Statistics lesson 11.1: Variance and Standard Deviation

Objectives:
- I can calculate measures of center and spread for data sets.
- I can use statistics to describe data sets or to compare or contrast data sets.

Mrs. Smith has two Algebra 2 classes. Each class takes a test, and the scores of each class are shown below:

1^{st} hour	80	90	82	88	85	79	91	84	86	87	83

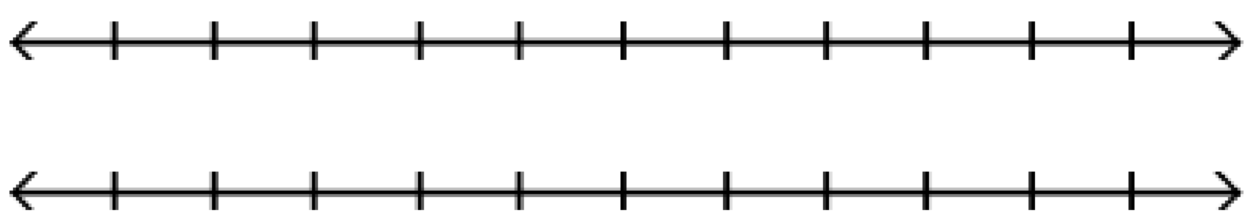
2^{nd} hour	100	81	100	85	85	70	89	85	100	55	85

Calculate the mean score for each class:

1^{st} hour: 2^{nd} hour:

Can she say that her classes performed equally well on the test? Why or why not?

Make a side-by-side dot plot showing the scores from each class:



What do the dotplots show about the distributions of scores?

Deviation: the difference between ____

Variance: roughly, the average of ____

Standard Deviation: square root of ____

Two data sets can have the same mean but a different standard deviation. This means:

Steps for finding standard deviation:

1.

2.

3.

4.

5.

6.

In summation notation:

variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

standard deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

<u>Example 1:</u> Finding Standard Deviation by Hand (no calculator)

4, 9, 11, 13, 18

mean**:**

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$

variance:

standard deviation:

sum:

<u>Your Turn: Find the standard deviation:</u>

8, 9, 10, 11, 12

mean**:**

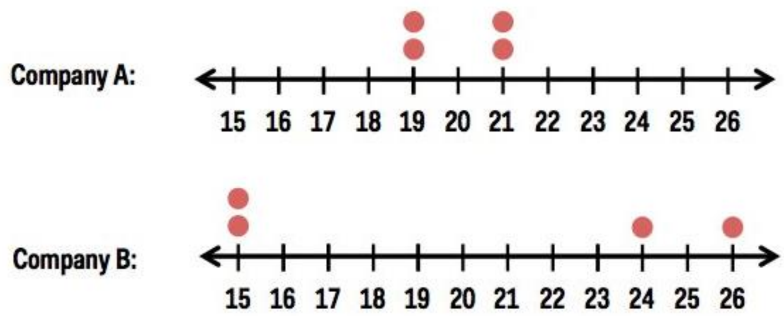
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$

variance:

standard deviation:

Sum:

Example:
Two companies sell used cars. Each keeps track of how many cars they sell during the months of November, December, January and February. The data is shown on the dotplots below.



Using your calculator, find the mean and standard deviation of the used car sales for each company.

Company A: mean: standard deviation:

Company B: mean: standard deviation:

Your turn:
Using your calculator, find the standard deviation for each of Mrs. Smith's Algebra 2 classes test scores.

1^{st} hour	80	90	82	88	85	79	91	84	86	87	83

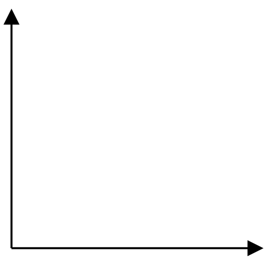
2^{nd} hour	100	81	100	85	85	70	89	85	100	55	85

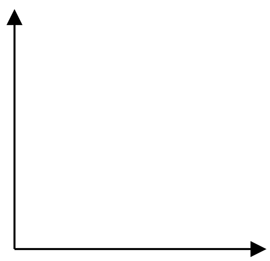
1^{st} hour: Standard deviation: **2^{nd} hour: Standard deviation:**

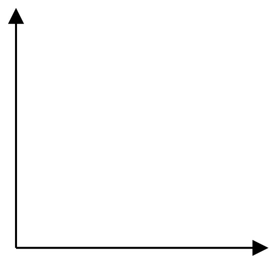
Statistics lesson 11.2: Distributions of data

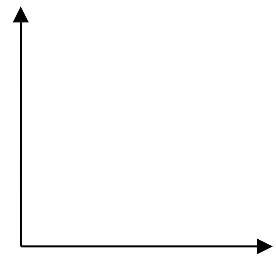
Objectives:
- I can determine the shape of a distribution.
- I can solve problems involving normally distributed data.

Common shapes for distributions of data:


uniform


skewed left


skewed right


bell-shaped

Example 1: Identify the most likely shape for the distribution of each data set:

a. Number of absences Mr. Kelly has in his first hour for one week:

DAY	# of ABSENCES
Monday	**2**
Tuesday	**1**
Wednesday	**5**
Thursday	**9**
Friday	**8**

b. The test scores for a very easy social studies test.

c. The SAT scores for 1000 randomly selected high school juniors

<u>Your turn:</u> Identify the most likely shape for the distribution of each data set:

a. The heights for all the students in your Algebra 2 class

b. The number of siblings for all the students at WOHS

Percentiles

Percentiles are a measure of the _______________ of a data value, compared to others in the data set.

p-th percentile: the pth percentile of a set of numbers is a value in the set such that p percent of the numbers are <u>less than or equal to</u> that value.

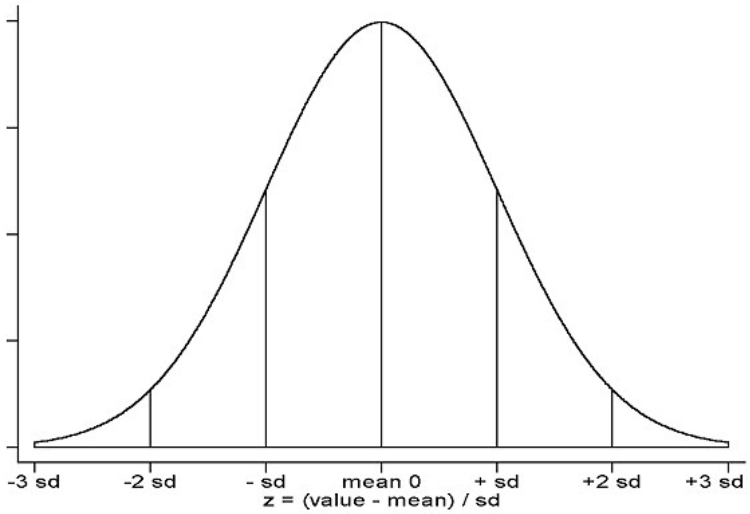
In other words, A percentile tells what percent of the values in the data set are less than or equal to the value you are considering.

Example 2: Johnny scored at the 87^{th} percentile on his SAT. What does this mean?

a. Johnny answered 87% of the questions on the SAT.
b. Johnny answered 87% of the questions correctly on the SAT
c. 87% of students who took the SAT scored better than Johnny
d. Johnny scored better than 87% of the students who took the SAT

Your turn: Samantha's doctor measured her height, and told her that she is at the 60^{th} percentile. Write a sentence explaining what this means.

<u>The standard Normal distribution:</u>



Facts:

____

____

____

____

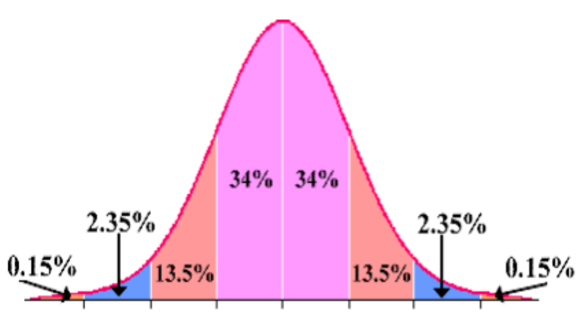
____

Empirical Rule: In a <u>normal</u> distribution,

_____% of the data lie within ____ standard deviation of the mean

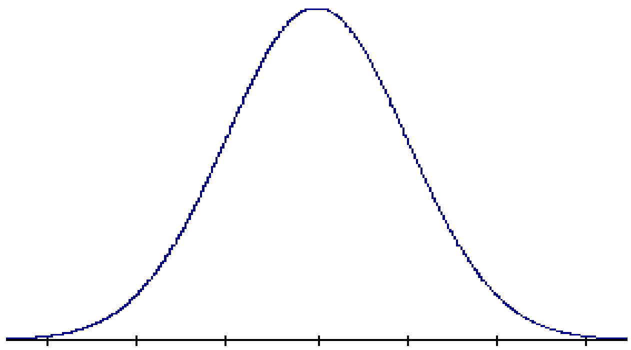
_____% of the data lie within ____ standard deviation of the mean

_____% of the data lie within ____ standard deviation of the mean



Example 1:
Students counted the number of candies in 150 small packages. They found that the number of candies per package was normally distributed with a mean of 23 candies per package and a standard deviation of 1 piece of candy.

a. Draw a sketch of the distribution of number of candies per package. Label the mean, as well as the points 1, 2 and 3 standard deviations from the mean.



b. What percent of the packages had between 22 and 24 candies?

c. What percent of the packages had more than 25 candies?

d. What percent of the packages had no more than 22 candies?

e. Out of the 150 bags, how many of them would have fewer than 21 candies?

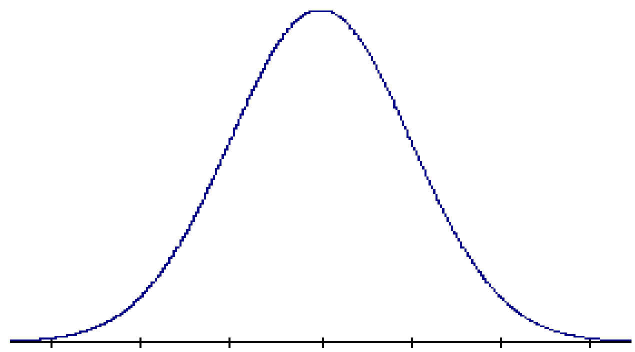
f. Between what two values does the middle 95% of candy bags lie?

g. How many candies would have to be in a bag for it to be at the 84th percentile?

Your Turn:

A fisherman counted the number of worms in 60 containers of bait. He found that the number of worms per container was normally distributed with a mean of 15 worms per container and a standard deviation of 2 worms.

a. Draw a sketch of the distribution of number of worms in the container. Label the mean, as well as the points 1, 2 and 3 standard deviations from the mean.



b. What percent of the containers had between 13 to 17 worms?

c. What percent of the containers had between 11 and 15 worms?

d. What percent of the containers had at least 17 worms?

e. What percent of the containers had at most 15 worms?

f. About how many containers had less than 11 worms or more than 19 worms?

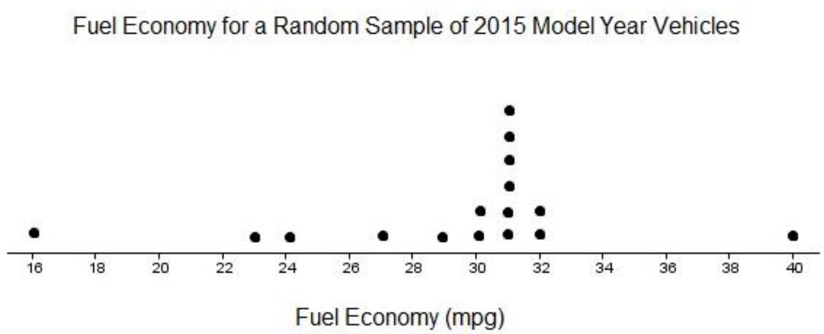
g. How many worms would a container have to have for it to be at the 16^{th} percentile?

Statistics lesson 11.3: Bivariate Data and Scatterplots

Objectives:
- I can construct a scatter plot for two quantitative variables.
- I can identify the explanatory and response variable in a bivariate data set.
- By looking at a scatter plot of a data set, I can identify the direction of the relationship between the variables and interpret the meaning of that direction in context.

When examining univariate data (one variable), we can use displays like dotplots, stemplots, box and whisker plots, etc.. to look for patterns in the data.

Example 1: Use the dotplot below to answer the questions.



What is the variable being measured?

How many vehicles were in the sample?

What is the median value in this data?

What is the mean?

What is the mode?

What is the range?

When working with bivariate data (____________ variables), we will concern ourselves with questions like:

1. **Are the two variables related?**
2. **What type of relationship exists?**
3. **If so, what is the strength of the relationship?**
4. **What kind of predictions can be made from the relationship?**

<u>Two types of variables:</u>

Categorical (Qualitative):

Examples:

Numerical (Quantitative):

Examples:

<u>Bivariate Data Where Both Variables are Quantitative</u>

- **Explanatory** variable (also called ____________________ or ____________________)
- **Response** variable (also called ______________________ or ____________________)

Example 2: In each of the following situations, identify the explanatory variable and the response variable.

a. The length of your bungee cord and the total distance you drop while bungee jumping

explanatory variable:

response variable:

b. number of calories in a candy bar and the amount of sugar in the candy bar

explanatory variable:

response variable:

c. the number of miles a vehicle has been driven and its price when sold.

explanatory variable:

response variable:

<u>Scatter plot</u>

- graph of 2 **quantitative variables** as ordered pairs (*x*, *y*) where *x* is the explanatory variable and *y* is the response variable

<u>Direction of the Relationship Between the 2 Variables</u>

- **Positive relationship** (positive association)

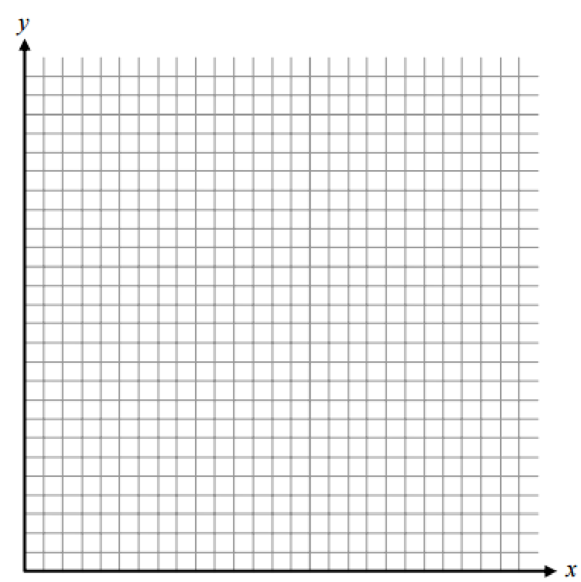
- **Negative relationship** (negative association)

Example 3:
Below is some data about the number of grams of sugar in some popular candy bars, along with the number of calories in each candy bar.

Name	Sugar (g)	Calories
Butterfinger Minis	45	450
Junior Mints	107	570
M&M'S	62	480
Milk Duds	44	370
Peanut M&M'S	79	790
Raisinettes	60	420
Reese's Pieces	61	580
Skittles	87	450
Sour Patch Kids	92	490
SweeTarts	136	680
Twizzlers	59	460
Whoppers	48	350

a. What is the explanatory (independent) variable?

b. What is the response (dependent) variable?



c. Construct a scatter plot of the data.

d. Describe the direction of the relationship. Are the variables positively or negatively associated?

e. What does this mean in the context of the situation?

Constructing a scatter plot on the graphing calculator:

1. Construct a scatter plot of the final grade vs. number of absences data set on your calculator. To do so, enter the number of absences in L_1, and the final grade in L_2. Remember: to enter data in a list, press **STAT EDIT**.

2. Access the **STAT PLOT** menu. To get to this feature, press **2nd Y=**.

3. If any plots other than Plot1 indicate that they're on, select that Plot and turn it off.

4. Select **Plot 1** by pressing **ENTER**. Turn Plot1 **On** by having On highlighted.

5. For graph **Type**, select the 1st graph in the upper left corner. The graph looks like several scattered points.

6. Next to **Xlist**, type **L_1**. Next to **Ylist**, type **L_2**.

7. Select which **Mark** you wish to use for your points. It's your decision.

8. Press the **ZOOM** key. Scroll down to option 9, **ZoomStat**. Using this, the calculator will resize your window automatically to best fit your data. With ZoomStat highlighted, press **ENTER.** You will now see the scatter plot on the screen.

Example 4: Each member of an Algebra 2 class ran a 40-yard sprint and then did a long jump (with a running start). The table below shows the sprint times (in seconds) and the long-jump distance (in inches).

Sprint time (Sec)	Long Jump Distance (in)
5.41	171
5.05	184
7.01	90
7.17	65
6.73	78
5.68	130
5.78	173
6.31	143
6.44	92
6.50	139
6.80	120
7.25	110

a. What is the explanatory (independent) variable?

b. What is the response (dependent) variable?

c. Construct a scatter plot of the data on your calculator. Sketch here:

d. Describe the direction of the relationship. Are the variables positively or negatively associated?

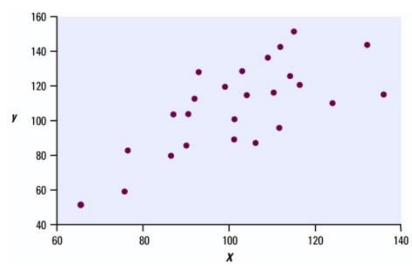
e. What does this mean in the context of the situation?

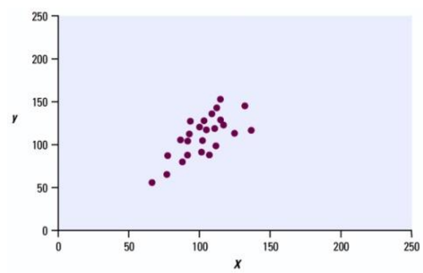
Statistics lesson 11.4: Correlation

Objectives:
- I can compute the correlation coefficient (*r*) for a set of quantitative data.
- I can describe what the correlation coefficient (*r*) measures.

Correlation: __________________ of a linear relationship

- A linear relationship is ____________ if the points lie __________ to a straight line.
- A linear relationship is _________ if the points are widely _________________ about a line.



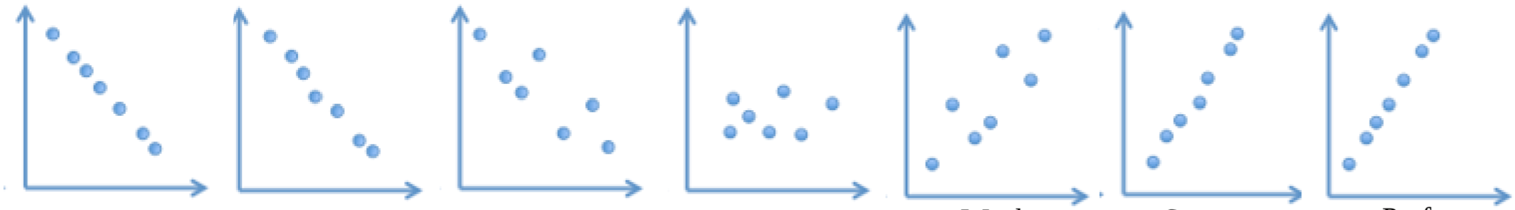


Our eyes are not a good judge of the strength of a linear relationship. The two scatterplots above show the same data set on a different scale. Which data set appears to be more strongly correlated?

correlation coefficient (*r*)

a number that indicates both the _____________ and ______________ of the __________ relationship between 2 variables

Facts/Properties About Correlation
1. Correlation requires both variables to be ______________________. 2. The correlation coefficient (*r*) is always a number between ______ and ______. 3. If *r* is *positive*, then there is a _______________ relationship between the variables. If *r* is *negative*, then there is a _______________ relationship between the variables. 4. The value of *r* will be exactly 1 or exactly -1 if ____. 5. The closer *r* is to _____, the ____________ the relationship between the variables.

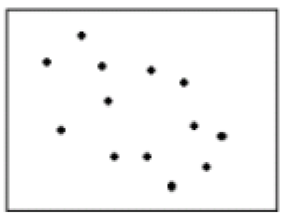


Perfect negative correlation | Strong negative correlation | Weak negative correlation | No correlation | Weak positive correlation | Strong positive correlation | Perfect positive correlation

r = ________ *r* = ________ *r* = ________

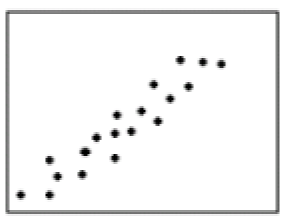
Example 1: Select the most likely value of the correlation coefficient for each scatterplot below.

A.



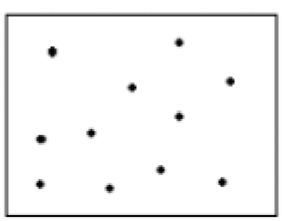
a. $r = 0.87$
b. $r = -0.54$
c. $r = 0$
d. $r = -1.15$

B.



a. $r = 0.92$
b. $r = -0.88$
c. $r = 1$
d. $r = 0$

C.



a. $r = -0.95$
b. $r = 0$
c. $r = -1$
d. $r = 0.95$

Example 2: Select the most likely value of the correlation coefficient (*r*) for each pair of variables.

A. x = outside temperature
y = ice cream sales

a. $r = 0.83$
b. $r = 0$
c. $r = 1$
d. $r = -0.92$

B. x = # of miles a car has been driven
y = the price of the car when sold

a. $r = 0$
b. $r = -0.74$
c. $r = -1$
d. $r = 0.95$

C. x = height of your classmates
y = classmates' scores on their last math test

a. $r = -1$
b. $r = 1$
c. $r = 0$
d. $r = 0.98$

Example 3: Look again at the data relating sugar and calories of popular candy bars. Put the data in to L_1 and L_2 in your calculator and make a scatterplot. Sketch below.

Name	**Sugar (g)**	**Calories**
Butterfinger Minis	45	450
Junior Mints	107	570
M&M'S	62	480
Milk Duds	44	370
Peanut M&M'S	79	790
Raisinettes	60	420
Reese's Pieces	61	580
Skittles	87	450
Sour Patch Kids	92	490
SweeTarts	136	680
Twizzlers	59	460
Whoppers	48	350

Based on looking at the scatterplot, will the value of *r* (the correlation coefficient) be positive or negative? __________________________ Why does this make sense?

Make a prediction on the value of *r*. ___________

Finding *r* with your calculator:

1. First, the **diagnostics** must be on.
2. Press **2nd**, **0**. This will bring you to the **catalog**.
3. Find **DiagnosticOn**. Press **ENTER** and **ENTER** again.
4. Now to find *r*: press **STAT**, **CALC**, **4: LinReg(*ax* + *b*)**. Complete the command **LinReg(*ax* + *b*) L_1, L_2.**
5. The value *r* at the bottom of the screen is the correlation coefficient. What value of *r* is shown on your calculator? ___________

Example 4: The following table gives the weights (in kg) and waist sizes (in cm) for 10 students.

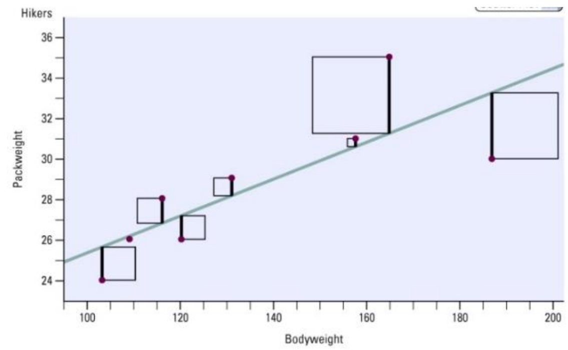
Name	Weight (kg)	Waist (cm)
Albert	87	101
Beth	65	71
Cindy	52	62
David	94	113
Emily	87	88
Frank	79	87
Gary	59	71
Helen	64	83
Ida	45	58
Jeremy	77	85

a. Calculate the value of the correlation coefficient.

b. What does the value of r tell you about the relationship between weight and waist size for these students?

Statistics lesson 11.5: Linear Regression

Objectives:
- I can find the equation of the least-squares regression line (LSRL) and use it to predict values.
- I can identify and interpret the slope and y-intercept of the LSRL in context.



Line of Best Fit/Least-Squares Regression Line (LSRL)

- *best fit* means that the sum of the squares of the vertical distances from each point to the line is as small as possible
- used to predict values of *y* for values of *x*

Try this applet: http://www.rossmanchance.com/applets/RegShuffle.htm

Equation of the Least Squares Regression Line

$$y = ax + b$$

a = ________________: tells us how much *y* is predicted to change for every increase of 1 unit in *x*

b = ________________: tells us what *y* is predicted to be when *x* = 0

Finding the Equation of the Least Squares Regression Line on the Calculator:

1. To get the equation of the LSRL, go to the **home screen (2^{nd} MODE)**. Type the following commands: **STAT CALC 4: LinReg(ax+b) L_1 , L_2 ENTER**

2. You will see the following information on the screen.

y = ax + b
a =
b =
r^2 =
r =

The value **a** is the **slope** of the LSRL.
The value **b** is the ***y*-intercept** of the LSRL.
The value *r* is the **correlation coefficient**.

Additional Option:

To have the calculator **graph the line on the scatter plot** you need one more command after what you just typed in. This is the sequence.

STAT CALC 4: LinReg(ax+b) L_1 , L_2 , VARS Y-VARS Function Y_1.

On the screen you will see **LinReg(ax+b) L_1, L_2, Y_1**. Press **ENTER**. If you now press **Y=**, you will see the equation of the LSRL entered. When you press **GRAPH**, the calculator will draw the LSRL on the scatter plot.

Example 1: Use the candy bar data again:

Name	Sugar (g)	Calories
Butterfinger Minis	45	450
Junior Mints	107	570
M&M'S	62	480
Milk Duds	44	370
Peanut M&M'S	79	790
Raisinettes	60	420
Reese's Pieces	61	580
Skittles	87	450
Sour Patch Kids	92	490
SweeTarts	136	680
Twizzlers	59	460
Whoppers	48	350

Construct a scatter plot of the data set on the calculator. Make a sketch below.

a. Find the equation of the least-squares regression line (LSRL). Round values to 3 decimal places. **STAT CALC 4: LinReg(ax+b) L_1 , L_2 ENTER**

b. What is the slope of the LSRL? Interpret the slope in context.

Slope = ____________. This means that for each additional gram of sugar a candy bar has, we can predict the number of calories in the candy bar will ________________ by _______________

c. What is the *y*-intercept of the LSRL? Interpret the *y*-intercept in context.

y-int = ____________. This means that if a candy bar has zero grams of sugar, we can

predict the number of calories the candy bar will have will be ________________.

d. A King-sized Butterfinger has about 50 grams of sugar. Use your LSRL equation to predict the number of calories in Butterfinger bar.

e. A King-sized Butterfinger actually contains about 514 calories. Was the prediction you made with the LSRL equation too high or too low? By how much? Why do you think this happened?

f. Use your LSRL equation to predict the number of calories in a candy bar that has 65 grams of sugar.

Example 2: Each member of an Algebra 2 class ran a 40-yard sprint and then did a long jump (with a running start). The table below shows the sprint times (in seconds) and the long-jump distance (in inches).

Sprint time (Sec)	Long Jump Distance (in)
5.41	171
5.05	184
7.01	90
7.17	65
6.73	78
5.68	130
5.78	173
6.31	143
6.44	92
6.50	139
6.80	120
7.25	110

a. Construct a scatter plot of the data on your calculator. Sketch here:

b. What is the value of the correlation coefficient (r)? __________

c. What does this value tell you about the relationship between the students' sprint times and long-jump distances? (Is the relationship positive or negative? Strong or weak?)

d. Find the equation of the least-squares regression line (LSRL). Round values to 3 decimal places.

e. What is the slope of the LSRL? Interpret the slope in context.

Slope = ____________. This means that ___________________________________

____

f. What is the y-intercept of the LSRL? Interpret the y-intercept in context.

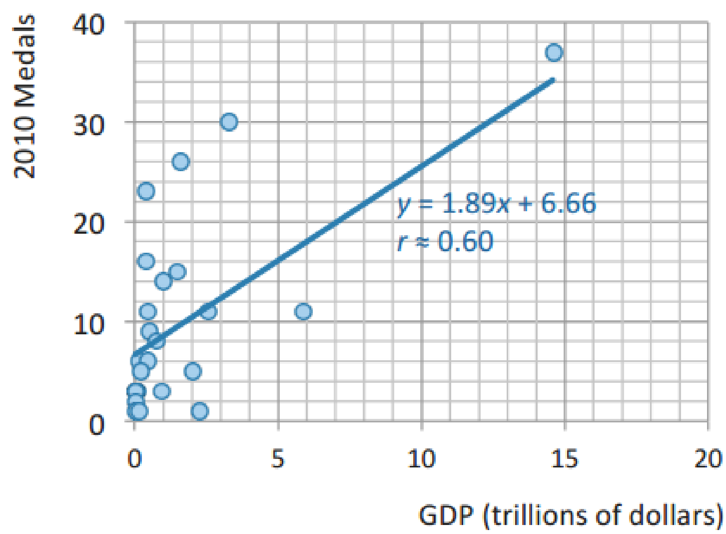
y-int = ____________. This means that ___________________________________

____

g. Use your LSRL equation to predict how far a student will jump if he/she runs the 40-yard sprint in 6.5 seconds.

h. One student in the class ran 40-yards in 6.5 seconds and jumped a distance of 139 inches. Was your prediction too high or too low? By how much?

Example 3: The scatterplot below shows a country's Gross Domestic Product (a measure of its wealth) and the number of medals it won in the 2010 Winter Olympics.

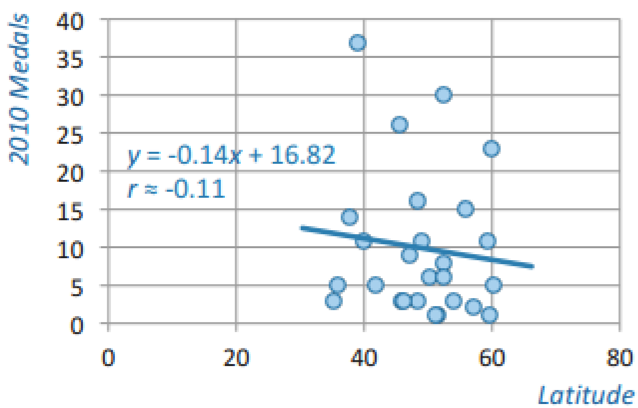


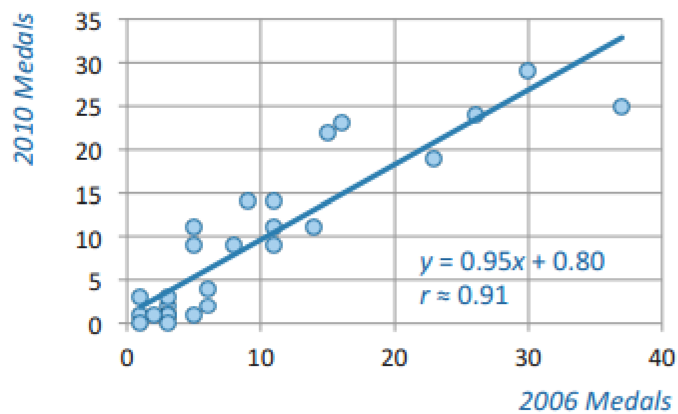
a. What does the data tell you about the relationship between GDP and performance in the Winter Games?

b. Germany's GDP in 2010 was approximately \$3.3 trillion. Based on this, how many medals would you predict that they won in the 2010 Winter Olympics? How confident are you in your prediction, and why?

c. Germany actually won 30 medals in 2010. How close was your prediction?

There are other factors besides a country's wealth that may predict success at the Olympics. For each country that has won a medal at the Winter games, two other data values were collected: the latitude of the country's capital and the number of medals the country won in 2006. Plots of these variables are shown below:





d. Based on the LSRL equations and the values of the correlation coefficients, which variable do you feel does the best job of predicting Olympic performance (GDP, latitude, or 2006 medals)? Why?

e. Use the best LSRL equation to predict the number of medals won by the Germany in 2010. Is this prediction better than your prediction in part b?

Latitude of German capital: 52.5º
Medals (2006): 29