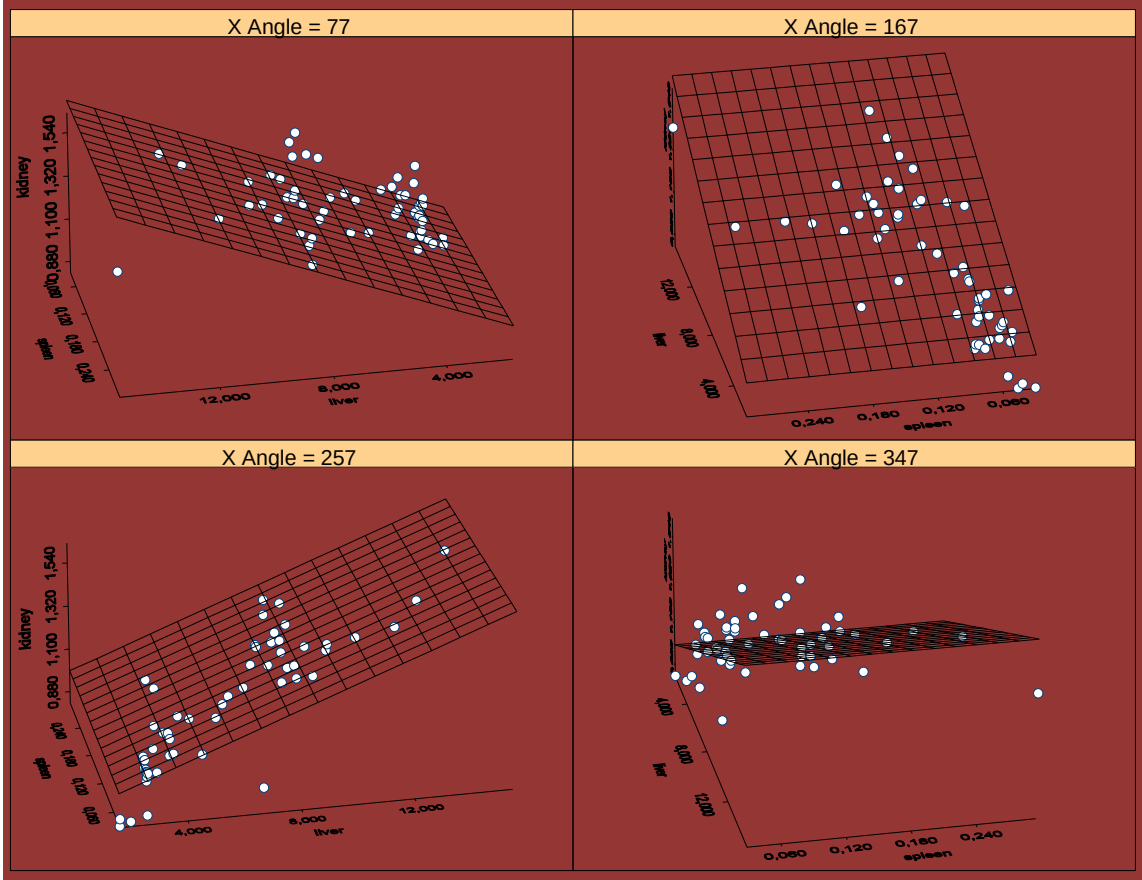
# BASIC STATISTICS




**Abd Basir A**

Department of Mathematics and Natural Sciences
Faculty of Teaching and Education Science
Mulawarman University Tahun 2021

# PREFACE

The basic statistics book, compiled to assist or facilitate students in learning, even practicing the basics of statistical data analysis. This book is used solely by the author in teaching the subject of basic statistics and as the main textbook for students in the course.

Some of the presentation of the material in this book still is direct, without any explanation by theoretical methods. Actually, the preparation is still under construction, material coverage is still not complete, is planned to include in the next edition. Similarly, some materials such as probability distributions, confidence interval estimation, one-way analysis of variance and factorial, and examination of assumptions such as normality, homogeneity of variances, multicollinearity, and autokorrelation, not covered at this time. The contents of this book include descriptive statistics and inferential statistics. Descriptive statistics only include the presentation of the data by frequency tables, histograms, frequency polygons, stem and leaf diagrams, boxplot diagrams, and the use of numerical measures to summarize data. While inferential statistics only include the normal distribution, hypothesis testing, simple linear regression analysis and multiple linear regression analysis. Nevertheless, the composition of the material in the first and second chapter comes with a worksheet, and the final chapters, a description of the problem solving combined with computerized results. Problem solving worksheets presented with examples of solving identical problems, intended to allow students to better understand the material presented in this book. Solving the results of computerized intended to allow students to recognize and trained to use the results of the data analysis by computer.

As authors, it is always a pleasure to formally record our appreciation for to help and support given by others. We also like to express our thanks to our friends or faculty colleagues, provided moral support and hospitality during intensive period for writing.

As authors, it is always a pleasure to formally record our appreciation for to help and support given by others. We also like to express our thanks to fellow lecturers, who provide moral support during intensive period for writing.

Abd. Basir A

# CONTENTS

C H A P T E R **1**

# PRESENTATION OF DATA WITH TABLE AND CHART

**Basic competence**

Summarize and present a mass of disorganized data trough frequency distribution tables, histograms, polygons, stem and leaf diagrams, and boxplots

**Indicators:**

1. Presents a set of data to the frequency distributian table
2. Making histograms a set of data
3. Making frequency polygons a set of data
4. Presents a set of data to the stem and leaf diagrams
5. Presents a set of data to the boxplots

# 1. PRESENTATION OF DATA WITH TABLE AND CHART

## 1.1 DESCRIBING DATA WITH FREQUENCY TABLE

Description of the data to the table often using frequency tables. Generally table frequency will be served through the grouping data. Partitioned data set over the first subset, called the class. Wide class called class intervals. Generally, the width of the class is created equal, although for the benefit or condition of the data, the class may not be the same width

**Table 1.1** Raw data : Mass of baggage (kg)

| | [,1] | [,2] | [,3] | [,4] | [,5] | [,6] | [,7] | [,8] | [,9] |
|---|---|---|---|---|---|---|---|---|---|
| [1,] | 8 | 51 | 20 | 37 | 19 | 40 | 11 | 43 | 55 |
| [2,] | 30 | 23 | 25 | 7 | 20 | 9 | 21 | 28 | - |
| [3,] | 55 | 55 | 40 | 34 | 49 | 33 | 7 | 7 | - |
| [4,] | 17 | 45 | 50 | 23 | 49 | 20 | 26 | 47 | - |
| [5,] | 27 | 20 | 46 | 42 | 39 | 12 | 24 | 34 | - |
| [6,] | 50 | 42 | 26 | 10 | 51 | 49 | 48 | 14 | - |
| [7,] | 21 | 25 | 27 | 32 | 20 | 12 | 35 | 27 | - |
| [8,] | 20 | 47 | 18 | 55 | 15 | 12 | 17 | 31 | - |

**Table 1.2** Ordered Array of baggage mass

| | [,1] | [,2] | [,3] | [,4] | [,5] | [,6] | [,7] | [,8] | |
|---|---|---|---|---|---|---|---|---|---|
| [1,] | 7 | 7 | 7 | 8 | 9 | 10 | 11 | 12 | |
| [2,] | 12 | 12 | 14 | 15 | 17 | 17 | 18 | 19 | |
| [3,] | 20 | 20 | 20 | 20 | 20 | 20 | 21 | 21 | |
| [4,] | 23 | 23 | 24 | 25 | 25 | 26 | 26 | 27 | |
| [5,] | 27 | 27 | 28 | 28 | 30 | 31 | 32 | 33 | |
| [6,] | 34 | 34 | 35 | 37 | 39 | 40 | 40 | 42 | |
| [7,] | 42 | 43 | 45 | 46 | 47 | 47 | 48 | 49 | |
| [8,] | 49 | 49 | 50 | 50 | 51 | 51 | 55 | 55 | 55 |

Determination of the class interval and the number of classes for a set of data is done by first reviewing range of the data, the smallest observed values (minimum score) and the largest observed values (maximum score). The range of the data is the difference between the minimum score with a maximum score. The more class then the description of the data will be more looks simple, but the missing information is also growing.

After the determination of the class and the class intervals, do taurus / Tolly to each observation from the data. The number of observations in a class is the frequency of the class. For certain purposes, in the frequency table are often added cumulative frequency column, relative frequency, and the midpoint of each class interval.

**Table 1.3** Frequency distribution

| **Classes (kg)** | **Frequencies ( $f_i$ )** | **Midpoint ( $x_i$ )** | **Cumulative Frequencies** | **Relative Frequencies ( %)** |
|---|---|---|---|---|
| 0 - 9 | 5 | 5 | 5 | (5/65) x 100 = 7,7 |
| 10 - 19 | 11 | 15 | 16 | (11/65)x100 = 16,9 |
| 20 - 29 | 20 | 25 | 36 | 30,8 |
| 30 - 39 | 9 | 35 | 45 | 13,8 |
| 40 - 49 | 13 | 45 | 58 | 20,0 |
| 50 - 60 | 7 | 55 | 65 | 10,8 |
| | N = 65 | 180 | -- | 100 |

Examples of data presentation with a frequency table is presented in Table 1.3. How to make a frequency table using a set of data about a set mass baggage in Table 1.1. The minimum score is 7 and maximum score is 55, so the range of the data is 48. Decision was made to the number of classes is 6, with a wide of class intervals 10. Determination of the number and width of classes does not need to use a specific formula, as written in several books. An important requirement is that the class is made to include all the data, and

considering how simple the desired presentation. The main thing that also needs is how to define a class interval, which determines the lower limit and upper limit of the class interval.

There are two ways that are often used to define the class interval. The selection is determined by the nature of the data used. For example, the variable age in terms of years. Suppose the class interval is written as "3-5". Children are included in this class is a child who have 3rd birthday until children who have not been 6th birthday. So the actual lower limit is 3 and the actual upper limit is 6. The midpoint of the class interval is 4.5. In contrast to the variable height using specific units, eg centimeters (cm). Relative error is half of the order of cm, which is 0.5 cm. Suppose the class interval is written "170-174", then the actual grade boundaries are 169.5 and 174.5. Two ways of defining this class intervals are shown respectively in Figure 1.1 and Figure 2.2

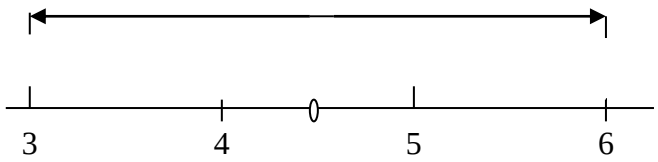



**Figure 1.1** Class intervals for the variable Age

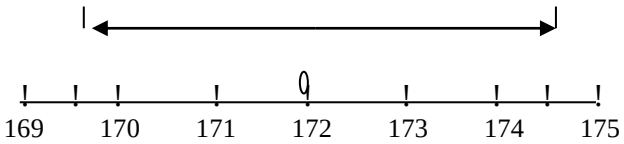



**Figure 1.2** Class intervals for the variable Height

## 1.2 HISTOGRAM AND FREQUENCY POLYGON

The actual limits of the class and the middle class becomes an important point when we want to demonstrate the distribution of the data with histograms and frequency polygons. Histogram for the grouped data is a bar charts with the area of each rectangle bar should be proportional to its frecuency. Histograms for grouped data in Table 1.3, are shown in Figure 1.3. The width of each column of the rectangle is 10 kg, following the class interval. The left side and right side of the bar coincides with the actual grade boundaries on the horizontal axis. Similarly, the center of each bar is located at the midpoint of the class interval.

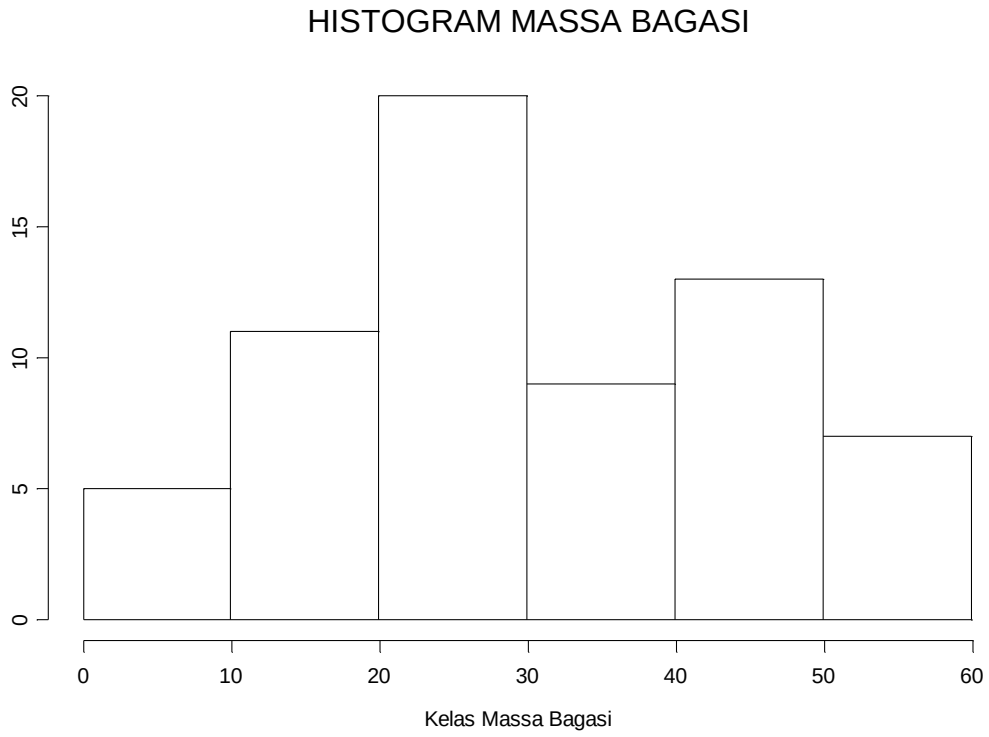



**Figure 1.3** Histogram of baggage mass

If the grouped data created unequal class intervals, eg in this example the last two classes are combined, or the width of the class interval to 20, then

height of the bar rectangle is reduced by half, which was reduced 0.5 from (13 + 7). High of last bar rectangle is 10. This method makes area of bar is proportional to its frequency. The results of this histogram is shown in Figure 4.

Polygon frequency is a plot of the frequency of each class by the midpoint of each class interval, by putting a line connecting each point of the plot. Furthermore, both ends of the line extended down to cut the horizontal axis. Line intersects the axis at a point which is equal to the width of the class interval, and distance is calculated from the midpoint of the edge of the two classes. Interval class is meant here is two interval classes left and right edge.

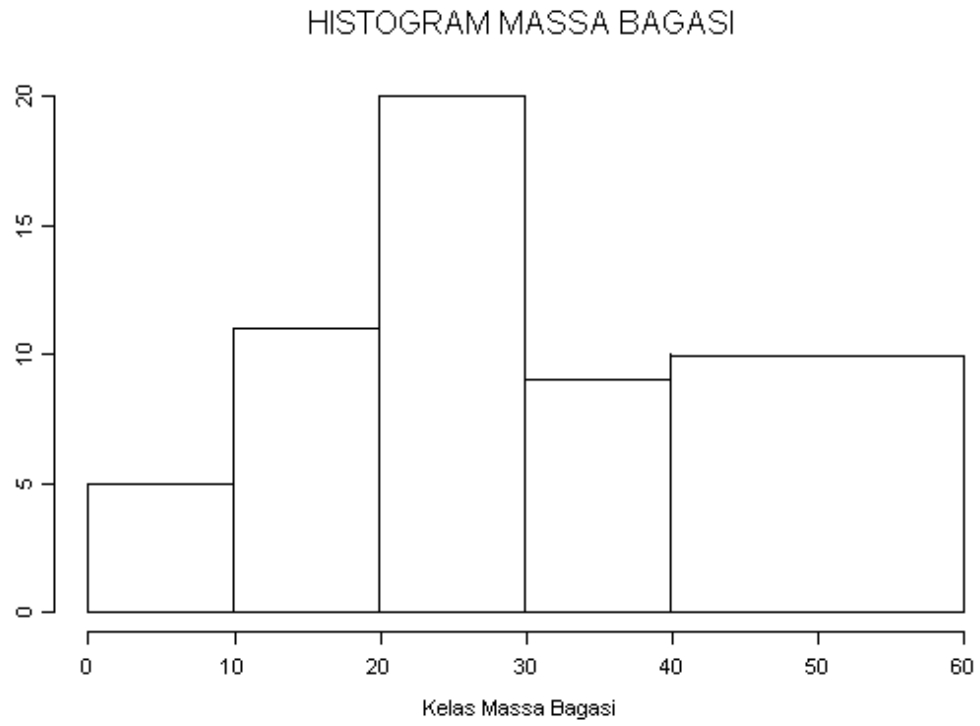



**Figure 1.4** Histograms with unequal class intervals

How to describe the frequency polygon for the above histogram, shown in Figure 1.5. The midpoints of the class intervals, ie 5, 15, 25, 35, 45, 55 respectively plotted with the frequency of each class, ie 5, 11, 20, 9, 13, 7.

Furthermore, the results of the plot points connected by lines. Both ends of line extended through the high-side or cut half of each rectangle bar (left and right), until it reaches the horizontal axis. Both ends of line cutting the horizontal axis at the point -5 and 65. Note that the area under the curve polygon, or the area bounded by the curve and the horizontal axis is equal to that enclosed by the histogram. Polygon become an alternative frequency histograms, especially when aiming to compare two sets of data on one graph. Results painting polygon for baggage mass data shown in Figure 1.6.

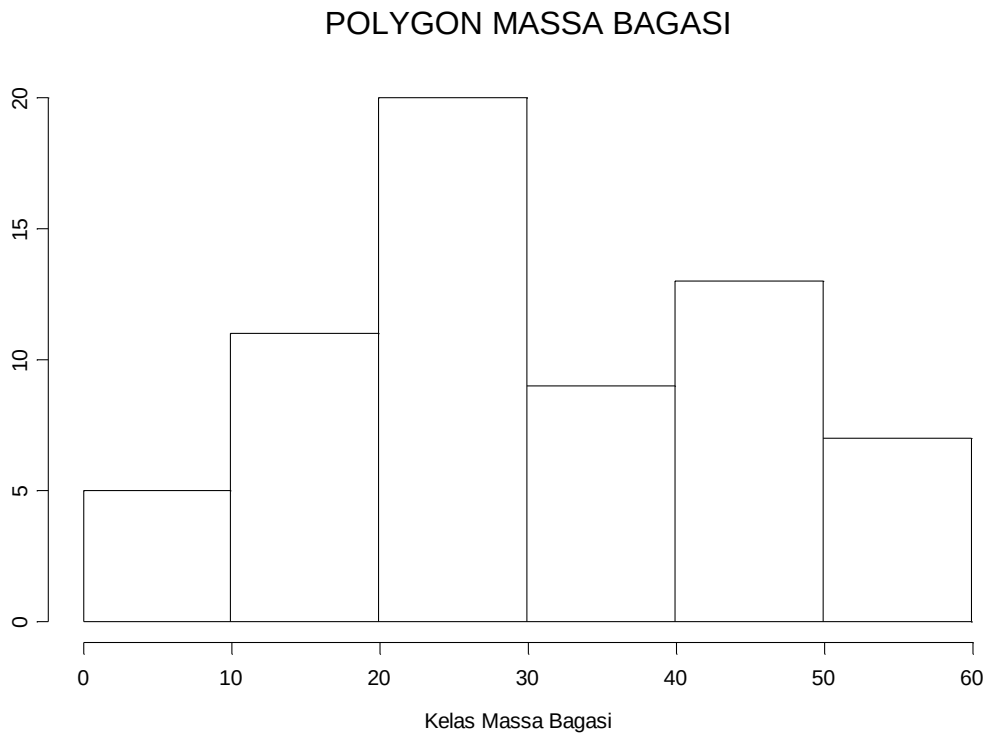



**Figure 1.5** Make a frequency polygon

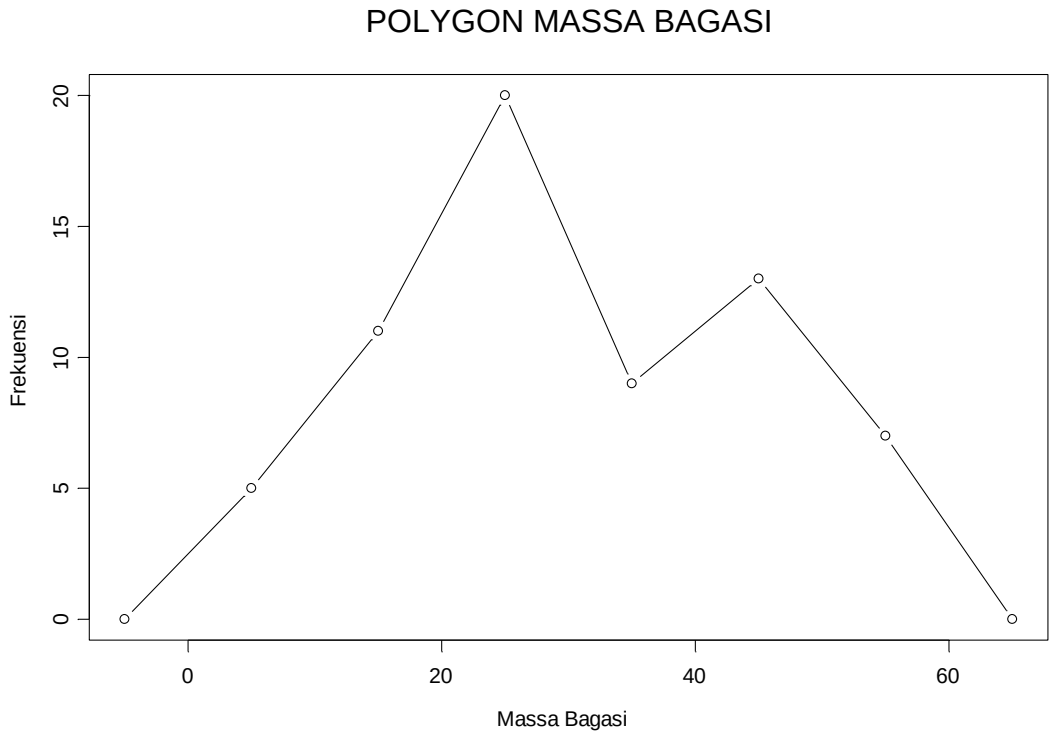



**Figure 1.6** Polygon of mass baggage

## 1.3 STEM AND LEAF DIAGRAMS

Alternatively instead of histograms, for the purpose showing the distribution of the data is stem and leaf diagram. Stem and leaf diagram are made directly from raw data, or without grouping the data. In the stem and leaf diagrams, information is wasted by the tolly in each class can be avoided. Observed values can be directly seen in the stem and leaf diagram, while displaying the shape of data distribution

Example stem and leaf diagram of baggage mass data presented in Figure 1.7. The diagram consists of two parts, namely stem and leaves. Then fitted with a cumulative frequency information in the first column. Figures on stem stating the number of 10s (numbers of 100s can also be expressed and so on) over which full data set is distributed. Leaves states the unit or gives the last digit. Each observation in the data set is divided into two parts, namely for

a stem and a leaf. Suppose the value of 48 observation, divided into the number 4 is placed on the stem which states forties numbers, and the number 8 is placed on the leaves which stated number eight units.

Let us focus on a stem view thirties, for example stem and leaf diagram in Figure 1.7, which is a view on the fourth line "29 3: 012 344 579". This presentation represents 9 observations, ie

30 31 32 33 34 34 35 37 39

Numbers thirties laid out on a stem by writing the number 3, then every unit value written on the leaves by writing "012 344 579". While the number 29 is written in the first column states the cumulative frequency is calculated starting from the bottom row

Cumulative frequency attached to the front of a diagram as an additional description of stem and leaf diagram. First of all, the cumulative frequency is calculated starting from the top line moves to the second row, and stopped after hitting a row or a branch that contains the median. In the raw or branch that contains median, is written how the frequency of the stem, which was written not cumulative frequency. Writing frequency is equipped with brackets. Furthermore, the cumulative frequency is calculated starting from the bottom row to row under the branch that contains the median.

```
 5       0 : 77789
 16      1 : 01222457789
(20)     2 : 00000011334556677788
 29      3 : 012344579
 20      4 : 0022356778999
 7       5 : 0011555
```

**Figure 1.7** Stem and leaf diagram

If on a stem and leaf diagram, there is one or several stem with striking frequency is very large, then the writing sequence numbers on the leaves will be very long towards the right edge of the page. Presentation of diagrams can be improved with spliting each stem into two parts. The first section to load the leaf : 0 1 2 3 4, and the second part to load leaf : 5 6 7 8 9. We can decide to split any branches into 2 parts, 5 parts, and 10 parts. For example stem and leaf diagram if stem divided into 2 parts shown in Figure 1.8.

```
5     0 : 77789
11    1 : 012224
16    1 : 57789
29    2 : 00000011334
(9)   2 : 556677788
29    3 : 012344
23    3 : 579
20    4 : 00223
15    4 : 56778999
7     5 : 0011
3     5 : 555
```

**Figure 1.8** Stem and leaf diagram were obtained by spliting stem into 2 parts

## 1.4 BOXPLOT DIAGRAM

In addition to the stem and leaf diagrams, describing of the data can also be presented with boxplot diagrams. Demonstration of Interquartile range appear more clearly through boxplot diagram, it is because he is equipped with a observations where are located outside of Interquartile range. Besides that, it also presented the results of the identification of observations that are outliers.

Interquartile range described as the width of the box. Around the box, there are a *inner fence* and a *outer fence* to help identify outliers. Boxplot diagram is shown in Figure 1.9.

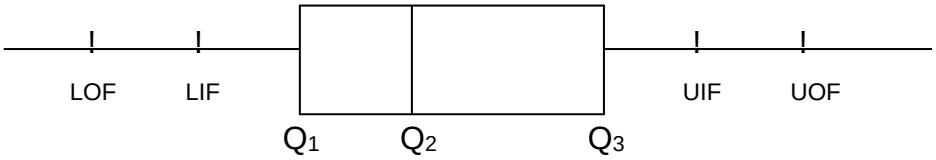



**Figure 1.9** Boxplot Diagram

Description :

$Q_1$ = First quartile

$Q_2$ = Second quartile

$Q_3$ = Third quartile

LOF = lower outer fence limit

$= Q_1 - 3\,(Q_3 - Q_1)$

LIF = lower inner fence

$= Q_1 - 3/2\,(Q_3 - Q_1)$

UIF = upper inner fence

$= Q_3 + 3/2\,(Q_3 - Q_1)$

UOF = upper outer fence

$= Q_3 + 3\,(Q_3 - Q_1)$

**Worked Example 1.1:**

Draw a boxplot diagram of the following data

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 21 | 40 | 42 | 50 | 60 | 64 | 72 | 78 |
| 80 | 84 | 85 | 86 | 86 | 90 | 90 | 90 |
| 92 | 98 | 100 | 114 | 120 | 140 | 160 | 215 |

**Worked solution:**

$Q_1$ lies at 6.25th observation, so that

$Q_1 = 64 + 0.25\,(72 - 64) = 66$

$Q_2$ lies at 12.5th observation, so that

$Q_2 = 86 + 0.5\ (86 - 86) = 86$

$Q_3$ lies at 18.75th observation, so that

$Q_3 = 98 + 0.75\ (100 - 98) = 99.5$

Interquartile range = 99.5 – 66 = 33.5

LIF $= Q_1 - 3/2\ (Q_3 - Q_1)$
$= 66 - 3/2\ (33.5) = 15.75$

LOF $= Q_1 - 3\ (Q_3 - Q_1)$
$= 66 - 3\ (33.5) = -34.5$

UIF $= Q_3 + 3/2\ (Q_3 - Q_1)$
$= 99.5 + 3/2\ (33.5) = 149.75$

UOF $= Q_3 + 3\ (Q_3 - Q_1)$
$= 99.5 + 3\ (33.5) = 200$

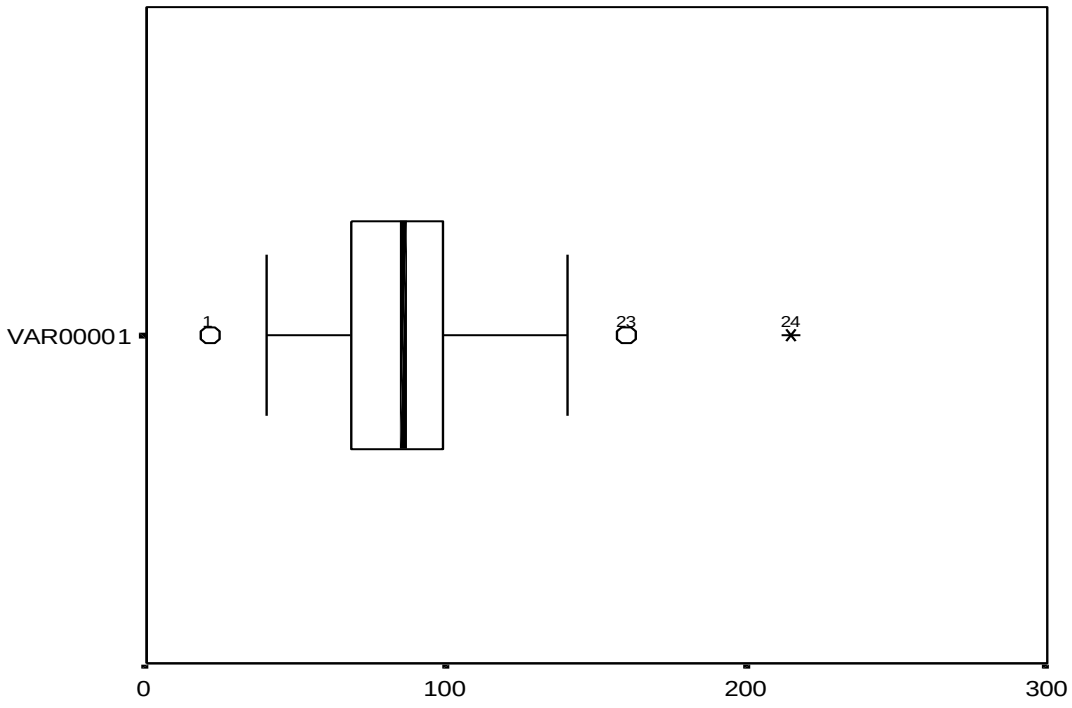



**Figure 1.10** Boxplot Diagram

## WORKED EXERCISES SHEET

**Problem 1.1 :**

In Table 1.4, there is a data set that records the height of 100 male students. Data has been sequenced from the observations of low to highest observation. Data are grouped according to class 8, as a class formed in the first column in Table 5.

**Table 1.4** Data on height of 100 male students

| | [,1] | [,2] | [,3] | [,4] | [,5] | [,6] | [,7] | [,8] | [,9] | [,10] |
|---|---|---|---|---|---|---|---|---|---|---|
| [1,] | 156 | 157 | 159 | 162 | 162 | 163 | 164 | 165 | 165 | 166 |
| [2,] | 166 | 166 | 166 | 167 | 167 | 168 | 168 | 168 | 169 | 169 |
| [3,] | 169 | 169 | 169 | 170 | 170 | 170 | 170 | 170 | 171 | 171 |
| [4,] | 171 | 171 | 171 | 172 | 172 | 172 | 172 | 172 | 173 | 173 |
| [5,] | 173 | 173 | 173 | 173 | 173 | 173 | 174 | 174 | 174 | 174 |
| [6,] | 174 | 174 | 174 | 174 | 175 | 175 | 175 | 175 | 175 | 175 |
| [7,] | 175 | 175 | 175 | 175 | 176 | 176 | 176 | 176 | 177 | 177 |
| [8,] | 177 | 177 | 177 | 178 | 178 | 179 | 179 | 179 | 179 | 179 |
| [9,] | 179 | 180 | 180 | 181 | 181 | 181 | 181 | 181 | 182 | 182 |
| [10,] | 182 | 183 | 183 | 184 | 186 | 186 | 188 | 188 | 193 | 194 |

a. Determine the range of data on height above. Are all the observations can be fit onto existing classes, such as grouping the first column in Table 5.
b. Perform Tolly to complete column 2 or the frequency of each class in Table 5. Furthermore, complete each its column.
c. In the class containing the median, explain the actual limits of class interval and middle point, through a presentation of images.
d. Draw a histogram and a polygon for a set of data on height that have been grouped.
e. Create a stem and leaf diagram for the set of data on height. Explain what happened in the stem170s, namely in its leaves

f. Create a stem and leaf diagram by separating each branch into 2 parts for the set of data on height.

**Solution :**

a. Range :
$Rg = X_{max} - X_{min}$

$Rg =$ ……. - ……. = …….

Comment : ……

……

……

**b.**

**Table 1.5** Frequency distribution

| Classes (kg) | Frequencies ( $f_i$ ) | Midpoint ( $x_i$ ) | Cumulative Frequencies | Relative Frequencies ( %) |
|---|---|---|---|---|
| 155 - 159 | 3 | 157 | 3 | |
| 160 - 164 | 4 | 162 | 7 | |
| 165 - 169 | …. | …. | …. | |
| 170 - 174 | …. | …. | …. | |
| 175 - 179 | …. | …. | …. | |
| 180 - 184 | | | | |
| 185 – 189 | | | | |
| 190 – 194 | | | | |
| | N = 100 | ….. | -- | 100 |

**c.** The median is the data ( n / 2)th observation

(…. / 2)th observation

( ….. )th observation

So the median lies in the " …… - ……. " class

**Figure of Excercise 1** The actual limits of class and middle point

**d.** Draw a histogram and a polygon:

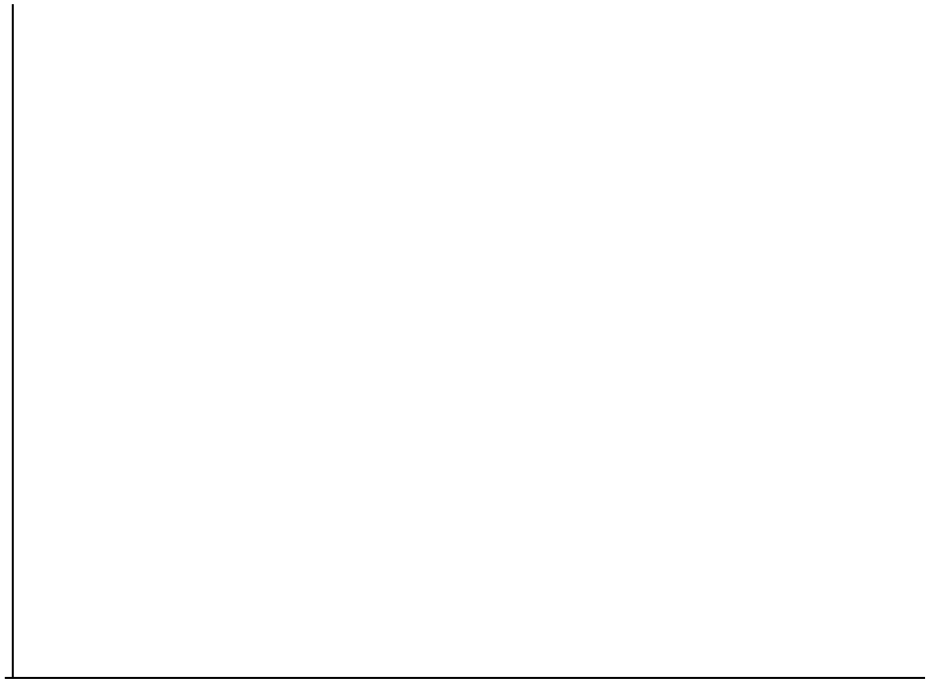


**Figure of Excercise 2** Histogram for high students

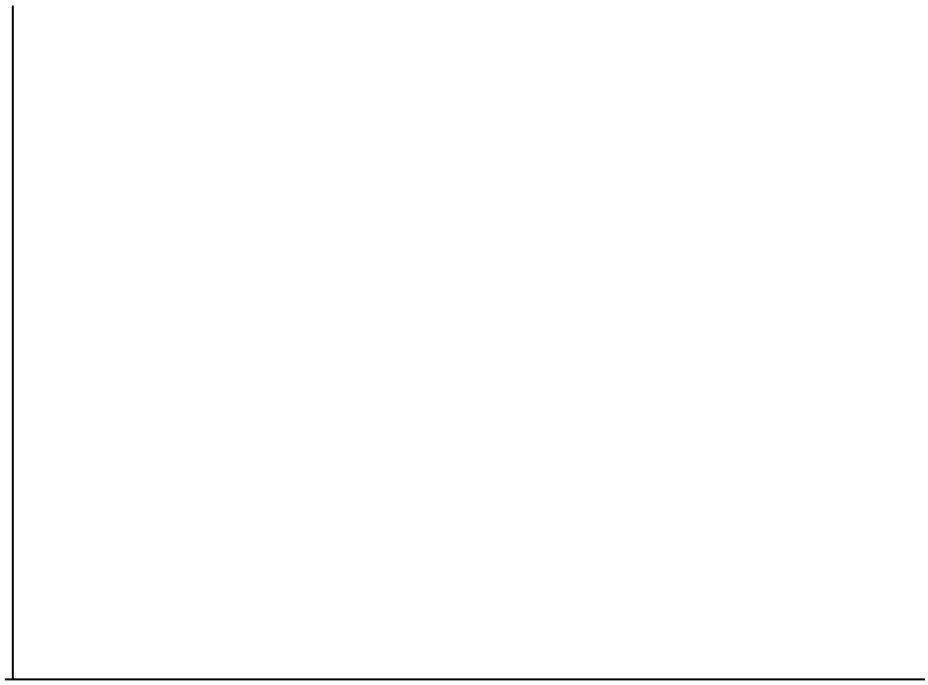


**Figure of Excercise 3** Polygon for high students

**e.** Here is a picture of the stem and leaf diagram :

```
2     15 : 679
...   16 : 22345566...
...   17 : ...
...   18 : ...
...   19 : 34
```

**Figure of Excercise 4** The stem and leaf diagram

Comments : ……

……

**f.** The following the stem and leaf diagram when use separated of stem to 2 parts :

```
2     15 : 679
7     16 : 2234
…     16 : 5566667788899999
…     17 : ...
…     17 : ...
…     18 : ...
…     18 : 6688
…     19 : 34
```

**Figure of Excercise 5** The stem and leaf diagram with stem split method

**Problem 1.2 :**

1. Here is the data on the test results of 50 students :

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 26 | 42 | 8 | 28 | 39 | 78 | 32 | 54 | 93 | 11 |
| 27 | 33 | 22 | 78 | 75 | 83 | 62 | 76 | 77 | 67 |
| 77 | 80 | 7 | 26 | 18 | 10 | 34 | 30 | 36 | 43 |
| 79 | 41 | 24 | 91 | 90 | 63 | 87 | 55 | 60 | 48 |
| 35 | 35 | 51 | 67 | 10 | 76 | 34 | 47 | 51 | 33 |

a. What is range of the data above.
b. Create a frequency table, complete with columns for the middle point, and a column for the cumulative frequency.
c. In the class containing the median, explain the actual limits of the class interval and the middle point, through a presentation of images.
d. Draw histogram and polygon of the data on the test results.
e. Create a the stem and leaf diagram for data set on the test results
f. Create a the stem and leaf diagram for data set on the test results, by separating each stem into 2 parts.

C H A P T E R **2**

# NUMERICAL MEASURES TO SUMMARIZE DATA

**Basic competencies**

1. Numerically describing a set of data using summary measures of central location.

   **Indicators:**

   1.1 Determine arithmetic mean, median, and modus of a set of ungrouped data

   1.2 Determine arithmetic mean, median, and modus of a set of grouped data

2. Numerically describing a set of data using summary measures of dispersion

   **Indicators:**

   2.1 Determine range and quartiles modified ranges, variance and standard deviation of a set of ungrouped data

   2.2 Determine range and quartiles modified ranges, variance and standard deviation of a set of grouped data

# 2. NUMERICAL MEASURES TO SUMMARIZE DATA

## 2.1 MEASURE OF CENTRAL LOCATION

There are four measures of central location that is often used, which are the arithmetic mean, median, mode, and geometric mean. The selection of which of one of the measures of central location are used, depending on the natural of the data and purpose of measurement.

### 2.1.1 MEASURE OF CENTRAL LOCATION FOR UNGROUP DATA

**a. Arithmetic Mean**

For the data set X1, X2, ..., Xn. Expressed as the arithmetic mean (read: X bar), defined as

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Where,

$\overline{X}$ = Mean value

n = Number of observations

**Worked Example 2.1 :**

For the following data set, find the value of mean

2, 6, 6, 7, 3, 1, 3

**Worked Solution** :

Mean is

$$\overline{X} = \frac{2+6+6+7+3+1+3}{7} = \frac{28}{7} = 4$$

**b. Median**

The median of a set of data is the middle value when the data are arranged from the smallest value to the largest valuein an array. Once the data is sorted, the median is defined as :

Med = The value of ((n + 1 ) / 2)th observations

Med = $X_{(n+1)/2}$

Where,

Med = Median

n = Number of observations

Median for odd-sized data can be directly identified, because the median is listed on the data. If the data size is even then the median is the average of the two middle observations in the ordered data.

**Worked Example 2.2 :**

Determine the median of the following data

a. 8, 4, 4, 3, 5, 9, 10, 8, 8, 7, 9
b. 5, 5, 4, 7, 8, 9, 8, 10, 10, 3, 4, 10

**Worked Solution** :

a. Ordered data : 3, 4, 4, 5, 7, 8, 8, 8, 9, 9, 10

Med = the value of the data in the ((11 +1) / 2)th observation

= the value of the data in the 6th observation

Med = $X_6$

= 8

b. Ordered data : 3, 4, 4, 5, 5, 7, 8, 8, 9, 10, 10, 10

Med = the value of the data in the ((12+1) / 2)th observation

= the value of the data in the 6.5th observation

Med = $X_6 + 0.5 (X_7 - X_6)$

= 7 + 0.5 (7-8) = 7.5

**c. Mode**

Mode is the most frequent events in an event. Thus the mode of a set of data is the data value that has the greatest frequency or the most frequent. Mode the data can not exist and can not be unique. If there are two modes of a set of data, the so-called bimodal. If there are three modes called trimodal.

**Worked Example 2.3 :**

Find the mode of the following data

Data : a. 6, 4, 3, 5, 5, 6, 6, 6, 7, 9, 8

b. 5, 6, 3, 4, 7, 7, 8, 8, 9, 8, 7

**Worked Solution** :

a. The value of the data that most often arises is the value of 6, with 4 frequencies so that the mode of this data is 6, written Mod = 6.

b. The are two of the most frequent data values appear with the greatest frequency, the value of 7 and a value of 8, both have frequency 3. Bimodal of this data is 7 and 8.

Other measures of central location, but is more often used as a measure of the spread are first quartile, third quartile 3, minimum value and maximum value. Quartiles together with the minimum value and maximum value is often termed a series five statistical or called robust statistical. Is robust because in addition to the measures of central located, as well as the size of the spread.

### 2.1.2 MEASURES OF CENTRAL LOCATION FOR GROUPED DATA

Grouped data is data that is summarized according to classes. First, the data set is partitioned into subsets, called class. Wide class called class intervals. Generally, the width of the class is created equal, although for the benefit or condition of the data, the class may not be the same width. Determination of the width of the class is done by first conducting a review of the range of the data, ie the width between the maximum value with the minimum value. A group of observations in a particular class is calculated amount, and expressed as the frequency of the class.

Example of the grouped data :

**Table 2.1** Grouped data

| **Classes** | **Frequencies ( $f_i$ )** | **Midpoint ( $x_i$ )** | **$f_i . x_i$** | **Cumulative Frequencies ( $F_i$ )** |
|---|---|---|---|---|
| 0 - < 5 | 2 | 2,5 | 5 | 2 |
| 5 - < 15 | 6 | 10 | 60 | 8 |
| 15 - < 25 | 20 | 20 | 400 | 28 |
| 25 - < 35 | 15 | 30 | 450 | 43 |
| 35 - < 45 | 8 | 40 | 320 | 51 |
| 45 - < 100 | 4 | 72,5 | 290 | 55 |
| Total | n = 55 | | 1.525 | |

The formula for determining measures of central location are :

Mean : $$\overline{X} = \frac{\sum_{i=1}^{p} f_i . x_i}{n}$$

Median : $$Md = L + \frac{f_c}{f_m} . W$$

Mode : $\text{Mod} = L + \frac{d_1}{d_1 + d_2} . W$

**Worked Example 2.4 :**

Calculate the mean, median, and mode for grouped data in Table 6

**Worked Solution** :

Mean : $\bar{X} = \frac{5 + 60 + \ldots + 290}{55} = \frac{1.525}{55} = 27.7273$

Median class is the class that contains the value that lies in the middle of the data, ie (n / 2)th observation, with n / 2 = 55/2 = 27.5.

In Table 6 the cumulative frequency column, from the first class to 2nd class, there are 8 observations, and until the 3rd class has been there are 28 observations. This means that the 27.5th data falls in the 3rd class, which is the class interval "15 - <25". This the 3rd class a median class has a lower limit of 15 and a frequency of 20.

Median : $\text{Md} = L + \frac{f_c}{f_m} . W$

Median : $\text{Md} = 15 + \left( \frac{27.5 - 8}{20} \right) .10 = 15 + \left( \frac{19.5}{20} \right) 10 = 24.75$

Mode : $\text{Mod} = L + \frac{d_1}{d_1 + d_2} . W$

Modus : $\text{Mod} = 15 + \frac{(20 - 6)}{(20 - 6) + (20 - 15)} . 10$

$= 15 + \frac{14}{14 + 5} . 10 = 22.07$

## 2.2 DISPERSION OF DATA

Known to some measure of the dispersion of the data, such as:

a. Range and modified Ranges
b. The average deviation
c. Standard deviation
d. Variance
e. Coefficient of variation

but only some of this measure of the dispersion are discussed in this book.

### 2.2.1 DISPERSION OF UNGROUP DATA

**a. Range**

Range of a set of data states that width of the data. For data of a sampling result, the range is calculated from the difference between the maximum value ($X_{max}$) with a minimum value ($X_{min}$).

In an ordered data set : $X_{min}$ , $X_2$ , $X_3$ , …, $X_{max}$.

Range ; $Rg = X_{max} - X_{min}$

**b. Modified Ranges**

Modified range based on various measures of the partitioning a set of ordered data, according to some of the same parts. The partitioning of data into several parts, can be the quartiles ( 4 parts), the deciles (10 parts), and the percentiles (100 parts).

Quartiles measure are defined as follows

1. First quartile (Q1) is the observation value that lies at the location to ((n +1) / 4)th in the ordered data. Partitioning the data on 25% down and 75% up.

2. Second quartile (Q2) is the observation value that lies in the location of (2 (n +1) / 4)th in the ordered data. Partitioning the data on 50% down and 50% up. Second quartile of a data set is the same as the median of the data.
3. Third quartile (Q3) is the observation value that lies in the location of the (3 (n +1) / 4)th in the ordered data. Partitioning the data on 75% down and 25% up.

**c. Variance**

For the set of data on population $X_1, X_2, \ldots, X_N$, variance of population can be calculated with the following formula.

$$\text{Variance} : \quad \sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}$$

For the set of data on sample $X_1, X_2, \ldots, X_n$, variance of sample can be calculated with the following formula.

$$\text{Variance} : \quad S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n - 1}$$

$$= \frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n - 1} \qquad \text{atau} \qquad S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\,\overline{X}^2}{n - 1}$$

**d. Standard Deviation**

Standard deviation is the square root of the variance, ie the set of data on population $X_1, X_2, \ldots, X_N$, standard deviation of population can be calculated with the following formula.

Standard deviation : $$\sigma = \sqrt{\frac{\sum X_i^2 - N\overline{X}^2}{N}}$$

For the sample data set $X_1, X_2, \ldots, X_n$ standard deviation of sample can be calculated by the following formula.

Standard deviation : $$S = \sqrt{\frac{\sum X_i^2 - n\overline{X}^2}{n - 1}}$$

**Worked Example 2.5 :**

For the following set of data sample : 4 7 8 9 22 26 29 37 40 48

Calculate :

a. Range
b. Interquartile range
c. Variance
d. Standard deviation

**Worked solution :**

a. Range : $Rg = X_{max} - X_{min} = 48 - 4$

$= 44$

b. Interquartile range :

$Q_1$ lies in ((n +1) / 4)th observation.

with (n+1)/4 = ( 10 +1)/ 4 = 2.75

So that $Q_1 = X_2 + 0.75 ( X_3 - X_2 )$

$Q_1 = 7 + 0.75 ( 8\text{-}7) = 7.75$

$Q_3$ lies in (3(n+1)/4)th observation

with 3(n+1)/4 = 3 ( 10 +1)/ 4 = 8.25

So that $Q_3 = X_8 + 0.25 ( X_9 - X_8 )$

$Q_3 = 37 + 0.25 ( 40\text{-}37) = 37.75$

Interquartile range $Q_3 – Q_1 = 37.75 – 7.75 = 30$

Range of data after trimmed 25% down and 25% up is 30.

c. Variance :

To calculate the variance, first performed the following preliminary count :

$$\sum X_i = 4 + 7 + \ldots + 48 = 230$$

$$\sum X_i^2 = 16 + 49 + 64 + 81 + 484 + 676 + 841 + 1369 + 1600 + 2304 = 7484$$

$$\text{Variance } S^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n - 1}$$

$$= \frac{7484 - ((230)^2/10)}{10 - 1}$$

$$= 243.78$$

d. Standard deviation :

The standard deviation is the square root of the variance. If you have obtained variance $S^2 = 243.78$ then the standard deviation is

$S = \sqrt{243.78} = 15.61$

### 2.2.2 DISPERSION OF GROUPED DATA

In grouped data, how to calculate the first quartile (Q1) and third quartile (Q3) equal to how to calculate the median.

$$Q_i = L_{q_i} + \frac{\frac{n}{4} i - F_{i-1}}{f_{q_i}} . W_{q_i}$$

Where,

$Q_i$ = (i)th quartile ( i = 1, 2 , 3 )

$Lq_i$ = lower limit of the (i)th quartile class interval

$F_{i-1}$ = cumulative frequency for a class below of the (i)th quartile class interval

$fq_i$ = frequency of the (i)th quartile class interval

$Wq_i$ = width of the (i)th quartile class interval

The formula for calculating the variance of the data grouped is

$$S^2 = \frac{\sum_{i=1}^{p} f_i (X_i)^2 - \left( \left( \sum_{i=1}^{p} f_i X_i \right)^2 / n \right)}{n-1}$$

The standard deviation for the grouped data is the square root of the variance for the grouped data, ie $S = \sqrt{S^2}$

**Worked Example 2.6 :**

Calculate the interquartile range, variance, and standard deviation for the grouped data in Table 2. 2.

**Table 2.2 The grouped data**

| **Classes** | **Frequencies ( $f_i$ )** | **Midpoint ( $x_i$ )** | **$f_i . x_i$** | **$f_i . (x_i)^2$** | **Cumulative Frequencies ( $F_i$ )** |
|---|---|---|---|---|---|
| 0 - < 5 | 2 | 2,5 | 5 | 12,50 | 2 |
| 5 - < 10 | 5 | 7,5 | 37,5 | 281,25 | 7 |
| 10 - < 15 | 8 | 12,5 | 100,0 | 1250,00 | 15 |
| 15 - < 20 | 6 | 17,5 | 105,0 | 1837,50 | 21 |
| 20 - < 25 | 4 | 22,5 | 90,0 | 2025,00 | 25 |
| 25 - < 45 | 1 | 35,0 | 35,0 | 1225,00 | 26 |
| Total | n = 26 | | 372,5 | 6631,25 | |

**Worked solution :**

Q1 lies in the (n/4)th observation, where n / 4 = 26/4 = 6.5 . In the column of cumulative frequencies, the 6.5th observation contained in the class " 5 - <10 ". thus,

$$Q_1 = L_{q_1} + \frac{f_{c_1}}{f_{q_1}} . W_{q_1}$$

$$Q_1 = 5 + \left( \frac{6.5 - 2}{5} \right) .5 = 5 + 4.5 = 9.5$$

$Q_3$ lies in the (3n/4)th observation, where 3n /4 = 3(26)/4 = 19.5 . In the column of cumulative frequencies, the 19.5th observation contained in the class ' 15 - < 20 '. thus,

$$Q_3 = L_{q_3} + \frac{f_{c_3}}{f_{q_3}} . W_{q_3}$$

$$Q_3 = 15 + \left( \frac{19.5 - 15}{6} \right) .5 = 15 + \left( \frac{4.5}{6} \right) 5 = 18.75$$

Interquartile range = $Q_3 - Q_1$ = 18.75 – 9.5 = 9.25

Variance :

$$S^2 = \frac{\sum_{i=1}^{p} f_i (X_i)^2 - \left( \left( \sum_{i=1}^{p} f_i X_i \right)^2 / n \right)}{n - 1}$$

$$= \frac{6631.25 - ((372.5)^2 / 26)}{26 - 1}$$

$$= 51.7789$$

Standard deviation:

The standard deviation for the grouped data is the square root of the variance for the grouped data, ie $S = \sqrt{51.7789} = 7.1958$

## WORKED EXERCISES SHEET

**Problem 2.1:**

A quality control inspector found the number of defective parts from production results for 15 production periods are as follows

3, 10, 9, 4, 6, 10, 12, 6, 10, 7, 11, 9, 1, 13, 3

a. Calculate the arithmetic mean of the data for the number of defective parts.

b. Find median and mode for data on the number of defective parts

**Worked Solution**:

a. $\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = ....$

b. Ordered Data : ……

Med = the value of the ((15 +1) / 2)th observation

= the value of the (.....)th observation

Med = ……

= …….

**Problem 2.2:**

Data on baggage masses in Table 1.3, are presented again in Table 2.2 below. Complete by filling in empty column , then calculate measures of central location :

a. Mean
b. Median
c. Mode

**Table 2.3** Frequency distribution

| **Classes (kg)** | **Frequencies ( $f_i$ )** | **Midpoint ( $x_i$ )** | **$f_i . x_i$** | **Cumulative Frequencies** |
|---|---|---|---|---|
| 0 - 9 | 5 | 5 | | 5 |
| 10 - 19 | 11 | 15 | | 16 |
| 20 - 29 | 20 | 25 | | 36 |
| 30 - 39 | 9 | 35 | | 45 |
| 40 - 49 | 13 | 45 | | 58 |
| 50 - 60 | 7 | 55 | | 65 |
| | N = 65 | 180 | | -- |

**Worked Solution**:

a. Mean : $\overline{X} = \dfrac{\sum_{i=1}^{p} f_i . x_i}{n}$

$$\overline{X} = \frac{(5 \text{ x } 5) + (11x15) + . \ . \ . + (7x55)}{65} = \frac{.....}{65} = .....$$

b. Median class is the class that contains the value that lies in the middle of the data, ie (n / 2)th observation, with n / 2 = ..... = .....

In Table 2.3 the cumulative frequency column, until the ..... class has been there are ..... observations. This means that the .....th data is in the ..... class, which is the class interval ".....". The median class has a lower limit of ..... and a frequency of .....

Median : $Md = L + \frac{f_c}{f_m} . W$

Median : Md = $\text{.....} + \left( \frac{\text{.....} - \text{.....}}{\text{.....}} \right) \text{.....} = \text{……..}$

c. Class that has the greatest frequency in Table 2.3
is ....., with frequency .....
Class before the mode class has a frequency of .....,
so that d1 = .... - .....
Class after the mode class has a frequency of ..... ,
so that d2 = .... - .....

Mode : $Mod = L + \frac{d_1}{d_1 + d_2} . W$

Mod = …….

**Problem 2.3:**

A basketball player contributed points to his team's in 12 matches as follows.

18, 3, 21, 15, 9, 34, 27, 10, 42, 6, 54, 62

Calculate:

a. Range
b. Interquartile range
c. Variance
d. Standard deviation

**Worked solution:**

a. Range: $Rg = X_{max} - X_{min}$ = ……. – ……..

= ……

b. Interquartile range :

$Q_1$ lies in ((n +1) / 4)th observation

with (n+1)/4 = ( …. +1)/ 4 = ……

so that $Q_1 = X_{...} + \ldots\ldots (X_{...} - X_{....})$

$Q_1$ = …… + …… ( …..-…….) = ……..

$Q_3$ lies in (3(n+1)/4)th observation

with 3(n+1)/4 = 3 ( …… +1)/ 4 = …….

So that $Q_3 = X_{....} + \ldots\ldots (X_{...} - X_{...})$

$Q_3$ = …… + ……. ( …….- ……..) = …….

Interquartile range $Q_3 - Q_1$ = ……. – …… = ……

c. Variance:

To calculate the variance, first performed the following preliminary count:

$\sum X_i = 18 + 3 + \ldots + 62 = \ldots\ldots\ldots$

$\sum X_i^2 = 18^2 + 3^2 + \ldots + 62^2 = \ldots\ldots$

$$\text{variance } S^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n - 1}$$

$$= \frac{\ldots\ldots\ldots - ((\ldots\ldots\ldots)^2 / 12)}{12 - 1}$$

$= \ldots\ldots\ldots$

d. Standard deviation :

The standard deviation is the square root of the variance. If you have obtained variance $S^2 = \ldots\ldots\ldots$ then the standard deviation is

$S = \sqrt{\ldots\ldots\ldots} = \ldots\ldots\ldots$

**Problem 2.4:**

Data on 100 student high recorded in Table 2.4. Sort observations from the lowest to the highest observation. Do Tolly to complement the contents of Table 2.5. Furthermore calculate:

a. Range
b. Interquartile range
c. Variance
d. Standard deviation

**Table 2.4** Raw data : Students height

```
       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

 [1,]  155  169  162  166  176  162  160  144  159   159
 [2,]  162  149  167  162  161  163  153  163  153   143
 [3,]  164  156  159  166  162  158  150  166  160   169
 [4,]  168  148  170  168  172  156  160  161  157   158
 [5,]  158  165  170  163  165  162  156  155  157   164
 [6,]  167  164  154  155  161  154  157  160  162   169
 [7,]  163  165  158  164  169  161  163  166  153   170
 [8,]  156  163  161  156  150  171  160  151  163   167
 [9,]  165  174  170  159  167  165  168  150  172   157
[10,]  165  160  154  164  158  160  170  160  164   159
```

**Table 2.5** Frequency distribution

| **Classes (cm)** | **Frequencies ( $f_i$ )** | **Midpoint ( $x_i$ )** | **Cumulative Frequencies** | **Relative Frequencies ( %)** |
|---|---|---|---|---|
| 140 - 144 | 2 | 142 | 2 | |
| 145 - 149 | 2 | 147 | 4 | |
| 150 - 154 | …. | …. | …. | |
| 155 - 159 | …. | …. | …. | |
| 160 - 164 | …. | …. | …. | |
| 165 - 169 | | | | |
| 170 - 174 | | | | |
| 175 - 179 | | | | |
| | N = 100 | ….. | -- | 100 |

**Worked solution:**

Q1 lies in the (n/4)th observation, where n / 4 = …./4 = …..

In the column of cumulative frequencies, the (.....)th observation contained in the class " .... - ..... ". thus,

$$Q_1 = L_{q_1} + \frac{f_{c_1}}{f_{q_1}} . W_{q_1}$$

$$Q_1 = .....+\left(\frac{.....-.....}{.....}\right)..... = .....+..... =.....$$

$Q_3$ lies in the (3n/4)th observation, where 3n /4 = ..... = ..... .
In the column of cumulative frequencies, the .....th observation contained in the class " ..... - ..... ". thus,

$$Q_3 = L_{q_3} + \frac{f_{c_3}}{f_{q_3}} . W_{q_3}$$

$$Q_3 = .....+\left(\frac{.....-.....}{.....}\right)..... = .....+..... =.....$$

Interquartile range = $Q_3 - Q_1$ = ……. – ……. = …….

Variance :

$$S^2 = \frac{\sum_{i=1}^{p} f_i (X_i)^2 - \left(\left(\sum_{i=1}^{p} f_i X_i\right)^2 / n\right)}{n-1}$$

$$= \frac{.....-((.....)^2 /.....)}{.....-1}$$

= ……

The standard deviation is the square root of the variance. If you have obtained variance $S^2$ = …….. then the standard deviation is

$S = \sqrt{.....} = .....$.

**Problem 2.5:**

Prove that three formulas for calculating the variance is the same

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n - 1},$$

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n - 1}, \text{ and}$$

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n - 1}$$

C H A P T E R **3**

# NORMAL PROBABILITY DIRTRIBUTION

**Basic competence**

Understanding the properties of the normal distribution, and can use the standard normal distribution table to find the probability of a set of random variables, and conversely can determine the value of a random variable based on probability which is given restrictions

**Indicators:**

1. Find the probability of a set of random variables, which is normally distributed, by use the standard normal distribution table.
2. Determine the value of a random variable, which is normally distributed, based on probability which is given restrictions
3. Using the properties of the normal distribution to solve problems related to the probability of an event.

# 3. NORMAL PROBABILITY DIRTRIBUTION

## 3.1 NORMAL CURVE

Continuous probability distribution are most commonly used in inferential statistics is the normal distribution. Graph of normal distribution is often called the normal curve, which is a bell-shaped curve that picture as shown in Figure 3.1

There are so many clusters of data encountered everyday that can be approached as a distribution in the form of a normal distribution or they had a probalility distribution resembles a normal distribution. As a result, the normal distribution and normal curve graph is very often used to cluster the data coming from various fields, such as industry, economy, agriculture, and for research. Normal distribution is often called the Gaussian distribution, in honor of Gauss (1777 - 1855), who managed to find the equation of the normal distribution through the study of error in repeated measurements of the same object. Previously, DeMoivre has managed to reduce the mathematical equation for the normal curve in 1733.

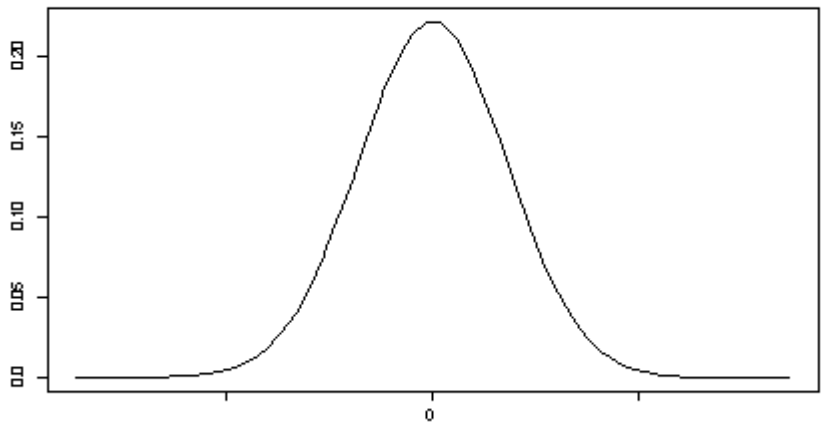



**Figure 3.1** The normal curve

A continuous random variable X which has a bell-shaped distribution as in Figure 3.1 is called a normal random variable. The mathematical equation for the probability distribution of normal random variable is determined by two parameters $\mu$ and $\sigma$, namely the mean and standard deviation. Therefore we symbolize the values of density function for X is $n(x;\ \mu,\ \sigma)$.

If X is a normal random variable with a mean value $\mu$ and variance $\sigma^2$, then the equation of the normal curve is

$$n(x;\ \mu,\ \sigma) = \frac{1}{\sqrt{2\pi}\ \sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \text{ untuk } -\infty < x < \infty, \qquad \textbf{( Z1 )}$$

whereas in this case $\pi$ = 3.14159… and e = 2.71828.

## 3.2 AREA UNDER THE NORMAL CURVE

Curves of any continuous probability distribution or density functions are such that the area under the curve was constrained by $x = x_1$ and $x = x_2$ is equal to the probability that a random variable X taking values between by $x = x_1$ and $x = x_2$. Thus, the normal curve in Figure 3.2. $P(x_1 < X < x_2)$ is expressed by the area shaded.

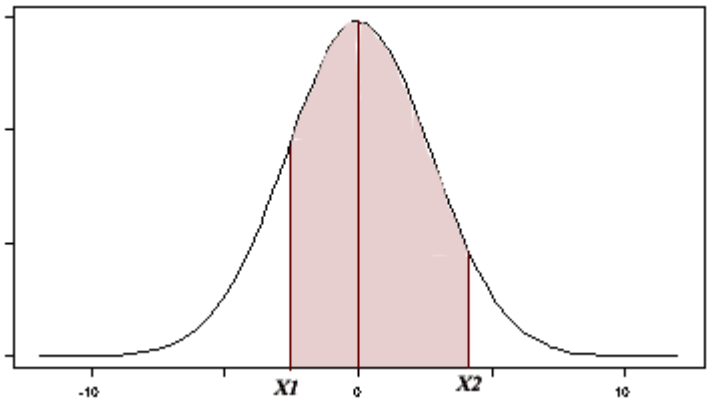



**Figure 3.2.** $P(x_1 < X < x_2)$ is expressed by the area shaded

The properties of the normal distribution such that, the proportion of all the observations of a normally distribution variable X that fall withing a range of n standard deviations on both side of the mean value is :

a. 68.26 percent of all X values fall withing the range of one standard deviation on both sides of the mean value, ie $\mu - 1\sigma$ to $\mu + 1\sigma$.

b. 95.44 percent of all X values fall withing the range of two standard deviations on both sides of the mean value, ie $\mu - 2\sigma$ to $\mu + 2\sigma$

c. 99.73 percent of all X values fall withing the range of three standard deviations on both sides of the mean value, ie $\mu - 3\sigma$ to $\mu + 3\sigma$.

The properties of a, b, and c above are visually shown in Figure 3.3 below

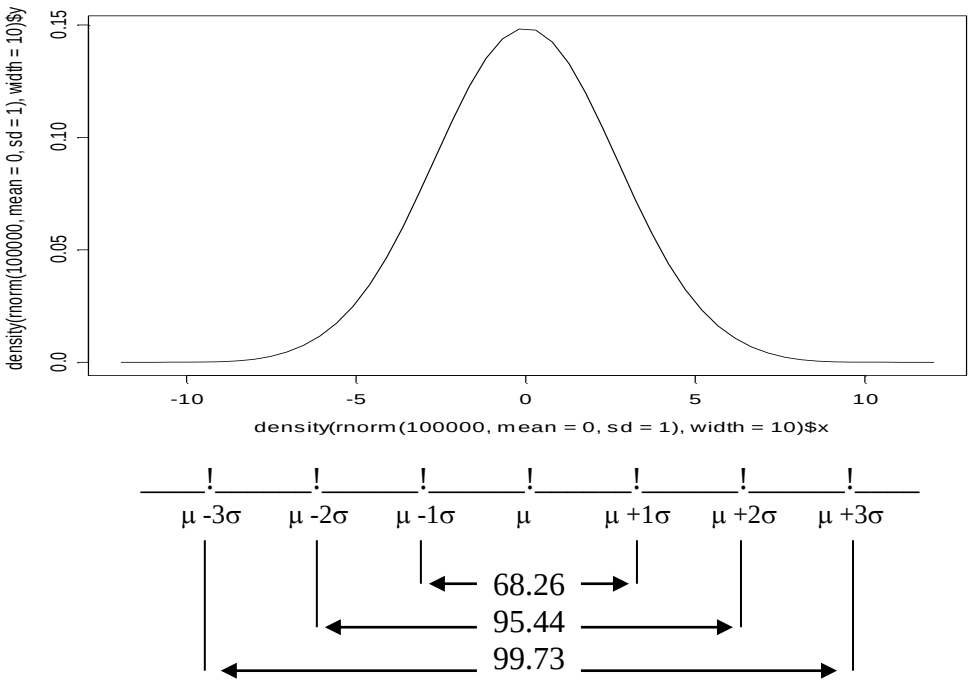



**Figure 3.3** The properties of the normal distribution

## 3.3 STANDARD NORMAL DISTRIBUTION Z

Not efficient if we always tried to organize a separate table for each of the normal curve for the possible each pair of values $\mu$ and $\sigma$. But we have to use the table if we want to avoid having to use the integral calculus. Fortunately, we can transform any observation from any normal random variable X be the value of the standard normal random variable Z with a mean value 0 ( $\mu$ =0 ) and the variance 1 ( $\sigma^2$ =1 ).

This can be done through the transformatio $$Z = \frac{X - \mu}{\sigma} \qquad \textbf{( Z2 )}$$

The mean value of Z is zero, because

$$E(Z) = \frac{1}{\alpha} E(X - \mu) = \frac{1}{\alpha}(\mu - \mu) = 0$$

While the variance is

$$\alpha_z^2 = \alpha_{(X-\mu)/\sigma}^2 = \alpha_{X/\sigma}^2 = \frac{1}{\sigma^2}\sigma_x^2 = \frac{\sigma^2}{\sigma^2} = 1.$$

**Definition**: Standard Normal Distribution. Distribution of the normal random variable with mean value 0 and a standard deviation 1 is called the standard normal distribution.

If X is between $x = x_1$ and $x = x_2$, the random variable Z will be among the values of equivalent

$$z_1 = \frac{x_1 - \mu}{\sigma} \qquad \text{and} \qquad z_2 = \frac{x_2 - \mu}{\sigma} \qquad \textbf{( Z3 )}$$

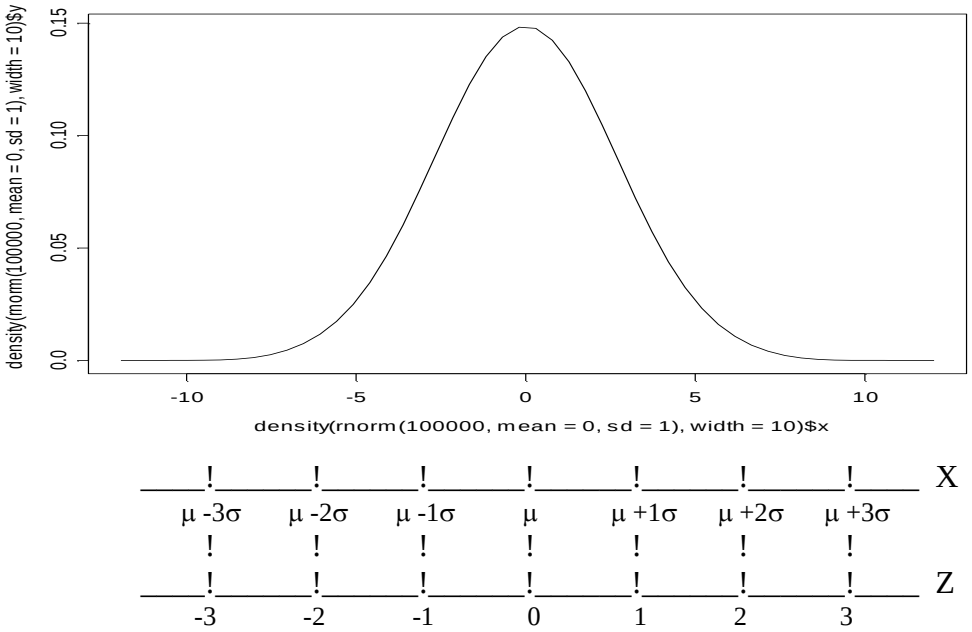



**Figure 3.4** Normal random variable X
and transformed into standard normal distribution Z

Distribution of the origin and distribution of the results of the transformation is illustrated in Figure 3.4. Because all values of X that falls between $x_1$ and $x_2$ have the equivalent values of z between $z_1$ and $z_2$, then the area under the normal curve X between $x = x_1$ and $x = x_2$ in Figure 3.1 is equal to the area under the standard normal curve between the values of Z the results of the transformation $z = z_1$ and $z = z_2$ .Thus

$$P(x_1 < X < x_2) = P(z_1 < Z < z_2)$$

## 3.4 USING THE STANDARD NORMAL DISTRIBUTION TABLE

In using the standard normal table, we reduce some tables on the area of the normal curve into only one, which is derived from the standard normal distribution. Table A.1 lists the area under the standard normal curve which is the value of P (0 < Z < z) for various values of z from 0 to 5.49. To illustrate the use of this table, let us count the probability that Z takes values between 0 to

1.74. First find the value of z is equal to 1.7 in the leftmost column, and then view along the raw until the column below 0.04, there we read 0.4591. Thus P(0 <Z <1.74) = 0.4591.

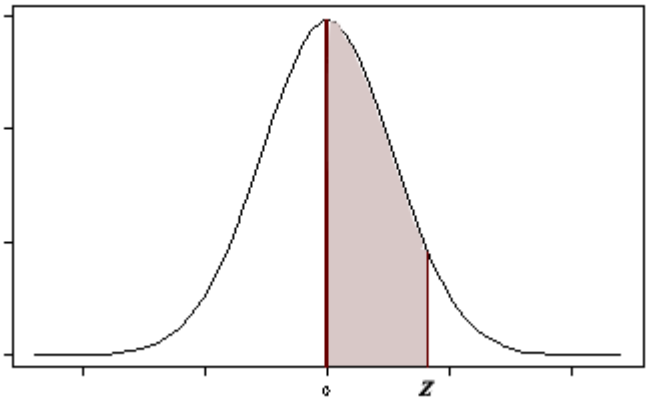



**Figure 3.5 Probability of Z between 0 to z**
**P ( 0 < Z < z )**

Sometimes, we are asked to find the value of z for a given probability value and the z value is between the values listed in Table A.1. For simplicity, we will take the value of z in the table which its probability value closest to probability value known. However, if the probability is known that falls right in the middle between the two tables, then the value of z we get an average of the two values of z matching both values in the table. For example, to find the value of z that generates probability for 0.2975, which lies between 0.2967 and 0.2995 in Table A.1, we will take z = 0.83, since 0.2975 is closer to 0.2967. But for a probability at 0.2981, which falls right in the middle of 0.2967 and 0.2995, we will take z = 0.835.

**Worked Example 3.1:** For a normal distribution with $\mu$ = 50 and $\sigma$ = 10, compute the probability that X taking a value between 40 and 62.

**Worked Solution:** The values of the equivalent of $x_1 = 40$ and $x_2 = 62$ is

$$z_1 = \frac{40-50}{10} = -1.0$$

$$z_2 = \frac{62-50}{10} = 1.2.$$

thus, $P\,(\,40 < X < 62\,) = P\,(\,-1.0 < Z < 1.2\,)$

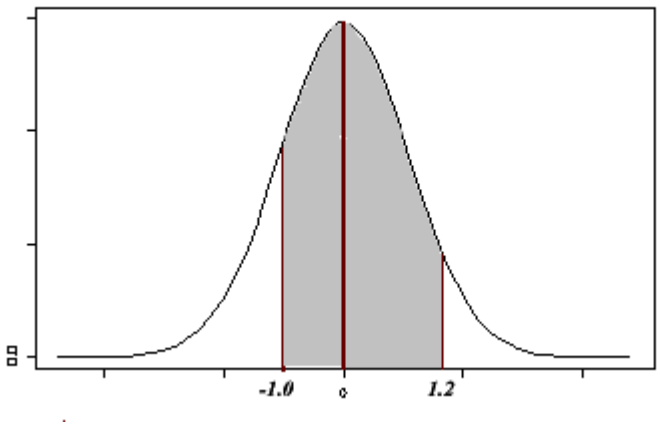



**Figure 3.6** ***P ( -1.0 < Z <1.2 )***

P (-1.0 <Z <1.2) is given by the dark areas in Figure 3.6. This area can be obtained by adding the area from -1 to 0 to the area from 0 to 1.2. Using Table A.1, we obtain as follows

P (45< X < 62 ) = P (-1.0< Z <1.2 )

= P ( -1.0 < Z < 0 ) + P ( 0 < Z < 1.2 )

= P ( 0 < Z < 1 ) + P ( 0 < Z < 1.2 ) ( symmetry properties of Z)

= 0.3413 + 0.3849

= 0.7262

**Worked Example 3.2:** For a normal distribution with $\mu$ = 20 and $\sigma$ = 5, compute the probability that a random variable X taking a value of :

a. between 8 to 16 b. X < 12

**Worked Solution:**

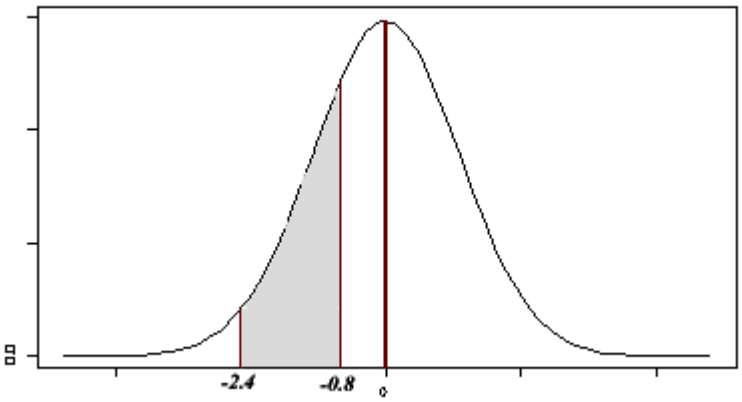



**Figure 3.7 P ( -2.4 < Z < -0.8 )**

a. $P(8 < X < 16) = P\left(\frac{8-20}{5} < Z < \frac{16-20}{5}\right)$

$= P(-2.4 < Z < -0.8)$

$= P(0.8 < Z < 2.4)$

$= P(0 < Z < 2.4) - P(0 < Z < 0.8)$

$= 0.4918 - 0.2881 = 0.2037$

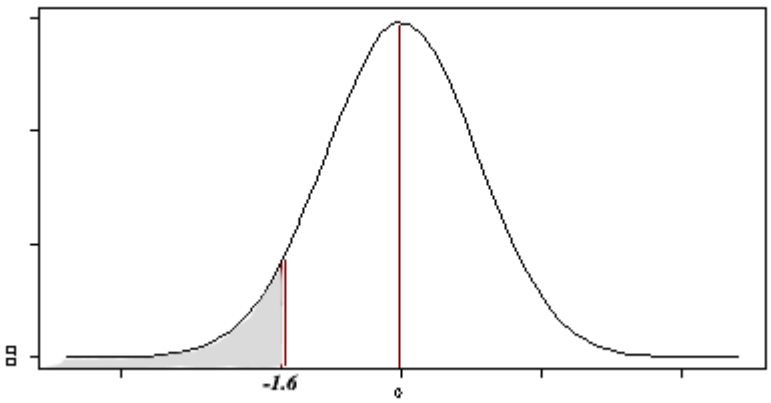



**Figure 3.8 P ( Z < -1.6 )**

b. $P( X < 12 ) = P\left( Z < \frac{12 - 20}{5} \right)$

$= P( Z < -1.6 )$

$= P( Z > 1.6 )$

$= 0.5 - P( 0 < Z < 1.6 )$

$= 0.5000 - 0.4452 = 0.0548$

**Worked Example 3:** Given a normal distribution with $\mu$ = 40 dan $\sigma$ = 6. Calculate the value of x that its area below 38%.

**Worked Solution:** Two exemplary previously completed work from the value of x to z values and then proceed to calculate the area desired. In this example we reverse the process, start with the opportunities or the area is known, then calculate the value of z and the last determines the value of x by altering the formula

$$z = \frac{x - \mu}{\sigma} \quad \text{make} \quad x = \sigma z + \mu$$

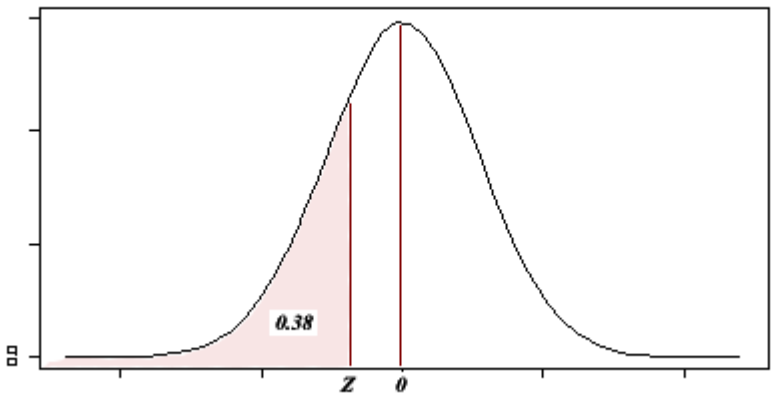



**Figure 3.9 P ( Z < z ) = 0.38**

Area on the left of z value is 0.38 or P(Z <z) = 0.38, shown in Figure 3.9. Because z lies on left of 0 then z value is negative. P (Z <z) = 0.5 - P (z <Z <0), then P(z <Z <0) = 0.12. If applied to the symmetry properties of the normal distribution, P(z <Z <0) = P(0 <Z <-z). From Table A.1 we get P(0 <Z <0.31) = 0.12, subsequently obtained - z = 0.31 or z = -0.31 thus

$x$ = (6)(-0.31) + 40

= 38.14

**EXERCISES 3**

1. If the random variable Z , which is standard normally distribution, use the standard normal table Z to find:

   a. P ( -2 < Z < 1,5 )
   b. P ( Z< 1.97 )
   c. P ( Z > 1.96 )

2. If the random variable X , which is normally distribution, has a mean 43 and a standard deviation of 5 , find the probability:

   a. P ( 40 < X < 49 )
   b. P ( X < 41 )
   c. P ( X > 50 )

3. Use the standard normal table Z to find the value of z, if the right side of the area $\alpha = 0.05$ ( $Z_{0.05}$ )

C H A P T E R **4**

# HYPOTHESIS TESTING

**Basic competence**

Capabilities in the testing of hypotheses as represent the main application of principles of statistical inference to solving problems and making conclusions

**Indicators:**

1. Perform steps hypothesis testing appropriately, covering formulation of a test hypothesis, determine and calculate test statistics, and establish critical region, in the testing hypoteses about equality or difference for the mean value of single population.
2. Infer and perform testing hypoteses about equality or difference for the mean values of the two populations.

# 4. HYPOTHESIS TESTING

The main aspect in the application of inferential statistics, after estimation is hypothesis testing. In statistics, hypothesis testing is an important part to making a decision. By testing the hypothesis of the researchers will be able to answer the questions posed, stating the rejection or acceptance of the hypothesis.

Hypothesis is a temporary answer before the experiment carried out, based on the results of the study of literature. Hypotheses often contain a statement that is neutral or common occurrence. Truth of the hypothesis is certainly never known except if carried out observations of the entire population. To do this it is extremely inefficient especially when the population size is very large.

Withdrawal of a random sample from a population, the observed characteristics and then compared with the hypothesis put forward is a step to test the hypothesis. If a random sample is an indication that supports the hypothesis, then the hypothesis is accepted. Conversely, if a random sample gives an indication that contrary to the hypothesis, then the hypothesis is rejected.

Definition of a hypothesis is accepted or rejected is not absolute. One hypothesis is rejected does not mean that the hypothesis is wrong, but the data does hint that there have been changes in the characteristics of the hypothesized population. Acceptance of the hypothesis means that is not enough evidence to accept the alternative hypothesis.

Statistical hypotheses divided into two statement, namely the null hypothesis ($H_0$) and the alternative hypothesis ($H_1$). Statement wishing rejected his truth set as the null hypothesis, while his opponent hypothesis set as a alternative hypothesis.

In the hypothesis testing are two kinds of errors, namely the type I error and the type II error. Type I error is a mistake to reject $H_0$ when in fact $H_0$ is true. Type II error is the error accept $H_0$ when in fact $H_1$ is true.

Probability of a type I error occurs is denoted by $\alpha$ (alpha), which is often also called the significance level. While probability of type II error occurs denoted by $\beta$ (beta). In statistical hypothesis testing, probability of the two types of errors are expected balanced and tailored to the problem under study. Probability value 1 - $\alpha$ is called the confidence interval, expressed probability accept $H_0$ when it is true. Probability value 1 - $\beta$ is called the power of the test, states the probability to reject $H_0$ when $H_0$ is false.

## 4.1 HYPOTHESIS TESTING FOR SINGLE POPULATION MEAN

The data of a sample taken from a population can be used to assess the characteristics of its population. One of the most interesting characteristics of the population to be studied is the mean value. There are three forms of the formulation of a test hypothesis for the mean value of single population, namely:

1 $H_0 : \mu = \mu_o$
$H_1 : \mu \neq \mu_o$

2 $H_0 : \mu \geq \mu_o$
$H_1 : \mu < \mu_o$ **( H1 )**

3 $H_0 : \mu \leq \mu_o$
$H_1 : \mu > \mu_o$

In the conduct of any hypothesis test, there are five steps to be followed :

a. Formulating a test hypothesis, namely the null hypothesis ($H_0$) and the alternative hypothesis ($H_1$). It is one of the formulas above.
b. Calculate the sample statistics, ie the mean value of the sample $\overline{X}$

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \textbf{( H2 )}$$

c. Determine and calculate the test statistic.

If the variance of the population ($\sigma^2$) is known or variances of the sample is stable to predict his the population variance ($\sigma^2$), then the test statistic used is the standard normal (z). Variance of sample ( $S^2$ ) is stable to predict the population variance ($\sigma^2$) when his sample size n ≥ 30. Further written $S^2 = \hat{\sigma}^2$. Statistical test used is the Z-standard normal :

$$z = \frac{\overline{X} - \mu_0}{\sigma_{\bar{x}}} = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \quad \text{or} \quad Z = \frac{\overline{X} - \mu_0}{\hat{\sigma} / \sqrt{n}} \qquad \textbf{( H3 )}$$

If the population variance is unknown and the sample size n < 30, then the test statistic used is the student-t, as follows

$$t = \frac{\overline{X} - \mu_0}{S_{\bar{x}}} = \frac{\overline{X} - \mu_0}{S / \sqrt{n}} \qquad \textbf{( H4 )}$$

with n-1 degrees of freedom.

d. Establish critical region to reject the null hypothesis. Determination of the critical region is highly dependent on three things: the formulation of a alternative hypothesis, test statistics are used and level of significance.

If $H_1$: $\mu \neq \mu_0$ then critical region

$$|z| > z_{\alpha/2} \text{ or } |t| > t_{\alpha/2,db=n-1} \qquad \textbf{( H5 )}$$

The rejection region is shown in Figure 4.1

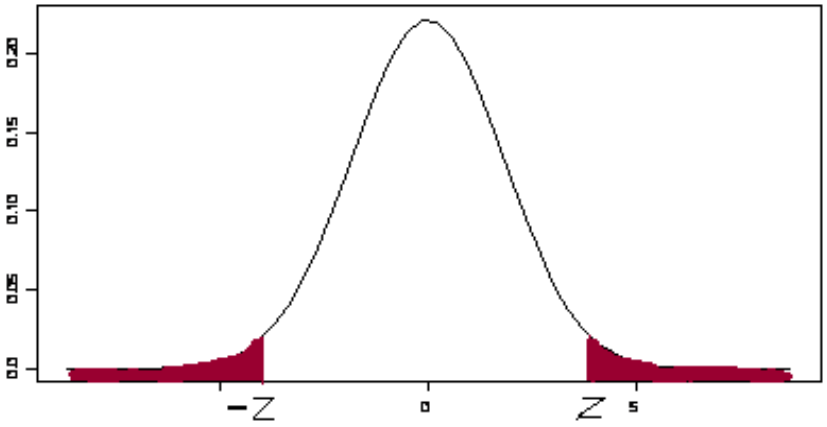



**Figure 4.1** The rejection region if used two-tailed test

If $H_1: \mu < \mu_0$ the critical region,

$$Z < -z_\alpha \quad \text{or} \quad t < -t_{\alpha,db=n-1} \qquad \textbf{( H6 )}$$

The rejection region or the rejection values is shown in Figure 4.2

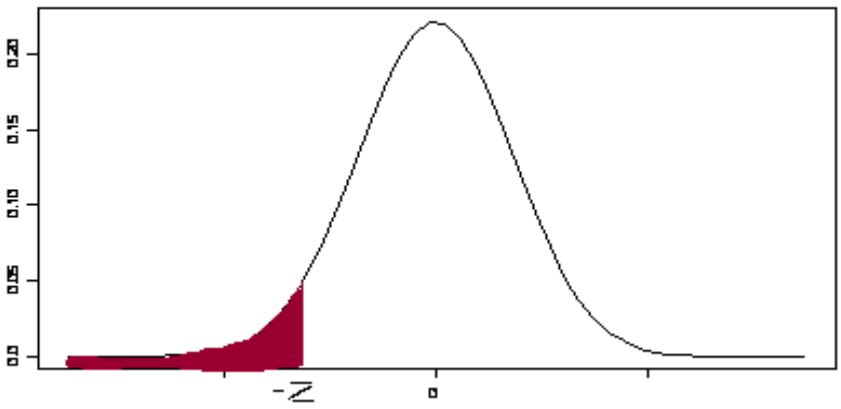



**Figure 4.2** The rejection region if used alternative hypothesis $\mathbf{H_1:}\ \mu < \mu_0$

Jika $H_1: \mu > \mu_0$ the critical region,

$$Z > z_\alpha \text{ or } t > t_{\alpha,db=n-1} \qquad \textbf{( H7 )}$$

The rejection region or the rejection values is shown in Figure 4.3

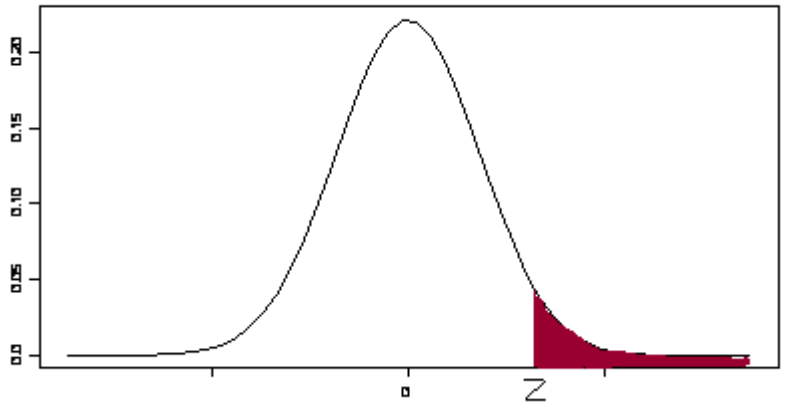



**Figure 4.3** The rejection region if used alternative hypothesis $\mathbf{H}_1$: $\mu > \mu_0$

**Worked Example 4.1:**

In an interesting collection of data with a sample from a population obtained (after the data are sorted) as follows :

| | | | | |
|---|---|---|---|---|
| 7.8910 | 8.7619 | 8.8901 | 9.1033 | 9.8530 |
| 9.8728 | 10.0135 | 10.3151 | 10.3221 | 10.8333 |
| 11.0250 | 11.5518 | 11.9744 | 12.3591 | 12.7180 |

If we know that the data is from a normally distributed population with variance is 2 and wanted to know whether the population still has a mean value of 10, then we will perform the following hypothesis test, if let's say you want to use significance level $\alpha$ = 0.05.

**Worked Solution:**

Test hypothesis:

$$H_0 : \mu = 10$$
$$H_1 : \mu \neq 10$$

significance level $\alpha$ = 0.05.

Since the population variance is known then the test statistic to be used is the Z statistic, namely :

$$Z = \frac{10.366 - 10}{\sqrt{2}/\sqrt{15}} = 1.002$$ and from tables : $|Z_{\alpha/2}| = |Z_{0.05/2}| = 1.96.$

Testing Results : Because $|Z| < |Z_{\alpha/2}|$, ie 1.002 < 1.96 then we accept $H_0$, meaning that the data is not enough significance to reject the hypothesis $H_0$.

**Worked Example 4.2:**

A factory battery (battery) claims that the average battery life to 55 hours. On the results of tests conducted on a production batch consisting of 40 batteries, gained an average of 50 hours of life, with a standard deviation of 11 734 hours. Perform hypothesis testing with a significance level of 1 percent that

a. Average battery life is 55 hours

b. Average battery life of less than 55 hours

**Worked Solution:**

Given : $\mu_0 = 55$; $\bar{x} = 50$; s = 11.734; n = 40

**a)**. Test hypothesis : $H_0 : \mu_0 = 55$

$H_1 : \mu_0 \neq 55$

Significance level : $\alpha$ = 0.01

Standard deviation for $\bar{X}$ : $s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{11.734}{\sqrt{(40)}} = 1.86$

Statistic Z : $z = \frac{\bar{X} - \mu_0}{s_{\bar{x}}} = \frac{50 - 55}{1.86}$

= -2.69

The critical value : Z-table = $z_{\alpha/2}$ = 2.58 (from Z-tables)

Test results : $|z| > z_{\alpha/2}$ , ie 2.69 > 2.58 . Furthermore, $H_0$ is rejected at the 1 percent significance level. It was concluded that the average battery life is not the same as 55 hours.

**b)**. Test hypothesis: $H_0: \mu_0 = 55$;
$H_1: \mu_0 < 55$

Significance level: $\alpha = 0.01$

Z-table = $z_{\alpha}$ = 2.33 ( from tables )

Test results: $z < -z_{\alpha}$, ie -2.69 < -2.33 . Furthermore, $H_0$ is rejected at the 1 percent significance level. It was concluded that the average battery life of less than 55 hours.

**Worked Example 4.3:**

The results of the last survey of the labor force states that in one year, the average number of days each employee absent due to illness was 15 days. A researcher using a random sample of 25 workers, and noted the absence of each worker in one year is as follows.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 5 | 25 | 10 | 0 | 3 | 50 | 12 | 14 | 40 |
| 12 | 32 | 8 | 4 | 47 | 20 | 14 | 16 | 10 |
| 1 | 22 | 58 | 5 | 23 | 18 | 9 | | |

Perform hypothesis testing at 5 percent significance level that
a. Average worker absent in a year of 15 days
b. Average worker absences greater than 15 days.

**Worked Solution:**

Given : $\mu_0 = 15$; n = 25

And $\bar{x} = 18.32$; s = 15.845

**(a)**. Test hypothesis : $H_0: \mu_0 = 15$;

$H_1: \mu_0 \neq 15$

Significance level : $\alpha = 0.05$

The critical value : $t = t_{\alpha/2;24} = 2.064$

(at the t-table, with (25-1) degrees of freedom)

Standard deviation for $\bar{X}$ : $s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{15.845}{\sqrt{(25)}} = 3.169$

$$t = \frac{\bar{X} - \mu_0}{s_{\bar{x}}} = \frac{18.32 - 15}{3.169}$$

$$= 1.05$$

Test results : $|t| < t_{\alpha/2,db=n-1}$, ie 1.05 < 2.064, then $H_0$ is accepted or not enough evidence to reject $H_0$. It was concluded that the average absences of workers in one year is equal to 15 days.

**b**. Test hypothesis : $H_0: \mu_0 = 15$

$H_1: \mu_0 > 15$

Significance level : $\alpha = 0.05$

The critical value : $t = t_{\alpha} = ……..$

Test results : ……..,ie……..., then $H_0$ is…….. It was concluded that ……..

## 4.2 HYPOTHESIS TESTING FOR TWO POPULATION MEANS

A research often aims to compare the average value of two independent populations. These two populations are parent populations from which two samples are drawn. Statistical method provides several ways to test differences in the average value of two populations. In parametric statistical method, the test can be carried out differences with the approach normal distribution (standard normal test-z), and the approach student-t distribution (Student's t-test). Both approaches this distribution can be done if the data is quantitative and continuous, so that the measurement data is feasible to apply the assumption of a normal distribution.

### 4.2.1 MEAN DIFFERENCE TEST USING STANDARD NORMAL TEST

**Approach a normal distribution**; The standard normal test-Z used if variance of the two populations are known or sample variance ( $S^2$ ) has stabilized to estimate population variance ( $\sigma^2$ ). Sample variance ( $S^2$ ) has stabilized to estimate population variance $\sigma^2$ when the sample size $n \geq 30$. Similarly, for sample size $n \geq 30$, mean ( $\overline{X}$ ) of the random variable X (unknown distribution) will form a normally distribution.

Hypothesis test to test the equality of the average of two independent populations (two-tailed test) is formulated as follows:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2 \qquad \text{( N1)}$$

Where,

$\mu_1$ = Mean of the first population

$\mu_2$ = Mean of the second population

To test the above hypothesis using approach the normal distribution, the Z value is calculated as follows:

**a.** If both population variance ( $\sigma_1^2$ and $\sigma_2^2$ ) is known, then

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\overline{X}_1 - \overline{X}_2}} \qquad (N2)$$

and

$$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}} \qquad (N3)$$

Where,

$\sigma_{\overline{X}_1 - \overline{X}_2}$ = Population standard deviation of difference between two means

$\sigma_1^2$ = Variance of the first population

$\sigma_2^2$ = Variance of the second population

$n_1$ = Number of observations in the first sample

$n_2$ = Number of observations in the second sample

**b**. If both the population variance is unknown but $n_1 \geq 30$ and $n_2 \geq 30$, then

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\hat{\sigma}_{\overline{X}_1 - \overline{X}_2}} \qquad (N4)$$

and

$$\hat{\sigma}_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{\hat{\sigma}_1^{\,2}}{n_1} + \frac{\hat{\sigma}_2^{\,2}}{n_2}} \qquad (N5)$$

Where,

$\hat{\sigma}_{\overline{X}_1 - \overline{X}_2}$ = Estimated value for the population standard deviation of difference between two means

$\hat{\sigma}_1^{\,2}$ = Estimated value for the first population variance

(use the value of the first sample variance $S_1^2$)

$\hat{\sigma}_2^2$ = Estimated value for the second population variance

(use the value of the second sample variance $S_2^2$)

$n_1$ = Number of observations in the first sample

$n_2$ = Number of observations in the second sample

**Testing criteria :**

If this test using significance level α and Z obtained from (N2) or (N4), then criteria to test the equality of the average of two independent populations (two-tailed test):

$H_0$ is accepted if $|$ Z-actual $| \leq Z_{\alpha/2}$-table , or

$Pr = [ P ( Z \leq -|z_{actual}|) + P ( Z \geq |z_{actual}|) ] \geq \alpha$ , otherwise

$H_0$ is rejected if $|$ Z-actual $| > Z_{\alpha/2}$-table , or

$Pr = [ P ( Z \leq -|z_{actual}|) + P ( Z \geq |z_{actual}|) ] < \alpha$.

To test whether the means of the first population is smaller than the mean of second population, *one-tailed test* was used, with the following formulation of the test hypothesis

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 < \mu_2 \qquad \text{( N6 )}$$

Testing the hypothesis (N6) with a normal distribution approach, using the same formula with the value of Z in testing hypotheses (N1), namely:

a. If both population variance ( $\sigma_1^2$ and $\sigma_2^2$ ) is known, then

$$\text{Z-hitung} = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\overline{X}_1 - \overline{X}_2}} \qquad \text{( N2 )}$$

b. If both the population variance is unknown but $n_1 \geq 30$ and $n_2 \geq 30$, then

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\hat{\sigma}_{\overline{X}_1 - \overline{X}_2}} \qquad (N4)$$

**Testing criteria:**

If this test use significance level $\alpha$ , then

$H_0$ is accepted if Z-actual $\geq$ - $Z_\alpha$-table or $Pr(Z \leq z_{actual}) \geq \alpha$ , otherwise

$H_0$ is rejected if Z-actual $<$ - $Z_\alpha$-tabel or $Pr(Z \leq z_{actual}) < \alpha$ .

**Worked Example 4.4:**

A golf ball factory introduced the latest production of golf balls (new type), and declared better than the old golf balls (old type). Distance of the two types of golf balls each having standard deviation : old type $\sigma_1 = 9.8$ meters and a new type $\sigma_2 = 7.1$ meters. A golfer wants to test the above statement by selected at random, hits 35 shots using balls of old types and hits 35 shots using balls of new types. Mean of distance in metres for each type respectively 20.1 meters and 23.6 meters. By using Significance level $\alpha = 0.05$, test the hypotheses that:

a. Both types of golf balls are not different
b. The old type is inferior to the new type.

**Worked Solution:**

| | | | |
|---|---|---|---|
| $\sigma_1$ | = 9.8 meter | $\sigma_2$ | = 7.1 meter |
| $\overline{X}_1$ | = 20.1 meter | $\overline{X}_2$ | = 23.6 meter |
| $n_1$ | = 35 | $n_2$ | = 40 |

(a) The test requires a two-tailed test:

Hypotheses: $H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

Significance level: $\alpha = 0.05$

Standard deviation: $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

$= \sqrt{\frac{(9.8)^2}{35} + \frac{(7.1)^2}{40}} \qquad = 2.0$

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

$$= \frac{20.1 - 23.6}{2.0} = -1.75$$

In the standard normal table; $\pm Z_{\alpha/2} = \pm 1.96$

Conclusion: $|Z| < Z_{\alpha/2}$, ie $1.75 < 1.96$, so not enough evidence to reject $H_0$. We conclude that there is no significant different of both types of golf balls.

b) This the test requires one-tailed test:

Hypotheses: $H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 < \mu_2$

Significance level: $\alpha = 0.05$

In the standard normal table; $-Z_\alpha = -1.64$

Conclusion: $Z < -Z_\alpha$-table, ie $-1.75 < -1.64$, so $H_0$ is rejected at the 5 percent significance level. We conclude that the old type is inferior to the new type.

**Worked Example 4.5:**

Data on test scores from the two models of learning, namely Student Teams Achievement Division (STAD) and Team Pairs Share (TPS) recorded consecutively in columns 2 and 3 in Table 4.1. Test the hypothesis that the average test scores between the two models: (a) did not differ, (b) TPS is better than STAD.

**Table 4.1**

| N0. | STAD $X_1$ | TPS $X_2$ | $X_1^2$ | $X_2^2$ |
|---|---|---|---|---|
| [1,] | 72 | 71 | 5184 | 5041 |
| [2,] | 67 | 78 | 4489 | 6084 |
| [3,] | 63 | 71 | 3969 | 5041 |
| [4,] | 70 | 66 | 4900 | 4356 |
| [5,] | 63 | 79 | 3969 | 6241 |
| [6,] | 69 | 75 | 4761 | 5625 |
| [7,] | 72 | 75 | 5184 | 5625 |
| [8,] | 64 | 72 | 4096 | 5184 |
| [9,] | 63 | 77 | 3969 | 5929 |
| [10,] | 62 | 77 | 3844 | 5929 |
| [11,] | 56 | 71 | 3136 | 5041 |
| [12,] | 68 | 79 | 4624 | 6241 |
| [13,] | 65 | 74 | 4225 | 5476 |
| [14,] | 68 | 71 | 4624 | 5041 |
| [15,] | 59 | 66 | 3481 | 4356 |
| [16,] | 69 | 78 | 4761 | 6084 |
| [17,] | 66 | 72 | 4356 | 5184 |
| [18,] | 64 | 71 | 4096 | 5041 |
| [19,] | 60 | 76 | 3600 | 5776 |
| [20,] | 61 | 77 | 3721 | 5929 |
| [21,] | 68 | 76 | 4624 | 5776 |
| [22,] | 65 | 77 | 4225 | 5929 |
| [23,] | 67 | 77 | 4489 | 5929 |
| [24,] | 71 | 70 | 5041 | 4900 |
| [25,] | 74 | 78 | 5476 | 6084 |
| [26,] | 64 | 73 | 4096 | 5329 |
| [27,] | 69 | 81 | 4761 | 6561 |
| [28,] | 65 | 66 | 4225 | 4356 |
| [29,] | 64 | 72 | 4096 | 5184 |
| [30,] | 68 | 72 | 4624 | 5184 |
| | $\sum X_1 = 1,976$ | $\sum X_2 = 2,218$ | $\sum X_1^2 = 130.646$ | $\sum X_2^2 = 164,456$ |

**Worked Solution:**

Means : $\overline{X}_1 = \frac{\sum_{i=1}^{30} X_i}{30} = \frac{1976}{30} = 65.867$ $\overline{X}_2 = \frac{\sum_{i=1}^{30} X_i}{30} = \frac{2218}{30} = 73.933$

Variance : $S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\,\overline{X}^2}{n-1}$

$S_1^{\,2} = \frac{(130646) - 30\,(65.867)^2}{30-1} = 17.016$ $S_2^{\,2} = \frac{(164456) - 30\,(73.933)^2}{30-1} = 16.271$

(a) The test requires a two-tailed test:

Hypotheses: $H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

Significance level: $\alpha = 0.01$

Standard deviation: $\hat{\sigma}_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{\hat{\sigma}_1^{\,2}}{n_1} + \frac{\hat{\sigma}_2^{\,2}}{n_2}}$

$= \sqrt{\frac{(17.016)}{30} + \frac{(16.271)}{30}} = 1.053$

$Z = \frac{\overline{X}_1 - \overline{X}_2}{\hat{\sigma}_{\overline{X}_1 - \overline{X}_2}}$

$= \frac{65.867 - 73.933}{1.053} = -7.66$

The test results:

a). In the standard normal table; $\pm Z_{\alpha/2} = \pm 2.57$

Conclusion: $|Z| > Z_{\alpha/2}$, ie 7.66 > 2.57, then $H_0$ is rejected at the 1 percent significance level. We conclude that there is a significant difference in the average test scores by two learning models, STAD and TPS.

b) This the test requires one-tailed test:

Hypotheses: $H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 < \mu_2$

Significance level: $\alpha = 0.01$

In the standard normal table; $-Z_\alpha$-table $= -2.33$

Conclusion: $Z < -Z_\alpha$-table, ie $-7.66 < -2.33$, so $H_0$ is rejected at the 1 percent significance level. We conclude that TPS is better than STAD.

**Solving using computerized:**

The data in columns 2 and 3 of Table 2 stored in the S-plus, called *Model of Learning*. Its data in the form of a matrix, and has the number of row 30 and column 30 and 2 respectively.

```
> t.test(Model of Learning [, 1], Model of Learning [, 2], alternative="two.sided", mu=0, paired=F,
 + var.equal=F, conf.level=.95)

Standard Two-Sample t-Test

data:  x: stad , and y: tps
t = -7.658,   df = 58,   p-value = 0.0000
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
     -10.175206      -5.958127

sample estimates:
 mean of x          mean of y
  65.86667          73.93333


> t.test(Model of Learning [, 1], Model of Learning [, 2], alternative="less", mu=0, paired=F,
 + var.equal=F, conf.level=.95)

Standard Two-Sample t-Test

data:  x: stad , and y: tps
t = -7.658,    df = 58,     p-value = 0.0000
alternative hypothesis: true difference in means is less than 0
```

```
95 percent confidence interval:
      NA       -6.305911

sample estimates:
 mean of x          mean of y
  65.86667          73.93333
```

**Comments:**

Although the above results do computerized with t test, but its the same as the Z-test. These results can be directly compared with the critical value of the Z-test.

a). The conclusion of the two-tailed test; Pr = [ P ( $Z \leq -|$z-actual$|$) + P ( $Z \geq |$z-actual$|$ ] < $\alpha$, ie 0.0000< 0.01. If used manual testing criteria, then $|$ Z-actual $|$ > $Z_{\alpha/2}$, ie 7.658 > 2.57, then $H_0$ is rejected at the 1 percent significance level. We conclude that there is a significant difference between the average test scores of STAD and TPS.

b). The conclusion of the two-tailed test; Pr ( $Z \leq z_{actual}$ ) < $\alpha$ , ie 0.0000 < 0.01, then $H_0$ is rejected at the 1 percent significance level. We conclude that TPS is better than STAD.

### 4.2.2 MEAN DIFFERENCE TEST USING T-STUDENT TEST

When dealing with situations where variance of the two populations ($\sigma_1$ and $\sigma_2$) are not known and sample sizes small ($n_1 < 30$ atau $n_2 < 30$ ), the tes procedure make use of the t distribution. The procedural steps are identical to those specified above for large independent samples, however to special case noted below.

1. Hypothesis test to test the equality of the mean of two independent populations (two-tailed test) is formulated as follows:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2 \qquad (T1)$$

Where,

$\mu_1$ = Mean of the first population

$\mu_2$ = Mean of the second population

2. To test whether the means of the first population is smaller than the mean of second population, *one-tailed test* was used, with the following formulation of the test hypothesis:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 < \mu_2 \qquad (T2)$$

To test the both hypotheses above, using approach the t-student distribution, the test statistical t-value is calculated as:

**a.** Special case, there are indications that both populations variances are equal. ( $\sigma_1^2 = \sigma_2^2$ ),

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{\overline{X}_1 - \overline{X}_2}} \qquad (T3)$$

where

$$S_{\overline{X}_1 - \overline{X}_2} = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (T4)$$

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^{\,2} + (n_2 - 1)S_2^{\,2}}{n_1 + n_2 - 2}} \qquad (T5)$$

Where:

$S_{\overline{X}_1 - \overline{X}_2}$ = The standard deviation of difference between two sample means.

$S_{pool}$ = The pooled variance of the samples.

$S_1^2$ = The variance the first sample

$S_2^2$ = The variance of second sample

$n_1$ = Number of observations in the first sample

$n_2$ = Number of observations in the second sample

The degrees of freedom for the t statistic in this special are simply given by

Df = $n_1 + n_2 - 2$

**b**. If both the variance of the population are not equal ( $\sigma_1^2 \neq \sigma_2^2$ ), then

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{\overline{X}_1 - \overline{X}_2}} \qquad (T6)$$

where

$$S_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{S_1^{\ 2}}{n_1} + \frac{S_2^{\ 2}}{n_2}} \qquad (T7)$$

Where :

$S_{\overline{X}_1 - \overline{X}_2}$ = The standard deviation of difference between two sample means.

$S_1^2$ = The variance the first sample

$S_2^2$ = The variance of second sample

$n_1$ = Number of observations in the first sample

$n_2$ = Number of observations in the second sample

The effective degrees of freedom for the t statistic given by expression

$$Db_{ef} = \frac{[(S_1^2/n_1) + (S_2^2/n_2)]^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_1 - 1}} \qquad (T8)$$

**Testing criteria:**

If this test use significance level $\alpha$ and the statistic t obtained from (T2) or (T4), then

**a**. Testing criteria for the *two-tailed test* hypothesis (T1) as follows:

$H_0$ is accepted if $|$ t-actual $| \leq t_{\alpha/2}$-table , or

$Pr = [ P ( t \leq -|t_{actual}|) + P ( t \geq |t_{actual}|) ] \geq \alpha$ , otherwise

$H_0$ is rejected if $|$ t-actual $| > t_{\alpha/2}$-table , or

$Pr = [ P ( t \leq -|t_{actual}|) + P ( t \geq |t_{actual}|) ] < \alpha$.

To test whether the means of the first population is smaller than the mean of second population, *one-tailed test* was used, with the following formulation of the test hypothesis

**b**. Testing criteria for the *one-tailed test* hypothesis (T2) as follows:

$H_0$ is accepted if t -actual $\geq -t_{\alpha}$-table or $Pr ( Z \leq t_{actual}) \geq \alpha$ , otherwise

$H_0$ is rejected if t -actual $< -t_{\alpha}$-tabel or $Pr ( Z \leq t_{actual}) < \alpha$ .

**Worked Example 4.6 :**

Data on the test scores of direct and cooperative learning models successively recorded at columns 2 and 3 in Table 4.2. Test the hypothesis that the two models of learning (a) did not differ, (b) the cooperative model is better than the direct model.

**Worked Solution :**

Means: $\overline{X}_1 = \frac{\sum_{i=1}^{25} X_i}{25} = \frac{1600}{25} = 64$ $\overline{X}_2 = \frac{\sum_{i=1}^{25} X_i}{25} = \frac{1918}{25} = 76.72$

**Table 4.2**

| Numb. | Direct | Cooperative | | |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_1^2$ | $X_2^2$ |
| [1,] | 61 | 76 | 3721 | 5776 |
| [2,] | 65 | 78 | 4225 | 6084 |
| [3,] | 69 | 66 | 4761 | 4356 |
| [4,] | 72 | 70 | 5184 | 4900 |
| [5,] | 68 | 86 | 4624 | 7396 |
| [6,] | 64 | 77 | 4096 | 5929 |
| [7,] | 60 | 84 | 3600 | 7056 |
| [8,] | 67 | 78 | 4489 | 6084 |
| [9,] | 61 | 83 | 3721 | 6889 |
| [10,] | 61 | 74 | 3721 | 5476 |
| [11,] | 65 | 75 | 4225 | 5625 |
| [12,] | 68 | 85 | 4624 | 7225 |
| [13,] | 65 | 79 | 4225 | 6241 |
| [14,] | 64 | 69 | 4096 | 4761 |
| [15,] | 62 | 75 | 3844 | 5625 |
| [16,] | 60 | 78 | 3600 | 6084 |
| [17,] | 56 | 75 | 3136 | 5625 |
| [18,] | 68 | 71 | 4624 | 5041 |
| [19,] | 57 | 78 | 3249 | 6084 |
| [20,] | 68 | 82 | 4624 | 6724 |
| [21,] | 69 | 79 | 4761 | 6241 |
| [22,] | 61 | 73 | 3721 | 5329 |
| [23,] | 61 | 60 | 3721 | 3600 |
| [24,] | 61 | 84 | 3721 | 7056 |
| [25,] | 67 | 83 | 4489 | 6889 |
| | $\sum X_1 = 1{,}600$ | $\sum X_2 = 1{,}918$ | $\sum X_1^2 = 102{,}802$ | $\sum X_2^2 = 148{,}096$ |

Variances: $$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\,\overline{X}^2}{n-1}$$

$$S_1^2 = \frac{(102802) - 25\,(64)^2}{25-1} = 16.75 \qquad S_2^2 = \frac{(148096) - 25\,(76.72)^2}{25-1} = 39.46$$

a. The test requires a two-tailed test:

Hypotheses: $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Significance level: $\alpha = 0.05$

If there is an indication that the variance is not the same, then standard deviation is found by computing:

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^{\,2}}{n_1} + \frac{S_2^{\,2}}{n_2}} = \sqrt{\frac{16.75}{25} + \frac{39.46}{25}} = 1.4994$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}} = \frac{64 - 76.72}{1.4994} = -8.48$$

Effective degrees of freedom: $Db_{ef} = \dfrac{[(S_1^2/n_1) + (S_2^2/n_2)]^2}{\dfrac{(S_1^2/n_1)^2}{n_1 - 1} + \dfrac{(S_2^2/n_2)^2}{n_1 - 1}}$

$$= \frac{((16.75/25) + (39.46/25))^2}{((16.75/25)^2/(25-1)) + ((39.46/25)^2/(25-1))}$$

$Db_{ef} = 41.264$

In Table t-student: $t_{\alpha/2;Df} = t_{0.025;41}$

$= 2.021$

Therefore $|t| > t_{\alpha/2;41}$, which is 8.48 > 2.021 then $H_0$ is rejected at the 5 percent significance level. It was concluded that there is a significant difference between the mean test scores of both models.

b) This the test requires one-tailed test:

Hypotheses: $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 < \mu_2$

Significance level: $\alpha = 0.05$

In Table t-student: $t_{\alpha/2;Df} = t_{0.05;41}$

$= -1.684$

Therefore t-actual < - $t_{0.05;41}$, which is - 8.48 < -1.684 then $H_0$ is rejected at the 5 percent significance level. It was concluded that the mean test scores of the direct learning lower than the mean test scores of the cooperative learning.

**Solving using computerized:**

The data in columns 2 and 3 of Table 2 stored in the S-plus, called TPS. Its data in the form of a matrix and has the number of row and column 25 and 2 respectively.

```
> var.test(TPS[,1],TPS[,2],alternative="two.sided", conf.level=.95)

        F test for variance equality

data:  TPS[, 1] and TPS[, 2]
F = 0.4245, num df = 24, denom df = 24, p-value = 0.0407
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1870554 0.9632639

sample estimates:
 variance of x  variance of y
     16.75          39.46
```

```
> t.test(TPS, TPS [, 2], alternative="two.sided", mu=0, paired=F,
 + var.equal=F, conf.level=.95)

Welch Modified Two-Sample t-Test

data:  x: direct in TPS , and y: cooperative  in TPS
t = -8.483,   df = 41.264,   p-value = 0
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.747645    -9.692355
sample estimates:
 mean of x     mean of y
      64           76.72


> t.test(TPS [, 1], TPS [, 2], alternative="less", mu=0, paired=F,
 + var.equal=F, conf.level=.95)

Welch Modified Two-Sample t-Test
```

```
data:  x: direct in TPS , and y: cooperative  in TPS
t = -8.483,   df = 41.264,   p-value = 0
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      NA       -10.19695
sample estimates:
 mean of x        mean of y
      64             76.72
```

**Comments:**

The results show three computerized test results, the test of homogeneity of variance (two-tailed F-test), the mean equality (two-tailed t-test), and the mean difference (one-tailed t-test).

1. The test of homogeneity of variance:

Used two-tailed F-test with significance level $\alpha = 0.05$ So $\alpha /2 = 0.025$. Print out shows that Pr = ( P( F< f-lower) + P( F > f-upper) ) = 0.0407 is smaller than $\alpha$. So the hypothesis $H_0$ is rejected at 5 percent of significance level. It was concluded that the variance between the two populations are not equal.

2. Test results of the mean equality:

Pr = [P ( $t \leq -|t_{actual}|$ ) + P ( $t \geq |t_{actual}|$ ) < $\alpha$, ie 0.00 < 0.05, then $H_0$ is rejected at the 5 percent significance level . It was concluded that there is a significant difference in the mean test scores by the two models.

3. Test results of mean difference:

Pr ( $t \leq t_{actual}$ ) < $\alpha$, ie 0.00 < 0.05, then $H_0$ is rejected at the 5 percent significance level. It was concluded that the mean test scores of the direct learning lower than the mean test scores of the cooperative learning.

**EXERCISES 4**

1. A survey of 210 households in the district of kutai kartanegara noted that the average monthly of household income is 1.2 million with a standard deviation of 0.12 million. The same survey conducted on 210 households in the district of kutai timur, noting that the average monthly of household income is 0.97 million with a standard deviation of 0.13 million. Test the hypothesis with a significance level of 5 percent that:
   a. Average of household income of kutai timur lower than kutai kartanegara.
   b. The Average household income are no different between the two districts.
2. A bulb plant produces two types of bulb, and claimed that the second type of bulb is better than the first type. From a drawn random sample of size 24 in both types of bulb, noting the average and standard deviation over the long lit bulb (hours), which bulb type I: $\overline{X}_1$ = 5,800 hours s = 31 hours, and type II: $\overline{X}_2$ = 6,000 hours, s = 30 hours. Test the hypothesis with a significance level of 1 percent that;
   a. The average of the lights up both type is equal.
   b. Second type longer than first type.
3. Data on the test scores of the expository and the cooperative learning models successively recorded in columns 2 and 3 of Table 4.3. Test the hypothesis that the two models of learning (a) did not differ, (b) the cooperative model is better than the direct model.

**Table 4.3**

| Numb. | Expository $X_1$ | Cooperative $X_2$ | $X_1^2$ | $X_2^2$ |
|---|---|---|---|---|
| [1,] | 59 | 68 | 3481 | 4624 |
| [2,] | 65 | 71 | 4225 | 5041 |
| [3,] | 59 | 72 | 3481 | 5184 |
| [4,] | 60 | 72 | 3600 | 5184 |
| [5,] | 60 | 72 | 3600 | 5184 |
| [6,] | 65 | 69 | 4225 | 4761 |
| [7,] | 61 | 79 | 3721 | 6241 |
| [8,] | 60 | 74 | 3600 | 5476 |
| [9,] | 66 | 68 | 4356 | 4624 |
| [10,] | 62 | 70 | 3844 | 4900 |
| [11,] | 66 | 68 | 4356 | 4624 |
| [12,] | 65 | 67 | 4225 | 4489 |
| [13,] | 59 | 71 | 3481 | 5041 |
| [14,] | 66 | 66 | 4356 | 4356 |
| [15,] | 64 | 76 | 4096 | 5776 |
| [16,] | 57 | 73 | 3249 | 5329 |
| [17,] | 61 | 68 | 3721 | 4624 |
| [18,] | 61 | 72 | 3721 | 5184 |
| [19,] | 60 | 75 | 3600 | 5625 |
| [20,] | 68 | 64 | 4624 | 4096 |
| [21,] | 69 | 70 | 4761 | 4900 |
| [22,] | 58 | 71 | 3364 | 5041 |
| [23,] | 62 | 82 | 3844 | 6724 |
| [24,] | 63 | 75 | 3969 | 5625 |
| [25,] | 60 | 74 | 3600 | 5476 |
| [26,] | 63 | 71 | 3969 | 5041 |
| | $\sum X_1 = 1{,}619$ | $\sum X_2 = 1{,}858$ | $\sum X_1^2 = 101{,}069$ | $\sum X_2^2 = 133{,}170$ |

C H A P T E R 5

---

# LINEAR REGRESSION MODEL

**Basic competencies**

1. Making statistical inference about the relationship between two variables or influence of a variable to another variable through capabilities in simple linear regression analysis.

**Indicators:**

Given a sample data set, students can do:

1.1 Estimate a the linear regression equation

1.2 Perform the significance test for the allegation regression equation by F-test

1.3 Checking the accuracy of a model linier relationship through its the coefficient of determination $R^2$

1.4 Infer and perform testing of the coefficients of a simple linear regression model

2. Making statistical inference about the relationship between one and a group variables or influence of a group variables to one another variable through capabilities in multiple linear regression analysis.

**Indicators:**

Given a sample data set, students can do:

2.1 Estimate the linear regression equation

2.2 Perform the significance test for a the allegation multiple linear regression model by F-test

2.3 Checking the accuracy of a linier relationship model through its the coefficient of determination $R^2$

2.4 Infer and perform partial testing for each regression coefficients of a multiple linear regression model

2.5 Infer dan perform sequential testing on the order of independent variables of a multiple linear regression model

2.6 Test the contribution subset of regressors to the sum of squares regression by defining a general linear hypothesis

# 5. LINEAR REGRESSION MODEL

Linear regression is a statistical analysis that models the relationship several variables by linear equations explicit form of relationship. Explicit form of the linear equation is a linear equation that puts a single variable is on one side of an equation.

Explicit variable in the model is a random variable, and the most likely to have behavior that depends on other variables. Variables which is the main concern is expressed as a dependent variable (response), with the symbol Y. As an example for these variables, can be a death caused by a disease, the level of prices according to market conditions, and the learning achievement of a teaching method.

Other variables in a model of linear equations are variables that might provide information about the behavior of dependent variables Y. These variables are placed as a predictor or independent variables in the model of linear equations. These variables are variables that are known fixed (not random), hereinafter referred to as independent variables, with the symbol X.

In general, this linear regression modeling aims to present how the average value of dependent variable "E(Y)" changes according to the change of each independent variable. It is assumed that the variance of Y is unaffected by changes in each independent variable. Furthermore, the linear regression equation is expressed as the seat of the expectation value of Y at each X value which is fixed. This expectation values have identical distribution and variance are equal.

## 5.1 THE SIMPLE LINEAR REGRESSION MODEL

### 5.1.1 MODEL AND ESTIMATION COEFFICIENTS

Simple linear regression model involves only one independent variable X. This model states constantly change the average value of the response

variable Y according to the change (increase or decrease) in the value of the independent variables X. This relationship is expressed in the form of a linear equation

$$E(Y_j) = \beta_0 + \beta_1 X_j \qquad \textbf{(R1)}$$

Where $\beta_0$ is the intercept, or the value of $E(Y_j)$ when X = 0, and $\beta_1$ is the slope or rate of change in $E(Y_j)$ per unit change in X.

Observations of the response variable Y, written $Y_j$, is assumed as a random observation from populations of random variables with the mean of each population given by $E(Y_j)$. The deviation of an observations Yj from its population mean $E(Y_j)$ is expressed as a random error $\varepsilon_j$. Furthermore, linear regression models were used modeled as follows.

$$Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j \qquad \textbf{(R2)}$$

Where,

j = Observational unit j =1, 2, . . . n

$Y_j$ = The value of the j-th observation for the variable response

$X_j$ = The value of the j-th observation of the explanatory variables.

$\varepsilon_j$ = j-th error of the model

Random error $\varepsilon_j$ have zero mean and are assumed to have common variance $\sigma^2$, and each error into j mutually independent. Because the only random elements in the model $Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j$ is $\varepsilon_j$, then this assumption imply that $Y_j$ also have common variance $\sigma^2$, and each $Y_j$ into j mutually independent. For the purposes of estimating of significance, $\varepsilon_j$ assumed

normally distributed, which causes $Y_j$ also normally distributed. Overall this assumption is stated in brief as follows.

$\varepsilon_j \sim NID\ (0, \sigma^2)$ , and imply that $Y_j \sim NID\ (\beta_0 + \beta_1 X_j, \sigma^2)$

Estimate of the regression coefficients using *least squares estimation procedure*, and for the purposes of testing significance or interval estimation, it is assumed to be normally, identically and independent distrubuted and with mean 0 and variance $\sigma^2$, or $\varepsilon_j \sim NID\ (0, \sigma^2)$. Estimation procedure with the least squares method is the amount of effort to minimize the (smallest) sum of squared deviations between $Y_j$ and its the estimated value, ie the sum of squares for $e_j = Y_j - \hat{Y}_j$ are minimum. This deviation is called residual, and will be obtained after gain estimation results coefficient $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$SSR = \sum (Y_j - \hat{Y}_j)^2$$

$$SSR = \sum (Y_j - (\hat{\beta}_0 + \hat{\beta}_1 X_j))^2$$

By using calculus, derivatives of the sum of squared residuals (SSR) according to each $\hat{\beta}_0$ and $\hat{\beta}_1$ equated to 0. Furthermore, the system of equations obtained is called the normal equation, ie

$$(n)\ \hat{\beta}_0 + (\sum X_j)\ \hat{\beta}_1 = \sum Y_j$$

$$(\sum X_j)\hat{\beta}_0 + (\sum X_j^2)\ \hat{\beta}_1 = \sum X_j Y_j \qquad \textbf{( R3 )}$$

The solution of the equation system will obtain the estimated value of each $\beta_0$ and $\beta_1$, ie

$$\hat{\beta}_1 = \sum( X_j - \overline{X})( Y_j - \overline{Y}) / \sum (X_j - \overline{X})^2$$

$$= [ \sum X_i Y_i - ( \sum X_i)( \sum Y_i) / n ] / [ \sum X_i^2 - ( \sum X_i)^2 / n ] \qquad \textbf{( R4 )}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Further allegations regression equation, namely

$$\hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j \qquad \textbf{( R5 )}$$

**Note:** Often the experimenter may become confused about the interpretation of $\beta_0$ or its estimate from experimental data. As we indicated earlier in this section, the linear model is the simplest empirical device for explaining how the data were produced. Often the characteristics of the model that are computed by the data are very much dependent on range of X in which the data were taken. In other word, there is a presumption that the model given by Eq.**(R2)** holds only in a confined region of X. If this region covers X=0, then the estimate of $\beta_0$ can certainly be interpreted as the mean Y at X=0. But if the data coverage is far away from the origin, then $\beta_0$ is merely a regression term that supplies little in the way of interpretation. Often, in fact, the estimate of $\beta_0$ turns out to be a value that seems quite unreasonable, or even impossible in the context of problem. The analyst must bear in mind that interpretation of $\beta_0$ is tantamount to extrapolating the model outside the range in which it was intended to be used.

### 5.1.2 ANALYSIS OF VARIANCE IN THE LINEAR REGRESSION

Analysis of variance on linear regression present the review of each partition of the sum of the squares on the term of equation which states: deviation of the estimated value of the actual observations, ie $e_j = Y_j - \hat{Y}_j$ or written as $Y_j = \hat{Y}_j + e_j$.

$$\sum Y_j^2 = \sum(\hat{Y}_j + e_j)^2$$

$$= \sum \hat{Y}_j^2 + \sum e_j^2 \qquad \text{“( The cross-product term } \sum \hat{Y}_j e_j = 0 \text{ )”}$$

$$= \sum \hat{Y}_j^2 + \sum ( Y_j - \hat{Y}_j)^2$$

$$SS(Total)_{uncorr} = SS(Model) + SS(Res) \quad \textbf{( R6 )}$$

Total sum of squares is partitioned over the explained sum of squares (SS (Model)) and the unexplained sum of squares (SS (Res)). The sum of squares on both sides of this partition is the sum of the squares that have not been corrected. This partition can then be made into a sum of squares corrected partition. Corrections were made on both sides of the equation by a correction factor $n\overline{Y}^2$ .

$$\sum Y_j^2 - n\overline{Y}^2 = (\sum \hat{Y}_j^2 - n\overline{Y}^2) + \sum (Y_j - \hat{Y}_j)^2$$

$$SS(Total) = SS(Reg) + SS(Res) \quad \textbf{( R7 )}$$

If the estimated value $\hat{\beta}_1$ is used, the sum of squares regression on this partition can be written as follows.

$$SS(Total) = SS(Reg) + SS(Res)$$

$$\sum Y_j^2 - n\overline{Y}^2 = (\sum \hat{Y}_j^2 - n\overline{Y}^2) + \sum (Y_j - \hat{Y}_j)^2$$

$$\sum Y_j^2 - n\overline{Y}^2 = (\hat{\beta}_1^2 \sum (X_j - \overline{X})^2) + \sum (Y_j - \hat{Y}_j)^2 \quad \textbf{( R8 )}$$

Degrees of freedom associated with the sum of the squares is determined by the sample size and the number of parameters in the model (p). Each degree of freedom of the corrected sum of squares is always reduced by 1 as a result of correction factors. Degrees of freedom associated with SS(total) is n - 1. Degrees of freedom associated with SS(Reg) is the degrees of freedom of the SS(model) minus one. Degrees of freedom associated with SS(Model) is equal to the number of parameters in the regression model, that is p = 2. Thus the degrees of freedom associated with SS(Reg) is p-1 = 2-1. Degrees of freedom associated with SS(Res) is the n-p = n-2.

The mean of the sum of squares (MS) is the sum of squares divided by their respective degrees of freedom. The calculation of each sum of squares with its mean (MS) is formulated as follows

$$\begin{aligned} SS(Total) &= \sum Y_j^2 - n\,\overline{Y}^2 \\ &= \sum Y_j^2 - (\sum Y_j)^2 / n \\ MS(Total) &= SS(Total)/ n\text{-}1 \end{aligned}$$

$$\begin{aligned} SS(Reg) &= \hat{\beta}_1^2 \sum (X_j - \overline{X})^2 \\ &= \hat{\beta}_1^2 \; [\sum X_i^2 - (\sum X_i)^2 / n \,] \qquad \textbf{( R9 )} \\ MS(Reg) &= SS(Reg)/ p\text{-}1 \end{aligned}$$

$$\begin{aligned} SS(Res) &= SS(Total) - SS(Reg) \\ MS(Res) &= SS(Res)/n\text{-}p \end{aligned}$$

The results of variance analysis is presented through the analysis of variance table. a value of F is also listed in the table analysis of variance. This is for testing purposes by using the approach of distribution F.

**Table 5.1** The analysis of variance

| **Source of variation** | **Sum of squares (SS )** | **Degrees of Freedom ( Df )** | **Mean Squares ( MS )** | **F** |
|---|---|---|---|---|
| Regression | SS(Reg) | $p$-1 | MS(Reg) | $F = \frac{MS(\text{Re}\,g)}{MS(\text{Re}\,s)}$ |
| Residual | SS(Res) | $n$-$p$ | MS(Res) | |
| Total | SS(Total) | $n$-1 | | |

In partitioning the total sum of squares, SS (Reg) represents the sum of the squares that can be explained or controlled, and SS (Res) represents the sum of squares unexplained. MS(Reg) estimate $\sigma^2 + \beta_1^2 \sum (X_j - \overline{X})^2$, and MS(Res)

estimate $\sigma^2$. If the model is correct then both are unbiased estimate. In the event the hypothesis $\beta_1 = 0$ is true, both the average sum of squares, ie MS (Reg) and MS (Res) estimate $\sigma^2$. If $\beta_1$ away from 0 then MS (Reg) increased greater than MS (Res). An analysis of variance results is quite good if the number of squares that is explained much greater than the unexplained sum of squares. The ratio between MS (Reg) to MS (Res) that large indicates $\beta_1$ is not equal to zero. If the assumption that the residual normally distributed is valid, and the hypothesis $\beta_1 = 0$ is true, then the ratio between MS (Reg) to MS (Res) follows the distribution F.

By using the F distribution approach in testing the significance of a allegation regression equation $\hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j$, then RSS (Reg) states the portion of the unexplained component that has been corrected, and SS(Res) states the portion of unexplained component. This method can also be used to test hypotheses:

$H_0 : \beta_1 = 0.$ **( R10 )**

$H_1 : \beta_1 \neq 0$

Furthermore, the formula for calculating the F statistic is

$F = MS(Reg) / MS(Res)$ **( R11 )**

where

MS (Reg) = Mean sum of squares regression

MS (Res) = Mean sum of squares residual

Which can be compared to the critical value $\alpha$, of the F distribution with 1 degrees of freedom in numerator and n-2 degrees of freedom in denominator.

**Testing criteria:**

If this test is used at the significance level of $\alpha$, then

Accept $H_0$ if $F \leq F_{(\alpha;1;n-2)}$-table or $Pr = P(F > F_{\text{-actual}}) \geq \alpha$ .

Reject $H_0$ if $F > F_{(\alpha;1;n-2)}$-table or $Pr = P(F > F_{\text{-actual}}) < \alpha$. **( R12 )**

### 5.1.3 THE COEFFICIENT OF DETERMINATION, $R^2$

A measure of how well a linear regression model to explain the relationship between independent variables with dependent variables, it can be seen from the value of koefisien of determination ($R^2$). In partitioning the total sum of squares, an indication that the model more appropriate if SS(Reg) greater approaching the total sum of squares. The coefficient of determination $R^2$ is defined as the proportion of SS (Reg) to SS(Total). Thus the coefficient of determination has a range of values from zero to one. $R^2$ value close to 1 means that the variation of the variable Y increasingly explained by its the linear relationship with the independent variables.

$$R^2 = SS(Reg) / SS(Total)$$

$$= \hat{\beta}_1^{\,2} \left[ \sum X_i^2 - (\sum X_i)^2 / n \right] / \sum Y_j^2 - n\overline{Y}^{\,2} \qquad \textbf{( R13 )}$$

### 5.1.4 COEFFICIENT TESTING

Investigation directly if there are significant changes in the variable Y by changes in variable X, made by testing the coefficient $\beta_1$. The magnitude of the coefficient $\beta_1$ is defined as the rate of change in the average value of Y by one unit change in X. To test whether $\beta_1$ is equal to zero or not, the test hypothesis formulated

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0 \qquad \textbf{( R14 )}$$

Testing this hypothesis using t-student distribution approach. Statistic t is calculated by

$$t = (\hat{\beta}_1 - 0) / S_{\hat{\beta}1} \qquad \textbf{( R15 )}$$

where

$$S_{\hat{\beta}1} = \sqrt{MS(Res) / \sum (Xj - \overline{X})^2}$$

$$= \sqrt{MS(Res) / (\sum X_j^2 - (\sum X_j)^2 / n)}$$

Where:

t = Statistic value t

$\hat{\beta}_1$ = Estimated value for the coefficient of the explanatory variables

$S_{\hat{\beta}1}$ = Standard deviation of the coefficient $\hat{\beta}_1$

T- statistic follows the t-student distribution and having n-2 degrees of freedom

**Criteria testing:**

The test requires a two-tailed test and if this test use significance level $\alpha$, then testing criteria as follows:

$H_0$ is accepted if $|$ t-actual $| \leq t_{\alpha/2}$-table , or

$Pr = [ P ( t \leq -|t_{actual}|) + P ( t \geq |t_{actual}|) ] \geq \alpha$ , otherwise

**( R16 )**

$H_0$ is rejected if $|$ t-actual $| > t_{\alpha/2}$-table , or

$Pr = [ P ( t \leq -|t_{actual}|) + P ( t \geq |t_{actual}|) ] < \alpha$.

**Worked Example 5.1 :**

A study was conducted to determine the relationship between the Increased consumption of gasoline (Y) with the level of car sales (X) in the Samarinda City. The data given in Table 5.2 were collected over twelve months.

**Table 5.2**

| | Increase Y | Car sales X | $Y^2$ | $X^2$ | X Y |
|---|---|---|---|---|---|
| 1 | 5.33 | 28.3 | 28.4089 | 800.89 | 150.839 |
| 2 | 6.00 | 29.0 | 36.0000 | 841.00 | 174.000 |
| 3 | 5.72 | 28.7 | 32.7184 | 823.69 | 164.164 |
| 4 | 5.55 | 28.2 | 30.8025 | 795.24 | 156.510 |
| 5 | 5.51 | 28.0 | 30.3601 | 784.00 | 154.280 |
| 6 | 5.12 | 27.8 | 26.2144 | 772.84 | 142.336 |
| 7 | 5.33 | 28.5 | 28.4089 | 812.25 | 151.905 |
| 8 | 5.80 | 28.9 | 33.6400 | 835.21 | 167.620 |
| 9 | 5.75 | 28.9 | 33.0625 | 835.21 | 166.175 |
| 10 | 5.91 | 29.2 | 34.9281 | 852.64 | 172.572 |
| 11 | 6.21 | 32.0 | 38.5641 | 1024.00 | 198.720 |
| 12 | 6.14 | 31.3 | 37.6996 | 979.69 | 192.182 |
| Σ Y= 68.37 | | Σ X= 348.8 | $\Sigma Y^2$ = 390.8075 | Σ $X^2$ = 10156.66 | ΣYX =1991.303 |

If the relationship between the two variables above want to be investigated by linear regression models,

a. Estimate the linear regression equation

b. Perform the significance test for the allegation regression equation through F test

c. Check the accuracy of this model linier relationship Y with X by the coefficient of determination $R^2$

d. Perform testing of the coefficient $\beta_1 = 0$

**Worked Solution:**

Means: $\overline{X} = \frac{\sum_{j=1}^{12} Xj}{12} = \frac{348.8}{12} = 29.07$ $\overline{Y} = \frac{\sum_{j=1}^{12} Y_j}{12} = \frac{68.37}{12} = 5.6975$

**a**. The calculation of the value of the coefficient $\hat{\beta}_1$ and $\hat{\beta}_0$ :

$$\hat{\beta}_1 = \sum( X_j - \overline{X})( Y_j - \overline{Y}) / \sum (X_j - \overline{X})^2$$

$$\hat{\beta}_1 = [ \sum X_j Y_j - ( \sum X_j)( \sum Y_j) / n ] / [ \sum X_j^2 - ( \sum X_j)^2 / n ]$$

$$= [1991.303 - (348.8)( 68.37) / 12 ] / [10156.66 - (348.8)^2 / 12 ]$$

$$= 0.2205$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$= 5.6975 - 0.2205 ( 29.07 )$$

$$= -0.7124$$

The allegation regression equation $\hat{Y}_j = -0.7128 + 0.2205 X_j$

**b**. Hypotheses:

$H_0 : \beta_1 = 0.$

$H_1 : \beta_1 \neq 0$

The calculation of the sum of squares:

SS(Total) $= \sum Y_j^2 - ( \sum Y_j)^2 / n$

$= ( 390.8075 )^2 - ( 68.37 )^2 / 12$

$= 1.2694$

SS(Reg) $= \hat{\beta}_1^2 [ \sum X_i^2 - ( \sum X_i)^2 / n ]$

$= (0.2205)^2 ( 10156.66 - (348.8)^2 / 12 )$

$= 0.8854$

RSS(Reg) $= 0.8854 / 1$

$= 0.8854$

SS(Res) $=$ SS(Total) – SS(Reg)

$= 1.2694 - 0.8854$

= 0.3840

RSS(Res) = 0.3840 /12 – 2

= 0.0384

$F_{hit}$ = RSS(Reg) / RSS(Res)

= 0.8854 / 0.0384 = 23.0559

In the F distribution table obtained: $F_{(0.05;1;10)}$-tabel = 4.96

**Table 5.3** The analysis of variance

| Source of variation | Sum of squares (SS ) | Degrees of Freedom | Mean Squares ( MS ) | F |
|---|---|---|---|---|
| Regression | 0.8854 | 1 | 0.8854 | |
| Residual | 0.3840 | 10 | 0.0384 | F =23.0559 |
| Total | 1.2694 | 11 | | |

Test results: $F > F_{(\alpha;1;n-2)}$, i.e. 23.056 > 4.96, then $H_0$ is rejected at the 5 percent significance level. This means that not all of the value of the parameter coefficients equal to zero. It was concluded that

" the regression equation $\hat{Y}_j$ = -0.7128 + 0.2205 $X_j$ is a significant relationship."

**c**. The coefficient of determination:

$R^2$ = SS(Reg) / SS(Total)

= 0.8854 / 1.2694

= 0.6975

The interpretation of  is that 69.75 % of the variation in the dependent variable is " explained" by its linear relationship with the independent variable.

**d**. Testing of the coefficient $\beta_1$

T-statistic testing on two-taild hypothesis is use to test partially coefficient. Because of its simple linear regression there is only one coefficient which tested then tested under these conditions, the t-statistic test the same as F-statistical test has been done above.

Hypotheses:

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Variance: $\hat{S}_{\beta 1} = \sqrt{MS(Res) / (\sum X_j^2 - (\sum X_j)^2 / n)}$

$= \sqrt{0.0384 / [10156.66 - (348.8) 2 / 12]}$

$= 0.0459$

t-value $= (\hat{\beta}_1 - 0) / \hat{S}_{\beta 1}$

$= 0.2250 / 0.0459 = 4.80$

In the t distribution table obtained: $t_{0.05/2;10} = 2.228$

Test results: $| t | > t_{\alpha/2;10}$., i.e. 4.80 > 2.228, then $H_0$ is rejected at the 5 percent significance level. This means that $\beta_1$ is significantly different to zero. It was concluded that the  rate of car sales affect the increased consumption of . gasoline.

**Computerized solution:**

**Table 5.4** Splus print out

```
*** Linear Model ***

Call: lm(formula = gasoline ~ car, data = riset1, na.action
       = na.exclude)

Residuals:
     Min      1Q  Median      3Q     Max
 -0.2982 -0.1504 0.04567 0.1123 0.3172

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) -0.7124  1.3361    -0.5332  0.6056
      Car    0.2205  0.0459     4.8017  0.0007


Residual standard error: 0.196 on 10 degrees of freedom

Multiple R-Squared: 0.6975

F-statistic: 23.06 on 1 and 10 degrees of freedom, the p-value is
0.0007218

Correlation of Coefficients:
      (Intercept)
Car   -0.9991

Analysis of Variance Table

Response: gasoline

Terms added sequentially (first to last)
          Df Sum of Sq   Mean Sq  F Value        Pr(F)
    Car    1 0.8854023 0.8854023 23.05599 0.0007218227
Residuals 10 0.3840227 0.0384023
```

**Comment:**

Solving of the above computerized presenting:

a. *In the coefficients part,* the estimations of both the value of the coefficients:

$\hat{\beta}_0 = -0.7124$ and $\hat{\beta}_1 = 0.2205$

So that the regression equation is $\hat{Y}_j = -0.7128 + 0.2205\, X_j$

b. *On the F-statistic*, F-statistics = 23.06 dan distributed F with the degrees of freedom of numerator and denominator respectively 1 and 10, giving the probability Pr (F> F-count) = 0.0007218. For significant level test $\alpha = 0.05$, this Pr < $\alpha$. The test results reject $H_0$. Even $H_0$ is rejected until a significant level of 0.0007219.

c. *In the Multiple R-Squared*, $R^2 = 0.6975$. This means that 69.75% of the variation in fuel consumption can be explained by its the linear relationship with the level of car sales

d. *In the coefficients part*, presents each the standard deviasion of value coefficients, value statistic t, and value of the probability Pr = [ P ( $t \leq -|t_{value}|$ ) + P ( $t \geq |t_{value}|$ ). For the coefficients testing $\hat{\beta}_1$, given Pr = 0.0007. So that $H_0$ is rejected at 0.01 significant level, even $H_0$ is rejected until a significant level of 0.00071.

**Worked Example 5.2 :**

Observations on the results of a chemical reaction to temperature variations are recorded as the following:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Temperature (Co) : | 125 | 125 | 125 | 150 | 150 | 150 | 175 | 175 | 175 | 200 | 200 | 200 |
| Reaction (Y%) : | 77 | 76 | 78 | 84 | 84 | 83 | 88 | 88 | 89 | 94 | 94 | 95 |

If the relationship between the two variables above, the chemical reaction as dependent variable (Y) and the temperature as independent variable (X), want to be investigated by linear regression model,

a. Estimate the linear regression equation

b. Perform the significance test for the allegation regression equation through F test

c. Check the accuracy of this model linier relationship Y with X by the coefficient of determination $R^2$

d. Perform testing of the coefficient $\beta_1 = 0$

**Worked Solution :**

$\Sigma X$= ……. $\Sigma X^2$ = …….. $\Sigma YX$ =……

$\Sigma Y$= ……. $\Sigma Y^2$ = ……..

Means : $\overline{X} = \frac{\sum_{j=1}^{12} Xj}{12} = \frac{.....}{12} = .....$ $\overline{Y} = \frac{\sum_{j=1}^{12} Y_j}{12} = \frac{.....}{12} = .....$

**a**. The calculation of the value of the coefficient $\hat{\beta}_1$ and $\hat{\beta}_0$ :

$\hat{\beta}_1 = \sum( X_j - \overline{X})( Y_j - \overline{Y}) / \sum (X_j - \overline{X})^2$

$\hat{\beta}_1 = [ \sum X_j Y_j - ( \sum X_j)( \sum Y_j) / n ] / [ \sum X_j^2 - ( \sum X_j)^2 / n ]$

= [……. – (……..)(……..) / …. ] / [……. - (……..) ² / ….. ]

= ……..

$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

= …… – ……. (…….)

= ……

The allegation regression equation $\hat{Y}_j$ = …….. + ……. $X_j$

**b**. Hypotheses:

$H_0$ : $\beta_1 = 0$.

$H_1$ : $\beta_1 \neq 0$

The calculation of the sum of squares:

SS(Total) $= \sum Y_j^2 - (\sum Y_j)^2 / n$

$= (\ldots\ldots)^2 - (\ldots\ldots)^2 / \ldots..$

$= \ldots\ldots.$

SS(Reg) $= \hat{\beta}_1^2 \ [\sum X_i^2 - (\sum X_i)^2 / n \ ]$

$= (\ldots\ldots\ldots)^2 \ (\ldots\ldots\ldots - (\ldots\ldots..)^2 / \ldots.. \ )$

$= \ldots\ldots\ldots$

MS(Reg) $= \ldots\ldots.. / 1$

$= \ldots\ldots..$

SS(Res) $=$ SS(Total) – SS(Reg)

$= \ldots\ldots\ldots - \ldots\ldots\ldots$

$= \ldots\ldots\ldots.$

MS(Res) $= \ldots\ldots.. / (\ldots\ldots. - 2)$

$= \ldots\ldots\ldots$

$F_{value}$ $=$ MS(Reg) / MS(Res)

$= \ldots\ldots.. / \ldots\ldots\ldots = \ldots\ldots\ldots$

In the F distribution table obtained: $F_{(0.05;1;10)}$ = .....

**Table 5.5** The analysis of variance

| **Source of variation** | **Sum of squares (SS )** | **Degrees of Freedom** | **Mean Squares ( MS )** | **F** |
|---|---|---|---|---|
| Regression | ..... | .... | ..... | F = ..... |
| Residual | ..... | .... | ..... | |
| Total | ..... | .... | | |

Test results: $F > F_{(\alpha;1;n-2)}$, ie ..... > ....., then $H_0$ is ..... at the 5 percent significance level.. It was concluded that.....

**c**. The coefficient of determination:

$R^2$ = SS(Reg) / SS(Total)

= …… / ……

= ……

The interpretation of is that ..... % of the variation in the dependent variable is " explained" by its linear relationship with the independent variable.

d. Testing of the coefficient $\beta_1$

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Variance: $S_{\hat{\beta}_1} = \sqrt{MS(Res) / (\sum X_j^2 - (\sum X_j)^2 / n)}$

$= \sqrt{\ldots\ldots\ldots\ldots/[\ldots\ldots\ldots\ldots-(\ldots\ldots\ldots)^2 \quad /\ldots\ldots]}$

$= \ldots\ldots\ldots$

t-value $= (\hat{\beta}_1 - 0) / S_{\hat{\beta}_1}$

$= \ldots\ldots\ldots / \ldots\ldots\ldots = \ldots\ldots\ldots$

In the t distribution table obtained: $t_{0.05/2;10} = \ldots\ldots\ldots\ldots$

Test results:....., then $H_0$ is .....at the 5 percent significance level. This means that $\beta_1$ is..... It was concluded that .....

## 5.2 THE MULTIPLE LINEAR REGRESSION MODEL

### 5.2.1 MODEL AND ESTIMATION COEFFICIENTS

Regresi ganda memodelkan hubungan antara suatu variabel terikat ( Y ) dengan beberapa variabel bebas ( Xi ). Model aditip linier bagi regresi ganda adalah:

Multiple regression models is a model of the relationship between a dependent variable (Y) with several independent variables (Xi). Linear additive model for multiple regression is:

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_p X_{pj} + \varepsilon_j \qquad \textbf{(G.1)}$$

In the matrix and vector algebra expression,

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \textbf{(G.2)}$$

or

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ . \\ . \\ . \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & X_{23} & \ldots & X_{2p} \\ 1 & X_{31} & X_{32} & X_{33} & \ldots & X_{3p} \\ . & . & . & . & & . \\ . & . & . & . & & . \\ . & . & . & . & & . \\ 1 & X_{n1} & X_{n2} & X_{n3} & \ldots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

( n x 1 ) ( n x (p+1) ) ( n x 1) ( n x 1 )

In association with regression modeling, the dependent variable Y is often called the response variable, and the independent variable Xi is called explanatory variables or regressors. Parameters $\boldsymbol{\beta}$ called regression coefficients, while the difference between the expected value of $\boldsymbol{Y}$ in the model ( $E(\boldsymbol{Y}) = \mathbf{X}\boldsymbol{\beta}$ ) with the actual value of $\boldsymbol{Y}$, which is $\boldsymbol{\varepsilon}$ called the error.

In the matrix and vector notation Eq.(**G.2**):

**The X matrix**; Each column **X** contains the value for a particular indevendent variable. The elements of a particular row of **X**, say row r, are the coefficients on the corresponding parameters in $\boldsymbol{\beta}$ which give. Notice that $\beta_0$ has the constant coefficien 1 for all observations; hence, the column vector **1** is the first column of **X**. The vectors $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$ are random vectors; the elements of these vectors are random variables. The matrix **X** is considered to be a matrix of known constants.

### 5.2.2 ASSUMPTIONS IN MULTIPLE REGRESSION

For estimation purposes, the above model it is assumed that the random vector **ε** have a multivariate normal distribution with mean vector **0** and variance-covariance matrix $\mathbf{I}\sigma^2$ . written brief $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$.

Mean vector **0** is a vector of size (n x 1), with all its elements 0. Variance-covariance matrix $\mathbf{I}\sigma^2$ is a matrix of size (n x n), the diagonal element is the variance $\sigma^2_{jj}$ (variance) of each random variable $\varepsilon_j$. While the nondiagonal elements (k, l) is covariance between $\varepsilon_k$ and $\varepsilon_l$.

Review the model $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, therefore **X** and $\boldsymbol{\beta}$ constant, then the rate of $\mathbf{X}\boldsymbol{\beta}$ in the model is a constant. By adding a vector of random error **ε**, causing $\boldsymbol{Y}$ is a random vector, with mean vector $\mathbf{X}\boldsymbol{\beta}$, and variance-covariance matrix $\mathbf{I}\sigma^2$, is written " $\boldsymbol{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2)$ ".

**2. Estimate of $\boldsymbol{\beta}$ or Model**

Estimation for the model $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}$, conducted through the estimate of the parameter $\boldsymbol{\beta}$. Estimator for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}$ . By using the method of least squares sum, the $\sum(Y - \mathbf{X}\hat{\beta})^2$ minimum, then we obtain the following normal equation

$$\mathbf{X'X}\hat{\boldsymbol{\beta}} = \mathbf{X'}\boldsymbol{Y} \qquad \text{(G.3)}$$

were,

$$\mathbf{X'X} = \begin{pmatrix} n & \sum X_{i1} & \sum X_{i2} & \cdots & \sum X_{ip} \\ \sum X_{i1} & \sum X_{i1}^2 & \sum X_{i1}X_{i2} & \cdots & \sum X_{i1}X_{ip} \\ \sum X_{i2} & \sum X_{i1}X_{i2} & \sum X_{i2}^2 & \cdots & \sum X_{i2}X_{ip} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ \sum X_{ip} & \sum X_{i1}X_{ip} & \sum X_{i2}X_{ip} & \cdots & \sum X_{ip}^2 \end{pmatrix}$$

$$\mathbf{X'Y} = \begin{pmatrix} \sum Y_i \\ \sum X_{i1}Y_i \\ \sum X_{i2}Y_i \\ . \\ . \\ . \\ \sum X_{ip}Y_i \end{pmatrix}$$

Unique solution to the normal equations (if any) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}(\mathbf{X'Y}) \qquad \text{(G.4)}$$

Unique solution of the normal equations exist, if the inverse of $\mathbf{X'X}$ exists (*nonsingular*). Nonsingular matrix $\mathbf{X}$ achieved if its full rank, scilicet there is no linear dependence between the independent variables (the columns are linearly independent vectors).

Having obtained the estimates of the regression coefficients or estimated values for vector $\boldsymbol{\beta}$, i.e $\hat{\boldsymbol{\beta}}$ the estimation of the regression equation has been obtained, i.e $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

**Worked Example 5.3 :**

Here in Table 5.6 is data about the results of environmental studies. The study measured the 4 variables, namely the concentration of ozone (Y), solar radiation ($X_1$), temperature ($X_2$), and the wind speed ($X_3$). If the relationship between the dependent variable ozone and independent variables want to be investigated by the multiple linear regression models, estimate the linear regression equation.

**Table 5.6** Data about the results of environmental studies

| N0 | Ozone | Radiation | Temperature | Wind |
|---|---|---|---|---|
| 1 | 3.45 | 190 | 67 | 7.4 |
| 2 | 3.30 | 118 | 72 | 8.0 |
| 3 | 2.29 | 149 | 74 | 12.6 |
| 4 | 2.62 | 313 | 62 | 11.5 |
| 5 | 2.84 | 299 | 65 | 8.6 |
| 6 | 2.67 | 99 | 59 | 13.8 |
| 7 | 2.00 | 19 | 61 | 20.1 |
| 8 | 2.52 | 256 | 69 | 9.7 |
| 9 | 2.22 | 290 | 66 | 9.2 |
| 10 | 2.41 | 274 | 68 | 10.9 |
| 11 | 2.62 | 65 | 58 | 13.2 |
| 12 | 2.41 | 334 | 64 | 11.5 |
| 13 | 3.24 | 307 | 66 | 12.0 |
| 14 | 1.82 | 78 | 57 | 18.4 |
| 15 | 3.11 | 322 | 68 | 11.5 |
| 16 | 2.22 | 44 | 62 | 9.7 |
| 17 | 1.00 | 8 | 59 | 9.7 |
| 18 | 2.22 | 320 | 73 | 16.6 |
| 19 | 1.59 | 25 | 61 | 9.7 |
| 20 | 3.17 | 92 | 61 | 12.0 |
| 21 | 2.84 | 13 | 67 | 12.0 |
| 22 | 3.56 | 252 | 81 | 14.9 |
| 23 | 4.86 | 223 | 79 | 5.7 |
| 24 | 3.33 | 279 | 76 | 7.4 |
| 25 | 3.07 | 127 | 82 | 9.7 |
| 26 | 4.14 | 291 | 90 | 13.8 |
| 27 | 3.39 | 323 | 87 | 11.5 |
| 28 | 2.84 | 148 | 82 | 8.0 |
| 29 | 2.76 | 191 | 77 | 14.9 |
| 30 | 3.33 | 284 | 72 | 20.7 |

**Worked Solution:**

Details of the calculation using the computerized results are presented as follows.

The matrix $\boldsymbol{X'X}$ is

```
       [,1]      [,2]   [,3]     [,4]
[1,]   30.0    5733.0   2085   354.70
[2,] 5733.0 1463179.0 411076 67085.70
[3,] 2085.0  411076.0 147243 24523.00
[4,]  354.7   67085.7  24523  4583.39
```

and the matrix $\boldsymbol{X'Y}$ is

```
          [,1]
[1,]    83.840
[2,] 17097.380
[3,]  5953.470
[4,]   973.641
```

The inverse of the $\mathbf{X'X}$ matrix, i.e the matrix $(\mathbf{X'X})^{-1}$ is

```
          [,1]         [,2]            [,3]           [,4]
[1,]  2.8559225951  6.173356e-004 -0.03535068677 -4.090973e-002
[2,]  0.0006173356  3.341412e-006 -0.00001807261  1.409908e-008
[3,] -0.0353506868 -1.807261e-005  0.00053385411  1.439104e-004
[4,] -0.0409097295  1.409908e-008  0.00014391044  2.613921e-003
```

Estimate of the regression coefficients using Eq.(**G.4**),

$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}(\mathbf{X'Y})$

So the regression coefficient vector $\hat{\boldsymbol{\beta}}$ is

```
       [,1]
[1,] -0.295271027
[2,]  0.001305816
[3,]  0.045605744
[4,] -0.027843496
```

Furthermore, the regression equation was obtained following

$$\hat{Y} = -0{,}2953 + 0{,}0013\, X_1 + 0.0456\, X_2 - 0{,}0278\, X_3$$

Results of computerized using Splus, displays the following output

**Tabel 5.7** The Splus output

```
*** Linear Model ***


Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) -0.2949  0.9986    -0.2953  0.7701
  radiation  0.0013  0.0011     1.2080  0.2379
temperature  0.0456  0.0137     3.3429  0.0025
       wind -0.0280  0.0302    -0.9268  0.3626


Residual standard error: 0.5909 on 26 degrees of freedom
Multiple R-Squared: 0.4583
F-statistic: 7.332 on 3 and 26 degrees of freedom, the p-value is 0.00102
```

### 5.2.3 FORMS OF LINEAR FUNCTION OF Y

Here it can be shown that the statistics (estimator) in multiple regression is a linear function of Y. These statistics include $\hat{\boldsymbol{\beta}}$ (coefficient estimators), $\hat{\boldsymbol{Y}}$ (predicted value), and $\boldsymbol{e}$ (residual).

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{Y}) \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\boldsymbol{Y} \qquad \text{(G.5)}\end{aligned}$$

Vector $\hat{\boldsymbol{\beta}}$ is a linear function of $\boldsymbol{Y}$, with coefficients $[(\mathbf{X}'\mathbf{X})^{-1}\ \mathbf{X}']$ .

The vector of estimated means of dependent variable $\boldsymbol{Y}$ for values the independent variables in the data set is computed as $\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. This is the simplest way to compute $\hat{\boldsymbol{Y}}$ and useful to express as a linear function of $\boldsymbol{Y}$.

$$\begin{aligned}\hat{\boldsymbol{Y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{Y}) \\ &= [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\boldsymbol{Y} \\ &= \mathbf{H}\boldsymbol{Y} \qquad \text{(G.6)}\end{aligned}$$

Vector $\hat{\boldsymbol{Y}}$ is a linear function of $\boldsymbol{Y}$, with coefficients $\mathbf{H} = [\ \mathbf{X}\ (\ \mathbf{X'X}\ )^{-1}\ \mathbf{X'}\ ]$. The matrix **H** called "hat matrix" is a matrix which is determined by the matrix **X**. This matrix is a very important role in the regression analysis. The matrix **H** is symmetric and idempotent matrix, ie $\mathbf{H'} = \mathbf{H}$ dan $\mathbf{H}\,\mathbf{H} = \mathbf{H}$.

$$\begin{aligned} \boldsymbol{e} &= \boldsymbol{Y} - \hat{\boldsymbol{Y}}\ . \\ &= \boldsymbol{Y} - [\ \mathbf{X}\ (\ \mathbf{X'X}\ )^{-1}\ \mathbf{X'}]\ \boldsymbol{Y} \\ &= \boldsymbol{Y} - \mathbf{H}\,\boldsymbol{Y} \\ &= [\mathbf{I} - \mathbf{H}]\ \boldsymbol{Y} \end{aligned} \qquad \text{(G.7)}$$

The vector $\boldsymbol{e}$ is a linear function of $\boldsymbol{Y}$, with coefficients $[\mathbf{I} - \mathbf{H}]$. Teh matrix $[\mathbf{I} - \mathbf{H}]$ is also symmetric and idempotent matrix.

### 5.2.4 DISTRIBUTION OF STATISTICS $\hat{\boldsymbol{\beta}}$ , $\hat{\boldsymbol{Y}}$ , AND $\boldsymbol{e}$

If the model is correct, then the expected value of $\boldsymbol{Y}$ is $\mathbf{X}\boldsymbol{\beta}$. Since statistics, $\hat{\boldsymbol{\beta}}$ , $\hat{\boldsymbol{Y}}$ , dan **e** is a linear function of $\boldsymbol{Y}$, and $\boldsymbol{Y}$ is a random vector is known, then $\hat{\boldsymbol{\beta}}$ , $\hat{\boldsymbol{Y}}$ , dan **e** are also random vectors. The properties of each statistic as a linear function of $\boldsymbol{Y}$ can be expressed as follows.

#### a. The Expectation

$$\begin{aligned} \mathrm{E}\,(\hat{\boldsymbol{\beta}}\ ) &= \mathrm{E}(\ (\ \mathbf{X'X}\ )^{-1}\ \mathbf{X'Y}\ ) \\ &= [(\ \mathbf{X'X}\ )^{-1}\ \mathbf{X'}\ ]\ \mathrm{E}\ (\ \mathbf{Y}\ ) \\ &= [(\ \mathbf{X'X}\ )^{-1}\ \mathbf{X'}]\ (\mathbf{X}\boldsymbol{\beta}) \\ &= [(\ \mathbf{X'X}\ )^{-1}(\ \mathbf{X'X})]\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

This show that $\hat{\boldsymbol{\beta}}$ is unbiased estimate of $\boldsymbol{\beta}$, if the chosen model is correct. If the chosen model is not correct, then the $(\ \mathbf{X'X}\ )^{-1}(\ \mathbf{X'X}) \neq \mathbf{I}$ and $\mathrm{E}\,(\hat{\boldsymbol{\beta}}\ )$ does not simplify to as in equation above.

$$
\begin{aligned}
E(\hat{\boldsymbol{Y}}) &= E(\mathbf{H}\,\boldsymbol{Y}) \\
&= \mathbf{H}\,E(\boldsymbol{Y}) \\
&= [\mathbf{X}\,(\mathbf{X'X})^{-1}\,\mathbf{X'}]\,\mathbf{X}\boldsymbol{\beta} \\
&= \mathbf{X}\,(\mathbf{X'X})^{-1}\,\mathbf{X'X}\,\boldsymbol{\beta} \\
&= \mathbf{X}\,\boldsymbol{\beta}
\end{aligned}
$$

Thus, if the model is correct, $\hat{\boldsymbol{Y}}$ is unbiased estimate of means of $\boldsymbol{Y}$.

$$
\begin{aligned}
E(\boldsymbol{e}) &= E([\mathbf{I} - \mathbf{H}]\,\boldsymbol{Y}) \\
&= [\mathbf{I} - \mathbf{H}]\,E(\boldsymbol{Y}) \\
&= [\mathbf{I} - \mathbf{H}]\,\mathbf{X}\boldsymbol{\beta} \\
&= [\mathbf{I}\,\mathbf{X} - \mathbf{H}\,\mathbf{X}]\,\boldsymbol{\beta} \\
&= [\mathbf{X} - \mathbf{X}]\,\boldsymbol{\beta} \\
&= \mathbf{0}
\end{aligned}
$$

Thus, $\boldsymbol{e}$ observed residuals are random variables with mean zero.

**b. Variance**

Discussion of the variance for a linear function of Y, is done by first review the idea of notation and matrix algebra. For $\boldsymbol{Y}$ as random vectors, eg variance-covariance matrix has written **Var (Y)**, and suppose a linear function of **u**, written $\mathbf{u} = \mathbf{a'}\,\boldsymbol{Y}$ .

$$\mathbf{Var\,(u)} = \mathbf{a'}\,[\mathbf{Var}\,(\boldsymbol{Y})]\,\mathbf{a}$$

As assumptions on estimation by ordinary least squares method, $\mathbf{Var\,(Y)} = \mathbf{I}\sigma^2$, so that

$\mathbf{Var\,(u)} = \mathbf{a'}\,(\mathbf{I}\sigma^2)\,\mathbf{a} = \mathbf{a'a}\,\sigma^2$. Notation **a'a** stating the sum of squares of the coefficients of linear function, that is $\sum a_i^2$

Form of a linear function of $\mathbf{u} = \mathbf{a'}\,\boldsymbol{Y}$, extended over several vector coefficients **a** simultaneously, i.e by a *k x n* matrix of coefficients **A**, becomes

$\mathbf{U}=\mathbf{A}\boldsymbol{Y}$. Furthermore, the definition of variance-covariance matrix for the random vector $\boldsymbol{Y}$ :

$$\mathbf{Var}\ (\boldsymbol{Y}) = E(\ [\ \boldsymbol{Y} - E(\boldsymbol{Y})][\ \boldsymbol{Y} - E(\boldsymbol{Y})]'\ )$$

Matrix multiplication results $[\ \boldsymbol{Y} - E(\boldsymbol{Y})][\ \boldsymbol{Y} - E(\boldsymbol{Y})]'$ size n x n, with main diagonal elements $(\ Y_i - E(Y_i)\ )^2$ and the nondiagonal elements $(\ Y_i - E(Y_i))(Y_j - E(Y_j))'$. Expectation value for the two groups of consecutive elements are the variance covariance.

If the definition of the variance-covariance matrix is applied to $\mathbf{U} = \mathbf{A}\ \boldsymbol{Y}$, then

$$\begin{aligned}\mathbf{Var\ (U)} &= E(\ [\mathbf{U} - E(\mathbf{U})][\ \mathbf{U} - E(\mathbf{U})]'\ ) \\ &= E(\ [\mathbf{A}\ \boldsymbol{Y} - E(\mathbf{A}\ \boldsymbol{Y})][\ \mathbf{A}\ \boldsymbol{Y} - E(\mathbf{A}\ \boldsymbol{Y})]'\ ) \\ &= E(\mathbf{A}\ [\ \boldsymbol{Y} - E(\boldsymbol{Y})][\ \boldsymbol{Y} - E(\boldsymbol{Y})]'\ \mathbf{A'}\ ) \\ &= \mathbf{A}\ E([\ \boldsymbol{Y} - E(\boldsymbol{Y})][\ \boldsymbol{Y} - E(\boldsymbol{Y})]')\ \mathbf{A'} \\ &= \mathbf{A}\ [\mathbf{Var\ (}\boldsymbol{Y}\mathbf{)}\ ]\ \mathbf{A'} \qquad \text{(G.8)}\end{aligned}$$

For $\mathbf{Var\ (}\boldsymbol{Y}\mathbf{)} = \mathbf{I}\sigma^2$,

$$\begin{aligned}\mathbf{Var\ (U)} &= \mathbf{A}\ [\mathbf{I}\sigma^2]\ \mathbf{A'} \\ &= \mathbf{A\ A'}\ \sigma^2 \qquad \text{(G.9)}\end{aligned}$$

Description: Elements of a diagonal matrix $\mathbf{AA'}$ is the sum of the squares of the coefficients of i-th linear function, then the results perkaliannya with $\sigma^2$ is variance a i-th linear function. Nondiagonal elements (i, j) is the cross product between the coefficients of the linear function, then the result of multiplying it by $\sigma^2$ is covariance between the two linear functions.

Variance of each statistic, $\hat{\boldsymbol{\beta}}$ , $\hat{\boldsymbol{Y}}$ , and $\boldsymbol{e}$ :

$\hat{\boldsymbol{\beta}} = [(\ \mathbf{X'X}\ )^{-1}(\ \mathbf{X'})]\ \boldsymbol{Y}$ , so that

$$\begin{aligned}\mathbf{Var}(\hat{\boldsymbol{\beta}}) &= [(\ \mathbf{X'X}\ )^{-1}(\ \mathbf{X'})]\ [\mathbf{Var\ (}\boldsymbol{Y}\mathbf{)}\ ]\ [(\ \mathbf{X'X}\ )^{-1}(\ \mathbf{X'})]' \\ &= (\ \mathbf{X'X}\ )^{-1}(\ \mathbf{X'\ X}\ )(\ \mathbf{X'X}\ )^{-1}\ \sigma^2 \\ &= (\ \mathbf{X'X}\ )^{-1}\ \sigma^2 \qquad \text{(G.10)}\end{aligned}$$

$\hat{\boldsymbol{Y}} = \mathbf{H}\,\boldsymbol{Y}$, so that

$$\mathbf{Var}(\hat{\boldsymbol{Y}}) = \mathbf{H}\,[\mathbf{Var}\,(\boldsymbol{Y})]\,\mathbf{H}'$$
$$= \mathbf{H}\,\mathbf{I}\,\mathbf{H}'\,\sigma^2$$
$$= \mathbf{H}\,\mathbf{H}'\,\sigma^2$$
$$= \mathbf{H}\,\sigma^2 \qquad \text{(G.11)}$$

Diagonal element is the variance for the predicted value $\hat{\boldsymbol{Y}}_{\boldsymbol{i}}$, i = 1, 2, ... n. Value prediction is used to estimate the means Y for various combinations of independent variables is given. For the value prediction of the future (predictive value) for various combinations of values of the independent variables is given, written $\hat{\boldsymbol{Y}}_{\boldsymbol{i}\ pred}$, then each variance increased by $\sigma^2$. Variance-covariance matrix for this prediction is

$$\mathbf{Var}\,(\hat{\boldsymbol{Y}}_{pred}) = (\mathbf{I} + \mathbf{H})\,\sigma^2 \qquad \text{(G.12)}$$

$\boldsymbol{e} = [\mathbf{I} - \mathbf{H}]\,\boldsymbol{Y}$, so that

$$\mathbf{Var}\,(\boldsymbol{e}) = [\mathbf{I} - \mathbf{H}]\,[\mathbf{Var}\,(\boldsymbol{Y})]\,[\mathbf{I} - \mathbf{H}]'$$
$$= [\mathbf{I} - \mathbf{H}]\,\mathbf{I}\,[\mathbf{I} - \mathbf{H}]'\,\sigma^2$$
$$= [\mathbf{I} - \mathbf{H}]\,[\mathbf{I} - \mathbf{H}]'\,\sigma^2$$
$$= [\mathbf{I} - \mathbf{H}]\,\sigma^2 \qquad \text{(G.13)}$$

Summary of the distribution of each random vectors can be expressed as follows:

$\boldsymbol{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2)$

$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$

$\hat{\boldsymbol{Y}} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{H}\sigma^2)$

$\mathbf{e} \sim N(\mathbf{0}, [\mathbf{I} - \mathbf{H}]\,\sigma^2)$

$\hat{\boldsymbol{Y}}_{pred} \sim N(\mathbf{X}\boldsymbol{\beta}, [\mathbf{I} + \mathbf{H}]\,\sigma^2)$

### 5.2.5 PARTITIONING OF THE SUM OF SQUARES

To match the linear additive model, the vector of observation for the dependent variable $\boldsymbol{Y}$, partitioned into vector $\hat{\boldsymbol{Y}}$ plus residual vector **e**. namely

$$\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \boldsymbol{e}$$

Similar partitions are used to sum of the squares of the dependent variable $\boldsymbol{Y}$.

$$\boldsymbol{Y}'\boldsymbol{Y} = (\hat{\boldsymbol{Y}} + \boldsymbol{e})'(\hat{\boldsymbol{Y}} + \boldsymbol{e})$$
$$= \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} + \hat{\boldsymbol{Y}}'\boldsymbol{e} + \boldsymbol{e}'\hat{\boldsymbol{Y}} + \boldsymbol{e}'\boldsymbol{e}$$

Substituting $\hat{\boldsymbol{Y}} = \mathbf{H}\boldsymbol{Y}$, dan $\boldsymbol{e} = [\mathbf{I} - \mathbf{H}]\boldsymbol{Y}$ gives

$$\boldsymbol{Y}'\boldsymbol{Y} = (\mathbf{H}\boldsymbol{Y})'(\mathbf{H}\boldsymbol{Y}) + (\mathbf{H}\boldsymbol{Y})'([\mathbf{I} - \mathbf{H}]\boldsymbol{Y}) + ([\mathbf{I} - \mathbf{H}]\boldsymbol{Y})'(\mathbf{H}\boldsymbol{Y}) + ([\mathbf{I} - \mathbf{H}]\boldsymbol{Y})'([\mathbf{I} - \mathbf{H}]\boldsymbol{Y})$$
$$= \boldsymbol{Y}'\mathbf{H}'\mathbf{H}\boldsymbol{Y} + \boldsymbol{Y}'(\mathbf{H}'[\mathbf{I} - \mathbf{H}])\boldsymbol{Y} + \boldsymbol{Y}'([\mathbf{I} - \mathbf{H}]'\mathbf{H})\boldsymbol{Y} + \boldsymbol{Y}'([\mathbf{I} - \mathbf{H}]'[\mathbf{I} - \mathbf{H}])\boldsymbol{Y}$$

Both **H** and **[I - H]** is symmetric and idempotent, so that **H′ H** = **H** and **[I - H]′ [I - H]** = **[I - H]**. The two middle term are zero because the two quadratic forms are orthogonal to each other, ie **H′ [I - H]** = **[H - H]** = **0**. Furthermore

$$\boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{Y}'\mathbf{H}\boldsymbol{Y} + \boldsymbol{Y}'[\mathbf{I} - \mathbf{H}]\boldsymbol{Y} = \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} + \boldsymbol{e}'\boldsymbol{e} \qquad \text{(G.14)}$$

$$\text{SS(Total)} = \text{SS(Model)} + \text{SS (Res)} \qquad \text{(G.15)}$$

Total sum of squares was partitioned into two sums of squares, which is the model sum of squares and residual sum of squares. Both of the sums of squares consecutive states as explained component and unexplained components of the model. SS (Model) = $\hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} = \boldsymbol{Y}'\mathbf{H}\boldsymbol{Y}$ has a defining matrix **H**, and SS (Res) = $\boldsymbol{e}'\boldsymbol{e} = \boldsymbol{Y}'[\mathbf{I} - \mathbf{H}]\boldsymbol{Y}$ has a defining matrix **[I - H]**.

If the two defining matrices are multiplied, **H [I - H]** = **0**, so the two sums of squares are mutually orthogonal, then forming additive partition. Degrees of freedom for both the sum of the squares of each is determined by the rank of

the defining matrices. Because of its two defining matrices are idempotent matrices, will rank equally with its trace, i.e

$Rank\ (\mathbf{H}) = tr(\mathbf{H}) = tr(\mathbf{X}\ (\ \mathbf{X'X}\ )^{-1}\ \mathbf{X'})$

$= tr(\ \mathbf{X'X}\ )^{-1}\ \mathbf{X'}\ \mathbf{X}\ )$ (theory in matrix algebra $tr(\mathbf{ABC}) = tr(\mathbf{BCA})$ )

$\mathbf{=}\ tr\ (\ \mathbf{I}_{p+1}) = p + 1$ ( p + 1 = number of columns of the matrix **X** )

$Rank(\mathbf{I - H}) = tr(\mathbf{I - H}) = tr(\mathbf{I}) - tr(\mathbf{H})$

$= n-(p+1)$

Degrees of freedom for SS(Model) is p +1, and the degrees of freedom for SS(Res) is n-(p +1).

How to count the sum of squares through the matrix notation are

SS(Total) $= \boldsymbol{Y'\ Y}$

SS(Model) $= \hat{\boldsymbol{Y}}\ '\ \hat{\boldsymbol{Y}} = (\mathbf{X}\ \hat{\boldsymbol{\beta}}\ )'(\mathbf{X}\ (\ \mathbf{X'X}\ )^{-1}\ \mathbf{X'}\boldsymbol{Y}) = \hat{\boldsymbol{\beta}}\ '\ \mathbf{X'}\boldsymbol{Y}$

SS(Sisa) $= \boldsymbol{Y'\ Y} - \hat{\boldsymbol{\beta}}\ '\ \mathbf{X'}\boldsymbol{Y}.$ (G.16)

In regression analysis, we need the knowledge of the contribution of a group of independent variables to the variation of $\boldsymbol{Y}$ around the mean value. The size information can be seen from the difference between the SS(Model), which contains independent variables with SS(Model) without independent variables. SS(Model) without independent variables is called a correction factor, and written SS($\mu$). The difference between the SS(model) with SS($\mu$) is called the sum of squares regression, written SS (Reg).

SS(Reg) = SS(Model) – SS($\mu$)

To get SS($\mu$), written in the linear additive model $\boldsymbol{Y} = \mathbf{X}^*\ \boldsymbol{\beta}\ +\ \boldsymbol{\varepsilon}$, where the matrix $\mathbf{X}^*\ = \mathbf{1}$. Matrix **1** is only a column vector with all elements 1 or the first column of the matrix **X**.

$\hat{\boldsymbol{\beta}}\ = ((\ \mathbf{X}^*)'\ \mathbf{X}^*\ )^{-1}\ (\mathbf{X}^*)'\boldsymbol{Y}\ = (\mathbf{1'\ 1}\ )^{-1}\ \mathbf{1'}\boldsymbol{Y}\ = (1/n)\ \mathbf{1'}\boldsymbol{Y} = \overline{Y}$ , so that

$$SS(\mu) = \hat{\boldsymbol{\beta}}' (\mathbf{X}^*)' \boldsymbol{Y} = (1/n) (\mathbf{1}'\boldsymbol{Y})' (\mathbf{1}'\boldsymbol{Y})$$
$$= (1/n) \boldsymbol{Y}' (\mathbf{1}\,\mathbf{1}')\boldsymbol{Y} \qquad \text{(G.17)}$$

because of $\mathbf{1}'\boldsymbol{Y} = \sum \mathbf{Y}_i$, then $SS(\mu) = n\ \overline{Y}^2$.

If we let $\mathbf{1}\,\mathbf{1}' = \mathbf{J}$, with **J** is a matrix of size (nxn) with all elements 1, then

$$SS(Reg) = SS(Model) - SS(\mu)$$
$$= \boldsymbol{Y}' \mathbf{H}\, \boldsymbol{Y} - \boldsymbol{Y}' (\mathbf{J/n})\boldsymbol{Y}$$
$$= \boldsymbol{Y}' (\mathbf{H} - \mathbf{J/n})\boldsymbol{Y} \qquad \text{(G.18)}$$

Degrees of freedom for $SS(\mu)$ is 1, so the degrees of freedom for SS (Reg) is p.

Partition sum of the squares on multiple linear regression is shown by the following table.

**Table 5.8** Analysis of variance summary for regression analysis.

| Source of variation | Degrees of Fredom | Sum of Squares Formula | Computational Formula |
|---|---|---|---|
| $Total_{corrt}$ | n-1 | $\boldsymbol{Y}'(\mathbf{I} - \mathbf{J})\boldsymbol{Y}$ | $\boldsymbol{Y}'\boldsymbol{Y} - n\,\overline{Y}^2$ |
| Model | p +1 | $\boldsymbol{Y}'\mathbf{H}\,\boldsymbol{Y}$ | $\hat{\boldsymbol{\beta}}'\mathbf{X}'\boldsymbol{Y}$ |
| Mean | 1 | $(1/n)\ \boldsymbol{Y}'(\mathbf{1}\,\mathbf{1}')\,\boldsymbol{Y}$ | $n\,\overline{Y}^2$ |
| Regression | p | $\boldsymbol{Y}'\ [\mathbf{H} - \mathbf{J}/n]\,\boldsymbol{Y}$ | $\hat{\boldsymbol{\beta}}'\mathbf{X}'\boldsymbol{Y} - n\,\overline{Y}^2$ |
| Residual | n-(p+1) | $\boldsymbol{Y}'\ [\mathbf{I} - \mathbf{H}]\,\boldsymbol{Y}$ | $\boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\boldsymbol{Y}$. |

To test the significance of the regression model, or whether a group of independent variables can provide information on the variation of $\boldsymbol{Y}$ around the middle value, formulated the following hypothesis test.

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

$H_1 : \beta_j \neq 0$ , For the smallest one of the j. $\quad j = 1, 2, \ldots p \qquad$ (G.19)

If $H_0$ is true, then the ratio between the mean of SS(Reg) with mean of SS(Res) will has F distribution with degrees of freedom of numerator is p, and the degrees of freedom of denominator is np-1 (proof can be found in Searle, 1971, and Kshirsagar, 1983). So that the statistic used to test the above hypothesis is

$$F = \frac{SS(\text{Reg})/p}{SS(\text{Res})/n-(p+1)} \qquad \text{(G.20)}$$

Testing criteria:

To a certain significance level $\alpha$ testing used, $H_0$ is accepted if

$F_{value} < F(\alpha, p, n\text{-}p\text{-}1)$ or $P(F > F_{actual}) > \alpha$ , otherwise $H_0$ is rejected.

**Worked Example 5.4:**

An expriment was conducted in order to study the zise of squid eaten by tuna. The regressor variables are characteristics of beak or mouth of the squid. The regressor variables and response considered for the study are: (X1) Rostral length in inches, (X2) Wing length in inches, (X3) Rostral to notch length, (X4) Notch to wing length, (X5) Width in inches, and (Y) weight in pounds. The study involved measurements and weight taken 22 specimen. The data are shown in Table 5.9. The model is given by

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \beta_4 X_{4j} + \beta_5 X_{5j} + \varepsilon_j$$

Test the significance of the regression model.

**Worked Solution:**

To test the significance of the regression model, formulated the following hypothesis test.

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

$H_1 : \beta_j \neq 0$ , For the smallest one of the j. $\quad j = 1, 2, 3, 4, 5.$

In Table 5.10 , given the statistic $F = \frac{SS(\text{Reg})/p}{SS(\text{Res})/n-(p+1)} = \frac{41.602}{0.4948} = 84.070$

**Table 5.9** Squid weigh and beak measurements

| X1 | X2 | X3 | X4 | X5 | Y |
|---|---|---|---|---|---|
| 1,31 | 1,07 | 0,44 | 0,75 | 0,35 | 1,95 |
| 1,55 | 1,49 | 0,53 | 0,90 | 0,47 | 2,90 |
| 0,99 | 0,84 | 0,34 | 0,57 | 0,32 | 0,72 |
| 0,99 | 0,83 | 0,34 | 0,54 | 0,27 | 0,81 |
| 1,05 | 0,90 | 0,36 | 0,64 | 0,30 | 1,09 |
| 1,09 | 0,93 | 0,42 | 0,61 | 0,31 | 1,22 |
| 1,08 | 0,90 | 0,40 | 0,51 | 0,31 | 1,02 |
| 1,27 | 1,08 | 0,44 | 0,77 | 0,34 | 1,93 |
| 0,99 | 0,85 | 0,36 | 0,56 | 0,29 | 0,64 |
| 1,34 | 1,13 | 0,45 | 0,77 | 0,37 | 2,08 |
| 1,30 | 1,10 | 0,45 | 0,76 | 0,38 | 1,98 |
| 1,33 | 1,10 | 0,48 | 0,77 | 0,38 | 1,90 |
| 1,86 | 1,47 | 0,60 | 1,01 | 0,65 | 8,56 |
| 1,58 | 1,34 | 0,52 | 0,95 | 0,50 | 4,49 |
| 1,97 | 1,59 | 0,67 | 1,20 | 0,59 | 8,49 |
| 1,80 | 1,56 | 0,66 | 1,02 | 0,59 | 6,17 |
| 1,75 | 1,58 | 0,63 | 1,09 | 0,59 | 7,54 |
| 1,72 | 1,43 | 0,64 | 1,02 | 0,63 | 6,36 |
| 1,68 | 1,57 | 0,72 | 0,96 | 0,68 | 7,63 |
| 1,75 | 1,59 | 0,68 | 1,08 | 0,62 | 7,78 |
| 2,19 | 1,86 | 0,75 | 1,24 | 0,72 | 10,15 |
| 1,73 | 1,67 | 0,64 | 1,14 | 0,55 | 6,88 |

With $Pr ( F > 84.07) = 0.000$

$F_{value} > F_{0.01;5;16}$ $(84.07 > 4.44)$ or $P( F > F_{actual} ) < 0.01$ , thus $H_0$ is rejected.

**Table 5.10** Analysis of variance summary for regression analysis

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Model | $\hat{\boldsymbol{\beta}}' \mathbf{X}'\boldsymbol{Y} = 595.164$ | 6 | | | |
| Mean | $n\overline{Y}^2 = 387.157$ | 1 | | | |
| Regression | $\hat{\boldsymbol{\beta}}' \mathbf{X}'\boldsymbol{Y} - n\overline{Y}^2 = 208.007$ | 5 | 41.601 | 84.070 | .000[b] |
| Residual | $\boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}' \mathbf{X}'\boldsymbol{Y} = 7.918$ | 16 | .495 | | |
| Total | $\boldsymbol{Y}'\boldsymbol{Y} - n\overline{Y}^2 = 215.925$ | 21 | | | |

a. Dependent Variable: Y

b. Predictors: (Constant), X5, X4, X3, X2, X1

### 5.2.6 PARTIAL REGRESSION COEFFICIENT TEST

Partial test is used to study the contribution of a single independent variable Xj to variations in the response variable Y on the regression model containing all independent variables. Tests conducted on coefficient of the variables, i.e $\beta_j$. The magnitude of the coefficient $\beta_j$ is defined as change in average of the j-th response variable due to per unit changes of the independent variable, with the other independent variables held constant.

In the decomposition of the sum of squares in partial and sequential, SS(Reg) written SS($\beta_1$ , $\beta_2$ , … , ,$\beta_p$ | $\beta_0$ ), so that the j-th coefficient partial sums of squares to be written

SS($\beta_j$ | $\beta_0$ , $\beta_1$ , ... , $\beta_{j-1}$, $\beta_{j+1}$, ... ,$\beta_p$ )

or

$\boldsymbol{R}$( $\beta_j$ | $\beta_0$ , $\beta_1$ , ... , $\beta_{j-1}$, $\beta_{j+1}$, ... ,$\beta_p$)

The partial test can use partial F-test, and can also use the two-way t-test. As indicated earlier, that the variance estimators of the regression coefficients to j is $S^2c_{jj}$, with $c_{jj}$ is a diagonal element to j = 0, 1, ..., p of the matrix ( $\mathbf{X'X}$ )$^{-1}$.

Formulation of hypotheses for the partial coefficient testing are:

$H_0$ : $\beta_j = 0$.

$H_1$ : $\beta_j \neq 0$ , j = 1, 2, … p (G.21)

Testing hypotheses above can be done with the F-test in the right direction, which its statistic

$$F = \frac{\hat{\beta}_j^2}{c_{jj} S^2} \qquad (G.22)$$

and has 1 and n-p-1 degrees of freedom.

F testing criteria:

To a certain significance level testing $\alpha$, $H_0$ is accepted if

$F_{actual} < F(\alpha, 1, n\text{-}p\text{-}1)$ or $P(F > F_{actual}) > \alpha$, otherwise $H_0$ is rejected.

For example, partial F-tests for linear regression models in worked example 5.4, are shown in Table 5.11 below. The results show that only the constant and coefficient $\beta_5$ are significantly different from zero.

**Table 5.11** Tests of between-subjects effects

Dependent Variable: Y

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 208,007[a] | 5 | 41,601 | 84,070 | ,000 |
| Intercept | 24,079 | 1 | 24,079 | 48,660 | ,000 |
| X1 | ,299 | 1 | ,299 | ,604 | ,449 |
| X2 | ,869 | 1 | ,869 | 1,756 | ,204 |
| X3 | ,078 | 1 | ,078 | ,158 | ,696 |
| X4 | ,983 | 1 | ,983 | 1,986 | ,178 |
| X5 | 4,352 | 1 | 4,352 | 8,795 | ,009 |
| Error | 7,918 | 16 | ,495 | | |
| Total | 603,081 | 22 | | | |
| Corrected Total | 215,925 | 21 | | | |

a. R Squared = ,963 (Adjusted R Squared = ,952)

Testing the above hypothesis $H_0 : \beta_j = 0$ may also to use two-way t-test with n-(p +1) degrees of freedom.

$$t = \frac{\hat{\beta}_j}{\sqrt{c_{jj} S^2}} \qquad (G.23)$$

t testing criteria:

To a certain significance level testing $\alpha$, $H_0$ is accepted if

$|t\text{-actual}| \leq t_{\alpha/2;n-p-1}$ , or

$Pr = [\, P(t \leq -|t_{actual}|) + P(t \geq |t_{actual}|)\,] \geq \alpha$, otherwise $H_0$ is rejected.

In this partial test, it should be noted that if the results of testing a variable coefficient is statistically significant at a model that involves all the independent variables, it is not necessarily significant to the model with only a subset of the independent variable.

Partial t-tests for linear regression models in worked example 5.4, are shown in Table 5.12 below. Just as the previous F-test results that only constant and coefficient $\beta_5$ are significantly different from zero.

**Table 5.12** Coefficients

| Model | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|
| | B | Std. Error | Beta | | |
| (Constant) | -6,512 | ,934 | | -6,976 | ,000 |
| X1 | 1,999 | 2,573 | ,223 | ,777 | ,449 |
| X2 | -3,675 | 2,774 | -,371 | -1,325 | ,204 |
| X3 | 2,524 | 6,347 | ,105 | ,398 | ,696 |
| X4 | 5,158 | 3,660 | ,366 | 1,409 | ,178 |
| X5 | 14,401 | 4,856 | ,671 | 2,966 | ,009 |

a. Dependent Variable: Y

### 5.2.7 SEQUENTIAL F-TEST

Sequential test is used to study the contribution of a variable in the model containing the preceding regressor variables. The order of entry, then, can have a profound effect on the results. If regressor 4 is adjusted for 1, 2, and 3, it is quite possible that its contribution to SS(Reg) will be quite different than

if it were adjusted for, say, only variable 1. In other word, *the appropriateness of regressor variable often depend on what regressor variables are in the model with it*.

The sum of squares, which is discussed on a partial test, not forming additives contribute to the SS (Reg),

$$SS(\beta_1 , \beta_2 , \dots , \beta_p \mid \beta_0 ) \neq SS(\beta_1 \mid \beta_0 , \beta_2 , \dots , \beta_p ) + SS(\beta_2 \mid \beta_0 , \beta_1 , \beta_3 , \dots , \beta_p ) + \dots$$
$$+ SS(\beta_j \mid \beta_0 , \beta_1 , \dots , \beta_{j-1}, \beta_{j+1}, \dots , \beta_p ) + \dots + SS(\beta_p \mid \beta_0 , \beta_1 , \beta_2 , \dots , \beta_{p-1} ).$$

But they form additive partitioning of the SS (Reg), scilicet:

$$SS(\beta_1 , \beta_2 , \dots , \beta_p \mid \beta_0 ) = SS(\beta_1 \mid \beta_0 ) + SS(\beta_2 \mid \beta_0 , \beta_1 ) + SS(\beta_3 \mid \beta_0 , \beta_1 , \beta_2 )$$
$$+ \dots + SS(\beta_p \mid \beta_0 , \beta_1 , \beta_2 , \dots , \beta_{p-1} ). \qquad \text{(G.24)}$$

The notation SS( . | . ) states "regression explained by ...", with the line vertikal denoting "in the presence of ...". For example, $SS(\beta_2 \mid \beta_0 , \beta_1 )$ is an increase in the regression sum of squares, when the regressor $X_2$ is added to a model that involving only $X_1$ and the constant term.

The sequential and partial sum of squares for worked example 5.4 are shown in Table 5.13 follows.

**Table 5.13** The sequential and partial sum of squares

| Sequential | | Partial | |
|---|---|---|---|
| $SS(\beta_1 \mid \beta_0 )$ | = 99.145 | $SS(\beta_1 \mid \beta_0 , \beta_2 , \beta_3 , \beta_4 , \beta_5)$ | =0.299 |
| $SS(\beta_2 \mid \beta_0 , \beta_1 )$ | = 0.127 | $SS(\beta_2 \mid \beta_0 , \beta_1 , \beta_3 , \beta_4 , \beta_5)$ | =0.869 |
| $SS(\beta_3 \mid \beta_0 , \beta_1 , \beta_2 )$ | = 4.120 | $SS(\beta_3 \mid \beta_0 , \beta_1 , \beta_2 , \beta_4 , \beta_5)$ | =0.078 |
| $SS(\beta_4 \mid \beta_0 , \beta_1 , \beta_2 , \beta_3 )$ | = 0.263 | $SS(\beta_4 \mid \beta_0 , \beta_1 , \beta_2 , \beta_3 , \beta_5)$ | =0.983 |
| $SS(\beta_5 \mid \beta_0 , \beta_1 , \beta_2 , \beta_3 , \beta_4)$ | = 4.352 | $SS(\beta_5 \mid \beta_0 , \beta_1 , \beta_2 , \beta_3 , \beta_4)$ | =4.352 |

Partitions of sequential sum of squares is very useful when we need information about the worth for a subset of regressors. For example, suppose, the regression model with p = 4 regressors, we can write

$SS(\beta_1, \beta_2, \beta_3, \beta_4 \mid \beta_0) = SS(\beta_1, \beta_2 \mid \beta_0) + SS(\beta_3, \beta_4 \mid \beta_0, \beta_1, \beta_2)$

Where both term on the right-hand side represents the partition with two degrees of freedom. $SS(\beta_1, \beta_2 \mid \beta_0)$ less useful for the inference on $\beta_1$ and $\beta_2$, since there has been no adjusment for $X_3$ dan $X_4$. However $SS(\beta_3, \beta_4 \mid \beta_0, \beta_1, \beta_2)$ would be vital in explaining the importance of $X_3$ dan $X_4$ Collectivelly. If we are interested in performing joint inference of the $\beta_1$ and $\beta_2$, then $SS(\beta_3, \beta_4 \mid \beta_0, \beta_1, \beta_2)$ can be calculated by

$$SS(\beta_3, \beta_4 \mid \beta_0, \beta_1, \beta_2) = SS(\beta_3 \mid \beta_0, \beta_1, \beta_2) + SS(\beta_4 \mid \beta_0, \beta_1, \beta_2, \beta_3) \qquad \text{(G.25)}$$

Furthermore, the statistic F can be used

$$F = \frac{SS(\beta_3, \beta_4 \mid \beta_0, \beta_1, \beta_2)/2}{MS(\text{Res})} \qquad \text{(G.26)}$$

with 2 degrees of freedom numerator and n-p-1 denominator. These statistics are used to test the hypothesis,

$H_0 : \beta_3 = \beta_4 = 0$

$H_1 : \beta_3 \neq 0$ or $\beta_4 \neq 0$

If we want to test the contribution of a variable $X_j$, that the model involving only the preceding regressor variables, namely $X_1$, $X_2$, .., $X_{j-1}$, the test statistic

$$F = \frac{SS(\beta_j \mid \beta_0, \beta_1, \ldots, \beta_{j-1})}{MS(\text{Res})} \qquad \text{(G.27)}$$

The test statistic has distribution F and successive degrees of freedom numerator and denominator 1, and n-p-1.

For example, sequential F-tests for linear regression model in worked example 5.4, are shown in Table 5.14 below. Of course, here there is a difference with the previous partial test results. This is due to the partial coefficient test, each coefficients of independent variable are tested when the model contains all

the other variables. While on sequeltial test, the coefficient of independent variables were tested in order to enter. For example, the coefficient $\beta_3$ was tested when the regressor $X_3$ is added to a model that involving only $X_1$ , $X_2$, and the constant term.

**Table 5.14** The sequential F-tests for linear regression model

Dependent Variable: Y

| Source | Type I Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 208,007[a] | 5 | 41,601 | 84,070 | ,000 |
| Intercept | 387,157 | 1 | 387,157 | 782,379 | ,000 |
| X1 | 199,145 | 1 | 199,145 | 402,440 | ,000 |
| X2 | ,127 | 1 | ,127 | ,256 | ,620 |
| X3 | 4,120 | 1 | 4,120 | 8,325 | ,011 |
| X4 | ,263 | 1 | ,263 | ,532 | ,476 |
| X5 | 4,352 | 1 | 4,352 | 8,795 | ,009 |
| Error | 7,918 | 16 | ,495 | | |
| Total | 603,081 | 22 | | | |
| Corrected Total | 215,925 | 21 | | | |

a. R Squared = ,963 (Adjusted R Squared = ,952)

Furthermore, suppose we want to use the sequential sum of squares to test on subsets of variables such as $H_0$: $\beta_4 = \beta_5 = 0$. Dengan menggunakan hasil pada tabel 5, The proper sum of squares is found by computing $SS(\beta_5, \beta_4 \mid \beta_0, \beta_1, \beta_2, \beta_3)$ and using the results in Table 5.12.

$$SS(\beta_5, \beta_4 \mid \beta_0, \beta_1, \beta_2, \beta_3) = SS(\beta_4 \mid \beta_0, \beta_1, \beta_2, \beta_3) + SS(\beta_5 \mid \beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$$
$$= 0.263 + 4.352$$
$$= 4.615$$

$$F = \frac{4.615/2}{0.495} = 64.663$$

The numerator and denominator degrees of freedom are 2 and 16 respectively. The hypothesis is rejected at the 0.025 level.

### 5.2.8 TESTING OF THE GENERAL LINEAR HYPOTHESIS

Partial coefficient test and a test of the coefficient subset, actually can be done through the establishment of general linear hypothesis. General linear hypothesis is defined as follows.

$H_0$ : $\mathbf{K'}\,\beta = \mathbf{m}$

$H_1$ : $\mathbf{K'}\,\beta \neq \mathbf{m}$ (G.28)

Where **K'** is a (k x (p+1)) matrix of coefficients defining k linear function of $\beta$ to be tested. Each row of the matrix **K'** consists of the coefficients of a linear function, and **m** is a (kx1) vector of constant.

Suppose $\beta' = (\beta_0, \beta_1, \beta_2, \beta_3)$ and we want to test the null hypothesis that the composition of $\beta_1 = \beta_2$ , $\beta_1 + \beta_2 = 2\beta_3$, dan $\beta_0 = 20$. Hypotheses are equivalent to

$$\begin{aligned} H_0 : \quad & \beta_1 - \beta_2 = 0 \\ & \beta_1 + \beta_2 - 2\beta_3 = 0 \\ & \beta_0 = 20 \end{aligned}$$

These three linear functions can be written in the form $\mathbf{K'}\,\beta = \mathbf{m}$ , by defining

$$\mathbf{K'} = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 1 & -2 \\ 1 & 0 & 0 & 0 \end{pmatrix} \text{ and } \mathbf{m} = \begin{pmatrix} 0 \\ 0 \\ 20 \end{pmatrix}$$

The least squares estimate for $\mathbf{K'}\,\beta - \mathbf{m}$ is obtained by subtituting the least squares estimate $\hat{\boldsymbol{\beta}}$ for $\beta$, to obtain $\mathbf{K'}\,\hat{\boldsymbol{\beta}} - \mathbf{m}$ . Under normality assumptions for $\boldsymbol{Y}$, and Eq. (G.5) that $\hat{\boldsymbol{\beta}}$ is linear function of $\boldsymbol{Y}$ , then

$E(\mathbf{K'}\,\hat{\boldsymbol{\beta}} - \mathbf{m}) = \mathbf{K'}\,E(\hat{\boldsymbol{\beta}}) - m = \mathbf{K'}\,\beta - \mathbf{m}$

If $H_0$ is true, then $\mathbf{K'}\,\beta - \mathbf{m} = \mathbf{0}$, and varince-covariance matrix

$$\begin{aligned} \text{Var}(\mathbf{K'}\,\hat{\boldsymbol{\beta}} - \mathbf{m}) &= \text{Var}(\mathbf{K'}\,\hat{\boldsymbol{\beta}}) - 0 \\ &= \mathbf{K'}\,\text{Var}(\hat{\boldsymbol{\beta}})\,\mathbf{K} \\ &= \mathbf{K'}\,(\mathbf{X'X})^{-1}\,\mathbf{K}\,\sigma^2. \end{aligned} \quad \text{(G.29)}$$

The sum of squares for liear hypotesis $H_0$ : $\mathbf{K'\beta = m}$ is computed by (Searle, 1971),

$$Q = (\mathbf{K'}\hat{\boldsymbol{\beta}} - \mathbf{m})' \ (\mathbf{K'}(\mathbf{X'X})^{-1}\mathbf{K})^{-1}(\mathbf{K'}\hat{\boldsymbol{\beta}} - \mathbf{m}) \qquad \text{(G.30)}$$

This is quadratic form in $\mathbf{K'}\hat{\boldsymbol{\beta}} - \mathbf{m}$ with defining matrix $\mathbf{A} = (\mathbf{K'}(\mathbf{X'X})^{-1}\mathbf{K})^{-1}$ . The defining matrix, except $1/\sigma^2$, is a inverse of the varince-covariance matrix of linear functions $\mathbf{K'}\hat{\boldsymbol{\beta}} - \mathbf{m}$. Thus, $tr(\mathbf{AV}) = tr(\mathbf{I}_k) = k\sigma^2$ . Furthermore, expectation of Q

$$E(Q) = k\sigma^2 + (\mathbf{K'}\beta - \mathbf{m})' \ (\mathbf{K'}(\mathbf{X'X})^{-1}\mathbf{K})^{-1}(\mathbf{K'}\beta - \mathbf{m}) \qquad \text{(G.31)}$$

With the assumption of normality, $Q/\sigma^2$ is distributed as a noncentral Chi-squares random variable with k degrees of freedom. In Rawlings (1988), F-test for hypothesis $H_0$ : $\mathbf{K'\beta = m}$ is

$$F = \frac{Q/k}{S^2} \qquad \text{(G.32)}$$

**Worked Example 5.5:**

For example problems using data from environmental studies in Table 5.6. By using the statistical software S-Plus, obtained matrix $(\mathbf{X'X})^{-1}$ as follows.

Matrix $(\mathbf{X'X})^{-1}$ :

```
          [,1]         [,2]            [,3]           [,4]
[1,]  2.8559225951  6.173356e-004 -0.03535068677 -4.090973e-002
[2,]  0.0006173356  3.341412e-006 -0.00001807261  1.409908e-008
[3,] -0.0353506868 -1.807261e-005  0.00053385411  1.439104e-004
[4,] -0.0409097295  1.409908e-008  0.00014391044  2.613921e-003
```

Test the contribution of 2 regressors, namely radiation ($X_1$) and wind ($X_3$) on the model that includes variable temperature ($X_2$).

**Worked Solution:**

The hypothesis tested is $H_0$: $\beta_1 = \beta_3 = 0$. General linear hypothesis is defined as follows.

.

$$\mathbf{K'}\,\boldsymbol{\beta} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

By matrix multiplication of $\mathbf{K'}\,(\,\mathbf{X'X}\,)^{-1}\,\mathbf{K}$, obtained a (2x2) submatrix extracted from the matrix $(\,\mathbf{X'X}\,)^{-1}$,

```
          [,1]          [,2]              [,3]          [,4]
[1,]
[2,]               3.341412e-006                   1.409908e-008
[3,]
[4,]               1.409908e-008                   2.613921e-003
```

From the results of previous work example 5.3, has obtained $\hat{\beta}_1 = 0.0013$ and $\hat{\beta}_3 = -0.0280$. Furthermore, the sum of squares Q can be calculated as follows

$$Q = (\mathbf{K'}\,\hat{\boldsymbol{\beta}})'\,(\mathbf{K'}\,(\,\mathbf{X'X}\,)^{-1}\,\mathbf{K})^{-1}\,(\mathbf{K'}\,\hat{\boldsymbol{\beta}})$$

$$= \begin{pmatrix} 0.0013 \\ -0.0280 \end{pmatrix}^{T} \begin{pmatrix} 3.341412e-006 & 1.409908e-008 \\ 1.409908e-008 & 2.613921e-003 \end{pmatrix}^{-1} \begin{pmatrix} 0.0013 \\ -0.0280 \end{pmatrix}$$

$$= 0.807$$

$$F = \frac{Q/k}{S^2} = \frac{0.807/2}{0.349195} = 1.1555$$

Which is much less than the critical value of F for $\alpha = 0.05$ and 2 and 26 degrees of freedom, $F_{0.05;2;26} = 3.37$ , There is no reason to reject $H_0$ that $\beta_1 = \beta_3 = 0$.

Output using *Statistical Analysis System* (*SAS*) software presented of the results which showing in Table 5.15.

**Table 5.15** *Output* of *SAS* for a general linear hypothesis testing

```
                          The SAS System

                    L Ginv(X'X) L'                Lb-c


            3.3414119E-6      1.4099083E-8      0.0013058159
            1.4099083E-8       0.002613921      -0.027843496


                Inv(L Ginv(X'X) L')                Inv()(Lb-c)


            299274.69475      -1.614241062      390.84258745
            -1.614241062      382.56703747      -10.65411167

Dependent Variable: Y
Test:          Numerator:      0.4035  DF:    2   F value:   1.1562
               Denominator:  0.348994  DF:   26   Prob>F:    0.3303
```

The second hypothesis illustration a case $\mathbf{m} \neq 0$. Suppose prior information suggested that the intercept $\beta_0$ for a group of mean of this radiation and wind should be 0.5 ($\beta_0 = 0.5$). We will construct a composite hypothesis by adding constraint $\beta_0 = 0.5$ to the two conditons in the first null hypothesis. The null null hypothesis is $H_0$ : $\mathbf{K'\beta - m = 0}$ where

$$\mathbf{K'\beta - m} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} - \begin{pmatrix} -0.5 \\ 0 \\ 0 \end{pmatrix}$$

For this hypotesis

$$\mathbf{K'}\hat{\boldsymbol{\beta}} \mathbf{- m} = \begin{pmatrix} -0.2949 - 0.5 \\ 0.0013 \\ -0.0280 \end{pmatrix} = \begin{pmatrix} -0.7949 \\ 0.0013 \\ -0.0280 \end{pmatrix}$$

and

$$(\mathbf{K'}\,(\,\mathbf{X'X}\,)^{-1}\,\mathbf{K})^{-1} = \begin{pmatrix} 2.8559225951 & 6.173356e-004 & -4.090973e-002 \\ 0.0006173356 & 3.341412e-006 & 1.409908e-008 \\ -0.0409097295 & 1.409908e-008 & 2.613921e-003 \end{pmatrix}^{-1}$$

Notice that causes the hypothesized $\beta_0 = 0.5$ to be subtracted from the estimated $\beta_0 = 0.2949$. The sum of squares for this composite hypothesis is

$$Q = (\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m})' \ (\mathbf{K}' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{K})^{-1} (\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{m})$$

$$= 1.8387$$

and has 3 degrees of freedom. Computed statistic F is

$$F = \frac{Q/k}{S^2} = \frac{1.8387/3}{0.349195} = 1.755$$

Which, again, is much less than the critical value of F for $\alpha = 0.05$ and (3;26) degrees of freedom. $F_{0.05;3;26} = 2.98$ , There is no reason to reject $H_0$ that $\beta_0 = 0.5$ and $\beta_1 = \beta_3 = 0$.

Output using *Statistical Analysis System* (*SAS*) software presented of the results which showing in Table 5.16.

**Table 5.16** *Output* of *SAS* for a composited hypothesis testing

```
                          The SAS System

                      L Ginv(X'X) L'   Lb-c


     2.8559225951     0.0006173356      -0.04090973     -0.795271027
     0.0006173356     3.3414119E-6     1.4099083E-8     0.0013058159
      -0.04090973     1.4099083E-8      0.002613921     -0.027843496



              Inv(L Ginv(X'X) L')    Inv()(Lb-c)


     0.4758460504     -87.94537601     7.4478046495     -0.700639963
     -87.94537601      315528.6684     -1378.109811     520.33414234
     7.4478046495     -1378.109811      499.1379201     -21.62032556

Dependent Variable: Y
Test:          Numerator:      0.6129  DF:    3   F value:   1.7561
               Denominator:  0.348994  DF:   26   Prob>F:    0.1803
```

**EXERCISES 5**

1. Data on the response Y is the amount of suspended solids and the pH of the cleansing tank in a coal cleasing system recorded in Table 5.17. If the relationship between the two variables want to be investigated by simple linear regression models,

   a. Estimate the linear regression equation

   b. Perform the significance test for the allegation regression equation through F test

   c. Check the accuracy of this model linier relationship Y with X by the coefficient of determination $R^2$

   d. Perform testing of the coefficient $\beta_1 = 0$

**Table 5.17** Data set of a coal cleasing system

| pH (X1) | Amount of suspended solids (Y) |
|---|---|
| 6,5 | 292 |
| 6,9 | 329 |
| 7,8 | 352 |
| 8,4 | 378 |
| 8,8 | 392 |
| 9,2 | 410 |
| 6,7 | 198 |
| 6,9 | 227 |
| 7,5 | 277 |
| 7,9 | 297 |
| 8,7 | 364 |
| 9,2 | 375 |
| 6,5 | 167 |
| 7,0 | 225 |
| 7,2 | 247 |
| 7,6 | 268 |
| 8,7 | 288 |
| 9,2 | 342 |

2. Data on the measurements results of several organs of the body are presented in Table 5.18. With linear regression model, want to investigate the influence of a group of independent variables, namely the size of the lungs (X1), heart (X2), liver (X3), spleen (X4) against Kidney size (Y). In Table 5.19, through the SAS output is presented the matrix $(\mathbf{X'X})^{-1}$ of linear regression model.

a. Find the estimated regression equation

b. Test the significance of the regression model by testing the following formulation of the test hypothesis

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

$H_1 : \beta_j \neq 0$ , For the smallest one of the j  j = 1, 2, 3,4

Find the sum of squares regression using

$SS(Reg) = SS(\beta_1 \mid \beta_0) + SS(\beta_2 \mid \beta_0, \beta_1) + SS(\beta_3 \mid \beta_0, \beta_1, \beta_2)$

$+ SS(\beta_4 \mid \beta_0, \beta_1, \beta_2, \beta_3)$

c. Perform partial testing for each regression coefficients, i.e test the following hypotheses.

$H_0 : \beta_j = 0.$

$H_1 : \beta_j \neq 0$ ,  j = 1, 2, 3, 4

d. Perform sequential testing on the order of independent variables $X_1$, $X_2$, $X_3$ , and $X_4$.

e. By defining a general linear hypothesis, test the contribution 3 regressors, namely the lungs ($X_1$), liver ($X_3$), and spleen ($X_4$), to the sum of squares regression, which is testing the hypothesis $H_0 : \beta_1 = \beta_3 = \beta_4 = 0$.

f. By defining a general linear hypothesis, test the hypothesis

$H_0$: $\beta_0 = 0$ dan $\beta_1 = \beta_3 = \beta_4 = 0$.

**Table 5.18** Data on the results of measurements of body organs

| | lung | heart | liver | spleen | kidney |
|---|---|---|---|---|---|
| 1 | 0.9695 | 1.3520 | 8.1855 | 0.2255 | 1.3480 |
| 2 | 1.1410 | 1.4994 | 8.5445 | 0.2705 | 1.3221 |
| 3 | 0.8720 | 1.2685 | 6.8230 | 0.0500 | 0.8970 |
| 4 | 0.7399 | 1.2702 | 9.0675 | 0.1130 | 1.2780 |
| 5 | 1.1393 | 1.4962 | 9.6220 | 0.1198 | 1.4150 |
| 6 | 1.1256 | 1.1914 | 9.7432 | 0.1393 | 1.3405 |
| 7 | 0.3672 | 0.6986 | 2.7266 | 0.0478 | 0.8199 |
| 8 | 0.3227 | 0.4193 | 1.7442 | 0.0448 | 0.8224 |
| 9 | 0.4972 | 0.4134 | 2.6617 | 0.0568 | 1.0697 |
| 10 | 0.5366 | 0.5298 | 3.2621 | 0.0751 | 0.9601 |
| 11 | 0.5219 | 0.6020 | 3.6231 | 0.0701 | 1.1635 |
| 12 | 0.4291 | 0.4230 | 2.0776 | 0.0383 | 0.8199 |
| 13 | 0.5354 | 0.5129 | 2.8608 | 0.0745 | 0.9825 |
| 14 | 0.4529 | 0.5715 | 2.8093 | 0.0559 | 1.0074 |
| 15 | 0.6532 | 0.5187 | 3.6025 | 0.0604 | 1.0751 |
| 16 | 0.9191 | 0.5479 | 2.8636 | 0.0687 | 0.9596 |
| 17 | 0.5620 | 0.5001 | 2.9257 | 0.0780 | 0.9589 |
| 18 | 0.5362 | 0.9423 | 2.6974 | 0.0543 | 1.0797 |
| 19 | 0.5516 | 0.7950 | 2.7466 | 0.0657 | 1.0101 |
| 20 | 0.5789 | 0.5051 | 3.1271 | 0.0755 | 1.0798 |
| 21 | 0.6553 | 0.5999 | 3.4968 | 0.0436 | 1.1976 |
| 22 | 0.5613 | 1.0344 | 3.1108 | 0.0668 | 1.2131 |
| 23 | 0.5370 | 0.7446 | 2.6780 | 0.0450 | 1.0426 |
| 24 | 0.6220 | 0.9023 | 8.3071 | 0.0997 | 1.3686 |
| 25 | 0.7738 | 0.7628 | 3.7972 | 0.0685 | 1.0629 |
| 26 | 1.0500 | 0.8981 | 4.8977 | 0.0821 | 1.0119 |
| 27 | 0.8416 | 1.5094 | 7.4928 | 0.1739 | 1.3171 |
| 28 | 0.8739 | 0.6723 | 2.6366 | 0.0536 | 1.0983 |
| 29 | 0.8260 | 0.9604 | 12.8040 | 0.1193 | 1.5933 |
| 30 | 0.7589 | 0.9483 | 5.3909 | 0.0828 | 1.1872 |
| 31 | 1.1778 | 1.3692 | 8.1259 | 0.1376 | 1.3538 |
| 32 | 1.0119 | 1.2080 | 7.1872 | 0.0784 | 1.4338 |
| 33 | 0.9820 | 1.2860 | 7.9245 | 0.1194 | 1.6396 |
| 34 | 0.6657 | 1.0690 | 4.4023 | 0.0752 | 1.2131 |
| 35 | 0.6190 | 1.0002 | 8.0259 | 0.1056 | 1.5638 |
| 36 | 0.7849 | 1.1167 | 5.9199 | 0.0952 | 1.2582 |
| 37 | 0.5092 | 0.8698 | 3.2649 | 0.0876 | 1.3482 |
| 38 | 0.5687 | 0.8791 | 3.6023 | 0.1796 | 1.1772 |
| 39 | 1.4364 | 0.9932 | 8.2075 | 0.1187 | 1.3165 |
| 40 | 1.0725 | 0.9484 | 6.9736 | 0.1043 | 1.4693 |
| 41 | 1.1626 | 1.2746 | 11.9148 | 0.1072 | 1.4999 |
| 42 | 0.9650 | 1.5450 | 6.5425 | 0.1062 | 1.2677 |
| 43 | 1.0239 | 1.0789 | 8.1096 | 0.1424 | 1.4012 |
| 44 | 0.3216 | 0.4987 | 1.6264 | 0.0293 | 0.8252 |
| 45 | 1.1653 | 1.6300 | 7.0897 | 0.1444 | 1.2884 |
| 46 | 0.8532 | 1.1439 | 4.0120 | 0.0769 | 1.2271 |
| 47 | 0.6354 | 0.8940 | 2.6877 | 0.0463 | 0.9939 |
| 48 | 0.6489 | 1.2394 | 8.7761 | 0.1145 | 1.4262 |
| 49 | 0.9620 | 0.8039 | 3.4576 | 0.0713 | 1.1646 |
| 50 | 0.7959 | 1.0385 | 8.3850 | 0.1990 | 1.3135 |
| 51 | 0.8099 | 1.1543 | 7.5313 | 0.0595 | 1.3867 |
| 52 | 0.9824 | 1.4903 | 7.9353 | 0.1567 | 1.3642 |
| 53 | 0.9592 | 1.0950 | 15.4923 | 0.2959 | 1.4378 |
| 54 | 0.9596 | 1.2763 | 10.9988 | 0.1617 | 1.3394 |
| 55 | 1.1194 | 1.4415 | 5.9829 | 0.1308 | 1.1359 |
| 56 | 1.0748 | 1.1714 | 8.6132 | 0.1283 | 1.2379 |

**Table 5.19** Output SAS for the matrix $(\mathbf{X'X})^{-1}$

X'X Inverse

| | INTERCEP | X1 | X2 |
|---|---|---|---|
| INTERCEP | 0.2143847334 | -0.168027136 | -0.095712472 |
| X1 | -0.168027136 | 0.6152818699 | -0.225741783 |
| X2 | -0.095712472 | -0.225741783 | 0.3972158583 |
| X3 | 0.006655296 | -0.015950671 | -0.011878562 |
| X4 | -0.094818419 | -0.038820519 | -0.428510383 |

| | X3 | X4 |
|---|---|---|
| INTERCEP | 0.006655296 | -0.094818419 |
| X1 | -0.015950671 | -0.038820519 |
| X2 | -0.011878562 | -0.428510383 |
| X3 | 0.005261954 | -0.131875573 |
| X4 | -0.131875573 | 12.971992938 |

3. The data in Table 5.19 consisted of 26 subjects were selected to study the effect of exercise activities (running and weight lifting), and body weight on HDL cholesterol. The subjects consisted of 8 people placed as the control group, 8 people in a group that took part in a relatively rigorous running program, and 10 peopole were placed in a program involving rigorous running and weightlifting. The weights and HDL cholesterol of the subjects were recorded after ten weeks of the program. If the linear regression model applied to each group, there will be three simple linear regression equation. Perform regression testing of the three equation, through test equality of coefficients (slope).

   As a result, the three following models are postulated.

   $Y_j = \beta_{01} + \beta_{1.1} X_{1j} + \varepsilon_j \quad j = 1, 2, \ldots 8 \quad$ (Control group)

   $Y_j = \beta_{02} + \beta_{1.2} X_{1j} + \varepsilon_j \quad j = 9, 10, \ldots 16 \quad$ (Running group)

   $Y_j = \beta_{03} + \beta_{1.3} X_{1j} + \varepsilon_j \quad j = 17, 18, \ldots 26 \quad$ (Running and lifting group)

   If the regression models and composition of data is designed with the **X** matrix and the $\boldsymbol{\beta}$ vector like as Eq.(G.33):

   a. Estimate of each regression coefficients
   b. Tests equality of the thee slopes through tensting the general linear hypothesis

$H_0$ : $\beta_{1.1} - \beta_{1.2} = 0$

$\beta_{1.1} - \beta_{1.3} = 0$

And defining matrix $\mathbf{K'\beta = m}$ ,where

$$\mathbf{K'} = \begin{pmatrix} 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix} \text{ dan } \mathbf{m} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

**Table 5.19** Effect of exercise activities on HDL cholesterol.

| | Groups | Weight (lb) | HDL Cholesterol (mg/decaliter) |
|---|---|---|---|
| Control | 0 | 163.5 | 75.0 |
| | 0 | 180.0 | 72.5 |
| | 0 | 178.5 | 62.0 |
| | 0 | 161.5 | 60.0 |
| | 0 | 127.0 | 53.0 |
| | 0 | 161.0 | 53.0 |
| | 0 | 165.0 | 65.0 |
| | 0 | 144.0 | 63.5 |
| Running | 1 | 141.0 | 49.0 |
| | 1 | 162.0 | 53.5 |
| | 1 | 134.0 | 30.0 |
| | 1 | 121.0 | 40.5 |
| | 1 | 145.0 | 51.5 |
| | 1 | 106.0 | 57.5 |
| | 1 | 134.0 | 49.0 |
| | 1 | 216.5 | 74.0 |
| Running and Weightlifting | 2 | 136.5 | 54.5 |
| | 2 | 142.5 | 79.5 |
| | 2 | 145.0 | 64.0 |
| | 2 | 165.0 | 69.0 |
| | 2 | 226.0 | 50.5 |
| | 2 | 122.0 | 58.0 |
| | 2 | 193.0 | 63.5 |
| | 2 | 163.5 | 76.0 |
| | 2 | 154.0 | 55.5 |
| | 2 | 139.0 | 68.0 |

The composition of the combined models;

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ . \\ . \\ . \\ Y_{16} \end{pmatrix} = \begin{array}{c} \begin{matrix} c1 & c2 & c3 & x1 & x2 & x3 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 163.5 & 0.0 & 0.0 \\ 1 & 0 & 0 & 180.0 & 0.0 & 0.0 \\ 1 & 0 & 0 & 178.5 & 0.0 & 0.0 \\ 1 & 0 & 0 & 161.5 & 0.0 & 0.0 \\ 1 & 0 & 0 & 127.0 & 0.0 & 0.0 \\ 1 & 0 & 0 & 161.0 & 0.0 & 0.0 \\ 1 & 0 & 0 & 165.0 & 0.0 & 0.0 \\ 1 & 0 & 0 & 144.0 & 0.0 & 0.0 \\ 0 & 1 & 0 & 0.0 & 141.0 & 0.0 \\ 0 & 1 & 0 & 0.0 & 162.0 & 0.0 \\ 0 & 1 & 0 & 0.0 & 134.0 & 0.0 \\ 0 & 1 & 0 & 0.0 & 121.0 & 0.0 \\ 0 & 1 & 0 & 0.0 & 145.0 & 0.0 \\ 0 & 1 & 0 & 0.0 & 106.0 & 0.0 \\ 0 & 1 & 0 & 0.0 & 134.0 & 0.0 \\ 0 & 1 & 0 & 0.0 & 216.5 & 0.0 \\ 0 & 0 & 1 & 0.0 & 0.0 & 136.5 \\ 0 & 0 & 1 & 0.0 & 0.0 & 142.5 \\ 0 & 0 & 1 & 0.0 & 0.0 & 145.0 \\ 0 & 0 & 1 & 0.0 & 0.0 & 165.0 \\ 0 & 0 & 1 & 0.0 & 0.0 & 226.0 \\ 0 & 0 & 1 & 0.0 & 0.0 & 122.0 \\ 0 & 0 & 1 & 0.0 & 0.0 & 193.0 \\ 0 & 0 & 1 & 0.0 & 0.0 & 163.5 \\ 0 & 0 & 1 & 0.0 & 0.0 & 154.0 \\ 0 & 0 & 1 & 0.0 & 0.0 & 139.0 \end{bmatrix} \end{array} \begin{pmatrix} \beta_{01} \\ \beta_{02} \\ \beta_{03} \\ \beta_{1.1} \\ \beta_{1.2} \\ \beta_{1.3} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ . \\ \varepsilon_{16} \end{pmatrix} \qquad \text{(G.33)}$$

# BIBLIOGRAPHY

Fleming, M. C, and Nellis, J.G., (1994), *Principles of Applied Statistics*, First edition, Routledge, New York

Kshirsagar, A. M., (1983), *A Course in Linear Models*, New York, Marcel Dekker, Inc.

Myers, R. H., (1990), *Classical and Modern Regression With Applications*, 2 nd edition, PWS-KENT Publishing Company, Boston

Rawlings, J. O., ( 1988 ), *Applied Regression Analysis*, Wadsworth and Brooks, California

Walpole, R. E., (1982), *Introduction to Statistics*, 3 rd edition, Macmillan Publishing Co. Inc.

SAS Institute Inc, (1998), *SAS User's Guide*: Statistics, Version 6. Cary, North Carolina: SAS Institute Inc.

Searle. S. R, (1971), *Linear Model*, New York: Wiley

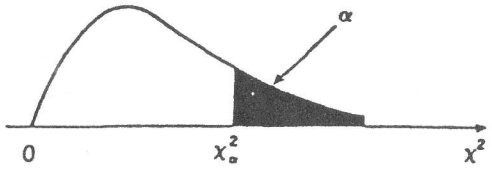
Searle. S. R, (1971), *Matrix Algebra Useful for Statistics*, New York: Wiley

Statistical Sciences, ( 1993 ), *S-Plus User's Manual*, Versi 3.2, MathSoft Inc., Massachusetts

# APPENDIX TABLES

## TABLE A7 $\chi^2$ DISTRIBUTION

Entries in the table give $\chi^2_\alpha$, where $\alpha$ is the area or probability in the tail of the distribution (the shaded area). For example, with ten degrees of freedom and $\alpha = 0.05$, $\chi^2_\alpha = 18.307$.

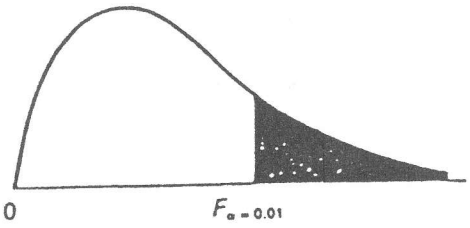



| Degrees of freedom | $\chi^2_{0.995}$ | $\chi^2_{0.990}$ | $\chi^2_{0.975}$ | $\chi^2_{0.950}$ | $\chi^2_{0.900}$ |
|---|---|---|---|---|---|
| 1 | 0.0000393 | 0.0001571 | 0.0009821 | 0.0039321 | 0.0157908 |
| 2 | 0.0100251 | 0.0201007 | 0.0506356 | 0.102587 | 0.210720 |
| 3 | 0.0717212 | 0.114832 | 0.215795 | 0.351846 | 0.584375 |
| 4 | 0.206990 | 0.297110 | 0.484419 | 0.710721 | 1.063623 |
| 5 | 0.411740 | 0.554300 | 0.831211 | 1.145476 | 1.61031 |
| 6 | 0.675727 | 0.872085 | 1.237347 | 1.63539 | 2.20413 |
| 7 | 0.989265 | 1.239043 | 1.68987 | 2.16735 | 2.83311 |
| 8 | 1.344419 | 1.646482 | 2.17973 | 2.73264 | 3.48954 |
| 9 | 1.734926 | 2.087912 | 2.70039 | 3.32511 | 4.16816 |
| 10 | 2.15585 | 2.55821 | 3.24697 | 3.94030 | 4.86518 |
| 11 | 2.60321 | 3.05347 | 3.81575 | 4.57481 | 5.57779 |
| 12 | 3.07382 | 3.57056 | 4.40379 | 5.22603 | 6.30380 |
| 13 | 3.56503 | 4.10691 | 5.00874 | 5.89186 | 7.04150 |
| 14 | 4.07468 | 4.66043 | 5.62872 | 6.57063 | 7.78953 |
| 15 | 4.60094 | 5.22935 | 6.26214 | 7.26094 | 8.54675 |
| 16 | 5.14224 | 5.81221 | 6.90766 | 7.96164 | 9.31223 |
| 17 | 5.69724 | 6.40776 | 7.56418 | 8.67176 | 10.0852 |
| 18 | 6.26481 | 7.01491 | 8.23075 | 9.39046 | 10.8649 |
| 19 | 6.84398 | 7.63273 | 8.90655 | 10.1170 | 11.6509 |
| 20 | 7.43386 | 8.26040 | 9.59083 | 10.8508 | 12.4426 |
| 21 | 8.03366 | 8.89720 | 10.28293 | 11.5913 | 13.2396 |
| 22 | 8.64272 | 9.54249 | 10.9823 | 12.3380 | 14.0415 |
| 23 | 9.26042 | 10.19567 | 11.6885 | 13.0905 | 14.8479 |
| 24 | 9.88623 | 10.8564 | 12.4011 | 13.8484 | 15.6587 |
| 25 | 10.5197 | 11.5240 | 13.1197 | 14.6114 | 16.4734 |
| 26 | 11.1603 | 12.1981 | 13.8439 | 15.3791 | 17.2919 |
| 27 | 11.8076 | 12.8786 | 14.5733 | 16.1513 | 18.1138 |
| 28 | 12.4613 | 13.5648 | 15.3079 | 16.9279 | 18.9392 |
| 29 | 13.1211 | 14.2565 | 16.0471 | 17.7083 | 19.7677 |
| 30 | 13.7867 | 14.9535 | 16.7908 | 18.4926 | 20.5992 |
| 40 | 20.7065 | 22.1643 | 24.4331 | 26.5093 | 29.0505 |
| 50 | 27.9907 | 29.7067 | 32.3574 | 34.7642 | 37.6886 |
| 60 | 35.5346 | 37.4848 | 40.4817 | 43.1879 | 46.4589 |
| 70 | 43.2752 | 45.4418 | 48.7576 | 51.7393 | 55.3290 |
| 80 | 51.1720 | 53.5400 | 57.1532 | 60.3915 | 64.2778 |
| 90 | 59.1963 | 61.7541 | 65.6466 | 69.1260 | 73.2912 |
| 100 | 67.3276 | 70.0648 | 74.2219 | 77.9295 | 82.3581 |

| Degrees of freedom | $\chi^2_{0.100}$ | $\chi^2_{0.050}$ | $\chi^2_{0.025}$ | $\chi^2_{0.010}$ | $\chi^2_{0.005}$ |
|---|---|---|---|---|---|
| 1 | 2.70554 | 3.84146 | 5.02389 | 6.63490 | 7.87944 |
| 2 | 4.60517 | 5.99147 | 7.37776 | 9.21034 | 10.5966 |
| 3 | 6.25139 | 7.81473 | 9.34840 | 11.3449 | 12.8381 |
| 4 | 7.77944 | 9.48773 | 11.1433 | 13.2767 | 14.8602 |
| 5 | 9.23635 | 11.0705 | 12.8325 | 15.0863 | 16.7496 |
| 6 | 10.6446 | 12.5916 | 14.4494 | 16.8119 | 18.5476 |
| 7 | 12.0170 | 14.0671 | 16.0128 | 18.4753 | 20.2777 |
| 8 | 13.3616 | 15.5073 | 17.5346 | 20.0902 | 21.9550 |
| 9 | 14.6837 | 16.9190 | 19.0228 | 21.6660 | 23.5893 |
| 10 | 15.9871 | 18.3070 | 20.4831 | 23.2093 | 25.1882 |
| 11 | 17.2750 | 19.6751 | 21.9200 | 24.7250 | 26.7569 |
| 12 | 18.5494 | 21.0261 | 23.3367 | 26.2170 | 28.2995 |
| 13 | 19.8119 | 22.3621 | 24.7356 | 27.6883 | 29.8194 |
| 14 | 21.0642 | 23.6848 | 26.1190 | 29.1413 | 31.3193 |
| 15 | 22.3072 | 24.9958 | 27.4884 | 30.5779 | 32.8013 |
| 16 | 23.5418 | 26.2962 | 28.8454 | 31.9999 | 34.2672 |
| 17 | 24.7690 | 27.5871 | 30.1910 | 33.4087 | 35.7185 |
| 18 | 25.9894 | 28.8693 | 31.5264 | 34.8053 | 37.1564 |
| 19 | 27.2036 | 30.1435 | 32.8523 | 36.1908 | 38.5822 |
| 20 | 28.4120 | 31.4104 | 34.1696 | 37.5662 | 39.9968 |
| 21 | 29.6151 | 32.6705 | 35.4789 | 38.9321 | 41.4010 |
| 22 | 30.8133 | 33.9244 | 36.7807 | 40.2894 | 42.7956 |
| 23 | 32.0069 | 35.1725 | 38.0757 | 41.6384 | 44.1813 |
| 24 | 33.1963 | 36.4151 | 39.3641 | 42.9798 | 45.5585 |
| 25 | 34.3816 | 37.6525 | 40.6465 | 44.3141 | 46.9278 |
| 26 | 35.5631 | 38.8852 | 41.9232 | 45.6417 | 48.2899 |
| 27 | 36.7412 | 40.1133 | 43.1944 | 46.9630 | 49.6449 |
| 28 | 37.9159 | 41.3372 | 44.4607 | 48.2782 | 50.9933 |
| 29 | 39.0875 | 42.5569 | 45.7222 | 49.5879 | 52.3356 |
| 30 | 40.2560 | 43.7729 | 46.9792 | 50.8922 | 53.6720 |
| 40 | 51.8050 | 55.7585 | 59.3417 | 63.6907 | 66.7659 |
| 50 | 63.1671 | 67.5048 | 71.4202 | 76.1539 | 79.4900 |
| 60 | 74.3970 | 79.0819 | 83.2976 | 88.3794 | 91.9517 |
| 70 | 85.5271 | 90.5312 | 95.0231 | 100.425 | 104.215 |
| 80 | 96.5782 | 101.879 | 106.629 | 112.329 | 116.321 |
| 90 | 107.565 | 113.145 | 118.136 | 124.116 | 128.229 |
| 100 | 118.498 | 124.342 | 129.561 | 135.807 | 140.169 |

*Source*: Reproduced from *Biometrika Tables for Statisticians*, Cambridge: Cambridge University Press, 1954, by permission of the Biometrika trustees

(b) $\alpha = 0.01$




| Denominator degrees of freedom ($v_2$) | Numerator degrees of freedom ($v_1$) 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 4,052 | 4,999.5 | 5,403 | 5,625 | 5,764 | 5,859 | 5,928 | 5,982 | 6,022 |
| 2 | 98.50 | 99.00 | 99.17 | 99.25 | 99.30 | 99.33 | 99.36 | 99.37 | 99.39 |
| 3 | 34.12 | 30.82 | 29.46 | 28.71 | 28.24 | 27.91 | 27.67 | 27.49 | 27.35 |
| 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 | 5.18 | 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 |
| 26 | 7.72 | 5.53 | 4.64 | 4.14 | 3.82 | 3.59 | 3.42 | 3.29 | 3.18 |
| 27 | 7.68 | 5.49 | 4.60 | 4.11 | 3.78 | 3.56 | 3.39 | 3.26 | 3.15 |
| 28 | 7.64 | 5.45 | 4.57 | 4.07 | 3.75 | 3.53 | 3.36 | 3.23 | 3.12 |
| 29 | 7.60 | 5.42 | 4.54 | 4.04 | 3.73 | 3.50 | 3.33 | 3.20 | 3.09 |
| 30 | 7.56 | 5.39 | 4.51 | 4.02 | 3.70 | 3.47 | 3.30 | 3.17 | 3.07 |
| 40 | 7.31 | 5.18 | 4.31 | 3.83 | 3.51 | 3.29 | 3.12 | 2.99 | 2.89 |
| 60 | 7.08 | 4.98 | 4.13 | 3.65 | 3.34 | 3.12 | 2.95 | 2.82 | 2.72 |
| 120 | 6.85 | 4.79 | 3.95 | 3.48 | 3.17 | 2.96 | 2.79 | 2.66 | 2.56 |
| ∞ | 6.63 | 4.61 | 3.78 | 3.32 | 3.02 | 2.80 | 2.64 | 2.51 | 2.41 |

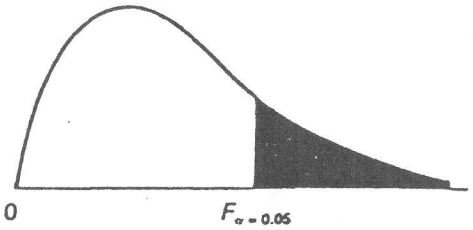
| Numerator degrees of freedom ($v_1$) 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ | Denominator degrees of freedom ($v_2$) |
|---|---|---|---|---|---|---|---|---|---|---|
| 6,056 | 6,106 | 6,157 | 6,209 | 6,235 | 6,261 | 6,287 | 6,313 | 6,339 | 6,366 | 1 |
| 99.40 | 99.42 | 99.43 | 99.45 | 99.46 | 99.47 | 99.47 | 99.48 | 99.49 | 99.50 | 2 |
| 27.23 | 27.05 | 26.87 | 26.69 | 26.60 | 26.50 | 26.41 | 26.32 | 26.22 | 26.13 | 3 |
| 14.55 | 14.37 | 14.20 | 14.02 | 13.93 | 13.84 | 13.75 | 13.65 | 13.56 | 13.46 | 4 |
| 10.05 | 9.89 | 9.72 | 9.55 | 9.47 | 9.38 | 9.29 | 9.20 | 9.11 | 9.02 | 5 |
| 7.87 | 7.72 | 7.56 | 7.40 | 7.31 | 7.23 | 7.14 | 7.06 | 6.97 | 6.88 | 6 |
| 6.62 | 6.47 | 6.31 | 6.16 | 6.07 | 5.99 | 5.91 | 5.82 | 5.74 | 5.65 | 7 |
| 5.81 | 5.67 | 5.52 | 5.36 | 5.28 | 5.20 | 5.12 | 5.03 | 4.95 | 4.86 | 8 |
| 5.26 | 5.11 | 4.96 | 4.81 | 4.73 | 4.65 | 4.57 | 4.48 | 4.40 | 4.31 | 9 |
| 4.85 | 4.71 | 4.56 | 4.41 | 4.33 | 4.25 | 4.17 | 4.08 | 4.00 | 3.91 | 10 |
| 4.54 | 4.40 | 4.25 | 4.10 | 4.02 | 3.94 | 3.86 | 3.78 | 3.69 | 3.60 | 11 |
| 4.30 | 4.16 | 4.01 | 3.86 | 3.78 | 3.70 | 3.62 | 3.54 | 3.45 | 3.36 | 12 |
| 4.10 | 3.96 | 3.82 | 3.66 | 3.59 | 3.51 | 3.43 | 3.34 | 3.25 | 3.17 | 13 |
| 3.94 | 3.80 | 3.66 | 3.51 | 3.43 | 3.35 | 3.27 | 3.18 | 3.09 | 3.00 | 14 |
| 3.80 | 3.67 | 3.52 | 3.37 | 3.29 | 3.21 | 3.13 | 3.05 | 2.96 | 2.87 | 15 |
| 3.69 | 3.55 | 3.41 | 3.26 | 3.18 | 3.10 | 3.02 | 2.93 | 2.84 | 2.75 | 16 |
| 3.59 | 3.46 | 3.31 | 3.16 | 3.08 | 3.00 | 2.92 | 2.83 | 2.75 | 2.65 | 17 |
| 3.51 | 3.37 | 3.23 | 3.08 | 3.00 | 2.92 | 2.84 | 2.75 | 2.66 | 2.57 | 18 |
| 3.43 | 3.30 | 3.15 | 3.00 | 2.92 | 2.84 | 2.76 | 2.67 | 2.58 | 2.49 | 19 |
| 3.37 | 3.23 | 3.09 | 2.94 | 2.86 | 2.78 | 2.69 | 2.61 | 2.52 | 2.42 | 20 |
| 3.31 | 3.17 | 3.03 | 2.88 | 2.80 | 2.72 | 2.64 | 2.55 | 2.46 | 2.36 | 21 |
| 3.26 | 3.12 | 2.98 | 2.83 | 2.75 | 2.67 | 2.58 | 2.50 | 2.40 | 2.31 | 22 |
| 3.21 | 3.07 | 2.93 | 2.78 | 2.70 | 2.62 | 2.54 | 2.45 | 2.35 | 2.26 | 23 |
| 3.17 | 3.03 | 2.89 | 2.74 | 2.66 | 2.58 | 2.49 | 2.40 | 2.31 | 2.21 | 24 |
| 3.13 | 2.99 | 2.85 | 2.70 | 2.62 | 2.54 | 2.45 | 2.36 | 2.27 | 2.17 | 25 |
| 3.09 | 2.96 | 2.81 | 2.66 | 2.58 | 2.50 | 2.42 | 2.33 | 2.23 | 2.13 | 26 |
| 3.06 | 2.93 | 2.78 | 2.63 | 2.55 | 2.47 | 2.38 | 2.29 | 2.20 | 2.10 | 27 |
| 3.03 | 2.90 | 2.75 | 2.60 | 2.52 | 2.44 | 2.35 | 2.26 | 2.17 | 2.06 | 28 |
| 3.00 | 2.87 | 2.73 | 2.57 | 2.49 | 2.41 | 2.33 | 2.23 | 2.14 | 2.03 | 29 |
| 2.98 | 2.84 | 2.70 | 2.55 | 2.47 | 2.39 | 2.30 | 2.21 | 2.11 | 2.01 | 30 |
| 2.80 | 2.66 | 2.52 | 2.37 | 2.29 | 2.20 | 2.11 | 2.02 | 1.92 | 1.80 | 40 |
| 2.63 | 2.50 | 2.35 | 2.20 | 2.12 | 2.03 | 1.94 | 1.84 | 1.73 | 1.60 | 60 |
| 2.47 | 2.34 | 2.19 | 2.03 | 1.95 | 1.86 | 1.76 | 1.66 | 1.53 | 1.38 | 120 |
| 2.32 | 2.18 | 2.04 | 1.88 | 1.79 | 1.70 | 1.59 | 1.47 | 1.32 | 1.00 | ∞ |

*Source*: From M. Merrington and C. M. Thompson (1943) 'Tables of percentage points of the inverted beta ($F$) distribution', *Biometrika* 33, 73–88. Reproduced by permission of the Biometrika trustees

## TABLE A6 *F* DISTRIBUTION

The table shows values of $F$ for various degrees of freedom, $v_1$ and $v_2$, from 1 to $\infty$, and for $\alpha = 0.05$ and $\alpha = 0.01$.

(a) $\alpha = 0.05$



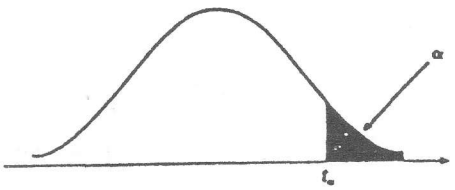

| Denominator degrees of freedom ($v_2$) | Numerator degrees of freedom ($v_1$) 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 161.4 | 199.5 | 215.7 | 224.6 | 230.2 | 234.0 | 236.8 | 238.9 | 240.5 |
| 2 | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.38 |
| 3 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 |
| 13 | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 |
| 14 | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 |
| 16 | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 |
| 17 | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 |
| 18 | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 |
| 19 | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 |
| 20 | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 |
| 21 | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 |
| 22 | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 |
| 23 | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.44 | 2.37 | 2.32 |
| 24 | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.30 |
| 25 | 4.24 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.28 |
| 26 | 4.23 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 |
| 27 | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.31 | 2.25 |
| 28 | 4.20 | 3.34 | 2.95 | 2.71 | 2.56 | 2.45 | 2.36 | 2.29 | 2.24 |
| 29 | 4.18 | 3.33 | 2.93 | 2.70 | 2.55 | 2.43 | 2.35 | 2.28 | 2.22 |
| 30 | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 |
| 40 | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 |
| 60 | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 |
| 120 | 3.92 | 3.07 | 2.68 | 2.45 | 2.29 | 2.17 | 2.09 | 2.02 | 1.96 |
| $\infty$ | 3.84 | 3.00 | 2.60 | 2.37 | 2.21 | 2.10 | 2.01 | 1.94 | 1.88 |

| Denominator degrees of freedom ($v_2$) | Numerator degrees of freedom ($v_1$) 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | $\infty$ | Denominator degrees of freedom ($v_2$) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 241.9 | 243.9 | 245.9 | 248.0 | 249.1 | 250.1 | 251.1 | 252.2 | 253.3 | 254.3 | 1 |
| 2 | 19.40 | 19.41 | 19.43 | 19.45 | 19.45 | 19.46 | 19.47 | 19.48 | 19.49 | 19.50 | 2 |
| 3 | 8.79 | 8.74 | 8.70 | 8.66 | 8.64 | 8.62 | 8.59 | 8.57 | 8.55 | 8.53 | 3 |
| 4 | 5.96 | 5.91 | 5.86 | 5.80 | 5.77 | 5.75 | 5.72 | 5.69 | 5.66 | 5.63 | 4 |
| 5 | 4.74 | 4.68 | 4.62 | 4.56 | 4.53 | 4.50 | 4.46 | 4.43 | 4.40 | 4.36 | 5 |
| 6 | 4.06 | 4.00 | 3.94 | 3.87 | 3.84 | 3.81 | 3.77 | 3.74 | 3.70 | 3.67 | 6 |
| 7 | 3.64 | 3.57 | 3.51 | 3.44 | 3.41 | 3.38 | 3.34 | 3.30 | 3.27 | 3.23 | 7 |
| 8 | 3.35 | 3.28 | 3.22 | 3.15 | 3.12 | 3.08 | 3.04 | 3.01 | 2.97 | 2.93 | 8 |
| 9 | 3.14 | 3.07 | 3.01 | 2.94 | 2.90 | 2.86 | 2.83 | 2.79 | 2.75 | 2.71 | 9 |
| 10 | 2.98 | 2.91 | 2.85 | 2.77 | 2.74 | 2.70 | 2.66 | 2.62 | 2.58 | 2.54 | 10 |
| 11 | 2.85 | 2.79 | 2.72 | 2.65 | 2.61 | 2.57 | 2.53 | 2.49 | 2.45 | 2.40 | 11 |
| 12 | 2.75 | 2.69 | 2.62 | 2.54 | 2.51 | 2.47 | 2.43 | 2.38 | 2.34 | 2.30 | 12 |
| 13 | 2.67 | 2.60 | 2.53 | 2.46 | 2.42 | 2.38 | 2.34 | 2.30 | 2.25 | 2.21 | 13 |
| 14 | 2.60 | 2.53 | 2.46 | 2.39 | 2.35 | 2.31 | 2.27 | 2.22 | 2.18 | 2.13 | 14 |
| 15 | 2.54 | 2.48 | 2.40 | 2.33 | 2.29 | 2.25 | 2.20 | 2.16 | 2.11 | 2.07 | 15 |
| 16 | 2.49 | 2.42 | 2.35 | 2.28 | 2.24 | 2.19 | 2.15 | 2.11 | 2.06 | 2.01 | 16 |
| 17 | 2.45 | 2.38 | 2.31 | 2.23 | 2.19 | 2.15 | 2.10 | 2.06 | 2.01 | 1.96 | 17 |
| 18 | 2.41 | 2.34 | 2.27 | 2.19 | 2.15 | 2.11 | 2.06 | 2.02 | 1.97 | 1.92 | 18 |
| 19 | 2.38 | 2.31 | 2.23 | 2.16 | 2.11 | 2.07 | 2.03 | 1.98 | 1.93 | 1.88 | 19 |
| 20 | 2.35 | 2.28 | 2.20 | 2.12 | 2.08 | 2.04 | 1.99 | 1.95 | 1.90 | 1.84 | 20 |
| 21 | 2.32 | 2.25 | 2.18 | 2.10 | 2.05 | 2.01 | 1.96 | 1.92 | 1.87 | 1.81 | 21 |
| 22 | 2.30 | 2.23 | 2.15 | 2.07 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.78 | 22 |
| 23 | 2.27 | 2.20 | 2.13 | 2.05 | 2.01 | 1.96 | 1.91 | 1.86 | 1.81 | 1.76 | 23 |
| 24 | 2.25 | 2.18 | 2.11 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.79 | 1.73 | 24 |
| 25 | 2.24 | 2.16 | 2.09 | 2.01 | 1.96 | 1.92 | 1.87 | 1.82 | 1.77 | 1.71 | 25 |
| 26 | 2.22 | 2.15 | 2.07 | 1.99 | 1.95 | 1.90 | 1.85 | 1.80 | 1.75 | 1.69 | 26 |
| 27 | 2.20 | 2.13 | 2.06 | 1.97 | 1.93 | 1.88 | 1.84 | 1.79 | 1.73 | 1.67 | 27 |
| 28 | 2.19 | 2.12 | 2.04 | 1.96 | 1.91 | 1.87 | 1.82 | 1.77 | 1.71 | 1.65 | 28 |
| 29 | 2.18 | 2.10 | 2.03 | 1.94 | 1.90 | 1.85 | 1.81 | 1.75 | 1.70 | 1.64 | 29 |
| 30 | 2.16 | 2.09 | 2.01 | 1.93 | 1.89 | 1.84 | 1.79 | 1.74 | 1.68 | 1.62 | 30 |
| 40 | 2.08 | 2.00 | 1.92 | 1.84 | 1.79 | 1.74 | 1.69 | 1.64 | 1.58 | 1.51 | 40 |
| 60 | 1.99 | 1.92 | 1.84 | 1.75 | 1.70 | 1.65 | 1.59 | 1.53 | 1.47 | 1.39 | 60 |
| 120 | 1.91 | 1.83 | 1.75 | 1.66 | 1.61 | 1.55 | 1.50 | 1.43 | 1.35 | 1.25 | 120 |
| $\infty$ | 1.83 | 1.75 | 1.67 | 1.57 | 1.52 | 1.46 | 1.39 | 1.32 | 1.22 | 1.00 | $\infty$ |

(continu

# TABLE A.2 t DISTRIBUTION

Entries in the table give $t_\alpha$ values, where $\alpha$ is the area or probability in the upper tail of the $t$ distribution. For example, with ten degrees of freedom and an area of 0.05 in the upper tail, $t_{0.05} = 1.812$.

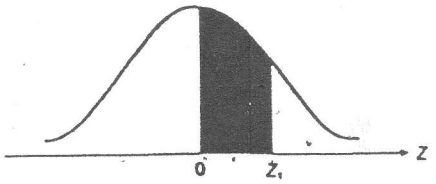



| *Degrees of freedom* | $t_{0.100}$ | $t_{0.050}$ | $t_{0.025}$ | $t_{0.010}$ | $t_{0.005}$ |
|---|---|---|---|---|---|
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 |
| 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 |
| 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.808 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 |
| 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 |
| 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 |
| 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 |
| 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 |
| 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 |
| 30 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 |
| 40 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 |
| 60 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 |
| 120 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 |
| ∞ | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 |

*Source*: Reproduced from *Biometrika Tables for Statisticians*, Cambridge: Cambridge University Press, 1954, by permission of the Biometrika trustees

# TABLE A.1 STANDARD NORMAL DISTRIBUTION

The entries in this table are the probabilities that a random variable having the standard normal distribution assumes a value between 0 and $Z_1$; the probability is represented by the area under the curve (the shaded area). Areas for negative values of z are obtained by symmetry.




| | *Second decimal place in Z* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
| 0.0 | 0.0000 | 0.0040 | 0.0080 | 0.0120 | 0.0160 | 0.0199 | 0.0239 | 0.0279 | 0.0319 | 0.0359 |
| 0.1 | 0.0398 | 0.0438 | 0.0478 | 0.0517 | 0.0557 | 0.0596 | 0.0636 | 0.0675 | 0.0714 | 0.0753 |
| 0.2 | 0.0793 | 0.0832 | 0.0871 | 0.0910 | 0.0948 | 0.0987 | 0.1026 | 0.1064 | 0.1103 | 0.1141 |
| 0.3 | 0.1179 | 0.1217 | 0.1255 | 0.1293 | 0.1331 | 0.1368 | 0.1406 | 0.1443 | 0.1480 | 0.1517 |
| 0.4 | 0.1554 | 0.1591 | 0.1628 | 0.1664 | 0.1700 | 0.1736 | 0.1772 | 0.1808 | 0.1844 | 0.1879 |
| 0.5 | 0.1915 | 0.1950 | 0.1985 | 0.2019 | 0.2054 | 0.2088 | 0.2123 | 0.2157 | 0.2190 | 0.2224 |
| 0.6 | 0.2257 | 0.2291 | 0.2324 | 0.2357 | 0.2389 | 0.2422 | 0.2454 | 0.2486 | 0.2517 | 0.2549 |
| 0.7 | 0.2580 | 0.2611 | 0.2642 | 0.2673 | 0.2704 | 0.2734 | 0.2764 | 0.2794 | 0.2823 | 0.2852 |
| 0.8 | 0.2881 | 0.2910 | 0.2939 | 0.2967 | 0.2995 | 0.3023 | 0.3051 | 0.3078 | 0.3106 | 0.3133 |
| 0.9 | 0.3159 | 0.3186 | 0.3212 | 0.3238 | 0.3264 | 0.3289 | 0.3315 | 0.3340 | 0.3365 | 0.3389 |
| 1.0 | 0.3413 | 0.3438 | 0.3461 | 0.3485 | 0.3508 | 0.3531 | 0.3554 | 0.3577 | 0.3599 | 0.3621 |
| 1.1 | 0.3643 | 0.3665 | 0.3686 | 0.3708 | 0.3729 | 0.3749 | 0.3770 | 0.3790 | 0.3810 | 0.3830 |
| 1.2 | 0.3849 | 0.3869 | 0.3888 | 0.3907 | 0.3925 | 0.3944 | 0.3962 | 0.3980 | 0.3997 | 0.4015 |
| 1.3 | 0.4032 | 0.4049 | 0.4066 | 0.4082 | 0.4099 | 0.4115 | 0.4131 | 0.4147 | 0.4162 | 0.4177 |
| 1.4 | 0.4192 | 0.4207 | 0.4222 | 0.4236 | 0.4251 | 0.4265 | 0.4279 | 0.4292 | 0.4306 | 0.4319 |
| 1.5 | 0.4332 | 0.4345 | 0.4357 | 0.4370 | 0.4382 | 0.4394 | 0.4406 | 0.4418 | 0.4429 | 0.4441 |
| 1.6 | 0.4452 | 0.4463 | 0.4474 | 0.4484 | 0.4495 | 0.4505 | 0.4515 | 0.4525 | 0.4535 | 0.4545 |
| 1.7 | 0.4554 | 0.4564 | 0.4573 | 0.4582 | 0.4591 | 0.4599 | 0.4608 | 0.4616 | 0.4625 | 0.4633 |
| 1.8 | 0.4641 | 0.4649 | 0.4656 | 0.4664 | 0.4671 | 0.4678 | 0.4686 | 0.4693 | 0.4699 | 0.4706 |
| 1.9 | 0.4713 | 0.4719 | 0.4726 | 0.4732 | 0.4738 | 0.4744 | 0.4750 | 0.4756 | 0.4761 | 0.4767 |
| 2.0 | 0.4772 | 0.4778 | 0.4783 | 0.4788 | 0.4793 | 0.4796 | 0.4803 | 0.4808 | 0.4812 | 0.4817 |
| 2.1 | 0.4821 | 0.4826 | 0.4830 | 0.4834 | 0.4838 | 0.4842 | 0.4846 | 0.4850 | 0.4854 | 0.4857 |
| 2.2 | 0.4861 | 0.4864 | 0.4868 | 0.4871 | 0.4875 | 0.4878 | 0.4881 | 0.4884 | 0.4887 | 0.4890 |
| 2.3 | 0.4893 | 0.4896 | 0.4898 | 0.4901 | 0.4904 | 0.4906 | 0.4909 | 0.4911 | 0.4913 | 0.4916 |
| 2.4 | 0.4918 | 0.4920 | 0.4922 | 0.4925 | 0.4927 | 0.4929 | 0.4931 | 0.4932 | 0.4934 | 0.4936 |
| 2.5 | 0.4938 | 0.4940 | 0.4941 | 0.4943 | 0.4945 | 0.4946 | 0.4948 | 0.4949 | 0.4951 | 0.4952 |
| 2.6 | 0.4953 | 0.4955 | 0.4956 | 0.4957 | 0.4959 | 0.4960 | 0.4961 | 0.4962 | 0.4963 | 0.4974 |
| 2.7 | 0.4965 | 0.4966 | 0.4967 | 0.4968 | 0.4969 | 0.4970 | 0.4971 | 0.4972 | 0.4973 | 0.4974 |
| 2.8 | 0.4974 | 0.4975 | 0.4976 | 0.4977 | 0.4977 | 0.4978 | 0.4979 | 0.4979 | 0.4980 | 0.4981 |
| 2.9 | 0.4981 | 0.4982 | 0.4982 | 0.4983 | 0.4984 | 0.4984 | 0.4985 | 0.4985 | 0.4986 | 0.4986 |
| 3.0 | 0.4987 | 0.4987 | 0.4987 | 0.4988 | 0.4988 | 0.4989 | 0.4989 | 0.4989 | 0.4990 | 0.4990 |
| 3.1 | 0.4990 | 0.4991 | 0.4991 | 0.4991 | 0.4992 | 0.4992 | 0.4992 | 0.4992 | 0.4993 | 0.4993 |
| 3.2 | 0.4993 | 0.4993 | 0.4994 | 0.4994 | 0.4994 | 0.4994 | 0.4994 | 0.4995 | 0.4995 | 0.4995 |
| 3.3 | 0.4995 | 0.4995 | 0.4995 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4997 |
| 3.4 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4998 |
| 3.5 | 0.4998 | | | | | | | | | |
| 4.0 | 0.49997 | | | | | | | | | |
| 4.5 | 0.499997 | | | | | | | | | |
| 5.0 | 0.4999997 | | | | | | | | | |

Source: Reprinted with permission from *Standard Mathematical Tables*, 15th edn. © CRC Press Inc., Boca Raton, FL

ISBN 978-979-98207-1-6