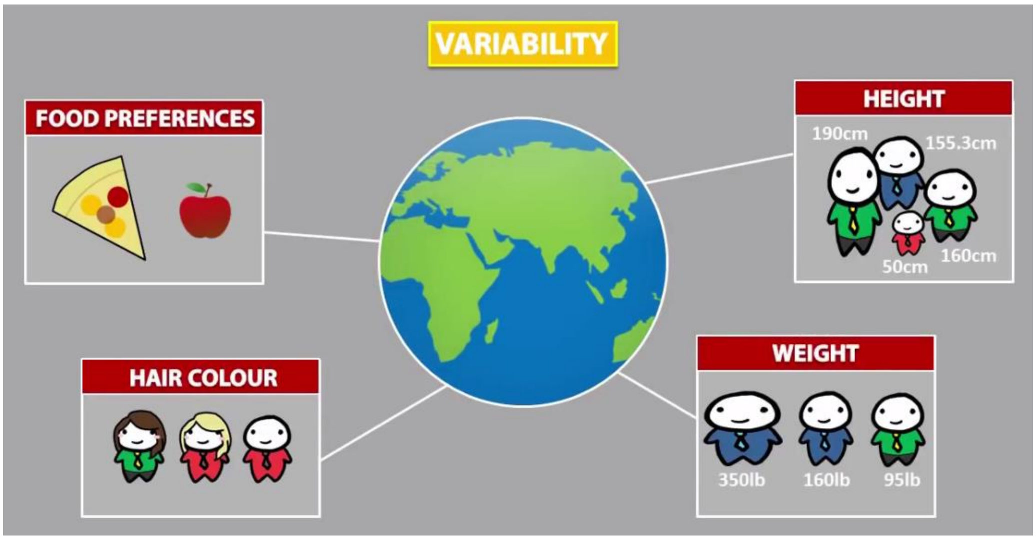
1. Maths - Statistics - Part - 1

Contents

1. Maths - Statistics - Part - 1

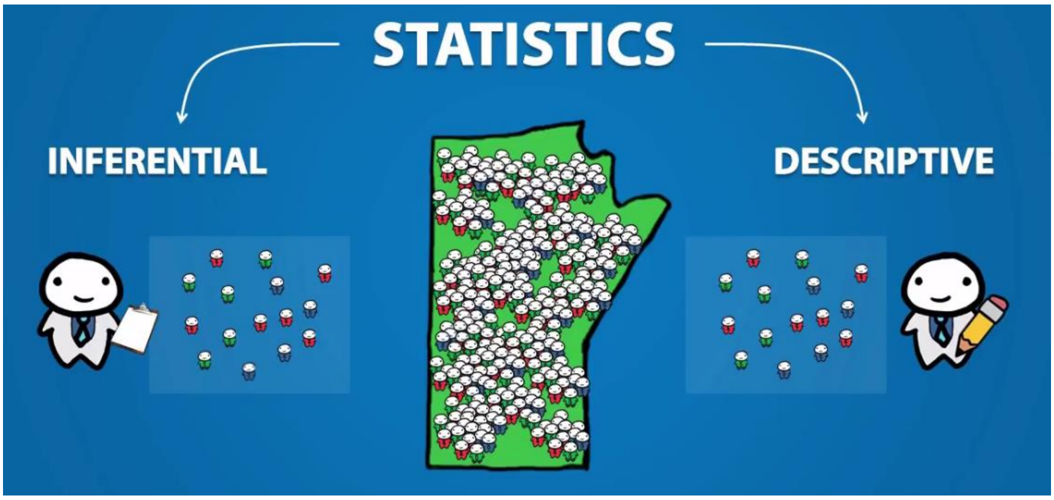
1. Statistics

- ✓ Statistics is the branch of mathematics dealing with the,
  - Data collection
  - Data Analysis,
  - Interpretation
  - Data presentation
  - Organizing the numerical data.



2. Types of statistics

- ✓ Inferential statistics
- ✓ Descriptive statistics



2.1. Inferential statistics

- ✓ Many times, a collection of the entire data is impossible.
- ✓ Hence a subset of the data points is collected.
- ✓ From the subset we can get conclusions about the entire population.
- ✓ This is called as inferential statistics.

2.2. Descriptive statistics

- ✓ These are used to summarize data, such as the mean, standard deviation & etc

Kind Info

- ✓ While applying statistics on data we can find underlying hidden relationships in between the variables

3. Population

- ✓ This point speaks about totality means all values or complete list of observations
- ✓ All the data points about the subject under study
    - o It can be people, vehicles, sales, cats, houses & etc

4. Sample

- ✓ A sample is a subset of a population, usually a small portion of the population that is being analysed.

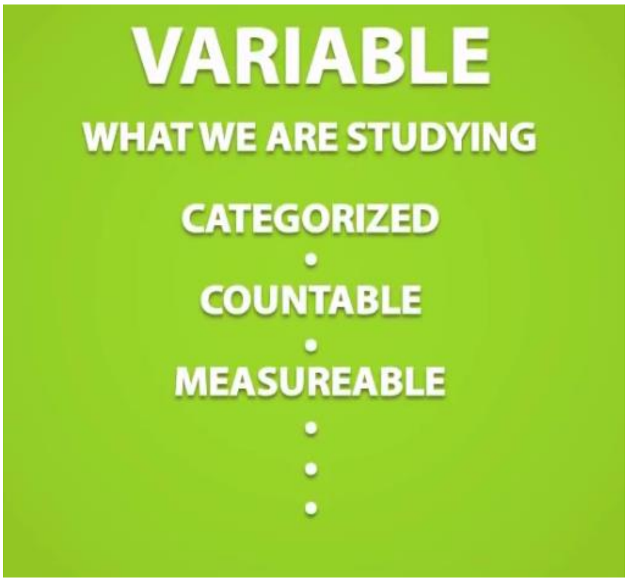
Info

- ✓ Usually, it is expensive to perform an analysis on an entire population.
- ✓ So, by analysing sample helps to draw the conclusions about a population.

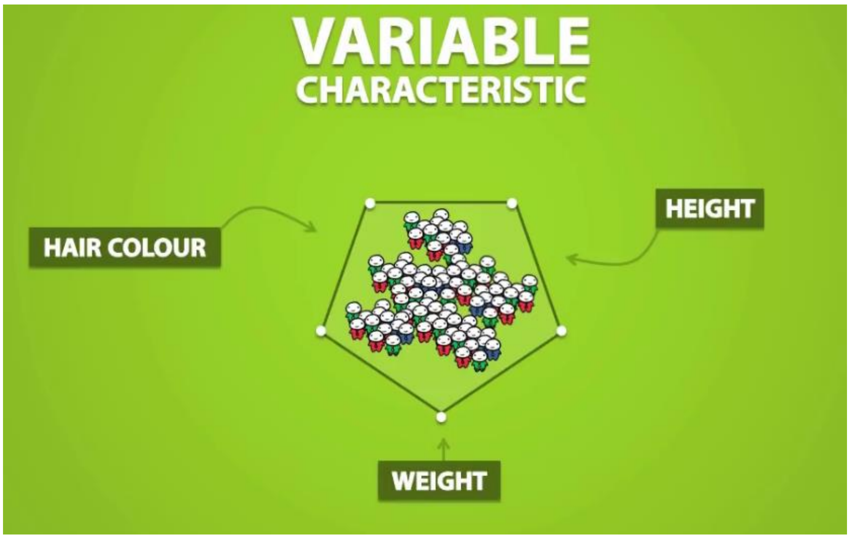
5. Variable

✓ In statistics what we are examining is a **variable**

✓ So, a variable is measurable, countable, categorized...



- ✓ People have different kind of heights, weights, and colors
- ✓ These are all variables



6. Variables means

- ✓ Variable represent the characteristic representation of study.
- ✓ A variable is an attribute
- ✓ A variable has a value
- ✓ A variable can store any type of data

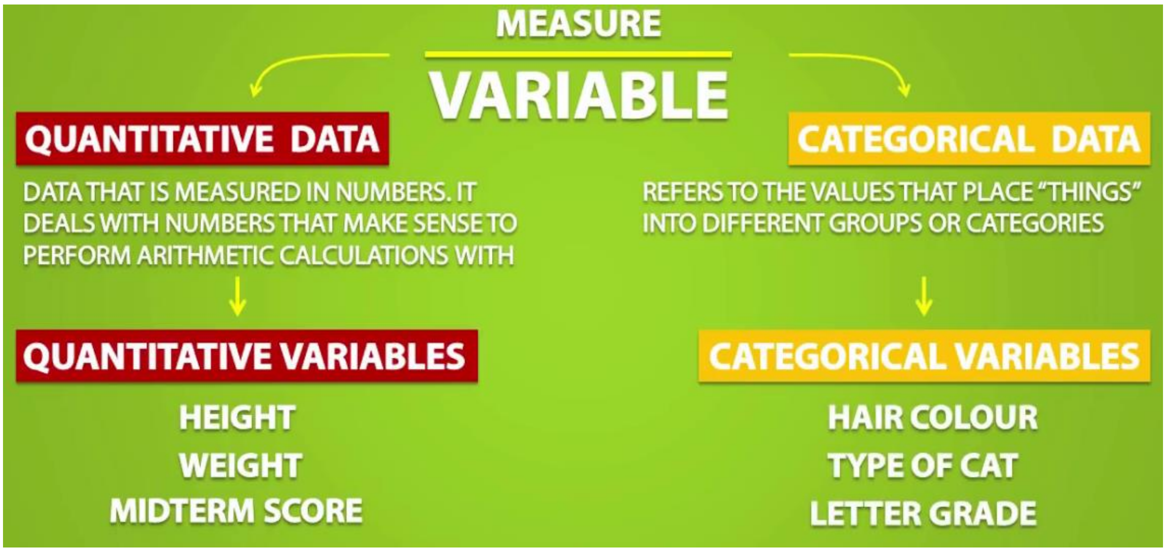
Example

✓ Person_name	=	Daniel
✓ Person_age	=	16
✓ Person_type	=	male
✓ Person_salary	=	16000
✓ Person_height	=	5.9

7. Types of variables

- ✓ Quantitative
  - Measured in numbers

- ✓ Categorical variable or Qualitative
  - Different group of categories



7.1. Quantitative variables

- ✓ Quantitative variables are numeric.
- ✓ We can do some kind of arithmetic calculations
- ✓ They represent a measurable quantity.
- ✓ Example
  - Height
  - Weight
  - Midterm score
  - Population of a city.
  - Salary and etc

7.2. Categorical or Qualitative variables

- ✓ Categorical variables stores values which are names or labels
- ✓ Example
  - Type of person : male or female
  - Review about food: good, bad, ok
  - Color of the bike : red, green, blue

8. Types of Categorical variable

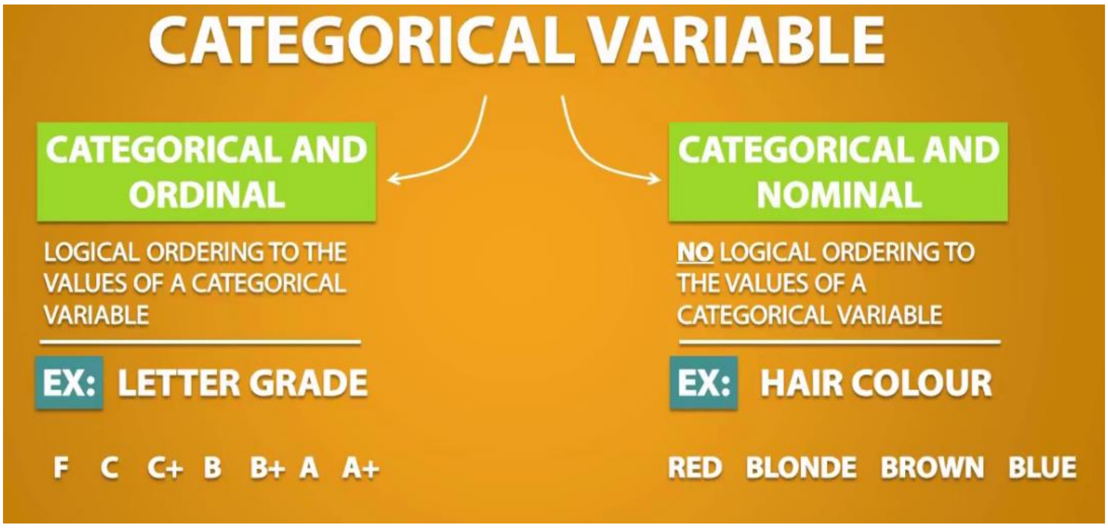
- ✓ There are two types of categorical variables
  - Ordinal
  - Nominal

8.1. Ordinal

- ✓ Logical order is possible to do analysis.
- ✓ We can logical order according to some dictionary format
- ✓ After this we can do some analysis
  - Example : Grades in exams

8.2. Nominal

- ✓ There is no logical ordering with respect to the actual values
  - Example : Hair color



9. Types of Quantitative variables

- ✓ Discrete variables
- ✓ Continuous variables

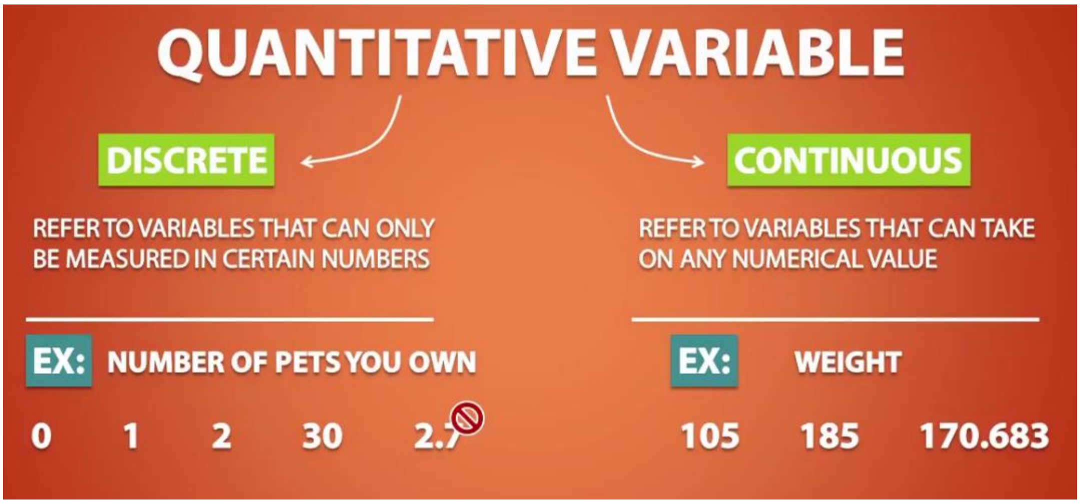
9.1. Discrete variables

- ✓ Discrete variables are like whole numbers.
- ✓ Example:
  - Number pets own
  - Number of people in a family
  - Number of bikes or cars

9.2. Continuous variables

- ✓ Continuous variables are like normal numbers includes floating point numbers.

- ✓ Example:
  - Weight
  - Salary
  - Bank balance



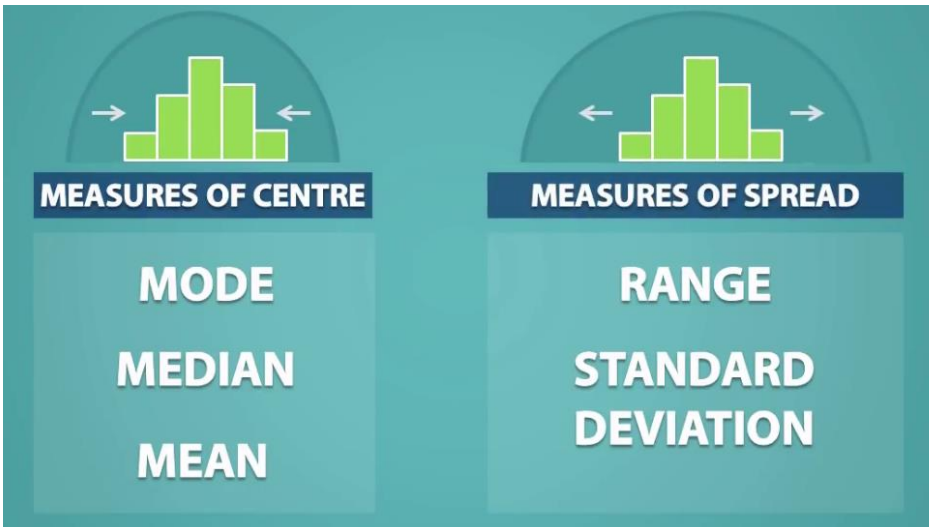
2. Maths - Statistics – PART – 2

Contents

2. Maths - Statistics – PART – 2

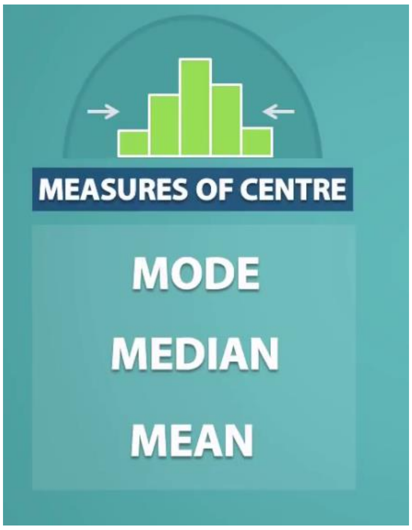
1. Usage of mode, median, mean, range & standard deviation?

- ✓ Identifying and describing like how the dataset got distributed
- ✓ These mode, median, mean, range and standard deviation gives the numerical information and distribution about the dataset
- ✓ These also explains about
    - o Measures of centre
    - o Measures of spread



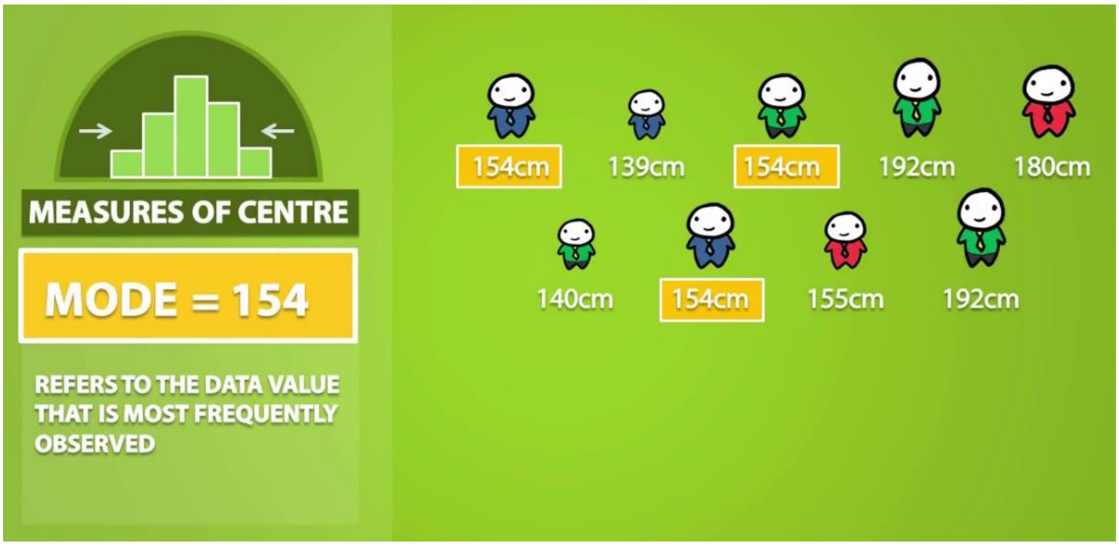
2. Measures of centre

- ✓ Mode
- ✓ Median
- ✓ Mean



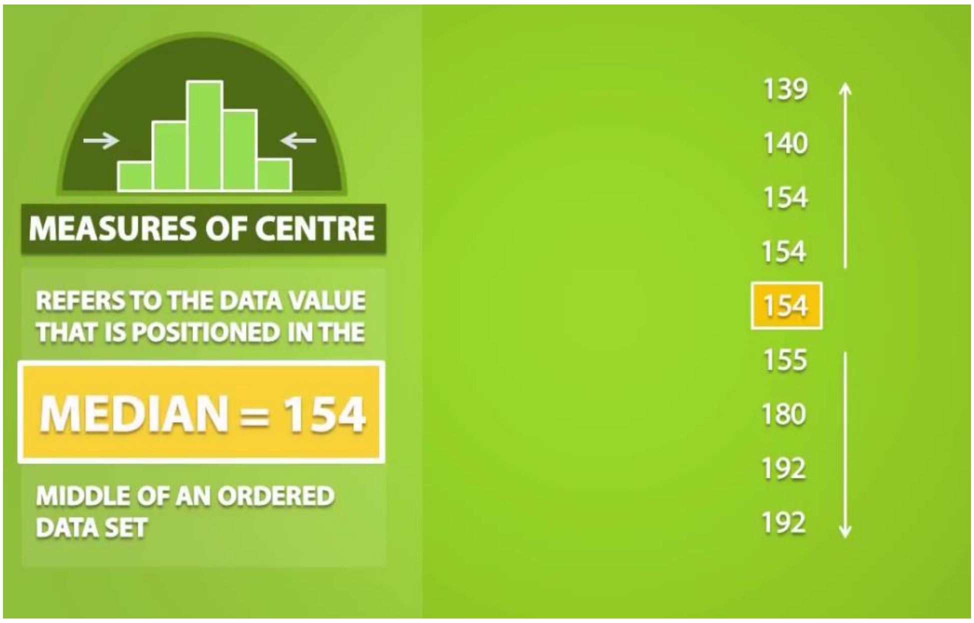
3. Mode

- ✓ Value which is most frequently observed
- ✓ Suppose we have taken random people heights and displayed as below
- ✓ Here 153 value is repeated in 3 times



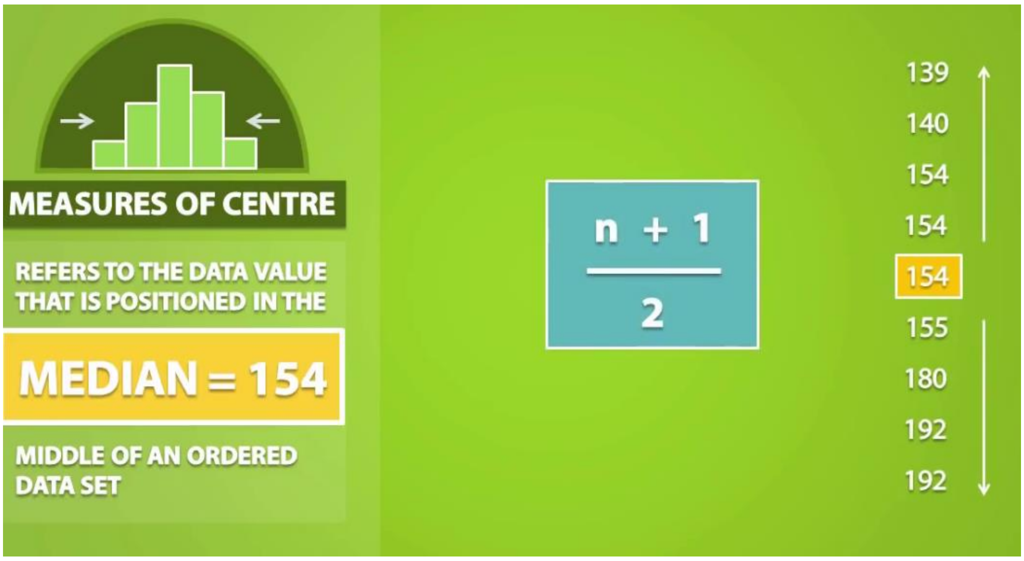
4. Median

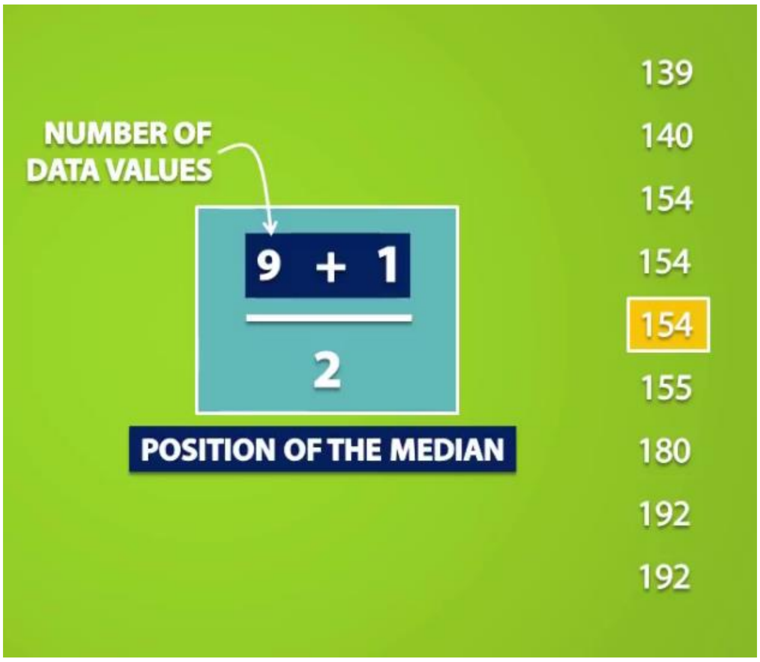
- ✓ Value that is positioned in the middle of an ordered dataset
- ✓ First we need to keep the data into an order
- ✓ We usually order the dataset into smallest to largest

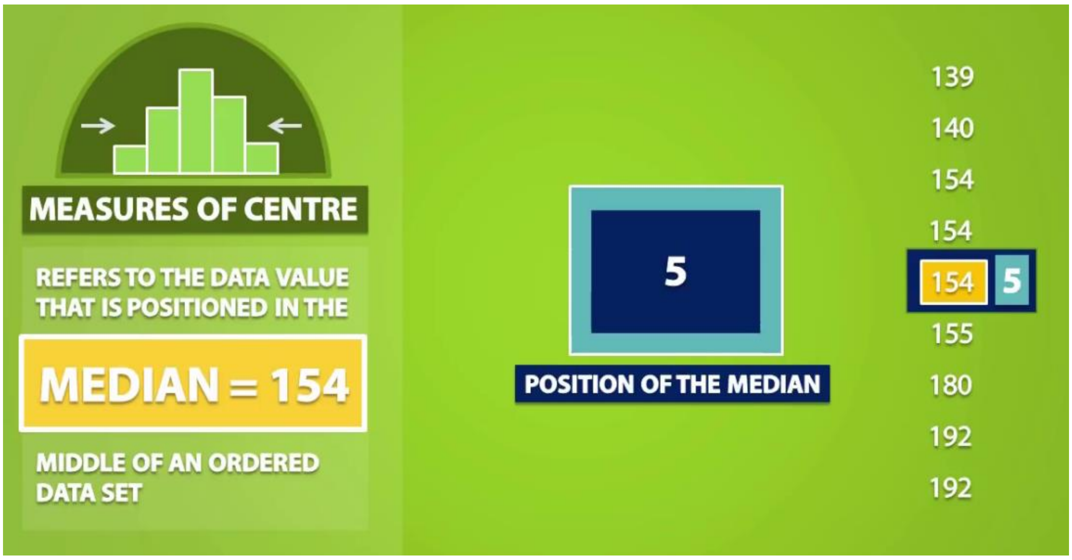


Special cases

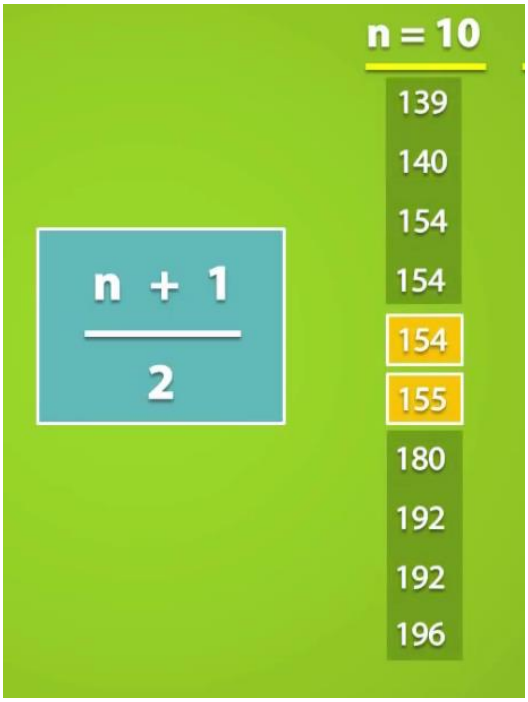
- ✓ If the dataset is extremely large then it might helpful use the below formula



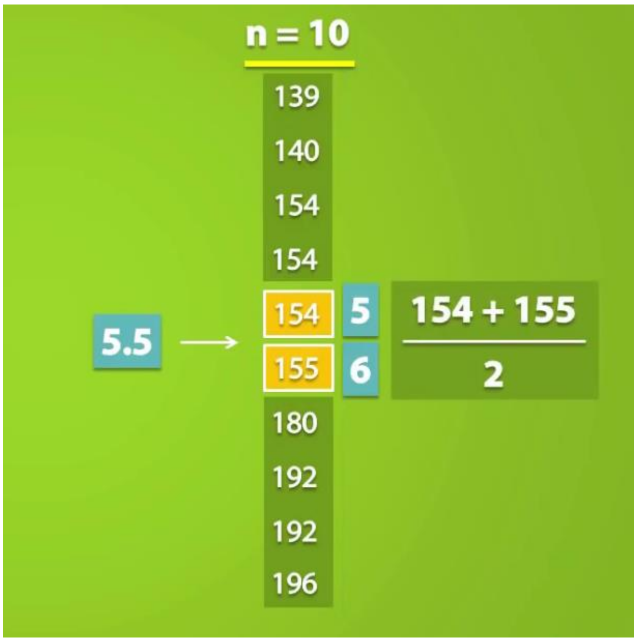



MEASURES OF CENTRE
REFERS TO THE DATA VALUE THAT IS POSITIONED IN THE
MEDIAN = 154
MIDDLE OF AN ORDERED DATA SET
5
POSITION OF THE MEDIAN
139
140
154
154
154
5
155
180
192
192

✓ If number of values are in odd or even number then we do have some special scenarios to find out the median value




n = 10
139
140
154
154
154
155
180
192
192
196
1
2
3
4
5
6
7
8
9
10
5.5


n = 10
139
140
154
154
154
155
180
192
192
196
5
6
5.5
154 + 155
2


n = 10
139
140
154
154
154
155
180
192
192
196
5
6
5.5
MEDIAN
154.5

6. Mean

- ✓ The mean is just another name of average
- ✓ Below is the formula which indicates mean of total values

$$\text{MEAN} = \frac{\sum x_i}{n}$$

SUMMATION OF ALL DATA VALUES

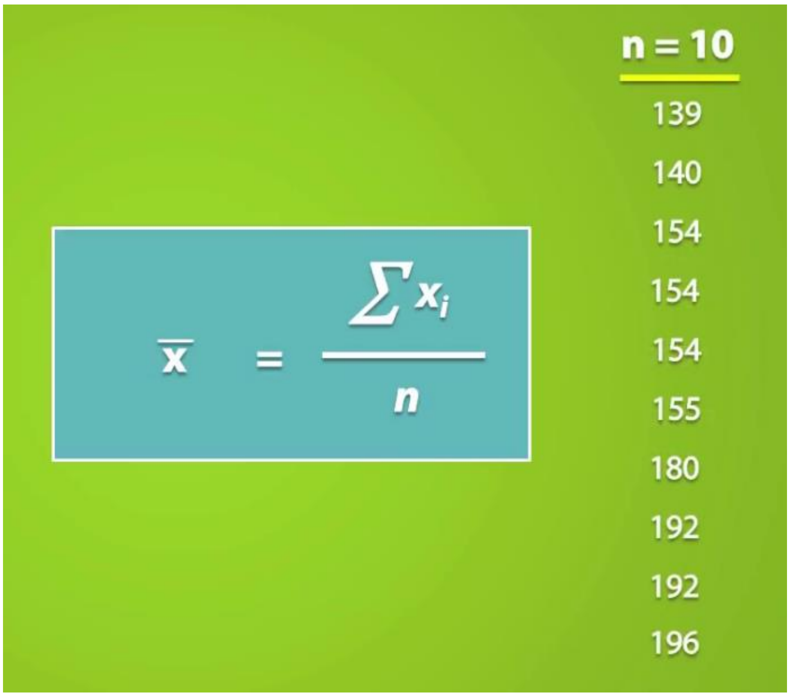
DIVIDED BY

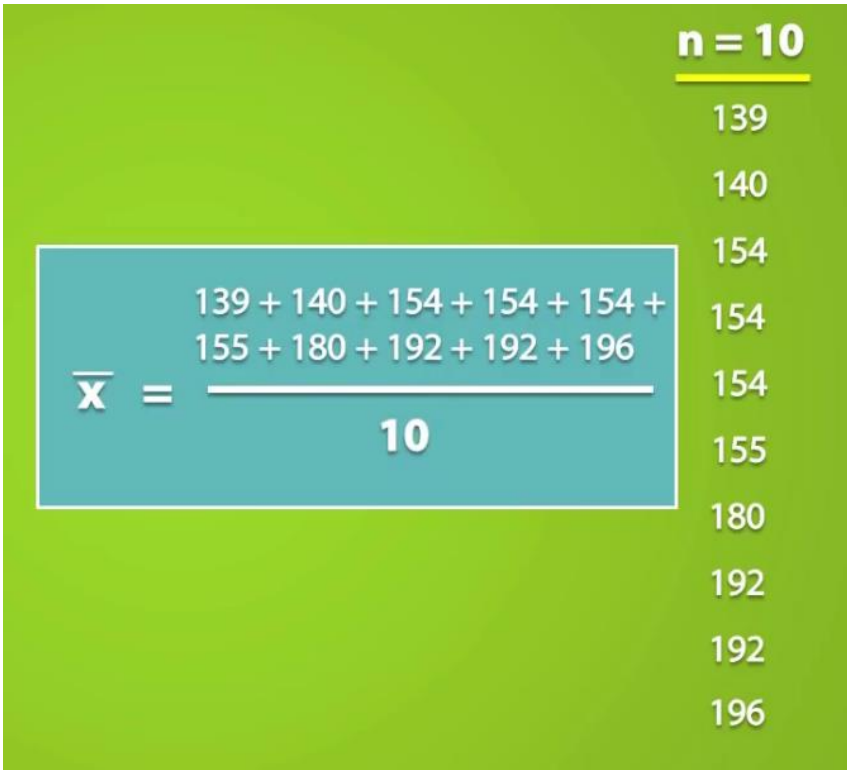
TOTAL NUMBER OF DATA VALUES

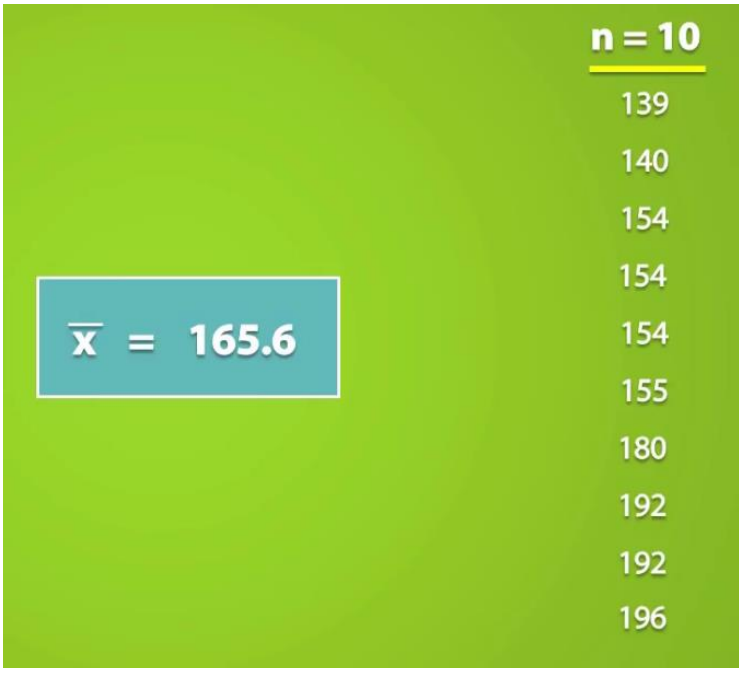
Sample mean

- ✓ Below is the formula which indicates mean of sample values

$$\overline{x} = \frac{\sum x_i}{n}$$


n = 10
139
140
154
154
154
155
180
192
192
196
$\bar{x} = \frac{\sum x_i}{n}$


n = 10
139
140
154
154
154
155
180
192
192
196
$\bar{x} = \frac{139 + 140 + 154 + 154 + 154 + 155 + 180 + 192 + 192 + 196}{10}$

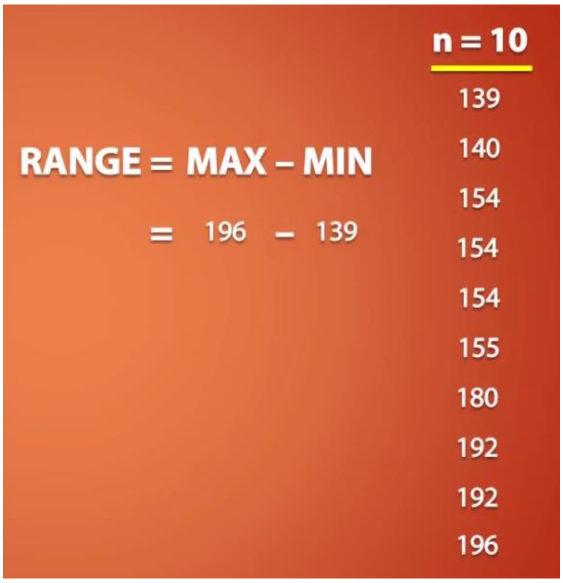

n = 10
139
140
154
154
154
155
180
192
192
196
$\overline{x}$ = 165.6

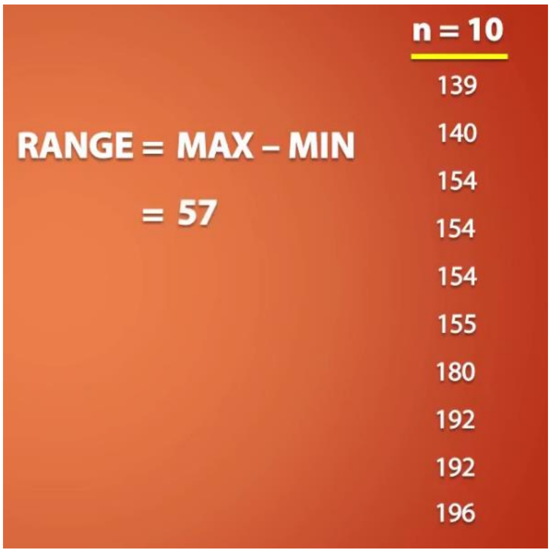
7. Measures of spread

- ✓ Range
- ✓ Standard deviation

8. Range

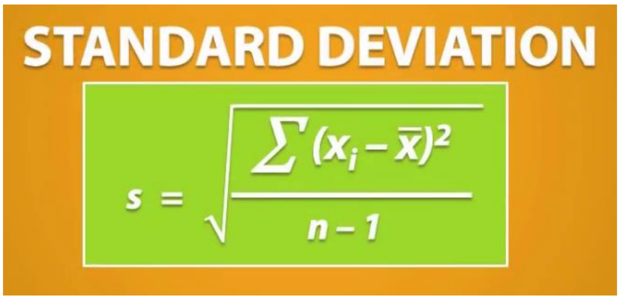
- ✓ Range means difference in between minimum value and maximum value
- ✓ It explains about the data is in between min and max values

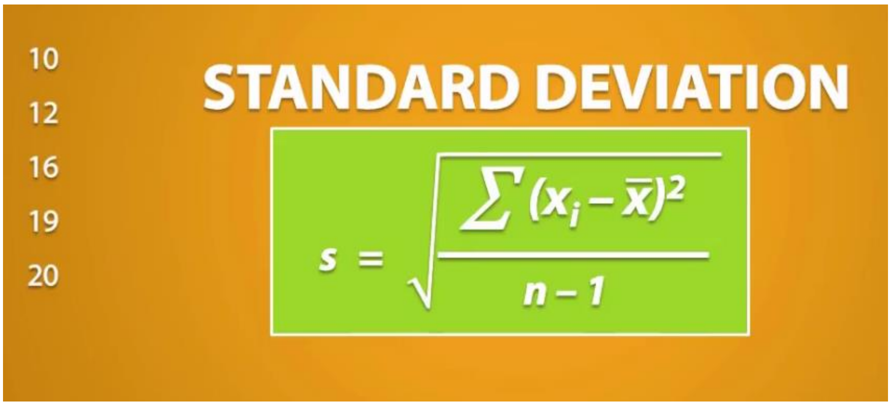


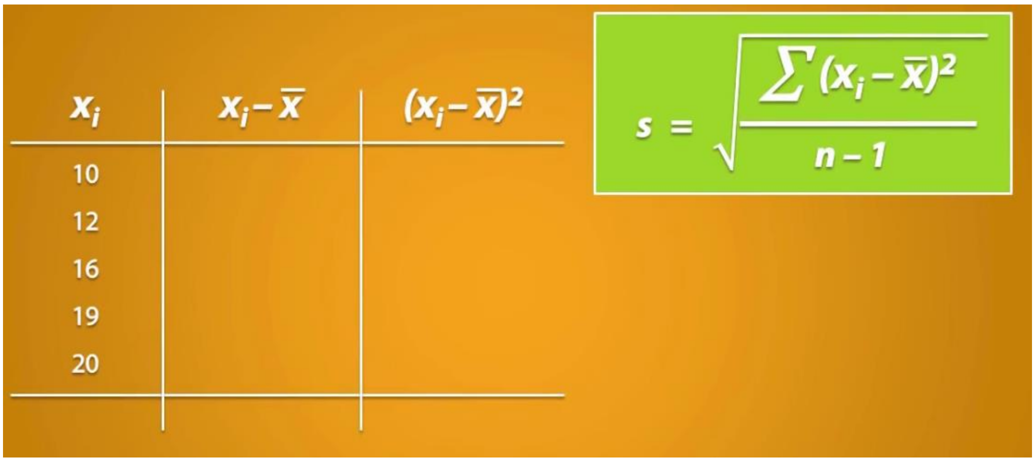


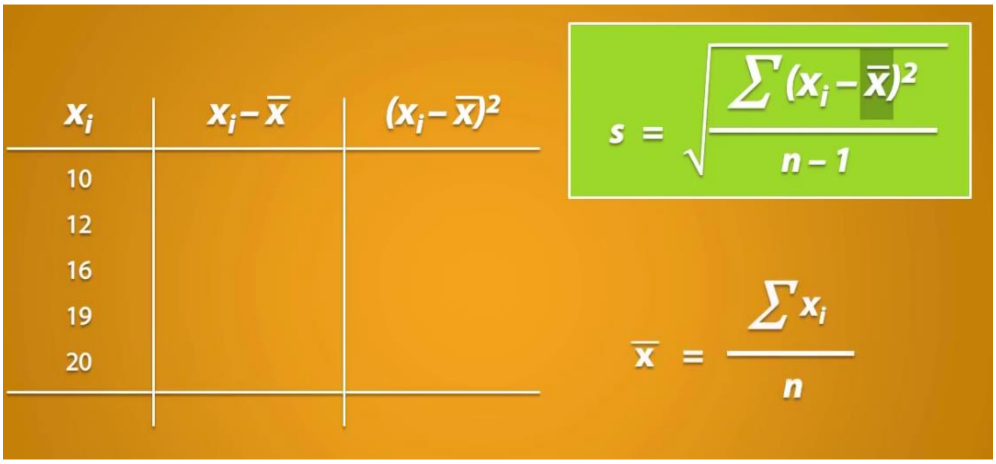
9. Standard Deviation

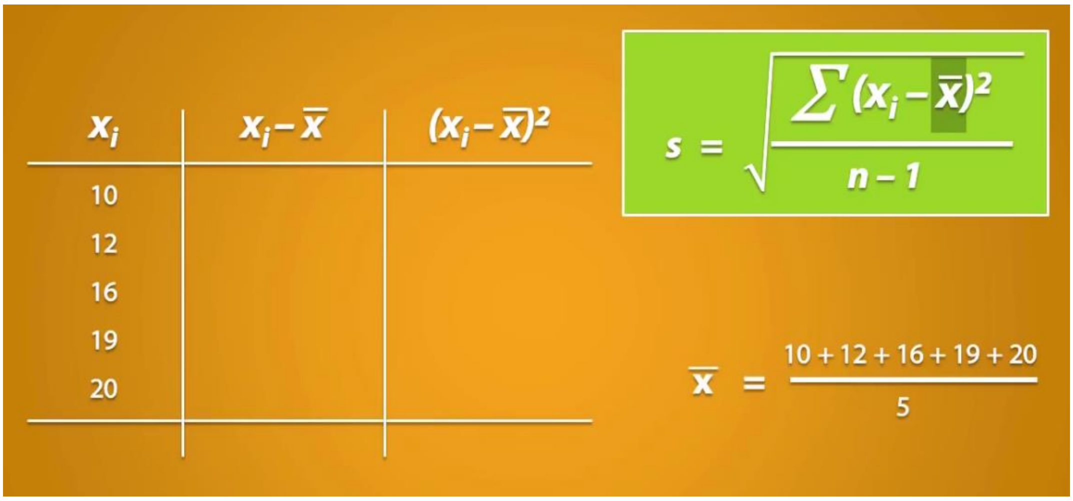
- ✓ The Standard Deviation is a measure of how spread out numbers.
- ✓ Formula is very simple,   It is the square root of the Variance

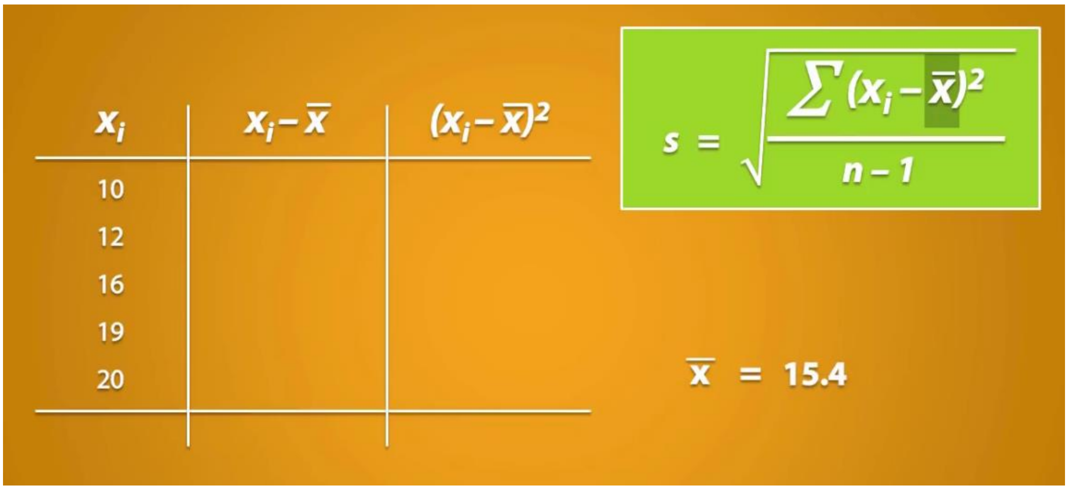


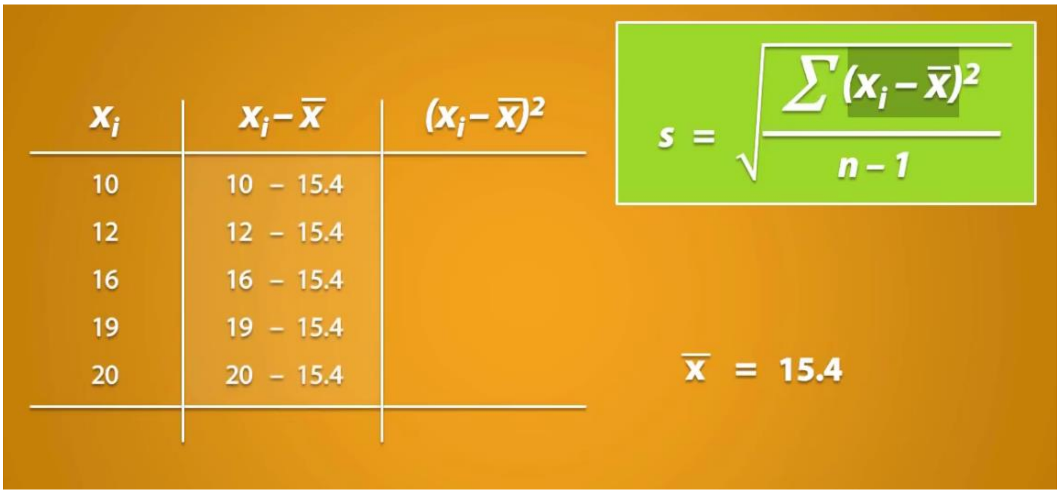





x_i
$x_i - \overline{x}$
$(x_i - \overline{x})^2$
10
12
16
19
20
$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$
$\overline{x} = \frac{\sum x_i}{n}$


x_i
$x_i - \overline{x}$
$(x_i - \overline{x})^2$
10
12
16
19
20
$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$
$\overline{x} = \frac{10 + 12 + 16 + 19 + 20}{5}$


x_i
$x_i - \overline{x}$
$(x_i - \overline{x})^2$
10
12
16
19
20
$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$
$\overline{x} = 15.4$


x_i
$x_i - \overline{x}$
$(x_i - \overline{x})^2$
10
12
16
19
20
10 – 15.4
12 – 15.4
16 – 15.4
19 – 15.4
20 – 15.4
$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$
$\overline{x} = 15.4$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
10	–5.4	
12	–3.4	
16	0.6	
19	3.6	
20	4.6	

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

$$\bar{x} = 15.4$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
10	–5.4	$(-5.4)^2$
12	–3.4	$(-3.4)^2$
16	0.6	$(0.6)^2$
19	3.6	$(3.6)^2$
20	4.6	$(4.6)^2$

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

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
10	–5.4	29.16
12	–3.4	11.56
16	0.6	0.36
19	3.6	12.96
20	4.6	21.16

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

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
10	–5.4	29.16
12	–3.4	11.56
16	0.6	0.36
19	3.6	12.96
20	4.6	21.16
		SUM = 75.2

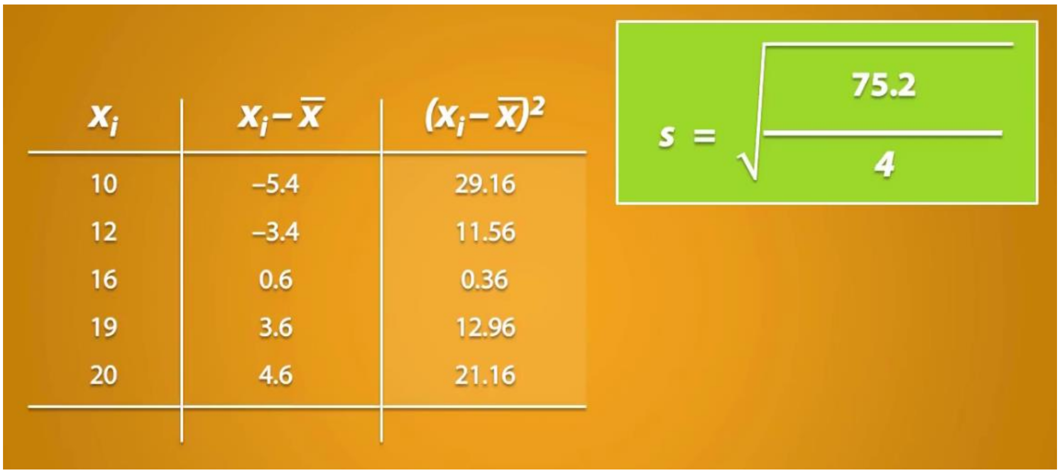
$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
10	–5.4	29.16
12	–3.4	11.56
16	0.6	0.36
19	3.6	12.96
20	4.6	21.16

$$s = \sqrt{\frac{75.2}{n-1}}$$

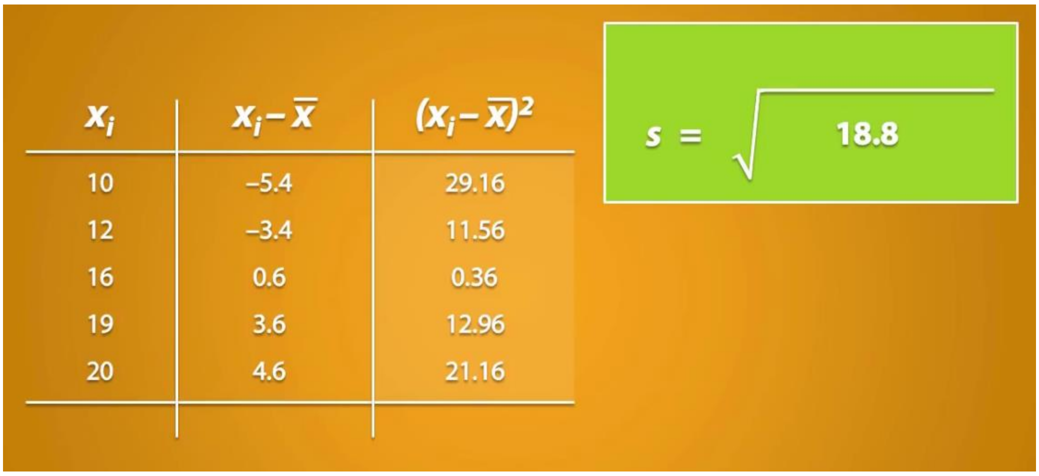
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
10	–5.4	29.16
12	–3.4	11.56
16	0.6	0.36
19	3.6	12.96
20	4.6	21.16

$$s = \sqrt{\frac{75.2}{5-1}}$$



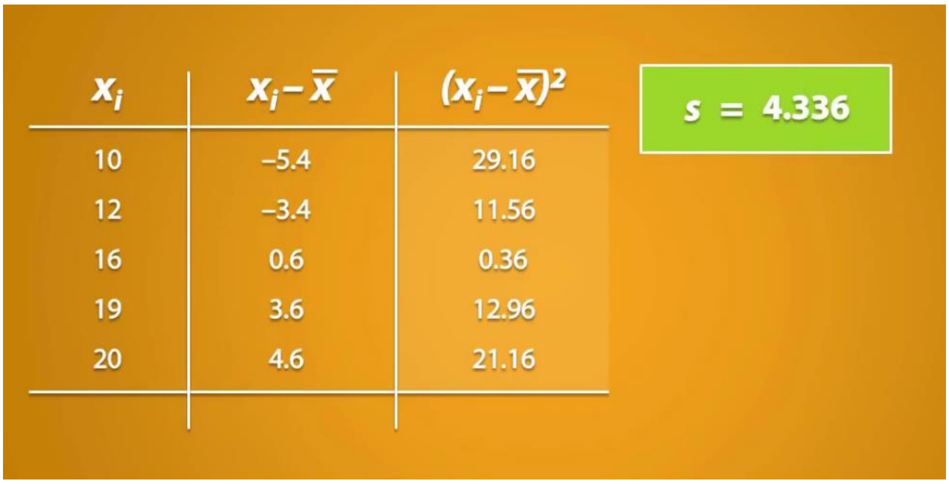
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
10	−5.4	29.16
12	−3.4	11.56
16	0.6	0.36
19	3.6	12.96
20	4.6	21.16

$$s = \sqrt{\frac{75.2}{4}}$$

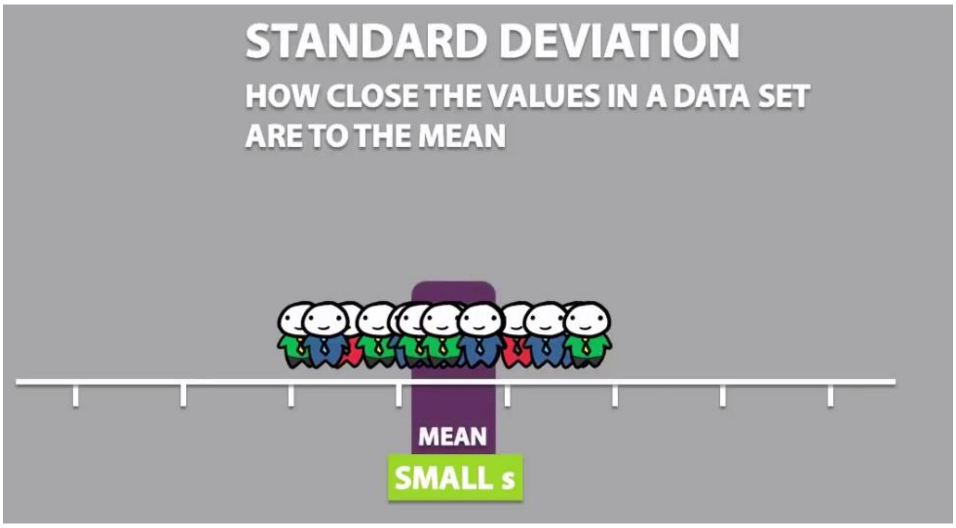


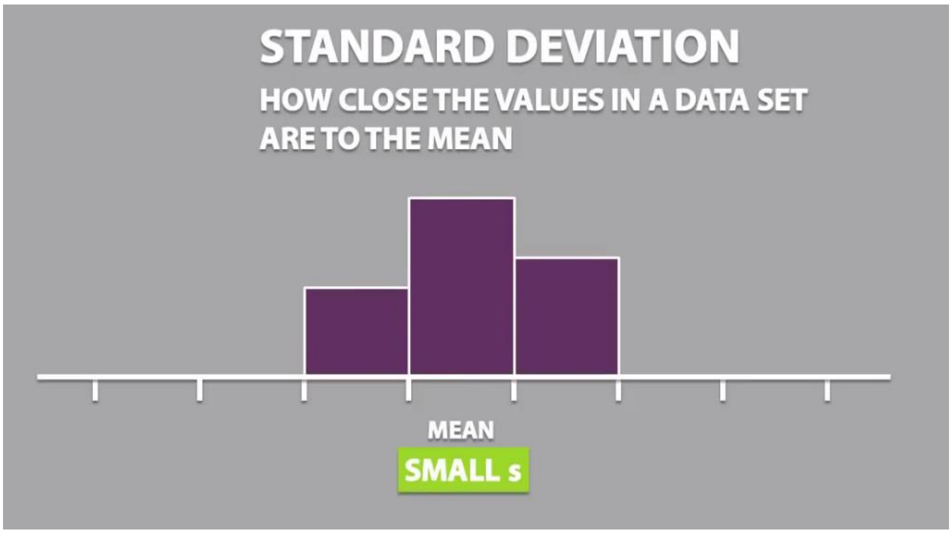
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
10	−5.4	29.16
12	−3.4	11.56
16	0.6	0.36
19	3.6	12.96
20	4.6	21.16

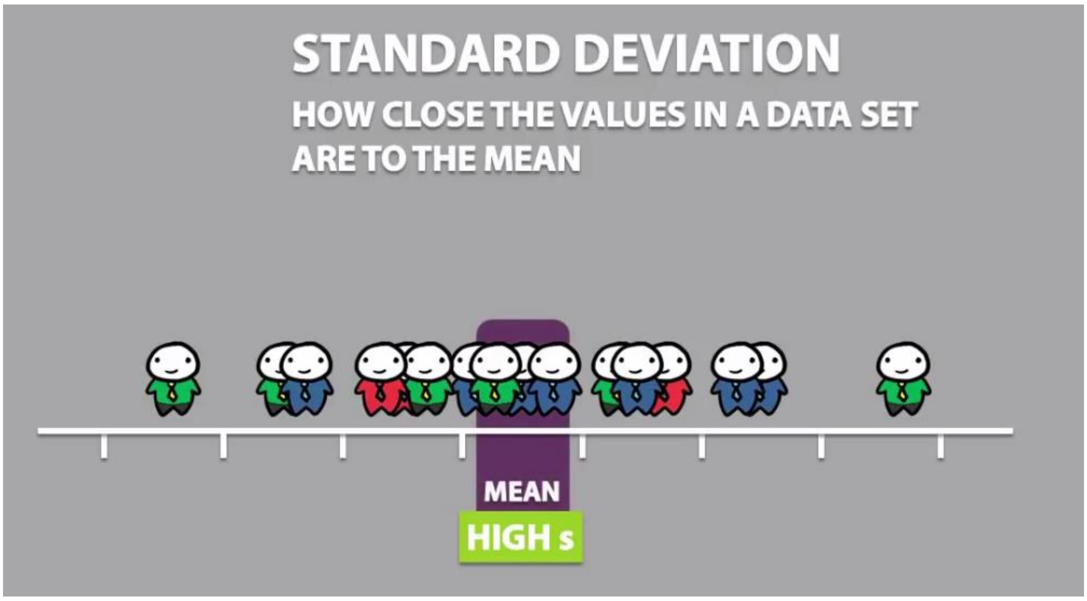
$$s = \sqrt{18.8}$$

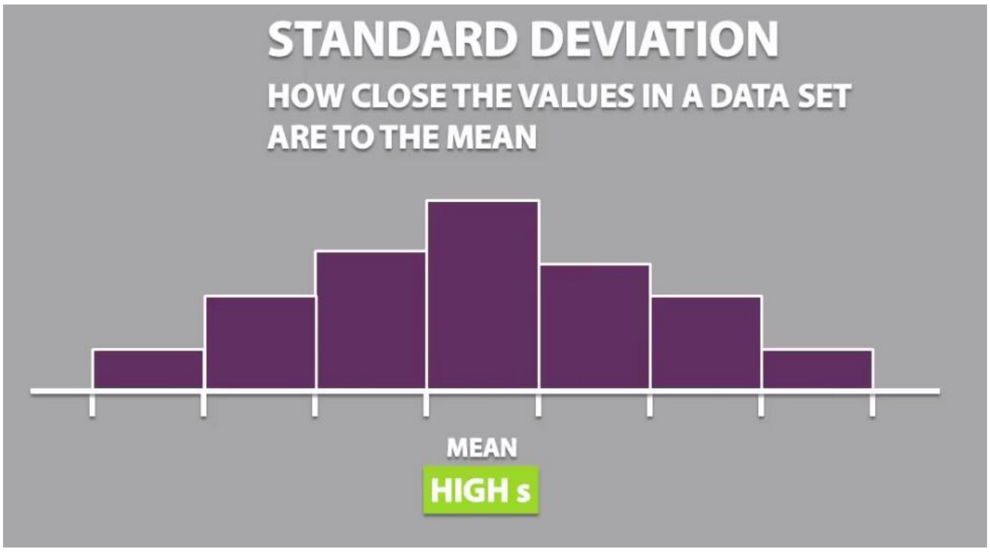


x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
10	−5.4	29.16
12	−3.4	11.56
16	0.6	0.36
19	3.6	12.96
20	4.6	21.16


STANDARD DEVIATION
HOW CLOSE THE VALUES IN A DATA SET ARE TO THE MEAN
MEAN
SMALL s

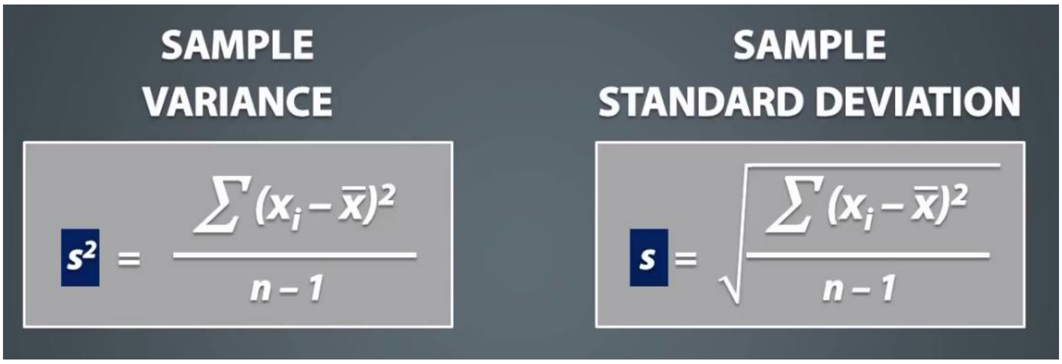

STANDARD DEVIATION
HOW CLOSE THE VALUES IN A DATA SET ARE TO THE MEAN
MEAN
SMALL s


STANDARD DEVIATION
HOW CLOSE THE VALUES IN A DATA SET ARE TO THE MEAN
MEAN
HIGH s


STANDARD DEVIATION
HOW CLOSE THE VALUES IN A DATA SET ARE TO THE MEAN
MEAN
HIGH s

10. Variance

- ✓ The average of squared differences from the mean.
- ✓ Variance is the average of squared differences from the mean
- ✓ By using this we can find how far the data points in a population are from the population mean.



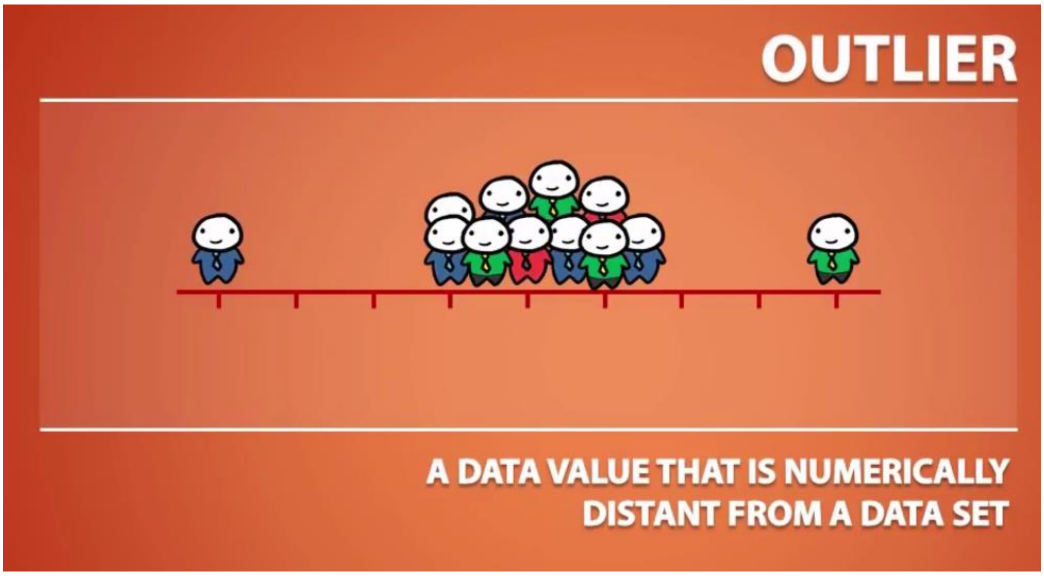
3. Maths - Statistics – PART – 3

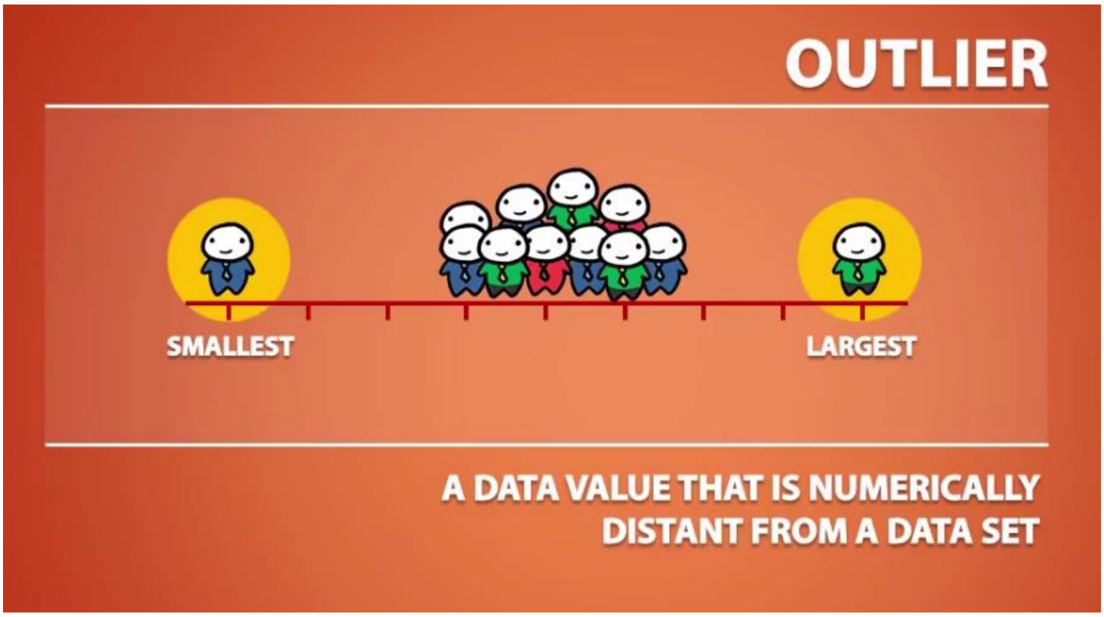
Contents

3. Maths - Statistics – PART – 3

1. What is an outlier?

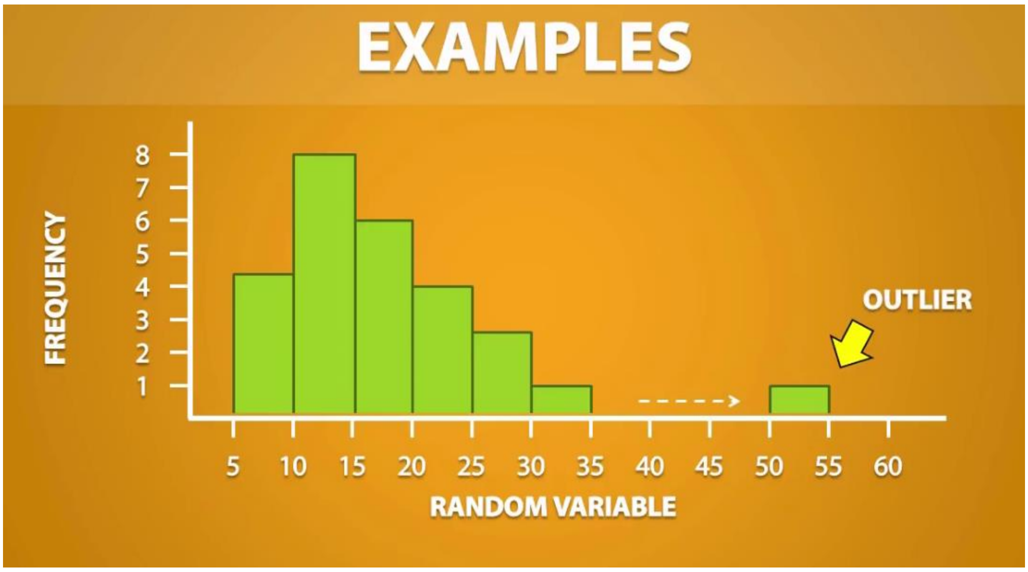
- ✓ An outlier is defined as a value which are very far from dataset
- ✓ An outlier is a data point that falls outside from main data points
- ✓ It can be largest value in dataset, smallest value in dataset





Examples

- ✓ Observe the below histogram, a point is far from value

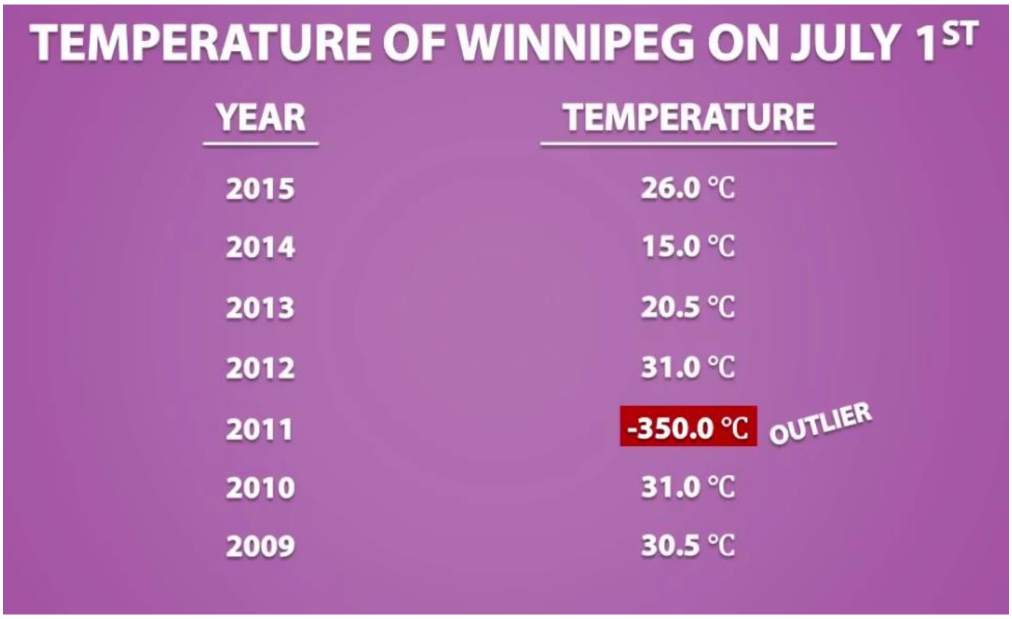


✓ Below example, 9000 is very larger value than other values

✓ Below example, 3 is very smaller value than other values

2. Surprising...!!!

- ✓ Outliers are data points but these are typical and surprising.
- ✓ These effects the measures of center and spread
- ✓ Observe the below example



TEMPERATURE OF WINNIPEG ON JULY 1ST

YEAR	TEMPERATURE
2015	26.0 °C
2014	15.0 °C
2013	20.5 °C
2012	31.0 °C
2011	-350.0 °C OUTLIER
2010	31.0 °C
2009	30.5 °C

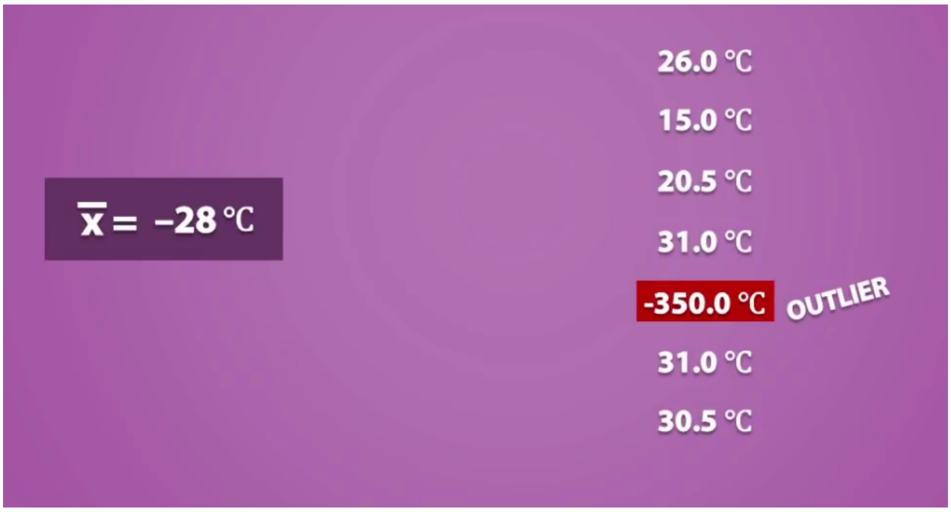
THE MEAN IS AFFECTED BY THE PRESENCE OF OUTLIERS

$$\bar{x} = \frac{\Sigma x_i}{n}$$

26.0 °C
15.0 °C
20.5 °C
31.0 °C
-350.0 °C
31.0 °C
30.5 °C

$$\bar{x} = \frac{26 + 15 + 20.5 + 31 + (-350) + 31 + 30.5}{7}$$

26.0 °C
15.0 °C
20.5 °C
31.0 °C
-350.0 °C
31.0 °C
30.5 °C

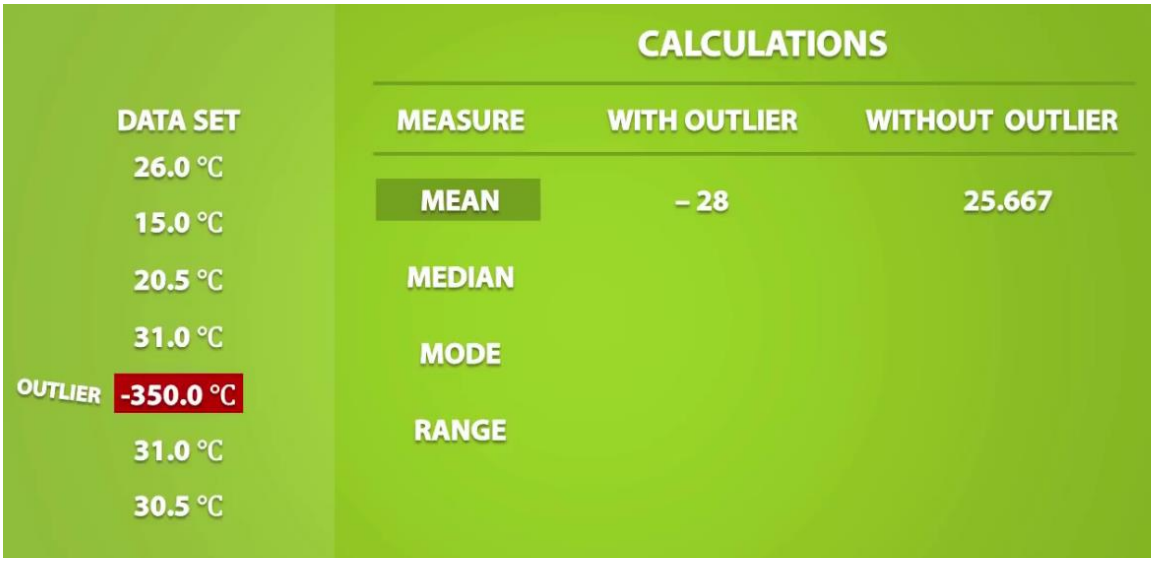

$\bar{x}$ = –28 °C
26.0 °C
15.0 °C
20.5 °C
31.0 °C
-350.0 °C
OUTLIER
31.0 °C
30.5 °C

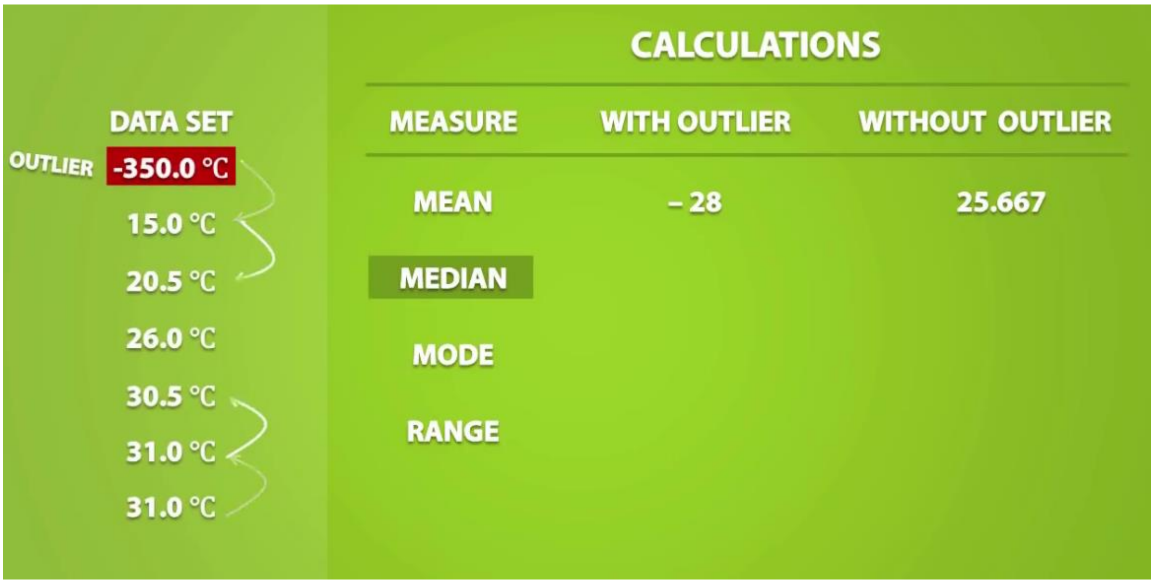
So

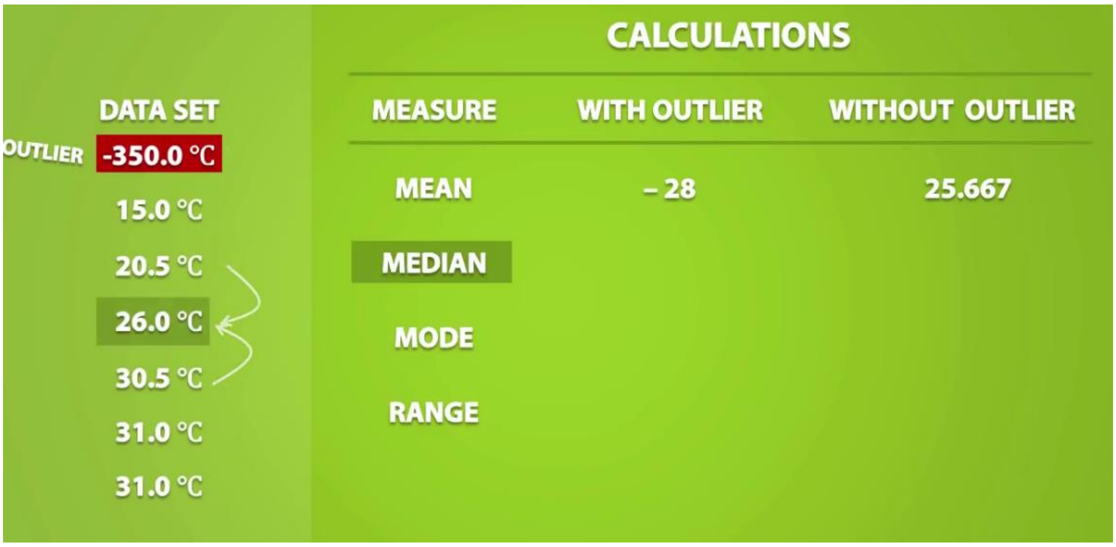

THE MEAN IS AFFECTED BY THE PRESENCE OF OUTLIERS

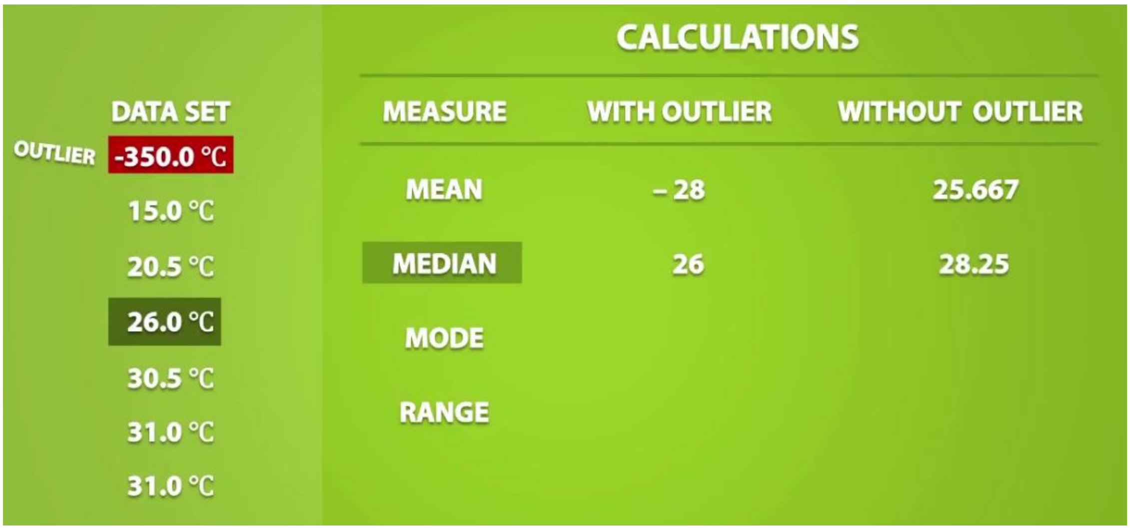
3. Checking mean, median, mode & Range

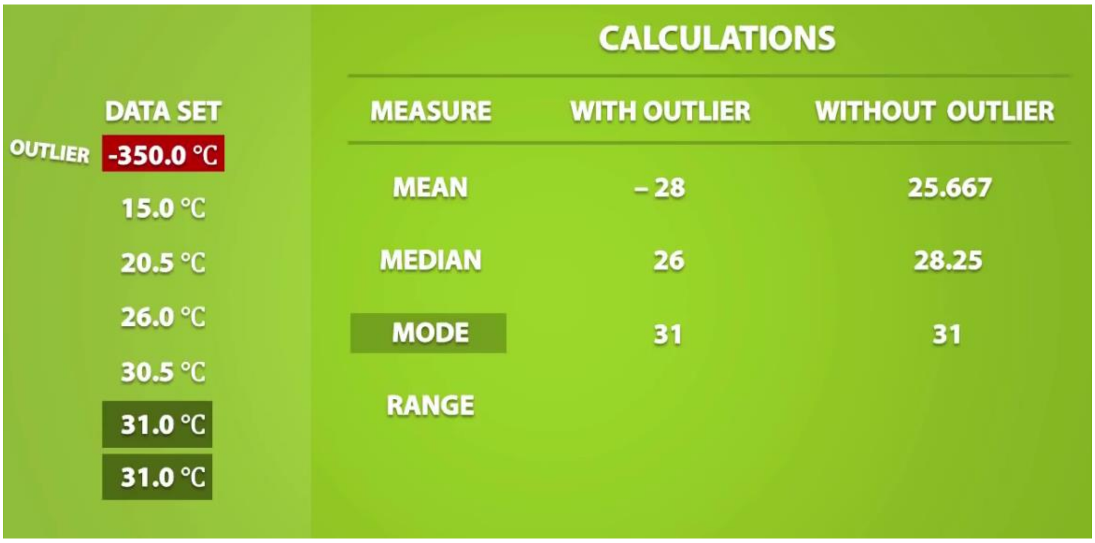
✓ Calculate mean, median, mode and Range for a dataset

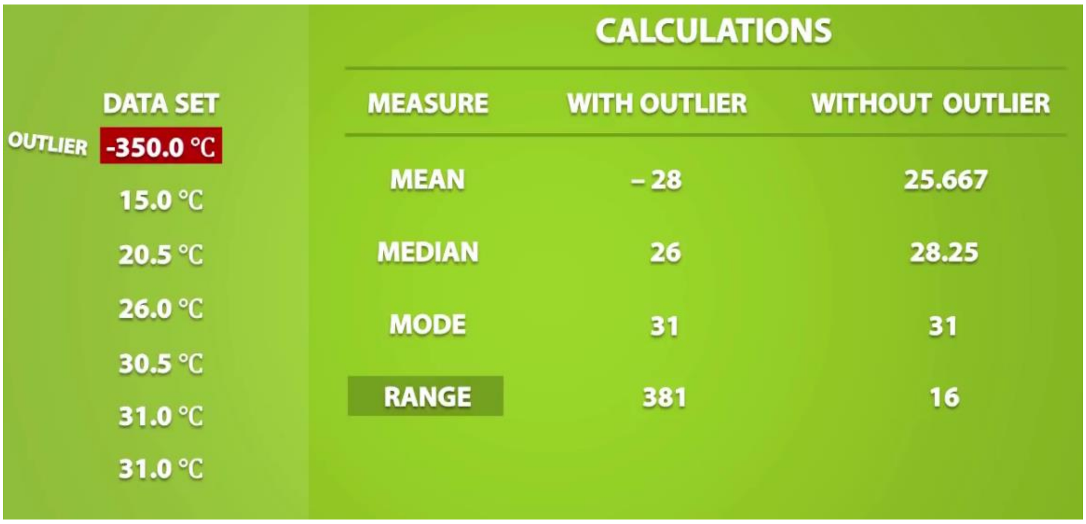





CALCULATIONS
DATA SET
OUTLIER
-350.0 °C
15.0 °C
20.5 °C
26.0 °C
30.5 °C
31.0 °C
31.0 °C
MEASURE
WITH OUTLIER
WITHOUT OUTLIER
MEAN
– 28
25.667
MEDIAN
MODE
RANGE


CALCULATIONS
DATA SET
OUTLIER
-350.0 °C
15.0 °C
20.5 °C
26.0 °C
30.5 °C
31.0 °C
31.0 °C
MEASURE
WITH OUTLIER
WITHOUT OUTLIER
MEAN
– 28
25.667
MEDIAN
26
28.25
MODE
RANGE

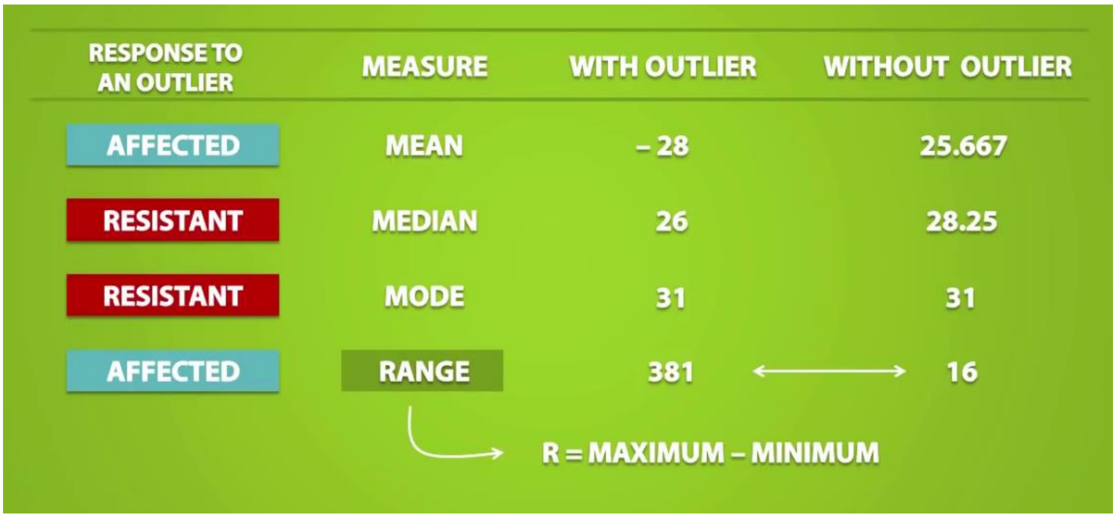

CALCULATIONS
DATA SET
OUTLIER
-350.0 °C
15.0 °C
20.5 °C
26.0 °C
30.5 °C
31.0 °C
31.0 °C
MEASURE
WITH OUTLIER
WITHOUT OUTLIER
MEAN
– 28
25.667
MEDIAN
26
28.25
MODE
31
31
RANGE


CALCULATIONS
DATA SET
OUTLIER
-350.0 °C
15.0 °C
20.5 °C
26.0 °C
30.5 °C
31.0 °C
31.0 °C
MEASURE
WITH OUTLIER
WITHOUT OUTLIER
MEAN
– 28
25.667
MEDIAN
26
28.25
MODE
31
31
RANGE
381
16

✓ Is outlier affects the calculations: Yes then observe the below table

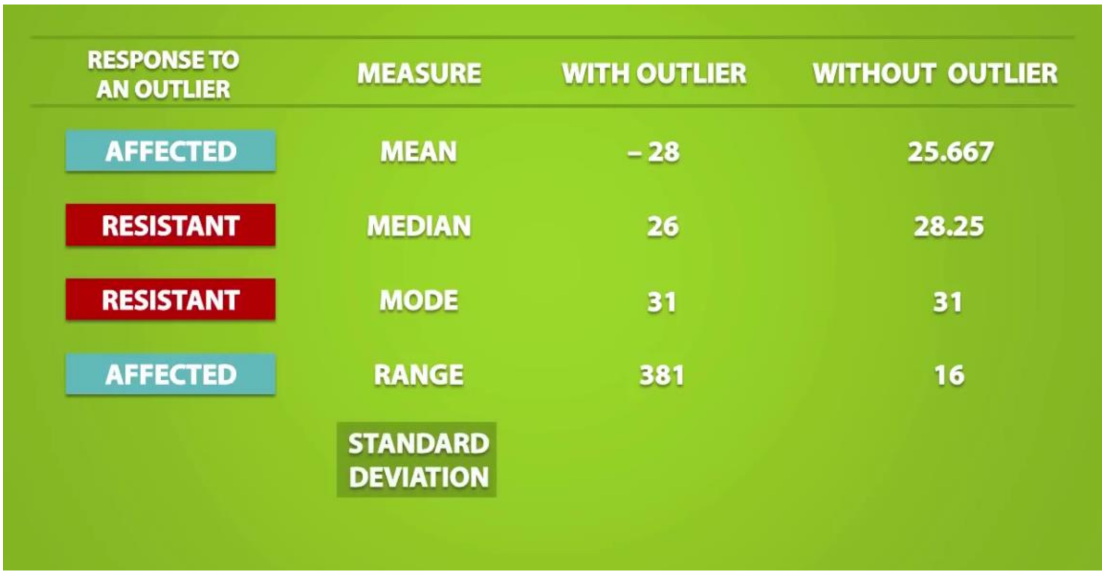
RESPONSE TO AN OUTLIER	MEASURE	WITH OUTLIER	WITHOUT OUTLIER
AFFECTED	MEAN	– 28	25.667
	MEDIAN	26	28.25
	MODE	31	31
	RANGE	381	16

RESPONSE TO AN OUTLIER	MEASURE	WITH OUTLIER	WITHOUT OUTLIER
AFFECTED	MEAN	– 28	25.667
RESISTANT	MEDIAN	26	28.25
RESISTANT	MODE	31	31
	RANGE	381	16

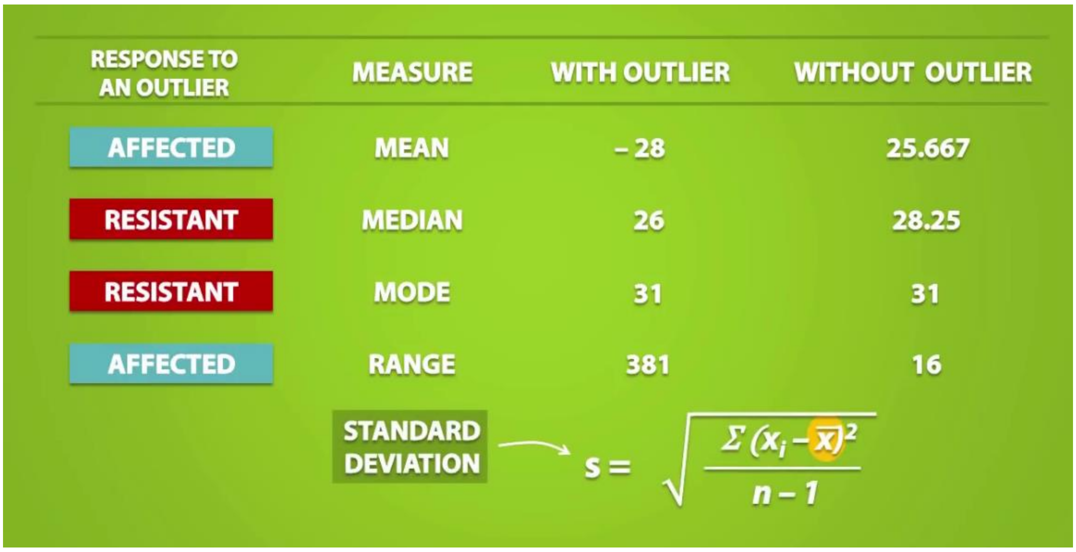


RESPONSE TO AN OUTLIER	MEASURE	WITH OUTLIER	WITHOUT OUTLIER
AFFECTED	MEAN	– 28	25.667
RESISTANT	MEDIAN	26	28.25
RESISTANT	MODE	31	31
AFFECTED	RANGE	381	16

R = MAXIMUM – MINIMUM

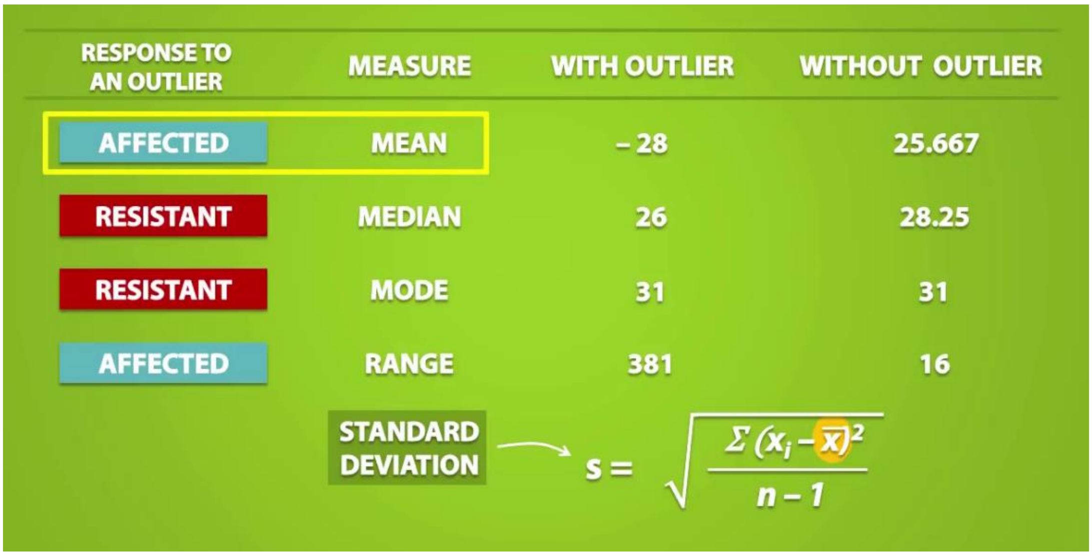


RESPONSE TO AN OUTLIER	MEASURE	WITH OUTLIER	WITHOUT OUTLIER
AFFECTED	MEAN	– 28	25.667
RESISTANT	MEDIAN	26	28.25
RESISTANT	MODE	31	31
AFFECTED	RANGE	381	16
	STANDARD DEVIATION		



RESPONSE TO AN OUTLIER	MEASURE	WITH OUTLIER	WITHOUT OUTLIER
AFFECTED	MEAN	– 28	25.667
RESISTANT	MEDIAN	26	28.25
RESISTANT	MODE	31	31
AFFECTED	RANGE	381	16

STANDARD DEVIATION $\rightarrow s = \sqrt{\dfrac{\Sigma (x_i - \bar{x})^2}{n-1}}$



RESPONSE TO AN OUTLIER	MEASURE	WITH OUTLIER	WITHOUT OUTLIER
AFFECTED	MEAN	– 28	25.667
RESISTANT	MEDIAN	26	28.25
RESISTANT	MODE	31	31
AFFECTED	RANGE	381	16

STANDARD DEVIATION $\rightarrow s = \sqrt{\dfrac{\Sigma (x_i - \bar{x})^2}{n-1}}$

RESPONSE TO AN OUTLIER	MEASURE	WITH OUTLIER	WITHOUT OUTLIER
AFFECTED	MEAN	– 28	25.667
RESISTANT	MEDIAN	26	28.25
RESISTANT	MODE	31	31
AFFECTED	RANGE	381	16
AFFECTED	STANDARD DEVIATION	$s = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}}$	

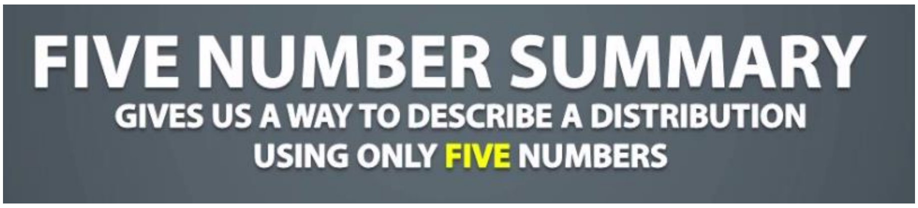
3. Maths - Statistics – PART – 4

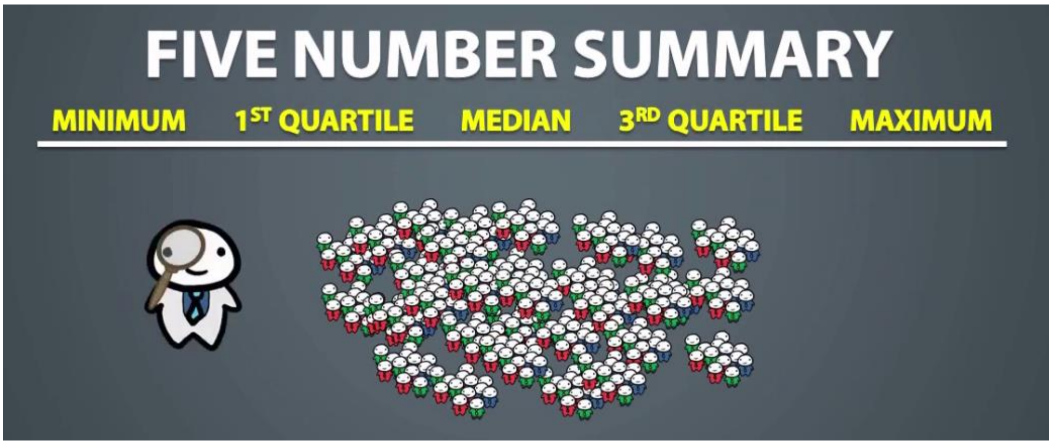
Contents

4. Maths - Statistics – PART – 4

1. What is five number summary?

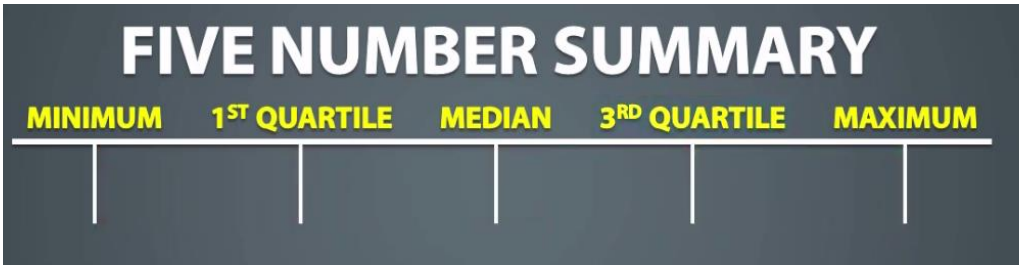
- ✓ The five number summary gives a way to describe the distribution



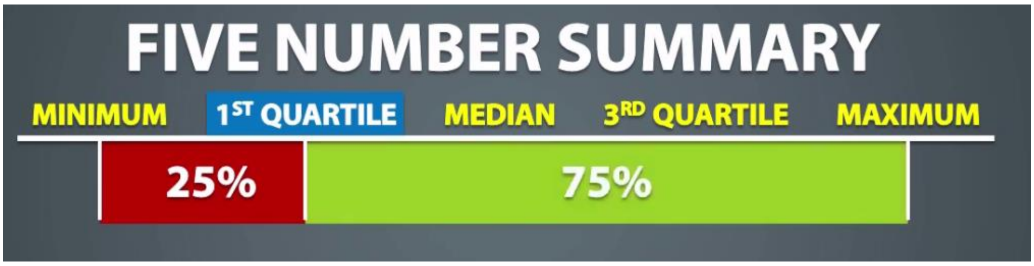


2. Explanation

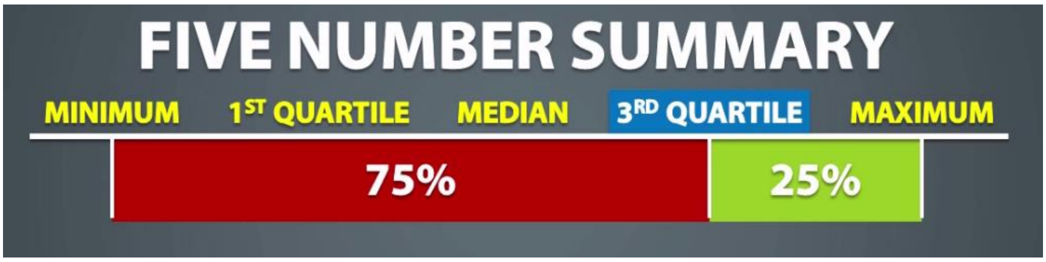
- ✓ So, if we took the data then we can find these five numbers in that data
- ✓ Minimum is smallest value in a dataset
- ✓ Maximum is largest value in a dataset
- ✓ Median is middle data value
    - o It is the point 50% data values is below the median and 50% data values is above the median



- ✓ The median of bottom half is called as 1st Quartile
    - o It is the position 25% of the data values are below this and 75% values are above the point
    - o The 1st Quartile is essentially as median of the median

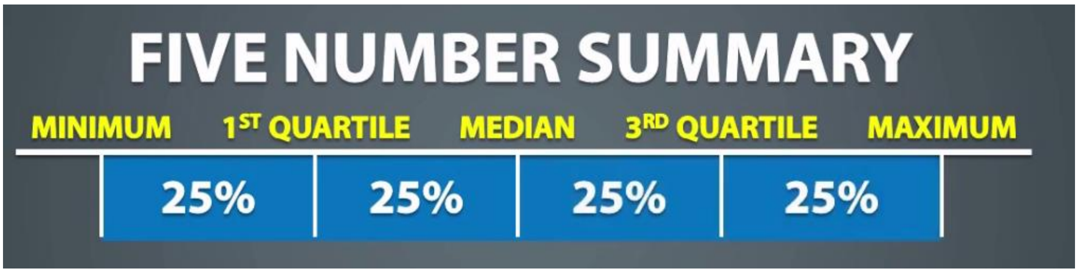


- ✓ The median of top half is called as 3rd Quartile
  - It is the position 75% of the data values are below this and 25% values are above the point

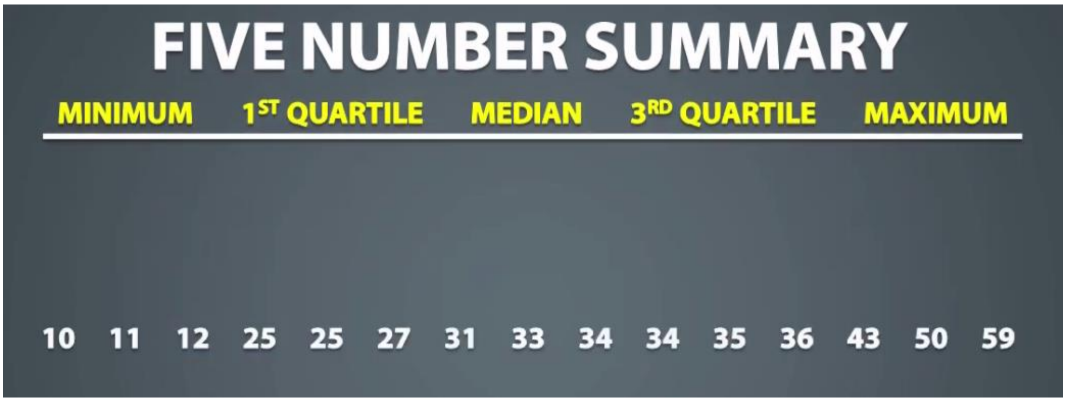


3. 4 equal quarters

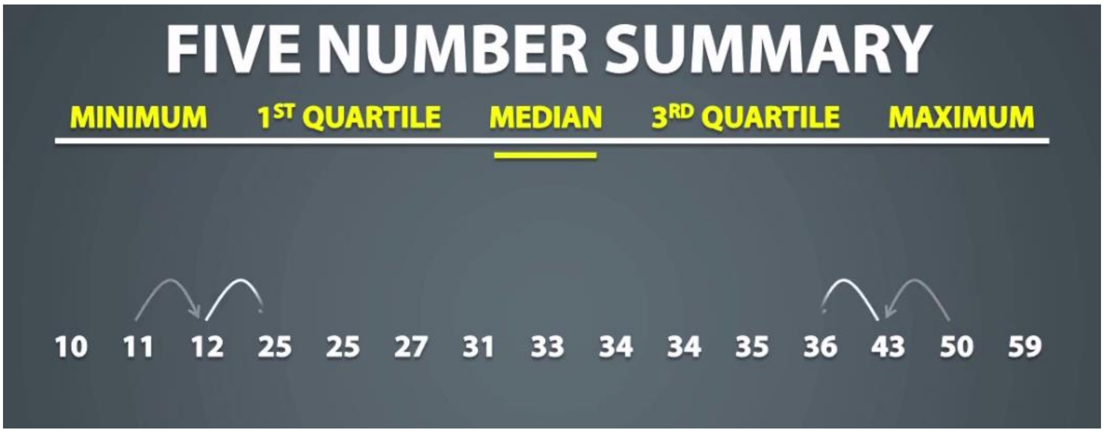
- ✓ The five number summary is divides the data into equal quarters

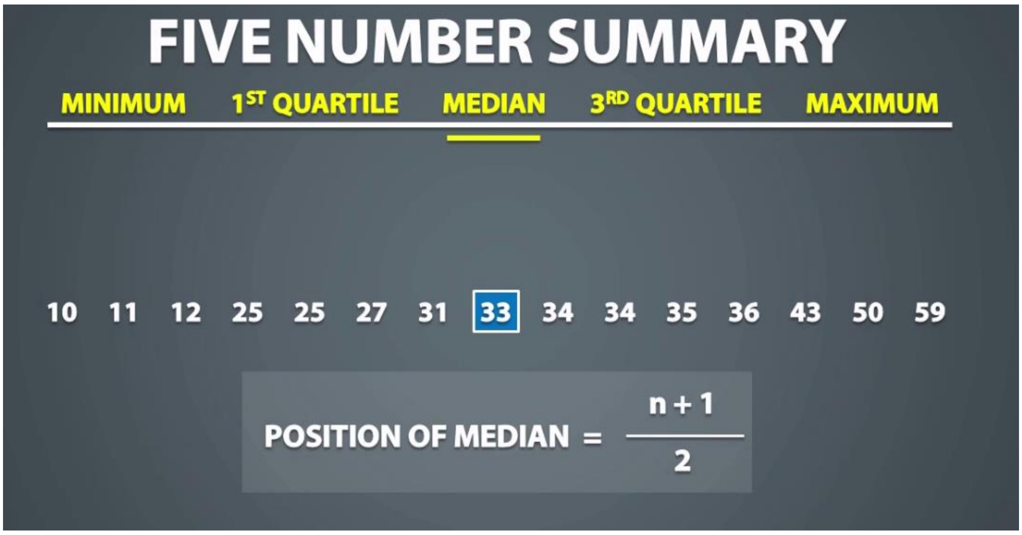


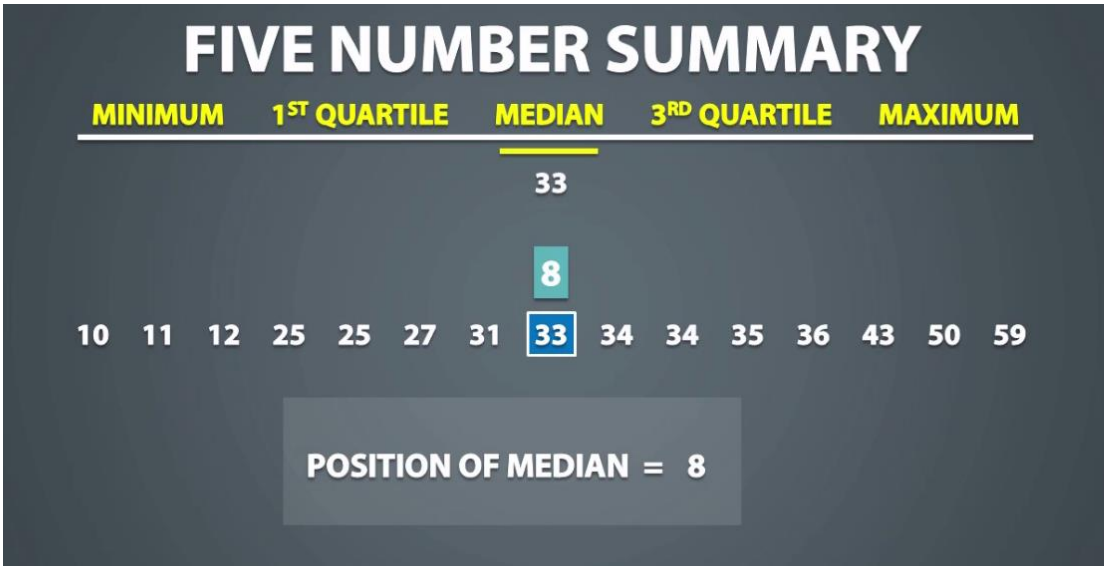
- ✓ Let's take one example to determine the five number summary



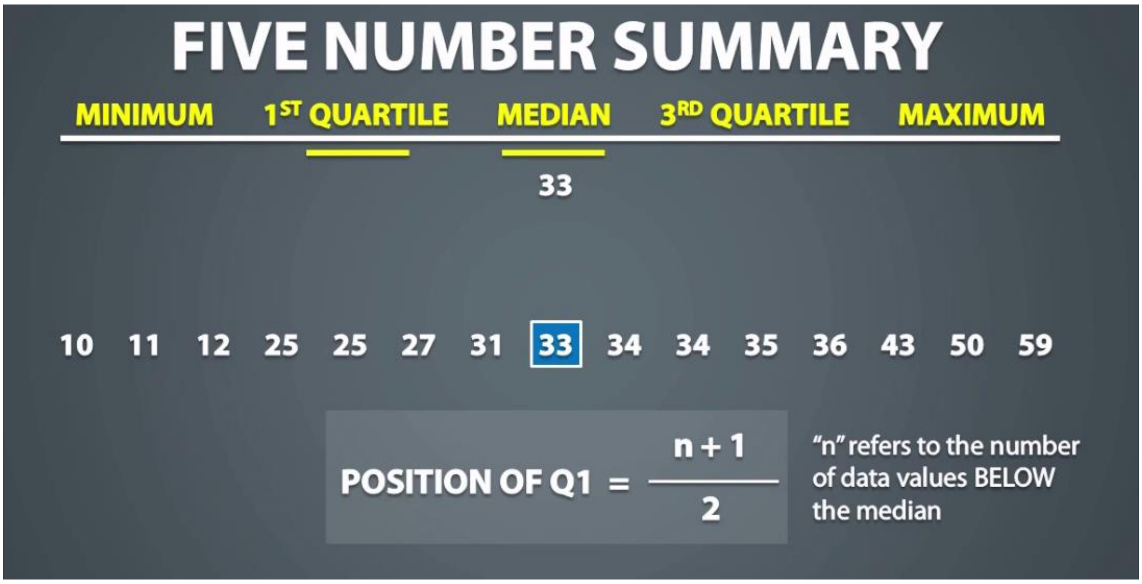
Median: Middle of the value

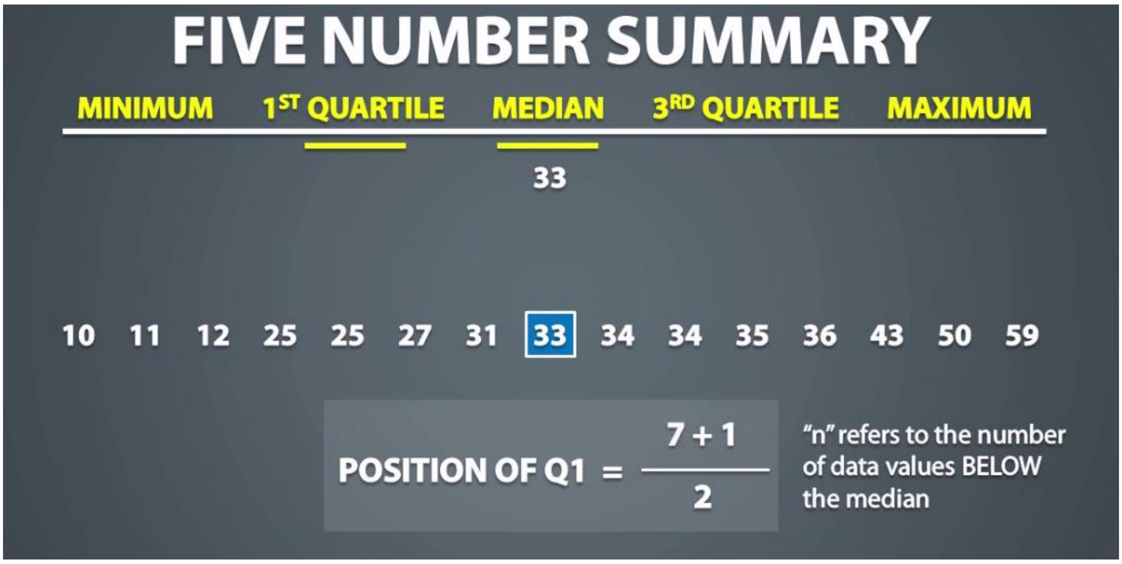


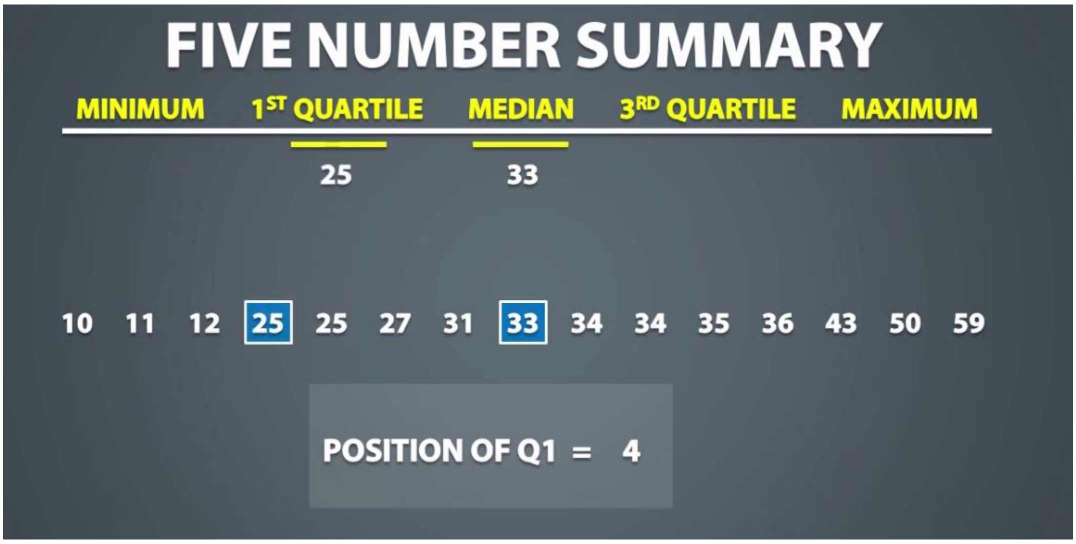




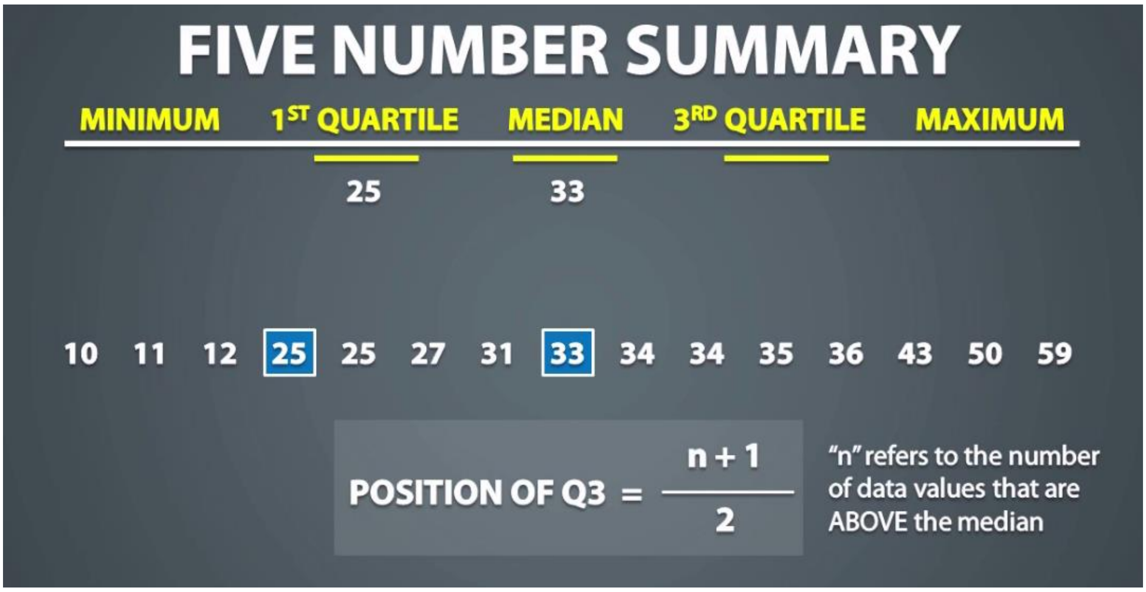
4. 1st Quartile

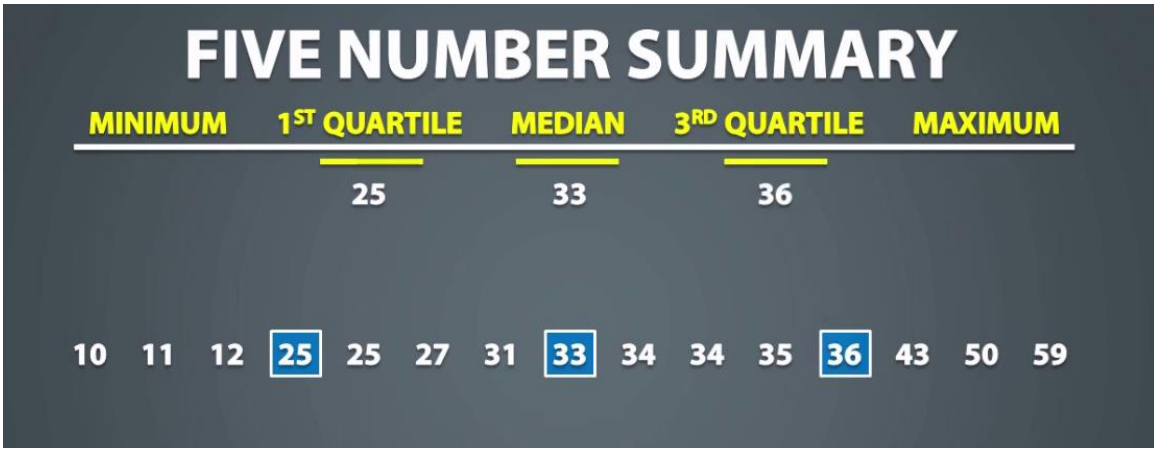



FIVE NUMBER SUMMARY
MINIMUM
1ST QUARTILE
MEDIAN
3RD QUARTILE
MAXIMUM
33
10 11 12 25 25 27 31 33 34 34 35 36 43 50 59
POSITION OF Q1 = (7 + 1) / 2
"n" refers to the number of data values BELOW the median

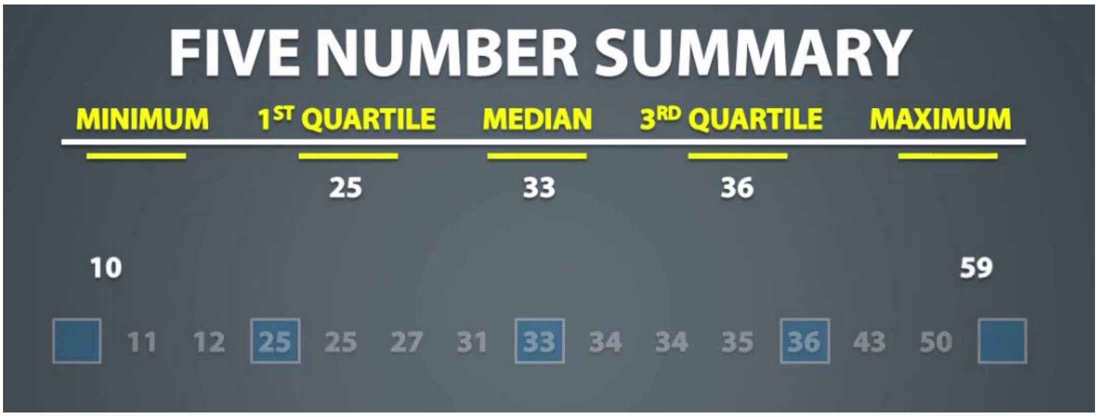

FIVE NUMBER SUMMARY
MINIMUM
1ST QUARTILE
MEDIAN
3RD QUARTILE
MAXIMUM
25
33
10 11 12 25 25 27 31 33 34 34 35 36 43 50 59
POSITION OF Q1 = 4

5. 3rd Quartile



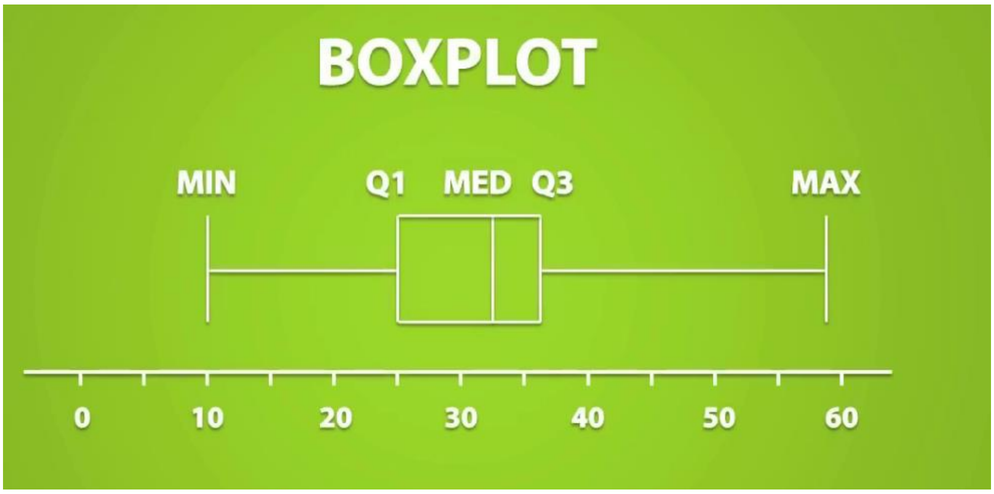


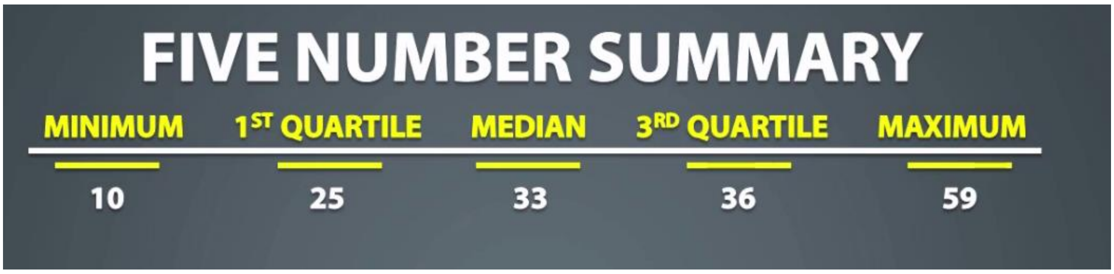
6. Smallest and largest

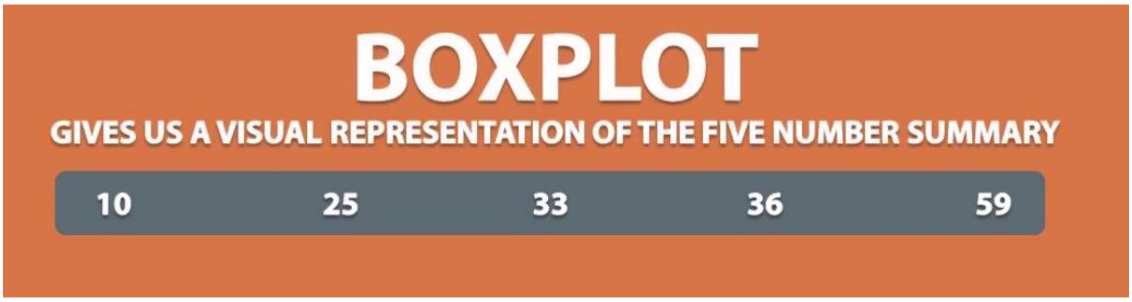


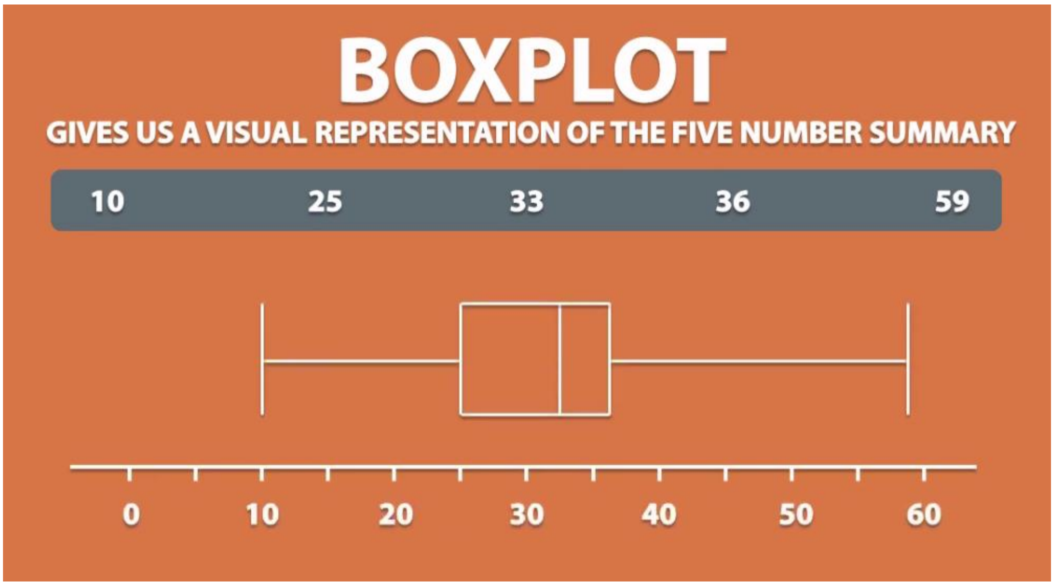
7. Box plot

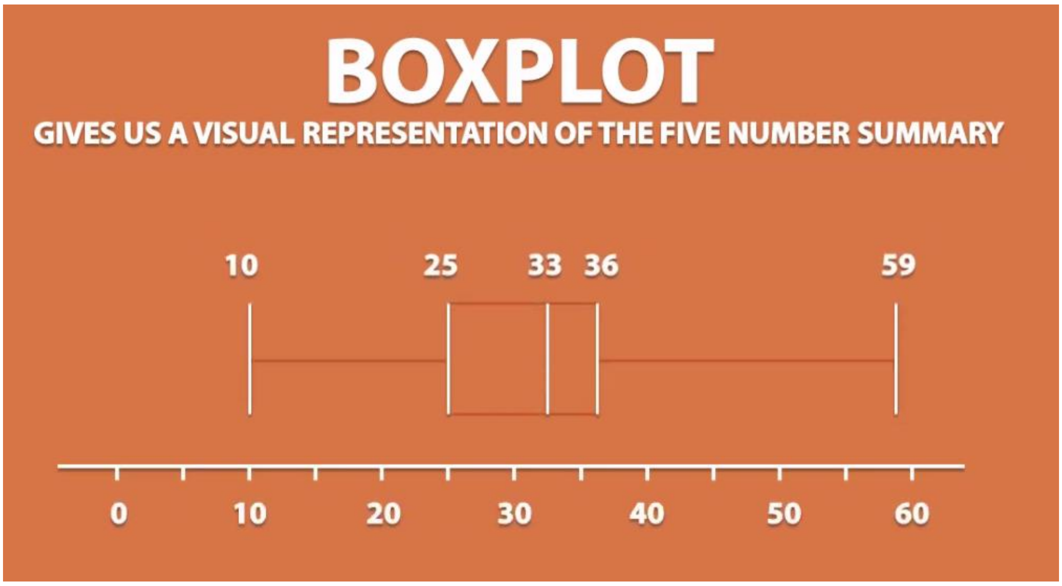
- ✓ We can take above five values summary and create a box plot
- ✓ A box plot gives us visual representation of the five numbers summary





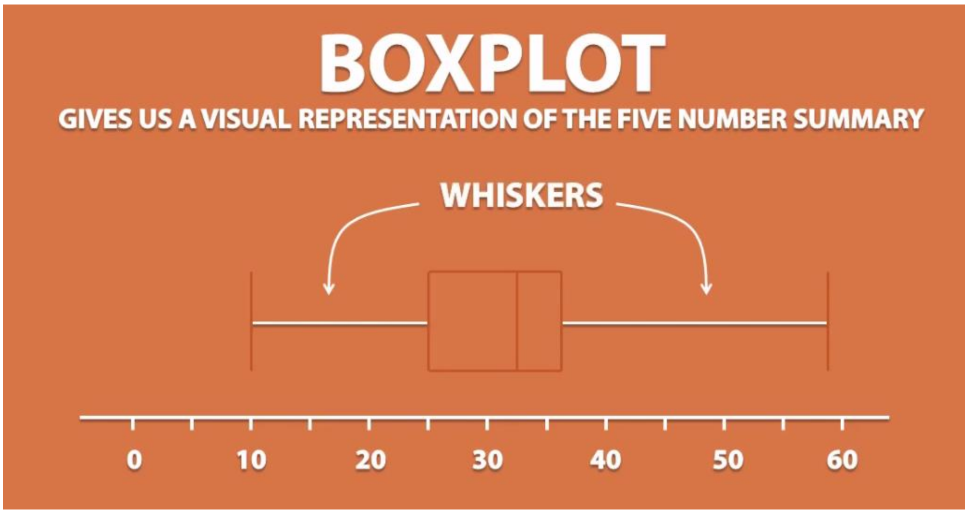



BOXPLOT
GIVES US A VISUAL REPRESENTATION OF THE FIVE NUMBER SUMMARY
10
25
33
36
59
0
10
20
30
40
50
60


BOXPLOT
GIVES US A VISUAL REPRESENTATION OF THE FIVE NUMBER SUMMARY
10
25
33
36
59
0
10
20
30
40
50
60

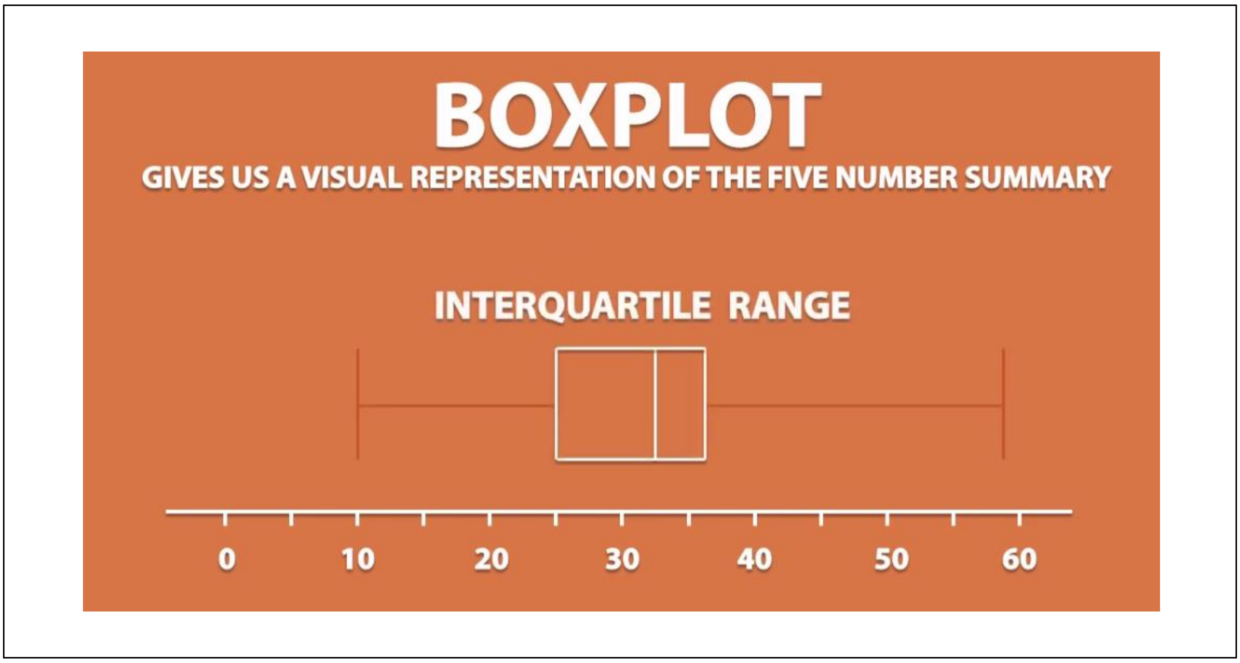
8. Whiskers

- ✓ The horizontal line that extends out from the box is called as whisker

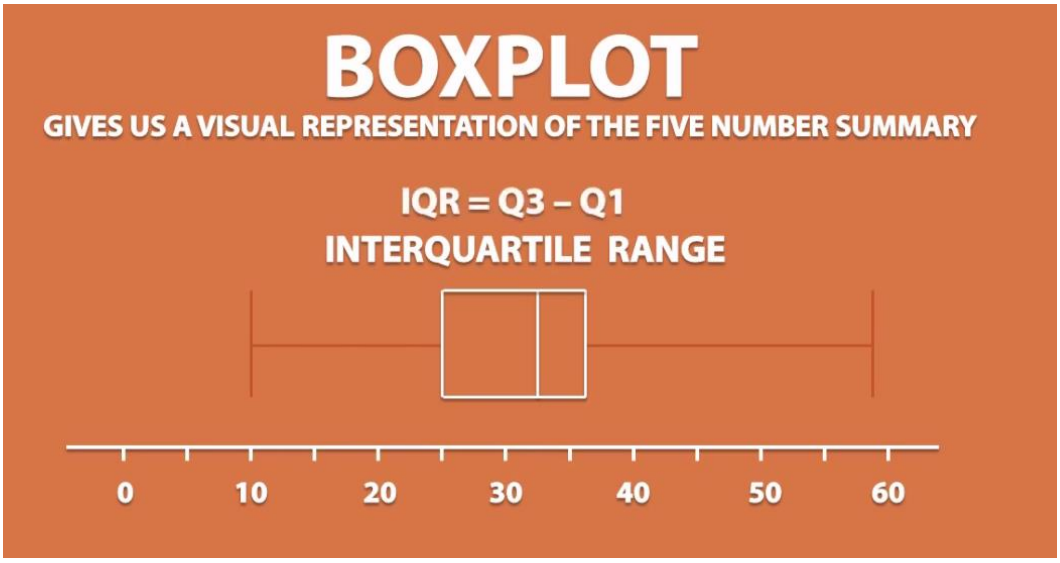


9. Interquartile range

- ✓ The actual box is called as Interquartile range

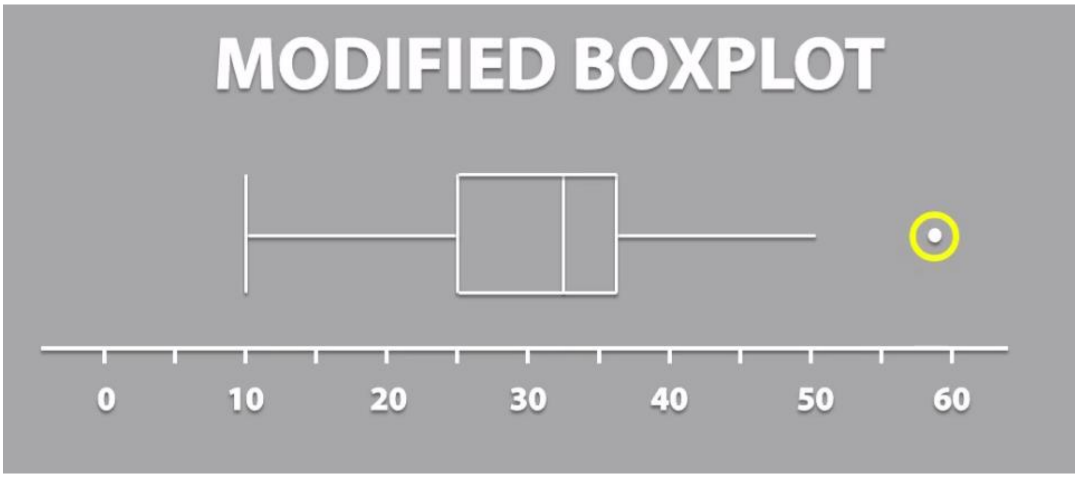


10. IQR = Q3 – Q1

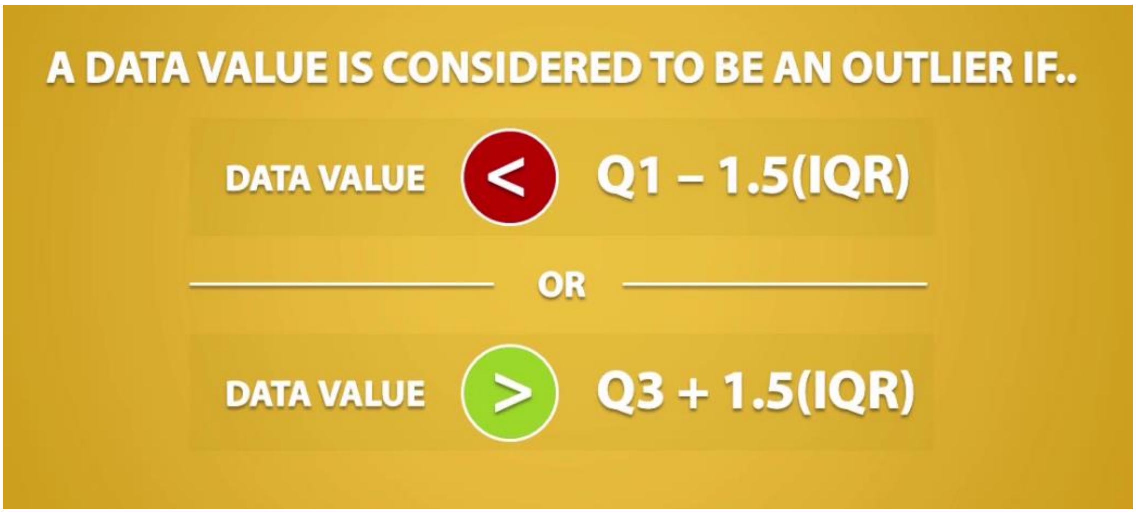

BOXPLOT
GIVES US A VISUAL REPRESENTATION OF THE FIVE NUMBER SUMMARY
IQR = Q3 – Q1
INTERQUARTILE RANGE
0
10
20
30
40
50
60

11. Outliers (modified box plot)

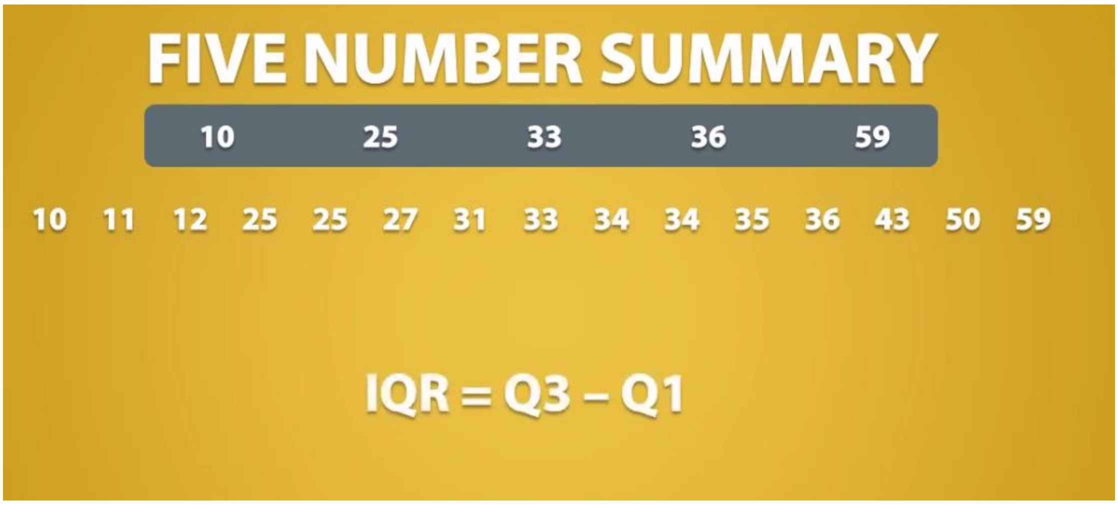
- ✓ A box plot with outlier is called as modified box plot

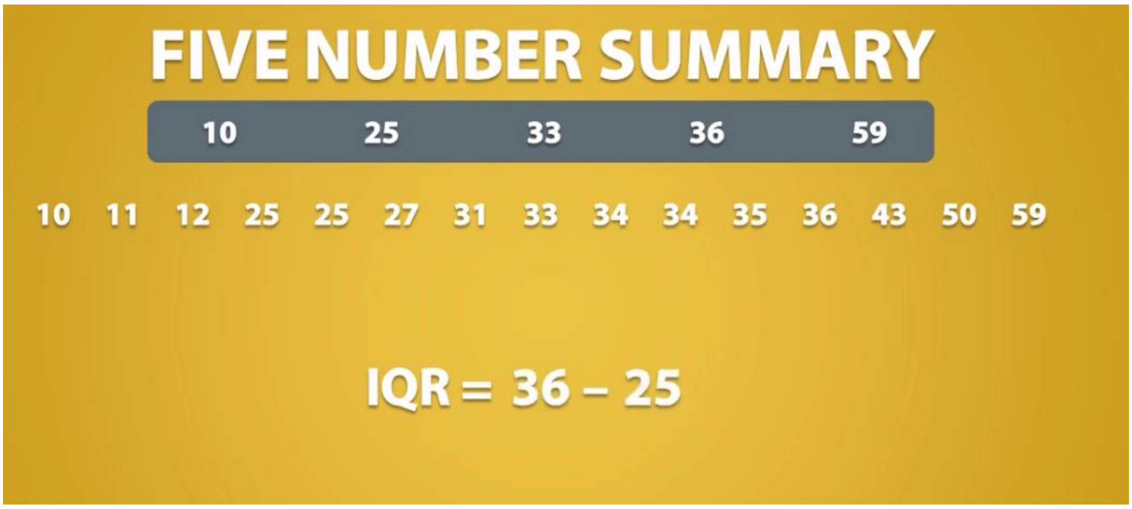


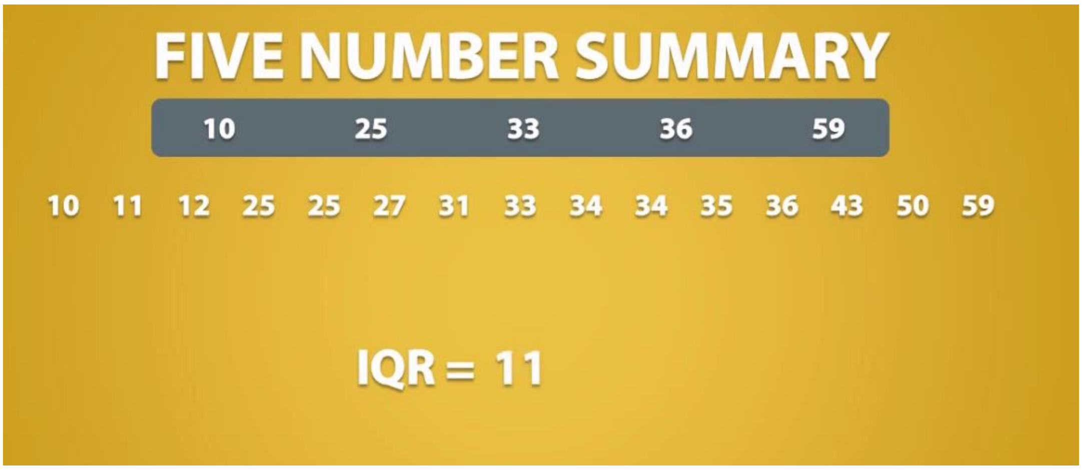
12. Outlier recognizing

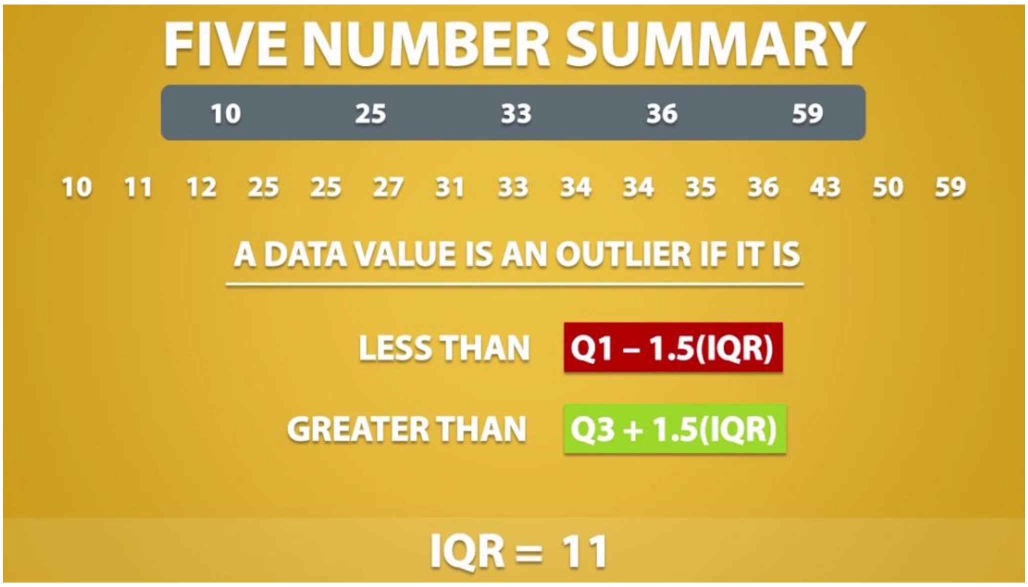


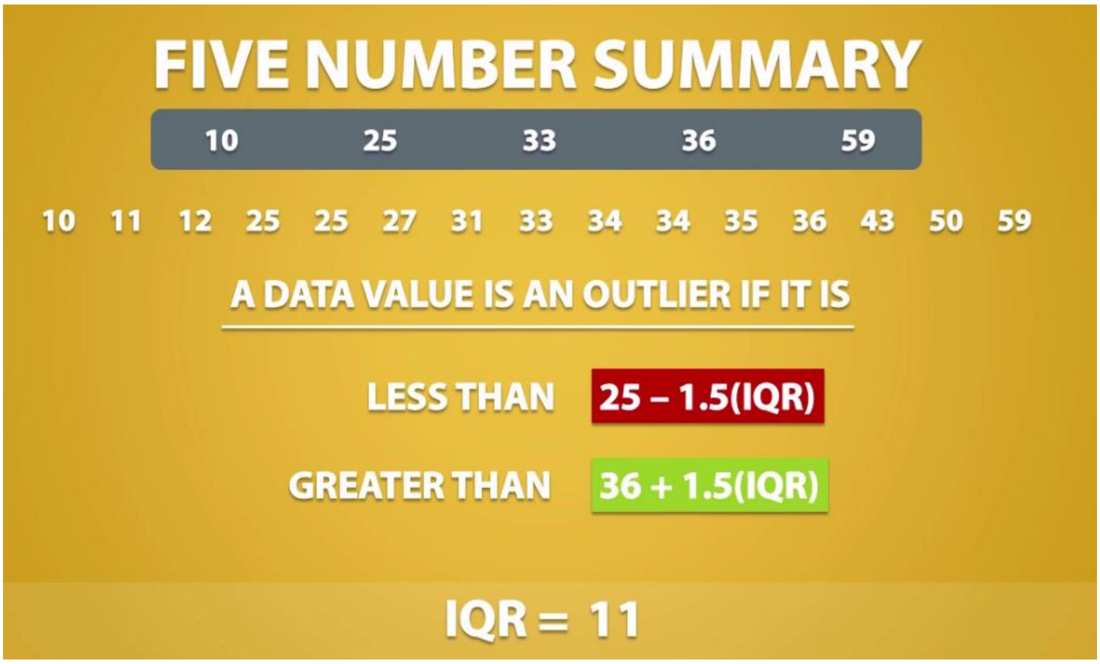
13. Checking outliers

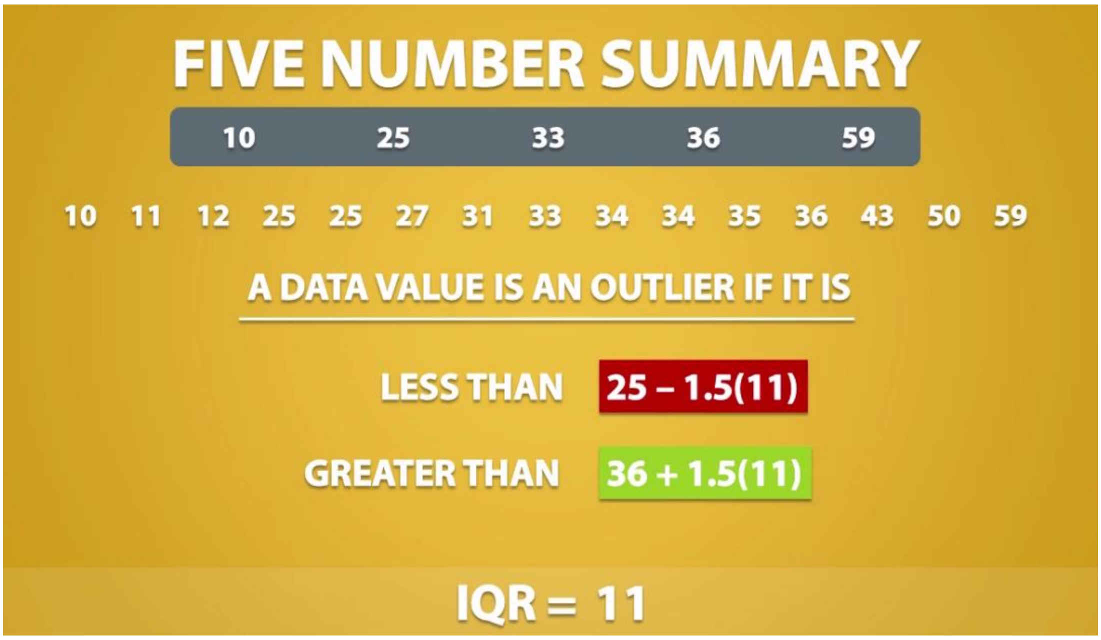


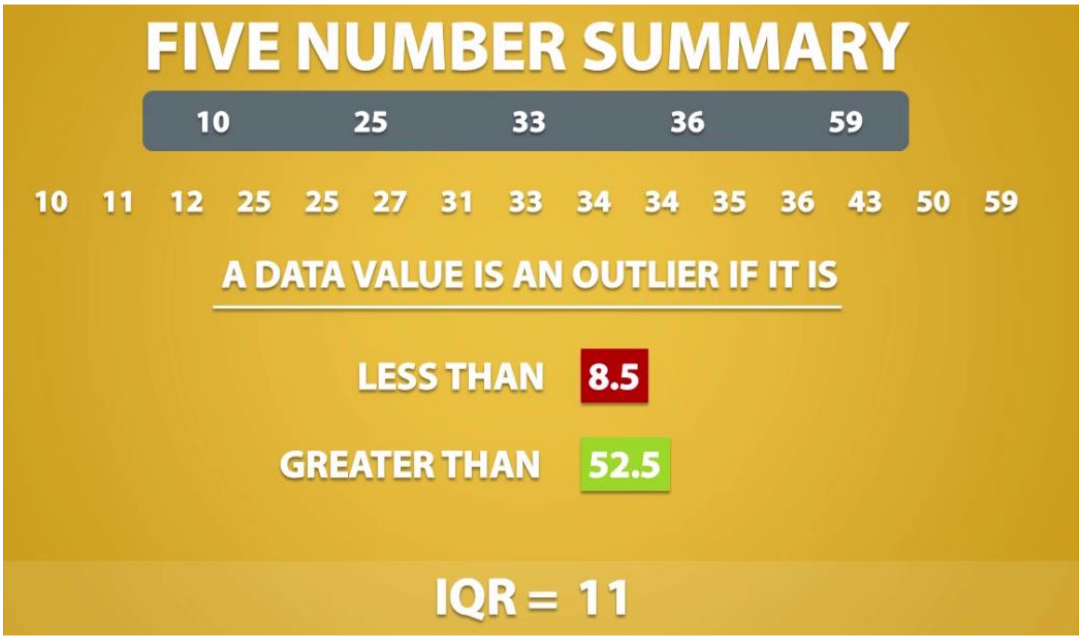


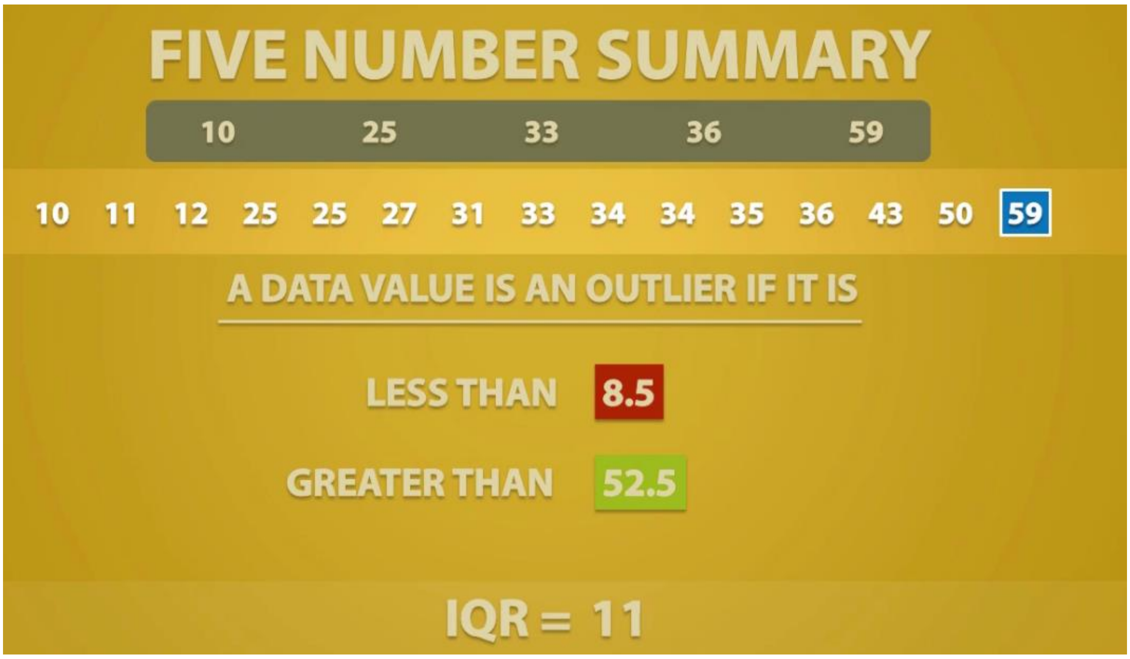

FIVE NUMBER SUMMARY
10 25 33 36 59
10 11 12 25 25 27 31 33 34 34 35 36 43 50 59
IQR = 11

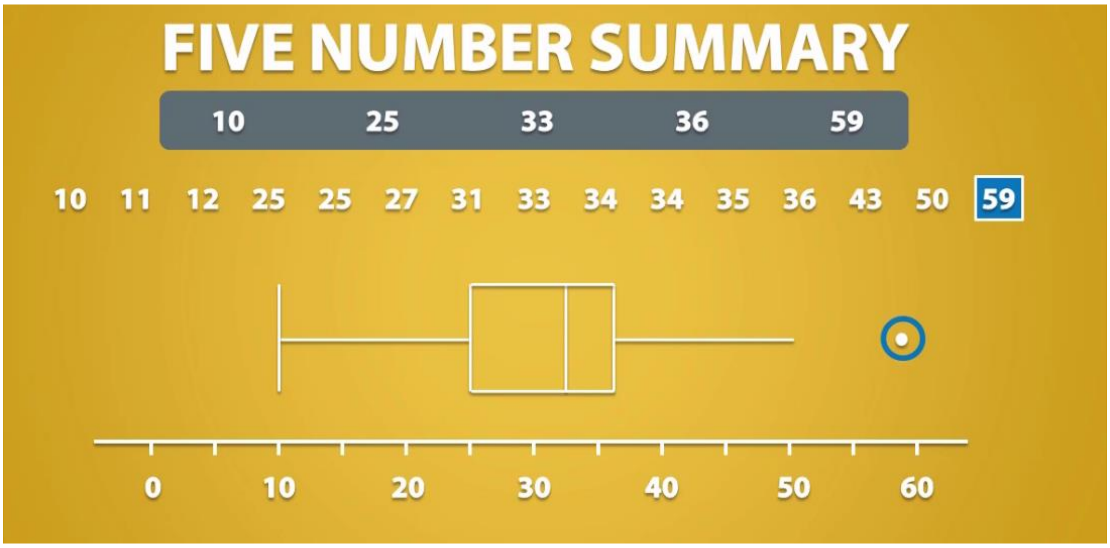

FIVE NUMBER SUMMARY
10 25 33 36 59
10 11 12 25 25 27 31 33 34 34 35 36 43 50 59
A DATA VALUE IS AN OUTLIER IF IT IS
LESS THAN Q1 – 1.5(IQR)
GREATER THAN Q3 + 1.5(IQR)
IQR = 11


FIVE NUMBER SUMMARY
10 25 33 36 59
10 11 12 25 25 27 31 33 34 34 35 36 43 50 59
A DATA VALUE IS AN OUTLIER IF IT IS
LESS THAN 25 – 1.5(IQR)
GREATER THAN 36 + 1.5(IQR)
IQR = 11


FIVE NUMBER SUMMARY
10 25 33 36 59
10 11 12 25 25 27 31 33 34 34 35 36 43 50 59
A DATA VALUE IS AN OUTLIER IF IT IS
LESS THAN 25 – 1.5(11)
GREATER THAN 36 + 1.5(11)
IQR = 11


FIVE NUMBER SUMMARY
10 25 33 36 59
10 11 12 25 25 27 31 33 34 34 35 36 43 50 59
A DATA VALUE IS AN OUTLIER IF IT IS
LESS THAN 8.5
GREATER THAN 52.5
IQR = 11


FIVE NUMBER SUMMARY
10 25 33 36 59
10 11 12 25 25 27 31 33 34 34 35 36 43 50 59
A DATA VALUE IS AN OUTLIER IF IT IS
LESS THAN 8.5
GREATER THAN 52.5
IQR = 11


FIVE NUMBER SUMMARY
10 25 33 36 59
10 11 12 25 25 27 31 33 34 34 35 36 43 50 59
0 10 20 30 40 50 60

5. Maths - Statistics – PART – 5

Contents

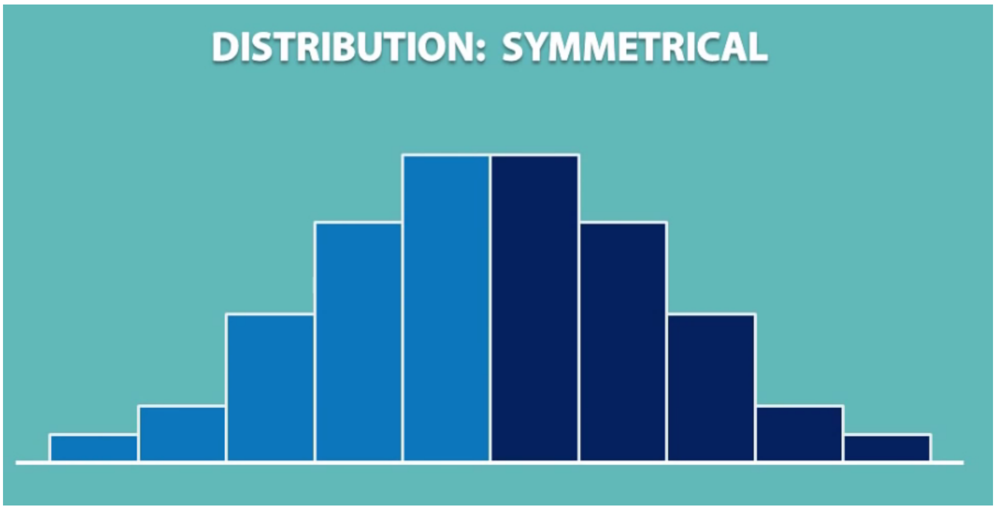
5. Maths - Statistics – PART – 5

1. Symmetry and Skewness

- ✓ This concept explains about shape of a distribution

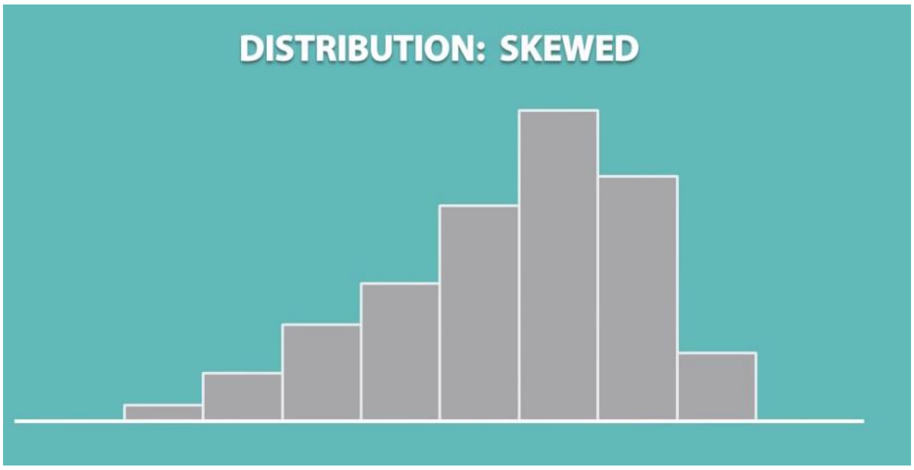
2. Symmetric distribution

- ✓ A distribution is called as symmetric, if it can be divided into two equal sizes of the same shape.
- ✓ Below histogram explains about symmetric distribution



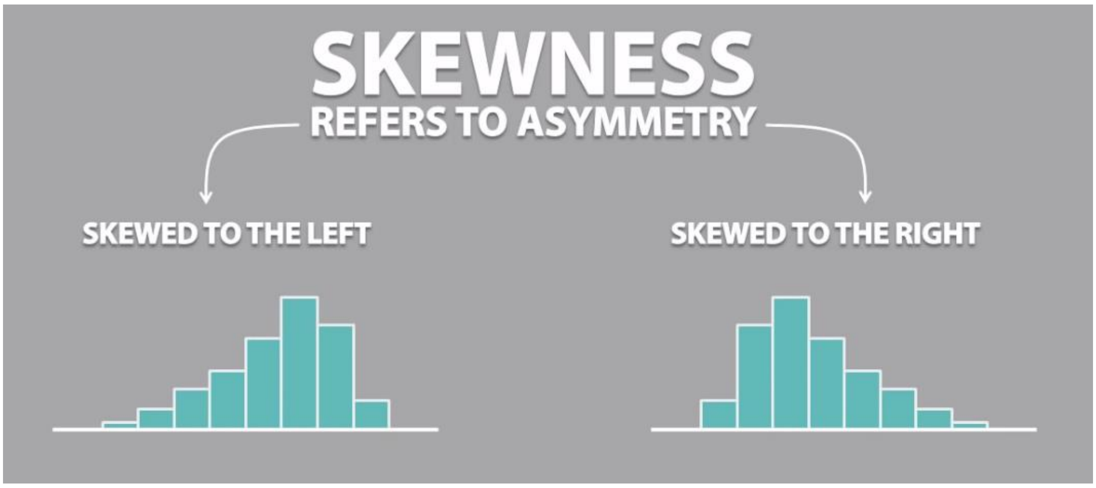
3. Distribution skewed or asymmetric distribution

- ✓ A distribution is called as skewed, if it cannot be divided into equal sizes.
- ✓ It's also called as asymmetric distribution
- ✓ Skewness refers to the asymmetry



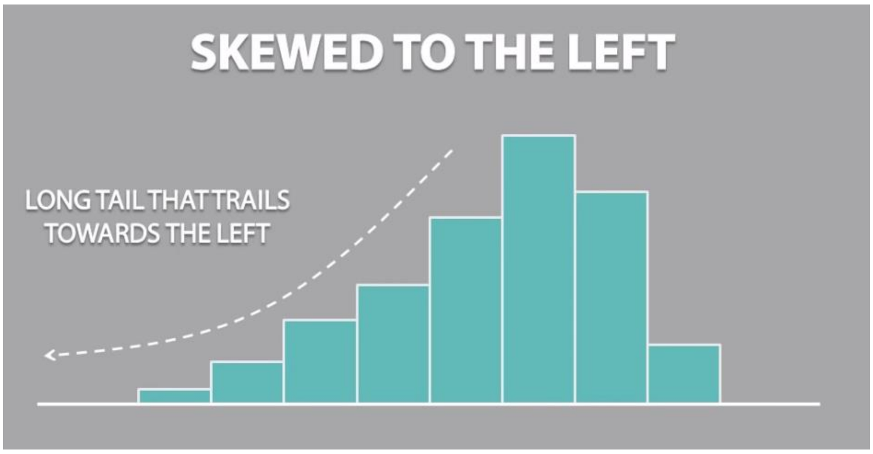
4. Types of Skewness

- ✓ We understand the skewness based on direction which the data points clustered
- ✓ There are two types
  - Skewness to the LEFT
  - Skewness to the RIGHT



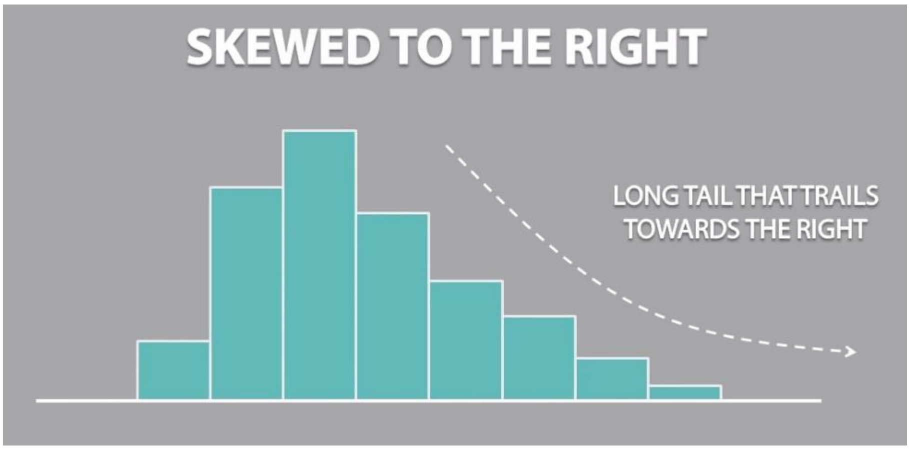
5. Skewness to the LEFT

- ✓ If the data is clustered at left hand side then it is called as skewness to the LEFT



6. Skewness to the RIGHT

- ✓ If the data is clustered at right hand side then it is called as skewness to the RIGHT

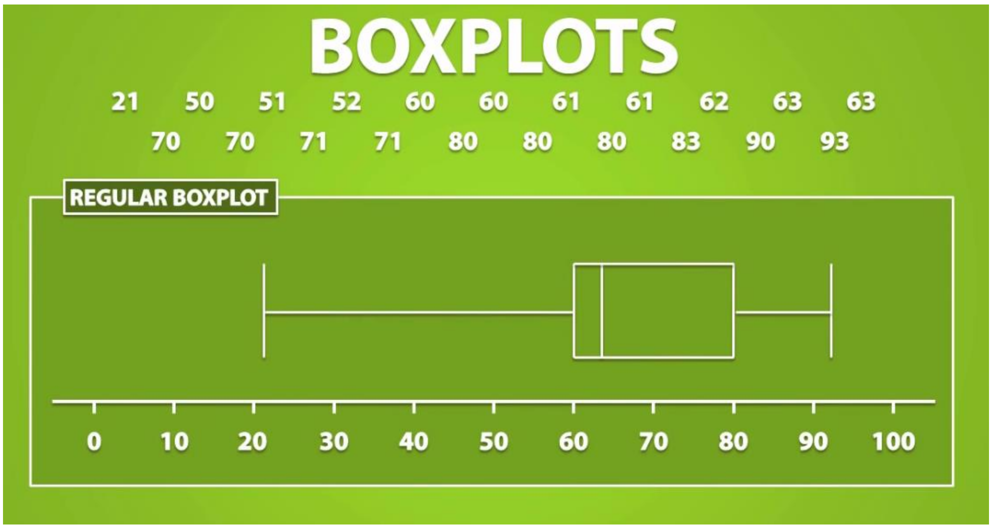


7. Boxplot

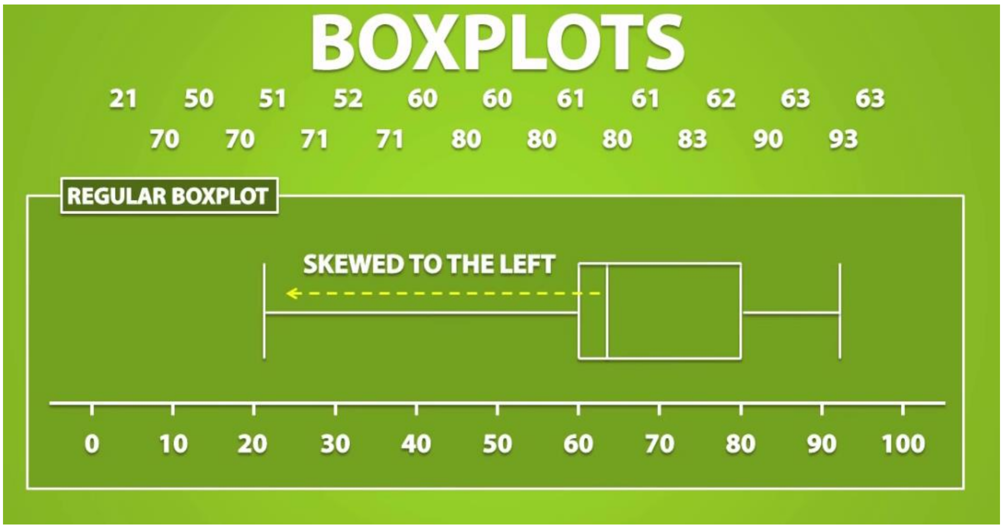
- ✓ We can determine skewness into the box plot
- ✓ The presences of outliers may effect to determine skewness

8. Boxplot skewness

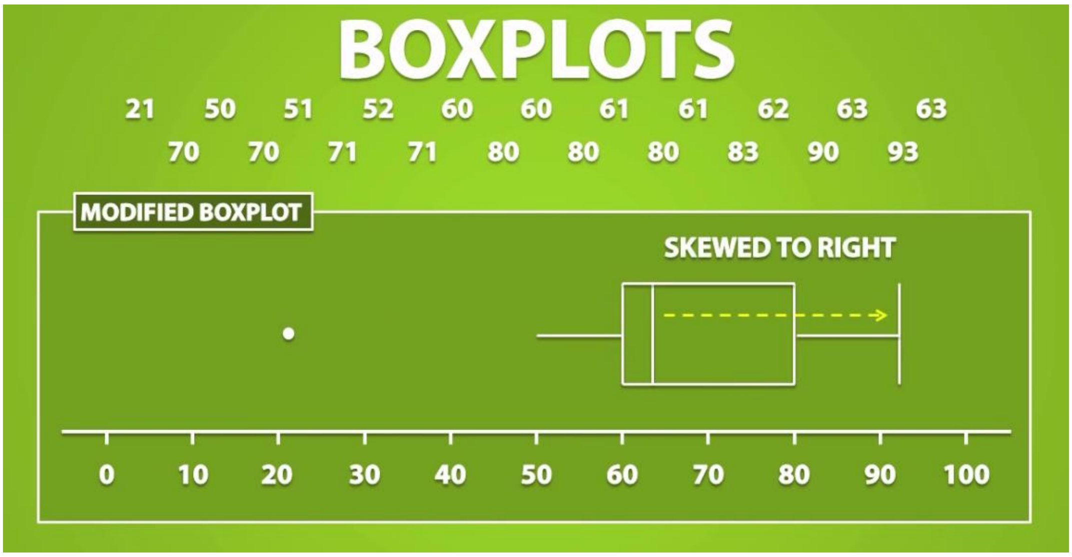
- ✓ When we construct the boxplot for the below data then we can draw below one



- ✓ According the boxplot we may think that this distribution is skewed to the left.



- ✓ But when we converted into modified box plot(because of outlier) then its directing to right side

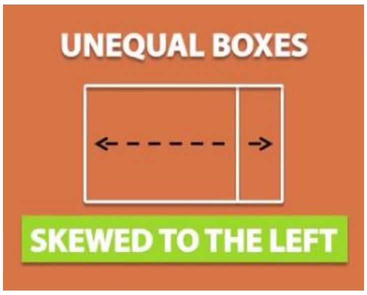


9. A strategy to find skewness in boxplot

- ✓ If we have unequal boxes, the side of the box is larger than that determines the skew

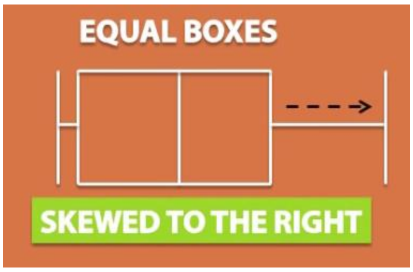
9.1. Case 1:

- ✓ If left side of the box larger than the right side, so its skewed to the left



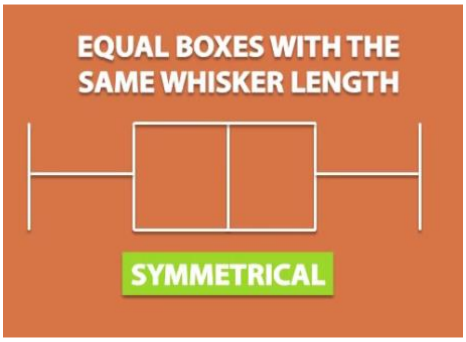
9.2. Case 2:

- ✓ If the boxes are equal in size then we need to consider the whisker size to determine this skew
- ✓ The larger whisker determines the skew
- ✓ In below case it's in right

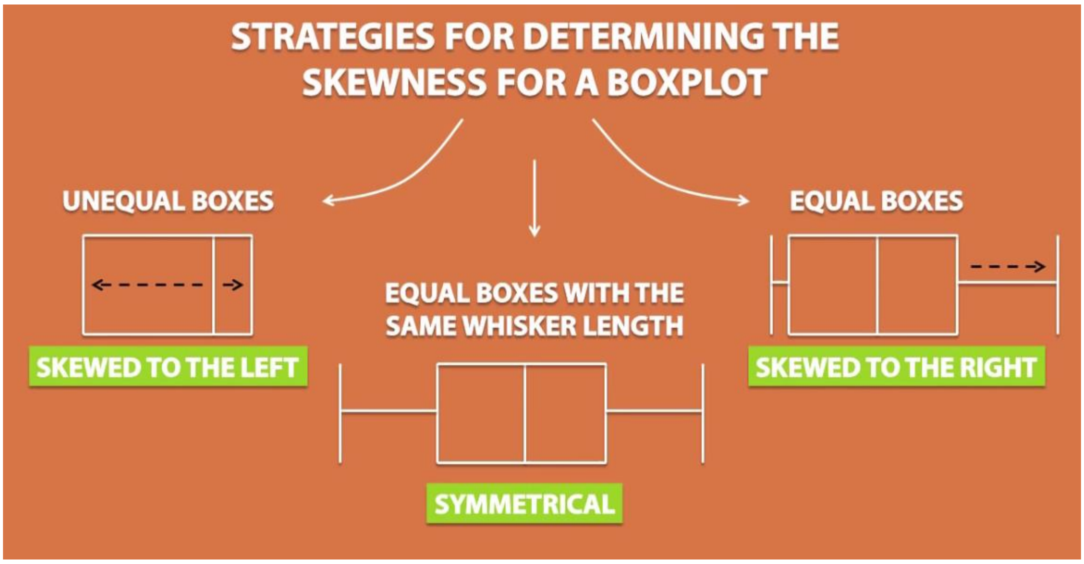


9.3. Case 3:

✓ If the boxes are equal with same whisker length then the distribution is said to be symmetrical



A summary

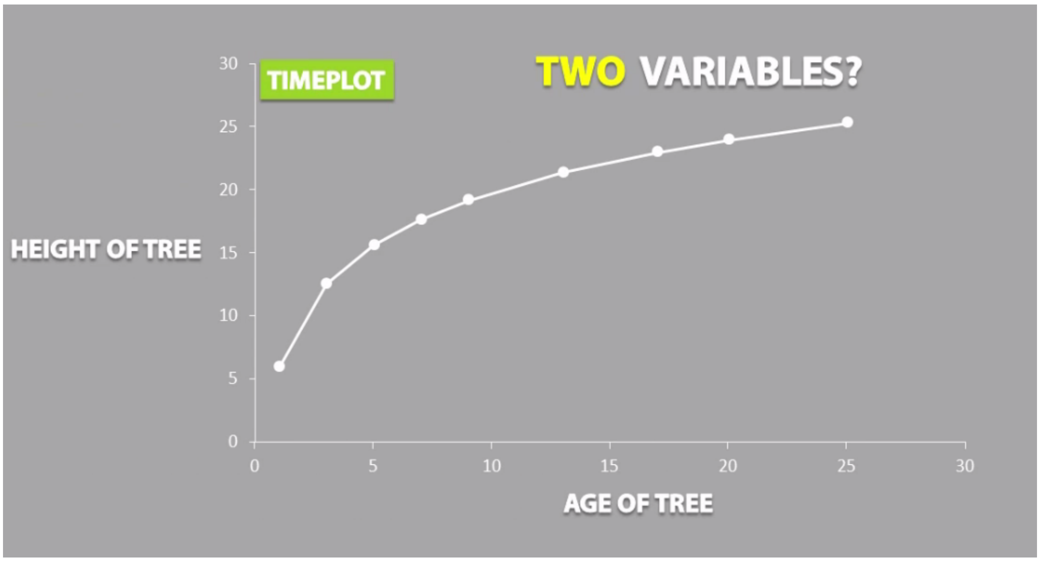


6. Maths - Statistics – PART – 6

Contents

6. Maths - Statistics – PART – 6

✓ This concept explains about how two variables are related each other

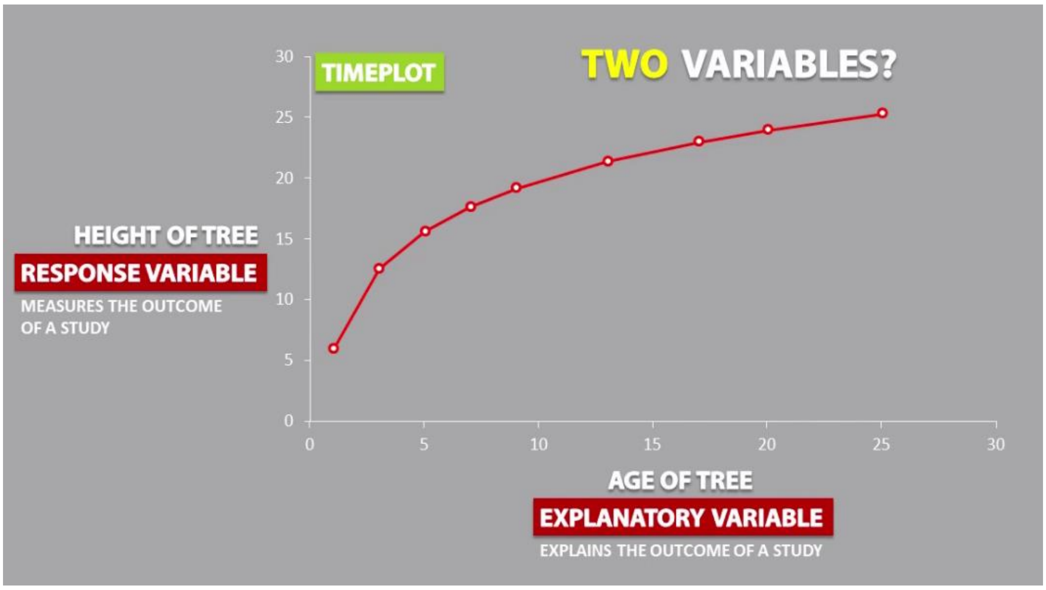


1. Explanatory variable

- ✓ This variables explains the outcome of the study
- ✓ Example is Age:
  - o As we are reaching older then the taller will be increase till to certain level
  - o **Age** is explains about **height**
- ✓ It is also called as Independent variable

2. Response variable

- ✓ This variables measures the outcome of the study
- ✓ It is also called as Dependent variable

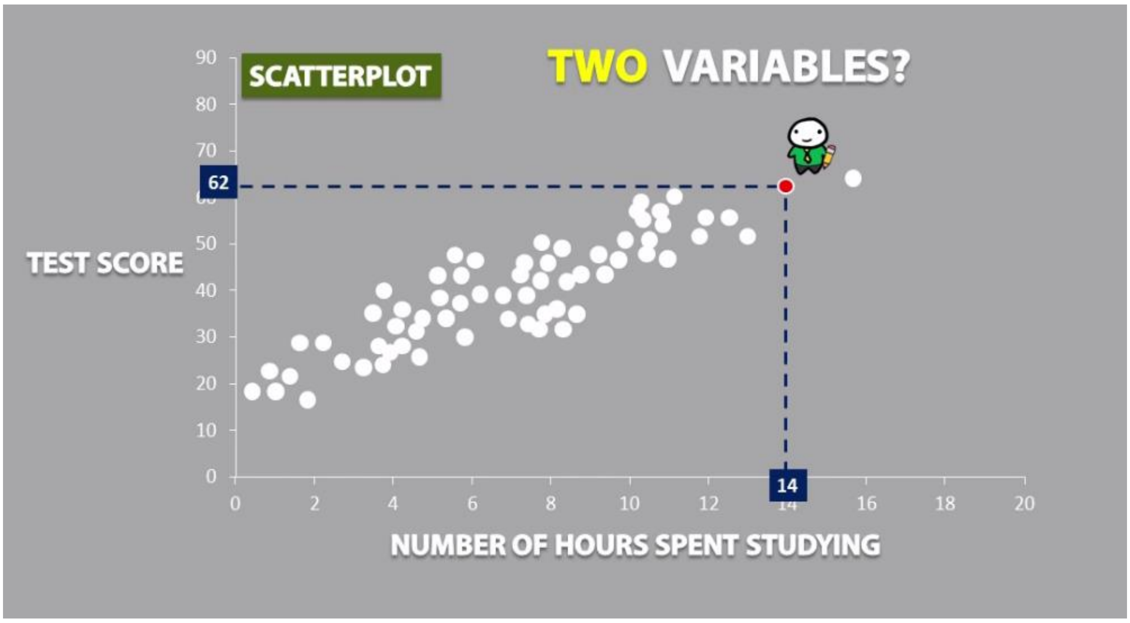


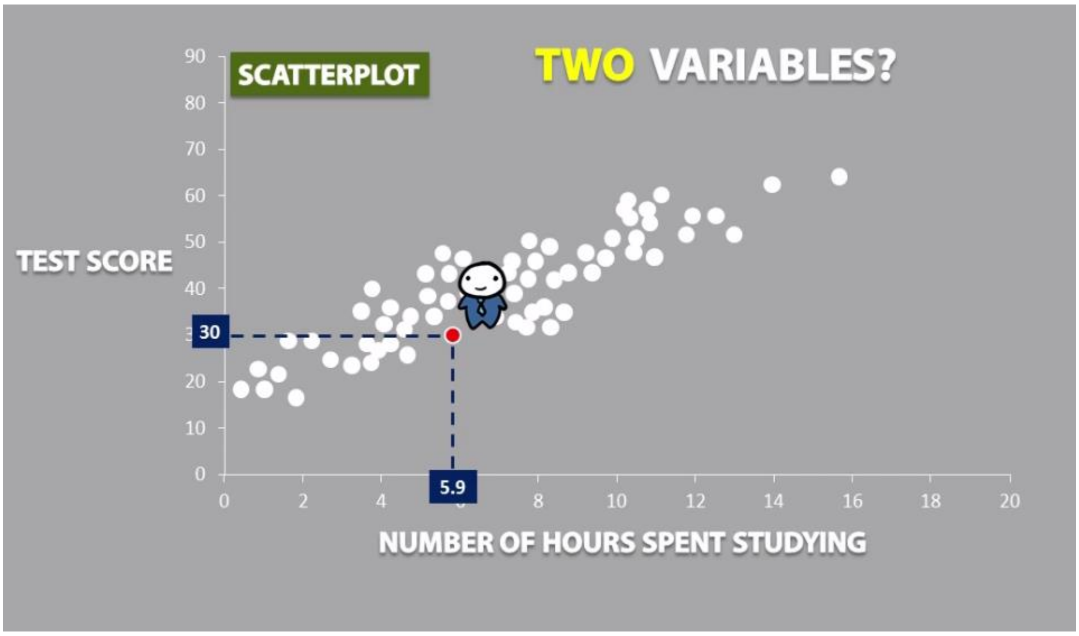
3. Scatter plot

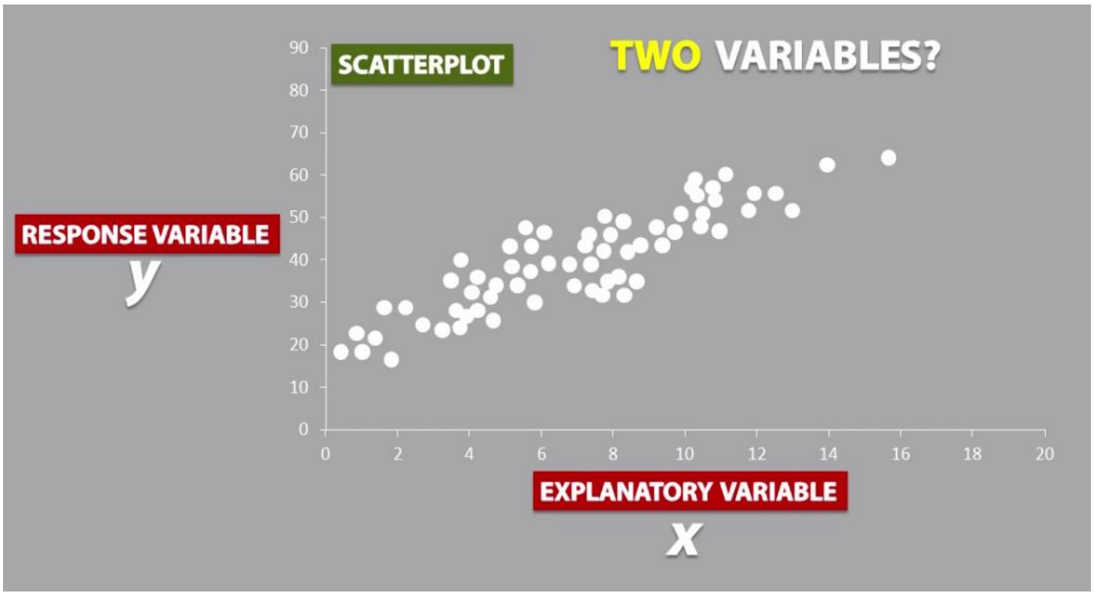
- ✓ Scatter plot is the good example which explains about one variable growth/down based on other variable

4. Example

- ✓ Scatter plot is the good example which explains about one variable growth/down based on other variable



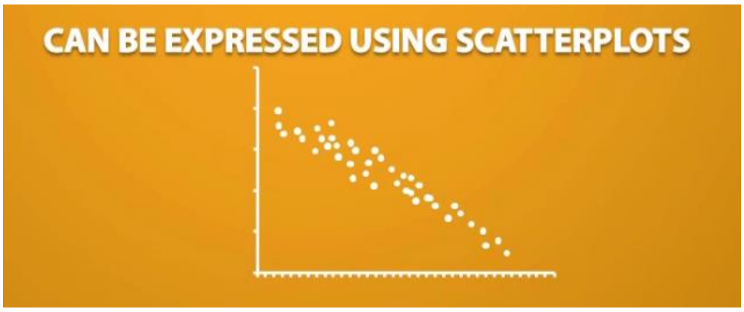

SCATTERPLOT
TWO VARIABLES?
TEST SCORE
30
5.9
NUMBER OF HOURS SPENT STUDYING
90
80
70
60
50
40
20
10
0
0
2
4
8
10
12
14
16
18
20


SCATTERPLOT
TWO VARIABLES?
RESPONSE VARIABLE
y
EXPLANATORY VARIABLE
x
90
80
70
60
50
40
30
20
10
0
0
2
4
6
8
10
12
14
16
18
20

5. Correlation

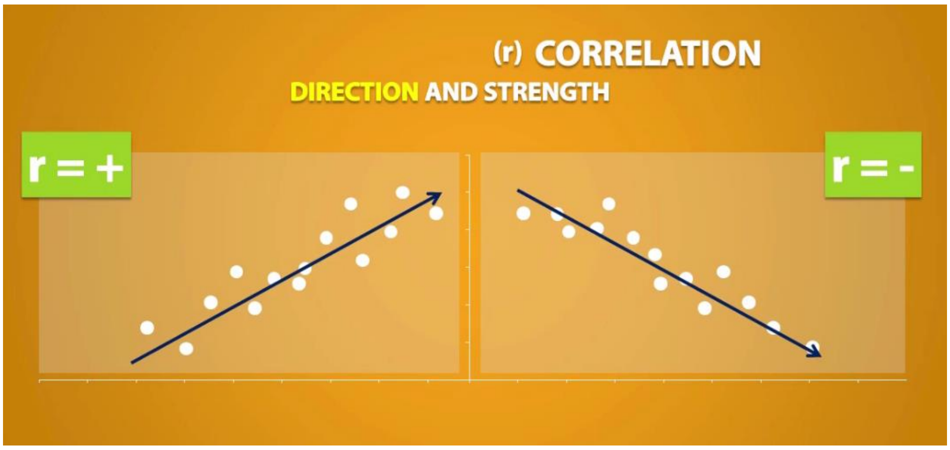
- ✓ It explains about the direction and strength of the linear relationship shared between two quantitative variables
- ✓ It is denoted as **r**

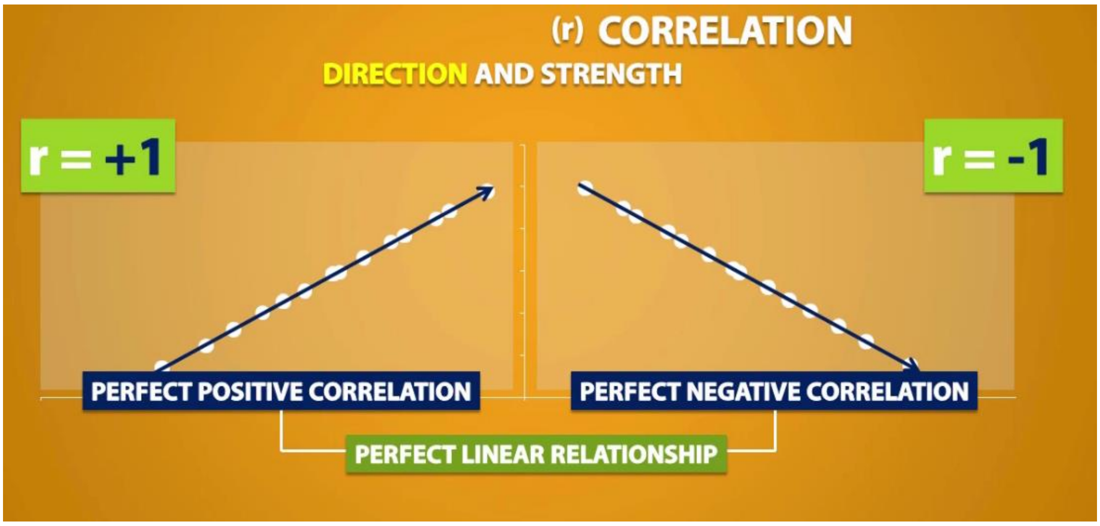
- ✓ Correlation can be expressed using scatter plots



6. Correlation – Direction & Strength

- ✓ Correlation speaks about the direction or slop of set of data points
- ✓ It explains about the direction can be upwards or downwards
  - If upwards then correlation is positive
  - If downwards then correlation is negative

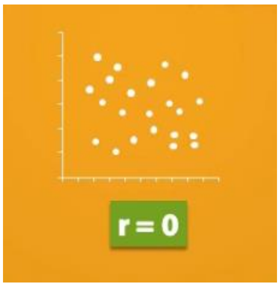




- ✓ Correlation measures the strength of the linear relationship
- ✓ So, correlation values can be in between **+1** and negative **-1**
- ✓ The strength of the linear relationship increased as r got close to positive **+1** or **-1**

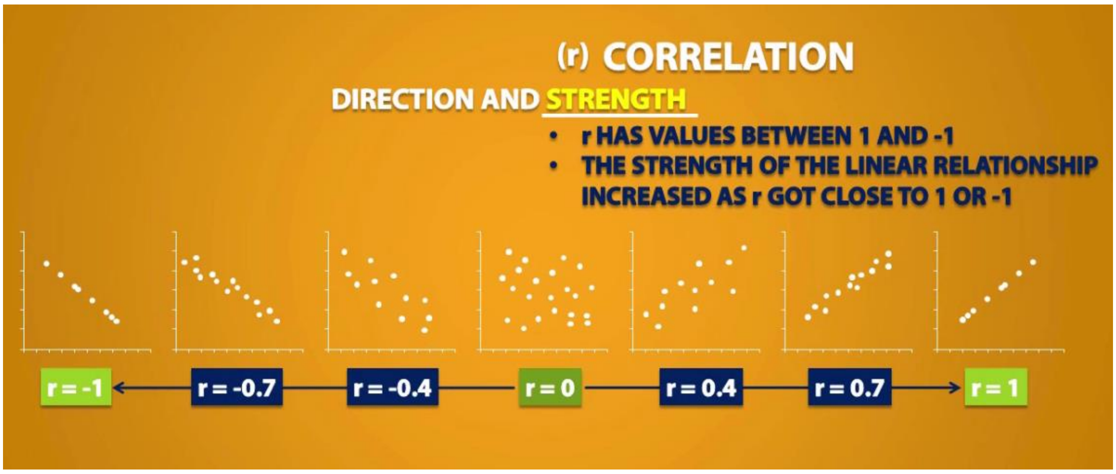
6.1. r = 0

- ✓ There is no correlation
- ✓ There is no linear relationship in between



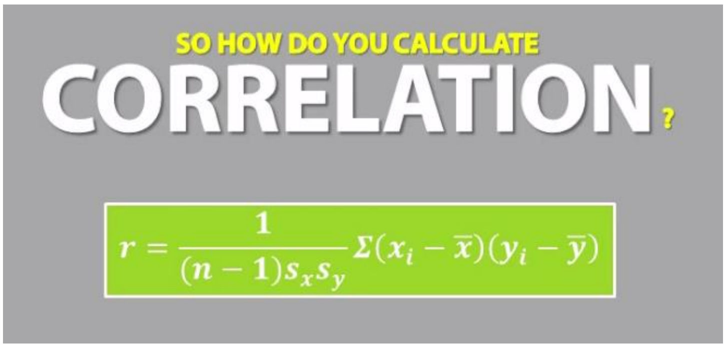
6.2. If r value is towards to +1 or -1

- ✓ In this case the linear relationship get stronger
- ✓ If r value gets close to +1 or -1 then the relationship is very stronger

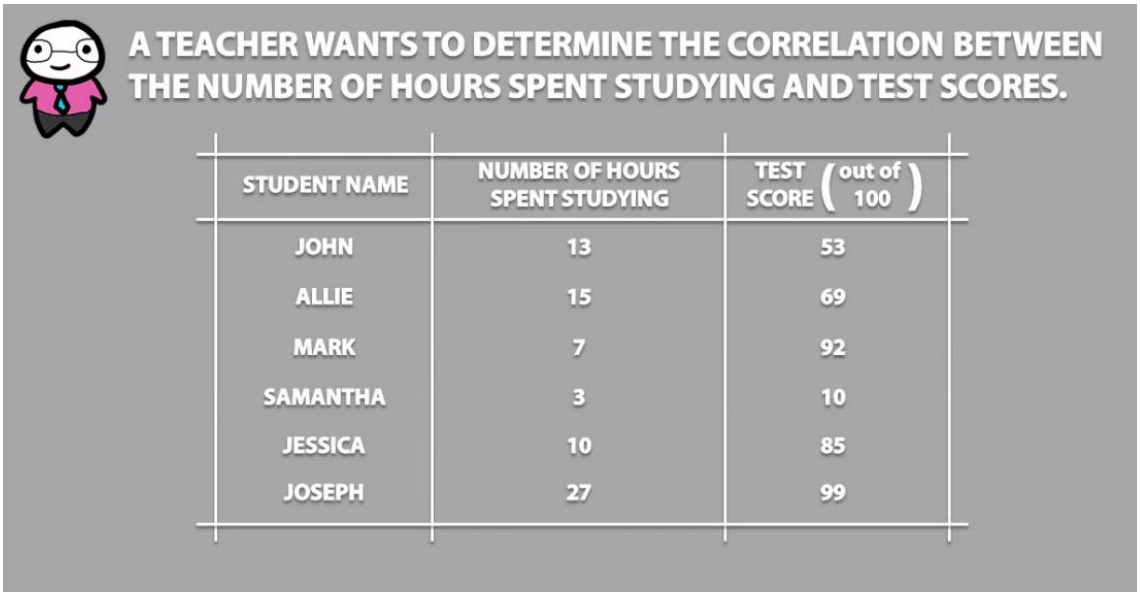


7. Calculate correlation

✓ We can calculate correlation by using formula

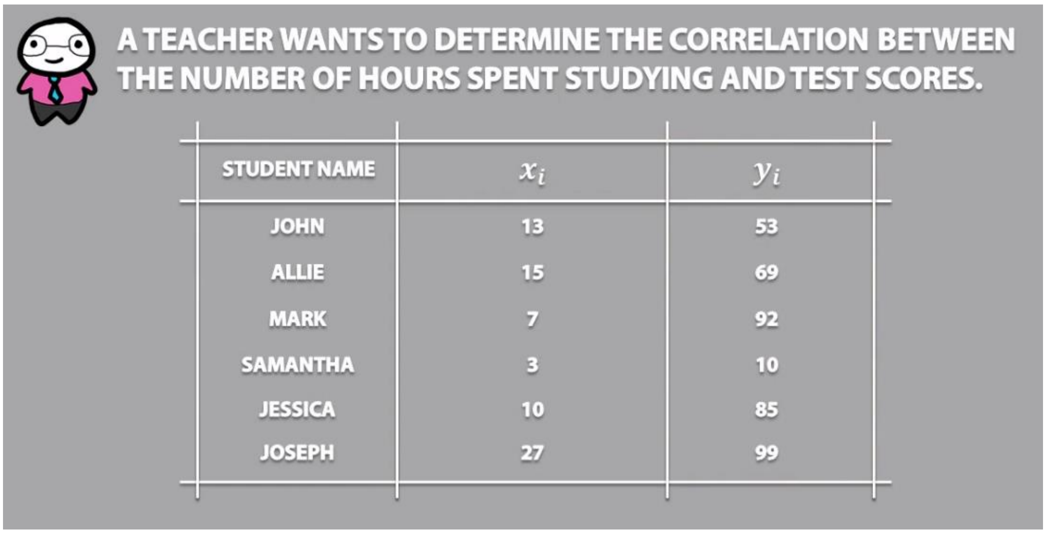


Example



A TEACHER WANTS TO DETERMINE THE CORRELATION BETWEEN THE NUMBER OF HOURS SPENT STUDYING AND TEST SCORES.

STUDENT NAME	NUMBER OF HOURS SPENT STUDYING	TEST SCORE (out of 100)
JOHN	13	53
ALLIE	15	69
MARK	7	92
SAMANTHA	3	10
JESSICA	10	85
JOSEPH	27	99



A TEACHER WANTS TO DETERMINE THE CORRELATION BETWEEN THE NUMBER OF HOURS SPENT STUDYING AND TEST SCORES.

STUDENT NAME	x_i	y_i
JOHN	13	53
ALLIE	15	69
MARK	7	92
SAMANTHA	3	10
JESSICA	10	85
JOSEPH	27	99

$$r = \frac{1}{(n-1)s_x s_y} \Sigma(x_i - \overline{x})(y_i - \overline{y})$$

x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})(y_i - \overline{y})$
13	53			
15	69			
7	92			
3	10			
10	85			
27	99			

$$r = \frac{1}{(n-1)s_x s_y} \Sigma(x_i - \overline{x})(y_i - \overline{y})$$

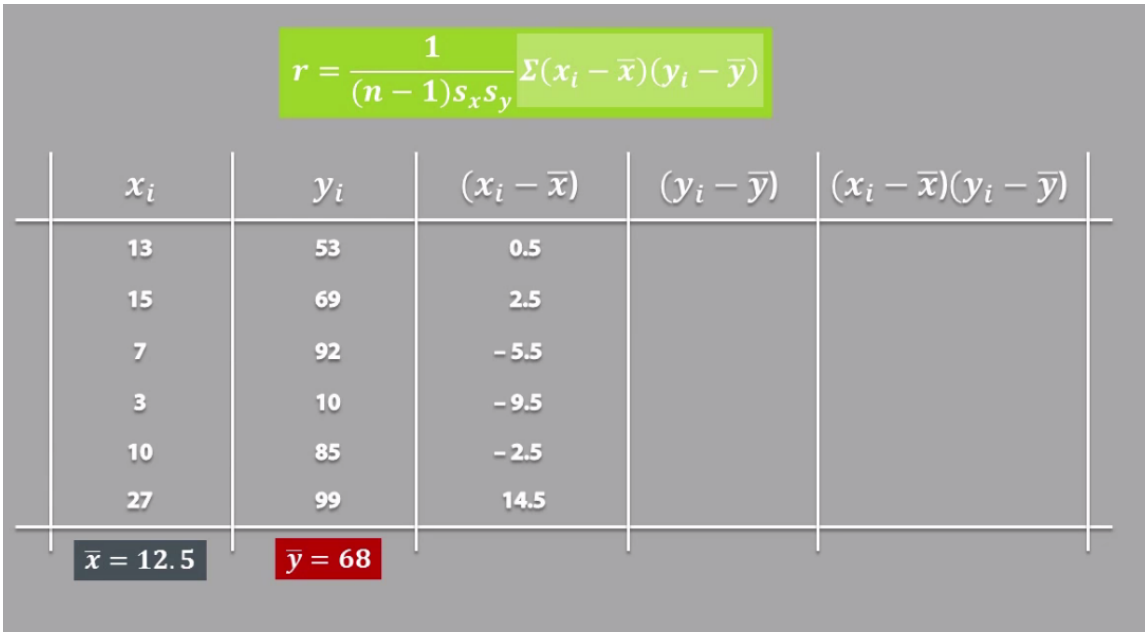
x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})(y_i - \overline{y})$
13	53			
15	69			
7	92			
3	10			
10	85			
27	99			
$\overline{x} = 12.5$	$\overline{y} = 68$			

$$r = \frac{1}{(n-1)s_x s_y} \Sigma(x_i - \bar{x})(y_i - \bar{y})$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
13	53	13 − 12.5		
15	69			
7	92			
3	10			
10	85			
27	99			
$\bar{x} = 12.5$	$\bar{y} = 68$			

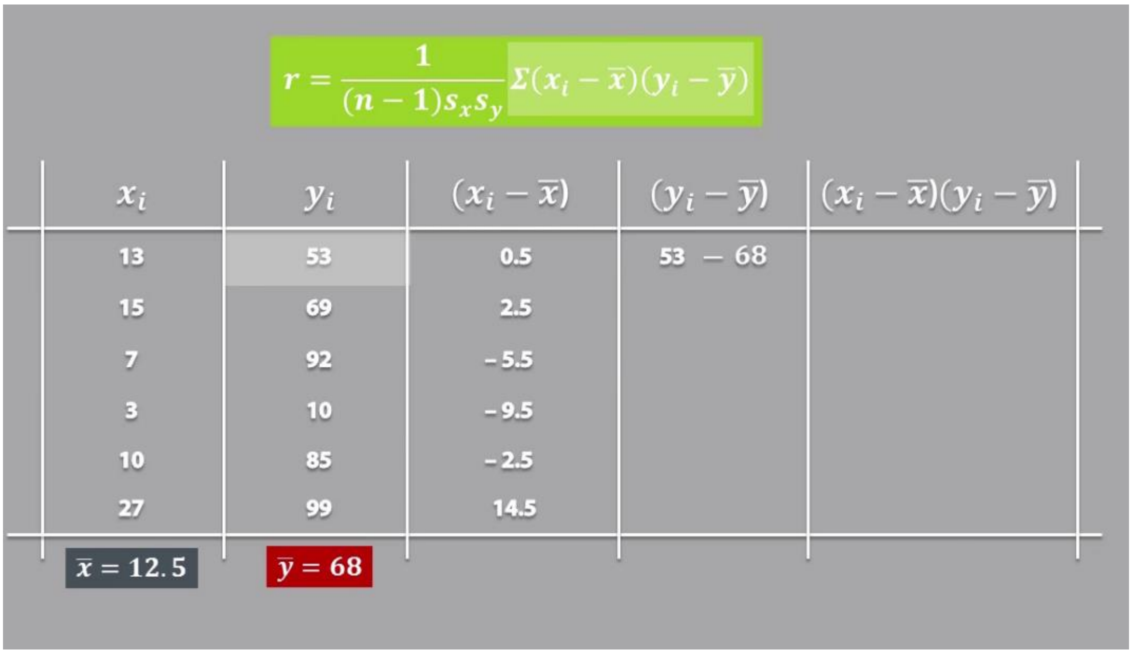
$$r = \frac{1}{(n-1)s_x s_y} \Sigma(x_i - \bar{x})(y_i - \bar{y})$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
13	53	0.5		
15	69	15 − 12.5		
7	92			
3	10			
10	85			
27	99			
$\bar{x} = 12.5$	$\bar{y} = 68$			



$$r = \frac{1}{(n-1)s_x s_y} \Sigma(x_i - \bar{x})(y_i - \bar{y})$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
13	53	0.5		
15	69	2.5		
7	92	−5.5		
3	10	−9.5		
10	85	−2.5		
27	99	14.5		
$\bar{x} = 12.5$	$\bar{y} = 68$			



$$r = \frac{1}{(n-1)s_x s_y} \Sigma(x_i - \bar{x})(y_i - \bar{y})$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
13	53	0.5	53 − 68	
15	69	2.5		
7	92	−5.5		
3	10	−9.5		
10	85	−2.5		
27	99	14.5		
$\bar{x} = 12.5$	$\bar{y} = 68$			

$$r = \frac{1}{(n-1)s_x s_y} \Sigma(x_i - \bar{x})(y_i - \bar{y})$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
13	53	0.5	−15	
15	69	2.5	1	
7	92	−5.5	24	
3	10	−9.5	−58	
10	85	−2.5	17	
27	99	14.5	31	
$\bar{x} = 12.5$	$\bar{y} = 68$			

$$r = \frac{1}{(n-1)s_x s_y} \Sigma(x_i - \bar{x})(y_i - \bar{y})$$

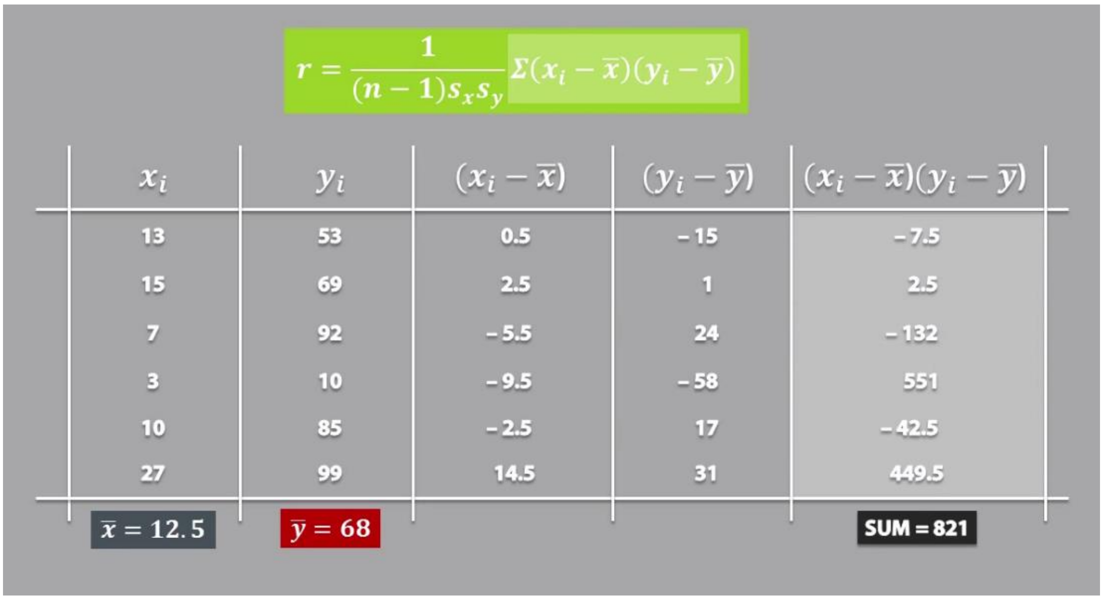
x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
13	53	0.5	−15	(0.5)(−15)
15	69	2.5	1	
7	92	−5.5	24	
3	10	−9.5	−58	
10	85	−2.5	17	
27	99	14.5	31	
$\bar{x} = 12.5$	$\bar{y} = 68$			

$$r = \frac{1}{(n-1)s_x s_y} \Sigma(x_i - \bar{x})(y_i - \bar{y})$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
13	53	0.5	−15	(0.5)(−15)
15	69	2.5	1	(2.5)(1)
7	92	−5.5	24	
3	10	−9.5	−58	
10	85	−2.5	17	
27	99	14.5	31	
$\bar{x} = 12.5$	$\bar{y} = 68$			

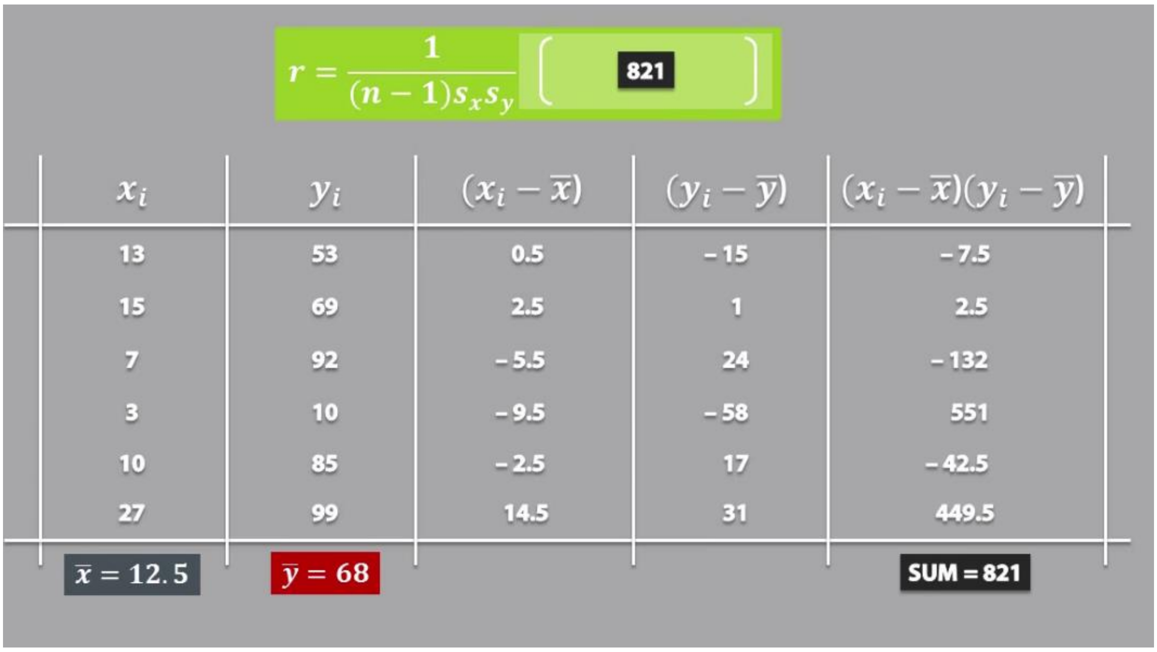
$$r = \frac{1}{(n-1)s_x s_y} \Sigma(x_i - \bar{x})(y_i - \bar{y})$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
13	53	0.5	−15	−7.5
15	69	2.5	1	2.5
7	92	−5.5	24	−132
3	10	−9.5	−58	551
10	85	−2.5	17	−42.5
27	99	14.5	31	449.5
$\bar{x} = 12.5$	$\bar{y} = 68$			



$$r = \frac{1}{(n-1)s_x s_y} \Sigma(x_i - \bar{x})(y_i - \bar{y})$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
13	53	0.5	−15	−7.5
15	69	2.5	1	2.5
7	92	−5.5	24	−132
3	10	−9.5	−58	551
10	85	−2.5	17	−42.5
27	99	14.5	31	449.5
$\bar{x} = 12.5$	$\bar{y} = 68$			SUM = 821



$$r = \frac{1}{(n-1)s_x s_y} \left(821 \right)$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
13	53	0.5	−15	−7.5
15	69	2.5	1	2.5
7	92	−5.5	24	−132
3	10	−9.5	−58	551
10	85	−2.5	17	−42.5
27	99	14.5	31	449.5
$\bar{x} = 12.5$	$\bar{y} = 68$			SUM = 821

$$r = \frac{1}{(6-1)s_x s_y}\left(821\right)$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
13	53	0.5	− 15	− 7.5
15	69	2.5	1	2.5
7	92	− 5.5	24	− 132
3	10	− 9.5	− 58	551
10	85	− 2.5	17	− 42.5
27	99	14.5	31	449.5
$\bar{x} = 12.5$	$\bar{y} = 68$			SUM = 821

$$r = \frac{1}{(6-1)s_x s_y}\left(821\right)$$

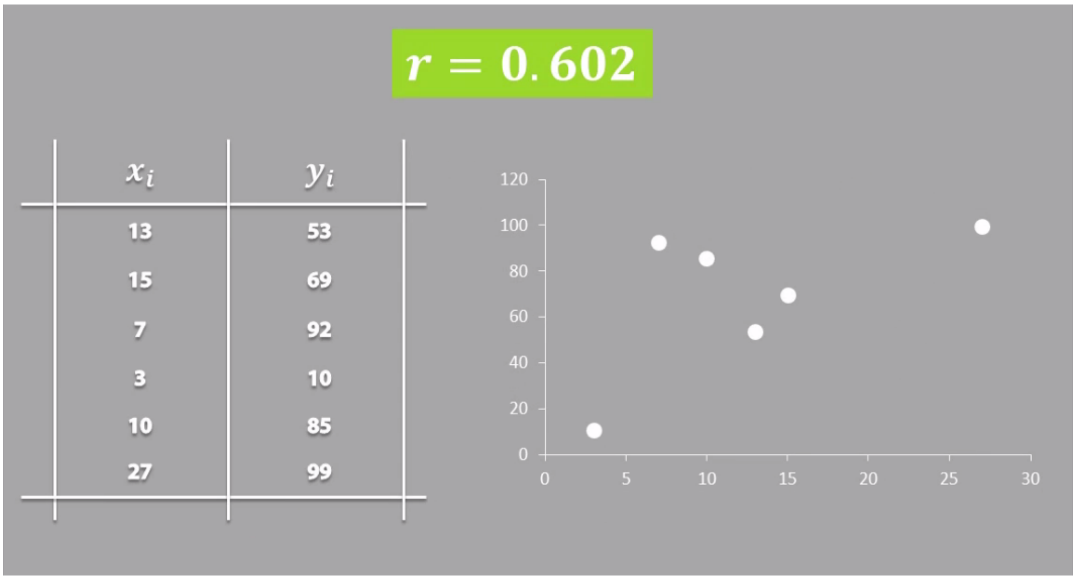
x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
13	53	0.5	− 15	− 7.5
15	69	2.5	1	2.5
7	92	− 5.5	24	− 132
3	10	− 9.5	− 58	551
10	85	− 2.5	17	− 42.5
27	99	14.5	31	449.5
$\bar{x} = 12.5$	$\bar{y} = 68$			SUM = 821
$s_x = 8.28$	$s_y = 32.91$			

$$r = \frac{1}{(6 - 1)(8.28)(32.91)}\left(821\right)$$

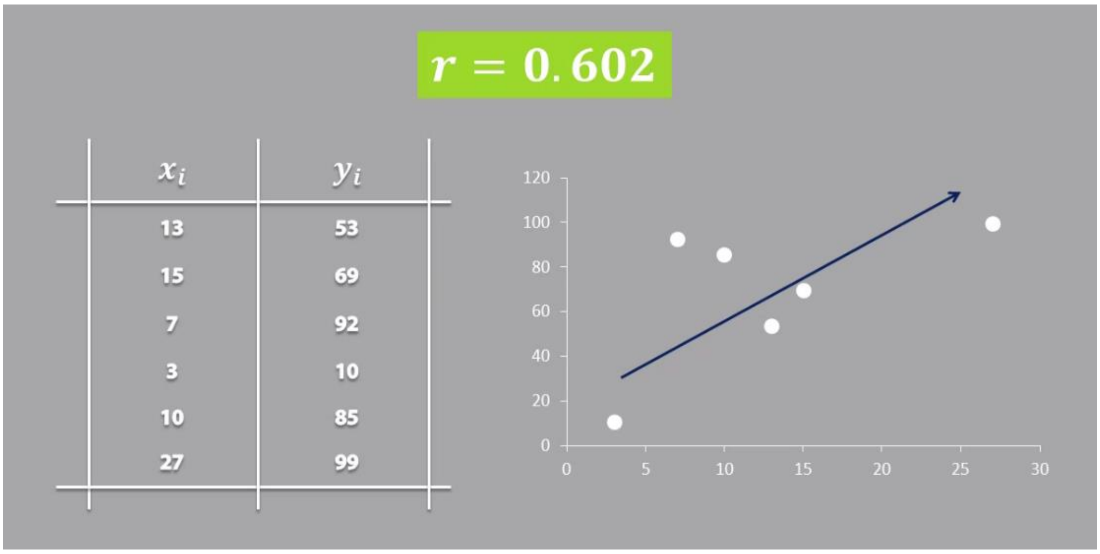
x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})(y_i - \overline{y})$
13	53	0.5	– 15	– 7.5
15	69	2.5	1	2.5
7	92	– 5.5	24	– 132
3	10	– 9.5	– 58	551
10	85	– 2.5	17	– 42.5
27	99	14.5	31	449.5
$\overline{x} = 12.5$ $s_x = 8.28$	$\overline{y} = 68$ $s_y = 32.91$			SUM = 821

$$r = 0.602$$

x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})(y_i - \overline{y})$
13	53	0.5	– 15	– 7.5
15	69	2.5	1	2.5
7	92	– 5.5	24	– 132
3	10	– 9.5	– 58	551
10	85	– 2.5	17	– 42.5
27	99	14.5	31	449.5
$\overline{x} = 12.5$ $s_x = 8.28$	$\overline{y} = 68$ $s_y = 32.91$			SUM = 821



- ✓ So, here its explains the correlation is +0.6, its upwards direction



7. Maths - Statistics – PART – 7

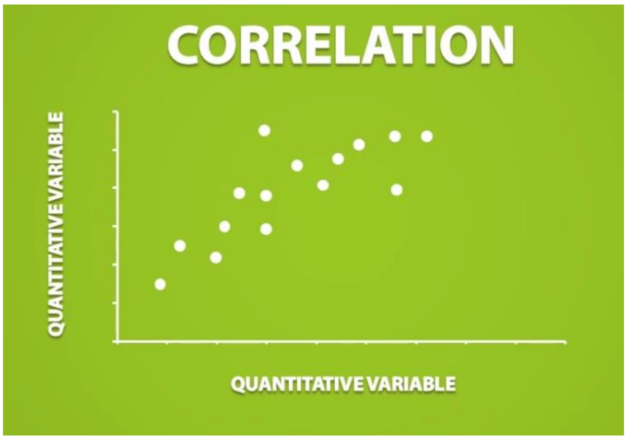
Contents

7. Maths - Statistics – PART – 7

- ✓ In this chapter we will discuss about Regression and R-Squared

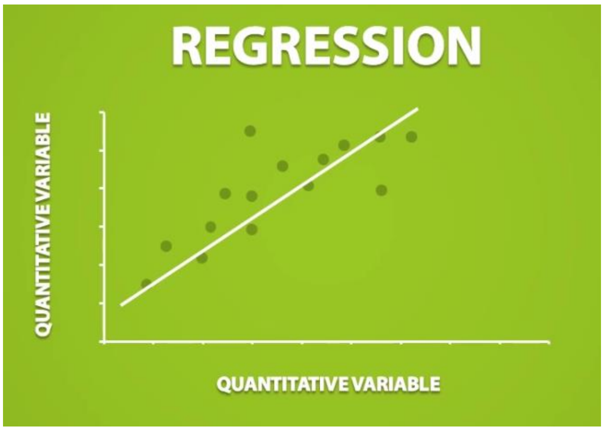
1. Correlation

- ✓ Correlation explains about how we measure the direction and strength of linear relationship in between quantitative variables



2. Regression

- ✓ Regression explains about how we can draw line in between the points
- ✓ This line represents about the pattern of the data
- ✓ This line is called as regression line

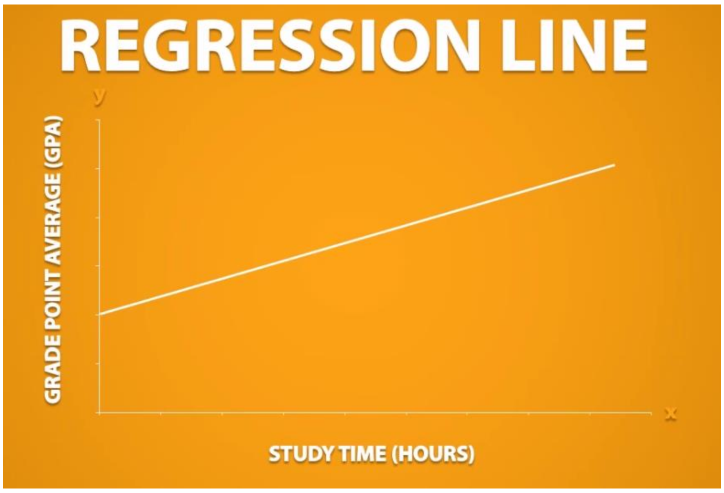


3. Regression line

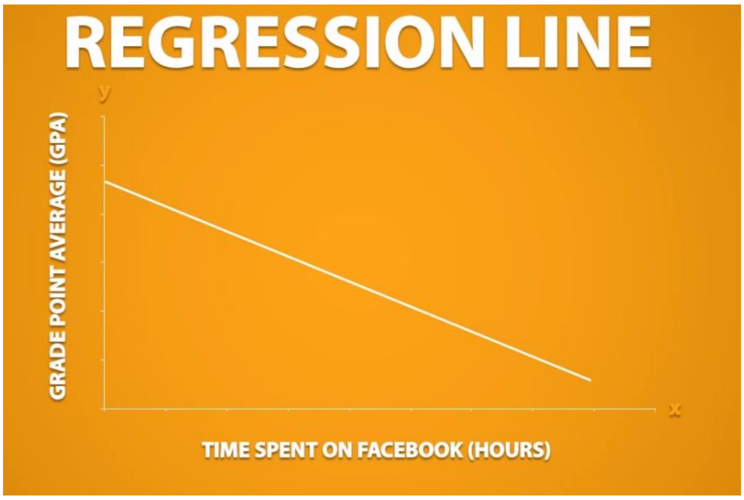
- ✓ Regression line predicts the change in **Y** when **X** increases by one unit
- ✓ Here change in Y can be increase/decrease

3.1. Example: Regression line

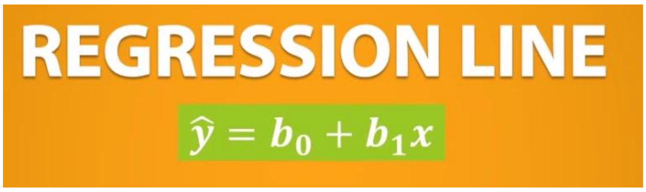
- ✓ If we study time (hours) on X axis and grade on Y axis then we should access some positive relationship in between these two variables.
- ✓ Generally speaking the more you study then we will get good grade ☺

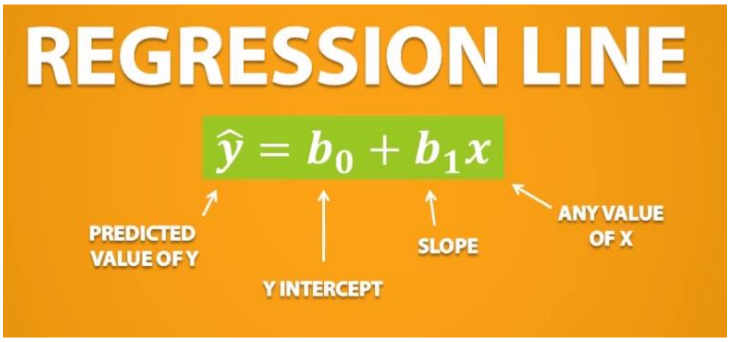


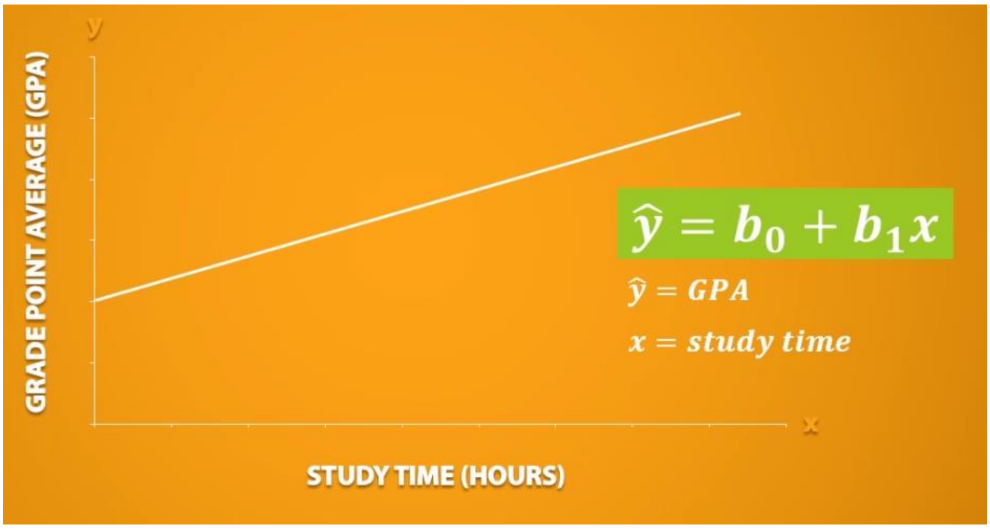
- ✓ Instead of studying if we spend more time on Facebook then we will get negative relationship

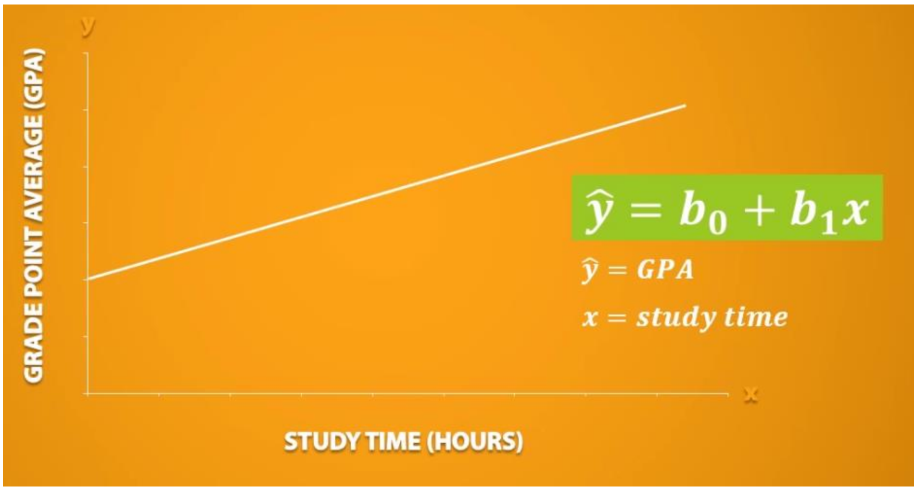


4.1. Regression line formula

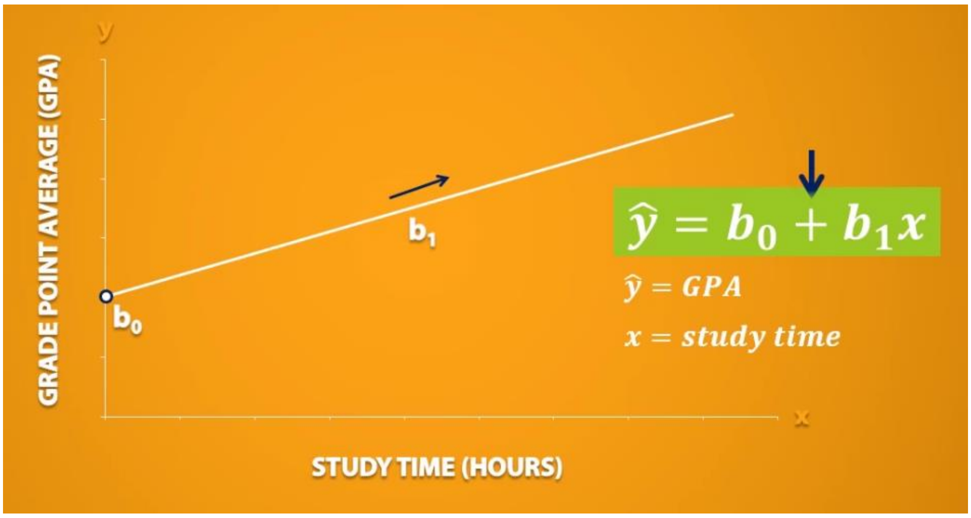








y
GRADE POINT AVERAGE (GPA)
$\hat{y} = b_0 + b_1 x$
$\hat{y} = GPA$
$x = study\ time$
x
STUDY TIME (HOURS)



y
GRADE POINT AVERAGE (GPA)
b_1
b_0
$\hat{y} = b_0 + b_1 x$
$\hat{y} = GPA$
$x = study\ time$
x
STUDY TIME (HOURS)

$$\hat{y} = b_0 + b_1 x$$

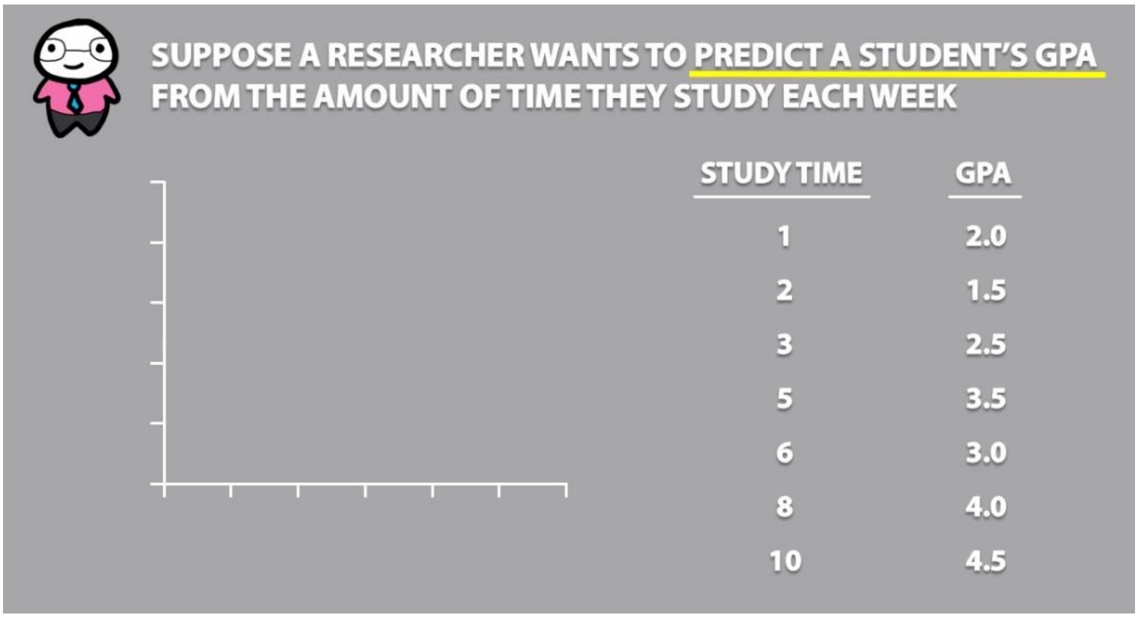
$$b_0 = \bar{y} - b_1 \bar{x}$$

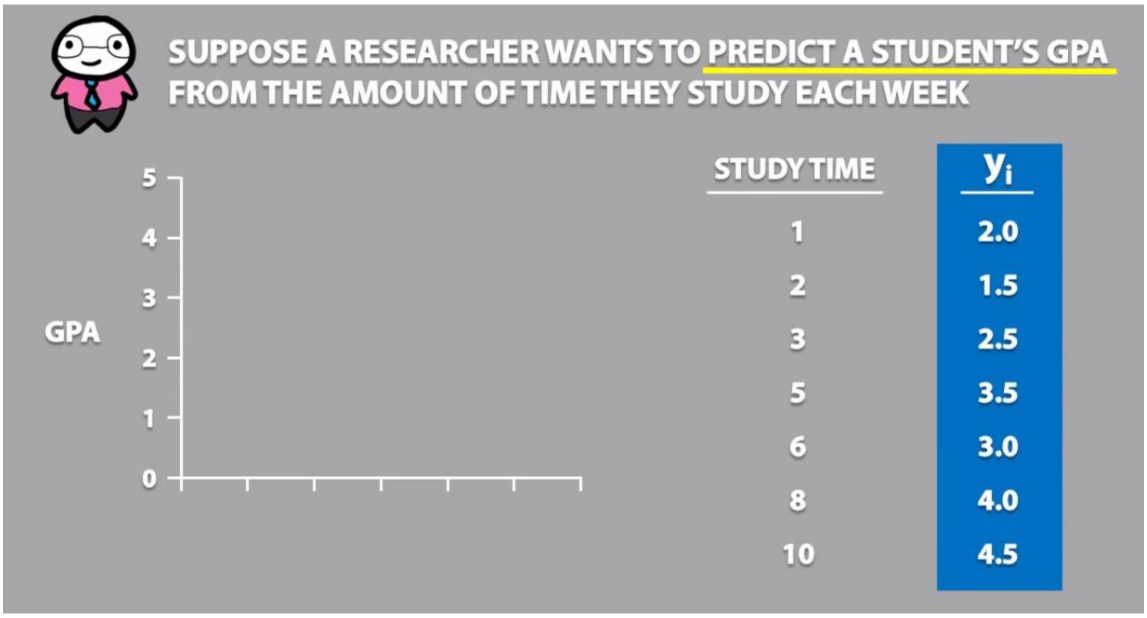
$$\hat{y} = b_0 + b_1 x$$

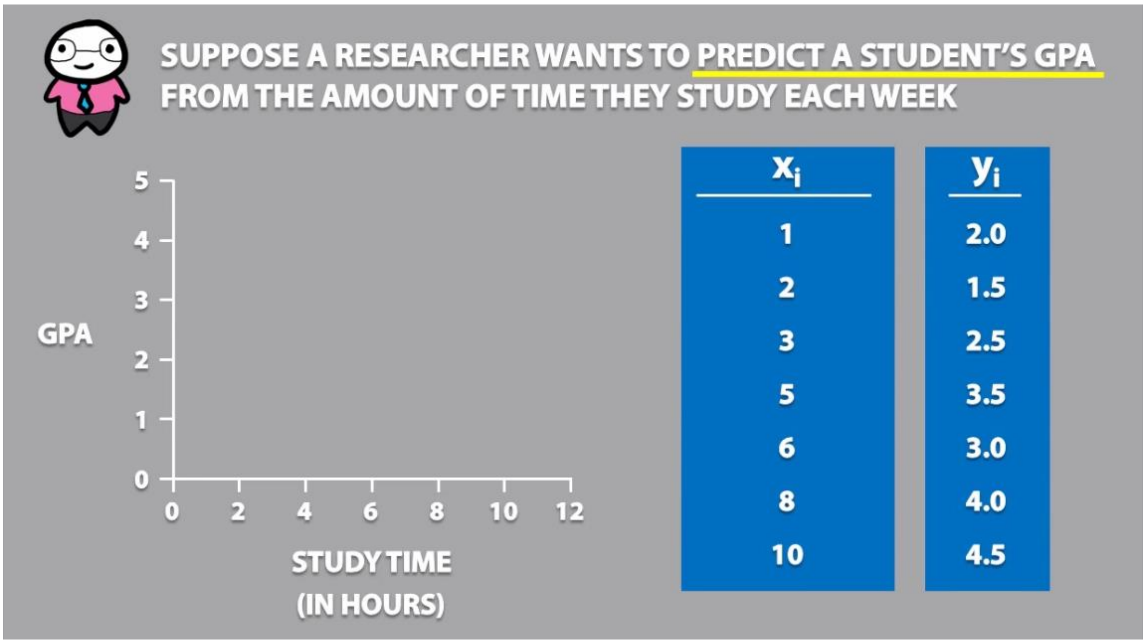
$$b_0 = \bar{y} - b_1 \bar{x} \qquad b_1 = r \frac{s_y}{s_x}$$

Scenario

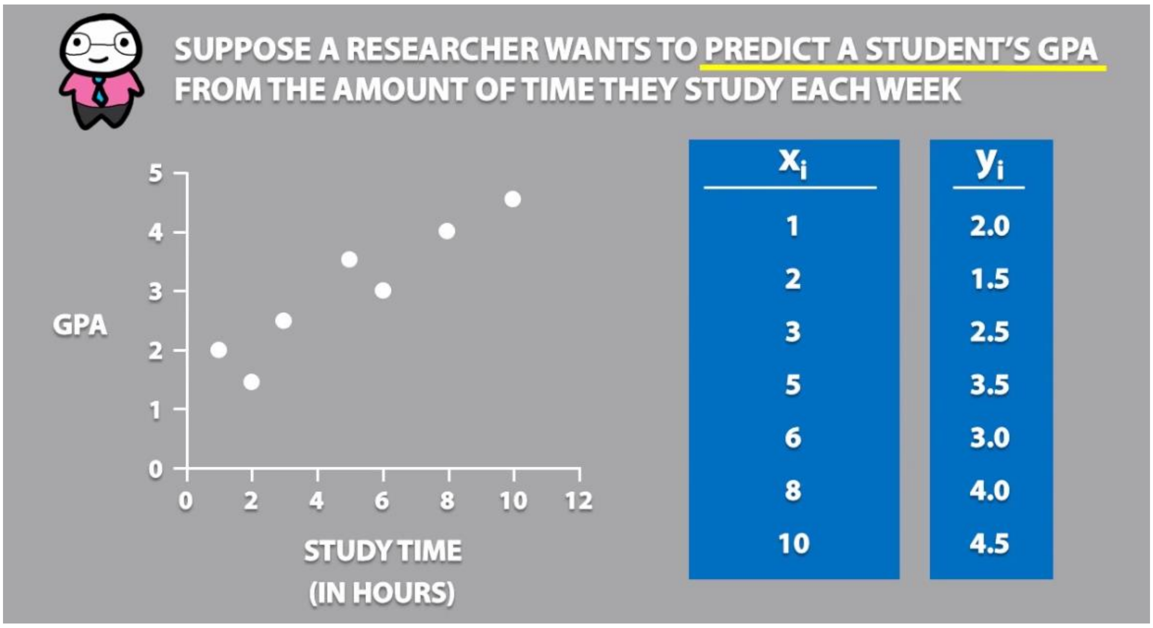
STUDENT	STUDY TIME	GPA
GEORGE	1	2.0
VANESSA	2	1.5
GATSBY	3	2.5
SOPHIA	5	3.5
EMMA	6	3.0
MELISSA	8	4.0
PATRICK	10	4.5

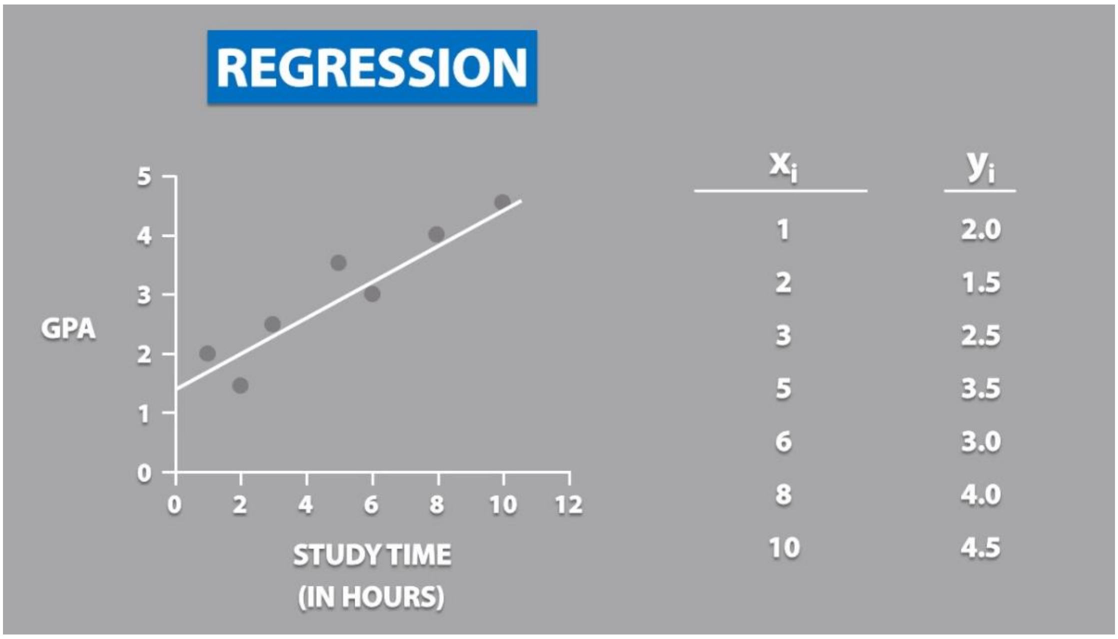

SUPPOSE A RESEARCHER WANTS TO PREDICT A STUDENT'S GPA FROM THE AMOUNT OF TIME THEY STUDY EACH WEEK
STUDY TIME
GPA
1
2.0
2
1.5
3
2.5
5
3.5
6
3.0
8
4.0
10
4.5

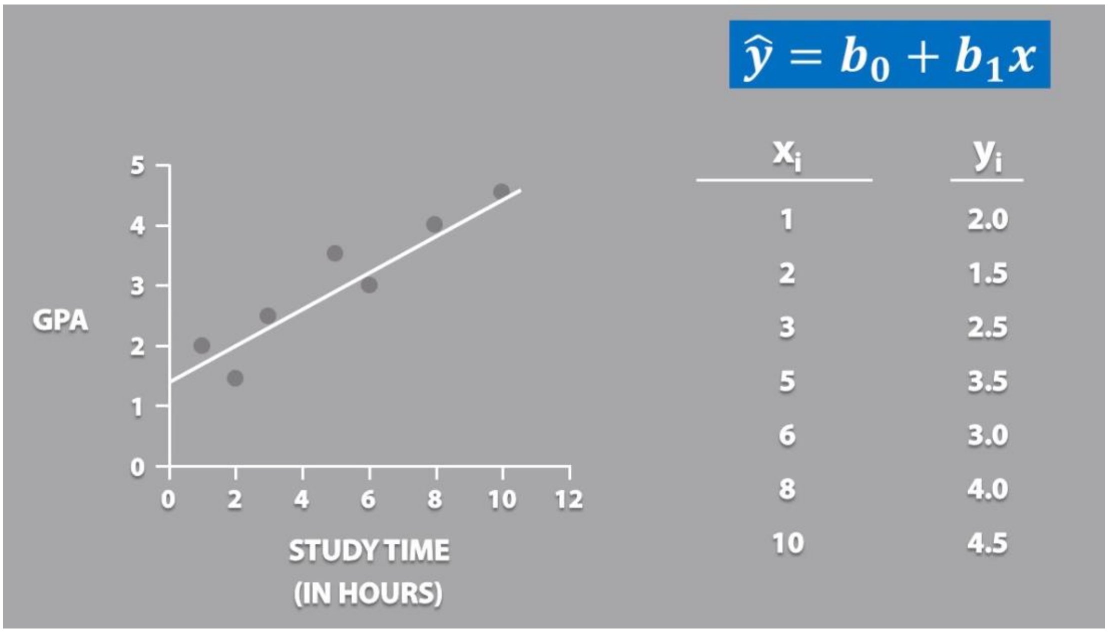

SUPPOSE A RESEARCHER WANTS TO PREDICT A STUDENT'S GPA FROM THE AMOUNT OF TIME THEY STUDY EACH WEEK
GPA
5
4
3
2
1
0
STUDY TIME
y_i
1
2.0
2
1.5
3
2.5
5
3.5
6
3.0
8
4.0
10
4.5

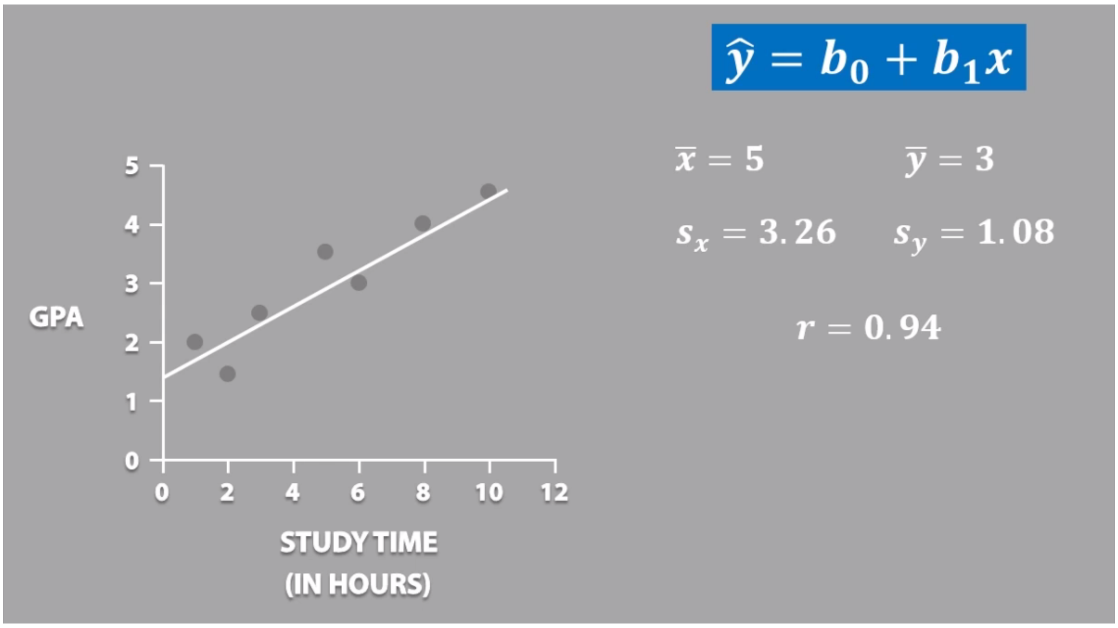


✓ Let's use scatter plot to plot this data

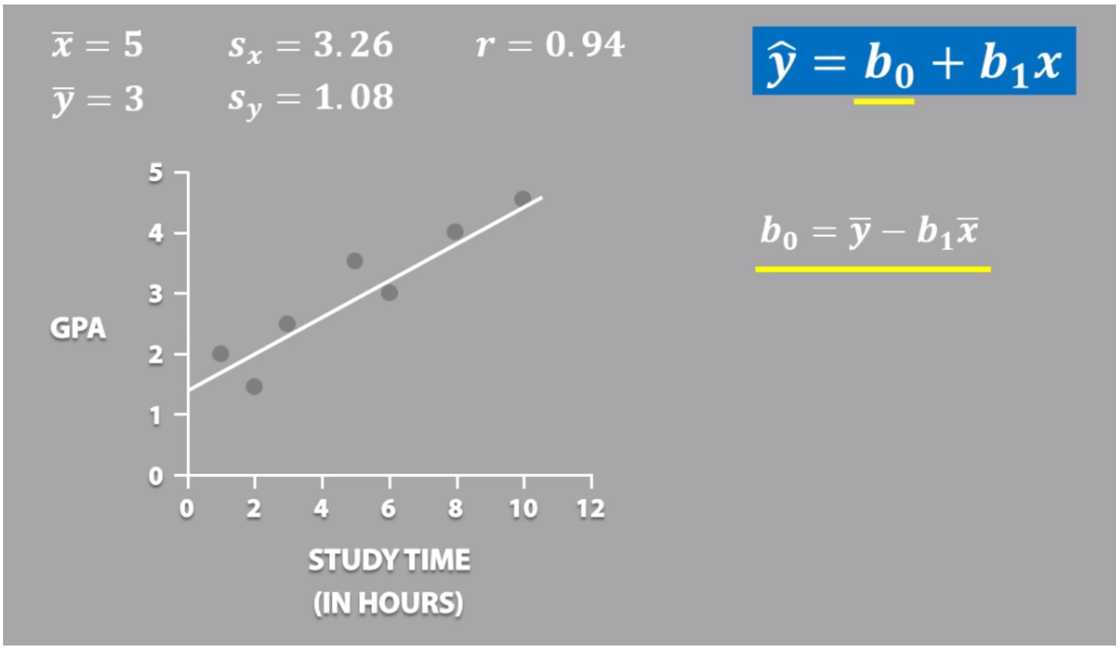



REGRESSION
GPA
STUDY TIME
(IN HOURS)
0 2 4 6 8 10 12
0 1 2 3 4 5
x_i y_i
1 2.0
2 1.5
3 2.5
5 3.5
6 3.0
8 4.0
10 4.5

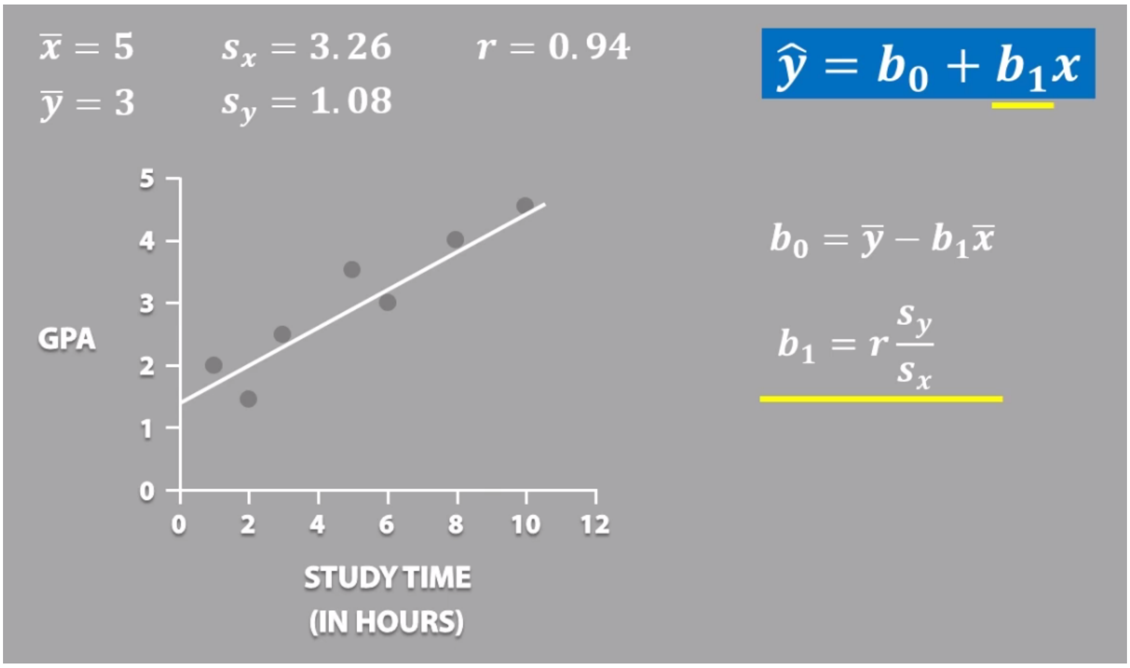

$\hat{y} = b_0 + b_1 x$
GPA
STUDY TIME
(IN HOURS)
0 2 4 6 8 10 12
0 1 2 3 4 5
x_i y_i
1 2.0
2 1.5
3 2.5
5 3.5
6 3.0
8 4.0
10 4.5

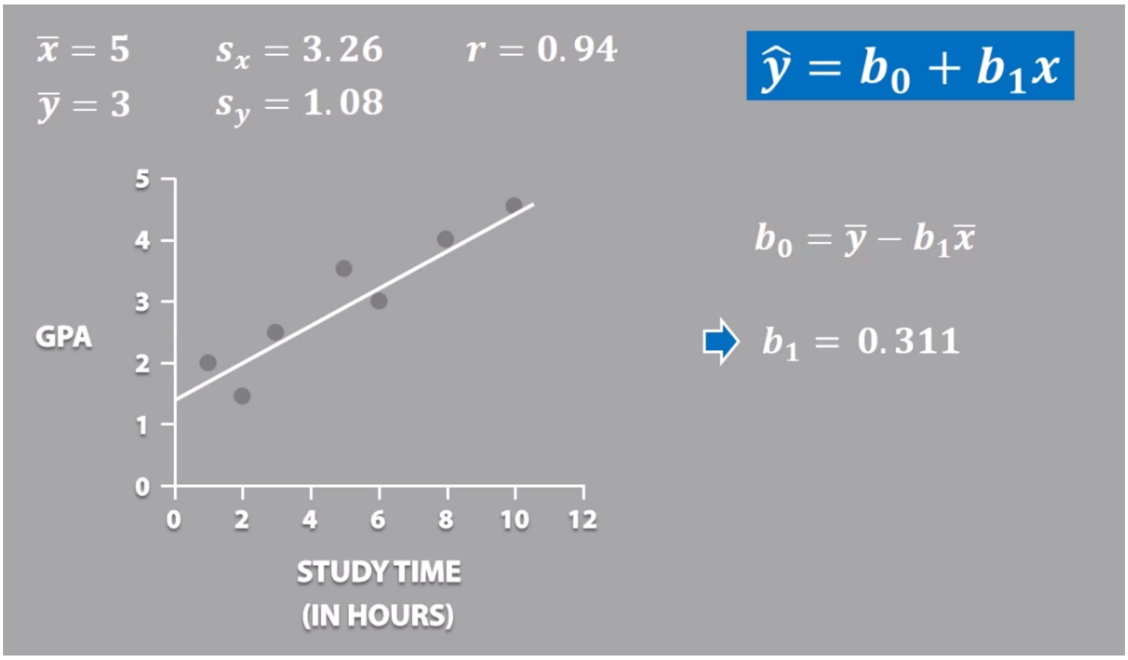


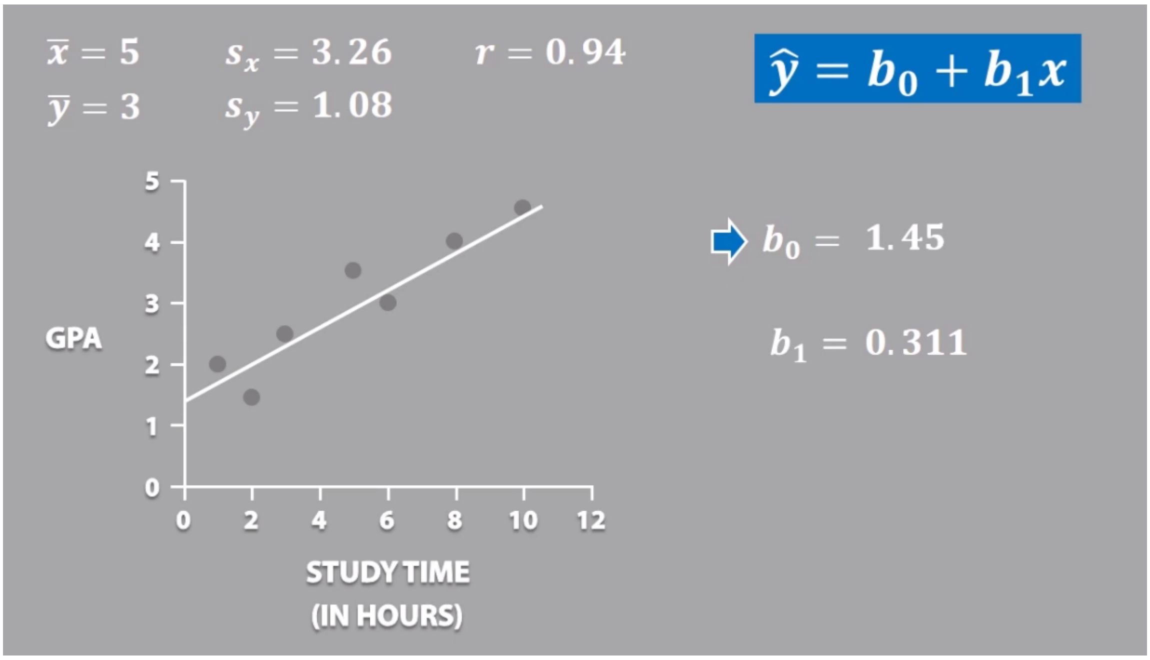
$\hat{y} = b_0 + b_1 x$
$\bar{x} = 5$
$\bar{y} = 3$
$s_x = 3.26$
$s_y = 1.08$
$r = 0.94$
GPA
0
1
2
3
4
5
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)

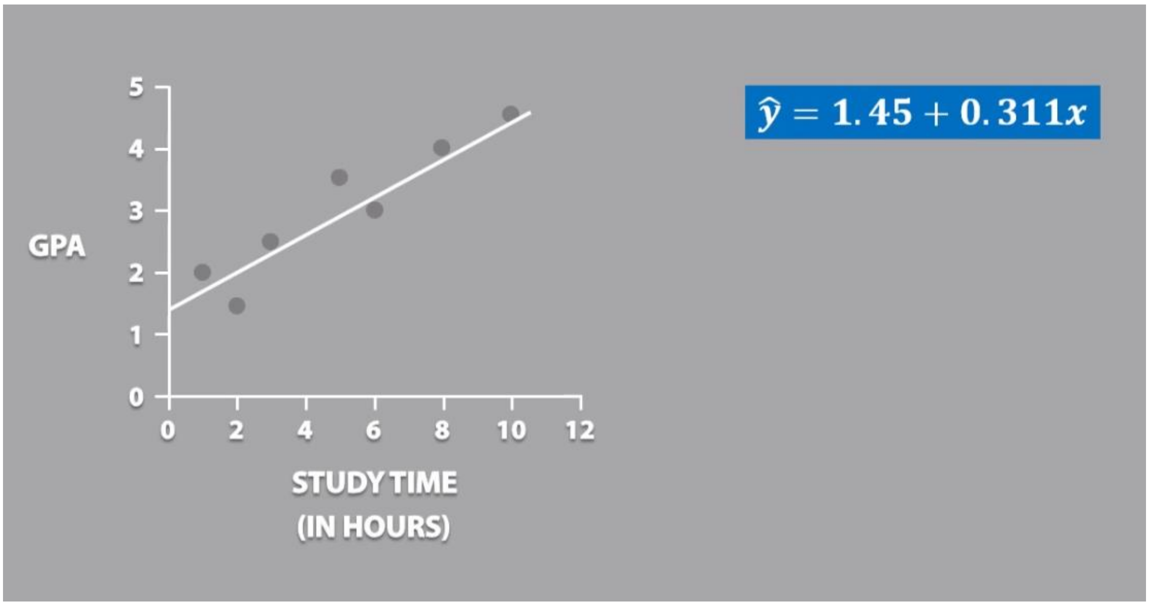


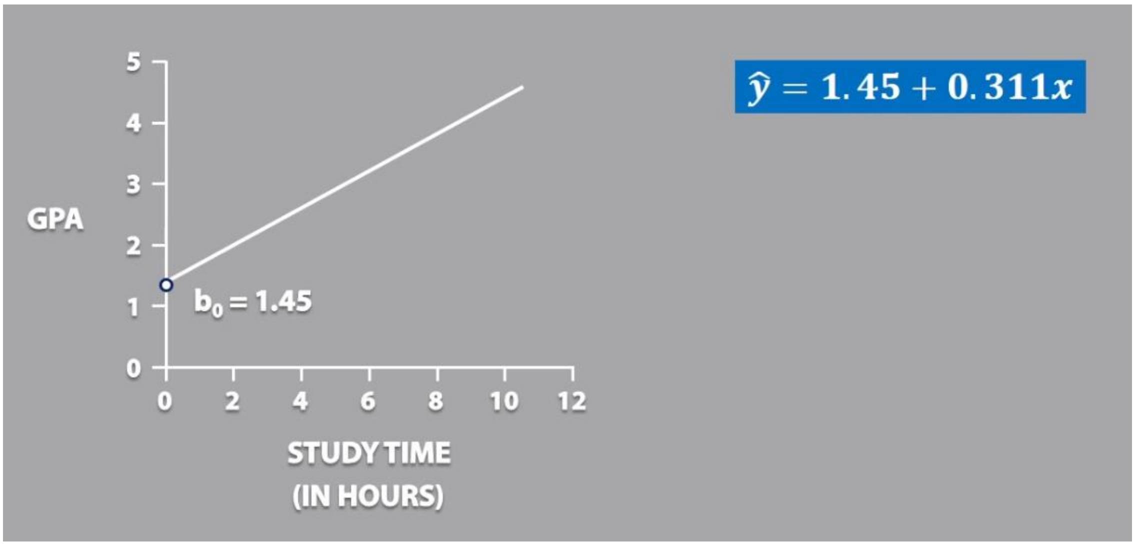
$\bar{x} = 5$
$s_x = 3.26$
$r = 0.94$
$\bar{y} = 3$
$s_y = 1.08$
$\hat{y} = b_0 + b_1 x$
$b_0 = \bar{y} - b_1 \bar{x}$
GPA
0
1
2
3
4
5
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)


$\bar{x} = 5$
$\bar{y} = 3$
$s_x = 3.26$
$s_y = 1.08$
$r = 0.94$
$\hat{y} = b_0 + b_1 x$
GPA
STUDY TIME (IN HOURS)
$b_0 = \bar{y} - b_1 \bar{x}$
$b_1 = r \frac{s_y}{s_x}$

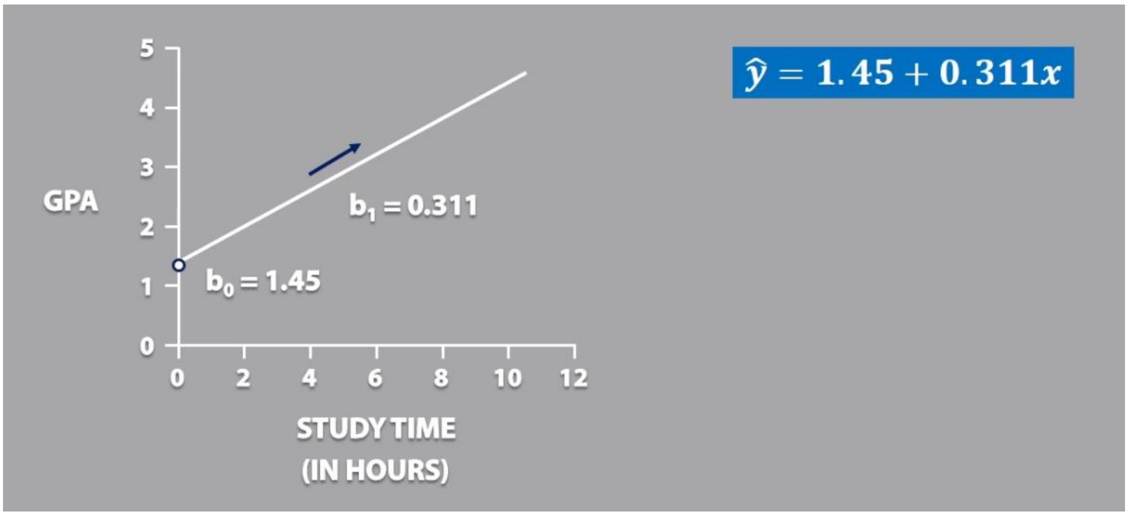

$\bar{x} = 5$
$\bar{y} = 3$
$s_x = 3.26$
$s_y = 1.08$
$r = 0.94$
$\hat{y} = b_0 + b_1 x$
GPA
STUDY TIME (IN HOURS)
$b_0 = \bar{y} - b_1 \bar{x}$
$b_1 = 0.311$


$\bar{x} = 5$
$\bar{y} = 3$
$s_x = 3.26$
$s_y = 1.08$
$r = 0.94$
$\hat{y} = b_0 + b_1 x$
GPA
5
4
3
2
1
0
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)
$b_0 = 1.45$
$b_1 = 0.311$


GPA
5
4
3
2
1
0
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)
$\hat{y} = 1.45 + 0.311x$



$\hat{y} = 1.45 + 0.311x$
GPA
5
4
3
2
1
0
$b_0 = 1.45$
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)

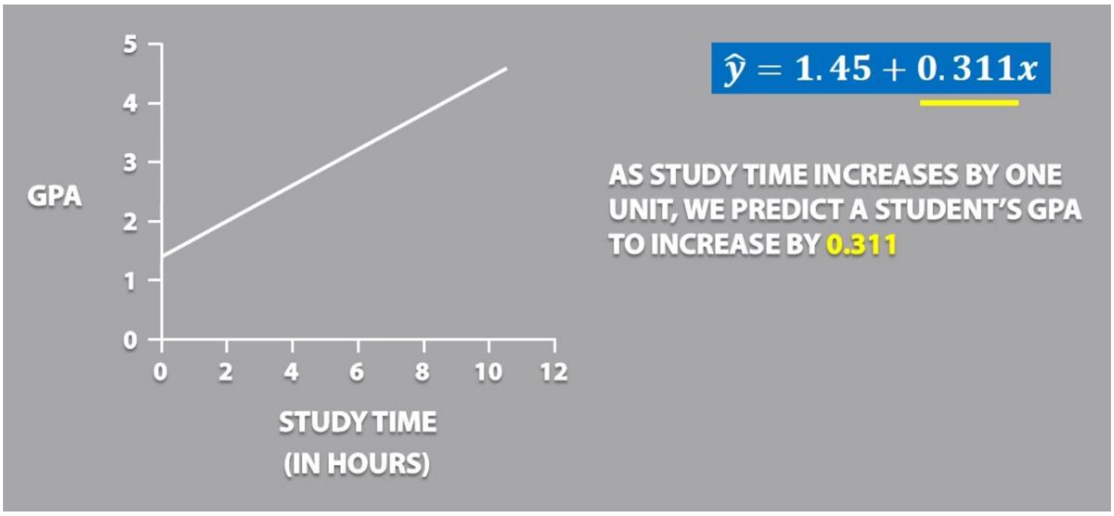


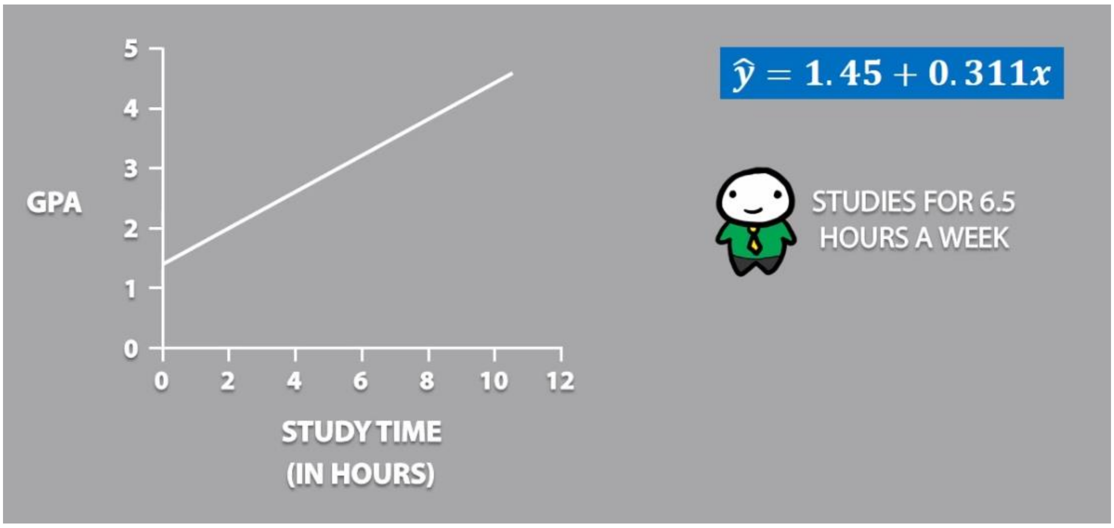
$\hat{y} = 1.45 + 0.311x$
GPA
5
4
3
2
1
0
$b_1 = 0.311$
$b_0 = 1.45$
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)

THE LINE OF LEAST SQUARES REGRESSION

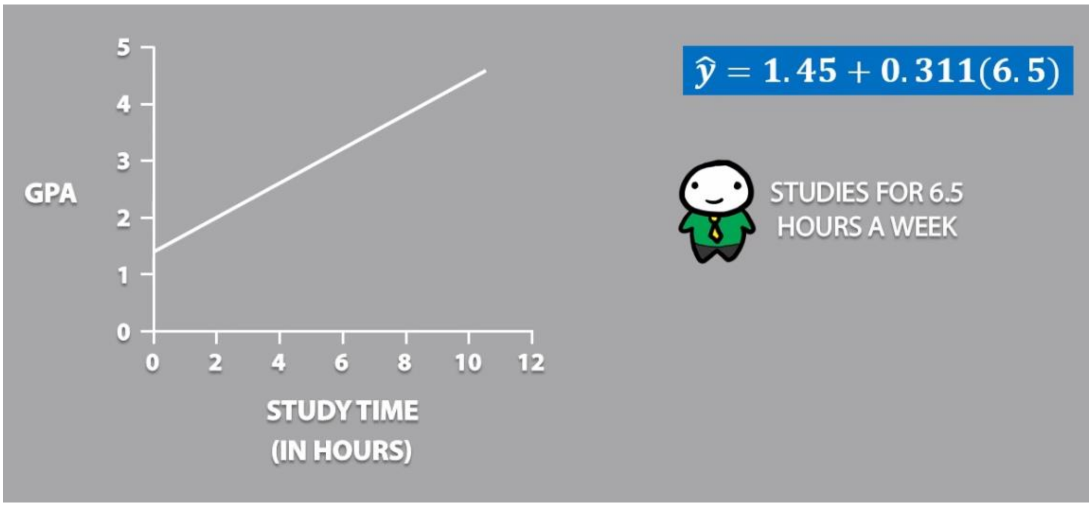
$$\hat{y} = b_0 + b_1 x$$

THE SLOPE OF A REGRESSION LINE PREDICTS THE CHANGE IN "Y" WHEN "X" INCREASES BY ONE UNIT

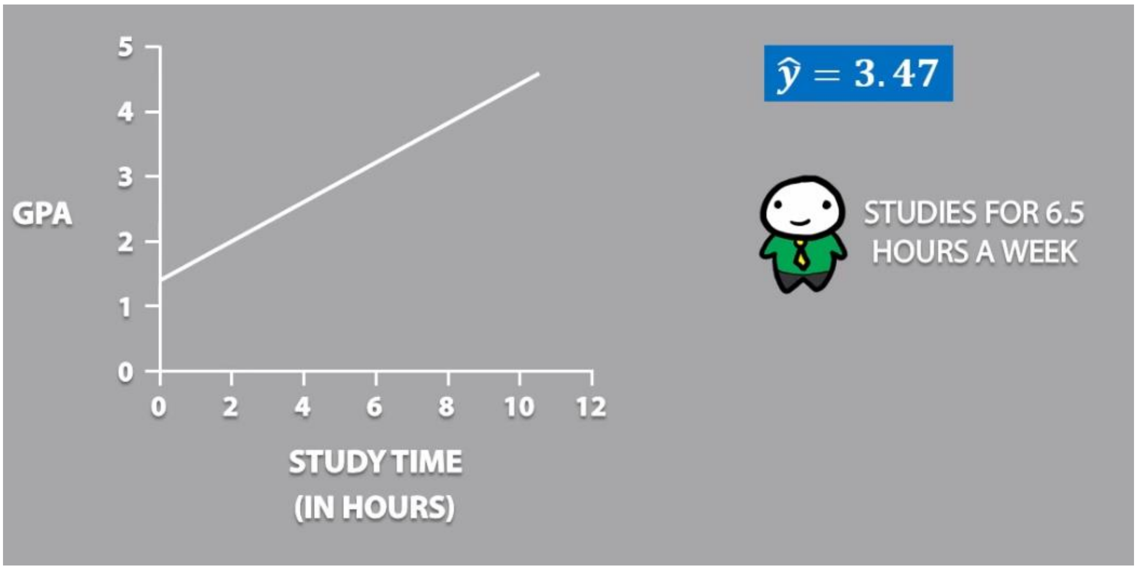




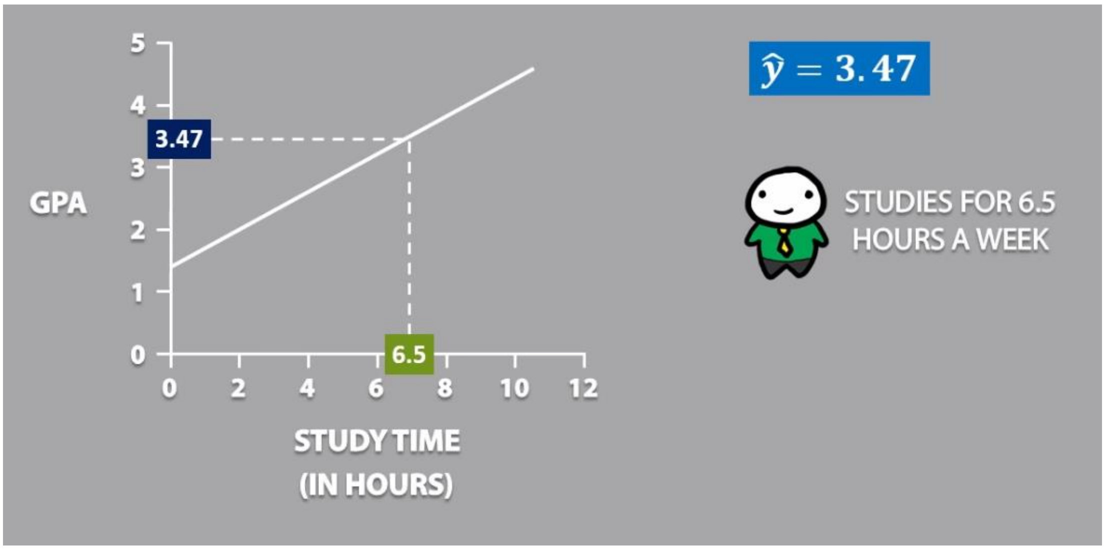
$\hat{y} = 1.45 + 0.311x$
GPA
5
4
3
2
1
0
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)
STUDIES FOR 6.5
HOURS A WEEK



$\hat{y} = 1.45 + 0.311(6.5)$
GPA
5
4
3
2
1
0
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)
STUDIES FOR 6.5
HOURS A WEEK



5
4
3
2
1
0
GPA
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)
$\hat{y} = 3.47$
STUDIES FOR 6.5
HOURS A WEEK



5
4
3.47
3
2
1
0
GPA
6.5
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)
$\hat{y} = 3.47$
STUDIES FOR 6.5
HOURS A WEEK

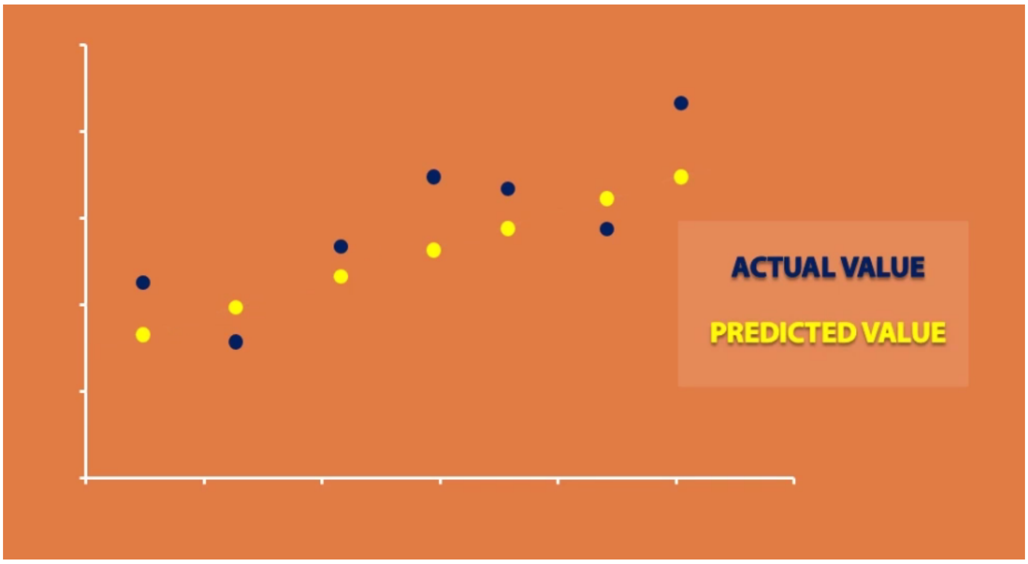
5. R – SQUARED

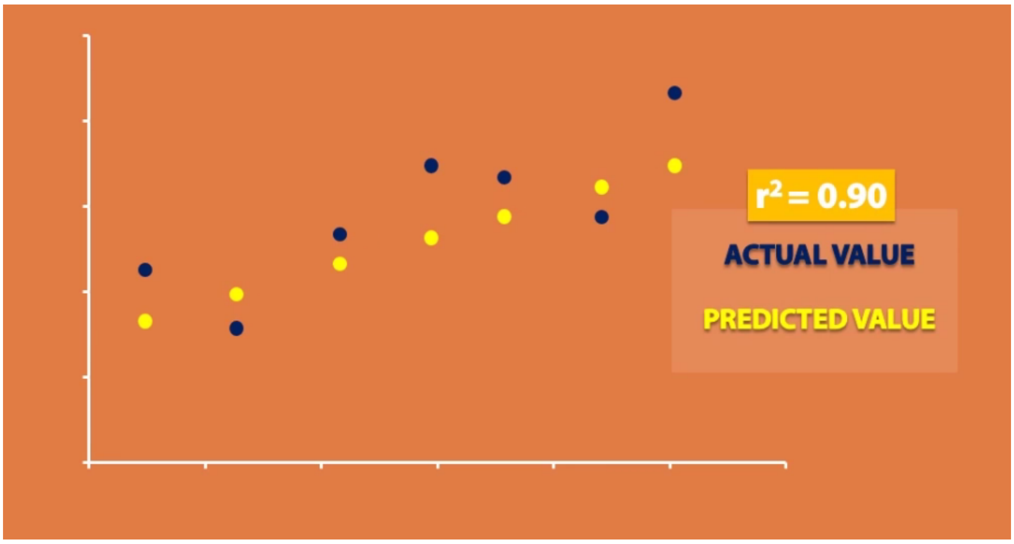
R-SQUARED

$$\text{R-SQUARED} = r^2$$

$$\text{R-SQUARED} = r^2 = r \times r$$


r
HAS VALUES BETWEEN -1 AND 1
MEASURES THE LINEAR RELATIONSHIP BETWEEN TWO QUANTITATIVE VARIABLES WITH RESPECT TO DIRECTION AND STRENGTH
r^2
HAS VALUES BETWEEN 0 AND 1
IS A MEASURE OF HOW CLOSE EACH DATA POINT FITS TO THE REGRESSION LINE
TELLS US HOW WELL THE REGRESSION LINE PREDICTS ACTUAL VALUES


ACTUAL VALUE
PREDICTED VALUE

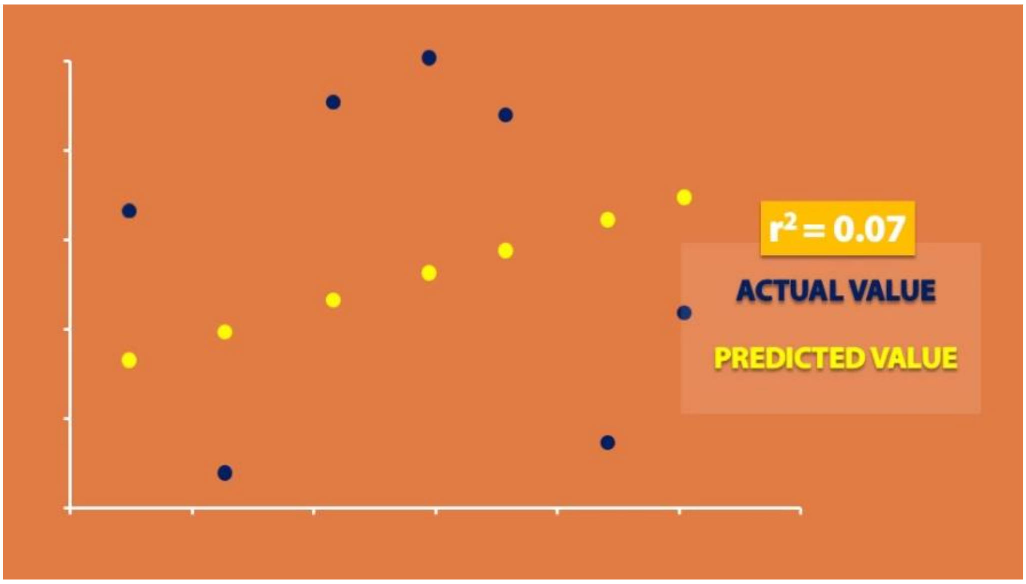

$r^2 = 0.90$
ACTUAL VALUE
PREDICTED VALUE

Case 1: r squared value is high

- ✓ A high value of r square value explains the actual values and predicted values are close together
- ✓ We can clear see the actual values are predicted values having very less distance in between them

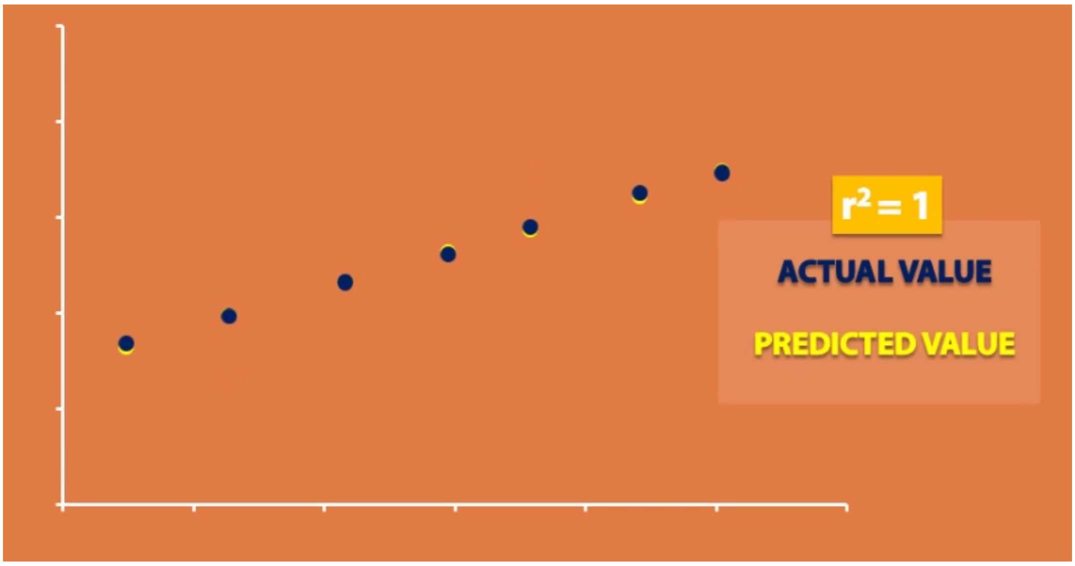
Case 2: r squared value is low

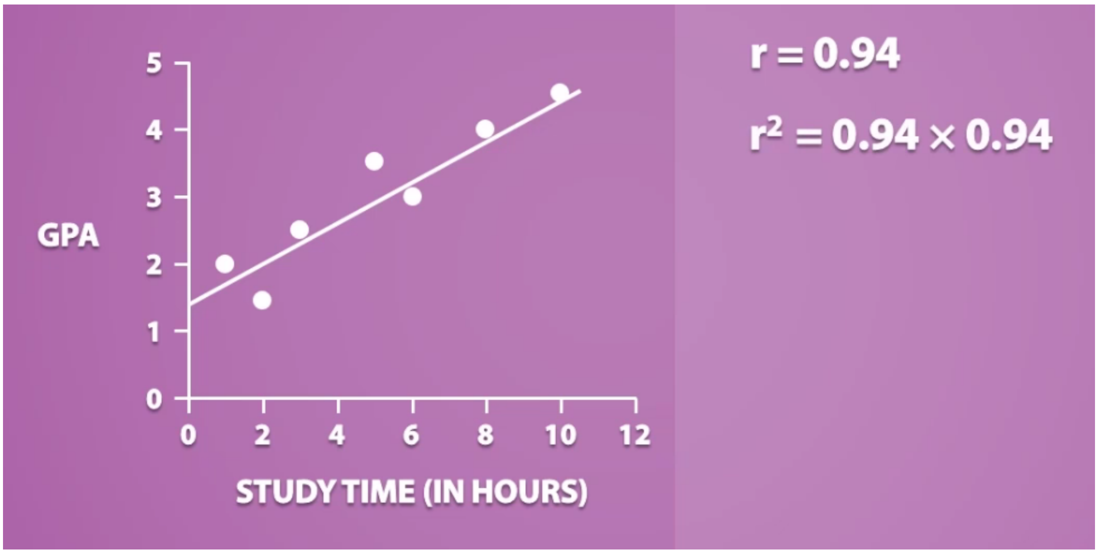
- ✓ A low value of r square value explains the regression line which the data points are does not fit the well
- ✓ We can clear see the actual values are predicted values having large distance in between them

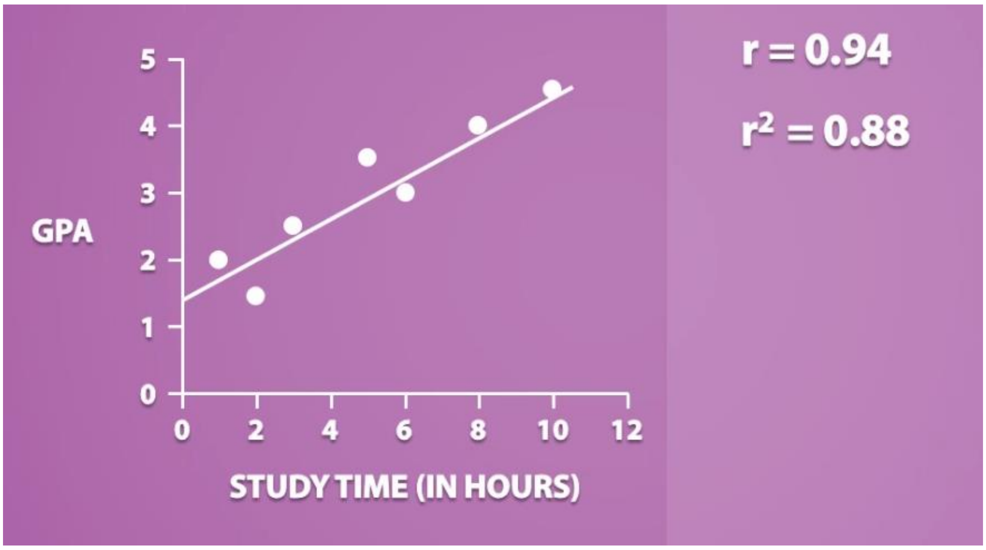


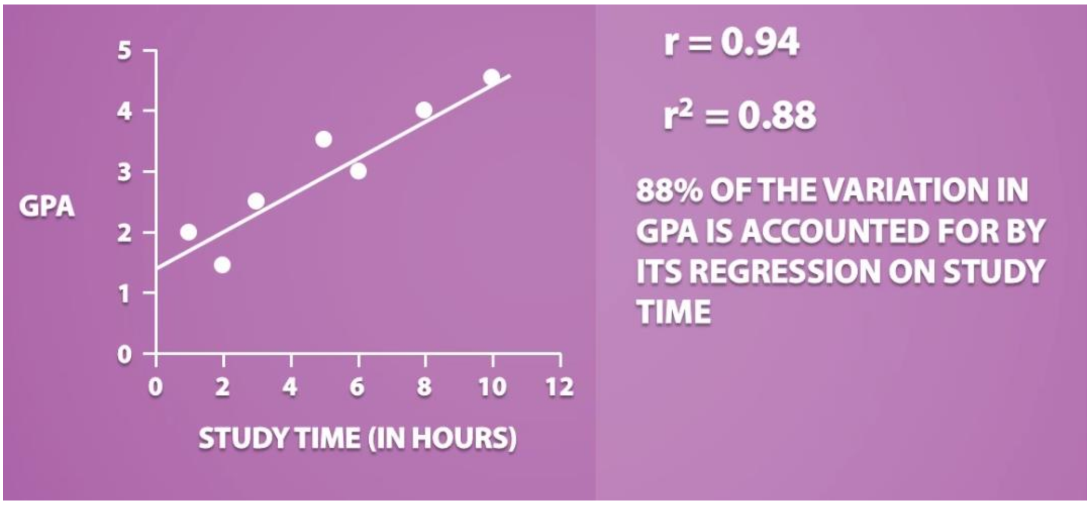
Case 3: r squared value is equals to one

- ✓ This means that we can predict value of y for any given value of x






r = 0.94
$r^2 = 0.88$
GPA
5
4
3
2
1
0
0
2
4
6
8
10
12
STUDY TIME (IN HOURS)


r = 0.94
$r^2 = 0.88$
88% OF THE VARIATION IN GPA IS ACCOUNTED FOR BY ITS REGRESSION ON STUDY TIME
GPA
5
4
3
2
1
0
0
2
4
6
8
10
12
STUDY TIME (IN HOURS)

8. Maths - Statistics – PART – 8

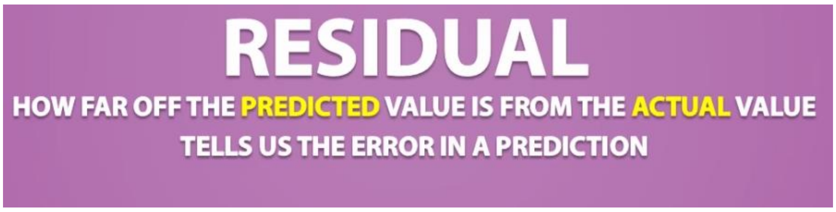
Contents

8. Maths - Statistics – PART – 8

- ✓ In this chapter we will discuss about residuals

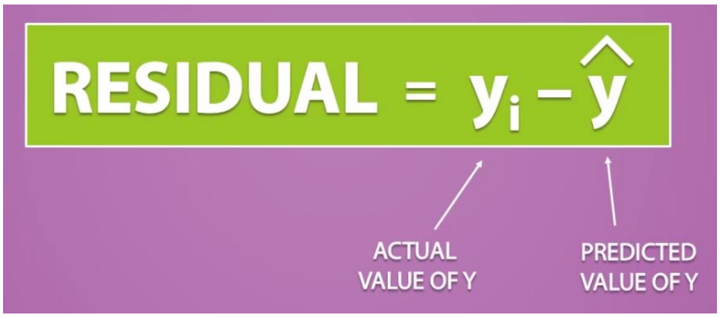
1. Residual

- ✓ It explains about how far the distance in between the predicted value from the actual value
- ✓ It basically tells the error in prediction

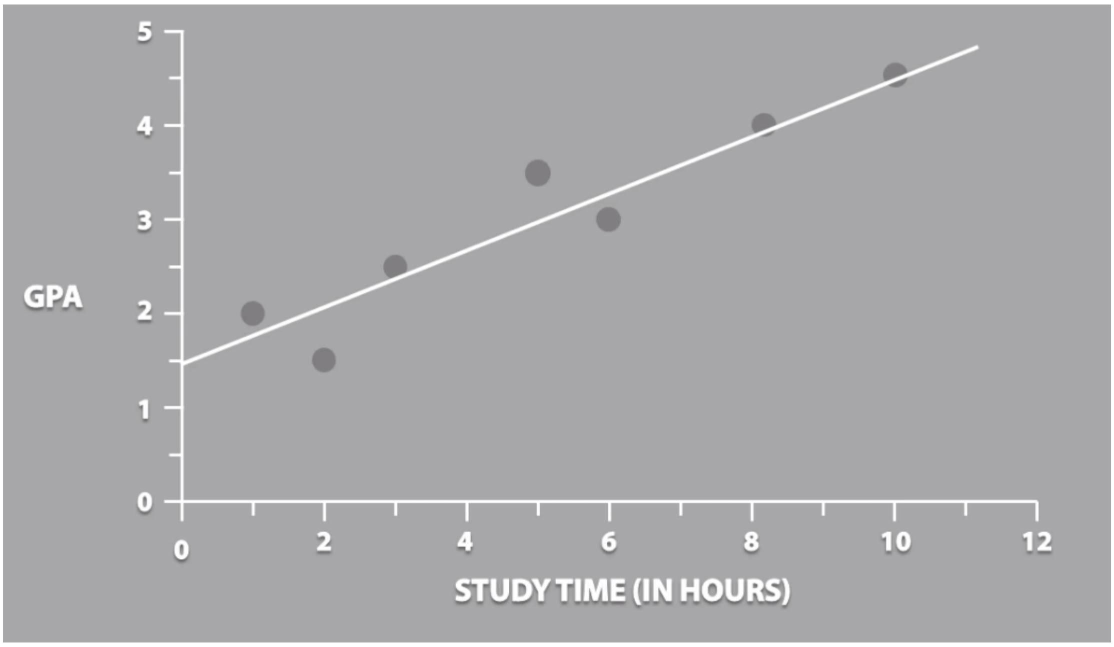


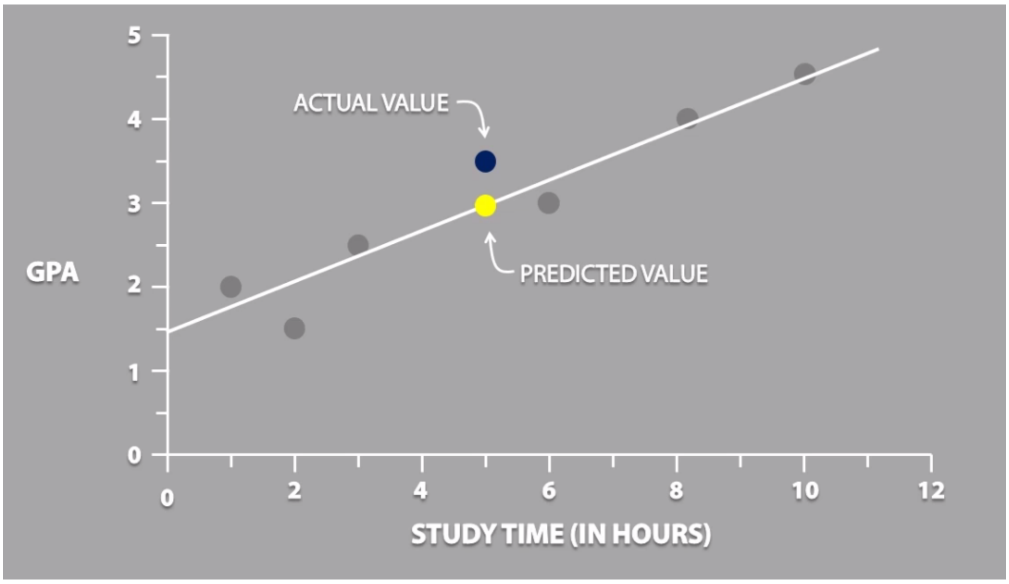
2. Formula

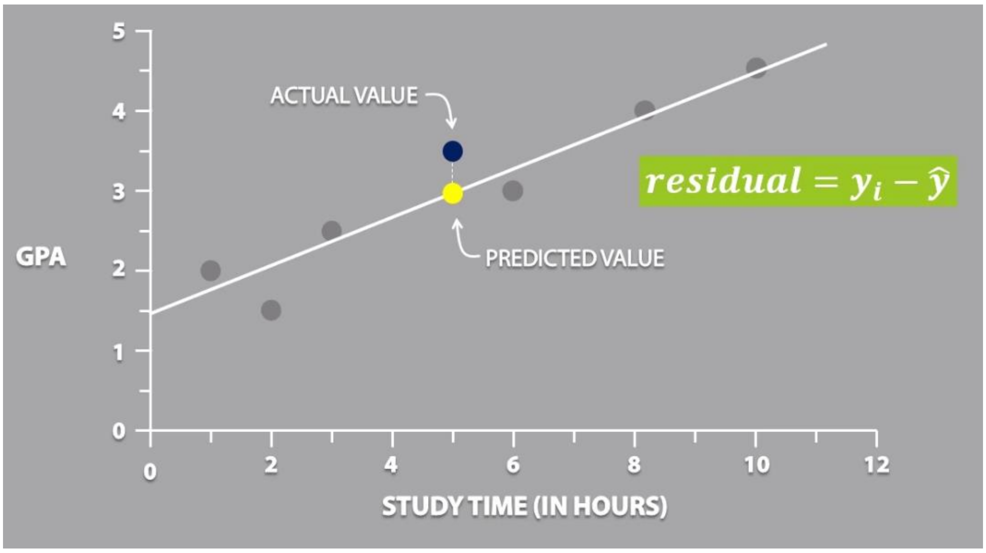
✓ The formula is, actual value of Y minus predicted value of Y

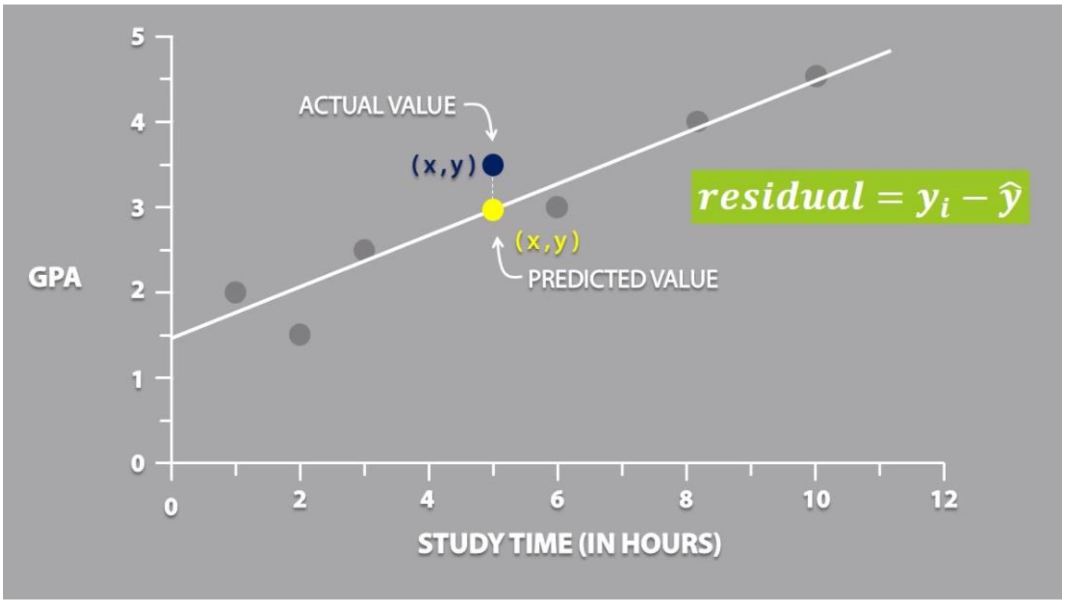


✓ From the previous example, study time on X axis and gpa on Y axis

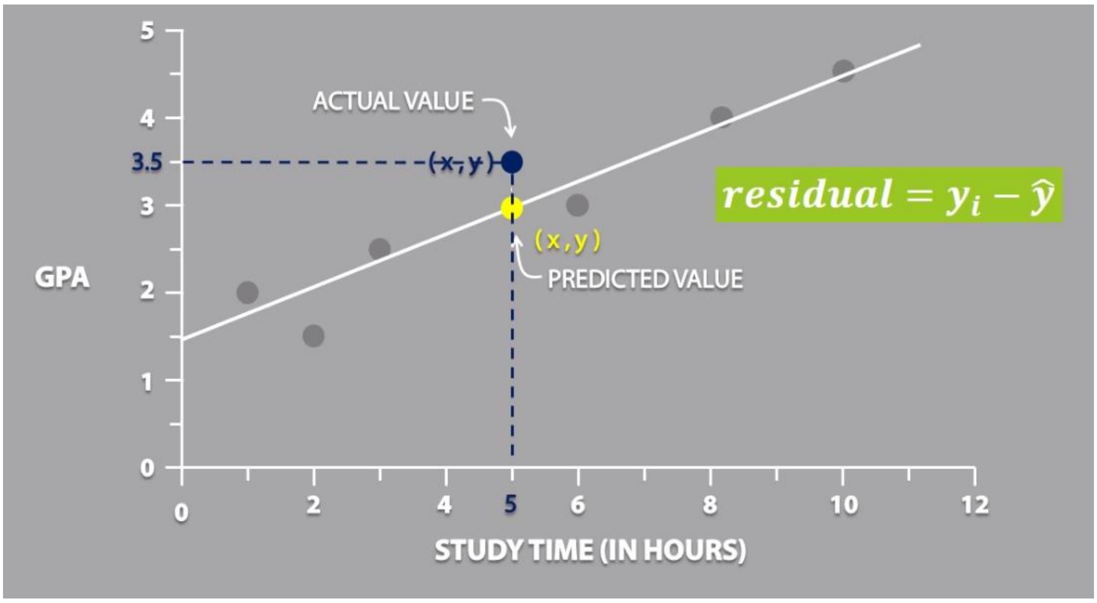



5
4
3
2
1
0
GPA
ACTUAL VALUE
PREDICTED VALUE
0
2
4
6
8
10
12
STUDY TIME (IN HOURS)

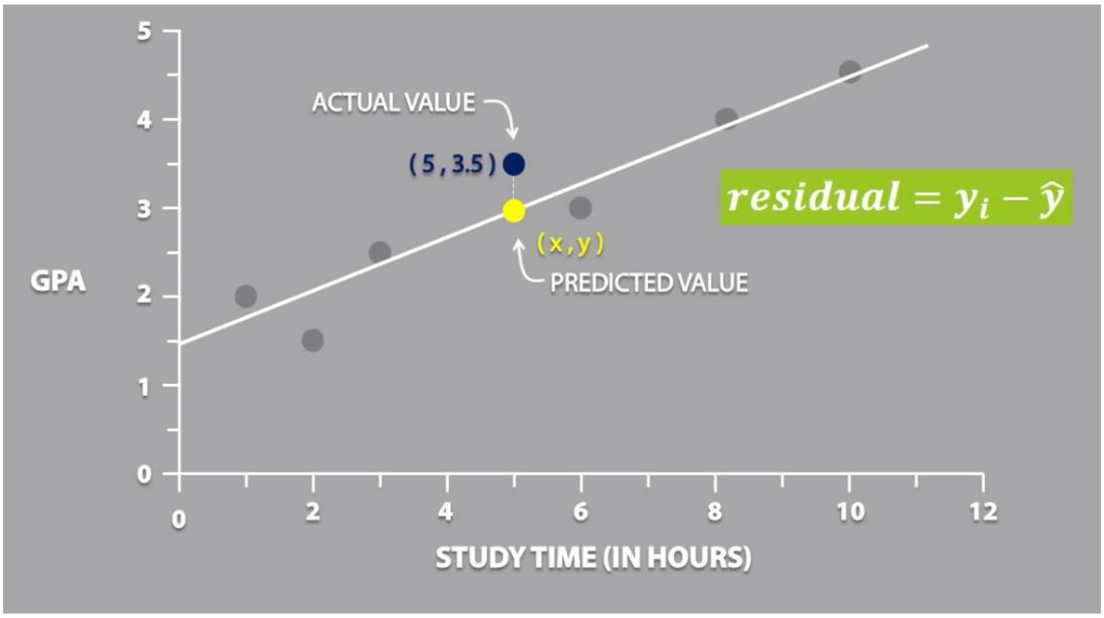

5
4
3
2
1
0
GPA
ACTUAL VALUE
PREDICTED VALUE
$residual = y_i - \hat{y}$
0
2
4
6
8
10
12
STUDY TIME (IN HOURS)

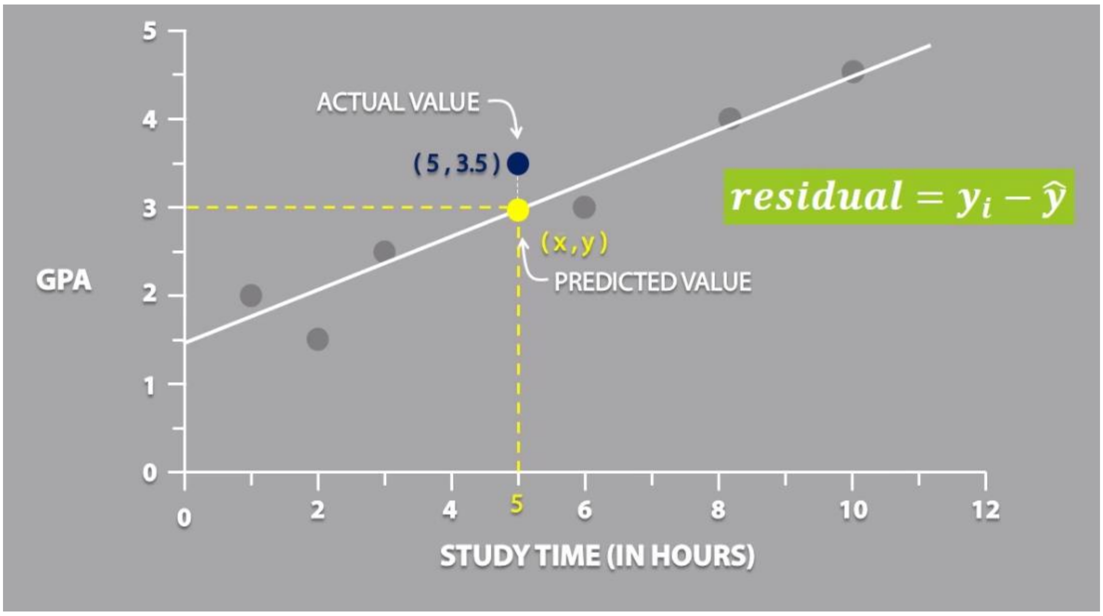


ACTUAL VALUE
(x,y)
(x,y)
PREDICTED VALUE
residual = $y_i - \hat{y}$
GPA
0
1
2
3
4
5
0
2
4
6
8
10
12
STUDY TIME (IN HOURS)

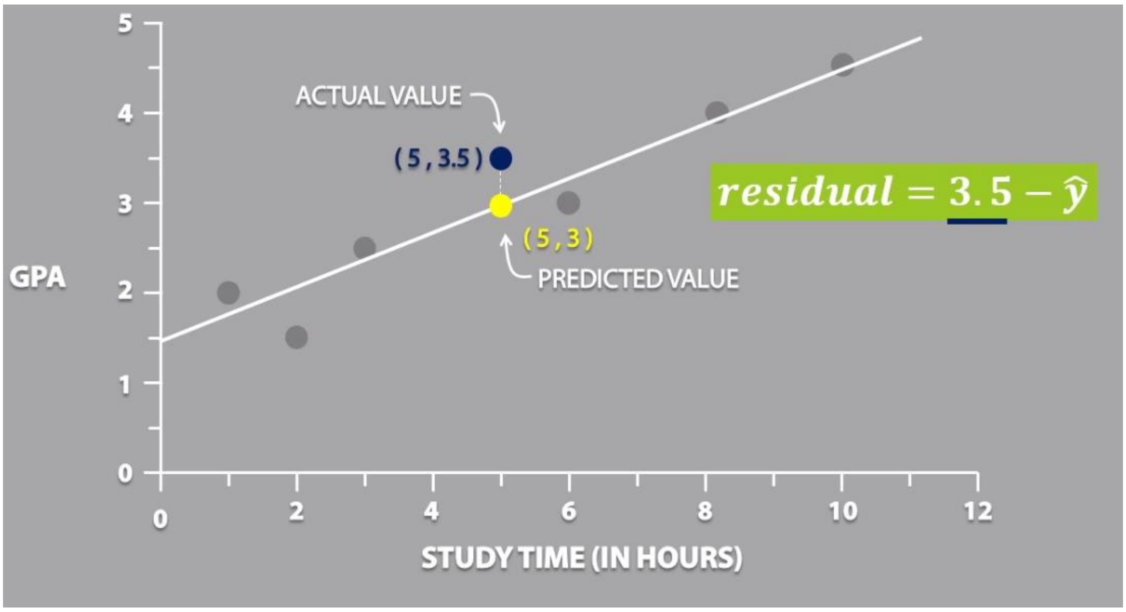


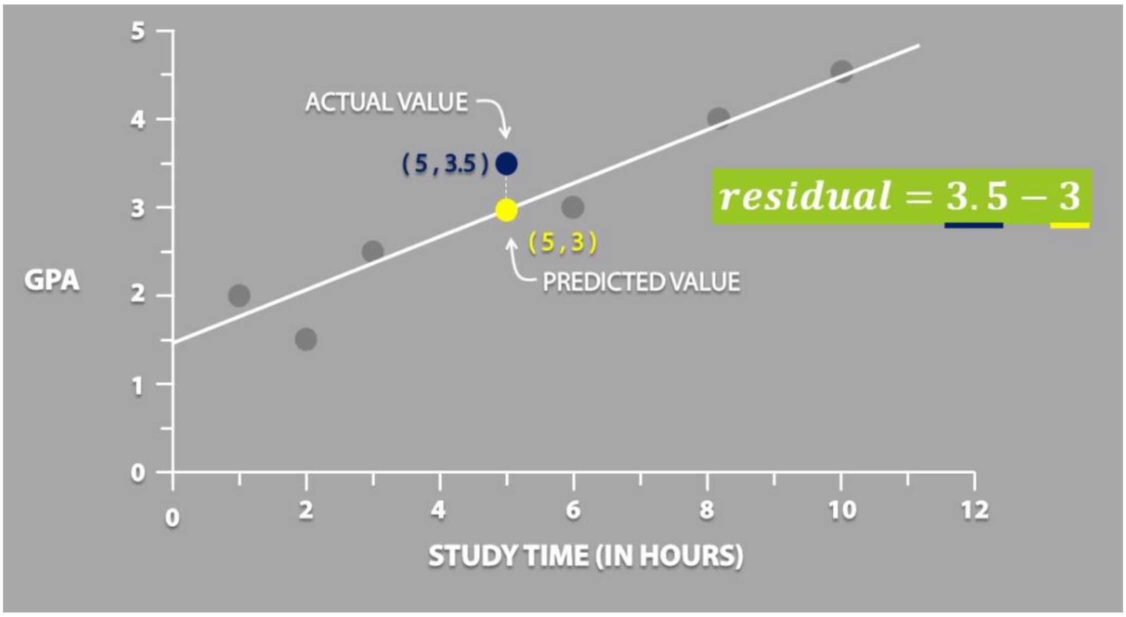
ACTUAL VALUE
(x,y)
(x,y)
PREDICTED VALUE
residual = $y_i - \hat{y}$
GPA
0
1
2
3
3.5
4
5
0
2
4
5
6
8
10
12
STUDY TIME (IN HOURS)


5
4
3
2
1
0
GPA
ACTUAL VALUE
(5, 3.5)
(x, y)
PREDICTED VALUE
$residual = y_i - \hat{y}$
0
2
4
6
8
10
12
STUDY TIME (IN HOURS)

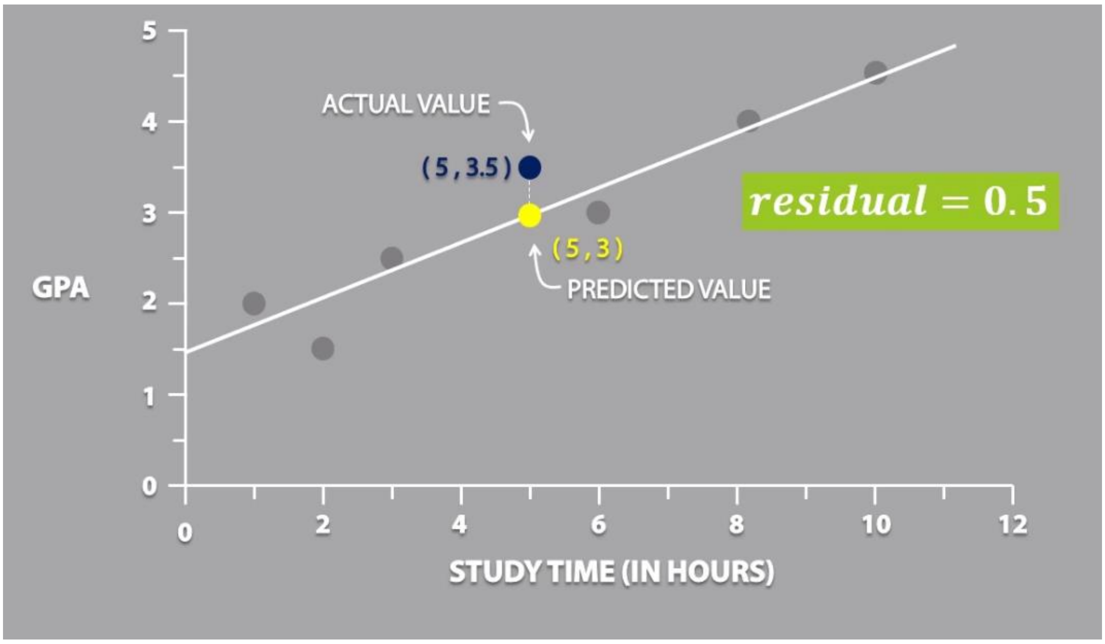

5
4
3
2
1
0
GPA
ACTUAL VALUE
(5, 3.5)
(x, y)
PREDICTED VALUE
$residual = y_i - \hat{y}$
0
2
4
5
6
8
10
12
STUDY TIME (IN HOURS)


ACTUAL VALUE
(5, 3.5)
(5, 3)
PREDICTED VALUE
$residual = y_i - \hat{y}$
GPA
STUDY TIME (IN HOURS)

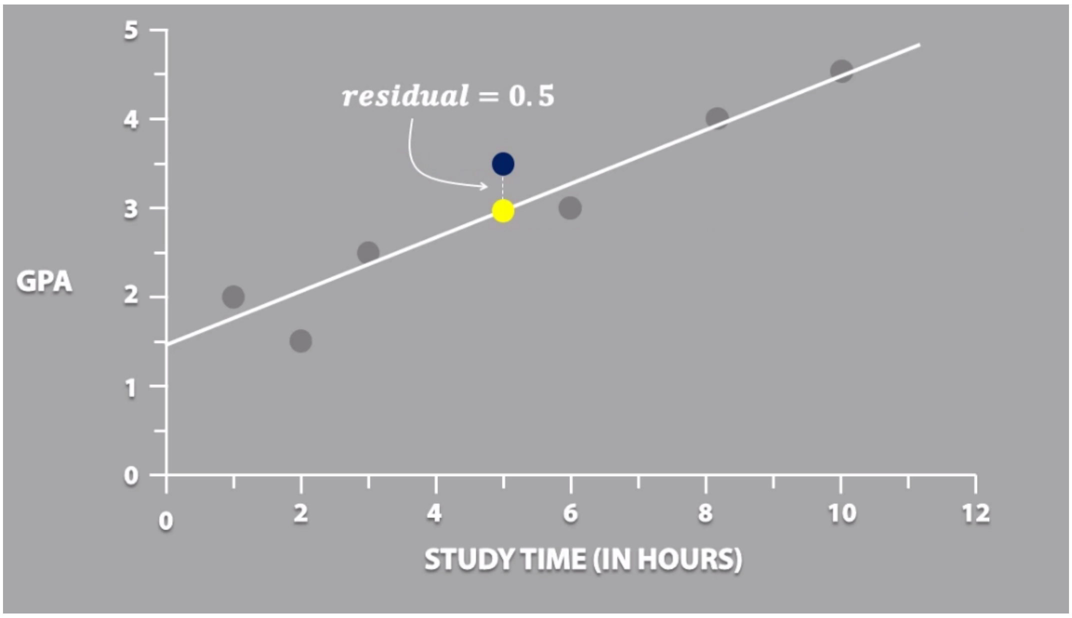

ACTUAL VALUE
(5, 3.5)
(5, 3)
PREDICTED VALUE
$residual = 3.5 - \hat{y}$
GPA
STUDY TIME (IN HOURS)



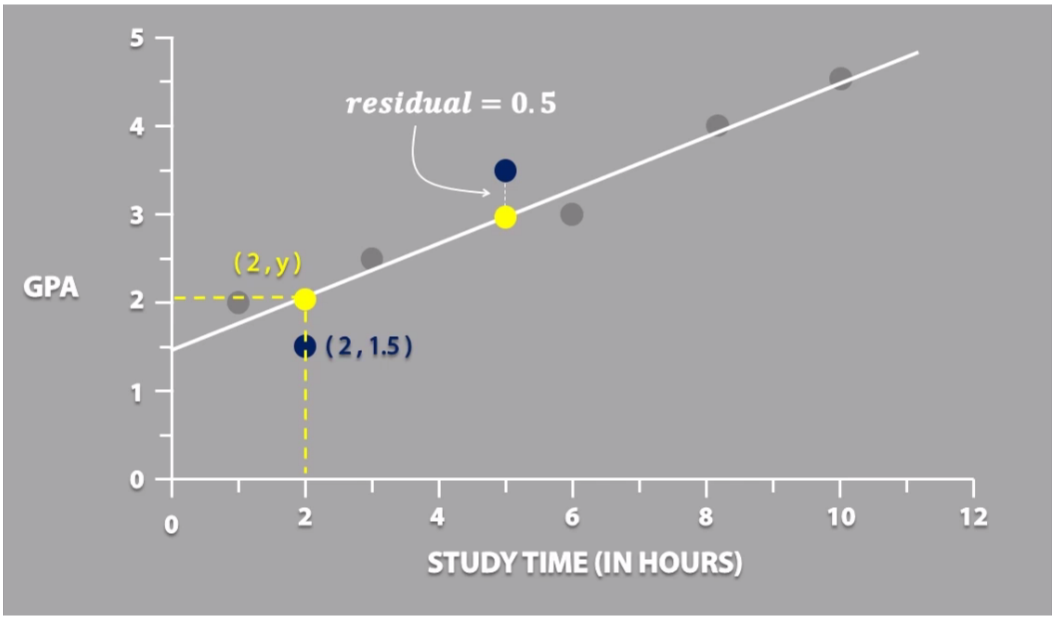
GPA
5
4
3
2
1
0
ACTUAL VALUE
(5,3.5)
(5,3)
PREDICTED VALUE
residual = 3.5 − 3
0
2
4
6
8
10
12
STUDY TIME (IN HOURS)



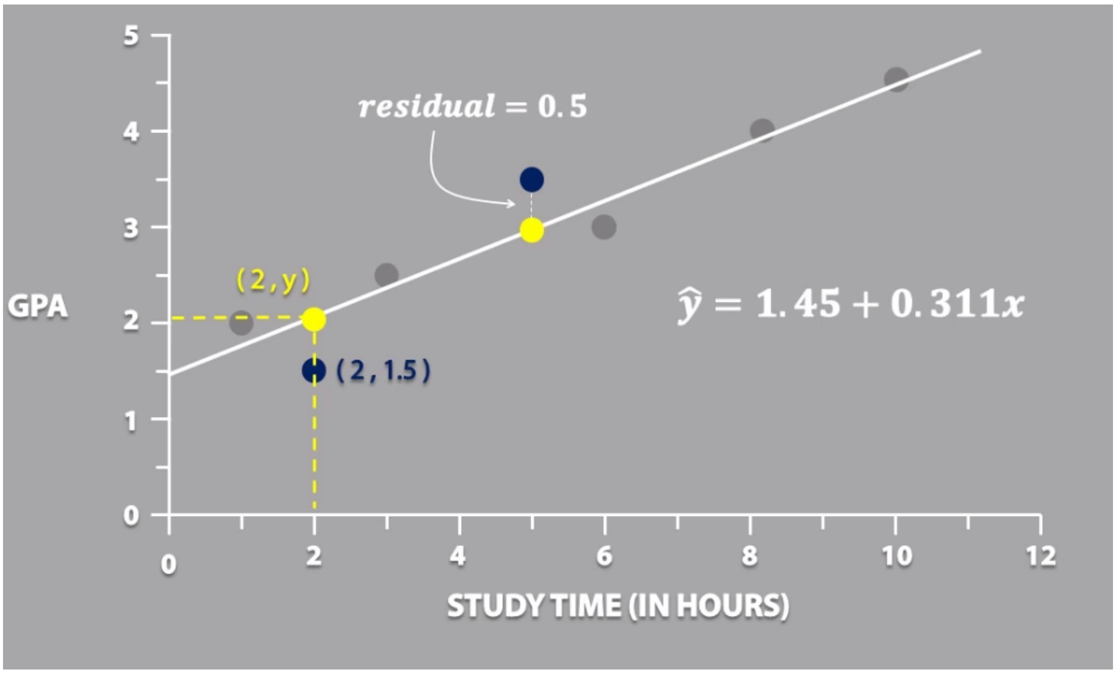
GPA
5
4
3
2
1
0
ACTUAL VALUE
(5,3.5)
(5,3)
PREDICTED VALUE
residual = 0.5
0
2
4
6
8
10
12
STUDY TIME (IN HOURS)



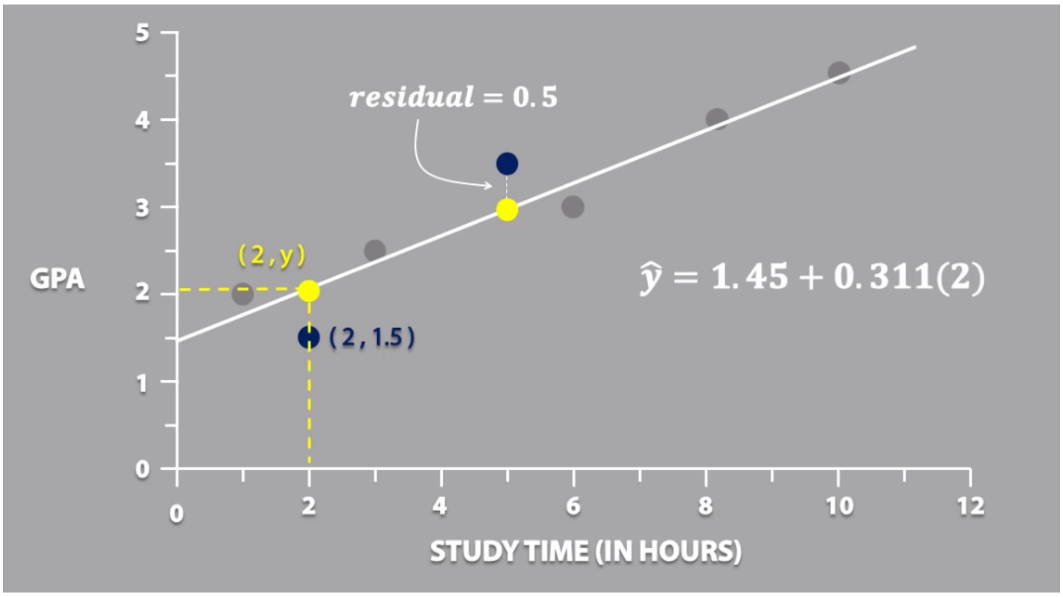
$residual = 0.5$
GPA
5
4
3
2
1
0
0
2
4
6
8
10
12
STUDY TIME (IN HOURS)



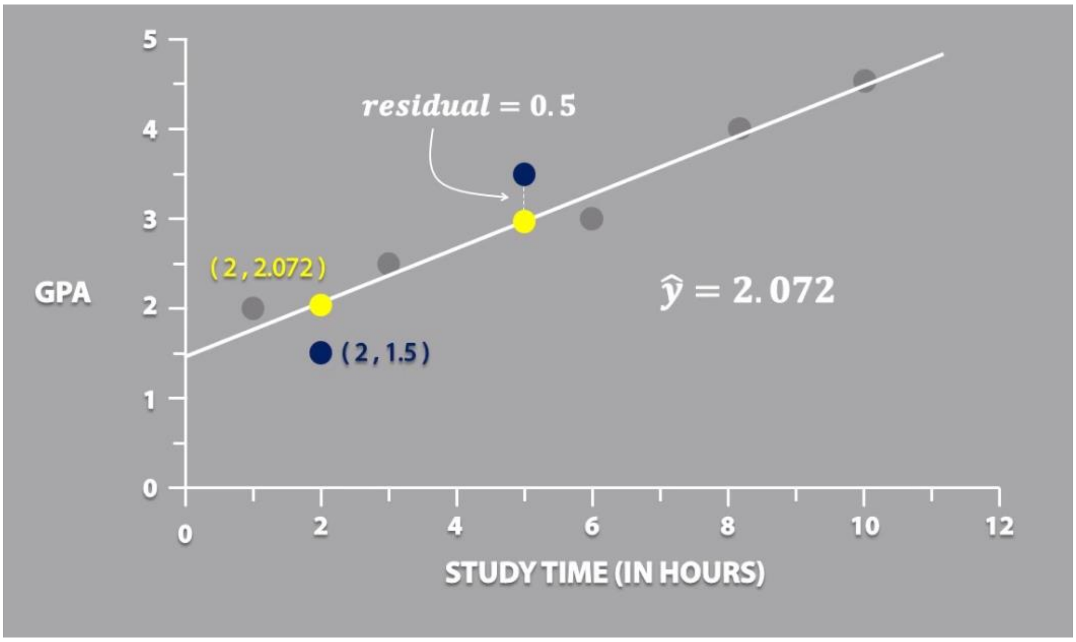
$residual = 0.5$
(2, y)
(2, 1.5)
GPA
5
4
3
2
1
0
0
2
4
6
8
10
12
STUDY TIME (IN HOURS)

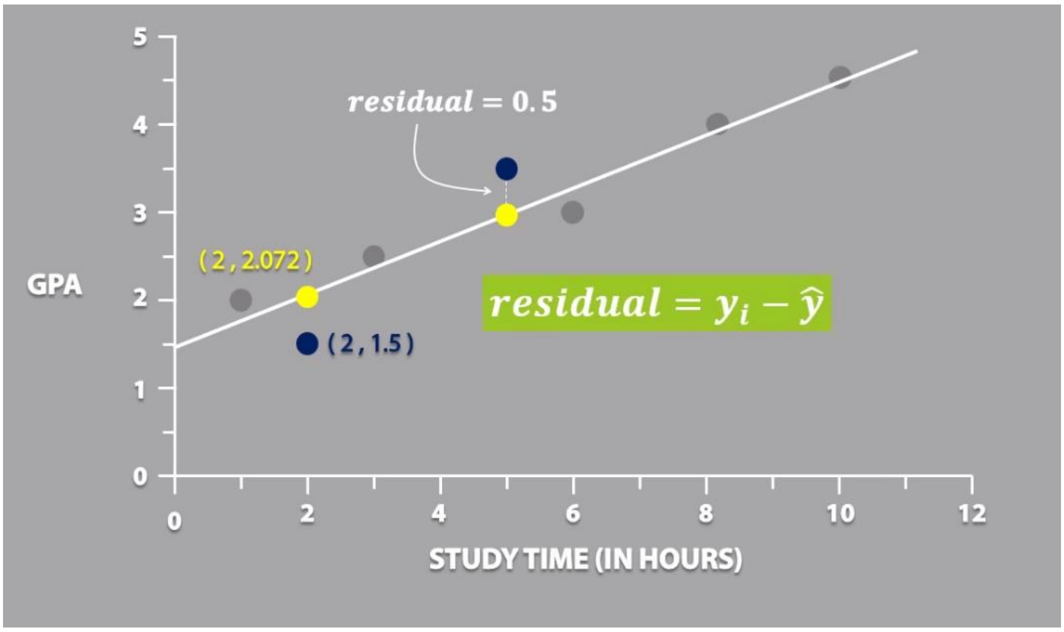


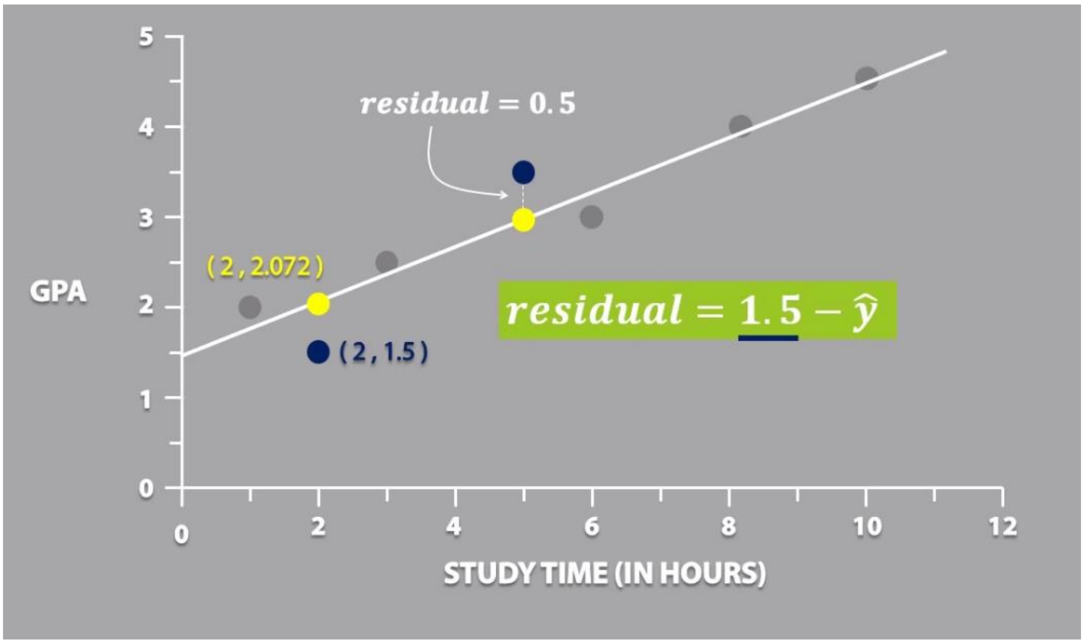
5
4
3
2
1
0
GPA
residual = 0.5
(2,y)
(2,1.5)
$\hat{y} = 1.45 + 0.311x$
0
2
4
6
8
10
12
STUDY TIME (IN HOURS)



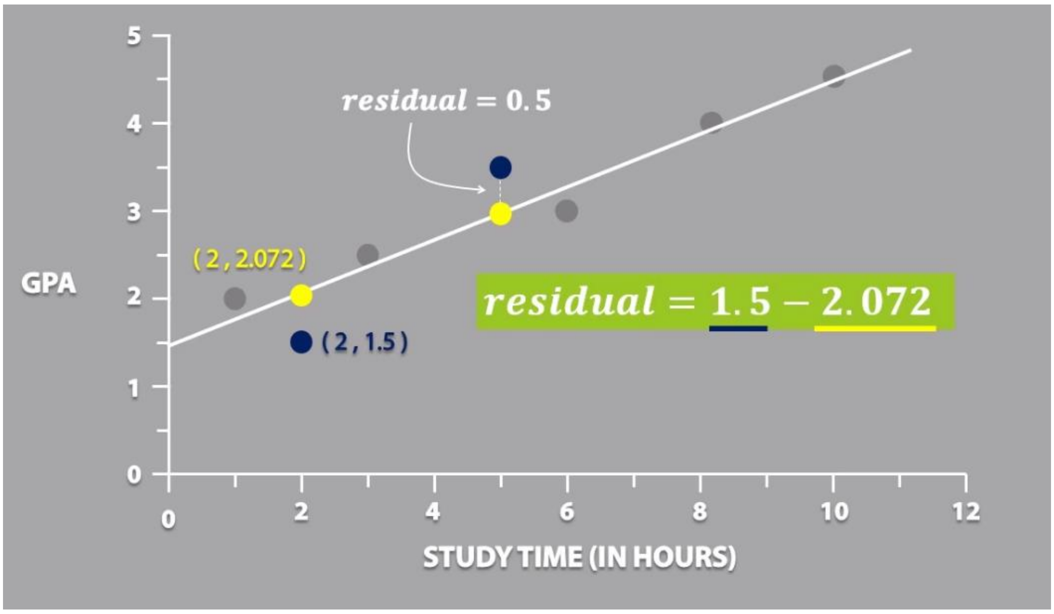
5
4
3
2
1
0
GPA
residual = 0.5
(2,y)
(2,1.5)
$\hat{y} = 1.45 + 0.311(2)$
0
2
4
6
8
10
12
STUDY TIME (IN HOURS)


residual = 0.5
GPA
(2, 2.072)
(2, 1.5)
ŷ = 2.072
STUDY TIME (IN HOURS)
0
1
2
3
4
5
6
8
10
12

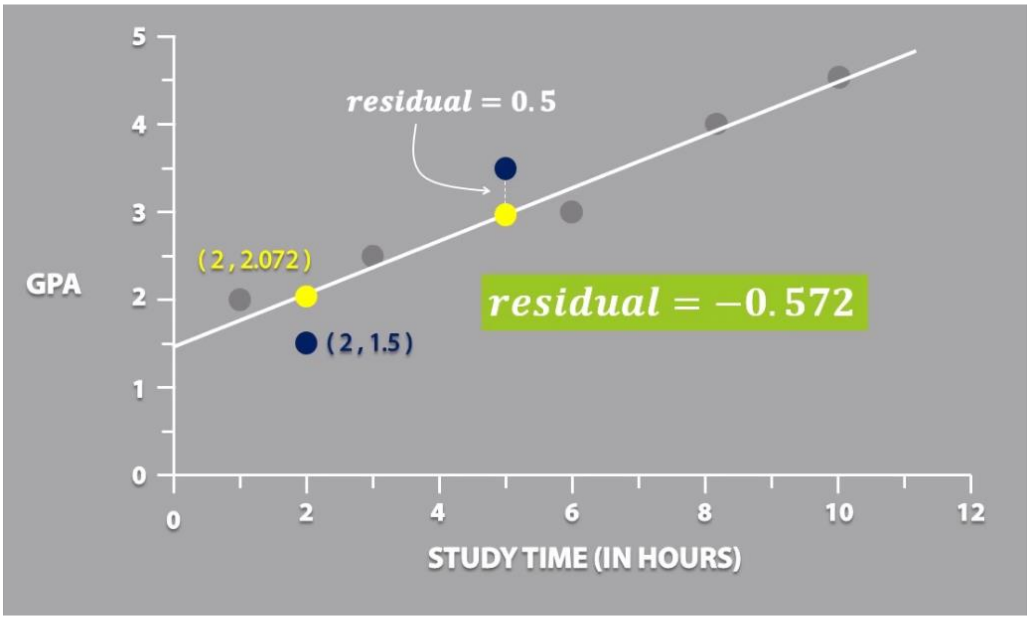

residual = 0.5
GPA
(2, 2.072)
(2, 1.5)
residual = y_i − ŷ
STUDY TIME (IN HOURS)
0
1
2
3
4
5
6
8
10
12



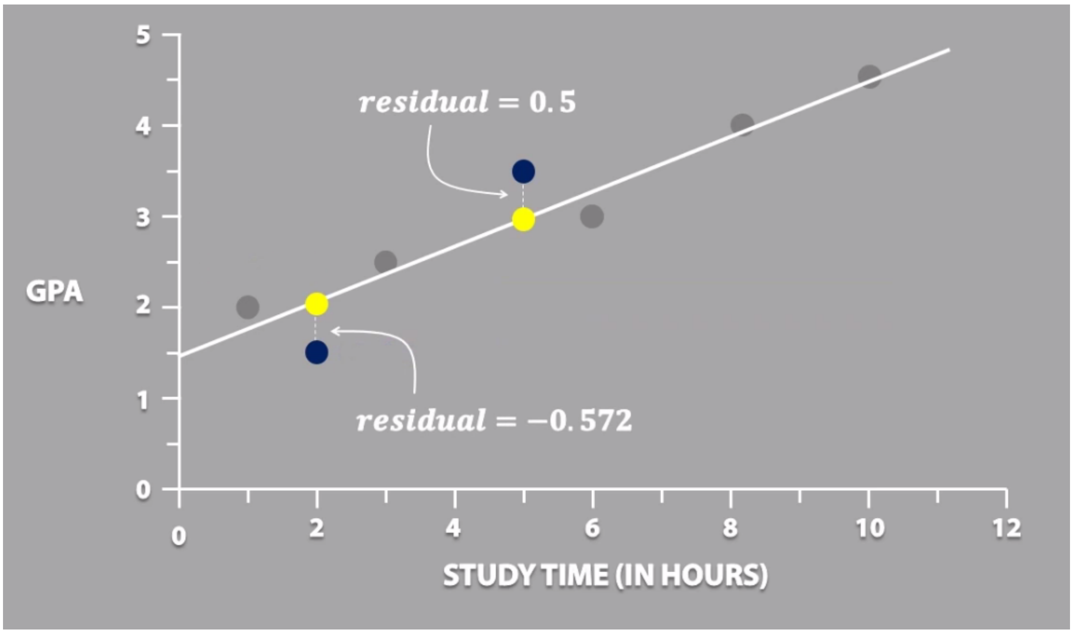
$residual = 0.5$
(2, 2.072)
(2, 1.5)
$residual = 1.5 - \hat{y}$
GPA
STUDY TIME (IN HOURS)



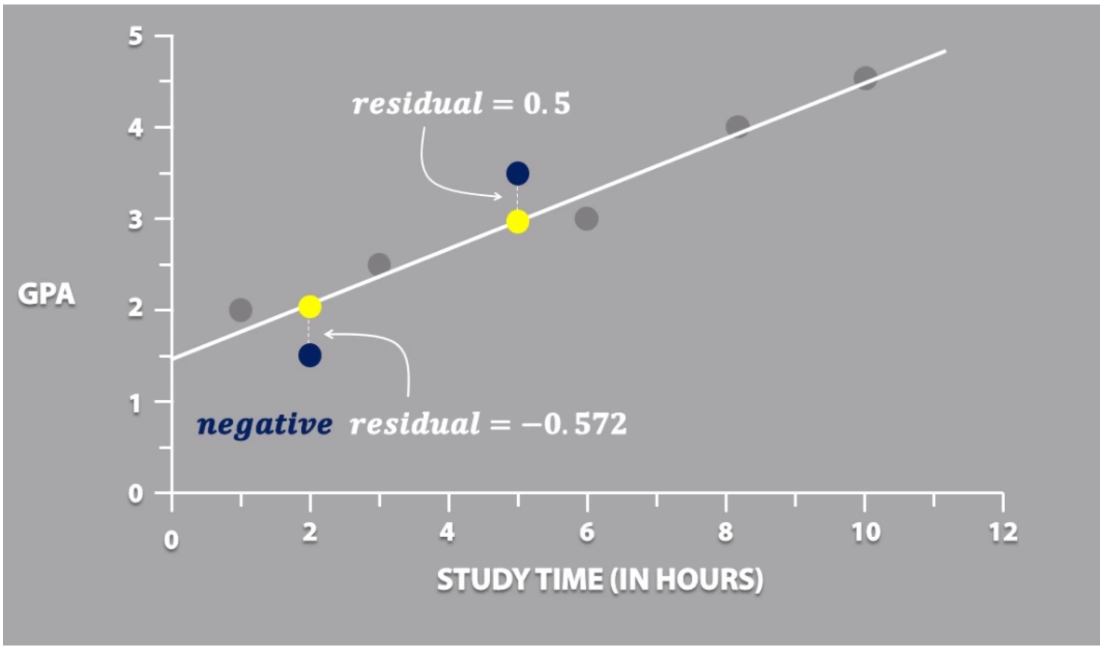
$residual = 0.5$
(2, 2.072)
(2, 1.5)
$residual = 1.5 - 2.072$
GPA
STUDY TIME (IN HOURS)



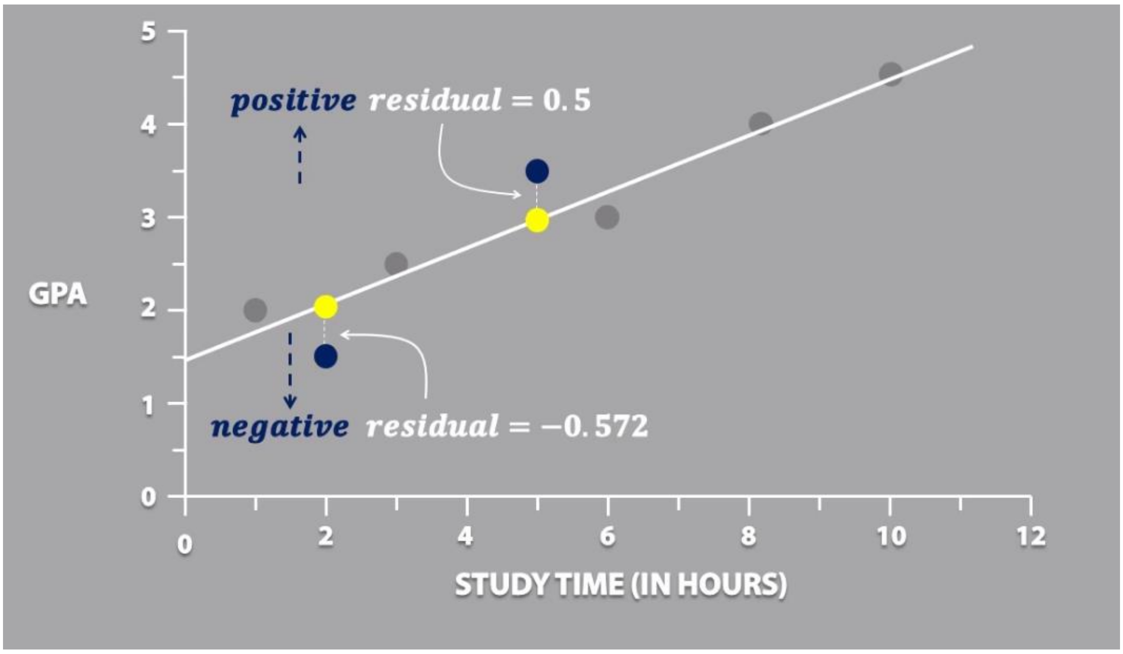
residual = 0.5
(2, 2.072)
(2, 1.5)
residual = −0.572
GPA
STUDY TIME (IN HOURS)



residual = 0.5
residual = −0.572
GPA
STUDY TIME (IN HOURS)


GPA
STUDY TIME (IN HOURS)
residual = 0.5
negative residual = −0.572
0
1
2
3
4
5
0
2
4
6
8
10
12

- ✓ A negative residual value means, it falls the below the regression line
- ✓ A positive residual value means, the actual value located into above regression line



10. Maths - Matrix

Contents

10. Maths - Matrix

1. Matrix

✓ A Matrix is an array of numbers:

$$\begin{bmatrix} 6 & 4 & 24 \\ 1 & -9 & 8 \end{bmatrix}$$

A Matrix

(This one has 2 Rows and 3 Columns)

2. Rows and columns

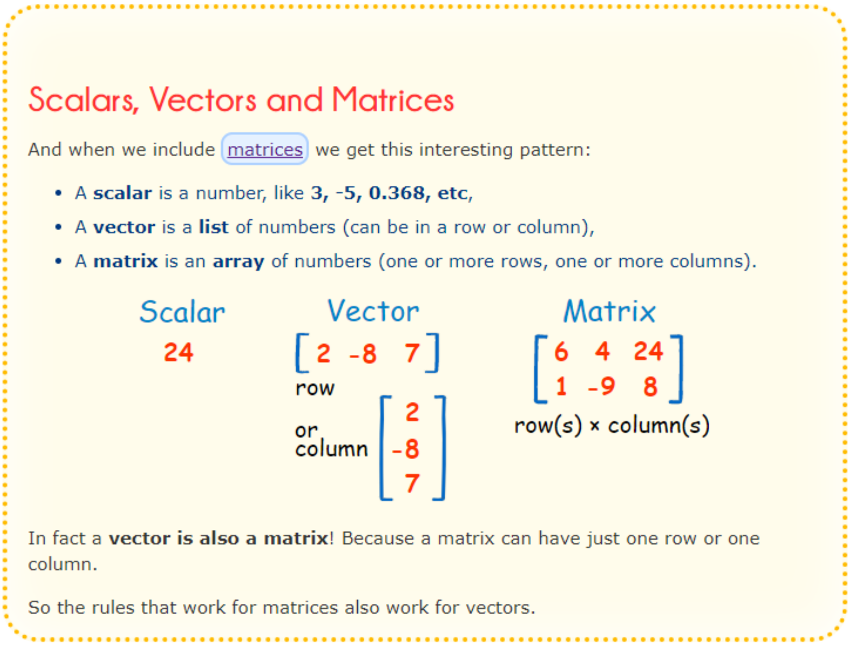
Rows and Columns

To show how many rows and columns a matrix has we often write **rows×columns**.

Example: This matrix is **2×3** (2 rows by 3 columns):

$$\begin{bmatrix} 6 & 4 & 24 \\ 1 & -9 & 8 \end{bmatrix}$$

3. Scalar, Vector and Matrix



Scalars, Vectors and Matrices

And when we include matrices we get this interesting pattern:

- A **scalar** is a number, like **3, -5, 0.368, etc**,
- A **vector** is a **list** of numbers (can be in a row or column),
- A **matrix** is an **array** of numbers (one or more rows, one or more columns).

In fact a **vector is also a matrix**! Because a matrix can have just one row or one column.

So the rules that work for matrices also work for vectors.

4. Adding matrix

- ✓ We can add two matrices
- ✓ The two matrices must be the same size, i.e. the rows must match in size, and the columns must match in size.
- ✓ Example: a matrix with 3 rows and 5 columns can be added to another matrix of 3 rows and 5 columns.
- ✓ But it could not be added to a matrix with 3 rows and 4 columns (the columns don't match in size)

3+4=7

$$\begin{bmatrix} 3 & 8 \\ 4 & 6 \end{bmatrix} + \begin{bmatrix} 4 & 0 \\ 1 & -9 \end{bmatrix} = \begin{bmatrix} 7 & 8 \\ 5 & -3 \end{bmatrix}$$

These are the calculations:

3+4=7	8+0=8
4+1=5	6−9=−3

5. Negative matrix

✓ The negative of a matrix is also simple:

$$-(2)=-2$$

$$-\begin{bmatrix} 2 & -4 \\ 7 & 10 \end{bmatrix} = \begin{bmatrix} -2 & 4 \\ -7 & -10 \end{bmatrix}$$

These are the calculations:

$-(2)=-2$	$-(-4)=+4$
$-(7)=-7$	$-(10)=-10$

6. Subtracting matrix

- ✓ To subtract two matrices: subtract the numbers in the matching positions:
- ✓ Note: subtracting is actually defined as the addition of a negative matrix: A + (−B)

3-4=-1

$$\begin{bmatrix} 3 & 8 \\ 4 & 6 \end{bmatrix} - \begin{bmatrix} 4 & 0 \\ 1 & -9 \end{bmatrix} = \begin{bmatrix} -1 & 8 \\ 3 & 15 \end{bmatrix}$$

These are the calculations:

3−4=−1	8−0=8
4−1=3	6−(−9)=15

7. Multiply by a Constant

- ✓ We can multiply a matrix by a constant *(the value 2 in this case)*:
- ✓ We call the constant a scalar, so officially this is called "scalar multiplication".

2x4=8

$$2 \times \begin{bmatrix} 4 & 0 \\ 1 & -9 \end{bmatrix} = \begin{bmatrix} 8 & 0 \\ 2 & -18 \end{bmatrix}$$

These are the calculations:

2×4=8	2×0=0
2×1=2	2×−9=−18

8. Multiply Matrices

✓ Multiplying a matrix by another matrix we need to do the "dot product" of rows and columns

To work out the answer for the **1st row** and **1st column**:

"Dot Product"

$$\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{bmatrix} \times \begin{bmatrix}7 & 8\\ 9 & 10\\ 11 & 12\end{bmatrix} = \begin{bmatrix}58 & \\ & \end{bmatrix}$$

The "Dot Product" is where we **multiply matching members**, then sum up:

$$(1, 2, 3) \bullet (7, 9, 11) = 1\times7 + 2\times9 + 3\times11$$
$$= 58$$

We match the 1st members (1 and 7), multiply them, likewise for the 2nd members (2 and 9) and the 3rd members (3 and 11), and finally sum them up.

Want to see another example? Here it is for the 1st row and **2nd column**:

$$\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{bmatrix} \times \begin{bmatrix}7 & 8\\ 9 & 10\\ 11 & 12\end{bmatrix} = \begin{bmatrix}58 & 64\\ & \end{bmatrix}$$

$$(1, 2, 3) \bullet (8, 10, 12) = 1\times8 + 2\times10 + 3\times12$$
$$= 64$$

We can do the same thing for the **2nd row** and **1st column**:

$$(4, 5, 6) \bullet (7, 9, 11) = 4\times7 + 5\times9 + 6\times11$$
$$= 139$$

And for the **2nd row** and **2nd column**:

$$(4, 5, 6) \bullet (8, 10, 12) = 4\times8 + 5\times10 + 6\times12$$
$$= 154$$

And we get:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \times \begin{bmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{bmatrix} = \begin{bmatrix} 58 & 64 \\ 139 & 154 \end{bmatrix}$$

DONE!

9. A real time example

Example: The local shop sells 3 types of pies.

- Apple pies cost **$3** each
- Cherry pies cost **$4** each
- Blueberry pies cost **$2** each

And this is how many they sold in 4 days:

	Mon	Tue	Wed	Thu
Apple	13	9	7	15
Cherry	8	7	4	6
Blueberry	6	4	0	3

Now think about this ... the **value of sales** for Monday is calculated this way:

Apple pie value + Cherry pie value + Blueberry pie value

$3×13 + $4×8 + $2×6 = $83

So it is, in fact, the "dot product" of prices and how many were sold:

($3, $4, $2) • (13, 8, 6) = $3×13 + $4×8 + $2×6
= $83

We **match** the price to how many sold, **multiply** each, then **sum** the result.

In other words:

- The sales for Monday were: Apple pies: **\$3×13=\$39**, Cherry pies: **\$4×8=\$32**, and Blueberry pies: **\$2×6=\$12**. Together that is \$39 + \$32 + \$12 = **\$83**
- And for Tuesday: **\$3×9 + \$4×7 + \$2×4** = **\$63**
- And for Wednesday: **\$3×7 + \$4×4 + \$2×0** = **\$37**
- And for Thursday: **\$3×15 + \$4×6 + \$2×3** = **\$75**

So it is important to match each price to each quantity.

Now you know why we use the "dot product".

And here is the full result in Matrix form:

$$\begin{bmatrix} \$3 & \$4 & \$2 \end{bmatrix} \times \begin{bmatrix} 13 & 9 & 7 & 15 \\ 8 & 7 & 4 & 6 \\ 6 & 4 & 0 & 3 \end{bmatrix} = \begin{bmatrix} \$83 & \$63 & \$37 & \$75 \end{bmatrix}$$

\$3x13 + \$4x8 + \$2x6

They sold **\$83** worth of pies on Monday, **\$63** on Tuesday, etc.

Examples

In General:

To multiply an **m×n** matrix by an **n×p** matrix, the **n**s must be the same, and the result is an **m×p** matrix.

$$m \times n \;\; \times \;\; n \times p \;\; \rightarrow \;\; m \times p$$

So ... multiplying a **1×3** by a **3×1** gets a **1×1** result:

$$\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 1\times4+2\times5+3\times6 \end{bmatrix} = \begin{bmatrix} 32 \end{bmatrix}$$

But multiplying a **3×1** by a **1×3** gets a **3×3** result:

$$\begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} = \begin{bmatrix} 4\times1 & 4\times2 & 4\times3 \\ 5\times1 & 5\times2 & 5\times3 \\ 6\times1 & 6\times2 & 6\times3 \end{bmatrix} = \begin{bmatrix} 4 & 8 & 12 \\ 5 & 10 & 15 \\ 6 & 12 & 18 \end{bmatrix}$$

10. Types of Matrix

A Matrix is an array of numbers:

$$\begin{bmatrix} 6 & 4 & 24 \\ 1 & -9 & 8 \end{bmatrix}$$

A Matrix

(This one has 2 Rows and 3 Columns)

The **Main Diagonal** starts at the top left and goes down to the right:

$$\begin{bmatrix} 7 & 6 & 4 \\ 4 & 2 & -2 \\ 3 & 0 & 9 \end{bmatrix}$$

11. Transpose matrix

- ✓ To "transpose" a matrix, swap the rows and columns.
- ✓ We put a "T" in the top right-hand corner to mean transpose:

$$\begin{bmatrix} 6 & 4 & 24 \\ 1 & -9 & 8 \end{bmatrix}^T = \begin{bmatrix} 6 & 1 \\ 4 & -9 \\ 24 & 8 \end{bmatrix}$$

Square

A **square** matrix has the same number of rows as columns.

$$\begin{bmatrix} 2 & 0 \\ 1 & 8 \end{bmatrix}$$

A square matrix (2 rows, 2 columns)

$$\begin{bmatrix} 6 & 4 & 24 \\ 1 & -9 & 8 \\ 3 & 0 & 7 \end{bmatrix}$$

Also a square matrix (3 rows, 3 columns)

Identity Matrix

An **Identity Matrix** has **1**s on the main diagonal and **0**s everywhere else:

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

A 3×3 Identity Matrix

- It is square (same number of rows as columns)
- It can be large or small (2×2, 100×100, ... whatever)
- Its symbol is the capital letter **I**

It is the matrix equivalent of the number "1", when we multiply with it the original is unchanged:

$$A \times I = A$$

$$I \times A = A$$

Diagonal Matrix

A diagonal matrix has zero anywhere not on the main diagonal:

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

A diagonal matrix

Scalar Matrix

A scalar matrix has all main diagonal entries the same, with zero everywhere else:

$$\begin{bmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

A scalar matrix

Zero Matrix (Null Matrix)

Zeros just everywhere:

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Zero matrix