



UNIVERSITY OF MINDANAO
College of Arts and Sciences Education

Physically Distanced but Academically Engaged

Self-Instructional Manual (SIM) for Self-Directed Learning (SDL)

Course/Subject: **Probability and Statistics (MTH 103L)**

Name of Teacher: **Rosyl S. Matin-ao, MAT**

THIS SIM/SDL MANUAL IS A DRAFT VERSION
ONLY, NOT FOR REPRODUCTION AND
DISTRIBUTION OUTSIDE OF ITS INTENDED
USE. THIS IS INTENDED ONLY FOR THE USE OF
THE STUDENTS WHO ARE OFFICIALLY
ENROLLED IN THE COURSE/SUBJECT.
EXPECT REVISIONS OF THE MANUAL.
NOT FOR REPRODUCTION AND COMMERCIAL USE.

Course Outline: Probabilities and Statistics (MTH 103L)

Course Facilitator:	ROSYL S. MATIN-AO
Email:	rosyl_matinao@umindanao.edu.ph
Student Consultation:	By Appointment
Mobile:	0999-544-7736
Phone:	-
Effectivity Date:	June 2020
Mode of Delivery:	Blended (On-line with face-to-face or virtual sessions)
Time Frame:	-
Student Workload	Expected Self-Directed Learning
Requisites:	None
Credit:	3 units (2 units lab & 1-unit lecture)
Attendance Requirements:	A minimum of 95% attendance is required at all scheduled Virtual or face to face sessions.

Course Outline Policy

Areas of Concern	Details
Contact and Non-contact Hours	This 5-unit course self-instructional manual is designed for blended learning mode of instructional delivery, i.e. online sessions through the LMS and the 2-days on-campus / onsite face-to-face review and final examination. The expected number of hours will be 162 including review and examination days. The face to face sessions shall include the summative assessment tasks (exams) since this course is crucial in the psychologist licensure examination.
Assessment Task Submission	<p>Submission of assessment tasks shall be on the 3rd, 5th, 7th, and 9th weeks of the term. The assessment paper shall be attached with a cover page indicating the title of the assessment task (<i>if the task is a performance</i>), the name of the course coordinator, date of submission, and the name of the student. The document should be emailed to the course coordinator.</p> <p>It is also expected that you already paid your tuition and other fees before the submission of the assessment task.</p> <p>If the assessment task is done in real-time through the features in the Blackboard Learning Management System, the schedule shall be arranged ahead of time by the course coordinator.</p>

<p>Turnitin Submission (if necessary)</p>	<p>To ensure honesty and authenticity, all assessment tasks are required to be submitted through Turnitin with a maximum similarity index of 30% allowed. This means that if your paper goes beyond 30%, the students will either opt to redo her/his paper or explain in writing addressed to the Course Facilitator the reasons for the similarity. In addition, if the paper has reached more than 30% similarity index, the student may be called for a disciplinary action in accordance with the University's OPM on Intellectual and Academic Honesty.</p> <p>Please note that academic dishonesty such as cheating and commissioning other students or people to complete the task for you have severe punishments (reprimand, warning, and expulsion).</p>
<p>Penalties for Late Assignments / Assessments</p>	<p>The score for an assessment item submitted after the designated time on the due date, without an approved extension of time, will be reduced by 5% of the possible maximum score for that assessment item for each day or part day that the assessment item is late.</p> <p>However, if the late submission of assessment paper has a valid reason, a letter of explanation should be submitted and approved by the Course Facilitator. If necessary, you will also be required to present/attach evidences.</p>
<p>Return of Assignments / Assessments</p>	<p>Assessment tasks will be returned to you two (2) weeks after the submission. This will be returned by email or via Blackboard portal.</p> <p>For group assessment tasks, the Course Facilitator will require some or few of the students for online or virtual sessions to ask clarificatory questions to validate the originality of the assessment task submitted and to ensure that all the group members are involved.</p>
<p>Assignment Resubmission</p>	<p>You should request in writing addressed to the Course Facilitator his/her intention to resubmit an assessment task. The resubmission is premised on the student's failure to comply with the similarity index and other reasonable grounds such as academic literacy standards or other reasonable circumstances e.g. illness, accidents financial constraints.</p>

Re-marking of Assessment Papers and Appeal	<p>You should request in writing addressed to the program coordinator your intention to appeal or contest the score given to an assessment task. The letter should explicitly explain the reasons/points to contest the grade. The program coordinator shall communicate with the students on the approval and disapproval of the request.</p> <p>If disapproved by the Course Facilitator, you can elevate your case to the program head or the dean with the original letter of request. The final decision will come from the dean of the college.</p>
Grading System	<p>ASSESSMENT METHOD</p> <p>Lecture (40%)</p> <ul style="list-style-type: none"> • 1 – 3 Exam (10% each) 30% • Final Exam 40% • LMS Activities 30% <p style="text-align: right;">TOTAL 100%</p> <p>Laboratory (60%)</p> <ul style="list-style-type: none"> • Written Exam 30% • Practical Exam 30% • Laboratory Activities 40% <p style="text-align: right;">TOTAL 100%</p>
Preferred Referencing Style	Depends on the discipline; if uncertain or inadequate, use the general practice of the APA 6th Edition.
Student Communication	<p>You are required to create a umindanao email account which is a requirement to access the BlackBoard portal. Then, the course coordinator shall enroll the students to have access to the materials and resources of the course. All communication formats: chat, submission of assessment tasks, requests etc. shall be through the portal and other university recognized platforms.</p> <p>You can also meet the course coordinator in person through the scheduled face to face sessions to raise your issues and concerns.</p> <p>For students who have not created their student email, please contact the course coordinator or program head.</p>
Contact Details of the Dean	<p>Khristine Marie D. Concepcion, Ph.D Email: artsciences@umindanao.edu.ph Phone: (082) 300-5456 / 305-0647 Local 134</p>
Contact Details of the Program Head	<p>Ronnie O. Alejan, MSAM Email: ronnie_alejan@umindanao.edu.ph Phone: (082) 300-5456 / 305-0647 Local 134</p>

Students with Special Needs	Students with special needs shall communicate with the Course Facilitator about the nature of his or her special needs. Depending on the nature of the need, the Course Facilitator, with the approval of the Program Head, may provide alternative assessment tasks or extension of the deadline of submission of assessment tasks. However, the alternative assessment tasks should still be in the service of achieving the desired course learning outcomes.
Blackboard LMS Helpdesk	blackboardclass@umindanao.edu.ph
Library Contact Details	Brigida E. Bacani Email: brigida_bacani@umindanao.edu.ph Phone: 300-5456 local 143
Well-being Welfare Support Held Desk Contact Details	Zerdszen P. Ranises – GSTC Facilitator Email: gstcmain@umindanao.edu.ph Phone: 0950-466-5431

Course Information: See download course syllabus in the Black Board LMS

CC's Voice: Hello students! Welcome to this course **MTH 103L: Probabilities and Statistics**. This course is an introduction to statistics and data analysis. It covers the following: reasons for doing statistics, collection, summarization and presentation of data, basic concepts in probability, point and interval estimation, and hypothesis testing.

Course Outcome: At the end of the course, you are expected to:

- 1) Apply different statistical techniques in the decision-making process.
- 2) Use computer software for data analysis and model building.

Let us begin!

Continuation of Week 6 - 7 Lessons

Big Picture in Focus

ULO – c. Test the difference between two means for dependent samples.

Testing the Difference between Two Means: Dependent Samples

In the previous lesson, the t test was used to compare two sample means when the samples were independent. In this lesson, a different version of the t test is explained. This version is used when the samples are dependent. Samples are **dependent samples** when the subjects are paired or matched in some way.

For example, suppose a medical researcher wants to see whether a drug will affect the reaction time of its users. To test this hypothesis, the researcher must pretest the subjects in the sample first. That is, they are given a test to ascertain their normal reaction times. Then after taking the drug, the subjects are tested again, using a posttest. Finally, the means of the two tests are compared to see whether there is a difference. Since the same subjects are used in both cases, the samples are *related*; subjects scoring high on the pretest will generally score high on the posttest, even after consuming the drug. Likewise, those scoring lower on the pretest will tend to score lower on the posttest. To take this effect into account, the researcher employs a t test, using the differences between the pretest values and the posttest values. Thus, only the gain or loss in values is compared.

Here are some other examples of dependent samples. A researcher may want to design an SAT preparation course to help students raise their test scores the second time they take the SAT. Hence, the differences between the two exams are compared. A medical specialist may want to see whether a new counselling program will help subjects lose weight. Therefore, the pre-weights of the subjects will be compared with the post weights.

Besides samples in which the same subjects are used in a pre-post situation, there are other cases where the samples are considered dependent. For example, students might be matched or paired according to some variable that is pertinent to the study; then one student is assigned to one group, and the other student is assigned to a second group. For instance, in a study involving learning, students can be selected and paired according to their IQs. That is, two students with the same IQ will be paired. Then one will be assigned to one sample group (which might receive instruction by computers), and the other student will be assigned to another sample group (which might receive instruction by the lecture discussion method). These assignments will be done randomly. Since a student's IQ is important to learning, it is a variable that should be controlled. By matching subjects on IQ, the researcher can eliminate the variable's influence, for the most part. Matching, then, helps to reduce type II error by eliminating extraneous variables.

Two notes of caution should be mentioned. First, when subjects are matched according to one variable, the matching process does not eliminate the influence of other variables. Matching students according to IQ does not account for their mathematical ability or their familiarity with computers. Since not all variables influencing a study can be controlled, it is up to the researcher to determine which variables should be used in matching. Second, when the same subjects are used for a pre-post study, sometimes the knowledge that they are participating in a study can influence the results. For example, if people are placed in a special program, they may be more highly motivated to succeed simply because they have been selected to participate; the program itself may have little effect on their success.

When the samples are dependent, a special t test for dependent means is used. This test employs the difference in values of the matched pairs. The hypotheses are as follows:

Two-tailed	Left-tailed	Right-tailed
$H_0: \mu_D = 0$	$H_0: \mu_D = 0$	$H_0: \mu_D = 0$
$H_1: \mu_D \neq 0$	$H_1: \mu_D < 0$	$H_1: \mu_D > 0$

Assumptions for the t Test for Two Means When the Samples Are Dependent

1. The sample or samples are random.
2. The sample data are dependent.
3. When the sample size or sample sizes are less than 30, the population or populations must be normally or approximately normally distributed.

Formulas for the t Test for Dependent Samples

$$t = \frac{\bar{D}}{s_D / \sqrt{n}}$$

with d.f. = $n - 1$ and where

$$\bar{D} = \frac{\sum D}{n} \quad \text{and} \quad s_D = \sqrt{\frac{n \sum D^2 - (\sum D)^2}{n(n-1)}}$$

Example 1: A sample of nine local banks shows their deposits (in billions of dollars) 3 years ago and their deposits (in billions of dollars) today. At $\alpha = 0.05$, can it be concluded that the average in deposits for the banks is greater today than it was 3 years ago?

Source: SNL Financial

Account	1	2	3	4	5	6	7	8	9
3 years ago	11.42	8.41	3.98	7.37	2.28	1.10	1.00	0.9	1.35
Today	16.69	9.44	6.53	5.58	2.92	1.88	1.78	1.5	1.22

Solution

STEP 1: State the hypothesis and identify the claim. Since we are interested to see if there has been an increase in deposits, the deposits 3 years ago must be less than the deposits today; hence, the differences must be significantly less 3 years ago than they are today. Hence the mean of the differences must be less than zero.

$$H_0: \mu_D = 0 \quad \text{and} \quad H_1: \mu_D < 0 \text{ (claim)}$$

STEP 2: Find the critical value. The degrees of freedom are $n - 1$, or $9 - 1 = 8$. The critical value for a left-tailed test with $\alpha = 0.05$ is -1.860 .

	Confidence intervals	80%	90%	95%	98%	99%
	One tail, α	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, α	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250
10		1.372	1.812	2.228	2.764	3.169

STEP 3: Compute the test value.

3 years ago (X_1)	Now (X_2)	$X_1 - X_2 = D$	D^2
11.42	16.69	$11.42 - 16.69 = -5.27$	$(-5.27)^2 = 27.7729$
8.41	9.44	$8.41 - 9.44 = -1.03$	$(-1.03)^2 = 1.0609$
3.98	6.53	$3.98 - 6.53 = -2.55$	$(-2.55)^2 = 6.5025$
7.37	5.58	$7.37 - 5.58 = 1.79$	$(1.79)^2 = 3.2041$
2.28	2.92	$2.28 - 2.92 = -0.64$	$(-0.64)^2 = 0.4096$
1.10	1.88	$1.10 - 1.88 = -0.78$	$(-0.78)^2 = 0.6084$
1.00	1.78	$1.00 - 1.78 = -0.78$	$(-0.78)^2 = 0.6084$
0.90	1.50	$0.90 - 1.50 = -0.60$	$(-0.60)^2 = 0.3600$
1.35	1.22	$1.35 - 1.22 = 0.13$	$(0.13)^2 = 0.1690$
-	-	$\sum D = -9.73$	$\sum D^2 = 40.5437$

Mean Difference

$$\bar{D} = \frac{\sum D}{n} = \frac{-9.73}{9} = -1.081$$

Standard Deviation

$$S_D = \sqrt{\frac{n \sum D^2 - (\sum D)^2}{n(n-1)}} = \sqrt{\frac{9(40.543) - (-9.73)^2}{9(9-1)}} = 1.937$$

Test value

$$t = \frac{\bar{D}}{\frac{S_D}{\sqrt{n}}} = \frac{-1.081}{\frac{1.937}{\sqrt{9}}} = \frac{-1.081}{\frac{1.937}{3}} = -\frac{1.081}{0.646} = -1.67$$

STEP 4: Make the decision. Do not reject the null hypothesis since the test value, -1.67, is greater than the critical value, -1.860.

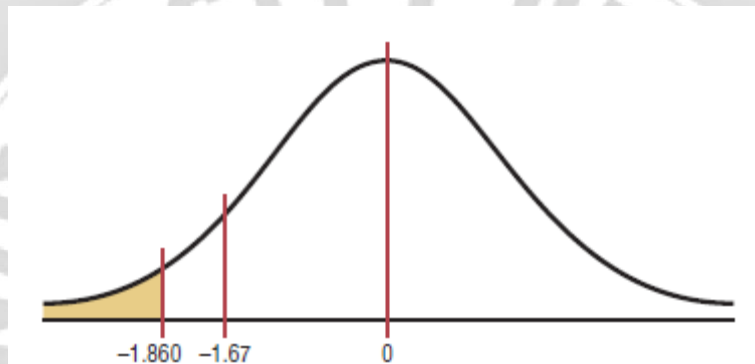


Figure 10-5

STEP 5: Summarize the results. There is not enough evidence to show that the deposits have increased over the last 3 years.

Example 2: A dietician wishes to see if a person's cholesterol level will change if the diet is supplemented by a certain mineral. Six subjects were pretested, and then they took the mineral supplement for a 6-week period. The results are shown in the table. (Cholesterol level is measured in milligrams per deciliter) Can it be concluded that the cholesterol level has been changed at $\alpha = 0.10$? Assume the variable is approximately normally distributed.

Subject	1	2	3	4	5	6
Before (X_1)	210	235	208	190	172	244
After (X_2)	190	170	210	188	173	228

Solution

STEP 1: State the hypotheses and identify the claim. If the diet is effective, the before cholesterol levels should be different from the after levels.

$$H_0: \mu_D = 0 \quad \text{and} \quad H_1: \mu_D \neq 0 \text{ (claim)}$$

STEP 2: Find the critical value. The degree of freedom is 5 ($d.f = 6 - 1 = 5$). At $\alpha = 0.10$, the critical values are ± 2.015 .

	Confidence intervals	80%	90%	95%	98%	99%
	One tail, α	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, α	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.152	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707

STEP 3: Compute the test value.

Before (X_1)	After (X_2)	$X_1 - X_2 = D$	D^2
210	190	$210 - 190 = 20$	$(20)^2 = 400$
235	170	$235 - 170 = 65$	$(65)^2 = 4225$
208	210	$208 - 210 = -2$	$(-2)^2 = 4$
190	188	$190 - 188 = 2$	$(2)^2 = 4$
172	173	$172 - 173 = -1$	$(-1)^2 = 1$
244	228	$244 - 228 = 16$	$(16)^2 = 256$
-	-	$\sum D = 100$	$\sum D^2 = 4890$

Mean Difference

$$\bar{D} = \frac{\sum D}{n} = \frac{100}{6} = 16.67$$

Standard Deviation

$$S_D = \sqrt{\frac{n \sum D^2 - (\sum D)^2}{n(n-1)}} = \sqrt{\frac{6(4890) - (100)^2}{6(6-1)}} = 25.4$$

Test value

$$t = \frac{\bar{D}}{\frac{S_D}{\sqrt{n}}} = \frac{16.67}{\frac{25.4}{\sqrt{6}}} = \frac{16.67}{10.37} = 1.61$$

STEP 4: Make the decision. The decision is to not reject the null hypothesis, since the test value 1.610 is in the noncritical region, as shown in Figure 9-6 below.

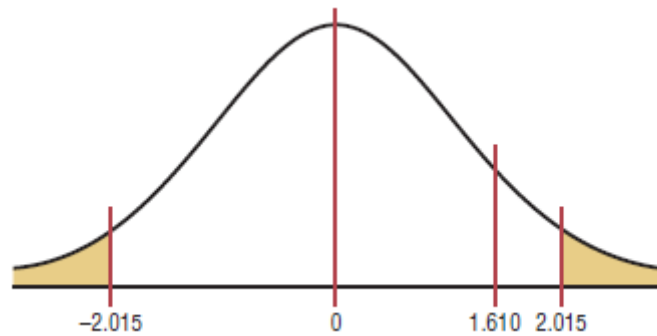


Figure 10-6

STEP 5: Summarize the results. There is not enough evidence to support the claim that the mineral changes a person's cholesterol level.

SELF-HELP: You can also refer to the sources below to help you understand the lesson.

1. Bluman, A. (2012). *Elementary Statistics: A Step by Step Approach 8th Edition*. McGraw-Hill Companies, Inc.
2. Pagano, R. (2009). *Understanding Statistics in the Behavioral Sciences 9th Edition*. Wadsworth Cengage Learning.

LET'S CHECK

ACTIVITY 1

Now that you know the most essential concepts on testing the difference between two dependent sample means, let us try to check your understanding of these concepts.

Use the traditional method in testing the hypothesis in the problems below. In each problem, state the following:

- State the hypotheses and identify the claim.
 - Find the critical value(s)
 - Find the test value
 - Make the decision
 - Summarize the result
- The manager of the cosmetics section of a large department store wants to determine whether newspaper advertising really does affect sales. For her experiment, she randomly selects 15 items currently in stock and proceeds to establish a baseline. The 15 items are priced at their usual competitive values, and the quantity of each item sold for a 1-week period is recorded. Then, without changing their price, she places a large ad in the newspaper, advertising the 15 items. Again, she records the quantity sold for a 1-week period. The results follow.

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. of Items Sold Before Ad	25	18	3	42	16	20	23	32	60	40	27	7	13	23	16
No. of Items Sold After Ad	32	24	7	40	19	25	23	35	60	43	28	11	12	32	28

Using $\alpha = 0.05$, two-tailed test. Analyze the data then interpret the result.

- Developmental psychologists at a prominent California university conducted a longitudinal study investigating the effect of high levels of curiosity in early childhood on intelligence. The local population of 3-year-olds was screened via a test battery assessing curiosity. Twelve of the 3-year-olds scoring in the upper 90% of this variable were given an IQ test at age 3 and again at age 11. The following IQ scores were obtained.

Student Number	1	2	3	4	5	6	7	8	9	10	11	12
IQ (Age of 3)	100	105	125	140	108	122	117	112	135	128	104	98
IQ (Age of 11)	114	116	139	151	106	119	131	136	148	139	122	113

Using $\alpha = 0.01$, two-tailed test, analyse the data and then interpret the result.

LET'S ANALYZE

ACTIVITY 1

Getting acquainted with the essential terms and concepts on testing the difference between two dependent sample means, it is now time for you to explain thoroughly your answers to the following questions.

1. Explain the role of the following statistical terms in testing the difference of two dependent samples.

- a. Hypotheses (Null and Alternative Hypotheses)

- b. Test of significance (One Tailed Test and Two Tailed Test)

- c. Level of significance

- d. Critical region

- e. Critical value


- f. Test value

IN A NUTSHELL

ACTIVITY 1

Based on the concept on testing the difference between two dependent sample means and the learning exercises that you have done, write your arguments or lessons learned below.

1.



2.

3.

4.

5.

Q & A LIST

Do you have any question for clarification?	
Questions / Issues	Answers
1.	1.
2.	2.
3.	3.
4.	4.
5.	5.

KEYWORDS INDEX

**Inferential
Statistics**

**Hypothesis
Testing**

**Difference between
Two Means**

**Dependent
Sample Mean**

**Traditional
Method**

BIG PICTURE

WEEK 8: UNIT LEARNING OUTCOMES (ULO): At the end of the chapter, you are expected to:

- Use the one-way ANOVA technique to determine if there is a significant difference among three or more means.
- Determine which means differ, using the Scheffé or Tukey test if the null hypothesis is rejected in the ANOVA.

Chapter 11: ANALYSIS OF VARIANCE

METALANGUAGE

To demonstrate all **ULOs**, some unfamiliar terms yet essential will be defined to understand this chapter. You will encounter these terms as we go through this chapter. Please refer to these definitions in case you will encounter difficulty in understanding some concepts.

- Between-Group Variance** – In analysis of variance, variability based on differences among the means of the various groups.
- F-distribution** – the statistical model used to test hypotheses when the analysis involves the comparison of variance estimates.
- Post hoc comparison** – a significance test involving two or more group means that was not planned prior to obtain the results.
- Within-group variance** – In analysis of variance, difference among the scores within groups.

ESSENTIAL KNOWLEDGE

Suppose a researcher wishes to see whether the means of the time it takes three groups of students to solve a computer problem using Fortran, Basic, and Pascal are different. The researcher will use the ANOVA technique for this test. The z and t tests should not be used when three or more means are compared, for reasons given later in this chapter.

For three groups, the F test can only show whether a difference exists among the three means. It cannot reveal where the difference lies—that is, between X_1 and X_2 , or X_1 and X_3 , or X_2 and X_3 . If the F test indicates that there is a difference among the means, other statistical tests are used to find where the difference exists. The most commonly used tests are the Scheffé test and the Tukey test, which are also explained in this chapter.

The analysis of variance that is used to compare three or more means is called a **one-way analysis of variance** since it contains only one variable. In the previous example, the variable is the type of computer language used. The analysis of variance can be extended to studies involving two variables, such as type of computer language used and mathematical background of the students. These studies involve a **two-way analysis of variance**.

Big Picture in Focus

ULO – a. Use the one-way ANOVA technique to determine if there is a significant difference among three or more means.

One-way Analysis of Variance

When an F test is used to test a hypothesis concerning the means of three or more populations, the technique is called **analysis of variance** (commonly abbreviated as **ANOVA**). At first glance, you might think that to compare the means of three or more samples, you can use the t test, comparing two means at a time. But there are several reasons why the t test should not be done.

First, when you are comparing two means at a time, the rest of the means under study are ignored. With the F test, all the means are compared simultaneously. Second, when you are comparing two means at a time and making all pairwise comparisons, the probability of rejecting the null hypothesis when it is true is increased, since the more t tests that are conducted, the greater is the likelihood of getting significant differences by chance alone. Third, the more means there are to compare, the more t tests are needed. For example, for the comparison of 3 means two at a time, 3 t tests are required. For the comparison of 5 means two at a time, 10 tests are required. And for the comparison of 10 means two at a time, 45 tests are required.

Assumptions for the F Test for Comparing Three or More Means

1. The populations from which the samples were obtained must be normally or approximately normally distributed.
2. The samples must be independent of one another.
3. The variances of the populations must be equal.

Even though you are comparing three or more means in this use of the F test, *variances* are used in the test instead of means.

With the F test, two different estimates of the population variance are made. The first estimate is called the **between-group variance**, and it involves finding the variance of the means. The second estimate, the **within-group variance**, is made by computing the variance using all the data and is not affected by differences in the means. If there is no difference in the means, the between-group variance estimate will be approximately equal to the within-

group variance estimate, and the F test value will be approximately equal to 1. The null hypothesis will not be rejected. However, when the means differ significantly, the between-group variance will be much larger than the within-group variance; the F test value will be significantly greater than 1; and the null hypothesis will be rejected. Since variances are compared, this procedure is called **analysis of variance (ANOVA)**.

For a test of the difference among three or more means, the following hypotheses should be used:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

H_1 : At least one mean is different from the others

As stated previously, a significant test value means that there is a high probability that this difference in means is not due to chance, but it does not indicate where the difference lies.

The degrees of freedom for this F test are $d.f.N. = k - 1$, where k is the number of groups, and $d.f.D. = N - k$, where N is the sum of the sample sizes of the groups $N = n_1 + n_2 + \cdots + n_k$. The sample sizes need not be equal. The F test to compare means is always right-tailed.

Examples 1 and 2 illustrate the computational procedure for the ANOVA technique for comparing three or more means. See the examples below.

Example 1: A researcher wishes to try three different techniques to lower the blood pressure of individuals diagnosed with high blood pressure. The subjects are randomly assigned to three groups; the first group takes medication, the second group exercises, and the third group follows a special diet. After four weeks, the reduction in each person's blood pressure is recorded. At 0.05, test the claim that there is no difference among the means. The data are shown.

Medication	Exercise	Diet
10	6	5
12	8	9
9	3	12
15	0	8
13	2	4

Solution

STEP 1: State the hypotheses and identify the claim.

$$H_0: \mu_1 = \mu_2 = \mu_3 \text{ (claim)}$$

H_1 : At least one mean is different from the others.

STEP 2: Find the critical value. Since $k = 3$ and $N = 15$,

$$d.f.N. = k - 1 = 3 - 1 = 2$$

$$d.f.D. = N - k = 15 - 3 = 12$$

The critical value is 3.89, obtained from Table C in Appendix with $\alpha = 0.05$.

d.f.D.: degrees of freedom, denominator	$\alpha = 0.05$																		
	d.f.N.: degrees of freedom, numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07



Note:

In finding the critical value in F-table, make sure to look at correctly the given α before looking at the N and D on the table. Critical values of F-table are classified according to the α .

STEP 3: Compute the test value, using the procedure outlined here.

- a. Find the mean and variance of each sample.

	Medication (X_1)	Exercise (X_2)	Diet (X_3)
Mean	$\bar{X}_1 = 11.8$	$\bar{X}_2 = 3.8$	$\bar{X}_3 = 7.6$
Variance	$s_1^2 = 5.7$	$S_2^2 = 10.2$	$S_3^2 = 10.3$

- b. Find the grand mean. The *grand mean*, denoted by \bar{X}_{GM} , is the mean of all values in the samples.

$$\bar{X}_{GM} = \frac{\sum X}{N} = \frac{10 + 12 + 9 + \dots + 4}{15} = \frac{116}{15} = 7.73$$

When samples are equal in size, find \bar{X}_{GM} by summing the \bar{X} 's and dividing by k , where k = the number of groups.

- c. Find the between-group variance, denoted by s_B^2 .

$$\begin{aligned}
 s_B^2 &= \frac{\sum n_i(\bar{X}_i - \bar{X}_{GM})^2}{k - 1} \\
 &= \frac{5(11.8 - 7.73)^2 + 5(3.8 - 7.73)^2 + 5(7.6 - 7.73)^2}{3 - 1} \\
 &= \frac{160.13}{2} = 80.07
 \end{aligned}$$

Note: This formula finds the variance among the means by using the sample sizes as weights and considers the differences in the means.

- d. Find the within-group variance, denoted by s_W^2 .

$$\begin{aligned}
 s_W^2 &= \frac{\sum(n_i - 1)s_i^2}{\sum(n_i - 1)} \\
 &= \frac{(5 - 1)(5.7) + (5 - 1)(10.2) + (5 - 1)(10.3)}{(5 - 1) + (5 - 1) + (5 - 1)} \\
 &= \frac{104.80}{12} = 8.73
 \end{aligned}$$

Note: This formula finds an overall variance by calculating a weighted average of the individual variances. It does not involve using differences of the means.

e. Find the F -test value.

$$F = \frac{s_B^2}{s_W^2} = \frac{80.07}{8.73} = 9.17$$

STEP 4: Make the decision. The decision is to reject the null hypothesis, since $9.17 > 3.89$.

STEP 5: Summarize the results. There is enough evidence to reject the claim and conclude that at least one mean is different from the others.

The numerator of the fraction obtained in step 3, part c, of the computational procedure is called the **sum of squares between groups**, denoted by SS_B . The numerator of the fraction obtained in step 3, part d, of the computational procedure is called the **sum of squares within groups**, denoted by SS_W . This statistic is also called the **sum of squares for the error**. SS_B is divided by $d.f. N$ to obtain the between-group variance. SS_W is divided by $N - k$ to obtain the within-group or error variance. These two variances are sometimes called **mean squares**, denoted by MS_B and MS_W . These terms are used to summarize the analysis of variance and are placed in a summary table, as shown below.

ANALYSIS OF VARIANCE SUMMARY OF TABLE				
Source	Sum of squares	d.f.	Mean square	F
Between	SS_B	$k - 1$	MS_B	
Within (error)	SS_W	$N - k$	MS_W	
Total				

In the table,

SS_B = sum of squares between groups

SS_W = sum of squares within groups

k = number of groups

$N = n_1 + n_2 + \cdots + n_k$ = sum of sample sizes for groups

$$MS_B = \frac{SS_B}{k - 1}$$

$$MS_W = \frac{SS_W}{N - k}$$

$$F = \frac{MS_B}{MS_W}$$

The totals are obtained by adding the corresponding columns. For Example 1, the **ANOVA summary table** is shown in the table below.

ANALYSIS OF VARIANCE SUMMARY OF TABLE FOR EXAMPLE 1				
Source	Sum of squares	d.f.	Mean square	<i>F</i>
Between	160.13	2	80.07	9.17
Within (error)	104.80	12	8.73	
Total	264.93	14		

Most computer programs will print out an ANOVA summary table.

Example 2: A state employee wishes to see if there is a significant difference in the number of employees at the interchanges of three state toll roads. The data are shown. At $\alpha = 0.05$, can it be concluded that there is a significant difference in the average number of employees at each interchange?

Pennsylvania Turnpike	Greensburg Bypass / Mon-Fayette Expressway	Beaver Valley Expressway
7	10	1
14	1	12
32	1	1
19	0	9
10	11	1
11	1	11

Source: Pennsylvania Turnpike Commission.

Solution

STEP 1: State the hypotheses and identify the claim.

$$H_0: \mu_1 = \mu_2 = \mu_3$$

H_1 : At least one mean is different from the others. (claim)

STEP 2: Find the critical value. Since $k = 3$ and $N = 18$, and $\alpha = 0.05$.

$$d.f.N. = k - 1 = 3 - 1 = 2$$

$$d.f.D. = N - k = 18 - 3 = 15$$

The critical value is 3.68

d.f.D.: degrees of freedom, denominator	The critical value is 3.68 $\alpha = 0.05$																		
	d.f.N.: degrees of freedom, numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92

STEP 3:

a. Find the mean and variance of each sample.

	Pennsylvania Turnpike (X_1)	Greensburg Bypass / Mon-Fayette Expressway (X_2)	Beaver Valley Expressway (X_3)
Mean	$\bar{X}_1 = 15.5$	$\bar{X}_2 = 4.0$	$\bar{X}_3 = 5.8$
Variance	$s_1^2 = 81.9$	$s_2^2 = 25.6$	$s_3^2 = 29.0$

b. Find the grand mean. The *grand mean*, denoted by \bar{X}_{GM} , is the mean of all values in the samples.

$$\bar{X}_{GM} = \frac{\sum X}{N} = \frac{7 + 14 + 32 + \cdots + 11}{18} = \frac{152}{18} = 8.4$$

c. Find the between-group variance, denoted by s_B^2 .

$$\begin{aligned} s_B^2 &= \frac{\sum n_i(\bar{X}_i - \bar{X}_{GM})^2}{k - 1} \\ &= \frac{6(15.5 - 8.4)^2 + 6(4 - 8.4)^2 + 6(5.8 - 8.4)^2}{3 - 1} \\ &= \frac{459.18}{2} = 229.59 \end{aligned}$$

d. Find the within-group variance.

$$\begin{aligned}
 s_W^2 &= \frac{\sum(n_i - 1)s_i^2}{\sum(n_i - 1)} \\
 &= \frac{(6 - 1)(81.9) + (6 - 1)(25.6) + (6 - 1)(29.0)}{(6 - 1) + (6 - 1) + (6 - 1)} \\
 &= \frac{682.5}{15} = 45.5
 \end{aligned}$$

e. Find the F -test value.

$$F = \frac{s_B^2}{s_W^2} = \frac{229.59}{45.5} = 5.05$$

STEP 4: Make the decision. Since $5.05 > 3.68$, the decision is to reject the null hypothesis.

STEP 5: Summarize the results. There is enough evidence to support the claim that there is a difference among the means. The ANOVA summary table for this example is shown in the table below.

ANALYSIS OF VARIANCE SUMMARY OF TABLE FOR EXAMPLE 2

Source	Sum of squares	d.f.	Mean square	F
Between	459.18	2	229.59	5.05
Within	682.5	15	45.5	
Total	1141.68	17		

When the null hypothesis is rejected in ANOVA, it only means that at least one mean is different from the others. To locate the difference or differences among the means, it is necessary to use other tests such as the Tukey or the Scheffé test.

SELF-HELP:

You can also refer to the sources below to help you understand the lesson.

1. Bluman, A. (2012). *Elementary Statistics: A Step by Step Approach 8th Edition*. McGraw-Hill Companies, Inc.

LET'S CHECK

ACTIVITY 1

Now that you know the most essential concepts on Analysis of Variance (ANOVA), let us try to check your understanding of these concepts.

- Find the F_{critical} for the following situations:
 - df (numerator) = 2 ; df (denominator) = 16 ; $\alpha = 0.005$
 - df (numerator) = 3 ; df (denominator) = 36 ; $\alpha = 0.01$
 - df (numerator) = 3 ; df (denominator) = 10 ; $\alpha = 0.025$
 - df (numerator) = 5 ; df (denominator) = 25 ; $\alpha = 0.05$
 - df (numerator) = 3 ; df (denominator) = 8 ; $\alpha = 0.10$
- The accompanying table is a one-way, independent groups ANOVA summary table with part of the material missing.

Source	Sum of Squares	Df	s^2	F
Between groups	1,253.68	3	?	?
Within groups	?	?		
Total	5,016.40	39		

- Fill the missing values.
- How many groups are there in the experiment?
- Assuming an equal number of subjects in each group, how many of subjects are there in each group?
- What is the value of F_{critical} using $\alpha = 0.05$?

For item no. 3: Use the traditional method in testing the hypothesis in the problems below. In the problem, state the following:

- State the hypotheses and identify the claim.
- Find the critical value(s)
- Find the test value
- Make the decision
- Summarize the result

3. A sleep researcher conducts an experiment to determine whether sleep loss affects the ability to maintain sustained attention. Fifteen individuals are randomly divided into the following three groups of five subjects each: group 1, which gets the normal amount of sleep (7–8 hours); group 2, which is sleep-deprived for 24 hours; and group 3, which is sleep-deprived for 48 hours. All three groups are tested on the same auditory vigilance task. Subjects are presented with half-second tones spaced at irregular intervals over 1-hour duration. Occasionally, one of the tones is slightly shorter than the rest. The subject's task is to detect the shorter tones. The following percentages of correct detections were observed:

Normal Sleep	Sleep-Deprived for 24-hours	Sleep-Deprived for 48-hours
85	60	60
83	58	48
76	76	38
64	52	47
75	63	50

Determine whether there is an overall effect for sleep deprivation, using the conceptual equations of the one-way ANOVA. Use $\alpha = 0.05$.

LET'S ANALYZE

ACTIVITY 1

Getting acquainted with the essential terms and concepts on Analysis of Variance (ANOVA), it is now time for you to explain thoroughly your answers to the questions found below.

1. Explain the usage of the following statistical tools in testing the hypotheses.
 - a. Testing One Sample Mean using t-test and z-test.

- b. Testing Two Independent Sample Mean.

- c. Testing Dependent Sample Mean.

- d. Analysis of Variance


2. What are the similarities found in all statistical tool mentioned above?

IN A NUTSHELL

ACTIVITY 1

Based on the concept on Analysis of Variance (ANOVA) and the learning exercises that you have done, write your arguments or lessons learned below.

1.



2.

3.

4.

5.

Q & A LIST

Do you have any question for clarification?	
Questions / Issues	Answers
1.	1.
2.	2.
3.	3.
4.	4.
5.	5.

KEYWORDS INDEX

Inferential Statistics	Hypothesis Testing	Analysis of Variance
Between-group Variance	Within-group Variance	Sum of squares between groups
Squares within group	Sum of squares for the error	Mean Squares

Big Picture in Focus

ULO – b. Determine which means differ using the Scheffé or Tukey test if the null hypothesis is rejected in the ANOVA.

The Scheffe Test and the Tukey Tests

When the null hypothesis is rejected using the F test, the researcher may want to know where the difference among the means is. Several procedures have been developed to determine where the significant differences in the means lie after the ANOVA procedure has been performed. Among the most commonly used tests are the **Scheffé test** and the **Tukey test**.

Scheffé test

To conduct the **Scheffé test**, you must compare the means two at a time, using all possible combinations of means. For example, if there are three means, the following comparisons must be done:

$$\bar{X}_1 \text{ versus } \bar{X}_2 \quad \bar{X}_1 \text{ versus } \bar{X}_3 \quad \bar{X}_2 \text{ versus } \bar{X}_3$$

Formula for the Scheffé Test

$$F_s = \frac{(\bar{X}_i - \bar{X}_j)^2}{s_w^2[(1/n_i) + (1/n_j)]}$$

where \bar{X}_i and \bar{X}_j are the means of the samples being compared, n_i and n_j are the respective sample sizes, and s_w^2 is the within-group variance.

To find the critical value F' for the Scheffé test, multiply the critical value for the F test by $k - 1$:

$$F' = (k - 1)(C.V.)$$

There is a significant difference between the two means being compared when F_s is greater than F' . The example below illustrates the use of the Scheffé test.

Example 1: Using the Scheffé test, test each pair of means in the previous Example 1 of One-way ANOVA to see whether a specific difference exists, at $\alpha = 0.05$.

Solution

Recall the following important data.

	Medication	Exercise	Diet
	10	6	5
	12	8	9
	9	3	12
	15	0	8
	13	2	4
	Medication (X_1)	Exercise (X_2)	Diet (X_3)
Mean	$\bar{X}_1 = 11.8$	$\bar{X}_2 = 3.8$	$\bar{X}_3 = 7.6$
Variance	$s_1^2 = 5.7$	$S_2^2 = 10.2$	$S_3^2 = 10.3$
n	$n_1 = 5$	$n_2 = 5$	$n_3 = 5$

ANALYSIS OF VARIANCE SUMMARY OF TABLE FOR EXAMPLE 1

Source	Sum of squares	d.f.	Mean square	F
Between	160.13	2	80.07	9.17
Within (error)	104.80	12	8.73	
Total	264.93	14		

a. For \bar{X}_1 versus \bar{X}_2 ,

$$F_S = \frac{(\bar{X}_1 - \bar{X}_2)^2}{s_{\bar{w}}^2[(1/n_1) + (1/n_2)]} = \frac{(11.8 - 3.8)^2}{8.73[(1/5) + (1/5)]} = 18.33$$

b. For \bar{X}_2 versus \bar{X}_3 ,

$$F_S = \frac{(\bar{X}_2 - \bar{X}_3)^2}{s_{\bar{w}}^2[(1/n_2) + (1/n_3)]} = \frac{(3.8 - 7.6)^2}{8.73[(1/5) + (1/5)]} = 4.14$$

c. For \bar{X}_1 versus \bar{X}_3 ,

$$F_S = \frac{(\bar{X}_1 - \bar{X}_3)^2}{s_{\bar{w}}^2[(1/n_1) + (1/n_3)]} = \frac{(11.8 - 7.6)^2}{8.73[(1/5) + (1/5)]} = 5.05$$

The critical value for the analysis of variance for Example 12–1 was 3.89, found by using Table C with $\alpha = 0.05$, $d.f.N. = k - 1 = 2$, and $d.f.D. = N - k = 12$. In this case, it is multiplied by $k - 1$ as shown.

The critical value for F' at $\alpha = 0.05$, with $d.f.N. = 2$ and $d.f.D. = 12$, is

$$F' = (k - 1)(C.V.) = (3 - 1)(3.89) = 7.78$$

Since only the F test value for part a (\bar{X}_1 versus \bar{X}_2) is greater than the critical value, 7.78, the only significant difference is between \bar{X}_1 and \bar{X}_2 , that is, between medication and exercise.



Note:

On occasion, when the F test value is greater than the critical value, the Scheffé test may not show any significant differences in the pairs of means. This result occurs because the difference may actually lie in the average of two or more means when compared with the other mean. The Scheffé test can be used to make these types of comparisons, but the technique is beyond the scope of this module.

Tukey Test

The **Tukey test** can also be used after the analysis of variance has been completed to make pairwise comparisons between means when the groups have the same sample size. The symbol for the test value in the Tukey test is q .

Formula for the Tukey Test

$$q = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{s_W^2/n}}$$

where \bar{X}_i and \bar{X}_j are the means of the samples being compared, n is the size of the samples, and s_W^2 is the within-group variance.

When the absolute value of q is greater than the critical value for the Tukey test, there is a significant difference between the two means being compared. The procedures for finding q and the critical value from Table N in Appendix C for the Tukey test are shown in the example below.

Example 1: Using the Tukey test, test each pair of means in the previous example 1 of One-way ANOVA to see whether a specific difference exists, at $\alpha = 0.05$.

Solution

a. For \bar{X}_1 versus \bar{X}_2 ,

$$q = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_w^2/n}} = \frac{11.8 - 3.8}{\sqrt{8.73/5}} = \frac{8}{1.32} = 6.06$$

b. For \bar{X}_1 versus \bar{X}_3 ,

$$q = \frac{\bar{X}_1 - \bar{X}_3}{\sqrt{s_w^2/n}} = \frac{11.8 - 7.6}{\sqrt{8.73/5}} = \frac{4.2}{1.32} = 3.18$$

c. For \bar{X}_2 versus \bar{X}_3 ,

$$q = \frac{\bar{X}_2 - \bar{X}_3}{\sqrt{s_w^2/n}} = \frac{3.8 - 7.6}{\sqrt{8.73/5}} = \frac{-3.8}{1.32} = -2.88$$

To find the critical value for the Tukey test, use Table D in Appendix. The number of means k is found in the row at the top, and the degrees of freedom for s_w^2 are found in the left column (denoted by v). Since $k = 3$, $d.f. = 12$, and $\alpha = 0.05$, the critical value is 3.77. See the figure below. Hence, the only q value that is greater in absolute value than the critical value is the one for the difference between \bar{X}_1 and \bar{X}_2 . The conclusion, then, is that there is a significant difference in means for medication and exercise. These results agree with the Scheffé analysis.

$\alpha = 0.05$				
$k \backslash v$	2	3	4	5
11	3.11	3.82	4.26	4.57
12	3.08	3.77	4.20	4.51
13	3.06	3.73	4.15	4.45
14	3.03	3.70	4.11	4.41
15	3.01	3.67	4.08	4.37

You might wonder why there are two different tests that can be used after the ANOVA. There are several other tests that can be used in addition to the Scheffé and Tukey tests. It is up to the researcher to select the most appropriate test. The Scheffé test is the most general, and it can be used when the samples are of different sizes. Furthermore, the Scheffé test can be used to make comparisons such as the average of X_1 and X_2 compared with X_3 . However, the Tukey test is more powerful than the Scheffé test for making pairwise comparisons for the means. A rule of thumb for pairwise comparisons is to use the Tukey test when the samples are equal in size and the Scheffé test when the samples differ in size.

SELF-HELP: You can also refer to the sources below to help you understand the lesson.

1. Bluman, A. (2012). *Elementary Statistics: A Step by Step Approach 8th Edition*. McGraw-Hill Companies, Inc.



LET'S CHECK

ACTIVITY 1

Now that you know the most essential concepts on Scheffé and Tukey tests, let us try to check your understanding of these concepts.

The following set of data values was obtained from a study of people's perceptions on whether the color of a person's clothing is related to how intelligent the person looks. The subjects rated the person's intelligence on a scale of 1 to 10. Group 1 subjects were randomly shown people with clothing in shades of blue and gray. Group 2 subjects were randomly shown people with clothing in shades of brown and yellow. Group 3 subjects were randomly shown people with clothing in shades of pink and orange. The results follow.

Group 1	8	7	7	7	8	8	6	8	8	7	7	8	8
Group 2	7	8	7	7	5	8	5	8	7	6	6	6	6
Group 3	4	9	6	7	9	8	5	8	7	5	4	5	4

1. Use the Tukey test to test all possible pairwise comparisons.
2. Are there any contradictions in the results?
3. Explain why separate t tests are not accepted in this situation.
4. When would Tukey's test be preferred over the Scheffé method? Explain.

LET'S ANALYZE

ACTIVITY 1

Getting acquainted with the essential terms and concepts on Scheffé and Tukey tests, it is now time for you to explain thoroughly your answers to the following questions.

1. What two tests can be used to compare two means when the null hypothesis is rejected using the one-way ANOVA F -test?

2. Explain the difference between the two tests used to compare two means when the null hypothesis is rejected using the one-way ANOVA F -test?

3. When the null hypothesis is accepted, can Scheffé and Tukey tests still be used? Explain.


4. Explain the difference between Scheffé and Tukey tests in terms of the process and usage.

IN A NUTSHELL

ACTIVITY 1

Based on the concept on Scheffé and Tukey tests and the learning exercises that you have done, write your arguments or lessons learned below.

1.



2.

3.

4.

5.

Q & A LIST

Do you have any question for clarification?	
Questions / Issues	Answers
1.	1.
2.	2.
3.	3.
4.	4.
5.	5.

KEYWORDS INDEX

**Inferential
Statistics**

**Hypothesis
Testing**

ANOVA

Scheffé Test

Tukey Test

BIG PICTURE

WEEK 9: UNIT LEARNING OUTCOMES (ULO): At the end of the chapter, you are expected to:

- Draw a scatter plot for a set of ordered pairs.
- Compute the correlation coefficient.
- Test the hypothesis $H_0: \rho = 0$.
- Compute the equation of the regression line.

Chapter 12: CORRELATION

METALANGUAGE

To demonstrate all **ULOs**, some unfamiliar terms yet essential will be defined to understand this chapter. You will encounter these terms as we go through this chapter. Please refer to these definitions in case you will encounter difficulty in understanding some concepts.

- Correlation coefficient** – a measure of the extent to which scores on one variable are related to scores on a second variable.
- Criterion** – the variable being predicted in a linear regression analysis.
– a dependent variable.
- Least squares regression line** – minimizes the sum of squared errors in prediction (the sum of squared deviations between the predicted scores and actual scores).
- Linear regression** – procedure for determining the straight line that will enable us to predict scores on one variable (the criterion) from scores on another variable (the predictor), while minimizing the amount of (squared) error.
- Pearson r** – a measure of the strength of the linear relationship between two variables, both of which are either continuous or dichotomous.
- Scatter plot** – a graph showing the relationship between two variables, wherein the score of each case on both variables is expressed as point in two dimensions.

ESSENTIAL KNOWLEDGE

Another area of inferential statistics involves determining whether a relationship exists between two or more numerical or quantitative variables. For example, a businessperson may want to know whether the volume of sales for a given month is related to the amount of advertising the firm does that month. Educators are interested in determining whether the

number of hours a student studies is related to the student's score on a particular exam. Medical researchers are interested in questions such as, is caffeine related to heart damage? Or is there a relationship between a person's age and his or her blood pressure? A zoologist may want to know whether the birth weight of a certain animal is related to its life span. These are only a few of the many questions that can be answered by using the techniques of correlation and regression analysis. **Correlation** is a statistical method used to determine whether a linear relationship between variables exists. **Regression** is a statistical method used to describe the nature of the relationship between variables, that is, positive or negative, linear or nonlinear.

The purpose of this chapter is to answer these questions statistically:

1. Are two or more variables linearly related?
2. If so, what is the strength of the relationship?
3. What type of relationship exists?
4. What kind of predictions can be made from the relationship?

To answer the first two questions, statisticians use a numerical measure to determine whether two or more variables are linearly related and to determine the strength of the relationship between or among the variables. This measure is called a **correlation coefficient**. For example, there are many variables that contribute to heart disease, among them lack of exercise, smoking, heredity, age, stress, and diet. Of these variables, some are more important than others; therefore, a physician who wants to help a patient must know which factors are most important.

To answer the third question, you must ascertain what type of relationship exists. There are two types of relationships: *simple* and *multiple*. In a **simple relationship**, there are two variables—an **independent variable**, also called an explanatory variable or a predictor variable, and a **dependent variable**, also called a response variable. A simple relationship analysis is called *simple regression*, and *there is one independent variable that is used to predict the dependent variable*. For example, a manager may wish to see whether the number of years the salespeople have been working for the company has anything to do with the amount of sales they make. This type of study involves a simple relationship since there are only two variables—years of experience and amount of sales.

In a **multiple relationship**, called **multiple regressions**, two or more independent variables are used to predict one dependent variable. For example, an educator may wish to investigate the relationship between a student's success in college and factors such as the number of hours devoted to studying, the student's GPA, and the student's high school background. This type of study involves several variables.

Simple relationships can also be positive or negative. A **positive relationship** exists when either variables increase or decrease at the same time. For instance, a person's height and weight are related; and the relationship is positive, since the taller a person is, generally, the more the person weighs. In a **negative relationship**, as one variable increases, the other variable decreases, and vice versa. For example, if you measure the strength of people over 60 years of age, you will find that as age increases, strength generally decreases. The word *generally* is used here because there are exceptions.

Finally, the fourth question asks what type of predictions can be made. Predictions are made in all areas and daily. Examples include weather forecasting, stock market analyses, sales predictions, crop predictions, gasoline price predictions, and sports predictions. Some

predictions are more accurate than others, due to the strength of the relationship. That is, the stronger the relationship is between variables, the more accurate the prediction is.

Big Picture in Focus

ULO – a. Draw a scatter plot for a set of ordered pairs.

Scatter Plot and Correlation

In simple correlation and regression studies, the researcher collects data on two numerical or quantitative variables to see whether a relationship exists between the variables. For example, if a researcher wishes to see whether there is a relationship between number of hours of study and test scores on an exam, she must select a random sample of students, determine the hours each studied, and obtain their grades on the exam. A table can be made for the data, as shown here.

Student	Hours of Study (X)	Grade % (Y)
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

As stated previously, the two variables for this study are called the independent variable and the dependent variable. The independent variable is the variable in regression that can be controlled or manipulated. In this case, the number of hours of study is the independent variable and is designated as the x variable. The dependent variable is the variable in regression that cannot be controlled or manipulated. The grade the student received on the exam is the dependent variable, designated as the y variable. The reason for this distinction between the variables is that you assume that the grade the student earns *depends* on the number of hours the student studied. Also, you assume that, to some extent, the student can regulate or *control* the number of hours he or she studies for the exam.

The determination of the x and y variables is not always clear-cut and is sometimes an arbitrary decision. For example, if a researcher studies the effects of age on a person's blood pressure, the researcher can generally assume that age affects blood pressure. Hence, the variable *age* can be called the *independent variable*, and the variable *blood pressure* can be called the *dependent variable*. On the other hand, if a researcher is studying the attitudes of husbands on a certain issue and the attitudes of their wives on the same issue, it is difficult to

say which variable is the independent variable and which the dependent variable is. In this study, the researcher can arbitrarily designate the variables as independent and dependent.

The independent and dependent variables can be plotted on a graph called a **scatter plot**. The independent variable x is plotted on the horizontal axis, and the dependent variable y is plotted on the vertical axis.

A **scatter plot** is a graph of the ordered pairs (x, y) of numbers consisting of the independent variable x and the dependent variable y .

The scatter plot is a visual way to describe the nature of the relationship between the independent and dependent variables. The scales of the variables can be different, and the coordinates of the axes are determined by the smallest and largest data values of the variables.

The procedure for drawing a scatter plot is shown in the examples below.

Example 1: Construct a scatter plot for the data shown for car rental companies in the United States for a recent year.

Company	Car (in ten thousands)	Revenue (in billions)
A	63.0	7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

Source: Auto Rental News.

Solution

STEP 1: Draw and label the x and y axes.

STEP 2: Plot each point on the graph, as shown in Figure 11–1.

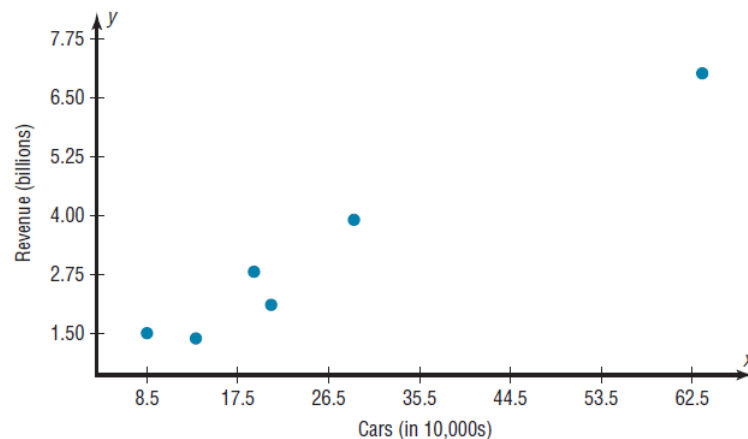


Figure 12-1

Example 2: Construct a scatter plot for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class. The data are shown here.

Student	Number of absences (X)	Final grade % (Y)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

Solution

STEP 1: Draw and label the x and y axes.

STEP 2: Plot each point on the graph, as shown in Figure 11–2.

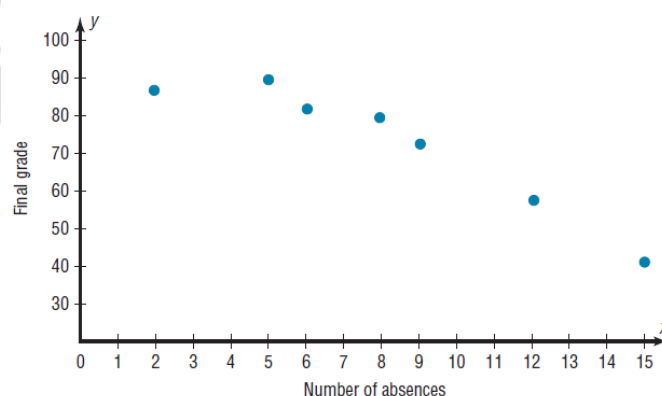


Figure 12-2

Example 3: A researcher wishes to see if there is a relationship between the ages and net worth of the wealthiest people in America. The data for a specific year are shown.

Person	Age (X)	Net wealth (Y) – in billion \$
A	73	16
B	65	26
C	53	50
D	54	21.5
E	79	40
F	69	16
G	61	19.6
H	65	19

Source: Forbes magazine.

Solution

STEP 1: Draw and label the x and y axes.

STEP 2: Plot each point on the graph, as shown in Figure 11–3.

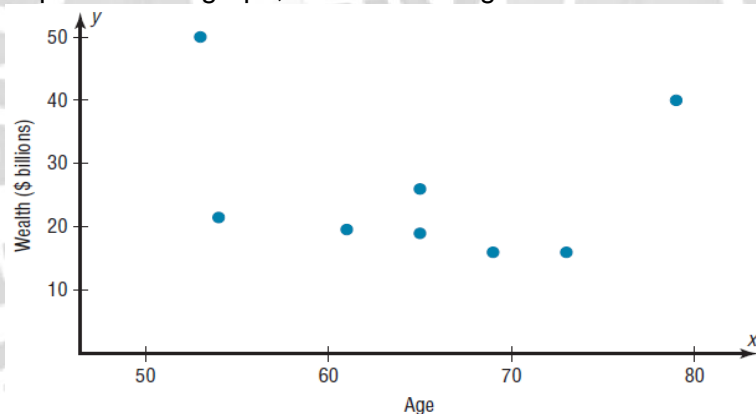


Figure 11-3

After the plot is drawn, it should be analyzed to determine which type of relationship, if any, exists. For example, the plot shown in Figure 11–1 suggests a positive relationship, since as the number of cars rented increases, revenue tends to increase also. The plot of the data shown in Figure 11–2 suggests a negative relationship, since as the number of absences increases, the final grade decreases. Finally, the plot of the data shown in Figure 11–3 shows no specific type of relationship, since no pattern is discernible.

Note that the data shown in Figures 11–1 and 11–2 also suggest a linear relationship, since the points seem to fit a straight line, although not perfectly. Sometimes a scatter plot, such as the one in Figure 11–4, shows a curvilinear relationship between the data. In this situation, curvilinear relationships are beyond the scope of this module.

Big Picture in Focus

ULO – b. Compute the correlation coefficient.

Correlation

Correlation Coefficient As stated in the Introduction, statisticians use a measure called the *correlation coefficient* to determine the strength of the linear relationship between two variables. There are several types of correlation coefficients. The one explained in this section is called the **Pearson product moment correlation coefficient (PPMC)**, named after statistician Karl Pearson, who pioneered the research in this area.

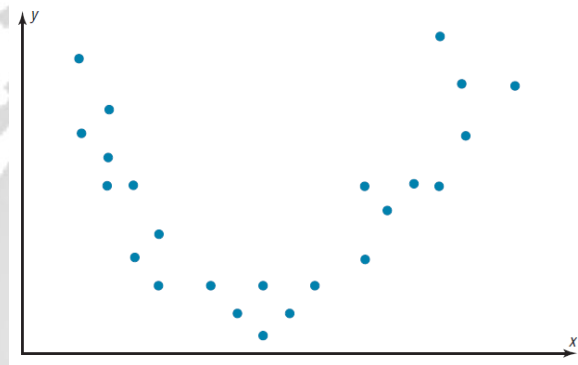


Figure 12-4

The **correlation coefficient** computed from the sample data measures the strength and direction of a linear relationship between two quantitative variables. The symbol for the sample correlation coefficient is r . The symbol for the population correlation coefficient is ρ (Greek letter rho).

The **range of the correlation coefficient** is from -1 to $+1$. If there is a **strong positive linear relationship** between the variables, the value of r will be close to $+1$. If there is a **strong negative linear relationship** between the variables, the value of r will be close to -1 . When there is no linear relationship between the variables or only a weak relationship, the value of r will be close to 0 . See Figure 11–5.

The graphs in Figure 11–6 show the relationship between the correlation coefficients and their corresponding scatter plots. Notice that as the value of the correlation coefficient increases from 0 to $+1$ (parts a , b , and c), data values become closer to an increasingly strong relationship. As the value of the correlation coefficient decreases from 0 to -1 (parts d , e , and f), the data values also become closer to a straight line. Again this suggests a stronger relationship.

There are several ways to compute the value of the correlation coefficient. One method is to use the formula shown here.

Formula for the Correlation Coefficient r

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

where n is the number of data pairs.

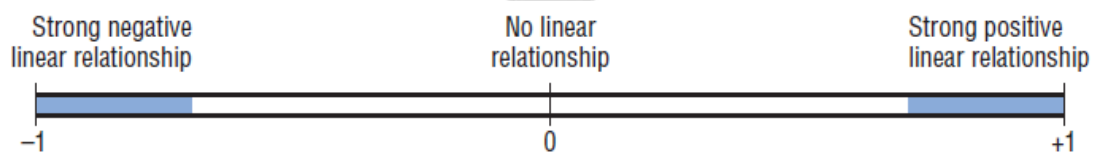


Figure 12-5

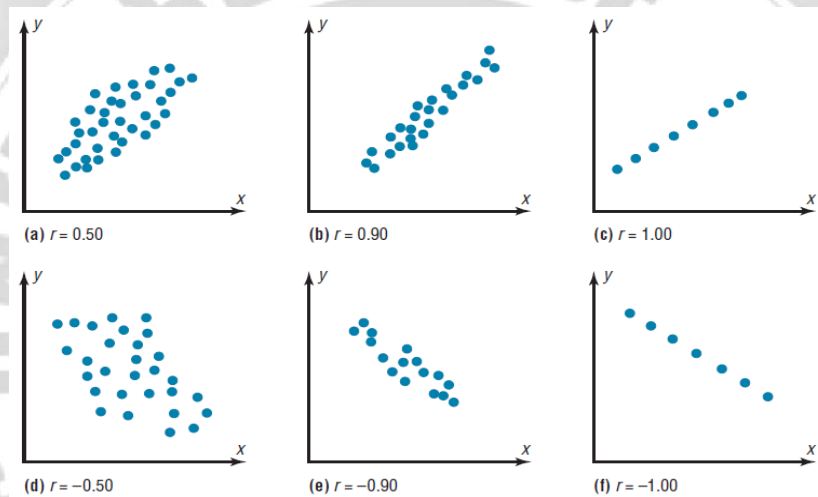


Figure 12-6

Assumptions for the Correlation Coefficient

1. The sample is a random sample.
2. The data pairs fall approximately on a straight line and are measured at the interval or ratio level.
3. The variables have a joint normal distribution. (This means that given any specific value of x , the y values are normally distributed; and given any specific value of y , the x values are normally distributed.)



Rounding Rule for the Correlation Coefficient

Round the value of r to three decimal places.

The formula looks somewhat complicated, but using a table to compute the values, as shown in the previous 4 examples, makes it somewhat easier to determine the value of r .

There are no units associated with r , and the value of r will remain unchanged if the x and y values are switched.

Example 1: Compute the correlation coefficient for the car rental companies in the United States for a recent year.

Company	Car (in ten thousands)	Revenue (in billions)
A	63.0	7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

Source: Auto Rental News.

Solution

STEP 1: Make a table as shown here.

Company	Car (X) (in ten thousands)	Revenue (Y) (in billions)	XY	X^2	Y^2
A	63.0	7.0			
B	29.0	3.9			
C	20.8	2.1			
D	19.1	2.8			
E	13.4	1.4			
F	8.5	1.5			

STEP 2: Find the values of XY , X^2 , and Y^2 and place these values in the corresponding columns of the table.

The completed table is shown.

Company	Car (X)	Revenue (Y)	XY	X ²	Y ²
A	63.0	7.0	63.0 x 7.0 = 441.00	(63.0) ² = 3969.0	(7.0) ² = 49.00
B	29.0	3.9	29.0 x 3.9 = 113.10	(29.0) ² = 841.00	(3.9) ² = 15.21
C	20.8	2.1	20.8 x 2.1 = 43.68	(20.8) ² = 432.64	(2.1) ² = 4.41
D	19.1	2.8	19.1 x 2.8 = 53.48	(19.1) ² = 364.81	(2.8) ² = 7.84
E	13.4	1.4	13.4 x 1.4 = 18.76	(13.4) ² = 179.56	(1.4) ² = 1.96
F	8.5	1.5	8.5 x 1.5 = 12.75	(8.5) ² = 72.25	(1.5) ² = 2.25
-	$\sum X = 153.8$	$\sum Y = 18.7$	$\sum XY = 682.77$	$\sum X^2 = 5,859.26$	$\sum Y^2 = 80.67$

STEP 3: Substitute in the formula and solve for r .

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

$$= \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{[(6)(5859.26) - (153.8)^2][(6)(80.67) - (18.7)^2]}} = 0.982$$

The correlation coefficient suggests a strong relationship between the number of cars a rental agency has and its annual revenue.

Example 2: Compute the value of the correlation coefficient for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class.

The data are shown here.

Student	Number of absences (X)	Final grade % (Y)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

Solution

STEP 1: Add a table for XY , X^2 and Y^2 .

STEP 2: Find the values of XY , X^2 and Y^2 ; place these values in the corresponding columns of the table.

Company	Number of absences (X)	Final grade % (Y)	XY	X^2	Y^2
A	6	82	$6 \times 82 = 492$	$(6)^2 = 36$	$(82)^2 = 6,724$
B	2	86	$2 \times 86 = 172$	$(2)^2 = 4$	$(86)^2 = 7,396$
C	15	43	$15 \times 43 = 645$	$(15)^2 = 225$	$(43)^2 = 1,849$
D	9	74	$9 \times 74 = 666$	$(9)^2 = 81$	$(74)^2 = 5,476$
E	12	58	$12 \times 58 = 696$	$(12)^2 = 144$	$(58)^2 = 3,364$
F	5	90	$5 \times 90 = 450$	$(5)^2 = 25$	$(90)^2 = 8,100$
G	8	78	$8 \times 78 = 624$	$(8)^2 = 64$	$(78)^2 = 6,084$
-	$\sum X = 57$	$\sum Y = 511$	$\sum XY = 3745$	$\sum X^2 = 579$	$\sum Y^2 = 38,993$

STEP 3: Substitute in the formula and solve for r .

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

$$= \frac{(7)(3745) - (57)(511)}{\sqrt{[(7)(579) - (57)^2][(7)(38,993) - (511)^2]}} = -0.944$$

The value of r suggests a strong negative relationship between a student's final grade and the number of absences a student has. That is, the more absences a student has, the lower is his or her grade.

Example 3: A researcher wishes to see if there is a relationship between the ages and net worth of the wealthiest people in America. The data for a specific year are shown.

Person	Age (X)	Net wealth (Y) – in billion \$
A	73	16
B	65	26
C	53	50
D	54	21.5
E	79	40
F	69	16
G	61	19.6
H	65	19

Source: Forbes magazine.

Solution

STEP 1: Add a table for XY , X^2 and Y^2 .

STEP 2: Find the values of XY , X^2 and Y^2 ; place these values in the corresponding columns of the table.

Company	Age (X)	Net wealth (Y)	XY	X^2	Y^2
A	73	16	$73 \times 16 = 1,168$	$(73)^2 = 5,329$	$(16)^2 = 256$
B	65	26	$65 \times 26 = 1,690$	$(65)^2 = 4,225$	$(26)^2 = 676$
C	53	50	$53 \times 50 = 2,650$	$(53)^2 = 2,809$	$(50)^2 = 2,500$
D	54	21.5	$54 \times 21.5 = 1,161$	$(54)^2 = 2,916$	$(21.5)^2 = 462.25$
E	79	40	$79 \times 40 = 3,160$	$(79)^2 = 6,241$	$(40)^2 = 1,600$
F	69	16	$69 \times 16 = 1,104$	$(69)^2 = 4,761$	$(16)^2 = 256$
G	61	19.6	$61 \times 19.6 = 1,195.60$	$(61)^2 = 3,721$	$(19.6)^2 = 384.16$
H	65	19	$65 \times 19 = 1,235$	$(65)^2 = 4,225$	$(19)^2 = 361$
-	$\sum_{i=1}^8 X = 519$	$\sum_{i=1}^8 Y = 208.1$	$\sum XY = 13,363.6$	$\sum X^2 = 34,227$	$\sum Y^2 = 6,495.41$

STEP 3: Substitute in the formula and solve for r .

$$\begin{aligned}
 r &= \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} \\
 &= \frac{8(13,363.6) - (519)(208.1)}{\sqrt{[8(34,227) - (519)^2][8(6,495.41) - (208.1)^2]}} \\
 &= \frac{-1095.1}{\sqrt{(4455)(8657.67)}} \\
 &= \frac{-1095.1}{6210.469} \\
 &= -0.176
 \end{aligned}$$

The value of r indicates a very weak negative relationship between the variables.

In Example 1, the value of r was high (close to 1.00); in Example 3, the value of r was much lower (close to 0). This question then arises, when is the value of r due to chance, and

when does it suggest a significant linear relationship between the variables? This question will be answered next.

Big Picture in Focus

ULO – c. Test the hypothesis $H_0: \rho = 0$.

The Significance of the Correlation Coefficient

As stated before, the range of the correlation coefficient is between -1 and $+1$. When the value of r is near $+1$ or -1 , there is a strong linear relationship. When the value of r is near 0 , the linear relationship is weak or non-existent. Since the value of r is computed from data obtained from samples, there are two possibilities when r is not equal to zero: either the value of r is high enough to conclude that there is a significant linear relationship between the variables, or the value of r is due to chance.

To make this decision, you use a hypothesis-testing procedure. The traditional method is like the one used in previous chapters.

STEP 1: State the hypotheses.

STEP 2: Find the critical values.

STEP 3: Compute the test value.

STEP 4: Make the decision.

STEP 5: Summarize the results.

The population correlation coefficient is computed from taking all possible (x, y) pairs; it is designated by the Greek letter ρ (rho). The sample correlation coefficient can then be used as an estimator of ρ if the following assumptions are valid.

1. The variables x and y are *linearly* related.
2. The variables are *random* variables.
3. The two variables have a *bivariate normal distribution*.

A bivariate normal distribution means that for the pairs of (x, y) data values, the corresponding y values have a bell-shaped distribution for any given x value, and the x values for any given y value have a bell-shaped distribution.

Formally defined, the **population correlation coefficient** ρ is the correlation computed by using all possible pairs of data values (x, y) taken from a population.

In hypothesis testing, one of these is true:

$H_0: \rho = 0$ This null hypothesis means that there is no correlation between the x and y variables in the population.

$H_1: \rho \neq 0$ This alternative hypothesis means that there is a significant correlation between the variables in the population.

When the null hypothesis is rejected at a specific level, it means that there is a significant difference between the value of r and 0. When the null hypothesis is not rejected, it means that the value of r is not significantly different from 0 (zero) and is probably due to chance.

Formula for the t Test for the Correlation Coefficient

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

with degrees of freedom equal to $n - 2$.

Although hypothesis tests can be one-tailed, most hypotheses involving the correlation coefficient are two-tailed. Recall that ρ represents the population correlation coefficient. Also, if there is no linear relationship, the value of the correlation coefficient will be 0. Hence, the hypotheses will be

$$H_0: \rho = 0 \quad \text{and} \quad H_1: \rho \neq 0$$

You do not have to identify the claim here, since the question will always be whether there is a significant linear relationship between the variables.

The two-tailed critical values are used. These values are found in Table B in Appendix. Also, when you are testing the significance of a correlation coefficient, both variables x and y must come from normally distributed populations.

Example 1: Test the significance of the correlation coefficient for the car rental companies in the United States for a recent year. Use $\alpha = 0.05$.

Solution

STEP 1: Recall the computed r value in the previous example then state the hypotheses.

$$H_0: \rho = 0 \quad \text{and} \quad H_1: \rho \neq 0$$

STEP 2: Find the critical values. Since $\alpha = 0.05$ and there are $6 - 2 = 4$ degrees of freedom, the critical values obtained from Table B are ± 2.776 , as shown in Figure 11–7.

	Confidence intervals	80%	90%	95%	98%	99%
	One tail, α	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, α	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032

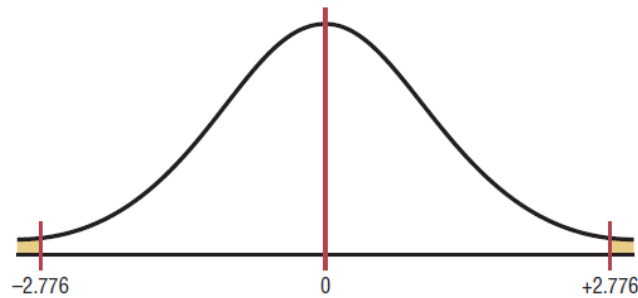


Figure 12-7

STEP 3: Compute the test value.

$$t = r \sqrt{\frac{n-2}{1-r^2}} = 0.982 \sqrt{\frac{6-2}{1-(0.982)^2}} = 10.4$$

STEP 4: Make the decision. Reject the null hypothesis, since the test value falls in the critical region, as shown in Figure 11-8.

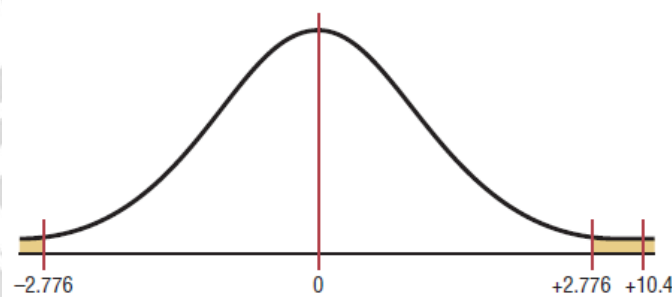


Figure 12-8

STEP 5: Summarize the results. There is a significant relationship between the number of cars a rental agency owns and its annual income.

SELF-HELP: You can also refer to the sources below to help you understand the lesson.

1. Bluman, A. (2012). *Elementary Statistics: A Step by Step Approach 8th Edition*. McGraw-Hill Companies, Inc.

LET'S CHECK

ACTIVITY 1

Now that you know the most essential concepts on correlation, let us try to check your understanding of these concepts.

1. A graduate student in developmental psychology believes there may be a relationship between birth weight and subsequent IQ. She randomly samples seven psychology majors at her university and gives them an IQ test. Next, she obtains the weight at birth of the seven majors from the appropriate hospitals (after obtaining permission from the students, of course). The data are shown in the following table.

Student	1	2	3	4	5	6	7
Birth Weight	5.8	6.5	8.0	5.9	8.5	7.2	9.0
IQ	122	120	129	112	127	116	130

- a. Construct a scatter plot of the data, plotting birth weight on the X axis and IQ on the Y axis. Does the relationship appear to be linear?
 - b. Assume the relationship is linear and compute the value of Pearson r .
2. A researcher conducts a study to investigate the relationship between cigarette smoking and illness. The number of cigarettes smoked daily, and the number of days absent from work in the last year due to illness is determined for 12 individuals employed at the company where the researcher works. The scores are given in the following table.

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Cigarettes Smoked	0	0	0	10	13	20	27	35	35	44	53	60
Days Absent	1	3	8	10	4	14	5	6	12	16	10	16

- a. Construct a scatter plot for these data. Does the relationship look linear?
- b. Calculate the value of Pearson r .
- c. Eliminate the data from subjects 1, 2, 3, 10, 11, and 12. This decreases the range of both variables. Recalculate r for the remaining subjects. What effect does decrease the range have on r ?

LET'S ANALYZE

ACTIVITY 1

Getting acquainted with the essential terms and concepts on correlation, it is now time for you to explain thoroughly your answers to the following questions.

1. Discuss the different kinds of relationship that are possible between two variables.

2. A study has shown that the correlation between fatigue and irritability is 0.53. Based on this correlation, the author concludes that fatigue is an important factor in producing irritability. Is this conclusion justified? Explain.

3. Is it possible to have the value of r greater than 2? Why or why not?

4. Explain the role of coefficient correlation in testing the hypothesis.

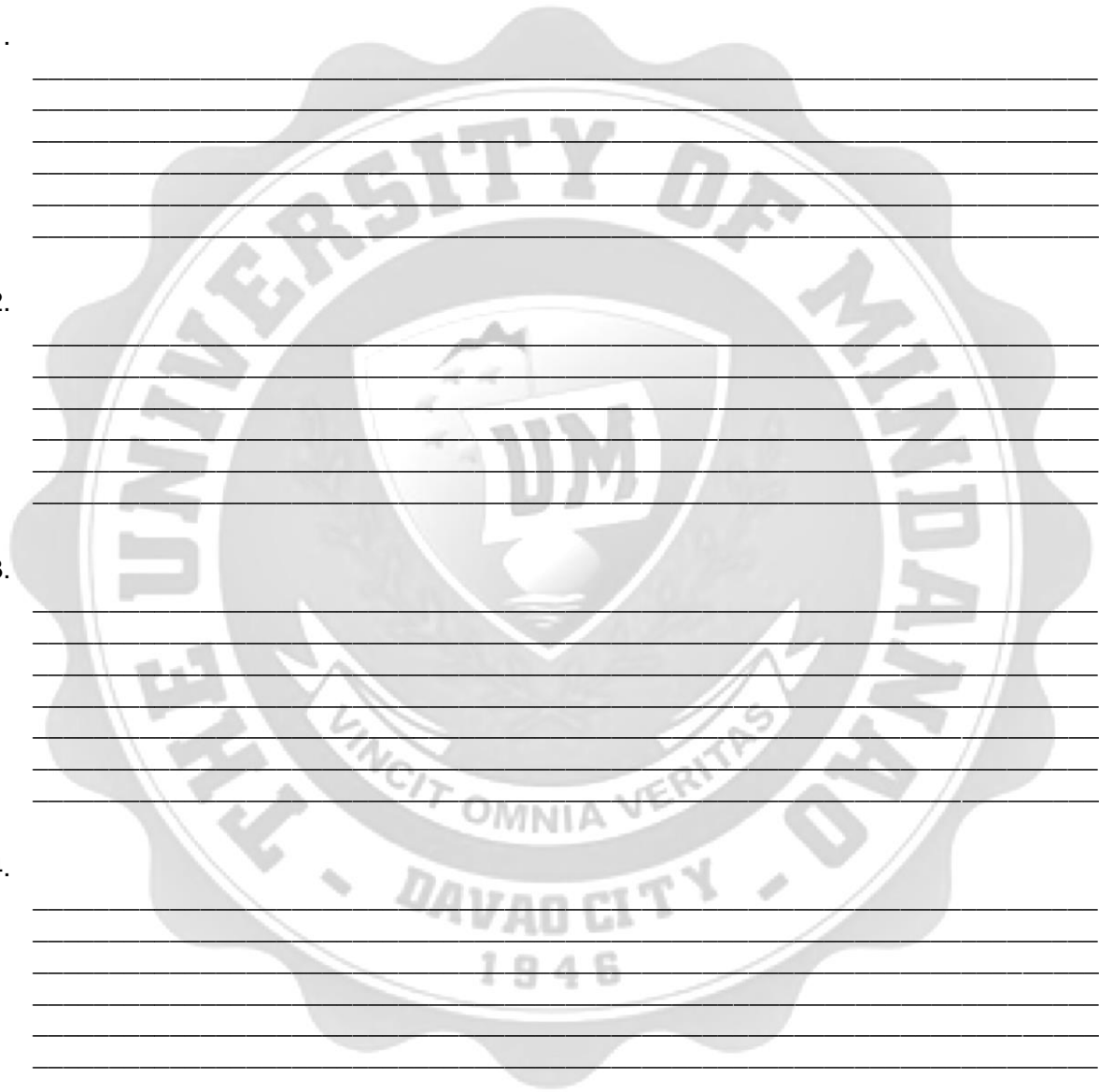
5. How does correlation differ to all statistical tools discussed in the previous chapters?

IN A NUTSHELL

ACTIVITY 1

Based on the concept on correlation and the learning exercises that you have done, write your arguments or lessons learned below.

1.



2.

3.

4.

5.

Q & A LIST

Do you have any question for clarification?	
Questions / Issues	Answers
1.	1.
2.	2.
3.	3.
4.	4.
5.	5.

KEYWORDS INDEX

Inferential Statistics	Correlation	Scatter plot
Coefficient Correlation	Simple relationship	P-value
Positive relationship	Negative relationship	Non-existence relationship

Chapter 13: REGRESSION

Big Picture in Focus

ULO – d. Compute the equation of the regression line.

METALANGUAGE

To demonstrate **ULO-d**, essential terms relevant to the study of regression are already defined in chapter 12. Please refer to those operational definition as to how the texts work and to establish common frame of reference.

ESSENTIAL KNOWLEDGE

In studying relationships between two variables, collect the data and then construct a scatter plot. The purpose of the scatter plot, as indicated previously, is to determine the nature of the relationship. The possibilities include a positive linear relationship, a negative linear relationship, a curvilinear relationship, or no discernible relationship. After the scatter plot is drawn, the next steps are to compute the value of the correlation coefficient and to test the significance of the relationship. If the value of the correlation coefficient is significant, the next step is to determine the equation of the **regression line**, which is the data's line of best fit. (Note: Determining the regression line when r is not significant and then making predictions using the regression line is meaningless.) The purpose of the regression line is to enable the researcher to see the trend and make predictions based on the data.

Line of Best Fit

Figure 12–1 shows a scatter plot for the data of two variables. It shows that several lines can be drawn on the graph near the points. Given a scatter plot, you must be able to draw the **line of best fit**. **Best fit** means that the sum of the squares of the vertical distances from each point to the line is at a minimum. The reason you need a line of best fit is that the values of y will be predicted from the values of x ; hence, the closer the points are to the line, the better the fit and the prediction will be. See Figure 12-2. When r is positive, the line slopes upward and to the right. When r is negative, the line slopes downward from left to right.

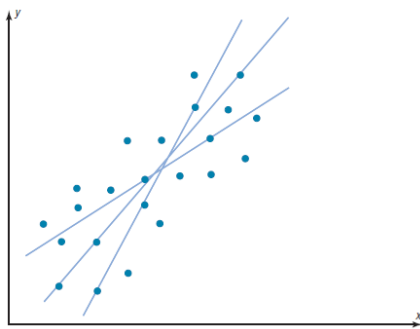


Figure 13 – 1

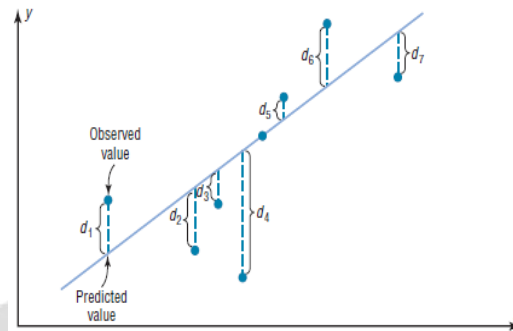


Figure 13 – 2

Determination of the Regression Line Equation

In algebra, the equation of a line is usually given as $y = mx + b$, where m is the slope of the line and b is the y intercept. In statistics, the equation of the regression line is written as $y' = a + bx$, where a is the y intercept and b is the slope of the line. See Figure 12–3.

There are several methods for finding the equation of the regression line. Two formulas are given here. *These formulas use the same values that are used in computing the value of the correlation coefficient.* The mathematical development of these formulas is beyond the scope of this module.

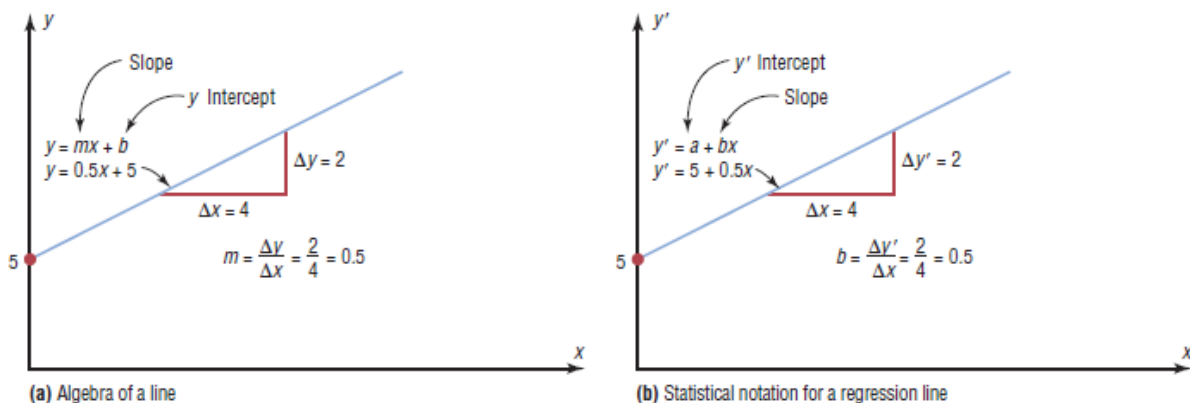


Figure 13-3

Formulas for the Regression Line $y' = a + bx$

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

where a is the y' intercept and b is the slope of the line.



Rounding Rule for the Intercept and Slope

Round the values of a and b to three decimal places.

Example 1: Find the equation of the regression line using the data found in the car rental companies in the United States for a recent year.

Solution

Recall the given data found in the previous problems in correlation.

Company	Car (X)	Revenue (Y)	XY	X ²	Y ²
A	63.0	7.0	63.0 x 7.0 = 441.00	(63.0) ² = 3969.0	(7.0) ² = 49.00
B	29.0	3.9	29.0 x 3.9 = 113.10	(29.0) ² = 841.00	(3.9) ² = 15.21
C	20.8	2.1	20.8 x 2.1 = 43.68	(20.8) ² = 432.64	(2.1) ² = 4.41
D	19.1	2.8	19.1 x 2.8 = 53.48	(19.1) ² = 364.81	(2.8) ² = 7.84
E	13.4	1.4	13.4 x 1.4 = 18.76	(13.4) ² = 179.56	(1.4) ² = 1.96
F	8.5	1.5	8.5 x 1.5 = 12.75	(8.5) ² = 72.25	(1.5) ² = 2.25
-	$\sum X = 153.8$	$\sum Y = 18.7$	$\sum XY = 682.77$	$\sum X^2 = 5,859.26$	$\sum Y^2 = 80.67$

The values needed for the equation are $n = 6$, $\sum X = 153.8$, $\sum Y = 18.7$, $\sum XY = 682.77$, and $\sum X^2 = 5,859.26$. Substituting in the formulas, you get

$$a = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{n(\sum X^2) - (\sum X)^2} = \frac{(18.7)(5859.26) - (153.8)(682.77)}{(6)(5859.26) - (153.8)^2} = 0.396$$

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2} = \frac{6(682.77) - (153.8)(18.7)}{(6)(5859.26) - (153.8)^2} = 0.106$$

Hence, the equation of the regression line $y' = a + bx$ is

$$y' = 0.396 + 0.106x$$

To graph the line, select any two points for x and find the corresponding values for y . Use any x values between 10 and 60. For example, let $x = 15$. Substitute in the equation and find the corresponding y' value.

$$y' = 0.396 + 0.106x$$

$$y' = 0.396 + 0.106(15)$$

$$y' = 1.986$$

Let $x = 40$; then

$$y' = 0.396 + 0.106x$$

$$y' = 0.396 + 0.106(15)$$

$$y' = 4.636$$

Then plot the two points (15, 1.986) and (40, 4.636) and draw a line connecting the two points. See Figure 12–4.

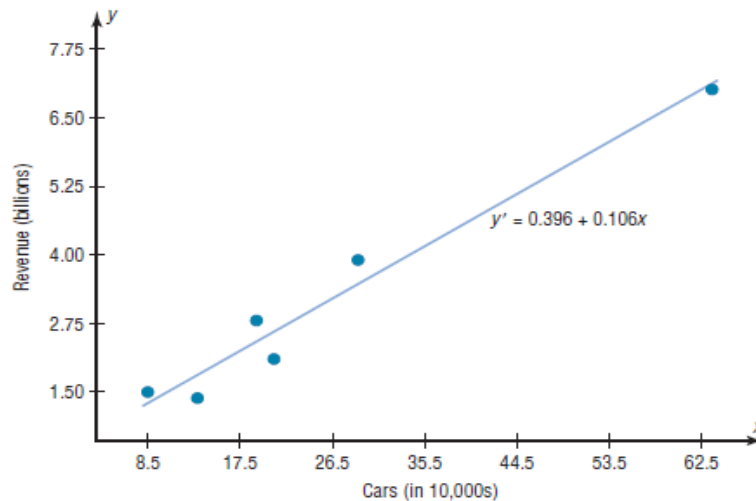


Figure 13-4

Example 2: Find the equation of the regression line using the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class.

Solution

Recall the given data found in the previous problems in correlation.

Company	Number of absences (X)	Final grade % (Y)	XY	X^2	Y^2
A	6	82	$6 \times 82 = 492$	$(6)^2 = 36$	$(82)^2 = 6,724$
B	2	86	$2 \times 86 = 172$	$(2)^2 = 4$	$(86)^2 = 7,396$
C	15	43	$15 \times 43 = 645$	$(15)^2 = 225$	$(43)^2 = 1,849$
D	9	74	$9 \times 74 = 666$	$(9)^2 = 81$	$(74)^2 = 5,476$
E	12	58	$12 \times 58 = 696$	$(12)^2 = 144$	$(58)^2 = 3,364$
F	5	90	$5 \times 90 = 450$	$(5)^2 = 25$	$(90)^2 = 8,100$
G	8	78	$8 \times 78 = 624$	$(8)^2 = 64$	$(78)^2 = 6,084$
-	$\sum X = 57$	$\sum Y = 511$	$\sum XY = 3745$	$\sum X^2 = 579$	$\sum Y^2 = 38,993$

The values need for the equation are $n = 7$, $\sum X = 57$, $\sum Y = 511$, $\sum XY = 3745$, and $\sum X^2 = 579$. Substituting in the formulas, you get

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} = \frac{(511)(579) - (57)(3745)}{(7)(579) - (57)^2} = 102.493$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{(7)(3745) - (57)(511)}{(7)(579) - (57)^2} = -3.622$$

Hence, the equation of the regression line $y' = a + bx$ is

$$y' = 102.493 - 3.622x$$

The graph of the line is shown in Figure 12-5.

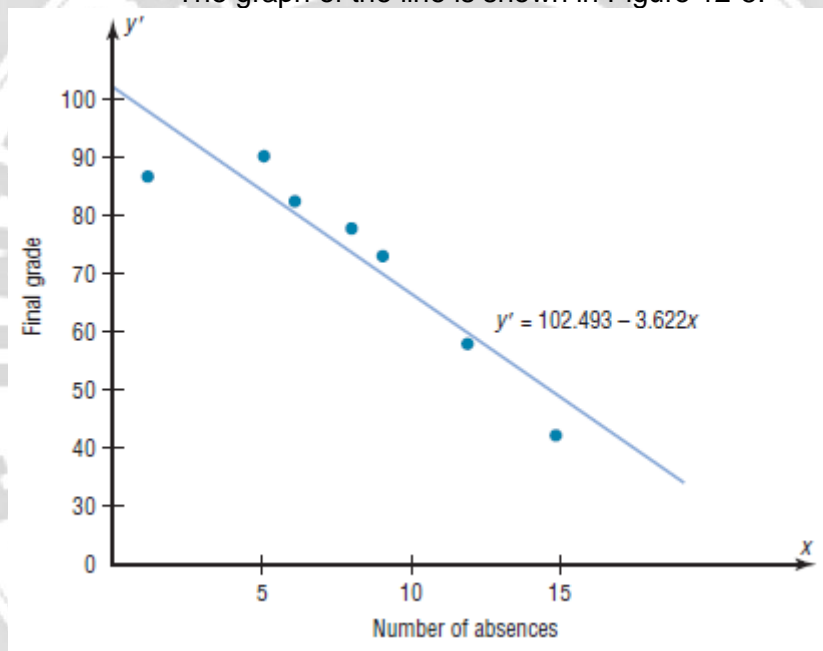


Figure 13-5

The regression line can be used to make predictions for the dependent variable. The method for making predictions is shown in Example 3.

The sign of the correlation coefficient and the sign of the slope of the regression line will always be the same. That is, if r is positive, then b will be positive; if r is negative, then b will be negative. The reason is that the numerators of the formulas are the same and determine the signs of r and b , and the denominators are always positive. The regression line will always pass through the point whose x coordinate is the mean of the x values and whose y coordinate is the mean of the y values, that is, (\bar{x}, \bar{y}) .

Example 3: Use the equation of the regression line to predict the income of a car rental agency that has 200,000 automobiles.

Solution

Since the x values are in 10,000s, divide 200,000 by 10,000 to get 20, and then substitute 20 for x in the equation.

$$y' = 0.396 + 0.106x$$

$$y' = 0.396 + 0.106(20)$$

$$y' = 2.516$$

Hence, when a rental agency has 200,000 automobiles, its revenue will be approximately \$2.516 billion.

The value obtained in Example 3 is a point prediction, and with point predictions, no degree of accuracy or confidence can be determined. The magnitude of the change in one variable when the other variable changes exactly 1 unit is called a **marginal change**. The value of slope b of the regression line equation represents the marginal change. For example, in Example 1 the slope of the regression line is 0.106, which means for each increase of 10,000 cars, the value of y changes 0.106 unit (\$106 million) on average.

Procedure Table

Finding the Correlation Coefficient and the Regression Line Equation

Step 1 Make a table, as shown in step 2.

Step 2 Find the values of xy , x^2 , and y^2 . Place them in the appropriate columns and sum each column.

x	y	xy	x^2	y^2
.
.
.
$\Sigma x =$	$\Sigma y =$	$\Sigma xy =$	$\Sigma x^2 =$	$\Sigma y^2 =$

Step 3 Substitute in the formula to find the value of r .

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

Step 4 When r is significant, substitute in the formulas to find the values of a and b for the regression line equation $y' = a + bx$.

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} \quad b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

SELF-HELP:

You can also refer to the sources below to help you understand the lesson.

1. Bluman, A. (2012). *Elementary Statistics: A Step by Step Approach 8th Edition*. McGraw-Hill Companies, Inc.



LET'S CHECK

ACTIVITY 1

Now that you know the most essential concepts on regression, let us try to check your understanding of these concepts.

A clinical psychologist is interested in the relationship between testosterone level in married males and the quality of their marital relationship. A study is conducted in which the testosterone levels of eight married men are measured. The eight men also fill out a standardized questionnaire assessing quality of marital relationship. The questionnaire scale is 0–25, with higher numbers indicating better relationships. Testosterone scores are in nanomoles / liter of serum. The data are shown below.

Subject Number	1	2	3	4	5	6	7	8
Relationship Score	24	15	15	10	19	11	20	19
Testosterone Level	12	13	19	25	9	16	15	21

- On a piece of graph paper, construct a scatter plot of the data. Use testosterone level as the X variable.
- Describe the relationship shown on the graph.
- Compute the value of Pearson r .
- Determine the least-squares regression line for predicting relationship score from testosterone level. Should b_Y be positive or negative? Why?
- Draw the least-squares regression line of part **d** on the scatter plot of part **a**.
- Based on the data of the eight men, what relationship score would you predict for a male who has a testosterone level of 23 nanomoles/liter of serum?

LET'S ANALYZE

ACTIVITY 1

Getting acquainted with the essential terms and concepts on regression, it is now time for you to explain thoroughly your answers to the following questions.

1. How does least-squares regression line is used for prediction?

2. Why the regression line of Y on X is not the same as the regression line of X on Y ? Explain.

3. How regression does is related to correlation?


4. Can regression still be used even if there is no relationship found between two variables?

IN A NUTSHELL

ACTIVITY 1

Based on the concept on regression and the learning exercises that you have done, write your arguments or lessons learned below.

1.



2.

3.

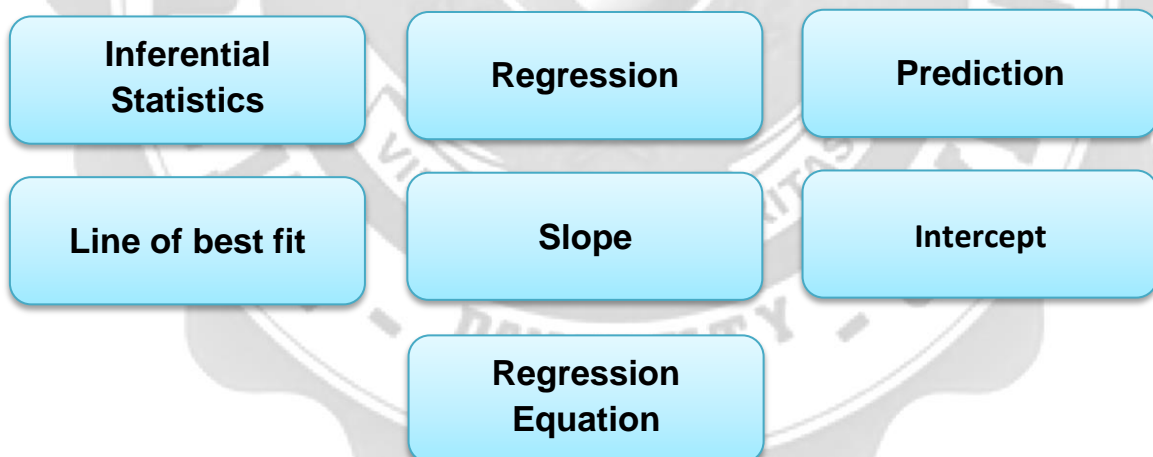
4.

5.

Q & A LIST

Do you have any question for clarification?	
Questions / Issues	Answers
1.	1.
2.	2.
3.	3.
4.	4.
5.	5.

KEYWORDS INDEX



ONLINE CODE OF CONDUCT

- 1) All teachers/Course Facilitators and students are expected to abide by an honor code of conduct, and thus everyone and all are exhorted to exercise self-management and self-regulation.
- 2) All students are likewise guided by professional conduct as learners in attending OBD or DED courses. Any breach and violation shall be dealt with properly under existing guidelines, specifically in Section 7 (Student Discipline) in the Student Handbook.
- 3) Professional conduct refers to the embodiment and exercise of the University's Core Values, specifically in the adherence to intellectual honesty and integrity; academic excellence by giving due diligence in virtual class participation in all lectures and activities, as well as fidelity in doing and submitting performance tasks and assignments; personal discipline in complying with all deadlines; and observance of data privacy.
- 4) Plagiarism is a serious intellectual crime and shall be dealt with accordingly. The University shall institute monitoring mechanisms online to detect and penalize plagiarism.
- 5) Students shall independently and honestly take examinations and do assignments unless collaboration is clearly required or permitted. Students shall not resort to dishonesty to improve the result of their assessments (e.g. examinations, assignments).
- 6) Students shall not allow anyone else to access their personal LMS account. Students shall not post or share their answers, assignment or examinations to others to further academic fraudulence online.
- 7) By enrolling in OBD or DED courses, students agree and abide by all the provisions of the Online Code of Conduct, as well as all the requirements and protocols in handling online courses.

Course prepared by:

ROSYL S. MATIN-AO, MAT
Faculty

Course reviewed by:

RONNIE O. ALEJAN, MSAM
Program Head

Approved by:

KHRISTINE MARIE D. CONCEPCION, Ph.D
Dean, CASE