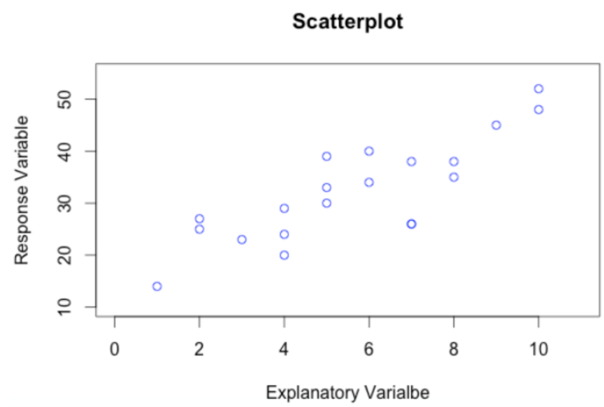
Scatterplots and Correlation

Objective 1 – Draw and interpret scatter diagrams

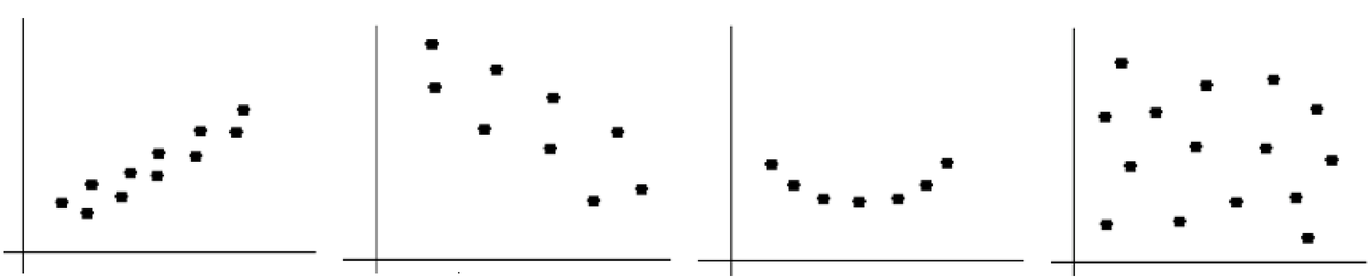
- ___________________________________ – a graph that shows the relationship between two quantitative variables measured on the same individual.

    - Each individual in the data set is represented by a point in the scatter diagram
    - The **explanatory** variable, *x*, is plotted on the ***horizontal*** axis
    - The **response** variable, *y*, is plotted on the ***vertical*** axis



After constructing a scatterplot, we want to interpret and describe the overall pattern:

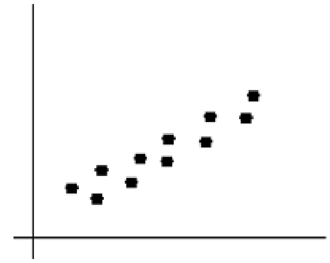
Types of relations in a scatterplot: linear, non-linear, no relation



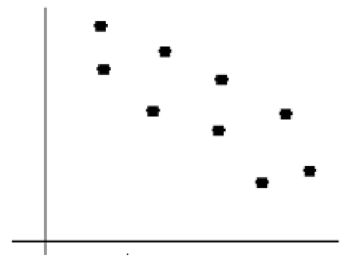
a) Linear b) Linear c) Non-linear d) No relation

Direction of a relation:

- There is a ____________________ positive linear relationship if, whenever the value of one variable increases, the value of the other variable also increases.



- There is a ____________________ negative linear relationship if, whenever the value of **one** variable **increases**, the value of **the other** variable **decreases.**



Example. The data shown to the final exam scores of 10 randomly selected students from a statistics class and the number of hours they studied for the exam. So, the number of hours studied is the explanatory variable, *x*, and final exam score is the response variable, *y*.

Hours	3	5	2	7	2	4	4	5	6	3
Score	65	80	60	92	66	78	85	90	90	71

a) Draw a scatter diagram of the data using your calculator, and then sketch it using appropriate scale and labels.

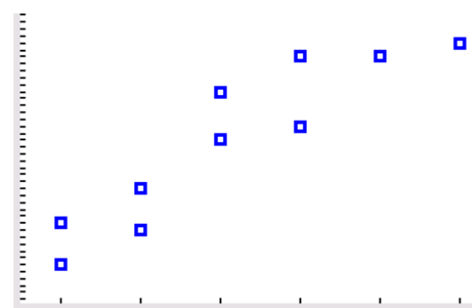
To **create the Scatterplot** on your calculator:

1. Input the data into L1 and L2
2. Hit 2nd, Stat Plot (Y=)
3. Cursor on **Plot 1** and hit Enter
4. Cursor on "**On**" hit Enter
5. Arrow **down** to **Type**, and select the **top-left option** (Scatterplot)
6. Input details from the problem:

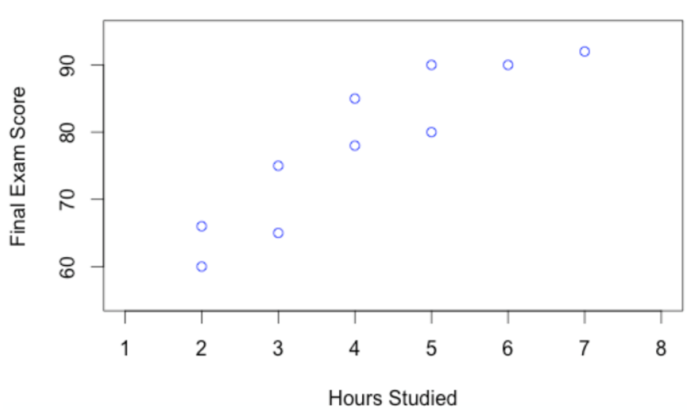
   Xlist: L1

   Ylist: L2

   Mark: any will work
7. Hit Zoom, arrow down to **ZoomStat** (option 9), and hit Enter

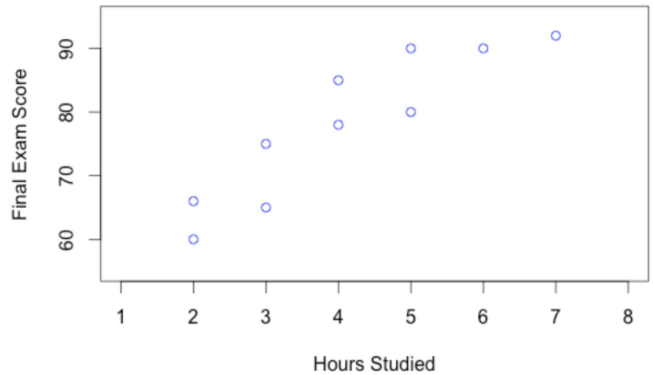


The final scatter plot with appropriate scale and labels should look as below:



b) Interpret the scatterplot by describing the type and direction of the relationship between number of hours studied and final exam score.

There is a **positive linear** relationship between number of hours studied and final exam score. That is, the more hours a student studies, the higher his or her final exam score will be.



Pause the video, and try this problem. Resume the video to show the answer to the problem.

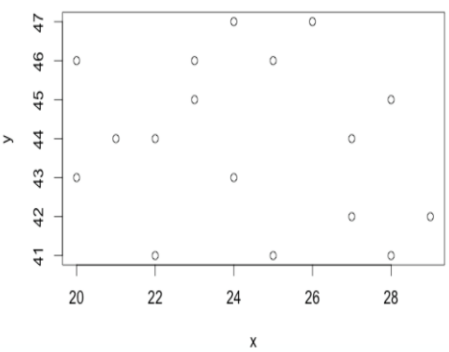
1) A sample of 9 employees was taken at a factory. The data below are their commute time to work and the number of absences.

Commute time (min), x	72	85	91	90	88	98	78	100	80
# of absences, y	3	7	10	10	8	15	4	15	5

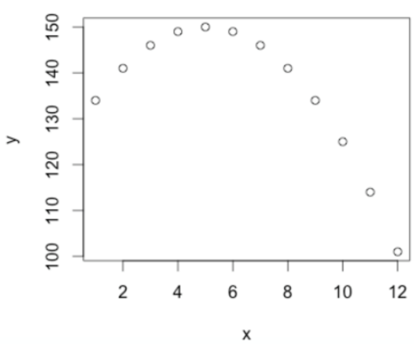
a. Draw a scatter diagram of the data using your calculator, and then sketch it.

b. Describe the type and direction of the relationship between commute time and number of absences.

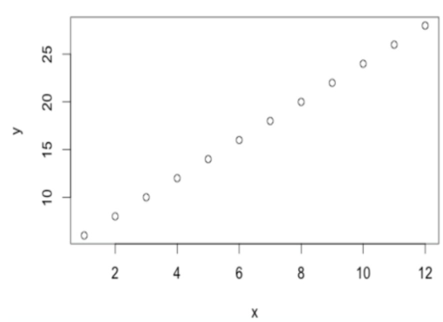
2) Given the scatterplots below,



a.



b.



c.

Describe the type and direction of the relationship between the variables.

Restart when you are ready to check your answers.

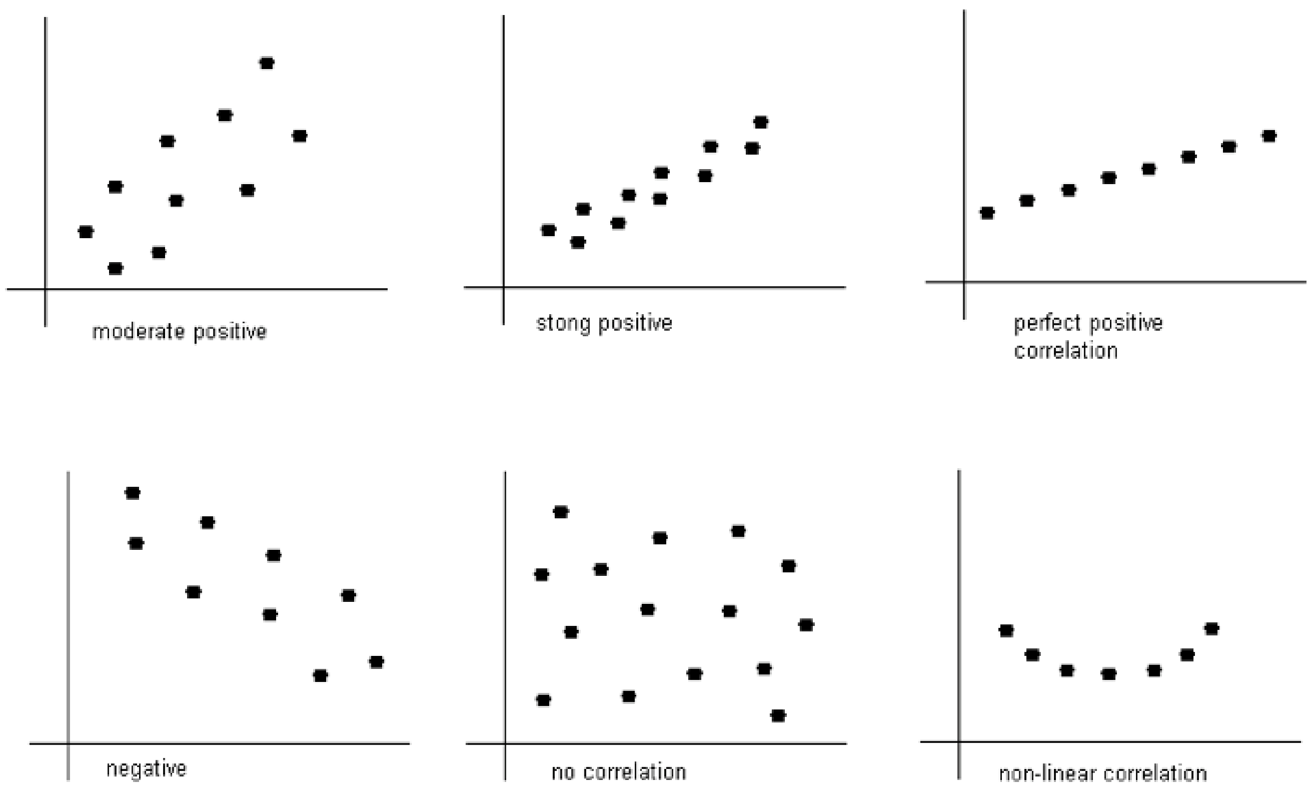
Objective 2 – Compute and Interpret the Linear Correlation Coefficient

- The ***linear correlation coefficient (r)*** – a number that describes the ***strength and direction*** of the linear relation between two quantitative variables.

Properties of the Linear Correlation Coefficient

- $-1 \leq r \leq 1$.
- If ____________________, then there is a **perfect positive** linear relation.
- If ____________________, then there is a **perfect negative** linear relation.
- The closer r is to +1 or -1, the ____________________ the linear relationship between the two variables
- If ***r* is close to 0**, then **little or** _________________ exists of a linear relation between the two variables
    - Note: r close to 0 does not imply no relation, just no **linear** relation

Quick Visual Examples:



Example. The data shown to the final exam scores of 10 randomly selected students from a statistics class and the number of hours they studied for the exam. Determine the linear correlation coefficient and use it to describe the linear relationship between hours studied and final exam score.

Hours	3	5	2	7	2	4	4	5	6	3
Score	65	80	60	92	66	78	85	90	90	71

To find the value of the correlation coefficient, *r*, with our calculator, **the first** thing we need to do is to turn **diagnostic on**.

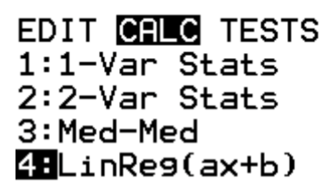
Hit 2nd, Catalog (#0 button), arrow down to **DiagnosticOn**

Hit Enter **twice**…your calculator should say, "**Done**"

If you ever do not get the value of *r*, it is likely because your calculator's diagnostic is not on.

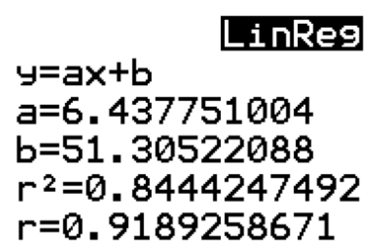
We have already entered the data in lists L1 and L2

Now hit Stat, arrow to **Calc**, arrow down to **LinReg(ax + b)** (option 4), Enter



Finally, input: **LinReg(ax + b) L1,L2** and hit Enter to execute.

The output will include our correlation coefficient.



For our example, we got to ***r* = 0.9189**.

Note that the value is positive and close to 1, so we can conclude that there is a strong, positive linear relationship between the number of hours a student studies and their final exam score.

Pause the video, and try this problem. Resume the video to show the answer to the problem.

1) A sample of 9 employees was taken at a factory. The data below are their commute time to work and the number of absences.

Commute time (min), x	72	85	91	90	88	98	78	100	80
# of absences, y	3	7	10	10	8	15	4	15	5

Find the linear correlation coefficient and use it to describe the strength and direction of the relationship between commute time and number of absences.

2) Given the linear correlation coefficient of a linear relation between two variables, describe the strength and direction of the linear relationship.

a. r = -0.87

b. r = 0.23

Restart when you are ready to check your answers.

Objective 3 – Compute and Interpret the Coefficient of Determination

- ____ – measures the proportion of total variation in the response variable that is ______________ by the least-squares regression line.
    - To determine R^2 for the linear regression model, simply **square the value of the linear correlation coefficient, ______.**

Properties of R^2:

- The coefficient of determination is a number between ___________, inclusive. That is, $0 \leq R^2 \leq 1$
- The closer R^2 is to 1, the better the regression line __________ the variation in the response variable.

Example. Find and interpret the coefficient of determination for the final exam data (previous example).

Recall, we found the correlation coefficient, $r = 0.9189$. So $\mathbf{R^2 = (0.9189)^2 = 0.8444}$

Interpretation: 84.44% of the ______________ in final exam scores is _____________ by the least-squares regression line.

Pause the video, and try this problem. Resume the video to show the answer to the problem.

1) A sample of 9 employees was taken at a factory. The data below are their commute time to work and the number of absences.

Commute time (min), *x*	72	85	91	90	88	98	78	100	80
# of absences, *y*	3	7	10	10	8	15	4	15	5

Find and interpret the coefficient of determination.

Restart when you are ready to check your answers.