SHS

AIRs - LM in

Statistics and Probability

Quarter 4 – Week 7, Module15

Correlation Analysis

Statistics and Probability
Quarter 4 – Week 7 Module 15: Correlation Analysis
First Edition, 2021

Development Team of the Module

Writers: LORENA D. LACHICA, T-II
RAQUEL D. DE GUZMAN, T-II

Editor: SDO La Union, Learning Resource Quality Assurance Team

Management Team:

ATTY. Donato D. Balderas, Jr.
Schools Division Superintendent
Vivian Luz S. Pagatpatan, PHD
Assistant Schools Division Superintendent

German E. Flora, PHD, *CID Chief*

Virgilio C. Boado, PHD, *EPS in Charge of LRMS*

Erlinda M. Dela Peña, EDD, *EPS in Charge of Mathematics*
Michael Jason D. Morales, *PDO II*

Claire P. Toluyen, *Librarian II*

Target

In your previous lesson, you have learned about population proportion. You have learned how to compute and draw conclusion about the population proportion based on the test-statistic value and the rejection region and have solved problems involving them.

This module will provide you with information and activities that will help you learn about correlation analysis.

After going through this module, you are expected to:

1. illustrate the nature of bivariate data (M11/12SP-IVg-2);
2. construct a scatter plot (M11/12SP-IVg-3);
3. describe shape(form), trend(direction), and variation(strength) based on a scatter plot (M11/12SP-IVg-4
4. calculate the Pearson's sample correlation coefficient (M11/12SP-IVh-2) and;
5. solve problems involving correlation analysis (M11/12SP-IVh-3)

Subtask:

1. define bivariate data
2. give examples of bivariate data
3. describe a scatter plot

Before going on, check how much you know about this topic. Answer the pretest below in a separate sheet of paper.

Pretest

Directions: Write the letter of the correct answer on a separate sheet of paper

1. Which of the following graph is used to see the relationship between bivariate data?

A. Box-and –whisker plot B. Histogram
C. Normal curve D. Scatterplot

2. Which of the following bivariate data are positively correlated?

A. Grades and IQ
B. B. Income and saving
C. Number of absences and grades
D. Size of the family and expenses

3. Which of the following bivariate data are negatively correlated?

A. Religion and spiritual belief
B. Study time and grades
C. Speed of the car and time to reach a place
D. Tuition fee & school performance

4. Which of the following bivariate data are not correlated to each other?

A. Height & test score
B. Math grade & math anxiety
C. Socio-economic status & expenses
D. Weight & swimming speed

5. What kind of correlation is shown when all the points lie on a straight line?

A. Negative correlation B. No correlation
C. Perfect correlation D. Positive correlation

6. Which of the following bivariate data are considered quantitative?

A. Family size & water consumption
B. Educational attainment & salary
C. Number of male teachers & type of school
D. Weight & dancing ability

7. Which of the following bivariate data are considered qualitative?

A. Family size & water consumption
B. Number of male teachers & type of school
C. Number of absences & grades
D. Daily allowance & food expenses

8. In a regression analysis, which of the following variable being explained or predicted

A. Continuous variable
B. Dependent variable
C. Discrete variable
D. Independent variable

9. Which of the following lines approximate the general direction of the points in a scatter plot?

A. Best fitting lines
B. Horizontal lines
C. Trend lines
D. Vertical lines

10. It is used to measure the linear relationship between two variables that are normally distributed.

A. Pearson's Product-Moment correlation coefficient
B. Spearman's Rank-order correlation coefficient
C. Both A & B
D. Neither A nor B

11. Which of the following is used as a comparative measure of association of two ordinal variables?

A. Coefficient Correlation
B. Normal Distribution
C. Population Parameter
D. Sampling Distribution

12. If the two variables have a strong negative correlation the value of ***r*** is close to what value?

A. 0 B. 1 C. -1 D. 0.5

13. Which of the following represents a positive correlation?

A. Anxiety and test performance
B. Stress and job satisfaction
C. Summative Scores and Grades
D. percentage free/ reduced lunch and FCAT scores

14. It is used to find out if the computed ***r*** is significant or not.

A. T-test B. Z-test C. Pearson r D. Spearman rho

15. Pearson's ***r*** ranges in value from_____.

A. 0 to 1 B. -1 to 1 C. 0 to -1 D. 1 to 10

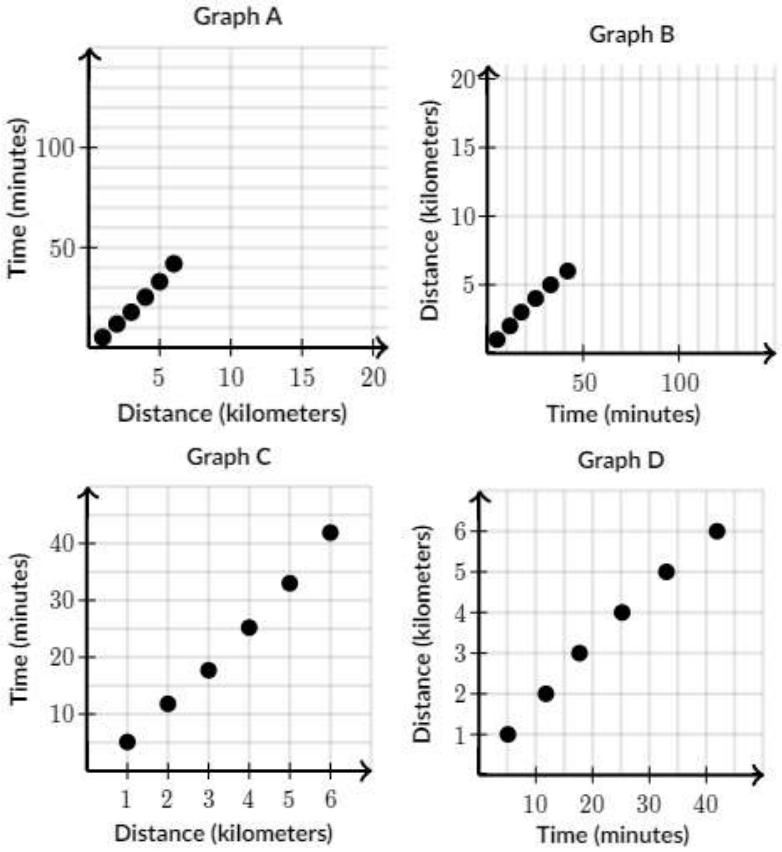
Jumpstart

Correlation analysis will help you understand how to determine if two variables are associated to each other and how strong is their relationship.

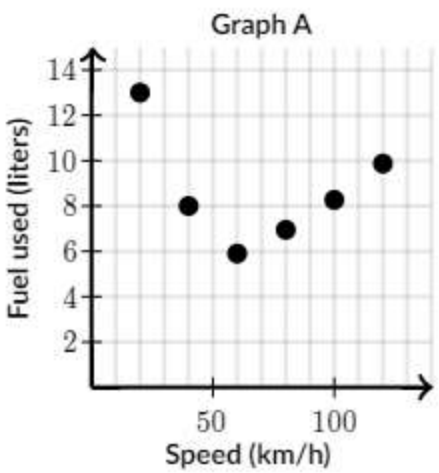
Activity 1. My Best!

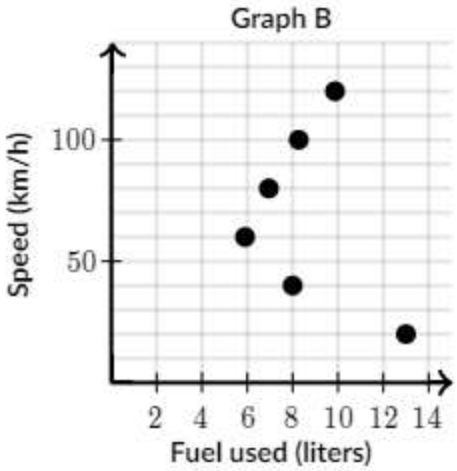
A. All of the scatterplots display the data correctly, but which of them displays the data best?

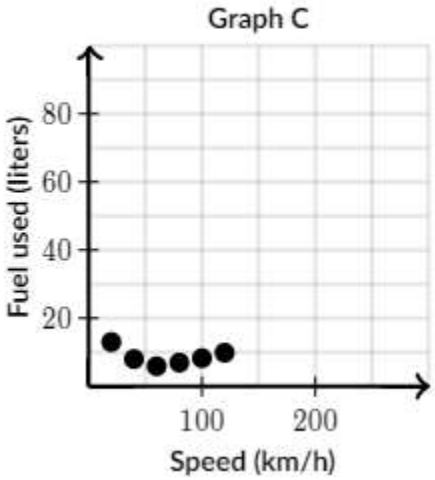
1.

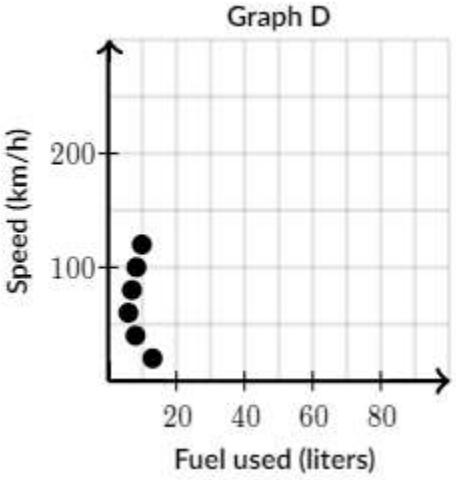


2.



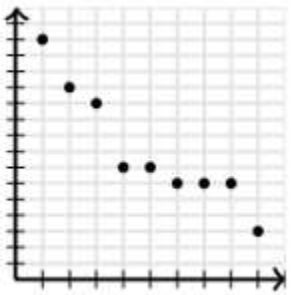


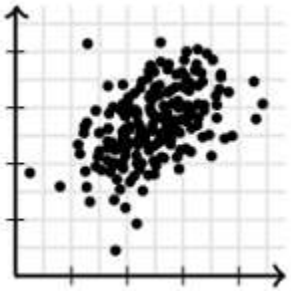


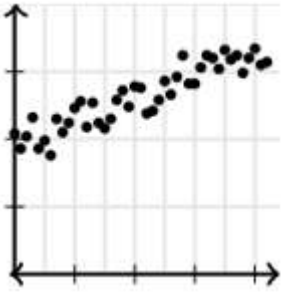


B. Choose the scatterplot that best fits the description.

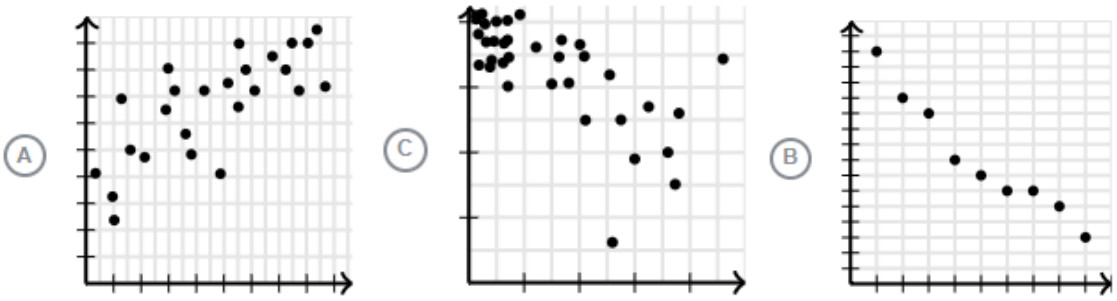
3. "There is a strong, positive, linear association between the two variables"

A 

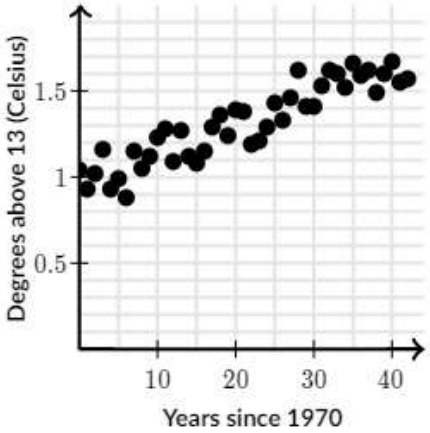
B 

C. 

4. “There is a moderately strong, negative, linear association between the two variables with a few potential outliers.”



5. Which statement is the best description of the association between the two variables?



A. As time went by, the average temperature tended to increase.
B. As the time, went by the average temperature tended to decrease.
C. There is no clear relationship between the time and average temperature.

Discover

Lesson 1: Bivariate data and Scatterplot

Bivariate data deals with two variables that can be explored to establish relationships. Bivariate data has **an independent variable** and **a dependent variable**. We use bivariate data to compare two sets of data and to discover any relationships between them.

Bivariate data examples

- age and heights of the babies and toddlers
- age and the systolic blood pressure
- IQ and academic performance
- Number of absences and grades

One of the best methods for graphing bivariate data that is through the **scatterplot**. A scatterplot is a type of data display that shows the relationship between two numerical variables. Each member of the dataset gets plotted as a point whose (x, y) coordinates relates to its values for the two variables. Scatter plots are used to observe relationships between variables.

A quick description of the association in a scatterplot should always include a description of the *form, direction,* and *strength* of the association, along with the presence of any *outliers*.
Form: Is the association linear or nonlinear?
Direction: Is the association positive or negative?
Strength: Does the association appear to be strong, moderately strong, or weak?
Outliers: Do there appear to be any data points that are unusually far away from the general pattern?

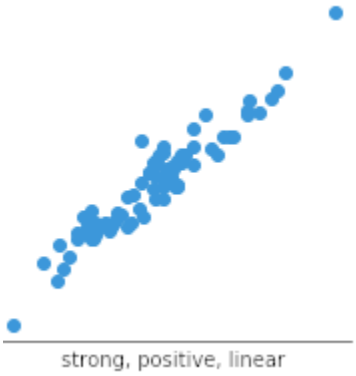


Figure 1.

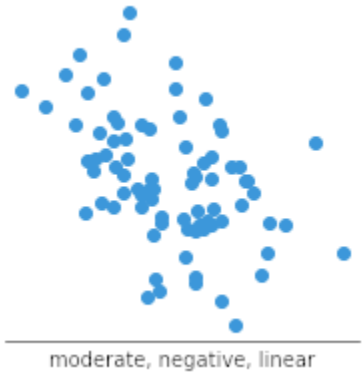


Figure 2.

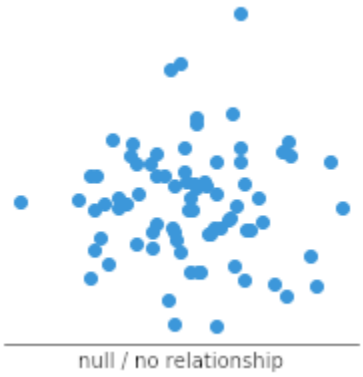


Figure 3

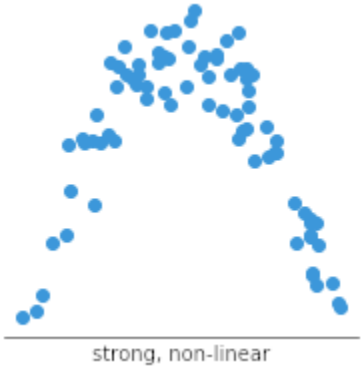


Figure 4

Figure 1.
Form: linear
Direction: positive (increases)
Strength: strong (points are clustered near the line line of fit= if drawn)
Outlier: none

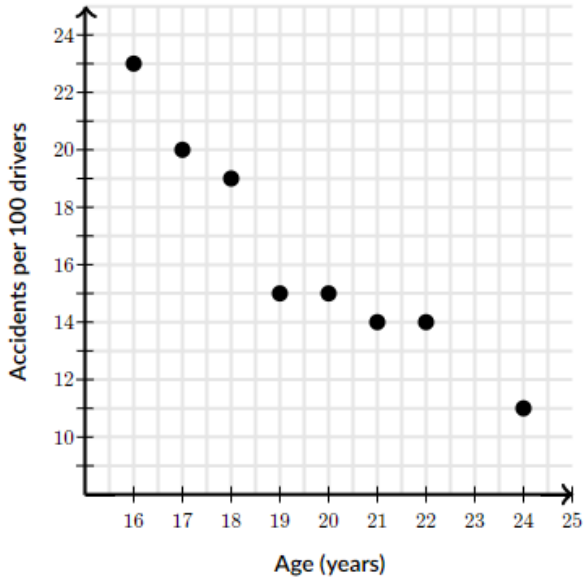
Figure 2
Form: linear
Direction: negative (decreases)
Strength: moderate (some points are away from the of line of fit - drawn)
Outlier: few

Figure 3
Form: non-linear
Direction: null
Strength: null
Outlier: null

Figure 4
Form: non-linear
Direction: negative
Strength: strong
Outlier: none

Example 1.

Let's describe this scatterplot, which shows the relationship between the age of drivers and the number of car accidents per 100 drivers in the year 2009.



This scatterplot shows a strong, negative, linear association between age of drivers and number of accidents. There don't appear to be any outliers in the data."

Example 2: **FRUITMAN CAN!**

A fruit man retailer repacked oranges in different ways and recorded the number of packs sold based on the numbers of oranges per pack. He noticed that very few packs are sold when there are 100 oranges in a pack, so he offers discount for packs of 100 candies based on the number of packs to be bought.

Number of pieces per pack	**Number of packs sold**	**Number of packs (100 pcs. per pack)**	**Discount per pack (in Pesos)**
10	30	3	2
20	27	4	4
30	24	5	6
40	21	6	8
50	18	7	10

60	15	8	12
70	12	9	14
80	9	10	16
90	6	11	18
100	3	12	20

Draw scatterplot for the following and interpret.

1. Number of oranges per pack and the number of packs sold.
2. Number of packs (100 pcs. per pack) to be bought and discount to be given.

Solution:

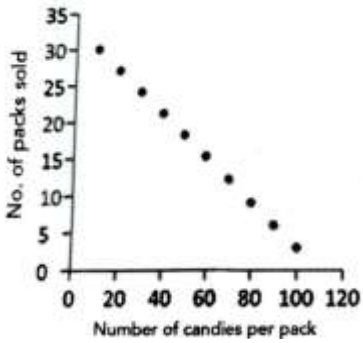


Figure 4

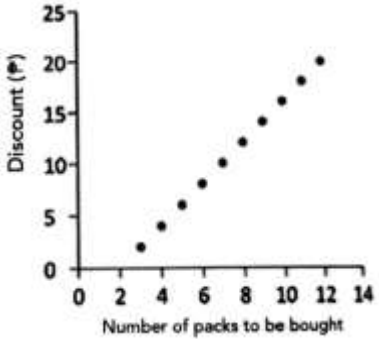


Figure 5

In Figure 4, we can say that there is perfect negative association between the number of oranges per pack and the number of packs sold. That is, the more oranges in a pack, the lesser the number of packs sold. On the other hand, a perfect positive association exists between the number of packs to be bought and the discount to be given, So, the more packs is bought, the bigger the discount to be availed.

Lesson 2: Coefficient of Correlation

Another way to get a qualitative measure for the strength and direction of association between two variables is to determine its correlation coefficient. The **Pearson product moment coefficient of correlation**, denoted by ***r*** with values between -1 and 1. It gives the following us information:

.

- ***The direction of association.*** If $-1 < r < 0$, then the variables are negatively correlated. That is, an increases in one variable implies a decrease in the other variable. If 0 < r < 1, then the variables *positively* correlated. This implies that an increase in one variable implies an increase in the other variable.
- ***The strength of association.*** The closer ***r*** is to -1, the stronger the negative correlation between the variables. The closer ***r*** is to 1, the stronger the positive correlation between the variables. As r gets, closer to 0 the association from either direction becomes weaker. Values of r very close to 0 suggest no correlation between the variables involved.

The following summarizes the correlation coefficient and the strength of relationships:

Computed r-value	Interpretation
0.00	No correlation/no relationship
±0.01 to ± 0.20	Very low correlation/almost Negligible correlation
±0.21 to ± 0.40	Slight correlation, definite but small relationship
±0.41 to + 0.70	Moderate correlation, substantial relationship
±0.71 to ± 0.90	High correlation, marked relationship
± 0.91 to ± 0.99	Very high correlation, very dependable relationship
±1.00	Perfect correlation, perfect relationship

The following formula are used to calculate the Pearson Product-Moment Correlation:

Formula 1: $r_p = \frac{\sum dxdy}{(N-1)(sd_x)(sd_y)}$

where: r_p= Pearson's ***r*** coefficient correlation; N= number of subjects

$\sum dxdy$= the sum of the product of the deviation of x and y variables

$sd_x = \sqrt{\frac{\sum(dx)^2}{N-1}}$ sd_x= standard deviation of x $sd_y = \sqrt{\frac{\sum(dy)^2}{N-1}}$ sd_y= standard deviation of y

$\mu_x = \frac{\sum x}{N}$ $\mu_y = \frac{\sum x}{N}$ $dx = x - \mu$ $dy = y - \mu$

Formula 2: $r = \frac{n\sum xy - (\sum x)(\sum y)}{[\sqrt{n(\sum x^2) - (\sum x)^2}][\sqrt{(n(\sum y^2) - (\sum y)^2}]}$

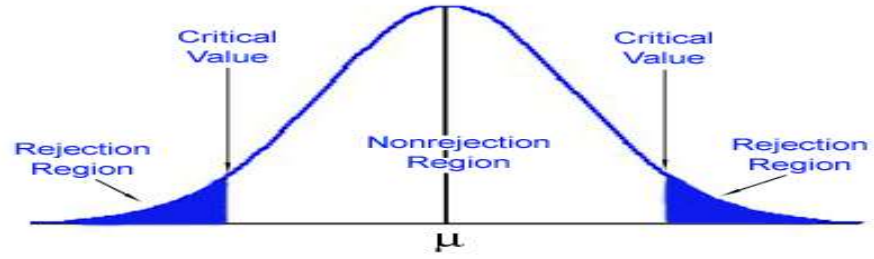
A test of significance for the coefficient of correlation should be used to find out if the computed **r** is significant or not. The test statistic follows the ***t***-

distribution with n-2 degrees of freedom. The formula is :

$$\boldsymbol{t} - value = r\sqrt{\frac{n-2}{1-r^2}}$$

where: t =t-test for correlation coefficient
r = correlation coefficient
n =number of paired samples

The test for correlation is two tailed, which means that the rejection region is divided into two equal parts.



Procedure for the Pearson Product-Moment Correlation Test:

1. Set up the hypothesis

H_0: The correlation in the population in zero or there is no significant correlation between the paired variables.
H_a: The correlation in the population is not zero or there is a significant correlation between the paired variables.

2. Set the level of significance.
3. Calculate the degrees of freedom and determine the critical value of ***t***.
4. Calculate the value of the Pearson ***r.***
5. Calculate the value of ***t*** and arrive at the statistical decision.

If $\boldsymbol{t}_{computed\ value} < critical\ value, accept\ H_o$
If $\boldsymbol{t}_{computed\ value} > critical\ value, reject\ H_o$

6. State the conclusion.

Illustrative Example

The data were taken from a population. In the relationship between the students' English Performance with their Mathematics Performance as shown in the table:

Eng(X)	Math(Y)
19	11
17	15
15	9
14	13
13	11
11	9
11	8

9	7
7	6
4	1

Step 1: Set up the hypothesis

H_0 : There is no significant relationship between the English and Mathematics performance.

H_a : There is significant relationship between the English and Mathematics performance.

Step 2: Set the level of significance.

The level of significance is at 0.05

Step 3: Calculate the degrees of freedom and determine the critical value of *t*.

Df= n-2= 10-2= 8 critical value= 2.31

Step 4: Calculate the value of the Pearson *r*.

Formula 1.

$\boldsymbol{r}_p = \frac{\sum dxdy}{(N-1)(sd_x)(sd_y)}$

Eng(X)	Math(Y)	dx $(X-\mu)$	$dx^2(X-\mu)^2$	dy $(Y-\mu)$	$dy^2(Y-\mu)^2$	$(d$
19	11	7	49	2	4	
17	15	5	25	6	36	
15	9	3	9	0	0	
14	13	2	4	4	16	
13	11	1	1	2	4	
11	9	-1	1	0	0	
11	8	-1	1	-1	1	
9	7	-3	9	-2	4	
7	6	-5	25	-3	9	
4	1	-8	64	-8	64	
$\sum X = 120$	$\sum Y = 90$		$\sum dx^2 = 188$		$\sum dy^2 = 138$	$\sum(dx$
$\mu_x = \frac{120}{10} = 12$	$\mu_x = \frac{90}{10} = 9$					

$sd_x = \sqrt{\frac{\sum(dx)^2}{N-1}} = \sqrt{\frac{188}{10-1}} = 4.57$

$sd_y = \sqrt{\frac{\sum(dy)^2}{N-1}} = \sqrt{\frac{138}{10-1}} = 3.92$

$r_p = \frac{\sum dxdy}{(N-1)(sd_x)(sd_y)}$

$r_p = \frac{140}{(10-1)(4.57)(3.92)}$

$\boldsymbol{r_p}$ = 0.87 which indicates very high positive correlation

Formula 2: $r = \frac{n\sum xy-(\sum x)(\sum y)}{[\sqrt{n(\sum x^2)-(\sum x)^2}][\sqrt{(n(\sum y^2)-(\sum y)^2}]}$

Example: The data were taken from a population. In the relationship between the students' English Performance with their Mathematics Performance as shown in the table:

Eng(X)	Math(Y)	X^2	Y^2	
19	11	361	121	
17	15	289	225	
15	9	225	81	
14	13	196	169	
13	11	169	121	
11	9	121	81	
11	8	121	64	
9	7	81	49	
7	6	49	36	
4	1	16	1	
$\sum X = 120$	$\sum Y = 90$	$\sum X^2 = 1628$	$\sum Y^2 = 1628$	$\sum$

$$r = \frac{n\sum xy-(\sum x)(\sum y)}{[\sqrt{n(\sum x^2)-(\sum x)^2}][\sqrt{(n(\sum y^2)-(\sum y)^2}]}$$

$$r = \frac{10(1220)-(120)(90)}{\sqrt{[10(1628)-(120)^2][10(948)-(90)^2}}$$

$$r = \frac{1400}{\sqrt{2594400}}$$

r = .87 which indicates very high positive correlation

Step 5: Calculate the value of *t* and arrive at the statistical decision.

$$t - value = r\sqrt{\frac{n-2}{1-r^2}}$$

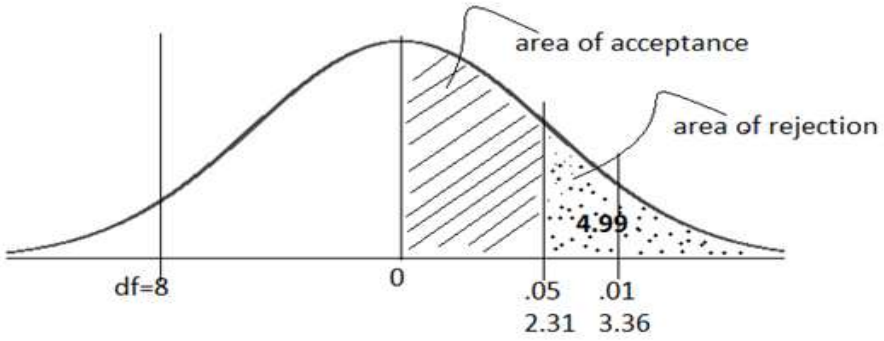
$$= 0.87\sqrt{\frac{10-2}{1-(.87)^2}}$$

$$=0.87\ (5.74)$$

$$= 4.99$$

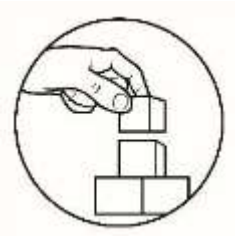
In testing the significance of ***r***- value, compare the computed ***t***-value with that of the tabular value at .05 alpha or at a ***p***-value set at .05 alpha or a two tailed test.

Calculated ***t***-value	Tabular ***t***-value	
	.05	.01
4.99	2.31	3.36



Step 6. State the conclusion.

The null hypothesis if there is no significant relationship between the English and Mathematics performance of the students was rejected. The researcher is confident positive that there is a significant relationship between the variables.



Explore

Here are some enrichment activities for you to work on to master and strengthen the basic concepts you have learned from this lesson.

Activity 2: KNOW MY CORRELATION!

Directions: For each of the following cases, identify whether there is positive correlation, negative correlation, or no correlation exists.

_____ 1. Total family income and family expenses
_____ 2. Mathematics grades and height of students
_____ 3. Company sales and advertising expenses
_____ 4. Number of absences and grades
_____ 5. Number of policemen and number of crimes recorded

_____6. Number of pages and price of the book

_____7. Weight and dancing ability

_____8. Educational attainment and salary

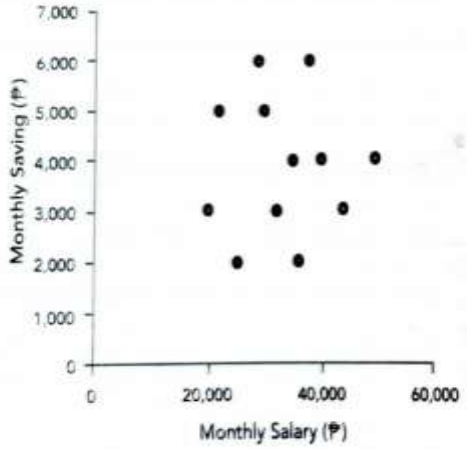
_____9. School allowance and academic performance

_____10. Size of the family and expenses

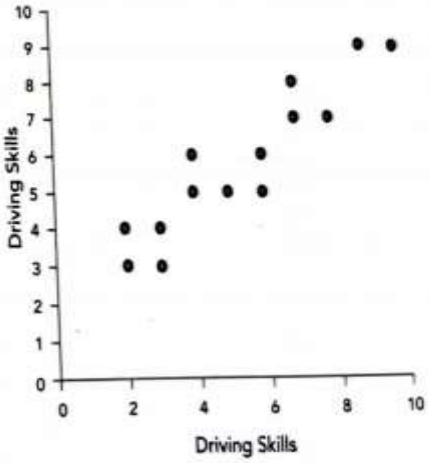
Activity 3: WHAT'S MY RELATIONSHIP?

Directions: Write the statement showing the relationship between the two variables.

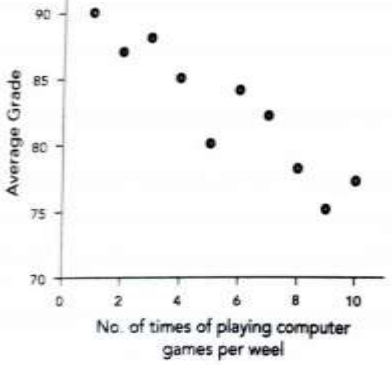
1.



2.

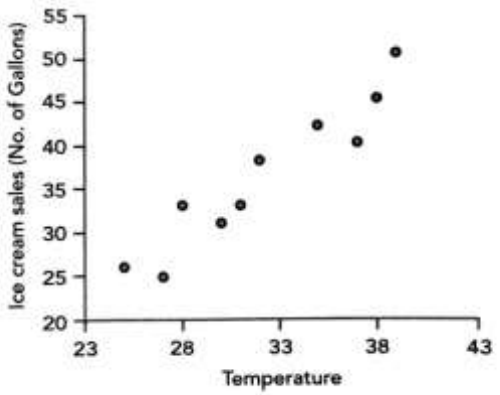


3.



4.

5.



Activity 4: APPLY WHAT YOU HAVE LEARNED!

A psychologist wanted to determine if there is a correlation between anxiety experienced by patients before an operation and religiosity. The patients completed an anxiety scale (high score=high anxiety) and also completed a checklist design to measure an individual's degree of religiosity (belief in a particular religion, regular attendance at religious services, number of times per week they regularly pray, etc.) (high score= greater religiosity). A data sample is provided below:

Anxiety before Operation	Religiosity
38	4
42	3
29	11
31	5
28	9
15	6
24	14
17	9
19	10
11	15
8	19
19	17
3	10
14	14
6	18

Perform the procedure for the Pearson Product-Moment Correlation Test.

Deepen

There is a popular belief that if you are good in English, you are not good in Mathematics and vice versa. This means a negative correlation, meaning, the higher your grade in English, the lower is your grade in Math, or the other way around. To prove that this is wrong, a Mathematics teacher gathered some data on the grades of students and came up with the table below:

Student	Grade in English	Grade in Math
1	75	76
2	88	89
3	94	95
4	75	73
5	86	85
6	83	84
7	79	78
8	77	76
9	76	76
10	93	94
11	91	90
12	89	88
13	85	86
14	84	85
15	84	83

Is the popular belief correct or just a misconception?
Justify and show your solution.

Very well done! You are now ready to take your posttest. You may again go over the lessons, activities and maps to review for the final assessment.
Good luck!

Gauge

Directions: Read each item carefully. Use a separate sheet for your answers. Write only the letter of the best answer for each test item.

1. What type of data uses two sets of variables that can change and are compared to find relationships?

   A. Bivariate B. Correlation C. Dependent D. Independent

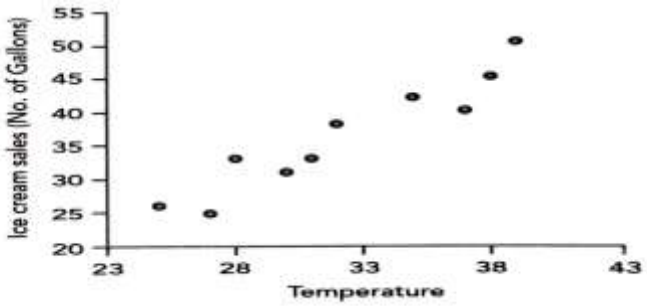
2. A condition or piece of data in an experiment that is controlled or influenced by an outside factor is considered what type of variable?

   A. Bivariate B. Correlation C. Dependent D. Independent

3. What type of relationship is evident between foot length and height?
   A. A strong negative linear relationship.
   B. A strong positive linear relationship
   C. A weak negative linear relationship.
   D. A weak positive linear relationship.

4. The correlation coefficient measures:
   A. Whether there is a relationship between two variables
   B. The strength of the relationship between two quantitative variables
   C. Whether a cause and effect relation exists between two variables
   D. The strength of the linear relationship between two quantitative variables.

5. What type of relationship between ice cream sales and temperature is shown on the scatter graph?
   A. A strong positive linear relationship.
   B. A strong negative linear relationship
   C. A weak positive linear relationship.
   D. A weak negative linear relationship.



6. Which of the following is true of the correlation ***r?***

A. It is a resistant measure of association

B. $-1 \leq r \leq 1$

C. If ***r*** is the correlation between X and Y, then $-r$ is the correlations between Y and X

D. Whenever all the data lie on a perfectly straight-line, the correlations ***r*** will always be equal to +1.0

7. When the values of two variables move in the same direction, correlation is said to be _____.

A. Linear B. Positive C. Non-linear D. Negative

8. When the values of two variables move in the opposite directions, correlation is said to be _____.

A. Linear B. Positive C. Non-linear D. Negative

9. When the amount of change in one variable leads to a constant ratio of change in the other variable, then correlation is said to be _____.

A. Linear B. Positive C. Non-linear D. Negative

10. Scatter diagram is also called _____.

A. Scatter chart B. Correlation graph

C. Both a and b D. None of these

11. What measure of correlation measures the strength and direction of the linear relationship of two variables and its association between interval and ordinal data?

A. Spearman's rho B. Pearson's ***r*** C. Chi-Square D. T-test

12. If there is no correlation or a weak correlation, what is the value of ***r***?

A. 0 B. -1 C. 1 D. 2

13. What statistical treatment is used to test the significant relationship between the number of hours of computer gaming and the students daily allowance?

A. Chi-Square test B. ANOVA C. Correlational rank D. T-test

14. What decision will you make if the computed ***t***-value is less than the critical value?

A. Accept the null hypothesis
B. Accept the Alternative hypothesis
C. Reject the null hypothesis
D. None of the Above

15. If there are 15 number of paired samples, what is the degree of freedom?

A. 14 B. 13 C. 12 D. 11

References

Printed Materials:

Calaca, Ninia I., Manalo, Ronald A., Noble, Nestor M. Statistics and Probability. Araneta Ave., Quezon City: Vibal Group, Inc. 2016. pp. 354-362

*DepED Material: Statistics and Probability Learner's Material

Melosantos, Luis Allan B., Antonio, Janice F., Robles, Susan J., Bruce, Ryan M.,Math Connections in the Digital Age Statistics and Probability. Sibs Publishing House. Inc. 2016. pp. 3-7

Website:

http://www.westmaths.com.au › Bivariate+Data+Test+ETA

bhsapstats.weebly.com /uploads/3/8/0/2/38020589 /bivariate_data_analysis_review_2017.

dept.stat.Isa.umich.edu/˜kshedden/ Courses/Stat401/Notes/401-bivariate-slides.

web.cortland.edu>STATS>corr.

https://www.khanacademy.org/math/ap-statistics/bivariate-data-ap/scatterplots-correlation/e/interpreting-scatter-plots