SCATTER PLOTS AND LEAST SQUARES LINE

In this unit you will learn about scatter plots. Scatter plots are used to show how two variables relate to each other by showing how closely the data points cluster to a line (line of best fit, least squares line). Scatter plots can be used to predict relationships between two sets of data like those related to weather.

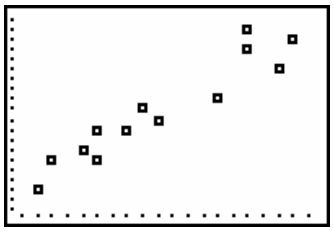
Scatter Plots

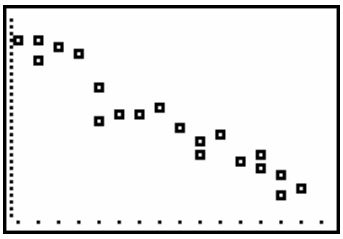
Line of Best Fit (Least Squares Line)

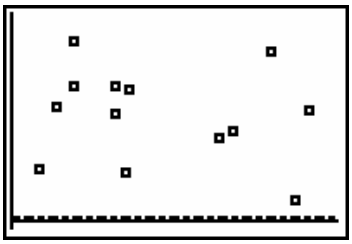
Scatter Plots

Scatter plots are an easy way of determining if there is a relationship between two variables. This relationship is called a **correlation**. A correlation is based on the slope of the line of best fit. (We will discuss how to find the line of best fit later in the unit).

There are three possible types of correlation; a) positive, b) negative, or c) no correlation. The illustrations below show the graph of each correlation.

a.) 

b.) 

c.) 

In graph "a" notice how the points cluster in a rise to the right. Recall that this suggests a **positive slope**. Graph "b" shows points that cluster in a fall to the right, which suggests a **negative slope** and graph "c" shows no cluster pattern and suggests the two variables have **no relationship** to each other.

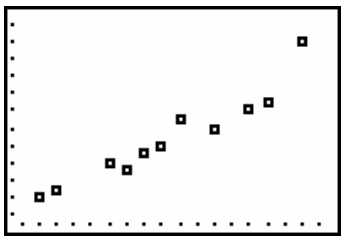
Let's take a look at a few examples and determine if each situation has a positive, negative, or no correlation.

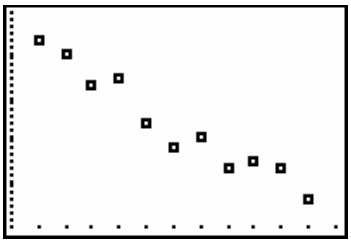
Example #1: Determine which scatter plot represents each situation.

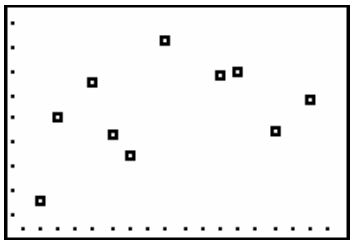
a.) your height and your hourly wage

b.) your height and your shoe size

c.) your age and the time needed to run 100 yards

1.) 

2.) 

3.) 

Scatter plot 1 shows a strong positive correlation. A **positive correlation** occurs when **both variables increase**. As you grow taller, your shoe size increases; therefore plot 1 represents situation (b).

Scatter plot 2 shows a strong negative correlation. A **negative correlation** occurs when **one variable increases as the other variable decreases**. In situation (c) as your age increases, the time it takes you to run 100 yards decreases.

The third scatter plot shows **no correlation** because the data points are **randomly scattered**. Your height has no relationship with your hourly wage; therefore this plot represents situation (a).

Line of Best Fit (Least Squares Line)

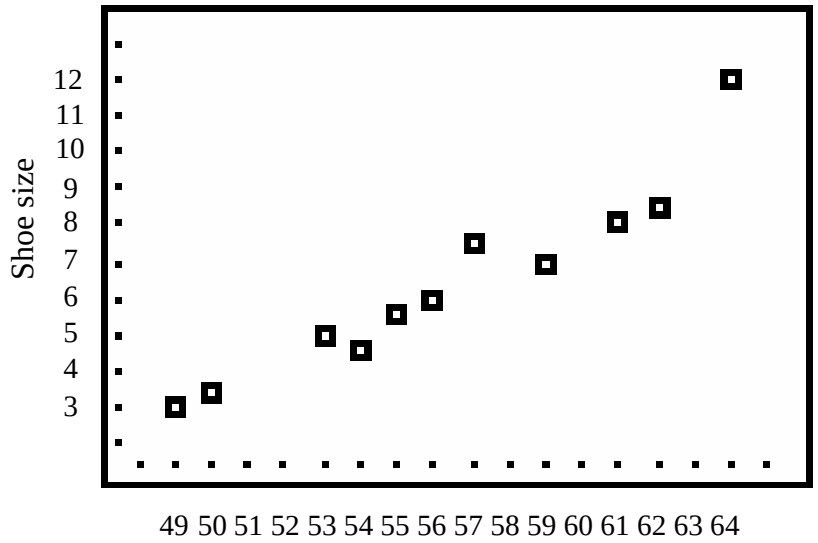
Earlier in this unit we talked about a line of best fit. Data in a scatter plot can be studied using a line of best fit, which represents the trend or behavior of the data. A line of best fit can be used to predict what the data might be for values not given.

Let's use the data given from the example above representing height in inches and shoe size to find the line of best fit for the scatter plot.

Example #1:

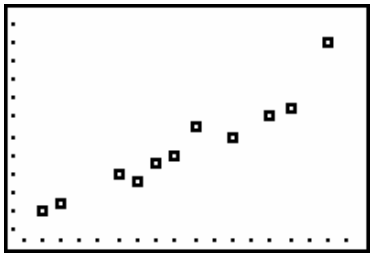
Height (inches)	49	50	53	54	55	56	57	59	61	62	64
Shoe size	3	3.5	5	4.5	5.5	6	7.5	7	8	8.5	12

1.) Plot the data as coordinates on a coordinate plane (height, shoe size). The height on the horizontal or *x*-axis and the shoe size on the vertical or *y*-axis.

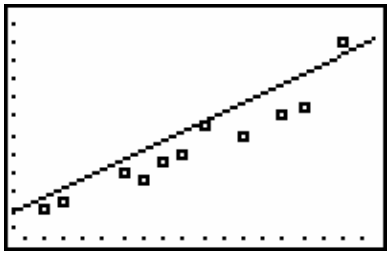


2.) Use some type of a straight edge, such as a clear ruler or piece of uncooked spaghetti to model the line that represents the trend of the data.

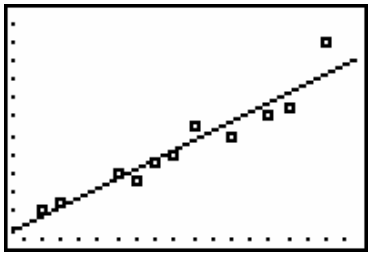
3.) To fit the line to the points, place your straightedge so that it best matches the overall trend-having the same number of points above and below the line. The examples below show what not to do and what should be done. ***The line does not have to go through any of the points as long as the general pattern or trend is represented.**



Scatter Plot



Line 1



Line 2

Line 1 goes through one of the points but completely ignores the others.

Line 2 is a closer representation of the data points as you can see it takes all points into consideration and there seems to be as many points above the line as there are below the line. Therefore line 2 is a better fit.

It will take you a while to practice this and it is a concept that has no exact answer. The idea of a line of best fit is to predict what the data will show for points that are not plotted.

After you have established the trend of the data, you need to choose two points that lie on your line $(x_{1,}y_1)$ and (x_2, y_2) so you can calculate the equation for the line of best fit.

Recall that if given two points on a line, you can find the equation of the line. Follow the directions below to refresh your memory.

1.) find the slope using the two points, do not choose two points that are side by side (it's a good idea to use a point from the first part of the line and the last part of the line)

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

2.) use one of the points and the slope you just found in the point-slope formula to generate the equation.

$$y - y_1 = m(x - x_1)$$

3.) solve the equation for *y* and you have the equation of the line of best fit.

For the example above, we will choose two points that look like they lie on the line. Let's choose (52, 4) and (62, 10).

1) find the slope using these two points

$$m = \frac{10-4}{62-52} = \frac{6}{10} = \frac{3}{5}$$

2) use one of the points and the slope in the point slope form

$$y - 4 = \frac{3}{5}(x - 52)$$

$$y - 4 = \frac{3}{5}x - 31.2$$

$$y = \frac{3}{5}x - 27.2$$

The equation of the line of best fit is $y = \frac{3}{5}x - 27.2$. You can now use this to determine data what is not shown on the graph.

For example, let's say that you wanted to know the shoe size of someone who is 68 inches tall. Since your scatter plot does not represent data for this height, you can use the equation $y = \frac{3}{5}x - 27.2$ to determine this.

Since you know the height of the person (68 inches), you will replace this for *x* in the equation and solve for *y*, which represents the shoe size.

$$y = \frac{3}{5}(68) - 27.2$$

$$y = 40.8 - 27.2$$

$y = 13.6$ approximately a shoe size of 13.5

The trend of the data shows us that if a person is 68 inches tall, they will wear a shoe size of approximately 13.5. Again this is just a predication based on the information we have in our scatter plot. It does not mean that a person 68 inches tall should or will wear a 13.5 size shoe.