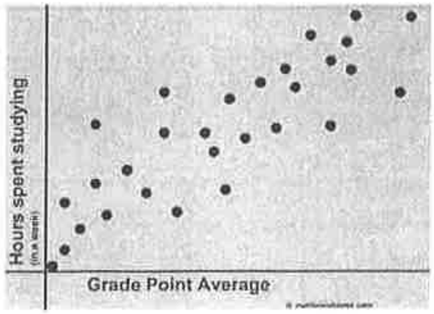
No association, Non-Linear, Moderate

F.)



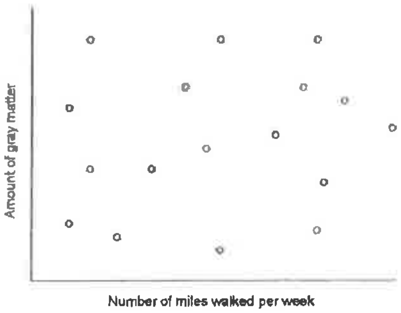
Positive, Linear, Strong

G.)



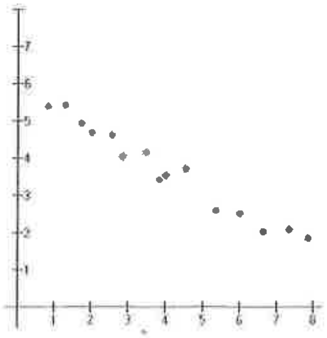
Positive, Linear, Moderate

H)



No association, Non-linear, Weak

I.)



Negative, Linear, Strong

Analyzing a Scatterplot:

Association	Correlation	Causation
The two variables show some sort of relationship (not necessarily linear)	The two variables have a <u>LINEAR</u> association	Cause-effect relationship Association ≠ Causation Correlation ≠ Causation * Causation cannot be proven statistically

Example 5: Decide whether you think each situation would have a positive correlation, negative correlation, or no correlation.

a. The number of loaves of bread baked and the amount of flour used.

Positive

b. A person's height and the number of letters in the person's name.

None

c. The number of mailboxes in a city and the number of firefighters in a city.

Positive

d. The price of a hamburger at a restaurant and the amount of hamburgers sold.

Negative

e. The amount of time you study for a test and the score you receive.

Positive

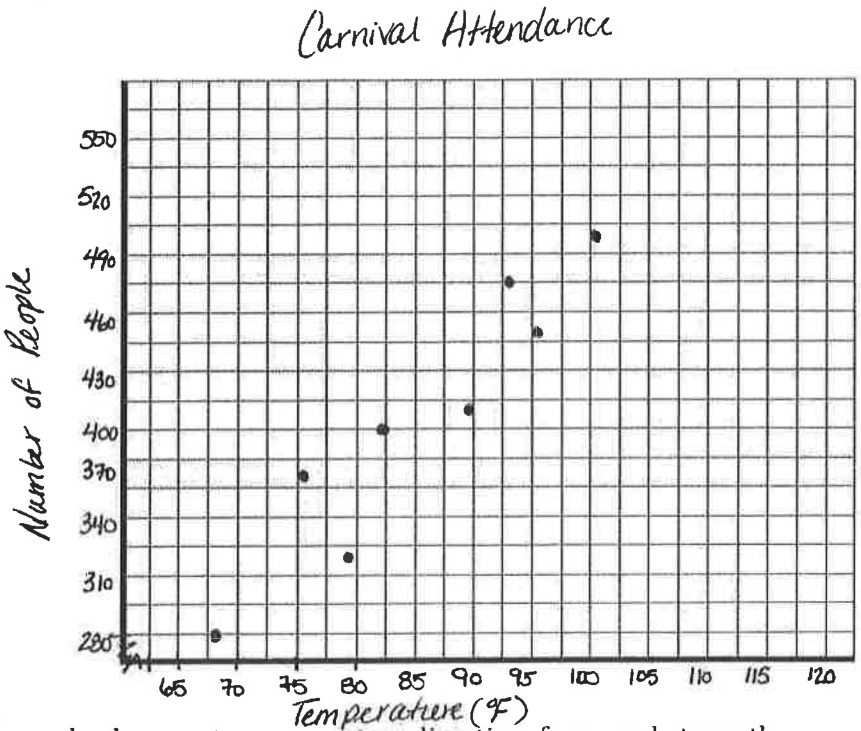
f. The amount of crime in the summer and the number of people who eat ice cream.

Positive

Example 6: The data table shows the temperature, in degrees Fahrenheit, and the number of people who attend a local carnival.

Day	Temperature (°F)	Number of People
1	68	280
2	75.2	360
3	96.8	450
4	89.6	420
5	82.4	400
6	100.4	500
7	93.2	475
8	78.8	320

a. Draw a scatter plot for temperature vs. number of people. Check the graph with your graphing calculator.



b. Describe the graph – be sure to comment on direction, form, and strength.

The graph shows a positive, moderate, linear association, which could be considered correlation.

c. Is there a correlation between temperature and number of people who attend the carnival? Explain using statistical concepts.

Possibly - there is a linear association so it is possible that a correlation exists between temperature and the number of people who attend a carnival.

It seems logical that an increase in temperature is paired with an increase in attendance.

Integrated Math 1 Honors
Unit 9: Statistics
9.8

Name: ______________________

Date: __________ Period: _____

Objective: To calculate the linear regression line for a scatterplot.

Warm up:

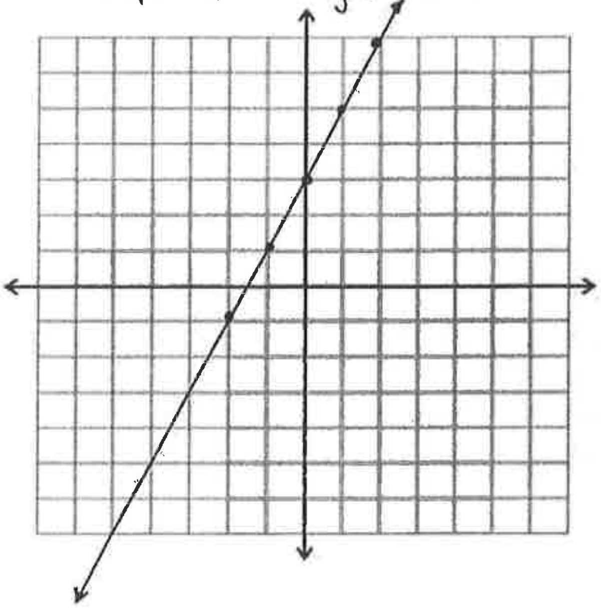
1. What is slope intercept form?

$y = mx + b$ ← y-intercept: (0,b)

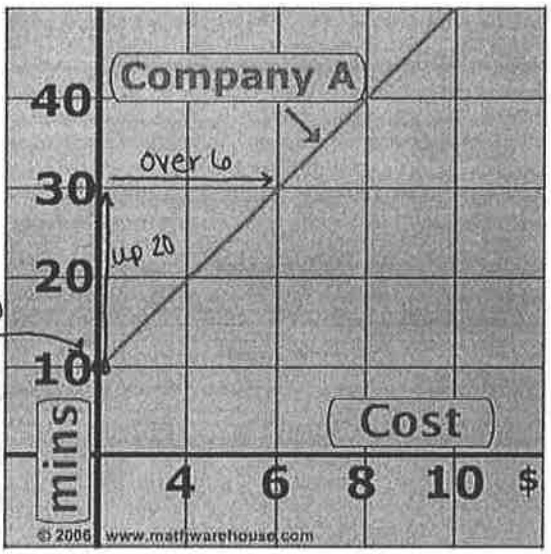
Slope = $\frac{rise}{run} = \frac{Y_2 - Y_1}{X_2 - X_1}$

2. Graph the line $y = 2x + 3$

Slope: $\frac{2}{1}$ y-int: (0,3)



3. What is the slope and y-intercept? Explain what these mean in context.



y-int: (0,10) The first ten minutes are free.

Slope: $\frac{20 \text{ min}}{\$6} = \frac{10}{3} = 3.\overline{3}$

Every 3.3 minutes costs $1

If a scatter plot appears to show a linear relationship (Correlation), then we can use our calculators to find the equation for a line that follows the general pattern of the data. You have probably heard the term "line of best fit." In statistics, this is called the ***least squares regression line*** *or* ***linear regression line.***

****THIS CANNOT BE DONE WITHOUT A GRAPHING CALCULATOR****

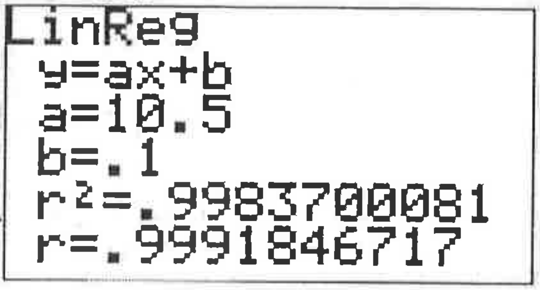
Finding the Linear Regression Line Using a Calculator:

1. Enter the x-values into L1
2. Enter the y-values into L2
3. 2nd → Y= → To create a scatter plot
4. STAT → CALC → 4: LinReg (ax+b)

Regression Equation:

predicted y-value → $\hat{y} = \underset{\text{Slope}}{a}x + \underset{\text{y-intercept}}{b}$

Something like this will pop up on your screen:



```
LinReg
 y=ax+b
 a=10.5
 b=.1
 r²=.9983700081
 r=.9991846717
```

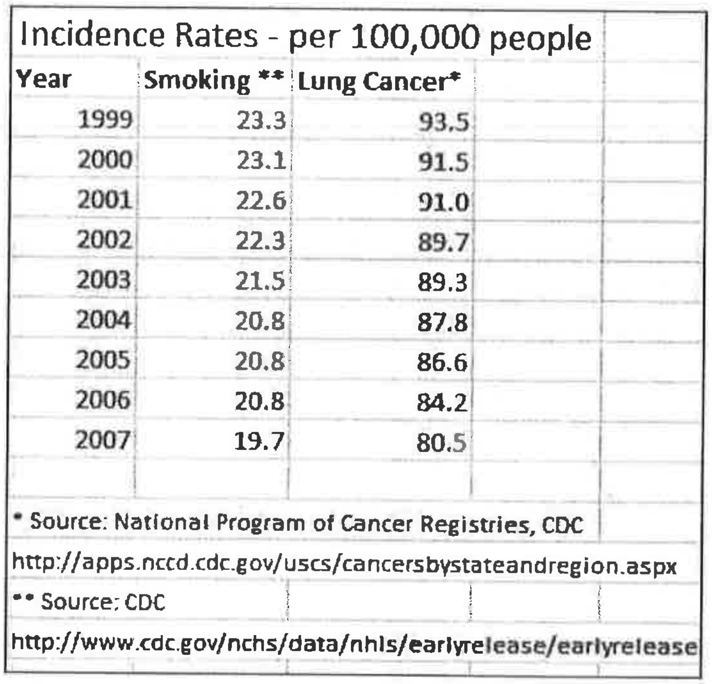
a: Slope

b: y-intercept

r²: % of variation accounted for by the model

r: Correlation Coefficient

Example 1: The following data shows smoking rates (per 100,000 people) vs. lung cancer rates (per 100,000 people) for the years 1999 through 2007.



Incidence Rates - per 100,000 people

Year	Smoking **	Lung Cancer*
1999	23.3	93.5
2000	23.1	91.5
2001	22.6	91.0
2002	22.3	89.7
2003	21.5	89.3
2004	20.8	87.8
2005	20.8	86.6
2006	20.8	84.2
2007	19.7	80.5

* Source: National Program of Cancer Registries, CDC

http://apps.nccd.cdc.gov/uscs/cancersbystateandregion.aspx

** Source: CDC

http://www.cdc.gov/nchs/data/nhis/earlyrelease/earlyrelease

a. Create a scatter plot in your calculator.

b. Describe the direction, form, and strength.

Strong (or moderate)
Positive
Linear

c. Calculate the linear regression equation.

$\hat{y} = 3.057x + 22.042$

d. Use your equation from part c to predict the lung cancer rate if the smoking rate is 17.3 per 100,000 people.

$\hat{y} = 3.057(17.3) + 22.042$

$= 74.93$ per 100,000 people.

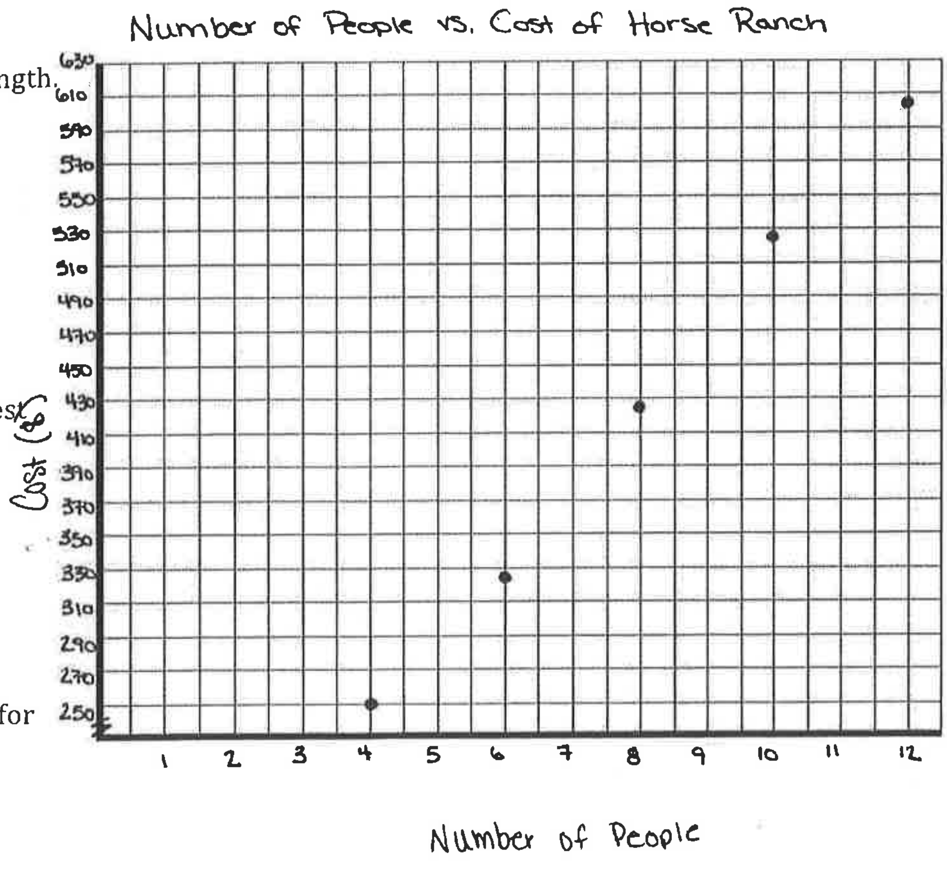
A linear regression equation can be very useful in predicting what could happen for an x-value that is not present in the original data. This is called extrapolation. However, we have to be careful when using the linear regression equation to predict:

1. It is only a prediction! It is not necessarily true.
2. The regression equation means nothing if the data is non-linear.
3. Your equation technically can only estimate y-values given x-values.
4. Predictions based off of values that are outliers are likely to be inaccurate.

Example 2: The following data shows the cost of visiting a horse ranch for one day and one night for different numbers of people.

Number of People	4	6	8	10	12
Cost	250	325	425	525	600

a. Draw a scatter plot of the data.



b. Describe the direction, form, and strength.

Strong, Positive, Linear

c. Calculate the equation of the line of best fit.

$\hat{y} = 45x + 65$

d. Use your equation to predict the cost for 15 people.

$\hat{y} = 45(15) + 65$

$= \$740$

e. Interpret the slope in context.

$\text{Slope} = \frac{45}{1} = \frac{\Delta \text{Cost}}{\Delta \text{people}}$

For every additional person, the cost increases by $45

f. Interpret the y-intercept in context.

There's a flat fee of $65

Important Aspects of Linear Regression

So far, we have just been using our own judgment based on a scatter plot to determine if a linear model is appropriate. There are many other more accurate ways to do this! This is great news for those cases that you feel are questionable. For this class, we will stick to checking three things:

1. Always check the Scatter plot of the data first first

If you determine that a linear model may be appropriate, find the equation of the line of best fit with your calculator. On the same screen as your slope and y-intercept, you will find two more things:

2. r^2: % of variation accounted for by the model (the bigger, the better!)

3. r: Correlation coefficient (the closer |r| is to 1, the better!)

These three things together should give you a very good idea of whether or not a linear model is appropriate for a set of data.

Let's think back to the weight, waist size, and body fat percentage data...

Perform a linear regression in your calculator for waist size vs. body fat percentage:

a: 2.22

b: -62.56

r^2: 0.787

r: 0.887

Waist (in) L1	Weight (lb) L2	Body Fat (%) L3	**Waist (in) L1**	Weight (lb) L2	Body Fat (%) L3
32	175	6	**33**	188	10
36	181	21	**40**	240	20
38	200	15	**36**	175	22
33	159	6	**32**	168	9
39	196	22	**44**	246	38
40	192	31	**33**	160	10
41	205	32	**41**	215	27
35	173	21	**34**	159	12
38	187	25	**34**	146	10
38	188	30	**44**	219	28

What does r^2 mean in this situation?

78.7% of variation is accounted for by the model. 21.3% comes from other factors (gender, age, diet, etc.)

What does r mean in this situation?

Moderate, positive correlation

What do these pieces tell us about using a linear model for this data?

A linear model is fairly accurate for interpreting the data.

Perform a linear regression in your calculator for weight vs. body fat percentage:

a: 0.2499

b: -27.376

r^2: 0.485

r: 0.697

What does r^2 mean in this situation?

48.5% of variation is accounted for by the model and 51.5% comes from other factors.

What does r mean in this situation?

Weak, positive correlation

What do these pieces tell us about using a linear model for this data?

A linear model could still be appropriate, but would not be a very good model for this data.

Which model is better: [✓ Waist size vs. Body fat] or [Weight vs. Body fat]? Explain using our NEW statistical concepts.

1. The scatterplot is more consistent and has a more obvious linear pattern.
2. The % of variation from outside sources is much smaller for waist size vs. body fat.
3. The r value is closer to 1, which indicates a stronger relationship.

Fun Facts about the r^2 and r values:

r^2 must be a number between 0 (0%) and 1 (100%)

- Higher percentages are better
- No set value we are looking for

r must be a number between -1 and 1

- Close to +1 means a strong, positive correlation
- Close to -1 means a strong, negative correlation.

Example 3: The data given below shows the height at various ages for a group of children.

Age (months)	18	19	20	21	22	23	24	25	26	27	28	29
Height (cm)	76	77.1	78.1	78.3	78.8	79.4	79.9	81.3	81.1	82	82.6	83.5

a. Find the line of best fit and interpret the slope and y-intercept in context.

$\hat{y} = .6339x + 64.9446$

Slope: Children grow .6339 cm per month

Y-Int: When a child is first born, they measure 64.9446 cm.

b. What is the residual of a 19-month old child that measures 78.2 cm tall? What does this mean in the context of the problem?

$\hat{y} = .6339(19) + 64.9446$

$= 76.9887$

Residual $= 78.2 - 76.9887$

$= \boxed{1.2113}$

This child is taller than what is to be expected.

c. Gene Poole's nephew is 2 years old (24 months) and is 80.2 cm. tall. Is this reasonable based off the data that was gathered? Explain.

$\hat{y} = .6339(24) + 64.9446$

$= 80.1582$

Residual $= 80.2 - 80.1582$

$= \boxed{.0418}$

This is reasonable. His nephew is only slightly taller than what is predicted.

d. Find the residual for an 18-month old child

$\hat{y} = .6339(18) + 64.9446$

$= 76.3548$

taken from table

Residual $= 76 - 76.3548$

$= -0.3548$

The child is shorter than predicted.

Integrated Math 1 Honors
Unit 9: Statistics
9.9

Name: ____________________________

Date: __________ Period: _____

Objective: To calculate and interpret the residual.

Warm up: Determine the upper and lower fence for the following data set and state any outliers:

61, 10, 32, 19, 22, 29, 36, 14, 49, 3

Q_1: 14 Q_3: 36
IQR: 22

Lower Fence: $14 - 1.5(22) = -19$

Upper Fence: $36 + 1.5(22) = 69$

No outliers

Calculating the Residual:

A **residual** is a measure of how well a line fits an individual data point. It is the vertical distance from the point to the best fit line.

$$Residual = Actual\ Value - Predicted\ Value$$

Example 1: The median age of men to marry can be predicted by $\hat{y} = 0.125x + 23.2$. This was calculated by knowing that the median age of men who tied the knot for the first time in 1970 was 23.2. In 1998, the median age was 26.7.

a. Use the equation to predict the median age of men who marry for the first time in 2005.

$2005 - 1970 = 35$

$\hat{y} = 0.125(35) + 23.2$

$= \boxed{27.575}$

b. If Harry R.M. Pitt got married at age 24 in 1998, calculate the residual and explain what this means in the context of the problem.

$\hat{y} = 0.125(28) + 23.2$
$= 26.7$

Residual $= 24 - 26.7 = \boxed{-2.7}$

Harry got married at an earlier age than what was expected.

Example 2: The following data shows the cost of visiting a horse ranch for one day and one night for different numbers of people.

Number of People	4	6	8	10	12
Cost	250	325	425	525	600

a. Determine the regression equation for the data.

$\hat{y} = 45x + 65$

b. If the Mr. and Mrs. Gusler and their two friends decide to go to the horse ranch and pay $268, calculate what the residual is and explain what it means in context.

$\hat{y} = 45(4) + 65$
$= 245$

Residual $= 268 - 245$
$= \boxed{23}$

They overpaid.