SUBJECT LEARNING

STATISTICS FOR EDUCATION

USED FOR S1 BACHELOR PROGRAM
ALL DEPARTMENT

Prof. Drs. YL Sukestiyarno, M. S., Ph. D.

2018

Hak Penerbitan pada UNNES PRESS
• Jl. Kelud raya No. 2 Semarang 50232
Telp/Fax.(024)8415032

STATISTICS FOR EDUCATION USED FOR S1 BACHELOR PROGRAM ALL DEPARTMENT: SUBJECT LEARNING

Penulis
Prof. Drs. YL Sukestiyarno, M. S., Ph. D.

Editor:
Putut Marwoto

Desain Cover:
Hasan Septia Saputra

STATISTICS FOR EDUCATION USED FOR S1 BACHELOR PROGRAM ALL DEPARTMENT: SUBJECT LEARNING/ Prof. Drs. YL Sukestiyarno, M. S., Ph. D.; Cet. I; UNNES Press, 2018
v+ Halaman, 14,7 x 21 cm

No Keanggotaan APPTI Penerbit UNNES Press 003.045.1.05.2018

ISBN: 978-602-285-173-8

Competencies

CHAPTER I
BASIC CONCEPTS

A. Description and Basic Competence

The description consists of the definition of statistics, statistical data, variables, and statistics for society.

The main instructional objectives are after the learning process students are able to:

1. define what statistics is
2. compare statistics and statistic data
3. apply statistics to daily life
4. differentiate variables and data
5. differentiate types of data
6. convert the data scale in the right rules
7. use data in appropriate analyses

B. Definition of statistics

The Word statistics is derived from Latin word "**Status**" or the Italian word "**Statista**", which means "**Political State**" or a Government. Shakespeare used the word Statist in his drama Hamlet (1602). In the past, statistics was used by rulers. The application of statistics was very limited but rulers and kings needed information about lands, agriculture, commerce, population of their states to assess their military potential, their wealth, taxation and other aspects of government.

Gottfried Achenwall used the word *statistik* at a German University in 1749 which had a meaning of "political science of different countries". In 1771 W. Hooper (Englishman) used the word statistics in his translation of Elements of Universal Erudition by Baron B.F Bieford. In his book, statistics was defined as the science which taught what the political arrangement of all the modern states of the known world is. There is a big gap between the old statistics and the modern one. However; old statistics is also used as a part of the present statistics.

For the last few centuries, statistic remained a part of mathematics as the original work done by mathematicians like Pascal (1623-1662), Bernaulli (1654 – 1705), De Moivre (1667-1754), Laplace (1749-1827), Gauss (1777-1855), etc. Until the early nineteenth century, statistics was mainly concerned to official statistics needed for the collection of information on revenue, population and area of land under cultivation etc. of a state or kingdom.

The word 'statistics' was used for the first time in Elements of Universal Erudition by Baron JF (Agarwal, 2006). Here statistics was defined as the science of political arrangement for all modern states. However; there is still other definition according to some experts such as:

1. Noether (1971) wrote completely that statistics was the science of the collection, classification, and measured evaluation of facts as a basis of inference. It is a body of techniques for acquiring accurate knowledge from incomplete information; a scientific system for the collection, organization, analysis, interpretation and a presentation of information which can be stated in numerical form.
2. Moore (1989) defined statistics as the science of collecting, organizing, and interpreting numerical facts.

Based on the above definitions, it can be drawn the functions of statistics in general which consists of:

1. collecting data
2. tabulating data
3. analyzing data
4. interpreting the results.

These four functions will be described adequately later. Further, their application and utility will obviously be clear from the discussion of the subject matter given in the body of this book.

C. Grouping Statistics

Statistics can be divided into two main subjects, namely: descriptive statistics and inferential statistics. The first concerns sequencing, grouping, sorting, and presenting data, in which one can find the relevant information. The second on the other hand concerns predicting and looking for relation between data. In processing the data, descriptive statistics is sometimes called qualitative statistics, while inferential statistics is called quantitative statistics.

There are some examples of descriptive data, such as when someone wants to describe:

1. Monalisa picture.
2. Fighting between two schools.
3. Tsunami Disaster in Aceh.
4. The situation when the National Examination is carried out, etc.

In inferential statistics, to make decision and predict the relation among events can be concluded right or wrong through examining sample data. The result can be used to estimate, conclude, predict, or generate broader result.

Some examples of inferential statistics data such as when someone wants to:

1. Compare the learning result of senior high school students in cities and countries.
2. Know the influence of student activities toward student's achievement.
3. Find the relation between the length of babies and their weight.

4. Find the difference of national examination result in several years.

D. Collection of Data

There is a lot of information in the media (TV, radio, internets, magazines, etc.) about unemployment, football match results, stock exchange, flight schedules, death report caused by smoking, etc. Those examples are related to statistic data. In this case, the data is already involved.

Information has to be collected from certain individuals directly or indirectly. Such a technique is known as survey method. This is commonly used in social sciences i.e., the problems relating to sociology, political science, psychology and various economic studies. In surveys, the required information is supplied by the individual under study or is based on measurements of certain units. Another way of collecting data is by experimentation i.e. an actual experiment is conducted on certain individuals or units about which the inference is to be drawn. Such experimental studies are common in agriculture, biology, chemistry, education etc.

Basic Definition

Statistic unit

A statistic unit is an individual object or human being which is researched, surveyed or taken as the data. Their characteristics must be identified to get more information about the problem being searched.

Variable

A variable is a measurable or countable or observable characteristic of objects which has various values. The value of variables can be divided into two groups namely discrete and continues. The example of variables are: students' height in class A, temperature in each lectures room of a university, motivation in learning process of student Biology department, gender in class A, etc. When a learning result is defined in class A of 3rd semester of English Department, it is called a variable, but if the average of a learning result is defined from the same class, it is not a variable. It is because the average has only one value or there is no variation data value.

Data

An expression of variable in values is called data. The existing data can be in a form of quantitative (numeric) or qualitative (attribute like good, defect, strong, less strong, weak, and so on). Quantitative data can be gained from the result of measuring and counting. On the other hand qualitative data can be gained from observation.

One example is the research entitled 'The orientation of students learning result at SMP 5 Semarang grade VII in algebra subject by using peer tutor strategy'. Based on the results of this research, it can be defined the unit statistic i.e. the students of SMP 5 Semarang grade VII and the variable can be in the form of: learning result which consists of students' creativity, skill or achievement.

Types of Data

This topic discusses a classification system that is often used to describe the measurement of concepts or variables that are used in social sciences and behavioural research. This classification system categorizes the variables as being measured on either nominal, ordinal, cardinal, interval, or ratio scale. After introducing the classification system and providing examples of variables which are typically measured on each type of scale, it is noted the implications of these measurement scales for the analysis of data. Specifically, the statistical tests discussions are most appropriate for data measured on each type of scale. Finally, it will be briefly considered some of the limits and criticisms of this classification system.

Quantitative and qualitative data can be grouped as follows:

1. Nominal scale

The data is obtained from an **observation** for which the result is qualitative data. A qualitative data can be quantified in numeric and be grouped in discrete data. The numeric symbol used does not represent sequence value; it is a symbol of categorical characteristics. For example in variable kind of gender, “boy” is represented with 1 and “girl” is 2. In this case 1 does not mean that it is lower than 2. Within a nominal measurement scale, there is no relative ordering of the categories -- the assignment of numeric scores to each category (male, female) is purely arbitrary. Another examples are the kind of the religions (1=Islam, 2=Christian, 3=Catholic, 4=Hindu, 5=Buddha), kind of marital status (1=single, 2=married, 3=divorced), etc.

2. Ordinal scale

It is the same as nominal scale in which the data is obtained from an **observation;** the numeric symbol however uses sequence value. For example in variable motivation to learn mathematics, it consists of 1=very bad, 2=bad, 3=enough, 4=good, 5=very good. In this case each number has sequence value, 2 is bigger than 1 and so on.

Although ordinal variables provide information concerning the relative position of participants or observations in this research study, ordinal variables do not tell anything about the absolute magnitude of the difference between 1st and 2nd or between 2nd and 3rd. It is known 1st is before 2nd, and 2nd is before 3rd, but how close 3rd was to 2nd or how close 2nd was to 1st is unpredictable. The other example of variable skill in learning process is 1=very unskilful, 2=unskilful, 3=quite skilful, 4=skilful, 5=very skilful. The other example of ordinal scale is student's creativity in learning, student's attitude in learning, and so on.

3. Cardinal scale

Data of this type is taken from **counting** a variable. The data is in the form of discrete value which can be presented in cardinal number. For example the number of chairs in each room of a school, the number of computer set in each classroom in a school, and so on.

4. Interval scale

This type of data is taken from **measuring** a variable. The measuring data is assumed in the group continue data. Quantitative attributes are all measurable on interval scales, as any difference between the levels of an attribute can be multiplied by any real number to exceed or equal another difference. A highly familiar example of interval scale measurement is temperature with the Celsius scale. In this particular scale, the unit of measurement is 1/100 of the difference between the melting temperature and the boiling temperature of water at atmospheric pressure. The "zero point" on an interval scale is arbitrary; and negative values can be used. In this case the measuring variable does not have absolute null. Another example is human weight in space.

5. Ratio scale

Ratio scale is almost the same as interval scale, for which the data is found by **measuring**. The difference is that a ratio scale has an absolute null. For example a mass of something, when this thing has zero value, it means there is no substance of this thing. In interval scales, there is no absolute zero point. Therefore, it is inappropriate to express interval level measurements as ratios; it would not be appropriate to say that 60 degrees is twice as hot as 30 degrees. The last type of measurement scales, ratio scales, do have a fixed zero point. Not only numbers or units on the scale are equal over all levels of the scale, but also a meaningful zero point which allows for the interpretation of ratio comparisons.

Time is an example of a ratio measurement scale. Not only the statement about the difference between three hours and five hours are the same as the difference between eight hours and ten hours (equal intervals), but it also can be said that ten hours is twice as long as five hours (a ratio comparison). Another example is variable: height (when something has zero height it means there is nothing), volume, speed, acceleration, and so on.

E. Application and Problem Solving

1. ***Which of the grouping data scales is appropriate to work in central tendency?***

The central tendency of a nominal scale is given by its mode; neither the mean nor the median can be defined. An ordinal attribute can be represented by its mode or its median, but the mean cannot be defined. But the cardinal, interval, and ratio scale can be represented by its mode, its median, or its arithmetic mean.

2. How can we find the data of variables in its characteristic scales?

It is easy to find the data of variables through counting and measuring. The characteristic of the variables can be counted or measured directly with appropriate equipments. For example: data of height and temperature are taken in metres and thermometer units respectively; data amount of people each class, sum of computers in each room can be counted directly. We have a little bit problems to convert the qualitative data from observable variables to symbolize in quantitative data. For example: To convert the nominal scale data of kind of gender is quite clear 1 for “boy” and 2 for “girl”. But converting an ordinal scale to quantitative score, it is not easy. Normally people use scoring for observable variable motivation like: 1=very bad, 2=bad, 3=enough, 4=good, 5=very good. We directly come to the question, what are the differences between 1 and 2, or between 2 and 3 and so on. In this case the observer must make a certain rule about scoring before he/she does the observation.

3. ***Many researchers work with inappropriate analyses***

In application, the qualitative or descriptive or non parametric analyses work appropriately in discrete data i.e. nominal, ordinal and cardinal scale. On the other side, the quantitative or parametric or inferential analyses work appropriately with continue data i.e. interval and ratio scale. Many researchers consider Likert-scale data (Questionnaire with the attribute 1 to 5) to possess ordinal qualities. However, leading research studies, for example in the education area, obtain measures such as means and standard deviations from Likert-scale data. They use the data for inappropriate analyses.

4. ***Is it possible to interpret the data which is taken through questionnaire or observation as an interval scale?***

There are several characteristic of variables that can be grouped in ordinal scale or in interval scale. It depends on the condition of the characteristic data. For example: the variable of students’ learning activity or motivation of student who joint the mathematics lecture. These data can be taken through observation or questionnaire. Normally, they belong to ordinal scale. But on the other case it can also be interpreted that these data are found with measurement. The conversion scoring scales are 1=very bad, 2=bad, 3=enough, 4=good, 5=very good in which each score has been given in rounding score neighborhood. For example, score 4 comes from the rounding score neighborhood 4. So it can be assumed that these data are continue. When we interpret observed variable like motivation as interval scale and analyze them in inferential statistics, a variable must be produced from more indicators characteristic observation. Inferential statistics need more variety scores to analyze data interval or ratio scales.

Note: we cannot interpret the nominal and cardinal scales into interval or ratio scales.

5. How can we think about transformation scale data to each other with formulas?

Some times it can happen that data scale is exchangeable one to another. Changing from the ratio or interval scale to ordinal scale is possible. For example the ratio scale variable learning achievement is converted in ordinal scale through the rules score that score less than 50 is bad (transformed by 1), score 50 to 70 is normal (transformed by 2) and score more than 70 is good (transform by 3). But we cannot do it backward. Transform the cardinal scale to interval or ratio scale is impossible. Even though there are special formulas, but there are not enough good theories to support the formulas.

5. ***Give example of several characteristic indicators of an observable variable, and how can we determine the convert scoring?***

We take the variable "learning activity and learning achievement in geometry by using discussing strategy in high school" as an example.

Characteristics indicators of learning activity for each student is given as follow,

- student is active in making question
- student is active in working together with friends
- student is active in discussing
- student is active in making conclusion
- and so on

We choose one of the indicators as example "*student is active in making question*". Here are the scoring rules for each student individually observed (these are exchangeable for other argument) :

Score 1: student has no idea to make a question

Score 2: student thinks in mind to make a question

Score 3: student can make only one easy question

Score 4: student can make one question in good question

Score 5: student can make more than one questions.

Example "*student is active in working together with friends*" :

Score 1: student is in group without working

Score 2: student is in group and give little reaction in working together

Score 3: student comes to question or answer in working together

Score 4: student is active and dominate others or make the others active

Score 5: student can organize the others well in working together

Below are examples of indicators for measuring variable "learning achievement in geometry". After the learning process the students are able to:

- show the formula of the area of circle
- compute the area of circle on the given radius
- compute the circumference of a circle on the given area

It is not difficult to make more questions to evaluate each indicator.

6. ***If we have to make research on the following topics:***

   a. The existence a library in a university.

   b. Comparison of the examination result of English between private and state school.

   c. Relation between the learning activity and achievement in teaching with open ended strategy.

   Mention the possibility research variables for those topics!

   The possible variables of each topic can be given as follows:

   a. The amount of each kind of grouping books, the sum of defect books in each room, the amount workers in each department, the time of student comes to library, etc.

   b. The learning achievement in English, the amount of success and failure of each school.

   c. The learning activity and learning achievement.

F. Exercise

1. Give the definition of statistics based on your appropriate consideration!
2. Explain clearly the functions and limitations of statistics!
3. Mention the various methods of collecting data! Describe each of them adequately!
4. Differentiate between the following data:

   a. nominal scale and ordinal scale

   b. interval scale and ratio scale

   c. discrete data and continue data.
5. Is it possible to transform scale data in question 4 for each other? Explain your opinion briefly?
6. What is the meaning “the conclusion analysis is unsure”?

7. You will lead a research to investigate the following interests:

   a. Population growth in your city in mid-2009.
   b. Number of students in each department in a university whose age is more than 25 years old.
   c. 4th semester Student's ability to use computer
   d. Observing the blood type of students in Language Faculty.

   Explain the upper observation by using appropriate statistics analysis and give example of variables on each observation and its classification of their data.

8. Determine the classification of the variables (continue or discrete) as follow,

   a. Value of the Biology test,
   b. Salary of each family,
   c. Aircraft speed,
   d. number of the books in library
   e. kind of playing card

10. If we have the following data variable:

   a. rate of interest
   b. length of the student' hair
   c. muscle strength of a sportsman
   d. the number of children in families
   e. the number of car production each year
   f. learning achievement of English
   g. kind of hobby each student

   Determine the kind of scale these data.

11. An observation with Questionnaire and change with Likert rule, can we obtain mean and standard deviation?

12. Give example of at least 3 characteristic indicators of observable variable of skill learning process.

13. Define the convert scoring of the indicators in question 12.

CHAPTER II
REPRESENTATION OF DATA

A. Description and Basic Competence

The description consists of the kind of representing data, how to make representing data, interpretation of representing data, application of representing data.

The main instructional objectives are after the learning process students are able to:

1. construct a table matrix row and column
2. construct a pie chart
3. construct a bar chart
4. construct a line diagram
5. construct a pictogram
6. make a steam and leaf display diagram
7. plot data in a plot diagram
8. choose an appropriate diagram to present data

B. **Kind of Presenting Data**

When the data is presented to the reader or an audience, whether it is on a news report or in a technical journal report, they are usually presented in the form of a diagram. Sometimes, a diagram is used simply to make the data more eye-catching; a list of numbers just does not grab any attention, the use of a diagram serves some further purpose. A diagram can be used to *summarize* large sets of data, or to focus attention on some *aspect* of the accuracy data, or to display a *trend* in the data over time. A good diagram enables the viewer to grasp in a single glance the relevant features of the data, features that would not be obvious from the raw numbers themselves.

The information which is collected through an enquiry or experiment may be presented in the form of tables or diagrams. Some time people show the data in unsuitable tabulation. The presenter should think that the reader must be able to understand their tabulating fast and clearly. The best choice of tabulator must be in accordance with the goal of presenting. It may come to the accuracy of data or comparing subject data or trend of data or the interesting presentation of data. Kind of presenting data can be grouped as follows:

1. table matrix row and column,
2. pie diagram,
3. bar diagram,
4. pictogram,
5. line diagram,
6. cartogram,
7. scatter plot diagram,
8. stamp and leaf diagram.

C. Table matrix row and column

A table matrix row and column is a systematic arrangement of data in rows and columns, which is easy to understand and makes data, fit for further analysis and draw conclusions.

Example 2.1: Table 2.1 shows the contrasts of Java population with the number of people living elsewhere in Indonesia (taken from Indonesia's 1990 census results)

Table 2.1: Population and Area Indonesia in 1990

Area	**Populations (thousands)**	**Total Area (m2)**
Java	118,300	2,286

Sumatra	41,400	183,025
Sulawesi	13,800	72,979
Kalimantan	10,400	208,124
Irian Jaya	1,000	162,946
All Others	9,800	60,622
Total	194,700	689,982

The next example shows that the data in big differences value can be also presented in table matrix. Table 2.2 shows the data prediction of the world population in the future.

People make a systematic arrangement of data in rows and columns which is aimed for the readers to understand easily and know exactly what number may be in decimal (score). It is of vital importance to choose the table with data accuracy. If we compare it with the other tabular presentation, the table is the best one on its accuracy.

Table 2.2: World Population Prospect

Country	**1950**	**2009**	**2015**	**2025**	**2050**
Australia	8,218,999	21,292,893	22,606,591	24,702,504	28,724,025
Brunei D	48,001	399,687	443,121	513,177	657,508
China	544,950,886	1,345,750,973	1,395,998,248	1,453,140,188	1,417,044,807
Germany	68,376,002	82,166,671	81,345,502	79,257,964	70,503,986
India	371,856,500	1,198,003,272	1,294,192,043	1,431,271,761	1,613,799,950
Indonesia	77,151,870	229,964,723	244,191,496	263,287,137	288,110,442
Israel	1,257,971	7,169,556	7,823,469	8,769,480	10,649,053
Malaysia	6,109,907	27,467,837	30,040,849	33,769,706	39,664,352
Palestine	1,004,800	4,277,360	5,090,124	6,553,091	10,264,625
Russia	102,702,461	140,873,647	137,983,426	132,345,350	116,097,030
Saudi Arabia	3,201,369	25,720,605	28,932,524	34,176,193	43,658,157
USA	157,813,040	314,658,780	332,334,019	358,734,625	403,931,520

Source: World Population Prospects: The 2008 Revision GeoHive

D. Pie Chart

A pie chart is a circle divided into component sectors according to the break-up of components given in percentage. If each component is represented by a separate circle, large figures would need large circles. But the percentages remove this difficulty. Moreover, in a pie diagram only one circle of any size can represent all the components. In the circle of a desired size, a radius, generally a horizontal line, is drawn and the calculated angles for various components are constructed one after another with the help of a protractor. Each sector is shaded differently by lines, dots or with different colours to look unique. A pie chart is a good to represent the component break up of a thing or commodity.

To construct a pie chart (sector diagram), we draw a circle with radius (square root of the total). The total angle of the circle is 360^0. The angles of each component are calculated by the formula

$$\text{Angle of sector} = \frac{componentPart}{Total} \times 360^0 .$$

People make a systematic arrangement of data in pie diagram which is aimed for the reader to understand easily and fast to compare each case. It is of vital importance to choose the pie diagram with comparing each component of data.

Example 2.2: We use the data from Table 2.1 and compare the population in several islands in Indonesia, to find the biggest or the fewest one. We can look in the circle diagram.

Percentage of population in different sectors is shown in column ii which is calculated as:

For sector 1, Percentage = $\frac{118300}{194700} x100\% = 60.8\%$

For sector 2, Percentage $= \frac{41400}{194700} x100\% = 21.3\%$

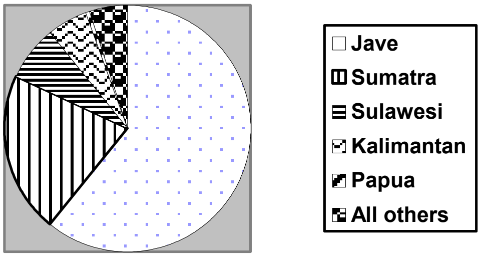
Similarly, other percentages are calculated. Angles equivalent to percentages are calculated as,

For sector 1, angle = $\frac{118300}{194700} x360^{o} = 218.7^{o}$

For sector 2, angle = $\frac{41400}{194700} x360^{0} = 76.5^{o}$.

The angles for other sectors have been calculated similarly. The Pie chart is drawn according to the method given in theory and displayed in figure 2.1.

Figure 2.1: Indonesian Population in 1990

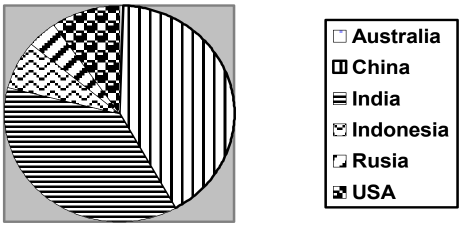


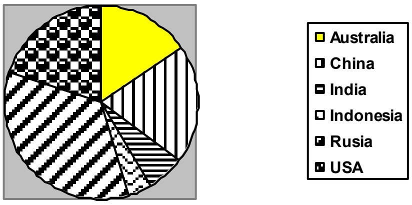
The data Table 2.3 take from part of Table 2.2. We will make the data Table 2.3 in pie chart. To compute the angles of each sector are similarly with the upper calculation. Next, Figure 2.2a and 2.2b show a comparison of the population and the area of 6 big countries in the world.

Table 2.3: The Population prospect of 6 big countries in 2009

Country	Population	Area sq.km
Australia	21,292,893	7,686,850
China	1,345,750,973	9,596,960
India	1,198,003,272	3,287,590
Indonesia	229,964,723	1,919,440
Russia	140,873,647	17,075,200
USA	314,658,780	9,826,630

Figure 2.2a: World population 6 countries in 2009 Figure 2.2b:World area 6 countries 2009





A pie chart can present only one subject. We cannot make pie chart more than one variable in each pie chart.

E. Bar Diagram

The bar diagram shows an aggregate value, whereas the component bar diagram gives the break up in parts which constitutes the aggregate each cases. In these types of diagrams, a bar is further sub divided into parts in proportional to the size of the sub-divisions. These sub divided rectangles are shaded differently by lines, dots, and colours etc. The goal to choose a bar diagram is a little bit similar with pie diagram, namely compare each cases in data.

A simple bar diagram is used to represent data involving only one variable classified on spatial, quantitative or temporal basis. In simple bar diagram, we make bars of equal width but variable length, i.e. the magnitude of a quantity is represented by the height or length of the bars. Following steps are undertaken in drawing a simple bar diagram:

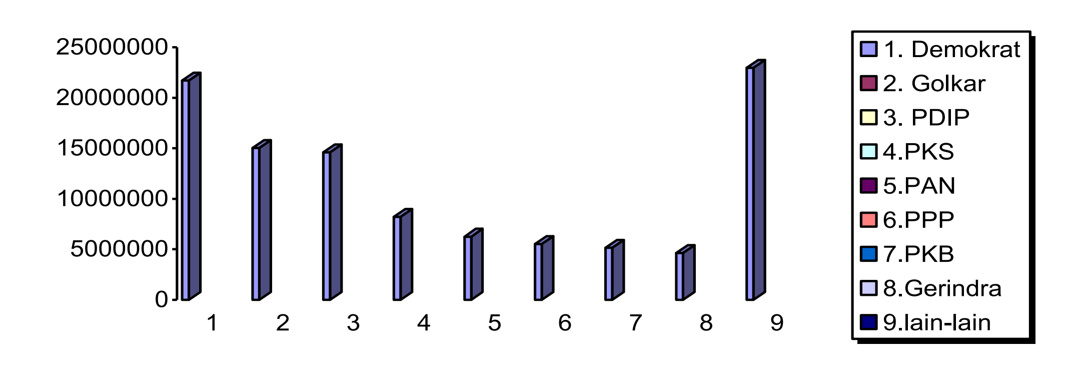
Example 2.3: Figure 2.3 shows the data from Table 2.4 about election in Indonesia 2009.

Table 2.4: 2009 Election results in Indonesia

No	Partai (Parties)	Suara (votes)	%
1	P Demokrat	21,703,137	20.8
2	Golkar	15,037,757	14.4
3	PDI-P	14,600,091	14
4	PKS	8,206,955	7.9
5	PAN	6,254,580	6
6	PPP	5,533,214	5.3
7	PKB	5,146,122	4.9
8	Gerindra	4,646,406	4.5
9	Others	22,971,523	22.2

Source: www.bps.go.id

Figure 2.3 : 2009 Election result in Indonesia



Multiple Bar Charts

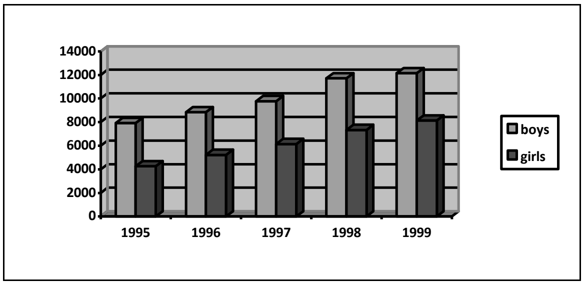
By multiple bars diagram, two or more sets of inter-related data are represented (multiple bar diagram facilities comparison between more than one phenomenon). The technique of simple bar chart is used to draw this diagram but the difference is that we use different shades, colours, or dots to distinguish between different phenomena. We use to draw multiple bar charts if the total of different phenomena is meaningless.

Example 2.4: Figure 2.4 presents the data Table 2.5 that presented about the students girls and boys in a district Indonesia for the years 1995 to 1999.

Table 2.5: Students in a district Indonesia 1995 to 1999

Years	**boys**	**girls**
1995	7930	4260
1996	8850	5225
1997	9780	6150
1998	11720	7340
1999	12150	8145

Figure 2.4: Students in a district Indonesia 1995-1999



F. Line Diagram

In this type of diagram, we have two variables under consideration. A variable is taken along X-axis and the other along Y-axis. The variety values are suitably scaled along the axes and all distances are measured from the origin. If the smallest value in the bivariate data or frequency distribution is at a distance from zero, the origin is shifted suitably to a value other than zero. The independent variable should be taken on X-axis and the dependent variable on Y-axis. The points are plotted and joined by line segments in order. These diagrams depict the trend of variability occurring in the data. Sometimes, two or more diagrams are drawn on the same diagram, paper taking the same scale so that the plotted diagrams are comparable.

People make systematic arrangement of data in line diagram with the goal that the reader can understand easy where the data make a trend. It is of vital importance to choose the line diagram with trend data.

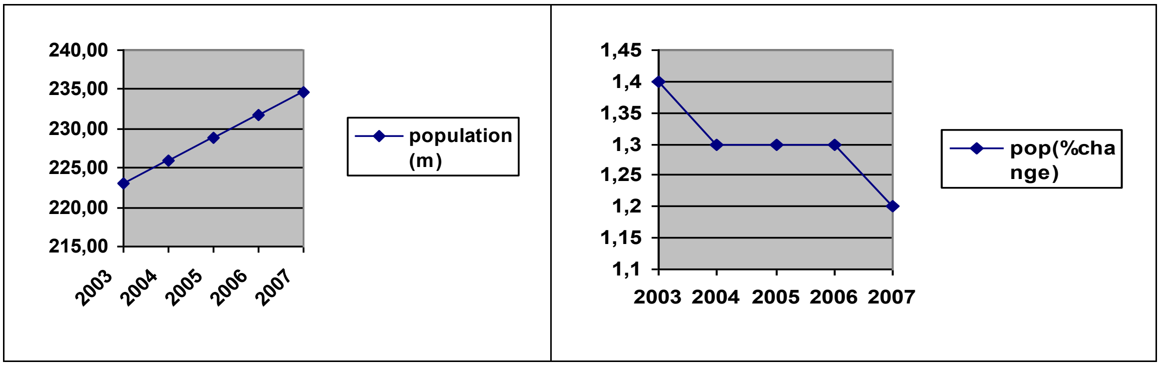
Example 2.5: We have information about population Indonesia is the fourth most populous country in the world after China, India and the US, with a population of around 235 million people. Indonesian population growth in 2003-07 was 1.3% per year. Table 2.6 shows the population in Indonesia 2003-07. Now we will make it in line diagram. The trend data can be looked at Figure 2.5.

Table 2.6: Population in Indonesia 2003-07

Years	2003	2004	2005	2006	2007
Population (million)	223,1	226,0	228,9	231,8	234,7
Population % change	1.4	1.3	1.3	1.3	1.2

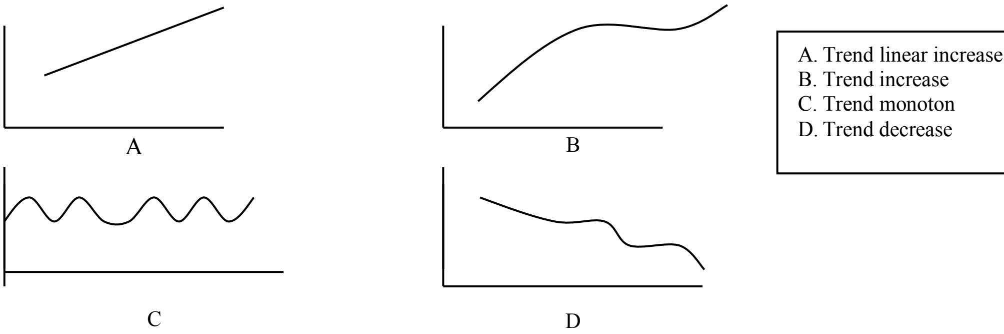
Source: US Census Bureau

Figure 2.5: Trend data population Indonesia 2003-2007



Through the line diagram, we can follow the trend of the data. That is an increase, decrease, monotone, fluctuate, or irregular fluctuation. Here are several diagrams in trend.

Figure 2.6: Trend Data



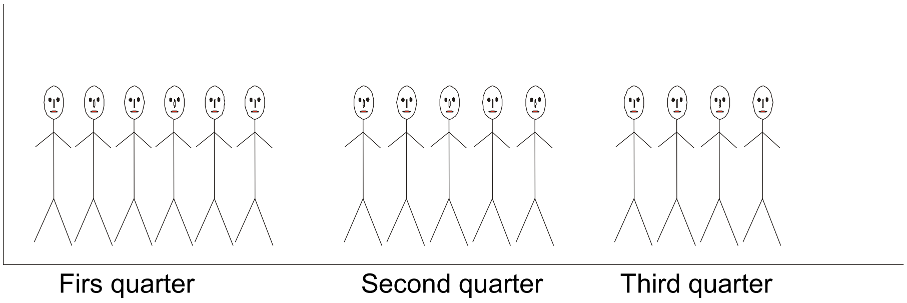
G. Pictograms

In pictorial diagrams, the magnitude of certain things is shown by their pictures. These diagrams are not abstract like the line graphs or bar diagrams. For instance, if we want to display the amount of cars production in Indonesia in 2000, and the picture of a car is to represent of 100 cars, two pictures of cars given for 200 cars and three pictures of 300 etc. Thus, the picture clearly shows the rise in the number of cars in that time. But pictogram is not frequently used as they can not be very accurate. The presenting data with pictogram reach to the interest for user to read information. The diagram presenters do not think about the accuracy of the amount.

Example2.6:

In 1990 the number of unemployment in Semarang, the first quarter is 30 people, the second quarter is 25 people, and the third quarter is 20 people. Pictogram Figure 2.7 shows that one person represents 5 people.

Figure 2.7: The Unemployment in Semarang 1990



Problems arise if the reference from the data will be described that one picture of 10 people, it means we have uncompleted pictures.

H. Steam and Leaf Display Diagram

Suppose that we have a set of n observations. This data will be presented in the steam and leaf display. The methodology of this kind of display is to divide each of the value into two parts. One part consists of one or more leading digits (highest and next highly placed value digits) as steam and rest of the digits as leaf. The stem values are listed out to the left of a vertical line and each leaf value corresponding to a steam is written in the horizontal line to the right of the steam in the order in which they are encountered in passing from one value to the other. For example we have score 53. The observation can be presented in the form of steam and leaf display as given by taking the digits in the tenth place as stem and in the unit place as leaf in which 5 is steam and 3 is leaf.

They are usually used when there are large amounts of numbers to analyze. Series of scores on sports teams, series of temperatures or rainfall over a period of time, series of classroom test scores are examples of when Stem and Leaf Plots could be used.

Example 2.7: Score row data test Statistics. The data Table 2.7 give the score of the test of statistics in July 2008.

Table 2.7: Score of Statistics test

79	49	48	74	81	98	87	80
80	84	90	70	91	93	82	78
70	71	92	38	56	81	74	73
68	72	85	51	65	93	83	86
90	35	83	73	74	43	86	88
92	93	76	71	90	72	67	75
80	91	61	72	97	91	88	81
70	74	99	95	80	59	71	77
63	60	83	82	60	67	89	63
76	63	88	70	66	88	79	75

The data (Table 2.7) can be presented by steam and leaf display as follows: The lowest value in the given data is 35 and the highest data 99. So we can take first digit as tenth as steam and the second digit as unit as leaf display. Thus, the steam and leaf display as given in Figure 2.8 bellow:

Figure 2.8: Score Statistics Test

Steam	Leaf
3	58
4	389
5	169
6	00133356778
7	0000111222233444455667899
8	0000111223334566788889
9	000111223335789

More Than One Set of Data

To compare two sets of data, we can use a 'back to back' stem and leaf plot. For instance, if we want to compare the scores of two sports teams, we will use the following Stem and Leaf Plot:

Scores		
Leaf	**Stem**	**Leaf**
Tigers		**Sharks**
0 3 7 9	3	2 2
2 8	4	3 5 5
1 3 9 7	5	4 6 8 8 9

What does this Stem and Leaf Plot Show?

The column is now in the middle and the ones column is to the right and left of the stem column. We can see that the sharks had more games with a higher score than the Tigers. The Sharks only had 2 games with a score in of 32. The Tigers had 4 games, a 30, a 33, a 37 and a 39. We can also see that the Sharks had the highest score of all - a 59, compared to the Tigers with a 57. Stem and Leaf Plots enable us to find medians, determine totals, and determine the modes.

I. Scatter Plot Diagram

A **scatter plot** is a type of mathematical diagram using Cartesian coordinates to display values for two variables for a set of data. The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis. A scatter plot is also called a *scatter chart*, *scatter diagram* and *scatter graph*.

Example 2.8: Figure 2.9 is an example of a **scatter plot diagram** for expression diagram of "*The relationship between running speed (x1) and arm muscle strength (x2) and long jump(y) result",* which first presented in Table 2.8.

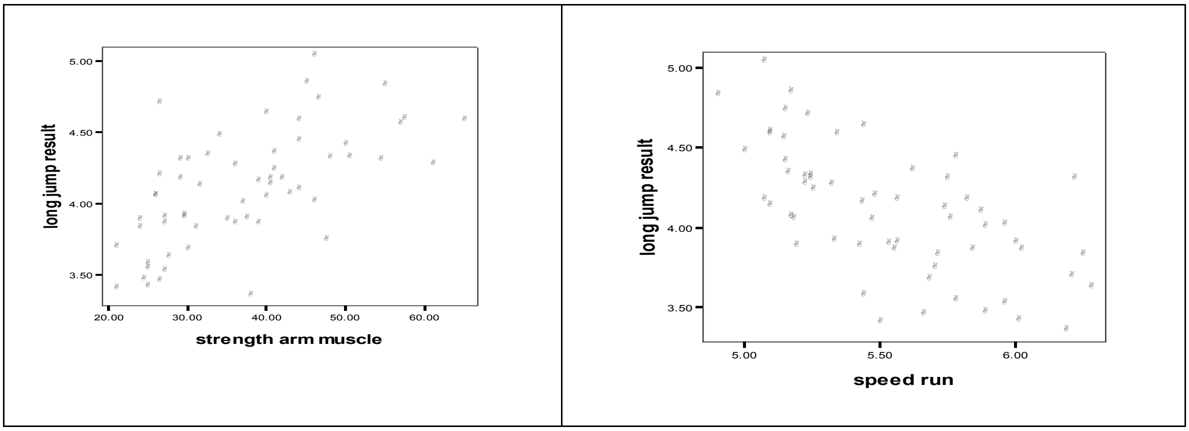
Table 2.8: Data of Running speed, arm muscle strength and long jump

o	X1	X2	Y
1	5.56	42.0	4.17
2	5.89	24.5	3.46
3	5.66	26.5	3.45
4	5.42	24.0	3.88
5	5.18	26.0	4.05
6	5.15	50.0	4.41
7	5.09	65.0	4.58
8	6.21	21.0	3.69
9	5.53	37.5	3.89
10	6.19	38.0	3.35
11	5.16	32.5	4.33
12	5.15	46.5	4.73
13	5.68	30.0	3.67
14	5.22	61.0	4.27
15	5.09	40.5	4.13
16	5.17	45.0	4.84
17	5.44	40.0	4.63
18	5.48	26.5	4.19
19	5.70	47.5	3.74
20	5.75	30.0	4.30
21	5.71	24.0	3.82
22	5.96	27.0	3.52
23	5.07	46.0	5.03
24	6.22	29.0	4.30
25	6.25	31.0	3.82
26	5.43	39.0	4.15
27	5.17	43.0	4.06
28	5.84	39.0	3.85
29	5.22	48.0	4.31
30	5.32	36.0	4.26
Continuum	X1	X2	Y
31	5.76	26.0	4.05
32	5.89	37.0	4.00
33	5.62	41.0	4.35
34	5.23	26.5	4.70
35	5.33	29.5	3.91
36	5.14	57.0	4.55
37	6.28	27.5	3.62
38	5.82	29.0	4.17
39	5.55	36.0	3.85
40	5.56	27.0	3.90
41	5.50	21.0	3.40
42	5.25	41.0	4.23
43	5.34	44.0	4.58
44	5.96	46.0	4.01
45	5.47	40.0	4.04
46	6.02	27.0	3.85
47	5.09	57.5	4.59
48	6.00	29.5	3.90
49	5.00	34.0	4.47
50	5.19	35.0	3.88
51	5.74	31.5	4.12
52	6.01	25.0	3.41
53	5.78	44.0	4.43
54	5.07	40.5	4.17
55	5.78	25.0	3.54
56	4.90	55.0	4.82
57	5.87	44.0	4.09
58	5.24	50.5	4.32
59	5.44	25.0	3.57
60	5.24	54.5	4.30

Data : *Thesis PPs UNNES, Sunarjo*

Scatter plot diagrams are used to evaluate the correlation or cause-effect relationship (if any) between two variables (e.g., running speed and long jump in a sport or strength arm muscle and long jump in a sport). When we think there is a cause-effect link between two indicators (e.g., running speed and long jump in a sport) then we can use the **scatter plot** to prove or disprove it. The points in Figure 2.9 are tightly clustered along the trend line. That mean that there is probably a *correlation* between running speed and long jump. It is the same situation that between strength arm muscle and long jump. But this relation is the opposite direction.

Figure 2.9: Relation between speed run, strength arm muscle, and long jump result



Scatter plot positive correlation Scatter plot negative correlation

J. Application and Problem Solving

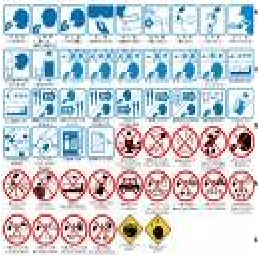
1. *How can we choose appropriate diagram to present data variables, and which area must be appropriate?*

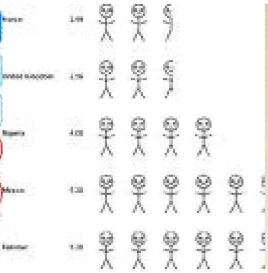
Presenting a special data variables, generally the presenter does not only give information with an arbitrary diagram. Presenter wants to present in a way so that the reader understand and receive information clearly and quickly before concluding the main subject of presentation. For example, the workers in BPS (Statistics central office) are impossible to present all the information with pictogram. They are sure to choose a table matrix row and column. Also the election's committee will choose pie diagram to give a report on the election result of each candidate. The answer of the above question depends on the goal of the presenter. They may want to show the accuracy data, trend data, interest presenting and so on.

Now, we list the using table and diagram in presenting data variables.

a. **Pictogram**: It is used more in an interesting presentation. It means the presentation of data should be interesting. By showing the Figure 2.10 pictogram presentation, at the first step, the reader will be interested in the picture. Then, the next step they want to know more information on it. Presenting with pictogram is appropriate in showing data exhibition, introducing new product, describing result of qualitative research, giving information for advertising; pictogram plays a large role in presenting data.

Figure 2.10: Pictogram





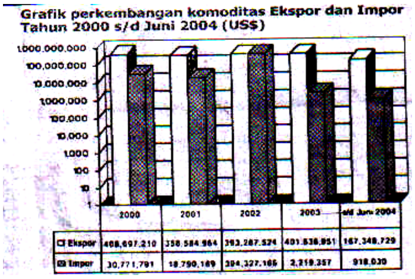
b. **Table matrix row and column**: Presenting data with table matrix row and column is stressed in accuracy data. It is not necessary that the presentation of data is interesting or not. Here, people need the real value of the data first. In this case the readers must pay attention exactly on the value of data. They need more minutes to find the hope data. This work is appropriate in data processing for office, general service centre, report of the experiment result, and so on. It is not appropriate for presenting data in exhibition. For example: If we need the Indonesian population in 2009. We can find exactly the number in Table 2.2 (Presenting Table matrix row and column), namely 229,964,723 people. But it is so difficult to find through another presentation data.

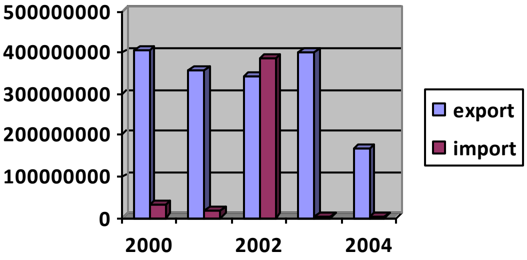
c. **Pie chart**: It is used more in comparing each case for characteristic in one variable. The readers with a minute look at the pie chart; they will direct give conclusion about whom/which one the best/ the worst is without understanding the accuracy data. The conclusion, what the message about comparison of the presentation data is, come from showing the differences of circle segment. Pie chart plays a large role in work area: give report the result of a competition and an experiment/observation in comparing. Example: we look at the diagram Figure 2.2a and want to know only comparison of the countries population. We can directly come to the conclusion that the biggest population is China, and the second is India. Indonesia takes place in number four and the last one is Australia. This comparing conclusion came without knowing first the accuracy data

d. **Bar chart**: Bar chart is used in comparing data also. Multiple bar diagram is used better by the efficient place presentation. If we take data in advertisement with multiple bar diagram cheaper than the other diagram. Bar diagram is also appropriate for showing the trend of data.

   Sometimes, we can play with bar chart to present data. By using the difference scale, it can be shown the difference interpretation about the message of presenting data. For example: look at two bar diagrams for showing the same data (Figure 2.11). With the difference scale of *y- apses*, we can have difference interpretation. The first one come to the conclusion that the trend data is monotone, there is no big difference each year. But the second one come to the conclusion that trend data is fluctuate.

Figure 2.11: Export Import 2000 -2004 in Central Java





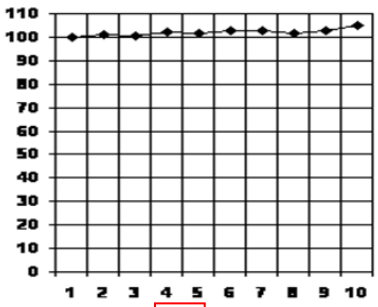
e. **Line diagram**: Trend data will be good if it is presented in line diagram. The motivation is like the pie diagram. Only for one minute reading, people can reach the conclusion of the message in data presentation. The trend of data may be decrease, increase, and monotone, fluctuate, or go irregularly. Normally, the works with variable that has relation with time then line diagram play a large role in presenting its data.

It is just the same as a bar chart. One data variable is made in two line diagram with different scale. We can come to difference conclusion. For example we have data Table 2.9 about produced condom in 10 years. Figures 2.12a and 2.12b show this data in different scale.

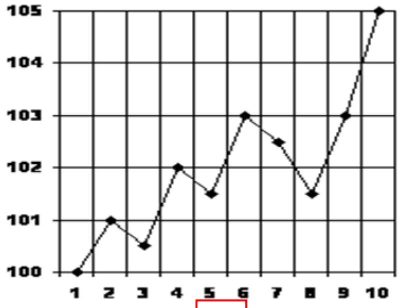
Table 2.9: Produced condom in 10 years (in thousand)

year	1	2	3	4	5	6	7	8	9	10
omset	100	101	100.5	102	101.5	103	102.5	101.5	103	105

Figure 2.12: Produced condom in 10 years (in thousand)



a



b

If we were the owner or the worker of this factory, we would like to show the picture 2.12b to make sure the leader of a bank for making a new credit. The picture 2.12b presented the increasing trend. But, if we are the user of the condom and we are unsatisfied. We should show picture 2.12a to make sure the policymakers, because the trend of data is monotone. There is no improving of the factory.

f. **Scatter plot diagram**: The scientists use scatter plot diagrams more if they want to present about the relationship among the variables or trend of error data, scatter plot data is needed.

K. Exercise

1. Assume that you will lead a research to investigate some interesting data as follows:

   a. Population growth in your city in middle year 2009.
   b. Number of students in each department in University aged more than 25 years old.
   c. Student's ability to use computer for 4th semester
   d. Observing the kind of blood of students in Language Faculty

   Which one do you choose for presenting of the research?

2. State the advantages of *steam and leaf* presentation with some examples.

3. Explain what a pie diagram and give its advantages!

4. You will present the data of these variables. Choose which one is appropriate: a table or a diagram? Give your reasons!
    a. The amount of people in Jakarta 2008 based on their religion
    b. The hobbies of the students in Physics department
    c. The relation between variables weigh and length of the babies
    d. The result of the Olympic Games
    e. The production of fish in a district from 2000 to 2009.
    f. The new product of teaching aid to teach languages.

5. What are the differences between pie diagram and bar diagram?

6. If you choose the table matrix row and column to present your data, explain your reason!

7. Look at the Table school candidates' prep times and test scores below!

Candidate	1	2	3	4	5	6	7	8
Days studied	7	9	5	1	8	4	3	6
Score earned	23	25	14	5	22	15	11	17

Make a scatter plot and give your interpretation.

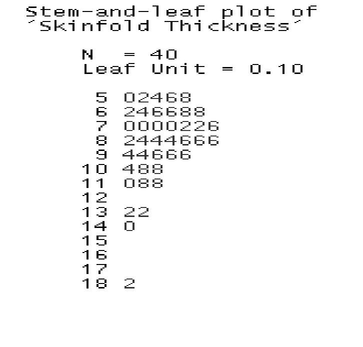
8. Look at the diagram stem and leaf bellow. Give all the real score.

9. Look at the scatter plot bellow! Interpret it and give your opinion!

10. Look at the line diagram below! Interpret it and give your opinion!

11. Look at the pie diagram below! Interpret it and give your opinion!

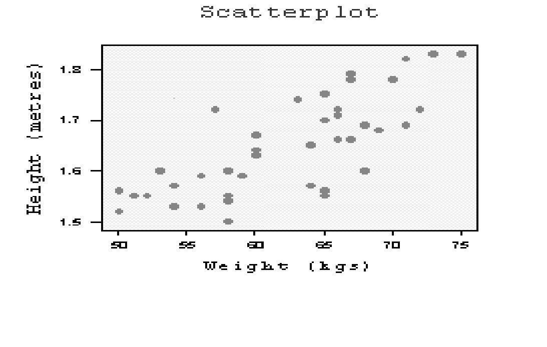
12. Look at the pie diagram. Interpret what is your opinion.

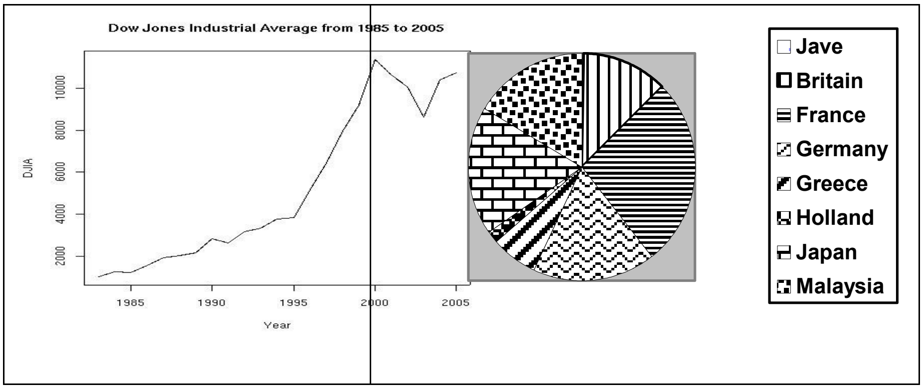


```
Stem-and-leaf plot of
'Skinfold Thickness'

      N   = 40
      Leaf Unit = 0.10

        5 02468
        6 246688
        7 0000226
        8 2444666
        9 44666
       10 488
       11 088
       12
       13 22
       14 0
       15
       16
       17
       18 2
```




Dow Jones Industrial Average from 1985 to 2005
DJIA
2000
4000
6000
8000
10000
1985
1990
1995
2000
2005
Year
Jave
Britain
France
Germany
Greece
Holland
Japan
Malaysia

CHAPTER III
VALUE OF CENTRAL TENDENCY

C. Description and Basic Competence

The description consists of the knowledge of statistical data, application of central tendency.

The main instructional objectives are after learning process students are able to:

8. Differentiate the main value central tendency in array data.
9. Compare among the arithmetic mean, geometric mean, and harmonic mean of array data.
10. Compare the computational among quartile, deciles, and percentile of array data.
11. Compare the meaning of median, quartile, percentile.
12. Compute the mode of array data.
13. Construct a table frequency distribution from array data to grouped data.
14. present frequency distribution in histogram and polygon frequency.
15. estimate the value of central tendency with ogive.
16. Compute the value of main central tendency in grouped data.

D. Value of Central Tendency in Array Data

1. Arithmetic Mean (AM)

Arithmetic mean is commonly called as average. The mean or average is defined as the sum of all the given elements divided by the total number of elements, with the formula:

$$\bar{x}_{am} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example 3.1: Calculate mean of the data 23, 3, 23, 46 and 45 as follows:

$$\bar{x} = \frac{23 + 3 + 23 + 46 + 45}{5} = 28.$$

Geometric Mean (GM)

When a positive value is repeated in either the means or extremes position of a proportion that value is referred to as a **geometric mean** (or **mean proportional**) between the other two values.

Example 3.2: Find the geometric mean between 4 and 25.

Let x = the geometric mean.

$\frac{4}{x} = \frac{x}{25}$ (Definition of geometric mean)

$x^2 = 100$ (Cross-Products Property)

$x = \sqrt{100}$

x = 10. The geometric mean between 4 and 25 is 10.

In general we define: The geometric mean is an average that is useful for sets of positive numbers that are interpreted according to their product and not their sum (as is the case with the arithmetic mean) e.g. rates of growth. The geometric mean is then computed by using the following formula:

$$\bar{x}_{gm} = \left(\prod_{i=1}^{n} x_i \right)^{1/n}.$$

Example 3.3: The geometric mean of six values: 34, 27, 45, 55, 22, 34 is:

$$\bar{x}_{gm} = (34.27.45.55.22.34)^{1/6} = 1{,}699{,}493{,}400^{1/6} \approx 34.545.$$

Harmonic Mean (HM)

The harmonic mean is an average which is useful for sets of numbers which are defined in relation to some unit, for example speed (distance per unit of time). The harmonic mean is then computed by using the following formula:

$$\bar{x}_{hm} = n.\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1} .$$

Example 3.4: The harmonic mean of the six values: 34, 27, 45, 55, 22, and 34 is

$$\bar{x}_{hm} = \frac{6}{\frac{1}{34}+\frac{1}{27}+\frac{1}{45}+\frac{1}{55}+\frac{1}{22}+\frac{1}{34}} = \frac{60588}{1835} \approx 33.018$$

Relationship Between AM, GM & HM

The relationship among AM, GM & HM can be generalized as follow:

AM > = GM > = HM.

Equality is possible only when all the elements of the given sample are equal.

Weighted Arithmetic Mean

In calculation of arithmetic mean, the importance of all the items was considered to be equal. However, there may be situations in which all the items under considerations are not equal importance. For example, we want to find average number of marks per subject test who appeared in different tests like task test, mid semester test, and semester examination. These tests do not have equal importance. Thus, arithmetic mean computed by considering relative importance of each items is called weighted arithmetic mean. To give due importance to each item under consideration, we assign a number which is called weight to each item in proportion to its relative importance. The weighted Arithmetic Mean is then computed by using the following formula:

$$\bar{X}_w = \frac{\sum wx}{\sum w}$$

where:

$\bar{X}_w$ Stands for weighted arithmetic mean

x Stands for values of the items and

w Stands for weight of the item.

Example 3.5:

A student obtained 50, 60, and 80 marks in the tests I, II, III respectively. Let's assume that the weights of 1, 2 and 4 are respectively for the above tests. Find the weighted arithmetic mean per subject.

We will find the weighted arithmetic mean as follows:

$$\bar{X}_w = \frac{\sum wx}{\sum w} = \frac{50(1)+60(2)+80(4)}{1+2+4} = 70.$$

2. Mode (Mo)

Mode is the most frequently occurring value in a data array.

Example 3.6: Find the mode of data: 11,3,5,11,7,3,11. We can arrange the numbers in ascending order: 3,3,5,7,11,11,11. In this case f(3)=2, f(5)=1, f(7)=1, and f(11)=3. The mode of this data is 11. It is possible the value of mode is not unique (can be more than one values), in which case it is multimodal. In the data set {1,1,2,3,3} there are two modes: 1 and 3.

3. Median (Me)

Median is the middle value of the given numbers or distribution in their ascending order or descending order, with the formula:

$$\mathrm{Me} = x_{(n+1\,1)/2} \text{ , if } n \text{ even number}$$

$$\mathrm{Me} = \frac{1}{2}(x_{1/2} + x_{(n/2)+1}), \text{ if } n \text{ odd number.}$$

To find the value of median, fist we arrange the numbers in ascending or descending order.

Let the data are: $x_1, x_2, \ldots, x_n$. Then the ascending data are $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

Example 3.7: Find the median of 4,5,7,2,1,8 [even]. We should arrange with descending data:1,2,4,5,7,8. So the position are n/2 = 6/2 = 3 and 4. The number at 3rd and 4th position are 4 and 5. The average is (4+5)/2= Median = 4.5.

4. **Generate Median (Quartile, Deciles, Percentile)**

Median is the middle values of the given descending or ascending numbers that divide the numbers in 50% on the left place and 50% on the right place. If it is generate this dividing in 4 parts we have quartile, in 10 parts we have deciles, in 100 parts we have percentile. Divided the array data in to 4 parts, then the scores limit are called first quartile =lower quartile (Q1), middle quartile = median = Q2, and third quartile = upper quartile (Q3). In deciles we have first up to ninth deciles (D1 to D9), and with the percentile we have first up to ninety-ninth percentile (P1 to P100).

The ascending/descending data are given: $x_1, x_2, \ldots, x_n$. To find the value of ith quartile/deciles/percentile first, we should find the position number of them. The position numbers:

$q_i = \dfrac{i(n+i)}{4}$ determine the value of $Q_i = x_m .+ t\,(x_{m+1} - x_m)$, i=1,2,3;

$d_i = \frac{i(n+i)}{10}$ determine the value of D$_i$ = x$_m$.+ t (x$_{m+1}$ – x$_m$), i=1,…,9;

$p_i = \frac{i(n+i)}{100}$ determine the value of P$_i$ = x$_m$.+ t (x$_{m+1}$ – x$_m$), i=1,…,99,

where: Q$_i$, D$_i$ and P$_i$ are the value of ith quartile, deciles and percentile respectively, m = down rounding number of q$_i$ or d$_i$ or p$_i$, m+1 = adding position m with 1, t = q$_i$ – m or t=d$_i$-m or t=p$_i$-m. In special case we have Q$_i$ = D$_{2.5i}$ = P$_{25i}$ and D$_i$= P$_{10i.}$

Example3.8: Find Median (Q$_2$), Deciles 7 (D$_7$) and Percentile 30 (P$_{30}$) of the data 1, 2, 2, 3, 7, 8, 9, 9, 11, 18, 20, 21.

Solution: Here, the number of cases n=12.

a. Calculate the value of Median:

   The position number of median is $q_2 = \frac{2(n+1)}{4} = \frac{2(12+1)}{4} = 6.5$. In this case, we have m=6, t=6.5 – 6 = 0.5, x$_6$=8, x$_7$=9, so that the value of Median:

   Me = Q$_2$ = x$_6$ + t (x$_{6+1}$-x$_6$)

   = 8 + 0.5(9 – 8) = 8.5.

b. Calculate the value of D$_7$:

   The position number of deciles 7 is $d_7 = \frac{7(n+1)}{10} = \frac{7(12+1)}{10} = 9.1$.

   In this case, we have m=9, t=0.1, x$_9$=11, x$_{10}$=18; so that the value of deciles 7:

   D$_7$ = x$_9$+t(x$_{9+1}$ –x$_9$) = 11 + 0.1(18 –11) =11.7.

b. Calculate the value of P$_{30}$:

The position number of percentile 30 is $p_{30} = \frac{30(n+1)}{100} = \frac{30(12+1)}{100} = 3.9$. In this case, we have m=3, t=0.9, x_3=2, x_4=3; so that the value of percentiles 30:

$P_{30} = x_3 + t(x_{3+1} - x_3) = 2 + 0.9\ (3-2) = 2.9$.

Quartile Deviation:

It is based on the lower quartile Q_1 and the upper quartile Q_3. The difference Q_3-Q_1 is called the Inter Quartile Range (IQR). The difference Q3-Q_1 divided by 2 is called semi-inter-quartile range or the quartile deviation. Thus

Quartile Deviation: QD = $\frac{Q_3 - Q_1}{2}$

A relative measure of dispersion based on the quartile deviation is called the coefficient of quartile deviation. It is defined as

Coefficient of Quartile Deviation $CQD = \frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$.

It is pure number free of any units of measurement. It can be used for comparing the dispersion in two or more than two sets of data.

Example 3.9: For the upper data example 3.8, find the quartile deviation and coefficient of quartile deviation.

Solution: We have:

$q_1 = \frac{1(12+1)}{4} = 3.25$ $q_3 = \frac{3(12+1)}{4} = 9.75$

Q_1= 2+0.25(3-2)=2.25 Q_3=11+0.75(18-11)=16.25.

QD = ½ (16.25 – 2.25) = 7, and CQD = $\frac{16.25 - 2.25}{16.25 + 2.25} = 0.76$.

Note: We have 2 different ways to determine the quartile or deciles. The firs way is to divide direct by the descending or ascending data in 4 partitions or 10 partitions, and the second way be found through determining the position like the upper formula. And the result is not always the same. For example we will find the low quartile of the data in example 3.8.

Firs way, through dividing:

1, 2, 2 | 3, 7, 8, | 9, 9, 11, | 18, 20, 21

Q1 Q2 Q3

The value of low and upper quartile:

Q1= ½(2+3)=2.5 and Q3=½(18+11)=14.5.

Second way with the positioning: From the calculation in example 3.8 we have found Q_1=2.25 and Q_3=16.25. The results in both ways are quite different.

To give solution in this case, we must give the question clearly whether with the dividing way or with determining position way to finding the value of quartile or deciles. If there is no information that means it is by default a dividing way.

C. Frequency Distribution in Grouped Data

1. Frequency Distribution

Before we discuss to calculate the value of central tendency in grouped data, here we will be discussed first about how to make data classification in group. The row data collected through surveys or experiments will be in a haphazard and unsystematic form. Such a data is not in appropriately form to draw right conclusions about the group or population under study. Hence, it becomes necessary to arrange or organize data in classification group.

The placement of data in different homogeneous group, formed on the basic of some characteristics or criteria, as called *classification of data*. For instance, the people may be divided into different age groups like 1-10, 11-20,21-30,31-40 etc. or may be classification with their monthly income like 250-500, 501-750, 751-1000, etc. Further these classified data can be presented in the form of well arranged tables. In short, a table is a systematic arrangement of data in rows and/or columns. As a matter of fact, the kind of classification or tabulation mostly depends on the type of information required for study and the type of further statistical treatment to be undertaken. Norms for an ideal classification are:

a. The classes should be complete and non overlapping.
b. Clarity of classes in another important properly.
c. one should use standardized classes so that the comparison of results can be possible from time to time.

A frequency distribution is a tabular arrangement of data into classes according to the size or magnitude along with corresponding class frequencies (the number of values fall in each class). Data presented in the form of frequency distribution is called *grouped data*.

The numerical raw data is arranged in ascending or descending order is called an *array*.

Quantitative classification means arranging data according to certain characteristic that has been measured. To determine the number of classes will depend upon the number of classes which will be arbitrarily decided keeping in view the quantum of data. A numerical formula as suggested by Sturges with the formula:

$k = 1 + 3.322 \log\ n$, n = total number of observations

k = number of classes

The formula size of the class interval (i) as follow:

$i = \dfrac{L-S}{k}$, L = largest observation and S = smallest observation.

In case of fractional results, the next higher whole number is taken as the size of the class interval.

The premise of data in the form of frequency distribution describes the basic pattern which the data assumes in the mass. Frequency distribution gives a better picture of the pattern of data if the number of items is large enough. Once the classes are formed, the frequencies for these classes from row data are expedited with the help of tally marks. A bunch of four tally marks is crossed by the fifth to make the counting simpler.

Example 3.10: The age of 30 men when they got marriage in a district in 2003 was reported as given below:

Table 3.1: The age of marriage man in district in 2003

20, 21, 23, 30, 31, 27, 28, 35, 31, 37
40, 23, 35, 42, 37, 32, 27, 25, 27, 38
31, 30, 26, 28, 29, 35, 41, 39, 28, 22

Frequency distribution can be formed in the manner described so far, using various class intervals. The width of the classes and the number of classes will be found out by Sturge's formula:

$$k = 1 + 3{,}322 \log 30 = 5.91 \approx 6, \text{ so that}$$

$$i = \frac{42-20}{6} = 3.67 \approx 4.$$

Decide the starting point: The lower class limits or class boundary should cover the smallest value in the raw data. It is a multiple of class interval. Hence, six classes with the width of 4 are to be taken in the frequency distribution with the help of tally marks and will show in table 3.2 bellow as,

Table 3.2: Frequency Distribution with Tally Marks

classes	Tally marks	Nr. of men
20-23	𝍸	5
24-27	𝍸	5
28-31	𝍸 ||||	9
32-35	||||	4
36-39	||||	4
40-43	|||	3

The distribution constituted by columns 1,2, and 3 in the table 3.2 is known as frequency distribution. The frequency distribution has helped to arrange the haphazard data in a systematic manner which is easy to handle for further treatment.

The important main definitions.

Class Limits:

The variant values of the classes or groups are called the class limits. The smaller value of the class is called lower class limit and larger value of the class is called upper class limit. Class limits are also called inclusive classes.

Example 3.11: Let us take the class 24-27, the smaller value 24 is lower class limit and larger value 27 is called upper class limit.

Class Boundaries:

The true values, which describe the actual class limits of a class, are called class boundaries. The smaller true value is called the lower class boundary and the larger true value is called the upper class boundary of the class. It is important to note that the upper class boundary of a class coincides with the lower class boundary of the next class. Class boundaries are also known as exclusive classes.

Example 3.12: Let's the class 24-27 and 28-31, a student whose marriage are between 27 years to 27.5 years would be included in the 24-27 class, a student whose marriage is 27.5 years to 28 years would be included in next class 28-31. Lower class boundary of 28-31 is 28 - ½(28-27)=27.5 (it is also as upper class boundary of 24-27).

Class Mark or Mid Point:

The class marks or mid point is the mean of lower and upper class limits or boundaries. So it divides the class into two equal parts. It is obtained by dividing the sum of lower and upper class limit or class boundaries of a class by 2.

Example 3.13: The class mark or mid point of the class 24-27 is (24+27)/2 = 25.5

Size of Class Interval (i)

The difference between the upper and lower class boundaries (not between class limits) of a class or the difference between two successive mid points is called size of class interval. In our example with the data table 3.2 the size of class interval,

i = 23.5-19.5=27.5-23.5=4.

Example 3.14: Arrange the marks of learning achievement of Biology in ascending order as:

12, 15, 21, 23, 26, 27, 30, 33, 34, 35, 36, 38, 39, 41, 42, 43, 43, 44, 46, 47, 47, 48, 48, 50, 50, 51, 51, 52, 52, 53, 54, 54, 55, 56, 56, 57, 58, 59, 59, 60, 62, 63, 64, 65, 65, 67, 68, 72, 75, 77

Minimum Value =12   Maximum = 77

Range = Maximum Value – Minimum Value = 77-12=65

$k = 1 + 3{,}322 \log 50 = 6.64 \approx approximate 7,$ and

$i = 65/7 = 9.28 \approx 10.$ Table 3.3 shows the frequency distribution with 7 class interval and size class interval i=10.

For finding the class boundaries, we take half of the difference between lower class limit of the 2nd class and upper class limit of the 1st class $\frac{20-19}{2} = \frac{1}{2} = 0.5.$ This value is subtracted from lower class limit and added in upper class limit to get the required class boundaries.

Table 3.3: Frequency distribution of learning achievement with k=7, i=10

Marks Class Limits C.L	Tally Marks	Number of Students f	Class Boundary C.B	Class Marks x
10-19	𝍷𝍷	2	9.5-19.5	14.5
20-29	𝍷𝍷𝍷𝍷	4	19.5-29.5	24.5
30-39	𝍸𝍷𝍷	7	29.5-39.5	34.5
40-49	𝍸𝍸	10	39.5-49.5	44.5
50-59	𝍸 𝍸𝍸𝍷	16	49.5;59.5	54.5
60-69	𝍸 𝍷𝍷𝍷	8	59.5-69.5	64.5
70-79	𝍷𝍷𝍷	3	69.5-79.5	74.5

2. Histogram

This type of diagrammatic representation is more suited for frequency distributions with continues classes (compare with the bar diagram appropriate with the discrete data). In this type of distribution the true upper limit of a class is the true lower limit of the following class. The magnitudes of the class intervals are plotted along the abscises and the frequencies along the ordinate according to the chosen scale. The rectangles are drawn on each class interval with height in proportion to its frequency. The number of such rectangles will be equal to the number of classes.

Example 3.15: We make histogram for the data Table 3.3 in the Example 3.14. Figure 3.1 shows this histogram, and is given as billow:

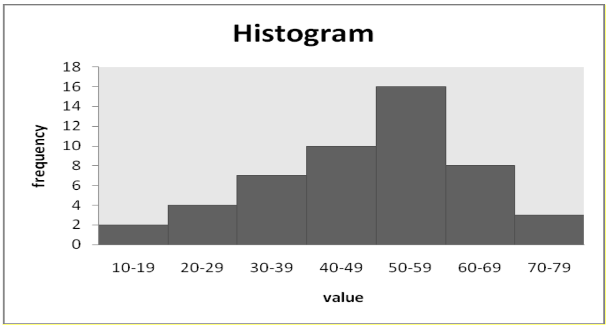


Figure 3.1: Histogram learning achievement of Biology

3. Frequency Polygon

By connecting the midpoints of the bars with lines, we produce a frequency polygon. The frequency polygon displays the same information as the histogram, but in a different form. If we remove the bars of the histogram, we obtain a frequency polygon graph, below.

In case of frequency distribution, the variety values are taken on the X axis and frequencies on the Y axis. The points are plotted on the graph paper and joined by line segments, in the order they are plotted. This graph is called a frequency polygon.

Example 3.16: The smoothened distribution of the age of marriage men are given in example 3.15 has been represented by a histogram. The frequency polygon has also been shown in the figure 3.2.

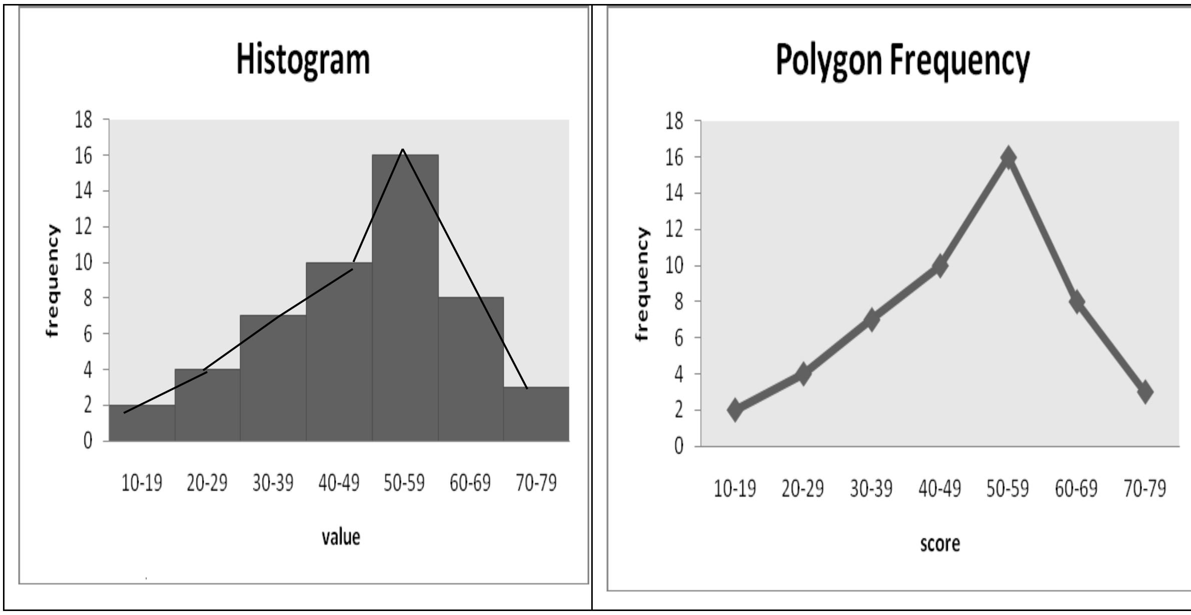


Figure 3.2: Polygon of learning achievement of Biology

It is worth noting how the points of the first and the last rectangles are joined to the abscissa. The frequency at the mid-point of a class before the first class interval and after the last class interval is zero. The area of the frequency polygon is equal to the area of the histogram in the case where class intervals are equal since the area of the histogram left out by the polygon is equal to the area encroached by the polygon outside the histogram.

4. Cumulative Frequency Distribution

The total frequency of all classes less than the upper class boundary of a given class is called the cumulative frequency of that class. “A table showing the cumulative frequencies is called a cumulative frequency distribution”. There are two types of cumulative frequency distributions.

Less than cumulative frequency distribution: It is obtained by adding successively the frequencies of all the previous classes including the class against which it is written. The cumulate is started from the lowest to the highest size. More than cumulative frequency distribution: It is obtained by finding the cumulate total of frequencies starting from the highest to the lowest class. Table 3.4 present the less than cumulative frequency distribution and more than cumulative frequency distribution for the frequency distribution.

Table 3.4: Cumulative Frequency lest than and more than

Class limit	f	C.B	Less than Marks	C.F	More than Marks	C.F
10-19	2	9.5-19.5	Less t 19.5	2	9.5 or more	50
20-29	4	19.5-29.5	Less t 29.5	6	19.5 or more	48
30-39	7	29.5-39.5	Less t 39.5	13	29.5 or more	44
40-49	10	39.5-49.5	Less t 49.5	23	39.5 or more	37
50-59	16	49.5-59.5	Less t 59.5	39	49.5 or more	27
60-69	8	59.5-69.5	Less t 69.5	47	59.5 or more	11
70-79	3	69.5-79.5	Less t 79.5	50	69.5 or more	3
	50					

When the points are plotted for variety value and their corresponding cumulative frequencies, and joined by a free hand smooth curve, the curve is S-Shaped. This S-Shaped curve is known as ogive. If we draw two cumulative frequency curves, one on the less than basis, and the other on the more than basis, they intersect at a point of which the X axis is the median value.

Figure 3.3 shows ogive curves on lest than basis and more than basis from the data Table 3.4. The staircase of cumulative frequency histogram has also been shown in the same graph.

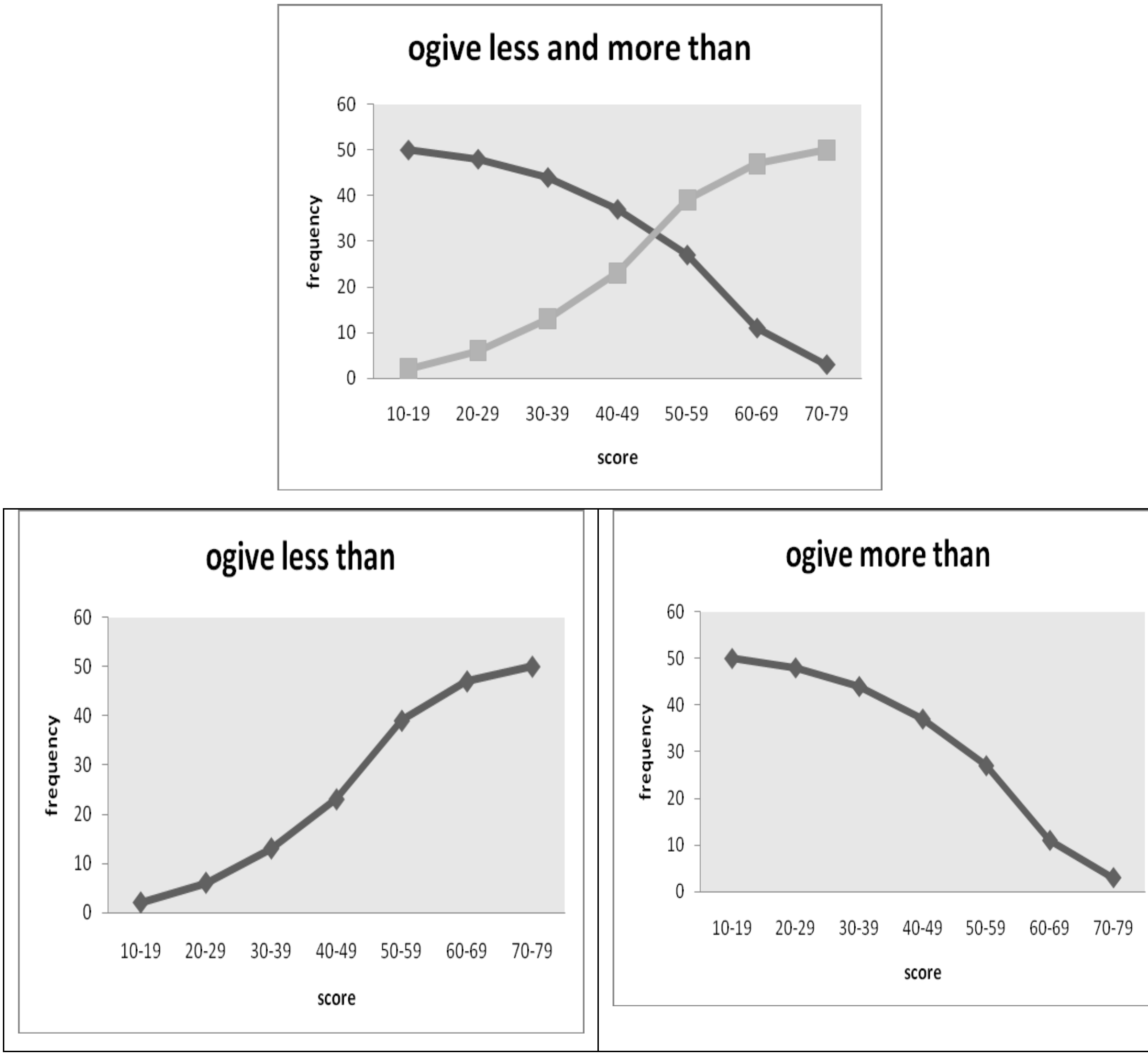


Figure 3.3 : Ogive about learning achievement of Biology

5. Frequency Distribution of Discrete Data

Discrete data is generated by counting. Every observation is exact. If we will present discrete data in frequency distribution, then we make each value data in grouping for the same value. The class limits in discrete data are true class limit; there are no class boundaries in discrete data.

Example 3.17: A teacher has counted for 52 students in classes A and B to come to school for first three months in second semester. The result of the experiment shows in Table 3.5 as given bellow:

Table 3.5: Presenting student in three months

74, 75, 77, 76, 72, 73, 74, 75, 74, 75, 74, 73, 76
78, 74, 75, 73, 74, 75, 75, 74, 75, 75, 72, 77, 78
77, 76, 75,74, 75, 78, 76, 75, 77, 75, 74, 77, 74
75, 75, 74, 75, 74, 76, 77, 75, 77, 76, 75,78, 76

The data will be presented in the form of frequency distribution. Table 3.6 shows the frequency distribution with tally marks.

Table 3.6: Tally of presenting students for three months

score	Tally marks	No of student
72	||	2
73	|||	3
74	𝍸 𝍸 ||	12
75	𝍸 𝍸 𝍸 ||	17
76	𝍸 ||	7
77	𝍸 ||	7
78	𝍸	4

D. **Value of Central Tendency in Grouped data**

1. Mean Arithmetic in Grouped data

When the data are arranged or given in the form of frequency distribution i.e. there are *k* variate value such that a value x_i has a frequency f_j (j = 1,2,…,k), the formula for the mean is,

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + ... + f_kx_k}{f_1 + f_2 + ... + f_k}$$

$$= \frac{\Sigma f_j x_j}{\Sigma f_j}, \qquad j=1,2,...,k.$$

Example 3.18: We use the data Table 3.6 to calculate the average of presenting students for three months.

$$\bar{x} = \frac{2(72) + 3(73) + ... + 4(78) + 0(79)}{2 + 3 + ... + 4 + 0}$$

$$= \frac{3909}{52} = 75.17 .$$

Thus, the students came for three months in second semester in average 75.15 days.

If the data continue are given with *k* class intervals i.e. the data are in the form as Table 3.2, than the formula of arithmetic mean is given as fallow:

$$\bar{x} = \frac{\Sigma f_j x_j}{\Sigma f_j}, \qquad f_j \text{ frequency j}^{\text{th}}\text{-class interval}$$

X_j the mid point (class mark) of the j^{th}-class interval.

Example 3.19: We use the data Table 3.2 the age of marriage man in district 2003. Table 3.7 shows the process of calculating mean.

Table 3.7:Calculating mean of marriage man with mid points formula

Classes interval	Mid points x_j	Frequency f_j	f_jx_j
20-23	21.5	5	107.5
24-27	25.5	5	127.5
28-31	29.5	9	265.5
32-35	33.5	4	134.0
36-39	37.5	4	150.0
40-43	41.5	3	124.2
	Sum	30	908.7

$$\bar{x} = \frac{908.7}{30} = 30.3.$$

So that the average of age for marriage man in district 2003 is 30.3 years old.

Coding formula: A linier transformation of data may be regarded as coding. In coding formula we shift the origin and change the scale. A change can involve either a change of origin or a change of scale or change of both, origin and scale together. We have frequency distribution with $x_1, x_2, \ldots, x_k$ mid point, *i* size class interval. Let x_0 an arbitrary chosen from mid point, the coding c_i are defined:

$$c_j = \frac{x_j - x_0}{i}, \qquad j=1,2,\ldots,k.$$

The mean arithmetic with coding formula is defined as follow:

$$\bar{x} = x_0 + i\frac{\Sigma f_j c_j}{\Sigma f_j}.$$

Example 3.20: We use data Table 3.7 to calculate the mean value with coding formula. Choose $x_0=x_i$, where x_i mid point belong to the interval which has the biggest frequency. In this case $x_0=29.5$, $c_1=(21.5-29.5)/4=-2$, $c_2=(25.5-29.5)/4=-1$, and so on. Table 3.8 presents the process of coding formula.

Table 3.8: Calculating mean of the age of marriage man with coding formula

Classes interval	Mid points x_j	Frequency f_j	Coding c_j	f_jc_j
20-23	21.5	5	-2	-10
24-27	25.5	5	-1	-5
28-31	29.5	9	0	0
32-35	33.5	4	1	4
36-39	37.5	4	2	8
40-43	41.5	3	3	9
	Sum:	30		6

$\bar{x} = x_0 + i\frac{\Sigma f_j c_j}{\Sigma f_j} = 29.5 + 4\frac{6}{30} = 30.3$ (it is the same value with the mean mid point formula).

Example 3.21: We use the data Table 3.3 of learning achievement of Biology to give another example for calculating arithmetic mean with coding formula. Choose $x_0=45.5$ (mid point belong to the interval which has the biggest frequency). So that, we show the process coding formula in Table 3.9 as bellow:

Table 3.9: Calculating mean of learning achievement with coding data

Classes x_i	Mid point x_i	Frequency f_i	Coding c_i	f_ic_i
10-19		2	-3	-6
20-29		4	-2	-8
30-39		7	-1	-7
40-49	45.5	10	0	0
50-59		16	1	16
60-69		8	2	16
70-79		3	3	9
	Sum	50		20

$$\bar{x} = x_0 + z\frac{\Sigma f_i c_i}{\Sigma f_i} = 45.5 + 10\frac{20}{50} = 49.5\ .$$

2. Mode in Grouped data

It is not difficult to find the mode of discrete data in frequency distribution. The variate value having the maximum frequency is the modal value. For instance, consider the discrete distribution in Table 3.10 that presents the sum of students in computer examination score.

Table 3.10: sum of students in computer examination score

Value (x)	: 3	4	5	6	7	8	9	10
Frequency (f)	: 4	7	3	7	4	3	2	1

Clearly the scores x = 4 and x=6 have maximum frequency, namely 7. Hence 4 and 6 are the mode values.

Another example, we use the data Table 3.6 about presenting student in three months. The score x = 75 has maximum frequency, namely 17. Hence 75 is the mode value.

Remark: An observation which has equal value, there is no mode. The value of mode is not unique.

If the data are given of continue data in frequency distribution, the process to calculate the mode through the explanation from Figure 3.4. We know clearly the mode class interval as the maximum frequency. If there are more than one class which have the same number of frequency (equally qualifying to be the mode class) then both of the classes will be the mode class. This is called bimodal. For example the Table 3.3 or Table 3.9 has the class mode interval is 50-59. The true value of mode we can follow the explanation Figure 3.4 as fallow.

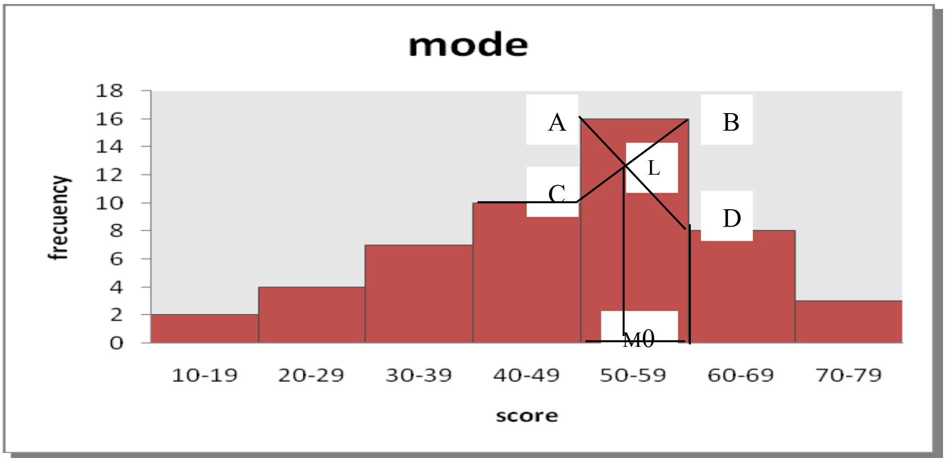


Figure 3.4: Mode of score statistics examination

If we draw a histogram for the given distribution, naturally the highest bar will possess the mode value. To find the exact mode value consider only three bars namely, the highest bar and the bars adjacent to it on both the sides. In the middle bar draw two diagonal lines joining the point A of the preceding bar to D and B of the following bar to C as shown in Fig 3.4. Suppose these diagonals AD and BC intersect each other at the point L. Draw a perpendicular line from L pm the axis of X which meets it at the point M_0. The distance of M_0 from origin on the aforesaid scale is the mode value. So that the formula of mode value is given:

$M_0 = L_0 + i\dfrac{d_1}{d_1+d_2}$, M_0 : modal value , i : size class interval

L_0 : the lower limit of the mode class

d_1 : difference of maximum frequency and preceding frequency

d_2 : difference of maximum frequency and following frequency

Example 3.22: We calculate the value of mode the data Table 3.9 about learning achievement in Biology. The class interval 50-59 has the biggest frequency, namely 16. So that the value of L_0 = 49.5, i=10, d_1=16-10=6, d_2=16-8=8. The value of mode is calculated as follow:

$$M_0 = L_0 + i\frac{d_1}{d_1+d_2} = 59.5 + 10\frac{6}{6+8} = 63.78.$$

Example 3.23: Another example to calculate the value of mode, come from the Table 3.11 about the weigh of baby in a hospital.

Table 3.11: Weigh of Children

Classes (weigh in kg)	Number of children
2.0-2.3	5
2.4-2.7	5
2.8-3.1	9
3.2-3.5	4
3.6-3.9	4
4.0-4.3	3

The interval 2.8-3.1 has the biggest frequency, namely 9. We can calculate first:

L_0 = lower class boundary of 2.8-3.1 = 2.8 – ½(2.8-2.7) = 2.75,

i = difference of upper and lower class boundary = 3.15 – 2.75 = 0.40,

d_1 = 9-5 = 4 and d_2 = 9-4=5. So that

$$M_0 = 2.75 + 0.4\frac{4}{4+5} = 2.93.$$

3. Median in Grouped data

a.

Example 3.24: Consider the discrete distribution.

Variate value (x)	: 3	4	5	6	7	9	10	13
Frequency (f)	: 4	4	3	7	4	3	2	1

First we make the cumulative distribution from the descending or ascending data as follow:

Variate value (x)	: 3	4	5	6	7	9	10	13
Frequency (f)	: 4	4	3	7	4	3	2	1
Cumulative freq.	: 4	8	11	18	22	25	27	28

The number N/2 = 28/2 = 14. The smallest cumulative frequency which contains value 14 is 18, so that the median is 6.

If the data are given of continue data in frequency distribution, the process to calculate the median through the explanation from Table 3.12.

Table 3.12: Cumulative Frequency for finding median

Classes	Frequency f_i	Cumulative frequency

$x_1 - x_2$	f_1	F_1
x_2-x_3	f_2	F_2
.	.	.
.	.	.
x_p-x_{p+1}	f_p	F_p
.	.	.
x_k-x_{k+1}	f_k	F_k
	N=Σf_j	

The procedure to find the value of median are:

1. Make in column 3 the cumulative frequency lest than.
2. Find the N/2 and see in which minimum of the cumulative frequency that N/2 is contained.
3. Suppose N/2 is contained in the minimum cumulative frequency F_p, then obviously the median class is the relation $x_p - x_{p+1}$.
4. Find the unique median value, we take the help of the interpolation. In this approach, it is assumed that the frequency of a class is uniformly distributed over the class interval. Let the cumulative frequency for the class just above the median class be F_{p-1}. Thus ($N/2-F_{p-1}$) is the frequency for the interval between the median and lower limit of the median class. The length of the interval for ($N/2-F_{p-1}$) is $\frac{1}{f_p} * i$, say M_d –median, L_0 –lower class boundary of class median, hence the median is defined:

$$M_d = L_0 + i\frac{N/2 - F_{p-1}}{f_p},$$

where M_d : median

L_0 : lower class boundary of class median

i : size class interval

N : sum of frequency

F_{p-}: cumulative frequency for the above of class median

f_p : frequency of the class median.

Example 3.25: We use data Table 3.2 the age of marriage man in district 2003. We will compute the value of median. Table 3.13 presents the process of finding median.

Table 3.13: Computing Median of the age of marriage men in grouping data

Classes x_i	Frequency f_i	Frequency cumulative
20-23	5	5
24-27	5	10
28-31	9	19
32-35	4	23
36-39	4	27
40-43	3	30
	N=30	

We have N/2 = 15. The number 15 is contained in the smallest cumulative frequency 19. Hence the corresponding value class median 28-31. In this case M_0=27.5, F_{p-1}=10, and f_p=9, so that median can be found as,

$$M_d = M_0 + z\frac{N/2 - F_{p-1}}{f_p} = 27.5 + 4\frac{15-10}{9} = 29.72 \cdot$$

There are 50% (or 15 couples) marriages in district 2003, when they got marriage more than 29.72 years old.

Example 3.26: The next example, we use the data Table 3.3 about the learning achievement. Table 3.14 shows the process to calculate the value of median.

Table 3.14: Computing median of learning achievement in grouped data

Classes x_i	Frequency f_i	Cum freq F_i
10-19	2	2
20-29	4	6
30-39	7	13
40-49	10	23
50-59	16	39
60-69	8	47
70-79	3	50
	50	

We have N/2=25, so that class median 50-59, M_0=49.5, F_{p-1}=23, f_p=16. Median can be computed as,

$$M_d = 49.5 + 10\frac{25-23}{16} = 50.75.$$

4. Quartile, Deciles and Percentile in Grouped data

We can generate the formula of median into quartile, deciles, and percentile in grouped data. The value of Q_m , D_m and P_m can be worked out in the following manner:

a. Make the cumulative frequency less than.

b. Find the N*t, where N=Σf_j for j=1,2,…,k, t=m/4 for quartile, t= m/10 for deciles, t=m/100 for percentile.

c. Search for the smallest cumulative frequency which contains this value N*t. Suppose N*t is contained in the cumulative frequency F_p, then obviously the median class is the relation $x_p - x_{p+1}$.

d. Let the cumulative frequency for the class just above the class be F_{p-1}. Thus (N*t-F_{p-1}) is the frequency for the interval between the Q_j (or D_j or P_j) and lower limit of its class. The length of the interval for (N*t-F_{p-1}) is $\frac{1}{f_p}*i$, say L_0 –lower class boundary, hence we define:

$$Q_m = L_0 + i\frac{Nm/4 - F_{p-1}}{f_p}, \quad m=1,2,3;$$

$$D_m = L_0 + i\frac{Nm/10 - F_{p-1}}{f_p}, \quad m=1,2,\ldots,9,$$

$$P_m = L_0 + i\frac{Nm/100 - F_{p-1}}{f_p}, m=1,2,\ldots,99.$$

Example 3.27: We use data Table 3.2 about the age marriage man to find Q_1, D_6, and P_{75}. Table 3.15 shows the process to find them.

Table 3.15: Computing Quartile, Deciles and Percentile

Classes x_i	Frequency f_i	Frequency cumulative
20-23	5	5
24-27	5	10
28-31	9	19
32-35	4	23
36-39	4	27
40-43	3	30
	30	

Calculate Q_1: We have N*1/4 = 7.5, the number 7.5 is contained in the smallest cumulative frequency 10. Hence the corresponding value class interval Q_1 is 24-27. In this case L_0=23.5, F_{p-1}=5, and f_p=5, so that Q_1 can be found as,

$$Q_1 = L_0 + i\frac{N/4 - F_{p-1}}{f_p} = 23.5 + 4\frac{7.5-5}{5} = 25.5\,.$$

Calculate D_6: We have N*6/10 = 18, the number 18 is contained in the smallest cumulative frequency 19. Hence the corresponding value class D_6 is 28-31. In this case L_0=27.5, F_{p-1}=10, and f_p=9, so that D_6 can be found as,

$$D_6 = L_0 + i\frac{N6/10 - F_{p-1}}{f_p} = 27.5 + 4\frac{18-10}{9} = 31.06.$$

Calculate P_{75}: We have N*75/100 = 22.5, the number 22.5 is contained in the smallest cumulative frequency 23. Hence the corresponding value class P_{75} is 32-35. In this case L_0=31.5, F_{p-1}=19, and f_p=4, so that P_{75} can be found as,

$$P_{75} = L_0 + i\frac{N75/100 - F_{p-1}}{f_p} = 31.5 + 4\frac{22.5-19}{4} = 35.0.$$

Here $P_{75} = Q_3$.

Using Mean, Median or Mode

Different measures of central tendency and fractiles (quantiles) have been discussed in this chapter. Out of mean, median, and mode; the mean (average) is the most commonly used in quantitative analysis. But the other two namely, the median and mode are the most commonly used in qualitative analysis. Median is a largely used in psychology, education and other social sciences. It is a suitable average for qualitative information like the attitude towards disabled people, beauty or intelligence of certain individuals, etc. Mode is a useful measure for manufacturers.

Which of these three measures of central tendency (mean, median, or mode) is chosen if we do something? Notice that not all measures of central tendency are appropriate for all kind of variables. For example,

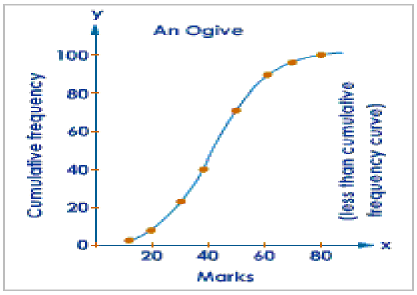
1. For **nominal data** (such as sex or race), the mode is the valid measure.
2. For **ordinal data** (such as motivation categories), the mode and median can be used.
3. For **interval**, **ratio data** (high, temperature), the third central tendency (mean, median, mode) can be used.

Exercise

1. What do you know by a measure or count of central tendency? Explain with examples.
2. Define the following and give one appropriate example of your own for the use of each:
   a. mode
   b. first quartile
   c. Harmonic mean.
   d. Median
3. We have a data from the result of the examination: 74, 81, 56, 96, 63, 55, 91, 93, 85, 51, 95, 69. Compute:
   a. third quartile
   b. Median
   c. Mean arithmetic
   d. Deciles 6
   e. Mode
4. Thirty students were tested to determine how long they used the time to run away from place A to place B. The results, to the

nearest minute, were listed as follows: 423, 369, 387, 411, 393, 394, 371, 377, 389, 409, 392, 408, 431, 401, 363, 391, 405, 382, 400, 381, 399, 415, 428, 422, 396, 372, 410, 419, 386, 390. Make a frequency distribution table in 10 classes intervals, the beginning first class is 360.

5. Make a histogram and polygon frequency for the work number 5.
6. We have diagram Ogive as fallow:



   a. Estimate the score of first, second, and third quartiles.
   b. If this ogive present about score examination, and the score of pass the exam is 65. Estimate how many students passed the exam?
7. Make an ogive for the work number 5.
8. With the data table frequency distribution in the work number 5, compute:
   a. Mean arithmetic with mid point and coding formulas
   b. Median
   c. Mode
   d. Deciles 4
   e. Percentile 80.
9. Find the arithmetic mean, mode, and median for the following distribution.

x_i	f_i
75	8
60	7
92	8
64	7
35	2

11. The result of test examination was given as follow:

Nm	score	f
1		2
2		3
3	53 – 61	5
4	62 – 70	10
5		15
6		18
7		9
8		3

Compute:

a. Mean
b. Median
c. Mode
d. Third quartile

12. Use the data from the work number 11.
    a. We will choose 15% students for the good score; find the lower score of it.
    b. We will give a remedial test for 20% students, which have bad score. Find the limit score of them.
13. What is the effect of reducing each observation of a decreasing series by 10 on the mean, median, mode, quartiles?
14. What do you think about the curve of polygon frequency trend negative position, trend to positive position, or it is in symmetry position.
15. The information regarding the no of children per family is given in the following table. Find the Mean, Median, Mode of the data

No of children	0	1	2	3	4	5
No of families	3	20	15	8	3	1

16. The frequency distribution of the marks obtained by 100 students in a test of Mathematics carrying 50 marks is given below. Find the Mean, median, Mode of data

Marks obtained	0 - 9	10 - 19	20 - 29	30 - 39	40 - 49
No of students	8	15	20	45	12

17.

CHAPTER IV
SPREAD AND SHAPE OF DATA

E. Description and Basic Competence

The description consists of the knowledge about spread and shape of data, interpretation the relation among the value of central tendency.

The main instructional objectives are after learning process students are able to:

1. Compute the range of array data
2. Compute coefficient range
3. Compare the computational the variance of array data and grouped data
4. Compare the computational the standard deviation of array data and grouped data
5. Compute the skewness of data
6. Interpret the skewness of data
7. Compute the kurtosis of data
8. Interpret the kurtosis of data
9. Construct a Box and Whisker Plot

B. Dispersion, Variance and Standard Deviation

Dispersion:

The word dispersion or spread has a technical meaning in statistics. The average measures the centre of the data. It is one aspect observations. Another feature of the observations is as to how the observations are spread about the centre. The observation may be close to the centre or they may be spread away from the centre. If the observation is close to the centre, we say that dispersion, variation is small. If the observations are spread away from the centre, we say dispersion is large.

In the last chapter, we concentrated upon a central tendency, which gives an idea of the whole mass that is a complete set of variate values. However, the information so obtained is neither exhaustive nor comprehensive, as the mean does not lead us to know whether the observations are close to each other or so far apart. Median is a positional average and has nothing to do with the variability of the observations in the series. Mode is the largest occurring value independent of other values of the set. This leads us to conclude that a measure of central tendensy alone is not enough to have a clear idea about data.

To clear this point consider the trhree sets a follows:

Set A	:	7	7	7	7	7
Set B	:	5	7	6	8	9
Set C	:	1	10	13	4	7

All the three sets A, B, and C have mean 7 and median is also 7. But by observation, it is apparent that the three sets differ remarkably from one another. Thus to have a clear picture of data, one needs to have a measure of dispersion or variability amongst observations in the set. Commonly used measures of dispersion are:

1. Range
2. Inter quartile range and quartile deviation
3. Mean deviation
4. Variance
5. Standard deviation
6. Coefficient of variation

Each of the measures of dispersion is presented in the subsequent discussion.

Range

Dispersion refers to the spread of the values around the central tendency. There are two common measures of dispersion, the range and the standard deviation. The **range** is simply the highest value (max) minus the lowest value (min), and it can show with the form as bellow:

$$Range = score \ \max - score \ \min .$$

Example 4.1: From the data **15,20,21,20,36,15,25,15;** the high value(score maximum) is 36 and the low (score minimum) is 15, so the range is 36 - 15 = 21.

A relative measure known as coefficient of range is given as,

Coefficient of Range (CR) = $\dfrac{\max - \min}{\max + \min}$.

Hence, from the data, the coefficient of range can be computed as,

$$\text{CR=} \ \frac{\max - \min}{\max + \min} = \frac{36 - 15}{36 + 15} = 0.41.$$

Properties:

1. It is the simplest measure and can be easily be understood.
2. Besides the above merit, it hardly satisfies any property of a good measure of dispersion, e.g it is based on two extreme values only, ignoring the others.

Inter Quartile Range and Quartile Deviation

Interquartile Range (IR)

The difference between the third quartile and first quartile is called inter quartile range, with the formula:

IQR= $Q_3 - Q_1$.

Quartile Deviation (QD)

This is a half of the interquartile range, I,e:

$$QD = \frac{Q_3 - Q_1}{2}.$$

Also the coefficient of quartile deviation is given by the formula:

$$CQD = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

Example 4.2: The value of Q_1 and Q_3 are worked out in Example 3.27 are $Q_1 = 25.5$ and $Q_3=35$. So that we can calculated:

IQR = $Q_3 - Q_1 = 35 - 25.5 = 9.5$.

$$QD = \frac{Q_3 - Q_1}{2} = \frac{9.5}{2} = 4.25.$$

$$CQD = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{35 - 25.5}{35 + 25.5} = 0.157.$$

Coefficient of quartile deviation is an absolute quantity and is useful to compare the variability among the middle 50% observations.

Properties:

1. It is a better measure of dispersion than range in the sense that it involves 50% of the mid values of a series of data rather than only two extreme values of a series.
2. Since it excludes the lowest and highest 25% values, it is not affected by the extreme values.
3. This measure does not take into account the individual values occurring between Q_1 and Q_3. It means that no idea about the variation of even 50% mid values is available from this measure. Anyhow, it provides some idea if the values are uniformly distributed between Q_1 and Q_3.
4. It is not considered a good measure of dispersion as it does not show the scattering of the central value. In fact, it is a measure of partitioning of distribution. Hence, it is not commonly used.

Mean Deviation

The measures of dispersion discussed so far are not satisfactory in the sense that they lack most of the requirements of a good measure. Mean deviation will give information more than range and quartile deviation. Mean deviation is the average of the absolute deviations taken from a central value, generally the mean or median.

Consider a set of n observations $x_1, x_2, \ldots, x_n$. Then the mean deviation is given:

$$MD = \frac{1}{n}\Sigma \,|\, x_i - a \,|$$

where i=1,2,…,n, a is a central value.

Let $d_i = |x_i - a|$, then $MD = \frac{1}{n}\Sigma d_i$

Example 4.3: Again let's take the set of scores: **15,20,21,20,36,15,25,15.** To compute the mean deviation, we first find the distance between each value and the mean with the formula $d_i = x_i - a \; or \; d_i = x_i - \bar{x}$. We calculate the mean:

$$\bar{x} = \frac{15 + 20 + \ldots + 15}{8} = 20.875.$$

The differences from the mean are:

d_1 =15 - 20.875 = -5.875,

d_2 = 20 - 20.875 = -0.875,

d_3 = 21 - 20.875 = +0.125,

d_4 = 20 - 20.875 = -0.875,

d_5 = 36 - 20.875 = 15.125,

d_6 =15 - 20.875 = -5.875,

d_7 = 25 - 20.875 = +4.125,

d_8 = 15 - 20.875 = -5.875.

We have problem that $\sum d_i = 0.$ So that we understand that the formula of mean deviation is taken the from the absolute value of the sum d$_i$. Now we calculate the value of MD.

$$MD = \frac{1}{n}\Sigma \mid d_i \mid = 38.75/8 = 4.84.$$

Variance

The mathematicians seldom use this formula *MD,* because in mathematics theory the absolute function is not differentiate. So that the negative signs are ignored, is removed by taking the square of the deviations from the mean, that is

$$l^2 = \frac{\Sigma(x_i - \bar{x})^2}{n}.$$

we calculate the square of distance:

$$d_1^2 = \text{-}5.875 * \text{-}5.875 = 34.515625,$$

$$d_2^2 = \text{-}0.875 * \text{-}0.875 = 0.765625,$$

$$d_3^2 = +0.125 * +0.125 = 0.015625,$$

$$d_4^2 = \text{-}0.875 * \text{-}0.875 = 0.765625,$$

$$d_5^2 = 15.125 * 15.125 = 228.765625,$$

$$d_6^2 = \text{-}5.875 * \text{-}5.875 = 34.515625,$$

$$d_7^2 = +4.125 * +4.125 = 17.015625,$$

$$d_8^2 = \text{-}5.875 * \text{-}5.875 = 34.515625.$$

Now, we calculate the value of l^2 as follow:

$$l^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{\Sigma d_i^2}{n} = \frac{350.875}{8} = 43.86.$$

Notice that $\boldsymbol{l^2}$ value still has a problem. Based on statistics theory, that l^2 is biased estimator. Unbiased estimator of variance population is the sum of square of distance divide by the number of scores minus 1. Here is,

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} \text{ or } s^2 = \frac{n\Sigma x_i^2 - (\Sigma x_i)^2}{n(n-1)}.$$

This formula is known as the variance of sample. Now we compute the variance sample for the data example 4.1 as follow:

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{\Sigma d_i^2}{n-1} = \frac{350.875}{8-1} = 50.125.$$

The **variance** is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range (as was true in this example where the single outlier value of 36 stands apart from the rest of the values. The variance shows the relation that set of scores has to the mean of the sample.

Standard Deviation

The drawbacks of variance are overcome in this measure of dispersion. The positive square root of the variance is called standard deviation, with the formula:

$$s = \sqrt{s^2}.$$

In simple words, we can say that standard deviation explains the average amount of variation on either side of the mean.

Now we compute the standard deviation from the data Example 4.1, the result of standard deviation is

$$s = \sqrt{s^2} = \sqrt{50.125}, \text{ s} = 7.079901129253.$$

Example 4.4: Here, is given another data example to find variance and standard deviation. The following nine measurements are the heights in cm in a sample of nine children in Kindergarten.

Height (x): 69, 66, 67, 69, 64, 63, 65, 68, 72

The sample variance of height of children in kindergarten can be computed as,

$$\bar{x} = \frac{69+66+...+72}{9} = 67,$$

$(x-\bar{x})$: 2, -1, 0, 2, -3, -4, -2, 1, 5

$(x-\bar{x})^2$: 4, 1, 0, 4, 9, 16, 4, 1, 25

$$s^2 = \frac{4+1+0+4+9+16+4+1+25}{9-1} = 8 \text{ and s=} \sqrt{8} = 2\sqrt{2}.$$

Now we use the second formula:

Table 4.1: Highs children in Kindergarten

x	x^2
69	4761
66	4356
67	4489
69	4761
64	4096
63	3969
65	4225
68	4624
72	5184
603	40465

$$s^2 = \frac{n\Sigma x_i^2 - (\Sigma x_i)^2}{n(n-1)}$$

$$= \frac{9(40465)-(603)^2}{9(9-1)} = 8$$

The standard deviation is considered to be the best measure of dispersion and is used widely.

Coeficient of variation (CV)

All the measures of dispersion discussed so far have units. If two series differ in their units of measurement, their variability can not be compared by any measure given so far. Also, the size of measures of dispersion depends upon the size of values. Hence in situations where either the two series have different units of measurements, or their means differ sufficiently in the size, the coefficient of variation should be used as a measure of dispersion.

It is a unitless measure of dispersion and also takes into account the size of the means of the two series. It is the best measure to compare the variability of two series or sets of observations. A series with less coefficient of variation is considered more consistent or stable.

Coefficient of variation of a series of variate values is the ratio of the standard deviation to the mean multiplied by 100, with the formula:

$CV = \frac{s}{\bar{x}}100.$ (This measure was given by Karl Pearson).

Example 4.5: We use the data from example 4.1, so that the value of CV is found as follow:

$$CV = \frac{s}{\bar{x}}100 = \frac{7.079}{20.875} \times 100 = 33.91.$$

And we use the data from example 4.4, so that the value of CV is:

$$CV = \frac{s}{\bar{x}}100 = \frac{2\sqrt{2}}{67} \times 100 = 4.22.$$

CV of the data from example 4.1 is greater than CV of the data from example 4.4, that mean the variability of the first data is more consistent or stable than the second data.

C. Variance and Standard Deviation in Grouped data

The data are given in the form of frequency distribution in which the variate value x_i has its corresponding frequency f_j (j=1,2,…,k). Table 4.2 presents the grouped data:

Table 4.2: Frequency distribution with mid values

Classes	Frequen cy f_i	Mid values
$x_1 - x_2$ x_2-x_3 . . x_p-x_{p+1} . x_k-x_{k+1}	f_1 f_2 . . f_p . f_k	x_1 x_2 . . x_p . x_k
	N=Σf_j	

If the data are given in the form of frequency distribution in which the variate value x_i has its corresponding frequency f_i , (i=1,2,3,…,k), the variance is defined as:

$$s^2 = \frac{\Sigma f_j (x_j - \bar{x})^2}{n-1} \text{ or } s^2 = \frac{n\Sigma f_j x_j^2 - (\Sigma f_j x_j)^2}{n(n-1)}$$

Example 4.5: Table 4.3 shows the fee in thousand rupiah of 84 administrators in an university. Find the variance and standard deviation of this data.

Table 4.3: Fee of administrators in an University

Fee in thousand	Number of workers (f_i)
20 – 29	7
30 – 39	9
40 – 49	16
50 – 59	21
60 – 69	14
70 – 79	9

80 – 89	4
90 – 99	3
100 – 109	1
sum	84

To find the variance with the above formula, first we find the average of data. Table 4.3 shows the process to compute the mean.

Table 4.3: Process of computing mean for the data fee

Fee	f_i	x	fx
20 – 29	7	24,5	171,5
30 – 39	9	34,5	310,5
40 – 49	16	44,5	712
50 – 59	21	54,5	1144,5
60 – 69	14	64,5	903
70 – 79	9	74,5	670,5
80 – 89	4	84,5	338
90 – 99	3	94,5	283,5
100 – 109	1	104,5	104,5
Sum	84		4638

From the Table 4.4, we have $n = \Sigma f = 84$, $\Sigma f_j(x_j - \bar{x})^2 = 27357.14$. So that the value of variance is calculated as,

$$s^2 = \frac{\Sigma f_j(x_j - \bar{x})^2}{n-1} = \frac{27357.14}{84-1} = 329.60.$$

Table 4.4: Process of computing variance of the data fee

Mid point	fj	X_j - $\bar{x}$	$(x_j - \bar{x})^2$	$f_j(x_j - \bar{x})^2$
24,5	7	-30,71	943,10	6601,72
34,5	9	-20,71	428,90	3860,14

44,5	16	-10,71	114,70	1835,37
54,5 $\bar{x} = 55.21$	21	-0,71	0,50	10,58
64,5	14	9,29	86,30	1208,26
74,5	9	19,29	372,10	3348,94
84,5	4	29,29	857,90	3431,62
94,5	3	39,29	1543,70	4631,11
104,5	1	49,29	2429,50	2429,50
sum	84			27357,14

The computation processes for finding the variance is so complicated, because first we must be calculated mean and then next to calculate variance. Now we use the next formula of variance direct without first to calculate the value of mean. Table 4.5 shows the process of finding variance as follow:

Table 4.5: Computing variance of fee with second formula

Mid point	f_i	x_i^2	$f_i x_i$	$f_i x_i^2$
24,5	7	600,25	171,5	4201,75
34,5	9	1190,25	310,5	10712,25
44,5	16	1980,25	712,0	31684,00
54,5	21	2970,25	1144,5	62375,25
64,5	14	4160,25	903,0	58243,50
74,5	9	5550,25	670,5	49952,25
84,5	4	7140,25	338,0	28651,00
94,5	3	8930,25	283,5	26790,75
104,5	1	10920,25	104,5	10920,25
sum	84		4638,0	283441,00

With the Table 4.5, we have $n = \Sigma f = 84$, $\Sigma f_j x_j^2 = 283441$, $\Sigma f_j x_j = 4638$.

So that the value of variance can be evaluated as,

$$s^2 = \frac{\Sigma f_j x_j^2 - (\Sigma f_j x_j)^2}{n(n-1)}$$

$$s^2 = \frac{84(283441.00) - (4638.0)^2}{84(84-1)} = 329.60 .$$

The computation processes for finding the variance here are large enough. Now we calculate the value of variance with coding of data (see chapter 3). The formula of variance with coding data is defined as follow,

$$s^2 = i^2 \left(\frac{n\Sigma f_j c_i^2 - (\Sigma f_j c_j)^2}{n(n-1)} \right).$$

where c_i coding of data, *i* size of class interval.

The work with this formula is a little bit complicated, but it does not need the large number.

We continue to work with the data from Table 4.2 to compute the value of variance. Table 4.6 shows the process coding data to compute the value of variace, as fallow:

Table 4.6: Compute variance with coding data

income	f_i	c_i	c_i^2	f_ic_i	$f_ic_i^2$
20 – 29	7	-3	9	-21	63
30 – 39	9	-2	4	-18	36
40 – 49	16	-1	1	-16	16
50 – 59	21	0	0	0	0
60 – 69	14	1	1	14	14
70 – 79	9	2	4	18	36
80 – 89	4	3	9	12	36
90 – 99	3	4	16	12	48
100 – 109	1	5	25	5	25
Sum	84			6	274

From the data Table 4.6, we have i=10, $\Sigma f_j c_j^2 = 274$, $\Sigma f_j c_j = 6$ and

$n = \Sigma f_j = 84.$

So that the value of variance can be calculated as,

$$s^2 = i^2\left(\frac{n\Sigma f_j c_i^2 - (\Sigma f_j c_j)^2}{n(n-1)}\right)$$

$$s^2 = 10^2\left(\frac{84(274) - 6^2}{84(84-1)}\right) = 329.60.$$

The three formulas of variance have been presented that all of them give the same result. To calculate the value of standard deviation is only take the square root of variance. Standard deviation is found:

$$s = \sqrt{s^2} = \sqrt{329.60} = 18.16.$$

D. Skewness and Kurtosis

Skewness and Kurtosis a fundamental task in many statistical analyses is to characterize the *location* and *variability* of a data set. A further characterization of the data includes skewness and kurtosis.

Now we should be able to look at real data sets and spot the three measures of central tendency. Use this activity to examine different variables. In chapter 3 we have discussed about histogram and polygon frequency. By connecting the midpoints of the bars with lines, we produce a frequency polygon. The frequency polygon displays the same information as the histogram, but in a different form. The graphic image of a histogram or frequency polygon tells us at a glance the group profile of the data. The incomprehensibility of a set of numbers is transformed into a meaningful visual portrait. This visual portrait displays two special characteristics: *skewness* and *kurtosis.*

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

In a normal distribution, the mean, median, and mode are all the same. In various other symmetrical distributions it is possible for the mean and median to be the same even though there may be several modes, none of which is at the mean. By contrast, in asymmetrical distributions the mean and median are *not* the same. Such distributions are said to be **skewed**, i.e., more than half the cases are either above or below the mean.

Definition of Skewness For univariate data $y_1, y_2, \ldots, y_n$ the formula for skewness is:

$$sk = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^3}{(n-1)s^3}$$

where $\bar{y}$ is the mean, s is the standard deviation, and n is the number of data points.

In case of data given in the form of a frequency distribution where the variate values $x_1, x_2, \ldots, x_n$ accur $f_1, f_2, \ldots, f_n$ times respectively, the formula for skewness is,

$$sk = \frac{\sum_{i=1}^{n} f(x_i - \bar{x})^3}{(n-1)s^3} .$$

The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly skewed right means that the right tail is relatively long to the left tail. Some measurements have a lower bound and are skewed right. For example, in reliability studies, failure times cannot be negative.

Alternative definition, Pearson defined the coefficient of skewness as,

$$C_{skew} = \frac{\bar{x} - Mo}{s} \text{ or the second formula } C_{skew} = \frac{3(\bar{x} - Q_2)}{s}.$$

Example 4.6: We use the data from table 3.6 about presenting students in three months, and reshow in table 4.7 as follow,

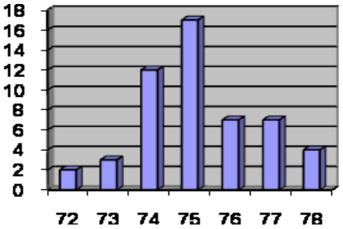
Table 4.7: Presenting students for three months

score	sum of student
72	2
73	3
74	12
75	17
76	7
77	7
78	4

Find the skewness with the all upper formulas.

Solution: First we will show the data in bar chart, then we can interpret the trend of data. Figure 4.1 show the bar chart of the data. With this figure we can predict that the value of skewness is not so far with the zero, because the diagram looks like a distribution normal.

Figure 4.1: Presenting students of three months



Now we calculate the value of skewness with the upper formulas. Table 4.8 shows the process of calculating of central tendency first.

Table 4.8: Process calculating central tendency

x	f	F	fx	$x_i-\bar{x}$	$f(x_i-\bar{x})^2$	$f(x_i-\bar{x})^3$	$f(x_i-\bar{x})^4$
72	2	2	144	-3.17	20.14	-63.90	202.75
73	3	5	219	-2.17	14.17	-30.79	66.90
74	12	17	888	-1.17	16.51	-19.37	22.72
75	17	34	1275	-0.17	0.51	-0.09	0.02
76	7	41	532	0.83	4.79	3.96	3.27
77	7	48	539	1.83	23.36	42.68	77.98
78	4	52	312	2.83	31.97	90.37	255.46
sum	52		3909		111.44	22.87	629.09

The arithmetic mean, mode and standard deviation are computed as bellow,

$$\bar{x}=\frac{\Sigma f_i x_i}{n}=\frac{3909}{52}=75.173$$

$$M_0 = x_{\max(f_i)}=75$$

$$s^2=\frac{\Sigma f_i(x_i-\bar{x})^2}{n-1}=\frac{111.44}{51}=2.185.\text{ and } s=\sqrt{s^2}=\sqrt{2.185}=1.478.$$

Now we calculate the value of median. The number n/2=26, the smallest cumulative frequency which contains 26 is 34, so that the median $=Q_2=75$.

We calculate the value of skewness with the all formulas of skewness as follows:

$$sk = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^3}{(n-1)s^3} = \frac{22.87}{51(1.478)^3} = 0.138,$$

$$C_{skew} = \frac{\bar{x} - Mo}{s} = \frac{75.173 - 75}{1.478} = 0.117,$$

$$C_{skew} = \frac{3(\bar{x} - Q_2)}{s} = \frac{3(75.173 - 75)}{1.478} = 0.351.$$

The value of skewness are small positive, so that we can conclude that the data are skewed small to the right.

Skewness and Central tendency

Now we learn more the relation between skewness central tendency. A distribution which has same value of mean median and mode (i.e. mean = median = mode) is known as a symmetrical distribution. But, not all symmetrical distributions have the same value of mean, median and mode. When values of mean, median and mode are not equal to each other the distribution is known as asymmetrical or skewed.

If the mean > median > mode, the distribution will be skewed to the right or positively skewed. If the mean < median < mode, the distribution will be skewed to the left or negatively skewed. In moderately skewed or asymmetrical distribution a very important relationship exists among these three measures of central tendency.

Example 4.7: The score test of chemistry of a district is given in Table 4.9.

Table 4.9: Scores in a Chemistry Test

score :	5	6	7	8	9	10
Freq :	3	5	8	10	13	14

We can find the values of mean, median and mode by the calculation with Table 4.10. The mean arithmetic and mode are computed as bellow,

$$\bar{x} = \frac{\Sigma f_i x_i}{n} = \frac{438}{53} = 8.26,$$

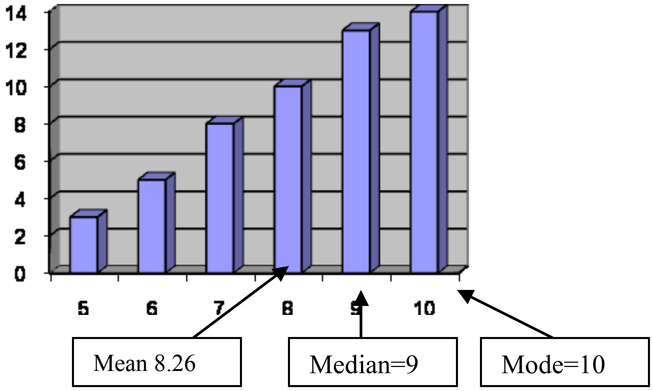
$$M_0 = x_{\max(f_i)} = 10 .$$

Now we calculate the value of median. The number n/2=26.5, the smallest cumulative frequency which contains 26.5 is 39, so that the median $= Q_2 = 9$.

Table 4.10: Process calculating central tendency and skewness of score test of Chemistry

x	f	F	fx	$x_i - \bar{x}$	$f(x_i - \bar{x})^2$	$f(x_i - \bar{x})^3$
5	3	3	15	-3.264	31.964	-104.335
6	5	8	30	-2.264	25.631	-58.034
7	8	16	56	-1.264	12.784	-16.162
8	10	26	80	-0.264	0.697	-0.184
9	13	39	117	0.735	7.039	5.179
10	14	53	140	1.735	42.184	73.225
sum	53		438		120.302	-100.31

A bar chart Figure 4.2 presents the given data score test of Chemistry and the position value of mean, median and mode.



Mean 8.26

Median=9

Mode=10

Figure 4.2: Score Chemistry Test

The bar graph above is an illustration of a special kind of data distribution. The distribution has the property that as the values increase the frequencies increase as well. This means that on the bar graph, the columns get taller as we look from left to right. In such a distribution, we say that the data is **skewed to the left (negative skew).**

Now, we compare the value of central tendency with the value of skewness. From the calculation in Table 4.10 we find the value of standard deviation and skewness as follows:

$$s^2 = \frac{\Sigma f_i (x_i - \bar{x})^2}{n-1} = \frac{111.44}{51} = 2.185. \text{ and } s = \sqrt{s^2} = \sqrt{2.185} = 1.478.$$

$$sk = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^3}{(n-1)s^3} = \frac{-100.31}{52(1.52)^3} = -0.548.$$

The value of skewness is negative, and it appropriate with the information from the value of tendency central.

In the example 4.7 we saw that for a specific distribution that was skewed to the left, the mode (10) was the greatest of the three measures of central tendency, the mean (8.26) was the least of the three measures of central tendency, and the median was in between. This illustrates a typical property of data that is skewed to the left. This relationship is summarized as follows: *mode* > *median* > *mean* (Note: there may be exceptions to this trend.).

Example 4.8: Now we compare the skew data with the illustration from these data table 4.11 about score of a sport competition.

Table 4.11: Score a Football Competition

Score :	0	50	75	100	150
Freq :	14	13	10	8	4

We can calculate the values of mean, median and mode by the calculation with Table 4.12.

Table 4.12: Process calculating central tendency and skewness of score a football competition

x	f	F	fx	$x_i - \bar{x}$	$f(x_i - \bar{x})^2$	$f(x_i - \bar{x})^3$
0	14	14	0	-57.143	45714.286	2612244.898
50	13	27	650	-7.143	663.265	-4737.609
75	10	37	750	17.857	3188.776	56942.420
100	8	45	800	42.857	14693.878	629737.609
150	4	49	600	92.857	34489.796	3202623.907
sum	49		2800		98750.000	1272321.429

The mean arithmetic and mode are computed as bellow,

$$\bar{x} = \frac{\Sigma f_i x_i}{n} = \frac{2800}{49} = 57.1,$$

$$M_0 = x_{\max(f_i)} = 0 .$$

Now we calculate the value of median. The number n/2=24.5, the smallest cumulative frequency which contains 24.5 is 50, so that the median $= Q_2 = 50$.

A bar chart Figure 4.3 presents the given data score of football competition and the position value of mean, median and mode.

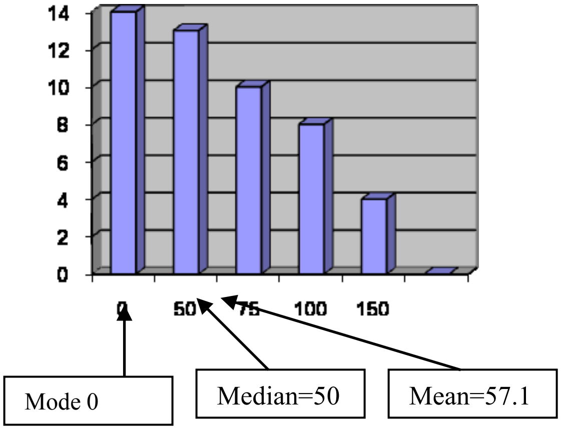


Figure 4.3: Score Football Competition

The bar graph above is an illustration of a special kind of data distribution. The distribution has the property that as the values decrease the frequencies decrease as well. This corresponds to the fact that on the bar graph, the bars get shorter as we look from left to right. A distribution that has those properties is called **skewed to the right (positive skew)**.

Now, we compare the value of central tendency with the value of skewness. From the calculation in Table 4.11 we calculate the value of standard deviation and skewness as follows:

$$s^2 = \frac{\Sigma f_i (x_i - \bar{x})^2}{n-1} = \frac{98750.000}{48} = 57.14. \qquad \text{and}$$

$$s = \sqrt{s^2} = \sqrt{2057.292} = 45.35.$$

$$sk = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^3}{(n-1)s^3} = \frac{1272321.428}{48(45.35)^3} = 0.284.$$

The value of skewness is positive, and it appropriate with the information from the value of tendency central.

In the upper example 4.8 we saw that for data skewed to the right, the three measures of central tendency had this numerical relationship: the mode (0) was the least, the mean (57.1) was the greatest, and the median was in between. This illustrates the following typical property of data skewed to the right. This relationship is summarized as follows: *mean* > *median* > *mode.* (Note: there may be exceptions to this trend.).

Example 4.9: The score of Aptitude test of a district is given in Table 4.13.

Table 4.13: Score of Aptitude test of a district

Value	Frequency
2	24
3	30
4	36
5	30
6	24

We can find the values of mean, median and mode by the calculation with Table 4.14.

Table 4.14: Process calculating central tendency and skewness of score of aptitude test

x	f	F	fx	$x_i - \bar{x}$	$f(x_i - \bar{x})^2$	$f(x_i - \bar{x})^3$
2	24	24	48	-2	96	-192
3	30	54	90	-1	30	-30
4	36	90	144	0	0	0
5	30	120	150	1	30	30
6	24	144	144	2	96	192
sum	144		576		252	0

The mean arithmetic and mode are computed as bellow,

$$\bar{x} = \frac{\Sigma f_i x_i}{n} = \frac{576}{144} = 4,$$

$$M_0 = x_{\max(f_i)} = 4.$$

Now we calculate the value of median. The number n/2=72, the smallest cumulative frequency which contains 72 is 4, so that the median $= Q_2 = 4$.

A bar chart Figure 4.4 presents the given data score of aptitude test and the position value of mean, median and mode.

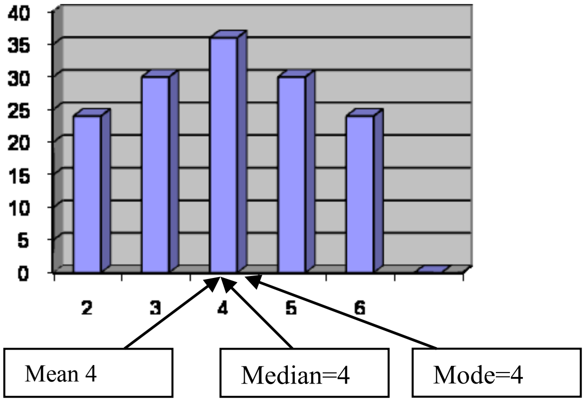


Figure 4.4: Score Aptitude Test

Now, we compare the value of central tendency with the value of skewness. From the calculation in Table 4.14 we calculate the value of standard deviation and skewness as follows:

$$s^2 = \frac{\Sigma f_i (x_i - \bar{x})^2}{n-1} = \frac{252}{143} = .1.76 \text{ and } s = \sqrt{s^2} = \sqrt{1.76} = 1.32.$$

$$sk = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^3}{(n-1)s^3} = \frac{0}{143(1.32)^3} = 0.$$

The value of skewness is zero, and it appropriate with the information from the value of tendency central.

In the previous example 4.9 we saw a distribution whose bar graph was symmetric, with the single highest bar in the middle. This kind of distribution is referred to symmetry of data. When graphed, normally distributed data will typically have an appearance similar to this: When data is normally distributed, the mean, median and mode will all be the same: *mean = median = mode.*

Below are some information that illustrates the relationship between mean, median, and mode in skewed distributions. In each information we will be asked to modify a histogram that satisfies certain conditions.

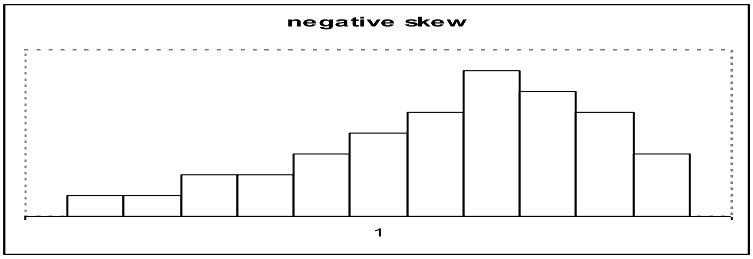
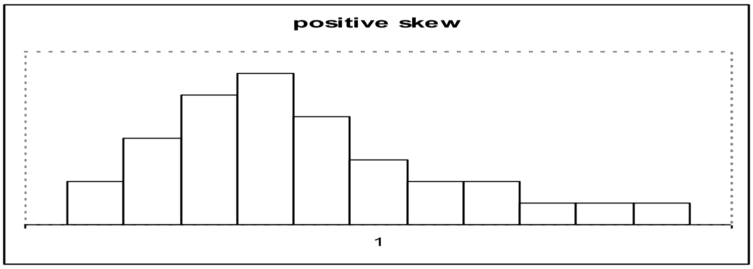
At this point, we should have created a symmetrical distribution, a negatively skewed distribution, and a positively skewed distribution. If we think about the three figures, we can deduce a general rule about the relationship between the symmetry of a distribution of scores and measures of central tendency. The rule is that, as the symmetry of a distribution increases, the three measures of central tendency converge on the same value. As the asymmetry or skewness of a distribution increases, the three measures of central tendency diverge systematically.

For a positively skewed distribution, the mean will always be the highest estimate of central tendency and the mode will always be the lowest estimate of central tendency (assuming that the distribution has only one mode). For negatively skewed distributions, the mean will always be the lowest estimate of central tendency and the mode will be the highest estimate of central tendency.

In any skewed distribution (i.e., positive or negative) the median will always fall in-between the mean and the mode. As previously discussed in the section on "choosing an appropriate measure of central tendency", when dealing with skewed distributions, researchers typically decide between the mean or median as the best estimate of central tendency. As distributions go from symmetrical to more skewed, the researcher is more likely to choose the median over the mean.

The skewness of a graph is in the direction of its “tail”. If the scores bunch up toward the high end, the graph has a **negative skew**. If the scores bunch up toward the low end, the graph has a **positive skew**. A distribution is said to be **normal** (or bell-shaped) if the scores bunch up in the middle and then taper off fairly equally on each side. Finally, a distribution is called a rectangular distribution if the scores are fairly evenly distributed throughout the graph.

The ***skewness*** of a curve describes how shaped or skewed it is. The three basic profiles of skewness are ***negative skew***, ***positive skew***, and ***rectangular distribution***.

The ***skewness*** of a curve describes how horizontally distorted a curve is from the familiar bell-shaped curve. A curve with *negative skew* has its left tail pulled outward to the left, to the negative end of the scale.	
A curve with *positive skew* has its right tail pulled outward to the right, to the positive end of the scale. A common mistake is to focus on the “mound	

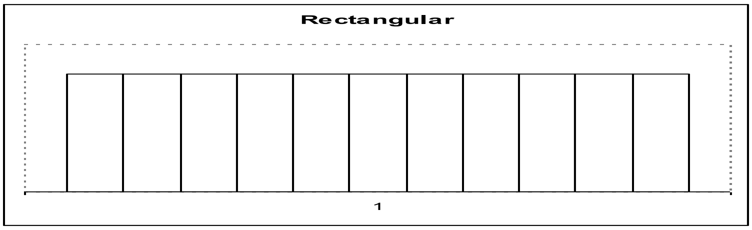
of scores" rather than the distorted tail.	
A distribution where all categories of scores have equal frequency is called a rectangular distribution.	

Figure 4.5: kind of skewness

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case.

Definition of Kurtosis: For non-variant data y_1, y_2, ..., y_n, the formula for kurtosis is:

$$ku = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^4}{(n-1)s^4},$$

where $\bar{y}$ the mean, s is is the standard deviation, and n is the number of data points.

In case of data given in the form of a frequency distribution where the variate values x_1, x_2, ..., x_n accur f_1, f_2, ..., f_n times respectively, the formula for kurtosis is,

$$ku = \frac{\sum_{i=1}^{n} f(x_i - \bar{x})^4}{(n-1)s^4}.$$

If the value of kurtosis $ku \approx 3$, then data are flat relative to a normal distribution. If the value of kurtosis $ku > 3$, then data are “peak” form. If the value of kurtosis $ku < 3$, then data are “flat” form.

Alternative Definition of Kurtosis. The kurtosis for a standard normal distribution is three. For this reason, some sources use the following definition of kurtosis (often referred to as "excess kurtosis"

$$kurtosis = ku = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^4}{(n-1)s^4} - 3.$$

This definition is used so that the standard normal distribution has a kurtosis of zero. In addition, with the second definition positive kurtosis indicates a "peaked" distribution and negative kurtosis indicates a "flat" distribution.

Example 4.10: We use the data example 4.6 to evaluate the value of kurtosis. From the Table 4.8 we have $\bar{x} = .75.173$, $M_0 = 75$, $s = 1.478$. $\sum f_i (x_i - \bar{x})^4 = 629.09$, so that we calculate the value of kurtosis:

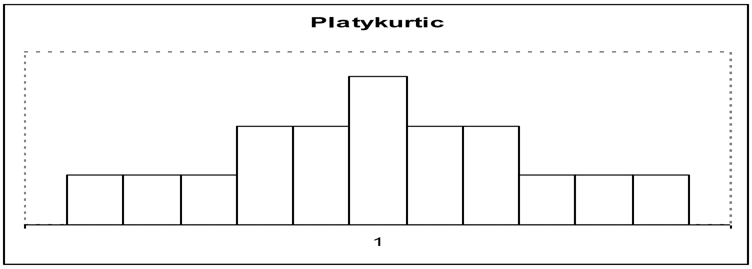
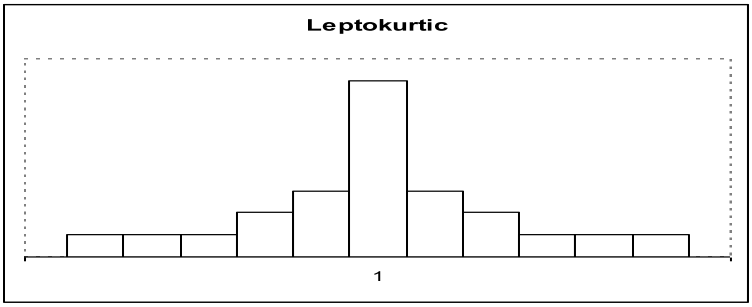
$$ku = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^4}{(n-1)s^4} = \frac{629.09}{51(1.478)^4} = 2.58.$$

And the alternative kurtosis formula can be calculated as,

$$ku = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^4}{(n-1)s^4} - 3 = 2.58 - 3 = -0.42.$$

The values of kurtosis showed that the data have flat distribution.

The ***kurtosis*** of a curve describes how flat or peaked it is. The three basic profiles of kurtosis are ***platykurtic*** (flat), ***leptokurtic*** (peaked), and ***mesokurtic*** (balanced).

A flat curve is called *platykurtic*. Think of the flatness of a plate and you’ll remember “platy-kurtic.” Notice that there are low frequencies for all the categories.	
A peaked curve is called *leptokurtic*. Think of the central frequencies “leaping” away from the others and you’ll remember “leap-tokurtic.” Notice that outer categories have lower frequencies while the central categories have high frequencies.	

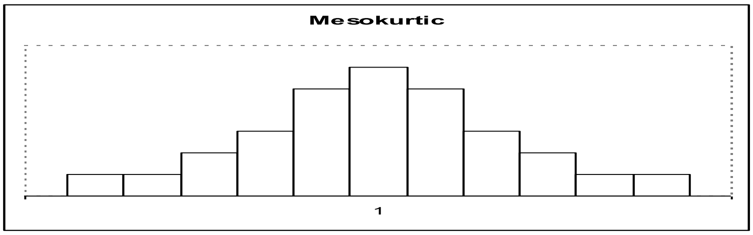
A curve that falls between platykurtic and leptokurtic is called *mesokurtic*. Think of medium (meso-) and you’ll remember mesokurtic. The familiar bell shaped curve is mesokurtic.	

Figure 4.6: kind of kurtosis

E. Spread and Shape data of Box and Whisker Plot

In 1977, John Tukey published an efficient method for displaying a five-number data summary. The graph is called a box plot (also known as a box and whisker plot). The **box plot** is a graphical display that simultaneously describes several important features of a data set, such as center, spread, departure from symmetry, and identification of unusual observations or outliers.

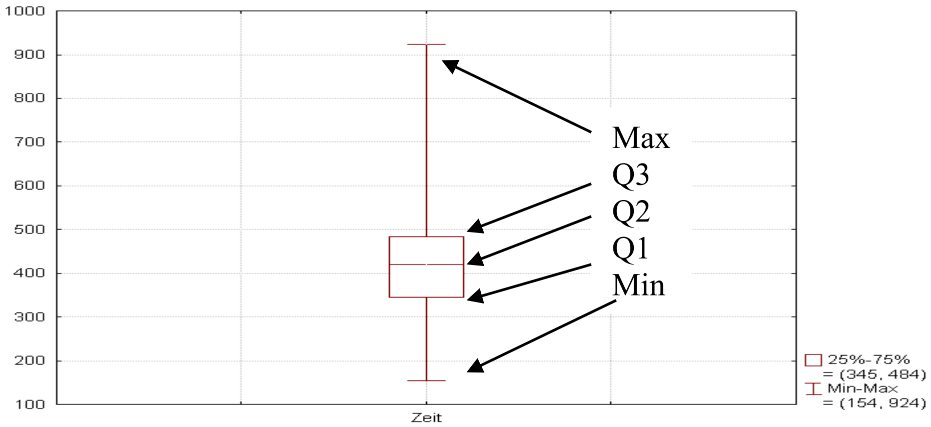
A box plot summarizes the following statistical measures:

1. the smallest observation (sample minimum)
2. lower quartile (Q1)
3. median (Q2)
4. upper quartile (Q3)

5. largest observation (sample maximum),

of the data on a rectangular box, aligned either horizontally or vertically. The box encloses the inter quartile range (IQR=Q_3-Q_1) with the left (or lower) edge at the first quartile, Q1, and the right (or upper) edge at the third quartile, Q3. A line is drawn through the box at the second quartile (which is the 50th percentile or the median), Q_2. A line, or **whisker,** extends from each end of the box. Figure 4.7 shows the expression of the upper information.

Figure 4.7: Box Plot Min, Max, Q1,Q2, and Q3



The lower whisker is a line from the first quartile to the smallest data point within 1.5 inter quartile ranges from the first quartile. The upper whisker is a line from the third quartile to the largest data point within 1.5 inter quartile ranges from the third quartile. Data farther from the box than the whiskers are plotted as individual points. A point beyond a whisker, but less than 3 inter quartile ranges from the box edge, is called an **outlier.** A point more than 3 inter quartile ranges from the box edge is called an **extreme outlier.** See Figure 4.8. Occasionally, different symbols, such as open and filled circles, are used to identify the two types of outliers.

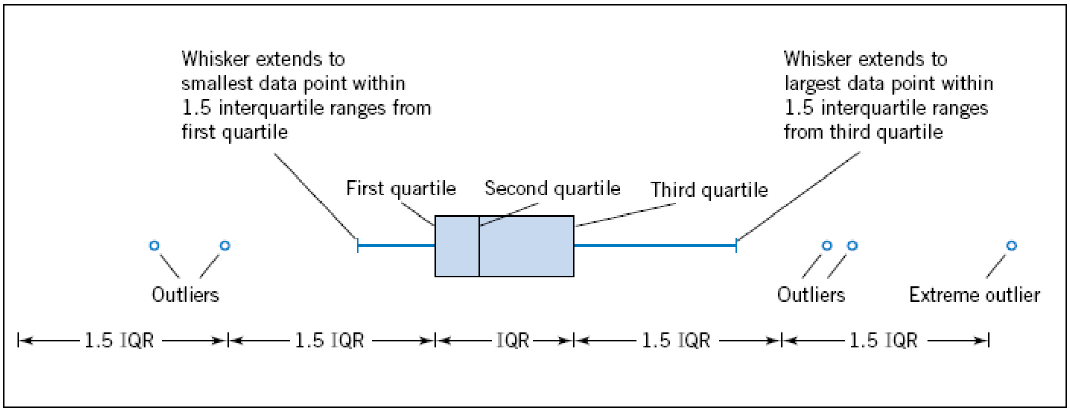

Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

Figure 4.8: Description of a Box Plot

Example 4.11: We use the data in Example 4.7 to make a box and Whisker plot. Table 4.15 shows the process of calculations the value of quartile.

Table 4.15: Process calculating Quartile of score test of Chemistry

x	f	F
5	3	3
6	5	8
7	8	16
8	10	26
9	13	39
10	14	53
sum	53	

Now we calculate the value of quartiles 1 to 3. The number n/4, n/2, and 3n/4 are 13.25, 26.5 and 39.75 respectively. The smallest cumulative frequency which contains 13.25, 26.5 and 39.75 are 16, 39 and 53 respectively. So that the value of Q_1=7, Q_2=9 and $Q_3 = 10$. And we have also value of minimum 5, maximum 10. Figure 4.9 presents the diagram of Box and Whisker Plot.

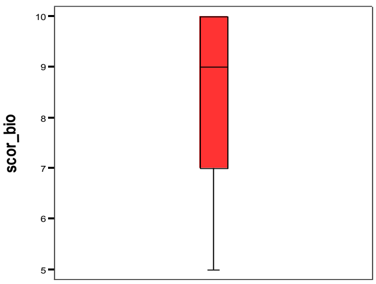


Figure 4.9: Box and Whisker Plot for score- Biology

The diagram shows that the data skewed to the left.

The boxplot is interpreted as follows:

1. Box and whisker are uniform in their use of the box: the bottom and top of the box are always the 25th and 75th percentile (the lower and upper quartiles, respectively)
2. The band near the middle of the box is always the 50th percentile (the median).
3. If the median line within the box is not equidistant from the hinges, then the data is shaped.
4. The ends of the whiskers indicate the minimum and maximum of all the data, unless outliers are present in which case the whiskers extend to a maximum of 1.5 times the inter-quartile range.
5. The points outside the ends of the whiskers are outliers or suspected outliers.

Any data outlier should be plotted as with a dot, small circle, or star, but occasionally this is not done. Some box plots include an additional dot or a cross is plotted inside of the box, to represent the mean of the data in addition to the median.

On some box plots, a crosshatch is placed on each whisker, before the end of the whisker. Quite rarely, box plots can be presented with no whiskers at all. Because of this variability, it is appropriate to describe the convention being used for the whiskers and outliers in the caption for the plot.

The unusual percentiles 2%, 9%, 91%, 98% are sometimes used for whisker cross-hatches and whisker ends to show the seven-number summary. If the data are normally distributed the locations of the seven marks on the box plot will be equally spaced.

Visualization

The box plot is a quick way of examining one or more sets of data graphically. Box plots may seem more primitive than a histogram, but they do have some advantages. They take up less space and are therefore particularly useful for comparing distributions between several groups or sets of data. Choice of number and width of bins techniques can heavily influence the appearance of a histogram, and choice of bandwidth can heavily influence the appearance of an estimate.

F. Exercise

1. Mansion the important measurements of dispersion or variability amongst observations in the set.
2. Define the following and give one appropriate example of your own for the use of each:
   a. Range
   b. Mean deviation
   c. Variance
3. We have a data from the result of the examination: 74, 81, 56, 96, 63, 55, 91, 93, 85, 51, 95, 69. Compute:

a. Coefficient of range

b. Quartile deviation

c. Coefficient of quartile deviation

d. Variance

e. Standard deviation.

4. Thirty students were tested to determine how long they used the time to run away from place A to place B. The results, to the nearest minute, were listed as follows: 423, 369, 387, 411, 393, 394, 371, 377, 389, 409, 392, 408, 431, 401, 363, 391, 405, 382, 400, 381, 399, 415, 428, 422, 396, 372, 410, 419, 386, 390. Use the frequency distribution table in exercise chapter 3 number 4, than find:

   a. Coefficient of range

   b. Quartile deviation

   c. Coefficient of quartile deviation

   d. Variance

   e. Standard deviation.

5. Make the Box Plot for the data number 4

6. Find the variance, standard deviation and coefficient of variance for the following distribution.

x_i	f_i
75	8
60	7
92	8
64	7
35	2

7. Make a diagram box plot for the data number 6.

8. The result of test examination was given as follow:

Nm	score	f
1		2
2		3
3	53 – 61	5

Compute:

e. Variance

f. Standard Deviation

g. Coefficient of variance

4	62 – 70	10
5		15
6		18
7		9
8		3

9. The information regarding the no of children per family is given in the following table. Find the variance and standard deviation of the data

No of children	0	1	2	3	4	5
No of families	3	20	15	8	3	1

10. Define the following of each:

    a. Skewness

    b. Kurtosis

11. From the data number 4, find the value of skewness and kurtosis. What is your interpret on it.

12. From the data number 9 find the value of skewness and kurtosis. What is your interpret on it.

CHAPTER V

TRANSFORMATION DATA AND TABLE DISTRIBUTION

F. Description and Basic Competence

The description consist of the knowledge about conversion data from the origin data of mean and standard deviation to the new mean and standard deviation, and reading the score table distribution.

The main instructional objectives are after the learning process students are able to:

1. convert data from original to standardized data
2. convert data from original to new mean and standard deviation
3. read the value of z-score
4. read the value of t-score
5. read the value of F-score
6. read the value of r-score.

G. Transforming data

1. Transform data from origin to standardized score

Anytime a distribution is changed by using a constant, it performs a linear transformation. We need to know how linear transformations affect the mean and standard deviation of a distribution as well. How does adding or subtracting a constant affect the mean and standard deviation? How does multiplying and dividing a constant affect the mean and standard deviation?

In the following example, we add a constant and see the changes to the mean and standard deviation.

Table 5.1: Adding Variable with constant 5

x	x+5
1	6
2	7
3	8
4	9
5	10
$\bar{x}=3,$ $s=1.41$	$\bar{x}=8,$ $s=1.41$

When adding or subtracting a constant from a distribution, the mean will be changed by the same amount as the constant. However, the standard deviation will remain unchanged. This fact is true because, again, we are just shifting the distribution up or down the scale. We do not affect the distance between values.

Table 5.2: Multiplying variable with 5

x	X*5
1	5
2	10
3	15
4	20
5	25
$\bar{x}=3,$ $s=1.41$	$\bar{x}=15,$ $s=7.91$

The effect is a little different when we multiply or divide it by a constant. For these transformations the mean will be changed by the same amount as the constant, but this time the standard deviation will change too. That is because when we multiply numbers together, for example, we change the distance between values rather than just shifting them up or down the scale.

The transformation is a linear transformation, as we have discussed previously. Transforming a raw score to a z-score will yield a number that expresses exactly on how many deviations from the mean a score lays. Here is the formula for transforming a raw data into z scores entails subtracting the mean from the row data and dividing by the standard deviation,

$z_i = \dfrac{x_i - \bar{x}}{s}$, x_i score ith of the data

$\bar{x}$: mean of the data

s : standard deviation.

Example 5.1: We have the data 47, 56, 75, 65, 71, 54, 53, 61, 77 81. Transform this data in standard score.

Solution: First we must compute the arithmetic mean and the standard deviation of the data. With the formula in chapter 3 we find mean $\bar{x} = 64$ and standard deviation s= 11.6. The standard scores can then be computed as follow,

$$z_1 = \frac{x_1 - \bar{x}}{s} = \frac{47 - 64}{11.6} = -1.47$$

$$z_2 = \frac{x_2 - \bar{x}}{s} = \frac{56 - 64}{11.6} = -0.69$$

$$z_3 = \frac{x_3 - \bar{x}}{s} = \frac{75 - 64}{11.6} = 0.95$$

$$z_4 = \frac{x_4 - \bar{x}}{s} = \frac{65 - 64}{11.6} = 0.09$$

$$z_5 = \frac{x_5 - \bar{x}}{s} = \frac{71 - 64}{11.6} = 0.60$$

$$z_6 = \frac{x_6 - \bar{x}}{s} = \frac{54 - 64}{11.6} = -0.86$$

$$z_7 = \frac{x_7 - \bar{x}}{s} = \frac{53 - 64}{11.6} = -0.95$$

$$z_8 = \frac{x_8 - \bar{x}}{s} = \frac{61 - 64}{11.6} = -0.26$$

$$z_9 = \frac{x_9 - \overline{x}}{s} = \frac{77 - 64}{11.6} = 1.12$$

$$z_{10} = \frac{x_{10} - \overline{x}}{s} = \frac{81 - 64}{11.6} = 1.47.$$

If we calculate the mean and standard deviation of standard score, we will find mean $\overline{x} = 0$ and standard deviation =1 (prove it!).

Interpret the meaning of a z-score.

- A z-score less than 0 represents an element less than the mean.
- A z-score greater than 0 represents an element greater than the mean.
- A z-score equal to 0 represents an element equal to the mean.

Using z-scores to compare values

If we compare two or more variables which have different scales, it is easy to convert the data in the standard score. For example: Andre gets a 66, 70 and 60 on his Statistics, English and Sports tests, respectively. We cannot say that Andre get the best score in English. In one case, maybe Andre's English score is 10 points above the mean, in other case, Andre's English score is 10 points bellow the mean. In an important sense, we must interpret Andre's grade relative to the average performance of the class. We will know exactly the position of Andre's score, if the other information is given like mean and standard deviation of each variable.

Suppose that the scores for all three of the tests were normally distributed and that

a. the distribution of the Statistics scores had a mean of 60 and a standard deviation of 6,
b. the distribution of English test scores had a mean of 75 and a standard deviation of 5,

c. the distribution of the Sport scores had a mean of 40 and a standard deviation of 10

Now, what statements can be made regarding Andre's relative performance on each of these three tests? Andre did the best on the English test; his raw score of 70 falls at a point one standard deviation above the mean. Andre's next best score was on the sport test; his raw score of 40 falls exactly at the mean of the distribution of scores. And finally there is Andre's performance on the sports test; his raw score of 66 falls at a point one standard deviation below the mean. Converting Andre’s raw scores to z scores, a scale that has a mean of 0 and a standard deviation of 1 will help to compare.

Statistics score: x=66, $\bar{x} = 60, s = 6 \Rightarrow z_{st} = \frac{66-60}{6} = 1,$

English score: x=70, $\bar{x} = 6, s = 6 \Rightarrow z_{In} = \frac{70-75}{5} = -1,$

Sports score: x=60, $\bar{x} = 40, s = 10 \Rightarrow z_{sp} = \frac{60-40}{10} = 2.$

So that, Andre’s score in Statistics has 1 step standard deviation higher of mean, in English has 1 step standard deviation lower of mean and in Sports has 2 steps higher of mean.

Transform data from original to new mean and standard deviation

While *z* scores are relatively simple to use. They have some computational disadvantages because a *z* score can be equal to 0 or can be negative, and certain types of data manipulation become awkward. Also, many examinees are disturbed by hearing test scores reported as negative numbers. For these reasons, as well as others, alternative standard score systems have been developed to linearly transform *z* scores (as well as raw scores) to a scale that does not contain negative numbers. Such systems are all "standardized" to the extent that both the mean and the standard deviation of the new scale have been arbitrarily set. The general formula for linearly converting a *z* score to a new standard score (*m*) is expressed as follow:

$m_i = x_0 + s_0 z_i$, m: the new standard score

x_0 : the new mean score

s_0: the new standard deviation score.

We will use Andre's scores to show this conversion score. Score variables will be converted to mean 50 and a standard deviation 10. Using the formula presented above, the new standardized score (m) would be calculated as follow:

Statistics score: x_0=50, $s_0 = 10, z = 1 \Rightarrow m_{st} = 50 + 10(1) = 60$,
English score: x_0=50, $s_0 = 10, z = -1 \Rightarrow m_{in} = 50 + 10(-1) = 40$,
Sport score: x_0=50, $s_0 = 10, z = 2 \Rightarrow m_{sp} = 50 + 10(2) = 70$.

Now, we are clear to show Andre's scores in comparison.

Example 5.2: Table 5.3 gives the result of test examination of 8 students in bachelor degree. A student is considered to have passed if he/she has a score of more than 55. Then we try to convert this score to a new standard score in mean 65 and standard deviation 10.

Table 5.3: Score exam for 8 students in master degree

Students	: A	B	C	D	E	F	G	H
Scores	: 3	4	5	3	8	9	2	3

We calculate first mean and standard deviation. With the formula in chapter 3 we find mean $\bar{x}$ = 4.6 and standard deviation s= 2.56. Then we can compute z scores and m scores for the students are shown in table 5.4 follows:

Table 5.4: Converting scores

Nr	Score	z	m
1	3	-0.63	59
2	4	-0.24	63

3	5	0.15	66
4	3	-0.63	59
5	8	1.32	78
6	9	1.71	82
7	2	-1.03	55
8	3	-0.63	59

Discussion:

The minimum score 2 and small score 3 become scores 55 and 66 respectively. It is an advantage value for the range scores 0 to 100, because it makes higher scores. But we can look at score maximum 9 and score 8 become 82 and 78 respectively. It is an disadvantage value for the range scores 0 to 100, because it makes lower scores.

To solve this problem, we try to work first with equality function in algebra. Back to our example we will convert our scores with the requirement that the minimum score become 55 and the maximum score become 95.

The evaluation is running as bellow,

Score 2 $\Rightarrow$ score 55 that mean $m_7 = x_0 + z_7 * s_0 \Rightarrow 55 = x_0 + (-1.03)s_0$ …….(1)

Score 9 $\Rightarrow$ score 95 that mean $m_6 = x_0 + z_6 * s_0 \Rightarrow 95 = x_0 + (1.71)s_0$ ……. (2).

Calculate x_0 and s_0 through the equality functions (1) and (2), we found x_0=70.0 and s_0=14.6. With the new standard score we can show in table 5.5 the all conversion score as follow,

Table 5.5: Converting data to z score and m score (x_0=70.0, s_0=14.6)

Nr	**score**	**z**	**m**
1	3	-0.63	61
2	4	-0.24	66
3	5	0.15	72
4	3	-0.63	61
5	8	1.32	89

6	9	1.71	95
7	2	-1.03	55
8	3	-0.63	61

Now, we can look at the new score better than the last one.

C. Table of z scores and Problems

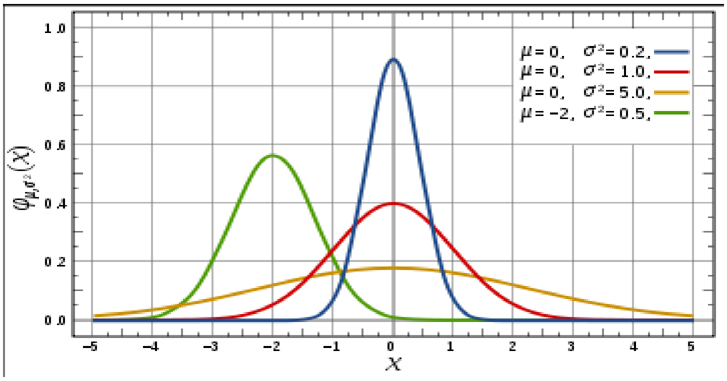
The most important distribution in statistics is the normal distribution, the distribution of a continuous random variable. Let a variable X has normal distribution. The normal distribution has the formula:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} ,$$

where e is the base of natural logarithms, μ is a mean population, and σ^2 is a variant population.

There is an infinite number of normal distributions. Four examples of the normal distribution with the difference mean and variance are show in Figure 5.1.

Figure 5.1: Norman distribution with the difference mean and variance



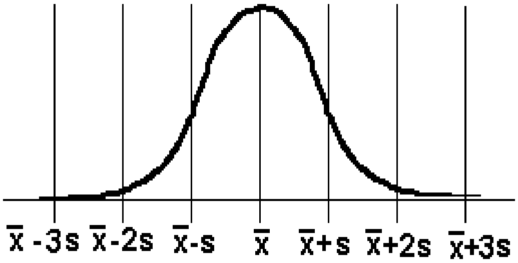
Every normal distribution has a bell-shaped curve; some normal distributions have a curve which is tall and narrow; while others have a curve which is short and wide. The exact shape of a normal distribution is determined by its mean and standard deviation.

Notice that for a standard normal distribution, μ = 0 and σ^2 = 1. The last part of the equation above shows that any other normal distribution can be regarded as a version of the standard normal distribution that has been stretched horizontally by a factor σ and then translated rightward by a distance μ. Thus, μ specifies the position of the bell curve's central peak, and σ specifies the "width" of the bell curve. The formula of standard normal distribution becomes:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}},$$

Figure 5.2 shows the normal standard distribution:

Figure 5.2: Normal standard distribution



Here are the properties in standard normal distribution:

- The total area under the normal curve is equal to 1.
- The probability that a normal random variable *X* equals any particular value is 0.

- The probability that *X* is greater than *a* equals the area under the normal curve bounded by *a* and plus infinity (as indicated by the *non-shaded* area in the figure 5.3).
- The probability that *X* is less than *a* equals the area under the normal curve bounded by *a* and minus infinity (as indicated by the *shaded* area in the figure 5.3).

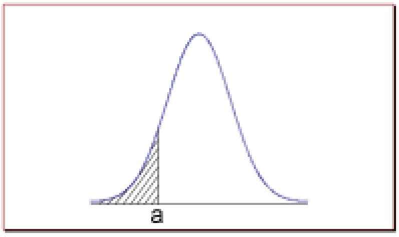


Figure 5.3: Area under the standard normal curve

Z scores are a special application of the transformation rules. The z score for an item, indicates how far and in what direction, that item deviates from its distribution's mean, expressed in units of its standard deviation distribution. The mathematics of the z score transformation is that if every item in a distribution is converted to its z score, the transformed scores will necessarily have a mean of zero and a standard deviation of one.

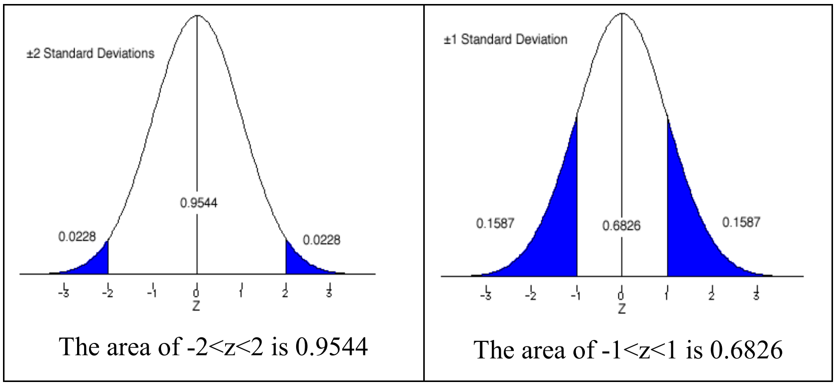
Z scores are sometimes called "standard scores". The z score transformation is especially useful when seeking to compare the relative standings of items from distributions with different means and/or different standard deviations.

Z scores are especially informative when the distribution to which they refer, is normal. In every normal distribution, the distance between the mean and a given z score cuts off a fixed proportion of the total area under the curve. Statisticians have provided us with tables indicating the value of these proportions for each possible z score.

The standard normal distribution has a mean of zero (mean = 0) and a variance of one ($s^2 = 1$) thus, of course, the standard deviation is one ($s = 1$). Since most distributions are observed rather than generated mathematically, there are features of normal distributions that can also be observed.

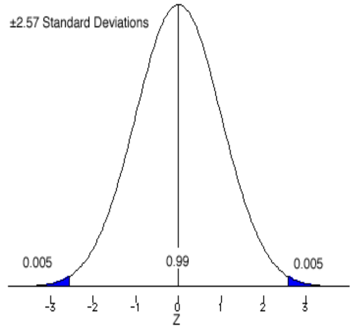
The area above the interval from z=0 to z=2 is 0.4772 and the area from z=0 to z=3 is 0.4987. These areas, and others, are listed in a table (Appendix Table I). If we double each of these numbers, we obtain the following important results. In a normal distribution 68.26% of the distribution will fall between plus and minus one standard deviation (±1.0s). Usually, this will be denoted using the standard normal (z-score) distribution as ±1.0z. Here are some other values which form the standard normal distribution, 95.44% between ±2.0z and 99.74% between ±3.0z. Figure 5.4 shows some example areas under the standard normal curve.

Figure 5.4: Areas under the standard normal curve

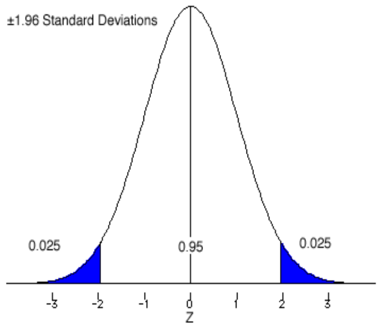


The area of -2<z<2 is 0.9544

The area of -1<z<1 is 0.6826



The area of -2.6<z<2.6 is 0.99



-1.96<z<1.96 area 0.95

The easy way to find the areas (probabilities) under the standard normal curve is to use the tables found in every statistics book. These tables can be formatted in several different ways. We will use a method displayed in Figure 5.5 specifically in the area between the mean (z = 0.0) and a given z-score.

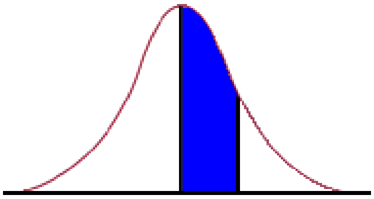
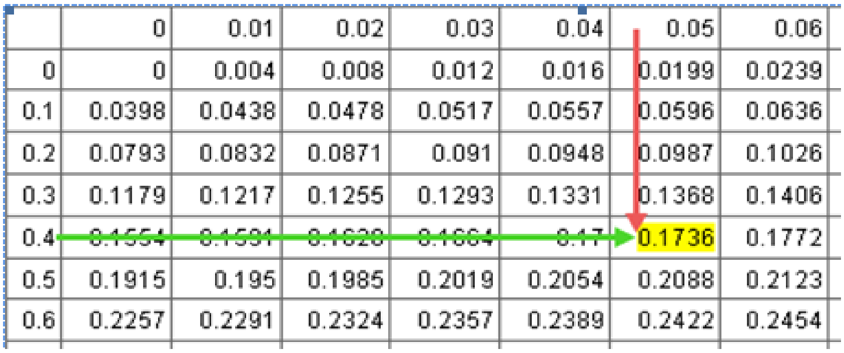


Figure 5.5: Areas between 0 & Z of the Standard Normal Distribution

To read the standard normal table (see Appendix Table 1), first find the row corresponding to the leading significant digit of the z-value in the column on the left hand side of the table. Then after locating the appropriate row, move to the column which matches the next significant digit.

Example: If our z-score = 0.45. Table 5.6 (part of Appendix Table I) shows the finding the value of area.

Table 5.6: Cumulative standard normal distribution



	0	0.01	0.02	0.03	0.04	0.05	0.06
0	0	0.004	0.008	0.012	0.016	0.0199	0.0239
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454

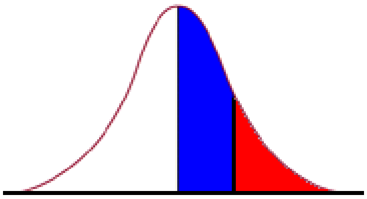
Follow the rows down to 0.4 and then across the columns to 0.05. The area is the highlighted box with a value of 0.1736.

Some Example z-score Problems

The 820 Students of grade 8 in a school in district took a standardized social studies test that was normally distributed and had a mean of 340 and a variance of 256. Here are the scores for four of the students: Mery scored 364, Yuta scored 356, Agus scored 344, and Ringo scored 332.

Note: Graphs are designed to give a rough estimation and are not drawn to exact scale or proportion.

1. How many students would be expected to score above Yuta?



First, compute the standard deviation: $s = \sqrt{s^2} = \sqrt{256} = 16$. Convert Yuta's raw score to a z-score:

$$z = \frac{x - \bar{x}}{s} = \frac{356 - 340}{16} = +1.0 .$$

Find area from the mean to a z-score of +1.0: .3413. The area above Yuta is the area from z=+1.0 to the right unlimited. It means the area above: .5000 - .3413 = .1587. Find the number of people: .1587 * 820 = 130.134 or 130 people.

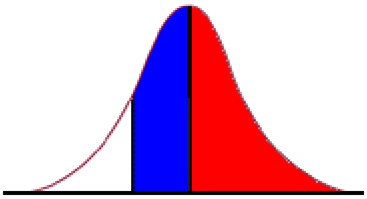
2. What proportion of the students would be expected to score above Mery?

Convert Mery's raw score to a z-score:

$$z = \frac{x - \bar{x}}{s} = \frac{364 - 340}{16} = +1.5.$$

Find area from the mean to a z-score of +1.5: .4332. The area above Mery is the area from z=1.5 to the right unlimited. That mean the area above: .5000 - .4332 = .0668. The proportion score above Mery is 6.68% of the students.

3. How many percents of the students would be expected to score above Ringo?

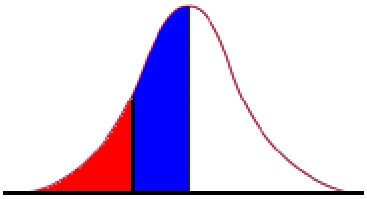


Convert Ringo's raw score to a z-score:

$$z = \frac{x - \bar{x}}{s} = \frac{332 - 340}{16} = -0.5.$$

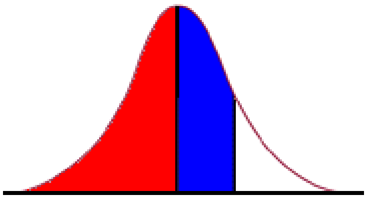
Find area from the mean to a z-score of -0.5: .1915. The area above Ringo is the area form z=-0.5 to the right unlimited. It means the area above: .5000 + .1915 = .6915. Therefore, the students score above Ringo is 69.15%.

4. How many students would be expected to score below Ringo?



Ringo's z-score: z = -0.5. Find area from the mean to a z-score of -0.5: .1915. Find the area below: .5000 - .1915 = .3085 or using information from 3: 1.0 - .6915 = .3085. Find number of people: .3085 * 820 = 252.97 or 253 people.

5. How many percents of students would be expected to score below Agus?



Convert Agus's raw score to a Z-score:

$$z = \frac{x - \bar{x}}{s} = \frac{344 - 340}{16} = +0.25\,.$$

Find the area from the mean to a Z-score of +0.25: .0987. The area below Agus is the area form z=-0.5 to the left negative unlimited. It means the area below: .5000 + .0987 = .5987. Convert to percent: 59.87%.

6. A z-score of 1.7 was found from an observation coming from a normal distribution with mean 14 and standard deviation 3. Find the raw score.

Solution: We have

$$1.7 = \frac{x - 14}{3}$$

To solve this we just multiply both sides by the denominator 3,

(1.7)(3) = x – 14

5.1 = x – 14

x = 19.1.

D. Table of t-score and problems

Let X$_1$, X$_2$, … X$_n$ *be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. The random variable:*

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}},$$

has a *t* distribution with n-1 degrees of freedom.

Several t distributions are shown in Figure 5.6. The general appearance of the t *distribution is similar to the standard normal distribution in that both distributions are symmetric and unimodal. As the number of degrees of freedom, the limiting form of the* t *distribution is the standard normal distribution. Generally, the degrees numbers of freedom for t are the degrees numbers of freedom associated with the estimated standard deviation.*

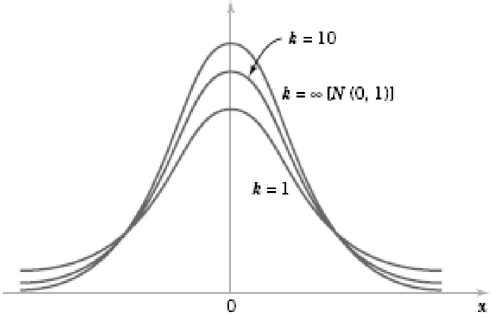


Figure 5.6: Probability density function of several distribution t

Appendix Table II provides **percentage points** of the t distribution. We will let t$_{\alpha,v}$ be the value of the random variable *T* with *v* degrees of freedom above which we find an area (or probability) α. Thus,

$t_{\alpha,v}$ is an upper-tail 100α percentage point of the *t* distribution with *v* degrees of freedom. This percentage point is shown in Figure 5.7.

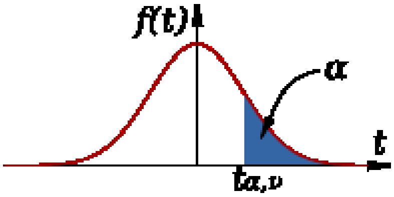
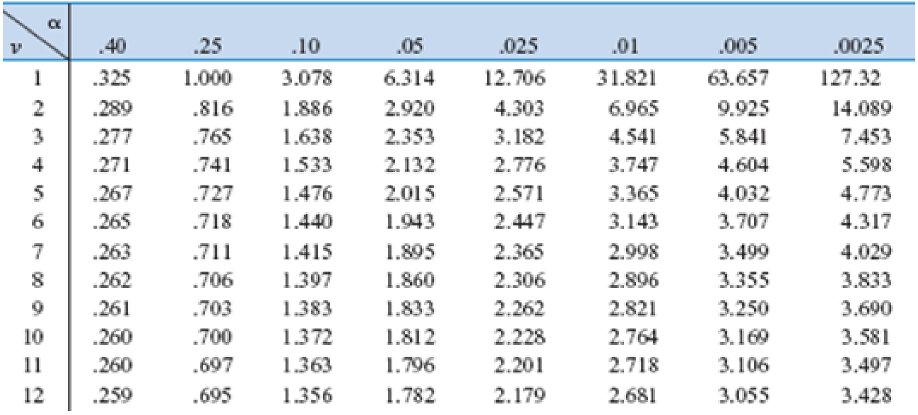


Figure 5.7: Percentage points of the t distribution

In the Appendix Table II theα values are the column headings, and the degrees of freedom are listed in the left column. To illustrate the use of the table look at Table 5.7 (part of Appendix Table II). The t-value with 10 degrees of freedom having an area of 0.05 to the right is $t_{0.05,10}$ =1.812.

Table 5.7: Percentage Points $t_{\alpha,k}$ of the *t*-Distribution



α ν	.40	.25	.10	.05	.025	.01	.005	.0025
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32
2	.289	.816	1.886	2.920	4.303	6.965	9.925	14.089
3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598
5	.267	.727	1.476	2.015	2.571	3.365	4.032	4.773
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.029
8	.262	.706	1.397	1.860	2.306	2.896	3.355	3.833
9	.261	.703	1.383	1.833	2.262	2.821	3.250	3.690
10	.260	.700	1.372	1.812	2.228	2.764	3.169	3.581
11	.260	.697	1.363	1.796	2.201	2.718	3.106	3.497
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.428

Since the t distribution is symmetric about zero, we have $t_{1-\alpha}$= $-t_{\alpha}$ (see Figure 5.8), that is, the t-value having an area of 1-α to the right (and therefore an area of α to the left) is equal to the negative of the *t*-value that has area α in the right tail of the distribution. Therefore, $t_{0.95,10}$ = -$t_{0.05,10}$ = -1.812. Finally, because *t* is the standard normal distribution, the familiar z_{α} values appear in the last row of Appendix Table II.

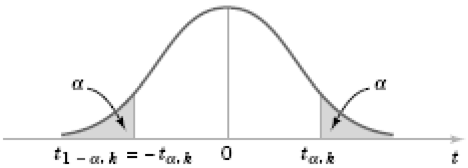


Figure 5.8: Percentage points of t distribution

Some Example t-score Problems

1. If the degree of freedom is 20, find the percentage points of t for a) α=10% and b) α=95%.

   Solution:

   a. $t_{10\%,20} = 1.325$.

   b. $t_{95\%,20} = -t_{5\%,20} = -1.725$.

2. If the degree of freedom is 15, find the areas under a distribution t for

   a. Percentage points of t distribution 2.947

   b. Percentage points of t distribution -2.947.

   Solution:

   a. $t_{\alpha,15}$=2.947 equivalent with α=0.005.

   b. $t_{\alpha,15}$=-2.947 equivalent with 1-α=0.005 or α=0.995.

3. A sport teacher claims that their students need average calories every special time of 20,000 calories, with a standard deviation of 1750 calories. Suppose a test 14 randomly-selected students. What is the probability that the average students in the test will be no more than 23,500 calories?

   Solution:

   One strategy would be a two-step approach:

   - Compute a t score, assuming that the mean of the sample test is 23,500 pounds.
   - Determine the cumulative probability for that t score.

We will follow that strategy here. First, we compute the t score: we have sample mean $\bar{x} = 23500$, mean population μ=20000, standard deviation sample s=1750 and sample size n=14. So that,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{23500 - 20000}{1750/\sqrt{14}} = 0.6265.$$

Now, we can determine the cumulative probability for the t score. We know the following:

- The t score is equal to 0.6265.
- The number of degrees of freedom is equal to 14-1=13.

Now, we are ready to use table t score. Since we have already computed the t score (0.6265) and the degrees of freedom (13), the appendix table II reports that the cumulative probability is near the 0.25. Therefore, there is approximately 25% students' need more than 23500 calories.

4. A school board administered an IQ test to 25 randomly selected teachers. They found that the average IQ score was 115 with a standard deviation of 11. Assume that the cumulative probability is 0.90. What population mean would have produced this sample result? Note: In this situation, a cumulative probability of 0.90 suggests that 90% of the random samples drawn from the teacher population will have an average IQ of 115 or less. This problem asks us to find the true population IQ for which this would be true.
Solution:
We know the following:

- The standard deviation is 11.
- The sample mean is 115.
- The number of degrees of freedom is 25-1=24.

- The cumulative probability is 0.90 or α=1-0.90=0.10. That is value of Percentage Points $t_{\alpha,k}$

First, find the value of percentage point t with α=10% and degree of freedom 24, namely $t_{10\%,24}$=1.318. Next, to find the population mean, formulate the formula t.

$T = \dfrac{\bar{x} - \mu}{s / \sqrt{n}}$ is equivalent with $T(s/\sqrt{n}) = \bar{x} - \mu$ or $\mu = \bar{x} - T(s/\sqrt{n})$.

Find the population mean μ=115 – (1.318)(11/$\sqrt{24}$) = 112.04.

If the true population mean is 112.04, we expect 90% of our samples to have a sample mean of 115 or less.

E. Table of χ^2 score and problems

Let $X1, X2, \dots, Xn$ be a random sample from a normal distribution with mean μ and variance σ^2, and let s^2 be the sample variance. Then the random variable

$$X^2 = \frac{(n-1)s^2}{\sigma^2}$$

has a chi-square (χ^2) distribution with *n-1* degrees of freedom.

Several chi-square distributions are shown in Figure 5.9. Note that the chi-square random variable is non-negative and that the probability distribution is skewed to the right. However, as *k* increases, the distribution becomes more symmetric. As the *k* goes to unlimited, the limited form of the chi-square distribution is the normal distribution.

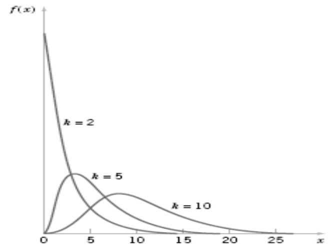


Figure 5.9: Probability density function of several distribution χ^2

The **percentage points** of the χ^2 distribution are given in Appendix Table III. Define $\chi^2{}_{\alpha,k}$ as the percentage point or value of the chi-square random variable with *k* degrees of freedom such that the probability that χ^2 exceeds this value is α. This probability is shown as the shaded area in Figure 5.10a.

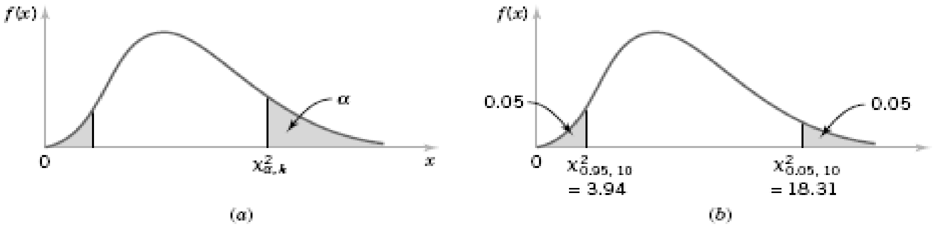
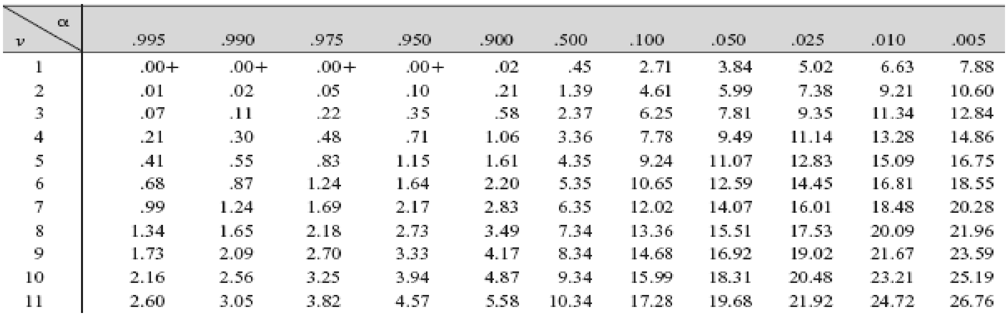


Figure 5.10: Percentage points of the χ^2 distribution

To illustrate the use of Appendix Table III, note that the areas α are the column headings and the degrees of freedom *k* are given in the left column. Table 5.8 (part of Appendix Table III) shows this illustration. Therefore, the value with 10 degrees of freedom having an area (probability) of 0.05 to the right is $\chi^2_{0.05,10} = 18.31$. This value is often called an **upper** 5% point of chi-square with 10 degree of freedom. The upper percentage point of the upper $\chi^2_{0.05,10} = 18.31$ and the **lower** percentage point $\chi^2_{0.95,10} = 3.94$ (see Figure 5.10b).

Table 5.8 : Percentage points χ^2



α ν	.995	.990	.975	.950	.900	.500	.100	.050	.025	.010	.005
1	.00+	.00+	.00+	.00+	.02	.45	2.71	3.84	5.02	6.63	7.88
2	.01	.02	.05	.10	.21	1.39	4.61	5.99	7.38	9.21	10.60
3	.07	.11	.22	.35	.58	2.37	6.25	7.81	9.35	11.34	12.84
4	.21	.30	.48	.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	.41	.55	.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	.68	.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76

Some Examples χ² score Problems

1. If the degree of freedom is 20, find the percentage points of χ^2 for a) α=10% and b) α=95%.

   Solution:

   a. $\chi^2{}_{10\%,20} = 28.41.$

   b. $\chi^2{}_{95\%,20} = 10.85.$

2. If the degree of freedom is 15, find the areas under a distribution χ^2 for

   a. Percentage points of χ^2 distribution 8.55

   b. Percentage points of χ^2 distribution 30.58.

   Solution:

   c. $\chi^2{}_{\alpha,15}$=8.55 equivalent with α=0.900.

   d. $\chi^2{}_{\alpha,15}$=30.58 equivalent with α=0.010.

3. The Coordinator national examination claims that the student score in the last 5 years, with a standard deviation score of 1. Assume that their claims are true. If we test a random sample of 9 students, what is the probability that the standard deviation in our sample will be less than score 0.95?

   Solution:

   We know the following:

   - The population standard deviation is equal to 1.
   - The sample standard deviation is equal to 0.95.
   - The sample size is equal to 9.
   - The degree of freedom is equal to 9-1=8.

   Given these data, we compute the chi-square statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(9-1)(0.95^2)}{1.0^2} = 7.22.$$

Now, we can determine the cumulative probability for the χ^2 score. We know the following:

- The χ^2 score is equal to 7.22.
- The number of degrees of freedom is equal to 9-1=8.

Now, we are ready to use table χ^2 score. Appendix table III reports that the cumulative probability is near the 0.500. Therefore, there is approximately 50% students have standard deviation less that 0.95.

F. Table of F score and problems

Let $X_1, X_2, \ldots, X_p$ be a random sample from a normal population with mean μ_1 and

variance σ_1^2, and let $Y_1, Y_2, \ldots, Y_q$, be a random sample from a second normal population with mean α_2 and variance σ_2^2. Assume that both normal populations are independent. Let s_1^2 and s_2^2 be the sample variances. Then the ratio is

$$F = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

has an *F* distribution with u=p-1, numerator degrees of freedom and v=q-1 denominator degrees of freedom.

Three F distributions are shown in Figure 5.11. The F random variable is nonnegative, and the distribution is skewed to the right. The F distribution looks very similar to the chi-square distribution; however, the two parameters *u* and *v* provide extra flexibility regarding shape.

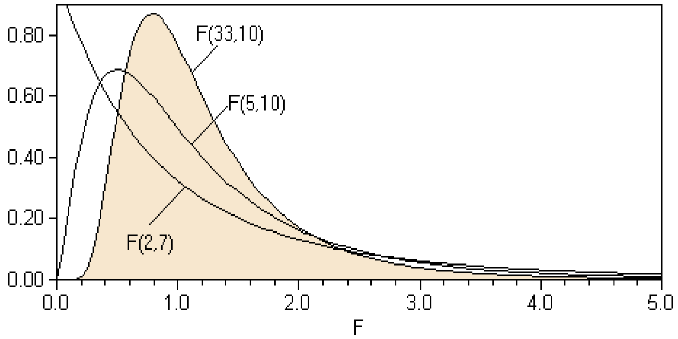


Figure 5.11: Distribution F with the Differences Degree of Freedom

The percentage points of the *F* distribution are given in Appendix Table IV. Let $f_{u,v}$, be the percentage point of the *F* distribution, with numerator degrees of freedom *u* and denominator degrees of freedom *v* such that the probability that the random variable *F* exceeds this value is α. This is illustrated in Figure 5.12.

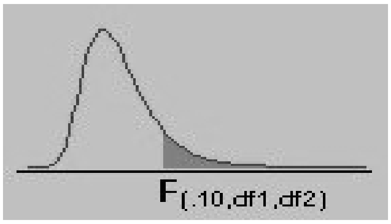
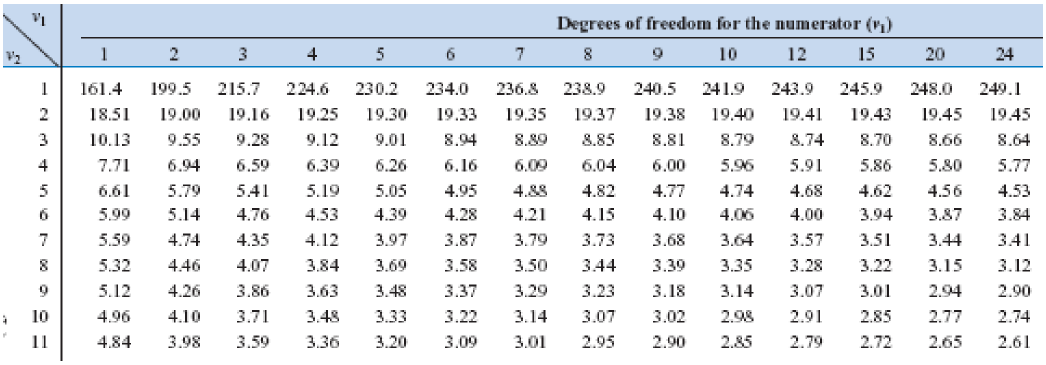


Figure 5.12: Percentage points of the F distribution

For example, if u=5, v=10, we find from Tabel 5.9 (part of Appendix Table IV) that $f_{0.05,5,10}$=3.33, that is, the upper percentage point of $F_{5,10}$ is $f_{0.05,5,10}$=3.33.

Table 5.9: Percentage Points $f_{\alpha,u,v}$ of the *F*-Distribution



ν_1 / ν_2	Degrees of freedom for the numerator (ν_1)													
	1	2	3	4	5	6	7	8	9	10	12	15	20	24
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61

Appendix Table IV contains only upper-tail percentage points (for selected value of $f_{\alpha,u,v}$ for $\alpha \leq 0.25$) of the F distribution. The lower – tail percentage points $f_{1-\alpha,u,v}$ can be found as follows:

$$f_{1-\alpha,u,v} = \frac{1}{f_{\alpha,v,u}}.$$

For example, to find the lower-tail percentage point $f_{0.95,5,10}$, note that:

$$f_{0.95,5,10} = \frac{1}{f_{0.05,10,5}} = \frac{1}{4.74} = 0.211.$$

Some Example F score Problems

1. If the degree of freedom numerator is 20 and the denominator is 15, find the percentage points of F for a) α=10% and b) α=95%.

   Solution:

   a. $f_{10\%,20,15} = 1.92$.

   b. $f_{95\%,20,15} = 1/f_{0.05,15,20} = 1/2.20 = 0.4545$.

2. If the degree of freedom numerator is 25 and the denominator is 45, find the areas under a distribution F for:

   a. Percentage points of F distribution 1.80

   b. Percentage points of F distribution 0.467.

   Solution:

   i. $f_{\alpha,25,45}$=1.80 (we choose the value around) equivalent with α=0.05.

   ii. $f_{\alpha,25,45}$=0.467 equivalent with $f_{1/\alpha,45,25}$ =1/467= 2.14, we find α=0.025 (the nearest).

3. Suppose we take independent random samples of size $n_1 = 11$ and $n_2 = 16$ from normal populations. If the cumulative probability of the f statistic is equal to 0.75, what is the percentage point of F distribution?

   Solution:

   We know the following:

   a. the sample size $n_1 = 11$, the degrees of freedom $v_1 = n_1 - 1 = 10$

   b. the sample size $n_2 = 16$, the degrees of freedom $v_2 = n_2 - 1 = 15$

   c. the cumulative probability is equal to 1-α= 0.75. So that α=0.25

   Now, we can determine the cumulative probability for the F-score: $F_{0.25,10,15} = 1.45$.

4. A study was presented about test of mental performance. A psychologist claimed that the standard deviation of man and women to give a biased information 10% and 15% respectively. Suppose we randomly select 7 women and 12 men from a population. The standard deviation of biased information in each sample for men and women is 12% and 14% respectively. Compute the percentage point of F distribution.

   Solution:

   The f statistic can be computed from the population and sample standard deviations, using the following equation:

$$F = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

   where σ_1 is the standard deviation of population 1, s_1 is the standard deviation of the sample drawn from population 1, σ_2 is the standard deviation of population 2, and s_1 is the standard deviation of the sample drawn from population 2.

As we can see from the equation, there are actually two ways to compute an f statistic from these data. If the women's data appears in the numerator, we can calculate an f statistic as follows:

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{0.15^2/0.12^2}{0.10^2/0.14^2} = 3.06.$$

For this calculation, the numerator degrees of freedom v_1 are 7 - 1 or 6; and the denominator degrees of freedom v_2 are 12 - 1 or 11.

On the other hand, if the men's data appears in the numerator, we can calculate an f statistic as follows:

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{0.10^2/0.14^2}{0.15^2/0.12^2} = 0.3265.$$

For this calculation, the numerator degrees of freedom v_1 are 12 - 1 or 11; and the denominator degrees of freedom v_2 are 7 - 1 or 6.

5. Find the cumulative probability associated with each of the f statistics from the problem 4, above.

Solution:

To solve this problem, we need to find the degrees of freedom for each sample. Then, we will use the F distribution to find the probabilities.

- The degree of freedom for the sample of women is equal to n - 1 = 7 - 1 = 6.
- The degree of freedom for the sample of men is equal to n - 1 = 12 - 1 = 11.

Therefore, when the women's data appear in the numerator, the numerator degrees of freedom v_1 is equal to 6; and the denominator degrees of freedom v_2 is equal to 11. And, based on the computations shown in the previous example, the f statistic is equal to 3.06. We look at these values into the percentage points of f in F distribution Table IV of the appendix and find that α= 0.05.

On the other hand, when the men's data appear in the numerator, the numerator degrees of freedom v_1 is equal to 11; and the denominator degrees of freedom v_2 is equal to 6. And, based on the computations shown in the previous example, the f statistic is equal to 0.3265. We calculate:

$$f_{\alpha,11,6} = \frac{1}{f_{1-\alpha,6,11}} = 0.3265 \iff f_{1-\alpha,6,11} = 3.06\,.$$

The last equation, we find 1-α=0.05 <=> α=0.95.

G. Exercise

1. The following nine measurements are the heights in cm in a sample of nine students.

   Heigiht (x) : 156 155 157 156 159 166 161 159
   158

   a. Find the mean and standard deviation of this data before and after added with 4.
   b. Find the mean and standard deviation of this data before and after multiplied with 5.
   c. What it your conclusion about mean and standard deviation for a and b.
2. Transform the data from the exercise number 1 in standard score.
3. Transform the data from the exercise number 1 in a new standard score with mean 70 and standard deviation 10.
4. If the data from the exercise number 1 will be change to the new presentation that the minimum and maximum score become 6 and 9 respectively. Find the other score in the new presentation.
5. For the set of 12 numbers given below,

a. Compute the mean and standard deviation of these number.

b. Transform the data to the standard score.

c. Transform the data to the new standard score mean 60 and standard deviation 5.

The number are: 6,7,4,8,5,4,8,6,2,7,9,4.

6. A variable x is distributed normally with mean 50 and standard deviation 10. Variable y is find from variable x by subtracting 40 and dividing by 10. What are the mean and standard deviation of variable y?

7. Find the area under the standard normal curve

a. below $z = .96$

b. below $z = 1.96$

c. between $z = 0$ and $z = 2.58$

d. between $z = 0$ and $z = 1.96$

e. between $z = 1.05$ and $z = 2.47$

f. to the right of $z=2.48$

g. to the right of $z = -1.96$

h. to the right of $z = 1.96$

i. between $z = -1.96$ and $z = 1.96$

j. between $z = -2.41$ and $z = 1.85$

k. between $z = -1.23$ and $z = -0.21$

8. Find the following probability or percentages of the interval

a. $z < 1.52$

b. $0 < z < 0.84$

c. $-1.54<z<2.19$

d. $0.59 < z < 1.83$

9. In a sample distribution with mean 16 and standard deviation 3, find the z-score of

a) 10

b) 6

c) 21

d) 19

10. You got an 80 on a history exam. (Mean 83, standard deviation 5). What was your z score?

11. The 400 students in a district was normally distributed and has mean =70 and standard deviation 15.

    a. How many students have score more than 75?

    b. How many students have score between 40 and 90?

    c. What proportion of the students would be expected to score less than 35.

12. Find the percentage points of t for the degree of freedom is 25

    a. α=10% and b. α=95% c. α=5% and b. α=90%

13. Find the areas a distribution t for sampling with the degree of freedom is 22

    a. Percentage points of t distribution 2.800.

    b. Percentage points of t distribution 2.000.

    c. Percentage points of t distribution -0.690.

    d. Percentage points of t distribution -3.500.

14. A biology teacher claims that their students need for sleeping 8.5 hours with standard deviation 0.8 hours. Suppose a test 20 randomly-selected students. What is the probability that the average students in the test will

    a. more than 6 hour?

    b. not less than 5 hours?

15. Find the percentage points of χ^2 for the degree of freedom is 25

    a. α=10% and b. α=95% c. α=5% and b. α=90%

16. Find the areas a distribution χ^2 for sampling with the degree of freedom is 26
    a. Percentage points of t distribution 13.70.
    b. Percentage points of t distribution 41.90.
17. A biology teacher claims that the students for sleeping have standard deviation 0.8 hour. Assume that his claims is true. If we test a random sample of 20 students, what is the probability that the standard deviation in this sample will be less than score 0.90?
18. Find the percentage point of F distribution for the degree of freedom numerator is 17 and the denominator is 25 a) α=10% and b) α=95%.
19. Find the area a F distribution for the degree of freedom numerator is 19 and the denominator is 40, for:
    a. Percentage points of F distribution 1.65
    b. Percentage points of F distribution 1.97.
20. Suppose we take independent random samples of size $n_1 = 26$ and $n_2 = 51$ from normal populations. If the cumulative probability of the *f* statistic is equal to 0.85, what is the percentage point of F distribution?

CHAPTER VI
INFERENTIAL STATISTICS
(ESTIMATION AND TEST HYPOTHESIS)

A. Introduction

The description consists of the knowledge of constructing confidence interval and test hypothesis for the mean sample data.

The main instructional objectives are after the learning process students are able to:

1. construct confidence interval for the population mean given sample data
2. construct the rule of test hypothesis
3. carry out hypothesis test for the population mean given one sample data using appropriate parametric procedures.
4. carry out hypothesis test for the population mean given two samples data using appropriate parametric procedures.

B. Basic Definition

Now we define the basic concepts in inferential statistics.

1. **Population:** A population is any entire collection of people, animals, plants or things from which we may collect data. It is the entire group we are interested in, which we wish to describe or draw conclusions about. In order to make any generalisations about a population, a sample, that is meant to be representative of the population, is often studied. For each population there are many possible samples. A sample statistic gives information about a corresponding population parameter. For example, the sample mean for a set of data would give information about the overall population mean.

2. **Sample**: A sample is a group of units selected from a larger group (the population). By studying the sample it is hoped to draw valid

conclusions about the larger group. A sample is generally selected for study because the population is too large to study in its entirety. The sample should be representative of the general population. This is often best achieved by random sampling. Also, before collecting the sample, it is important that the researcher carefully and completely defines the population, including a description of the members to be included.

Example 6.1*:* The population for a study of competence behaviour consist of all students of The State University of Semarang in 2009. The sample might be 5 students from each department.

Sampling theory takes a whole lecture on its own. Since any result produced from the sample can be used to estimate the corresponding result for the population it is absolutely essential that the sample taken is as representative as possible of that population. Common sense rightly suggests that the larger the sample the more representative it is likely to be but also the more expensive it is to take and analyse.

We will not study sampling in this chapter but just give a list of the main methods below. More details can be found in another lecture.

a. Simple Random Sampling
b. Systematic Sampling
c. Stratified Random Sampling
d. Multistage Sampling
e. Cluster Sampling
f. Quota Sampling.

It is usually neither possible nor practical to examine every member of a population so we use the data from a sample, taken from the same population, to estimate the 'something' we need to know about the population itself. The sample will not provide us with the exact 'truth' but it is the best we can do. We also use our knowledge of samples to estimate limits within which we can expect the 'truth' about the population to lie and state how confident we are about this estimation. In other words instead of claiming that the mean cost of buying a book is, say, exactly we say 45$ that it lies between 40$ to 50$.

3. **Parameter**: A parameter is a value, usually unknown (and which therefore has to be estimated), used to represent a certain population characteristic. For example, the population mean is a parameter that is often used to indicate the average value of a quantity. Within a population, a parameter is a fixed value, which does not vary. Each sample drawn from the population has its own value of any statistic that is used to estimate this parameter. For example, the mean of the data in a sample is used to give information about the overall mean in the population from which that sample was drawn. Parameters are often assigned Greek letters (e.g. μ), whereas statistics are assigned Roman letters (e.g. $\bar{x}$).

4. **Statistical inference**: Statistical Inference makes use of information from a sample to draw conclusions (inferences) about the population from which the sample was taken. Statistical inference consists of estimation of parameters and testing of hypothesis. The goal of statistical inference is to go beyond the data at hand and make inferences about population parameters. In order to use inferential statistics, it is assumed that random selection was carried out (i.e., some form of randomization must is assumed).

C. Estimation Problems

In statistics, **estimation** refers to the process by which one makes inferences about a population, based on information obtained from a sample. Statisticians use sample statistics to estimate population parameters. For example, sample means are used to estimate population means; sample proportions, to estimate population proportions.

An estimate of a population parameter may be expressed in two ways:

1. Point estimation of parameters

A point estimation of a population parameter is a single value of a statistic. From the sample, a value is calculated which serves as a point estimation for the population parameter of interest. For example: Suposed that a random variable X is normally distributed with an unknown mean μ. We may have several different choices for the point estimator of the parameter μ. If we wish to estimate the mean μ, we might consider the sample mean, the sample median, or perhaps the average of the smallest and largest observations in the sample as point estimators. In order to decide which point estimator of this particular parameter is the best one to use (unbiased estimator), we need to learn more their statistical properties some criteria for comparing estimators.

The sample mean is a point unbiased estimator of the unknown population mean μ. Therefore, $\hat{\mu} = \bar{x}$. After the sample has been selected, the numerical value $\bar{x}$ is the point estimate of μ.

Example 6.2: If we have sample data Table 6.1 about learning achievement of 20 students, we can calculate the estimates point of parameter μ.

Table 6.1: The learning achievement of Students

80	80	78	76	81	84	80	75	88	80
88	76	76	56	49	80	67	82	78	74

The point estimate of μ is $\bar{x} = \frac{80+80+...+74}{20} = \frac{1527}{20} = 76.35 \quad or \quad \hat{\mu} = 76.35$.

Similarly, if the population variance σ^2 is also unknown, a point unbiased estimator for σ^2 is the sample variance s^2, and the numerical value $s^2 = 92.16$ calculated from the sample data is called the point estimate of σ^2.

Estimation problems occur frequently in education. We often need to estimate:

a. The mean μ of a single population.

b. The variance σ^2 (or standard deviation σ) of a single population.

c. The proportion *p* of items in a population that belong to a class of interest

d. The difference in means of two populations, μ_1-μ_2.

e. The difference in two population proportions, p_1-p_2.

Reasonable point estimates of these parameters are as follows:

a. For μ, the estimate is $\hat{\mu} = \bar{x}$, the sample mean.

b. For σ^2, the estimate is $\hat{\sigma}^2 = s^2$, the sample variance.

c. For *p*, the estimate is $\hat{p} = x/n$, the sample proportion, where *x* is the number of items in a random sample of size *n* that belong to the class of interest.

d. For μ_1-μ_2, the estimate is $\hat{\mu}_1 - \hat{\mu}_2 = \bar{x}_1 - \bar{x}_2$, the difference between the sample means of two independent random samples.

e. For p_1-p_2, the estimate is $\hat{p}_1 - \hat{p}_2$, the difference between two sample proportions computed from two independent random samples.

2. *Confidence Intervals*

We have already illustrated previously how a parameter can be estimated from sample data. However, it is important to understand how good the estimation is obtained. For example, supposed that we estimate the mean of a result study to be $\hat{\mu} = \bar{x}$ =65. Now because of sampling variability, it is almost never the case that $\hat{\mu} = \bar{x}$. The point estimation says nothing about how close $\hat{\mu}$ is to μ. Is the process of mean likely to be between 60 and 70? Or is it likely to be between 67 and 79? The answer to these questions affects our decisions regarding to this process. Bounds that represent an interval of plausible values for a parameter are an example of an interval estimate. Surprisingly, it is easy to determine such intervals in many cases, and the same data that provided the point estimation are typically used.

An interval estimation is defined by two numbers, between which a population parameter is said to lie. For example, $a < x < b$ is an interval estimation of the population mean μ. It indicates that the population mean is greater than *a* but less than *b*. An interval estimation for a population parameter is called a **confidence interval.** We cannot be certain that the interval contains the true, unknown population parameter—we only use a sample from the full population to compute the point estimate and the interval. However, the confidence interval is constructed so that we have high confidence which contains the unknown population parameter. Confidence intervals are widely used in education and the sciences.

Statisticians use a **confidence interval** to express the precision and uncertainty associated with a particular sampling method. A confidence interval consists of three parts.

- A confidence level.
- A statistic.

- A margin of error.

The confidence level describes the uncertainty of a sampling method. The statistic and the margin of error define an interval estimate that describes the precision of the method. The interval estimate of a confidence interval is defined by the *sample statistic* ± *margin of error*.

For example, we might say that we are 95% confident that the true population mean falls within a specified range. This statement is a confidence interval. It means that if we used the same sampling method to select different samples and compute different interval estimates, the true population mean would fall within a range defined by the *sample statistic* ± *margin of error* 95% of the time.

Confidence intervals are preferred to point estimates, because confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.

Confidence Level

The probability part of a confidence interval is called a **confidence level**. The confidence level describes how strongly we believe that a particular sampling method will produce a confidence interval that includes the true population parameter. Here is how to interpret a confidence level. Suppose we collected many different samples, and computed confidence intervals for each sample. Some confidence intervals would include the true population parameter; others would not. A 95% confidence level means that 95% of the intervals contain the true population parameter; a 90% confidence level means that 90% of the intervals contain the population parameter; and so on.

Margin of Error

In a confidence interval, the range of values above and below the sample statistic is called the margin of error. For example, suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the vote. The newspaper states that the survey had a 5% margin of error and a confidence level of 95%. These findings result in the following confidence interval: We are 95% confident that the independent candidate will receive between 25% and 35% of the vote.

Note:

a. Many public opinion surveys report interval estimation, but not confidence intervals. They provide the margin of error, but not the confidence level. To clearly interpret survey results we need to know both. We are much more likely to accept survey findings if the confidence level is high (say, 95%) than if it is low (say, 50%).

b. How do we interpret a confidence interval? If 100 similar samples were taken and analysed then, for a 95% confidence interval, we are confident that 95 of the intervals calculated would include the true population mean. In practice we tend to say that we are 95% confident that our interval includes the true population value. Note that there is only one true value for the population mean; it is the variation between samples which gives the range of confidence intervals.

Confidence interval for the Population Mean, μ (σ, is known)

To describe a confidence interval with an arbitrary confidence level, we will use the expression $(1-\alpha)100\%$ for the confidence level. This complex expression is used because it makes the formulas for the limits of the confidence interval relatively simple. For example, 99=(1-0.01)100, so that α=0.01. The lower and upper limits of the (1-α)100% confidence interval for the mean, when σ is known, are

Lower limit = statistic - margin of error =: $\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Upper limit = statistic + margin of error =: $\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Note that the percentiles of the normal distribution appear in these formulas and that the area in the upper tail of the normal distribution is α/2. For the 99% confidence interval, α=1% and α/2 = 0.5% = 0.005. Since $z_{0.005}$=2.58, the 99% confidence interval is:

$$(\bar{x} - 2.58\frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58\frac{\sigma}{\sqrt{n}}).$$

Another example: For the 80% confidence interval, α=20% and α/2 = 010% = 0.10. Since $z_{0.10}$=1.28, the 80% confidence interval is:

$$(\bar{x} - 1.28\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.28\frac{\sigma}{\sqrt{n}}) .$$

This interval is not the 80% confidence interval. The interval is centered on the unknown population mean, μ. Confidence intervals are centered on an observed sample mean $\hat{\mu} = \bar{x}$.

What is the probability that a confidence interval includes μ? The answer is simple: the probability is 0.80. The interval includes μ for precisely those sample means (such as same $\bar{x}_1, \bar{x}_2, ..., \bar{x}_p$) that fall in the original interval about the population mean extending from $\bar{x} - 1.28\frac{\sigma}{\sqrt{n}}$ *to* $\bar{x} + 1.28\frac{\sigma}{\sqrt{n}}$.

The argument depends on the fact that the confidence intervals have exactly the same width as the interval about the population mean.

Since the probability that a sample mean falls in the interval about μ is 80%, it follows that the probability that a confidence interval of the same width includes μ is also 80%. Conversely, if the sample mean (such as $\bar{x}_3$) is outside the interval about the population, then the confidence interval drawn about the sample mean will not include μ. The probability of this occurrence is 20%.

Example 6.3: For the small school as a whole it is known that the standard deviation of the score examination is 1.50. A random sample of 10 students gave a mean score of 6.15. Assuming the same standard deviation, calculate the 95% confidence interval for the average score. Determine the student has score 6.50 include then that interval.

Solution: For the 95% confidence interval, α=5% and α/2 = 2.5% = 0.025, so that $z_{0.025}$=1.96. As we actually know the population standard deviation we do not need to estimate it from the sample standard deviation. The value of standard error is:

$$\frac{\sigma}{\sqrt{n}} = \frac{1.50}{\sqrt{10}} = 0.93.$$

<u>The value of confidence Interval</u>: $\mu \;=\; \bar{x} \pm z\frac{\sigma}{\sqrt{n}}$

$$\mu = 6.15 \pm 1.96 \times \frac{1.50}{\sqrt{10}} \quad = 6.15 \pm 0.93.$$

The confidence limits for the population mean, μ , are 5.22 and 7.08 or

$$5.22 < \mu < 7.08.$$

This interval includes the score 6.50.

Confidence interval for the Population Mean, μ (σ, is unknown)

In most situations the population standard deviation is not known and therefore the standard error must be estimated. If standard deviation of a population is unknown, then we estimate it from the sample standard deviation, s. We must be used the t-table to compensate for the probable error in estimating s value from the sample standard deviation. Recall, that the number egrees of freedom is (n-1). The degree of freedom is indicated as a subscript of the t in the formulas for the limits of the (1-α)100% confidence interval.

Lower limit = statistic - margin of error =: $\bar{x} - t_{n-1,\alpha/2}\dfrac{s}{\sqrt{n}}$

Upper limit = statistic + margin of error =: $\bar{x} + t_{n-1,\alpha/2}\dfrac{s}{\sqrt{n}}$.

Note: Consider the data of example 6.2 : $\bar{x} = 76.35$, s^2=92.16 or $s = \sqrt{92.16} = 9.6$, n=20. The estimated standard error is $\dfrac{s}{\sqrt{n}} = \dfrac{9.6}{\sqrt{20}} = 2.15$. In order to compute the 95% confidence interval for the mean, we need $t_{n-1,0.025}$. Since $t_{n-1,0.025} = t_{19,0.025} = 2.093$, the 95% confidence interval is:

$$76.35 - (2.093)(2.15), 76.35 + (2.093)(2.15) = (71.8, 80.8).$$

Example 6.4: Find the 99% confidence interval for the mean value of all the scores of the result in table 6.1. If Rudy has score 56, does he belong to this interval?

<u>Confidence Interval</u>: $\mu = \bar{x} \pm t\dfrac{s}{\sqrt{n}}$.

From Example 6.2 we have $\bar{x}$=76.35, s=9.60, n=20. Degrees of freedom = (20 – 1) = 19; 99% confidence; so that we find $t_{99\%,19}$= 2.539.

Confidence interval

$$\mu = \bar{x} \pm t\frac{s}{\sqrt{n}} = 76.35 \pm 2.539\frac{9.60}{\sqrt{20}} = 76.35 \pm 5.45, \ \ or$$

$$70.90 < \mu < 81.80.$$

This interval does not include 56.0, so that Rudy does not belong to the interval.

D. Hypothesis Testing Population Parameters

Hypothesis testing is the branch of inferential statistics that is concerned with how well the sample data support a null hypothesis and when the null hypothesis can be rejected in favour of the alternative hypothesis. Pay attention for several information as follows:

1. First note that the null hypothesis is usually the prediction that there is no relationship or no difference in the population.
2. The alternative hypothesis is the logical opposite of the null hypothesis and says there is a relationship of difference in the population.
3. We use hypothesis testing when we expect a relationship or difference to be present; in other words, we usually hope to “nullify” the null hypothesis and tentatively accept the alternative hypothesis.
4. Here is the key question that is answered in hypothesis testing: "*Is the value of my sample statistic unlikely enough (assuming that the null hypothesis is true) for me to reject the null hypothesis and tentatively accept the alternative hypothesis*?"
5. Note that it is the null hypothesis that is directly tested in hypothesis testing (not the alternative hypothesis).

Note: If we expect the null to be true, we can use the estimation approach described previously. A 95% confidence interval is equivalent to a two-tail test of hypotheses at $\alpha = 0.05$. If the hypothesized population mean falls outside of the confidence then the null hypothesis is rejected. For example: The hypothesis that the sample comes from a population with a mean of 33.02. The confidence interval is (26.48, 29.03). The hypothesized population mean is 33.02, which falls outside of the 95% confidence interval. Therefore, reject the null hypothesis at $\alpha = 0.05$.

Look at the example 6.5, it is sure to notice that the null hypothesis has the equality sign in it and the alternative hypothesis has the "not equals" sign in it. We can also see in the example that hypotheses can be tested for many different kinds of research questions such as questions about means, standard deviation, and so on.

Example 6.5: Null and alternative hypothesis

Research Question	Verbal Null (H_0) Hypothesis	Symbolic H_0 hypothesis	Verbal Alternative (H_1) Hypothesis	Symbolic H_1 hypothesis
Do Teachers score on the GRE verbal than the national average?	The teacher population GRE verbal mean is equal to the national average of 476	H_0:$\mu_{GRE.V}$ =476	The teacher popula tion GRE verbal mean is different from the national average of 476	H_0:$\mu_{GRE.V} \neq 4$
Do males or females tend to score better on the GRE verbal?	The male and female population means are not different	H_0:$\mu_M = \mu_F$	The male and female population means are different	H_0:$\mu_M \neq \mu_F$
Is there a correlation between GPA(X) and starting salary (y)?	The population correlation between GPA and starting salary is equal to zero	H_0:ρ_{xy} =0	The population corre lation between GPA and starting salary is not equal to zero	H_0:$\rho_{xy} \neq 0$
Is there a relation ship between GRE verbal (x) and starting salary (y)?	The population regression coefficient is equal to zero	H_0:β_{xy} =0	The population regression coefficient is not equal to zero	H_0:$\beta_{xy} \neq 0$

The decision whether reject hypothesis null or accept the hypothesis null, we still need more information as like bellows:

1. Earlier it is mentioned that we reject the null hypothesis when the *probability* of our result assuming a true null is very small. It means that we reject the null when the evidence would be unlikely under the assumption of the null.
2. In particular, we set a significance level (also called the α level) to use in our research study, which is the point at which we would consider a result to be very unlikely. Then, if our probability value is less than or equal to our significance level, we reject the null hypothesis.
3. It is essential that we understand the difference between the probability value_(also called the p-value) and the significance level (also called the α level).

The next idea is for us to realize that we will either make a correct decision about statistical significance or we will make an error whenever we conduct a hypothesis test. This idea is shown below and in Table 6.2.

Table 6.2: The four Possible Outcomes in Hypothesis Testing

		The True (but Unknown) Status of the null Hypothesis	
		The null hypothesis is true (It should not be rejected)	The null hypothesis is false (It should be rejected)
Your Decision	Fail to reject the null hypothesis	Type A Correct decision	Type II Error (false negative)
	Reject the null hypothesis	Type I Error (false positive)	Type B Correct decision

Remember that if the null hypothesis is true, it should not be rejected, but if the null hypothesis is false, it should be rejected. The problem is that we will not know if the null hypothesis is true or false. We only have the probabilistic evidence obtained from our sample data. Here is more detail the explanation about that table.

1. Looking at the top of the table (i.e., above the two columns) we will see that the null hypothesis is either *true* or *not true* in the empirical world.
2. If we look at the side of the table (i.e., beside the two rows) we will see that we must make a decision to either *fail to reject* or to *reject* the null hypothesis.
3. When the null is false we want to reject it, but when it is true we do not want to reject it.
4. The four logical possibilities of hypothesis testing are shown in the table.

5. When the null hypothesis is true we can make the correct decision (i.e., fail to reject the null) or we can make the incorrect decision (rejecting the true null). The incorrect decision is called a Type I error or a "false positive" because we have erroneously concluded that there is an effect or relationship in the population.
6. When the null hypothesis is false we can also make the correct decision (i.e., rejecting the false null) or we can make the incorrect decision (failure to reject the false null). The incorrect decision is called a Type II error or a "false negative" because we have erroneously concluded that there is no effect or relationship in the population.
7. We need to memorize the definitions of Type I and Type II errors, and after working with many examples of hypothesis testing they will become easier to ponder.
8. Exercise: In guidance counselling schools, a student is presumed to be innocent (i.e., that is the null hypothesis). Explain the idea of Type I and Type II errors here. Which error has occurred when an innocent person is found guilty? Which error has occurred when a guilty person is found innocent by the psychologist? (The answers are below.)

Type I Error (α)

In a hypothesis test, a type I error occurs when the null hypothesis is rejected when it is in fact true; that is, H_0 is wrongly rejected. For example, work in the upper process guidance counselling.

H_0 : the student presumed is innocent.

A type I error would occur if we concluded that the student is guilty when in fact the student is found innocent. A type I error is often considered to be more serious, and therefore more important to avoid, than a type II error. The hypothesis test procedure is therefore adjusted so that there is a guaranteed 'low' probability of rejecting the null hypothesis wrongly; this probability is never 0. This probability of a type I error can be precisely computed as P(type I error) = significance level = α.

Researchers choose often significances levels equal to 0.01; 0.05; or 0.10; but any value between 0 and 1 can be used.

Type II Error (β)

The null hypothesis H_0 may be false but it may be accepted. It is an error and is called Type-II error. The value of the test-statistic may fall in the acceptance region when H_0 is in fact false. Suppose the hypothesis being tested is H_0: $\theta = \theta_0$ and H_0 is false and true value of θ is θ_1. If the difference between θ_0 and θ_1 is very large then the chance is very small that θ_0 (wrong) will be accepted. In this case the true sampling distribution of the statistic will be quite away from the sampling distribution under H_0. There will be hardly any test-statistic which will fall in the acceptance region of H_0 . When the true distribution of the test-statistic overlaps the acceptance region of H_0, then H_0 is accepted though H_0 is false. If the difference between θ_0 and θ_1 is small, then there is a high chance of accepting H_0. This action will be an error of Type-II.

The probability of making Type II error is denoted by β. Type-II error is committed when H_0 is accepted while H_1 is true. The value of β can be calculated only when we happen to know the true value of the population parameter being tested.

The **significance level** is just that point at which we would consider a result to be "rare." *We are the one who decides on the significance level* to use in our research study. A significance level is not an empirical result; it is the level that we set so that we will know what probability value will be small enough for us to reject the null hypothesis.

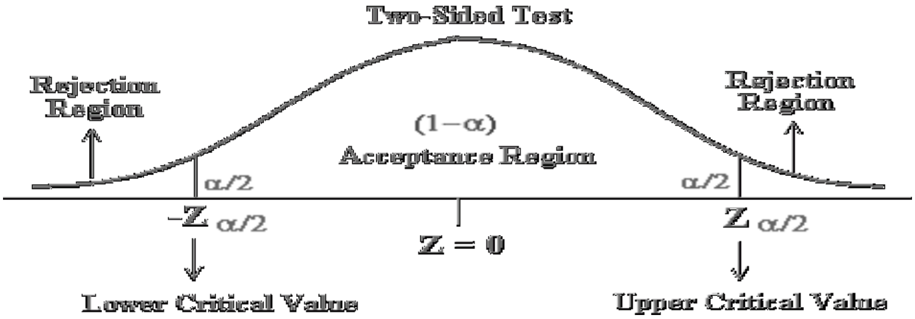
The significance level that is usually used in education is .05 (5%). It boils down to this: *if our probability value is less than or equal to the significance level (e.g., .05) then we will reject the null hypothesis and tentatively accept the alternative hypothesis. If not (i.e., if it is > .05) then we will fail to reject the null*. We just compare our probability value with our significance level.

We must memorize the definitions of probability value and significance level right away because they are at the heart of hypothesis testing. At the most simple level, the process just boils down to seeing whether we probability value is less than (or equal to) our significance level. If it is, we are happy because we can reject the null hypothesis and make the claim of statistical significance.

Two Tailed Test

When the rejection region is taken on both ends of the sampling distribution, the test is called two-sided test or two-tailed test. When we are using a two-sided test, half of the rejection region equal to $\alpha/2$ is taken on the right side and the other half equal to $\alpha/2$ is taken on the left side of the sampling distribution. Suppose the sampling distribution of the statistic is a normal distribution and we have to test the hypothesis $H_0:\theta=\theta_0$ against the alternative hypothesis $H_1:\theta\neq\theta_0$ which is two – sided. H_0 is rejected when the calculated value of Z is greater than $Z_{\alpha/2}$ or it is less than $-Z_{\alpha/2}$. Thus the critical region is $Z<Z_{\alpha/2}$ or $Z>-Z_{\alpha/2}$, it can also be written as $-Z_{\alpha/2}<Z<Z_{\alpha/2}$. When H_0 is rejected, then H_1 is accepted. Two-sided test is shown in the figure 6.1.

Figure 6.1: The interval two tailed test

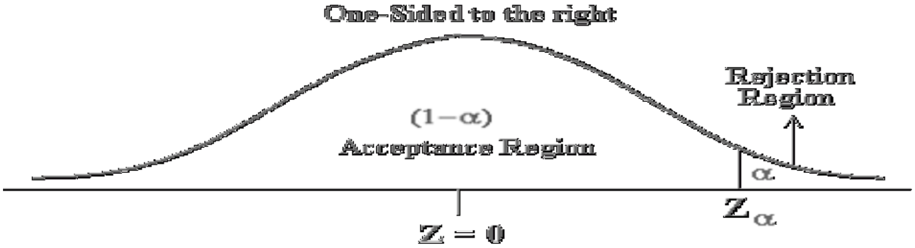


One Tailed Test

When the alternative hypothesis H$_1$ is one-sided like $\theta > \theta_0$ or $\theta < \theta_0$, then the rejection region is taken only on one side of the sampling distribution. It is called one-tailed test or one-sided test. When H$_1$ is one-sided to the right like $\theta > \theta_0$, the entire rejection region equals to α , is taken in the right end of the sampling distribution.

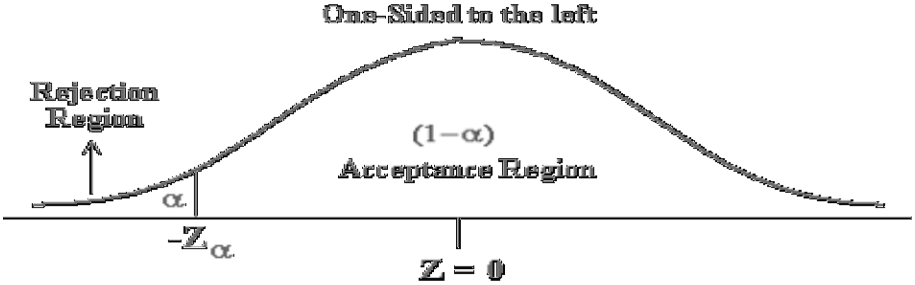
The test is called one-sided to the right. The hypothesis H$_0$ is rejected if the calculated value of a statistic, say Z falls in the rejection region. The critical value is Z_α which has the area equal to α to its right. The rejection region and acceptance region are shown in Figure 6.2. The null hypothesis H$_0$ is rejected when Z(calculated) > Z_α.

Figure 6.2: The interval one tailed test (right)



If the alternative hypothesis is one-sided to the left like $\theta < \theta_0$, the entire rejection region equal to α is taken on the left tail of the sampling distribution. The test is called one-sided or one-tailed to the left. The critical value is $-Z_\alpha$ which cuts off the area equal to α to its left. The critical region is $Z < -Z_\alpha$ and is shown in Figure 6.3.

Figure 6.3: The interval one tailed test (left)



For some important values of α, the critical values of Z for two-tailed and one tailed tests are given below:

Table 6.3: Critical Value of Z

α	Two-side test	One-sided to the right	One-sided to the left
10%	-1.645 and +1.645	+1.282	-1.282
5%	-1.96 and +1.96	+1.645	-1.645
2%	-2.326 and +2.326	+2.054	-2.054
1%	-2.575 and +2.575	+2.326	-2.326

Note: To make an easy in thinking in one tailed test, we always write hypothesis null in equality and alternative hypothesis in inequality what we will test.

Hypothesis Test of Mean for one sample with known standard deviation population

Hypothesis test for mean one sample consists of four steps:

1. **State the hypothesis**. The firs step is to state the null hypothesis and an alternative hypothesis.

   H$_0$: $\mu = \mu_0$

   H$_1$: $\mu \neq \mu_0$ (two tailed), $\mu > \mu_0$ (right one tailed), $\mu < \mu_0$ (left one tailed)

2. **Formulate an analysis plan**. Determine the significance level α and choosing tailed test.

3. **Analyse sample data**: For this analysis we use the sample data to find the statistic test and its associated z-score. The test statistic is a z score, it is defined by,

$$z = \frac{\mu - \mu_0}{\sigma / \sqrt{n}}.$$

In which μ is the hypothesized value of population mean in the null hypothesis, μ_0 is the sample mean, σ is the standard deviation of population distribution and n is the sample size.

4. **Interpret results**. We accept H_0 for two tailed test, if $-z_{\alpha/2} < z < z_{\alpha/2}$ for two tailed, $z > z_\alpha$ for right one tailed test, and $z < z_\alpha$ for left one tailed test, other wise reject H_0.

Example 6.6: A physic student has researched about energy-efficient lawn mower engine. He will prove the claim from the producers that the engine will run continuously for 5 hours (300 minutes) on a single gallon of regular gasoline and has standard deviation 10 minutes. He has observed and tested simple random sample of 50 engines. The engines run for an average of 295 minutes. Test the null hypothesis with the mean run time is 300 minutes against the alternative hypothesis with the mean run time is not 300 minutes. Use a 0.05 level of significance. Assume that run times for the population of engines are normally distributed. What is the conclusion?

Solution: This research gives information that the standard deviation is known namely σ=10 minutes the average μ=300. In this case we use the analyse data with z-score for two tailed test. The sampling mean μ_0=295, n=50. The step solution is as follows:

1. **State the hypothesis**.

H_0: $\mu = 300$

H$_1$: $\mu \neq 300$

2. **Formulate an analysis plan**. For this analysis is two tailed test and we use significance level 5%.
3. **Analyse sample data**.

$$z = \frac{\mu - \mu_0}{\sigma / \sqrt{n}} = \frac{300 - 295}{10/\sqrt{50}} = 3.54.$$

4. **Interpret results**. Since we have a two tailed test and $\alpha = 0.05$, the interval value accepting H$_0$ is ($-z_{0.025} < z < z_{0.025}$)=($-1.96<z<1.96$). The computing value of z= 3.54 does not belong to this accepting interval, therefore we reject H$_0$. It means the claim is not right or the average of running engine is not 300 minutes.

Hypothesis Test of Mean for one sample with unknown standard deviation population

Hypothesis test for mean one sample consists of four steps:

1. **State the hypothesis**. The firs step is to state the null hypothesis and an alternative hypothesis.

   H$_0$: $\mu = \mu_0$

   H$_1$: $\mu \neq \mu_0$ (two tailed), $\mu > \mu_0$ (right one tailed), $\mu < \mu_0$ (left one tailed)
2. **Formulate an analysis plan**. Determine the significance level α and choosing tailed test.
3. **Analyse sample data**: For this analysis we use the sample data to find the statistic test and its associated t-score. The test statistic is a t score, it is defined by,

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}.$$

where $\bar{x}$ is the hypothesized value of sample mean in the null hypothesis, μ_0 is the sample mean, s is the standard deviation of sample distribution and n is the sample size.

4. **Interpret results**. We accept H_0 for two tailed test, if $-t_{\alpha/2,n-1} < t < t_{\alpha/2,n-1}$ for two tailed, $t > t_{\alpha,n-1}$ for right one tailed test, and $t < t_{\alpha,n-1}$ for left one tailed test, other wise reject H_0. The degree of freedom is equal to the sample size (n) minus one. Thus, DF=n-1.

Example 6.7: Semarang Club Elementary School has 300 students. The principal of the school thinks that the average IQ of students at Semarang Club is at least 110. To prove her point, she administers an IQ test to 20 randomly selected students. Among the sampled students, the average IQ is 108 with a standard deviation of 10. Based on these results, should the principal accept or reject her original hypothesis? Assume a significance level of 0.01.

Solution: This research gives information that the standard deviation is unknown. In this case we use the analyse data with t-score for one tailed test (the average IQ at least 110). The sampling mean x=295, n=20. The solution steps are as follow:

1. **State the hypothesis**.
   H_0: $\mu = 110$
   H_1: $\mu > 110$
2. **Formulate an analysis plan**. For this analysis we use right one tailed test and significance level 0.01%.
3. **Analyse sample data**.
   $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{108 - 110}{10/\sqrt{20}} = -0.89.$
4. **Interpret results**. Since we have a one tailed test, degrees of freedom DF=n-1=19 and α =1%, so that $t_{1\%,19}$=2.539. The interval value of accepting H_0 is t<2.539. The computing value of t= -0.89

is belongs to the accepting interval, so that we accept H_0. It mean, that the average IQ of students at Semarang Club is lest than or equal 110.

Hypothesis Test of the Difference between Two Means

To test the difference of two samples we make assume that the both samples are homogeneity.

The assumption with homogeny come to the t test distribution. To comparing means 2 samples consists of four steps:

1. **State the hypothesis**. The first step is to state the null hypothesis and an alternative hypothesis.

   H_0: $\mu_1 = \mu_2$

   H_1: $\mu_1 \neq \mu_2$ (two tailed), $\mu_1 > \mu_2$ (right one tailed), $\mu_1 < \mu_2$ (left one tailed)

2. **Formulate an analysis plan**. For this analysis, determine the significance level.

3. **Analyse sample data**: For this analysis we use the samples data to find the test statistic and its associated t-score. we use the formula as follow,

   $$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/n_1 + 1/n_2}},$$

   in which $\bar{x}_1$ mean of first sample, $\bar{x}_2$ mean of second sample, s is determined with $s^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$ and n_1 and n_2 samples sizes of first and second samples.

4. **Interpret results**. The degree of freedom DF=n_1+n_2-2. Since we have a two- tailed test and α=0.05, the interval value accepting H_0 for homogeny sampling

is $-t_{1-\alpha/2,DF} < t < t_{1-\alpha/2,DF}$, other wise reject H$_0$. Since we work in one tailed test the interval value accepting H$_0$ is $t < t_{1-\alpha,DF}$ (right test), $t > -t_{1-\alpha,DF}$ (left test), other wise reject H$_0$.

Example 6.8: Within a school district, students were randomly assigned to one of two Biology teachers - Ms. Ani and Ms. Sri. Both of them gave their students' assignment. Ms. Ani had 30 students, while Ms. Sri had 25 students. At the end of the year, each class took the same standardized test. Ms. Ani's students had an average test score of 78, with a standard deviation of 10; and Mrs. Sri' students had an average test score of 85, with a standard deviation of 15. Test the hypothesis whether Ms. Ani and Ms. Sri are equally effective teachers. Use a 0.10 level of significance. (Assume that student performance is approximately normal.)

Solution: The solution to this problem with difference t test. We work through those steps below:

Difference Test with t Distribution

Data of the samples shows $\bar{x}_1$= 78, $\bar{x}_2 = 85$, n$_1$=30 and n$_2$=25. The solution to this problem takes four steps as fallow,

1. **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

   H$_0$: μ_1 - μ_2 = 0

   H$_1$: μ_1 - $\mu_2 \neq 0$

   Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the difference between sample means is too big or if it is too small.

2. **Formulate an analysis plan**. For this analysis, the significance level is 0.10. By using sample data, we will conduct a two-sample t-test of the null hypothesis.

3. **Analyze sample data**. By using sample data, we compute the mix variances s^2 and t-score:

$$s = \sqrt{s^2} = \sqrt{156.60} = 12.51$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/n_1 + 1/n_2}} = \frac{78-85}{12.51\sqrt{1/30+1/25}} = -7.63.$$

4. **Interpret results**. Since we have a two tailed test, degrees of freedom DF= n_1+n_2-2 = 30+25-2= 53 and α =10%, so that $t_{5\%,53}$=1.67. The interval value of accepting H_0 is -1.67<t<1.67. The computing value of t= -7.63 does not belong to the accepting interval, so that we reject H_0. It means the mean for both teachers are difference.

Example 6.9: The two classes A and B were sampled both experiment and control form the research. The results were analysed with the following results: experiment class A: $\bar{x}_1$ = 67, s_1 = 12.0, n_1 = 30 and control class B: $\bar{x}_2$ = 65, s_2 = 5.3, n_2 = 30. Has the result experiment class A been better than the result of control class B? Calculate it with the level of significant 5%!

Solution: The solution to this problem with difference t test. We work through those steps below:

Difference t Test Distribution: Data of the samples show $\bar{x}_1$= 67, $\bar{x}_2 = 65$, n_1=30 and n_2=30. The solution to this problem takes four steps as fallow:

1. **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

   H_0: $\mu_1 - \mu_2 = 0$

   H_1: $\mu_1 - \mu_2 > 0$

2. **Formulate an analysis plan**. The significance level is 0.05. These hypotheses constitute a one-tailed test. Using sample data and homogeneity assumption, we will conduct a one-sample t-test and using t test.

3. **Analyze sample data**. Using sample data, we compute the t-score:

$$s^2 = \frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2} = \frac{12^2(30-1)+5.3^2(30-1)}{30+30-2} = 86.045$$

$$s = \sqrt{s^2} = \sqrt{86.045} = 9.27$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/n_1 + 1/n_2}} = \frac{67-65}{9.27\sqrt{1/30+1/30}} = 0.84.$$

h. **Interpret results**. Since we have a two tailed test, degrees of freedom DF= n_1+n_2-2=30+30-2= 58 and α =5%, so that $t_{5\%,58}$=1.67. The interval value of accepting H_0 is t<1.67. The computing value of t= 0.84 belongs to the accepting interval, so that we accept H_0. It means both classes are not different.

E. Exercise

1. An example of a data set for a matched-pairs t-test might look like this:

Pre test	8	13	22	25	29	31	35	38	42	52
Post test	31	37	45	28	50	37	49	25	36	69

   a. Find the estimation point of percentage of peoples who do not have increased value from pre test to post test.

   b. Find the estimation point of mean for pre test and post test

   c. Find the estimation point of standard deviation for pre test and post test

2. Do the pre-test and the post-test in number 1 reach the score success 45? Use the confidence interval 5%.
3. Compare the data between pre-test and post-test in number 1 with the confidence interval 5%.
4. Supposed that an organization has 5,000 members. Prior to their renewal driving membership, 75 members were randomly selected and surveyed to find out their priorities for the coming year. The mean average age of the sample was 53.1 and the unbiased standard deviation was 4.2 years. What is the 90% confidence interval around the mean?
5. After selecting a random sample of 18 people from a very large population, we want to determine if the average age of the sample is representative of the average age of the population. From previous research, we know that the mean age of the population is 32.0. For our sample, the mean age was 28.0 and the unbiased standard deviation was 3.2. Is the mean age of our sample significantly different from the mean age in the population?
6. Two samples were taken from the population. One sample had 25 subjects with the standard deviation 4.5 on some key variables. The other sample had 12 subjects with a standard deviation of 6.4 on the some key variables. Is there a significant difference between the variances of the two samples?
7. Two new product formulas were developed and tested. A twenty-point scale was used to measure the level of product approval. Six subjects tested the first formula. They gave it a mean rating of 12.3 with a standard deviation of 1.4. Nine subjects tested in the second formula, and they gave it a mean rating of 14.0 with a standard deviation of 1.7. The question we might ask is whether the observed difference between the two formulas is reliable.

CHAPTER VII
PREDICTING THE RELATIONSHIP BETWEEN VARIABLES

H. Description and Basic Competence

Description: The knowledge about the relationship between variables and the interpretation of value of relation given sample data.

Instructional objectives

After learning process students are able to:

1. distinguish dependent and independent variables
2. calculate the correlation coefficient
3. interpret correlation coefficient appropriately
4. calculate constant and slope coefficient for linear regression
5. use constant and slope coefficient to graph the 'best fitting' line
6. make reasonable predictions

I. Correlation and Linearity

Correlation is one of the most common and useful statistics. A correlation is a single number that describes the degree of relationship between two variables. Let us work through an example to show first how this relationship is with a picture. We assume that we want to look at the relationship between two variables, x and y. Look at the data in Table 7.1. Data of variable x are collected for 10 cases. Data of variables y_1, y_2 and y_3 are selected double of data x, minus double of data x plus 40, and arbitrary respectively.

Table 7.1: ***Relation variable x and variable y_i, i=1,2,3***

Variable x	*1*	*3*	*4*	*5*	*7*	*8*	*10*	*11*	*14*	*16*
Variable y_1	*2*	*6*	*8*	*10*	*14*	*16*	*20*	*22*	*28*	*32*
Variable y_2	*38*	*34*	*32*	*30*	*26*	*24*	*20*	*18*	*12*	*8*
Variable y_3	*5*	*7*	*22*	*6*	*16*	*14*	*32*	*2*	*10*	*5*

We plot the relation between x and y_1, x and y_2, x and y_3 respectively. These show in Figure 7.1a; 7.1b; 7.1c respectively. The values of each correlation coefficient (r) are 1, -1 and 0 respectively by Product Moment Correlation formula (the formula will be given in the next page).

Figure 7.1a:***The extreme correlation value of x and y_1 is 1***

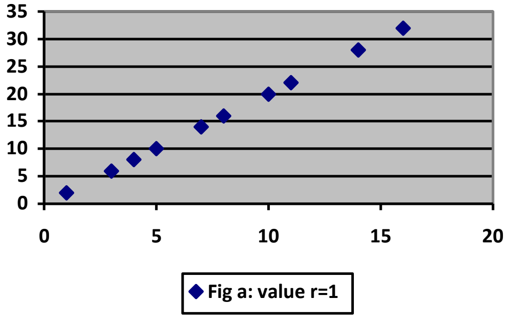


Figure 7.1a illustrates perfect positive correlation. The two variables of interest are on the x and y axis, respectively. By graphing through this way, it is apparent that a (positive) linear relationship exists between the two variables. It is important to note that a strong (or even perfect) correlation does not imply causation, as other variables may affect the relationship between the two variables of interest.

Figure 7.1b: Extreme correlation value of x and y_2 is -1

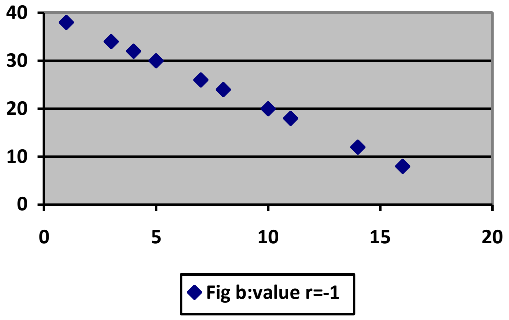


Figure 7.1b illustrates perfect negative correlation. The two variables of interest are on the x and y axis, respectively. By graphing through this way, it is apparent that a (negative) linear relationship exists between the two variables, i.e. the variables “move together”.

Figure 7.1c: Extreme correlation value of x and y_3 is near 0

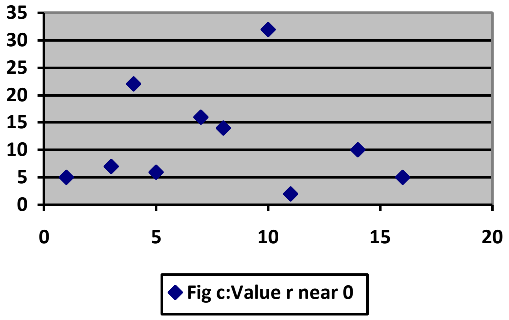
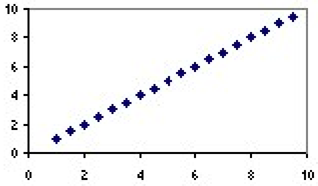


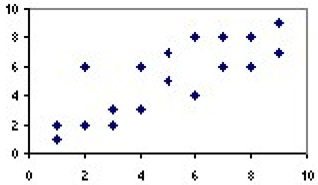
Figure 7.1c illustrates minimal correlation. The two variables of interest are on the x and y axis, respectively. By graphing through this way, it is difficult to establish any visual linear relationship between the two variables. In fact, this set of data points has a slight negative correlation (-0.024).

The scatter plots Figure 7.2 show how different patterns of data produce different degrees of correlation.

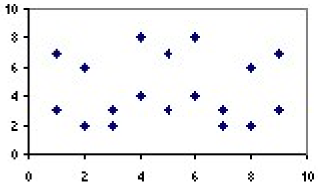
Figure 7.2: Different value of correlation coefficient



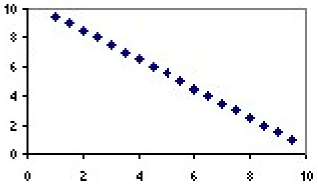
Maximum positive correlation (r = 1.0)



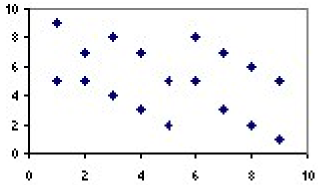
Strong positive correlation (r = 0.80)



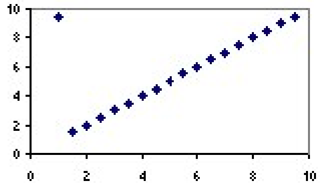
Zero correlation (r = 0)



Minimum negative correlation (r = -1.0)



Moderate negative correlation (r = -0.43)



Strong correlation & outlier (r = 0.71)

Several points are evident from the scatter plots:

1. When the slope of the line in the plot is negative, the correlation is negative; and vice versa.
2. The strongest correlations (r = 1.0 and r = -1.0) occur when data points fall *exactly* on a straight line.
3. The correlation becomes weaker as the data points become more scattered.
4. If the data points fall in a random pattern, the correlation is equal to zero.
5. Correlation is affected by outliers. Compare the first scatter plot with the last scatter plot. The single outlier in the last plot greatly reduces the correlation (from 1.00 to 0.71).

How to Calculate a Correlation Coefficient

If we look at different statistics textbooks, we are likely to find different-looking (but equivalent) formulas for computing a coefficient correlation. In this chapter, we present several formulas that we may encounter.

The most common formula for computing a product-moment correlation coefficient (r) is given below.

Product-moment coefficient correlation

The correlation r between two variables is:

$$r = \frac{\Sigma\chi\gamma}{\sqrt{\Sigma\chi^2}\sqrt{\Sigma\gamma^2}},$$

Where Σ is the summation symbol, $\chi = \bar{x} - x_i$, x_i is the value for observation i, $\bar{x}$ is the mean x value, $\gamma = \bar{y} - y_i, y_i$ is the value for observation i, $\bar{y}$ is the mean y value. So that we can write this formula as follow:

$$r_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2}\sqrt{\Sigma(y_i - \bar{y})^2}}.$$

We reformulate this formula as given bellow:

$$r_{xy} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{\{n\Sigma x_i^2 - (\Sigma x_i)^2\}\{n\Sigma y_i^2 - (\Sigma y_i)^2}},$$

n is the number of observations.

The interpretation of the sample correlation coefficient depends on how the sample data is collected. With a simple random sample, the sample correlation coefficient is an unbiased *estimate* of the population correlation coefficient.

Fortunately, we will rarely have to compute a correlation coefficient by hand. Many software packages (e.g., Excel, SPSS) and most graphing calculators have a correlation function that will do the job for us.

Example 7.1:

We will compute correlation coefficient data Table 7.2 (the modified data from Table 7.1)

Table 7.2: ***Relation variable x and variable y***

Variable x	***1***	***3***	***4***	***5***	***7***	***8***	***10***	***11***	***14***	***16***
Variable y	***2***	***6***	***8***	***10***	***14***	***16***	***20***	***22***	***15***	***32***

First, we should compute the mean of the data, and then to compute r with the second formula we need Table 7.3 to make the calculation easier. Means of variable x and variable y are

$\bar{x} = 7.9$ and $\bar{y} = 14.5$ respectively. The value of coefficient correlation r is as follow,

$$r_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2}\sqrt{\Sigma(y_i - \bar{y})^2}} = \frac{346.5}{\sqrt{212.9}\sqrt{686.5}} = 0.91.$$

Table 7.3: Calculating coefficient correlation with the second formula

x	**y**	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	2	-6.9	-12.5	86.25	47.61	156.25
3	6	-4.9	-8.5	41.65	24.01	72.25
4	8	-3.9	-6.5	25.35	15.21	42.25
5	10	-2.9	-4.5	13.05	8.41	20.25
7	14	-0.9	-0.5	0.45	0.81	0.25
8	16	0.1	1.5	0.15	0.01	2.25

10	20	2.1	5.5	11.55	4.41	30.25
11	22	3.1	7.5	23.25	9.61	56.25
14	15	6.1	0.5	3.05	37.21	0.25
16	32	8.1	17.5	141.75	65.61	306.25
Sum: 79	145			346.5	212.9	686.5

This formula is used rarely, because it is quite complicated to compute manually. Normally, we take the third formula which is a little bit easier than the second one. The third formula is computed directly without the first computing mean. Let us compute the value of r with the third formula through Table 7.4 as follows,

Table 7.4: Calculating coefficient correlation

x	y	xy	x^2	y^2
1	2	2	1	4
3	6	18	9	36
4	8	32	16	64
5	10	50	25	100
7	14	98	49	196
8	16	128	64	256
10	20	200	100	400
11	22	242	121	484
14	15	210	196	225
16	32	512	256	1024
79	145	1492	837	2789

Here n=10, so the value of coefficient correlation r is as follows,

$$r_{xy} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{\{n\Sigma x_i^2 - (\Sigma x_i)^2\}\ \{n\Sigma y_i^2 - (\Sigma y_i)^2}} = \frac{10(1492) - (79)(145)}{\sqrt{\{10(837) - 79^2\}\ \{10(2789) - 145^2\}}} = 0.91.$$

How to Interpret a Coefficient Correlation

The sign and the absolute value of a coefficient correlation describe the direction and the magnitude of the relationship between two variables.

1. The value of a coefficient correlation ranges between -1 and 1.
2. The greater the absolute value of a coefficient correlation is, the stronger the *linear* relationship will be.
3. The strongest linear relationship is indicated by a coefficient correlation of -1 or 1.
4. The weakest linear relationship is indicated by a coefficient correlation which is equal to 0.
5. A positive correlation means that if one variable gets bigger, the other variable tends to get bigger.
6. A negative correlation means that if one variable gets bigger, the other variable tends to get smaller.

Keep in mind that the Pearson product-moment coefficient correlation only measures linear relationships. Therefore, a correlation of 0 does not mean zero relationship between two variables; rather, it means zero *linear* relationship. (It is possible for two variables to have zero linear relationship and a strong curvilinear relationship at the same time.)

Hypothesis Test of the Correlation between Two Variables

We will formally go through the steps described in the chapter 6 to test the significance of a correlation using the logical reasoning and creativity data. Hypothesis test for correlation consists of four steps:

1. **State the hypothesis**.

   The first step is to state the null hypothesis and an alternative hypothesis.

   H$_0$: $\rho = 0$

   H$_1$: $\rho \neq 0$ (two tailed) ($\rho < 0$ *or* $\rho > 0$) (one tailed)

   Notice the hypotheses are stated in terms of population parameters. The null hypothesis specifies an exact value which implies no correlation.

2. **Formulate an analysis plan**.

   For this analysis, determine the significance level and the alternative test (one tailed or two tailed)!

3. **Analyse sample data**.

   For this analysis, we use sample data to find the statistic test and its t-score associated. The appropriate statistic test for these hypotheses is:

$$T_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

where r correlation coefficient samples x and y, n number of observations.

The form of *T$_0$* has the *t* distribution with n-2 degrees of freedom.

4. **Interpret results**.

   We reject H$_0$ for two tailed test and one tailed test, if $|t_0| > t_{\alpha/2,n-2}$, and $|t_0| > t_{\alpha,n-2}$ respectively. where $t_{\alpha,n-2}$ is the value of percentage points of distribution t with the significant level α and the degrees of freedom DF=n-2, other wise accept H$_0$.

Example 7.2: The paper "The Effective Learning of Mathematics : 'Central Tendency' using Integrated and Discovery Strategy Based on Technological Application" (Sukestiyarno, 2008) measures the variable x: skill learning process, and y: learning achievement by the students. The data observation show in Table 7.5. Examine: How strong is the relation between variable x and variable y?

Table 7.5: Research data about skill learning process (x) and learning achievement (y)

x	85	68	87	89	78	81	84	78	70	96	84	75	79	72	82	89	78	87	75
y	85	50	90	90	75	95	85	80	65	100	75	65	80	80	80	95	80	95	75

***Solution**:* We will test the relation between variable skill learning process (x) and variable learning achievement (y). In this case we use the analysed data with r-score. The steps solutions are as follow:

1. **State the hypothesis**.

   H$_0$: $\rho = 0$ (weak relation)

   H$_1$: $\rho > 0$ (not weak in positive relation)

2. **Formulate an analysis plan**. For this, we use one tailed test analysis and significance level 5%.

3. **Analyse sample data**. To compute the value of coefficient correlation r, follow the step like the previous calculation. To compute the value of r, we need the process through Table 7.6.

   The number of observation n=20, so the value of coefficient correlation r is as follow,

$$r_{xy} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{\{n\Sigma x_i^2 - (\Sigma x_i)^2\}\,\{n\Sigma y_i^2 - (\Sigma y_i)^2}}$$

$$r_{xy} = \frac{20(133505) - (1626)(1625)}{\sqrt{\{20(133190) - 1626^2\}\,\{20(134775) - 1625^2\}}} = 0.842.$$

Next, we find the value of t_0 :

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.842(\sqrt{20-2})}{\sqrt{1-0.842^2}} = 6.62.$$

4. **Interpret results**. Since we have a one-tailed test and α=0.05 DF=n-2=18, so the value percentage points $t_{18,5\%}$=1.734. The interval value accepting H_0 is (t<1.734). The computing value of t= 6.62 does not belong to this accepting interval. So we reject H_0. It means that the relation between skill learning process and learning achievement is not weak in positive relation (in this case r=0.842 is strong relation).

Table 7.6: Correlation skill learning process and learning achievement

x	y	xy	x^2	y^2
85	85	7225	7225	7225
68	50	3400	4624	2500
87	90	7830	7569	8100
89	90	8010	7921	8100
78	75	5850	6084	5625
81	95	7695	6561	9025
84	85	7140	7056	7225
78	80	6240	6084	6400
70	65	4550	4900	4225
96	100	9600	9216	10000
84	75	6300	7056	5625
75	65	4875	5625	4225
79	80	6320	6241	6400
72	80	5760	5184	6400
82	80	6560	6724	6400
89	95	8455	7921	9025
78	80	6240	6084	6400
87	95	8265	7569	9025
75	75	5625	5625	5625
89	85	7565	7921	7225
1626	1625	133505	133190	134775

J. Simple Linear Regression

Many problems in education involve exploring the relationships between two or more variables. **Regression analysis** is a statistical technique that is very useful for these types of problems. For example, in a learning process, supposed that the yield of the product is related to the process-activity or –skill learning, regression analysis can be used to build a model to predict yield at a given activity level.

As an illustration, consider once more the data in Table 7.5. In this table *y* is the learning achievement produced through test after a learning process, and *x* is the skill of learning process that are observed in the learning process.

Figure 7.3 presents a **scatter diagram** of the data in Table 7.5.

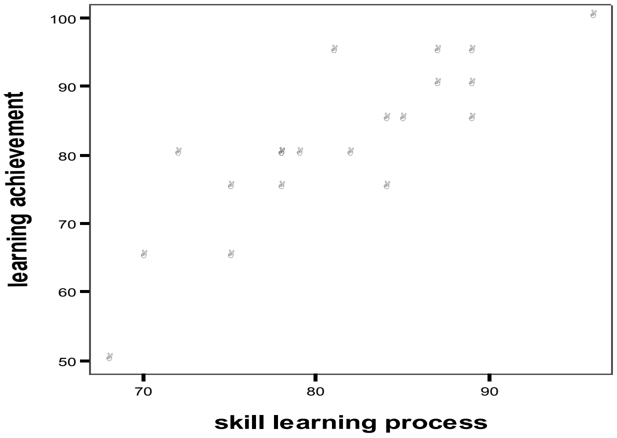


Figure 7.3: The scatter plot of learning achievement versus skill learning process

Inspection of this scatter diagram indicates that although no simple curve will pass exactly through all the points, there is a strong indication that the points lie scattered randomly around a straight line. Therefore, it is probably reasonable to assume that the mean of the random variable *Y* is related to *x* by the straight-line.

The case of **simple linear regression** considers a single **regressor** or **predictor** *x* and a dependent or **response variable** *Y*. Supposed that the true relationship between *Y* and *x* is a straight line and that the observation *Y* at each level of *x* is a random variable, we assume that each observation, *Y*, can be described by the model

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

where ε is a random error with zero mean and (unknown) variance σ^2. The random errors corresponding to different observations are also assumed to be uncorrelated random variables. β_0 is a parameter constant, β_1 is the parameter coefficient regression, x is the value of the independent variable, and Y is the value of the dependent variable.

Given a random sample of observations, with the ordinary least square (OLS) strategy the population regression line is estimated by:

$$\hat{y} = b_0 + b_1 x$$

where b_0 is a constant, b_1 is the regression coefficient, x is the value of the independent variable, and ŷ is the *predicted* value of the dependent variable. We compute for b_0 and b_1 "by hand". Here are the equations,

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$b_0 = \bar{y} - b_1\bar{x}.$$

Normally, we will use a computational tool - a software package (e.g., Excel or SPSS) to find b_0 and b_1. You enter the *X* and *Y* values into your program or calculator, and the tool solves for each parameter.

Now, we will fit a simple linear regression model to the learning achievement data in Table 7.5. The following quantities may be computed from Table 7.6:

Table 7.6: Relation skill learning process and learning achievement

x	y	xy	x^2
85	85	7225	7225
68	50	3400	4624
87	90	7830	7569
89	90	8010	7921
78	75	5850	6084
81	95	7695	6561
84	85	7140	7056
78	80	6240	6084
70	65	4550	4900
96	100	9600	9216
84	75	6300	7056
75	65	4875	5625
79	80	6320	6241
72	80	5760	5184
82	80	6560	6724
89	95	8455	7921
78	80	6240	6084
87	95	8265	7569
75	75	5625	5625
89	85	7565	7921
1626	1625	133505	133190

Here n=20, then the value of b_1 and b_0 are as follows,

$$b_1 = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} = \frac{20(133505) - (1626)(1625)}{20(133190) - (1626)^2} = 1.374,$$

$$b_0 = \bar{y} - b_1\bar{x} = 81.25 - (1.374)(81.3) = -30.387.$$

The fitted simple linier regression model (with the coefficients reported to three decimal places) is,

$\hat{y}$ = -30.387 + 1.374x.

This model is plotted in figure 7.4, along with the sample data.

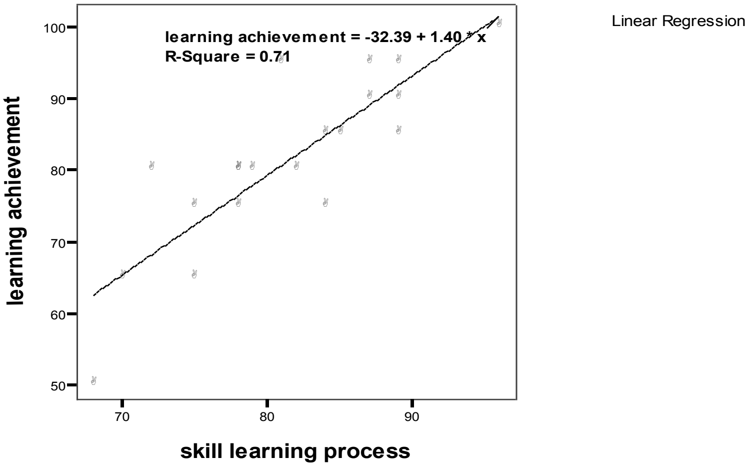


Figure 7.4: Scatter plot of learning achievement versus skill learning process and regression

Properties of the Regression Line

When the regression parameters (b0 and b1) are defined as described before, the regression line has the following properties.

1. The line minimizes the sum of squared differences between observed values (the *y* values) and predicted values (the ŷ values computed from the regression equation).
2. The regression line passes through the mean of the *X* values (x) and the mean of the *Y* values (y).
3. The regression constant (b0) is equal to the y intercept of the regression line.
4. The regression coefficient (b1) is the average change in the dependent variable (*Y*) for a 1-unit change in the independent variable (*X*). It is the slope of the regression line.

The least squares regression line is the only straight line that has all of these properties.

The Coefficient of Determination

The **coefficient of determination** (denoted by R^2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

1. The coefficient of determination ranges from 0 to 1.
2. An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
3. An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.
4. An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*; an R^2 of 0.20 means that 20 percent is predictable; and so on.

The formula for computing the coefficient of determination for a linear regression model with one independent variable is given bellow,

$$R^2 = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2}$$

where $\hat{y} = b_0 + b_1x$, $\acute{y}$ is mean of variable dependent Y, y_i is the Y value of observation.

In special case simple linier regression model $\mathbf{R^2} = \mathbf{r^2}$ (square of the correlation coefficient between x und y. For example the value of the coefficient determination from the data example 7.2 is,

$$R^2 = r^2 = (0.842)^2 = 0.708 = 70.8\%.$$

Test of the Simple Regression and the Interpretation

An important part of assessing the adequacy of a linear regression model is testing statistical hypotheses about the model parameters and constructing certain confidence intervals hypothesis. To test hypotheses about the slope of the regression model, we must make the additional assumption that the error component in the model, ε, is normally distributed. Thus, the complete assumptions are that the errors are normally and independently distributed with zero mean and variance σ^2.

The procedure for testing in a linear regression is the same as our last discussion in chapter 6. A hypothesis test for regression consists of four steps:

1. **State the hypothesis**.

   The first step is to state the null hypothesis and an alternative hypothesis.

   H_0: $b_1 = 0$ (equation is not linear or there is no relation x to y)

   H_1: $b_1 \neq 0$ (equation is linear or there is relation x to y)

2. **Fix a level of significant.**

   For this analysis, determine the significance level.

3. **Analyse sample data**

   For this analysis, we use the sample data to find the statistic test and its associated F-score. The appropriate statistic test for these hypotheses is:

   $$F_0 = \frac{R^2(n-2)}{1-R^2}$$

   Where n is the number of observations and R^2 is coefficient determination. The form of F_0 has the F distribution with 1,n-2 degrees of freedom.

4. **Interpret results**.

We accept H$_0$, if $f_0 < f_{\alpha,1,n-2}$, where f$_{\alpha,1,n-2}$ percentage points of distribution F with the significant level α and the degrees of freedom numerator 1 and denominator n-2, otherwise, H$_0$ is rejected.

If from the test we reject the null hypothesis, we can use the value of coefficient determination R^2 as the explanation about predictable of dependent variable.

Example 7.3: We will test the significance of regression using the data learning achievement from Table 7,5. In this case we look for the influence of skill learning as independent variable process (x) to learning achievement as dependent variable (y).

Solution: The steps of test hypothesis are:

1. **State the hypothesis**.

   H$_0$: $b_1 = 0$ (equation is not linear or there is no relation x to y)

   H$_1$: $b_1 \neq 0$ (equation is linear or there is relation x to y)

2. **Fix a level of significant.**

   We will use the significance level α=5%.

3. **Analyse sample data**:

   We will compute the coefficient determination from data Table 7.5. Recall from the calculating coefficient correlation that r=0.842, so that R^2=r^2=(0.842)^2 =0.709. Next we find the value of f$_0$:

$$f_0 = \frac{R^2(n-2)}{1-R^2} = \frac{0.709(20-2)}{1-0.709} = 43.85.$$

4. **Interpret results**.

   We use α=0.05, degrees of freedom 1,18, so that the value percentage points f$_{1,18,5\%}$=4.41. The interval value accepting H$_0$ is

($f<4.41$). The computing value of $f_0 = 43.85$ does not belong to this acceptance interval, therefore we reject H_0. It means the equation regression is linear or there is a linear relationship between skill learning process (x) and learning achievement (y).

We have R^2=0.709=70.9%. It means variable x (skill learning process) affects variable y (learning achievement) as 70.2%. 70.2% of the variance y is predictable from x. Therefore 29.8% variance y is predictable from another variable x.

How to Use the Regression Equation

Here are the steps in how we can use the regression equation. Choose a value for the independent variable (*x*), and then perform the computation. So we have an estimated value (ŷ) for the dependent variable. In our example, the independent variable is the students' scores on the skill learning process. The dependent variable is the students' score on the learning achievement. If a student made an 80 on the skill learning process, by using the regression model, we would predict learning achievement ŷ as follows:

$$\hat{y} = -30.387 + 1.374x =-30.387+ 1.374(80) = 79.53.$$

The learning achievement 79.53 may be interpreted as an estimation of the true population mean of learning achievement when x=80, or as an estimation of a new observation when x=80. These estimation is, of course, a subject to error, which means that it is unlikely that a future observation on learning achievement would be exactly 79.53 when the skill learning process is 80.

Warning: When we use a regression equation, do not use values for the independent variable that are outside the range of values used to create the equation. That is called **extrapolation**, and it can produce unreasonable estimations. In this example, the skill learning processes used to create the regression equation ranged from 68 to 96. Therefore, only use values inside the range to estimate learning achievements. Using values outside the range (less than 68 or greater than 96) is problematic.

Statistics: Residuals and Outliers

A linear regression model is not always appropriate for the data. We can assess the appropriateness of the model by examining residuals, outliers, and influential points.

Residuals

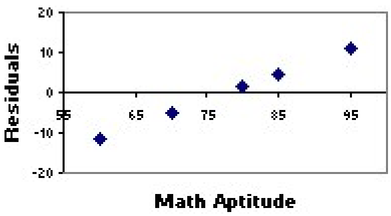
The difference between the observed value of the dependent variable (*y*) and the predicted value ($\hat{y}$) is called the **residual** (*e*). Each data point has one residual. Residual = Observed value - Predicted value $e = y - \hat{y}$. Both the sum and the mean of the residuals are equal to zero. That is, $\Sigma\ e = 0$ and $\bar{e} = 0$.

A **residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

Below, the table on the left presents results from a hypothetical regression analysis, and the chart on the right displays those results as a residual plot. In the chart, the independent variable (x) is math aptitude.

The residual plot shows a non-random pattern - negative residuals on the low end of the X axis and positive residuals on the high end. This indicates that a non-linear model will provide a much better fit to the data. Or it may be possible to "transform" the data to allow us to use a linear model. We discuss linear transformations in the next lesson.

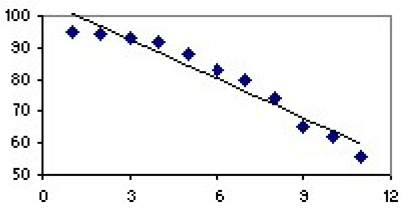
x	95	85	80	70	60
y	85	95	70	65	70
ŷ	74.05	90.49	68.71	70.159	81.59
e	10.95	4.51	1.29	- 5.159	- 11.59



Outliers

Data points that diverge from the overall pattern and have large residuals are called outliers. Outliers limit the fit of the regression equation to the data. This is illustrated in the scatter plots below. The coefficient of determination is bigger when the outlier is not present.

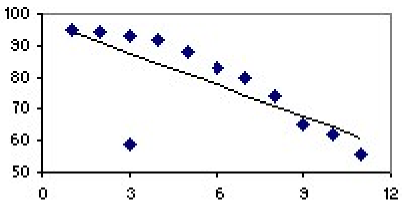
Without Outlier



Regression equation: ŷ = 104.78 - 4.10x

Coefficient of determination: R^2 = 0.94

With Outlier



Regression equation: ŷ = 97.51 - 3.32x

Coefficient of determination: R^2 = 0.55

Exercise

1. ***Express what is the meaning of correlation between two variables.***

2. ***What is the meaning that coefficient correlation in the extreme conditions.***
3. ***The following scores were obtained in a study of 20 subjects who were measured on two scales: a measure of self reported loneliness (L) and a measure of dissastisfaction with existing social relationships (D).***

Subject	*L*	*D*
1	***2***	***6***
2	***3***	***6***
3	***7***	***4***
4	***7***	***6***
5	***9***	***5***
6	***9***	***3***
7	***4***	***2***
8	***5***	***3***
9	***7***	***7***
10	***8***	***9***
11	***8***	***7***
12	***6***	***8***
13	***4***	***9***
14	***1***	***2***
15	***2***	***3***
16	***1***	***3***
17	***2***	***2***
18	***4***	***4***
19	***6***	***5***
20	***6***	***4***

a. ***Make a scatterplot of the data.***

b. ***Compute the value of coefficient correlation***

c. ***Test whether the Pearson correlation between these measures is significantly different from zero.***

4. ***The following data for two variable x and y, were obtained for the sample of 20 subjects.***

Subject	*L*	*D*	*Subject*	*L*	*D*
1	***44***	***28***	***``11***	***34***	***13***
2	***37***	***27***	***12***	***54***	***23***
3	***32***	***38***	***13***	***46***	***14***
4	***31***	***30***	***14***	***35***	***15***
5	***49***	***26***	***15***	***64***	***36***
6	***42***	***29***	***16***	***54***	***22***
7	***50***	***23***	***17***	***34***	***23***
8	***39***	***24***	***18***	***54***	***17***
9	***43***	***31***	***19***	***33***	***18***
10	***34***	***36***	***20***	***45***	***24***

a. ***Make a scatterplot of the data.***
b. ***Find the value of intercept and coefficient of regression***
c. ***Find the regression equation***
d. ***Find the value of coefficient correlation and coefficient of determination***
e. ***Test whether the coefficient of regression is significantly different from zero.***
f. ***What is your conclusion about the upper analysis.***

5. We use the data from Sunarnyo 2004 (see chapter 2) "*The relationship between running speed and arm muscle strength and long jump result*

x1: Running speed, x2: arm muscle strength and y: long jump

o	X1	X2	Y
1	5.56	42.0	4.17
2	5.89	24.5	3.46
3	5.66	26.5	3.45
4	5.42	24.0	3.88
5	5.18	26.0	4.05
6	5.15	50.0	4.41

Continu	X1	X2	Y
31	5.76	26.0	4.05
32	5.89	37.0	4.00
33	5.62	41.0	4.35
34	5.23	26.5	4.70
35	5.33	29.5	3.91
36	5.14	57.0	4.55

7	5.09	65.0	4.58	37	6.28	27.5	3.62
8	6.21	21.0	3.69	38	5.82	29.0	4.17
9	5.53	37.5	3.89	39	5.55	36.0	3.85
10	6.19	38.0	3.35	40	5.56	27.0	3.90
11	5.16	32.5	4.33	41	5.50	21.0	3.40
12	5.15	46.5	4.73	42	5.25	41.0	4.23
13	5.68	30.0	3.67	43	5.34	44.0	4.58
14	5.22	61.0	4.27	44	5.96	46.0	4.01
15	5.09	40.5	4.13	45	5.47	40.0	4.04
16	5.17	45.0	4.84	46	6.02	27.0	3.85
17	5.44	40.0	4.63	47	5.09	57.5	4.59
18	5.48	26.5	4.19	48	6.00	29.5	3.90
19	5.70	47.5	3.74	49	5.00	34.0	4.47
20	5.75	30.0	4.30	50	5.19	35.0	3.88
21	5.71	24.0	3.82	51	5.74	31.5	4.12
22	5.96	27.0	3.52	52	6.01	25.0	3.41
23	5.07	46.0	5.03	53	5.78	44.0	4.43
24	6.22	29.0	4.30	54	5.07	40.5	4.17
25	6.25	31.0	3.82	55	5.78	25.0	3.54
26	5.43	39.0	4.15	56	4.90	55.0	4.82
27	5.17	43.0	4.06	57	5.87	44.0	4.09
28	5.84	39.0	3.85	58	5.24	50.5	4.32
29	5.22	48.0	4.31	59	5.44	25.0	3.57
30	5.32	36.0	4.26	60	5.24	54.5	4.30

a. ***Test the correlation between x1, y and x2, y with the confidence interval 5%.***

b. ***Test with regression the independent variable x2 give influence to variable dependent y, with the confidence interval 5%.***

Appendix

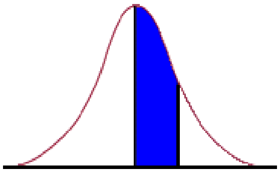


Table I: Cumulative Standard Normal Distribution

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0	0.004	0.008	0.012	0.016	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	0.091	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	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	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	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	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	0.334	0.3365	0.3389
1	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	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	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	0.437	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	0.475	0.4756	0.4761	0.4767
2	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.483	0.4834	0.4838	0.4842	0.4846	0.485	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.489
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	0.492	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.494	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	0.496	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.497	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	0.498	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	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.499	0.499

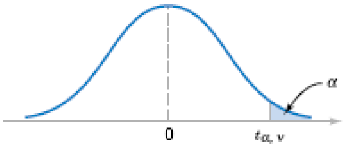


Table II Percentage Points $t_{\alpha,\nu}$ of the *t*-Distribution

α ν	.40	.25	.10	.05	.025	.01	.005	.0025	.001	.0005
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	.289	.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	.258	.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	.256	.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	.256	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	.255	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	.254	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

ν = degrees of freedom.

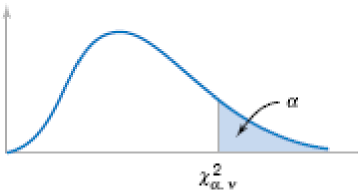


Table III Percentage Points χ^2 of the Chi-Squared Distribution

ν \ α	.995	.990	.975	.950	.900	.500	.100	.050	.025	.010	.005
1	.00+	.00+	.00+	.00+	.02	.45	2.71	3.84	5.02	6.63	7.88
2	.01	.02	.05	.10	.21	1.39	4.61	5.99	7.38	9.21	10.60
3	.07	.11	.22	.35	.58	2.37	6.25	7.81	9.35	11.34	12.84
4	.21	.30	.48	.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	.41	.55	.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	.68	.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.28	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17

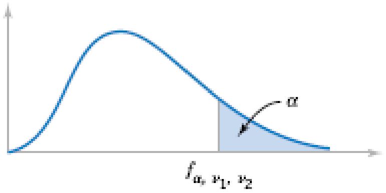


Table IV Percentage Points $f_{\alpha,v1,v2}$ of the F Distribution

$f_{0.25,\nu_1,\nu_2}$

Degrees of freedom for the denominator (ν_2)

ν_1	Degrees of freedom for the numerator (ν_1)																		
ν_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.41	9.49	9.58	9.63	9.67	9.71	9.76	9.80	9.85
2	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.41	3.43	3.43	3.44	3.45	3.46	3.47	3.48
3	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47
4	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08
5	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.88	1.88	1.88	1.88	1.87	1.87	1.87
6	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.76	1.76	1.75	1.75	1.75	1.74	1.74	1.74
7	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.70	1.69	1.68	1.68	1.67	1.67	1.66	1.66	1.65	1.65	1.65
8	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63	1.62	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58
9	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.54	1.53	1.53
10	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.53	1.52	1.52	1.51	1.51	1.50	1.49	1.48
11	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.47	1.46	1.45
12	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.45	1.44	1.43	1.42
13	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40
14	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.41	1.41	1.40	1.39	1.38
15	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45	1.44	1.43	1.41	1.41	1.40	1.39	1.38	1.37	1.36
16	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34
17	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33
18	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33	1.32
19	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.33	1.32	1.30
20	1.40	1.49	1.48	1.47	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.29
21	1.40	1.48	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28
22	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39	1.37	1.36	1.34	1.33	1.32	1.31	1.30	1.29	1.28
23	1.39	1.47	1.47	1.45	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28	1.27
24	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38	1.36	1.35	1.33	1.32	1.31	1.30	1.29	1.28	1.26
25	1.39	1.47	1.46	1.44	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.32	1.31	1.29	1.28	1.27	1.25
26	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37	1.35	1.34	1.32	1.31	1.30	1.29	1.28	1.26	1.25
27	1.38	1.46	1.45	1.43	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.33	1.32	1.31	1.30	1.28	1.27	1.26	1.24
28	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.31	1.30	1.29	1.28	1.27	1.25	1.24
29	1.38	1.45	1.45	1.43	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.32	1.31	1.30	1.29	1.27	1.26	1.25	1.23
30	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35	1.34	1.32	1.30	1.29	1.28	1.27	1.26	1.24	1.23
40	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.31	1.30	1.28	1.26	1.25	1.24	1.22	1.21	1.19
60	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30	1.29	1.27	1.25	1.24	1.22	1.21	1.19	1.17	1.15
120	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.26	1.24	1.22	1.21	1.19	1.18	1.16	1.13	1.10
∞	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24	1.22	1.19	1.18	1.16	1.14	1.12	1.08	1.00

$f_{0.10,v_1,v_2}$

Degrees of freedom for the denominator (v_2)

v_2 \ v_1	Degrees of freedom for the numerator (v_1)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.03	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

$f_{0.05,\nu_1,\nu_2}$

Degrees of freedom for the denominator (ν_2)

ν_1 / ν_2	Degrees of freedom for the numerator (ν_1)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.55	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

$f_{0.025,v_1,v_2}$

v_1 / v_2	Degrees of freedom for the numerator (v_1)																		
Degrees of freedom for the denominator (v_2)	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

$f_{0.01,v_1,v_2}$

Degrees of freedom for the denominator (v_2)

v_1 / v_2	Degrees of freedom for the numerator (v_1)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.00	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.46
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.36	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.59
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

REFERENCES

Agarwal, BL. 2006. *Basic Statistics.* Fourth Edition. New Age International Limited Publishers. New Delhi.

Antodajan. 1986. *Pengantar Metode Statistik I.* Edisi 2, LP3ES Jakarta

Bluman. 2001. *Elementary Statistics: A Step by Step Approach*. McGraw-Hill.

Draper, NR and Smith, H. 1981. *applied Regression Analysis.* Second Edition. John Wiley and Sons New York.

Harper, WM. 1971. *Statistics.* Macdonald & Evans LTD, London.

Herzberg, PA. 1983. *Principles of Statistics.* John Wiley & Sons Canada.

Holmes, P. 1979. *Stochastik in der Schule.* Band 1 Nummer 2. Fachbereich Statistik Universität Dortmund.

Iversen, GR and Gergen, M. 1997. *Statistics: The Conceptual Approach.* Springer Verlag Berlin.

Krämer, W. 1998. *Statistik Verstehen,* 3. auflage. Frankfurt/Main:Campus Verlag.

Krämer, W. 1994a. *Überzeugt man mit Statistik.* Frankfurt/Main: Campus Verlag.

Krämer, W. 1994b. *So Lügt Man mit dem Statistik.* Frankfurt/Main: Campus Verlag.

Rencher, AC. 2000. *Linear Models in Statistics.* John Wiley & Sons Canada.

Rinne,H., 1997, *Taschenbuch der Statistik,2.* ueberarbeitete und erweitererte Auflage, Verlag Harri Deutsch.

Sudjana, 1992, *Metode Statistika,* Edisi 5, Tarsito Bandung.