V SEMESTER- BCA

UNIT- IV
Statistical Inference and Hypothesis Testing

What is Hypothesis?

A hypothesis is an assumption that is made based on some evidence. This is the initial point of any investigation that translates the research questions into predictions. It includes components like variables, population and the relation between the variables.

Example

The field of business, decision makers are continually attempting to find answers to questions such as the following:

■ What container shape is most economical and reliable for shipping a product?

■ Which management approach best motivates employees in the retail industry?

■ What is the best way to link client databases for fast retrieval of useful information?

■ Which indicator best predicts the general state of the economy in the next six months?

■ What is the most effective means of advertising in a business-to-business setting?

Hypotheses are tentative explanations of a principle operating in nature.*

Types of Hypotheses

Three types of hypotheses that will be explored here:

1. Research hypotheses

2. Statistical hypotheses

3. Substantive hypotheses.

Research Hypotheses

Research hypotheses are most nearly like hypotheses defined earlier. A research hypothesis is a statement of what the researcher believes will be the outcome of an experiment or a study.Before studies are undertaken, business researchers often have some idea or theory based on experience or previous work as to how the study will turn out. These ideas, theories, or notions established before an experiment or study is conducted are research hypotheses. Some examples of research hypotheses in business might include:

■ Older workers are more loyal to a company.

■ Companies with more than $1 billion in assets spend a higher percentage of their annual budget on advertising than do companies with less than $1 billion in assets.

■ The price of scrap metal is a good indicator of the industrial production index six months later.

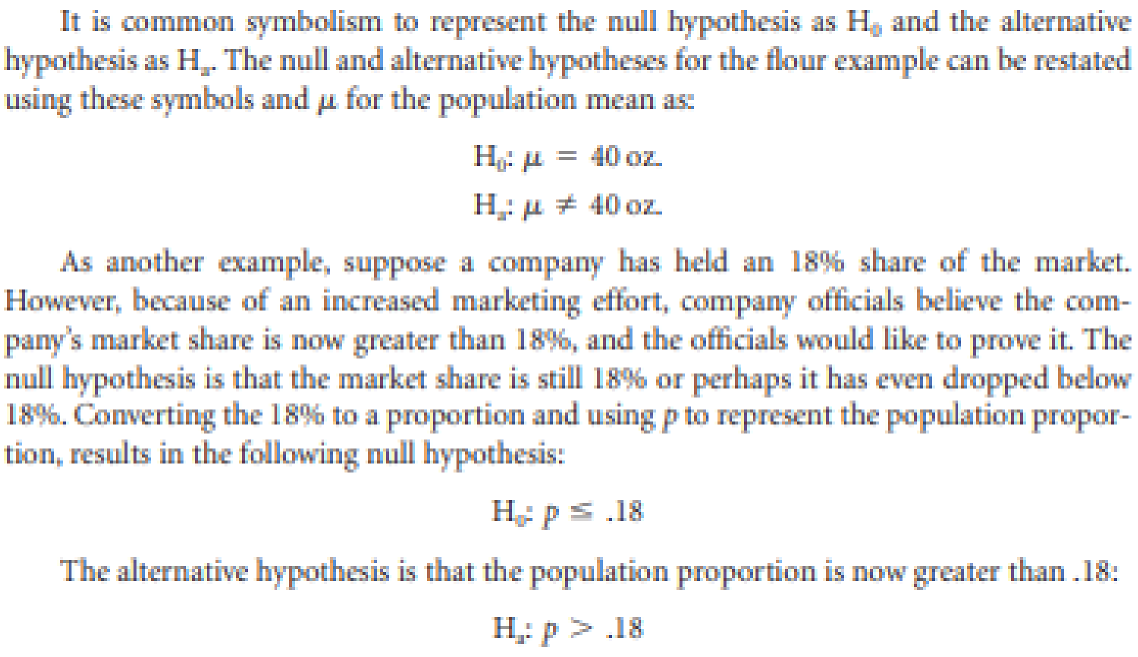
Statistical Hypotheses

In order to scientifically test research hypotheses, a more formal hypothesis structure needs to be set up using statistical hypotheses. Suppose business researchers want to "prove" the research hypothesis that older workers are more loyal to a company. A "loyalty" survey instrument is either developed or obtained. If this instrument is administered to both older and younger workers, how much higher do older workers have to score on the "loyalty" instrument (assuming higher scores indicate more loyal) than younger workers to prove the research hypothesis? What is the "proof threshold"? Instead of attempting to prove or disprove research hypotheses directly in this manner, business researchers convert their research hypotheses to statistical hypotheses and then test the statistical hypotheses using standard procedures.

All statistical hypotheses consist of two parts, a null hypothesis and an alternative hypothesis. These two parts are constructed to contain all possible outcomes of the experiment or study. **Generally, the null hypothesis states that the "null" condition exists; that is, there is nothing new happening, the old theory is still true, the old standard is correct, and the system is in control**.

The alternative hypothesis, on the other hand, states that the new theory is true, there are new standards, the system is out of control, and/or something is happening.

As an example, suppose flour packaged by a manufacturer is sold by weight; and a particular size of package is supposed to average 40 ounces. Suppose the manufacturer wants to test to determine whether their packaging process is out of control as determined by the weight of the flour packages. The null hypothesis for this experiment is that the average weight of the flour packages is 40 ounces (no problem). The alternative hypothesis is that the average is not 40 ounces (process is out of control).



It is common symbolism to represent the null hypothesis as H_0 and the alternative hypothesis as H_a. The null and alternative hypotheses for the flour example can be restated using these symbols and μ for the population mean as:

$$H_0: \mu = 40 \text{ oz.}$$
$$H_a: \mu \neq 40 \text{ oz.}$$

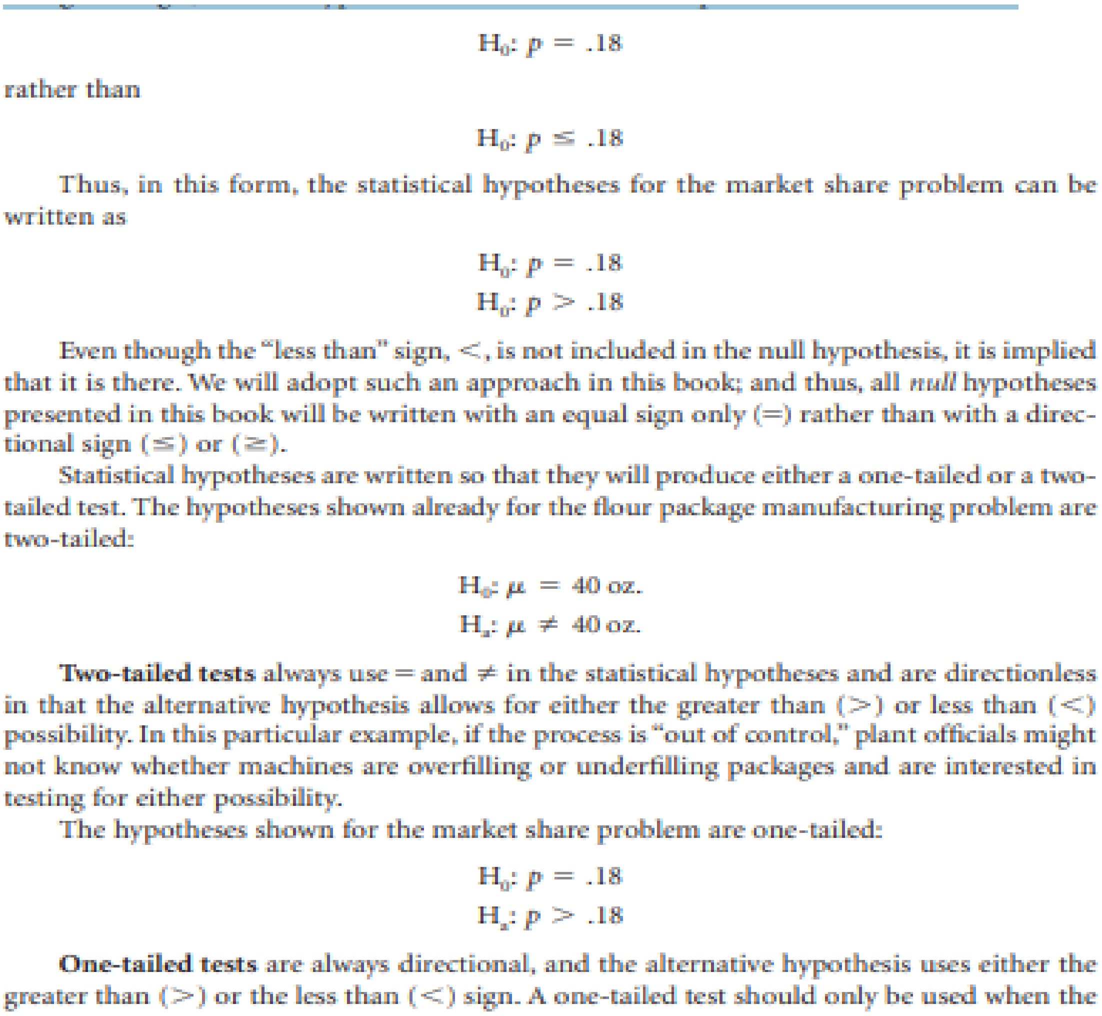
As another example, suppose a company has held an 18% share of the market. However, because of an increased marketing effort, company officials believe the company's market share is now greater than 18%, and the officials would like to prove it. The null hypothesis is that the market share is still 18% or perhaps it has even dropped below 18%. Converting the 18% to a proportion and using p to represent the population proportion, results in the following null hypothesis:

$$H_0: p \leq .18$$

The alternative hypothesis is that the population proportion is now greater than .18:

$$H_a: p > .18$$

Note that the "new idea" or "new theory" that company officials want to "prove" is stated in the alternative hypothesis. The null hypothesis states that the old market share of 18% is still true..



$$H_0: p = .18$$

rather than

$$H_0: p \leq .18$$

Thus, in this form, the statistical hypotheses for the market share problem can be written as

$$H_0: p = .18$$
$$H_0: p > .18$$

Even though the "less than" sign, <, is not included in the null hypothesis, it is implied that it is there. We will adopt such an approach in this book; and thus, all *null* hypotheses presented in this book will be written with an equal sign only (=) rather than with a directional sign (≤) or (≥).

Statistical hypotheses are written so that they will produce either a one-tailed or a two-tailed test. The hypotheses shown already for the flour package manufacturing problem are two-tailed:

$$H_0: \mu = 40 \text{ oz.}$$
$$H_a: \mu \neq 40 \text{ oz.}$$

Two-tailed tests always use = and ≠ in the statistical hypotheses and are directionless in that the alternative hypothesis allows for either the greater than (>) or less than (<) possibility. In this particular example, if the process is "out of control," plant officials might not know whether machines are overfilling or underfilling packages and are interested in testing for either possibility.

The hypotheses shown for the market share problem are one-tailed:

$$H_0: p = .18$$
$$H_a: p > .18$$

One-tailed tests are always directional, and the alternative hypothesis uses either the greater than (>) or the less than (<) sign. A one-tailed test should only be used when the

In business research, the conservative approach is to conduct a two-tailed test because sometimes study results can be obtained that are in opposition to the direction that researchers thought would occur. For example, in the market share problem, it might turn out that the company had actually lost market share; and even though company officials were not interested in "proving" such a case, they may need to know that it is true. It is recommended that, if in doubt, business researchers should use a two-tailed test.

Substantive Hypotheses

In testing a statistical hypothesis, a business researcher reaches a conclusion based on the data obtained in the study. If the null hypothesis is

rejected and therefore the alternative hypothesis is accepted, it is common to say that a statistically significant result has been obtained.

For example, in the market share problem, if the null hypothesis is rejected, the result is that the market share is “significantly greater” than 18%. The word significant to statisticians and business researchers merely means that the result of the experiment is unlikely due to chance and a decision has been made to reject the null hypothesis. However, in everyday business life, the word significant is more likely to connote “important” or “a large amount.” One problem that can arise in testing statistical hypotheses is that particular characteristics of the data can result in a statistically significant outcome that is not a significant business outcome.

Type I and Type II Errors

Because the hypothesis testing process uses sample statistics calculated from random data to reach conclusions about population parameters, it is possible to make an incorrect decision about the null hypothesis. In particular, two types of errors can be made in testing hypotheses: Type I error and Type II error.

A Type I error is committed by rejecting a true null hypothesis. With a Type I error, the null hypothesis is true, but the business researcher decides that it is not.

As an example, suppose the flour-packaging process actually is “in control” and is averaging 40 ounces of flour per package. Suppose also that a business researcher randomly selects 100 packages, weighs the contents of each, and computes a sample mean. It is possible, by chance, to randomly select 100 of the more extreme packages (mostly heavy weighted or mostly light weighted) resulting in a mean that falls in the rejection region. The decision is to reject the null hypothesis even though

the population mean is actually 40 ounces. In this case, the business researcher has committed a Type I error.

The notion of a Type I error can be used outside the realm of statistical hypothesis testing in the business world**. For example, if a manager fires an employee because some evidence indicates that she is stealing from the company and if she really is not stealing from the company, then the manager has committed a Type I error.**

As another example, suppose a worker on the assembly line of a large manufacturer hears an unusual sound and decides to shut the line down (reject the null hypothesis). If the sound turns out not to be related to the assembly line and no problems are occurring with the assembly line, then the worker has committed a Type I error.

The probability of committing a Type I error is called alpha (α) or level of significance. Alpha equals the area under the curve that is in the rejection region beyond the critical value(s). The value of alpha is always set before the experiment or study is undertaken. As mentioned previously, common values of alpha are .05, .01, .10, and .001.

A Type II error is committed when a business researcher fails to reject a false null hypothesis. In this case, the null hypothesis is false, but a decision is made to not reject it.

Suppose in the case of the flour problem that the packaging process is actually producing a population mean of 41 ounces even though the null hypothesis is 40 ounces. A sample of 100 packages yields a sample mean of 40.2 ounces, which falls in the nonrejection region. The business decision maker decides not to reject the null hypothesis. A Type II error has been committed. The packaging procedure is out of control and the hypothesis testing process does not identify it.

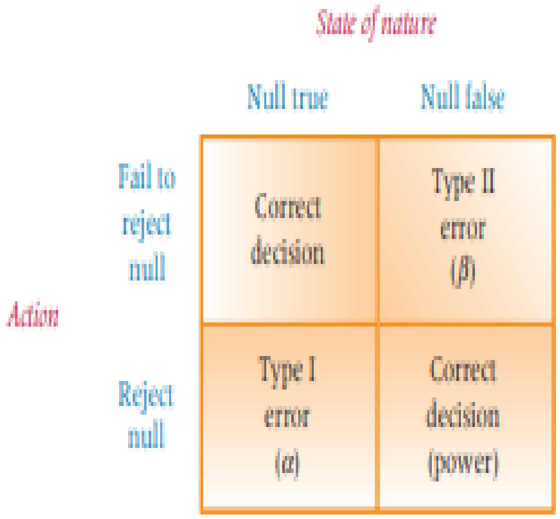
Suppose in the business world an employee is stealing from the company. A manager sees some evidence that the stealing is occurring but lacks enough evidence to conclude that the employee is stealing from the company. The manager decides not to fire the employee based on theft. The manager has committed a Type II error.

Consider the manufacturing line with the noise. Suppose the worker decides not enough noise is heard to shut the line down, but in actuality, one of the cords on the line is unraveling, creating a dangerous situation. The worker is committing a Type II error

The probability of committing a Type II error is beta (β). Unlike alpha, beta is not usually stated at the beginning of the hypothesis testing procedure. Actually, because beta occurs only when the null hypothesis is not true, the computation of beta varies with the many possible alternative parameters that might occur.

FIGURE 9.5

Alpha, Beta, and Power



		State of nature	
		Null true	Null false
Action	Fail to reject null	Correct decision	Type II error (β)
	Reject null	Type I error (α)	Correct decision (power)

Power, which is equal to 1 -β , is the probability of a statistical test rejecting the null hypothesis when the null hypothesis is false. Figure 9.5 shows the relationship between α, β, and power.

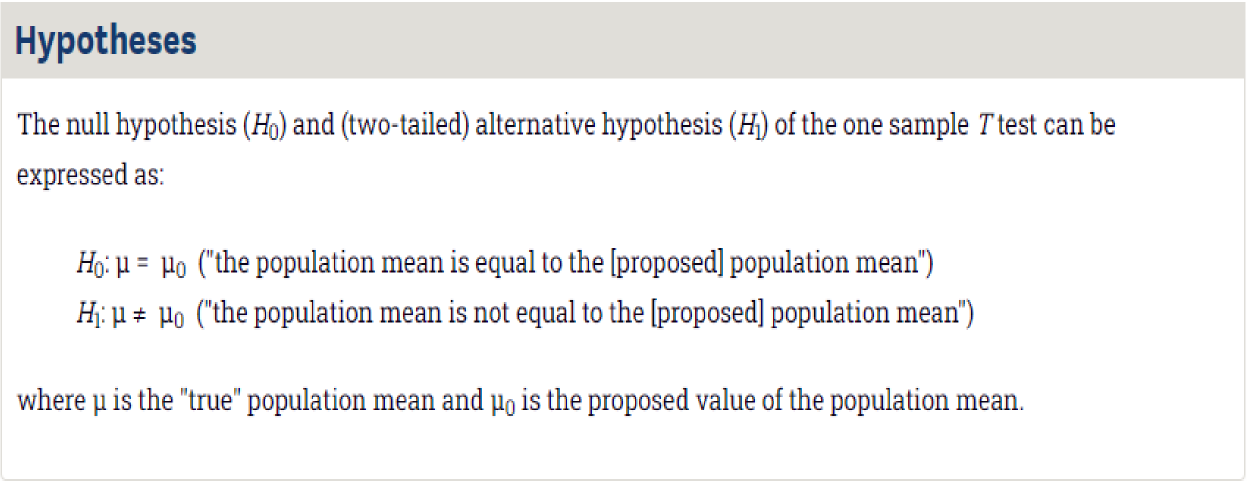
One Sample t Test

The One Sample *t* Test examines whether the mean of a population is statistically different from a known or hypothesized value. The One Sample *t* Test is a parametric test.

Common Uses

The One Sample *t* Test is commonly used to test the following:

- Statistical difference between a mean and a known or hypothesized value of the mean in the population.
- Statistical difference between a change score and zero.
- This approach involves creating a change score from two variables, and then comparing the mean change score to zero, which will indicate whether any change occurred between the two time points for the original measures. If the mean change score is not significantly different from zero, no significant change occurred.



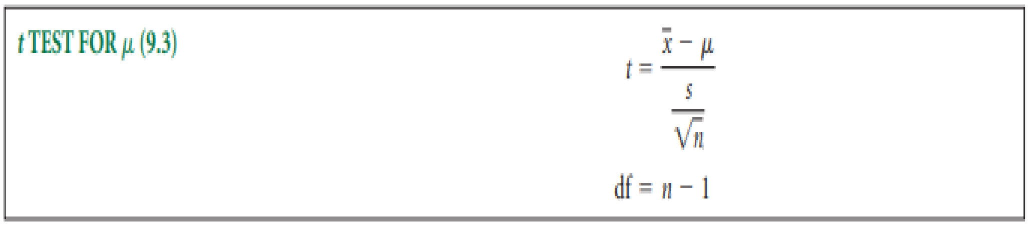
Hypotheses

The null hypothesis (H_0) and (two-tailed) alternative hypothesis (H_1) of the one sample *T* test can be expressed as:

H_0: $\mu = \mu_0$ ("the population mean is equal to the [proposed] population mean")
H_1: $\mu \neq \mu_0$ ("the population mean is not equal to the [proposed] population mean")

where μ is the "true" population mean and μ_0 is the proposed value of the population mean.

The formula for testing such hypotheses follows.



t TEST FOR μ (9.3)

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$\text{df} = n - 1$$

For example, imagine a company wants to test the claim that their batteries last more than 40 hours. Using a simple random sample of 15 batteries yielded a mean of 44.9 hours, with a standard deviation of 8.9 hours. Test this claim using a significance level of 0.05.

$$H_0 : \mu = 40$$
$$H_a : \mu > 40$$

$$\hat{x} = 44.9, \;\; \mu = 40 \;\; s = 8.9, \;\; n = 15, \; df = n - 1 \rightarrow df = 15 - 1 = 14$$

$$test\ statistic: \;\; t = \frac{44.9 - 40}{\left(\frac{8.9}{\sqrt{15}}\right)} = 2.13$$

t 0.05,14=1.761

2.13>1.761 Reject Null Hypothesis

1. A random of sample size 20 is taken resulting in sample mean of 25.51 and a sample standard deviation of 2.1933.Assume data is normally distributed use this information and α =0.05 to test the following hypothesis.

$$H_0: \mu = 25 \text{ pounds}$$
$$H_a: \mu \neq 25 \text{ pounds}$$

The test is to determine whether the machine is out of control, and the shop supervisor has not specified whether he believes the machine is producing plates that are too heavy or too light. Thus a two-tailed test is appropriate. The following hypotheses are tested.

H_0: μ = 25 pounds
H_a: $\mu \neq$ 25 pounds

Anα of .05 is used. Figure 9.11 shows the rejection regions. Because n = 20, the degrees of freedom for this test are 19 (20 - 1). The t distribution table is a one-tailed table but the test for this problem is two tailed, so alpha must be split, which yields α/2 = .025, the value in each tail. (To obtain the table t value when conducting a two-tailed test, always split alpha and use α/2.) The table t value for this example is 2.093. Table values such as this one are often written in the following form:

$t_{.025,19} = 2.093$

Figure 9.12 depicts the t distribution for this example, along with the critical values, the observed t value, and the rejection regions. In this case, the decision rule is to reject the

FIGURE 9.11

Rejection Regions for the Machine Plate Example

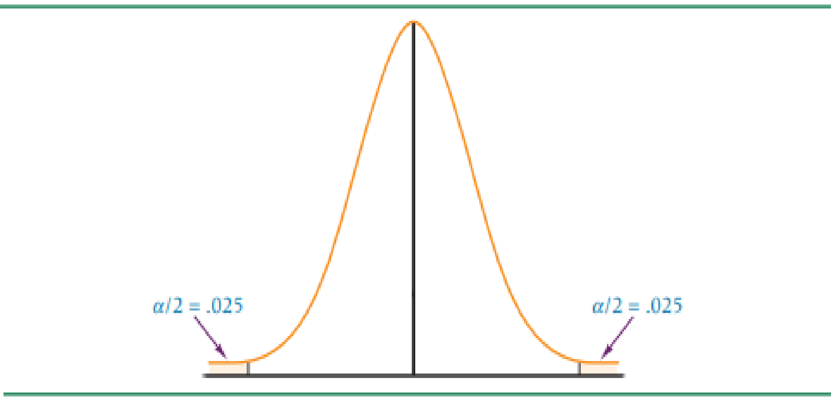
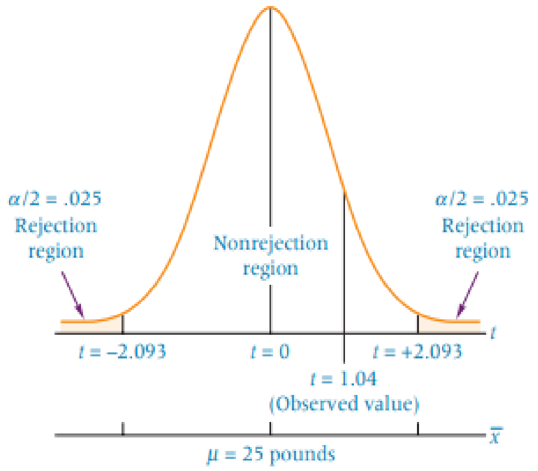


FIGURE 9.12

Graph of Observed and Critical *t* Values for the Machine Plate Example



null hypothesis if the observed value of t is less than -2.093 or greater than +2.093 (in the tails of the distribution). Computation of the test statistic yields

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{25.51 - 25.00}{\frac{2.1933}{\sqrt{20}}} = 1.04 \text{ (observed } t \text{ values)}$$

Because the observed t value is +1.04, the null hypothesis is not rejected. Not enough evidence is found in this sample to reject the hypothesis that the population mean is 25 pound.

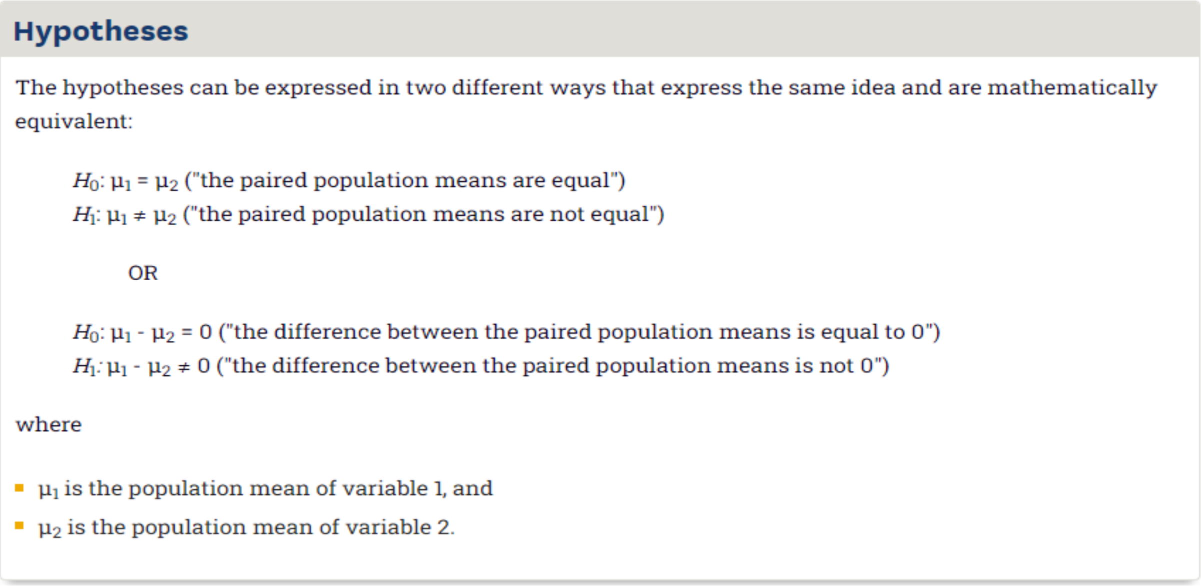
Paired Samples t Test

The Paired Samples *t* Test compares the means of two measurements taken from the same individual, object, or related units. These "paired" measurements can represent things like:

Common Uses

The Paired Samples *t* Test is commonly used to test the following:

- Statistical difference between two time points
- Statistical difference between two conditions
- Statistical difference between two measurements
- Statistical difference between a matched pair



Hypotheses

The hypotheses can be expressed in two different ways that express the same idea and are mathematically equivalent:

H_0: $\mu_1 = \mu_2$ ("the paired population means are equal")
H_1: $\mu_1 \neq \mu_2$ ("the paired population means are not equal")

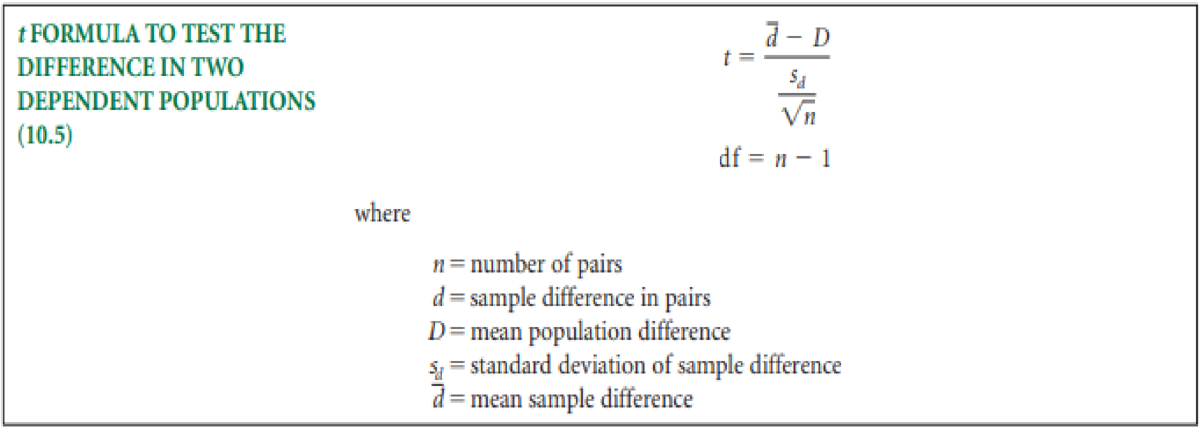
OR

H_0: $\mu_1 - \mu_2 = 0$ ("the difference between the paired population means is equal to 0")
H_1: $\mu_1 - \mu_2 \neq 0$ ("the difference between the paired population means is not 0")

where

- μ_1 is the population mean of variable 1, and
- μ_2 is the population mean of variable 2.

Formula 10.5 is used to test hypotheses about dependent populations.



t FORMULA TO TEST THE DIFFERENCE IN TWO DEPENDENT POPULATIONS (10.5)

$$t = \frac{\bar{d} - D}{\frac{s_d}{\sqrt{n}}}$$

$$\text{df} = n - 1$$

where

n = number of pairs
d = sample difference in pairs
D = mean population difference
s_d = standard deviation of sample difference
$\bar{d}$ = mean sample difference

Suppose a stock market investor is interested in determining whether there is a significant difference in the P/E (price to earnings) ratio for companies from one year to the next. Assume α=.01 Assume that differences in P/E ratios are normally distributed in the population. n=9

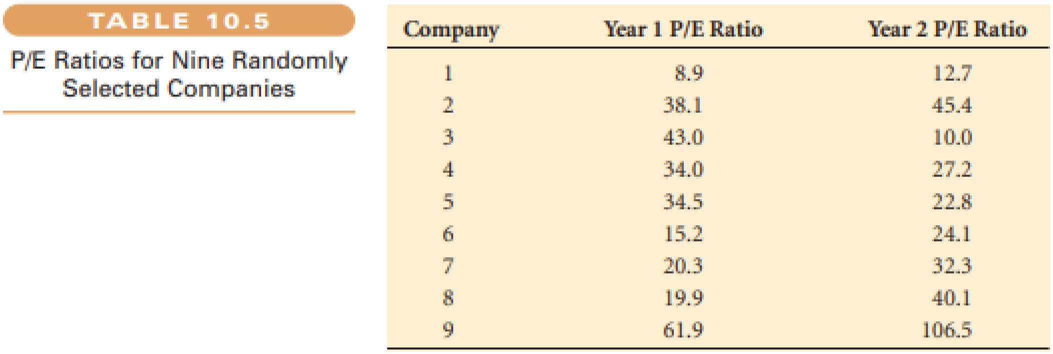


TABLE 10.5

P/E Ratios for Nine Randomly Selected Companies

Company	Year 1 P/E Ratio	Year 2 P/E Ratio
1	8.9	12.7
2	38.1	45.4
3	43.0	10.0
4	34.0	27.2
5	34.5	22.8
6	15.2	24.1
7	20.3	32.3
8	19.9	40.1
9	61.9	106.5

These data are related data because each P/E value for year 1 has a corresponding year 2 measurement on the same company. Because no prior information indicates whether P/E ratios have gone up or down, the hypothesis tested is two tailed. Assume α=.01 Assume that differences in P/E ratios are normally distributed in the population.

HYPOTHESIZE:

STEP 1.

$$H_0\colon D = 0$$
$$H_a\colon D \neq 0$$

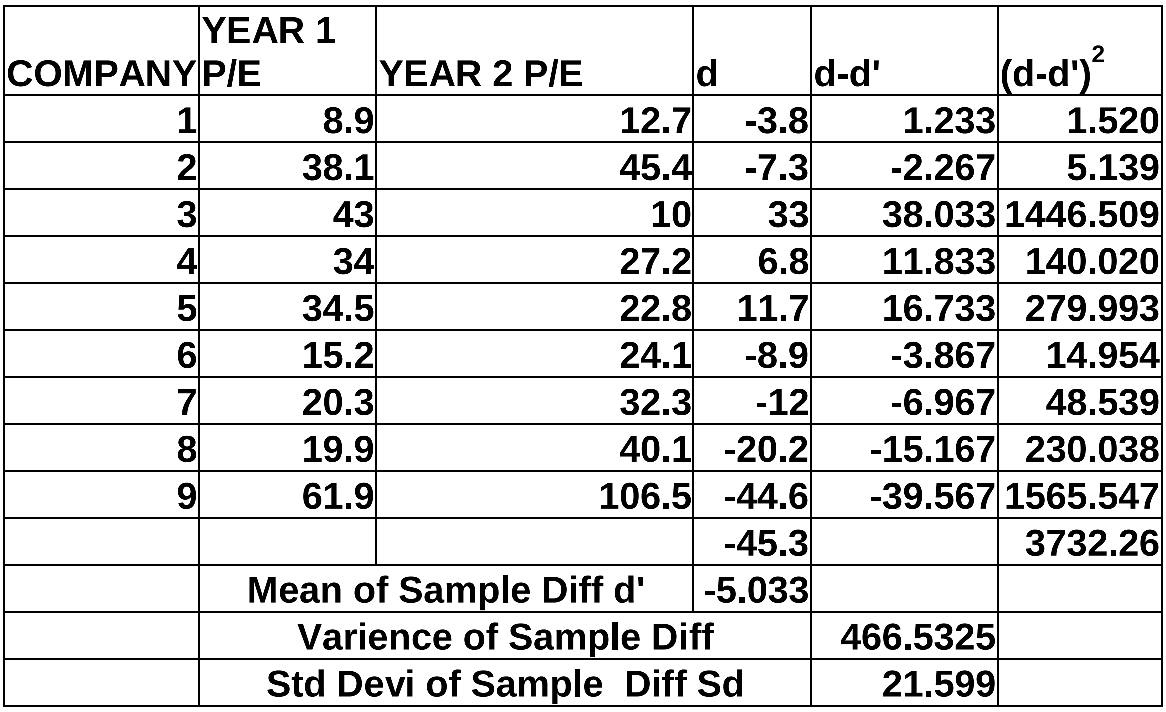
TEST:

STEP 2. The appropriate statistical test is

$$t = \frac{\bar{d} - D}{\frac{s_d}{\sqrt{n}}}$$

STEP 3.α=.01

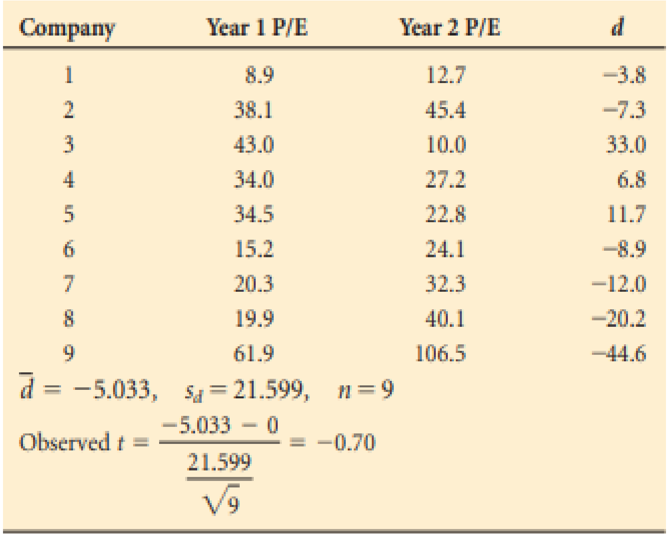
STEP 4. Becauseα=.01 and this test is two tailed,α/2=.005 is used to obtain the table t value. With nine pairs of data, n = 9, df = n - 1 = 8. The table t value is



COMPANY	YEAR 1 P/E	YEAR 2 P/E	d	d-d'	$(d-d')^2$
1	8.9	12.7	-3.8	1.233	1.520
2	38.1	45.4	-7.3	-2.267	5.139
3	43	10	33	38.033	1446.509
4	34	27.2	6.8	11.833	140.020
5	34.5	22.8	11.7	16.733	279.993
6	15.2	24.1	-8.9	-3.867	14.954
7	20.3	32.3	-12	-6.967	48.539
8	19.9	40.1	-20.2	-15.167	230.038
9	61.9	106.5	-44.6	-39.567	1565.547
			-45.3		3732.26
	Mean of Sample Diff d'		-5.033		
	Varience of Sample Diff			466.5325	
	Std Devi of Sample Diff Sd			21.599	

TABLE 10.6

Analysis of P/E Ratio Data



Company	Year 1 P/E	Year 2 P/E	d
1	8.9	12.7	−3.8
2	38.1	45.4	−7.3
3	43.0	10.0	33.0
4	34.0	27.2	6.8
5	34.5	22.8	11.7
6	15.2	24.1	−8.9
7	20.3	32.3	−12.0
8	19.9	40.1	−20.2
9	61.9	106.5	−44.6

$\bar{d} = -5.033, \quad s_d = 21.599, \quad n = 9$

$$\text{Observed } t = \frac{-5.033 - 0}{\frac{21.599}{\sqrt{9}}} = -0.70$$

$t_{.005,8}$ = ±3.355.If the observed test statistic is greater than 3.355 or less than -3.355, the null hypothesis will be rejected.

STEP 5. The sample data are given in Table 10.5.

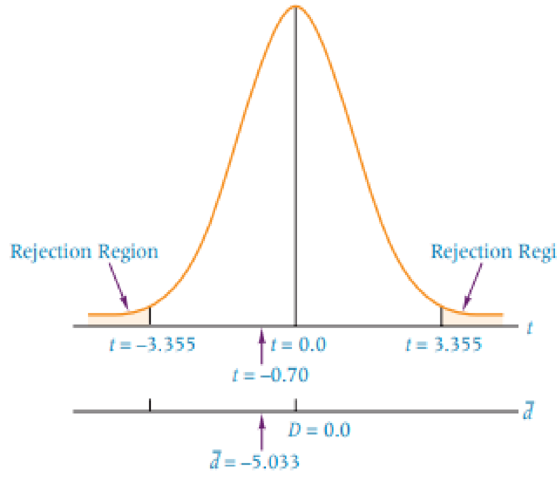
STEP 6. Table 10.6 shows the calculations to obtain the observed value of the test statistic, which is t=-0.70

ACTION: STEP 7. Because the observed t value is greater than the critical table t value in the lower tail(t=-0.70>t=-3.355), it is in the nonrejection region.

BUSINESS IMPLICATIONS: STEP 8. There is not enough evidence from the data to declare a significant difference in the average P/E ratio between year 1 and year 2. The graph in Figure 10.9 depicts the rejection regions, the critical values of t, and the observed value of t for this example.

FIGURE 10.9

Graphical Depiction of P/E Ratio Analysis



Paired Samples t-test: Example

Suppose we want to know whether or not a certain training program is able to increase the max vertical jump (in inches) of college basketball players.To test this, we may recruit a simple random sample of 20 college basketball players and measure each of their max vertical jumps. Then, we may have each player use the training program for one month and then measure their max vertical jump again at the end of the month.

- sample mean of the differences = **-0.95**
- sample standard deviation of the differences = **1.317**

To determine whether or not the training program actually had an effect on max vertical jump, we will perform a paired samples t-test at α = 0.05 using the following steps:

Step 1: Calculate the summary data for the differences.

- **x_{diff}:** sample mean of the differences = **-0.95**
- **s:** sample standard deviation of the differences = **1.317**
- **n:** sample size (i.e. number of pairs) = **20**

Step 2: Define the hypotheses.

We will perform the paired samples t-test with the following hypotheses:

- **H_0: D=0 OR (**$\mu_1 - \mu_2$)=0(the two population means are equal)
- **H_1: D>0 OR (**$\mu_1 - \mu_2$)>0 (the two population means are not equal)

Step 3: Calculate the test statistic *t*.

t = d' / (Sd/√n) = -0.95 / (1.317/√20) = **-3.226**

Step 4: Calculate the Critical t value

Degreesof freedom = n-1 = 20-1 = 19

α = 0.05 one tailed

t 0.05,19=1.729

Step 5: Draw a conclusion.

-3.226<1.729

Accept Null hypothesis

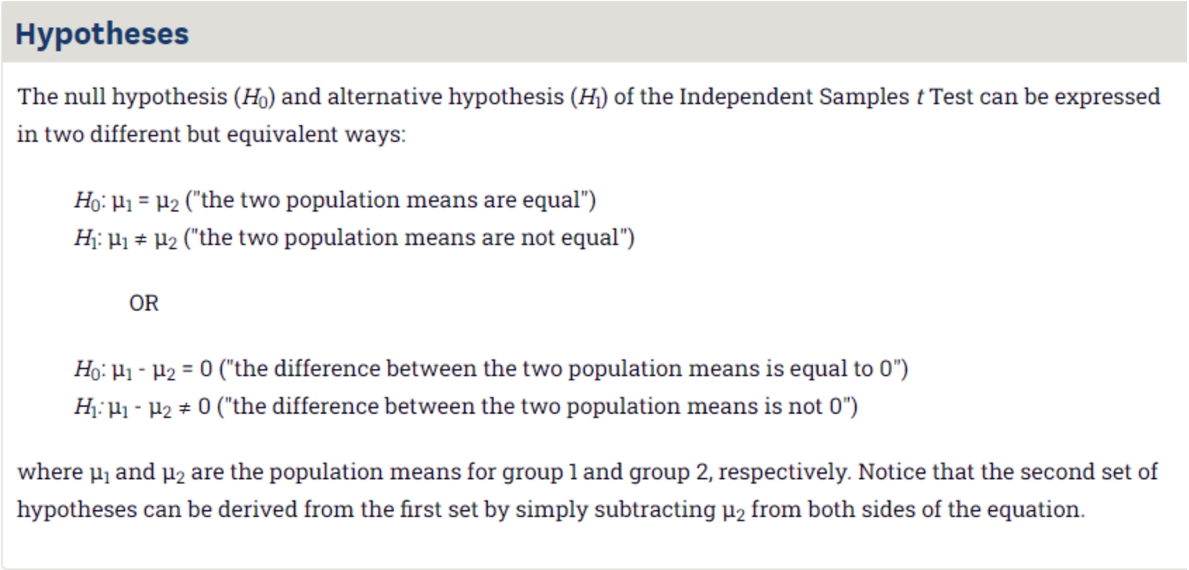
Independent Samples t Test

The Independent Samples *t* Test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. The Independent Samples *t* Test is a parametric test.

Common Uses

The Independent Samples *t* Test is commonly used to test the following:

- Statistical differences between the means of two groups
- Statistical differences between the means of two interventions
- Statistical differences between the means of two change scores



Hypotheses

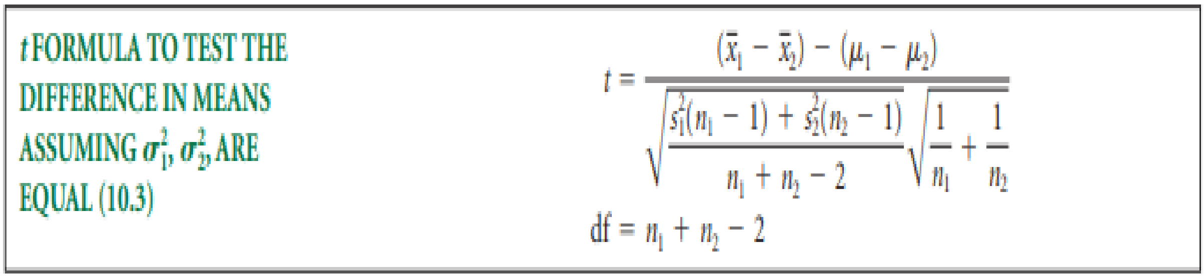
The null hypothesis (H_0) and alternative hypothesis (H_1) of the Independent Samples *t* Test can be expressed in two different but equivalent ways:

H_0: $\mu_1 = \mu_2$ ("the two population means are equal")
H_1: $\mu_1 \neq \mu_2$ ("the two population means are not equal")

OR

H_0: $\mu_1 - \mu_2 = 0$ ("the difference between the two population means is equal to 0")
H_1: $\mu_1 - \mu_2 \neq 0$ ("the difference between the two population means is not 0")

where μ_1 and μ_2 are the population means for group 1 and group 2, respectively. Notice that the second set of hypotheses can be derived from the first set by simply subtracting μ_2 from both sides of the equation.

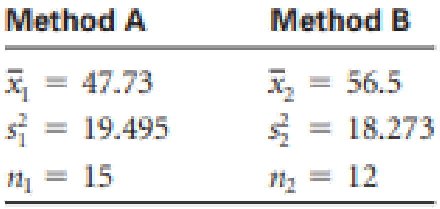


t FORMULA TO TEST THE DIFFERENCE IN MEANS ASSUMING σ_1^2, σ_2^2 ARE EQUAL (10.3)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

$$\text{df} = n_1 + n_2 - 2$$

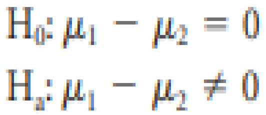
To test the difference in the two methods, the managers randomly select one group of 15 newly hired employees to take the three-day seminar (method A) and a second group of 12 new employees for the two-day DVD method (method B). Table shows required data Using α= .05, the managers want to determine whether there is a significant difference in the mean scores of the two groups.



Method A	Method B
$\bar{x}_1 = 47.73$	$\bar{x}_2 = 56.5$
$s_1^2 = 19.495$	$s_2^2 = 18.273$
$n_1 = 15$	$n_2 = 12$

HYPOTHESIZE:

STEP 1. The hypotheses for this test follow.



$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

TEST:

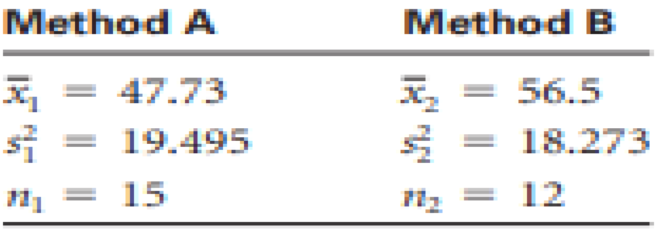
STEP 2. The statistical test to be used is formula 10.3.

STEP 3. The value of alpha is .05.

STEP 4. Because the hypotheses are = and ≠this test is two tailed. The degrees of freedom are 25 (15 + 12 - 2 = 25) and alpha is .05. The t table requires an alpha value for one tail only, and, because it is a two-tailed test, alpha is split from .05 to .025 to obtain the table t value: **$t_{.025,25}$ = ±2.060**

The null hypothesis will be rejected if the observed t value is less than -2.060 or greater than +2.060.

STEP 5. The sample data are given in Table 10.2. From these data, we can calculate the sample statistics. The sample means and variances follow.



Method A	Method B
$\bar{x}_1 = 47.73$	$\bar{x}_2 = 56.5$
$s_1^2 = 19.495$	$s_2^2 = 18.273$
$n_1 = 15$	$n_2 = 12$

Note: If the equal variances assumption can no

STEP 6. The observed value of t is

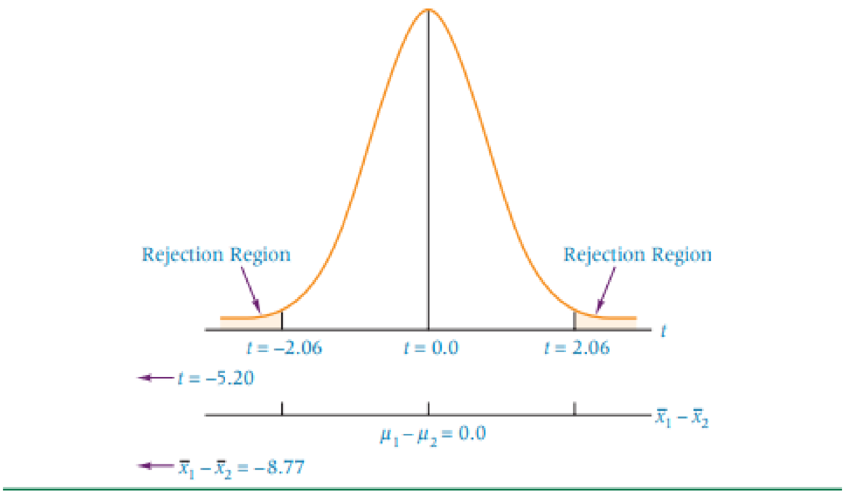
$$t = \frac{(47.73 - 56.50) - (0)}{\sqrt{\frac{(19.495)(14) + (18.273)(11)}{(15 + 12 - 2)}}\sqrt{\frac{1}{15} + \frac{1}{12}}} = -5.20$$

ACTION: STEP 7. Because the observed value, t = -5.20, is less than the lower critical table value, t = -2.06, the observed value of t is in the rejection region. The null hypothesis is rejected. There is a significant difference in the mean scores of the two tests.

BUSINESS IMPLICATIONS: STEP 8. Figure 10.6 shows the critical areas and the observed t value. Note that the computed t value is -5.20, which is enough to cause the managers of the Hernandez Manufacturing Company to reject the null hypothesis.

FIGURE 10.6

t Values for the Training Methods Example



Chi-Square Test of Independence

The Chi-Square Test of Independence determines whether there is an association between categorical variables (i.e., whether the variables are independent or related). It is a nonparametric test.

Common Uses

The Chi-Square Test of Independence is commonly used to test the following:

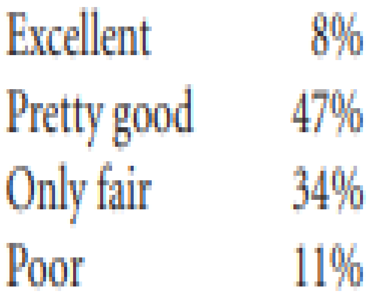
- Statistical independence or association between two or more categorical variables.

The Chi-Square Test of Independence can only compare categorical variables. It cannot make comparisons between continuous variables or between categorical and continuous variables. Additionally, the Chi-Square Test of Independence only assesses *associations* between categorical variables, and can not provide any inferences about causation.

1. Suppose a store manager wants to find out whether the results of this consumer survey apply to customers of supermarkets in her city. To do so, she interviews 207 randomly selected consumers as they leave

supermarkets in various parts of the city.. Now the manager can use a chisquare goodness-of-fit test to determine whether the observed frequencies of responses from this survey are the same as the frequencies that would be expected on the basis of the national survey.(α= .05).

Expected %



Excellent	8%
Pretty good	47%
Only fair	34%
Poor	11%

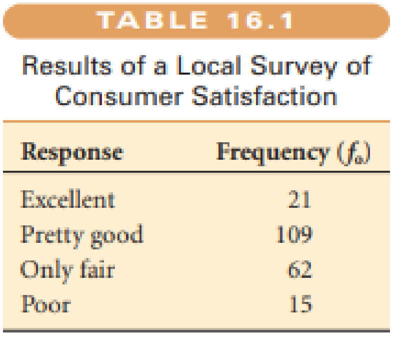
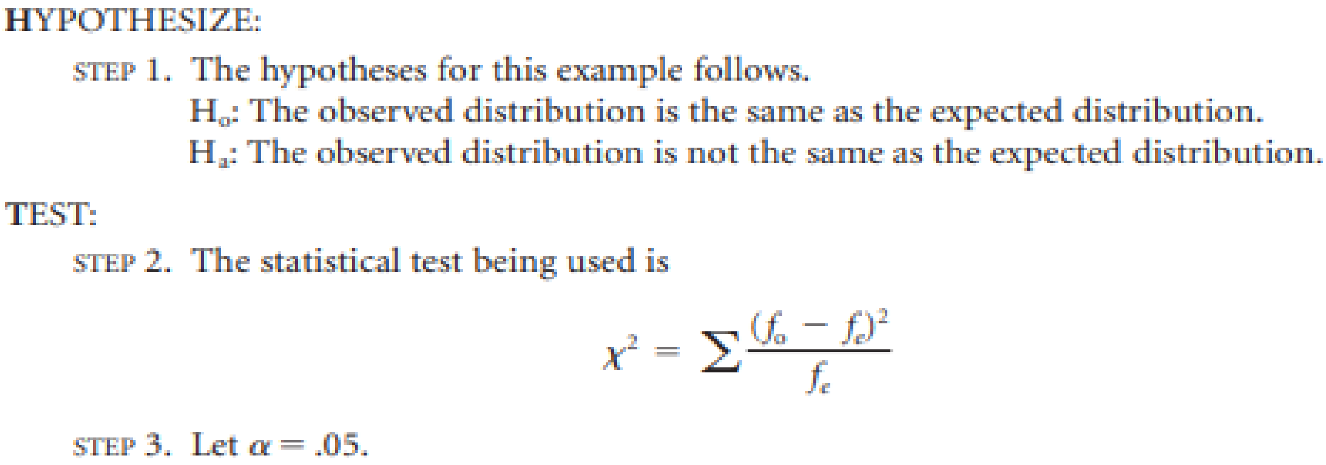


TABLE 16.1

Results of a Local Survey of Consumer Satisfaction

Response	Frequency (f_o)
Excellent	21
Pretty good	109
Only fair	62
Poor	15



HYPOTHESIZE:

STEP 1. The hypotheses for this example follows.
H_o: The observed distribution is the same as the expected distribution.
H_a: The observed distribution is not the same as the expected distribution.

TEST:

STEP 2. The statistical test being used is

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

STEP 3. Let $\alpha = .05$.

STEP 4. Chi-square goodness-of-fit tests are one tailed because a chi-square of zero indicates perfect agreement between distributions. Any deviation from zero difference occurs in the positive direction only because chi-square is determined by a sum of squared values and can never be negative. With four categories in this example (excellent, pretty good, only fair, and poor), k = 4. The degrees of freedom are k - 1 because the expected distribution is given: k - 1 = 4 - 1 = 3. For α= .05 and df = 3, the critical chisquare value is

$$\chi^2_{.05,3} = 7.8147$$

After the data are analyzed, an observed chi-square greater than 7.8147 must be computed in order to reject the null hypothesis.

STEP 5. The observed values gathered in the sample data from Table 16.1 sum to 207. Thus n = 207. The expected proportions are given, but the expected frequencies must be calculated by multiplying the expected proportions by the sample total of the observed frequencies, as shown in Table 16.2

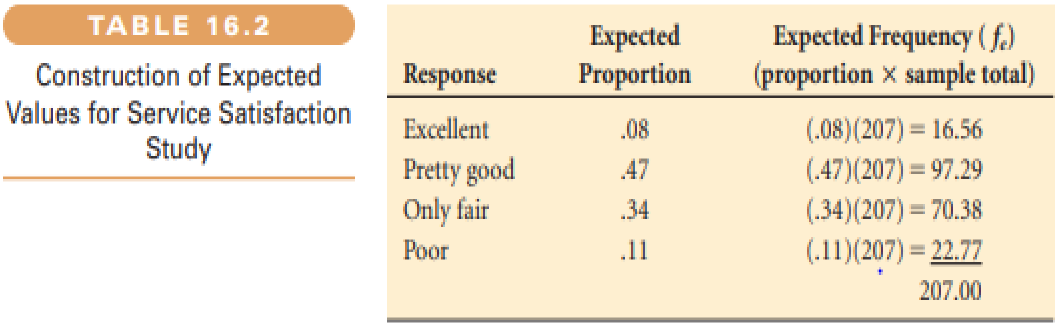


TABLE 16.2

Construction of Expected Values for Service Satisfaction Study

Response	Expected Proportion	Expected Frequency (f_e) (proportion × sample total)
Excellent	.08	(.08)(207) = 16.56
Pretty good	.47	(.47)(207) = 97.29
Only fair	.34	(.34)(207) = 70.38
Poor	.11	(.11)(207) = 22.77
		207.00

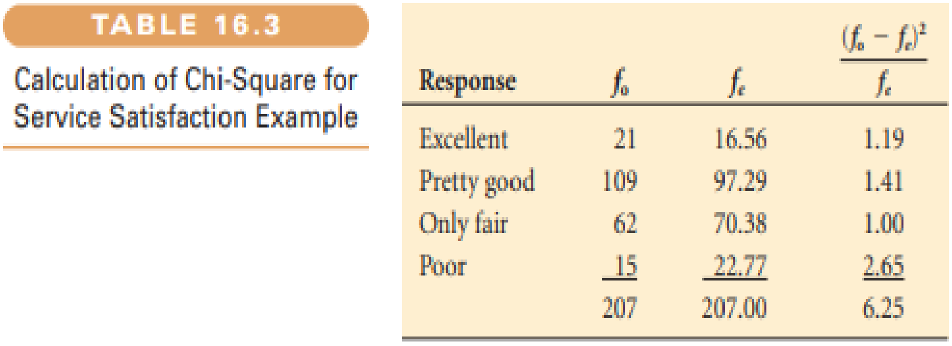


TABLE 16.3

Calculation of Chi-Square for Service Satisfaction Example

Response	f_o	f_e	$\frac{(f_o - f_e)^2}{f_e}$
Excellent	21	16.56	1.19
Pretty good	109	97.29	1.41
Only fair	62	70.38	1.00
Poor	15	22.77	2.65
	207	207.00	6.25

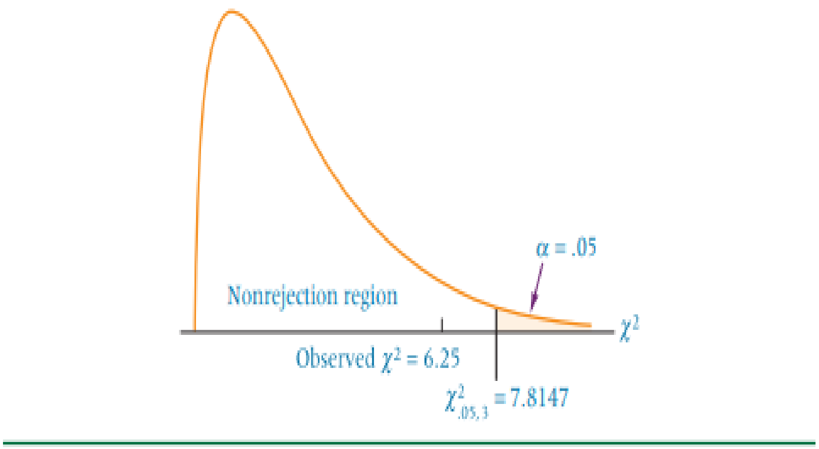
STEP 6. The chi-square goodness-of-fit can then be calculated, as shown in Table 16.3.

ACTION: STEP 7. Because the observed value of chi-square of 6.25 is not greater than the critical table value of 7.8147, the store manager will not reject the null hypothesis.

BUSINESS IMPLICATIONS: STEP 8. Thus the data gathered in the sample of 207 supermarket shoppers indicate that the distribution of responses of supermarket shoppers in the manager's city is not significantly different from the distribution of responses to the national survey. The store manager may conclude that her customers do not appear to have attitudes different from those people who took the survey.

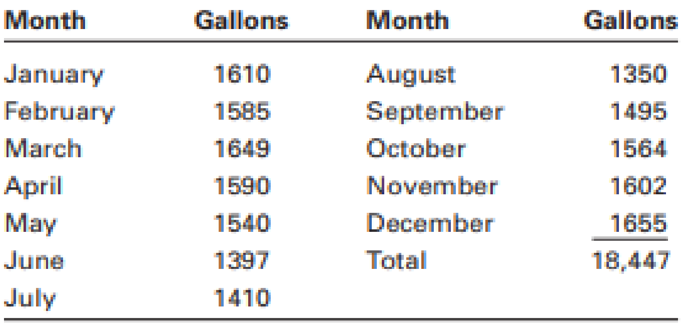
FIGURE 16.1

Minitab Graph of Chi-Square Distribution for Service Satisfaction Example



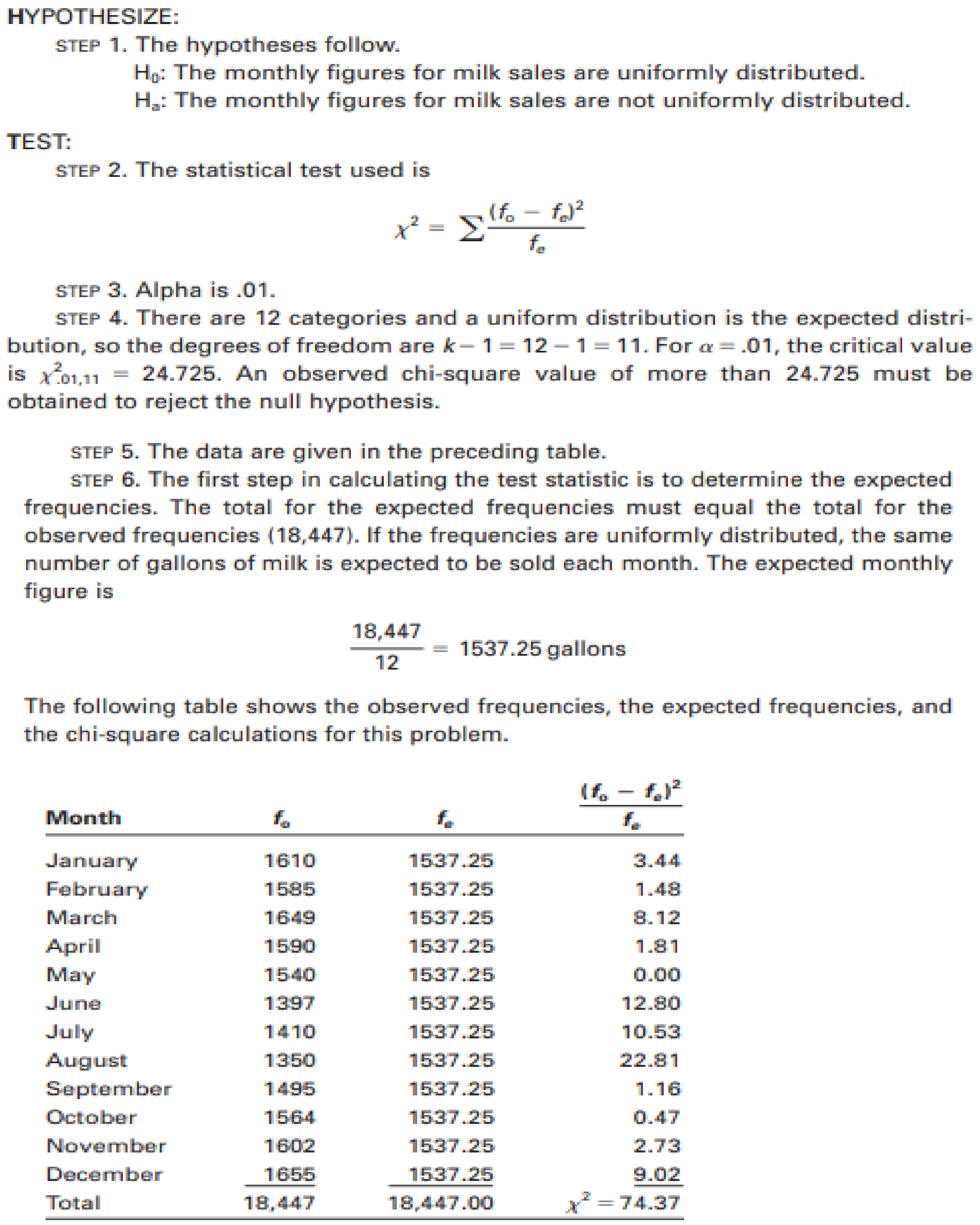
Problem1:

Dairies would like to know whether the sales of milk are distributed uniformly over a year so they can plan for milk production and storage. A uniform distribution means that the frequencies are the same in all categories. In this situation, the producers are attempting to determine whether the amounts of milk sold are the same for each month of the year. They ascertain the number of gallons of milk sold by sampling one large supermarket each month during a year, obtaining the following data. Use α=.01 to test whether the data fit a uniform distribution.



Month	Gallons	Month	Gallons
January	1610	August	1350
February	1585	September	1495
March	1649	October	1564
April	1590	November	1602
May	1540	December	1655
June	1397	Total	18,447
July	1410		

Solution:



HYPOTHESIZE:

STEP 1. The hypotheses follow.

H_0: The monthly figures for milk sales are uniformly distributed.

H_a: The monthly figures for milk sales are not uniformly distributed.

TEST:

STEP 2. The statistical test used is

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

STEP 3. Alpha is .01.

STEP 4. There are 12 categories and a uniform distribution is the expected distribution, so the degrees of freedom are $k - 1 = 12 - 1 = 11$. For $\alpha = .01$, the critical value is $\chi^2_{.01,11} = 24.725$. An observed chi-square value of more than 24.725 must be obtained to reject the null hypothesis.

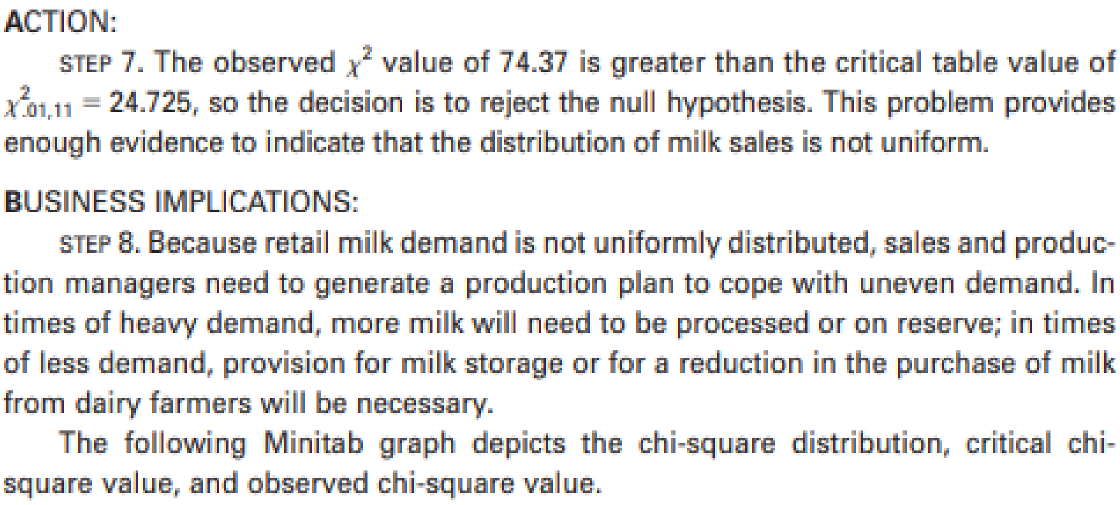
STEP 5. The data are given in the preceding table.

STEP 6. The first step in calculating the test statistic is to determine the expected frequencies. The total for the expected frequencies must equal the total for the observed frequencies (18,447). If the frequencies are uniformly distributed, the same number of gallons of milk is expected to be sold each month. The expected monthly figure is

$$\frac{18{,}447}{12} = 1537.25 \text{ gallons}$$

The following table shows the observed frequencies, the expected frequencies, and the chi-square calculations for this problem.

Month	f_o	f_e	$\frac{(f_o - f_e)^2}{f_e}$
January	1610	1537.25	3.44
February	1585	1537.25	1.48
March	1649	1537.25	8.12
April	1590	1537.25	1.81
May	1540	1537.25	0.00
June	1397	1537.25	12.80
July	1410	1537.25	10.53
August	1350	1537.25	22.81
September	1495	1537.25	1.16
October	1564	1537.25	0.47
November	1602	1537.25	2.73
December	1655	1537.25	9.02
Total	18,447	18,447.00	$\chi^2 = 74.37$



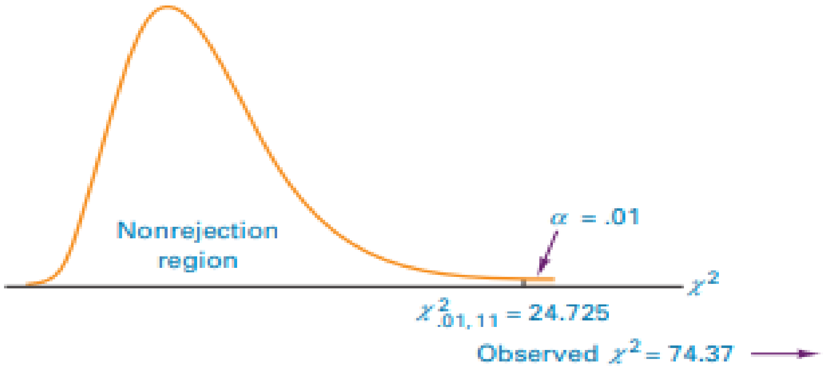
ACTION:

STEP 7. The observed χ^2 value of 74.37 is greater than the critical table value of $\chi^2_{.01,11} = 24.725$, so the decision is to reject the null hypothesis. This problem provides enough evidence to indicate that the distribution of milk sales is not uniform.

BUSINESS IMPLICATIONS:

STEP 8. Because retail milk demand is not uniformly distributed, sales and production managers need to generate a production plan to cope with uneven demand. In times of heavy demand, more milk will need to be processed or on reserve; in times of less demand, provision for milk storage or for a reduction in the purchase of milk from dairy farmers will be necessary.

The following Minitab graph depicts the chi-square distribution, critical chi-square value, and observed chi-square value.



One-Way ANOVA

One-Way ANOVA ("analysis of variance") compares the means of two or more independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. One-Way ANOVA is a parametric test.

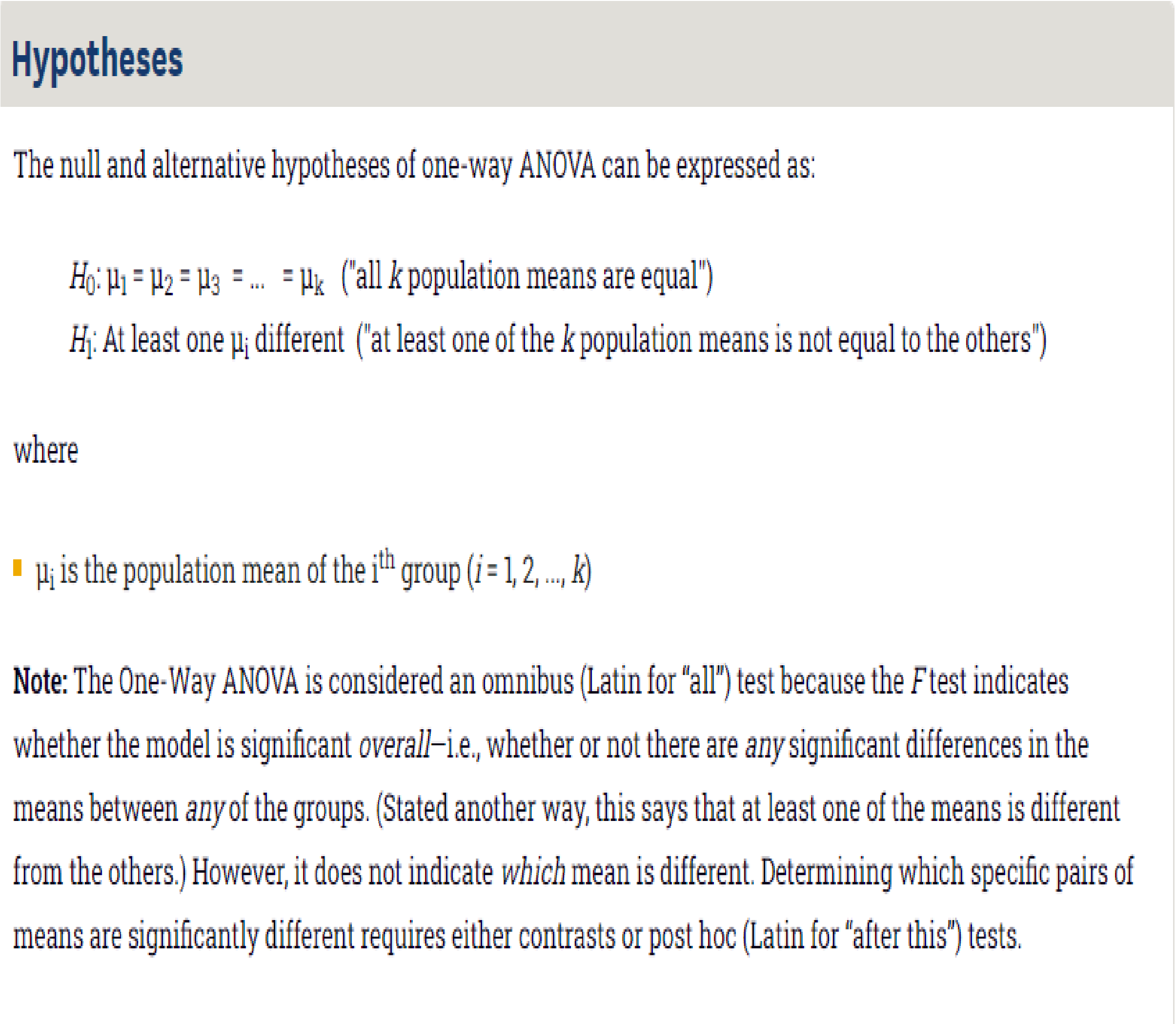
Common Uses

The One-Way ANOVA is often used to analyze data from the following types of studies:

- Field studies
- Experiments
- Quasi-experiments

The One-Way ANOVA is commonly used to test the following:

- Statistical differences among the means of two or more groups
- Statistical differences among the means of two or more interventions
- Statistical differences among the means of two or more change scores



Hypotheses

The null and alternative hypotheses of one-way ANOVA can be expressed as:

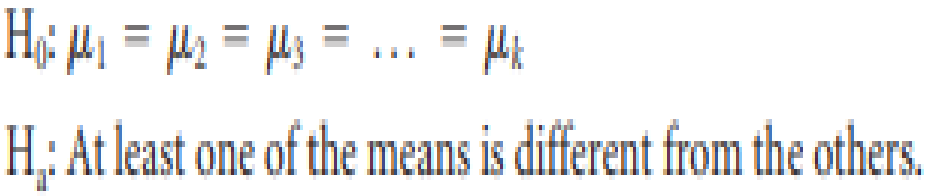
H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$ ("all *k* population means are equal")

H_1: At least one μ_i different ("at least one of the *k* population means is not equal to the others")

where

- μ_i is the population mean of the i^{th} group ($i = 1, 2, \ldots, k$)

Note: The One-Way ANOVA is considered an omnibus (Latin for "all") test because the *F* test indicates whether the model is significant *overall*—i.e., whether or not there are *any* significant differences in the means between *any* of the groups. (Stated another way, this says that at least one of the means is different from the others.) However, it does not indicate *which* mean is different. Determining which specific pairs of means are significantly different requires either contrasts or post hoc (Latin for "after this") tests.



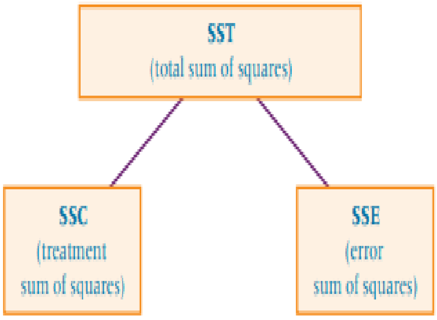
H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

H_a: At least one of the means is different from the others.

Testing these hypotheses by using one-way ANOVA is accomplished by partitioning the total variance of the data into the following two variances.

1. The variance resulting from the treatment (columns)
2. The error variance, or that portion of the total variance unexplained by the treatment

FIGURE 11.3 Partitioning Total Sum of Squares of Variation



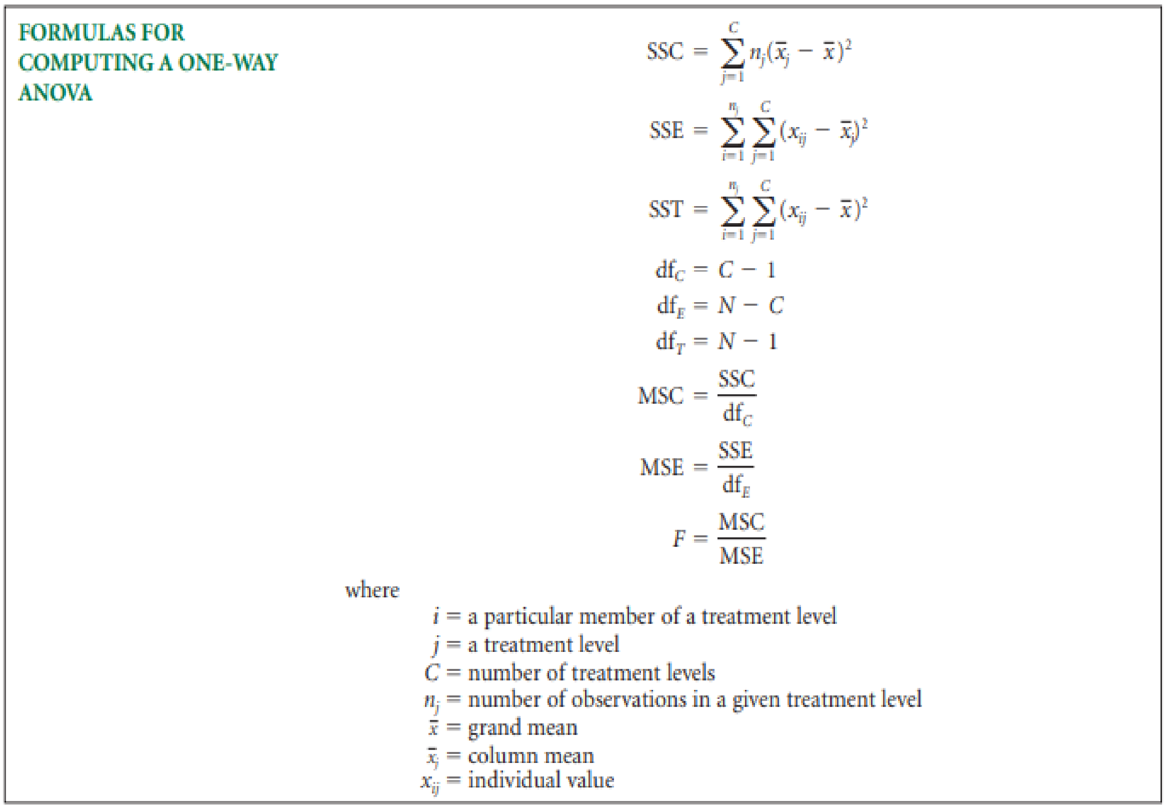
As part of this process, the total sum of squares of deviation of values around the mean can be divided into two additive and independent parts.

$$\begin{array}{ccccc} SST & = & SSC & + & SSE \end{array}$$

$$\sum_{i=1}^{n_j}\sum_{j=1}^{C}(x_{ij}-\bar{\bar{x}})^2 = \sum_{j=1}^{C} n_j(\bar{x}_j-\bar{\bar{x}})^2 + \sum_{i=1}^{n_j}\sum_{j=1}^{C}(x_{ij}-\bar{x}_j)^2$$

where

SST = total sum of squares
SSC = sum of squares column (treatment)
SSE = sum of squares error
i = particular member of a treatment level
j = a treatment level
C = number of treatment levels
n_j = number of observations in a given treatment level
$\bar{\bar{x}}$ = grand mean
$\bar{x}_j$ = mean of a treatment group or level
x_{ij} = individual value



FORMULAS FOR COMPUTING A ONE-WAY ANOVA

$$\text{SSC} = \sum_{j=1}^{C} n_j(\bar{x}_j - \bar{x})^2$$

$$\text{SSE} = \sum_{i=1}^{n_j}\sum_{j=1}^{C}(x_{ij} - \bar{x}_j)^2$$

$$\text{SST} = \sum_{i=1}^{n_j}\sum_{j=1}^{C}(x_{ij} - \bar{x})^2$$

$$\text{df}_C = C - 1$$

$$\text{df}_E = N - C$$

$$\text{df}_T = N - 1$$

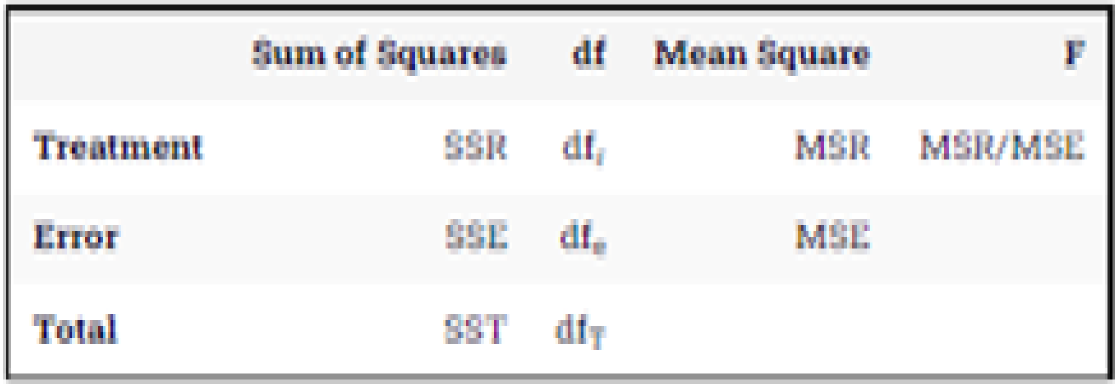
$$\text{MSC} = \frac{\text{SSC}}{\text{df}_C}$$

$$\text{MSE} = \frac{\text{SSE}}{\text{df}_E}$$

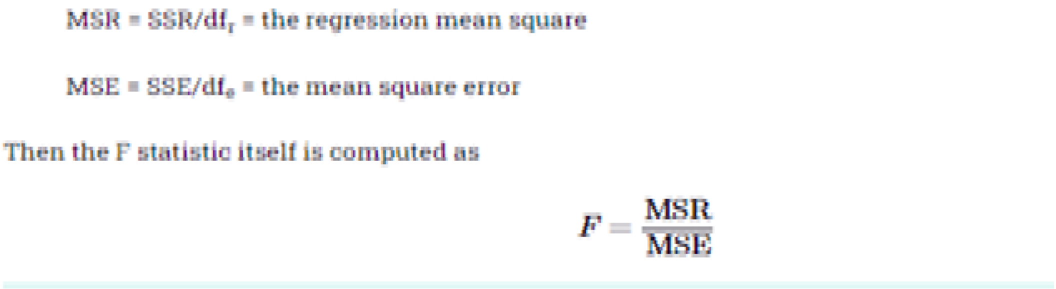
$$F = \frac{\text{MSC}}{\text{MSE}}$$

where

i = a particular member of a treatment level
j = a treatment level
C = number of treatment levels
n_j = number of observations in a given treatment level
$\bar{x}$ = grand mean
$\bar{x}_j$ = column mean
x_{ij} = individual value



	Sum of Squares	df	Mean Square	F
Treatment	SSR	df_r	MSR	MSR/MSE
Error	SSE	df_e	MSE	
Total	SST	df_T		



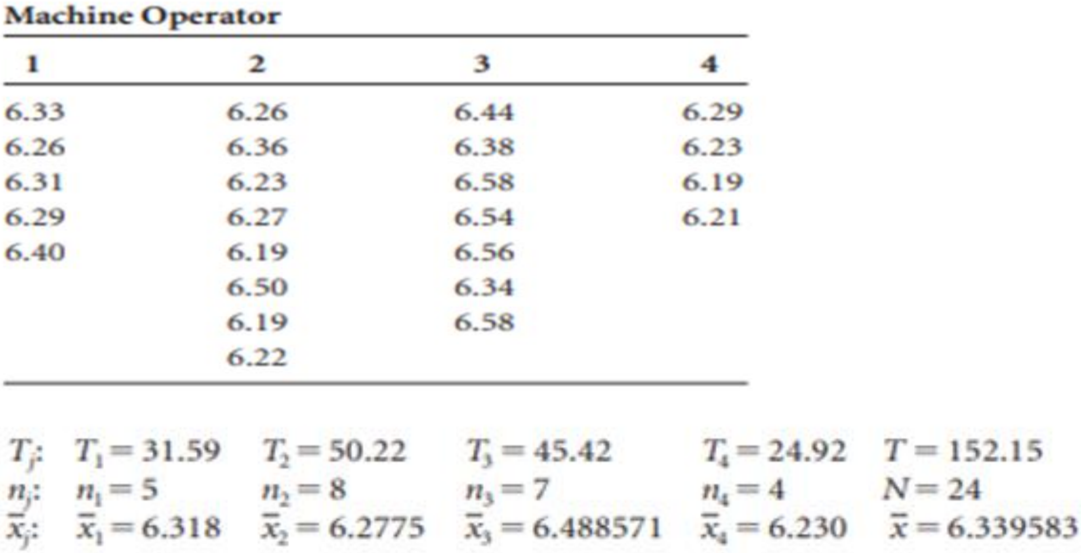
MSR = SSR/df_r = the regression mean square

MSE = SSE/df_e = the mean square error

Then the F statistic itself is computed as

$$F = \frac{\text{MSR}}{\text{MSE}}$$

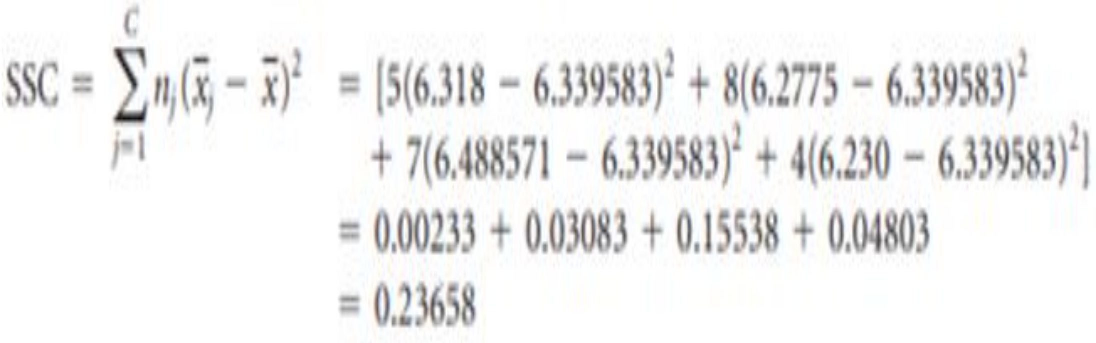
1. Construct ANOVA table for following data.



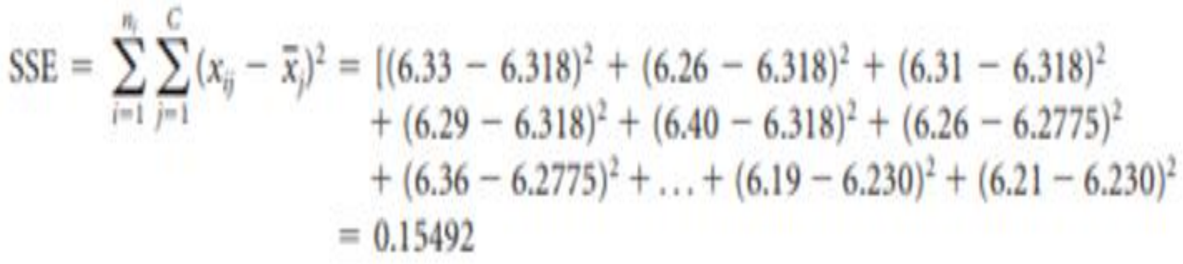
Machine Operator

1	2	3	4
6.33	6.26	6.44	6.29
6.26	6.36	6.38	6.23
6.31	6.23	6.58	6.19
6.29	6.27	6.54	6.21
6.40	6.19	6.56	
	6.50	6.34	
	6.19	6.58	
	6.22		

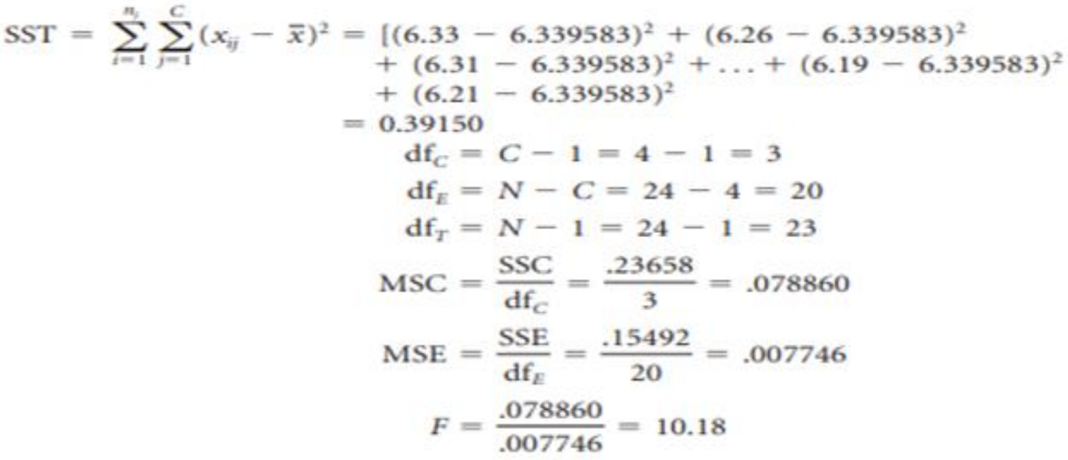
T_j:	$T_1 = 31.59$	$T_2 = 50.22$	$T_3 = 45.42$	$T_4 = 24.92$	$T = 152.15$
n_j:	$n_1 = 5$	$n_2 = 8$	$n_3 = 7$	$n_4 = 4$	$N = 24$
$\bar{x}_j$:	$\bar{x}_1 = 6.318$	$\bar{x}_2 = 6.2775$	$\bar{x}_3 = 6.488571$	$\bar{x}_4 = 6.230$	$\bar{x} = 6.339583$



$$\begin{aligned} SSC = \sum_{j=1}^{C} n_j(\bar{x}_j - \bar{x})^2 &= [5(6.318 - 6.339583)^2 + 8(6.2775 - 6.339583)^2 \\ &\quad + 7(6.488571 - 6.339583)^2 + 4(6.230 - 6.339583)^2] \\ &= 0.00233 + 0.03083 + 0.15538 + 0.04803 \\ &= 0.23658 \end{aligned}$$



$$\begin{aligned} SSE = \sum_{i=1}^{n_j} \sum_{j=1}^{C} (x_{ij} - \bar{x}_j)^2 &= [(6.33 - 6.318)^2 + (6.26 - 6.318)^2 + (6.31 - 6.318)^2 \\ &\quad + (6.29 - 6.318)^2 + (6.40 - 6.318)^2 + (6.26 - 6.2775)^2 \\ &\quad + (6.36 - 6.2775)^2 + \ldots + (6.19 - 6.230)^2 + (6.21 - 6.230)^2 \\ &= 0.15492 \end{aligned}$$



$$
\begin{aligned}
\text{SST} = \sum_{i=1}^{n_j}\sum_{j=1}^{C}(x_{ij} - \bar{x})^2 &= [(6.33 - 6.339583)^2 + (6.26 - 6.339583)^2 \\
&\quad + (6.31 - 6.339583)^2 + \ldots + (6.19 - 6.339583)^2 \\
&\quad + (6.21 - 6.339583)^2 \\
&= 0.39150
\end{aligned}
$$

$$\text{df}_C = C - 1 = 4 - 1 = 3$$

$$\text{df}_E = N - C = 24 - 4 = 20$$

$$\text{df}_T = N - 1 = 24 - 1 = 23$$

$$\text{MSC} = \frac{\text{SSC}}{\text{df}_C} = \frac{.23658}{3} = .078860$$

$$\text{MSE} = \frac{\text{SSE}}{\text{df}_E} = \frac{.15492}{20} = .007746$$

$$F = \frac{.078860}{.007746} = 10.18$$

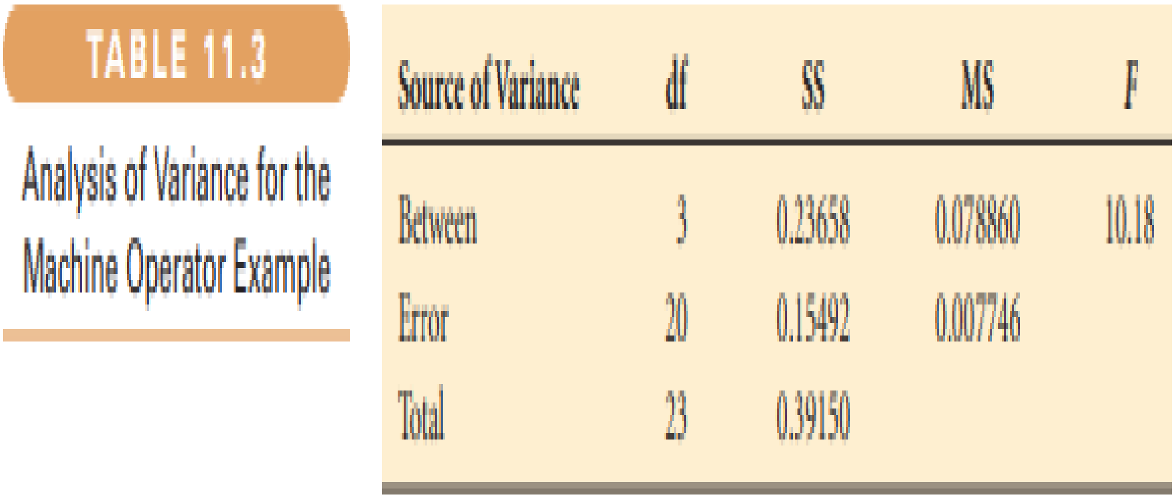
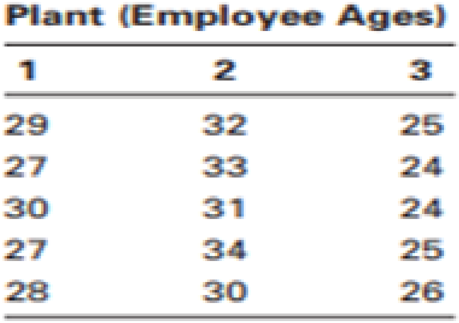


TABLE 11.3

Analysis of Variance for the Machine Operator Example

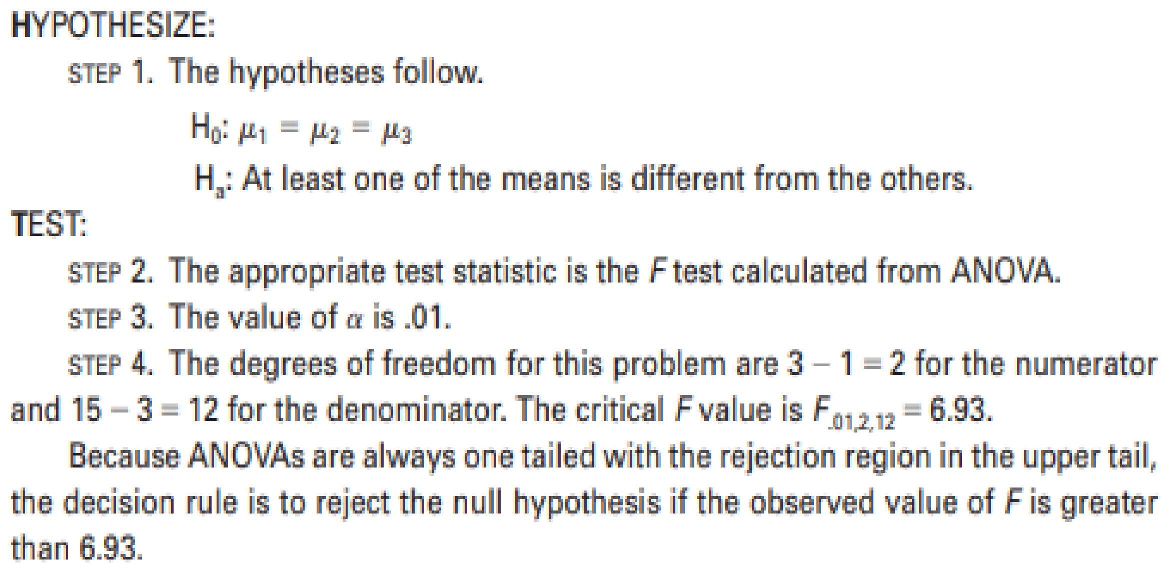
Source of Variance	df	SS	MS	F
Between	3	0.23658	0.078860	10.18
Error	20	0.15492	0.007746	
Total	23	0.39150		

A company has three manufacturing plants, and company officials want to determine whether there is a difference in the average age of workers at the three locations. The following data are the ages of five randomly selected workers at each plant. Perform a one-way ANOVA to determine whether there is a significant difference in the mean ages of the workers at the three plants. Use α = .01 and note that the sample sizes are equal.



Plant (Employee Ages)

1	2	3
29	32	25
27	33	24
30	31	24
27	34	25
28	30	26



HYPOTHESIZE:

STEP 1. The hypotheses follow.

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

H_a: At least one of the means is different from the others.

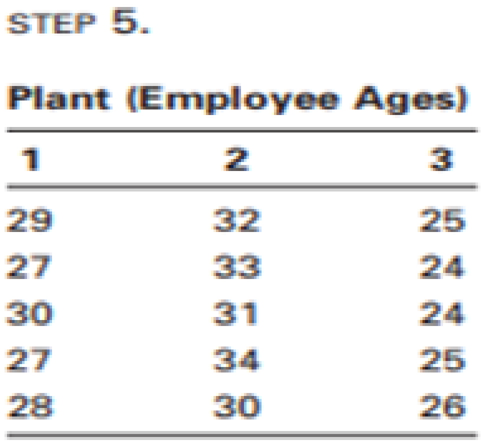
TEST:

STEP 2. The appropriate test statistic is the *F* test calculated from ANOVA.

STEP 3. The value of α is .01.

STEP 4. The degrees of freedom for this problem are 3 − 1 = 2 for the numerator and 15 − 3 = 12 for the denominator. The critical *F* value is $F_{.01,2,12} = 6.93$.

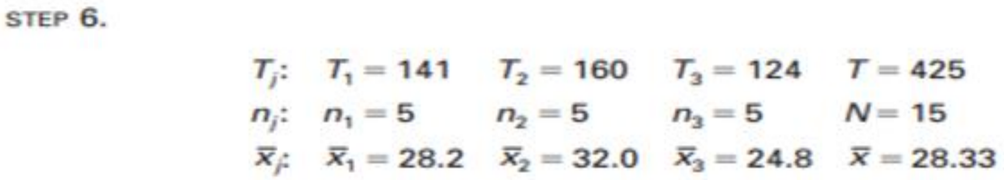
Because ANOVAs are always one tailed with the rejection region in the upper tail, the decision rule is to reject the null hypothesis if the observed value of *F* is greater than 6.93.



STEP 5.

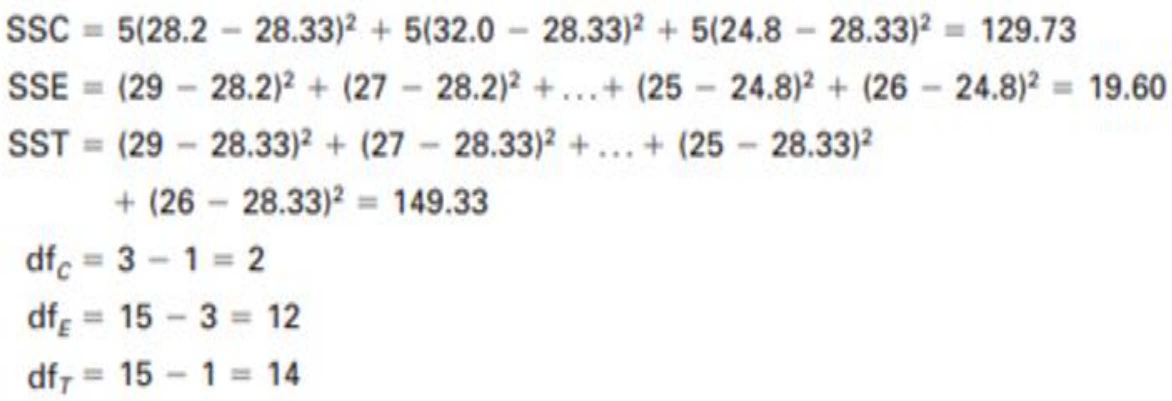
Plant (Employee Ages)

1	2	3
29	32	25
27	33	24
30	31	24
27	34	25
28	30	26



STEP 6.

T_j:	$T_1 = 141$	$T_2 = 160$	$T_3 = 124$	$T = 425$
n_j:	$n_1 = 5$	$n_2 = 5$	$n_3 = 5$	$N = 15$
$\bar{x}_j$:	$\bar{x}_1 = 28.2$	$\bar{x}_2 = 32.0$	$\bar{x}_3 = 24.8$	$\bar{x} = 28.33$



$$\text{SSC} = 5(28.2 - 28.33)^2 + 5(32.0 - 28.33)^2 + 5(24.8 - 28.33)^2 = 129.73$$

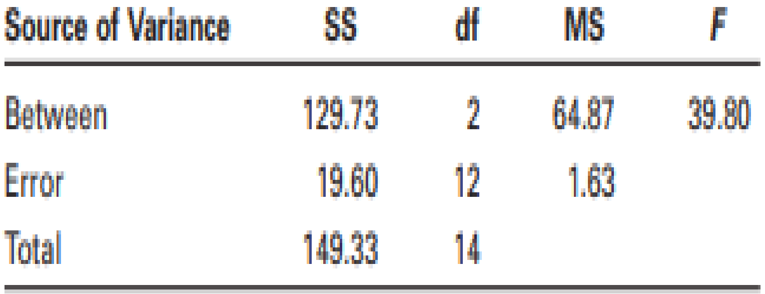
$$\text{SSE} = (29 - 28.2)^2 + (27 - 28.2)^2 + \ldots + (25 - 24.8)^2 + (26 - 24.8)^2 = 19.60$$

$$\text{SST} = (29 - 28.33)^2 + (27 - 28.33)^2 + \ldots + (25 - 28.33)^2 + (26 - 28.33)^2 = 149.33$$

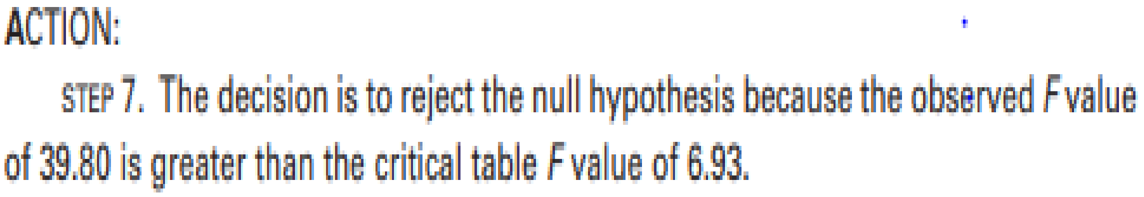
$$\text{df}_C = 3 - 1 = 2$$

$$\text{df}_E = 15 - 3 = 12$$

$$\text{df}_T = 15 - 1 = 14$$

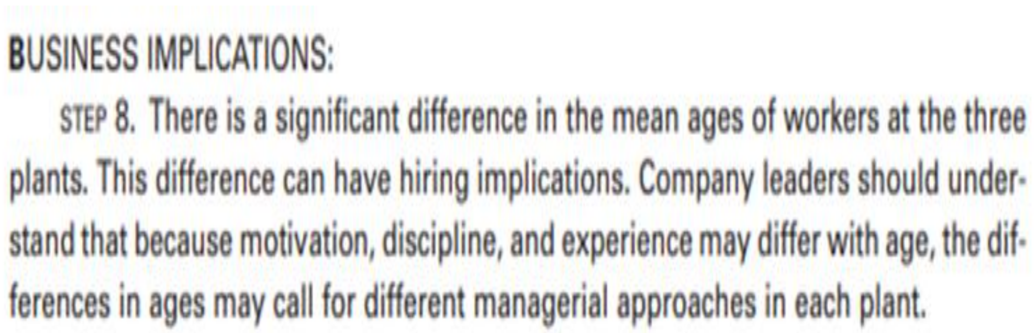


Source of Variance	SS	df	MS	*F*
Between	129.73	2	64.87	39.80
Error	19.60	12	1.63	
Total	149.33	14		



ACTION:

STEP 7. The decision is to reject the null hypothesis because the observed *F* value of 39.80 is greater than the critical table *F* value of 6.93.



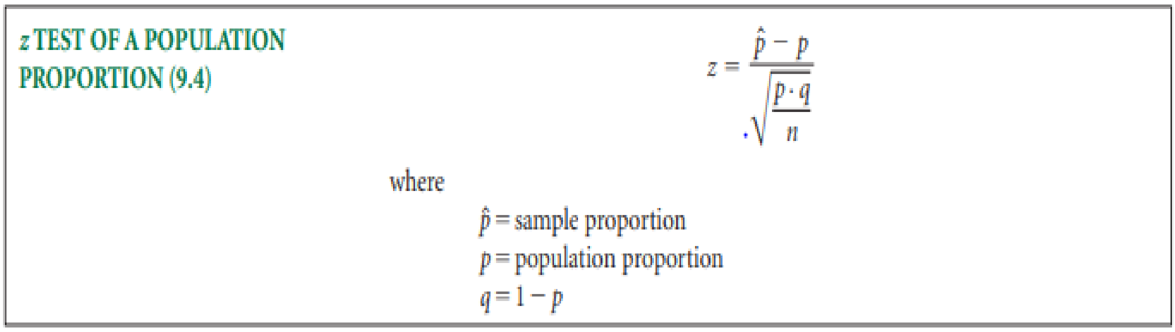
BUSINESS IMPLICATIONS:

STEP 8. There is a significant difference in the mean ages of workers at the three plants. This difference can have hiring implications. Company leaders should understand that because motivation, discipline, and experience may differ with age, the differences in ages may call for different managerial approaches in each plant.

One Sample Proportion Test

Data analysis used in business decision making often contains proportions to describe such aspects as market share, consumer makeup, quality defects, on-time delivery rate, profitable stocks, and others.

As an example, suppose a company held a 26% or .26, share of the market for several years. Due to a massive marketing effort and improved product quality, company officials believe that the market share increased, and they want to prove it.



z TEST OF A POPULATION PROPORTION (9.4)

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

where

$\hat{p}$ = sample proportion
p = population proportion
$q = 1 - p$

A manufacturer believes exactly 8% of its products contain at least one minor flaw. Suppose a company researcher wants to test this belief. The null and alternative hypotheses are

H_0: $p = .08$
H_a: $p \neq .08$

The business researcher randomly selects a sample of 200 products, inspects each item for flaws, and determines that 33 items have at least one minor flaw. Calculating the sample proportion.(α=.10)

This test is two-tailed because the hypothesis being tested is whether the proportion of products with at least one minor flaw is .08. Alpha is selected to be .10. Figure 9.15 shows the distribution, with the rejection regions and $z_{.05}$. Because is divided for a two-tailed test, the table value for an area of (1/2)(.10) = .05 is $z_{.05} = \pm 1.645$

For the business researcher to reject the null hypothesis, the observed z value must be greater than 1.645 or less than -1.645. The business researcher randomly selects a sample of 200 products, inspects each item for flaws, and determines that 33 items have at least one minor flaw. Calculating the sample proportion gives

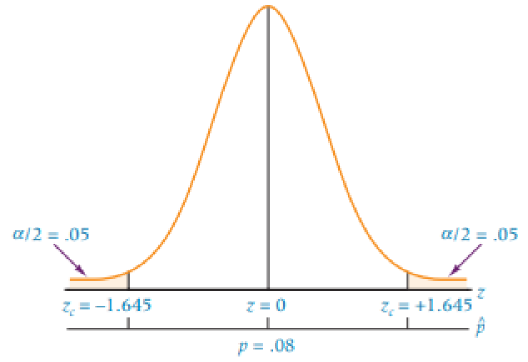
$$\hat{p} = \frac{33}{200} = .165$$

The observed z value is calculated as:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.165 - .080}{\sqrt{\frac{(.08)(.92)}{200}}} = \frac{.085}{.019} = 4.43$$

FIGURE 9.15

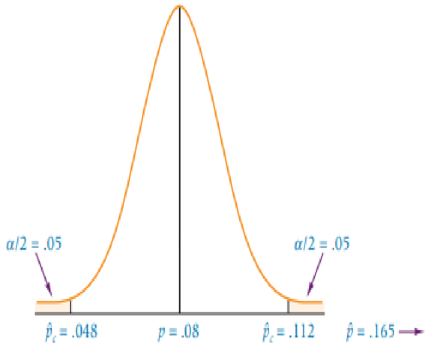
Distribution with Rejection Regions for Flawed-Product Example



The observed value of z is in the rejection region (observed z = 4.43> table z $_{.05}$ = +1.645), so the business researcher rejects the null hypothesis. He concludes that the proportion of items with at least one minor flaw in the population from which the sample of 200 was drawn is not .08. With α=.10 , the risk of committing a Type I error in this example is .10 .

FIGURE 9.17

Distribution Using Critical Value Method for the Flawed-Product Example



$$z_{\alpha/2} = \frac{\hat{p}_c - p}{\sqrt{\frac{p \cdot q}{n}}}$$

$$\pm 1.645 = \frac{\hat{p}_c - .08}{\sqrt{\frac{(.08)(.92)}{200}}}$$

$$\hat{p}_c = .08 \pm 1.645\sqrt{\frac{(.08)(.92)}{200}} = .08 \pm .032$$

$$= .048 \text{ and } .112$$

A survey of the morning beverage market shows that the primary breakfast beverage for 17% of Americans is milk. A milk producer in Wisconsin, where milk is plentiful, believes the figure is higher for Wisconsin. To test this idea, she contacts a random sample of 550 Wisconsin residents and asks which primary beverage they consumed for breakfast that day. Suppose 115 replied that milk was the primary beverage. Using a level of significance of .05, test the idea that the milk figure is higher for Wisconsin.

- HYPOTHESIZE:
  STEP 1. The milk producer's theory is that the proportion of Wisconsin residents who drink milk for breakfast is higher than the national proportion, which is the alternative hypothesis. The null hypothesis is that the proportion in Wisconsin does not differ from the national average. The hypotheses for this problem are

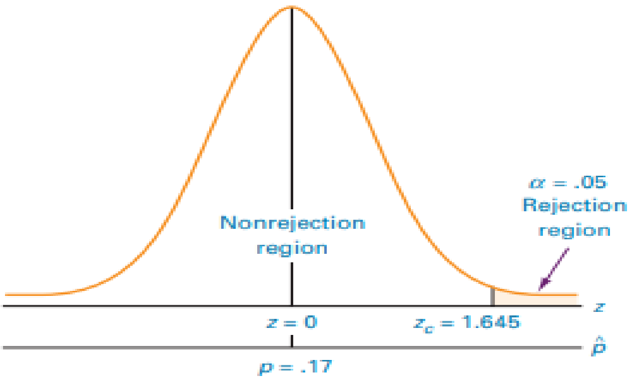
$$H_0: p = .17$$
$$H_a: p > .17$$

STEP 2. The test statistic is

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

STEP 3. The Type I error rate is .05.

STEP 4. This test is a one-tailed test, and the table value is $z_{.05}$ =+1.645The sample results must yield an observed z value greater than 1.645 for the milk producer to reject the null hypothesis. The following diagram shows $z_{.05}$ and the rejection region for this problem.



STEP 5. $n = 550$ and $x = 115$

$$\hat{p} = \frac{115}{550} = .209$$

STEP 6.

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.209 - .17}{\sqrt{\frac{(.17)(.83)}{550}}} = \frac{.039}{.016} = 2.44$$

ACTION:

STEP 7. Because z=2.44 is beyond $z_{.05}$ =1.645 in the rejection region, the milk producer rejects the null hypothesis. The probability of obtaining a z>=2.44 by chance is .0073.

BUSINESS IMPLICATIONS:

STEP 8. If the proportion of residents who drink milk for breakfast is higher in Wisconsin than in other parts of the United States.

$$z_{.05} = \frac{\hat{p}_c - p}{\sqrt{\frac{p \cdot q}{n}}}$$

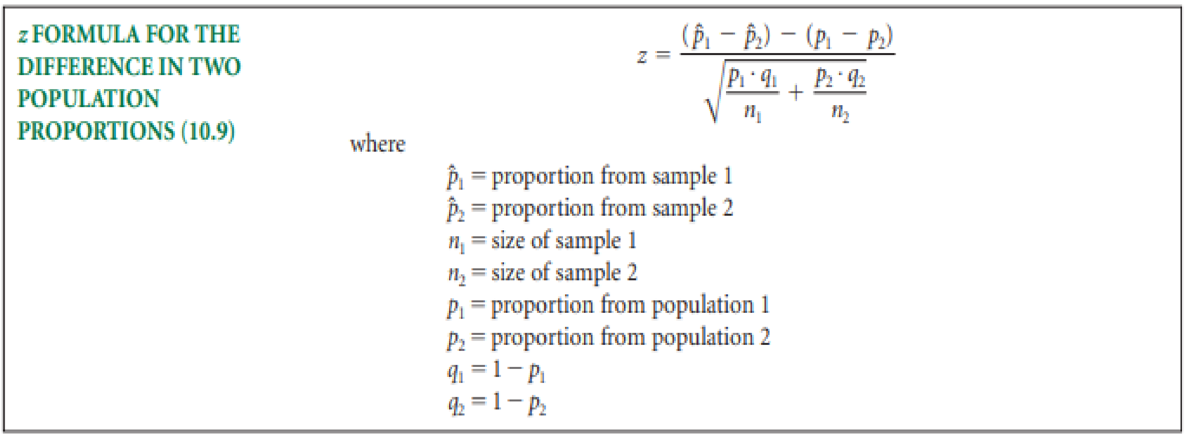
$$1.645 = \frac{\hat{p}_c - .17}{\sqrt{\frac{(.17)(.83)}{550}}}$$

$$\hat{p}_c = .17 + 1.645\sqrt{\frac{(.17)(.83)}{550}} = .17 + .026 = .196$$

Two Population Proportions, p_1-p_2

Sometimes a researcher wishes to make inferences about the difference in two population proportions. This type of analysis has many applications in business, such as comparing the market share of a product for two different markets, studying the difference in the proportion of female customers in two different geographic regions, or comparing the proportion of defective products from one period to another.

$\hat{p}_1$-$\hat{p}_2$ This statistic is computed by taking random samples and determining **$\hat{p}$** for each sample for a given characteristic, then calculating the difference in these sample proportions.

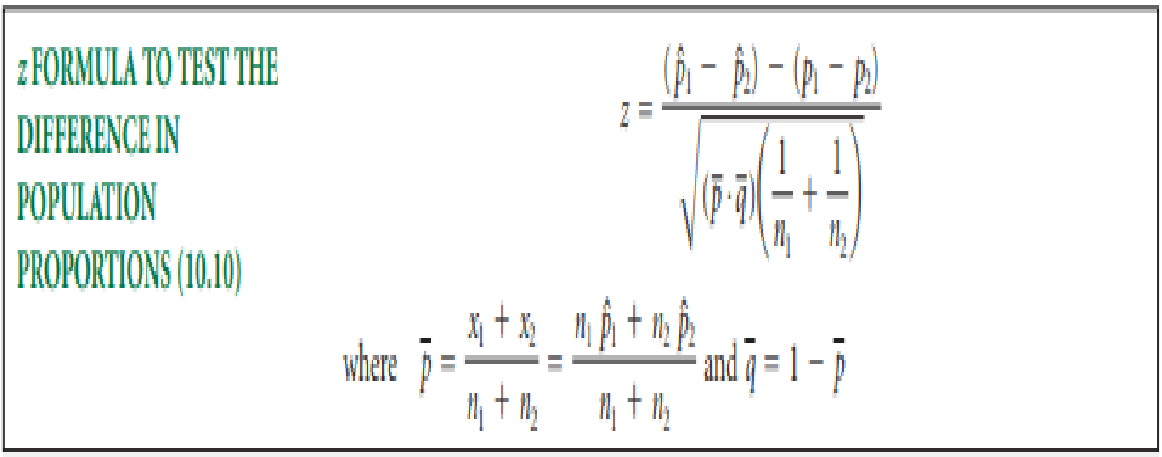


z FORMULA FOR THE DIFFERENCE IN TWO POPULATION PROPORTIONS (10.9)

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 \cdot q_1}{n_1} + \frac{p_2 \cdot q_2}{n_2}}}$$

where

$\hat{p}_1$ = proportion from sample 1
$\hat{p}_2$ = proportion from sample 2
n_1 = size of sample 1
n_2 = size of sample 2
p_1 = proportion from population 1
p_2 = proportion from population 2
$q_1 = 1 - p_1$
$q_2 = 1 - p_2$



z FORMULA TO TEST THE DIFFERENCE IN POPULATION PROPORTIONS (10.10)

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{(\bar{p} \cdot \bar{q})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ and $\bar{q} = 1 - \bar{p}$

A group of researchers attempted to determine whether there was a difference in the proportion of consumers and the proportion of CEOs who believe that fear of getting caught or losing one’s job is a strong influence of ethical behavior. In their study, they found that 57% of consumers said that fear of getting caught or losing one’s job was a strong influence on ethical behavior, but only 50% of CEOs felt the same way.

Suppose these data were determined from a sample of 755 consumers and 616 CEOs. Does this result provide enough evidence to declare that a significantly higher proportion of consumers than of CEOs believe fear of getting caught or losing one's job is a strong influence on ethical behavior? α=0.10

HYPOTHESIZE: STEP 1. Suppose sample 1 is the consumer sample and sample 2 is the CEO sample. Because we are trying to prove that a higher proportion of consumers than of CEOs believe fear of getting caught or losing one's job is a strong influence on ethical behavior, the alternative hypothesis should be $p_1>p_2=0$.The following hypotheses are being tested

$$H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 > 0$$

Where

p1 is the proportion of consumers who select the factor

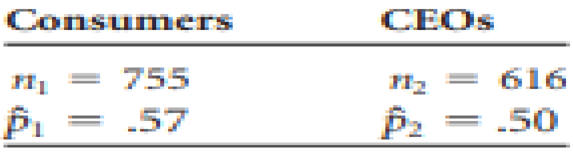
p2 is the proportion of CEOs who select the factor TEST:

STEP 2. The appropriate statistical test is formula 10.10.

STEP 3. Let α=0.10

STEP 4. Because this test is a one-tailed test, the critical table z value is $z_{0.10}$ =1.28. If an observed value of z of more than 1.28 is obtained, the null hypothesis will be rejected. Figure 10.12 shows the rejection region and the critical value for this problem.

STEP 5. The sample information follows



Consumers	CEOs
$n_1 = 755$	$n_2 = 616$
$\hat{p}_1 = .57$	$\hat{p}_2 = .50$

STEP 6.

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{(755)(.57) + (616)(.50)}{755 + 616} = .539$$

If the statistics had been given as raw data instead of sample proportions, we would have used the following formula.

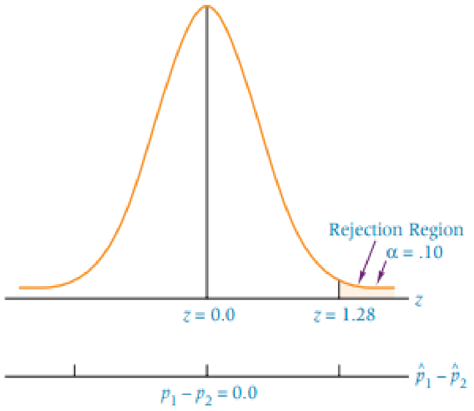
$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

The observed z value is

$$z = \frac{(.57 - .50) - (0)}{\sqrt{(.539)(.461)\left(\frac{1}{755} + \frac{1}{616}\right)}} = 2.59$$

FIGURE 10.12

Rejection Region for the Ethics Example



ACTION: STEP 7. Because z =2.59 is greater than the critical table z value of 1.28 and is in the rejection region, the null hypothesis is rejected.

BUSINESS IMPLICATIONS: STEP 8. A significantly higher proportion of consumers than of CEOs believe fear of getting caught or losing one’s job is a strong influence on ethical behavior

Suppose you decide to test this result by taking a survey of your own and identify female entrepreneurs by gross sales. You interview 100 female entrepreneurs with gross sales of less than $100,000, and 24 of them define sales/profit as success. You then interview 95 female entrepreneurs with gross sales of $100,000 to $500,000, and 39 cite sales/profit as a definition of success. Use this information to test to determine whether there is a significant difference in the proportions of the two groups that define success as sales/profit. Use α=.01

HYPOTHESIZE:

STEP 1. You are testing to determine whether there is a difference between two groups of entrepreneurs, so a two-tailed test is required. The hypotheses follow.

$$H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 \neq 0$$

STEP 2. The appropriate statistical test is formula 10.10.

STEP 3. Alpha has been specified as .01.

STEP 4. With $\alpha = .01$, you obtain a critical z value from Table A.5 for $\alpha/2 = .005$, $z_{.005} = \pm 2.575$. If the observed z value is more than 2.575 or less than −2.575, the null hypothesis is rejected.

STEP 5. The sample information follows.

Less than $100,000	$100,000 to $500,000
$n_1 = 100$	$n_2 = 95$
$x_1 = 24$	$x_2 = 39$
$\hat{p}_1 = \frac{24}{100} = .24$	$\hat{p}_2 = \frac{39}{95} = .41$

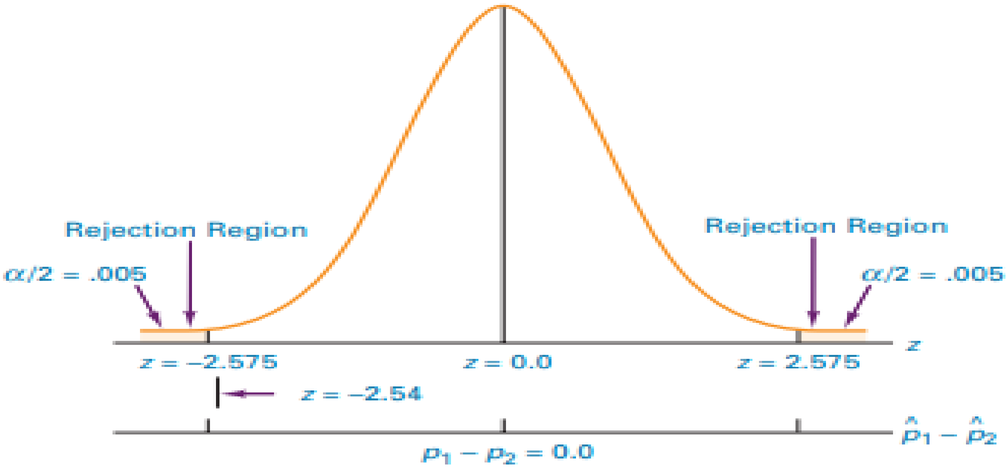
where

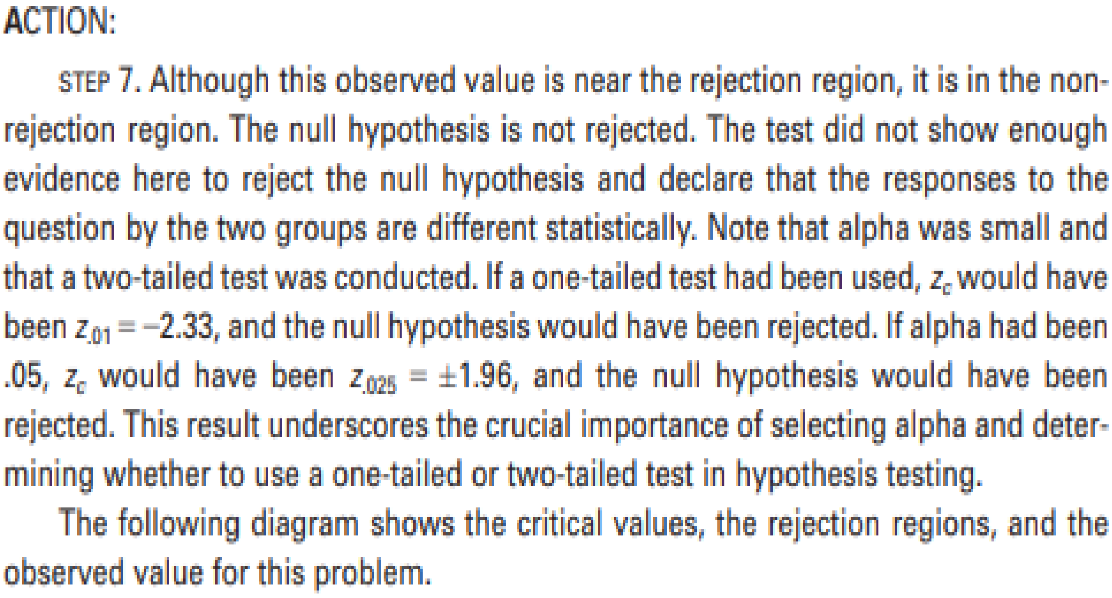
$$\overline{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{24 + 39}{100 + 95} = \frac{63}{195} = .323$$

x = the number of entrepreneurs who define sales/profits as success

STEP 6. The observed z value is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{(\overline{p} \cdot \overline{q})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(.24 - .41) - (0)}{\sqrt{(.323)(.677)\left(\frac{1}{100} + \frac{1}{95}\right)}} = \frac{-.17}{.067} = -2.54$$





ACTION:

STEP 7. Although this observed value is near the rejection region, it is in the nonrejection region. The null hypothesis is not rejected. The test did not show enough evidence here to reject the null hypothesis and declare that the responses to the question by the two groups are different statistically. Note that alpha was small and that a two-tailed test was conducted. If a one-tailed test had been used, z_c would have been $z_{.01} = -2.33$, and the null hypothesis would have been rejected. If alpha had been .05, z_c would have been $z_{.025} = \pm 1.96$, and the null hypothesis would have been rejected. This result underscores the crucial importance of selecting alpha and determining whether to use a one-tailed or two-tailed test in hypothesis testing.

The following diagram shows the critical values, the rejection regions, and the observed value for this problem.

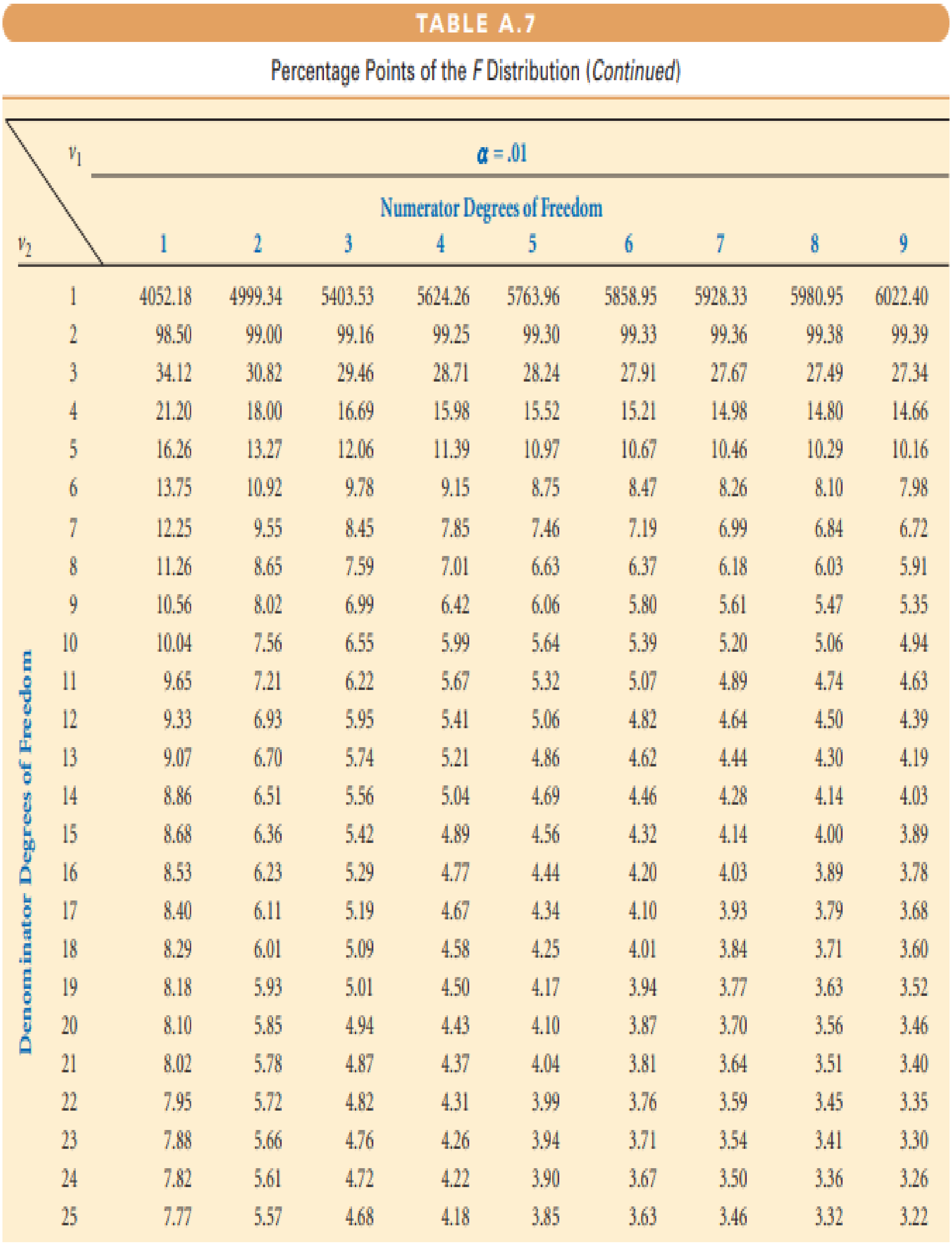


TABLE A.7

Percentage Points of the *F* Distribution (*Continued*)

v_2 \ v_1	$\alpha = .01$								
	Numerator Degrees of Freedom								
Denominator Degrees of Freedom	1	2	3	4	5	6	7	8	9
1	4052.18	4999.34	5403.53	5624.26	5763.96	5858.95	5928.33	5980.95	6022.40
2	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22

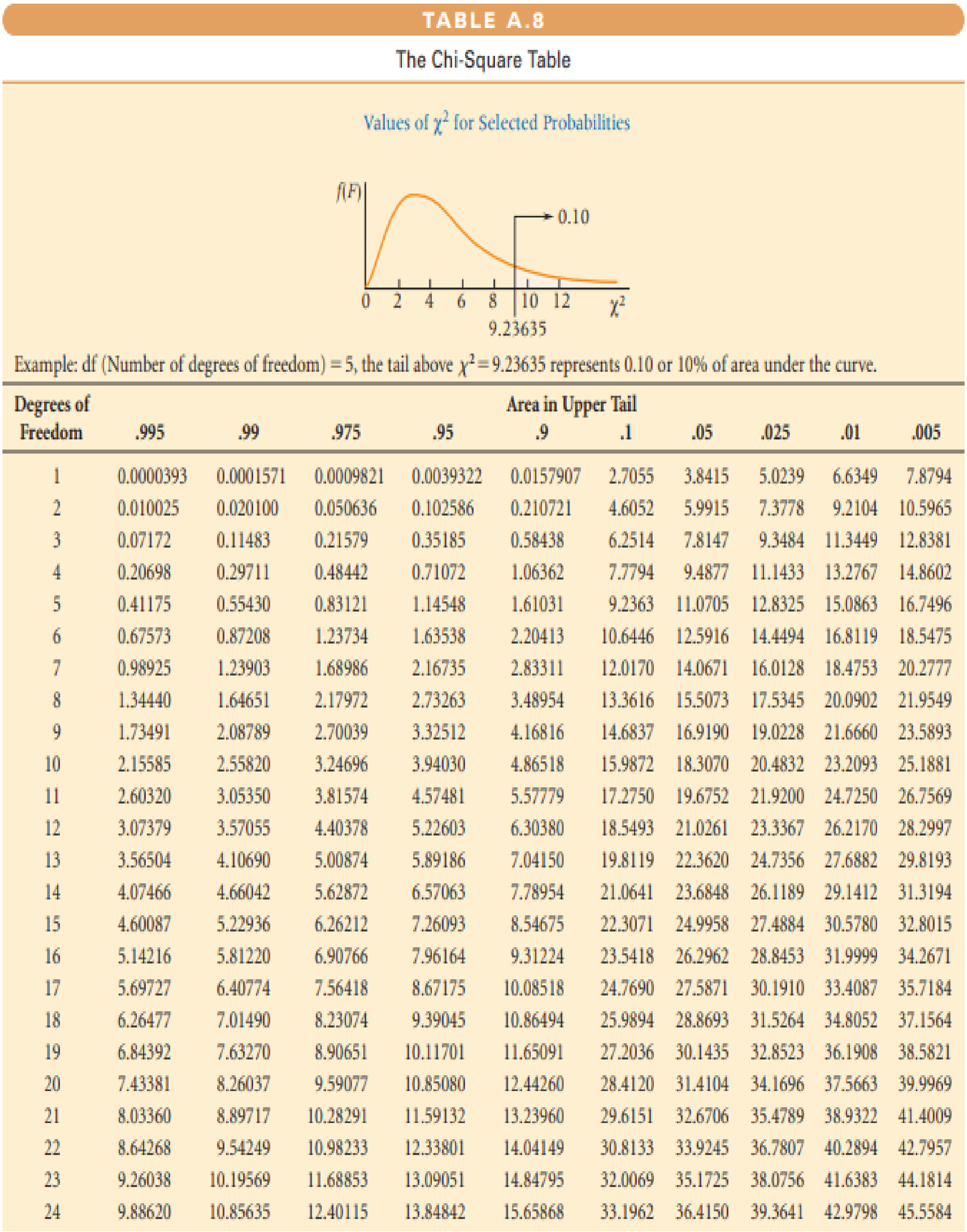


TABLE A.8

The Chi-Square Table

Values of χ^2 for Selected Probabilities

f(F)
0.10
0
2
4
6
8
10
12
χ^2
9.23635

Example: df (Number of degrees of freedom) = 5, the tail above $\chi^2 = 9.23635$ represents 0.10 or 10% of area under the curve.

Degrees of					Area in Upper Tail					
Freedom	.995	.99	.975	.95	.9	.1	.05	.025	.01	.005
1	0.0000393	0.0001571	0.0009821	0.0039322	0.0157907	2.7055	3.8415	5.0239	6.6349	7.8794
2	0.010025	0.020100	0.050636	0.102586	0.210721	4.6052	5.9915	7.3778	9.2104	10.5965
3	0.07172	0.11483	0.21579	0.35185	0.58438	6.2514	7.8147	9.3484	11.3449	12.8381
4	0.20698	0.29711	0.48442	0.71072	1.06362	7.7794	9.4877	11.1433	13.2767	14.8602
5	0.41175	0.55430	0.83121	1.14548	1.61031	9.2363	11.0705	12.8325	15.0863	16.7496
6	0.67573	0.87208	1.23734	1.63538	2.20413	10.6446	12.5916	14.4494	16.8119	18.5475
7	0.98925	1.23903	1.68986	2.16735	2.83311	12.0170	14.0671	16.0128	18.4753	20.2777
8	1.34440	1.64651	2.17972	2.73263	3.48954	13.3616	15.5073	17.5345	20.0902	21.9549
9	1.73491	2.08789	2.70039	3.32512	4.16816	14.6837	16.9190	19.0228	21.6660	23.5893
10	2.15585	2.55820	3.24696	3.94030	4.86518	15.9872	18.3070	20.4832	23.2093	25.1881
11	2.60320	3.05350	3.81574	4.57481	5.57779	17.2750	19.6752	21.9200	24.7250	26.7569
12	3.07379	3.57055	4.40378	5.22603	6.30380	18.5493	21.0261	23.3367	26.2170	28.2997
13	3.56504	4.10690	5.00874	5.89186	7.04150	19.8119	22.3620	24.7356	27.6882	29.8193
14	4.07466	4.66042	5.62872	6.57063	7.78954	21.0641	23.6848	26.1189	29.1412	31.3194
15	4.60087	5.22936	6.26212	7.26093	8.54675	22.3071	24.9958	27.4884	30.5780	32.8015
16	5.14216	5.81220	6.90766	7.96164	9.31224	23.5418	26.2962	28.8453	31.9999	34.2671
17	5.69727	6.40774	7.56418	8.67175	10.08518	24.7690	27.5871	30.1910	33.4087	35.7184
18	6.26477	7.01490	8.23074	9.39045	10.86494	25.9894	28.8693	31.5264	34.8052	37.1564
19	6.84392	7.63270	8.90651	10.11701	11.65091	27.2036	30.1435	32.8523	36.1908	38.5821
20	7.43381	8.26037	9.59077	10.85080	12.44260	28.4120	31.4104	34.1696	37.5663	39.9969
21	8.03360	8.89717	10.28291	11.59132	13.23960	29.6151	32.6706	35.4789	38.9322	41.4009
22	8.64268	9.54249	10.98233	12.33801	14.04149	30.8133	33.9245	36.7807	40.2894	42.7957
23	9.26038	10.19569	11.68853	13.09051	14.84795	32.0069	35.1725	38.0756	41.6383	44.1814
24	9.88620	10.85635	12.40115	13.84842	15.65868	33.1962	36.4150	39.3641	42.9798	45.5584

Correlation and Regression

Correlation

Meaning:

- **Correlation is a statistical technique to ascertain the association or relationship between two or more variables.**
- **Correlation analysis is a statistical technique to study the degree and direction of relationship between two or more variables.**
- **A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other.**

Uses of correlations:

1. Correlation analysis helps inn deriving precisely the degree and the direction of such relationship.
2. The effect of correlation is to reduce the range of uncertainty of our prediction. The prediction based on correlation analysis will be more reliable and near to reality.
3. The measure of coefficient of correlation is a relative measure of change.

Types of Correlation:

Correlation is described or classified in several different ways. Three of the most important are:

I. Positive and Negative

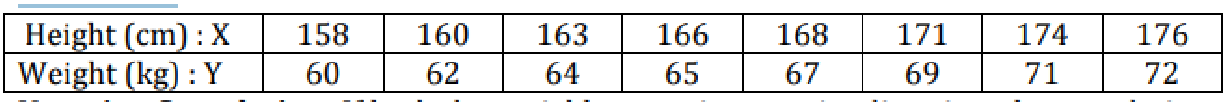
II. Simple, Partial and Multiple

III. Linear and non-linear

I. Positive, Negative and Zero Correlation:

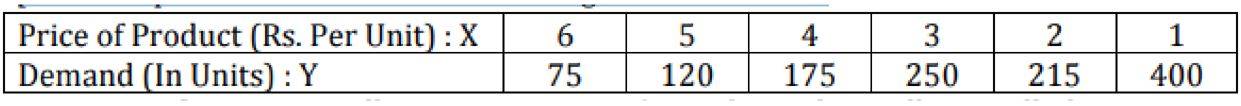
Whether correlation is positive (direct) or negative (in-versa) would depend upon the direction of change of the variable.

Positive Correlation: If both the variables vary in the same direction, correlation is said to be positive. It means if one variable is increasing, the other on an average is also increasing or if one variable is decreasing, the other on an average is also deceasing, then the correlation is said to be positive correlation. For example, the correlation between heights and weights of a group of persons is a positive correlation.



Height (cm) : X	158	160	163	166	168	171	174	176
Weight (kg) : Y	60	62	64	65	67	69	71	72

Negative Correlation: If both the variables vary in opposite direction, the correlation is said to be negative. If means if one variable increases, but the other variable decreases or if one variable decreases, but the other variable increases, then the correlation is said to be negative correlation. For example, the correlation between the price of a product and its demand is a negative correlation.



Price of Product (Rs. Per Unit) : X	6	5	4	3	2	1
Demand (In Units) : Y	75	120	175	250	215	400

Zero Correlation: Actually it is not a type of correlation but still it is called as zero or no correlation. When we don't find any relationship between the variables then, it is said to be zero correlation. It means a change in value of one variable doesn't influence or change the value of other variable. For example, the correlation between weight of person and intelligence is a zero or no correlation.

II. Simple, Partial and Multiple Correlations:

The distinction between simple, partial and multiple correlation is based upon the number of variables studied.

Simple Correlation: When only two variables are studied, it is a case of simple correlation. For example, when one studies relationship between the marks secured by student and the attendance of student in class, it is a problem of simple correlation.

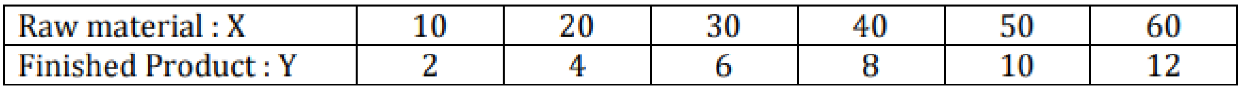
Partial Correlation: In case of partial correlation one studies three or more variables but considers only two variables to be influencing each other and the effect of other influencing variables being held constant. For example, in above example of relationship between student marks and attendance, the other variable influencing such as effective teaching of teacher, use of teaching aid like computer, smart board etc are assumed to be constant.

Multiple Correlations: When three or more variables are studied, it is a case of multiple correlation. For example, in above example if study covers the relationship between student marks, attendance of students, effectiveness of teacher, use of teaching aids etc, it is a case of multiple correlation.

III. Linear and Non-linear Correlation:

Depending upon the constancy of the ratio of change between the variables, the correlation may be Linear or Non-linear Correlation.

Linear Correlation: If the amount of change in one variable bears a constant ratio to the amount of change in the other variable, then correlation is said to be linear. If such variables are plotted on a graph paper all the plotted points would fall on a straight line. For example: If it is assumed that, to produce one unit of finished product we need 10 units of raw materials, then subsequently to produce 2 units of finished product we need double of the one unit.



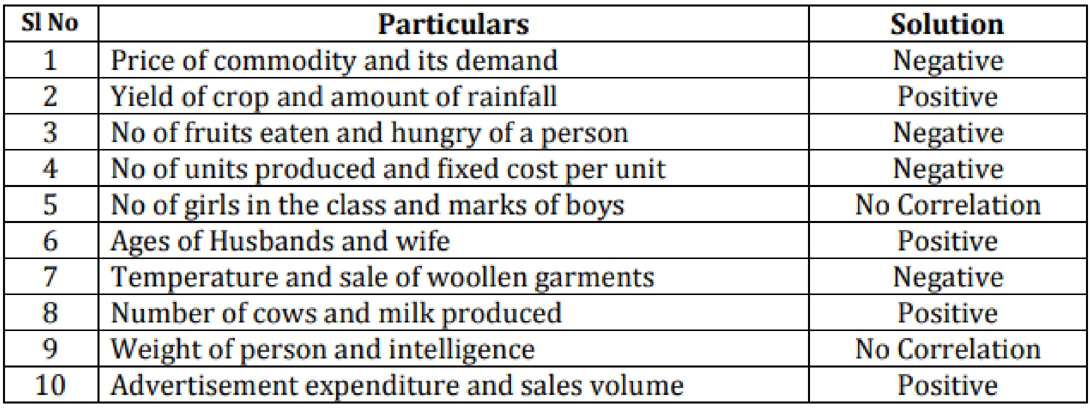
Raw material : X	10	20	30	40	50	60
Finished Product : Y	2	4	6	8	10	12

Non-linear Correlation: If the amount of change in one variable does not bear a constant ratio to the amount of change to the other variable, then correlation is said to be non-linear. If such variables are plotted on a graph, the points would fall on a curve and not on a straight line. For example, if we double the amount of advertisement expenditure, then sales volume would not necessarily be doubled.

ADVERTISEMENT EXPENSES: X	**10**	**20**	**30**	**40**	**50**
SALES VOLUME: Y	**2**	**5**	**8**	**4**	**12**

Illustration 01:

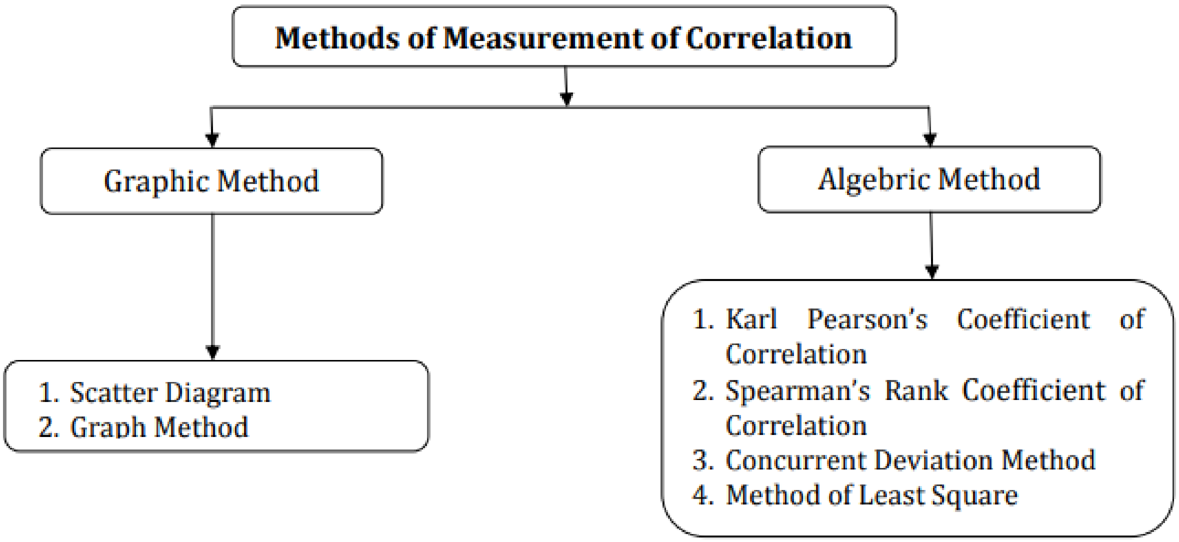
State in each case whether there is (a) Positive Correlation (b) Negative Correlation (c) No Correlation



Sl No	Particulars	Solution
1	Price of commodity and its demand	Negative
2	Yield of crop and amount of rainfall	Positive
3	No of fruits eaten and hungry of a person	Negative
4	No of units produced and fixed cost per unit	Negative
5	No of girls in the class and marks of boys	No Correlation
6	Ages of Husbands and wife	Positive
7	Temperature and sale of woollen garments	Negative
8	Number of cows and milk produced	Positive
9	Weight of person and intelligence	No Correlation
10	Advertisement expenditure and sales volume	Positive

Methods of measurement of correlation:

Quantification of the relationship between variables is very essential to take the benefit of study of correlation. For this, we find there are various methods of measurement of correlation, which can be represented as given below:


Methods of Measurement of Correlation
Graphic Method
Algebric Method
1. Scatter Diagram
2. Graph Method
1. Karl Pearson's Coefficient of Correlation
2. Spearman's Rank Coefficient of Correlation
3. Concurrent Deviation Method
4. Method of Least Square

Among these methods we will discuss only the following methods:

1. Scatter Diagram

Scatter Diagram:

This is graphic method of measurement of correlation. It is a diagrammatic representation of bivariate data to ascertain the relationship between two variables. Under this method the given data are plotted on a graph paper in the form of dot. i.e. for each pair of X and Y values we put dots and thus obtain as many points as the number of observations. Usually an independent variable is shown on the X-axis whereas the dependent variable is shown on the Y-axis. Once the values are plotted on the graph it reveals the type of the correlation between variable X and Y. A scatter diagram reveals whether the movements in one series are associated with those in the other series.

• Perfect Positive Correlation: In this case, the points will form on a straight line rising from the lower left hand corner to the upper right hand corner.

• Perfect Negative Correlation: In this case, the points will form on a straight line declining from the upper left hand corner to the lower right hand corner.

- High Degree of Positive Correlation: In this case, the plotted points fall in a narrow band, wherein points show a rising tendency from the lower left hand corner to the upper right hand corner.
- High Degree of Negative Correlation: In this case, the plotted points fall in a narrow band, wherein points show a declining tendency from upper left hand corner to the lower right hand corner.
- Low Degree of Positive Correlation: If the points are widely scattered over the diagrams, wherein points are rising from the left hand corner to the upper right hand corner.
- Low Degree of Negative Correlation: If the points are widely scattered over the diagrams, wherein points are declining from the upper left hand corner to the lower right hand corner.
- Zero (No) Correlation: When plotted points are scattered over the graph haphazardly, then it indicate that there is no correlation or zero correlation between two variables.

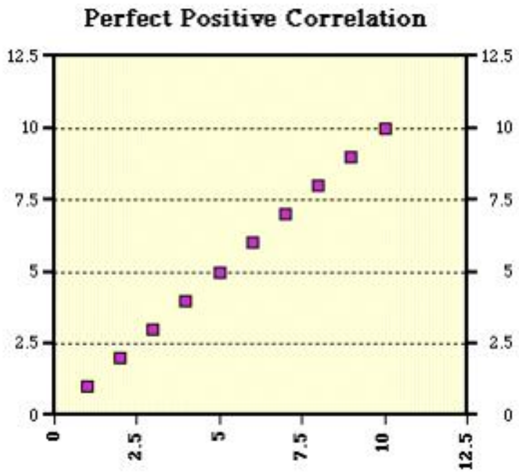


Diagram – I

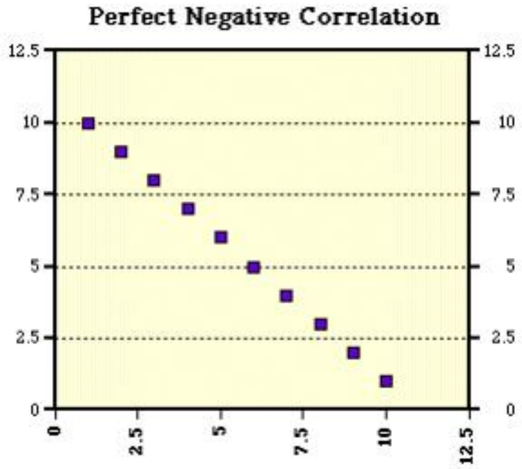


Diagram – II

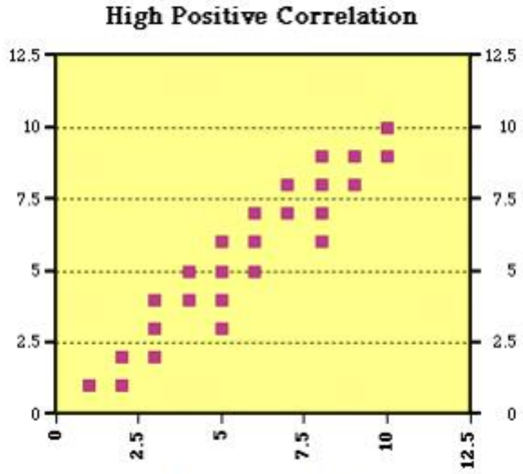


Diagram – III

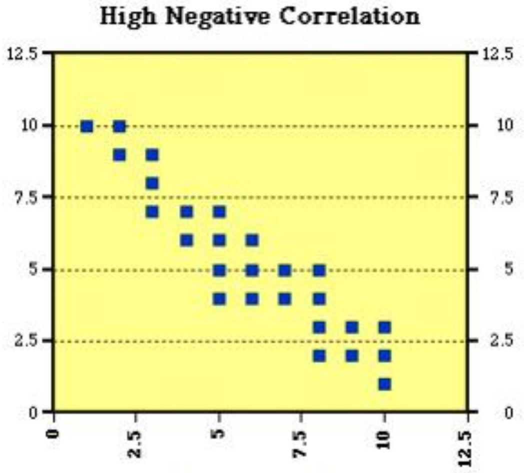


Diagram – IV

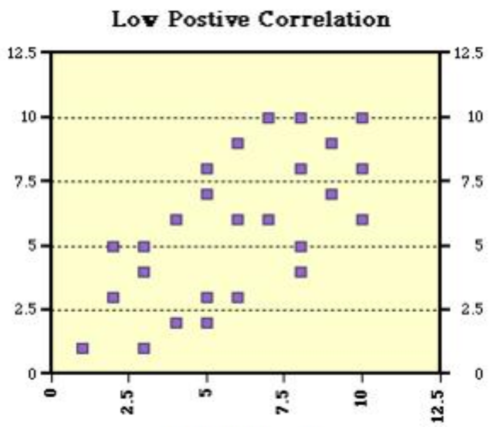


Diagram – V

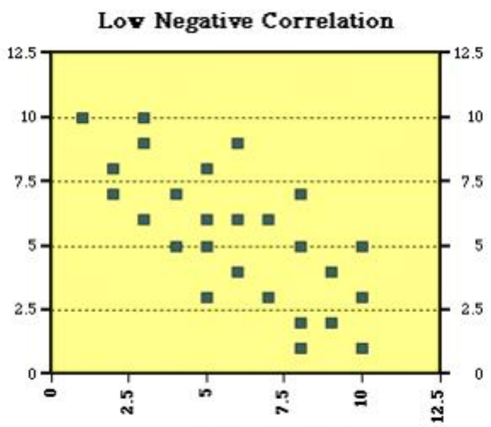


Diagram – VI

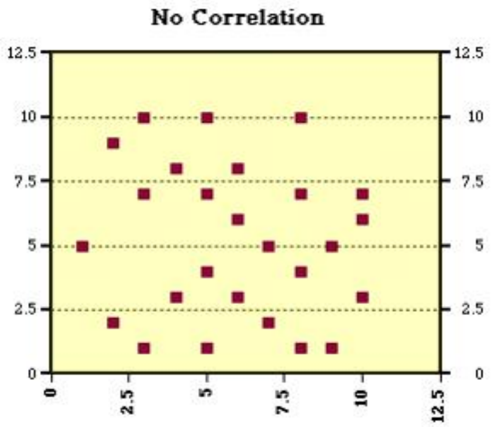


Diagram – VII

Illustration 02:

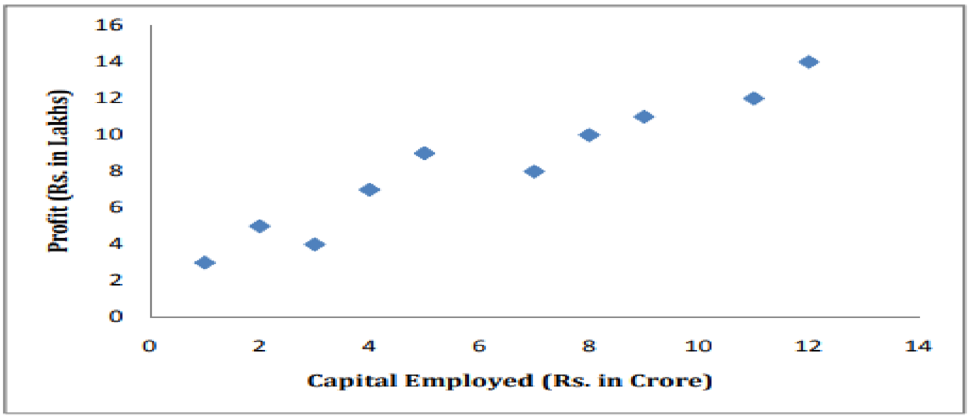
Given the following pairs of values:

Capital Employed (Rs. In Crore)	1	2	3	4	5	7	8	9	11	12
Profit (Rs. In Lakhs)	3	5	4	7	9	8	10	11	12	14

(a)Draw a scatter diagram (b)Do you think that there is any correlation between profits and capital employed? Is it positive or negative? Is it high or low?

Solution:

From the observation of scatter diagram we can say that the variables are positively correlated. In the diagram the points trend toward upward rising from the lower left hand corner to the upper right hand corner, hence it is positive correlation. Plotted points are in narrow band which indicates that it is a case of high degree of positive correlation.



Correlation coefficient

DIRECT METHOD

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} \quad \text{where} \quad X = x - \bar{x}, \quad Y = y - \bar{y}$$

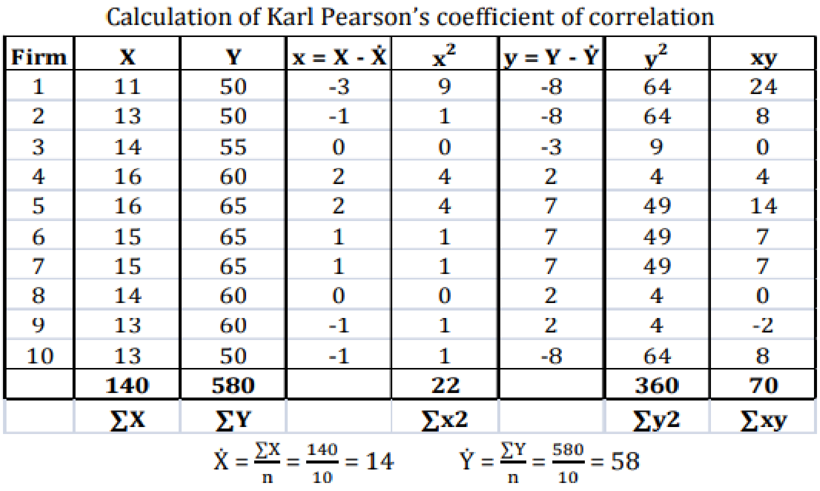
- Interpretation of coefficient of correlation

1. A positive value of γ indicates positive correlation.
2. A Negative value of γ indicates negative correlation.
3. If γ =+1then the correlation is perfect positive
4. If γ =-1then the correlation is perfect negative.
5. γ =0 then the variables are non-correlated.
6. If γ >=0.5 then correlation will be high degree of positive.
7. If γ >=-0.5 then correlation will be high degree of negative
8. If γ <0.5 then correlation will be Low degree of positive.
9. If γ<-0.5 then correlation will be low degree of negative

DIRECT METHOD

From following information find the correlation coefficient between advertisement expenses and sales volume using Karl Pearson's coefficient of correlation method.

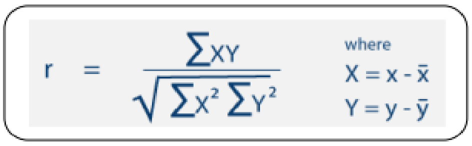
Firm	1	2	3	4	5	6	7	8	9	10
Advertisement Exp. (Rs. In Lakhs)	11	13	14	16	16	15	15	14	13	13
Sales Volume (Rs. In Lakhs)	50	50	55	60	65	65	65	60	60	50



Calculation of Karl Pearson's coefficient of correlation

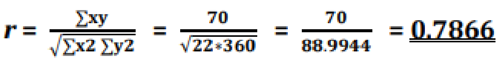
Firm	X	Y	x = X - Ẋ	x^2	y = Y - Ẏ	y^2	xy
1	11	50	-3	9	-8	64	24
2	13	50	-1	1	-8	64	8
3	14	55	0	0	-3	9	0
4	16	60	2	4	2	4	4
5	16	65	2	4	7	49	14
6	15	65	1	1	7	49	7
7	15	65	1	1	7	49	7
8	14	60	0	0	2	4	0
9	13	60	-1	1	2	4	-2
10	13	50	-1	1	-8	64	8
	140	**580**		**22**		**360**	**70**
	∑X	**∑Y**		**∑x2**		**∑y2**	**∑xy**

$\dot{X} = \frac{\sum X}{n} = \frac{140}{10} = 14$ $\dot{Y} = \frac{\sum Y}{n} = \frac{580}{10} = 58$

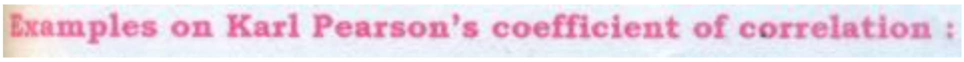


$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

where
$X = x - \bar{x}$
$Y = y - \bar{y}$

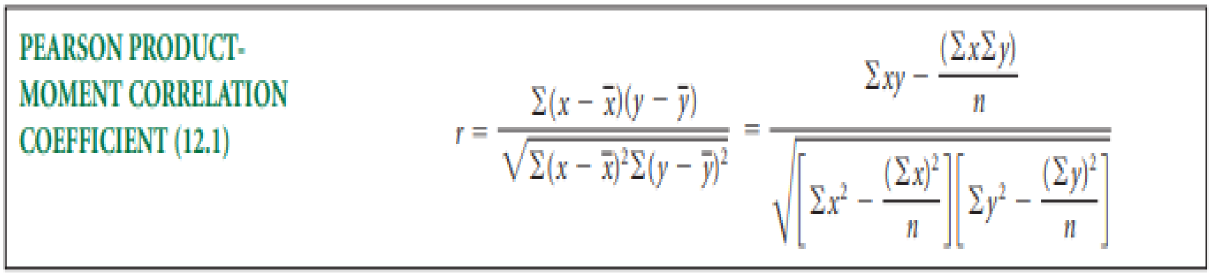


$$r = \frac{\sum xy}{\sqrt{\sum x2 \sum y2}} = \frac{70}{\sqrt{22 * 360}} = \frac{70}{88.9944} = \underline{0.7866}$$



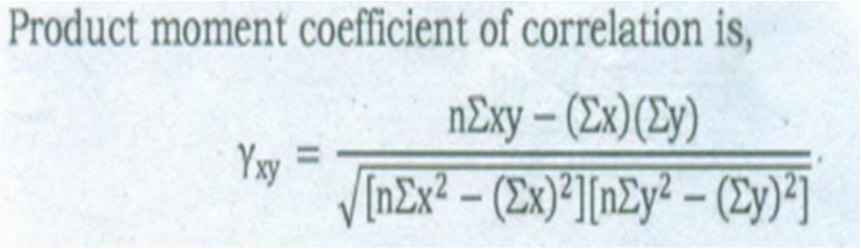
Examples on Karl Pearson's coefficient of correlation :

PRODUCT MOMENT COEFFICIENT OF CORRELATION



PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT (12.1)

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

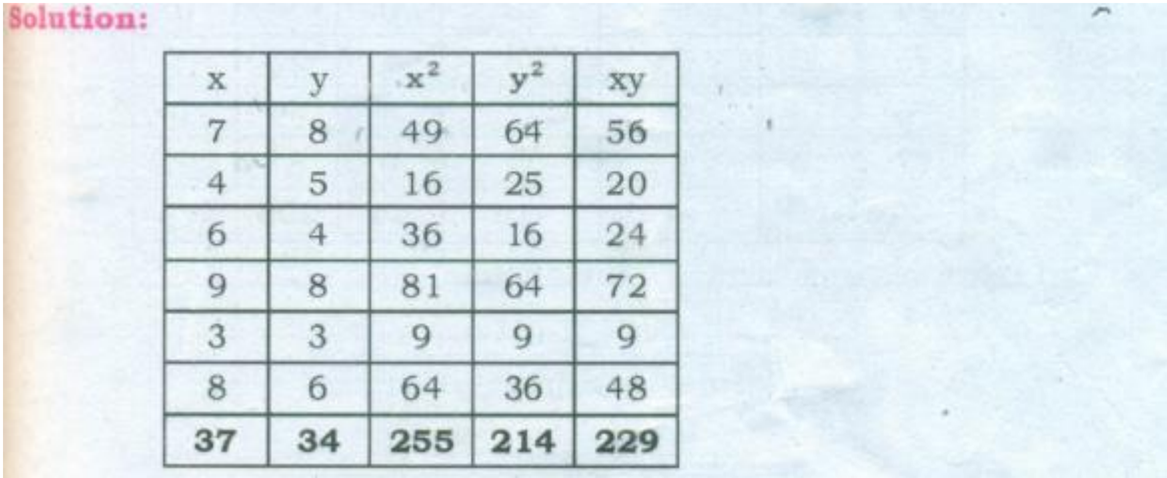


Product moment coefficient of correlation is,

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

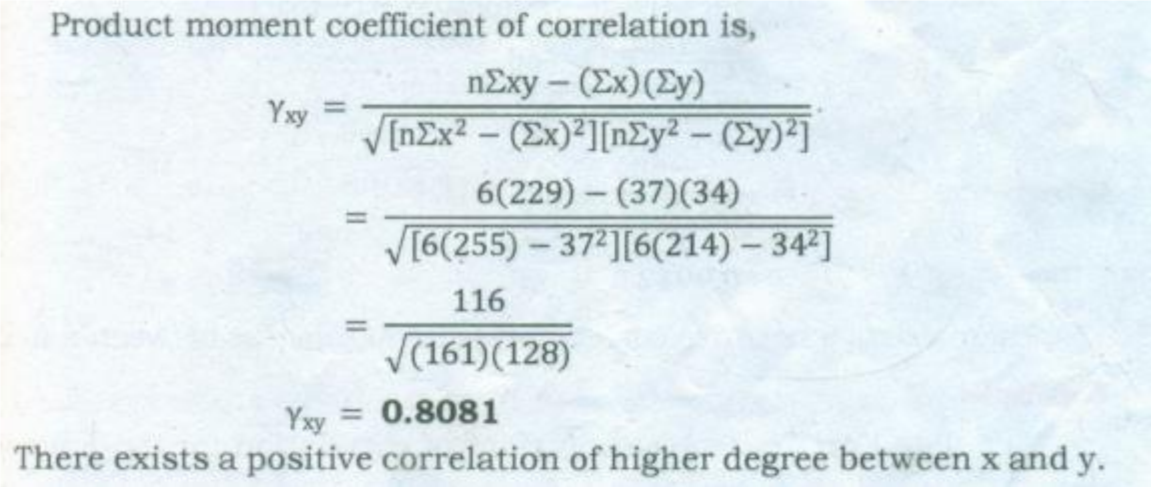
Calculate the Karl Pearson’s product moment of coefficient of correlation.

Student	1	2	3	4	5	6
Statistics(x)	7	4	6	9	3	8
Mathematics(y)	8	5	4	8	3	6



Solution:

x	y	x^2	y^2	xy
7	8	49	64	56
4	5	16	25	20
6	4	36	16	24
9	8	81	64	72
3	3	9	9	9
8	6	64	36	48
37	**34**	**255**	**214**	**229**



Product moment coefficient of correlation is,

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

$$= \frac{6(229) - (37)(34)}{\sqrt{[6(255) - 37^2][6(214) - 34^2]}}$$

$$= \frac{116}{\sqrt{(161)(128)}}$$

$$\gamma_{xy} = \mathbf{0.8081}$$

There exists a positive correlation of higher degree between x and y.

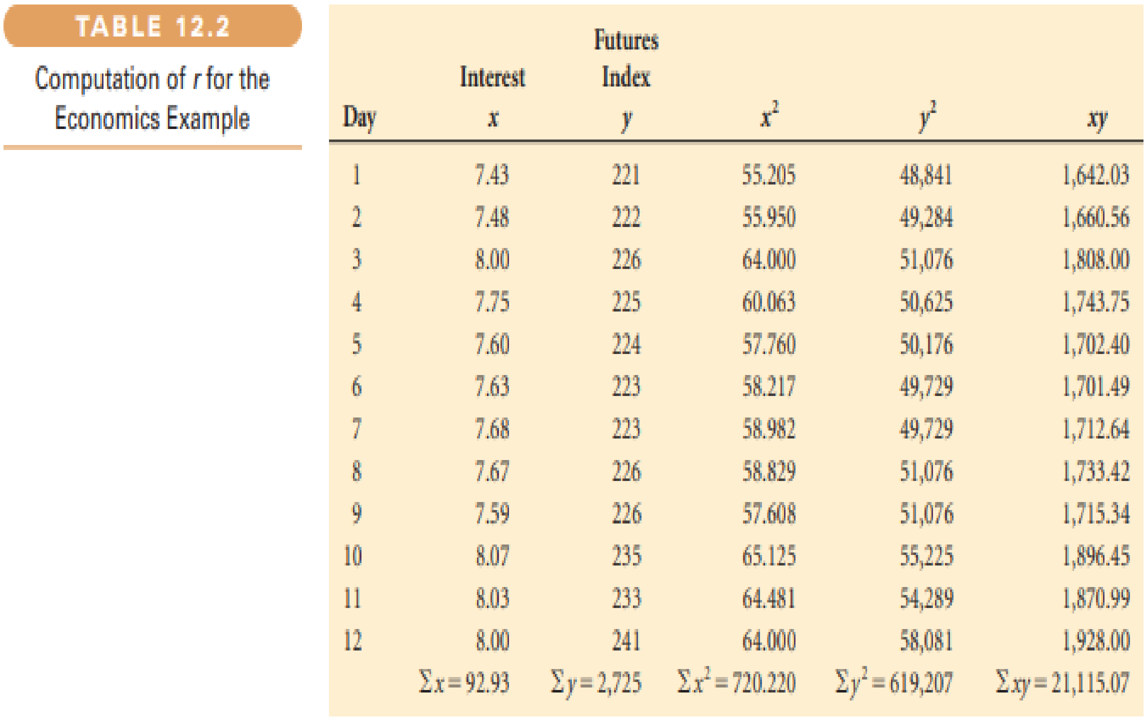
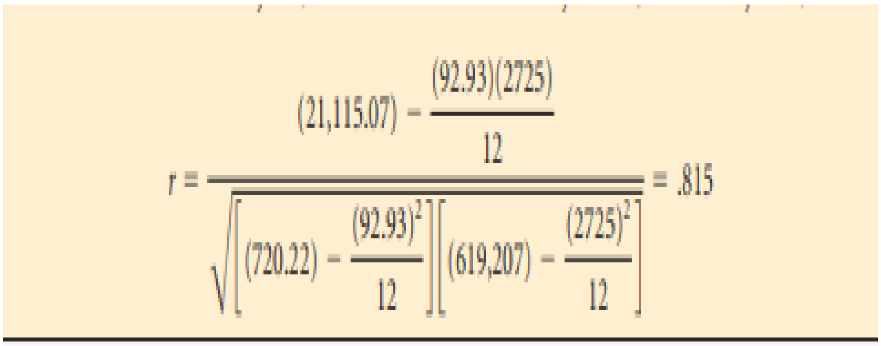


TABLE 12.2

Computation of *r* for the Economics Example

Day	Interest *x*	Futures Index *y*	x^2	y^2	*xy*
1	7.43	221	55.205	48,841	1,642.03
2	7.48	222	55.950	49,284	1,660.56
3	8.00	226	64.000	51,076	1,808.00
4	7.75	225	60.063	50,625	1,743.75
5	7.60	224	57.760	50,176	1,702.40
6	7.63	223	58.217	49,729	1,701.49
7	7.68	223	58.982	49,729	1,712.64
8	7.67	226	58.829	51,076	1,733.42
9	7.59	226	57.608	51,076	1,715.34
10	8.07	235	65.125	55,225	1,896.45
11	8.03	233	64.481	54,289	1,870.99
12	8.00	241	64.000	58,081	1,928.00
	$\Sigma x = 92.93$	$\Sigma y = 2{,}725$	$\Sigma x^2 = 720.220$	$\Sigma y^2 = 619{,}207$	$\Sigma xy = 21{,}115.07$



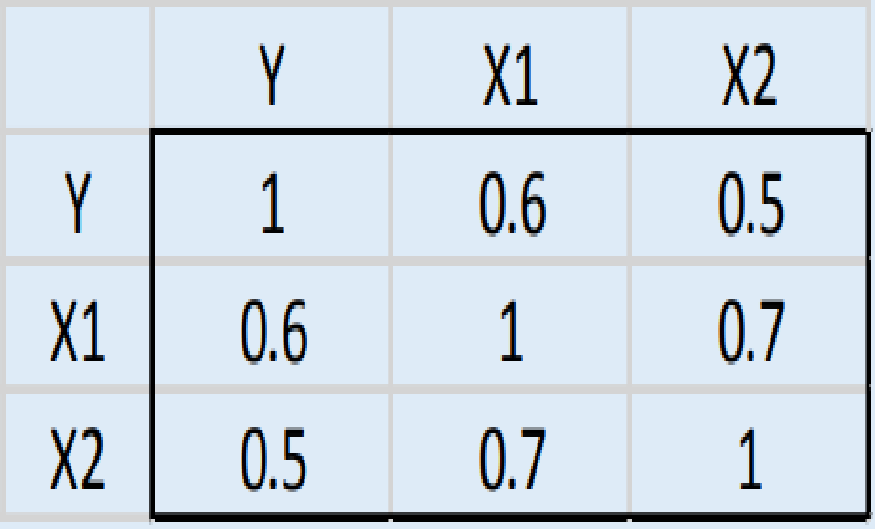
$$r = \frac{(21{,}115.07) - \dfrac{(92.93)(2725)}{12}}{\sqrt{\left[(720.22) - \dfrac{(92.93)^2}{12}\right]\left[(619{,}207) - \dfrac{(2725)^2}{12}\right]}} = .815$$

Correlation matrix

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses

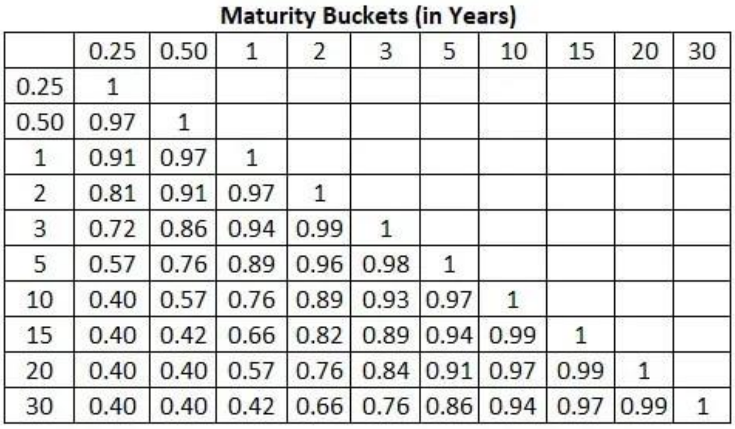
An example of a correlation matrix

Typically, a correlation matrix is “square”, with the same variables shown in the rows and columns. I've shown an example below. This shows correlations between the stated importance of various things to people. The line of 1.00s going from the top left to the bottom right is the main *diagonal*, which shows that each variable always perfectly correlates with itself. This matrix is symmetrical, with the same correlation is shown above the main diagonal being a mirror image of those below the main diagonal.



	Y	X1	X2
Y	1	0.6	0.5
X1	0.6	1	0.7
X2	0.5	0.7	1

The above table is a correlation matrix between different Bonds issued by the Government with different residual maturity stated in the form of years in both horizontal and vertical buckets. It enables us to interpret that a bond with 0.25 years to maturity and a bond with 0.5 years to maturity has a correlation coefficient of 0.97 in their price movements and similarly for other maturity bonds.



Maturity Buckets (in Years)

	0.25	0.50	1	2	3	5	10	15	20	30
0.25	1									
0.50	0.97	1								
1	0.91	0.97	1							
2	0.81	0.91	0.97	1						
3	0.72	0.86	0.94	0.99	1					
5	0.57	0.76	0.89	0.96	0.98	1				
10	0.40	0.57	0.76	0.89	0.93	0.97	1			
15	0.40	0.42	0.66	0.82	0.89	0.94	0.99	1		
20	0.40	0.40	0.57	0.76	0.84	0.91	0.97	0.99	1	
30	0.40	0.40	0.42	0.66	0.76	0.86	0.94	0.97	0.99	1

Applications of a correlation matrix

There are three broad reasons for computing a correlation matrix:

1. T**o summarize a large amount of data where the goal is to see patterns**. In our example above, the observable pattern is that all the variables highly correlate with each other.

2. **To input into other analyses**. For example, people commonly use correlation matrixes as inputs for exploratory factor analysis, confirmatory factor analysis, structural equation models, and linear regression when excluding missing values pairwise.

3. **As a diagnostic when checking other analyses**. For example, with linear regression, a high amount of correlations suggests that the linear regression estimates will be unreliable.

REGRESSION

Meaning:

Regression analysis is a statistical tool to study the nature and extent of functional relationship between two or more variables and to estimate (or predict) the unknown values of dependent variable from the known values of independent variable.

The variable that forms the basis for predicting another variable is known as the Independent Variable and the variable that is predicted is known as dependent variable. For example, if we know that two variables price (X) and demand (Y) are closely related we can find out the most probable value of X for a given value of Y or the most probable value of Y for a given value of X. Similarly, if we know that the amount of tax and the rise in the price of a commodity are closely related, we can find out the expected price for a certain amount of tax levy.

Uses of Regression Analysis:

1. It provides estimates of values of the dependent variables from values of independent variables.

2. It is used to obtain a measure of the error involved in using the regression line as a basis for estimation.

3. With the help of regression analysis, we can obtain a measure of degree of association or correlation that exists between the two variables.

4. It is highly valuable tool in economies and business research, since most of the problems of the economic analysis are based on cause and effect relationship.

Linear Regression

Regression lines and regression equations are used synonymously. Regression equations are algebraic expression of the regression lines. Let us consider two variables: X & Y. If y depends on x, then the result comes in the form of simple regression. If we take the case of two variable X and Y, we shall have two regression lines as the regression line of X on Y and regression line of Y on X. The regression line of Y on X gives the most probable value of Y for given value of X and the regression line of X on Y given the most probable value of X for given value of Y. Thus, we have two regression lines. However, when there is either perfect positive or perfect negative correlation between the two variables, the two regression line will coincide, i.e. we will have one line. If the variables are independent, r is zero and the lines of regression are at right angles i.e. parallel to X axis and Y axis.

Therefore, with the help of simple linear regression model we have the following two regression lines

1. Regression line of Y on X: This line gives the probable value of Y (Dependent variable) for any given value of X (Independent variable).

   Regression line of Y on X : $Y - \dot{Y} = b_{yx} (X - \dot{X})$

   OR : $Y = a + bX$

2. Regression line of X on Y: This line gives the probable value of X (Dependent variable) for any given value of Y (Independent variable).

   Regression line of X on Y : $X - \dot{X} = b_{xy} (Y - \dot{Y})$

   OR : $X = a + bY$

In the above two regression lines or regression equations, there are two regression parameters, which are “a” and “b”. Here “a” is unknown constant and “b” which is also denoted as “b_{yx}” or “b_{xy}”, is also another unknown constant popularly called as regression coefficient. Hence, these “a” and “b” are two unknown constants (fixed numerical values) which determine the position of the line completely. If the value of either or both of them is changed, another line is determined. The parameter “a” determines the level of the fitted line (i.e.

the distance of the line directly above or below the origin). The parameter "b" determines the slope of the line (i.e. the change in Y for unit change in X).

If the values of constants "a" and "b" are obtained, the line is completely determined. But the question is how to obtain these values. The answer is provided by the method of least squares. With the little algebra and differential calculus, it can be shown that the following two normal equations, if solved simultaneously, will yield the values of the parameters "a" and "b".

Two normal equations:

X on Y	Y on X
$\sum X = Na + b\sum Y$ $\sum XY = a\sum Y + b\sum Y^2$	$\sum Y = Na + b\sum X$ $\sum XY = a\sum X + b\sum X^2$

This above method is popularly known as direct method, which becomes quite cumbersome when the values of X and Y are large. This work can be simplified if instead of dealing with actual values of X and Y, we take the deviations of X and Y series from their respective means. In that case:

Regression equation Y on X:

$Y = a + bX$ will change to $(Y - \dot{Y}) = b_{yx} (X - \dot{X})$

Regression equation X on Y:

$X = a + bY$ will change to $(X - \dot{X}) = b_{xy} (Y - \dot{Y})$

In this new form of regression equation, we need to compute only one parameter i.e. "b". This "b" which is also denoted either "b_{yx}" or "b_{xy}" which is called as regression coefficient.

Illustration 01:

Find the two regression equation of X on Y and Y on X from the following data:

X	:	10	12	16	11	15	14	20	22
Y	:	15	18	23	14	20	17	25	28

Solution:

Calculation of Regression Equation

X	Y	X^2	Y^2	XY
10	15	100	225	150
12	18	144	324	216
16	23	256	529	368
11	14	121	196	154
15	20	225	400	300
14	17	196	289	238
20	25	400	625	500
22	28	484	784	616
120	**160**	**1,926**	**3,372**	**2,542**
∑X	**∑Y**	**$\sum X^2$**	**$\sum Y^2$**	**∑XY**

Here N = Number of elements in either series X or series Y = 8

Now we will proceed to compute regression equations using normal equations.

Regression equation of X on Y: **X = a + bY**

The two normal equations are:

$$\sum X = Na + b\sum Y$$
$$\sum XY = a\sum Y + b\sum Y^2$$

Substituting the values in above normal equations, we get

$$120 = 8a + 160b \quad \text{..... (i)}$$
$$2542 = 160a + 3372b \quad \text{..... (ii)}$$

Let us solve these equations (i) and (ii) by simultaneous equation method

Multiply equation (i) by 20 we get 2400 = 160a + 3200b

Now rewriting these equations:

$$\begin{array}{lcccl} 2400 & = & \cancel{160a} & + & 3200b \\ 2542 & = & \cancel{160a} & + & 3372b \\ \underline{(-)} & & \underline{(-)} & & \underline{(-)} \\ -142 & = & & & -172b \end{array}$$

Therefore now we have -142 = -172b, this can rewritten as 172b = 142

Now, $b = \frac{142}{172} = 0.8256$ (rounded off)

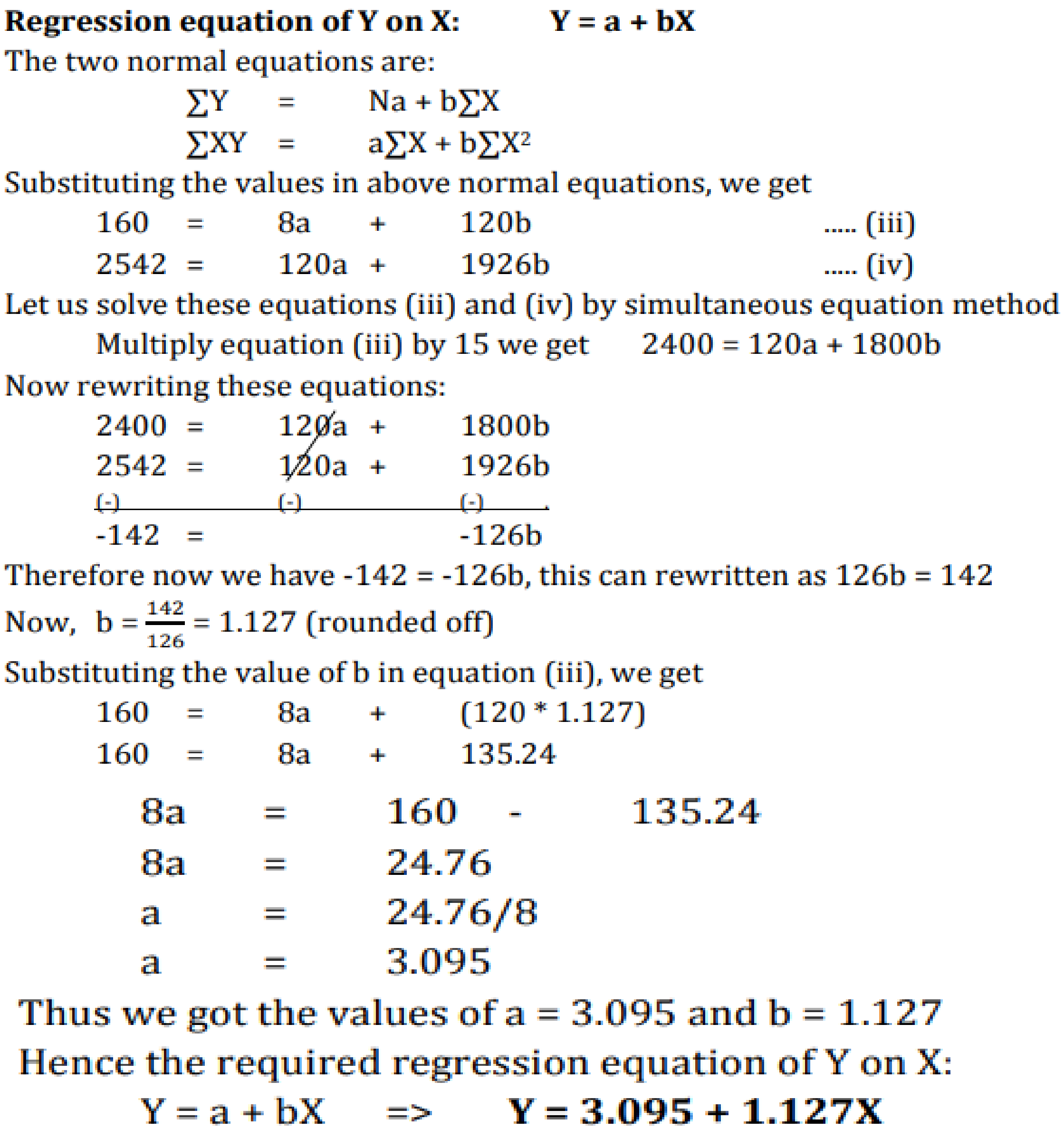
Substituting the value of b in equation (i), we get

$$\begin{array}{lcl} 120 & = & 8a + (160 * 0.8256) \\ 120 & = & 8a + 132 \text{ (rounded off)} \\ 8a & = & 120 - 132 \\ 8a & = & -12 \\ a & = & -12/8 \\ a & = & -1.5 \end{array}$$

Thus we got the values of a = -1.5 and b = 0.8256

Hence the required regression equation of X on Y:

X = a + bY => **X = -1.5 + 0.8256Y**



Regression equation of Y on X: **Y = a + bX**

The two normal equations are:

$$\sum Y = Na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Substituting the values in above normal equations, we get

$$160 = 8a + 120b \qquad \text{..... (iii)}$$

$$2542 = 120a + 1926b \qquad \text{..... (iv)}$$

Let us solve these equations (iii) and (iv) by simultaneous equation method

Multiply equation (iii) by 15 we get 2400 = 120a + 1800b

Now rewriting these equations:

$$2400 = 12\cancel{0}a + 1800b$$

$$2542 = 1\cancel{2}0a + 1926b$$

$$(-) \qquad (-) \qquad (-)$$

$$-142 = \qquad -126b$$

Therefore now we have -142 = -126b, this can rewritten as 126b = 142

Now, $b = \frac{142}{126} = 1.127$ (rounded off)

Substituting the value of b in equation (iii), we get

$$160 = 8a + (120 * 1.127)$$

$$160 = 8a + 135.24$$

$$8a = 160 - 135.24$$

$$8a = 24.76$$

$$a = 24.76/8$$

$$a = 3.095$$

Thus we got the values of a = 3.095 and b = 1.127

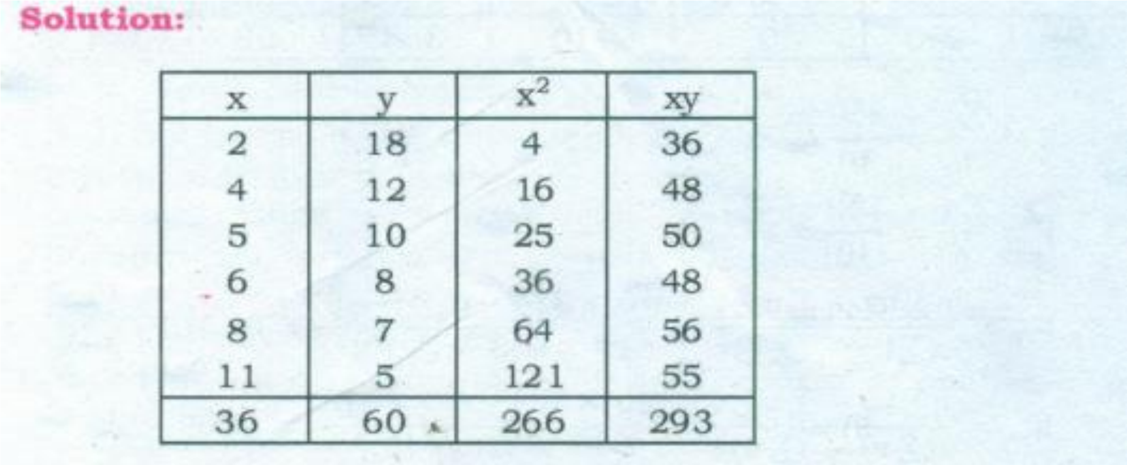
Hence the required regression equation of Y on X:

Y = a + bX => **Y = 3.095 + 1.127X**

Illustration 02:

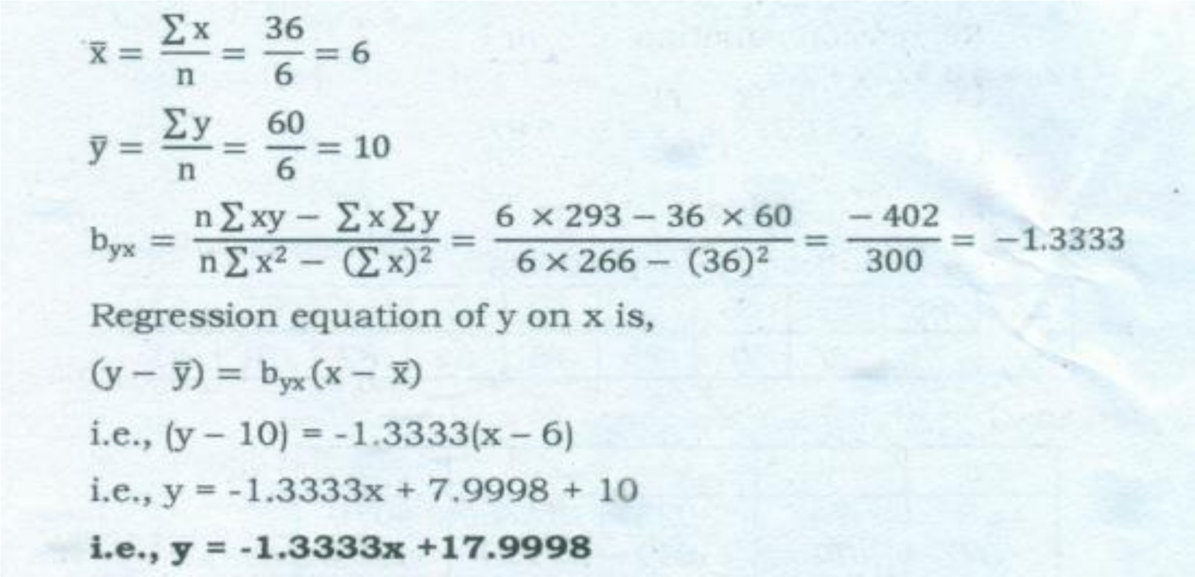
Compute the regression equation of y on x from the following data.

X	2	4	5	6	8	11
Y	18	12	10	8	7	5



Solution:

x	y	x^2	xy
2	18	4	36
4	12	16	48
5	10	25	50
6	8	36	48
8	7	64	56
11	5	121	55
36	60	266	293



$$\bar{x} = \frac{\sum x}{n} = \frac{36}{6} = 6$$

$$\bar{y} = \frac{\sum y}{n} = \frac{60}{6} = 10$$

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{6 \times 293 - 36 \times 60}{6 \times 266 - (36)^2} = \frac{-402}{300} = -1.3333$$

Regression equation of y on x is,

$(y - \bar{y}) = b_{yx}(x - \bar{x})$

i.e., $(y - 10) = -1.3333(x - 6)$

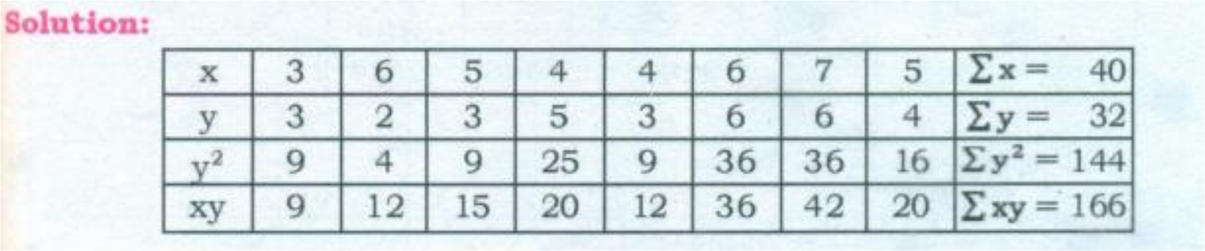
i.e., $y = -1.3333x + 7.9998 + 10$

i.e., $y = -1.3333x + 17.9998$

Illustration 03:

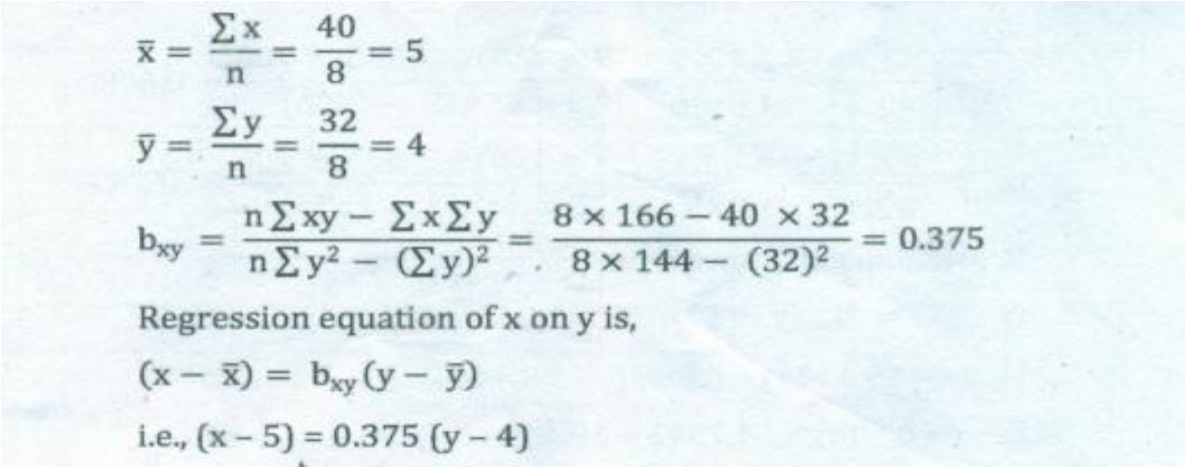
Find the regression equation of x on y and predict the value of x when y is 9.

X	3	6	5	4	4	6	7	5
Y	3	2	3	5	3	6	6	4



Solution:

x	3	6	5	4	4	6	7	5	$\sum x =$ 40
y	3	2	3	5	3	6	6	4	$\sum y =$ 32
y^2	9	4	9	25	9	36	36	16	$\sum y^2 = 144$
xy	9	12	15	20	12	36	42	20	$\sum xy = 166$



$$\bar{x} = \frac{\sum x}{n} = \frac{40}{8} = 5$$

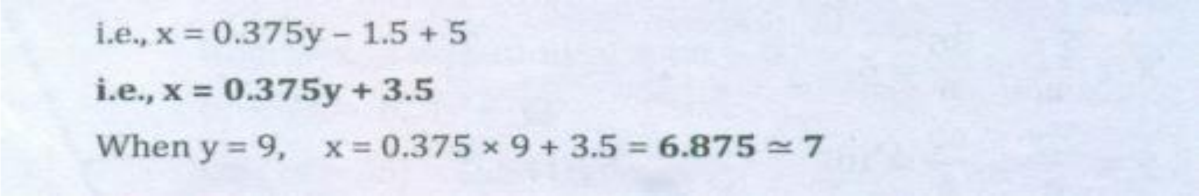
$$\bar{y} = \frac{\sum y}{n} = \frac{32}{8} = 4$$

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2} = \frac{8 \times 166 - 40 \times 32}{8 \times 144 - (32)^2} = 0.375$$

Regression equation of x on y is,

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

i.e., $(x - 5) = 0.375\,(y - 4)$



i.e., $x = 0.375y - 1.5 + 5$

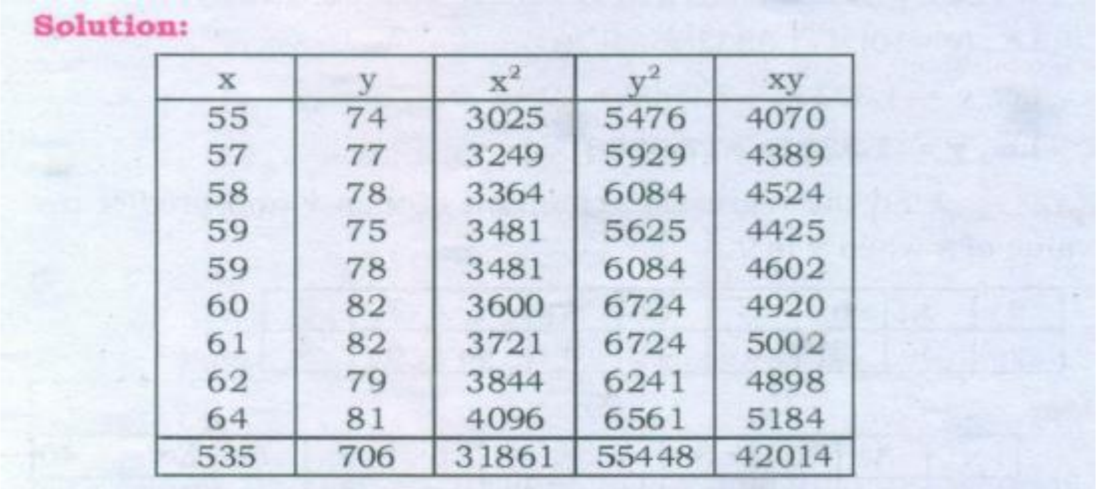
i.e., $x = 0.375y + 3.5$

When $y = 9$, $x = 0.375 \times 9 + 3.5 = \mathbf{6.875 \simeq 7}$

<u>Illustration 04:</u>

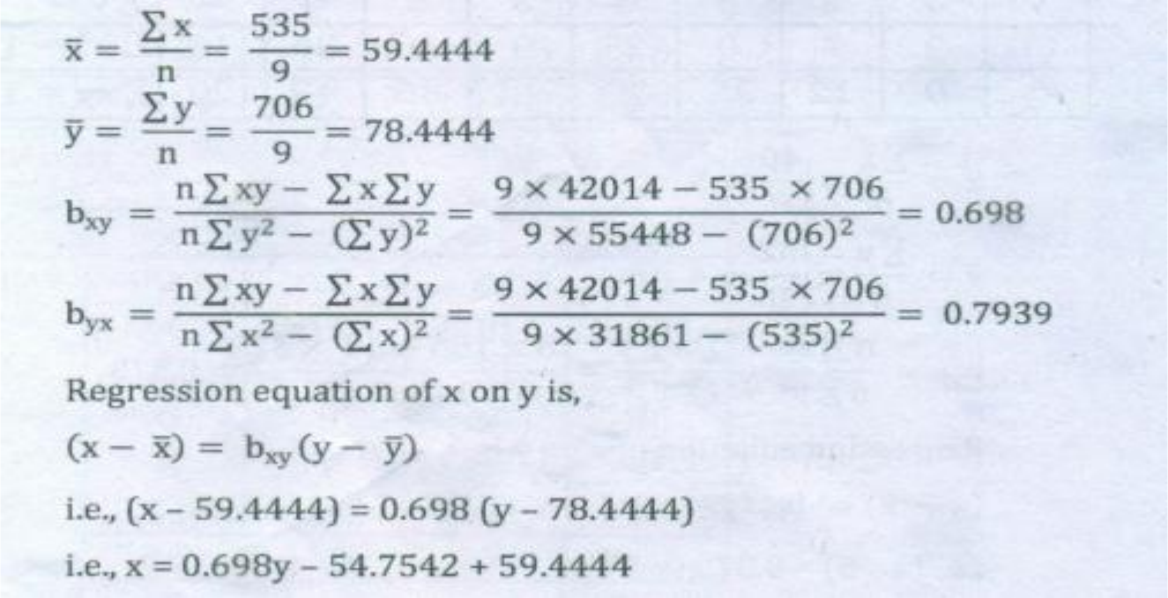
Find the two regression lines from the following data.

X	55	57	58	59	59	60	61	62	64
Y	74	77	78	75	78	82	82	79	81



Solution:

x	y	x^2	y^2	xy
55	74	3025	5476	4070
57	77	3249	5929	4389
58	78	3364	6084	4524
59	75	3481	5625	4425
59	78	3481	6084	4602
60	82	3600	6724	4920
61	82	3721	6724	5002
62	79	3844	6241	4898
64	81	4096	6561	5184
535	706	31861	55448	42014



$$\bar{x} = \frac{\sum x}{n} = \frac{535}{9} = 59.4444$$

$$\bar{y} = \frac{\sum y}{n} = \frac{706}{9} = 78.4444$$

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2} = \frac{9 \times 42014 - 535 \ \times 706}{9 \times 55448 - \ (706)^2} = 0.698$$

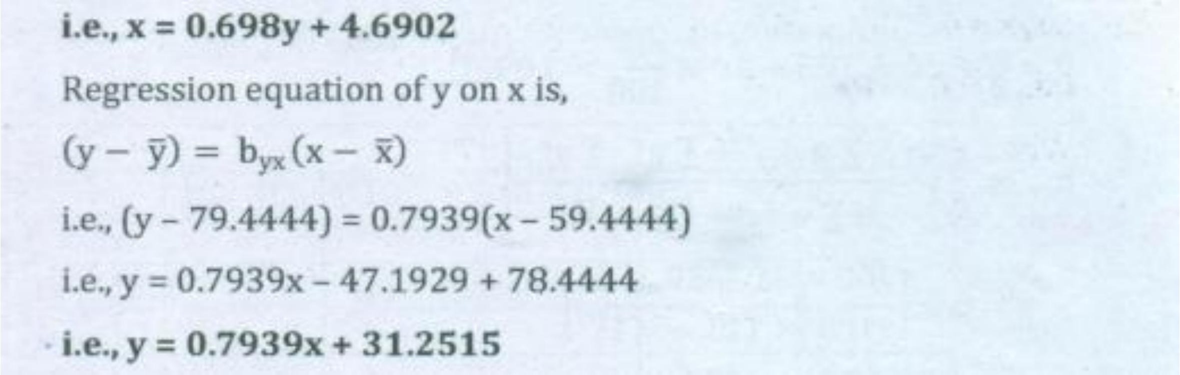
$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{9 \times 42014 - 535 \ \times 706}{9 \times 31861 - \ (535)^2} = 0.7939$$

Regression equation of x on y is,

$(x - \bar{x}) = b_{xy}(y - \bar{y})$

i.e., $(x - 59.4444) = 0.698\ (y - 78.4444)$

i.e., $x = 0.698y - 54.7542 + 59.4444$



i.e., $x = 0.698y + 4.6902$

Regression equation of y on x is,

$(y - \bar{y}) = b_{yx}(x - \bar{x})$

i.e., $(y - 79.4444) = 0.7939(x - 59.4444)$

i.e., $y = 0.7939x - 47.1929 + 78.4444$

i.e., $y = 0.7939x + 31.2515$