# Statistics and Probability

Quarter 4 (week 5)

# Learning Competencies

- calculates the Pearson's sample correlation coefficient
- solves problems involving correlation analysis
- identifies the independent and dependent variables
- calculates the slope and y-intercept of the regression line
- interprets the calculated slope and y-intercept of the regression line

# Calculates the Pearson’s sample correlation coefficient

Module 18

The **Pearson's sample correlation coefficient** (also known as Pearson r ), denoted by **r**, is a test statistic that measures the strength of the linear relationship between two variables. To find r, the following formula is used:

The **correlation coefficient (r)** is a number between -1 and 1 that describes both the strength and the direction of correlation. In symbol, we write **$-1 \leq r \leq 1$.**

**Example:**

Teachers of Pag-asa National High School instilled among their students the value of time management and excellence in everything they do. The table below shows the time in hours spent in studying **(X)** by six Grade 11 students and their scores in a test **(Y)**. Solve for the **Pearson's sample correlation coefficient r.**

| | | | | |
|---|---|---|---|---|
| | 2 | 3 | 4 | 5 |
| | 10 | 10 | 15 | 25 |

| X | Y | XY | $X^2$ | $Y^2$ |
|---|---|---|---|---|
| 1 | 5 | 5 | 1 | 25 |
| 2 | 10 | 20 | 4 | 100 |
| 3 | 10 | 30 | 9 | 100 |
| 4 | 15 | 60 | 16 | 225 |
| 5 | 25 | 125 | 25 | 625 |
| 6 | 30 | 180 | 36 | 900 |
| | | | | |

The value of r is a positive number. Therefore, we can say accurately that there is a positive correlation between hours spent in studying and their scores in a test.

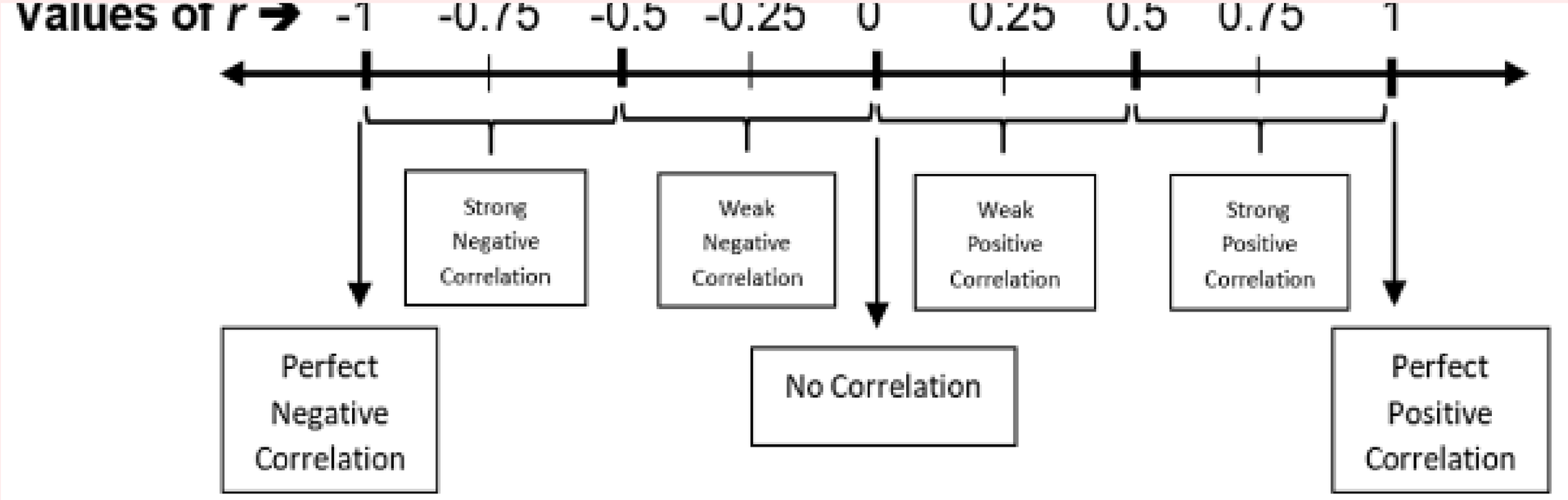
**Note: For consistency of our answer, round your final answer into two decimal places.**

# Solves problems involving correlation analysis

Module 19

**Correlation** is used to determine the existence, strength, and direction of the relationship between two variables. Correlation coefficient r is a number between -1 and 1 that describes both the strength and the direction of correlation. In symbol, we write **$-1 \leq r \leq 1$.**

You identified the trends, estimated the values of r, and based on the values, you chose the correct descriptions of the strength of the correlation. So now, we will interpret r value by looking at the scale that gives both strength and direction of correlation.



Using the correlation scale, we can determine the strength of the correlation coefficient r. For example, you have r = 0.63 which means that there is a “strong positive correlation” between the two variables. To interpret, we can simply state it this way: “As x values increase, y values also increase and vice versa.”

In interpreting the linear relationship of two bivariate data, refer to the value of r and the scale presented above. We can state our interpretation in different ways. In order for you to solve problems involving correlation analysis, you must know how to calculate the value of r and interpret this value using the scale. Since computing for r value is a necessary skill, you may go back to the previous lesson if you feel that you haven’t mastered it yet. Otherwise, proceed to the following examples of solving for r.

**Scenario:** Filipino employees are known for being persistent and hardworking. That is why they truly value every single cent of their salary. Here are some situations showing the relationship between the salary and spending of a Filipino employee.

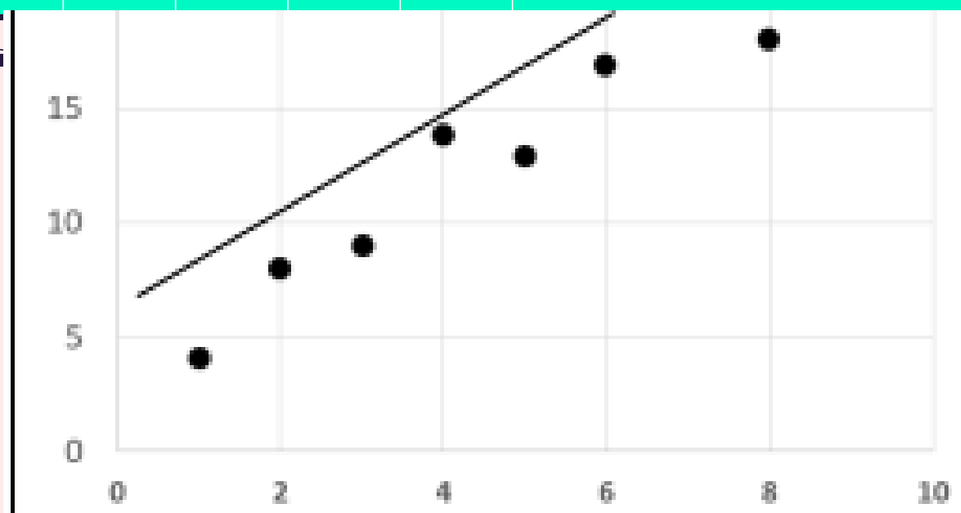
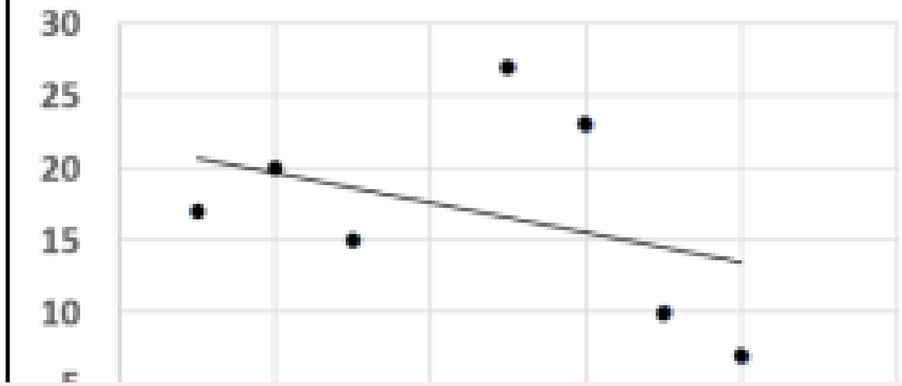
**Situation 1:** There is a survey wherein the correlation coefficient r between salary and spending of employee was found to be 0.97.

**Interpretation:** There is a “strong positive correlation” between salary and spending of employees.

**Situation 2**: In another survey, the correlation coefficient r between salary and spending of employee was found to be -0.81.

**Interpretation:** There is a “strong negative correlation” between salary and spending of employees.

| Bivariate Data | Computed Pearson's *r* | Interpretation |
|---|---|---|
| Temperature and the number of hot chocolate products sold | -0.781 | There is a strong negative correlation between the temperature and the number of hot chocolate products sold. |
| Amount of coffee intake and number of hours you stay awake | 0.426 | There is a weak positive correlation between the amount of coffee intake and number of hours you stay awake. |
| Height and salary of | 0 | There is no correlation between the |

| | | |
|---|---|---|
|  | Strong Positive Correlation | be in the range between 0.5 and 1.<br><br>We can say 0.8 or 0.75 as long as it is within the range in the correlation scale. |
|  | Weak Negative Correlation | The value of *r* should be in the range between 0 and -0.5 |

The **closeness of the points** around the trend line determines the **strength** of the correlation. The **closer the points to the trend line, the stronger the correlation of the variables is.**

# Identifies the independent and dependent variables

Module 20

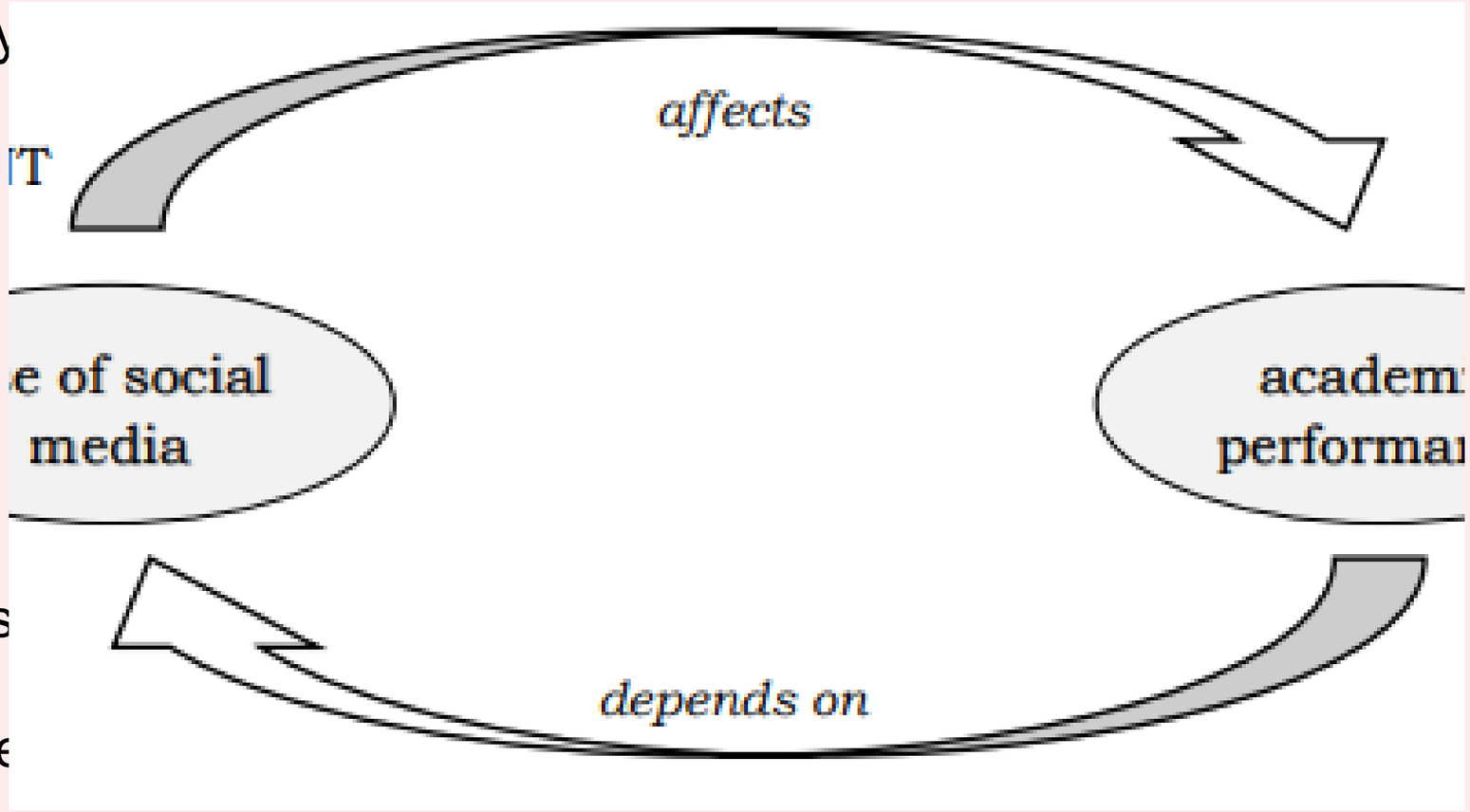
# What is the difference between the two variables?

**Dependent variable** depends on other variables or factors. It is something that is influenced and affected. It is also associated with the word effect or outcome.

**Independent variable** affects the dependent variable. It is something you have control over, one which you can choose and manipulate. However, in some cases, you may not be able to manipulate the independent variable. It is commonly known as the cause or the reason behind changes.

For example, the researcher wants to determine the effects of the use of social media on the academic performance of students in Mathematics.

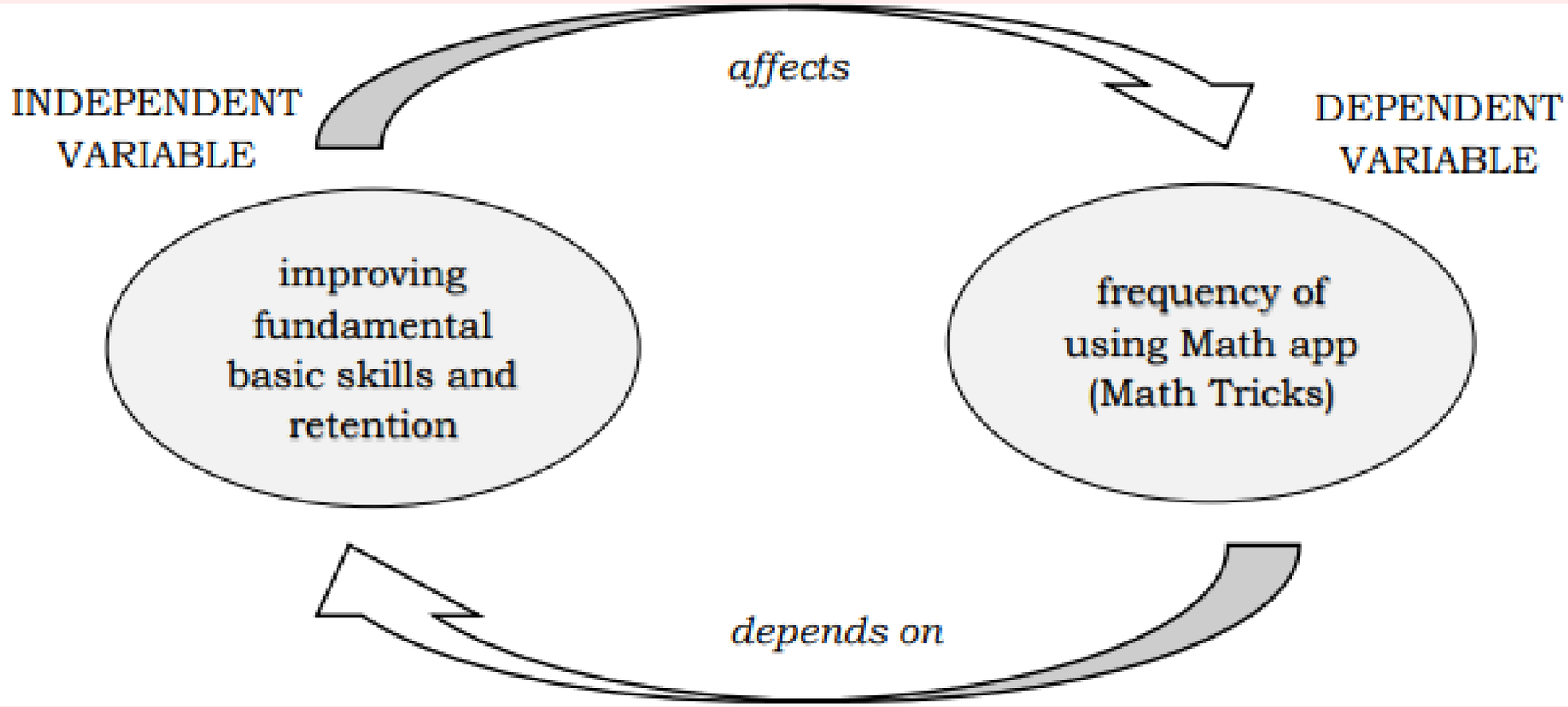
The bivariate data in the study are the use of social media and academic performance. Academic performance depends on the use of social media, or we can say that academic performance is affected by the use of social media. Therefore, **independent variable** here is the use of social media and the **dependent variable** is the academic performance.




Independent variables happen first and y (dependent variable) is the result of x (independent variable).

# Research Title: Math App (Math Tricks) for Improving Fundamental Basic Skills and Retention of Grade 11 Students of Mapalad Integrated High School

Increasing or improving your fundamental basic skills and retention depends on how you frequently practice or use the Math app called Math Tricks.



# Calculates the slope and y-intercept of the regression line

Module 21

Unequal slopes and y-intercepts are solved because two different pairs of points are used. Since the slopes and y-intercepts are unequal, two different equations are obtained using the algebraic method.

Whenever there are more than two points of data, it is usually impossible to find one line that passes through all points. However, a best-fit line that is a good approximation of the data can usually be found. This best-fit straight line used to predict the value of y for a given value of x is called the regression line.

For statistics, there is a simpler way to find the equation of the best-fit line. The equation of the best-fit line is also called **equation of the regression line** or simply regression equation.

e equation of the regression line is $\hat{y} = bx + a$
where: $b$ = slope of the regression line
$a$ = y-intercept of the regression line

Examples: a. = $2x + 3$ $b = 2$; $a = 3$; The slope is 2 and y-intercept is 3.
b. = $x$ 2 − 1 $b$ = 1 2 ; $a = -1$; The slope is 1 2 and y-intercept is -1.

The slope (b) can be calculated using the following formula:

The y-intercept (a) can be calculated using the following formula:

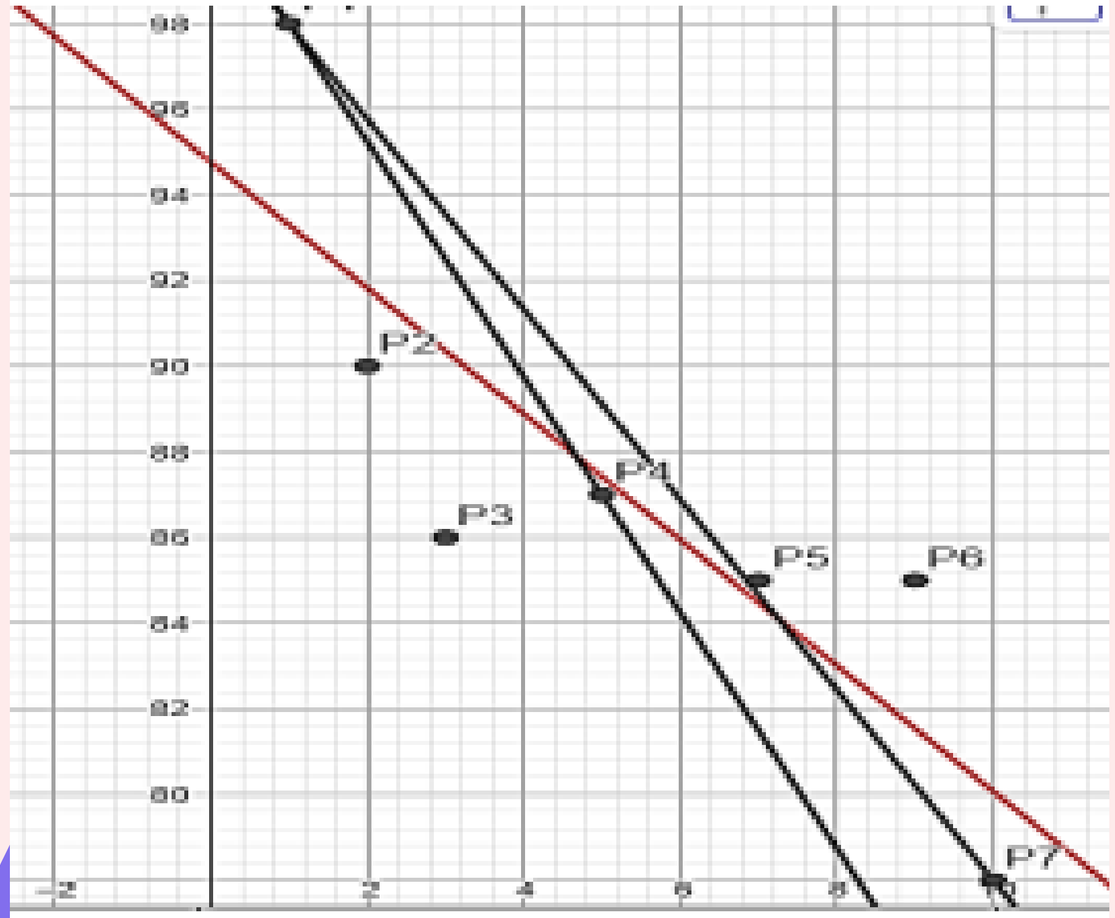
| | Number of Absences (x) | General Average (y) | xy | $x^2$ |
|---|---|---|---|---|
| 1 | 1 | 98 | 98 | 1 |
| 2 | 2 | 90 | 180 | 4 |
| 3 | 3 | 86 | 258 | 9 |
| 4 | 5 | 87 | 435 | 25 |
| 5 | 7 | 85 | 595 | 49 |
| 6 | 9 | 85 | 765 | 81 |
| 7 | 10 | 78 | 780 | 100 |
| | | | | |

94.77

This is the y-intercept.

This is the slope

## Best-Fit Line




This graph shows the number of absences (x) and general average (y). The two computed lines in the previous activity are shown here. This also shows the best-fit line obtained from calculating the slope and y- intercept. By simply looking at the graphs of the lines, you can see that only one line will be best fit. That is the line with equation

$= -1.47x + 94.77$

# Interprets the calculated slope and y-intercept of the regression line

Module 22

As shown in the activity, you need to consider the slope of the regression equation to determine the relationship between variables x and y.

Recall: $\hat{y} = bx + a$

$\hat{y}$ = predicted values of $y$

$a$ as the y-intercept of the regression equation

It is the value of $y$ when $x$ is equal to 0.

$b$ as the slope of the regression equation

It is the estimated rate of change of $y$ per unit change in $x$

Based on the slope, you can determine the relationship between the variables and then, interpret its meaning related to the situation. The sign of the slope (b) is used to identify the trend or direction of relationship between the dependent and independent variable.

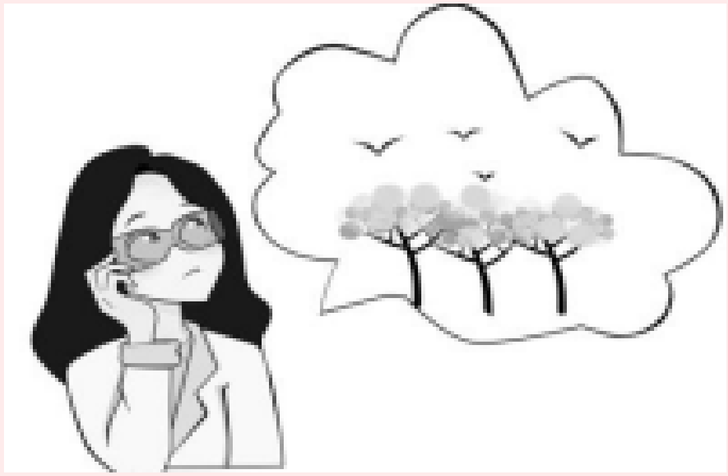
| Relationship | Interpretation |
|---|---|
| A positive slope means a positive relationship between variables *x and y*. | When x-values increase, y-values also increase. *OR* When x-values decrease, y-values also decrease. |
| A negative slope means a negative relationship between variables *x and y*. | When x-values decrease, y-values increase. *OR* When x-values increase, y-values decrease. |
| A zero slope means no relationship between variables *x and y*. | There is no relationship between x and y-variables. |

The y-intercept can also be interpreted to determine its relevance on a specific situation.

- Substitute the value of y into the equation.
- Analyze if the values of the dependent and independent variables depict a possible situation between the two variables.

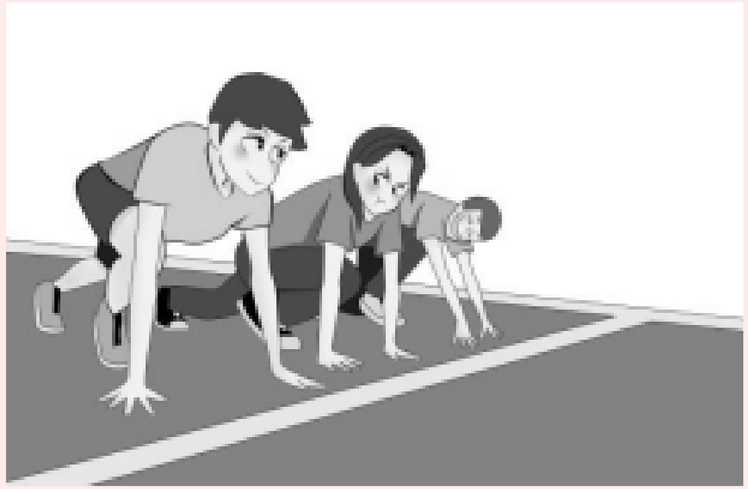
**Note:** Unlike the slope which is easily determined, analysis of the situation is needed to determine if the y-intercept is relevant with respect to the situation.

**Example 1:** A biologist wants to study the relationship between the number of trees per square meter (x) and the number of birds per square meter (y). She came up with the equation of the regression line:



$= 3x + 4$

| Slope and Y-Intercept | Interpretation |
|---|---|
| **The slope is 3. The positive slope indicates a positive relationship between x and y.** | For every additional tree, you can expect an average of 3 additional birds per square meter. |
| **The y-intercept is 4.** | It means that the average number of birds per square meter in an area with no (0) trees is 4. This value is relevant because the number of trees can be 0 and it is possible to have a bird in an area without trees. |



**Example 2:** A survey was done to know the relationship between the age in years of a young person (x) and the time in minutes a person can run one mile (y). Data from children between the ages of 8 and 15 were collected. The regression equation is $= -x + 15$

| Slope and Y-Intercept | Interpretation |
|---|---|
| **The slope is -1. The negative slope indicates a negative relationship between x and y.** | From the situation, as a young person ages 1 year, his run time decreases by 1 minute. |
| **The y-intercept is 15.** | This is not relevant because a zero-year-old child cannot run one mile. It is also stated in the problem that the domain is between 8 and 15. The age of a person cannot be 0. Thus, getting the y-intercept is a violation to the problem but not on the regression equation. |

# Thank You!