Course Title
Course Code: MAT 212
Course Unit: 2 Units

Module 1 Descriptive statistics and probability
Title of the Units

1. Introduction to statistics
2. Summary and Display of Data
3. Descriptive Analysis
4. Introduction to probability

End of Module Assessment

Module 2: Inferential statistics
Title of the Units

1. Test if statistical Hypotheses
2. Correlation and Regression Analysis

End of Module Assessment

Module 1: Descriptive Statistics and Probability

Unit 1: Introduction to Statistics

At the end of this lesson, you should be able to

(I) Define some basic terms used in statistics

(II) Understand what is data and the branches

1.1 Introduction

Statistics is the science of collecting, classifying, presenting, and interpreting data. Our society has developed into one where science and technology affect everything around us. Statistics is one of the most important of these scientific tools. Virtually all facets of our lives are affected by statistics. Statistics has become a necessary element in most academic fields including the sciences, engineering, business, political science, economics, psychology, sociology, education, medicine, nursing, and other health-related areas.

Statistics is the universal language of the sciences. Statistics is more than just a "kit of tools". As potential users of statistics; we need to master the "art" of using these tools correctly. Careful use of statistical methods enables us to:

(1) Accurately describe the findings of scientific research
(2) Make decisions and
(3) Make estimations

The field of statistics can be roughly subdivided into two areas: descriptive statistics and inferential statistics. Descriptive statistics is what most people think of when they hear the word *statistics*. It includes the collection, presentation, and description of data. The term inferential statistics refers to the technique of interpreting the values resulting from the descriptive techniques

and then using them to make decisions and draw conclusions about the population.

1.2 INTRODUCTION OF BASIC TERMS

Some basic terms that will be used throughout this book are presented.

Population: A collection, or set, of individuals or objects whose properties are to be analyzed. The concept of a population is the most fundamental idea in statistics. The population of concern must be carefully defined and is considered fully defined only when its membership list of elements is specified. The set of "all students who take Algebra course in year two" is an example of a well-defined population. Another is the set of "all lecturers in University of Lagos".

Population involves not only people but also a collection of animals, manufactured objects, or whatever. There are two types of population, finite and infinite. When the members in a population can be physically counted, the population is said to be finite. It is infinite when the membership is uncountable. The number of students in University of Lagos is a finite population. The set of all registered voters in Nigeria is a very large finite population. On the other hand, the population of all stars in the sky and the population of all sands at the seashore all over the world are infinite.

Sample

A sample is a subset of a population. A subset consists of the individuals, objects, or measurements selected by the sample collector from the population. For example, a set of males in the

department of mathematics is a subset of the number of students in the department. A set of Toyota cars parked at the faculty car park is a subset of cars parked at the faculty car park.

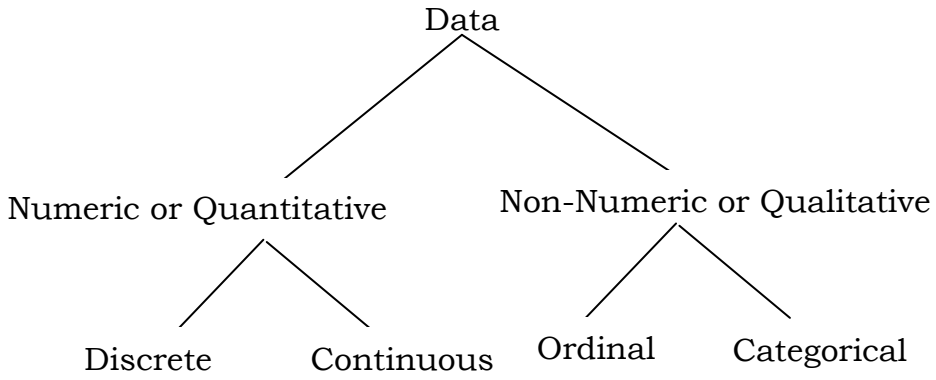
Variable

A variable is a characteristic of interest about each individual element of a population or sample. A student's name, matriculation number, year and department are all variables.

Data

Data are raw facts or unprocessed information. There are basically two types of data:

(1) Data obtained from numeric or quantitative information and
(2) Data obtained from non-numeric or qualitative information. These classifications are given below:



Non-numeric data are values that cannot be quantified. For example, matriculation number, tribe, country, etc. Data in this form are either categorical or ordinal. Examples of ordinal non-numeric data are students' height, Age group, while sex, country, tribe, etc are examples of categorical non-numeric data.

Numeric data can be subdivided into two classifications:

(1) Discrete numeric data and

(2) Continuous numeric data. Counts will always yield discrete numeric data, e.g. the number of students in a school. A measure of a quantity will usually be continuous, e.g. weight of weight lifters.

Statistic

A statistic is a quantity whose numerical values can be obtained from data. A statistic is a value that describes a sample. Most sample statistics are found with the aid of formulas. For example, mean, median, mode etc.

Experiment

A planned activity whose results yield set of data is known as experiment.

1.3 DATA COLLECTION

One of the first problems a statisticians faces is obtaining data. Data can be collected directly from respondents or from established data bank. Data collected directly from the source or respondents are known as primary data and those from established data bank are known as secondary data.

Primary data collection for statistical analysis in an involved process and includes the following important steps.

1. Defining the objectives of the survey or experiment. Example: estimating the average height of female students in UNILAG.
2. Defining the target population
3. Defining the strategy and method to be used for data-collection and data measuring.
4. Ascertaining the appropriate descriptive or inferential data-analysis to employ.

There are two methods used to collect data. These are *experiments* and *surveys*. In an experiment, the investigator controls or modifies the environment and observes the effect on the response variable. This is common in laboratories. In a survey, data are obtained by sampling some population of interest. Various methods that might be used in order to obtain sample data from surveys are presented below. When selecting a sample for a survey; it is necessary to construct a *sampling frame*. A sampling frame is a list of the elements that belongs to the population from which the sample is drawn. An example is a list of all students in year one, Mathematics department.

1.4 EXERCISES

14.2 Identify each of the following as examples of

1. Non-numeric 2.Discrete 3.Continuous variables:

a) The hair colour of people in a concert show.
b) The number of hours required to heal a patient of a disease.
c) The length of time required answering a telephone call at a certain business center.
d) The number of pages per job coming off a computer printer.
e) The kind of trees used as Christmas tree.
f) The number of voters in a community.
g) Whether a statement is true or false.
h) The number of books in a library.

1.4.3 Define and explain the following terms:

a) Population b) Sample c) Statistic
d) Statistics e) Variable f) Data
f) Experiment

CHAPTER TWO

SUMMARY AND DISPLAY OF DATA

2.1 FREQUENCY DISTRIBUTION

Listing large set of data does not present much of a picture to the reader. Sometimes we want to condense the data into a more manageable form. This can be accomplished with the aid of a *Frequency distribution.*

Let us demonstrate the concept of a frequency distribution by using the following set.

1	5	3	4	1	3	2	5
2	4	1	3	2	0	1	2
1	2	0	2	1	4	5	3

Let x represent these data values, we can use a frequency distribution to represent this set of data by listing the x values with their frequencies in Table 2.1.

Table 2.1 Frequency distribution

X	0	1	2	3	4	5
F	2	6	6	4	3	3

In the case where many different entries for x and several low frequencies, it often makes sense to combine the data in groups or *classes*. Let us demonstrate this with this example:

55	60	61	35	41	43	50	78	72	83
45	70	76	31	49	65	79	83	41	86
53	62	52	47	38	57	64	78	47	54
43	73	85	48	66	48	85	86	82	48
56	84	37	57	57	45	95	45	73	39

The following guidelines and terminology will be used to group continuous-type data into classes of equal length. These guidelines can also be used for sets of discrete data that have a large range.

1. Determine the largest (maximum) and smallest (minimum) observations. The *range* is the difference,
   R = maximum – minimum

2. A frequency distribution should have a minimum of 5 classes and a maximum of 20. For small data sets, use between 5 and 10 classes. For large data sets, use up to 20 classes.

3. Each data entry must fall into one and only one class.

4. There should be no gaps. Moreover, if there are no entries for a particular class, that class must still be included with a frequency of 0.

5. The first interval should begin about as much below the smallest value as the last interval ends above the largest.

6. The intervals are called *class intervals* and the boundaries are called *class boundaries***.**

7. The *class limits* are the smallest and largest possible observed values in a class.

8. The *class mark* is the midpoint of a class.

We set up the following classes for the above data 30 – 39, 40 – 49, 50 – 59, etc. We now create a summary table below in Table 2.2.

Table 2.2: Frequency distribution

Class	Class limits	Tally	Frequency	Class Mark	Relative Frequency
1	30 – 39	~~IIII~~	5	35	10%
2	40 – 49	ÌÑI ÌÑI III	13	45	26%
3	50 – 59	~~IIII~~ IIII	9	55	18%
4	60 – 69	ÌÑ I	6	65	12%
5	70 – 79	~~IIII~~ III	8	75	16%
6	80 – 89	~~IIII~~ III	8	85	16%
7	90 – 99	I	1	95	2%
			50		100%

Tables like this show us how the data are spread out or distributed; we call this a *frequency distribution table* or simply a *frequency distribution*.

The relative frequency for a class is the number of entries in the class divided by the total number of entries. For example the relative frequency of class 50 – 59 is

$$\frac{9}{50} \times 100\% = 18\%$$

The next type of tabular display is known as a *cumulative frequency distribution*, which (as its name suggests) contains a column for the running cumulative total of frequencies for all

classes. The cumulative frequency of a class is the total of all class frequencies up to and including the present class.

The cumulative frequency distribution of the example given above is as follows:

Table 2.3: Cumulative Frequency Table

Class	Class limits	Frequency	Cumulative Frequency	Relative Cum. Frequency
1	30 – 39	5	5	10%
2	40 – 49	13	18	36%
3	50 – 59	9	27	54%
4	60 – 69	6	33	66%
5	70 – 79	8	41	82%
6	80 – 89	8	49	98%
7	90 – 99	1	50	100%

Relative Cumulative Frequency is also called *Percentage Cumulative Frequency*. For example the Relative Cumulative Frequency for class 60 – 69 is

$$\frac{33}{50} \times 100\% = 66\%$$

2.2 GRAPHIC PRESENTATION OF DATA

One of the most helpful ways to become acquainted is to use an initial exploratory technique that will result in a pictorial representation of the data. The displays visually reveal patterns of behaviour of the variable being studied. There are several graphic (pictorial) ways to describe data. The type of data and the idea to be presented determines the method used.

Data can be presented graphically in many ways as, line graph, dot plot display, bar chart, pie chart, histogram, cumulative frequency curve (Ogive) and stem-and-leaf display.

2.2.1 Dot Plot Display

Dot plots display the data of a sample by representing each piece of data with a dot positioned along a scale. This scale can be either horizontal or vertical. The frequency of the values is represented along the other scale. They are usually used to represent the frequency distribution of a discrete variable. The dot plot display is a convenient technique to use as you first begin to analyze the data. It results in a picture of the data as well as sorts the data into numerical order.

Example 2.1: A random sample of 20 children took their weights in kilogram in a hospital and are presented below:

23	22	26	28	22	29	30	25	26	27
21	23	27	26	25	29	30	26	25	28

Construct a dot plot of these data.

Solution

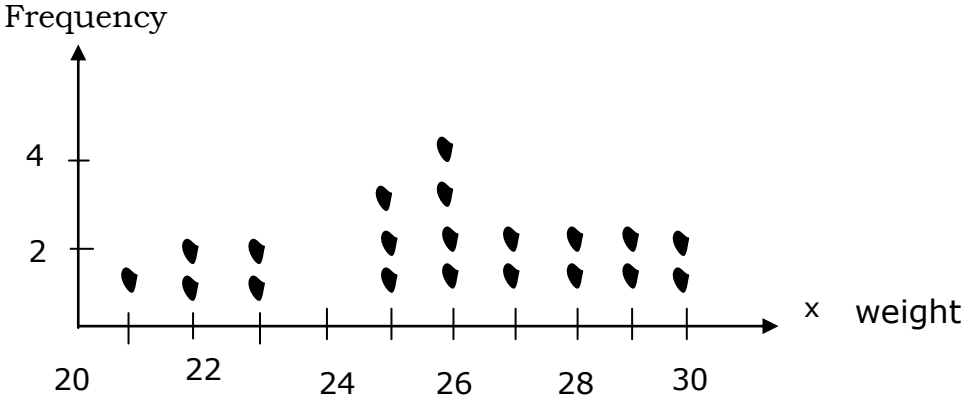
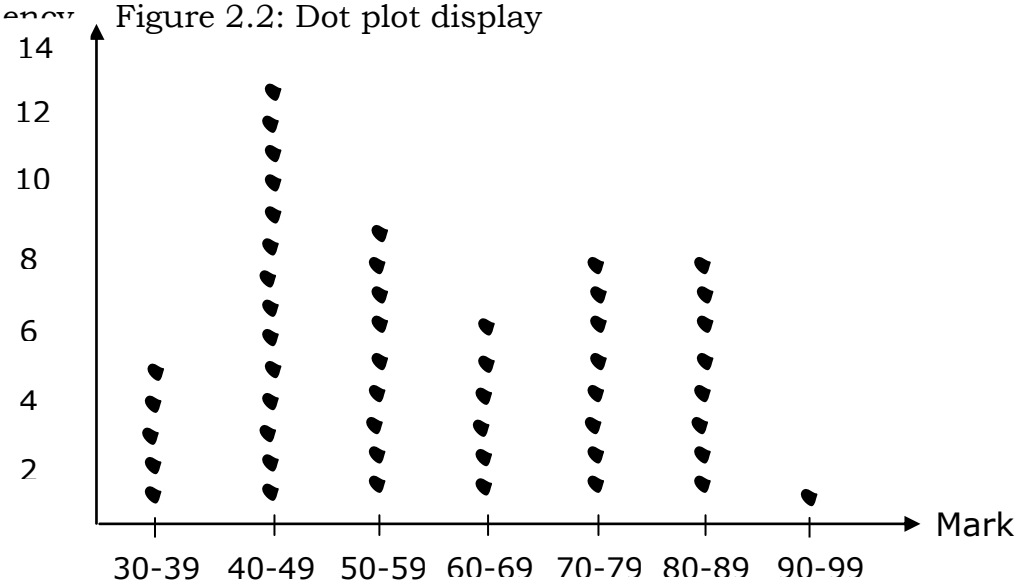


Figure 2.1 Dot plot of weights of children

Example 2.2: Use Table 2.2 to construct a Dot plot display

Solution

Figure 2.2: Dot plot display



2.2.2 Bar Chart

To construct a bar chart, we start with horizontal and vertical axes. We label the quantity being studied horizontally from left to right. The markings along the horizontal axis should correspond to the limits of the classes in the frequency distribution. The corresponding frequency in each class is measured vertically upward. A vertical bar is then drawn across each class interval with height equal to the frequency for that class. We could also draw a bar chart by using the relative frequencies instead of the frequencies for each class. The relative frequencies are measured along the vertical axis as percentages.

Example 2.3: Use table 2.2 to construct a frequency bar chart and a bar chart.

Solution

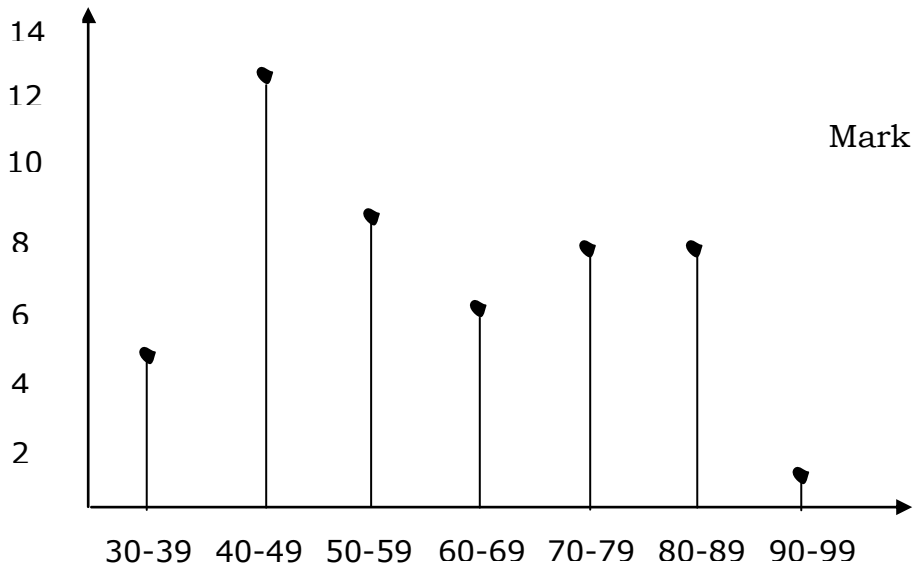


Figure 2.3: Frequency bar chart

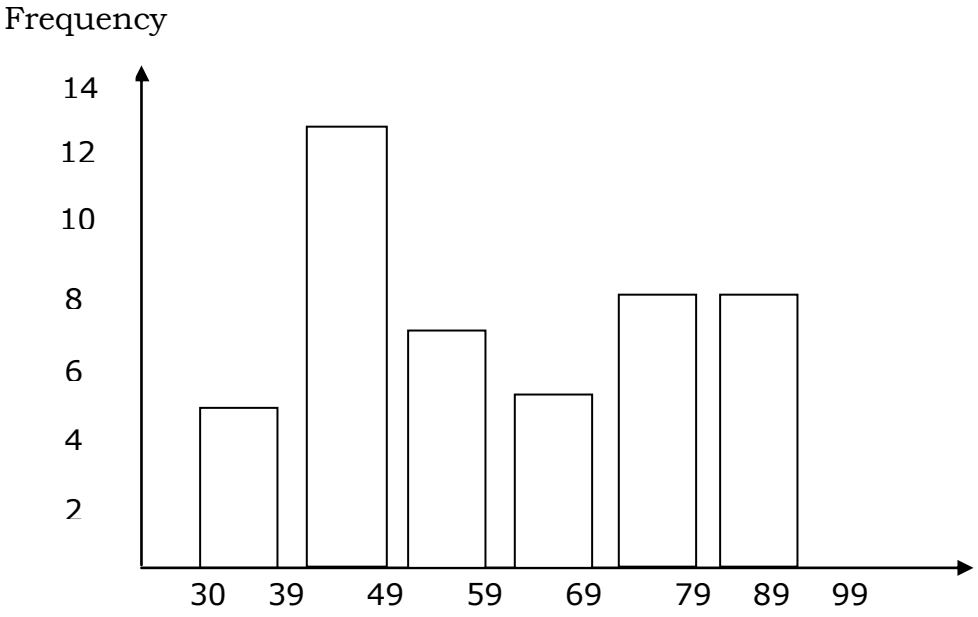


Figure 2.4: Bar chart

Example 2.4: A computer anxiety questionnaire was given to 300 children in a computer course. One of the questions was " I enjoy using computer." The responses to this particular question were

Table 2.4

Response	Strongly Agree	Slightly	Slightly Disagree	Strongly

	Agree		Agree	Disagree		Disagree
Number	60	85	40	50	35	30

Solution

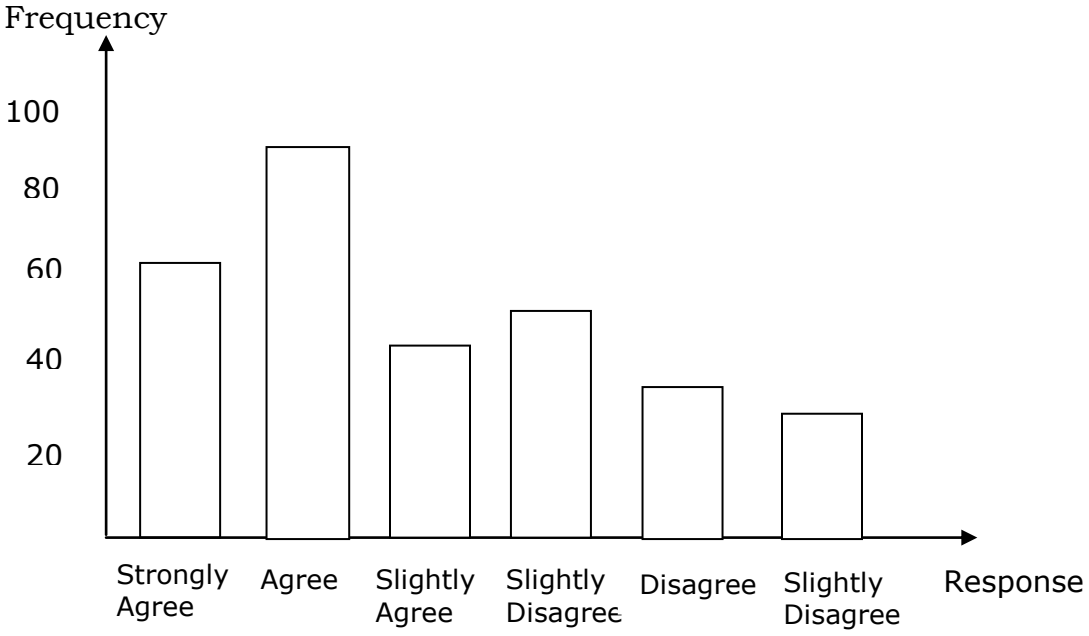


Figure 2.5: Bar chart of responses

Example 2.5: The following table shows the intake through JAMB by the Faculty of Science of a certain University in three consecutive years.

Table 2.5

Department	**2002**	**2003**	**2004**
Botany	43	40	35
Chemistry	28	35	42

Zoology	45	40	35
Computer Science	33	25	28
Physics	40	35	38
Mathematics	35	42	45
Biology	37	40	42
Total	**261**	**257**	**265**

Draw (i) a component bar chart.
(ii) multiple bar chart department by department for the three years.

Solution

(i)

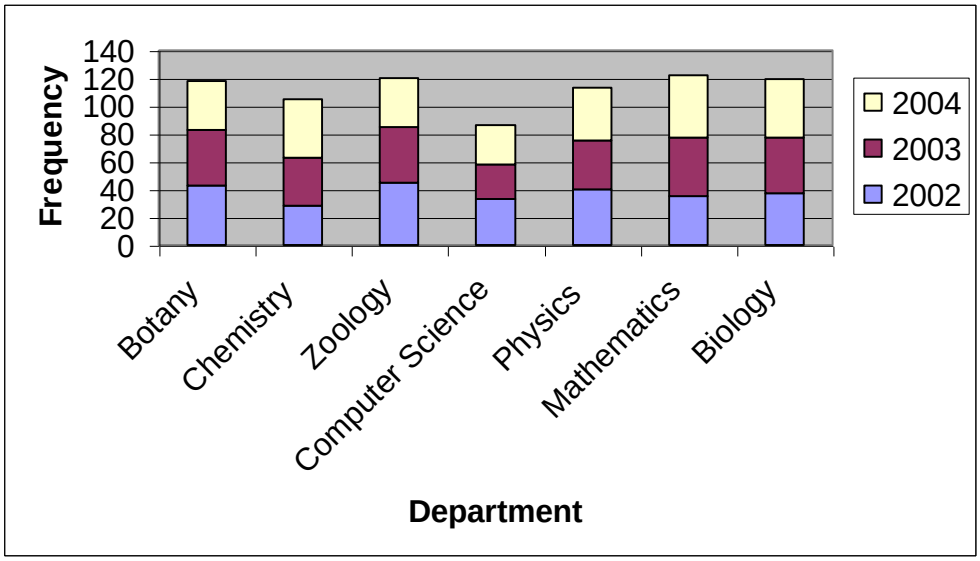


Figure 2.6: Component bar chart for JAMB Admission

(ii)

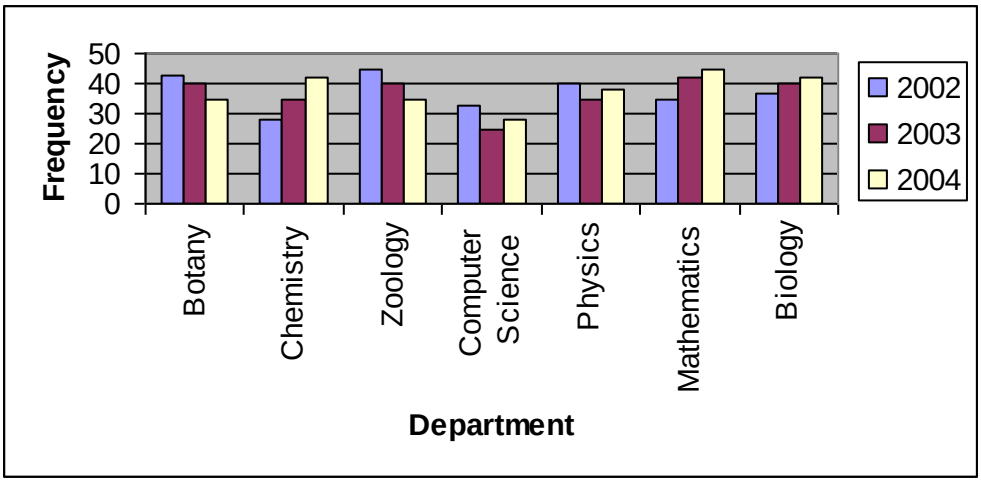


Figure 2.7: Multiple bar chart for JAMB Admission

2.2.3 Pie Chart

The pie chart (circle graph) is used to display relative frequencies rather than actual frequencies for the data. We draw a circle and then divide it into a series of wedges or slices to represent each class in the relative frequency distribution. The size of each slice is proportional to the percentage of the data that fall into the corresponding class.

Example 2.6 Represent the question in example 2.4 in a pie chart

Solution

$$\text{Angle for each class} = \frac{\text{Number in the class}}{\text{Total number of observations}} \times 360^0$$

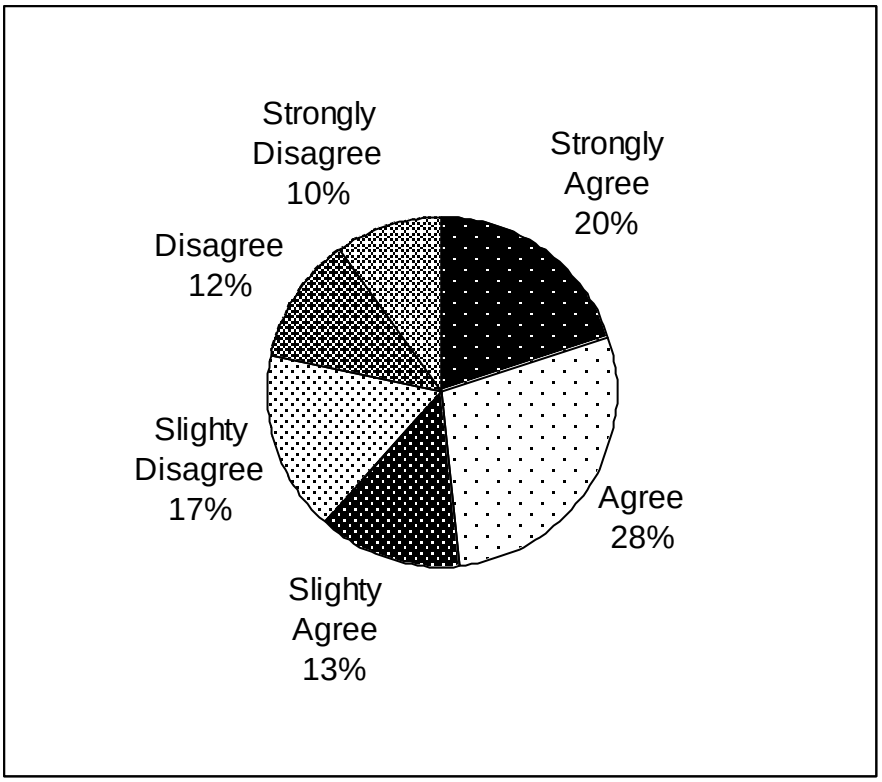


Figure 2.8: Pie chart for response

Response	**Number**	**Angles**
Strongly Agree	60	72^0
Agree	85	102^0
Slighty Agree	40	48^0
Slighty Disagree	50	60^0
Disagree	35	42^0
Strongly Disagree	30	36^0
Total	**300**	**360^0**

2.2.4 Histogram

The *histogram* is a type of bar chart representing an entire set of data. A histogram is made up of the following components:

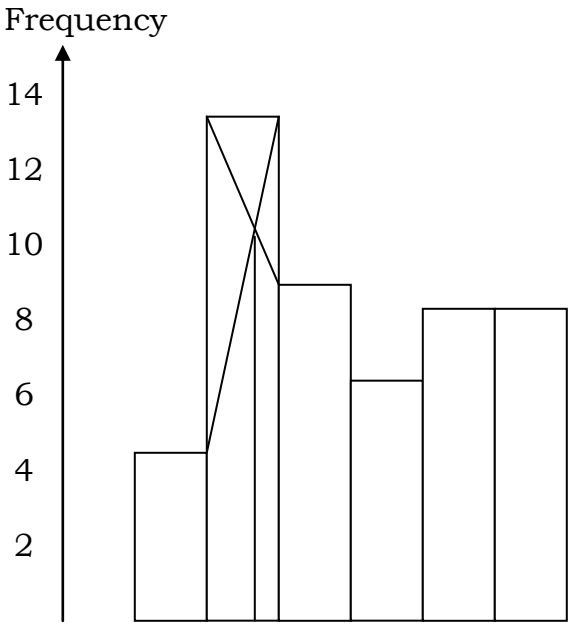
a. A title, which identified the population of concern.

b. A vertical scale, which identifies the frequencies in the various classes.
c. A horizontal scale, which identifies the variable.

Values for the class boundaries, class limits, or class marks may be labeled along the x-axis. Use whichever one of these sets of class numbers best represents the variable. Using the Table 2.2. We draw the histogram.

Table 2.6

Class	Class limits	Frequency	Class boundaries	Class center
1	30 – 39	5	29.5 – 39.5	34.5
2	40 – 49	13	39.5 – 49.5	44.5
3	50 – 59	9	49.5 – 59.5	54.5
4	60 – 69	6	59.5 – 69.5	64.5
5	70 – 79	8	69.5 – 79.5	74.5
6	80 – 89	8	79.5 – 89.5	84.5
7	90 – 99	1	89.5 – 99.5	94.5



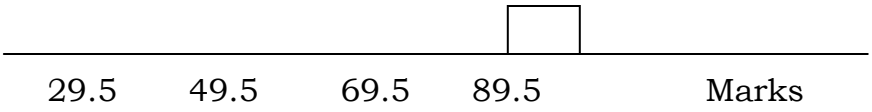


Figure 2.9: Histogram of Marks

In the histogram, a single vertical line between the first two boxes replaces the gap between 39 and 40. However, it is not clear what the vertical line should represent – is it 39 or 40 or what? To resolve this ambiguity, we agree that the vertical line represents 39.5, which is the class boundary between the two classes. In the same way, the next vertical line represents 49.5 and so forth.

Another type of graphical display is the *frequency polygon*. To construct this type of graph, we first determine the measurement corresponding to the midpoint of each class. This value is called the *class mark*, or *class center*, or *class midpoint* and is given by

$$\text{Class center} = \frac{\text{lower limit + upper limit}}{2}$$

For example, in the class 29.5 to 39.5, the class center is

$$\text{Class center} = \frac{29.5 + 39.5}{2} = 34.5$$

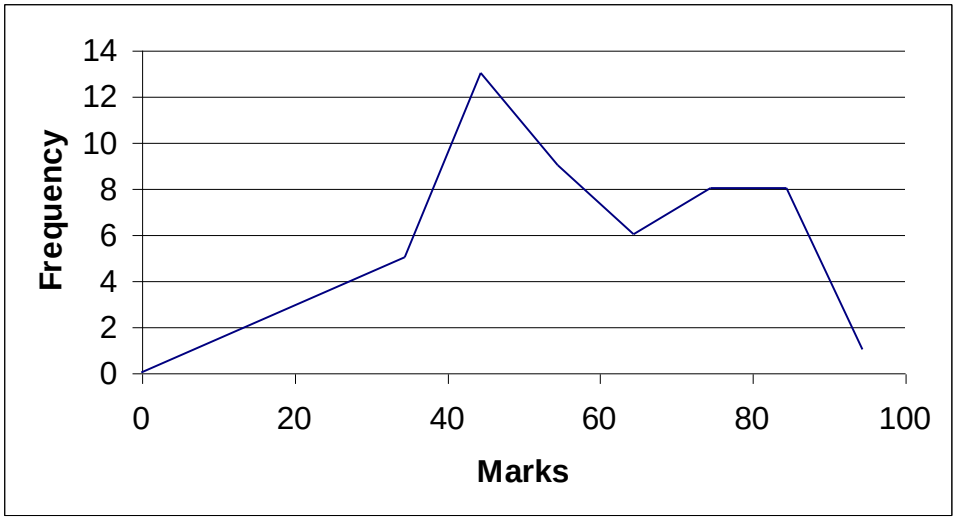


Figure 2.10: Frequency polygon

The *mode* is the value of the piece of data that occurs with the greatest frequency. From Figure 2.9, the mode is 46. To obtain this, use a ruler to connect both right and left edges of the tallest bar to the bars on both sides of the tallest bar, then locate their point of intersection and trace this down to the horizontal axis using a vertical broken line. Where this line meets the x-axis is the mode.

The *modal class* is the class with the highest frequency. A data set with two modes is called *bimodal.* A data set with three modes is called *trimodal*; if there are more than three modes, it is called *multimodal*.

We now present a *relative frequency histogram*. Note that the total area of this histogram is equal to one. The shape of this and that of the histogram is the same. The relative frequency histogram g(x) is

$$\frac{\text{Number in a class}}{\text{Total number of observation x class interval}}$$

Example 2.7
The following 30 gains were recorded to the nearest 1 million Naira of some private entrepreneurs.

1	1	1	1	1	1	1	1	1	1
3	3	3	3	4	4	6	8	9	10
12	12	13	14	28	32	34	36	39	40

Construct the relative frequency histogram.

Solution
Let c_0, c_1, c_2, c_3, and c_4 represent the class boundaries. Let c_o = 0.5 and c_1 = 3.5 with 14 observations in between; c_2 = 10.5 with 6 observations; c_3 = 29.5 with 5 observations and c_4 = 40.5 with 5 observations. This yield the following relative frequency histogram:

$$g(x)=\begin{cases}\dfrac{14}{(30)(3)}, & 0.5 < x \le 3.5\\[2ex] \dfrac{6}{(30)(7)}, & 3.5 < x \le 10.5\\[2ex] \dfrac{5}{(30)(19)}, & 10.5 < x \le 29.5\\[2ex] \dfrac{5}{(30)(11)}, & 29.5 < x \le 40.5\end{cases}$$

It is important to note in the case of unequal lengths among class intervals that the areas, not the heights, of the rectangles are proportional to the frequencies.

g(x)

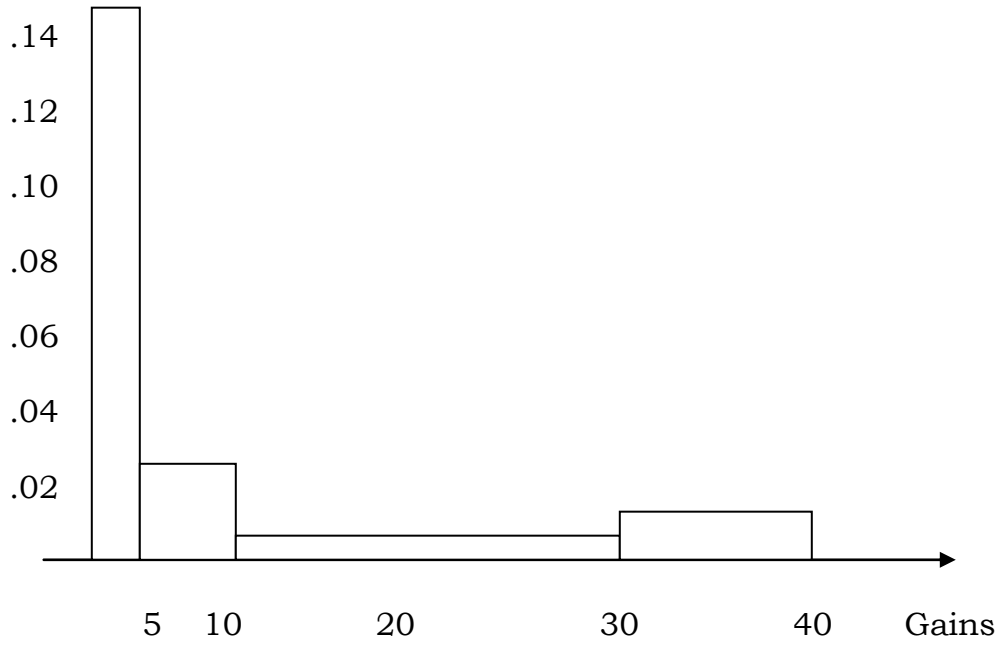


Figure 2.11: Relative frequency histogram

2.2.5 Cumulative Frequency Curve (Ogive)

A frequency distribution can easily be converted to a *cumulative frequency distribution* by replacing the frequencies with cumulative frequencies. This was shown in Table 2.3. The same information can be presented by using a *relative cumulative frequency distribution* (See Table 2.3). This combines the cumulative frequency idea and the relative frequency idea.

The vertical scale represents either the cumulative frequencies or the relative cumulative frequencies. The horizontal scale represents the upper class boundaries. Until the upper class boundary of a class has been reached, you cannot be sure you have accumulated all the data in that class. Therefore, the horizontal scale for an Ogive is always based on the upper class boundaries.

Every Ogive starts on the left with a relative frequency of zero at the lower class boundary of the first class and ends on the right with a relative frequency of 100% at the upper class boundary of the last class.

Example 2.8

Prepare an Ogive from Table 2.3

a. Give the estimates of the quartiles
b. Find the median
c. Estimate the 30 and 70 percentiles
d. Obtain the Range, Interquartile range and semi interquartile range
e. What number of students scored marks between 60% and 80%?
f. What will be the pass mark if 60% of the student failed?

Solution

Frequency

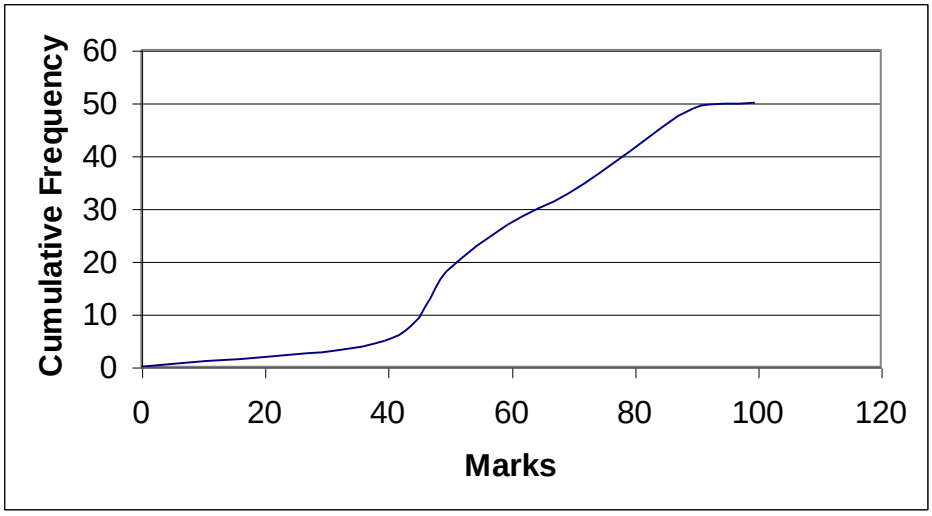


Figure 2.12: Cumulative frequency curve

a) Quartles

Q_1 = 25th percentile = 46.5
Q_2 = 50th percentile = 59.5
Q_3 = 75th percentile = 72

b) median is the 50th percentile and it is equal to 59.5

c) 30th percentile = 49.5
70th percentile = 68

d) i. Range = Highest mark - Lowest mark
= 95 – 31 = 64
from the raw data in section 2.1

ii. Interquartile range = $Q_3 - Q_1$
= 72 - 46.5 = 25.5

iii. Semi-Interquartile range = $\frac{Q_3 - Q_1}{2}$

$= \frac{72 - 46.5}{} = \frac{25.5}{}$

$$= 12.75 \quad \begin{matrix} 2 & 2 \end{matrix}$$

e) At 60% mark this intercept the curve at cumulative frequency of 25 students and at 80% mark this intercept the curve at cumulative frequency of 43. Therefore, the number of students that scored between 60% and 80% mark are 43 – 25 = 18 students

f) If 60% of the students failed, the pass mark will be from the 60^{th} percentile mark. Trace this to the curve and the pass mark will be 67.

2.2.6 Stem–and-Leaf

Stem-and-Leaf display combines the visual impact of the histogram or bar chart with the detail of the original list of data entries. This technique, very simple to create and use, is a combination of a graphic technique and a sorting technique. The data values themselves are used to do this sorting. The *stem* is the leading digit(s) of the data, and the *leaf* is the trailing digit(s). For example, the numerical data value 325 might be split 32 – 5 as shown:

Leading Digits	Trailing Digits
32	5

Example 2.9
Construct the stem-and-leaf of the following sets of data

i.
52	33	44	48	49	36	50	61	65	72
68	55	60	53	33	41	68	70	82	85
48	51	37	45	58	65	43	45	61	81

ii
1.6	1.9	3.5	4.9	8.2	7.5	3.3	3.8	4.5	5.2
2.7	4.8	5.7	6.2	7.8	3.4	5.7	8.3	4.1	1.6
2.7	3.1	2.4	1.8	4.5	7.1	3.3	2.5	5.6	1.8

Solution

i. In a stem-and-leaf plot, we consider all entries.

Let's look at the group of entries in the

30s:	33	33	36	37				
40s:	41	44	48	49	48	45	43	45
50s:	52	50	55	53	51	58		
60s:	61	65	68	60	68	65	61	
70s:	72	70						
80s:	82	85	81					

We separate the last digit of each entry from the primary numbers 30, 40, 50, 60, 70, 80 and we display the results in ascending order of the values:

Stem	Leaf
3	3 3 6 7
4	1 3 4 5 5 8 8 9
5	0 1 2 3 5 8
6	0 1 1 5 5 8 8
7	0 2
8	1 2 5

ii. For this data set, the stem is the whole number including the decimal point while the leaf is the trailing decimal digit. The groups entries are:

1 :	1.6	1.9	1.6	1.8	1.8	
2 :	2.7	2.7	2.4	2.5		
3 :	3.5	3.8	3.3	3.4	3.1	3.3
4 :	4.9	4.5	4.8	4.1	4.5	
5 :	5.2	5.7	5.7	5.6		
6 :	6.2					
7 :	7.5	7.8	7.1			
8 :	8.2	8.3				

The corresponding stem-and-leaf is:

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

Unfortunately, not all sets of data can be organized into a stem-and-leaf plot. First, there should not be too much spread in the data e.g. from 1 – 10000. Similarly, if there is very little spread. Further, the numbers in the data should not be extremely large. For example, if the values were in hundreds of thousands, such as 345,005 and 582,281, then just separating the last digit would be meaningless.

2.2.7 Line graph

Line graphs are diagrammatical representation of the relation between two variables x and y. The co-ordinate points of these variables are joined together to have the line graph.

Example 2.10
Draw a line graph to represent the information below:

Before	14	20	21	24	22	25	26
After	16	24	23	25	30	27	34

Solution

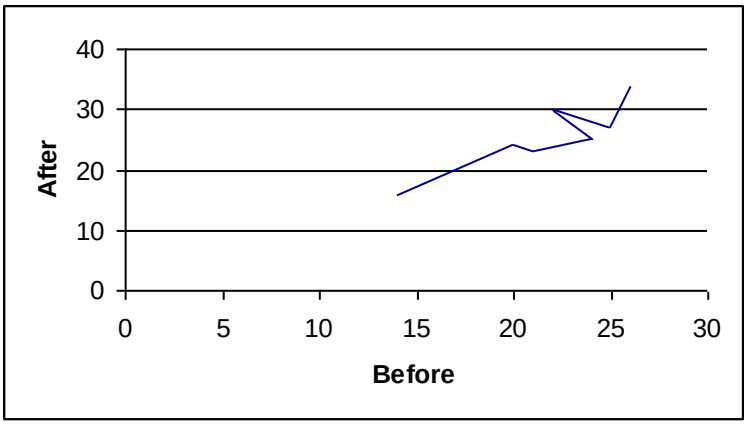


Figure 2.13: Line graph

2.3 EXERCISES

2.3.3 The following 45 amounts are the fees that Fast Delivery charged for delivering small freight items last Monday morning:

2.57	4.21	1.05	3.06	4.50	5.05	3.45	2.15	0.92
3.12	2.67	0.76	4.13	5.93	4.15	2.03	0.57	1.85
4.10	3.41	1.86	2.53	1.46	3.85	5.12	3.24	1.89
2.51	0.95	1.24	2.21	5.86	3.57	2.18	4.29	3.50
0.91	0.82	1.47	4.25	3.81	2.48	1.27	5.35	3.33

i. Classify these data into a grouped frequency distribution by using classes of 0.01 – 1.00, 1.01 – 2.00, . . ., 5.01 – 6.00

ii. Find the class width

iii. For the class 4.01 – 5.00, name the value of:
(a) the class center
(b) the class limit,
(c) the class boundaries

iv. Construct a relative frequency histogram of these data.

2.3.4 The incomes of 80 employees of a company are recorded as follows in ₦'000 per annum.

430	650	730	450	357	370	680	880	720	500
555	600	710	375	481	639	700	850	650	400
885	730	650	480	537	390	495	755	800	450
633	741	839	395	485	631	737	810	561	492
453	439	810	750	653	495	849	675	800	795
385	411	865	721	846	666	713	874	815	873
555	414	312	481	672	411	813	817	361	845
315	481	618	535	621	781	432	537	615	811

Use an appropriate class interval to construct the frequency distribution. Draw a histogram to represent the data and frequency polygon on your histogram. Estimate the mode from your histogram.

2.3.5 The following table shows the frequency distribution of marks of 200 students in a mathematics examination.

Mark	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
Frequency	4	26	52	16	36	34	20	12

i) Draw a cumulative frequency curve and estimate the Quartiles.

ii) Calculate the interquartile and semi-interquartile range from your graph.

iii) Find the pass mark if only 20% of the students should pass.

iv) How many of the students scored between 60% and 85%.

2.3.6 Use the table in Exercise 2.3.5 to answer the following questions

i. Draw a bar chart of the frequency distribution
ii. Draw a pie chart of the frequency distribution
iii. Draw a histogram of the frequency distribution

CHAPTER THREE

DESCRIPTIVE ANALYSIS

Descriptive Analysis is of two parts namely:

i. Measures of location or central tendency
ii. Measures of dispersion, variation or spread

Measures of central tendency are numerical values that tend to locate in some sense the middle of a set of data. The term *average* is often associated with these measures. Each of the several measures of central tendency can be called the average value. They are the mean, median, and mode.

Once the middle of a set of data has been determined, our search for information immediately turns to the *measures of dispersion* (spread). The measures of dispersion include the range, variance, and standard deviation. These numerical values describe the amount of spread, or variability, that is found among the data.

3.1 THE SIGMA (Σ) NOTATION

The Greek capital letter sigma (Σ) is used in Mathematics to indicate the summation of a set of addends. Each of these addends must be of the form of the variable following Σ. For example,

i. Σx means sum the variable x
ii. Σ (x – 3) means sum the set of addends that are 3 less than the values of each x

When large quantities of data are collected, it is usually convenient to index the response so that at a future time its source will be known. This indexing is shown on the notation by using i (or j or

k) and affixing the index of the first and last addend at the bottom and top of the Σ. For example,

Means to all consecutive values of square of x's starting with the source: number 1 and proceeding to source number 4

Example 3.1

Find (i) Σx (ii) Σx^2 (iii) $(\Sigma x)^2$

x	1	2	4	6	5	7	3
x^2	1	4	16	36	25	49	9

Solution

$$\Sigma x = 1 + 2 + 4 + 6 + 5 + 7 + 3 = 28$$

$$\Sigma x^2 = 1 + 4 + 16 + 36 + 25 + 49 + 9 = 140$$

$$(\Sigma x)^2 = (28)^2 = 784$$

Example 3.2

Simplify

$\sum_{i=1}^{3} (3x_i + 1)$ and find its value when $x_1 = x_2 = x_3 = 1$

Solution

$$\sum_{i=1}^{3} (3x_i + 1) = (3x_i + 1) + (3x_2 + 1) + (3x_3 + 1)$$

$$= (3x_1 + 3x_2 + 3x_3) + (1 + 1 + 1)$$

$$= 3 \sum_{i=1}^{3} x_i + 3$$

$$= 3 (1 + 1 + 1) + 3 = 9 + 3 = 12$$

3.2 MEAN

To find the *mean*, $\bar{x}$ (read "x bar"), you will add all the values of the variable x and divide by the number of these values, n. We express this in formula form as

$$\text{Sample mean} \quad = \quad \bar{x} \quad = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (3.1)$$

Example 3.3: Find the mean of the following numbers 2, 3, 4, 2, 3, 2, 4, 8

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2 + 3 + 4 + 2 + 3 + 2 + 4 + 8}{8}$$

$$= \frac{28}{8} = 3.5$$

When the sample data has the form of a frequency distribution, we will need to make a slight adaptation in order to find the mean. Consider the frequency distribution of Table 3.1.

Table 3.1: ungrouped frequency distribution

x	1	2	3	4	5
f	4	8	5	4	7

To calculate the mean $\bar{x}$ using the above formula; we have

$\Sigma x = 1 + 1 + 1 + 1 + 2 + 2 + \ldots + 2 + 3 + \ldots + 3 + 4 + \ldots + 4 + 7 + \ldots + 7$

$\Sigma x = 4\,(1) + 8\,(2) + 5\,(3) + 4\,(4) + 7\,(5)$

$= 86$

$= \Sigma fx$

Therefore, the mean of a frequency distribution may be found by dividing the sum of the data, Σfx, by the sample size, Σf. We can rewrite formula (3.1) for use with a frequency distribution as:

$$\bar{x} = \frac{\Sigma x f}{\Sigma f} \qquad (3.2)$$

Table 3.2

x	f	xf
1	4	4
2	8	16
3	5	15
4	4	16
5	7	35
Total	28	86

$$\bar{x} = \frac{\Sigma x f}{\Sigma f} = \frac{86}{28} = 3.07$$

3.2.1 Mean of Grouped Data

The class centers (mark) are now being used as representative values for the observed data.

Example 3.4: What is the mean of this distribution?

Table 3.3

Mark	**Frequency (f)**	**Class center (x)**	**fx**
30 – 39	5	34.5	172.5
40 – 49	10	44.5	445
50 – 59	15	54.5	817.5
60 – 69	10	64.5	645
70 – 79	5	74.5	372.5
Total	**Σf = 45**	**Σfx =**	**2452.5**

$$\text{Mean} = \bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{2452.5}{45} = 54.5$$

3.2.2 Using Assumed Mean

The method of using an assumed mean makes strenuous calculations of large numbers to be easier. For the ungrouped data, we use

$$\bar{x} = A + \frac{\Sigma d_i}{N} = A + \frac{\Sigma d}{N} \qquad (3.3)$$

and for the grouped data, we use

$$\bar{x} = A + \frac{\Sigma f_i d_i}{\Sigma f_i} = A + \frac{\Sigma fd}{\Sigma f} \qquad (3.4)$$

where A is the assumed mean, $d_i = x_i - A$ are the deviation of x_i from A.

Example 3.5: Using the data in Example 3.4, let A = 44.5

Table 3.4

Mark	Frequency (f)	Class centre (x)	d = x – A	fd
30 – 39	5	34.5	-10	-50
40 – 49	10	44.5	0	0
50 – 59	15	54.5	10	150
60 – 69	10	64.5	20	200
70 – 79	5	74.5	30	150
Total	Σf = 45			Σfd = 450

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f} = 44.5 + \frac{450}{45}$$

$$= 44.5 + 10 = 54.5$$

which is the same as in previous example

3.2.3 Harmonic Mean

This is the reciprocal of the average of reciprocals. It is usually represented by $\bar{x}_H$ and defined by

$$\bar{x}_H = \left[\frac{1}{N}\sum_{j=1}^{n}\frac{1}{X_j}\right]^{-1} = \frac{1}{\frac{1}{N}\sum_{j=1}^{n}\frac{1}{X_j}} = \frac{N}{\sum_{j=1}^{n}\frac{1}{X_j}} \qquad (3.5)$$

Example 3.6: Find the Harmonic mean for the following data 2, 5, 3, 6, 7.

Solution

$$\bar{x}_H = \frac{5}{1/2 + 1/5 + 1/3 + 1/6 + 1/7} = \frac{5}{0.5 + 0.2 + 0.33 + 0.167 + 0.143}$$

$$= \frac{5}{1.34} = 3.73$$

3.2.4 Geometric Mean

This is the nth root of the product of the n numbers in a data set. This is usually represented by $\bar{x}_G$ and defined by

$$\bar{x}_G = \sqrt[n]{X_1 x X_2 x \ldots x X_n} = (X_1 x X_2 x \ldots x X_n)^{1/n} \qquad (3.6)$$

Example 3.7: Find the Geometric mean for the data above in Example 3.6

$$\bar{x}_G = \sqrt[5]{2x5x3x6x7} = \sqrt[5]{1260} = 4.17$$

3.2.5 Arithmetic Mean

This has been dealt with earlier. It is represented by $\bar{x}_A$ and defined as stated in (3.1)

The arithmetic mean of the sample above is

$$\bar{x} = \frac{\Sigma x}{n} = \frac{23}{5} = 4.6$$

Note: that the expression

$$\bar{x}_H \leq \bar{x}_G \leq \bar{x}_A \qquad (3.7)$$

is true for any data

3.3 MEDIAN

The *median* is the value of the data that occupies the middle position when the data are ranked in order according to size.

The *depth* (number of positions from either end), or position, of the median is determined by the formula.

$$\text{Depth of median} = \frac{n + 1}{2} \qquad (3.8)$$

If the number of measurement n is an odd number, the *median* is the middle value. If the number of measurement n is an even number, the *median* is the average of the middle two values. For example, lets find the median of these numbers 2, 4, 6, 8, 9.

In our example, n = 5, and therefore the depth of the median is

$$\text{depth} = \frac{5 + 1}{2} = 3$$

That is, the median is the third number from either end in the ranked data, i,e, median is 6

Lets look at these data 4, 6, 7, 8, 10, 12. Here n = 6, and therefore the median depth is

$$\text{depth} = \frac{6 + 1}{2} = 3.5$$

This is to say that the median is halfway between the third and fourth pieces of data. To find the number halfway between any two

values, add the two values together and divide by 2. In this case, add 7 and 8, then divide by 2. The median is 7.5

For grouped data, the median is obtained by interpolation and given by

$$\text{Median} = L_1 + \left[\frac{\frac{N}{2} - F_b}{F_m}\right]C \qquad (3.9)$$

Where L_1 is lower class boundary of the median class,

C - Size (width) of the median class interval,
N - Total frequency,
F_b - Sum of frequencies of all classes below the median class.
F_m - Frequency of median class.

Example 3.8: Find the median mark in the table below:

Marks	30-39	40-49	50-59	60-69	70-79
Frequency	5	10	15	10	5

Solution

Table 3.5

Mark	Class Boundaries	Frequency	Cumulative Frequency
30 – 39	29.5 – 39.5	5	5
40 – 49	39.5 – 49.5	10	15
50 – 59	49.5 – 59.5	15	30
60 – 69	59.5 – 69.5	10	40
70 – 79	69.5 – 79.5	5	45
Total		$\Sigma f = 45$	

Median class - 50 – 59

$L_1 = 49.5$
$N = 45$
$C = 59.5 - 49.5 = 10$
$F_b = 15$
$F_m = 15$

$$\text{Median} = L_1 + \left[\frac{\frac{N}{2} - F_b}{F_m}\right]C$$

$$= 49.5 + \left[\frac{\frac{45}{2} - 15}{15}\right]x10$$

$$= 49.5 + \frac{(22.5 - 15)}{15} \times 10$$

$$= 49.5 + 0.5 \times 10 = 49.5 + 5 = 54.5$$

Note: The mean for the data is the same as the median due to symmetry of data.

3.4 MODE

The *mode* for a set of data is the value that occurs most frequently.

Example 3.9: Find the modes of the following data. 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4

Solution

The values with the highest number of occurrence are 2 and 4. They both have equal frequency of 4. That is, we have a bimodal case.

For group data, the mode is obtained by

$$\text{Mode} = L_1 + \frac{f_1 + f_0}{2f_1 + f_0 + f_2}(L_2 - L_1) \qquad (3.10)$$

where

f_0 = the frequency of the group before the group that appears most often,

f_1 = the frequency of the group that appears most often,

f_2 = the frequency of the group after the group that appears most often,

L_1 = the lower limit of the group with f_1 and

L_2 = the upper limit of the group with f_1

OR

$$\text{Mode} = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)C \qquad (3.11)$$

Where

L_1 = lower class boundary of modal class,

Δ_1 = excess of modal frequency over next lower class

Δ_2 = excess of modal frequency over next higher class and

C = size of modal class interval

Example 3.10. Find the mode of the data given in example 3.8. Using the two methods given above.

Solution

Method I - (3.10)

$$\text{Mode} = 49.5 + \frac{15 + 10}{2(15) + 10 + 10}(59.5 - 49.5)$$

$$= 49.5 + \frac{25}{50} \times 10$$

$$= 49.5 + 5 = 54.5$$

Method II - (3.11)

$L_1 = 49.5$, $\Delta_1 = 15 - 10 = 5$

$\Delta_2 = 15 - 10 = 5$ C = 59.5 - 49.5 = 10

$$\text{Mode} = 49.5 + \left(\frac{5}{5+5}\right)x10$$

$$= 49.5 + \frac{5}{10} \text{ x } 10$$

$$= 49.5 + 5 = 54.5$$

Note: The mean = Mode = Median of the data considered above due to symmetry of data.

Also Mean – Mode = 3 (Mean - Median) (3.12)

3.5 DECILES, PERCENTILES, QUARTILES

when observations are ordered from small to large, the resulting ordered data are called the *order statistics* of the sample. Lets have the following data

24	31	31	40	45	47
48	48	48	49	50	50
50	50	50	50	51	53
53	56	60	70	71	76

We give ranks to these ordered statistics and use the rank as the subscript on x. The first order statistic x_1= 24 has rank 1; the second order statistic x_2 = 31 has rank 2, the third order statistic x_3 = 31 has rank 3, …; and the 24^{th} order statistic x_{24} = 76 has rank 24. It is clear here that $x_1 \le x_2 \le \ldots. \le x_{24}$.

From these order statistics, it is rather easy to find the *sample percentiles*. If 0 < p < 1, the (100p)th sample percentile has approximately np sample observations less than it and also n(1-p) sample observation greater than it. One way of achieving this is to take the (100p)th sample percentile as the (n+1)pth order statistic, provided that (n+1)p is an integer. If (n+1)p is not an integer but is equal to r plus some proper fraction, say a/b, use a weighted average of the rth and the (r+1)st order statistics. That is, define the (100p)th sample percentile as

$$\Pi_p = x_r + (a/b)(x_{r+1} - x_r) = (1 - a/b)x_r + (a/b)\, x_{r+1} \qquad (3.13)$$

Note: that this is simply a linear interpolation between x_r and x_{r+1}
For illustration, consider the 24 ordered examination scores. With p = ½, we find the 50th percentile by averaging the 12th and 13th order statistics, since (n+1)p = 25/2 = 12.5

$$\Pi_{0.50} = (½)\, x_{12} + (½)\, x_{13} = (50 + 50)/2 = 50$$

With p = ¼, we have (n+1)p = 25/4 =6.25; and thus the 25th sample percentile is

$$\Pi_{0.25} = (1 - 0.25)\, x_6 + 0.25\, x_7$$
$$= (0.75)\ (47) + (0.25)\ (48) = 35.25 + 12$$
$$= 47.25$$

With p = ¾, so that (n+1) p = (25) (3/4) = 18.75, the 75th sample percentile is

$$\Pi_{0.75} = (1- 0.75)\, x_{18} + (0.75)\, x_{19}$$
$$= (0.25)\ (53) + (0.75)\ (53)$$
$$= (13.25 + 39.75 = 53$$

Note: that approximately 50%, 25% and 75% of the sample observation are less than 50, 47.25, 53, respectively.

As already discussed in chapter two, 50th percentile is the median of the sample. The 25th, 50th, and 75th percentiles are the first, second, and third quartiles of the sample, denoted as Q_1, Q_2, and Q_3, respectively. The 10th, 20th, 30th, ……, 90th percentiles are the *deciles* of the sample. So note that the 50th percentile is also the median, the second quartile, and the fifth deciles.

For example, the 2th and 9th deciles would be calculated as thus:
(n+1)p = (25)(2/10)= 5 for the second deciles and
(n+1)p = (25)(9/10) = 22.5 for the ninth deciles.

$$\Pi_{0.20} = (1- 0)\, x_5 + 0x_6 = x_5 = 45$$

and

$$\Pi_{0.90} = (1- 0.5)\, x_{22} + 0.5x_{23} = (0.5)\ 60 + (0.5)\ (70)$$
$$= 30 + 35 = 65$$

3.6 BOX-AND-WHISKER DIAGRAM

This is a graphical means for displaying the five-number summary of a set of data (smallest, first quartile, median or second quartile, third quartile and the largest) that is called a *box-and-whisker* diagram or more simply as a *box plot*. The three values used – Q_1, Q_2 and Q_3 – are sometimes called *hinges*.

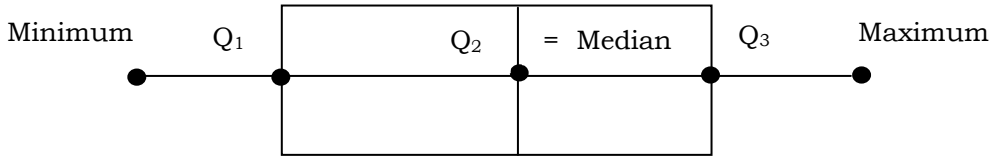
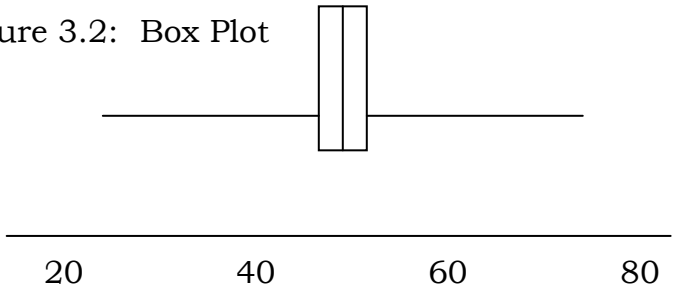


Figure 3.1 Box Plot

To construct a horizontal box-and-whisker diagram, draw a horizontal axis that is scaled to the data. Above the axis draw a rectangle box with the left and right sides drawn at Q_1 and Q_3 with a vertical line segment drawn at the median, Q_2 = median. A left whisker is drawn as a horizontal line segment from the minimum to the midpoint of the left side of the box, and a right whisker is drawn as a horizontal line segment from the midpoint of the right side of the box to the maximum. Note that the length of the box is equal to the interquartile range ($Q_3 - Q_1$). The left and right whiskers contain the first and fourth quarters of the data.

Example 3.11: Draw the Box Plot of the data in section 3.5
Solution
The five number summary are minimum = 24
$Q_1 = 47.25$, $Q_2 = 50$, $Q_3 = 53$, and the maximum = 76

Figure 3.2: Box Plot



Example 3.12: Let x denote the concentration of acid on milligrams per liter. Twenty observations of x are:

115	116	117	118	118	118	119	121	122	125
126	128	129	129	130	131	131	133	133	134

(a) Find the mid range, interquartile range and median
(b) Draw a box-and-whisker diagram.

Solution

a) Midrange = average of the extremes

$$= \frac{x_1 + x_n}{2} = \frac{115 + 134}{2} = \frac{249}{2}$$

$$= 124.5$$

With p = ¼, we have (n + 1)p = 21/4 = 5.25 and the 25th sample percentile is

$Q_1 = \Pi_{0.25} = (1 - 0.25)\, x_5 + (0.25)\, x_6$

$= (0.75)\,(118) + (0.25)\,118 = 118$

with p = ½, we have (n + 1)p = 21/2 = 10.5 and the 50th sample percentile is

$Q_2 = \Pi_{0.50} = (1 - 0.5)\, x_{10} + 0.5x_{11}$

$= (0.5)\,(125) + (0.5)\,(126) = 62.5 + 63$

$= 125.5$

With p = ¾, we have (n + 1)p = 21 x ¾ = 15.75 and the 75th sample percentile is

$Q_3 = \Pi_{0.75} = (1 - 0.75)\, x_{15} + 0.75x_{16}$

$= 0.25x_{15} + 0.75x_{16} = (0.25)\,(130) + (0.75)\,(131)$

$= 32.5 + 98.25 = 130.75$

Interquartile range $= Q_3 - Q_1 = 130.75 - 118$

$= 12.75$

Median $= Q_2 = 125.5$

b)

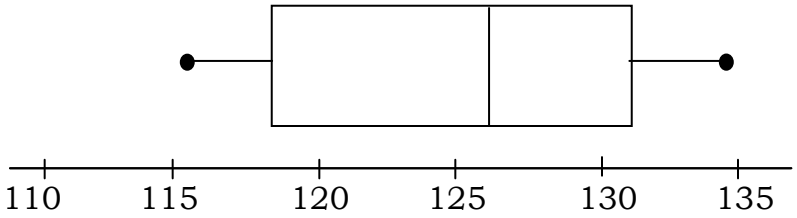


Figure 3.3 box plot

Tukey suggested a method for defining outliers that is resistant to the effect of one or two extremes values and makes use of the interquartile range. In a box-and-whisker diagram, construct *inner fences* to the left and right of the box at a distance of 1.5 times the interquartile range. *Outer fences* are constructed in the same way at a distance of 3 times the interquartile range. Observations that lie between the inner and outer fences are called *suspected outliers*. Observations that lie beyond the outer fences are called *outliers*.

3.7 MEAN ABSOLUTE DEVIATION (MAD)

This is the average amount by which values in a distribution differ from the mean.

Mean Absolute Deviation for ungrouped data

$$MAD = \frac{\sum_{i=1}^{n} | x_i - \bar{x} |}{n} \qquad (3.14)$$

Mean Absolute Deviation of ungrouped data with frequency and of Group Data

$$MAD = \frac{\sum_{i=1}^{n} f| x_i - \bar{x} |}{\sum f} \qquad (3.15)$$

Example 3.13: Find the mean deviation for the following data:
3 4 5 8 15

First, we find the mean of the data

$$\text{Mean} = \bar{x} = \frac{\sum x}{n} = \frac{3 + 4 + 5 + 8 + 15}{5}$$
$$= \frac{35}{5}$$
$$= 7$$

x	3	4	5	8	15
$\lvert x - \bar{x} \rvert$	4	3	2	1	8

$$\text{MAD} = \frac{\sum \lvert x - \bar{x} \rvert}{n} = \frac{4 + 3 + 2 + 1 + 8}{5}$$
$$= \frac{18}{5}$$
$$= 3.6$$

This implies that the average distance that this piece of data is from the mean is 3.6.

Example 3.14: Find the mean absolute deviation of the following data.

Mark	2	4	5	7	8	9
Frequency	4	6	8	1	4	2

Solution

Table 3.6

Mark (x)	Frequency (f)	fx	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$	$f\lvert x - \bar{x} \rvert$
2	4	8	-3.16	3.16	12.64
4	6	24	-1.16	1.16	6.96
5	8	40	-0.16	0.16	1.28
7	1	7	1.84	1.84	1.84
8	4	32	2.84	2.84	11.36
9	2	18	3.84	3.84	7.68
Total	$\Sigma f = 25$	$\Sigma fx = 129$			41.76

$$\text{Mean} = \frac{\sum fx}{\sum f} = \frac{129}{25}$$
$$= 5.16$$

$$\text{Mean} = \frac{\sum f|x - \bar{x}|}{\sum f} = \frac{41.76}{25}$$
$$= 1.6704$$

Example 3.15: The following distribution of commuting distances was obtained for a sample of employees.

Table 3.7

Distance (Kilometer)	Frequency
1.0 – 2.9	2
3.0 – 4.9	6
5.0 – 6.9	12
7.0 – 8.9	50
9.0 –10.9	35
11.0 –12.9	15
13.0 –14.9	5

Find the mean deviation for the commuting distances.

Solution

Table 3.8

Distance (kg)	**f**	**Class center (x)**	**fx**	**$x - \bar{x}$**	**$\lvert x - \bar{x} \rvert$**	**$f\lvert x - \bar{x} \rvert$**
1.0 – 2.9	2	1.95	3.9	-6.8	6.8	13.6
3.0 – 4.9	6	3.95	23.7	-4.8	4.8	28.8
5.0 – 6.9	12	5.95	71.4	-2.8	2.8	33.6
7.0 – 8.9	50	7.95	397.5	-0.8	0.8	40
9.0 – 10.9	35	9.95	348.25	1.2	1.2	42
11.0 –12.9	15	11.95	179.25	3.2	3.2	48
13.0 –14.9	5	13.95	69.75	5.2	5.2	26
Total	**$\sum f = 125$**		**$\sum fx = 1093.75$**			**$f\lvert x - \bar{x} \rvert = 232$**

$$\text{Mean} = \frac{\sum fx}{\sum f} = \frac{1093.75}{125}$$
$$= 8.75$$

$$\text{Mean} = \frac{\sum f|x - \bar{x}|}{\sum f} = \frac{232}{125}$$
$$= 1.856$$

3.8 VARIANCE AND STANDARD DEVIATION

Variance is a useful measure of the spread of the original values about the mean. When we are concerned with a population, the variance is written in terms of the Greek letter σ (lower case sigma) and is denoted by σ^2. Thus, we can summarize the above calculations with the following formula:

$$\text{Population variate } \sigma^2 = \frac{\sum (x - \mu)^2}{N} \quad \text{OR} \quad \frac{N(\sum x^2) - (\sum x)^2}{N^2} \tag{3.16}$$

where N is the size of the population.

However, a far more useful measure of the spread or variability in a set of data is the *standard deviation*, which is defined as the square root of the variance.

$$\text{Standard Deviation (SD)} = \sqrt{Variance} \tag{3.17}$$

Since the standard deviation is the square root of the variance σ^2, the standard deviation is denoted by σ and is found from the formula.

$$\text{Population standard deviation } \sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \tag{3.18}$$

$$\text{OR} \quad \sqrt{\frac{N(\sum x^2) - (\sum x)^2}{N^2}}$$

One special advantage of working with the standard deviation is that it is measured in the same units as the original data. Thus, if the original set of numbers represent weights of a certain type of item, then both the mean and standard deviation are measured in weights.

The larger that σ is for a set of numbers, the greater the spread or variability among those numbers. The smaller the value of σ, the smaller the amount of variation in the data.

All the above ideas for the variance and standard deviation were developed in the context of a population. Very similar ideas exist for the variance and standard deviation of a sample drawn from a population, with one significant difference. When we deal with a sample, we cannot average the sum of the squared deviations, $(x - \bar{x})^2$, over the entire set of data. Instead, it is necessary to make the following modification:

$$\text{Sample variance } (s^2) = \frac{\sum(x-\mu)^2}{n-1} \quad \text{OR} \quad \frac{n(\sum x^2) - (\sum x)^2}{n(n-1)} \qquad (3.19)$$

and

$$\text{Sample standard deviation } (s) = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \qquad (3.20)$$

$$\text{or } \sqrt{\frac{n(\sum x^2) - (\sum x)^2}{n(n-1)}}$$

That is, instead of dividing by n data points, we divide by n-1. Just as σ^2 and σ represent the variance and standard deviation of a population, respectively, we use the symbols s^2 and s to stand for the variance and standard deviation, respectively, of a sample.

Variance and standard deviation with frequency counts and of Group data are

$$\sigma^2 = \frac{\sum f(x - \bar{x})^2}{\sum f} \quad \text{or} \quad \frac{\sum fx^2 - \frac{(\sum fx)^2}{\sum f}}{\sum f} \tag{3.21}$$

and

$$s^2 = \frac{\sum f(x - \bar{x})^2}{\sum f - 1} \quad \text{or} \quad \frac{\sum fx^2 - \frac{(\sum fx)^2}{\sum f}}{\sum f - 1} \tag{3.22}$$

Standard deviations σ- and s are the square roots of (3.21) and (3.22), respectively.

Variation and standard deviation using Assumed Mean are given below where d = x – A (assumed mean)

$$\sigma^{-2} = \frac{\sum fd^2 - \frac{(\sum fd)^2}{\sum f}}{\sum f} \tag{3.23}$$

$$s^2 = \frac{\sum fd^2 - \frac{(\sum fd)^2}{\sum f}}{\sum f - 1} \tag{3.24}$$

Standard deviation σ- and s are the square roots of (3.23) and (3.24), respectively.

Variance and Standard Deviation using Assumed Mean and Scaling Factor.

The foregoing calculations can be made simpler by further scaling down of d to h = d/c, where c is the regular increment in the x values. The formulas are given below:

$$\bar{x} = A + \frac{\sum fh}{\sum f} . c \tag{3.25}$$

$$\sigma^{-2} = \left[\frac{\sum fh^2 - \frac{(\sum fh)^2}{\sum f}}{\sum f} \right] x\, c^2 \qquad (3.26)$$

and

$$s^2 = \left[\frac{\sum fh^2 - \frac{(\sum fh)^2}{\sum f}}{\sum f - 1} \right] x\, c^2 \qquad (3.27)$$

The standard deviations σ^- and S are the square roots of (3.26) and (3.27), respectively.

Example 3.16: Find the mean and standard deviation σ^- of the data: 4 6 8 9 10 12

Solution

Using $\frac{\sum (x - \bar{x})^2}{N}$

Table 3.9

x	$x - \bar{x}$	$(x - \bar{x})^2$
4	-4.33	18.75
7	-1.33	1.77
8	-0.33	0.11
9	0.67	0.45
10	1.67	2.79
12	3.67	13.47
$\sum x = 50$		$\sum (x - \bar{x})^2 = 37.34$

Mean $= \frac{\sum x}{N} = \frac{50}{6} = 8.33$

$$\text{Variance} = \frac{\sum(x - x)^2}{N} = \frac{37.34}{6}$$

$$= 6.22$$

$$\text{S.D} = \sqrt{Variance} = 2.494$$

$$\text{Using} = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}$$

Table 3.10

x	x^2
4	16
7	49
8	64
9	81
10	100
12	144
$\sum x = 50$	$\sum x^2 = 454$

$$\text{Mean} = \frac{\sum x}{N} = \frac{50}{6} = 8.33$$

$$\text{Variance} = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N} = \frac{454 - \frac{50^2}{6}}{6}$$

$$= \frac{454 - 416.7}{6} = \frac{37.33}{6} = 6.22$$

$$\text{S.D.} = \sqrt{Variance} = \sqrt{6.22} = 2.494$$

Example 3.17: Find the mean and standard deviation (σ) for the following grouped frequency distribution.

Table 3.11

Class limits	f
2 – 5	7
6 – 9	15

10 – 13	22
14 – 17	14
18 – 21	2

Solution

Method I

Using $\frac{\sum fx^2 - (\sum fx)^2/\sum f}{\sum f}$

Table 3.12

Class limits	Class Center (x)	f	fx	fx^2
2 – 5	3.5	7	24.5	85.75
6 – 9	7.5	15	112.5	843.75
10 –13	11.5	22	253	2909.5
14 –17	15.5	14	217	3363.5
18 –21	19.5	2	39	760.5
Total		$\sum f = 60$	$\sum fx = 646$	$\sum fx^2 = 7963$

Mean $= \frac{\sum fx}{\sum f} = \frac{646}{60} = 10.77$

Variance $= \frac{\sum fx^2 - \frac{(\sum fx)^2}{\sum f}}{\sum f} = \frac{7963 - \frac{646^2}{60}}{60}$

$= \frac{7963 - 6955.27}{60} = \frac{1007.73}{60} = 16.8$

S.D. $= \sqrt{16.8} = 4.099$

Method II

Using Assumed mean Let A = 11.5

Table 3.13

Class limits	Class Centre (x)	Frequency (f)	d = x – A	fd	fd^2
2 – 5	3.5	7	-8	-56	448
6 – 9	7.5	15	-4	-60	240
10 –13	11.5	22	0	0	0
14 –17	15.5	14	4	56	224
18 –21	19.5	2	8	16	128
Total		$\sum f = 60$		$\sum fd = -44$	$\sum fd^2 = 1040$

Mean $= \bar{x} = A + \frac{(\sum fd)}{\sum f} = 11.5 + \frac{-44}{60}$

$= 11.5 - 0.73 = 10.77$

Variance $= \sigma^{-2} = \frac{\sum fd^2 - \frac{(\sum fd)^2}{\sum f}}{\sum f}$

$= \frac{1040 - \frac{(-44)^2}{60}}{60} = \frac{1040 - \frac{1936}{60}}{60}$

$= \frac{1040 - 32.27}{60} = \frac{1007.73}{60}$

$= 16.8$

S.D $= \sqrt{16.8} = 4.099$

Method III

Table 3.14

Class limits	Class Centre (x)	Frequency (f)	d = x – A	h	fh	fh^2
2 – 5	3.5	7	-8	-2	-14	28
6 – 9	7.5	15	-4	-1	-15	15
10 –13	11.5	22	0	0	0	0
14 –17	15.5	14	4	1	14	14
18 –21	19.5	2	8	2	4	8
					-11	65

Let A = 11.5, h = d/c where c is the regular increment in x values from one class to another e.g. 7.5 – 3.5 = 4, 11.5 – 7.5 = 4, and so on.

$$\text{Mean} = A + \frac{\sum fh}{\sum f} \cdot c = 11.5 + \frac{-11}{60} \times 4$$

$$= 11.5 - \frac{44}{60}$$

$$= 11.5 - 0.73 = 10.77$$

$$\text{Variance} = \sigma^{-2} = \frac{[\sum fh^2 - \frac{(\sum fh)^2}{\sum f}]\, c^2}{\sum f}$$

$$= \frac{[65 - \frac{(-11)^2}{60}] \times 4^2}{60} = \frac{[65 - 2.02] \times 4^2}{60}$$

$$= 1.05 \times 4^2 = 16.8$$

$$\text{S.D} = \sqrt{16.8} = 4.099$$

3.9 COEFFICIENT OF VARIATION

In order to compare the relative amounts of variation in populations having different means, the *coefficient of variation*,

symbolized by CV, has been developed. This is simply the standard deviation expressed as a percentage of the mean. Its formula is

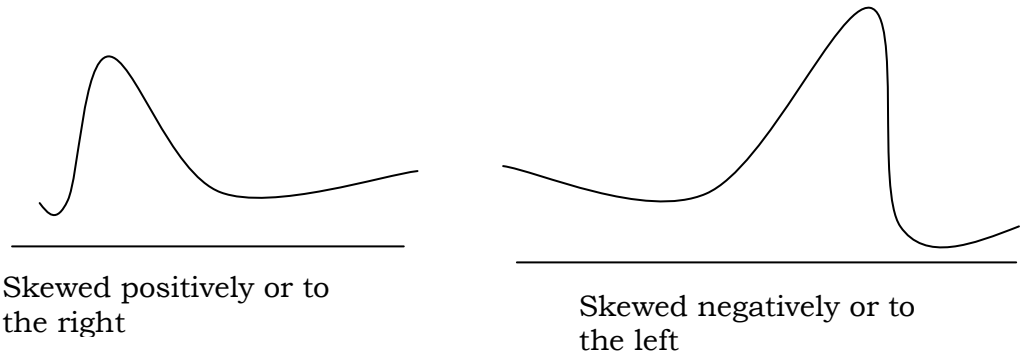
$$C.V = \frac{S.D}{Mean} \times 100\% \qquad (3.28)$$

The one with smallest C.V among the variables is preferred to be better than others.

3.10 SKEWNESS AND KURTOSIS

When observed frequency distribution depart from symmetry, it is useful to have a statistic that measures the nature and amount of departure. One is *skewness* while the other is *Kurtosis*.

When a distribution is symmetrical about the mean, the skewness is equal to zero. If the probability histogram has a longer "tail" to the right than to the left, the measure of skewness is positive, and we say that the distribution is skewed positively or to the right. If the probability histogram has a longer tail to the left than to the right, the measure of skewness is negative, and we say that the distribution is skewed negatively or to the left.



Skewed positively or to the right

Skewed negatively or to the left

Figure 3.4 Skewness

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{\bar{x} - \text{mode}}{s} \qquad (3.29)$$

also known as the Pearson's coefficient of relative skewness and can also be defined as:

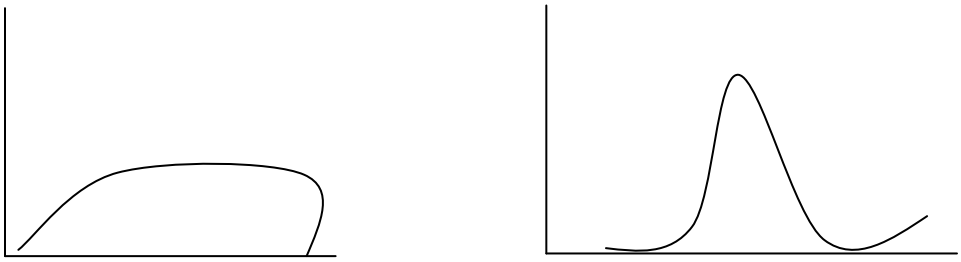
$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}} = \frac{3(\bar{x} - \text{median})}{s} \qquad (3.30)$$

Using mean – mode = 3(mean – median)

Kurtosis is the degree of peakness of a distribution relative to a normal (symmetry) distribution. A *leptokurtic* curve has more items near the mean and at the tails, with fewer items in the intermediate regions relative to a normal distribution with the same mean and variance. A *platykurtic* curve has fewer items at the mean and at the tails than the normal curve but has more items in intermediate regions. A bimodal distribution is an extreme platykurtic distribution. The measure of kurtosis is based on both quartiles and percentiles and is given by

$$K = \frac{\text{Interquartile Range } (Q_3 - Q_1)}{\Pi_{0.9} - \Pi_{0.1}} \qquad (3.31)$$

This is also known as percentile coefficients of kurtosis



Platykurtic Leptokurtic

Figure 3.5 Kurtosis

Example 3.18: Use example 3.17 to find the coefficient of variation and skewness.

Solution

$$C.V = \frac{S.D}{Mean} \times 100\% = \frac{4.099}{10.77} \times 100\%$$

$$= 38.06\%$$

$$Skewness = \frac{Mean - Mode}{S.D} = \frac{10.77 - 11.5}{4.099}$$

$$= \frac{-0.73}{4.99} = -0.18$$

3.11 EXERCISES

3.11.3 Find (a) $\sum x^2$, (b) $(\sum x)^2$, (c) $\sum x \sum y$, (d) $\sum y^2$, (e) $(\sum y)^2$ for the data shown below:

x	3	4	5	6	7
Y	7	8	9	10	11

3.11.5 The weights, in pounds, of a group of people signing up at a hotel are:

125	141	141	132	155	160	185	165	172	148
131	154	162	148	135	181	172	133	141	135

Find (i) the mean, median and mode of the weights.
(ii) the quartiles, skewness and kurtosis

3.11.8 Estimate the mean, standard deviation and median for the following set of data:

Class boundaries	Frequency
151 – 160	50
161 – 170	60
171 – 180	30
181 – 190	35
191 – 200	25
201 – 210	17
211 – 220	13

CHAPTER FOUR

INTRODUCTION TO PROBABILITY

The application of probability is evident in most areas of human endeavour. For example, the chance of an accident occurring on a road, probability of getting a head when a coin is tossed, chance of a top politician winning an election, e.t.c. are examples of probability. Therefore, we must be able to assess the degree of uncertainty, in any given situation, and this is done mathematically by using *probability*

4.1 PROBABILITY OF EVENTS

We begin by defining some terminology that we are using in this chapter and in subsequent ones.

Experiment: Any process that yields a result or an observation.

Outcome: A particular result of an experiment

Sample space: The set of all possible outcomes of an experiment.

Sample point: The individual outcomes in a sample space.

Event: Any subset of the sample space. If A is an event, then n(A) is the number of sample points that belong to event A.

Probability of an event is a measure of the likelihood of that event occurring. If an experiment has a finite number of outcomes which are equally likely, then the probability that an event A will occur is given by

$$P(A) = \frac{\text{number of ways A can occur}}{\text{Total number of possible outcomes}} \qquad (4.1)$$

Example 4.1: A die is tossed once and the outcome could be any of these: The sample space is

$$S = \{1, 2, 3, 4, 5, 6\}$$

Example 4.2: Lets toss a coin twice and the outcome for each of toss in recorded. The sample space is shown here in two different ways.

Diagram Representation

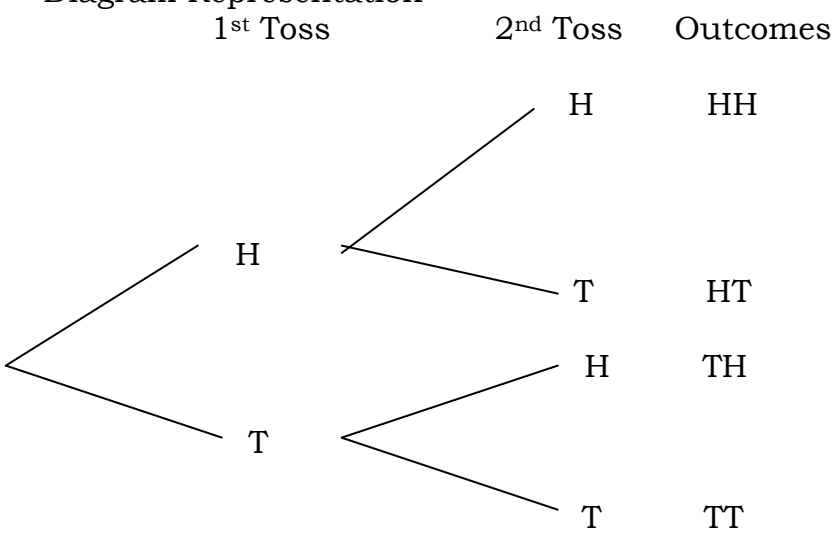


Figure 4.1: Tree Diagram

Listing

$S = \{HH, HT, TH, TT\}$

$n(S) = 4$

Example 4.3: Lets toss a coin thrice and the outcome for each toss is recorded.

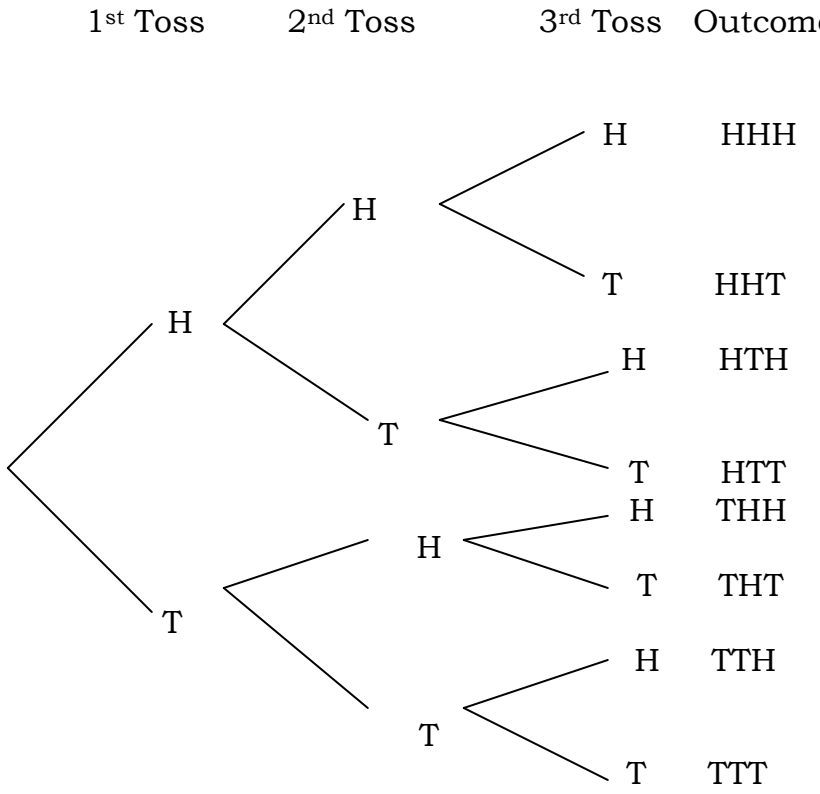


Figure 4.2 Tree Diagram

S = { HHH, HHT, HTH, THH,HTT,THT,TTH,TTT }
n(S) = 8

Example 4.4. Two dice are rolled and the sum of the numbers appearing are observed.

Table 4.1

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

The sample space S = {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}

x	2	3	4	5	6	7	8	9	10	11	12	
n(x)	1	2	3	4	5	6	5	4	3	2	1	36

with a total of 36-point sample space.

4.2 PERMUTATIONS AND COMBINATONS

4.2.1 Permutation

Permutation is a special arrangement of a group of objects in some order. Any other arrangement of the same objects is a different permutation. The key words for permutation are *order* or *arrangement*. For example, lets arrange n people in order. There are n possible chances for the first person, n-1 remaining possible chances for the second person, n-2 remaining possible chances for the third person, e.t.c, that is,

The number of possible arrangement = n x (n –1) x (n – 2) x ….x 1)
= n! (n factorial)

Example 4.5

0! = 1
1! = 1
2! = 2 x 1 = 2
3! = 3 x 2 x 1 = 6
4! = 4 x 3 x 2 x 1 = 24
5! = 5 x 4 x 3 x 2 x 1 = 120
6! = 6 x 5 x 4 x 3 x 2 x 1 = 720

$$nPr = \frac{n!}{(n-r)!}$$

this is the number of permutations of n objects taken r at a time.

Example 4.6: In how many ways can three people be seated on 6 seats in a row?

Solution

Arranging 3 people on 6 seats $= {}^6P_3$

$${}^6P_3 = \frac{6!}{(6-3)!} = \frac{6!}{3!} = \frac{6 \times 5 \times 4 \times 3!}{3!}$$

$$= 6 \times 5 \times 4 = 120 \text{ ways}$$

Example 4.7: How many distinct arrangements can be made using all the letters of the word Economics.

Solution

From the word Economics, o = 2, c = 2, and total letters = 9

$$\therefore \text{ Total arrangement} = \frac{\text{(Number of letter)!}}{\text{(Frequency of letters)!}}$$

$$= \frac{9!}{2!\,2!} = \frac{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2}{2 \times 2}$$

$$= 90720$$

Example 4.8: How many different numbers of six digits can be formed using digits 4, 4, 6, 6, 6, 6.

Solution

Total digits (n) = 6
4 has frequency = 2
6 has frequency = 4

Total numbers that can be formed $= \frac{5!}{2!\,4!}$

$$= \frac{6 \times 5 \times 4!}{2 \times 1 \times 4!}$$

$$= 15$$

Example 4.9

A plate number is to be made so that it contains four letters and four digits. Two letters begin the plate number and two letters end it. In how many ways can this number be made so that the first digits is not zero when

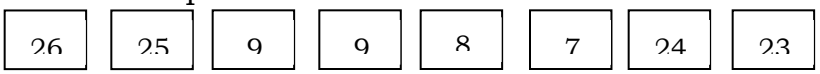
i. Both letters and digits cannot be repeated
ii. Both letters and digits can be repeated

Solution

For four letters and four digits, together will result to 8 boxes.

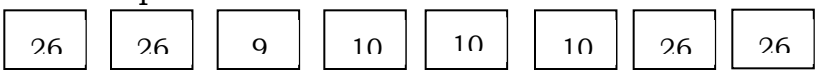
There are 26 alphabets a, b, c, …. z.
There are 10 digits 0, 1, 2, ……, 9.

i. Without Repetition



number of ways = 26 x 25 x 9 x 9 x 8 x 7 x 24 x 23
= 1,627,516,800 number plates

ii. With Repetition



number of ways = 26 x 26 x 9 x 10 x 10 x 10 x 26 x 26
= 26^4 x 10^3 x 9 = 4,112,784,000 number plates

4.2.2 Combination

Combination is any collection of a group of objects without regard to order. Problems involving combinations, where order is not relevant, are very similar to problems involving combinations, where order is critical. The only difference between permutations and combinations is whether order matters.

$$nC_r = \frac{n!}{(n-r)!\, r!}$$

is the number of possible combinations of n objects taken r at a time.

Example 4.10: Find the number of ways in which three students can be selected from five students.

Solution

3 students can be chosen from 5 students in $5C_3$ ways

$$= \frac{5!}{(5-3)!\, 3!} = \frac{5!}{2!\, 3!} = \frac{5 \times 4 \times 3!}{2!\, 3!}$$

$$= 5 \times 2 = 10 \text{ ways}$$

Example 4.11

A Mathematics examination consists of 8 questions out of which candidates are to answer 5. In how many ways can each candidate select if

a) There is no compulsory question
b) The first 3 questions are compulsory
c) At least 3 out of the first 4 questions are compulsory

Solution

a. From 8 questions to answer 5 questions, if there is no compulsory question,

we have $8C_5 = \frac{8!}{(8-5)!\, 5!} = \frac{8!}{3!\, 5!}$

$$= \frac{8 \times 7 \times 6 \times 5!}{3 \times 2 \times 5!}$$

$$= 56 \text{ ways}$$

b. If the first 3 questions are compulsory, then a candidate can choose 2 more questions from the remaining 5,

$$5C_2 = \frac{5!}{(5-2)!\,2!} = \frac{5!}{3!\,2!} = \frac{5 \times 4 \times 3!}{3!\,2!}$$

$$= 10 \text{ ways}$$

c. At least 3 out of the first 4 questions are compulsory means the candidate may answer 3 out of the first 4 compulsory questions and 2 from the remaining 4 questions or all the 4 first compulsory questions and 1 from the remaining 4 questions.

$$4C_3 \times 4C_2 + 4C_4 \times 4C_1$$

$$= \frac{4!}{(4-3)!\,3!} \times \frac{4!}{(4-2)!\,2!} + \frac{4!}{(4-4)!\,4!} \times \frac{4!}{(4-1)!\,1!}$$

$$= \frac{4!}{1!\,3!} \times \frac{4!}{2!\,2!} + \frac{4!}{0!\,4!} \times \frac{4!}{3!\,1!}$$

$$= 4 \times 6 + 1 \times 4 = 24 + 4 = 28 \text{ ways}$$

Example 4.12: From a gathering of 100 people of which 40 are men, a committee of 15 is to be formed. In how many ways can this be done so that (i) 3 men are there? (ii) no man is included?

Solution

Total number of people = 100
Men = 40, Women = 100 – 40 = 60

i. 3 men in the committee means 12 women in the committee

$$40C_3 \times 60C_{12} = \frac{40!}{(40-3)!\,3!} \times \frac{60!}{(60-12)!\,12!}$$

$$= \frac{40!}{37!\,3!} \times \frac{60!}{48!\,12!} = \frac{40 \times 39 \times 38}{6} \times \frac{60!}{48!\,12!}$$

$$= \quad 40 \times 13 \times 19 \times \frac{60!}{48! + 2!} = \frac{9880 \times 60!}{48!\ 12!}$$

ii. If no man is included, it means the whole of the committee members are women. We have 60 women in the gathering.

$$60C_{15} \times 40C_0 = 60C_{15} = \frac{60!}{(60-15)!\ 15!} = \frac{60!}{45!\ 15!}$$

Example 4.13: A bag contains 2 white and 3 red balls. In how many ways can 3 balls be chosen if

a. at least one ball must be white?
b. at least one ball must be red?

Solution

White balls = 2, red balls = 3
Total balls = 5

a. To choose at least one white ball means one white or more is to be chosen. That is, 1W and 2R or 2W and 1R
   $= 2C_1 x 3C_2 + 2C_2 x 3C_1$
   $= 2x3 + 1x3 = 6 + 3 = 9$ ways

b. to chose at least one red means one red or more. That is, 1R and 2W or 2R and 1W or 3R and no white

$$= \quad 3C_1 \times 2C_2 + 3C_2 \times 2C_1 + 3C_3$$
$$= \quad 3 \times 1 + 3 \times 2 + 1 = 3 + 6 + 1 = 10 \text{ ways}$$

4.3 LAWS OF PROBABILITY

A probability is always a numerical value between zero and one.

Property I

$$O \leq P(A) \leq 1$$

Property II

$$\sum_{alloutcomes} P(x) = 1$$

Property 2 states that if we add up the probabilities of each of the sample points in the sample space, the total probability must equal one.

Example 4.14: Find the probability that a head will appear when two coins are tossed.

Solution

The sample space = { HH, HT, TH, TT}
Let event A be the occurrence of one head.

$$P(\text{a head will appear}) = \frac{\text{numbers of A in sample space}}{\text{number in sample space}}$$
$$= 2/4 = 0.5$$

Example 4.15: Two dice are rolled and the sum of the numbers appearing are observed. Find the possibility of getting (i) a total of 5 (ii) a total of 12.

Solution
From example 4.3, the sample space is given as
(i) P (of getting a total of 5) = $\frac{\text{numbers with total of 5}}{\text{number in sample space}}$

$$= \frac{4}{36}$$
$$= 1/9$$

ii. P (of getting a total of 12) = 1/36

COMPLEMENT OF AN EVENT

The set of all sample points in the sample space that do not belong to event A. The complement of event A is denoted by A^1 or A^c.

This also implies that $P(A) + P(A^1) = 1$

Example 4.16: Find the probability that at least a tail will appear when two coins are tossed.

Solution

Let A be the occurrence of no tail then A^1 will be the occurrence of at least one tail will appear.

Sample space = {HH, HT, TH, TT}

P(A) = ¼
Therefore, $p(A^1) = 1 - P(A) = 1 - ¼ = ¾$

Combined events are formed by combining several simple events. For example, the probability of either event A or event B will occur is P(A or B) or $P(A \cup B)$; the probability that both events A and B will occur is P(A and B) or $P(A \cap B)$.

MUTUALLY EXCLUSIVE EVENTS

These are events defined in such a way that the occurrence of one event precludes the occurrence of any of the other events. (In short, if one of them happens, the others cannot happen.)

Consider an experiment in which two dice are tossed. Three events are defined.

A: the sum of the numbers on the two dice is 5
B: the sum of the numbers on the two dice is 8.
C: each of the two dice shows the same number.
Table 4.2

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Events A and B are mutually exclusive, because the sum on the two dice cannot be both 5 and 8 at the same time. It is clearly seen in Table 4.2 that events A and B do not intersect at a common sample point. Therefore, they are mutually exclusive. Point (4, 4) satisfies C, and the total of the two 4s satisfies B.

ADDITION LAW

If events A and B are mutually exclusive, then

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) \quad (4.3)$$
$$\text{i.e. } P(A \cap B) = 0$$

If the events A and B are not mutually exclusive then

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (4.4)$$

In other words, if two events A and B are mutually exclusive, then we can find the probability that either A or B will occur by simply adding the individual probabilities. If they are not mutually exclusive, then there is some overlap between them. Thus, to find the probability that either A or B will occur, we must subtract the probability of the duplication P(both A and B) from the sum P(A) + P(B).

This can be expanded to consider more than two mutually exclusive events:

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \quad (4.5)$$

INDEPENDENCE

Two events A and B are *independent* if the probability of second event B is not affected by the occurrence or nonoccurrence of the first event A. If A and B are independent events, then

$$P(\text{Both A and B}) = P(A \cap B) = P(A) \times P(B) \quad (4.6)$$

That is, when A and B are independent events, the probability that both hold is just the product of the individual probabilities. This formula can be extended. If A, B, C, ……, Z are independent events then,

$$P(A \text{ and } B \text{ and } C \text{ and } \ldots \text{ and } Z) = P(A \cap B \cap C \cap \ldots \cap Z) \qquad (4.7)$$

$$= P(A).\ P(B).\ P(C) \ldots\ldots.\ P(Z)$$

Example 4.17: Let two events A and B be defined on the same sample space. Suppose P(B) = 0.2 and P(A ∪ B) = 0.75. Find P(A) such that

i. A and B are independent
ii. A and B are mutually exclusive

Solution

i. If A and B are independent, then
$P(A \cap B) = P(A)\ P(B)$

Thus $P(A \cup B) = P(A) + P(B) - P(A)\ P(B)$
$0.75 = P(A) + 0.2 - P(A) \times 0.2$
$= 0.2 + 0.8\ P(A)$

$0.8\ P(A) = 0.75 - 0.2 = 0.55$
$P(A) = \frac{0.55}{0.8}$
$= 0.6875$

ii. If A and B are mutually exclusive, then
$P(A \cap B) = 0$

Thus $P(A \cup B) = P(A) + P(B)$
$0.75 = P(A) + 0.2$
$P(A) = 0.75 - 0.2 = 0.55$

Example 4.18: Find the probability of getting:
(i) 2 heads (ii) 1 head (iii) no head if a coin is tossed twice.

Solution

P(getting of head) = ½

i. P(getting two heads) = P(1st is head) P(2nd is head)
= ½ x ½ = ¼

ii. P(getting 1 head) = P(1st is head and 2nd is tail) or
P(1st is tail and 2nd is head)
= P(H T) + P(T H)
= ½ x ½ + ½ + ½ = ¼ + ¼ = ½

iii. P(getting no head) = P(1st is tail and 2nd is tail)
= ½ x ½ = ¼

CONDITIONAL PROBABILITY

Most work we have considered so far involved independent events. As a result, getting the probability of such an event was reasonably straightforward. However, when the events are dependent, then solving them become more complicated.

Conditional Probability written as P(B/A) is the probability of an event B given that a "previous" event A has occurred. The conditional probability of B given A is

$$P(B/A) = \frac{P(\text{both A and B})}{P(A)} = \frac{P(A \cap B)}{P(A)} \qquad (4.8)$$

If two events are independent events then

$$P(B/A) = P(B) \text{ or } P(A/B) = P(A) \qquad (4.9)$$

Example 4.19: A coin is tossed thrice. Find the probability that there are two heads (i) given that at least one is a tail (ii) given that the first is a tail.

Solution

S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}

i) Sample space for at least one tail

S = {HHT, HTH, THH, HTT, THT, TTH, TTT} having 7 elements. n(2 heads) = 3

Therefore, the conditional probability that there are 2 heads, given that at least one is a tail, is

P(2 heads/at least one is a tail) = 3/7

ii. Sample space for the first being tail
= {THH, THT, TTH, TTT} having 4 elements. n(2 heads) = 1.

Therefore, the conditional probability that there are two heads given that the first is a tail is

P(2 heads/first is a tail) = ¼

USING THE FORMULA

i. Let A: at least one is a tail
B: 2 heads appear

The conditional probability is

P(B/A) = P(2 heads appear/at least one is a tail)

$$= \frac{P(\text{2 heads appear and at least one is a tail})}{P(\text{at least one is a tail})}$$

Using the original sample space of all 8 equally likely possible outcomes, we see that
P(at least one is a tail) = 7/8 and
P(2 heads appear and at least one is a tail) = 3/8

Therefore,

$$P(B/A) = P(\text{2 heads appears/at least one is a tail}) = \frac{3/8}{7/8}$$

$$= 3/8 \times 8/7 = 3/7$$

which is the same result as we obtained above.

ii. P(the first is a tail) = 4/8
P(2 heads appear and the first is a tail) = 1/8

Therefore,
P(2 heads appear/the first is a tail)

$$= \frac{P(\text{2 heads appear and the first is a tail})}{P(\text{the first is a tail})}$$

$$= \frac{1/8}{4/8}$$

$$= 1/8 \times 8/4 = \frac{1}{4}$$

Example 4.20: A die to tossed twice. Find (i) probability of getting a sum of 9 (ii) probability of getting a sum of 9 giving that the number on the 2nd toss is larger than the number on the first toss.

Solution

+	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

i. P(getting a sum of 9) $= \frac{4}{36} = \frac{1}{9}$

iii. Let A = "sum of 9", B = "2nd toss number larger than the first toss".

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$P(B)$ = P(2nd toss number larger than the first toss) = $\frac{15}{36}$

$P(A \cap B)$ = P(sum of 9 and 2nd toss number larger than the first toss) = $\frac{2}{36}$

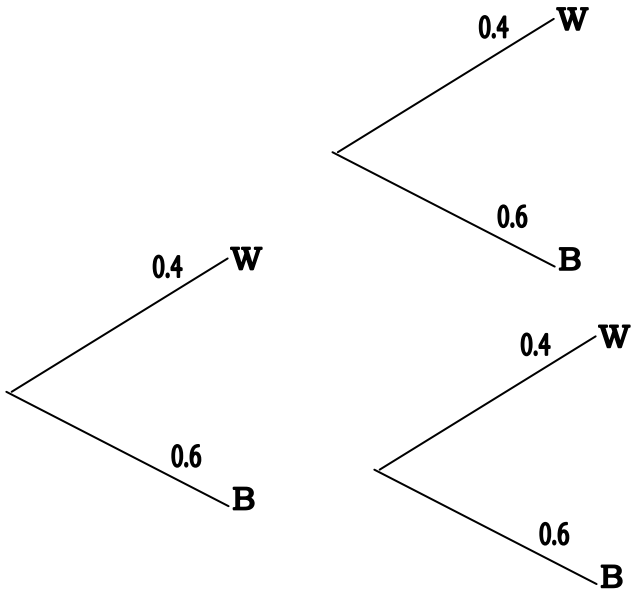
Therefore,

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{2/36}{15/36} = 2/15$$

Example 4.21: A bag contains 10 balls. Four are white and 6 are black. Draw the tree diagram when two balls are drawn with (i) replacement (ii) no replacement

Solution

(i) with replacement Figure 4.3 Tree Diagram



P(2nd white) = P(2nd is white and 1st is white) or
P(2nd is white and 1st is black)

= P(1st is white) – P(2nd is white given the 1st is white)
+ P(1st is black). P(2nd is white given the 1st is black)

= 0.4 x 0.4 + 0.6 x 0.4
= 0.16 + 0.24 = 0.4

ii without replacement

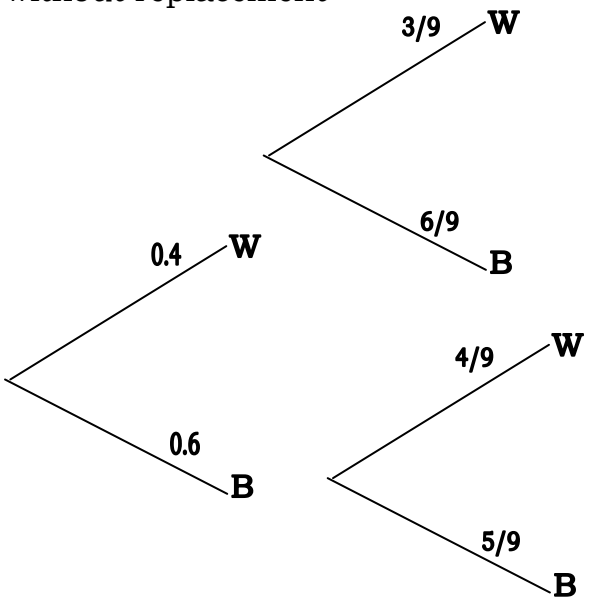
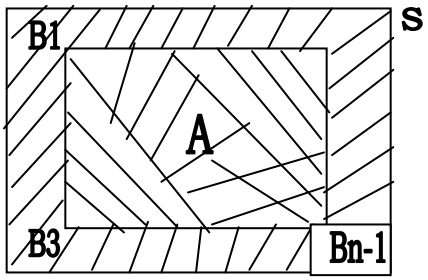


Figure 4.4 Tree Diagram

P(2nd white) = 0.4 x 3/9 + 0.6 x 4/9
= 0.133 + 0.267 ≅ 0.4

Let B_1, B_2, …. B_n for a partition of a sample space S. Let A ∈ S, then



$$P(A) = P(A/B_1).\ P(B_1) + P(A/B_2)\ P(B_2) + \ldots\ldots + P(A/B_n)\ P(B_n)$$

That is,

$$P(A) = \sum_{i=1}^{n} P(A/B_i).P(B_i) \qquad (4.10)$$

i = 1

4.4 BAYES'S THEOREM

Bayes's theorem was developed by an English Presbyterian minister, Reverend Thomas Bayes (1702 – 1761). This is an expanded form for conditional probabilities.

$$P(B_i/A) = \frac{P(B_i).\ P(A/B_I)}{\sum [\ P(B_i).\ P(A/B_i)]} \qquad (4.11)$$

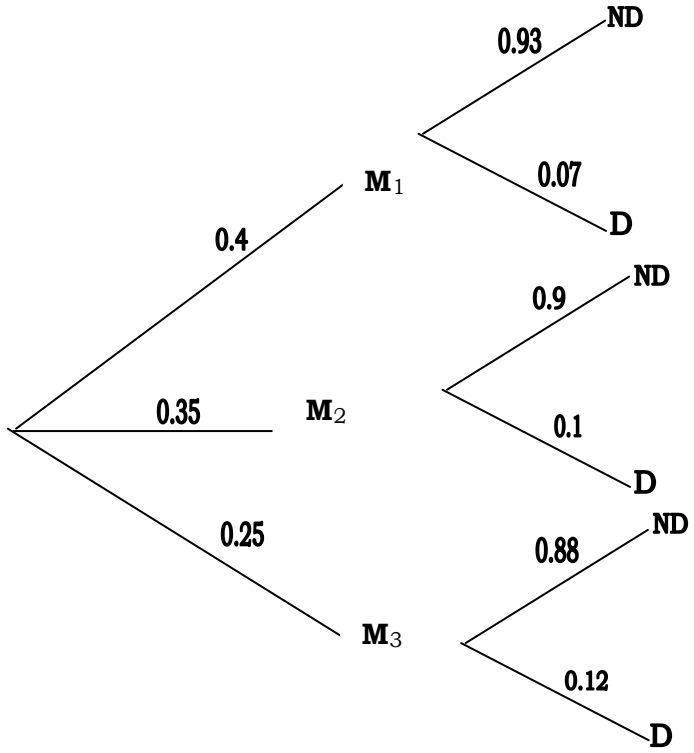
where B_1, B_2, ….. B_n is an all – inclusive set of possible outcomes given A

Example 4.22: A product is being produced by three machines M_1, M_2 and M_3. These machines produce 40%, 35% and 25% of the product respectively. Accordingly, the respective defective products produced by these machines M_1, M_2 and M_3 are 7%, 10% and 12% respectively. Find

i. the probability that a part selected at random from the finished product is defective

ii. the probability of the defective product was produced by machine M_1, M_2, or M_3.

Solution



M_1 - Machine 1 ND-Non defective product
M_2 - Machine 2 D-Defective product
M_3 - Machine 3

$P(M_1) = 0.4$ $P(D/M_1) = 0.07$
$P(M_2) = 0.35$ $P(D/M_2) = 0.1$
$P(M_3) = 0.25$ $P(D/M_3) = 0.12$

i. $P(D) = P(M_1)\,P(D/M_1) + P(M_2)\,P(D/M_2) + P(M_3)\,P(D/M_3)$
$= 0.4 \times 0.07 + 0.35 \times 0.1 + 0.25 \times 0.12$
$= 0.028 + 0.035 + 0.03 = 0.093$

ii. $P(M_i/D) = \dfrac{P(M_i \cap D)}{\sum P(M_i \cap D)} = \dfrac{P(M_i)\,P(D/M_i)}{\sum P(M_i)\,P(D/M_i)}$

Thus

$$P(M_i/D) = \frac{0.4 \times 0.07}{0.93} = \frac{0.028}{0.093} \approx 0.3011$$

$$P(M_2/D) = \frac{0.35 \times 0.1}{0.93} = \frac{0.035}{0.093} \approx 0.3763$$

$$P(M_3/D) = \frac{0.25 \times 0.12}{0.093} = \frac{0.035}{0.093} \approx 0.3226$$

4.5 EXERCISES

4.5.1 Two unbiased dice are thrown once. What is the probability that

i. the sum is 8?
ii. the sum is 8 given that a 3 appears?
iii. at least one 3 is thrown?

4.5.2 How many different license plates are possible if a country uses:

a) Two letters followed by a four-digit integer (leading zeros permissible and the letters and the digits can be repeated)?

b) Three letters followed by a three-digit integer (leading zeros not permissible and the letters and the digits cannot be repeated)?

4.5.8 Suppose that $P(A) = 0.45$, $P(B) = 0.6$, and $P([A \cup B]^1) = 0.2$

a. Find $P(A \cap B)$
b. Give $P(A/B)$
c. Give $P(B/A)$

4.5.9 A drawer contains five black, seven brown and six-yellow socks. Two socks are selected at random from the drawer.

a) Compute the probability that both socks are different colors.
b) Compute the probability that both socks are the same color.
c) Compute the probability that both socks are yellow if it is known that they are the same color.

4.5.13 Beans seeds from supplier A have a 60% germination rate and those from supplier B have a 70% germination rate. A seed packaging company purchases 45% of their bean seeds from supplier A and 55% from supplier B and mixes these seeds together.

a) Find the probability that a seed selected at random from the mixed seeds will germinate say P(G).
b) Given that a seed germinates, find the probability that the seed was purchased from supplier B.

END OF MODULE ASSESSMENT

4.5.19A box contains 45 biros, of which 5 are defective. If 2 biros are selected without replacement, find the following probabilities:

a) P(both are non-defective)
b) P(both are defective)
c) P(exactly one is defective)

4.5.10 Let A and B be independent events with P(A) = 0.6 and P(B) = 0.3. Compute

a. $P(A \cap B)$, (b) $(A \cup B)$, and (c) $P(A^1 \upsilon B^1)$

3.11.10 Use the table below to find

a) mean, median, mode
b) the coefficient of skewness for the data and comment of the degree of symmetry of the data.
c) the 10^{th} and 90^{th} percentile

d) the 4^{th} and 8^{th} deciles

Profit	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	15	13	24	17	5	16	15

2.3.7 The following table shows the number admitted into the postgraduate programme for 2 years.

Department	**2003**	**2004**
Chemical Engn.	41	37
Electrical Engn.	40	48
Surveying	35	25
Geography	45	48
Mechanical Engn.	50	45
Geology	35	45

Draw a component and multiple bar charts for this data

CHAPTER TEN

TEST OF STATISTICAL HYPOTHESES

In this chapter, we will find out how an hypothesis test is used to make a statistical decision about the mean μ, standard deviation σ, and the binomial probability of "success" p and learn about the basic concepts of hypothesis testing.

10.1 DEFINITIONS AND CONCEPTS

Hypothesis is a statement that something is true. There are two types of hypothesis, the null hypothesis and the alternative hypothesis.

Let us assume, that the mean age of 100 students on a University Campus is 33 years. This claim or statement about the mean of a population is known as the *null hypothesis* and is denoted by H_0. For the above illustration, we write

$$H_0 : \quad \mu = 33$$

Since. $H_0 : \mu = 33$ specifies the distribution completely it is called a *simple hypothesis*: thus $H_0 : \mu = 33$ is a *simple null hypothesis*.

In addition, we have an *alternative hypothesis*, denoted by H_1, which states our suspicion about the population mean μ. We might suspect that the mean age is less than 33 years. In this case, we write the alternate hypothesis as

$$H_1 : \mu < 33$$

Since $H_1 : \mu < 33$ does not completely specify the distribution. It is a *composite hypothesis* because it is composed of many simple hypotheses.

Another possibility is to say the mean age is actually greater than the claimed value. In such a case, the alternate hypothesis takes the form.

$H_1 : \mu > 33$

In addition, we may make no indication of whether the value is less or greater than, we may simply say the mean is different from the claim, and we therefore write.

$H_1 : \mu \neq 33$

However, in any given problem, there is just one alternative hypothesis. An appropriate one from the three is chosen based on the context of the study being conducted.

The first two possibilities for an alternate hypothesis

$H_1 : \mu < 33$

and

$H_1 : \mu > 33$

are known as *one tailed tests*. The third possibility

$H_1 : \mu \neq 33$

is known as a *two-tailed test*.

At the conclusion of the hypothesis test, we decide which one of the two possible *decisions* to take. We decide to "fail to reject H_0" or we decide to oppose the null hypothesis and say we "reject H_0". Our rule of rejecting H_0 and accepting H_1 and otherwise accepting H_0 is called a *test of a statistical hypothesis*.

Whenever we make a judgment in a hypothesis-testing situation, we can make either the correct decision or an incorrect one. A correct decision arises in two ways:

1. Reject a null hypothesis which is wrong or

2. Fail to reject the null hypothesis that is valid

Similarly, there are two ways to make an incorrect decision:

1. Reject the null hypothesis which is valid or
2. Fail to reject the null hypothesis that is wrong

A *type I error* is the error we make when we wrongly reject a valid null hypothesis, while a *type II error* is the error in accepting an incorrect null hypothesis.

The four cases that can occur are:

Table 10.1

	Accept H_0	Reject H_0
H_0 true	Correct decision	Type I error
H_0 false	Type II error	Correct decision

It is not always possible to make correct decisions. The best one can do is to control the risk, or probability, with which an error occurs. The probability assigned to the type I error is called alpha, α(known also as level of significance) and the probability of the type II error is called beta, β. We shall only consider alpha in this textbook. Typically, α = 0.10, 0.05, 0.02 or 0.1 or the percentage equivalents 10%, 5%, 2% or 1%.

The *test criteria* consist of

1. specifying a level of significance, α
2. determining a test statistic,
3. determining the critical region(s)
4. determining the critical value(s).

A *test statistic* is a random variable used to make the decision whether to "fail to reject H_0" or "reject H_0". The *critical region* is the set of values of the test statistic that will cause us to reject the null hypothesis. The cutoff value(s) for the test statistic that defines the rejection region is (are) known as the critical value(s) for the test statistic.

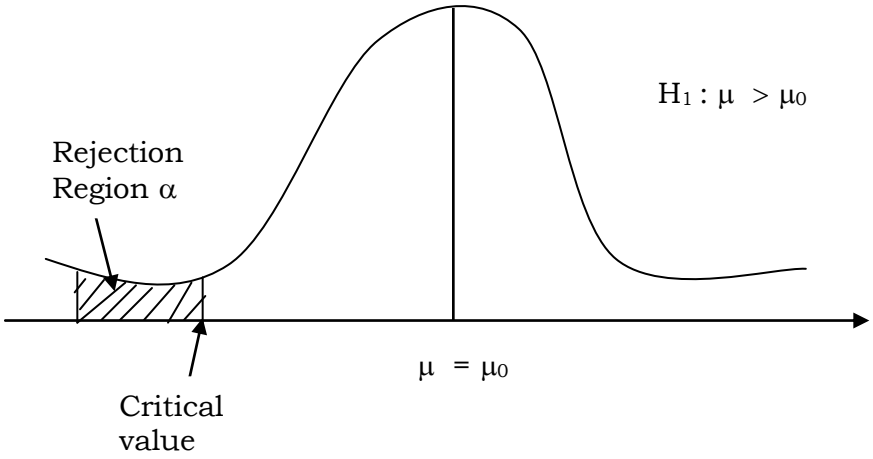


Figure 10.1 A one-tailed test

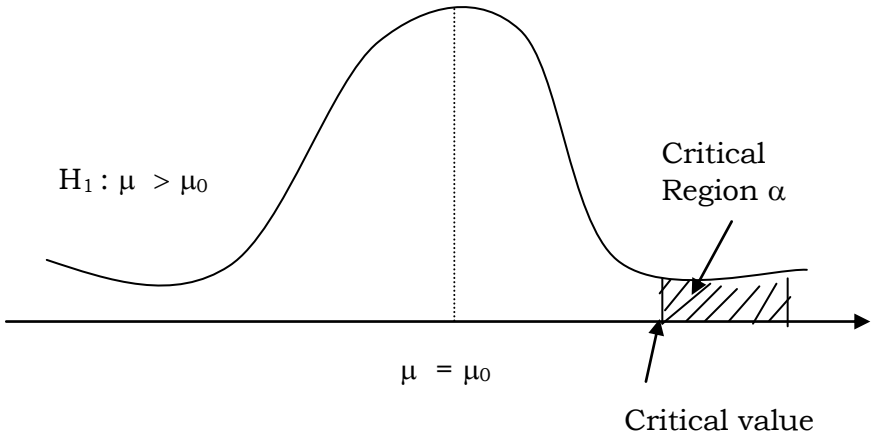


Figure 10.2 One tailed test

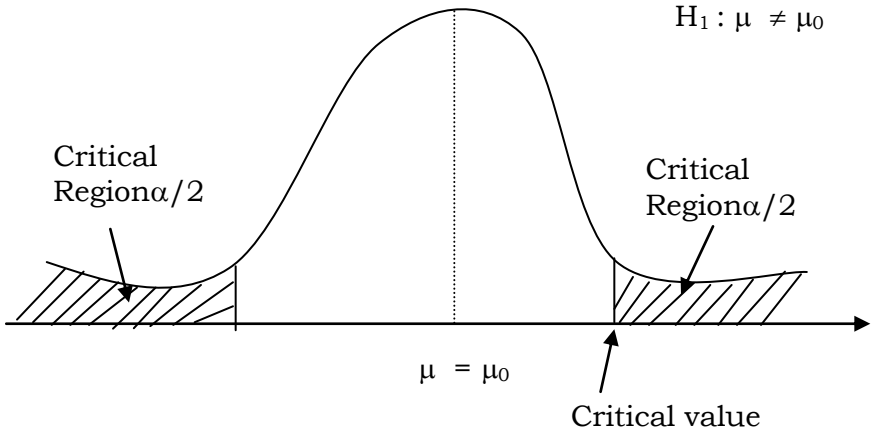


Figure 10.3. Two-tailed test.

If the test statistic falls within the critical region, we reject H_0. If the test statistic does not fall within the critical region, we will fail to reject H_0. The set of values that are not in the critical region is called the *non-critical region* or, sometimes, the *acceptance region*.

Hypothesis testing: A four-step model.

Step 1: State the null hypothesis (H_0) and the alternative hypothesis (H_1)

Step 2: Determine the test criteria:

a. the level of significance,α, of the test
b. the test statistics appropriate
c. the critical region(s)
d. the critical value(s)

Step 3: Compute the value of the observed test statistic

Step 4: Determine the results.

a. compare the calculated result of the test statistic to the critical value(s)
b. Make a decision about H_0

c. conclude about H_1

10.4 THE CHI-SQUARE DISTRIBUTION

Chi-square distribution had been used in chapter 8 and 10 to estimate and to test the value of the variance (or standard deviation) of a single population. Other applications of chi-square distribution are considered in this section. The Goodness-of-fit test and analysis of contigency table will be looked into.

This section deals with the analysis of categorical data. These are data whose values are categories of responses. Example are

* Sex (male or female)
* Marital status (married, single, divorced, separated, widowed)
* Religion – (Islam, Christianity, Traditional, Jewish or

others)

* Income
* Ethnicity (Hausa, Yoruba, Ibo, others)
* Level of Education (Univeristy, Polythenic, College of Education, Vocation institute, Secondary or Primary) or many other possible characteristics.

The *goodness-of-fit* test provides us with a means of determining whether a set of data is a good fit to a theoretical model or prediction while determining whether two categorical variables are true or otherwise is the study of analysis of contingency tables.

10.4.1 Goodness-of-fit Test

Suppose that we have a number of *cells* into which n observations have been sorted.

Table 10.1

	K Categories				
	1st	2nd	3rd	.....	kth
Observed Frequency	O_1	O_2	O_3	.....	O_k

The observed frequencies in each cell are denoted by O_1, O_2, O_3, …. O_k and the sum of all observed frequencies is equal to $O_1 + O_2 +$ …. $O_k = n$, where n is the sample size. These observed frequencies are compared with the *expected or theoretical frequencies*, denoting by E_1, E_2, E_3, ….., E_k, for each of these cells. Again, the sum of these expected frequencies must be exactly $E_1 + E_2 +$ ….. $+ E_k = n$.

We then decide whether the observed frequencies seem to agree or seem to disagree with the expected frequencies. The calculated value of the test statistics is given by

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (10.7)$$

In general, the number of degrees of freedom for any goodness-of-fit test is given by

df = number of categoreis – 1

The critical value of chi-square, χ^2_α (df) can be found in Table V of the Appendix

Example 10.20: A die is rolled 420 times with following outcomes

1	2	3	4	5	6
63	67	66	73	72	79

Based on these data, test whether the die is fair at the $\alpha = 0.05$ level of significance.

Solution

Step 1: H_0: The die is fair
H_1: The die is not fair

Step 2: $\alpha = 0.05$, df = k – 1 = 6 – 1 = 5 (where k is the number of cells). The critical value is

χ^2_α (df) = $\chi^2_{0.05}$ (5) = 11.1

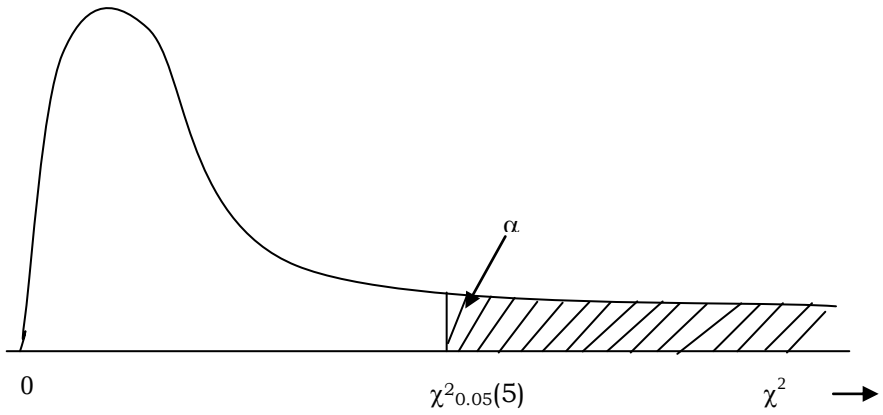


Figure 10.45

Step 3: To calculate the test statistic we use the formula

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)}{E_i}$$

Table 10.2

Number	**Observed (O_i)**	**Expected (E_i)**	**$O_i - E_i$**	**$(O_i - E_i)^2$**	**$(O_i - Ei)^2 / E_i$**
1	63	70	-7	49	0.7
2	67	70	-3	9	0.129
3	66	70	-4	16	0.229
4	73	70	3	9	0.129
5	72	70	2	4	0.057
6	79	70	9	81	1.157
Total	**420**	**420**	**0**		**2.401**

$\chi^{2*} = 2.401$

The test statistic falls within the non-critical region.

Step 4: Fail to reject H_0

The observed frequencies are not significantly different from those expected of a fair die at 5% level of significance.

Example 10.21: Blood types in the general population are distributed as follows:- 20% have type O, 45% have type A, 20% have type B and 15% have type AB. A group of 300 people from a certain area are tested and found to have the following frequencies for the different blood types

Type	0	A	B	AB
Frequency	75	50	90	85

On the basis of these results, can we conclude that people from this area have a different distribution of blood types at the $\alpha = 0.05$ level of significance?

Solution

Step 1: H_0: There is no difference in the distribution of blood types.

H_0: The is difference in the distribution of blood types.

Step 2: $\alpha = 0.05$, df = k – 1 = 4 – 1 = 3

The critical value is given as $\chi^2_{0.05}(3) = 7.82$

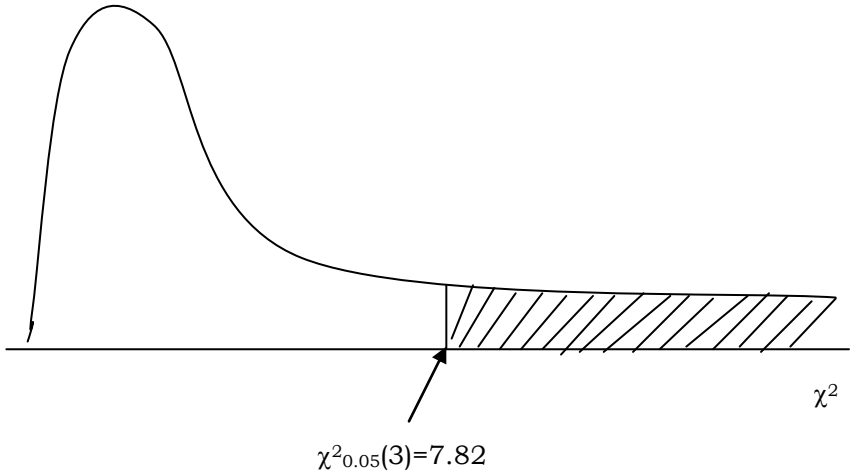


Figure 10.46

Step 3:

Table 10.4

Number	**Observed (O_i)**	**Expected (E_i)**	**$O_i - E_i$**	**$(O_i - E_i)^2$**	**$(O_i - E_i^2/ E_i$**
1 (0)	75	60	15	225	3.75
2 (A)	50	135	-85	7225	53.52
3 (B)	90	60	30	900	15
4 (AB)	85	45	40	1600	35.56
Total	**300**	**300**	**0**		**107.83**

$\chi^{2*} = 107.83$

The test statistic falls within the critical region.

Step 4: Reject H_0. There is significant difference in the distribution of blood type in the area.

10.4.2 Contingency Tables

The arrangement of data into a two-way classificiation is known as a *contigency table*. Contigency table test whether the data of two variables are independent or dependent.

To obtain the expected frequency for a cell, we multiply the column total for the cell by the fraction formed by dividing the corresponding row total by the grand total.

$$\text{Expected frequency} = \frac{\text{row total x column total}}{\text{grand total}}$$

$$E_{ij} = \frac{R_i \times C_j}{n}$$

Table 10.5 Contingency Table

	C_1	C_2	…….	C_c	Row Total
R_1	O_1,	O_2,	…….	O_{1c}	R_1
R_2	O_2,	O_2,	…….	O_{2c}	R_2
R_r	O_{r1}	O_{r2}	…….	O_{rc}	R_r
Column Total	C_1	C_2		C_c	Grand Total

The number of degrees of freedom for the chi-square distribution associated with a particular contigency table. We take the product of 1 less than the number of rows and 1 less than the number of columns in the contigency table.

$$df = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

$= (r - 1) \times (c - 1)$

The chi-square test statistic is as given in formula 10.7

Example 10.22: A study is conducted comparing the proportions of people of different age groups who prefer several major brands of soft drink with these results.

	Below 25yrs	Between 25 and 45yrs	Above 45 yrs
Coca-Cola	45	60	45
Pepsi-Cola	35	80	95
Seven-Up	40	45	55

Test whether there is any difference in the propportions of people preferring the different brands based on age at 5% level of significance.

Solution

Step 1: H_0: The proportions of people preferring different brands of soft drink is independent of the age

H_1: The preference is not independent of the age

Step 2: To determine the critical value of chi-square, we need to know the degrees of freedom. The degrees of freedom is

$$(r - 1) \times (c - 1) = (3 - 1)(3 - 1) = 2 \times 2 = 4$$

The critical value is

χ^2_{α} (df) = $\chi^2_{0.05}$ (4) = 9.49

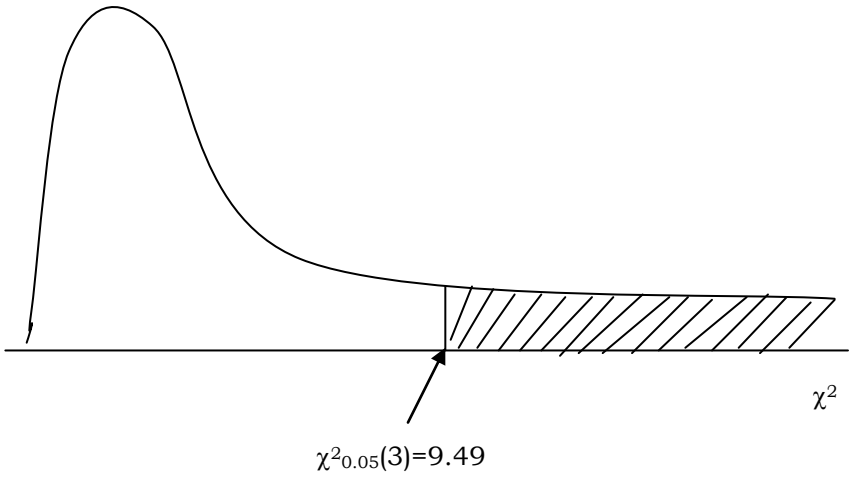


Figure 10.47

Step 3: There is need to calculate the expected value for each cell before we can compute the value of the chi-square.

Table 10.6 Expected Table

χ = ages	**X < 25**	**25 ≤ X ≤ 45**	**X > 45**	**Total**
Coca-Cola	$\frac{150 \times 120}{500} = 36$	$\frac{150 \times 185}{500} = 55.5$	$\frac{150 \times 195}{500} = 58.5$	150
Pepsi-Cola	$\frac{210 \times 120}{500} = 50.4$	$\frac{210 \times 185}{500} = 77.7$	$\frac{210 \times 195}{500} = 81.9$	210
Seven-Up	$\frac{140 \times 120}{500} = 33.6$	$\frac{140 \times 185}{500} = 51.8$	$\frac{140 \times 195}{500} = 54.6$	140
Total	**120**	**185**	**195**	**500**

Table 10.7

Number	Observed (O)	Expected (E)	O – E	$(O – E)^2$	$(O - E^2/ E$
1	45	36	9	81	2.25
2	60	55.5	4.5	20.25	0.37
3	45	58.5	-13.5	182.25	3.12
4	35	50.4	-15.4	237.25	4.71
5	80	77.7	2.3	5.29	0.07
6	95	81.9	13.1	171.61	2.10
7	40	33.6	6.4	40.96	1.22
8	45	51.8	-6.8	46.24	0.89
9	55	54.6	0.4	0.16	0.00
Total	**500**	**500.0**	**0**		**14.73**

$\chi^{2*} = 14.73$

The test statistic falls within the critical region.

Step 4: Reject H_0. That is, at 5% level of significance, the proportions of people preferring different brands of soft drink is not independent of the age. This implies that the ages of people determine what soft drinks they take.

Note: When performing the chi-square analysis for a contingency table, it is recommended that the expected frequencies E. in each cell is greater than or equal to 5. If any of the expected frequencies is less than 5, then the chi-square distribution may not be a good approximation and the results become invalid.

10.5 ANALYSIS OF VARIANCE

Researchers often want to compare more than two means. Yields of several maize hybrids, results due to three or more treatments or teaching techniques and so on. We now introduce a method known as Analysis Of Variance (or ANOVA) which allows us to test all the means simulteneously to see if there is any difference among them. The single-factor or one-way ANOVA will be

considered. The main idea in a one-way ANOVA is to divide the total variation into two components, that is,

i) Variation due to *treatment* or factor, that is, factor effect, and
ii) Variation due to random error

Table 10.8 One-factor random samples

		Observations			Mean	
	x_1:	x_{11}	x_{12}	.....	x_1n_1	$\bar{x}_1.$
	x_2:	x_{21}	x_{22}	.....	x_2n_2	$\bar{x}_2.$
Treatments	.	.	.	.		.
(factor)	.	.	.	.		.
	.	.	.	.		.
	x_m	x_{m1}	x_{m2}	.....	x_mn_n	$\bar{x}_m.$
					Grand Mean:	$\bar{x}..$

We summarize the steps needed to solve any ANOVA.

Step1: Calculate SS (Factor)

$$= \sum_{i=1}^{m} (n_i [\bar{x}_i. - \bar{x}..)^2$$

$$= \frac{\sum (\text{Row total})^2}{\text{number of entries per row}} - \frac{(\text{grand total})^2}{\text{total sample size}}$$

Step2: Calculate SS (Factor) $= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}..)^2$

$$= \sum x_{ij}^2 - \frac{(\text{grand total})^2}{\text{total sample size}}$$

Step 3: Find SS (Error)

SS (Error) = SS (Total) – SS (Factor)

Step 4: Calculate df (Factor)
df (Factor) = m – 1

Step 5: Calculate df (Total)
df (Total) = m x n – 1

Step 6: Find df (Error)
df (Error) = df (Total) – df (Factor)

Step 7: Calculate MS (Factor)
$$\text{MS (Factor)} = \frac{\text{SS (Factor)}}{\text{df (Factor)}}$$

Step 8: Calculate MS (Error)
$$\text{MS (Error)} = \frac{\text{SS (Error)}}{\text{df (Error)}}$$

Step 9: Calculate the F value
$$F = \frac{\text{MS (Factor)}}{\text{MS (Error)}}$$

Step 10: Draw a conclusion. Compare this test statistic to the critical value of F found in the F – distribution table (Table VI), using the appropriate numbers of degrees of freedom and the appropriate significance level α.

Table 10.9 Analysis of variance table

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	Mean Square (MS)	F - Ratio	F - Tab
Traatment (Factor)	SS (Factor)	m - 1	MS (Factor) = $\frac{\text{SS (Factor)}}{\text{m - 1}}$	$\frac{\text{MS (Factor)}}{\text{MS (Error)}}$	F_α (m – 1, n – m)
Error	SS (Error)	n – m	MS (Error) = $\frac{\text{SS (Error)}}{\text{n - m}}$		
Total	SS (Total)	n – 1			

All ANOVA hypothessis tests involving the F-distribution are one-tailed tests, we reject the null hypothesis that the means are equal only if the F-statistic is too large. That is, we reject H_0, when Ftab > Fratio.

When H_0 is true, we may regard x_{ij}, i (factor) = 1, 2, ..., m, j = 1, 2, ..., n_i, as a random sample of size $n = n_1 + n_2 + ... + ni$ from the normal distribution with common variance.

Example 10.23: Construct the ANOVA table and compare the F value to the critical value at the $\alpha = 0.01$ level of significance.

A	B	C
40	55	60
53	63	55
60	41	58

Solution

H_0: $\mu_A = \mu_B = \mu_C$
H_1: $\mu_A \neq \mu_B \neq \mu_C$

Table 10.10

A	40	53	60	153
B	55	63	41	159
C	60	55	58	173
				485

m = 3 (Factor A, B and C)
$n = n_A + n_B + n_C = 3 + 3 + 3 = 9$

Step 1:

$$SS\ (Factor) = \frac{\sum (\text{Row total})^2}{\text{number of entries per row}} - \frac{(\text{grand total})^2}{\text{total sample size}}$$

$$= \frac{(153)^2 + (159)^2 + (173)^2}{3} - \frac{(485)^2}{9}$$

$$= \frac{78619}{3} - \frac{235225}{9}$$

$$= 26206.33 - 26136.11$$

$$= 70.22$$

Step 2:

$$SS\ (Total) = \sum x_{ij}^2 - \frac{(grand\ total)^2}{total\ sample\ size}$$

$$= (40)^2 + (53)^2 + (60)^2 + ... + (58)^2 - \frac{(485)^2}{9}$$

$$= 26673 - 26136.11$$

$$= 536.89$$

Step 3:

SS (Error) = SS (Total) - SS (Factor)

$$= 536.89 - 70.22 = 466.67$$

Step 4: df (Factor) = m – 1 = 3 –1 = 2

Step 5: df (Total) = n – 1 = 9 – 1 = 8

Step 6: df (Error) = df (Total) – df (Factor)

$$= 8 – 2 = 6$$

Step 7:

$$MS\ (Factor) = \frac{SS\ (Factor)}{df\ (Factor)}$$

$$= \frac{70.22}{2} = 35.11$$

Step 8:

$$\text{MS (Error)} = \frac{\text{SS (Error)}}{\text{df (Error)}}$$

$$= \frac{466.67}{6} = 77.78$$

Step 9:

$$F = \frac{\text{MS (Factor)}}{\text{MS (Error)}} = \frac{35.11}{77.78}$$

$$= 0.029$$

Step 10:
$F_{tab} = F_{0.01}$ (m – 1, n – m) = $F_{0.01}$ (2,6) = 10.9. Two degrees of freedom for the numerator and 6 degrees of freedom for the denominator. We discover that $F_{ratio} < F_{tab}$, therefore, we fail to reject the null hypothesis. The conclusion is that there may not be any difference among the three population means.

Table 10.11

Source of variation	df	SS	MS	F_{ratio}	F_{tab}
Factor	2	70.22	35.11	0.029	$F_{0.01}$ (2, 6) = 10.9
Error	6	466.67	77.78		
Total	8	536.89			

Example 10.24: A statewide teachers' Union is studying the average class size for primary school classes in four local government area in the state. The respondents to a survey provide the following data on such classes.

Local Government A	35	25	42	37	15
Local Government B	15	20	18	19	
Local Government C	24	21	25	27	
Local Government D	26	24	23		

Determine whether there is any difference in the average class sizes in the different local government areas at the $\alpha = 0.05$ level of significance.

Solution

$H_0: \mu_A = \mu_B = \mu_C = \mu_D$
$H_1: \mu_A \neq \mu_B \neq \mu_C \neq \mu_D$

Table 10.12

LG A	35	25	42	37	15	154
LG B	15	20	18	19		72
LG C	24	21	25	27		97
LG D	26	24	23			73
						396

m = 4 (Factors)

$n = n_A + n_B + n_C + n_D = 5 + 4 + 4 + 3 = 16$

Step 1

$$SS(\text{Factor}) = \frac{\sum (\text{Row total})^2}{\text{number of entries per row}} - \frac{(\text{grand total})^2}{\text{total sample size}}$$

$$= \frac{(154)^2}{5} + \frac{(72)^2}{4} + \frac{(97)^2}{4} + \frac{(73)^2}{3} - \frac{(396)^2}{16}$$

$$= 4743.2 + 1296 + 2352.25 + 1776.33 - 9801$$

$$= 366.78$$

Step 2:

$$SS(\text{Total}) = \sum x_{ij}^2 - \frac{(\text{grand total})^2}{\text{total sample size}}$$

$$= 35^2 + 25^2 + 42^2 + 37^2 + 15^2 + \ldots\ldots + 23^2 - \frac{(396)^2}{16}$$

$$= 10670 - 9801$$

$$= 869$$

Step 3:

SS (Error) = SS (Total) - SS (Factor)

= 869 - 366.78 = 502.22

Step 4: df (Factor) = m – 1 = 4 –1 = 3

Step 5: df (Total) = n – 1 = 16 – 1 = 15

Step 6: df (Error) = df (Total) – df (Factor)

= 15 – 3 = 12

Step 7:

$$\text{MS (Factor)} = \frac{\text{SS (Factor)}}{\text{df (Factor)}}$$

$$= \frac{366.78}{3} = 122.26$$

Step 8:

$$\text{MS (Error)} = \frac{\text{SS (Error)}}{\text{df (Error)}}$$

$$= \frac{502.22}{12} = 41.86$$

Step 9:

$$F_{ratio} = \frac{\text{MS (Factor)}}{\text{MS (Error)}} = \frac{122.26}{41.85}$$

$$= 2.92$$

Step 10:

$F_{tab} = F_{0.05}(3, 12) = 3.49$

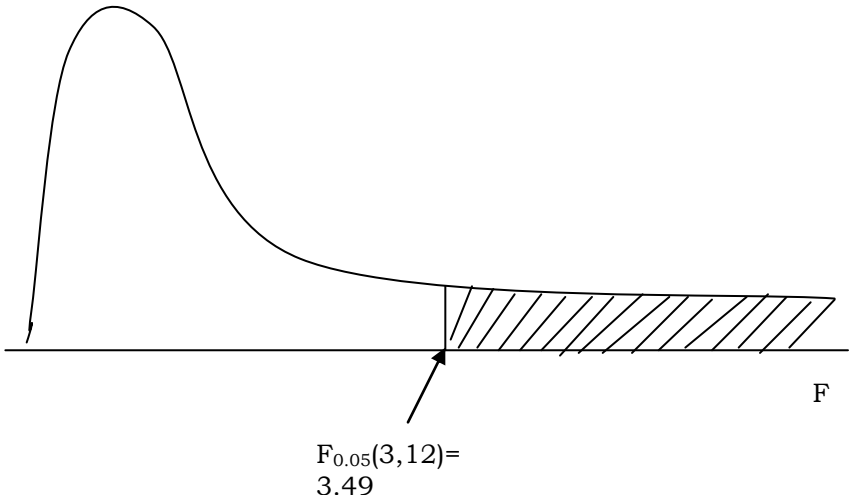


Figure 10.48

It is observed that the F test statistic falls in the non-critical region. So, we fail to reject the null hypothesis at $\alpha = 0.05$ level of significance. Therefore, there is no significant difference in the average class sizes in the different primary schools from the different local government area.

Table 10.13

Source of variation	df	SS	MS	F_{ratio}	F_{tab}
Factor	3	366.78	122.26	2.92	$F_{0.05}(3, 12) = 3.49$
Error	12	502.22	41.86		
Total	15	869			

10.6 EXERCISES

10.6.3 The state environmental protection group is responsible for monitoring the level of pollutants at five local governments in the state to ensure the environment is safe. At each local government, a variety of different readings are taken. The results for a day are as follows:

LG A	15	10	11	12	14	17
LG B	11	12	15	18	17	9
LG C	14	13	10	9	8	6
LG D	22	21	24	21	15	12
LG E	14	11	12	13	9	14

Based on these readings, test if there is any difference among the mean pollutant levels at the five local governments at $\alpha = 0.025$ level of significance.

10.6.4 A stock broker keeps track of the number of stock purchase orders he receives from his clients during a certain week. The results are as shown.

Monday	Tuesday	Wednesday	Thursday	Friday
105	109	120	135	140

Test if there is any difference in the number of orders received based on the day of the week at the $\alpha = 0.025$ level of significance.

10.6.11 A study is conducted asking people in different cities to express their degree of statifaction with the city in which they live. The results are as follows:

City	Unhappy	Somewhat Happy	Very Happy
Lagos	61	105	145
Ibadan	47	125	131
Kano	59	108	153

Onitsha	43	130	142
Abuja	50	112	89

Test whether there is any difference in the proportions of people who express satistication living in each of these five cities at 0.05 significance.

10.6.32 A study was conducted to determine the media credibility for reporting news. Test whether the media credibility and age are independent and give the appropriate p–value for the test.

Age	Newpaper	Radio	Television
Under 35	43	37	25
35 – 54	25	44	30
Over 54	36	38	22

CHAPTER ELEVEN

CORRELATION AND REGRESSION ANALYSIS

In this chapter, we consider situations in which there are two distinct, but possibly related, quantities for each individual. In particular, we consider ways in which we can determine whether the two different quantities are related to each other and, if so, how can we find a relationship and use it in a predictive manner.

Let's look at an example of the scores of some students in an examination. The data is given as

Table 11.1

Students	Exam. I	Exam. II
1	70	50
2	60	50
3	40	40
4	50	60
5	40	80
6	70	70
7	70	40
8	50	50

The graphical display of this table in Figure 11.1 is known as a *scattergram* or *scatterplot*. There seems to be a linear relationship between the two variables. What we need is a way of measuring the degree of relationship or *correlation* between the two quantities. The measure is known as the *correlation coefficient* and is

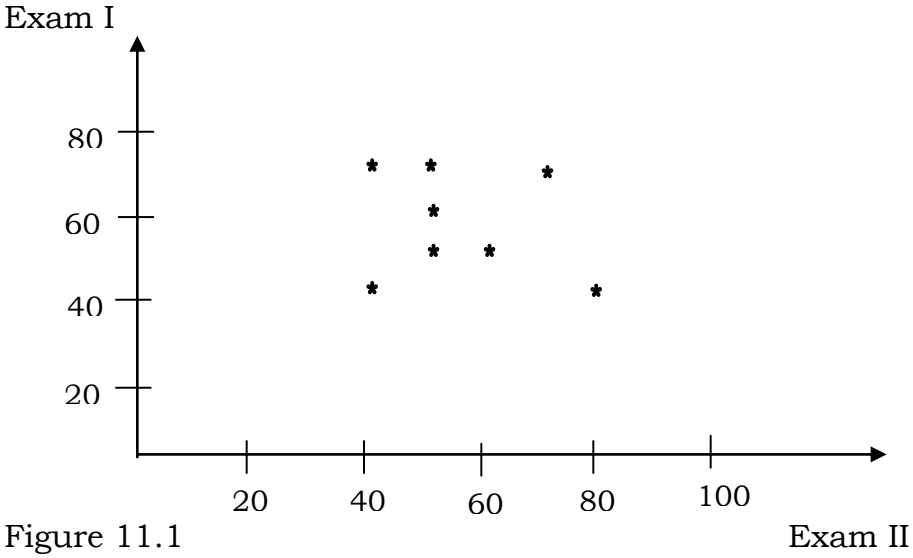
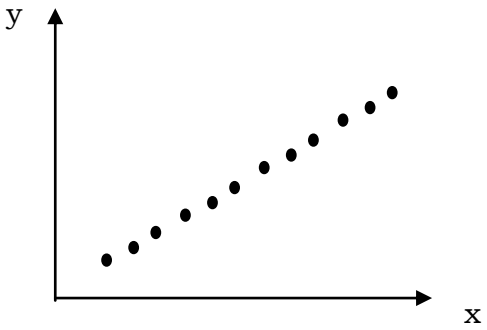


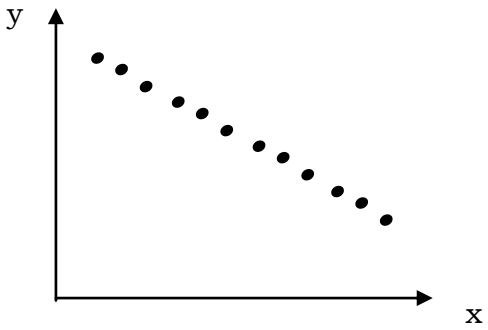
Figure 11.1

denoted by *r*. The value of this quanity is between –1 and 1.

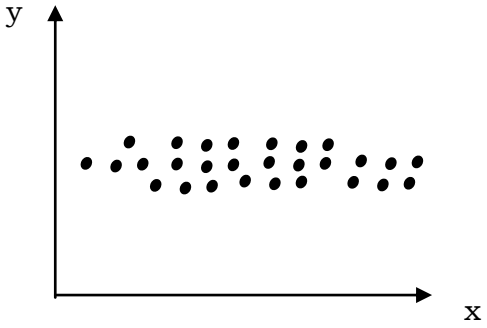
When the value of r is equal to 1, then is a *strong positive correlation* (or *perfect direct*. See Figure 11.2a). On the other hand, if the value of r is equal to –1, then there is a *strong negative correlation* (or *indirect perfect*. See Figure 11.2b). If r is equal to zero, then there is no correlation between the two variables. (See Figure 11.2c). If r is positive but moderately small, then there is a slight positive correlation between the two variables (or direct correlation see Figure 11.2d). Finally, if the correlation coefficient is some what negaitve, there is a slight *negative correlation* (or indirect correlation. See Figure 11.2e).



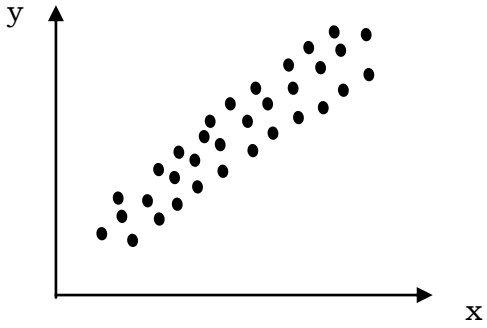
(a) Perfect Direct



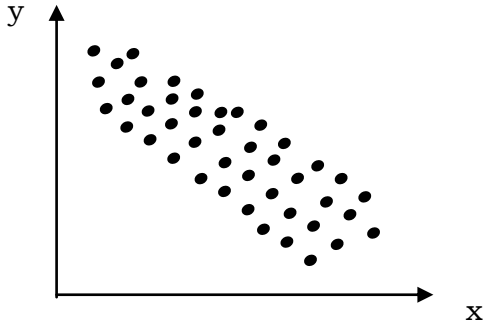
(b) Perfect indirect



(c) Zero correlation



(d) Direct



(e) Indirect

Figure 11.2

When there is high degree of correlation between two variables, then there is a linear relationship between them. However, it is evident that we cannot connect all the points in a scalterplot with a single straight line. So, one finds the particular line which comes closest, in some sense, to all the points in the seatterplot, since several different lines may be drawn which may seem close. The line that is closest to all points is known as the line of *Best fit*. The particular line is known as the *regression line* or the *least squares line*.

11.1 CORRELATION

11.1.1 Pearson's Product Moment

One measure of linear depedency is the *covariance*. The covariance of x and y is defined as the sum of the products of the distances of all values of x and y from the mean values of x and y, divided by n – 1

$$\text{Covariance } (x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{x})}{n - 1}$$

or

$$\frac{n(\sum xy) - (\sum x)(\sum y)}{n - 1}$$

Therefore, the coefficient of linear correlation is

$$r = \frac{\text{Covar } (x, y)}{s_x \times s_y} \qquad (11.2)$$

Where s_x and s_y are the standard deviation of x and y, respectively. This formula is also referred to as *Pearson's Product Moment*, *r*.

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2}\sqrt{\sum (y - \bar{y})^2}} \qquad (11.3)$$

or

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\ \sqrt{n(\sum y^2) - (\sum y)^2}} \qquad (11.4)$$

In this book, we shall use formula 11.4.

Example 11.1: Determine the correlation for the data given in Table 11.1

Solution

Table 11.2

Exam. I (y)	**Exam. II (x)**	**xy**	**x^2**	**y^2**
70	50	3500	2500	4900
60	50	3000	2500	3600
40	40	1600	1600	1600
50	60	3000	3600	2500
40	80	3200	6400	1600
70	70	4900	4900	4900
70	40	2800	1600	4900
50	50	2500	2500	2500
$\sum y$ = 450	**440**	**24500**	**25600**	**26500**

$n(\sum xy) - (\sum x)(\sum y) = 8(24500) - (440)(450)$

$= 196000 - 198000$

$= -2000$

$n(\sum x^2) - (\sum x)^2 = 8(25600) - (440)^2$

$= 204800 - 193600$

$= 11200$

$n(\sum y^2) - (\sum y)^2 = 8(26500) - (450)^2$

$= 212000 - 202500$

$= 9500$

The correlation coefficient r is

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

$$= \frac{-2000}{\sqrt{11200}\ \sqrt{9500}} = \frac{-2000}{105.83 \times 97.47}$$

$$= -0.194$$

It is evident that there is not strong negative correlation between Exam. I and Exam. II.

Example 11.2: Consider the acompanying bivariate data:

	A	B	C	D	E	F	G	H
X	21	26	27	49	20	25	30	25
Y	14	15	13	16	17	18	14	11

a. Draw a scatter diagram for the data
b. Calculate the covariance
c. Calculate s_x and s_y
d. Calculate the correlation between x and y

Solution

a.

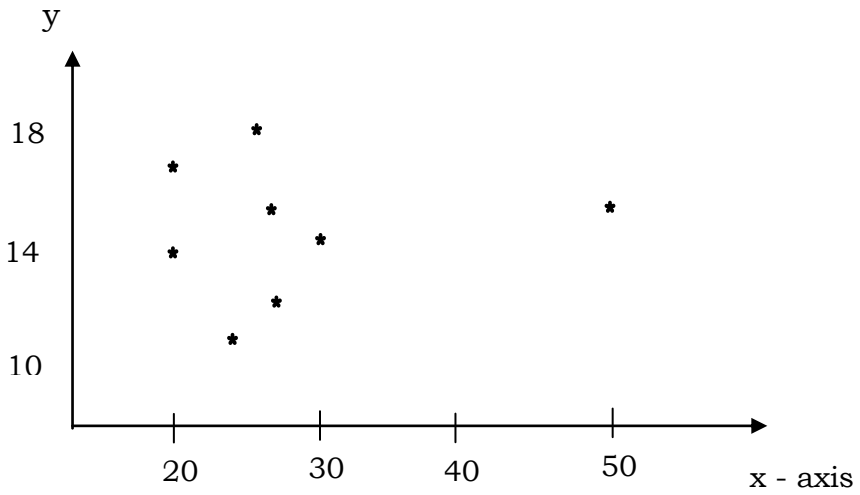


Figure 11.3

b. Covariance (x, y) = $\frac{n\,(\sum xy) - (\sum x)\,(\sum y)}{n - 1}$

Table 11.3

X	Y	XY	X^2	Y^2
21	14	294	441	196
26	15	390	676	225
27	13	351	729	169
49	16	784	2401	256
20	17	340	400	289
25	18	450	625	324
30	14	420	900	196
25	11	275	625	121
$\sum x = 223$ $\sum y = 118$ $\sum xy = 3304$ $\sum x^2 = 6797$ $\sum y^2 = 1776$				

$$\text{Cov}(x, y) = \frac{8(3304) - (223)\,(118)}{8 - 1}$$

$$= \frac{26432 - 26314}{7} = 118/7 = 16.86$$

c. $$s_x = \sqrt{\frac{n(\sum x^2) - (\sum x)^2}{n-1}} = \sqrt{\frac{8(6797) - (223)^2}{8-1}}$$

$$= \sqrt{\frac{54376 - 49729}{7}} = \sqrt{\frac{4647}{7}}$$

$$= \sqrt{663.86} = 25.77$$

$$s_y = \sqrt{\frac{n(\sum y^2) - (\sum y)^2}{n-1}} = \sqrt{\frac{8(1776) - (118)^2}{8-1}}$$

$$= \sqrt{\frac{14208 - 13924}{7}} = \sqrt{\frac{284}{7}}$$

$$= \quad 6.37$$

d. Correlation (r) $= \dfrac{\text{Cov}(x, y)}{s_x \cdot s_y}$

$$= \frac{16.86}{25.77 \times 6.37} = 0.103$$

11.1.2 Spearman's Rank Correlation

Another method for finding the correlation coefficient is the *Spearman's Rank Correlation Coefficient*. This method uses the ranks of the variables instead of the actual values. The formula is given as

$$r_{sp} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \qquad (11.5)$$

where $d = r_x - r_y$ (difference between the rank of x and y) and n the number of pairs of observation in the sample.

Example 11.3: Use Rank correlation method to determine the correlation coefficient of the following data.

X	55	10	70	75	25
Y	60	95	70	35	80

Solution
Table 11.4

X	Y	r_x	r_y	$d = r_x - r_y$	d^2
55	60	3	2	1	1
10	95	1	5	-4	16
70	70	4	3	1	1
75	35	5	1	4	16
25	80	2	4	-2	4
					38

$$r_{sp} = 1 - \frac{6 \sum d^2}{n (n^2 - 1)}$$

$$= 1 - \frac{6 (38)}{5 (5^2 - 1)} = 1 - \frac{228}{120}$$

$$= 1 - 1.9 = - 0.9$$

Thus there is a strong indirect relationship between the two variables x and y.

11.1.3 Hypothesis Test for Correlation

It is necessary to confirm if the linear correlation coefficient, r, calculated indicates that there is a linear dependency between two variables in the population. To confirm this, we perform a hypothesis test. In this hypothesis test, the null hypothesis typically asserts that there is no correlation.

$H_0: \rho = 0$

While the alternate hypothesis states that

$H_1: \rho \neq 0$

So that we have a two-tailed test, where ρ (the lowercase Greek letter rho) is the *linear correlation coefficient for the population*. Most frequently it is two-tailed. However, when we suspect that there is only a positive or only a negative correlation, we should use a one-tailed test. The alternative hypothesis of a one-tailed test is $\rho > 0$ or $\rho < 0$.

The critical region for the test is on the right when a positive correlation is expected and on the left when a negative correlation is expected. The test statistic used to test the null hypothesis is the calculated value of r from the sample. Critical values for r are found in Table VII of the Appendix at the intersection of the

column identified by the appropriate value of α and the row identified by the degrees of freedom. The number of degrees of freedom for the r statistic is 2 less than the sample size, df = n –2. As the number of pairs of points used increases, the critical value for r decreases.

Example 11.4: Use the result of Example 11.3 to determine if there is any correlation between the two variables at $\alpha = 0.1$ significance level.

Solution

Step 1: $H_0: \rho = 0$
$H_1: \rho \neq 0$

Step 2: $\alpha = 0.1$, df = n – 2 = 5 – 2 = 3

We found r = -0.9

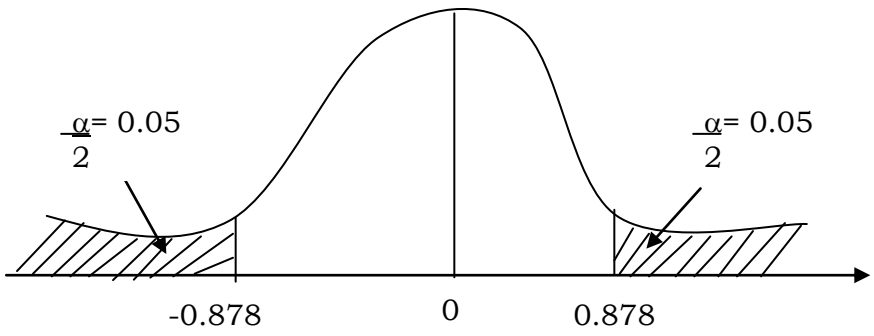


Figure 11.4

The critical values (-0.878 and 0.878) were obtained from Table VII

Step 3: The calculated value of r, r* = -0.9, falls within the critical region.

Stpe 4: We reject H_0. That is, at 0.1 level of significance, variables x and y are correlated.

When we have established that there is a linear relationship between two variables x and y, then the next stop is to attempt to fit a straight line through that set of points provided (fit the line of best fit or regressin line).

11.2 LINEAR REGRESSION ANALYSIS

From the equation of a straight, we have

$$y = mx + c$$

where y is the depedent or predicted variable

m is the gradient (slope) of the line

c is the intercept of the line on y – axis

x is the independent, or input variable

The linear model used to explain the bahaivour of linear bivariate data in the population is

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (11.6)$$

and

$$\hat{y} = b_0 + b_1 x \qquad (11.7)$$

is the regressin line from the sample data.

β_0 is the y – intercept and β_1 is the slope. **ε** (lower case Greek letter epsilon) *is the random experimental error* in the observed value of y at a given value of x.

The *regression line* from the sample data gives us b_0, which is our *estimate of* β_0 and b_1, which is our *estimate of* β_1 . The error (residual) ε is approximated by

$$e = y - \hat{y} \tag{11.8}$$

The difference between the observed value of y and the predicted value of y, $\hat{y}$, at a given value of x.

$$b_0 = \bar{y} - b_1\bar{x} \tag{11.9}$$

where $\bar{x}$ and $\bar{y}$ are the mean of x and y respectively

$$b_1 = \frac{s_{xy}}{s_{xx}} = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \tag{11.10}$$

Example 11.5: A small company is interested in analyzing the effects of advertising on its sales over a 6 month period, it finds the following results

X	6	7	10	12	15	20
Y	10	11	15	18	21	35

Where x represents the money spent on advertising (in thousands of Naira) and y represents the sales (in hundred of thousands of Naira). Determine the linear regression equation for sales as a function of advertising.

Solution

The means of x and y are

$$\bar{x} = \frac{\sum x}{n} = \frac{70}{6} = 11.67 \quad \bar{y} = \frac{\sum y}{n} = \frac{110}{6} = 18.33$$

Table 11.5

X	Y	XY	X^2	Y^2
6	10	60	36	100
7	11	77	49	121
10	15	150	100	225
12	18	216	144	324
15	21	315	225	441
20	35	700	400	1225
$\sum x = 70$	**$\sum y = 110$**	**$\sum xy = 1518$**	**$\sum x^2 = 954$**	**$\sum y^2 = 2436$**

The Slope

$$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$= \frac{6 \times 1518 - (70) \times (110)}{6 \times 954 - (70)^2}$$

$$= \frac{9108 - 7700}{5724 - 4900} = \frac{1408}{824} = 1.71$$

$b_0 = \bar{y} - b_1 \bar{x} = 18.33 - (1.71)(11.67)$

$= 18.33 - 19.96 = -1.63$

Therefore, the equation of the regression line for this set of data is

$$\hat{y} = -1.63 + 1.71x$$

When a numerical value is substituted for x in the regression or *forecast equation* $\hat{y}$, the resulting value of $\hat{y}$ is the *forecast* value of y for that given value of x. For example, when x = 25, we have

$\hat{y} = -1.63 + (1.71) \times (25) = -1.63 + 42.75$

$= 41.12$

11.3 THE STANDARD ERROR OF THE ESTIMATE

Before we can make any inferences about a regression line, we must be sure that the distributions of y are approximately normal and that the variances of the distributions of y at all values of x are the same.

The *variance of error e* is given as

$$s_e^2 = \frac{\sum(y - \hat{y})^2}{n - 2} \qquad (11.11)$$

where n – 2 is the number of degrees of freedom.

Similarly,

$$s_e^2 = \frac{\sum(y - b_0 - b_1x)^2}{n - 2} \qquad (11.12)$$

Since $\hat{y} = b_0 + b_1x$

Formular (11.12) can be rewritten as

$$s_e^2 = \frac{\sum(y^2) - (b_0)(\sum y) - (b_1)(\sum xy)}{n - 2} \qquad (11.13)$$

This quantities (Formular 11.11 – 11.13) is known as the *standard error of the estimate* or the sum of squares for error (SSE).

Example 11.6: In a regression problem where x is the number of days and y is the number of kilogram lost, we obtain the following:

$n = 5$ $\quad \sum xy = 140$
$\sum x = 21$ $\quad \sum x^2 = 132$
$\sum y = 25$ $\quad \sum y^2 = 158$

i. Find the equation of the line of best fit.
ii. Find s_e

Solution

i. $b_1 = \frac{(5)(140) - (21)(25)}{5(132) - (21)^2}$

$= \frac{700 - 525}{660 - 441} = \frac{175}{219} = 0.799$

$\bar{x} = \frac{\sum x}{n} = \frac{21}{5} = 4.2$

$\bar{y} = \frac{\sum y}{n} = \frac{25}{5} = 5$

$b_0 = \bar{y} - b_1\bar{x} = 5 - (0.799)(4.2)$

$= 5 - 3.356 = 1.644$

Therefore, the line of best fit is given as

$\hat{y} = 1.644 + 0.799x$

$$s_e = \sqrt{\frac{\sum y^2 - (b_o)(\sum y) - (b_1)(\sum xy)}{n-2}}$$

$$= \sqrt{\frac{158 - (1.644)(25) - (0.799)(140)}{5-2}}$$

$$= \sqrt{\frac{158 - 41.1 - 111.86}{3}} = \sqrt{\frac{5.04}{3}}$$

$$= \sqrt{1.68} = 1.3$$

The variance of y about the line of best fit is the same as the variance of the error e.

11.4 HYPOTHESIS TESTING CONCERNING THE SLOPE OF THE REGRESSION LINE

It is not just enough to accept the equation of the line of best fit. We need to determine whether we can use the equation to predict y. That is done by testing the null hypothesis.

$H_0 : \beta_1 = 0$ against
$H_1 : \beta_1 \neq 0$

That is, β_1 (the slope of the relationship in the population) is zero. If $\beta_1 = 0$, then the linear equation will be of no real use in predicting y.

Usually the criterion on n for normality is expressed as $n - 2 > 30$ but for small samples where $n < 32$, we use a t-test with $n - 2$ degrees of freedom provided that the underlying distribution is approximately normal.

$$\sigma^2_{b_1} = \frac{\sigma_\varepsilon^2}{\sum (x - \bar{x})^2} = \frac{\sigma_\varepsilon^2}{n\sum x^2 - (\sum x)^2} \qquad (11.14)$$

An appropriate estimator for $\sigma^2_{b_1}$ is obtained by replacing σ_ε^2 by s_e^2, the estimate of the variance of the error about the regression line is

$$s^2_{b_1} = \frac{s_e^2}{\sum (x - \bar{x})^2} = \frac{s_e^2}{n\sum x^2 - (\sum x)^2} \qquad (11.15)$$

The test statistic we use in the hypothesis test is the value of b_1 for the slope of the regression line based on the sample data. That is,

$$t = \frac{b_1 - \beta_1}{S_{b_1}} \qquad (11.16)$$

Example 11.7: Does the sample of Example 10.5 present sufficient evidence at 5% level of significance to reject the null hypothesis (slope is zero) in favour of the alternative hypothesis that the slope is positive?

Solution

Step 1: $H_0 : \beta_1 = 0$

The alternative hypothesis can be either one-tailed or two-tailed. If we suspect that the slope is positive (that, we would expect sales, y, to increase as advertisement, x, increased), a one-tailed test is appropriate.

$H_1: \beta_1 > 0$

Step 2: The test statistic is t. The number of degrees of freedom is n – 2, df = 6- 2 = 4. The critical value of t at 0.05 level of significance is

$t_\alpha\,(df) = t_{0.05}\,(4) = 2.132$

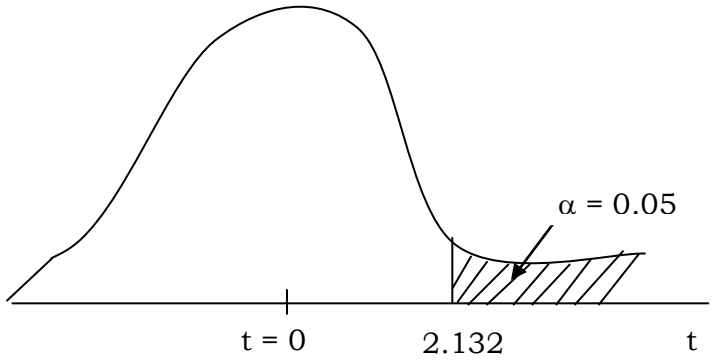


Figure 11.5

Step 3:

$$S_e = = \sqrt{\frac{\sum y^2 - (b_o)(\sum y) - (b_1)(\sum xy)}{n-2}}$$

$$= \sqrt{\frac{2436 - (1.63)(110) - (1.71)(1518)}{6-2}}$$

$$= \sqrt{\frac{158 - 179.3 - 2595.78}{4}} = \sqrt{\frac{19.52}{4}}$$

$$= \sqrt{4.88} = 2.209$$

$$s_{b_1} = \frac{S_e}{\sqrt{n\sum x^2 - (\sum x)^2}} = \frac{2.209}{\sqrt{6(954) - (70)^2}}$$

$$= \frac{2.209}{\sqrt{5724 - 4900}} = \frac{2.209}{\sqrt{824}}$$

$$= \frac{2.209}{28.705} = 0.077$$

The test statistic

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{1.71 - 0}{0.077}$$

$$= 22.21$$

$$t^* = 22.21$$

The test statistic falls within the critical region.

Step 4: Reject H_0. At 5% significance level, we conclude that the slope of the line of best fit in the population is greater than zero. The evidence indicates that there is a linear relationship between sales (y) and advertisement (x).

The slope β_1 of the regression line of the population can be estimated by means of a *confidence interval*. The confidence interval is given by

$$b_1 \pm t_{\alpha/2}\,(n-2).\, s_{b_1} \qquad (11.17)$$

Therefore, the 95% confidence interval of the population's slope, β_1 for Example 11.5 is

$$1.71 \pm t_{0.025}(4) \times 0.077$$

$$= 1.71 \pm 2.776 \times 0.077 = 1.71 \pm 0.21375$$

$$= [1.49625, 1.92375]$$

$\cong$ 1.50 to 1.92 at 95% confidence interval for β_1

11.5 ESTIMATING CONFIDENCE INTERVAL FOR REGRESSION

The next thing to do when the line of best fit is appropriate is to use the equation to make prediction. There are two different quantities that we can estimate.

Case 1: the mean of the population y values at a given value of x, written μ_{y/x_o}

Case 2: the individual y value selected at random that will occur at a given value of x, written y_{x_o}.

The best *point estimate*, or prediction, for both μ_{y/x_o} and y_{x_o} is $\hat{y}$. The confidence intervals for μ_{y/x_o} and y_{x_o} are constructed in the same way as previous cases.

The confidence interval estimate of the *mean value of y* at a given value of x, μ_{y/x_o} is

$$\hat{y} \pm t_{\alpha/2}\,(n-2).\, s_e \,.\, \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x - \bar{x})^2}} \qquad (11.18)$$

or

$$\hat{y} \pm t_{\alpha/2}\,(n-2).\, s_e \,.\, \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{n\sum x^2 - (\sum x)^2}} \qquad (11.19)$$

Example 11.8: Construct a 95% confidence interval for the mean sale of the company in Example 11.5 if it spends ₦40,000 in advertising.

Solution

Find $\hat{y}_{x_o}$, when $x_0 = 40$

$$\hat{y} = -163 + 1.71x = -1.63 + (1.71)(40)$$
$$= -1.63 + 68.4 = 66.77$$

Find s_e

$$s_e = 2.209 \text{ (from example 11.7)}$$

The critical value is

$$t_{\alpha/2}(n-2) = t_{0.025}(4) = 2.132$$

95% confidence interval

$$= 66.77 \pm (2.132)(2.209).\ \sqrt{\frac{1}{6} + \frac{(40-11.67)^2}{824}}$$

$$= 66.77 \pm 4.7096 \text{ x } \sqrt{\frac{1}{6} + \frac{802.589}{824}}$$

$$= 66.77 \pm 4.7096 \text{ x } \sqrt{0.1667 + 0.974}$$

$$= 66.77 \pm 4.7096 \text{ x } \sqrt{1.14067}$$

$$= 66.77 \pm 4.7096 \text{ x } 1.06802$$

$$= 66.77 \pm 5.03 = [61.74,\ 71.8]$$

Therefore, 61.74 to 71.8, 95% confidence interval for $\mu_{y/x_o} = 40$.

The formula for the interval estimate of the value of a *single randomly* selected y is

$$\hat{y} \pm t_{\alpha/2}(n-2).\ s_e.\sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{n\sum x^2 - (\sum x)^2}} \qquad (11.20)$$

Example 11.9: Determine a 95% confidence interval for the sale of the company if the expenses on advertisement, x = 40

Solution

$x_0 = 40$
$\hat{y} = 66.77$
$s_e = 2.209$
$t_{0.025}(4) = 2.132$

The 95% confidence interval for the single value

$$= 66.77 \pm (2.132)(2.209).\ \sqrt{1+\frac{1}{6}+\frac{(40-11.67)^2}{824}}$$

$$= 66.77 \pm (4.7096) \times \sqrt{1 + 1.14067}$$

$$= 66.77 \pm (4.7096) \times \sqrt{2.14067}$$

$$= 66.77 \pm (4.7096) \times (1.46310)$$

$$= 66.77 \pm 6.891 = [59.88,\ 73.66]$$

Note that the confidence interval for the single point is much longer than the confidence interval for $\mu_{y/x_o} = 40$

11.6 CURVILINEAR REGRESSION

One extension of the regression we have considered is to introduce the idea of *non-linear regression*. Instead of drawing a line we are used to, we draw a *regression curve* which best fits the data set. When the scatter plot shows a non-linear relationship them any of the following models may be adopted.

i. $y = Ax^2 + Bx + C$ (parabola or Quadratic curve)

ii. $y = Ax^3 + Bx^2 + Bx + C$ (Cubic curve)

iii. $y = Ax^4 + Bx^3 + Cx^2 + Dx + E$ (Quartic curve)

iv. In general,
$y = A_1x^n + A_2x^{n-1} + \ldots\ldots + A_n$ (n^{th} Degree curve)

v. $y = \frac{1}{Ax + B}$ or $\frac{1}{y} = Ax + B$ (Hyperbola curve)

vi. $y = AB^x$ or $\log Y = \log A + x \log B$ (Exponential curve)

vii $y = Ax^B$ or $\log y = \log A + B \log x$ (Geometric curve)

viii. $y = AB^x + G$ (Modified exponential curve)

ix. $y = Ax^B + G$ (Modified geometric curve)

x. $y = \frac{1}{AB^x + G}$ or $\frac{1}{y} = AB^x + G$ (Logistic curve)

Multivariate linear regression analysis or simply multivariate regression analysis involves considering a single variable, y, which depends on a set of n different variables, often denoted by x_1, x_2, …, x_n. It is also possible to measure the degree of correlation between the quantity y and all the individual variables. If the y and the x's variables are correlated and the relationship between them is linear, then we relate y to the other variables by using a linear equation of the form.

$$y = A_1x_1 + A_2x_2 + \ldots\ldots + A_nX_n + B$$

This equation could be used as a predictor after all the coefficients, A_1, A_2, … A_n and B must have been estimated

11.7 EXERCISES

11.7.1 Draw the scatter plot and calculate the correlation coefficient for the following sets of data.

i.

x	10	15	17	25	35
y	70	60	50	40	31

ii.

x	1	3	7	10	15	17
y	10	15	18	25	31	39

iii.

x	2	8	15	17	28	38
y	0.5	0.3	0.1	-0.4	-0.7	-0.9

11.7.5 Draw the scatter plot and calculate the regression equation for the following sets of data:

i.

x	10	15	20	25	30
y	100	90	75	45	21

ii.

x	6	10	18	25	36	45
y	3.5	6.1	8.7	10.5	15.8	20.7

END OF MODULE ASSESEMENT

11.7.11 The following set of scores was randomly selected from a teacher's class list. Let x be the test score and y the final examination.

Students	X	Y
1	25	45
2	35	50
3	12	21
4	22	40
5	15	28
6	17	37
7	10	40
8	15	49
9	18	29
10	30	50
11	35	45
12	31	50
13	16	28
14	5	22
15	13	33

i. Draw a scatter diagram for these data.

ii. Draw a regression line (by eye) and its equation.

iii. Calculate the equation of the line of best fit.

iv. Estimate the value of the coefficient of linear correlation.

v. Draw the line of best fit on your graph.
How does it compare with your estimate?

vi. Calculate the linear correlation coefficient.
How does it compare with your estimate?

vii. Test the significance of r at $\alpha = 0.10$.

viii. Find a 95% confidence interval estimate for the true value of ρ.

ix. Find the standard deviation of the y values about the regression line.

x. Calculate a 95% confidence interval estimate for the true value of the slope, β_1.

xi. Test the significance of the slope at $\alpha = 0.05$.

xii. Using 95% confidence interval, estimate the mean final Examination that all students with 35 test score will obtain.

xiii. Using the 95% confidence interval, predict the final examination score of a student, knowing that his test score is 21.

10.6.31 A random sample of 100 women who were tested for cholesterol were classified according to age and cholesterol level and grouped in the following contignecy table.

	Cholesterol level		
Age	<180	180 – 210	>210
< 50	8	6	18
≥ 50	10	22	36

Test the null hypothesis H_0: age and cholesterol level are independent attributes of classification. What is your conlcusion if $\alpha = 0.05$?

10.6.28 Construct an ANOVA table and state your conclusion using the following data at $\alpha = 0.01$ significance level.

X_1:	21	35	37	40	33	34	35	41

X_2:	11	31	40	41	33	30	15	21
X_3:	15	17	23	43	15	41	43	44
X_4:	11	18	33	23	20	33	41	41
X_5:	24	20	41	24	30	35	43	42