THE UNIVERSITY OF ZAMBIA

INSTITUTE OF DISTANCE EDUCATION

INTRODUCTION TO STATISTICS:

MAT 2602 MODULE

© Copyright

Acknowledgements

The Institute of Distance Education of the University of Zambia wishes to thank Mr Urban N. Haankuku for writing this module **MAT 2602 Introduction to Statistics**

Table of Contents

2. Introduction to Statistics; (2016)., David Lane, Rice University, Publisher: Saylor Foundation **255**

About this introduction to probability Module

The introduction to statistics has been produced by The University of Zambia, Institute of Distance education.

How this statistics Module is structured

The course overview

The course overview gives you a general introduction to the course. Information contained in the course overview will help you determine:

- If the course is suitable for you.
- What you will already need to know.
- What you can expect from the course.
- How much time you will need to invest to complete the course.

The overview also provides guidance on:

Study skills.

Where to get help.

Course assignments and assessments.

Activity icons.

Units.

We strongly recommend that you read the overview *carefully* before starting your study.

The course content

The course is broken down into units. Each unit comprises:

- An introduction to the unit content.
- Unit outcomes.
- New terminology.
- Core content of the unit with a variety of learning activities.
- A unit summary.
- Exercises and Answers to the exercises, as applicable.

Resources

For those interested in learning more on this subject, we provide you with a list of additional resources at the end of this statistics Module; these may be books, calculator, articles or web sites.

Your comments

After completing this module, we would appreciate it if you would take a few moments to give us your feedback on any aspect of this course. Your feedback might include comments on:

- Module content and structure.
- Module reading materials and resources.
- Module unit Activities

- Module Answers.
- Module duration.
- Module support (assigned tutors, technical help, etc.)

Your constructive feedback will help us to improve and enhance this course.

Module overview

Welcome to Introduction to statistics Module.

This Module is designed to give students background knowledge/ information of statistics concepts, knowledge and skills in statistics. The module gives a background of what is data and how you would manipulate data into a more meaningful information that can be used for decision making and planning. You will learn preliminary analysis under descriptive statistics, data presentations for both discrete and continuous data, estimation theory, statistical inference and linear regression analysis.

The module has a number of unit activities at the end of each unit for students to practice. Answers to these activities are provided to give students feedback. Statistics is a practical subject so you are advised to try all activities before you come for residential schools. During the residential schools you are expected to come for consultations on areas you found concepts difficult to understand.

Statistics Module —is this course for you?

This course is intended for people who are studying Mathematics for their Bachelor's degree in Mathematics Education, Demography and Eonomics.

Pre requisite: In order for you to study this module well you should have done Module MAT 1100 (Foundation Mathematics).

Course outcomes

By the end of this Module, you should be able to you will be able to:

- Make a distinction between discrete random variables and continuous random variables.
- present data in graphical methods
- Carry out basic analysis using descriptive statistics

- Carry out statistical inference using estimation theory, hypotheses testing methods and linear regression analysis

Timeframe

It is recommended that you take 120 Hours to complete this module

Study skills

As an adult learner your approach to learning will be different from the way you used to learn in your school days: you will choose what you want to study, you will have professional and/or personal motivation for doing so and you will most likely be fitting your study activities around other professional or domestic responsibilities.

Essentially you will be taking control of your learning environment. As a consequence, you will need to consider performance issues related to time management, goal setting, stress management, etc. Perhaps you will also need to reacquaint yourself with areas such as essay planning, coping with exams and using the web as a learning resource.

Your most significant considerations will be *time* and *space* i.e. the time you dedicate to your learning and the environment in which you engage in that learning.

We recommend that you take time now—before starting your self-study—to familiarize yourself with these issues. There are a number of excellent resources on the web. A few suggested links are:

- http://www.how-to-study.com/

  The "How to study" web site is dedicated to study skills resources. You will find links to study preparation (a list of nine essentials for a good study place), taking notes, strategies for reading text books, using reference sources, test anxiety

- http://www.ucc.vt.edu/stdysk/stdyhlp.html

  This is the web site of the Virginia Tech, Division of Student Affairs. You will find links to time scheduling (including a "where does time go?" link), a study skill checklist, basic

concentration techniques, control of the study environment, note taking, how to read essays for analysis, and memory skills ("remembering").

- http://www.howtostudy.org/resources.php

Another "How to study" web site with useful links to time management, efficient reading, questioning/listening/observing skills, getting the most out of doing ("hands-on" learning), memory building, tips for staying motivated and developing a learning plan.

The above links are our suggestions to start you on your way. At the time of writing these web links were active. If you want to look for more go to www.google.com and type "self-study basics", "self-study tips", "self-study skills" or similar search form.

Need help?

When you need help in this module you can contact the Director, UNZA IDE or visit **www.mathstutor.com**

Assessments

1.	Continuous Assessment	30%
1.1	Assignments/Quizzes	10%
1.2	Tests	20%
2.	Final Examination	70%
Total		100%

Getting around this Statistics Module.

Margin icons

While working through this module you will notice the frequent use of margin icons. These icons serve to “signpost” a particular piece of text, a new task or change in activity. They have been included to help you to find your way around.

A complete icon set is shown below. We suggest that you familiarize yourself with the icons and their meaning before starting your study.

 Activity	 Assessment	 Assignment	 Case study
 Discussion	 Group activity	 Help	 Note it!
 Outcomes	 Reading	 Reflection	 Study skills
 Summary	 Terminology	 Time	 Tip

UNIT 1 STATISTICS

1.1 Unit Introduction

Welcome to Unit 1 of this Module in which you will learn that Statistics is a science that deals with the methods of data: Collections, Compilation, Presentation, analysis, and Interpretation of results, Conclusion and above all making decisions based on the study of the data. Why statistics? This field of study is used to study experiments and another natural phenomenon, to solve a particular problem or find explanations of why specific events happen.

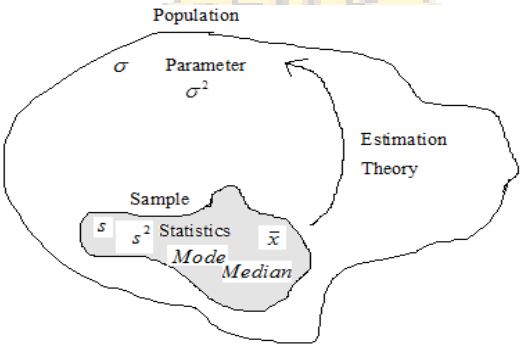


Figure 1.1 Pictorial picture of research

Figure 1.1 an illustration of inference from sample statistics to population parameters. Statistical analysis is critical in Epidemiology, Agriculture, Engineering, Environment, Education, Finance, economics etc. Figure1. 1 illustrates the schematic diagram of what is involved in statistics.

1.2 Unit Aim

The aim of this course is to give you concepts and skills of statistics.

1.3 Unit Objectives

By the end of the unit you should be able to:

- Calculate required statistics on any given data,
- Distinguish between discrete and continuous data
- Calculate statistics from any given data

- Calculate measures of central tendencies and measures of dispersion of data

Terminology

- $\bar{x}$ - The mean or average
- $\sum$ - Summation
- σ - Standard deviation
- f - Frequency

1.4 Unit Time required

You need 30 hours for this unit

1.5 Unit Topics

1.5.1 Types of Data

There are many examples of types of data that exist and table 1 gives examples of types of data which gives basic distinction between quantitative variables (for which one asks "how much?") and categorical variables (for which one asks "what type?"). The focus of this unit is on multivariate analysis of discrete data. Here you will deal with data which are discretely measured responses such as counts, proportions, nominal variables, ordinal variables, discrete interval variables with few values, continuous variables grouped into a small number of categories, etc. You will learn basic statistical methods and discuss issues relevant for the analysis of some discrete distribution.

Quantitative variables can be continuous or discrete. Examples of continuous variables, such as height, weight, this is the type of data that can in theory take any value within a given range. Examples of discrete variables are: number of children in a family, number of attacks of asthma per week.

Categorical variables are either nominal (unordered) or ordinal (ordered). Examples of nominal variables are male/female, alive/dead, blood group O, A, B, AB. For nominal variables with more

than two categories the order does not matter. For example, one cannot say that people in blood group B lie between those in A and those in AB. Sometimes, however, people can provide ordered responses, such as grade of breast cancer, or they can "agree", "neither agree nor disagree", or "disagree" with some statement. In this case the order does matter and it is usually important to account for it.

Table 1.1 Examples of types of data

Examples of types of data	
Quantitative	
Continuous	**Discrete**
Blood pressure, height, weight, age	Number of children Number of attacks of asthma per week
Categorical	
Ordinal (Ordered categories)	**Nominal (Unordered categories)**
Grade of breast cancer Better, same, worse Disagree, neutral, agree	Sex (male/female) Alive or dead Blood group O, A, B, AB

Variables shown at the left of Table 1.1 can be converted to one's further to the right by using "cut off points". For example, blood pressure can be turned into a nominal variable by defining "hypertension" as a diastolic blood pressure greater than 90 mmHg, and "norm tension" as blood pressure less than or equal to 90 mmHg. Height (continuous) can be converted into "short", average" or "tall" (ordinal). Data can be classified into two major types. These are discrete Data and continuous Data. In the study of statistics, your interests are to study the several methods of statistics of which the ultimate goal is to estimate the population parameters of interests.

1.5.2 Frequency Distribution

The frequency distribution is a statistical table which shows the value of a variable in order of magnitude, either individually or in class interval, along with the corresponding frequencies side by side. The data pertaining to a quantitative phenomenon can be classified in four ways:

- The set or series of individual observations- ungrouped (raw) or grouped (arrayed) data.
- Discrete frequency distribution.
- Continuous frequency distribution.

The data given in Example 1.1 are called the raw or ungrouped data which does not give us any useful information. Your objective is to express the huge data in a suitable condensed form which will highlight the significant facts and comparisons and furnish more useful information without sacrificing any information of interest about the important characteristics of the distribution. A better presentation of above raw data would be to arrange them in an ascending or descending order of magnitude which is called arraying of data. However, this method is better than raw data but does not reduce the volume of the data. This is the distribution of the ungrouped data.

Choice of class interval

It is important that the class interval is chosen so that the histogram gives a clear representation of the data. Choice of too large a class interval will lead to loss of detail. On the other hand too small a class interval would destroy the original point of grouping. A convenient rule of thumb is to choose a class interval so that the average frequency is about 5. That is number of classes $\cong \frac{1}{5}(total\ frequency)$. If the values given are many and the difference between the smallest and largest is greater than 10, you can group the data into suitable class intervals. Class intervals are discussed in elementary mathematics as : (a, b) as open interval, [a, b] as closed interval, (a, b) as open-closed interval and [a, b) as closed - open interval. Continuous data uses [a, b) closed-open intervals.

Example 1.1: Consider the concentration of an antibody in blood serum from 50 donors, given below:

Table 1.2 Concentration of antibody in blood serum

11.6	15.8	17.0	14.2	16.2	17.3	15.7	17.3	12.0	11.5
8.4	13.2	12.5	15.7	14.0	10.8	9.9	14.0	15.0	10.2
19.2	12.0	8.2	13.8	12.6	9.6	16.0	13.7	11.5	14.8
17.5	15.2	7.2	17.0	5.2	9.4	7.5	15.5	10.7	14.2
15.3	7.8	11.2	10.0	12.1	10.5	12.2	8.6	9.9	16.1

Tally marks are used in order to track the individual values of the data set. Using the class interval of 4.0 – 5.9, etc as discrete group class interval this is shown in column 1 and 3.95-5.95, 5.95-7.95 etc as grouped continuous class interval shown as true class limits.

Table 1.3 Grouped frequency table of concentration of an antibody in blood serum

Class interval (g/	tally marks	frequency (f)	Mid class value(g/l)	Trueclass limits
4.0 - 5.9	|	1	4.95	3.95 - 5.95
6.0 - 7.9	|||	3	6.95	5.95 - 7.95
8.0 - 9.9	𝍸 ||	7	8.95	7.95 - 9.95
10.0 - 11.9	𝍸 ||||	9	10.95	9.95 - 11.95
12.0 - 13.9	𝍸 ||||	9	12.95	11.95 - 13.95
14.0-15.9	𝍸 𝍸 ||	12	14.95	13.95 - 15.95
16.0 - 17.9	𝍸 |||	8	16.95	15.95 - 17.95
18.0 - 19.9	|	1	18.95	17.9 - 19.95
		50		

The table above gives the first column as discrete grouped data, and this can be used for constructing bar chart, while the second column is the continuous grouped data, and this can be used for constructing a histogram, in which a frequency polygon can be constructed by joining the mid points. The midpoint value calculated as (upper lines + low)/2 of column 1 and these mid points are used to calculate the statistics from the frequency distributions. What this mean is that for Example 1.1 above is represented by 8.95. If you have 5.96 do not belong to the first interval or second interval in the first column it is discrete. But using True class limit it belongs to the second interval.

While preparing the frequency distribution the following points must be kept in mind:

1. The class interval should be uniform i.e. it should be of equal width. A comparison of different frequency distributions is facilitated if the same class interval is used for all. The class interval should be an integer as far as possible.
2. The class interval should be so chosen that all the observations should be reflected by the frequency distribution.
3. The class interval should be continuous open-end. Classes less than a or greater than b should be avoided. These classes create difficulty in analysis and interpretation.
5. There should not be too many or too small number of classes. The number of classes should never be less than 6 and not more than 30 i.e. the number of class intervals should lie between 6 and 30. With less number of classes, the accuracy may be lost, and with more number of classes the computations become tedious. The optimum number of classes is generally considered as 10.

Advantages of grouping

(i) First advantage of grouping is that in subsequent calculations, much labour is saved in numerical computation by treating all individuals in a class interval as having the value at the centre of that interval.

(ii) The second advantage of grouping is where the observed sample is of moderate size and from a large population. In such a case the frequency table is more likely to exhibit a rise or fall of frequency against class interval.

1.1 UNIT ACTIVITY

1. Which of the following observations are discrete and which are continuous?

(a) The throws on a die (b) The height of a person

(c) The length of a train journey (d) the weight of an egg

(e) the age of a car

2. For the following classes, give the true class limits and mid-class values.

(a) Length in mm (measured to nearest 0.1 mm) 20.0 - 20.4, 20.5 - 20.9, 21.0 - 31.4, 21.0 - 21.9

(b) Age in years 1, 2, 3, 4, 5

(c) Weight in Kg (to nearest 1Kg) 58, 59, 60, 61, 62, 63

(d) Height in CM (to nearest 1 cm) 140, 150, 160, 170 and less than 180

3. For the table 1.4 below: give the true class limits, mid-class values, illustrate by a Histogram and a cumulative frequency Curve.

Table 1.4 Frequency distribution of heights

Height in cm(nearest cm)	Frequency
120 – 129	2
130 – 139	12
140 – 149	17
150 – 159	18
160 – 160	7
170 - 179	1

4. For the table given below, gives the population of Zambia analysed by sex and age at 30 December 2017. Compare the figures for males and females graphically

5. The lives of 50 electric lamps in hours to the nearest hour, are given below: Form grouped frequent table and illustrate by (a) a Histogram (b) a cumulative frequency curve.

724	695	716	730	689	700	689	726	662	681	676	732
676	697	710	694	715	738	696	696	682	699	714	707
697	710	660	703	717	692	698	684	695	682	721	708

722 692 717 656 697 701 699 705 680 702 690 663
695 670

6. The numbers of pupils on the register of grant aided schools in Zambia are given below. Show this information on a histogram.

Table 1.5 Frequency distribution of registered pupils of grant aided school.

Age (years)	Number ('000)
2-4	197
5-10	3677
11	596
12	572
13	595
14	172
15-18	276

7. The following data pertain to first lactation milk yield (in kg) of 100 Daily cows

Table 1.6 Frequency distribution of first lactation milk yield (in kg) of 100 Daily cows

1630	1648	1663	1665	1671	1677	1680	1687	1690
1787	1788	1790	1800	1862	1855	1815	1835	1845
1974	1998	2000	2000	2005	2031	2045	2045	2050
2168	2171	2180	2187	2200	2218	2245	2323	2372
2063	2069	2085	2098	2100	2100	2100	2105	2117
1736	1743	1760	1765	1763	1767	1775	1775	1776
1695	1754	1698	1700	1742	1732	1711	1713	1718
1854	1850	1855	1856	1857	1860	1863	1863	1875
1890	1900	1910	1912	1915	1918	1928	1916	1915
1950	1958	1951	1960	1963	1968	1965	1967	1970

The data given in example 1 are called the raw or ungrouped data which does not give us any useful information. Our objective will be to express the huge data in a suitable condensed form which will highlight the significant facts and comparisons and furnish more useful information without sacrificing any information of interest about the important characteristics of the distribution.

(a) Construct a frequency table

(b) Present the given data as a histogram

(c) Present the given data as a frequency polygon

1.5.3 Descriptive Statistics

Descriptive Statistics is the preliminary stage of analysis of statistical data. This involves the calculations of values such as the median, mode, mean, variance, range etc. These values are categorized as: measurers of central tendencies and measures of dispersion. In calculating these measurers, we need to consider whether the given data is in discrete, ungrouped, discrete grouped or continuous ungrouped or continuous grouped. Formulas adhere to the arrangement of the type of data.

If you have a set of observations, it is useful to have a single value which represents all the observations, and this value is called the central tendency and also you should have an idea of how these values are spread called measures of dispersion.

Median: this is a middle value of the observations when the observations are arranged in order.

Note that if the number of observation is an odd number, then the median is one of the observations, but if the number of observations is even, the median is not one of the observations.

Example 1.2 Find the median for the observations: 2, 2, 3, 3, 4, 4, 5, 6, 7, n = 9 observations

(a) 2, 2, 3, 3, 4, 4, 5, 6, 7

The median is one of the observations since n = 9 (odd).

Therefore, the median = 4 Median

(b) 2, 2, 3, 4, 5, 7, 8, 9, 11, 13

The median is between 5 and 7 and it is not one of the observations since n = 10 (even)

Therefore; the median is $\frac{5+7}{2} = 6$

In general the median, M, of n observations is defined as follows: If the observations are arranged in order according to size, then, for odd values of n, the median is the ½ $(n+1)^{th}$ observation; for even values of n, the median is half the sum of the ½ ^{nth} and $(½ n + 1)^{th}$ observations.

Median from a frequency table for a discrete variable.

If the data are grouped the median can be easily found from the cumulative frequency table. Table 1.4 gives the scores obtained for 50 throws of a die. The 25^{th} value and the 26th value are both 3, giving a median of 3.

Table 1.4 Cumulative frequency table for 30 throws of a die.

Score	frequency	cumulative frequency
1	7	7
2	15	22
3	10	32
4	3	35
5	9	40
6	6	50

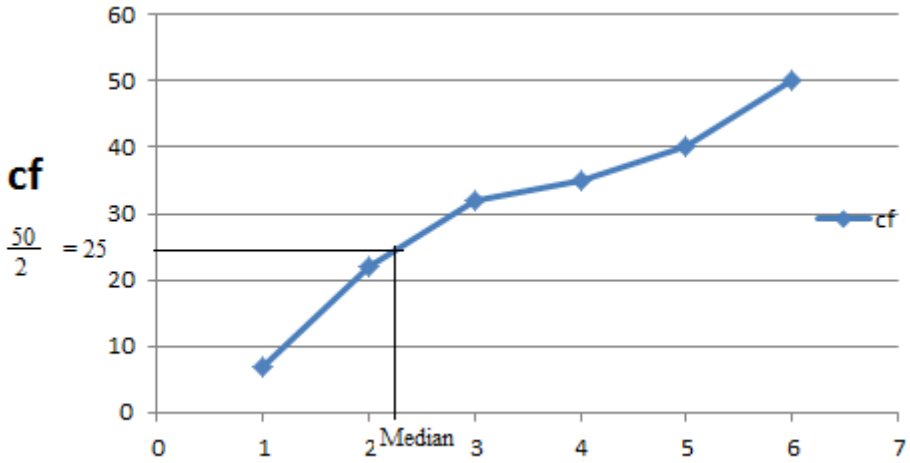


Figure 1.1 Median of a continuous variety

To find the median divide total frequency by 2 and draw a straight line from the cumulative frequency locating the position of the value obtained up to the curve and drop to the x-axis and read the median as shown above in figure 1.1.

Example 1.3 Table 1.5 gives the data of civic of anti's. Find the median civic antis using the cumulative curve.

Table 1.5 gives the data of civic of anti's

class interval	frequency	cumulative frequency	Upper true class limit
4.0 -5.9	1	1	5.95
6.0 - 7.9	3	4	7.95
8.0 - 9.9	7	11	9.95
10.0 - 11.9	9	20	11.95
12.0 - 13.9	10	30	13.95
14.0 - 15.9	11	41	15.95
16.0 - 17.9	8	49	17.95
18.0 - 19.9	1	50	19.95

Using the upper true class limit you can construct a cumulative curve

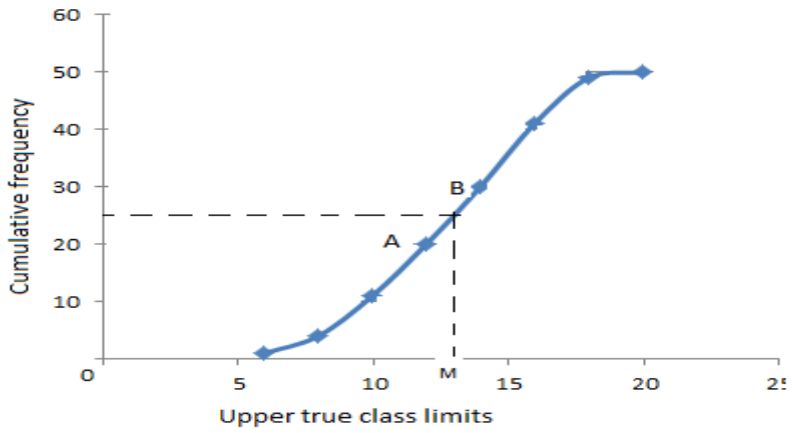


Figure 1.2 Cumulative curve of civic anti's

When using graphical methods answering statistics question you must construct your tables or charts using graph papers.

The mode

Mode: is the most occurring observation in the set of values, for ungrouped data. If data are grouped into classes, you call it the modal class, which is the class with the highest frequency. In statistics we can have more than one mode in any set of values. If there is one mode, you call this unimodal, those with two modes, you call it bimodal, and those with more than two modes is called multimodal.

Example 1.4 Find the mode for each of the following observations:

(a) 3, 7, 4, 9, 6 . No Mode

(b) 2, 2, 2, 3, 3, 5, 7, 9, modes : 2

(c)

x	f
0	2
1	3
2	2
3	1

Mode = 1

(d)

C.I	f
1 - 4	1
5 - 9	3
10 - 14	2
15 - 19	1

Modal Class = 5 - 9

(e)

C.I.	f
1 - 4	1
5 - 9	2
10 - 14	2
15 - 19	1

Modal classes = 5 - 9 & 10 - 14

Mean

One of the important objectives of statistics is to find out various numerical values which explains the inherent characteristics of a frequency distribution? The first of such measures is averages. The averages are the measures which condense a huge unwieldy set of numerical data into single numerical values which represent the entire distribution. The inherent inability of the human mind to remember a large body of numerical data compels us to few constants that will describe the data. Averages provide us the gist and give a bird's eye view of the huge mass of unwieldy numerical data. Averages are the typical values around which other items of the distribution congregate. The mean at time called arithmetic mean is the most common measure of central tendencies and used widely in everyday life. At times it is miscalculated, misused by many users. The uses of the mean in statistical analysis is numerous. This is value in statistical analysis which is used widely.

Definition The mean is defined as the central value of any given observations. This is calculated by adding all the observations and divide by the number of observations. Since it is calculated by using all observations it sits on the Centre of the observations that means some value are below the meanwhile others are above the mean. The mean is not half of the observation are below and half of the other observation are above, it depends on the magnitude of each value to the mean. But if we sum the difference between those below and those above it is zero. This will be made clear when we discuss the standard deviation.

Notation The following notation is being used here;

Mean, $\sum_{i=1}^{n}$ summing up the observations from i = 1,2, ….., n,

n = the number of observations. If the variates takes n values, they can be presented as $x_1, x_2, ..., x_n$, and $\bar{x}$ =mean is given by;

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

Equation (1) is the formula for calculating the mean from observations which is not grouped.

Example 1.5 Calculate the mean for the observations: 71, 58, 67, 82, 74, 66, 47, 90, 67, 82.

There are 10 observations so n = 10 the values are observations give.

The mean is = $\frac{71+58+67+82+74+66+47+90+67+82}{10} = \frac{704}{10} = 70.4$ .That is the mean is 70.4.

Equation (1) is used for calculating the mean from the ungrouped data.

Note that the observations given are integers but there is no reason why the mean should be an integer, it can be any real number. The mean can be calculated from a set of ungrouped data as the Example 1.5 given above or it can be calculated from grouped data (grouped as individual observations or grouped in class intervals) as discussed in the next sections.

Calculation of the mean from a frequency table

The table below shows the scores observed for 50 tosses of a die. Instead of recording each score, a column of frequency is added to make the tabulation easy and its frequency multiplies each score. The total of this column is the total of the scores and its mean is calculated. This is an example of grouped data.

Table 1.6 The scores observed for 50 tosses of a die

score	frequency(f)	scorexfrequency
1	7	7
2	15	30
3	10	30
4	3	12
5	9	45
6	6	36
	50	**160**

The mean $= \frac{160}{50} = 3.8$

If the data is grouped into classes, we can calculate the mean as follows:

Table 1.7

class	True class	lid-class valu	frequency	
interval	**limits**	**x**	f	**fx**
4.0-5.9	3.95 - 5.95	4.95	1	4.95
6.0-7.9	5.95 - 7.95	6.95	3	20.85
8.0-9.9	7.95 - 9.95	8.95	7	62.65
10.0-11.9	9.95 - 11.95	10.95	9	98.55
12.0-13.9	11.95 - 13.95	12.95	9	116.55
14.0-15.9	13.95 - 15.95	14.95	12	179.4
16.0-17.9	15.95 - 19.95	16.95	8	135.6
18.0-19.9	17.95 - 19.95	18.95	1	18.95
			50	**637.5**

Mean $\frac{637.5}{50} = 12.75$

Note that data can be grouped individually just like data presented in table 1.8 below. The scores represented as x in column 1 and their corresponding frequency under f column. The score of 0 has been observed 3 times while the score of 2 has been observed 5 times.

Table 1.8 Distribution of X values

x	f
0	3
1	2
2	5
3	1

From this table 1.8 the mean can be calculated using the formula below

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} \qquad (2)$$

This is the formula for calculating the mean from a grouped data

To calculate the mean from table 1.9 consider the column fx which is the total of the scores. If you add 0 three times the answer is 0 (fx=0x3 = 0), if you add 2 five times the answer is 10 (fx=2x5 = 10).

Table 1.9

x	f	fx
0	3	0
1	2	2
2	5	10
3	1	3
	11	15

Using the formula given in equation (2) the mean from table 1.9 is calculated as:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{15}{11} = 1.3$$

That is, the mean of the given observation is 1.3.

Data can be grouped using class intervals as illustrated in table 1.10 to read this table take an interval 1 – 5 under the class interval column, it has a corresponding value of 1 under frequency column. This means there is one value in this interval, interval 11 – 15 has the frequency of 5 which means there are 5 values in this interval but we do not know what these values are. So these 5 values are approximated by the mid value of this interval called class mark (x). This is calculated as: class mark (x) = $\frac{upper\ \lim it - lower \lim it}{2}$. Then we can calculate the mean using the formula: $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$

Table 1.10 Distribution of grouped data

		class mark	
Class interval	frequency	x	fx
1 - 5	1	3	3
6 - 10	3	8	24
11 - 15	5	13	65
16 - 20	2	18	36
21 - 25	1	23	23
	Total	**65**	**151**

Therefore, $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{151}{65} = 2.32$

That is the mean of the given table of values is 2.32

It should be observed that the class intervals given in table 1.10 are discrete class intervals. This table can be used to calculate the mean as illustrated above and you can represent this data as a bar chart. Table 1.10 can be made continuous by closing the gaps between intervals as follows: Introduce an exact class limits by calculating

$\text{TCL} = \frac{upper\text{lim}it + lowe\text{lim}it}{2} = \frac{5.9 + 6.0}{2} = \frac{11.9}{2} = 5.95$, 5.95 will be the upper limit for interval one and lower limit for the next interval and this interval is closed-open interval meaning if 5.95 appears it should be included in the interval because it is closed. Table 1.11 below illustrate this where the first column is discrete and the second column is continuous, the third column is the class marks, the frequency column and the last column is the total of the scores.

Table 1.11

class	True class	Mid-class valu	frequency	
interval	limits	x	f	fx
4.0-5.9	3.95 - 5.95	4.95	1	4.95
6.0-7.9	5.95 - 7.95	6.95	3	20.85
8.0-9.9	7.95 - 9.95	8.95	7	62.65
10.0-11.9	9.95 - 11.9	10.95	9	98.55
12.0-13.9	11.95 - 13.	12.95	9	116.55
14.0-15.9	13.95 - 15.	14.95	12	179.4
			41	**482.95**

The mean is calculated as;

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{482.95}{41} = 11.78$$

Calculation of the mean using the assumed mean.

At times calculating the mean may be extremely tedious even with a calculator. This can be simplified by using assumed mean in which you take an arbitrary origin of the data. The formula used is as follows:

$$\bar{x} = \frac{A\sum_{i=1}^{n} f_i + B\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i}$$

$$\bar{x} = A + B\bar{u}$$

This method can be generalized as follows:

If we have observations $x_1, x_2, ..., x_n$ which are converted to $u_1, u_2, ..., u_n$, using A as arbitrary origin and B as unit, then;

$$x_1 = A + Bu_1$$
$$x_2 = A + Bu_2$$
$$.....$$
$$x_n = A + Bu_n$$

And $\bar{x} = A + B\bar{u}$

The u_i s at times are called coded data.

Example 1.6 Using table 1.12 calculate the mean using the arbitrary origin of A = 12.95 and B = 2.

Table 1.12 Distribution of observations in class intervals

		Mid			
class	True class	class value	frequency	coded value	
interval	limits	x	f	u_i	$f_i u_i$
4.0-5.9	3.95 - 5.95	4.95	1	-4	-4
6.0-7.9	5.95 - 7.95	6.95	3	-3	-9
8.0-9.9	7.95 - 9.95	8.95	7	-2	-14
10.0-11.9	9.95 - 11.9	10.95	9	-1	-9
12.0-13.9	11.95 - 13.	12.95	9	0	0
14.0-15.9	13.95 - 15.	14.95	12	1	12
16.0-17.9	15.95 - 19.	16.95	8	2	16
18.0-19.9	17.95 - 19.	18.95	1	3	3
			50		-5

$$\bar{u} = \frac{-5}{50} = -0.1$$

$$\bar{x} = A + B\bar{u}$$
$$= 12.95 - 2x0.1$$
$$= 12.75$$

The following table gives the distribution of 100 accidents during seven days of the week in a given month. During a particular month there were 5 Fridays and Saturdays and only four each of other days. Calculate the average number of accidents per day.

Table 1.13

Days:	Sun.	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	Total
Number of accidents:	20	22	10	9	11	8	20	= 100

Solution: Calculation of Number of Accidents per Day

The formula for computation of arithmetic mean according to the short cut method is = where A is assumed mean, dx = (X – A) and N = ∑f.

You can solve this example by short-cut method as given below:

Table 1.14

Calculation of Average Accidents per Day

Day	X	$dx = \bar{X} - A$ (where A = 10)	f	f dx
Sunday	20	+ 10	4	+ 40
Monday	22	+ 12	4	+ 48
Tuesday	10	+ 0	4	+ 0
Wednesday	9	− 1	4	− 4
Thursday	11	+ 1	4	+ 4
Friday	8	− 2	5	− 10
Saturday	20	+ 10	5	+ 50
			30	**+ 128**

f dx = $\bar{X}$ **– A**

128/30 = $\bar{X}$ **- 10**

$\bar{X}$ = 14 accidents per day

Skewness

The set of observations from a frequency distribution can either be symmetrical that is bell-shaped in which case: mean = median = mode, Otherwise, the data is said to be skewed.

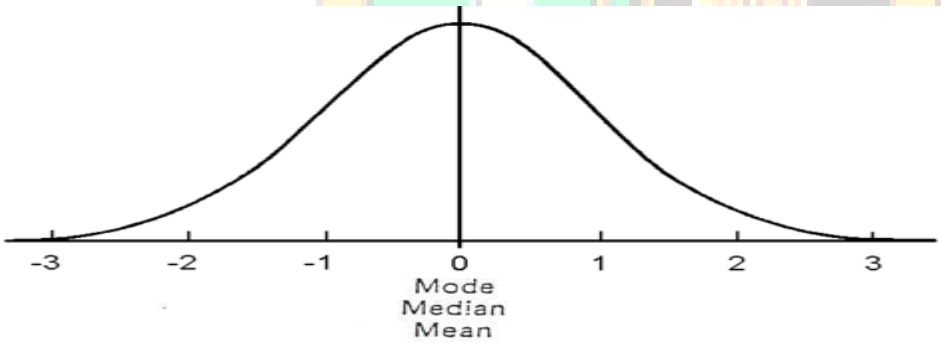


Figure 1.3 Normal distribution

If the observations are not symmetrical then they are skewed, either to the left or to the right. A distribution is skewed if one of its tails is longer than the other. The first distribution shown has a positive skew. That is it has a long tail in the positive direction. It is right for a distribution that

is skewed to the right. For right-skewed distribution, the mean is higher than the median. Also it be observed that the tail of the distribution on the right is longer than on the left-hand side.

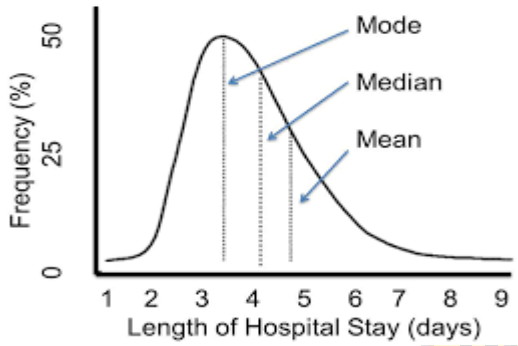


Figure 1.4 Skewed data to the right

The distribution below it has a negative skew since it has a long tail in the negative direction. A distribution which is skewed to the left has its mean smaller than its median

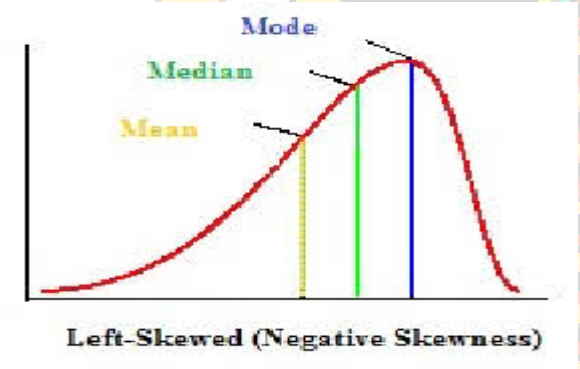


Figure 1.5 Skewed data to the left

In probability and **statistics**, **skewness** measures the symmetry of the **distribution** of a probability of a real-valued function of a random variable about its mean **w**hich can be either positive or negative.

Weighted mean

Suppose the journey from Lusaka to Livingstone which is 480km cost K160 and a journey from Lusaka to Ndola a distance of 320km cost K120. If our interest is to find the average cost per km overall, that is Ndola to Livingstone, it is wrong to add the cost per journey and divide by total distance. Since the two journeys are of different distances and each distance has its own cost we need to use weighted averages in calculating overall cost per km. To do this we use weighted average to calculate overall cost per km weighted in the ratios of their distances. Therefore, the example given above can be calculated as:

Table 1.15 Table for weighted data

destination	Distance	Cost (K)	cost per km(p/km)
Ndola	320	120	0.38
Livinstone	480	160	0.33

The weight mean $= \frac{320x0.38 + 480x0.33}{320 + 480} = 0.35$

That means the overall cost per km is 0.35

In general if the values $x_1, x_2, ..., x_n$ are given and the weights are $w_1, w_2, ..., w_n$ then

Weighted mean $= \frac{w_1x_1 + x_2w_2 + ... + w_nx_n}{w_1 + w_2 + ... + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

The weighted mean is used in financial calculations such as; knowing the average amount of time left before mortgages in a mortgage-backed security expire

Moving Averages

In the study of times series data (when time effect is evident) data such as rainfall data, labour statistics data (unemployment data), population data etc. may contain some seasonal variations, trends and cyclic variations. Data given daily, monthly, yearly, a five or ten year moving average

or twelve months moving average can be calculated. A moving average is used to reveal the general trends and this is generally done in economics, businesses and time series data analysis.

Example 1.7 The table below gives the quarterly unemployment figure for Zambia throughout nearly three years. This is usually dealt in time series analysis, where a seasonal variation produces a zig-zag effect which to some extent masks any general trend.

Table 1.16 Number of unemployment in Zambia

Year	Month	No of unemployment
2014	July	1471
	October	1368
	January	1464
	April	1341
2015	July	1456
	October	1430
	January	1586
	April	1452
2016	July	1549
	October	1518
2017	January	1622

The annual variation can be removed by calculating the four quarterly moving averages. This is done by averaging the values over four quarters and plotting the average obtained in the middle of the range used. For example the average for the first four quarters is

$\frac{1}{4}(1471+1368+1464+1341) = 1411$ which plotted mid-way between Oct 2014 and Jan 2015.

The next moving average can be found most simply from 1535.25 + (value for July 2015 – value for July 2014) and so on.

Calculation of M. average

Table 2.17 Calculations of moving averages

Years	Months	No. of unemployment	quartely Totals	Moving average
2014	July	1471		
	October	1368	5644	1411
	January	1464	5629	1407.25
	April	1341	5691	1422.75
2015	July	1456	5813	1453.25
	October	1430	5924	1481
	January	1586	6017	1504.25
	April	1452	6105	1526.25
2016	July	1549	6141	1535.25
	October	1518		
2017	January	1622		

Moving averages can be calculated from different types of data, but their underlying purpose remains the same: that is to help technical traders track the trends of financial assets by smoothing out the day-to-day price fluctuations, or noise. It is a tool used for smoothing out price data by creating a continually updated average price. The average is calculated over a specific period, like 10 days, 20 minutes, 30 weeks or any period the trader chooses.

1.5.4 Measure of dispersion

When data is available, we are interested in the preliminary analysis the measures of central tendencies of the observations which were discussed in the measures of central tendencies discussed above. The next interest would be how spread is these observations from the measures of central tendencies. These values are called measures of dispersion. Here we discuss measures of dispersion and provides an interpretation of these values concerning measures of central tendencies and their uses in everyday life.

The range

If we want to know how scattered the observations are; we can use them by finding the difference between the smallest and largest values.

Range = highest value- smallest value

Interpretation: If the range is small then the values are clustered together. However, if the range is large, it means the values are widely scattered around. If the range is zero, it means, we have only one value.

Example 1.8 The numbers are 7, 4, 9, 8, 2. Find the range, IQR, variance, standard deviation and coefficient of variation.

Solution: In order these numbers are 2, 4, 7, 8, 9. n = 5

Table 1.18

Index	x	x^2	$(x-\bar{x})$	$(x-\bar{x})^2$
1	2	4	2 – 6 = -4	16
2	4	16	4 – 6 = -2	4
3	7	49	7 – 6 = 1	1
4	8	64	8 – 6 = 2	4
5	9	81	9 - 6 = 3	9
	30	214	0	34

i) Range: $range = 9-2 = 7$.

ii) Variance: First, find the sample mean. $\bar{x} = \frac{\sum x}{n} = \frac{30}{5} = 6$. The preferred method for finding the variance is the computational formula, which has the advantage that the $(x-\bar{x})$ and $(x-\bar{x})^2$ columns are not needed.

iii) $\text{Var(x)} = \sigma^2 = \frac{1}{n}\sum (x-\bar{x})^2 = \frac{34}{5} = 6.8$

IQR: To find the first quartile $p = .25$ so $position = p(n+1) = .25(6) = 1.5 = a.b$. So $a = 1$ and $.b = .5$. Thus $1Q = x_{.75} = x_a + .b(x_{a+1} - x_a) = x_1 + .5(x_{1+1} - x_1) = x_1 + .5(x_2 - x_1)$

$= 2 + .5(4-2) = 2+1 = 3$. To find the third quartile $p = .75$ so $position = p(n+1) = .75(6) = 4.5 = a.b$. So $a = 4$ and $.b = .5$. Thus $3Q = x_{.25} = x_a + .b(x_{a+1} - x_a) = x_4 + .5(x_{4+1} - x_4) = x_4 + .5(x_5 - x_4) = 8 + .5(9-8)$

$= 8 + 0.5 = 8.5$. Finally, $IQR = 3Q - 1Q = x_{.25} - x_{.75} = 8.5 - 3 = 5.5$.

iv) Var(x) = $\sigma^2 = \frac{1}{n}\sum(x-\bar{x})^2 = \frac{34}{5} = 6.8$, standard deviation is square root of variance.

Standard deviation is σ. Therefore $\sigma = 2.61$

v) Coefficient of variation: $C = \frac{std.deviation}{mean} = \frac{\sigma}{\bar{x}} = \frac{2.61}{6} = 0.435$ or 43.5%.

Inter-quartile range

Interquartile range for ungrouped data

The other measure of dispersion in the inequality range which extends the idea of the median. The median is just the middle value of the observations when the observations are ordered. The median of the first set of the observation is called the first quarter, called the lower quartile denoted by Q_1, and the median of the second half is called the upper quartile denoted by Q_3. In find Q_1 and Q_3, you observe whether the number of observation n is either odd or even.

: - Interquartile range = $Q_3 - Q_3$

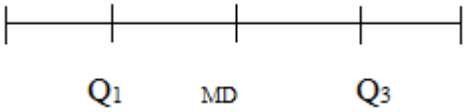
The IQR is an indicator of how observation is spread. When drawing box-whisker plots, the IQR is used. This displays the variability of observation if IQR is small, it means the observations are clustered around, but if the IQR is large then the observations are scattered around.

Example 1.9 Find the interquartile range for the observations given below

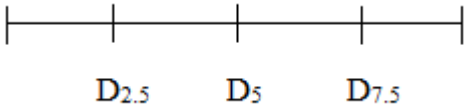
1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4,

Interquartile range for grouped data

The median for grouped data is found by using the cumulative curve. From the cumulative curve which should be drawn on the graph paper, can be used to estimate the median, Q_1, and Q_3. The lower quartile (Q_1) is found by reading off the income corresponding to a cumulative frequency of one-quarter the total frequency and the upper quartile is found by reading off the corresponding to a cumulative frequency which is three-quarters of the total frequency when finding Q_1 and Q_3, it means the data set when ordered is partitioned into four equal parts as Q_1, median and Q_3.



You can partition Data into ten equal parts, and this gives deciles as Q_1, median, and Q_3.



Where D_5 = Median = ½ $(n+1)^{th}$ value if n is odd and ½ $(n^{th} + (n + 1)^{th})$ value if n is even

$D_{2.5}$ is $\frac{2.5}{10}$ $(n+1)^{th}$ value if n is odd

$D_{7.5}$ is $\frac{7.5}{10}$ $(n+1)^{th}$

If the data is partitioned into 100 equal parts, this is called percentiles, with

P_{50} = Md = D_5, P_{75} = Q_3, P_{25} = Q_1

Example 1.10 Find the inter quartile range of the following data set given below:

Table 1.19 Frequency distribution of observations

X	F	CF	
0	2	2	
1	3	5	
2	7	12	n = 17, 0dd
3	4	16	
4	1	17	

The Standard Deviation

Dispersion is the extent to which the magnitudes or quantities of the items differ, the degree of diversity. The word dispersion may also be used to indicate the spread of the data. Measure of dispersion is a value that indicates the extent to which all other values are dispersed about the central value in a particular distribution. The standard deviation can be calculated from an ungrouped data or grouped data.

Ungrouped data

In calculating the standard deviation from an ungrouped data the following formulas are being used:

a) Mean = $\bar{x} = \frac{1}{n}\sum_i x$ (b) standard deviation = $\sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n}}$

Example 1.11 Find the standard deviation for the ungrouped data 8, 9, 10, 10, 11

	x	$x-\bar{x}$	$(x-\bar{x})^2$
	8	-1.6	2.56
	9	-0.6	0.36
	10	0.4	0.16
	10	0.4	0.16
	11	1.4	1.96
Total	48		5.2
$\bar{x}$	9.6		
sd	1.02		

$$\sigma = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n}}$$

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = 9.6 \quad \sigma = \sqrt{\frac{\sum_i (x_i-\bar{x})^2}{n}} = 1.02$$

Grouped data: There are many ways of writing the formula for the standard deviation. The one above is for a basic list of numbers. The formula for the variance when the data is grouped is as follows. The standard deviation can be found by taking the square root of this value.

(a) $Mean = \bar{x} = \dfrac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$, (b) $Variance = \sigma^2 = \dfrac{\sum_{i=1}^{n} f_i (x-\bar{x})^2}{\sum_{i=1}^{n} f_i}$ or

$Variance = \sigma^2 = \dfrac{\sum_{i=1}^{n} f_i x_i^{\,2}}{\sum_{i=1}^{n} f_i} - \left(\dfrac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}\right)^2$, These formulas can be used in calculating the mean and variance of any grouped data. Note that data can be grouped as individual values or grouped in intervals.

Example 1.12 The table shows marks (out of 10) obtained by 20 people in a test. Calculate (i) the mean marks, (ii) the standard deviation of the marks.

Table 1.20 Frequency distribution of marks obtained in a test

Mark (x)	Frequency (f)
1	0
2	1
3	1
4	3
5	2
6	5
7	5
8	2
9	0
10	1

Solution

	Mark (x)	Frequency (f)	fx	$x - \bar{x}$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
	1	0	0	-4.9	24.01	0
	2	1	2	-3.9	15.21	15.21
	3	1	3	-2.9	8.41	8.41
	4	3	12	-1.9	3.61	10.83
	5	2	10	-0.9	0.81	1.62
	6	5	30	0.1	0.01	0.05
	7	5	35	1.1	1.21	6.05
	8	2	16	2.1	4.41	8.82
	9	0	0	3.1	9.61	0
	10	1	10	4.1	16.81	16.81
Total	**55**	**20**	**118**			**67.8**
	Mean	5.9				
	Var(x)	3.39				

(i) $\bar{x} = \dfrac{\sum fx}{\sum f} = \dfrac{118}{20} = 5.9$

(ii) $\sigma = \sqrt{\dfrac{\sum f(x-\bar{x})^2}{\sum f}} = \sqrt{\dfrac{67.8}{20}} = \sqrt{3.39} = 1.84$

Example 1.13 The frequency distribution of the lengths of 100 leaves from a certain species of plant is given in below in table 1.21:

Table 1.21 Frequency distribution of the lengths of 100 leaves

length (mm)	Frequency
20 – 24	6
25 – 29	10
30 – 34	18
35 – 39	25
40 – 44	22
45 – 49	15
50 – 54	4

Find the range and standard deviation of heights.

Solution The range is L = 52 and s =22 Range = 30

	length (mm)	Frequency	Class Mark(x)	fx	$x - \bar{x}$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$	Cf
	20 – 24	6	22	132	-15.4	237.16	1422.96	6
	25 – 29	10	27	270	-10.4	108.16	1081.6	16
	30 – 34	18	32	576	-5.4	29.16	524.88	34
	35 – 39	25	37	925	-0.4	0.16	4	59
	40 – 44	22	42	924	4.6	21.16	465.52	81
	45 – 49	15	47	705	9.6	92.16	1382.4	96
	50 – 54	4	52	208	14.6	213.16	852.64	100
Totals		100		3740			5734	
mean	37.4							
variance	57.34							
sd	7.57							

(i) $\bar{x} = \dfrac{\sum fx}{\sum f} = \dfrac{3740}{100} = 37.4$

(ii) $\sigma = \sqrt{\dfrac{\sum f(x-\bar{x})^2}{\sum f}} = \sqrt{\dfrac{5734}{100}} = \sqrt{57.34} = 7.57$

Finding the mean and standard deviation of a combined sample observations

Sometimes we have the mean and standard deviation of two or more sets of values, and we want to find the mean & standard deviation of the combined set of values. This presents no problem if the original data are available, it is possible to find the combined mean and standard deviation if the sample means & standard deviation are available.

Example 1.14: Below is a table showing the mean and standard deviation of the marks obtained by two classes in a test. Calculate

(a). the mean and b) the standard deviation for the two classes combined.

Table 1.22 Mean and standard deviation Combined tables

	mean	Standard deviation	Class size
Class A	67	5	15
Class B	58	8	18

The method is based on finding $\sum x$ and $\sum x^2$ for each sample and hence finding $\sum x$ and $\sum x^2$ for the sample combined. Using the formula $\bar{x} = \frac{\sum x}{n}$ you have $n\bar{x} = \sum x$.You use subscripts to distinguish between two samples $\sum x_A = n_A \bar{x}_A$ =15x67=1005

$\sum x_B = n_B \bar{x}_B = 18x58 = 1044$

Combining the samples

$\sum x = 1005 + 1044 = 2049,\ n = 15 + 18 = 33$

Giving a mean for the combined sample of $\bar{x} = \frac{2049}{33} = 62.1$

Note that in effect you have calculated a weighted mean of .x

(a) From equation $\sigma^2 = \frac{\sum x^2}{n} - \bar{x}^2$ this can be rearranged to give

$\sum x^2 = n(\bar{x} + \sigma^2)$ for the two samples separately you have

$\sum {x_A}^2 = n(\bar{x}_A + {\sigma_A}^2) = 15(67^2 + 5^2) = 67710$

$\sum {x_B}^2 = n(\bar{x}_B + {\sigma_B}^2) = 18(58^2 + 8^2) = 61704$ and the sample combined

$$\sum x^2 = 67710 + 61704 = 129414$$

The variance for the combined sample is given by:

$$\sigma^2 = \frac{\sum x^2}{n} - \bar{x}^2 = \frac{129414}{33} - \left(\frac{2049}{33}\right)^2 = 66.355$$

Interpretation of the measure of dispersion

In a statistical sense, dispersion has two meanings: first it measures the variation of the items among themselves, and second, it measures the variation around the average. If the difference between the value and average is high, then dispersion will be high, otherwise it will be low. This displays how observations are spread about the mean.

1.5.5 Presentation of Data

Preliminary analysis of data involves data presentations in pictorial forms for non-mathematics audience to understand what is being presented. Two types of presentation can be displayed.

Discrete data: This type of data can be presented using: scatter plots, bar chart, pie chart etc. These charts can be drawn from frequency distributions.

Example 1.15 The frequency table of number of members per family. Construct a scatters diagram.

Table 1.23 Frequency table of number of members per family

Family No.	Number of members
1	5
2	6
3	0
4	2
5	4

Solution. The graph below shows the scatter diagram where individual points are shown in x-y plane.

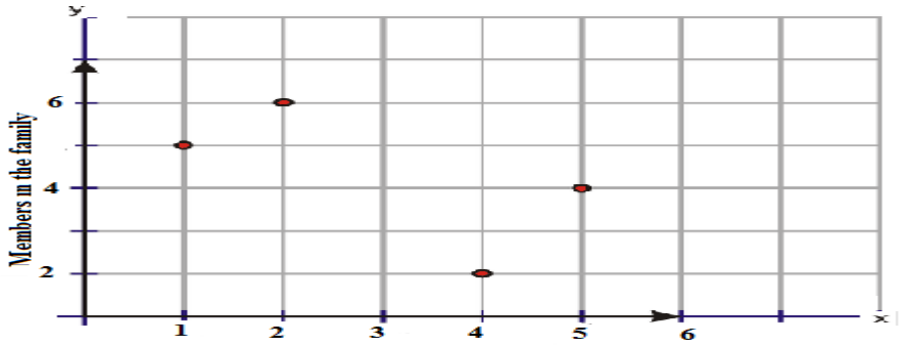


Figure 1.6 Number of family versus number of members

Bar graph

Example 1.16 The students in IDE math class have birthdays in each of the four seasons. Table 1.24 shows the four seasons, the number of students who have birthdays in each season, and the percentage (%) of students in each group. Construct a bar graph showing the number of students.

Table 1.24 Seasons number of student's proportions of population

Spring	8	24%
Summer	9	26%
Autumn	11	32%
Winter	6	18%

Solution

Using the data from table 1.24 above, you can construct a bar graph showing the percentages.as follows:

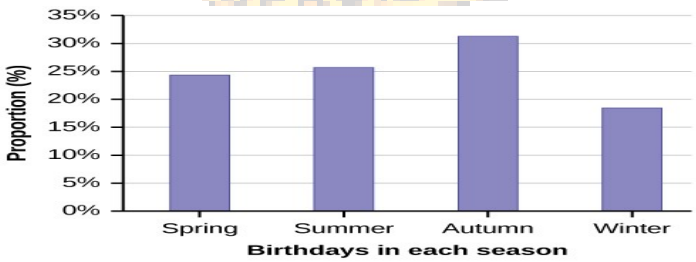


Figure 1.7 Seasons number of student's proportions of population

Example 1.17 The table below shows the number of members in each family

Table 1.25 number of members in each family

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

Construct a bar chart to illustrate the number of members in each family.

Solution. Figure 1.8 below illustrates on how to construct a bar graph which shows bars which are the same and for family 3 there is zero member.

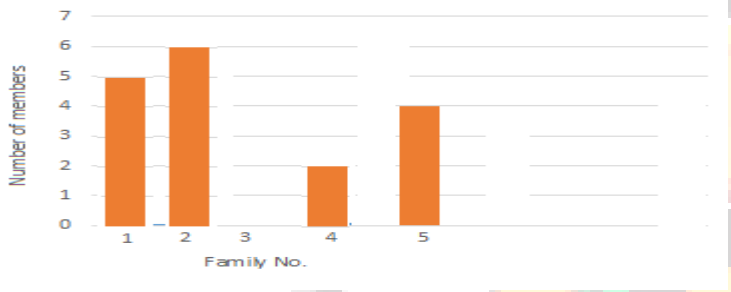


Figure 1.8 Bar graph of number of members per Family

Example 1.18 The table below is a frequency distribution on types of jobs people are involved in. Construct a bar chart for this distribution.

Table 1.26 Frequency table of part-time jobs

Part-Time jobs	frequency
Baby sitting	8
Fast foods	14
Delivery	12
Lawns Maintenance	10
Grocery Store	13
Tutoring	5

Solution You can read the values of the part time jobs from the Bar charts as:

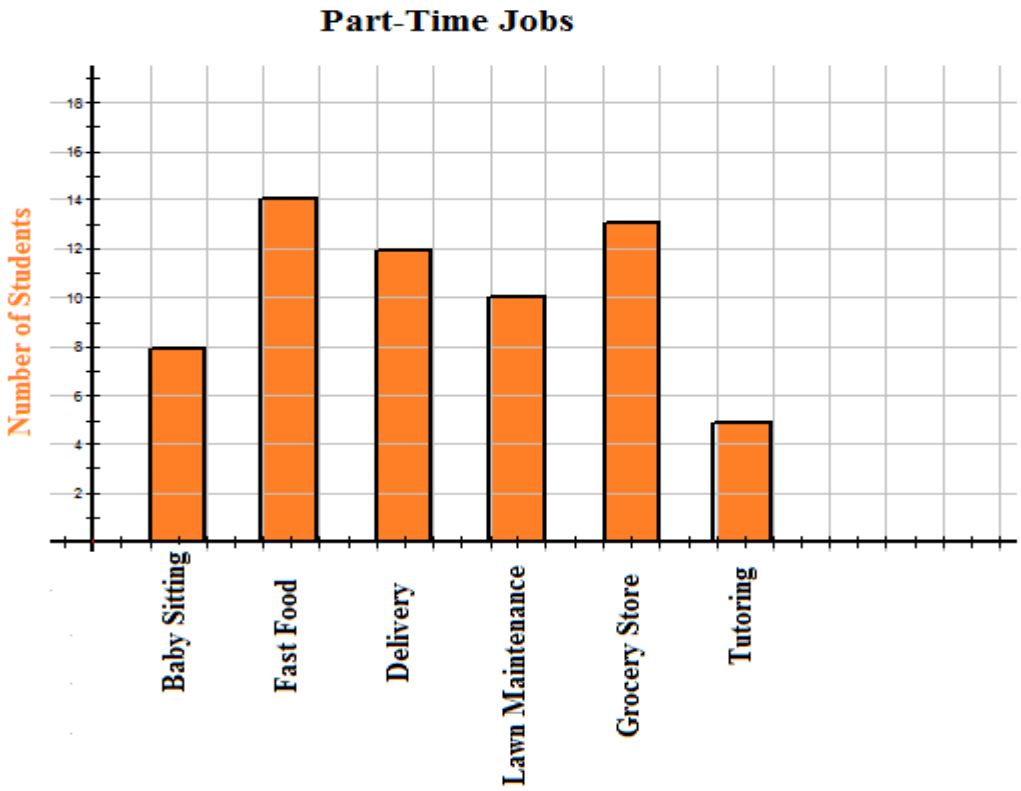


Figure 1.9 Frequency distribution on types of jobs

Example 1.19 Suppose, of the 140 children, 20 lived in owner occupied houses, 70 lived in council houses and 50 lived in private rented accommodation. Figures from the census suggest that for this age group, throughout the county, 50% live in owner occupied houses, 30% in council houses, and 20% in private rented accommodation. Type of accommodation is a

categorical variable, which can be displayed in a bar chart. We first express our data as percentages: 14% owner occupied, 50% council house, 36% private rented. You then display the data as a bar chart. The sample size should always be given.

Solution The bar chart of housing data for 140 children and comparable census data

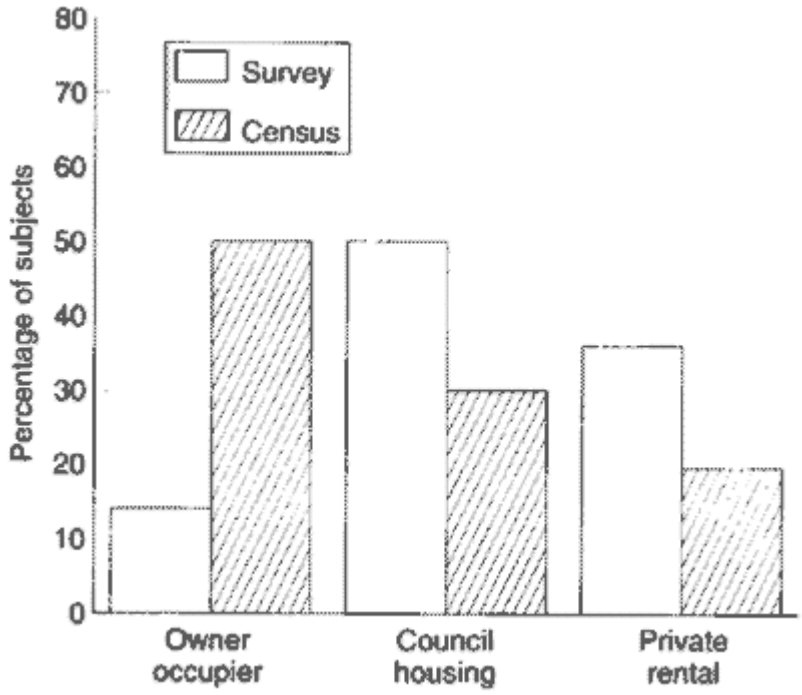


Figure 1.10 The bar chart of housing data for 140 children and comparable census data

This example shows you that you can present two or more variables on the same bar chart.

Presentation of continuous data: For continuous **data** you can present data as; a line graph, histogram, cumulative curve (Ogive curve), frequency polygon.

Line graph: In constructing line graph you start by plotting points and join these points with straight lines. In joining straight line you can use a ruler and not by free hands.

Example 1.20 The graph in figure 1.11 represents the average temperatures during the months in 2018 at IDE offices. This is an example of a line graph which is continuous. You can easily tell this by looking at the graph and seeing the data points connected together.

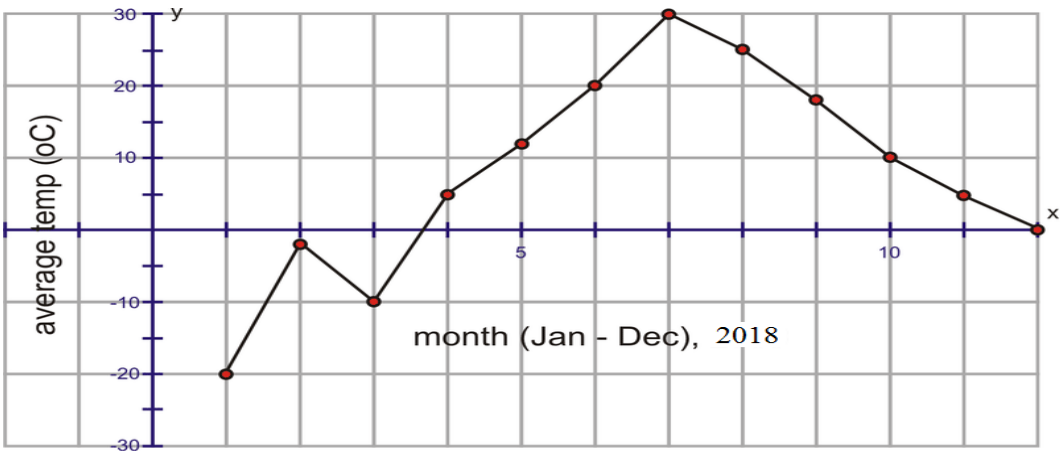


Figure 1.11 Months-temperature graph for IDE offices

The graph above is continuous, hence you can read any temperature from the graph within the given range. This is an example of a line graph.

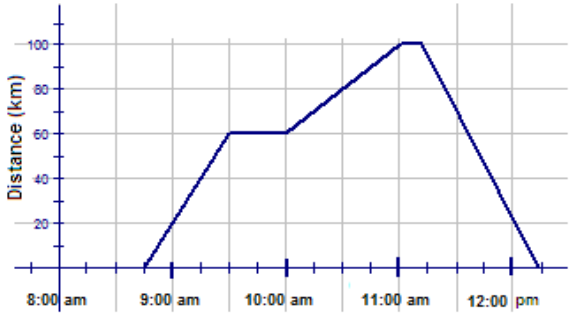


Figure 1.12 Time – distance graph

The graph shown above is time - distance graph which is a line graph. As you can see from the graph, there is no break in the line. In other words, you can choose any time between 8:45 am and 12:15 pm, even one involving a fraction of a second, and there will be a corresponding distance in km. Therefore, this is an example of a line graph which represents continuous data.

Histogram: You can construct histogram for continuous data by making sure the bars are joined and not separate bars as examples in bar graphs. These bars should be of the same width and the values of the height should be the frequency values as shown below:

Example 1.21 Consider the concentration of an antibody in blood serum from 50 donor, given in table 2.27(continued) below:

Table 1.27(cont') Concentration of antibody in blood serum

11.6	15.8	17.0	14.2	16.2	17.3	15.7	17.3	12.0	11.5
8.4	13.2	12.5	15.7	14.0	10.8	9.9	14.0	15.0	10.2
19.2	12.0	8.2	13.8	12.6	9.6	16.0	13.7	11.5	14.8
17.5	15.2	7.2	17.0	5.2	9.4	7.5	15.5	10.7	14.2
15.3	7.8	11.2	10.0	12.1	10.5	12.2	8.6	9.9	16.1

Using the class interval of 4.0 – 5.9, etc as discrete ungroup class interval and 3.95-5.95, 5.95-7.95 etc as grouped continuous class interval.

Table 1.28 Grouped frequency table of data

Class interval (g/l)	True class limits(g/l	Mid class value(g/l)	Frequency
4.0 - 5.9	3.95 - 5.95	4.95	1
6.0 - 7.9	5.95 - 7.95	6.95	3
8.0 - 9.9	7.95 - 9.95	8.95	7
10.0 - 11.9	9.95 - 11.95	10.95	9
12.0 - 13.9	11.95 - 13.95	12.95	9
14.0-15.9	13.95 - 15.95	14.95	12
16.0 - 17.9	15.95 - 17.95	16.95	8
18.0 - 19.9	17.9 - 19.95	18.95	1
			50

The table 1.28 above gives the first column as discrete grouped data, and this can be used for constructing bar chart, while the second column is the continuous grouped data, and this can be used for constructing a histogram, in which a frequency polygon can be constructed by joining

the mid points. The midpoint value calculated as (upper lines + low)/2 of column 1. What this means is that for Example, 1.21 above is represented by 8.95.

Results for continuous data such as the data above can be represented by a histogram as shown below. The divisions between the bars fall on the true class limits and the heights of the bars, and consequently their areas, are proportional to frequency.

The frequency histogram would look like

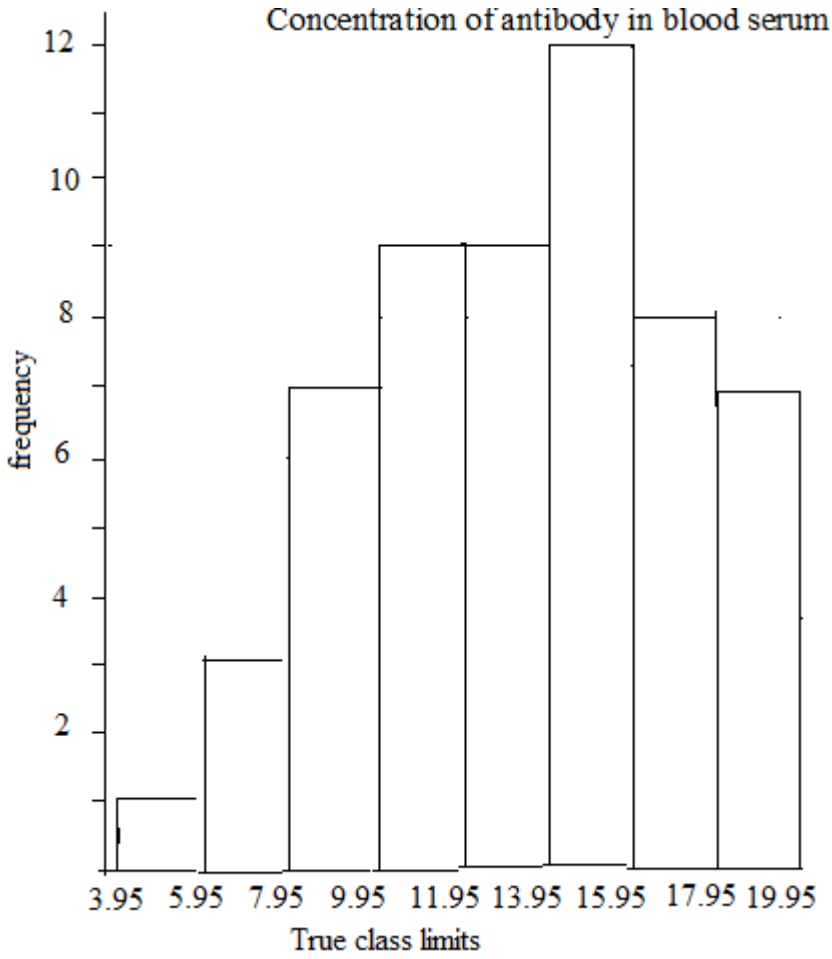


Figure 1.13 Concentration of antibody in blood serum

Example 1.22: The table below is the frequency table on the nose length of IDE statistics students.

Table 1.29 the nose length of IDE statistics

Class interval	Frequency
27.5-32.5	2
32.5-37.5	5
37.5-42.5	17
42.5-47.5	21
47.5-52.5	11
52.5-57.5	3
57.5-62.5	0
62.5-67.5	0
67.5-72.5	1
	60

Construct a frequency table showing the relative frequency and density height

Construct a histogram using the density height using 5mm classes cantered at 30, 35, …,70

Solution

Table 1.30 Nose length

Class interval	Tally	Frequency	Relative Frequency	Density height
27.5-32.5	||	2	0.033	0.0066
32.5-37.5	||||	5	0.083	0.0166
37.5-42.5	|||| |||| |||| ||	17	0.283	0.0566
42.5-47.5	|||| |||| |||| |||| |	21	0.350	0.0700
47.5-52.5	|||| |||| |	11	0.183	0.0366
52.5-57.5	|||	3	0.050	0.010
57.5-62.5		0	0	0
62.5-67.5		0	0	0
67.5-72.5	|	1	0.017	0.0034
		60	0.999 (rounding)	

From table 1.30 you can construct a histogram using frequency column, relative frequency column or density column see figure 1.14. You can construct a histogram using frequency values or relative frequency from table 1.30. Note that a density histogram is just a modified relative frequency histogram. A density histogram is defined so that:

- the area of each rectangle equals the relative frequency of the corresponding class, and
- the area of the entire histogram equals 1.

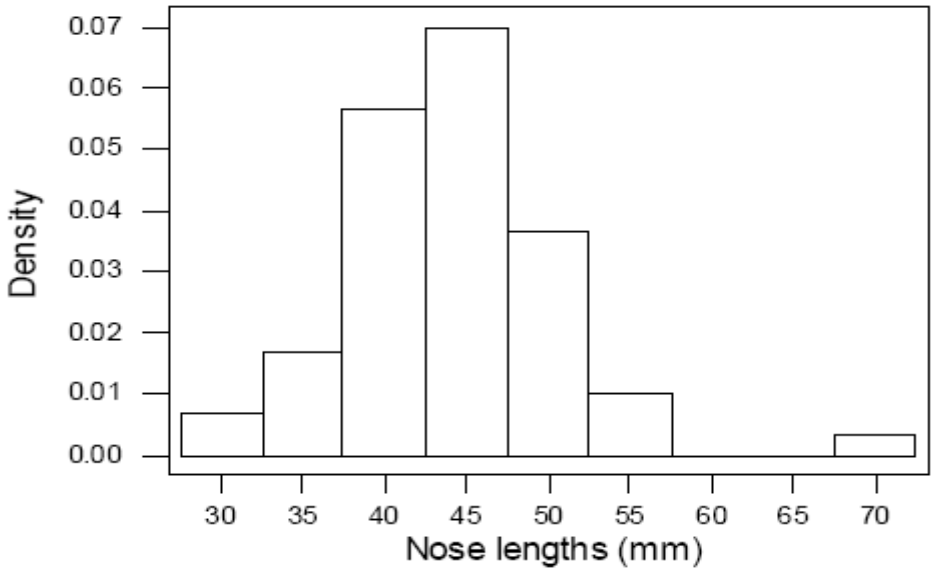


Figure 1.14 Histogram of the length of Nose for IDE students

Frequency Polygon

A frequency polygon is a graphical form of representation of data. It is used to depict the shape of the data and to depict trends. It is usually drawn with the help of a histogram but can be drawn without it as well. A histogram is a series of rectangular bars with no space between them and is used to represent frequency distributions.

Steps to Draw a Frequency Polygon

- Mark the class intervals for each class on the horizontal axis. We will plot the frequency on the vertical axis.
- Calculate the class mark for each class interval. The formula for class mark is:

  Class mark = (Upper limit + Lower limit) / 2

- Mark all the class marks on the horizontal axis. It is also known as the mid-value of every class.
- Corresponding to each class mark, plot the frequency as given to you. The height always depicts the frequency. Make sure that the frequency is plotted against the class mark and not the upper or lower limit of any class.

- Join all the plotted points using a line segment. The curve obtained will be kinked.
- This resulting curve is called the frequency polygon.

Note that the above method is used to draw a frequency polygon without drawing a histogram. You can also draw a histogram first by drawing rectangular bars against the given class intervals. After this, you must join the midpoints of the bars to obtain the frequency polygon. Remember that the bars will have no spaces between them in a histogram.

Example 1.23: Construct a frequency polygon using the data given below:

Table 1.31 Frequency distribution of observations

Test Scores	Frequency
49.5-59.5	5
59.5-69.5	10
69.5-79.5	30
79.5-89.5	40
89.5-99.5	15

Solution: We first need to calculate the cumulate frequency from the frequency given.

Table 1.32 Continuous data Frequency distribution of observations

Test Scores	Frequency	Cumulative Frequency
49.5-59.5	5	5
59.5-69.5	10	15
69.5-79.5	30	45
79.5-89.5	40	85
89.5-99.5	15	100

You now start by plotting the class marks such as 54.5, 64.5, 74.5 and so on till 94.5. Note that you will also plot the previous and next class marks to start and end the polygon, i.e. you plot

44.5 and 104.5 as well. Then, the frequencies corresponding to the class marks are plotted against each class mark. Like you can see below, this makes sense as the frequency for class marks 44.5 and 104.5 are zero and touching the x-axis. These plot points are used only to give a closed shape to the polygon. The polygon looks like this:

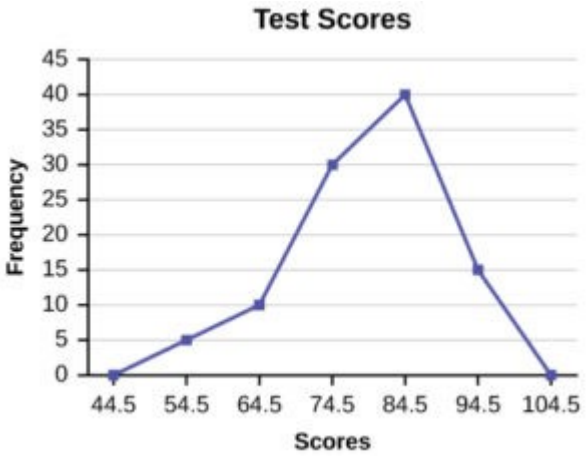


Figure 1.15 Frequency polygon

Note that the points plotted have been joined with straight lines. You need to complete by joining the x-axis otherwise it will not be a polygon.

Cumulative curve

This is another way of representing data graphically. The cumulative frequency for a particular value of the variety is frequency of observations, which accumulates, constructed using the upper true class limits. The frequency is the number of times an event occurs within a given scenario. Cumulative frequency is defined as the running total of frequencies. It is the sum of all the previous frequencies up to the current point. It is easily understandable through a Cumulative frequency table.

Example 1.24: The table below gives the frequency distribution number of students marks. Construct a cumulative frequency curve.

Table 1.33: Frequency Table for the No. of students

Marks	Frequency (No. of Students)	Cumulative Frequency
0 – 5	2	2
5 – 10	10	12
10 – 15	5	17
15 – 20	5	22

Cumulative Frequency is an important tool in Statistics to tabulate data in an organized manner. Whenever you wish to find out the popularity of a certain type of data, or the likelihood that a given event will fall within certain frequency distribution, a cumulative frequency table can be most useful. Say, for example, the Census department has collected data and wants to find out all residents in the city aged below 45. In this given case, a cumulative frequency table will be helpful.

Cumulative Frequency Curve

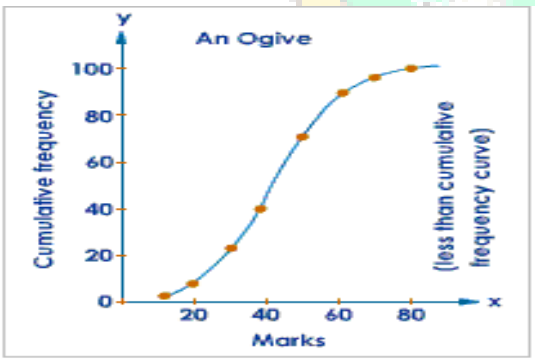


Figure 1.16: Cumulative curve of the marks of students

A curve that represents the cumulative frequency distribution of grouped data on a graph is called a Cumulative Frequency Curve or an Ogive. Representing cumulative frequency data on a graph is the most efficient way to understand the data and derive results.

Box and Whisker plots: This type of data presentation is commonly used in time series. This shows the variability of data say in each month of the year. If the box is small that means the variability of the inter quartile occurs within a short time. If the range is small then the values are clustered together. But if the range is large it means the values are widely scattered around. If the range is zero, it means, we have only one value. : **- Interquartile range = $Q_3 - Q_1$**

The IQR is an indicator how observation are spread. When drawing box – whisker plots, the IQR is used. The box plots it looks like

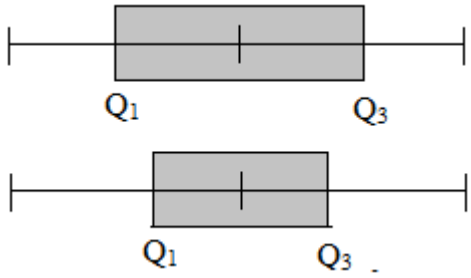


Figure 1.17 Box and whisker plots

These box plots are used when comparing variability between data sets. The smaller the box the more closely the data from the mean. This displays the variability of observation. If IQR is small, it means the observations are clustered around the median, but if the IQR is large than the observations are scattered away from the median.

1.2 UNITY ACTIVITY

1. Give the mean, median and mode(s) for the following values, which gives the best measures of central tendency? 2, 4, 6, 3,1,2, 1,1, 5,4,4,2,5,6,8,15,7.
2. The mark distribution in the table below where obtained from two groups of children taking the same school examination.

   (a) Plot these distribution in the way that seems best to you, in order both to illustrate them separately and to show up the difference between them.

   (b) Estimate the median marks for the two groups

Table 1.34 The marks of test results for two groups

Marks	0-29	30-39	40-49	50-59	60-69
Group A	4	2		4	11
Group B	2	4		4	5

3. The sunshine figures for Lusaka town for the year 2015-2017 are given in table2.35 below: Each figure representing the mean hours of sunshine per day over the month.

   (a) Draw a graph to illustrate this information

   (b) Superimpose a second graph showing the twelve –month moving average

Table 1.35 The sunshine figures for Lusaka

	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
2015	1.2	2.8	5.0	5.6	6.6	5.4	8.5	6.4	5.2	3.8	1.8	1.4
2016	1.8	2.5	4.5	5.5	7.8	5.1	5.1	4.9	3.3	3.6	1.9	0.7
2017	1.5	2.7	3.1	5.2	6.7	9.6	4.5	4.7	3.7	2.7	2.1	1.7

4. The table below shows the cumulative distribution of gross weekly earnings of male full time workers in Zambia Sugar company in April 2014. Estimate for the distribution

   (a) The median earnings of a manual worker,

   (b) The median earnings of a non- manual worker

   (c) The proportion of manual workers earning less than the median earnings of non-manual workers

Table 1.36 Distribution of gross weekly earnings of male full time workers in Zambia Sugar Company

Gross weekly earning (K)	35	40	45	50	55	60	70	80	100	200
Manual workers ('000)	0.1	0.3	0.7	1.3	2.0	2.7	4.0	4.9	5.7	6.1
Non-manual workers ('000)	0.1	0.2	0.4	0.6	0.9	1.2	1.8	2.4	3.2	4.1

(5) Table 1.37 gives further unemployment figures for Zambia. Draw a graph to illustrate this information and superimpose the four-quarterly moving average.

Table 1.37 Unemployed people in Zambia. Figures in thousands.

Year	January	April	July	October
1983	3021	3094		
1984	3200	3108	3101	3225
1985	3341	3273	3235	3277
1986	3408			

(6) Table 1.38 is a frequency table for the number of words in a sentence for a paragraph in a book. Calculate the mean number of words in a sentence.

Table 1.38 Frequency table for the number of words in a sentence

Number of words	Number of sentences
5-9	9
10-14	10
15-19	8
20-24	11
25-29	8
30-34	4
35-39	4
40-44	1
45-49	1

(7) For a project a student collected data on the hair-care habits of fellow students. She asked 50 boys and 50 girls how many times they had washed their hair in the previous week. The results are shown in Table 1.39.

Table 1.39 Hair-care habits of fellow students

Number of washes	1	2	3	4	5	6	7
Number of girls	2	22	22	4	0	0	0
Number of boys	0	10	19	13	2	2	4

Find the mean, median and mode for each group.

(8) A factory uses five raw materials *A,B,C,D,E* to manufacture a flash-gun. The masses of the materials used in its production are in the ratios 1:1:4:3:1 respectively. The prices of the materials, in kwacha per ton, in the years 1978 and 1980 are given in Table 1.40.

Table 1.40 The masses of the materials used

Raw material	*A*	*B*	*C*	*D*	*E*
1978	4	3	2	5	3
1980	8	5	3	9	8

Taking 1978 as the base year, calculate an index number for the total cost of the raw materials used for the manufacture of the flash-gun in 1980.

1.0 UNIT SUMARY

A frequency table gives the frequency with which each value of an observation occurs in a group. For a discrete data it can be represented by a bar chart. Continuous data can be displayed graphically by

(a) Histogram in which equal areas represent equal frequencies.

(b) A frequency polygon which is obtained by forming the mid points at the top of the histogram blocks

(c) A cumulative frequency curve which gives the number of observations which are less than a stated value of the observed value.

The Standard Deviation. This is the most used measure of spread of observations notation

The standard deviation is calculated as:

1. ungrouped data $\sigma = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$

2. grouped data $\sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}}$

3. Variance is calculated as: $\sigma^2 = \frac{\sum (x - \bar{x})^2}{n}$ for ungrouped data

and $\sigma^2 = \frac{\sum f(x - \bar{x})^2}{\sum f}$ for grouped data.

Your interest here is to learn how to calculate the mean and standard deviation from grouped data and ungrouped data. Their uses and applications are numerals in the next units.

UNIT 2 SAMPLING

2.1 Unit Introduction

Welcome to Unit 2 in which you will be introduced to procedures of sampling. Sampling is a process used in statistical analysis in which a predetermined number of observations are taken

from a larger population. The methodology used to sample from a larger population depends on the type of analysis being performed,

2.2 Unit Aim

The aim of this Unit is to introduce you procedures of Sampling in which you will learn how to make inferences about the whole population.

2.3 Unit Objectives

By the end of the Unit you should be able to:

- Carry out sampling correctly
- Obtain the optimum results, i.e., the maximum information about the characteristics of the population with the available sources at our disposal in terms of time, money and manpower by studying the **sample** values only.
- Estimate means and proportions of the population

Terminology

S – Sample standard deviation biased

$\hat{S}$ - Sample standard deviation which is unbiased

n – Sample size

2.4 Unit Time required

You need 20 hours for this unit

2.5 Unit Topics

2.5.1 The need for sampling

In many cases it is not possible to obtain information about all Members of a population, for the following reasons:

(1) The collecting of the information may destroy the sample, e.g. testing fireworks or electric fuses.
(2) The population may be infinite, e.g. the measurements of a physical constant such as gravity using a particular apparatus.
(3) It may be impracticable to make a measurement for every member of the population, e.g. measuring the length of ants of a particular species.
(4) Even if a measurement could be made for each member of a population, considerations of time and expense usually dictate otherwise.

Random sampling

In order for a sample to be representative of the whole population each member of the population must have an equal chance of being chose. A sample chosen in this way is called a random sample.

The simplest method of selecting a random sample is by using a table of **random numbers** (see Table A3). Such tables are now normally compiled electronically but could be made using any device which gives the digits 0 to 9 with equal probability. It is a prism whose cross-section is a decagon and whose faces are labelled 0 to 9.

Suppose we wish to select two days at random from the month of August using a random number table. Each day is allocated a number 01, 02, 03, etc., up to 31. Note that each day must have the same number of digits in its number so that each number has an equal probability of being chosen. Any starting position can be chosen in Table A3. Suppose you obtain the pairs of digits: (46), (51), 06, (59), (60), 16. The numbers shown in brackets do not correspond to any members of our population and are rejected. Pairs of digits are taken until we have sufficient to give a sample of the required size. In this case the random sample will consist of August 6th and August 16th.

To reduce the amount of numbers which has to be rejected, you can allocate more than one number to each member of the population, on a cyclic basis, thus;

August 1st 01, 32, 64

August 2nd 02, 33, 65 etc.

Each member of the population must have the same number of numbers allocated to it.

Using this method the first two random numbers we obtained before, i.e. 46 and 51, correspond to August 15th and August 20th.

Periodic sampling

This is a method of sampling where every nth member of the population is chosen. It is quicker and easier than using random numbers and might be appropriate for, say, selecting names from an electoral register. In some situations it is not suitable, since, for example, choosing every tenth item produced by machine might coincide with a periodicity of the machine.

Stratified random sampling

As its name implies, this method involves dividing the population into strata. A random sample is then selected from each stratum. The size of each sample is in proportion to the size of the stratum from which it is taken. The advantage of stratified random sampling is that the accuracy of the mean is greater than for an unstratified sample of the same size. Sometimes, however the differences between the strata may be so great that calculation of a mean for the whole population may seem inappropriate.

Drawing a random sample from a discrete distribution

Random numbers can be used to simulate the drawing of a sample from a given distribution. Suppose we wish to choose a random sample of five from a Binomial distribution for which $p = \frac{1}{3}$, n = 3.

The possible values of the random variable and the corresponding probabilities are given in Table 2.1. We can use random numbers to draw a sample if we assign numbers so that the probability of selecting x = 0 is 8/27 etc. these are shown in the third column of Table 2.1, using a cyclic method as before to use as many digits as possible. From the table of random numbers we obtain the pairs of digits. (94), 68, 81, (97), 25, 39, 68

And the corresponding values of r are 1, 3, 2, 1, 1

Other discrete distributions can be treated similar way. Table 2.2 shows the method for a Poisson distribution with = 0.3. In this case the probabilities have to be rounded off, here (arbitrarily) to four decimal places, and so the values of the variable > 5 have been

Table 2.1 Random sample from a Binomial distribution

X	P(X=x)	Random digits
0	$\left(\frac{2}{3}\right)^3 = 8/27$	**01-08, 28-35, 55-62**
1	$\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right) = 12/27$	**09-20, 36-47, 63-74**
2	$\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^2 = 6/27$	**21-26, 48-53, 75-80**
3	$\left(\frac{1}{3}\right)^3 = 1/27$	**27, 53 , 81**

Table 2.2 Drawing a random sample from a Poisson distribution

X	P(X = x)	Cumulative	Random numbers
0	0.7408	0.7408	0000-7407
1	0.2222	0.9630	7408-9629
2	0.0333	0.9963	9630-9962
3	0.0033	0.9996	9963-9995
4	0.0003	0.9999	9996-9998
>5	0.0001	1.0000	9999

Grouped together. The random numbers start from 0000 so that they all have four digits. If from the table of random numbers we obtain, for example, the four digit number 7452, the corresponding value of the variable is 1.

Drawing a random sample from a continuous distribution

Suppose we wish to simulate drawing a random sample from a Normal distribution. In this case the variety is continuous and so can be selected with varying degrees of accuracy. The probability of getting a value of the standard deviate below a certain value is found from Table A1. For example the probability of a standard deviate z where $z < 0.65$ is 0.7422 and this probability could be represented by assigning it the random numbers 0000 to 7421. Similarly, the probability that $z < 0.64$ is 0.7389 and could be represented by the random numbers 0000 – 7388. Then the probability $0.64 < z < 0.65$ would be represented by the random numbers 7389 to 7421. This suggests that to select a random sample from a Normal distribution we can first select four-figure random numbers and then convert them to F(z), the area under the standardized Normal probability distribution, by putting a decimal point in front. From Table A1 the corresponding value of z is found and hence x, the value of the variable. This is indeed the method used apart from an important provision. The four-figure random numbers are 0000 to 9999 and if we take the corresponding value of F(z) as 0.0000 to 0.9999, the values of z will not be symmetrically distributed about their mean, 0. To avoid this we add 0.00005 to each F(z) giving a range of 0.00005 to 0.999 95 which is symmetrically about the mean value of F(z), i.e. 0.5 and gives values of z symmetrical about 0.

Example 2.1

Select a random sample of four values of the variable from a normal distribution mean 10, s.d. 2. (The measurements should be correct to 1 decimal place.) The method is set out in Table 2.3.

Four 4-digit random numbers are taken from Table A3. They are converted to values of F(z) adding a decimal point and 0.00005. From Table A1 the range in which Z lies is found and the values of z are converted to values of x using $z = \frac{(x - \mu)}{\sigma}$, with $\mu = 10$, $\sigma = 2$. (If necessary the range of Z can be reduced by using interpolation in Table A.1) Correct to one decimal place the four randomly chosen values of X are 6.6, 12.8, 14.3 and 11.1.

Table 2.3 Drawing a random sample from a Normal distribution

Random numbers	F(x)	Range of Z	Range of Z
0452	0.04525	-1.70 <z < - 1.69	6.60 < x < 6.62
9197	0.9197	1.40 <z < 1.41	12.80 < x < 12.82
9847	0.98475	2.16<z< - 1.41	12.80 < x < 12.82

A similar method can be used for other continuous distributions. Random numbers are taken and converted to a value of the cumulative distribution function in the same way as they were for the Normal distribution. Then the corresponding value of the variable is found using the c.d.f. For example, a continuous Uniform distribution over the range 1 to 3 has probability density function $f(x) = \frac{1}{2}$, 1< x < 3, and cumulative distribution function $F(x) = \frac{1}{2}(x-1)$. Using the random number 6432 gives 0.643 25 for the value of F(x) so that 25 = ½ (x – 1) giving x = 2,2865

Practical Sampling

In discussing and comparing sampling schemes the following criteria should be borne in mind:

(i) the randomness of the sample,

(ii) time,

(iii) cost,

(iv) Convenience to the person being questioned.

Consider, for example, some of the ways in which a survey might be made in a small town to find out whether parents of children under five consider the nursery school facilities satisfactory. A truly random sample is one in which all of the parents have an equal chance of being chosen. This could be achieved by selecting names from the electoral register using random numbers and interviewing those chosen. This has the disadvantages that (i) the time and expense involved in travelling to peoples' homes would be considerable, especially since more than one visit would be

required if they were out, and (ii) a large number of those chosen would not have children under five and further names would have to be chosen to replace them.

The latter disadvantage could be overcome by selecting the sample from a list of parents with children under five, possibly obtainable from the Local Health Authority. The sampling could be stratified by dividing the parents into groups according to which district of the town they live in so that different income (and possibly ethnic) groups are fairly represented.

An attractive way of overcoming the first disadvantage mentioned above, i.e. the time and expense involved in traveling, would seem to be offered by selecting names at random from the telephone directory. This method is not satisfactory since (i) many people, generally those less well off, do not have a phone, and (ii) people are suspicious of being phoned by strangers.

One of the simplest ways of obtaining a sample would be to stop people with small children in the town centre. This method is quick and cheap but has the disadvantage that the sample is not necessarily random. In practice, however, a compromise may have to be made between obtaining a random sample and considerations of time, cost and convenience.

2.5.2 Sampling distribution

Estimators

The main purpose of taking a sample is to obtain information about the parameters of the population from which the sample is drawn. For example, the mean of a sample gives us an estimate of the mean of the population.

If we took another sample from the population and calculated *its* mean we should be most unlikely to obtain the same value as we did for the first sample. In fact, if we continued taking samples and calculating their means, these means would have a frequency distribution of their own. If we consider all possible values that the mean can take when all possible random samples (of a given size) are drawn from the population then we can form the probability distribution of the sample mean. The random variable defined by this probability distribution, in this case the sample mean, is called an estimator. The probability distribution of an estimator is called its sampling distribution**.** In this unit you look at sampling distributions, in particular that of the mean, and the properties which an estimator needs to give a ‘good’ estimate of a population parameter.

An unbiased estimate of the mean

One of the first requirements of an estimator is that it should be unbiased. This means that the expected value of the sampling distribution of the estimator should be equal to the parameter which is being estimated. It seems intuitively obvious that the variable, X, the mean of a sample, should give an unbiased estimate of u, the population mean. For a sample size n, $\overline{X}$ is calculated from n independent observations of X, which you will call $x_1, x_2, ..., x_n$ you can show that $\overline{X}$ is an unbiased estimate of μ as follows:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

$$E(X) = E\{(X_1 + X_2 + ... + X_n)/n\}$$

$$E\left\{\frac{x_1}{n} + \frac{x_2}{n} + ... + \frac{x_n}{n}\right\}$$

$$= \left\{\frac{E(x_1)}{n} + \frac{E(x_2)}{n} + ... + \frac{E(x_n)}{n}\right\}, \qquad (1)$$

$$= \frac{\mu}{n} + \frac{\mu}{n} + ... + \frac{\mu}{n}$$

$$E(\overline{X}) = \mu$$

2.5.3 The sampling distribution of the mean

You know that each sample will give a different estimate of the population mean. In order to find out how close an estimate will be to the population parameter you need to study the sampling distribution of the mean in more detail and to find out how it is related to the population distribution. You will start by considering a particular numerical example.

Consider a large population which consists of equal numbers of the digits 1, 2 and 3.

The mean of the population is given by

$$\mu = xp(X = x)$$

$$= \frac{1}{3} * 1 + \frac{1}{3} * 2 + \frac{1}{3} * 3$$

$$= 2$$

and its variance by

$$\sigma^2 = \sum_{all\,x} x^2 p(X = x) - \left[xp(X = x)\right]^2$$

$$= \frac{1}{3} * 1^2 + \frac{1}{3} * 2^2 + \frac{1}{3} * 3^2 - 2^2$$

$$= \frac{14}{3} - 4$$

$$= \frac{2}{3}$$

If you take samples, size two, from the population, the possible pairs of values are (1, 1), (1, 2), (2, 1), (2, 2), (2, 3), (3, 2), (3, 3), (1, 3), (3, 1). (Since the population is large you need not concern ourselves with whether or not sampling is with replacement but the theory of this unit is concerned only with sampling with replacement.) Figure 2.1 gives the sample space for the means of the samples. The sample means x with their corresponding probabilities are given in Table 2.1 together with the calculation of $\sum_{all\,\bar{x}} \bar{x}p(\bar{X} = \bar{x})$ and $\sum_{all\,\bar{x}} \bar{x}^2 p(\bar{X} = \bar{x})$.

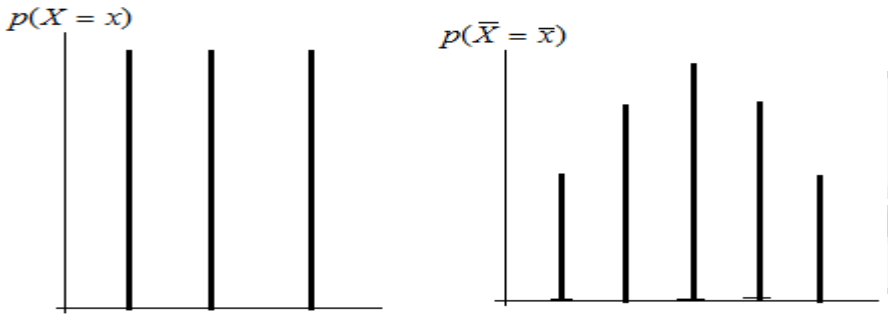
You have mean of sampling distribution $= \sum_{all\,\bar{x}} \bar{x}p(\bar{X} = \bar{x}) = 2$, which, as you expected, is the mean of the original or parent population

	1	2	3
1	1	$\frac{3}{2}$	2
2	$\frac{3}{2}$	2	$\frac{5}{2}$
3	2	$\frac{5}{2}$	3

Figure 2.1 Sample space for the example in Section 2.5.3

Table 2.4 Probability distribution

$\bar{x}$	$p(\bar{X}=\bar{x})$	$\bar{x}p(\bar{X}=\bar{x})$	$\bar{x}^2 p(\bar{X}=\bar{x})$
1	1/9	1/9	1/9
3/2	2/9	1/3	1/9
2	3/9	2/3	4/3
5/2	2/9	5/9	25/18
3	1/9	$\frac{1/3}{2}$	$\frac{1/13}{3}$



Parent population sampling distribution of the mean

Figure 2.2 Probability distribution

Probability distribution for the parent population and the means of the sample of the example in Section 2.5.3 variance of sampling distribution $= \sum_{all\,\bar{x}} \bar{x}^2 p(\bar{X}=\bar{x}) - \left[\sum \bar{x}p(\bar{X}=\bar{x})\right]^2$

$$= \frac{13}{3} - 2^2$$

$$= \frac{1}{3}$$

Figure 2.2 illustrates the probability distribution of the parent population and the sampling distribution of the mean. The diagram shows that the values of the sample mean cluster more closely about the mean that the values of the variable in the parent population. This is reflected

in the values for the variance: the variance of the sampling distribution of the mean is half that of the parent population. This is a particular example of a theorem which is as follows:

Theorem 2.1

If all possible random samples, size n, are drawn (with replacement) from a population, mean μ and s.d. σ, then the means of the samples have a probability distribution known as the sampling distribution of the mean, with mean μ and s.d. $\sigma/\sqrt{n}$. The standard deviation of the sampling distribution of the mean is known as the **standard error (s.e.) of the mean**.

Proof of Theorem 2.1

You have already shown in Section 2.5.3 that E($\bar{X}$) = μ and E($\bar{X}$) = μ and E($\bar{X}$) is the mean of the sampling distribution of the mean. The variance of the sampling distribution is var ($\bar{X}$). You have

$$Var(\bar{X}) = \text{var}\left\{\sum_{i=1}^{n}\frac{X_1}{n}\right\}$$

$$= \frac{1}{n^2}\text{var}\left\{\sum_{1}^{n}X_i\right\}$$

$$= \frac{1}{n^2}\{\text{var}(X_1) + \text{var}(X_2) + ... + \text{var}(X_n)\}$$

$$= \frac{1}{n^2} * n\sigma^2$$

$$= \frac{\sigma^2}{n} \quad (2)$$

Thus the s.d. of the sampling distribution of the mean $\frac{\sigma}{\sqrt{n}}$.

If the sampling is without replacement and the population size is N then equation (2) becomes

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

it can be seen that if n is very much smaller than N then this formula gives approximately the same result as equation (2). In this book it is assumed that the sample size is always much less than the

population size so that even if sampling is without replacement (the more usual situation), equation (2) may still be used.

Whether the sampling is with or without replacement var ($\overline{X}$) decreases as the sample size increases. This confirms what we would expect intuitively; the larger the sample size, the closer the sample mean is likely to be to the population mean.

2.1 UNIT ACTIVITY

(1) A large population consists of equal numbers of the digits 1 and 3. Find the mean and variance of this population.

Find the probability distributed of the mean of samples size three taken from this population and verify that its mean is equal to the population mean and its variance is equal to one-third of the population variance.

(2) The discrete random variance J has the distribution given in Table 2.6

(a) Find the mean μ and variance σ^2 of the distribution. Random samples size two are taken from the distribution. By considering all possible samples, obtain the probability distribution of the mean of such samples. Verify that this distribution has mean μ and variance $½\sigma^2$.

(b) What would be the mean and variance of the distribution of the mean of random samples of size three from the original distribution?

Table 2.6 The discrete random variance

J	- 2	- 1	0	1	2
P(J = j)	0.1	0.2	0.4	0.2	0.1

2.5.4 An unbiased estimator of variance

Just as the sample mean varies from sample to sample, so does the sample variance. Here you are concerned with using a sample to obtain an unbiased estimate of population variance, σ^2.

You might expect the random variable $\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{n}$ to be an unbiased estimator of σ^2 and this can be proved as follows by finding n its expected value:

$$E\left[\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{n}\right]=E\left[\frac{(X_1-\mu)^2}{n}+\frac{(X_2-\mu)^2}{n}+...+\frac{(X_n-\mu)^2}{n}\right]$$

$$=E\left[\frac{(X_1-\mu)^2}{n}\right]+E\left[\frac{(X_2-\mu)^2}{n}\right]+...+E\left[\frac{(X_n-\mu)^2}{n}\right]$$

$$=\frac{1}{n}E\left[(X_1-\mu)^2\right]+\frac{1}{n}E\left[(X_2-\mu)^2\right]+...+\frac{1}{n}E\left[(X_n-\mu)^2\right]$$

$$=\frac{1}{n}\sigma^2+\frac{1}{n}\sigma^2+....+\frac{1}{n}\sigma^2$$

$$=\sigma^2$$

Unfortunately we are not usually in the position of requiring an *estimate* of σ when we know μ. In most cases we have only an estimate of μ, i.e. $\bar{x}$. Using $\bar{x}$ we can calculate the s.d., s, of the sample, from the formula

$$s^2=\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n}$$

However, we cannot use $s^2=\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n}$ to give an unbiased estimate of variance, since the sum of the squares of the deviation of the xi's from $\bar{x}$ is less than the sum of the squares of the deviations from μ and consequently s^2 under estimates σ^2.

You can find an unbiased estimator of σ^2 as follows:

$$\sigma^2 = E\left\{\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{n}\right\}$$

$$= E\left\{\sum\frac{\left[(X_i-\bar{X})-(\mu-\bar{X})\right]^2}{n}\right\}$$

$$= E\left\{\sum\frac{\left[(X_i-\bar{X})^2-2(X_i-\bar{X})(\mu-\bar{X})+(\mu-\bar{X})^2\right]}{n}\right\}$$

$$= E\left\{\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n}-2(\mu-\bar{X})\sum_{i=1}^{n}\frac{(X_i-\bar{X})}{n}+n*\frac{(\mu-\bar{X})^2}{n}\right\}$$ The second term is zero since

$$\sum(X_i-\bar{X})=\sum X_i-\sum\bar{X}=n\bar{X}-n\bar{X}=0$$

So you have $\sigma^2 = E\left\{\sum\frac{(X_i-\bar{X})^2}{n}\right\}+E\left\{(\mu-\bar{X})^2\right\}$ (3)

The first term is $E(S^2)$ where S^2 is the random variable $\sum\frac{(X_i-\bar{X})^2}{n}$. The second term, $E[(\mu-\bar{X})^2]$ or $E\left\{(\bar{X}-\mu)\right\}^2$, is var($\bar{X}$) which was shown in the proof of Theorem 1 to be $\frac{\sigma^2}{n}$. Thus

$$\sigma^2 = E(S^2)+\frac{\sigma^2}{n}$$

Rearranging,

$$\sigma^2 = \frac{n}{n-1}E(S^2)$$

$$= E\left\{\frac{n}{n-1}S^2\right\}$$ (4)

Substitute for S,

$$\sigma^2 = E\left\{\left(\frac{n}{n-1}\right)\left(\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n}\right)\right\}$$

$$\sigma^2 = E\left\{\sum\frac{(X_i-\bar{X})^2}{n-1}\right\}$$ (5)

Equation (5) gives us an unbiased estimator of variance, which you shall denote by $\hat{S}^2$, Where

$$\hat{S}^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1} \qquad (6)$$

and $\hat{S}$ and S are related by $\hat{S} = \sqrt{\left(\frac{n}{n-1}\right)} S$ (7)

The term n – 1 in the denominator of equation (6) is referred to as the number of **degrees of freedom** and is given the symbol ν. The reason for this name is as follows. If you knew μ then the variance could be calculated from an *independent* deviations, using $\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$.

If instead you measure the deviation from $\bar{x}$, then you have only n – 1 independent deviations, as when (n – 1) deviations are given the last deviation can be deduced using the fact that $\sum_{i=1}^{n} (x - \bar{x}) = 0$. One degree of freedom has been 'lost' since only n – 1 of the deviations can be varied independently. Equation (7) shows that, for large n, there is only a small error in takings S^2 rather than $\hat{S}^2$ as an estimate of o^2.

Example 2.2

Five measurements of the volume of acid required in a titration are 25.1, 25.2, 25.2, 25.0, 25.5 cm. Use the results in table 2.7 to obtain estimate for the mean and s.d of the volume of acid required.

First you need to calculate x and s are shown in Table 2.7 and the calculations that follows.

Table 2.7 Five measurements of the volume of acid

x	f_i	u_i	f_iu_i	$f_iu_2^{\ 2}$
25.0	1	0	0	0
25.1	1	1	1	1
25.2	2	2	4	8
25.5	1	5	5	25
	5		10	34

Using A = 25.0, B = 0.1 (that is using thee assumed mean of 25.0)

$$\bar{x} = A + B\frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i}$$

$$= 25 + 0.1 * \frac{10}{5} = 25.2$$

$$s^2 = B^2\left\{\frac{\sum_{i=1} f_i u_i^{\ 2}}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i}\right)\right\}$$

$$s^2 = 0.1^2\left\{\frac{34}{5} - \left(\frac{10}{5}\right)^2\right\} = 0.0280$$

The unbiased estimate of μ is $\bar{x}$ = 25.2 cm.

The unbiased estimate of σ^2 is $\hat{s}^2 = \frac{n}{n-1}s^2 = \frac{5}{4} * 0.0280 = 0.0350$

So $\hat{s}$ = 0.187 cm^3

2.5.5 Relative efficiency and consistency

If you have two estimators which both give an unbiased estimate of a population parameter, you would prefer to use the one for which the sampling distribution is more closely clustered about the true value of the parameter, i.e. the sampling distribution with the smaller variance. This is

the more **efficient** estimator. The ratio of the variances of the two sampling distributions gives a measure of **relative efficiency** on a scale between 0 and 1. For example, both the mean and the median give unbiased estimates of the mean of a Normal distribution. However, it can be shown that for large samples, the sampling distribution of the median is Normal with standard error $\frac{1.25\sigma}{\sqrt{n}}$, and so the median is a less efficient estimator than the mean, which has a standard error of $\frac{\sigma}{\sqrt{n}}$.

A good estimator of variance should also be **consistent**. This means that the larger n the closer the statistic is likely to be to the parameter it estimates. If G is an estimator of y then this is achieved by having

E(G) $\rightarrow$ y (8)

var(G) $\rightarrow 0$ (9)

as n $\rightarrow \infty$

You have already shown that $\bar{X}$ as an estimator of μ satisfies the first of these criteria irrespective of the value of n, and it satisfies the second since var($\bar{X}$) = $\frac{\sigma^2}{\sqrt{n}}$ (see equation (2)).

Example 2.3

A discrete random variable X can take values 0, 1 and 2 only, with respective probabilities $\frac{1}{2}\vartheta$, $1 - \vartheta$ and $\frac{1}{2}\vartheta$, where ϑ is an unknown number between 0 and 1. Let X_1 and X_2 denote two randomly observed values of X. List the possible values of $\{X_1, X_2\}$ that may arise and calculate the probability of each possibility of each possibility; verify that your probabilities sum to unity. By calculating the value of $(X_1 - X_2)^2$ for each possible $\{X_1, X_2\}$ determine the sampling distribution of $(X_1 - X_2)^2$. Hence show that $Y = \frac{1}{2}(X_1 - X_2)^2$ is an unbiased estimator of ϑ and express its sampling variance in terms of 0. Since ϑ is the probability that X will not take the value 1, another possible estimator of ϑ is the proportion of sample values not equal to

1; for a sample of two observations this estimator is given by $Z = \frac{1}{2}N$, where N is the number of observations (0, 1 and 2) not equal to 1. State, giving your reasons, which is Y and Z you would prefer as the estimator of ϑ. Table 2.8 sets out the possible values (x_1, x_2) together with their probabilities. Then you should check that the sum of these probabilities is 1. The third column gives the value of $(x_1 - x_2)^2$ for each {x_1, x_2}. Table 2. 9 gives the sampling distribution of $(X_1 - X_2)^2$. The fourth and fifth columns of this table give $yP(Y = y)$ and y^2 P(Y = y) which are required for

Table 2.8 discrete random variable X

(x_1, x_2)	$P(x_1, x_2)$	$(x_1 - x_2)^2$
0,0	$\frac{1}{4}\vartheta^2$	0
0,1	$\frac{1}{2}\vartheta(1-\vartheta)$	1
0,2	$\frac{1}{4}\vartheta^2$	4
1,0	$\frac{1}{2}\vartheta(1-\vartheta)$	1
1,1	$(1-\vartheta)^2$	0
1,2	$\frac{1}{2}\vartheta(1-\vartheta)$	1
2,0	$\frac{1}{4}\vartheta^2$	4
2,1	$\frac{1}{2}\vartheta(1-\vartheta)$	1
2,2	$\frac{1}{4}\vartheta^2$	0

Table 2.9 Probability discrete random variable

(x_1, x_2)	Y	P(Y = y)	P(Y = y)	P(Y = y)
0	0	$\frac{1}{2}\vartheta^2 + (1-\vartheta)^2$	0	0
1	$\frac{1}{2}$	$\frac{1}{2}\vartheta(1-\vartheta)$	$\vartheta(1-\vartheta)$	$\frac{1}{2}\vartheta(1-\vartheta)$
4	2	$\frac{1}{2}\vartheta^2$	ϑ^2	$2\vartheta^2$
			ϑ	$\frac{3}{2}\vartheta^2 + \frac{1}{2}\vartheta$

Calculating the mean and variance of the sampling distribution of Y (where $Y = \frac{1}{2}X_1 - X_2)^2$.

You find $E(Y) = \sum_{all\,x} yP(Y = y) = \vartheta$, showing that Y is an unbiased estimator of ϑ. The sampling variance of Y is given by

$$\sum y^2 p(Y = Y) - [\sum yp(y = y)]^2 = \frac{3}{2}\vartheta^2 + \frac{1}{2}\vartheta - \vartheta^2$$

$$= \frac{1}{2}\vartheta^2 + \frac{1}{2}\vartheta$$

$$= \frac{1}{2}\vartheta(\vartheta + 1)$$

N, the number of observations in a sample of two which are not equal to 1, is Binomially distributed with n = 2, p = ϑ, q = 1 – ϑ. Thus the mean and variance of the sampling distribution of N are mean = np = 2ϑ, variance= npq = 2ϑ (1 - ϑ)

Since Z = ½ N, the mean and variance of its sampling distribution are mean = ½ x 2ϑ = ϑ (see equation (4)) variance = $(½)^2$ x 2ϑ (1 – ϑ) = ½ ϑ (1 – ϑ).

Since the sampling distribution of Z Has the smaller variance, Z is the preferred estimator.

2.2 UNIT ACTIVITY

(1) Each trial of a random experiment has probability p (0 < p < 1) of yielding a success. In n_1 independent trials of the experiment the number of successes obtained was r_1

Write down an unbiased estimate, p_1 of p and find its standard error in terms of n_1 and p. In a further n_2 independent trials of the same experiment the number of successes obtained was r_2. Let p_2 denote the unbiased estimate of p from these n_2 trials. Verify that $\frac{1}{2}(p_1 + p_2)$ is an unbiased estimate of p, and find its standard effort in terms of n_1, n_2 and p.

Determine the range of values of the ratio n_1/n_2 for which the estimate $\frac{1}{2}(p_1 + p_2)$ is to be preferred to each of p_1 and p_2

(2) Explain the terms (a) unbiased estimator, (b) consistent estimator, and indicate how possession of these properties helps to ensure a 'good' estimator.

Let $X_1, X_2, ..., X_n$ be a random sample from some population with distribution f(x) and let $S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$, for n > 1

Show that S^2 is an unbiased and a consistent estimator for σ^2, the population variance.

(3) Explain what is meant by the sampling distribution of estimator.

Explain what you understand by (a) an unbiased estimator, (b) a consistent estimator, (c) the relative efficiency of two estimators of the same parameter.

In order to estimate the mean μ of a population, random observation x_1, x_2, x_3 are taken of are taken of a random variable X which has variance σ^2. Find the relative efficiency of the two estimators μ_1 and μ_2 where

$$\hat{\mu}_1 = \frac{x_1 + x_2 + x_3}{3}, \quad \hat{\mu}_2 = \frac{x_1 + x_2 + x_3}{3}$$

(4) A random sample of n_1 observations is made from a population with unknown mean u and variance σ^2. For this sample the mean x_1 and variance s_1^2 are calculated. A second sample, size n_2, has mean x_2 and variance s_2^2. Show that an unbiased estimate of the population mean μ is given by

$$\frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

and an unbiased estimate of the population variance σ^2 is given by

$$\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

(5) To estimate the mean of a population, two observations x_1 and x_2 are made of the random variable X, which has mean μ and variance o^2. Show that $\hat{\mu} = kx_1 + (1-k)x_2$ is an unbiased estimator of μ and find the value of k for which this estimator is most efficient.

2.5.6 The central limit theorem

If a sample, size n, is taken from a normal population, the sampling distribution of the mean is also normal. Even if the original population is not normal, it can be shown that, as n increases, the sampling distribution of the mean approaches a normal distribution.

This result is called the central limit theorem and is of fundamental importance. It is usually assumed that for values of n > 30 the sampling distribution of the mean is Normal.

Example 2.4

What is the probability that the mean of 100 digits taken from a random number table is greater than 5.0? Random numbers form a discrete Uniform distribution. By symmetry mean = 4.5

The variance for a Uniform distribution where the variable takes the values 1, 2, 3, . . . n was shown to be $\frac{1}{12}(n^2 - 1)$. The numbers 0, 1, 2, . . . , 9 will have the same

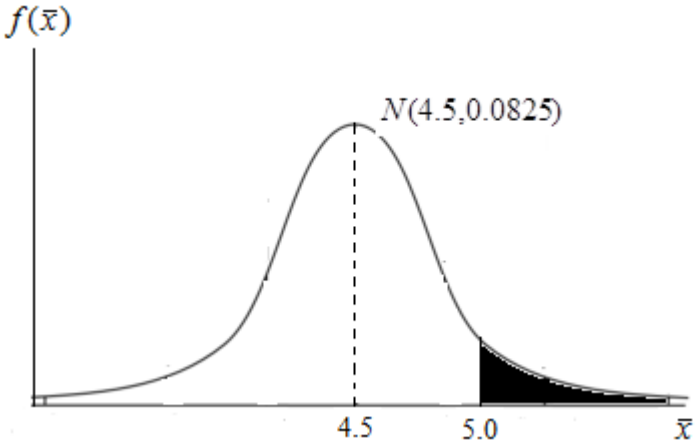


Figure 2.3 Graph to illustrate Example 2.4

Variance as 1, 2, 3, . . . , 10, since a change of origin does not alter the variance. Thus variance of random numbers = $\frac{1}{12}(10^2 - 1) = 8.25$. Using equation (2), the variance of the sampling distribution of the mean of 100 random digits is $\frac{\sigma^2}{n} = \frac{8.25}{100} = 0.0825$, and the mean of the sample distribution of the mean is 4.5. The central limit theorem tells us that the sampling distribution of the mean is Normal as shown in Figure 2.3. The s.d. of this distribution is $\sqrt{0.0825} = 0.287$. The probability that the mean of 100 members is greater than 5 is given by the shaded area. $z = \frac{5.0 - 4.5}{0.287} = 1.74$

Required probability = $p(z > 1.74)$

= 1 – P(Z < 1.74)

= 1 - 0.9591

= 0.0409

2.3 UNIT ACTIVITY

(1) The body length of a certain species of insect is normally distributed with mean 3.1 mm, s.d. 0.2 mm.

(a) What is the probability that n insect chosen at random has a body length greater than 3.12 mm?

(b) What is the probability that the mean body length of a sample of (i) 9 insects, and (ii) 100 insects is greater than 3.12 mm?

(2) Explain what is meant by the standard error of the mean.

For each of the following large populations (a) – (c), three possible standard deviations (i)-(iii) are given. In each case, selected the standard deviation you believe to be the most reasonable and use it to obtain a range of values within which the mean of a sample of 100 values is like to lie with 95% probability.

(a) The weights in kg of adult males. Mean is 66 kg, s.d:

(i) 1, (ii) 8, (iii) 16.

(b) The salaries of typists. Mean is 795, s.d:

(i) 110, (ii) 425, (ii) 1410.

(c) The number of matches in a box of nominal contents 50. Mean is 51.6, s.d:

(i) 0.6, (ii) 5.3, (iii) 20.2.

(3) In a certain examination with a very large entry, the percentage marks obtained by the male candidates were found to follow a Normal distribution with a mean of 54 and a standard deviation of 16. Let $\overline{X}$ denote the mean of the percentage marks scored by a random sample of four male candidates. What is the sampling distribution of $\overline{X}$? Calculate the probability that $\overline{X}$ will exceed 70 and the value c such that there is a probability of 0.95 that $\overline{X}$ will be within c marks of the mean mark of 54.

In the same examination the percentage marks obtained by the female candidates were found to follow a Normal distribution with a mean of 59 and a standard deviation of 20. Let $\overline{Y}$ denote the mean of the percentage marks scored by a random sample of five female

candidates. What is the sampling distribution of $\overline{Y} - \overline{X}$? Calculate the probability that the value of $\overline{Y}$ will be greater than the value of $\overline{X}$.

2.5.7 An unbiased estimator of population proportion

If the probability that a member of a population possesses an attribute is p then we know from that X, the number of members of sample size n which possess this attribute, is Binomially distributed, B(n, p). We might expect P s = X/n, the proportion in the sample which shows the attribute, to be an unbiased estimator of p and this can be shown as follows:

$$\begin{aligned} E(P_s) &= E(\frac{X}{n}) \\ &= E(X)/n \\ &= np/n \\ &= p \end{aligned} \qquad (10)$$

You can also find the variance of the sampling distribution of P_s since

$$\begin{aligned} Var(P_s) &= Var(X/n) \\ &= Var(X)/n^2 \\ &= np(1-p)/n^2 \\ &= p(1-p)/n \end{aligned} \qquad (11)$$

When n is large the sampling distribution of X tends to a Normal distribution. Thus the sampling distribution of P_s also tends to a Normal distribution with mean p/n and variance $\sqrt{\{p(1-p)/n\}}$. The standard deviation of this sampling distribution, $\sqrt{\{p(1-p)/n\}}$, is called the standard error of a proportion.

2.4 UNIT ACTIVITY

(1) The number of days that each of five employees (A, B , C, D, E) in an office was absent from work during a year is shown in Table 2.10

Table 2.10

Employee	A	B	C	D	E
Number of days absent	10	6	0	4	0

Calculate the mean u and the variance σ^2 f the numbers of days these employees were absent from work.

(a) Three of these employees are selected at random without replacement. Let X denote the mean number of days absent for the chosen three employees.

(i) Determine the sampling distribution of X and display it in a table.

(ii) Determine whether or not X is an unbiased estimator of μ.

(iii) Find the variance of X and verify that it is equal to one-half of the variance of the sample mean if the three employees are chosen with replacement.

(2) A random sample x_1, x_2 is drawn from a distribution with mean μ and standard deviation σ. state the mean and standard deviation of the distribution of (a) $x_1 + x_2$, (b) $x_1 - x_2$, (c) x

A student's performance is equally good in two subjects. The marks she might be expected to score in each subject may be treated as independent observations drawn from a Normal distribution with mean 45 and standard deviation 5. Two procedures might be used to decide whether to give the student an overall pass. One is to demand that she passes separately in each subject, the pass mark being 40; the other is to require that her mean mark in the two subjects exceeds 40. Find the probability that the standard will obtain an overall pass by each of these procedures.

(3) The random variable $X_1, X_2, ..., X_n$ are independent and each has a Normal distribution with mean u and variance 1. The random variable X is defined to be $(X_1, X_2, ..., X_n)/n$. Determine, in terms of n, the value u which is such that, when u = 0, the probability of X exceeding v is 0.05. For this value of v it is desired that the probability of X being less than v when u = 0.2 should be at most 0.10. Calculate the smallest possible value of n which satisfies this requirement.

(4) Explain what is meant by the sampling distribution of the mean, and discuss briefly the properties of the distribution.

By considering all possible outcomes, find the sampling distribution of the total score obtained when two unbiased dice are thrown. Find the mean and variance of the distribution. If 2n dice are thrown, and n is large, what distribution is a close approximation to that of the total score?

(5) Explain what is meant by the expectation of a random variable.

A random sample is drawn from a large population in which a proportion p have a certain rare disease. Sampling continues until a predetermined number a of the sample are found to have the disease, and at this stage the sample size is r. find the probability distribution of r, and show that $\frac{a-1}{r-1}$ is an unbiased estimate of p, i.e. that $E\left[\frac{a-1}{r-1}\right] = p$

2.0 UNIT SUMMARY

Unbiased estimate of population mean, μ

Unbiased estimate of popular variance, σ^2,

Unbiased estimate of population proportion, p, $P_S = {}^{x}\!/_{n}$

The sampling distribution of the mean has mean μ and variance σ^2 where μ and σ^2 are the population mean and variance respectively. When n is large this sampling distribution approximates to a normal distribution; this result is known as the central limit theorem.

The sampling distribution of the sample proportion has mean p and variance $p(1-p)/n$, where p is the population proportion. When n is large this sampling distribution approximates to a Normal distribution.

The sampling distribution of the mean

Using the table of random numbers, Table A3, as the population, take 200 samples of ten digits each. Calculate the mean of each sample and present the results in a frequency table with classes 1.4- 2.0, 2.1, 2.7, 2.88 3.4, 3,5-4.1, 4.2- 4.8 etc. (These classes have been chosen because they are symmetrically distributed about the expected mean.) Illustrate by a frequency polygon.

Calculate:

(i) the mean variance for the original population, i.e. a discrete Rectangular distribution for which P(r) = 1/10 for r = 0, 1, 2, . . . ,9;

(ii) the mean and variance for the frequency table of the sample means.

Verify that (a) the mean of the means of the samples is approximately equal to the mean of the original population, (b) the variance of the distribution of the means is approximately equal to one-tenth of the variance of the original population (since in this case n = 10). (The agreement is only approximate since not all of the possible samples are considered.)

The population from which the sample were drawn has mean 4.5, variance 99/12 so that, using the central limit theorem, we expect the sampling distribution of the mean to be approximately Normal with mean 4.5, variance $\frac{99}{12*10} = 0.825$. Using these parameters calculate the expected frequencies for classes 1.4- 2.0, 2.1- 2.7, 2.7 etc, for the means of 200 samples. Compare with the observed frequencies by drawing a frequency polygon using the same axes as before.

UNIT 3 CONFIDENCE LIMITS

3.1 Unit Introduction:

Welcome to Unit 3 in which you will learn how to construct confidence interval (CI). You will learn types of interval estimates, computed from the statistics of the observed data that might contain the true value of an unknown population parameter. Most commonly, the 95% confidence level is used. However, other confidence levels can be used, for example, 90% and 99%. An interval estimate gives a range of values which has a certain probability of containing the population parameter. This unit describes the way in which some interval estimates are calculated from the appropriate sampling distribution.

In the previous unit you learnt that the mean of a sample gave an unbiased estimate of the population mean. Such an estimate which is the form of a single value is sometimes called a point estimate. This is to distinguish it from an alternative form of giving an estimate called an interval estimate. An interval estimate gives a range of values which has a certain probability of containing the population parameter. In this unit you will study methods of constructing confidence intervals from the appropriate sampling distributions.

3.2 Unit Aim

The aim of this Unit is teach you on how to construct confidence interval (CI).

3.3 Unit Objectives

By the end of the unit you should be able to:

- Get good estimates of the unknown population parameter.
- Construct confidence intervals.

Terminology

$\hat{s}$: Sample standard deviation, this is unbiased estimator of the population standard deviation

s : Sample standard deviation biased

σ^2 : Population variance

σ : Population standard deviation

$\vec{x}$: Sample mean

μ : Population mean

$Z_{\alpha/2}$: Degree of confidence, table value

3.4 Unit Time required

You need 20 hours for this unit

3.5 Unit Topics

3.5.1 Confidence limits of the mean (σ known)

You will start by calculating an interval estimate for the population mean, μ in the situation where the population standard deviation, σ, is known but μ is not. If n is large, the central limit theorem tells you that the sampling distribution of X will be normal, as shown in Figure 1, with mean μ and s.d. $\sigma/\sqrt{n}$. You can use this distribution to find arrange of values within which $\overline{X}$ will lie with a certain probability. Suppose you wish to calculate a range of values for $\overline{X}$ so that $\overline{X}$ lies within this range with 95% probability. Let $\bar{x}_u$ and $\bar{x}_l$ be the upper and lower values of this range respectively. They are shown in Figure 1 placed symmetrically on either side of u. the shaded area is 0.95, so that the area above $\bar{x}_u$ is 0.025 and0 below it 0.975. Using Table A1, the standard deviates corresponding to $\bar{x}_l$ and $\bar{x}_u$ are – 1.96 and 1.96 respectively, so that

$$1.96 = \frac{\bar{x}_U - \mu}{\sigma/\sqrt{n}}, \text{ giving } x_U = \mu + \frac{1.96\sigma}{n} \text{ and } -1.96 = \frac{\bar{x}_L - \mu}{\sigma/\sqrt{n}}, \text{ giving } x_L = \mu - \frac{1.96\sigma}{n}$$

Thus there is a 95% probability that $\bar{X}$ lies in the range $\mu - \frac{1.96\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{1.96\sigma}{\sqrt{n}}$

This inequality can be rearranged to give an interval which includes μ with probability of 95%

$$\bar{X} - \frac{1.96\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{1.96\sigma}{\sqrt{n}}$$

In this expressing $\bar{X}$ is a random variable. If, for a particular sample, $\bar{X}$ takes the value $\bar{x}$, you can calculate an interval $\bar{x} - \frac{1.96\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{1.96\sigma}{\sqrt{n}}$. (1)

This interval is called the 95% confidence interval of the mean and its end-points are called the 95% confidence limits of the mean. μ is fixed, and so for a particular sample the interval calculated in (1) either does or does not include μ and we do not know which is the case. What we *do* know is that if we calculate confidence intervals by this method then is a probability of 95% that the interval does include μ. Other confidence limits are sometimes used, e.g. 99 and can be found by using the appropriate value of the standard deviate. The 99% confidence limits are

$$\bar{x} - \frac{2.58\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{2.58\sigma}{\sqrt{n}} \quad (2)$$

Notice that the confidence interval must increase if we are to become more confident that it contains μ. Equations (1) and (2) can also be used for a small sample from a Normal population, since in this case the sampling distribution is Normal.

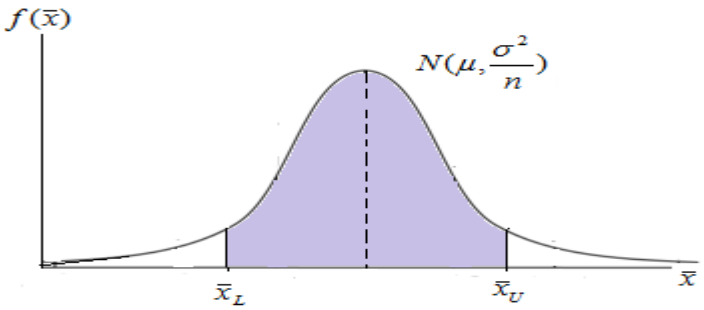


Figure 3.1 Graph to illustrate normal distribution

Example 3.1

The standard deviation for a method of measuring the concentration of nitrate ions in water is known to be 0.05 ppm. If 100 measurements give a mean of 1.13 ppm, calculate the 95% confidence limits for the true mean.

Solution

Using $\bar{x}$ = 1.13, n = 100 and σ = 0.05,

$$1.13 - \frac{1.96 * 0.05}{100} < \mu < 1.13 + \frac{1.96 * 0.05}{\sqrt{100}}$$

$$1.12 < \mu < 1.14$$

3.1 UNIT ACTIVITY

(1) A machine produces washers whose diameter is known to be normally distributed with a s.d. of 0.04 mm. in order to find the mean diameter of the washers produced, a random sample of nine washers is taken whose mean diameter is found to be 3.14 mm. Calculate (a) 95%, (b) 98% confidence limits for the mean diameter of washers produced by the machine.

(2) The masses of sweets produced by a machine are known to have a standard deviation of 0.5 g. A sample of 50 sweets has a mean mass of 15.21 g. Calculate a 99% confidence interval for the mean mass of sweets produced by the machine. Why is it not necessary to assume that the masses are normally distributed in this calculation?

3.5.2 Confidence limits for a large sample (σ unknown)

In practice it is unusual to know σ^2 but not μ. However, you can obtain an unbiased estimate $\hat{s}^2$ of σ^2 from the sample. The estimate varies from sample to sample but for large samples the variation is so small compared with $\hat{s}^2$ itself that we can replace σ^2 by $\hat{s}^2$ in equation (1). This gives for the 95% confidence limits of the mean

$$\bar{x} - \frac{1.96\hat{s}}{\sqrt{n}} < \mu < \bar{x} + \frac{1.96\hat{s}}{\sqrt{n}} \qquad (3)$$

Since $\hat{s} = s\sqrt{\frac{n}{n-1}}$ (see equation (7)), this expression can also be written as

$$\bar{x} - \frac{1.96s}{\sqrt{n-1}} < \mu < \bar{x} + \frac{1.96s}{\sqrt{n-1}} \qquad (4)$$

Example 3.2

Fifty measurements of the acceleration due to gravity, g, had a mean value of 9.8 ms -2 and a s.d. of 0.75 s^{-2}. What are the 95% confidence limits for g?

Solution

Using equation (2) with $\bar{x}$ = 9.8, s = 0.75 and n = 50, the 95% confidence limits of the mean are

$$9.8 - \frac{1.96*0.75}{\sqrt{49}} < \mu < 9.8 + \frac{1.96*0.75}{\sqrt{49}}$$

Therefore, the 95% confidence interval is given by: 9.6 < μ < 10.0. That is you are 95% confidence that the true mean for the population is in the interval [9.6, 10]

Example 2.3

In the previous example how many measurements would be necessary to reduce the 95% confidence limits to $9.7 < \mu < 9.9$?

You have

$$\frac{1.96s}{\sqrt{n-1}} = 0.1$$

Rearranging

$$n - 1 = (1.96*0.75/0.1)^2$$

n = 217

3.2 UNIT ACTIVITY

(1) Fifty children were selected at random from the pupils at a school and each was asked how many hours a week he or she spent in watching television. The mean of the sample was 17.2 hours and the standard deviation was 5.3 hours. Calculate the (a) 95%, (b) 98% confidence interval for the mean number of hours spent a week in watching television for the population consisting of all the children in the school.

(2) Fifty boxes of matches were selected at random from a large carton of such boxes. The numbers of matches in each of the 50 boxes were counted and the mean and standard deviation of these numbers were found to be 48 and 0.5 respectively. Between what limits would you expect the mean for all of the boxes in the carton to lie with 0.95 probability?

(3) On 1 January 100 new Eternity light bulbs were installed in a certain building, together with a device which records how long each light bulb is in use. By 1 March all 100 bulbs had failed, and the recorded lifetimes, t (in hours of use since 1 January) are summarised in Table 3.1.

Obtain values for the sample mean and the sample standard deviation for this set of data. (Assume measurements are to the nearest second.) Obtain values for the sample mean and the sample standard deviation for this set of data. (Assume measurements are to the nearest second.)

Assuming that the bulbs constituted a random sample of Eternity light bulbs, obtain a symmetric 99% confidence interval for the mean lifetime of Eternity light bulbs.

Table 3.1 Random sample of Eternity light bulbs

Time	frequency
$0 < t < 50$	31
$50 < t < 100$	24
$100 < t < 150$	20
$150 < t < 200$	13
$200 < t < 300$	11
$300 < t < 500$	1

Table 3.2 The diameters of the heads of 100 rivets

Diameter (mm)	11.1	11.2	11.3	11.4	11.5	11.6	11.7
f	1	6	24	33	22	12	2

(4) The random variable X takes values x_i with associated frequencies f_i (i = 1, 2, . . ., n) and the mean of these values is $\bar{x}$, $N = \sum f$. If c is a constant prove the formula

$$\frac{1}{N}\sum f(x-\bar{x})^2 = \frac{1}{n}\sum f(x-c)^2 - (\bar{x}-c)^2$$

The diameters of the heads of 100 rivets from a production line were measured correct to the nearest 0.1 mm and the results were as in Table 3.2.

Using an assumed mean of 11.4, calculate the sample mean and standard deviation of the diameters.

Determine 95% confidence limits for the population mean diameter, stating any assumptions that you make.

3.5.3 Confidence limits for a small sample from a Normal population (σ unknown)

If small samples, size n, are drawn from a Normal population, their means are normally distributed with mean μ and s.d. $\frac{\sigma}{\sqrt{n}}$. However you can no longer use the confidence limits derived for large samples because this derivation rested on the fact that $\frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}}$ was normally distributed. For large samples, even when you did not know σ, you could assume that $(\bar{x}-\mu)/(\hat{s}/\sqrt{n})$ was still approximately Normal. This is no longer true for small samples because the variation of s from sample to sample is too large to be ignored. The statistic $\frac{(\bar{x}-\mu)}{\hat{s}/\sqrt{n}}$ is called 't' and its distribution is called the **Student t-Distribution** Its form depends on the 'degrees of freedom' ν, mentioned in connection with the Estimation of σ (see Section 2.5.4) and, as in that case, we have ν = n – 1.

For large n, the t-distribution

Approximates to the Normal distribution, a property which was used in Section 3.5.2. As n decreases the Distribution remains bell-shaped but becomes more spread out as shown in Figure 2.2. This spread Reflects the uncertainty introduced because of the variation in $\hat{s}$.

To denote the dependence of t on the degrees of freedom you write

$$t_{n-1} = \frac{\bar{x} - \mu}{\hat{s}/\sqrt{n}} \qquad (5)$$

Since $\hat{s} = \sqrt{\left(\frac{n}{n-1}\right)}s$

you can also write $t_{n-1} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$ (6)

Because n can take any value, tables for the t-distribution are not given in the form of cumulative probability values. Instead they are given in the form shown in Table A4. This table gives what are called the percentage points of the t-distribution for different values of n.

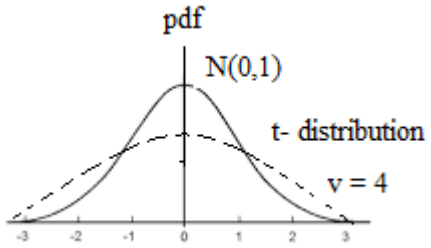


Figure 3.2 Graph comparing the standard Normal distribution and the t-distribution with v = 4.

The percentage points give a value of t outside which t lies with a certain probability. For example, if n = 8 (== v = 7) then, using Table A4, 5% of the values of t lie outside the range – 2.36 to + 2.36 and consequently 95% of the values lie within this range. In general if we require the α % confidence limits for n- 1 degrees of freedom, the appropriate value of t will lie in the row for v = n – 1 and the column for (100 – α)%. We can denote this value by $t_{n-1,(100-\alpha)\%}$. Equation (2) for the 95% confidence limits becomes

$$\bar{x} - t_{n-1,,5\%} * \frac{s}{\sqrt{n-1}} < \mu < \bar{x} + t_{n-1,5\%} * \frac{s}{\sqrt{n-1}} \qquad (7)$$

and in general the α % confidence limits will be

$$\bar{x} - t_{n-1,,(100-\alpha)\%} * \frac{s}{\sqrt{n-1}} < \mu < \bar{x} + t_{n-1,(100-\alpha)\%} * \frac{s}{\sqrt{n-1}} \qquad (8)$$

The last line of the t – table, v = ∞, gives the percentage points of the Normal distribution.

Example 2.4

Ten measurements of the zero error on an ammeter yield the results 0.13, - 0.09, 0.06, 0.15, - 0.02, + 0.03, + 0.01, - 0.02, - 0.07, + 0.05 A. Calculate the 95% confidence limits of the mean zero error. (Assume the errors are normally distributed.)

$$\bar{x} = \frac{1}{10}[0.13 - 0.09 + 0.06 + 0.15 - 0.02 + 0.03 + 0.01 - 0.02 - 0.07 + 0.05] \quad \bar{x} = 0.023$$

You now calculate the sample standard deviation:

$$s^2 = \frac{1}{10}[0.013^2 + 0.09^2 + 0.06^2 + 0.15^2 + 0.02^2 + 0.03^2 + 0.01^2 + 0.02^2 + 0.07^2 + 0.05^2] -$$

$$\left(\frac{0.23}{10}\right)^2$$

$$s = 0.0742$$

We have no = 10, v = 9. From Table A4, 9.5% = 2.26. Using equation (7), the 95% confidence limits of the mean are

$$0.023 - 2.26 * \frac{s0.742}{\sqrt{9}} < \mu < 0.023 + 2.26 * \frac{0.0742}{\sqrt{9}} \qquad -0.033 < \mu < 0.097$$

3.3 UNIT ACTIVITY

(1) The diameters of 25 steel rods are measured and found to have a mean of 0.980 cm and a standard deviation of 0.015 cm. Find 99% confidence limits for the population mean

(2) The masses, in grams, of thirteen ball bearings taken at random from a batch are
21:4, 23.1, 25.9, 24.7, 23.4, 24.5, 25.0, 26.9, 26.4, 25.8, 23.2, 21.9
Calculate a 95% confidence interval for the mean mass of the population, supposed Normal, from which these masses were drawn.

(3) A random sample of sic eggs taken from a day's production of a poultry farm has the following masses, measured in grams: 51.2, 52.6, 53.1, 53.2, 53.2, 54.7. Making suitable

assumption, which should be stated, find 95% confidence limits for the mean mass of the eggs produced that day.

(4) Twelve cotton threads are taken at random from a large batch, and the breaking strengths are found to be 7.41, 7.01, 8.34, 8.29, 8.08, 6.60, 6.59, 7.39, 4.72, 8.65, 8.51 and 8.01 respectively. Assuming the breaking strengths of the threads form a Normal distribution, find 95% confidence limits for the mean breaking strength of threads in the batch.

3.5.4 Confidence interval of a population from a large sample

In Section 2.5.7 you saw that the sampling distribution of *Ps,* the proportion of a sample possessing a given attribute, was Normal with mean *p/n* and variance *p(1 – p)/n*, where p is the proportion of the population possessing the attribute and *n* is the sample size. We can develop an argument similar to that of Section 3.5.1 to give an equation corresponding to equation (1) for the 95% confidence interval for *p*: $p_s - 1.96\sqrt{[p(1-p)/n]} < P < p_s + 1.96\sqrt{[p(1-p)/n]}$

Since we do not know the value of *p* in order to estimate the standard error of the proportion, $\sqrt{[p(1-p)/n]}$, you must replace *p* by p_s in this term to give as the approximate 95% confidence interval of a proportion

$$p_s - 1.96\sqrt{[p_s(1-p_s)/n]} < P < p_s + 1.96\sqrt{[p_s(1-p_s)/n]} \qquad (9)$$

Example 3.5

An opinion poll is taken as to how an electorate will vote in a forthcoming referendum. Out of a random sample of 100, 40 says *'yes'* and 60 say *'no'*. What is the 95% confidence interval for the proportion of the population who will vote *'yes'*?

We have $n = 100$ and $p_s = 40/100 = 0.4$. Substituting these values in equation (1) gives the 95% confidence interval of the population proportion voting *'yes'* as

$0.4 - 1.96\sqrt{[0.4 + 0.6/100]} < P < 0.4 + 1.96\sqrt{[0.4 * 0.6/100]}$ $\quad 0.3 < p < 0.5$

It should be noted that, provided the sample is sufficiently small compared with the population for *p* to be regarded as constant, the accuracy with which we can estimate a proportion depends only on the absolute size of the sample and not on its size relative to the whole population.

Example 3.6

How large should the sample have been in the previous example to reduce the confidence interval to 1%?

The confidence limits required are 0.4 ± 0.005 giving $1.96\sqrt{\left(\frac{0.4*0.6}{n}\right)} = 0.005$, n = 36 880

3.4 UNIT ACTIVITY

(1) Out of a random sample of 50 children from a school, 24 were found to have been vaccinated against whooping-cough. Calculate the 95% confidence limits for the proportion of children at the school who have been vaccinated against whooping-cough.

(2) At a school of 1200 pupils it is found that a random sample of 40 pupils contains five who are left-handed. Find the 95% confidence limits for

(a) the proportion of pupils in the school who are left-handed,

(b) the number of pupils in the school who are left-handed.

(3) In a sample of 400 shops taken in 1972, it was discovered that 136 of them sold carpets at below the list prices which had been recommended by manufacturers.

(a) Estimate the percentage of all carpet-selling shops selling below list price.

(b) Calculate the 95% confidence limits for this estimate, and explain briefly what this means.

(c) What size sample would have to be taken in order to estimate the percentage to within $\pm$ 2%?

(4) Table 3.3 shows the number of cases of and deaths from diphtheria at the City of Ndola Hospital in the years 1900- 10 under antitoxin and ordinary treatments.

Obtain an estimate (with an estimate of the standard error) of the probability that a patient treated with antitoxin died.

Table 3.3 Deaths from diphtheria at the City of Ndola

Treatment	Cases	Deaths
Antitoxin	228	37
Ordinary	337	28

Give an approximate 95% symmetric confidence interval for this probability. It was required to estimate this probability with a standard error of 0.01. Estimate, to the nearest 100, the number of patients that should have been treated.

Give an approximate 95% symmetric confidence interval for the probability that a patient treated with the ordinary treatment died.

Comment on the relative effectiveness of the antitoxin treatment and the ordinary treatment, stating, in coming to your conclusions, any assumption you have made about the allocation of patients to the different treatments.

(5) (a) When an object is weighed on a chemical balance the readings obtained are subject to random errors which are known to be independent and Normally distributed with mean zero and standard deviation 1 mg. A certain object is to be weighed nine times on such a balance and the mean of the nine readings is to be calculated.

Find the probability that the mean of the nine readings will be within 0.5 mg of the true weight of the object.

(b) Another weighing device is undergoing tests to determine its accuracy. A certain object of known true weight 50 mg was weighed ten times on this device and the readings in mg were 49, 51, 49, 52, 49, 50, 52, 51, 49, 48

(i) Calculate an unbiased estimate of the variance of the errors in readings using this device.

(ii) Calculate 95% confidence limits for the mean error in readings using this device.

(6) Discuss briefly the relative merits of estimating an unknown parameter by means of either a single value or a confidence interval.

(a) Crates of bananas are packed in the West Indies with a nominal net mass of 55 kilograms. However, on their arrival in Liverpool, this has usually decreased due to ripening and shrinkage. A large batch of such crates has just arrived aboard *SS Gauss*. A random sample of twelve crates is selected and their net masses are recorded in kilograms as listed below. 56.4, 52.1, 49.5, 56.4, 48.1, 54.5, 47.8, 58.0, 48.4, 53.9, 46.7, 56.0
Assuming that this sample came from an underlying Normal population with variance 16, calculate a 95% confidence interval for the mean mass of the population.

(b) The drained masses, in kilograms, of ten catering size tins of peaches taken at random from a batch are 2.57, 2.05, 1.65, 2.62, 2.44, 1.48, 2.31, 1.58, 2.60, 1.85
Assuming that this sample came from an underlying Normal population, calculate a 95% confidence interval for the mean of the population.

(7) (a) In order to calculate a confidence interval for the mean mass of packets of butter produced by a machine, a random sample of ten packets is taken. These have masses (measured in kg) of $x_1, x_2, ..., x_{10}$ such that

$$\sum_{i=1}^{10} x_i = 2.57, \sum_{i=1}^{10} x_i^2 = 06610$$

Calculate 95% confidence limits for the mean.

(b) If it is known that the standard deviation of the mass of a packet of butter is 0.008kg, what is the least number of packets that would need to be sampled in order to give a 95% confidence interval for the mean mass whose width is less than 0.002kg?

(8) A random sample of 600 was chosen from the adults living in a town in order to investigate the number x of days of work lost through illness. Before taking the sample it was decided that certain categories of people would be excluded from the analysis of the number of working days lost although they would not be excluded from the sample. In the sample 180 were found to be from these categories. For the remaining 420 members of the sample x = 1260 and $x^2 = 46000$.

(a) Estimate the mean number of days lost through illness, for the restricted population, and give a 95% confidence interval for the mean.

(b) Estimate the percentage of people in the town who fall into the excluded categories, and give a 99% confidence interval for the this percentage.

(c) Give two examples, with reasons, of people who might fall into the excluded categories.

(9) (a) A random sample of n observations from a population distribution had the values $x_1, x_2, ..., x_{10}$, whose mean is x. Show that for any value of c,

$$\sum_{i=1}^{n}(x_i - c)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = n(c - \bar{x})^2$$

Hence find the value of c for which $\sum_{i=1}^{n}(x_i - c)^2$ is a minimum.

(b) A random sample of 12 values from a normal distribution, whose mean μ and variance σ^2 are unknown, were such that

$$\sum_{i=1}^{12} x_i = 5472, \ \sum_{i=1}^{12}(x_i - 450)^2 = 1620$$

(i) Calculate unbiased estimate of μ and σ^2.

(ii) Determine a 95% confidence interval for μ.

(iii) Given that (451, 463) was a 95% confidence interval for μ based on another random sample of 12 values from the same Normal distribution, deduce the corresponding unbiased estimate of μ and σ^2 from this sample

(10) A random sample of 500 fish is taken from a lake, marked, and returned to the lake. After a suitable interval a second sample of 500 is taken and 25 of these are found to be marked. By considering the number of marked fish in the second sample, estimate the number of fish in the lake and, by considering a confidence interval for the proportion of marked fish in the lake, obtain a 95% confidence interval for the number of fish.

(11) (a) A school dental service wishes to estimate the average number of teeth with fillings of the 600 pupils at a Secondary School. Explain how you would select a sample of pupils for a dentist to inspect.

(b) A survey of 3000 randomly chosen households in England revealed that 250 had moved during the previous year. Estimate 95% confidence limits for the percentage of households moving during the year considered.

(12) An estimate is required of the proportion of a large number of consumers who are likely to purchase a particular brand of butter. Determine the smallest sample size that should be taken in each of the following situations:

(a) The population proportion is known to be in the range from 0.1 and 0.2 and it is required that there should be a probability of at least 0.99 that the difference between the sample proportion and the population proportion is less than 0.02.

(b) Nothing is known about the value of the population proportion and it is required that there should be a probability of at least 0.95 that the difference between the sample proportion and the population proportion is less than 0.03.

(9) When s independent sets of n Binomial trials produce $r_1, r_2, ..., r_s$ successes the formula $\hat{p} = \sum \frac{r_i}{sn}$ is used to find an estimate $\hat{p}$ of p, the probability of success in a single trial. Obtain the variance of p and show that approximate 95% confidence limits for p are given by the roots of the equation $sn(sn+3.84)p^2 - 2(\sum r_i + 1.92)snp + (\sum r_i)^2 = 0$ if 1.96^2 is taken as 3.84.

In one set of observations n = 10, s = 5 and r_i takes the values 1, 2, 1, 3 and 1. Determine whether p = 0.1 lies within the confidence limits given by the equation.

(13) The continuous variable X has probability density function

$$f(x) = \begin{cases} \frac{2(a+3-x)}{9}, & for\ a \le x \le a+3 \\ 0, & elsewhere \end{cases}$$

In this *a* is an unknown constant. In a single observation X is found to be 5. Find an unbiased estimate for *a* and limits within which a lies with 95% confidence.

A sample of 50 values of X is found to have a mean of 5.1. Find an unbiased estimate for a based on this sample and limits within which a will lie with approximately 95% confidence.

3.5.5 Confidence Intervals for Variances

You checked off the estimation of a number of population parameters already. Let's check off a few more! In this unit, you will derive $(1-\alpha)100\%$ confidence intervals for:

(1) a single population variance: σ

(2) the ratio of two population variances: $\frac{\sigma_X^2}{\sigma_Y^2}$ or $\frac{\sigma_Y^2}{\sigma_X^2}$

Along the way, we'll take a side path to explore the characteristics of the probability distribution known as the *F*-distribution.

3.5.6 Confidence interval of one Variance

Let's start right out by stating the confidence interval for one population variance.

Theorem. If $X_1, X_2, ..., X_n$ are normally distributed and $a = \chi^2_{1-\alpha/2n-1}$ and $b = \chi^2_{1-\alpha/2n-1}$

, then a $(1-\alpha)$% confidence interval for the population variance σ^2 is:

$$\left(\frac{(n-1)s^2}{b} \leq \sigma^2 \leq \frac{(n-1)s^2}{a} \right)$$

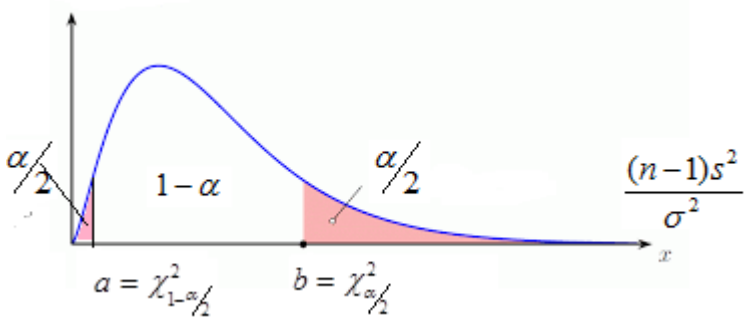
And a $(1-\alpha)$% confidence interval for the population standard deviation σ **is:**

$$\left(\frac{\sqrt{n-1}s}{\sqrt{b}} \leq \sigma \leq \frac{\sqrt{n-1}s}{\sqrt{a}} \right)$$

Proof. We learned previously that if $X_1, X_2, ..., X_n$ are normally distributed with mean μ and population variance σ^2, then:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Then, using the following picture as a guide:



with $a = \chi^2_{1-\alpha/2}$ and $b = \chi^2_{1-\alpha/2}$, we can write the following probability statement:

$$P\left[a \le \frac{(n-1)S^2}{\sigma^2} \le b\right] = 1-\alpha$$

Now, as always it's just a matter of manipulating the quantity in the parentheses. That is:

$$a \le \frac{(n-1)S^2}{\sigma^2} \le b$$

Taking the reciprocal of all three terms, and thereby changing the direction of the inequalities, we get:

$$\frac{1}{a} \ge \frac{\sigma^2}{(n-1)S^2} \ge \frac{1}{b}$$

Now, multiplying through by $(n-1)S^2$, and rearranging the direction of the inequalities, we get the confidence interval for σ^2:

$$\left(\frac{(n-1)S^2}{b} \le \sigma^2 \le \frac{(n-1)S^2}{a}\right)$$

as was to be proved. And, taking the square root, we get the confidence interval for σ:

$$\left(\frac{\sqrt{n-1}s}{\sqrt{b}} \le \sigma \le \frac{\sqrt{n-1}s}{\sqrt{a}}\right)$$, as was to be proved.

Example 3.7

A large candy manufacturer produces, packages and sells packs of candy targeted to weigh 52 grams. A quality control manager working for the company was concerned that the variation in the actual weights of the targeted 52-gram packs was larger than acceptable. That is, he was concerned that some packs weighed significantly less than 52-grams and some weighed significantly more than 52 grams. In an attempt to estimate σ, the standard deviation of the weights of all of the 52-gram packs the manufacturer makes, he took a random sample of $n = 10$ packs off of the factory line. The random sample yielded a sample variance of 4.2 grams. Use the random sample to derive a 95% confidence interval for σ.

Solution. First, we need to determine the two chi-square values with $(n-1) = 9$ degrees of freedom. Using the table in the back of the text book, we see that they are:

$$a = \chi^2_{1-\alpha/2,n-1} = \chi^2_{0.975,9} = 2.7 \quad \text{and} \quad a = \chi^2_{1-\alpha/2,n-1} = \chi^2_{0.025,9} = 19.02$$

Now, it's just a matter of substituting in what we know into the formula for the confidence interval for the population variance. Doing so, we get:

$$\left(\frac{9(4.2)}{19.02} \le \sigma^2 \le \frac{9(4.2)}{2.7}\right)$$

Simplifying, you get:

$$\left(1.99 \le \sigma^2 \le 14.0\right)$$

You can be 95% confident that the variance of the weights of *all* of the packs of candy coming off of the factory line is between 1.99 and 14.0 grams-squared. Taking the square root of the confidence limits, we get the 95% confidence interval for the population standard deviation σ:

$$1.41 \le \sigma \le 3.74$$

3.5 UNITY ACTIVITY

1. A statistician chooses 27 randomly selected dates, and when examining the occupancy records of a particular motel for those dates, finds a standard deviation of 5.86 rooms rented.

If the number of rooms rented is normally distributed, find the 95% confidence interval for the population standard deviation of the number of rooms rented.

3.0 UNIT SUMMARY

Confidence interval of the mean

(a) σ known, large samples from any population and small samples from Normal populations: $\bar{x} - \frac{z\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{z\sigma}{\sqrt{n}}$ where z takes the values shown in Table 4

Table 3.4 Critical values of the Z- table

Confidence Interval	95%	98%	99%
z	1.96	2.33	2.58

(b) σ not known, large samples from any population: $\bar{x} - \frac{zs}{\sqrt{n-1}} < \mu < \bar{x} + \frac{zs}{\sqrt{n-1}}$ with values of z as in Table 3.4

(c) σ not known, small samples from a Normal population:

$\bar{x} - \frac{t_{n-1,(100-\alpha)\%}s}{\sqrt{n-1}} < \mu < \bar{x} + \frac{t_{n-1,(100-\alpha)\%}s}{\sqrt{n-1}}$. with values of t obtained from Table A4.

Confidence interval of a proportion

$p_s - z\sqrt{[p_s(1-p_s)]} < p < p_s + z\sqrt{[p_s(1-p_s)]}$, with values of z as in Table 4

UNIT 4 HYPOTHESES TESTING

4.1 Unit Introduction

Welcome to Unit 4 in which you learn how to conduct Hypothesis Testing. Hypotheses testing is a statistical procedure used to determine whether the assumed statement about a problem is true or not. Simply.

4.2 Unit Aim

The aim of this unit is to determine whether there is enough statistical evidence in favour of a certain belief, or hypothesis, about a parameter.

4.3 Unit Objectives

By the end of the unit you should be able to:

- Make an inference about the population of interest on the basis of a random sample taken from that population.
- quantify evidence against a particular hypothesis being true

Terminology

H_o **:** This is a null hypothesis normally to be proved. It is a positive statement about your claim. This can be written in symbols or in a statement form

H_1 : This is an alternative hypothesis. This is a negative statement.

$\bar{x}$: The sample mean

σ : The population standard deviation

n : Sample size

μ : Population mean

4.4 Unit Time required

You need 20 hours for this unit

4.5 Unit Topics

4.5.1 Setting up a hypothesis

A person claims that she can tell whether the tea or the milk is put into a cup of tea first. To test her claim we ask her to perform a series of trials. In each trial she tastes two cups of tea which are identical in every respect except that one has the tea put in first and the other the milk put in first.

The two cups are presented in a random order and she has to identify the cup with the milk put in first. If she is correct seven times out of eight, ought we to accept her claim?

You may feel that if she is right more than a certain number of times, we should accept her claim; otherwise we should reject it. To decide what should be our boundary line we need to set up a mathematical model. If we are sceptical of her claim and think she is guessing we would expect the probability of success at a single trial to be ½. Consequently the probability of X successes in eight trials would be binomially distribute with p = ½, n = 8. Using this hypothesis we can calculate the probability of getting the observed result, i.e. not more than one failure. Such a hypothesis is called a null hypothesis. We should also set up an alternative hypothesis which states, as its name implies, an alternative to the null hypothesis. In this case it will be that the taster is not guessing, i.e. p > ½. We do not need to specify an exact value of the parameter for the alternative hypothesis since it is not used to calculate probabilities.

Denoting the null hypothesis by H_o and the alternative hypothesis by H_1 we write

Model: Binomial

H_o : p = ½

H_1 : p > ½

Using the null hypothesis we can calculate the probability of making no more than one mistake by calculating the probability of seven or more correct guesses. The Binomial distribution gives this probability as

$$P(X > 7) = P(X = 7) + P(X = 8)$$

$$= (8)(½)^7(½) + (½)^8$$

$$= 0.035$$

This means that if the taster is guessing there is a probability of only 0.035 that she will be correct seven or more times out of eight, and so such a result is unlikely. To decide whether to accept or reject the null hypothesis we have to be more precise about what we mean by 'unlikely'. The usual convention is that events with a probability of 0.05 or less are 'unlikely'. Using this convention the above result is unlikely to arise by *chance* if the null hypothesis is true and so we reject the null hypothesis.

If the taster had been correct only six times out of eight, the required probability is the

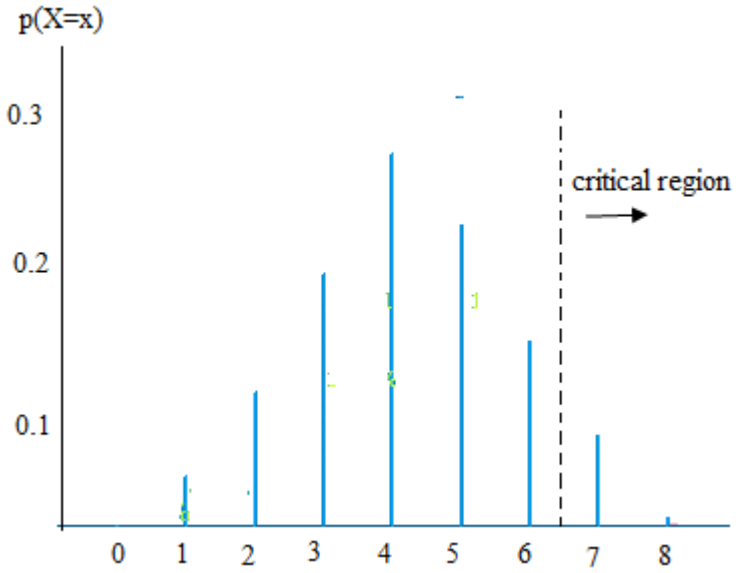


Figure 4.1 Diagram showing the critical region for the example in Section 4.5.1

Probability that the taster makes no more than two mistakes, and, assuming the null hypothesis is true, this probability is

$$p(X \geq 6) = \binom{8}{6}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^2 + \binom{8}{7}\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)^8$$

$$p(X \geq 6) = 0.145$$

Using our convention, this would not be deemed unlikely and H_o retained. Thus the possible outcomes of the experiment can be divided into two sets: X < 7 and X > 7 according to whether or not H_o is retained. Those values of the variety for which H_o rejected form the **critical** region. This is shown diagrammatically in Figure 4.1. A result which falls in the critical region is said to be significant. Further, if, on the null hypothesis, the probability of failing in the critical region is < 0.05 of 5%, then the result is said to be 'significant at the 5% level', and 5% is called the significance level of the test. Other levels of significance are sometimes used, notably 1% and 0.1%.

The significance test described above is called a one-sided or one-tailed test because the alternative hypothesis specifies that the change in p, if any, occurs in one direction only, i.e. an increase in

this case, and so the critical region falls in one 'tail' of the distribution. The following example uses a two-sided or two-tailed test.

Example 4.1

A coin is tested for bias by tossing it twelve times and counting the number of heads. If the result is two heads, is there evidence, at the 2% significance level, that the coin is biased? What is the critical region for this test?

Again we start by assuming that the coin is not biased, i.e. p = ½. Before we start the test we have no idea in which direction, if any, the coin is biased and so the alternative hypothesis specifies a change of p in either direction. We have:

Model: Binomial

H_o : p = ½

H_1 p = ½

If the null hypothesis is true then the number of heads, X, is binomially distributed. B(12, ½). In this case either a high or a low number of heads could lead us to reject the null hypothesis and so you require;

$P(X,2 \ or \ X > 10) = 2 * P(X < 2)$ (since the distribution is symmetrical)

$$= 2 * \left\{ \left(\frac{1}{2}\right)^{12} \right\} + \binom{12}{1}\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^{11} + \binom{12}{2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{10} \Bigg\}$$

$= 0.0386 = 3.86\% > 2\%$

The probability of the observed (or a more extreme) result is greater than 2% and so the result is not significant at the 2% level. The null hypothesis is retained: there is no evidence at this significance level that the coin is biased.

The critical region gives those values of X for which the null hypothesis will be rejected. Table 4.1 shows the probability distribution for B(12, ½).

Table 4.1 Probability distribution for X where X is B(12, ½)

x	P(X = x)
0	0.00024
1	0.00293
2	0.01611
3	0.05371
4	0.12085
5	0.19336
6	0.22559
7	0.19336
8	0.12085
9	0.05371
10	0.01611
11	0.00293
12	0.00024

From this table you can see that

$P(X < 1) + P(X > 11) = 0.006\,34 = 0.6\% < 2\%$

$P(X < 2) + P(X > 10) = 0.038\,56 = 3.9\% > 2\%$

and so the critical region, which is shared between the two tails of the distribution, is $X < 1$ and $X > 11$.

It is most important to realise that we can rarely prove or disprove a null hypothesis. For example, when we perform a test at the α % significance level there is a probability of α % that the result falls in the critical region when the null hypothesis is true. This means that there is a probability of α % that a true null hypothesis will be rejected. In Example 4.1 this means that there is a 2% risk of saying the coin is biased when it is not. If the test had been performed at the 5% significance level the observed result would have been significant and the null hypothesis rejected but in this case the risk of rejecting a true null hypothesis has risen to 5%. The risk of errors involved in significance testing is discussed further in Section 4.5.2.

4.1 UNIT ACTIVIT

(1) The tea-tasting experiment is modified so that at each trial the taster is offered four cups of tea. One of these is 'milk-first' and the others are 'tea-first'. Again the cups are presented in a random order. She has to identify the milk-first cup. She makes the correct

identification three times out of ten. Test at the 5% level her claim that she can identify the milk-first tea correctly, stating clearly your null and alternative hypothesis. How many times would she have to be correct before you would accept her claim, using the same level of significance?

(2) In a multiple choice test a student has to choose between three answers for each question in a test consisting of ten questions. If he gets six questions right test, at the 5% level, the null hypothesis that he is guessing.

(3) The national average pass rate for an exam is 70%. A teacher finds that six out of her twelve pupils pass. Is there evidence that this group did significantly worse than average?

(4) It is suspected that a die is biased towards a 6. This is test by throwing the die eight times. The result is three 6s. Is there evidence at the 2% significance level that the die is biased towards a 6?

4.5.2 Significance test on the mean of a Normal distribution (σ known)

You often wish to test whether a sample is drawn from population of specified mean, μ. You will start by considering the situation in which the population is normally distributed with known standard deviation σ.

Example 4.2

A machine should be set to produce bags of sugar whose weights are normally distributed with μ = 1000 g, σ = 5 g. To check the setting, a sample of nine bags is taken and the mean weight is found to be 1003 g. Is the machine correctly set? (Assume o cannot change.) You must choose a null hypothesis that assumes the mean given for the machine is correct, so that you can calculate the probability of observing a sample whose mean is 1003 g. you take

Model: Normal

H_o: μ = 1000 g, σ = 5 g

Since you are interested in differences either way from μ = 1000 g, this is a two-tailed test and

H_1 : $\mu \neq$ 1000 g

According to H_o the sampling distribution for the mean, $\bar{X}$, of a sample of nine bags is Normally distributed with u = 1000 g, s.d. = $\sigma/\sqrt{n}$ = 5 / $\sqrt{9}$ = 1.67 g. For the observed sample mean, 1003, you have: Z = (1003 – 1000)/1.67 = 1.80

Since you are carrying out a two-sided test we require

P(z < 1.80) + P(z > 1.80 = 2 x P(z > 1.80) (since the distribution is symmetrical)

= 2 x [1 - P(z < 1.80)]

= 2 x (1 – 0.9641)

= 2 x 0.0359

= 0.0718 = 7.18%

If you adopt a 5% significance level for our test then this result is not significant: there is no evidence that machine setting is incorrect.

A quicker method of arriving at the same result is to work with the critical region for the statistic z rather than to calculate an exact value of the probability, as was done above. Figure 4.2 shows the standard Normal distribution with the critical region for a test at the 5% significance level. Since the test is two-sided the critical region is divided equally between the two tails of the distribution. If a result falls in the critical region then the null hypothesis is rejected. For a two-sided test, this happens when |z| > 1.96. In the calculation above you found that z = 1.80: this result does not fall in the critical region and so there is no reason to reject the null hypothesis.

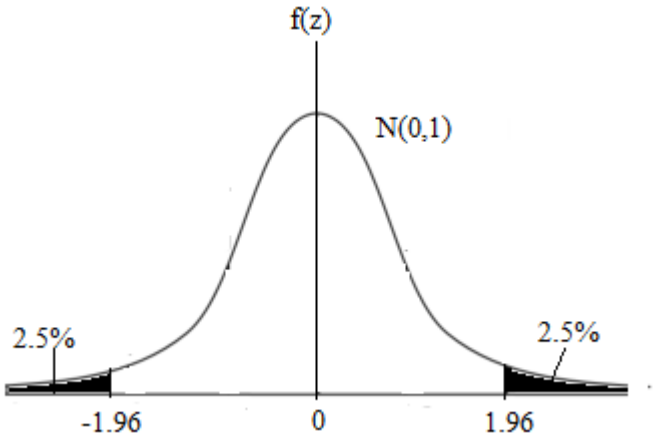


Figure 4.2 Graph showing the critical region for example 4.2

Suppose the previous example had been altered so that the question asked had been 'Is the machine set too high?' We now need to carry out a one-tail test for which we have:

Model: Normal

H_o : $\mu = 1000$ g, $\sigma = 5$ g

H_1 : $\mu > 1000$ g,

The critical region is shown in figure 3 and is given by z > 1.64. The previous calculation gave z = 1.80. This value is significant at the 5% level. you reject the null hypothesis and accept the alternative hypothesis: the machine is set too high.

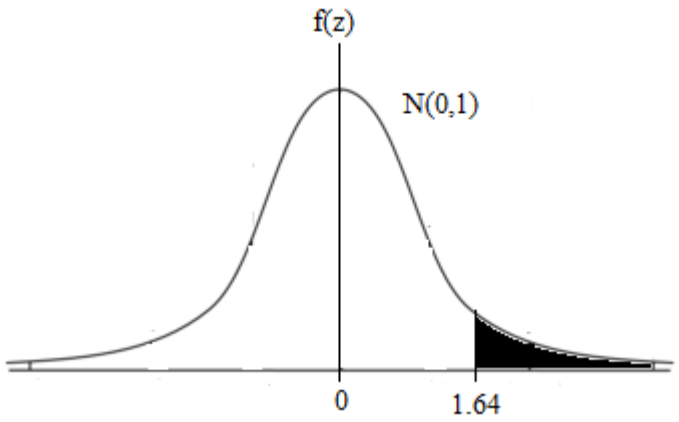


Figure 4.3 Graph showing the critical region for a one-tail test

The critical values of z can be conveniently found from the last row, v = ∞, of Table A4. For a two-tail test at the $(100-\alpha)\%$ level the critical value is given in the --% column; for a one-tail test in the $2\alpha\%$ column.

In practice you would not carry out a one-and a two-tail test on the same set of results. The null and alternative hypotheses are defined before the data are collected and the choice between a one-and a two-tailed test depends on prior knowledge. For example, if we wished to see whether a catalyst increased the rate of a reaction then you know before you start that we are looking for an increase. If the actual results showed a decrease then there would be no need to carry out the test.

The method given in this section is applicable even if the population from which the sample is drawn is not Normal *provided that the sample is large*. This is a consequence of the central limit theorem.

4.2 UNIT ACTIVITY

(1) A machine designed to produce rope with a breaking strain of 1000 N. The breaking strain is known to be normally distributed with s.d. 21 N. A new material is introduced into the rope which is hoped will increase the breaking strain. A random sample of nine pieces of rope had a mean breaking strain of 1012 N. Is there evidence at the 5% significance level that the mean breaking strain has increased?

(2) A machine filling bottle of orange squash is adjusted to deliver 0.725 litre with a standard deviation of 0.010 litre. A sample of 50 bottles is checked and the mean quantity is found to be 0.721. Determine whether this should be taken to be significantly different from 0.725 at the 5% level.

In a later check the manufacturer decides on a more stringent test, using the 2% level. If he now takes a sample of 40 bottles and the mean turns out to be 0.722, is this mean significantly differently from 0.725?

(3) A factory produces washers of mean thickness 3 mm with a standard deviation of 0.2 mm. Each day, as a routine check, the thickness of a random sample of 100 washers is measured and the mean calculated and recorded. Every 30 days, these means are assembled into a frequency table. What would you expect the mean and standard deviation of this table to be?

On one day, the mean of the sample of 100 washers is 3.036 mm. Is this significantly different from the expected value?

(4) Explain the difference between a one-tail and a two-tail significance test.

A test of mental ability has been constructed such that, for adults in Great Britain, the test score is normally distributed with mean 100 and standard deviation 15. A doctor wishes to test whether sufferers from a particular disease differ from the general population in

their performance on this test. He chooses a random sample of ten from his patients. Their scores on the test are :119, 131, 95, 107, 125, 90, 123, 89, 103, 103.
What would you conclude?

(5) The mean number of letters per word in a dictionary of the English language is 6.8 and the standard deviation of the number of letters per word is 2.5.
Work out the mean number of letters per word for the words in question 3 and test whether it appears to come from this dictionary population. Is this a valid and test of the hypothesis that the question is worded is typical English?

(6) Bill has used a particular type of razor for shaving for a long time. The length of time (in seconds) that he takes to shave is normally distributed with mean 240 and s.d. 20. He changes to a new razor and finds that his shaving times on nine consecutive days are 210, 220, 230, 220, 250, 230, 260, 210, 240. Assuming that he s.d. of his shaving time has not changed, test whether his mean shaving time has changed with the new razor.

(7) The mass of packs of unwrapped chocolates is normally distributed with mean 508 g and s.d. 4 g. You am told that a pack of chocolates weighs 520 g but you are not told whether the chocolates are unwrapped or not. Carry out a significance test to show that there is evidence that the chocolates are wrapped.

4.5.3 Significance tests (σ unknown): large samples

If σ is unknown but the sample is *large*, then the test statistic z can still be used, with σ replaced by $\hat{s}$ (as in Section 3.5.2). This gives

$$z = \frac{\bar{x} - \mu}{\hat{s}/\sqrt{n}} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$$

Example 4.3

One hundred measurements of a variety gave $\sum_{i=}^{100} x_i = 151$, $\sum_{i=1}^{100} x_i^2 = 390$. Could these measurements have come from a population, mean 1.5?

Sample s.d., $s = \sqrt{\left\{\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2\right\}}$

$$s = \sqrt{\left\{\frac{390}{100} - \left(\frac{1.51}{100}\right)\right\}^2}$$

$$= 1.27$$

$H_o : \mu = 1.5$

$H_1 : \mu \neq 1.5$

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} = \frac{1.51 - 1.5}{1.27/\sqrt{99}} = 0.078$$

This value of z is not significant at the 5% level since the critical region is $|z| > 1.96$. H_o is retained: the Measurements could have come from a population mean 1.5.

4.3 UNIT ACTIVITY

(1) A random sample of the mean marrying in a town had the age distribution given in Table 4.2.

Table 4.2 Distribution of marrying age

Age	Under 18	18-24	25-31	32-38	39-45	46-52	Over 52	Total
Coded value		0	1	2	3	4		
frequency	0	101	62	25	10	2	0	200

By using the suggested coding, or otherwise, estimate the mean and variance of the population from which the sample is drawn. (Quote your results to two decimal places). The mean age of men at marriage in the large country in which the town is situated is 25.90 years. Test whether the men in the town differ significantly from the men in the whole

country in the age at which they marry. (State your null hypothesis and the significance level you use.)

(2) The times taken by a salesman to travel between two shops on 50 occasions averaged 64 minutes with a s.d. of 6.3 minutes. He claims that this provided evidence that the journey takes longer since he changed to a smaller car, as before he averaged 60 minutes. Do you agree with the salesman?

(3) Explain briefly what is meant by a significance level of α %.

A student was asked to mark, by eye, the centre of a line drawn on a sheet of paper. She repeated this 50 times, using a new test sheet each time. The signed deviations from the centre, d, were measured in millimeters and the quantities $\sum d = -60.0$, $\sum d^2 = 197.44$ calculated. Is there evidence that the student does not locate the mean position of her bisection marks on the centre of the line?

(4) A pilot wishes to check whether the average depth of a section of river mouth is still 6.1 fathoms as his chart suggests. He takes 30 soundings at random points and finds that the mean depth is 5.8 fathoms with a s.d. of 1.1 fathoms. Test at the 5% significance level whether the mean depth has changed.

(5) A machine which fills orange squash bottles should be set to deliver 725 ml. A sample of 50 bottles is checked and the mean quantity is found to be 721 ml and the sample s.d. is 13 ml. Does this differ significantly from 725 ml at the 1% level?

(6) A new surgical technique has been developed in an attempt to reduce the time that patients have to spend in hospital after a particular operation. In the past the mean time spent in hospital was 5.3 days. For the first 40 patients on whom the new technique was used the mean time spent in hospital was 5.0 days and the sample s.d. was 0.4 days. Is there evidence that the new technique has decreased the time spent in hospital?

4.5.4 Significance tests (σ unknown): small samples

For small samples the statistic

$$t_{n-1} = \frac{\bar{x} - \mu}{\hat{s}/\sqrt{n}} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$$

must be used. It should only be used if the population from which the sample is drawn is normal.

Example 4.4

A manufacturer claims that his light bulbs have an average lifetime of 1500 hours. A purchaser decides to check this claim and finds that for six bulbs the lifetimes are 1472, 1486, 1401, 1350, 1610, 1590 hours. Does this evidence support the manufacturer's claim?

You have to assume the lifetimes of the light of the light bulbs are normally distributed.

Then you have ;

Model: Normal

$H_o : \mu = 1500h$

$H_1 : \mu \neq 1500h$

Two-tail test

Using Table A4 the critical region is $|t_5| > 2.57$ for a test at the 5% significance level.

The reader should check that for the observed lifetimes

$\bar{x}$ = 1484.8 h, s = 93.2 h

$$t_{n-1} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$$

$$t_5 = \frac{1484.8 - 1500}{93.2/\sqrt{5}} = -0.36$$

This result is not significant at the 5% level. The null hypothesis is retained and the manufacturer's claim is vindicated.

4.4 UNIT ACTIVITY

(1) Packets of breakfast cereal claim that the minimum net weight of the contents is 450 g. The weights of the contents of seven packets are: 445, 453, 447, 451, 440, 460, 449 Is there any evidence at the 5% significance level that the packets are underweight?

(2) The lives of six candles are found to be 8.1, 8.7, 9.2, 7.8, 8.4, 9.4 hours. Estimate the population mean and show that the estimate of the population variance is 0.388.
The manufacturer claims that the average life is 9½ hours. Making a suitable assumption concerning the nature of the distribution of the life of a candle, carry out a statistical test of the manufacturer's claim. Give full details of your test.

(3) A student titrates 10 ml of 0.1 M acid against 0.1 M alkali five times and obtains the following results for the volume of alkali: 9.88, 10.18, 10.23, 10.39, 10.25 ml
Is there any evidence that these results show a bias from the expected value of 10 ml?

(4) Eight volunteers tested a food which the manufacturer claimed would help people to slim. At the end of the test they had lost 1, 2, 0, 1, - 2, 0, 3, 3 kg respectively. Assuming that these losses are a random sample from a Normal distribution, carry out a t-test to determine whether the mean loss differs significantly from zero, and comment on the manufacturer's claim.

4.5.5 Difference between two means for large samples

Suppose we take large samples from two populations as indicated below:

Population 1	*Population 2*
s.d. = σ_1	s.d. = σ_2
Sample size = n_1	Sample size = n_2
Sample mean = $\bar{x}_1$	Sample mean = $\bar{x}_2$
Sample s.d. = s_1	Sample s.d. = s_2

$\bar{X}_1$ will be normally distributed (since n_1 is large) with mean μ_1 and s.d. $\sigma_1 / \sqrt{n_1}$ and $\bar{X}_2$ will be normally distributed with mean μ_2 and s.d. $\sigma^2 / \sqrt{n_2}$. The difference between the means, $\bar{X}_1 - \bar{X}_2$, will also be normally distributed with mean $\mu_1 - \mu_2$ and s.d. $\sqrt{\left\{\left(\sigma_1^2 / n_1\right) + \left(\sigma_2^2 / n_2\right)\right\}}$

(see Section 4.5.4). You frequently with to test whether two samples come from populations with the same mean in which case $\mu_1 - \mu_2 = 0$.

Example 4.5

A firm employs 300 women and 100 men. The mean number of days absent last year for the women was 5.3 with a s.d. of 2.2 and for the men the corresponding figures were 6.2 and 2.9. Is the difference between the means significant?

Model: Normal

$H_o : \mu_1 = \mu_2$ Therefore $\mu_1 - \mu_2 = 0$

$H_1 : \mu_1 \neq \mu_2$

Two-tail test

The test statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left\{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right\}}}$$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left\{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right\}}}$$

Since from . $\mu_1 - \mu_2 = 0$

If σ_1 and σ_2 are not known, then, because the samples are large, σ_1 and σ_2 can be replaced by $\hat{s}$ and $\hat{s}$. Since $\hat{s} = \sqrt{\left(\frac{n}{n-1}\right)}s$ this gives

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left\{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}\right\}}}$$

$$z = \frac{5.3 - 6.2}{\sqrt{\left\{\frac{2.2^2}{299} + \frac{2.9^2}{99}\right\}}}$$

z = - 2.83

The critical region for z is $|z| \geq 1.96$ so that z is significant at the 5% (two-tail) level and we reject the null hypothesis. There is a significant difference between the mean absences of men and women .The theory for small samples is beyond the scope of this book.

4.5 UNIT ACTIVITY

(1) It is found that over a certain period at one telephone exchange 200 subscribers taken at random made a total of 13 248 calls. During the same time, a random sample of 300 subscribers at another exchange made a total of 20 922 calls. The standard deviation of the number of calls made by a subscriber at either exchange in the period is 8.
Is there any evidence of a difference between the subscribers at the two exchanges in their average frequency of calls?

Find 95% confidence limits for the mean number of calls made in the period per subscriber at each exchange.

(2) Samples of leaves were collected from two oak trees A and B. The number of galls was counted on each leaf and the mean and standard deviation of the number of galls per leaf was calculated with the results given in Table 4.3

Table 4.3 Samples of leaves

Tree	A	B
Sample size	60	80
Mean	11.4	10.7
s.d.	2.6	3.1

Assuming Normal distributions, do the data provide evidence at the 5% significance level of different population means for the two trees?

(3) In an investigation into the effectiveness of a particular course in speed reading a group of 500 students was split into two groups, A and B, of sizes 300 and 200 respectively, thought to have been chosen at random.

Those in group A were given no special instruction; those in group B were given a course in speed reading. Each student was asked to read the same passage and the time taken was measured. The results were

Group A: mean time 78.4 s, variance 14 s^2

Group B: mean time 77.4 s, variance 15 s^2

Carry out a significance test to see if there is evidence that the course has improved reading speed. State carefully your null hypothesis, alternative hypothesis and final conclusion.

You learn later that, of the original 500 students, 200 students had decided for themselves that they wanted to take the course in speed reading and that these students became group B. Discuss briefly how this might affect your previous conclusion.

4.5.6 Testing if two samples come from the same population

In this case we wish to test whether $\bar{x}_1 - \bar{x}_2$ differs significantly from zero as in the previous example.

However, since our null hypothesis is now that the two samples are from the *same* population, we have $\sigma_1 = \sigma_2 = \sigma$ (say) and the s.d. of $\bar{x}_1 - \bar{x}_2$ is therefore $\sigma\sqrt{\left\{ 1/n_1 + 1/n_2 \right\}}$. If we know o we can use the test statistic

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}}} \qquad (1)$$

More usually σ is not known. We need to make an unbiased estimate s of it from both the samples and use the test statistic t. it can be shown that, if s_1 and s_2 are the standard deviations of the two samples, the most efficient unbiased estimate of the s.d of the population is given by

$$\hat{s} = \sqrt{\left\{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\right\}}$$

The denominator represents the number of degrees of freedom of this estimate so that the test statistic

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{n_1 s_1^2 + n_2 n_2^2}{n_1 + n_2 - 2}\right)\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}}} \qquad (2)$$

can be used provided that the population are Normal and/or the samples are large

Example 4.6

A market gardener decides to test a new pesticide, which the manufacturer claims increases the yield, by applying it to one of his two orchards. The treated orchard contains twenty trees and the mean and s.d of the yield per tree are 98 kg and 10 kg respectively. The untreated orchard contains fifteen trees and the corresponding values are 94 kg and 8 kg.

Test whether these results are consistent with the yields being drawn from the same population.

Model: Normal

$H_o : \mu_1 = \mu_2, \sigma_1 = \sigma_2 = \sigma$

H_1 : Samples come from different populations

$\bar{x}_1 = 98$ kg, $\bar{x}_2 = 94$ kg

$s_1 = 10$ kg, $s_2 = 8$ kg

$n_1 = 20$, $n_2 = 15$

The unbiased estimate of variance using equation (2) is

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{n_1 s_1^2 + n_2 n_2^2}{n_1 + n_2 - 2}\right)\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}}} = \sqrt{\left\{\frac{20*10^2 + 15*8^2}{20+15-2}\right\}} = 9.47$$

Using equation (3),

$$t_{33} = \frac{98-84}{9.47\Big/\sqrt{\left\{\frac{1}{20} + \frac{1}{15}\right\}}} = 1.24$$

This is not significant at the 5% level (two-tail) and H_o is retained. The results are consistent with the samples being drawn from the same population. Strictly speaking, before such a test is made,

a test should first be made on the variances of the samples to see whether a pooled estimate of variance is justified (see Section 4.5.4). If there is a significant difference between the variances, then the analysis described is no longer valid.

4.6 UNIT ACTIVITY

(1) In a butter-packing plant the quantity of butter packed in a day using a certain type of machine is a normal variable with a standard deviation of 39 kg. Two packers A and B average 1518 kg and 1499 kg per day respectively over a 26-day month. Is the performance of the two operatives significantly different?

A third packer C averages 1480 kg per day for her first 26-day month. Is this a significantly worse performance than B's?

(For each test you should state clearly your null and alternative hypothesis.)

Table 4.4 Measurements of butter-packing plant the quantity

Source	Length in mm							
Zambia	12.3	12.7	13.1	10.8	11.3	11.8	12.4	13.2
N. Africa	10.6	9.8	11.5	10.0	11.1			

(2) The lengths of the femur in samples of Mus homunculus from two sources (Zambia and North Africa) are given in Table 4.4.

The mean length of the femur is known to be characteristic of each breed of Mus homunculus. Test whether the data are consistent with the assumption that Mus homunculus in Zambia and N. Africa are of the same breed.

(3) In an experiment 22 mice were divided into two groups, one of which was given a special diet. After an interval the gains in weight of the mice were measured, with the results given in Table 4.5.

The variance of weight increases may be assumed to be the same for both populations of which these groups are samples, those being fed a special diet and those being ordinarily

fed. Calculate an estimate of this common variance from the combined results of both groups, and discuss whether the difference in the observed weight increases is significant.

Table 4.5 Experiment of 22 mice

	Number	Mean gain	Variance
Special diet	10	20.6	4.51
Control group	12	8.2	3.17

4.5.7 Paired tests

The paired sample t-test, sometimes called the dependent sample t-test, is a statistical procedure used to determine whether the mean difference between two sets of observations is zero. In a paired sample t-test, each subject or entity is measured twice, resulting in pairs of observations.

Example 4.7

Table 4.6 gives the times taken (in minutes) by eight typists to type the same number of words using two different typewrites. Do the data indicate any difference in speeds for the two typewriters?

Table 4.6 Times taken (in minutes) by eight typists

Typist	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
X, time using typewriter 1	6.3	4.5	7.1	8.4	3.7	3.9	4.7	5.2
Y, time using typewriter 2	5.1	4.4	6.2	7.3	4.5	4.0	3.6	5.1

In this case it would be incorrect to calculate the means for typewriter 1 and typewriter 2 as in Section 4.5.6 since the variations between the typists could swamp any difference due to the machines. Instead we compute the difference, D, for each typist, since if there is no difference between the machines then we would expect the mean difference, D, to be zero. Provided the original populations are Normal, $\overline{D}$ will also be approximately normally distributed.

Table 4.7 shows the calculation of the mean *d*, and the s.d., s_d, of the sample. (Note that var(D) = var(X) + var(Y) since the values of X and Y used to calculate each d_i are not independent. In fact var(D) < var(X) + var(Y).)

$$\bar{d} = \frac{3.6}{8} = 0.45 \text{ min.}$$

$$s_d = \sqrt{\left\{\frac{5.34}{8} - \left(\frac{3.6}{8}\right)^2\right\}}$$

$$= 0.68 \text{ mim}$$

If μ_D is the mean of the population from which D is drawn you have:

Model Normal.

$H_o: \mu_D = 0$

$H_1: \mu_D \neq 0$

Two-tail test

Using *t* as the test statistic, since the s.d s_d is estimated from the sample, you have

Table 4.7 Calculation for paired t-test

Typist	$d_i = x_i - y_i$	d_i^2
A	1.2	1.44
B	0.1	0.01
C	0.9	0.81
D	1.1	1.21
E	-0.8	0.64
F	-0.1	0.01
G	1.1	1.21
H	0.1	0.01
	3.6	5.34

$$t_{n-1} = \frac{O - d}{s_d / \sqrt{n-1}}$$

$$= \frac{-0.45}{0.67/\sqrt{7}}$$

$$= -1.75$$

This value of *t* is not significant at the 5% (two-tail) level and H_o ***r***etained. The data do not indicate a difference in speed for the two typewriters.

4.5.8 The sign test

This test is sometimes used instead of the paired *t*-test. Consider again the data in example 4.1 (see Table 4.8). A plus sign in the last row indicate that typewriter 1 was faster and a minus sign that typewriter 2 was faster. If we adopt the null hypothesis that there is no difference between the two typewriters we would expect an equal number of plus and minus signs. In fact we have two minus signs out of eight. The number of minus signs will be binomially distributed with, on our null hypothesis, p = ½, n = 8. The test is two-tailed since we are only concerned with whether the typewriters differ in speed.

Model: Binomial

H_o : p = ½

H_1 : p ≠ ½

Two-tail test

The probability of observing two or less minus signs is

Table 4.8 Times taken (in minutes) by eight typists

Typist	A	B	C	D	E	F	G	H
Time using Typewriter 1	6.3	4.5	7.1	8.4	3.7	3.9	4.7	5.2
Time using Typewriter 1	5.1 +	4.4 +	6.2 +	7.3 +	4.5 -	4.0 -	3.6 +	5.1 +

$$\left(\frac{1}{2}\right)^8 + \binom{8}{1}\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right) + \binom{8}{2}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^2 = 0.145$$

Thus the probability of observing < 2 or > 6 minus signs is 0.29 and the observed number of minus signs is not significant at the 5% level. H_o is retained (as it was when the paired t-test was performed). The sign test can be used whether or not the populations are Normal.

The sign test is an example of a non-parametric test. Non-parametric tests make no assumptions about the distribution from which the sample is drawn. In particular they do not assume that the population distribution is Normal as many of the tests described in this unit do.

4.7 UNIT ACTIVITY

(1) Ten marksmen shot at targets with two types of rifle. Their scores out of 100 were as in Table 4.9. Apply the sign test to the hypothesis that he rifles are equally good. Also use the t-test for paired values to test the same hypothesis. What assumption is made in applying this test, but not in applying the sign test?

Table 4.9 Ten marksmen

Marksman	A	B		D	E	F	G	H	I	J
Rifle 1	93	99	90	87	85	94	88	91	96	79
Rifle 2	89	93	86	92	78	90	91	87	92	86

(2) During negotiations between the union and management of a large firm two alternative offers are put to the union side:

(i) old rates plus 20% increase across the board,

(ii) new rates based on a production bonus scheme.

The union statistician, Percy Glum, considers these offers and calculates the pay that 25 typical employees would receive from each offer. These are summarised in Table 4.10.

Table 4.10 negotiations between the union and management

Employee	1	2	3	4	5	6	7	8	9	10	11	12	13
Old rates+20%(K)	56	43	59	62	38	49	53	37	71	53	47	39	37
New rate(K)	67	58	58	75	47	51	52	49	75	59	56	41	42

Employee	14	15	16	17	18	19	20	21	22	23	24	25
Old rates+20%(K)	68	27	68	75	42	53	61	56	58	35	46	37
New rate(K)	65	31	72	84	45	54	65	61	57	39	49	39

By use of the sign test, or otherwise, determine whether the new rates will lead, on average, to an increase of more than the 20% on the old rates.

4.5.9 Significance of a proportion (large samples)

In this unit you return to the situation considered at the beginning of the unit where you have a variable which is binomially distributed, the difference being that you will consider large samples so that the Normal approximation can be used.

Example 4.8

On a national basis the success rate for people taking their driving test for the first time is 40%. A driving instructor claims that his record is superior because, of the 50 pupils of his who took the test for the first time last year, 25 passed. Is his claim justified?

You have

Model: Binomial approximated by Normal

$H_o : p = 0.4$

$H_o : p > 0.4$

One-tail test

On the null hypothesis the number of people, X, who pass first time in a sample of 50, is a variable which is binomially distributed with

$\mu = np = 50 * 0.4 = 20$

$\sigma = \sqrt{npg} = \sqrt{(50 * 0.4 * 0.6)} = \sqrt{12}$

Since *n* is large, X is approximately normally distributed with mean 20, s.d. 3.46. The Number in the sample who pass first time is 25. Since the variance of population is known we can use the test statistic *z*. As we have a *single* measurement of the variable X,

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{25 - 20}{\sqrt{12}} = 1.44$$

This value of z is not significant at the 5% level (one-tail). H_o is retained and the instructor's claim is not justified. Strictly speaking a continuity correction of $\pm\frac{1}{2}$ should be applied. In this case this would lead to a value of $z = \frac{24.5 - 20}{\sqrt{12}} = 1.30$. However, this correction is frequently omitted.

The solution to this problem can also be given in terms of proportions since the sampling distribution of the proportion in the sample, p_s, *is* $N[p, p(1-p)/n]$. We have

$$z = -\frac{p_s - p}{\sqrt{\left[\frac{p(1-p)}{n}\right]}}$$

$$z = -\frac{0.5 - 0.4}{\sqrt{\left[\frac{0.4 * 0.6}{50}\right]}} = 1.44$$

which, of course, gives the same result as before. In this case the continuity correction, if applied, is $\pm\frac{1}{2}n$

4.8 UNIT ACTIVITY

(1) After a survey a market research company asserted that 75% of TV viewers watched a certain programme. Another company interviewed 75 viewers and found that 51 watched the programme and 24 did not. Does this provide evidence at the 5% level of significance that the first company's figure of 75% was incorrect?

(2) In a multiple choice examination paper, a candidate has to select which of four possible answers to a question is the correct one. On a paper with 100 questions he gets 34 correct

answers. Explain carefully, with supporting calculations, whether you regard this result as contradicting the supposition that his answers are obtained entirely by guesswork.

(3) Explain what is meant by a standard error.

A certain method of scaling examination marks is supposed to fix the quartiles at 25 and 75. Out of 1000 candidates, 541 have marks between these limits. Is this evidence that the method has failed?

(4) You are engaged as an expert witness for the prosecution in a Court case in which a gaming club is accused of running an unfair roulette wheel. The evidence is that out of 3700 trial spins, zero (on which the club wins) turned up 140 times. There are 37 possible scores on a trial spin, labelled 0 t0 36, and these should have equal probability. Test whether there is evidence that the wheel is biased. Explain briefly what this test of significance means, bearing in mind you have to convince a non-mathematical jury.

(5) If births are equally likely on any day of the week then the proportion of babies born at the weekend should be 2/7. Out of a random sample of 100 children it was found that 23 were born at the weekend. Does this provide evidence that the proportion of babies born at the weekend differs from 2/7?

4.5.10 Difference between two proportions (large samples)

Example 4.9

In a random sample of 500 people from a certain town there are 270 men, of whom 160 are smokers, and 230 women, of whom 110 are smokers. Is there evidence that the men of the town are more likely to smoke than the women? If p_1 is the probability that a man from the town smokes and p_2 is the corresponding probability for women, then

H_o: $p_1 = p_2 = p$ (say)

H_1: $p_1 > p_2$

One-tail test

The numbers of men and women who smoke are each binomially distributed but since the number in the sample is large, the Binomial distribution can be approximated by the Normal distribution. Our best estimate of p is found by combining all the data:

$$\hat{p} = \frac{number\ who\ smoke}{number\ in\ sample} = \frac{160+110}{500} = 0.54$$

According to H_o, the observed proportion of men who smoke, P_1, has s.d. $\sqrt{pq/n_1}$, where n_1 is the number of men in the sample (see Section 4.5.10). Similarly for women, the observed proportion, P_2, has s.d. $\sqrt{pq/n_2}$ where n_2 is the number of women in the sample. Thus $P_1 - P_2$ has s.d. σ_{1-2} given by

$$\sigma_{1-2} = \sqrt{\left\{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}}$$

Using our estimate of p, $\hat{p} = 0.54$, this gives for an estimate $\hat{s}_{1-2}$ of σ_{1-2}

$$\hat{s}_{1-2}\sqrt{0.54*0.46\left[\frac{1}{270}+\frac{1}{230}\right]} = 0.0447$$

According to H_o the mean value of $P_1 - P_2$ is 0. The observed value is

$$\frac{160}{270}+\frac{110}{230} = 0.114$$

Using the test statistic z since n is large

$$z = \frac{0.114-0}{0.0447} = 2.55$$

This value is significant at the 5% level (one-tail). H_o is rejected and there is evidence that men are more likely to smoke than women.

4.9 UNITY ACTIVITY

(1) A television rental organisation supplies its service mechanics with small vans for use when visiting customers. In 1975, the London branch used vans of type *A* and 15 out of a total of 60 spent at least one day off the road being repaired. The Edinburgh branch used vans of type *B* in the same year and the corresponding figures were 20 out of a total of 40. Is there a significant difference between the two proportions?

A manager of the organisation wishes to use these figures to compare the reliability of the vans of the two types to help him decide whether the organisation should use one type of

van rather than the other in both cities. Comment on (a) what may be inferred about relative reliability from the significance test, (b) what extra information, if any, would be useful when comparing reliability.

(2) Test whether the proportion of patients who die when they receive antitoxin treatment is significantly different from the proportion who die when they receive ordinary treatment for the data given in unit activities 3.54, question (4).

(3) The average number of flaws per 100 metre length of a yarn produced by a machine has been found to be seven. A new machine is installed and the first 100 metre length has three flaws. Does this provide evidence that the new machine is better than the old?

(4) In an intensive survey of Nguzu land it was found that the average number of plants of a particular species was 26 per square metre. After a hard winter the number of these plants found growing in a randomly chosen area of one square metre was 15. Is there evidence that the hard winter has tended to kill off this species of plant?

4.5.11 Significance test using the Poisson distribution

The ideas developed in this unit can also be applied when the Poisson distribution is a suitable model for the population from which the sample is drawn.

Example 4.10

Over a number of years the average number of breakdowns of an office photocopies has been six per month. A new photocopier is installed and the number of breakdowns in the first month is reduced to one. Is this evidence that the new photocopier is more reliable than the old one? Encouraged by this result the office decides to keep the photocopier and the total number of breakdowns for the first year is 50. Does this confirm the idea that the new photocopier is more reliable?

Assuming that breakdowns are events which are randomly distributed in time, we have

Model: Poisson

H_o: $\lambda = 6$

H_1: $\lambda < 6$

$p(X \leq 1) = e^{-6} + 6 * e^{-6}$

$= 0.0174 = 1.74\% < 5\%$

The result is significant at the 5% level and there is evidence that the new photocopier is more reliable than the old. Again adopting the null hypothesis that the new machine has the same reliability as the old one, the mean number of breakdowns per year will be Poisson distributed with mean = 12 x 6 = 72. Since this is large, the number of breakdowns per year will be approximately Normally distributed with mean and variance both equal to 72. We have

Model: Poisson approximated by Normal

H_o: $\lambda = 72$

H_1: $\lambda < 72$

You calculate the test statistic

$$z = \frac{50 - 72}{\sqrt{72}} = -2.59 < -1.64$$

This result is also significant at the 5% level and suggests that the new photocopier is more reliable than the old one. (Strictly speaking a continuity correction should be applied).

4.5.12 Type I and Type II errors

When we test a null hypothesis by applying a significance test to a sample we can never be *certain* that our conclusion is correct. All we know is that our decision is probably correct. For example if we reject a null hypothesis it is because we obtained a value of the test statistic in the critical region and this event is unlikely if the null hypothesis is true. you can make two types of error:

Type I: the null hypothesis is rejected when it is in fact true.

Type II:the null hypothesis is retained when it is in fact false.

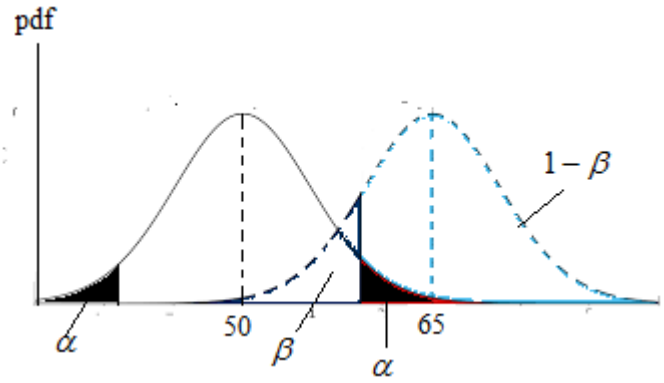


Figure 4.4 Diagram showing the probabilities of making Type I and Type II errors for the example in Section 4.5.12

The probability of making a Type I error is easily calculated since it is the probability that the test Statistic lies in the critical region and is thus equal to the significance level of the test.

The probability of a Type II error cannot be calculated unless we specify a particular value of the Parameter for the alternative hypothesis. Suppose we are sampling from a Normal distribution with known standard deviation, o equal to 5 and wish to test $H_o : \mu = 50$ against $H_1 : \mu = 65$ by taking a single measurement. The solid curve in figure 4 show the distribution of a single measurement, X, if the null hypothesis is true and the broken curve shows the distribution if the alternative hypothesis is true.

At the 5% level of significance the probability of making a Type I error is given by the solid black area corresponding to |z| > 1.96. The upper critical value of X, i.e. the value of X above which H_o is rejected, is given by

$$z = \frac{x - \mu}{\sigma}$$

$$1.96 = (x - 50)/5$$

$$x = 59.8$$

You can now calculate the probability of a type II error which is given by the hatched area. You have

$$z = (59.8 - 65)/5$$

$$= 1.04$$

$$P(Z < -1.04) = P(Z > 1.04)$$

$= 1 - P(Z < -1.04$

$= 1 - 0.8508$

$= 0.1492$

$= 15\%$

You can reduce the first error by using a lower significance level, e.g. 1%. This will move the upper critical value of X to the right in figure 15.4 which unfortunately increases the probability of a Type II error.

Similarly reducing the Type II error will increase the Type I error. If, however, we are prepared to take a larger sample and use it's mean to calculate the test statistic then both errors can be reduced because the standard deviation of the sampling distributions is reduced. If, for example, we take samples size four then the standard error of the mean is $5/\sqrt{4} = 2.5$. If the upper critical value of X is kept as 59.8 you have

$$P(\text{Type I error}) = P[Z > (59.8 - 50)/2.5]$$
$$= P(Z > 3.92)$$
$$= 0.005\%$$

$$P(\text{Type II error}) = P[Z > (59.8 - 65)/2.5]$$
$$= (Z < 2.08)$$
$$= 1.9\%$$

In an experiment the critical value and the sample size can be chosen to give acceptable values for both types of error. An example is given in question (3) of the next exercise.

Example 4.11

A box is known to contain either H_o ten white counters or ninety black counters or (H_1) fifty white counters and fifty black counters. In order to test hypothesis H_o against hypothesis H_1, four counters are drawn at random from the box without replacement. If all four counters are black H_o is accepted. Otherwise it is rejected. Find the size of the Type 1 and Type II errors for this test.

Type I error

H_o is rejected when it is in fact true. H_o is rejected if less than four black counters are obtained. The Probability of this event is most easily calculated by finding the probability that all four counters are black.

$$P(\text{four black}) = = \frac{90}{100} * \frac{89}{99} * \frac{88}{98} * \frac{87}{97} = 0.652$$

Therefore the probability of a Type I error is 1 - 0.652 = 0.348.

Type II error

H_o is retained when it is false and H_1 is true. If H_1 is true the probability of selecting four black counters is $= \frac{50}{100} * \frac{49}{99} * \frac{48}{98} * \frac{47}{97} = 0.059$

This is the probability of a Type II error.

If the probability of a Type II error occurring for a specific alternative hypothesis is β, then 1 – β gives the probability that we reject a false null hypothesis, i.e. we make the correct decision. This probability is known as the **power** of a test. In the previous example, with the specified alternative hypothesis, the Power was 1 – 0.059 = 0.941.

4.10 UNIT ACTIVITY

(1) You are provided with a coin which may be biased. In order to test this you are allowed to toss it twelve times and count the number, *r*, of heads to decide. If the coin is really fair you wish to have at least a 95% chance of saying so. For what value of *r* should you decide the coin is fair?

If you adopt you procedure with a coin which is actually biased two to one in favour of heads, what is the probability that you decide the coin is biased?

(2) A sample of twenty items is taken from what is believed to be a Normal population with mean 24 and standard deviation 5. Obtain the range of values within which the mean of the sample must lie in order that the sample may be accepted as coming from the

population if the significance level of the test is 5%. What is the probability that we shall reject the conclusion that the sample comes from this population when it in fact does so? Assuming that he standard deviation is correct but the mean of the population is in fact 25, what is the probability that we shall accept the original hypothesis when it is untrue in this particular way?

(3) (a) outcomes: success and failure. To test the null hypothesis that the probability, p, of a success is 0.25, an experiment consisting of ten independent trials is performed. If six or more successes are observed the null hypothesis is rejected. Find the significance level of the test. If p is actually 0.5, find the power of the test.

(b) The life of a particular type of projector bulb has a mean value of 50 hours and standard A machine produces nails whose lengths are normally distributed with mean u cm and standard deviation 0.1 cm. When the machine is working correctly u = 3.0, but occasionally the machine goes wrong, in which case u = 3.05, the standard deviation remaining 0.1 cm. In order to decide whether the machine is working correctly, the lengths of the nails in a sample of n nails are measured, and the sample mean x is found. If the value of x exceeds a predetermined value v then it is concluded that the machine has gone wrong; otherwise the machine is presumed to be working correctly. It is required that there should be probability of no more than 5% of presuming that the machine has gone wrong when in fact it is working correctly, and that there should also be a probability of no more than 10% of presuming the machine to be working correctly when it has gone wrong. Determine appropriate values of n and v, if n is to be as small as possible.

(4) State what you understand by the critical region and the power of a test.

(a) A trial has two possible deviation of 10 hours. It is thought that a different design of bulb will have a longer life, but it is expected that the standard deviation will not change. A random sample of 25 bulbs of the new type is obtained. The sample mean x, which may be assumed to be normally distributed, is to be used to test for an increase in mean life. If the significance level of the test is 1%, find the critical region for the sample mean.

(b) Find the power of the test when the mean life for the improved bulb is (i) 56 hours,

(ii) 52 hours.

5. (a) In an industrial manufacturing process, when production is under control, one unit in fifty is defective. A sample of 100 units is taken at random every half-hour. Sketch a control chart for the number of defectives found in a sample, showing the 1 in 40 and 1 in 1000 lines.

(b) Something goes wrong with the process, so that one unit in twenty is defective. Calculate the probability that the number in random samples of 100 will go outside (i) the 1 in 40 line, (ii) the 1 in 1000 line.

(c) What is the probability that the process will continue out of control for two hours without

any sample exceeding the 1 in 1000 limit?

(6) The mean of a random sample of n observations, $x_1, x_2, \ldots, x_n$ from a Normal distribution is x. It is proposed to test the null hypothesis that the population mean is u_o, against some alternative hypothesis. State how your decision as to whether to use the t-distribution for such a test would be influenced by (i) the value of n, (ii) whether the population standard deviation is known to you.

You are given that $x = 12.9$, $n = 10$, $\chi_i^2 = 1683$.

(a) Test the null hypothesis that the population mean is 12 against the alternative hypothesis that the mean is not 12.

(b) Determine a two-sided, symmetric, 95% confidence interval for the mean and explain carefully what this confidence interval means.

(7) Over a period of years the average number of road deaths during the New Year Festival period in Ndola has been 27. This year there were 35, a rise of approximately 30% and the country's press was full of gloomy comment on declining driving standards. Make suitable calculations to demonstrate that the probability of there being at least the observed number of road deaths is about 7½%.

Write an account of your argument and conclusions as if addressing an intelligent reader but one having no knowledge of statistics or probability theory.

(8) What are the conditions under which a variable might be distributed according to the Binomial distribution?

An investigator suspects that there might be an unnecessary rounding off of weights recorded by a spring balance. In a spot check, he finds that, out of ten recorded weights, six end in the digits 0 or 5. How strong is the evidence that his suspicions are justified?

(9) Explain the meaning of the term "confidence interval'.

The expressions $\sum x_i^2$ $(x_1 - x)^2/n$ and $\sum x_i^2$ $(x_1 - x)^2/(n - 1)$ are both used in connection with variance for a set of observations $x_1, x_2, \ldots, x_n$. Explain the circumstances in which each is used and hence distinguish between them.

For such a set of 81 independent observations from a normal distribution, $\sum x_i = -36$ and $\sum x_i^2 = 736$. Construct a 95% confidence interval for the mean of their probability distribution, and use it to test whether this mean is likely to be zero.

Explain, either by carrying out the appropriate calculations, or otherwise, how you would conduct this test without using a confidence interval.

(10) Explain how student's t-distribution may be used to test the hypothesis that a random sample of n observations is derived from a population whose mean is μ_o assuming the population to be Normal.

Table 4.11 Test in geography before and after a series of tape-slide

Pupil	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Mark before tape-sliders	3	15	14	18	10	3	5	8	9	11
Mark after tape-slides	4	18	14	16	12	6	5	9	10	15

Ten pupils took a test in geography before and after a series of tape-slide sequences on that subject. Their marks (out of 20) were in Table 4.11.

Test, at the 5% level of significance, the hypothesis that the pupils' performance is not affected by the series of tap-slide sequences.

State the assumptions made in applying the test.

(11) Explain what is meant by a 95% confidence interval for a population mean.

A factory manufacturing ammeters tests them for zero errors in their calibration.

From past routine tests, it is known that the standard deviation of these errors is 0.3.

A batch of nine ammeters, taken from one worker's production, has zero errors of 1.0, -0.1, -0.3, 1.6, 0.5, 0.4, 0.5, 0.2, - 0.2. Test whether there is evidence of a bias in the ammeters produced by this worker and establish a 95% confidence interval for the mean zero of his ammeters.

(12) Two alternative hypotheses concerning the probability density function of a random variable are

$$H_o:\ f(x)\begin{cases} 2x, & 0<x<1 \\ 0 & elsewhere \end{cases}$$

$$H_1:\ f(x)\begin{cases} 2(1-x)x, & 0<x<1 \\ 0 & elsewhere \end{cases}$$

Give a sketch of the probability density function for each case.

The following test procedure is decided upon. A single observation of X is made and if X exceeds a particular value c, where 0 < c < 1, then H_o is accepted, otherwise H_1 is accepted.

Find the value of c if the probability of accepting H_1 given that H_o is true is 1/9. With this value of *c*, find the probability of accepting H_o given that H_1 is true.

(13) A blindfold subject was given a sample of butter and one of margarine and asked to state which was butter. This was repeated to give six tests in all, the samples being presented in a random order on each occasion. The subject correctly identified the butter five times out of six. Find the probability of five or six correct identifications by a subject who is not able to distinguish between butter and margarine. State, with reasons, whether the subject's performance would lead you to believe that she could identify the butter.

The experiment was carried out simultaneously with a total of 24 subjects. The total number of correct identifications was 83 out of a total of 144 trials. Test if the group as a whole could identify the butter significantly better than might be expected.

(14) Explain what is meant by the sampling distribution of the mean. What shape would you expect this to have for large samples?

A machine is designed to produce rods 2 cm long with a standard deviation of 0.02 cm. The lengths may be taken as normally distributed. The machine is moved to a new position in the factory, and, in order to check whether the setting for the mean length has altered, the lengths of the first nine rods produced are measure. The standard deviation may be considered to be unchanged. If these lengths, in cm, are as given below, test whether the setting has been altered or not.

2.04, 1.97, 1.99, 2.03, 2.04, 2.10, 2.01, 1.98, 2.07

(15) An engineer wishes to compare the results obtained by two methods of measuring the breaking strains of a certain type of wire rope. Describe carefully the design of an experiment for this purpose which would require a two-sample *t*-test for its analysis.

In such an experiment the measurements of breaking strain (x) obtained from six observations using the first method gave $x = 492$, $x^2 = 43\ 850$, whilst eight observations using the second method gave $x = 608$, $x^2 = 50\ 766$. Test if the measurements came from the same population.

(16) In the course of a survey concerning the proportion of lest-handed children the figure in Table 4.12 were obtained from two schools.

Show that an approximate 95% confidence interval for the population proportion, *p*, of left-handed children derived from the data from school 1 is $0.25 < p < 0.32$, and calculate a corresponding interval for school 2.

Table 4.12 proportion of lest-handed children

	Number of Children	Proportion left-handed
School 1	620	0.284

School 2 475 0.341

==========

Explain briefly what is wrong with the following argument: 'Since these two confidence intervals overlap, we cannot reject, at the 5% significant level, the hypothesis that the populations from which the children in the two schools are samples each have the same proportion of left-handed children.'

Calculate the overall proportion of left-handed children in both schools, and show that the observed difference in proportions is significant at the 5% level.

(17) Experimental data concerning a variable X, which measures the reliability of a certain electronic component, are as follows: $x = 1164.2$, $x_2 = 13911.6$, $n = 100$.

Calculate the sample mean and standard deviation from these figures. Explain whether, on the evidence of this sample, you would reject the hypothesis that the mean value of X is 12.

Figures collected over a long period have established that the mean and standard deviation of X are 12 and 2 respectively. After a change in the manufacturing process it is expected that the mean will have been *increased*, but it may be assumed that the standard deviation remains equal to 2. A sample of n values of X is taken, with sample mean m: if m is greater than some critical value it will be accepted that the mean has in fact increased, but it m is less than the critical value the increase is not established.

State carefully appropriate null and alternative hypotheses for the situation, and find, in terms of n, the critical value for a 1% significance level.

(18) (a) In one county in England, a random sample of 225 twelve year old boys and 250 twelve year old girls was given an arithmetic test. The average mark for the boys was 57 with a standard deviation of 12, whilst the average for the girls was 60 with a standard deviation of 15. Assuming that the distributions are Normal, does this provide evidence at the 2% level that twelve year old girls are superior to twelve year old boys at arithmetic?

(b) An IQ test which had been standardised giving a mean of 100 and a standard deviation of 12 was given to a random sample of 50 children in one area. The average mark obtained was 105.

Does this provide evidence, at the 5% level, that children from this area are generally more intelligent?

(19) Describe what is meant by a random sample of observations of a random variable. A battery manufacturer claims that his batteries have a mean life of 8 hours when used in a particular model of calculating machine. Describe how you would use a significance test to examine this claim on the basis of a large random sample of observed lifetimes. You should explain clearly the hypothesis you would consider, the choice of significance level, and the details of the test. State what is meant by concluding that the null hypothesis is rejected at the 5% significance level.

A random sample of 121 batteries has a mean life of 7.56 hours and, for this sample, $s^2 = 5.30$ hours^2. Test whether these data would lead to rejection of the manufacturer's claim, at the 5% significant level.

(20) A chemical is delivered in batches to a factory for use in a production process. It is important that the percentage of manganese in the chemical should not decrease significantly in successive batches. The first batch delivered is to act as a control. Ten determinations of percentage manganese are made, with results as follow: Control batch (% manganese) 3.3, 3.7, 3.5, 4.1, 3.4, 3.5, 4.0, 3.2, 3.7.Show that the control mean is 3.62, and estimate the standard deviation. The results of the ten determinations for a later batch are: Batch X (% manganese) 3.2, 3.6, 3.1, 3.4, 3.0, 3.4, 2.8, 3.3, 3.6. Determine whether the mean of this batch is significantly smaller than 3.62. Use a 5% level of significance, and justify any assumptions you make.

4.0 UNIT SUMMARY

Test of	Test statistic	Value if

Mean σ known $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ Normal population and/or large sample

Mean σ unknown $z = \dfrac{\bar{x} - \mu}{\hat{s} / \sqrt{n}} = \dfrac{\bar{x} - \mu}{s / \sqrt{n-1}}$ large sample

Mean σ unknown $t_{n-1} = \dfrac{\bar{x} - \mu}{\hat{s} / \sqrt{n}} = \dfrac{\bar{x} - \mu}{s / \sqrt{n-1}}$ Normal population

Equal population means, $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left\{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right\}}}$,normal population and/or large sample σ_1, σ_2 known

Equal population means, σ_1, σ_2 un known $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left\{\dfrac{s_1^2}{n_1 - 1} + \dfrac{s_2^2}{n_2 - 1}\right\}}}$ large samples

Same population, σ unknown $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\left\{\dfrac{1}{n_1} + \dfrac{1}{n_2}\right\}}}$ normal population and/or large samples

Same population, σ unknown $z = \dfrac{\bar{x}_1 - \bar{x}_2}{\hat{s}\sqrt{\left\{\dfrac{1}{n_1} + \dfrac{1}{n_2}\right\}}}$ large samples

Same population, σ unknown $t_{n_1+n_2-2} = \dfrac{\bar{x}_1 - \bar{x}_2}{\hat{s}\sqrt{\left\{\dfrac{1}{n_1} + \dfrac{1}{n_2}\right\}}}$ normal population

Where $\hat{s}^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

UNIT 5 THE χ^2 –TEST. GOODNESS OF FIT

5.1 Unit Introduction

Welcome to Unit 5 in which you will learn Chi-Square goodness of fit test which is a non-parametric test that is used to find out how the observed value of a given phenomenon is significantly different from the expected value. In Chi-Square goodness of fit test, sample data is divided into intervals. Chi-Square Test is used in Market Research when:

- They need to estimate how closely an observed distribution matches an expected distribution. This is referred to as a “goodness-of-fit” test.
- They need to estimate whether two random variables are independent.

The goodness of fit test is a statistical hypothesis test to see how well sample data fit a distribution from a population with a normal distribution. In other words, it tells you if your sample data represents the data you would expect to find in the actual population.

5.2 Unit Aim

The aim of Chi square test for testing goodness of fit is to decide whether there is any difference between the observed (experimental) value and the expected (theoretical) value.

5.3 Unit Objectives

By the end of the unit you should be able to:

- apply when you have one categorical variable from a single population
- Determine whether sample data are consistent with a hypothesized distribution.
- compare the observed sample distribution with the expected probability distribution
- Determine how well theoretical distribution (such as normal, binomial, or Poisson) fits the empirical distribution.

Terminology

O_i : Observed values

E_i : Expected values

X^2 : Chi square distributions

5.4 Unit Time required

You need 20 hours for this unit

5.5 Unit Topics

5.5.1 An experiment

Table 5.1 shows the results of an experiment in which four coins were thrown 160 times. The results are divided into classes according to the number of heads obtained each time. For comparison the theoretical frequencies predicted by the Binomial distribution have been calculated, assuming that the coins are unbiased and using the parameters $p = ½$, $n = 4$. These frequencies are usually called the **expected frequencies**.

As we would expect there is not exact agreement between the observed and expected frequency for each class, since the frequency for each class is a random variable which varies from sample to sample. Our problem is to test whether the observed frequencies differ significantly from those calculated using the null hypothesis that the distribution is Binomial with $p = ½$, $n = 4$. Let O_i denote the observed and E_i the expected frequency in the ith class. As a first step we might calculate the discrepancy between the two values and sum over the classes, as shown in the third column of Table 5.2 This, of course, produces the result zero. We can avoid this by squaring the discrepancies, as shown in the fourth column of Table 5.2. To add the values in this column as it stands would mean that the difference between 10 and 15 would be given the same weight as the difference between 35 and 40 although the percentage difference in the first case in much greater. It can be shown (using mathematics beyond the scope of this book) that the appropriate statistic to use is

Table 5.1 Frequency **distribution of the number of heads when four coins are tossed**

Number of heads	Observed frequency	Expected frequency
0	15	$160 \times \binom{4}{0}(½)^4 = 10$
1	46	$160 \times \binom{4}{0}(½)^3\,(½) = 40$

		(1)	
2	54	160 x (4)(½)^2 (½)^2	= 60
		(2)	
3	35	160 x (4)(½)(½)^3	= 40
	(3)		
4	10	160 x (4)(½)^4	= 10
	(4)		
	-----		----
	160		160

Table 5.2 Calculation of X^2 for data in Table 5.1

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
15	10	+ 5	25	2.5
46	40	+ 6	36	0.9
54	60	- 6	36	0.6
35	40	- 5	25	0.625
10	10	0	0	0
		0		4.625

$\sum\left\{\frac{(O_i - E_i)^2}{E_i}\right\}$, whose calculation is shown in the last column of Table 5.2. You shall call this statistics X^2. The square emphasizes that it is always a positive quantity.

5.5.2 Degrees of freedom

If the experiment described in Section 5.1 was repeated with the same coins, different values of O_i and consequently of O_i would have been obtained. In other words, X^2 has a sampling distribution. This sampling distribution is approximately the same as a theoretical distribution

known as the X^2 **(or chi-squared) distribution** ('chi' is pronounced as the 'ki' in 'kite'). (X^2 is often used to denote X^2 as well as the sampling distribution of X^2.) It can be shown that the form of the X^2-distribution depends on the 'degrees of freedom' (and consequently it is often written X^2). You have already met the term 'degrees of freedom' in Section 2.5.2. In the present instance v is the number of expected frequencies which can be varied independently. In the example in Table 5.2 there are five expected frequencies but, since the total frequency is fixed at 160, which four frequencies are given, the fifth is known, and so we have $v = 4$. The fact that the total frequency must equal 160 is known as a **constraint**. In general, v is given by degrees of freedom = number of classes – number of constraints

Figure 5.1 shows the form of the X^2 –distribution for some different values of v.

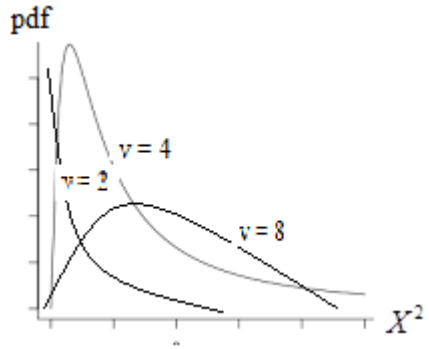


Figure 5.1 The x^2 –distribution for various values of v

5.5.3 Significance testing using the X^2 –distribution

As the discrepancy between the observed and expected frequencies increases so does the value of X^2. To decide if a value of calculated using a given null hypothesis is significant we use Table A6. This gives the percentage points of the X^2 – distribution for different values of v. it is a one-tall table and gives the probability that X^2v exceeds a certain value. This probability is given by the shaded area in Figure 2. For example, the probability is 5% that X²⅔ > 7.81, 99.5% that $\chi^2_{14} > 4.07$, etc. Using this table we are in a position to calculate if the value of $X^2 = 4.625$ calculated in Table 2 is significant. The critical region is $\chi^2_4 > 9.49$ at the 5% significance level, so the value of X^2 obtained is not significant and we retain the null hypothesis that the

distribution of throws is Binomial with p = ½ n = 4. Although in a X^2 –test we normally test if X^2 exceeds the critical value, very low values of X^2 should regarded with suspicion since they also have a low probability. For example there is only a 5% probability that $\chi_4^2 < 0.711$. Such a low value of X^2 is almost too good to be true and may indicate that the sample is not random, or the data are fictitious.

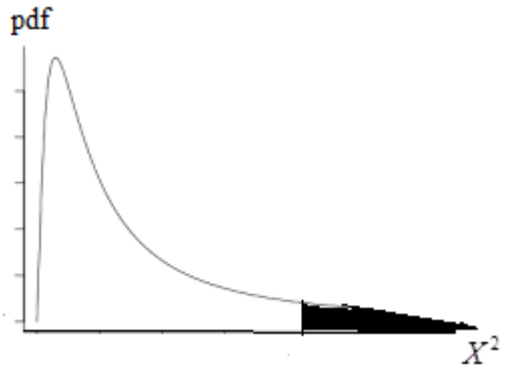


Figure 5.2 Schematic diagram showing the critical region for a X^2 –test

Conditions in which the x^2 –test is applicable

The X^2 – test should only be used when the sampling distribution of X^2 approximates closely to a X^2 – distribution. The conditions for this are

(i) the total frequency is not less than 50,

(ii) the expected class frequencies are not less than 5.

If the second condition is not met, it can be complied with by combining one or more classes. The agreement between the X^2 and X^2 distributions can be improved when v = 1 by making an adjustment known as **Yates' correction**, which involves reducing each value of $|O_i - E_i|$ by ½.

Example 5.1

On a national basis the success rate for people taking their driving test for the first time is 40%. A driving instructor's records show that for the 50 pupils of his who took the test for the first time last year, 25 passed. Does his success rate differ from the national success rate?

Table 5.3 Driving test results

	O_i	E_i	$O_i - E_i$	$\lvert O_i - E_i \rvert - ½$	$\{\lvert O_i - E_i \rvert - ½\}^2/E_i$
Pass	25	20	+5	4½	0.0125
Fail	25	30	-5	4½	0.675
			0		$X^2 = 1.6875$

You take the null hypothesis that the instructor's results do not differ from the national success rates, that is

$H_o : p = 0.4$

$H_1 : p = 0.4$

This gives Table 5.3 for the observed and expected frequencies. Notice that to use all the information we include the frequencies for pass and fail. In this case $v = 1$, since we have two classes and one restriction (that the total frequency is 50). Since $v = 1$, Yates' correction has been applied. The critical region at the 5% level as $\chi_1^2 > 3.84$. The value of X^2 obtained is not significant at the 5% level and there is no evidence that the driving instructor's success rate differs from the national rate.

A problem similar to this was solved in Section 5.5.2 using z but the question asked was 'Is the instructor's success rate *better* than the national one?' so that a one-tail test was used. The X^2 – test is basically a two-test since X^2 is always positive and takes no account of the sign of the difference between the observed and expected frequencies. For a one-tail test the z-method is preferable.

5.1 UNIT ACTIVITY

(1) The following table purports to be a sample of 100 random digits. Test whether the frequencies of the digits significantly from expectation.

11584	42689	08394	57019	33922	22413	21138	83541	53216	74935
51186	49197	30157	28543	51328	49788	31489	18971	28719	97121

(2) Two fair dice are thrown 432 times. Find the expected frequencies of scores of 2, 3, 4, . . ., 12 . Two players *A* and *B* are each given two dice and told to throw them 432 times, recording the results.

The frequencies reported are given in Table 4.

Table 5.4 Two fair dice are thrown 432 times

Scores	2	3	4	5	6	7	8	9	10	11	12
A's frequency	18	33	28	54	62	65	66	42	30	27	7
B's frequency	14	22	34	51	58	73	63	45	38	25	9

Is there any evidence that either pair of dice is biased? What can be said about *B*'s alleged results?

5.5.4 Test a distribution for normality

In section 9.5 a normal distribution was fitted to data. You are now in a position to test how good the fit is. Although the Normal distribution is continuous, the data were divided into classes. We can use the observed and expected frequencies in each class to calculate X^2 and test the null hypothesis that the observed values come from a Normal distribution with a specified mean and s.d. Table 5.5 gives the observed and expected frequencies taken from Table 5.4.

The first four and the last three classes have been combined so that all E_i are > 5. The table shows the calculation of X^2. In this example there are six classes and three constraints: the total frequency is 50, the mean (698.3) and the s.d. (18.6) are calculated from the data. So we have v = 6 –3 = 3

The critical region is $\chi_3^2 > 7.81$. Thus the value of X^2 is not significant and the null hypothesis is retained. If the values for the mean and s.d. had been given rather than calculated from the data, then there would have been only one constraint.

Table 5.5 Calculation of X^2 for data in Table 5.4

Class	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
>649.5	0 }	0.2 }		
649.5 – 659.5	1 }	0.7 }		
659.5 – 669.5	3 } 7	2.1]	- 0.8	0.082
669.5 – 679.5	3	4.8		
679.5 – 689.5	7	8.2	- 1.2	0.176
689.5 – 699.5	15	10.2	4.8	2.259
699.5 – 709.5	7	10.1	3.1	0.951
709.5 – 719.5	7	7.3	0.3	0.012
719.5 – 729.5	4 }	4.1 }		
729.5 – 739.5	3 } 7	1.6 }6.4	0.3	0.056
> 739.5	0 }	0.7 }		
			0	$\overline{X^2 = 3.536}$

5.5.5 Testing the fit of a Binomial distribution

The example given in Section 5.5.1 was one in which a distribution was tested to see if it was Binomial with a given value of p. In a Binomial distribution was fitted to data and p was not given but calculated from the observed frequencies. The observed and expected frequencies obtained are reproduced in Table 6. You wish to test the null hypothesis that the observed frequencies are binomially distributed with $p = 0.49$, $n = 4$.

Table 5.6 Calculation of X^2 for data in Table 5.3

Numbers of girls	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
0	8	6.8	1.2	0.212
1	25	26.0	- 1.0	0.038
2	37	37.5	- 0.5	0.007
3	23	24.0	- 1.0	0.024
4	7	5.8	+ 1.2	0.248
	1.00	100.1	- 0.1	$X^2 = 0.547$

In this case, besides the constraint that the total frequency is 100, there is the added constraint that p was calculated from the data so that $v = 5 - 2 = 3$

The critical region is $X^2 > 7.81$. Thus X^2 is not significant at the 5% level , the null hypothesis is retained and the observed distribution conforms to a Binomial distribution the $p = 0.49$, $n = 4$.

5.5.6 Testing the fit of a Poisson distribution

In a Poisson distribution was fitted to data: the observed and expected frequencies are reproduced in Table 5.7. We wish to test the null hypothesis that the data are Poisson distributed with a mean equal to the mean of the data, which was found to be 0.37.

The calculation of X^2 is shown in Table 5.7. There are two constraints: first, that the total frequency is 100 and, second, that the mean is calculated from the data. There are three classes since the last three have been combined, giving. $v = 3 - 2 = 1$ since $v = 1$, Yates' correction has been applied.

The critical region is $x^2 > 3.84$. Thus X^2 is not significant at the 5% level and we retain the null hypothesis that the data are Poison distributed with a mean of 0.37. If a value for the mean and been given rather than calculated from the data, then there would have been only one constraint.

Table 5.7 Calculation *of* X^2 *for the data in Table 5.2*

Numbers of girls O_i	E_i	$O_i - E_i$	$\lvert O_i - E_i \rvert - ½$	$\{\lvert O_i - E E_i \rvert - ½\}^2 \lvert E$i
0	71	69.1	+ 1.9	1.4	0.028
1	23	25.6	- 2.6	2.1	0.172
2	4	4.7 } 5.3			
3	2	0.6 } 5.3	+ 0.7	0.2	0.008
>4	0	0.0 } 5.3			
		0			$X^2 = 0.208$

5.2 UNIT ACTIVITY

(1) Two dice were thrown 216 times, and the number of 6s at each throw were counted. The results were as in Table 5.8.

Test the hypothesis that the distribution is Binomial with the parameter $p = ⅙$.

Explain how the test would be modified if the hypothesis to be tested is that the distribution is Binomial with the parameter p unknown. (Do not carry out the test.)

Table 5.8 Two dice were thrown 216 times

Number of 6*s*	0	1	2	Total
Frequency	130	76	10	216

(2) A man kept count of the number of letter he received each day over a period of 78 days (excluding Sundays). The frequencies of the number of letters per day and of the Poisson distribution with the same mean (0.88) and total frequency are given in Table 5.9

Do the observations show a significant departure from a Poisson distribution?

Table 5.9 Number of letter he received each day over a period of 78 days

Letter Per day	Number of days	Poisson frequencies
0	33	32.4
1	26	28.4
2	14	12.6
3 or more	5	4.6

(3) For a period of six months 100 similar hamsters were given a new type of feedstuff. The gains in mass are recorded in Table 5.10

Table 5.10 Types of feedstuff

Gain in mass (g) x	Observed frequency f
$-\infty < x < -10$	3
$-10 < x < -5$	6
$-5 < x < 0$	9
$0 < x < 5$	15
$5 < x < 10$	24
$10 < x < 15$	16
$15 < x < 20$	14
$20 < x < 25$	8
$25 < x < 30$	3
$30 < x < \infty$	2

It is though that these data follow a Normal distribution, with mean 10 and variance 100. Use the X^2 –distribution at the 5% level of significance to test this hypothesis.

Describe briefly how you would modify this test if the mean and variance were unknown.

Table 2.11 Types of feedstuff

x	0	1	2	3	4	5	6	7	8 or more
f	3	7	12	10	8	5	3	2	0

(4) A group of students are required to carry out an experiment in which the results are expected to have a Poisson distribution. There results for 50 replications of the experiment are as in Table 2.11.

Estimating the mean of the Poisson distribution by the mean of these observations, obtain the appropriate estimated frequencies and test how well these data fit the expected form. State what you might suspect about the students' results.

5.5.7 Contingency tables

Sometimes we wish to test if two attributes of a population are associated, i.e. if they occur together. A common example is whether inoculation is associated with the prevention of disease. In this case one

Attribute is whether or not a person is inoculated and the other attribute is whether or not they are attacked by the disease. Table 12 gives the (fictitious) results for a random sample of 100, where the total frequency has been divided into rows and columns according to attribute. This is known as a **contingency table**. On the null hypothesis that inoculation has no effect in preventing the disease, the marginal totals can be used to calculate the expected frequencies as follows. First we have

$$P(\text{inoculated}) = \frac{70}{100} \qquad P(\text{attacked}) = \frac{35}{100}$$

$$P(\text{not inoculated}) = \frac{30}{100} \qquad P(\text{not attacked}) = \frac{65}{100}$$

Since according to our null hypothesis the events attacked and not attacked are independent of the events inoculated and not inoculated we have, using the product law.

P (inoculated and attacked)) $= \frac{70}{100} * \frac{35}{100}$

Table 5.12 Contingency table for inoculation and disease

	Attacked	Not attacked	
Inoculated	20	50	70
Not inoculated	1	15	30
	35	65	100

Table 5.13 Calculation of expected frequencies for Table 5.12

	Attacked	Not attacked	
Inoculated	$\frac{70 \times 35}{100} = 24.5$	$\frac{70 \times 65}{100} = 45.5$	70
Not inoculated	$\frac{30 \times 35}{100} = 10.5$	$\frac{30 \times 65}{100} = 19.5$	30
	35	65	100

Table 5.14 Calculation of X^2 for data in Table 5.12

O_i	E_i	$O_i - E_i$	$\lvert O_i - E_i \rvert - ½$	$\{\lvert O_i - E E_i \rvert - ½\}^2 / Ei$
20	24.5	-4.5	4	0.65
50	45.5	+4.5	4	0.35
15	10.5	+4.5	4	1.52
15	19.5	-4.5	4	0.82
				$X^2 = 3.34$

P(inoculated and not attacked) $= \frac{70}{100} * \frac{35}{100}$

P (not inoculated and attacked) $= \frac{30}{100} * \frac{65}{100}$

P (not inoculated and not attacked)= $\frac{30}{100} * \frac{65}{100}$

The expected frequencies are found by multiplying these probabilities by 100, and are shown in Table 13. Notice that the row and column totals are the same as before. It can be seen that the expected frequencies could be calculated more directly using the rule

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

You are now in a position to calculate X^2. How many degrees of freedom will X^2 have? The answer is v = 1 since in the calculation of the expected frequencies, when one frequency was found, the other frequencies could be found by subtraction from the marginal totals. Table 14 shows the calculations of X^2 using Yates' correction. This value of X^2 is not significant at the 5% level. The null hypothesis is retained and there is no evidence that the inoculation is effective against the disease.

5.5.8 Larger contingency tables

Table 5.15 shows a random sample of houses classified by region and type. Does the type of housing vary between regions? Taking as our null hypothesis that the type of housing and the region are independent, the expected frequencies are calculated in Table 5.16.

Table 5.15 Contingency table for housing and region.

	Detached	Semi-detached	Terraced	
Kabwata	95	297	242	634
Woodlands	175	417	362	954
Chelston	703	994	861	2558
	973	1708	1465	4146

Table 5.16 Expected frequencies for data in Table 5.15

	Detached	Semi-detached	Terraced	
Kabwata	$\frac{973 \times 634}{4146} = 149^*$	$\frac{1708 \times 634}{4146} = 261^*$	$\frac{1465 \times 634}{4146} = 224$	634
Woodlands	$\frac{973 \times 954}{4146} = 224^*$	$\frac{1708 \times 954}{4146} = 393^*$	$\frac{1465 \times 634}{4146} = 337$	954
Chelston	$\frac{973 \times 2558}{4146} = 600^*$	$\frac{1708 \times 2558}{4146} = 1054^*$	$\frac{1465 \times 2558}{4146} = 904$	2558
	973	1708	1465	4146

Table 5.17 Calculation of X^2 for data in Table 5.15

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2/E_i$
95	149	- 54	19.6
297	261	36	5.0
242	224	18	1.4
175	224	- 49	10.7
417	393	24	1.5
362	337	25	1.9
703	600	103	17.7
994	1054	- 60	3.4
861	904	- 43	2.0
		0	$X^2 = 63.2$

There are four degrees of freedom, since, when the frequencies marked with an asterisk have been calculated, the other expected frequencies can be found by subtracting from the marginal totals. The calculation of X^2 is given in Table 5.17. The critical region is $\chi_4^2 > 9.49$ so the value of X^2 is significant at the 5% level: in fact it is significant at the 0.1% level. The null hypothesis is rejected and the result indicates that there is a strong association between type of housing and region.

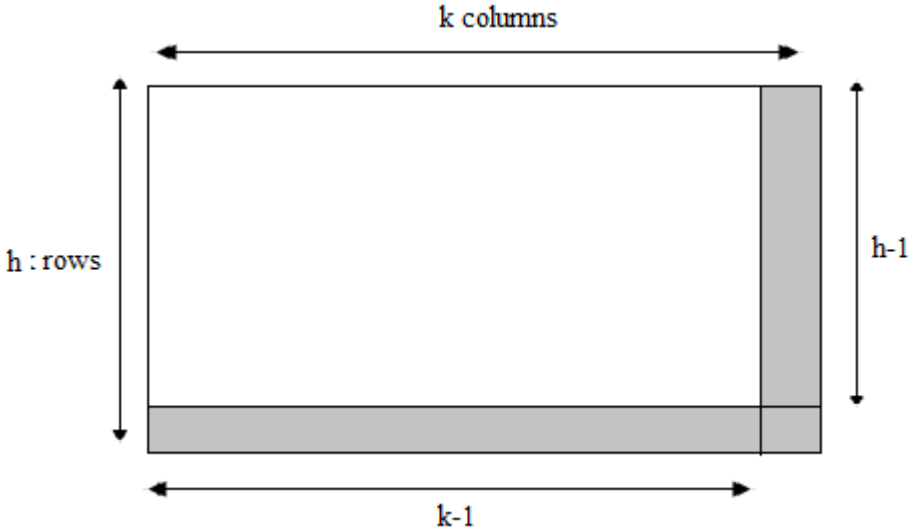


Figure 5.3. Diagram to illustrate the degree of freedom of an *h x k* contingency table

A table with *h* rows and *k* columns is called an *h* x *k* contingency table. Its degrees of freedom are $(k - 1)(h - 1)$, since, as indicated in Figure 3, the frequencies in the last row and column can be calculated by subtraction of the other frequencies from the marginal totals.

5.3 UNIT ACTIVITY

(1) Explain why the X^2 statistic calculated from a 2 x 2 contingency table only has one degree of freedom. Out of 100 rats fed with diet *A*, 65 showed signs of vitamin deficiency, while out of 100 rats fed with diet *B*, only 53 suffered from vitamin deficiency. Do the proportions of vitamin-deficient rats on the two diets differ significantly?

(2) Table 5.18 summarises the incidence of cerebral tumours in 141 neurosurgical patients. the expected frequencies on the hypothesis that there is no association between the type and site of a tumour. Use the X^2-distribution to test this hypothesis.

Table 5.18 Summarises the incidence of cerebral tumours in 141 neurosurgical patients

Site of tumour	Type of tumour: Benign	Malignant	Others
Frontal lobes	23	9	6
Temporal lobes	21	4	3
Elsewhere	34	24	17

Table 5.19 Oral tests are conducted by three examiners separately

	A	*B*	*C*	*Total*

Credit	10	5	13	28
Pass	31	38	28	97
Fail	29	20	26	75
	70	63	67	200

(3) Oral tests are conducted by three examiners separately. The numbers of candidates in the categories credit, pass, fail are as shown in Table 5.19.

Use a X^2 –test to examine the hypothesis that the examiners do not differ in their standards of awards. State the assumptions made.

5.5.9 Relationship between X^2 and Normal distribution

If X^1, X^2, . . ., X_n are *independent* N (0, 1) distributions, then χ_1^2 is said to be a X^2 random variable with *n* degrees of freedom. In Particular if X is N(0, 1) then X^2 is a X^2 random variable with one degree of freedom. In Section 5.5.2 we stated that the statistic has a sampling distribution which is X^2. We will now prove that this is true for the particular case in which there are two classes. Let the probability of being in the first class be *p* and the probability of being in the second class be *q* where $p + b = 1$. If the observed frequencies are n_i and n_2, the total frequency is $n = n_1 + n^2$. The expected frequencies will be *np* and $n(1 – p)$. This information is summarized in Table 20.

Then

$$\chi^2 = \sum \left\{ \frac{(O_i - E_i)^2}{E_i} \right\}$$

$$= \frac{(n_1 - np)^2}{np} + \frac{(n_2 - np)^2}{np}$$

$$= \frac{(n_1 - np)^2}{np} + \frac{(n_2 - (n - np))^2}{np}$$

$$= = \frac{(n_1 - np)^2}{np} + \frac{(np - n_1)^2}{np}$$

$$= \frac{(np - n_1)^2}{npq}(p + q)$$

$$= \frac{(np - n_1)}{npq}, (\text{since} \quad p + q = 1)$$

Table 5.20 Observed and expected frequencies for two classes

Observed frequency		Expected frequency
Class 1	n_1	np
Class 2	n_2	nq

n_1 is Binomially distributed since we have n trials and a probability p that the result of a trial is in class 1. If n is large then the distribution of n_1 can be approximated by a Normal distribution with mean np and s.d. $\sqrt{npq}$ Thus $\frac{np - n_1}{\sqrt{npq}}$ is a standardized Normal deviate z and $X^2 = z^2$. From the definition given at the beginning of this section $\chi_1^2 = z^2$. Thus in this case $X^2 = \chi_1^2$. This is consistent, with the rule for calculating the degrees of freedom since in this case there are two classes and one constraint : $n_1 + n_2 = n$.

5.4 UNIT ACTIVITY

(1) In a large town, the number of road accidents reported daily over 300 working days gave the results in Table 5.21

Table5.21 Number of accidents reported

Number of accidents reported in a day (x)	0	1	2	3	4	5	6	7	8
Number of days (f)	17	43	69	50	28	13	8	4	

Compare these frequencies with those of a Poisson distribution having the same mean and total frequency. Comment on the agreement.

(2) (a) A large haulage company employers drivers to work a 'round-the-clock' system .The drivers' union is concern that some periods are more dangerous than others. The management contests this, claiming that no one period is more dangerous than any other. The following data on the incidence of traffic accidents by time of days, for this group over the previous two years is presented (Table 5.22). Are these results consistent with the management's claim?

Table 5.22 Haulage company employers' drivers work hours

Time of day (24 hr clock)	00.01–04.00	04.01–08.00	08.00–12	12.01–16.00	16.01–20.00	20.01–24.00	Total
Number of accidents	14	16	24	22	24		20

(b) The same company wishes to examine whether there is an association between accident proneness and coloured blindness. The results for a group of 80 drivers (with a minimum of five years' employment) are as in Table 5.23.

Table 5.23 Records of accidents due to Coloured blind

		Coloured blind	
		No	Yes
Accidents during Last five years	None	22	5
	One or more	38	15

Is there any evidence of an association between colour blindness and accident proneness?

(3) Explain how to calculate the degrees of freedom for the X^2-statistic in

(a) a goodness-of-fit test,

(b) a test of no association of the two factors in an h x k contingency table.

An ecologist collected organisms of a particular species from three beaches and counted the number of females in each sample (the remainder were males).

Table 5.24 Ecologist counts of organisms

Beach	1	2	3
Number of females	44	86	110
Total number in sample	100	200	200

Test if the proportion of females differed significantly between the beaches.

Find the percentage of females at each beach and comment on the results

(4) In routine test of seeds of a certain flower, batches of hundreds are allowed to germinate and the colour of the resulting blooms is noted. 25% are expected to be red. It is suspected that a certain laboratory assistant may have been recording data carelessly. In eight successive batches, he returns the number of red blooms (out of 100 in each case) as 18, 20, 26, 21, 37, 16, 30, 32. Does this evidence support the suspicion?

(5) Table 5.25 gives the distribution for the number of heavy rainstorms reported by 330 weather stations in the United States of America over a one-year period.

(a) Find the expected frequencies of rainstorms given by the Poisson distribution having the same mean and total as the observed distribution.

(b) Use the X^2-distribution to test the adequacy of the Poisson distribution as a model for these data.

Table 5.25 Distribution for the number of heavy rainstorms

Number of rainstorms (x)	0	1	2	3	4	5	More than 5
Number of stations (f) reporting x rainstorm	101	114	74	28	10	2	0

(5) A doctor made a comparison of ointment *A* and ointment *B* for the cure of a skin disease. She chose at random which ointment to use for each patient and assessed the result after two weeks. Of the 95 patients receiving ointment *A,* 85 were cured, while of the 105 patients receiving *B*, 78 were cured. Is there evidence that one ointment is butter than the other? Suppose ointment *A* had been used by doctor *C* on her patients and ointment *B* had been used by doctor *D* on his patients and the same numbers of patients and cures resulted. Would this affect your interpretation of the data?

(6) Given that a random variable which has a X^2 –distribution with n degrees of freedom can be represented as the sum of the squares of n independent standardised normal variables, show that the sum of two independent X^2 variables is itself a X^2 variable and explain, with justification, how the degrees of freedom of the three variables are related.

In making a scientific measurement a technician has to estimate the position of a needle between two successive divisions of a scale in order to obtain the final digit of his reading. Table 5.26 shows the final digit recorded in 200 such measurements. State what frequencies you would have expected, and justify your choice. Test whether the data are consistent with your expectation, and comment as fully as possible on the result of your analysis.

Table 5.26 Measurements of readings

Final digit	0	1	2	3	4	5	6	7	8	9
Frequency	11	27	16	23	13	31	14	18	24	23

(7) A chain store offers rain hats for sale in three colours: red, blue and green. A sales manager wonders if the colour preference of customers differs in London and a county town. In the London store 48 hats were sold in a week and of these 19 were red, 14 blue and 15 green. In the county two branch 32 hats were sold and of these 6 were red, 16 blue and 10 green. Is there evidence of differential colour preference between London and the county town? Are any colours significantly preferred to the others?

5.0 UNIT SUMMARY

Chi-squared (X^2) **test** for frequencies, test statistic X^2 where

$$\chi^2 = \sum \left\{ \frac{(O_i - E_i)^2}{E_i} \right\}$$

$$X^2 = \frac{(O_i - E_i)^2}{Ei}$$

If v = 1, use **Yates' correction**:

$$\chi^2 = \sum \left\{ \frac{(|o_i - E_i| - 0.5)^2}{E_i} \right\}$$

X^2 has a χ^2_v distribution with v degrees of freedom:

Type of test	Conditions	v
Binomial fit, n classes	(a) p know	(a) n - 1
	(b) p unknown	(b) n - 2
Poisson fit, n classes	(a) known	(a) n - 1
	(b) unknown	(b) n - 2
Normality, n classes	(a) u. o known	(a) n -1
	(b) u. o unknown	(b) n - 3
Contingency table,	h x k	(h – 1)(k – 1)

UNIT 6 CORRELATION ANALYSIS

6.1 Unit Introduction

Welcome to Unit 6 in which you learn how to carry out the correlation analysis using statistical methods. You will learn what correlation is and how to calculate and interpret the correlation coefficients.

6.2 Unit Aim

The aim of this Unit is to teach you on how to carry out correlation analysis..

6.3 Unit Objectives

By the end of the Unit you should be able to:

- measure the degree of association between numerical variables
- Establish if there are possible connections between variables.
- investigating the relationship between two quantitative, continuous variables

Terminology

$\sum$ **:** Summation

r : Correlation coefficient

6.4 Unit Time required

You need 20 hours for this unit

6.5 Unit Topic

6.5.1 Scatter diagrams

In the unit you will develop a theory to test whether two *attributes* of a population are associated with each other: the theory of this unit is concerned with the relationship between two variates for a population. Table 6.1 gives the marks, X and Y, obtained by a group of students in each of two mathematics examinations.

As we would expect, each student tends to obtain similar marks in both papers. This is illustrated graphically in Figure 6.1. Such a graph is known as a scatter diagram. Each

Table 6. 1 Marks obtained in a Mathematics examination

Student	X, mark in paper 1	Y, mark in paper 2
A	42	31
B	84	83
C	50	42
D	42	60
E	33	28
F	50	63
G	69	59
H	81	92
I	50	73
J	*35*	*40*

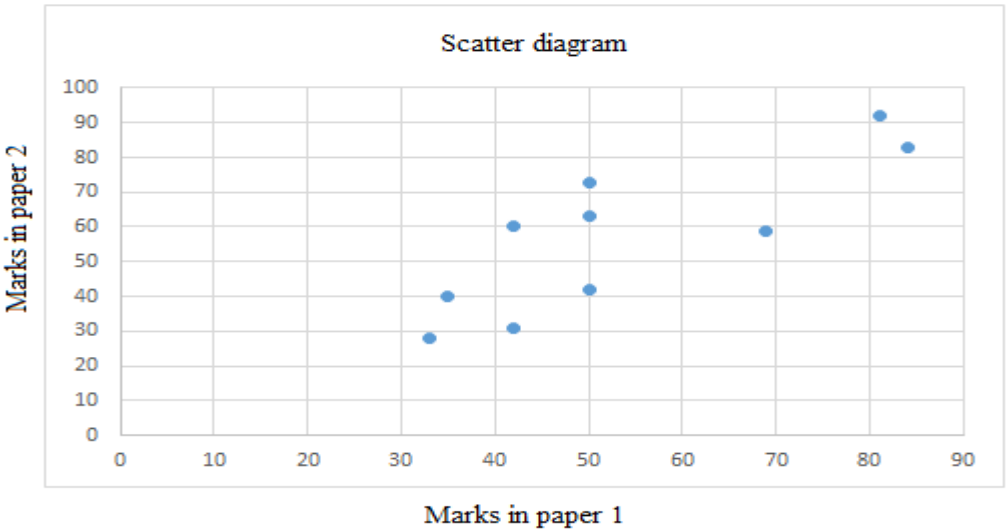


Figure 6.1 Scatter diagram for the data in Table 6.1

Figure 6.1 shows a scatter diagram of X and Y. You just plot as coordinates of (x,y) and the relationship between X and Y and compare your graph with standard graphs in figure 6.4. The data shown in Table 6.1 are shown on a scatter diagram in figure 6.2. In figure 6.2 the graph shows an inverse correlation.

The data in Table 6.3 gives the scores obtained on each of two dice when they were thrown together 15 times, and the corresponding scatter diagram is shown in figure 6.3.

In this case there is no obvious correction.

Table 6.2 Number of working horses and number of tractors in Zambia (in ten thousands) (Data from An Agricultural Statistics,*)*

Year	Number of horses	Number of tractors
1938	67	6
1942	49	10
1946	44	18
1948	38	23
1950	29	30
1952	21	33
1954	15	39
1956	11	43

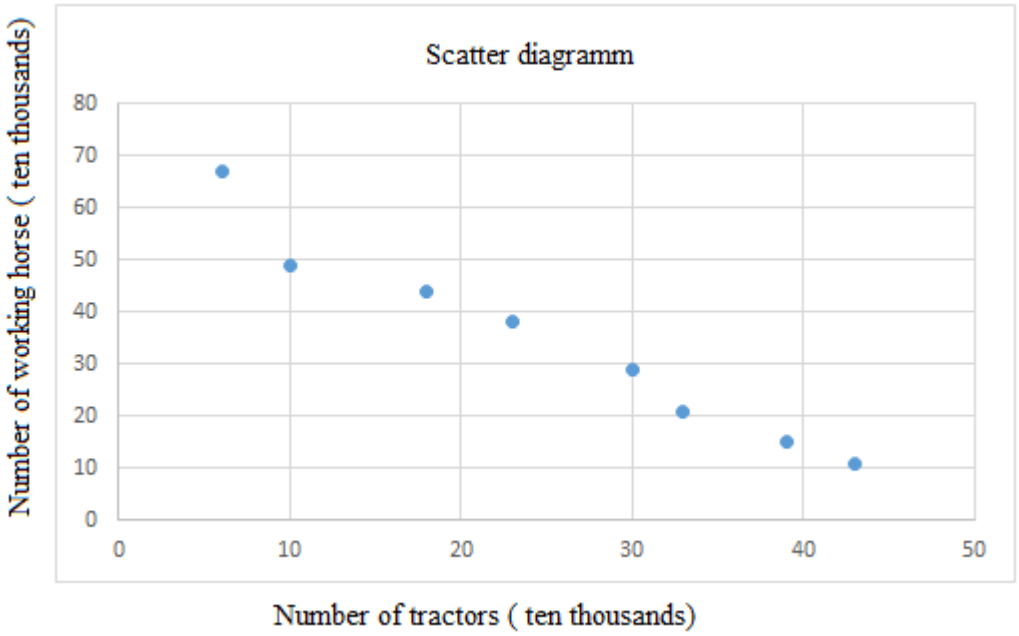


Figure 6.2 Scatter diagram for the data in Table 6.2

Table 6.3 Scores of two dies

Score on first die	Score on second die
2	5
4	2
5	4
5	1
3	4
6	3
4	2
5	3
3	5
6	1
5	4
6	6
2	3
3	2
1	2

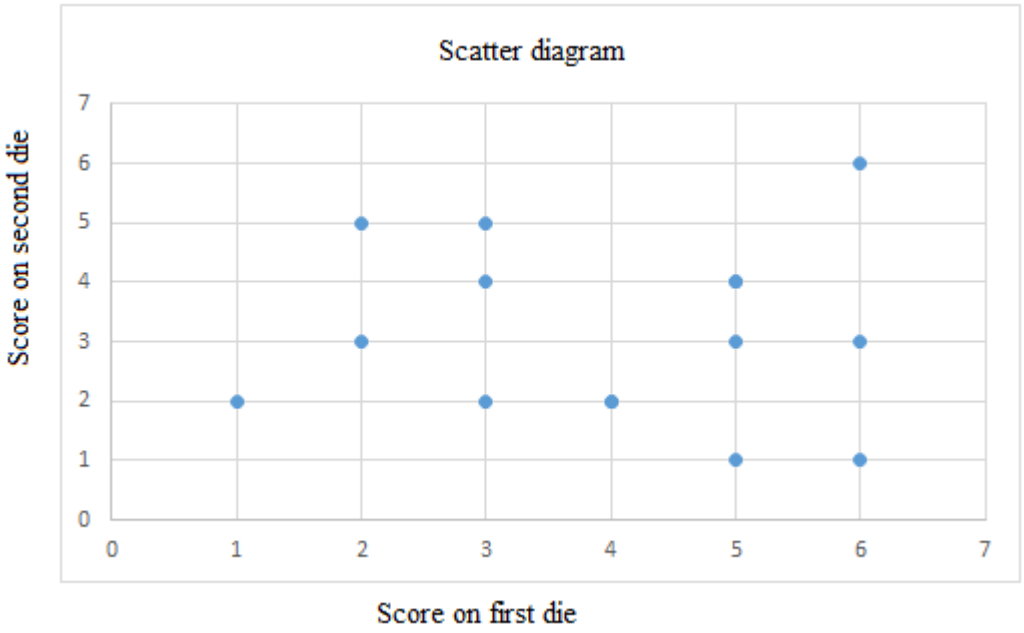


Figure 6.3. Scatter diagram for the data in Table 6.3

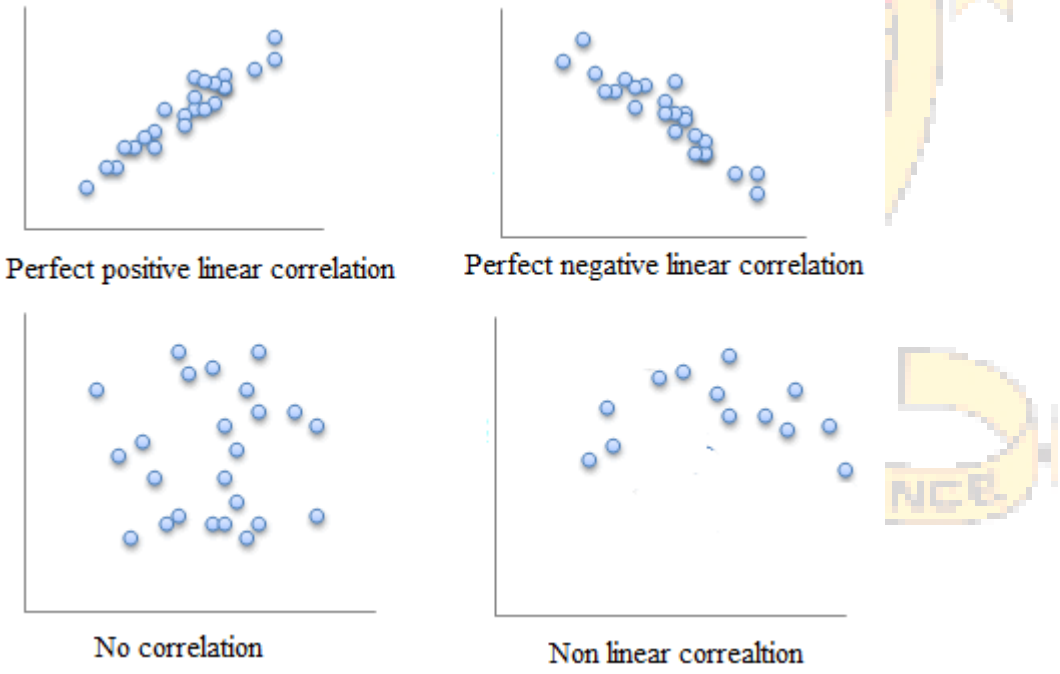


Figure 6.4 Standard scatter diagrams

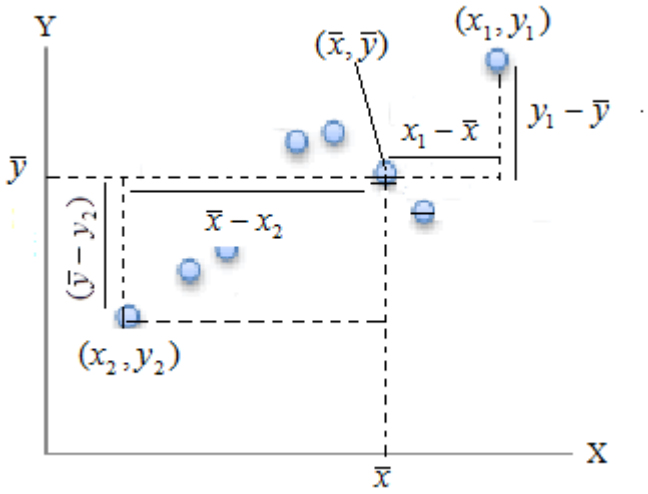


Figure 6.5. Diagram showing why cov(*X*, *Y*) is positive for positive correlation

6.5.2 Measurement of correlation

The scatter diagrams in Figure 6.4 show a difference in the degree of correlation which can be readily appreciated by eye. How can we measure the degree of correlation quantitatively? You know an expression $E(X-\mu_X)(Y-\mu_Y)$ which is termed as covariance, cov(X, Y), of two random variables X and Y, and it was proved there that cov(X, Y) is zero if X and Y are independent. If X and Y are not independent cov(X, Y) differs from zero. This can be seen by reference to Figure 6.5. For the point (x_1, y_1), $x_1 - \bar{x}$ *and* $y_1 - \bar{y}$ are both positive and therefore their product is positive also. For the point (x_2, y_2), $x_2 - \bar{x}$ *and* $y_2 - \bar{y}$ are both negative so that their product is positive. In fact for the majority of points $(x_i - \bar{x})(y_i - \bar{y})$ is positive leading to a positive estimate for cov(X, Y). Similarly for an inverse correlation such as that shown in Figure 6 the product $(x_i - x)(y_i - y)$ is generally negative and the estimated value of cov(X , Y) is also negative. Only in cases where there is no correlation will cov(X, Y) be zero since then positive and negative values of $(x_i - \bar{x})(y_i - \bar{y})$ are equally likely.

It can be shown that the unbiased estimate of cov(X, Y) is

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

This depends not only on the degree of correlation but also on the spread of values of X and of Y. To form a basis of comparison the covariance must be standardised so that it does not depend on the scales on the X and Y axes. This can be done by making $x_i - \bar{x}$ into a standardised deviate by dividing by $\hat{s}_x$, the unbiased estimate of σ_x, and correspondingly for Y. This gives as an expression for the degree of correlation

$$\frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{\hat{s}_X} * \frac{(y_i-\bar{y})}{\hat{s}_Y}$$

This statistic is called *r* and is known as the (product moment) **correlation coefficient**.

Substituting

$$\hat{s}_X = \sqrt{\left\{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}} \text{ and } \hat{s}_Y = \sqrt{\left\{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2\right\}}$$

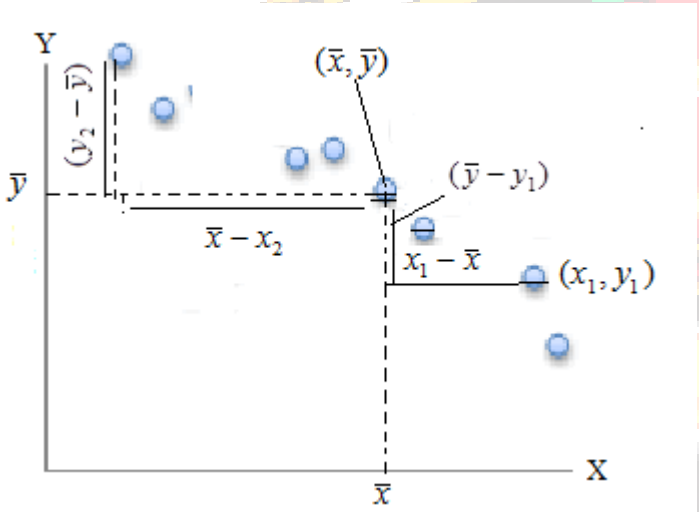


Figure 6.6. Diagram showing why cov(X, Y) is negative for negative correlation.

This becomes

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2\right]}} \qquad (1)$$

To simplify calculation this formula can be rewritten using the following relationships (see equation (3.7.1)):

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2/n$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2/n$$

$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ can also be rewritten in an alternative form, since

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}\bar{x}y_i-\sum_{i=1}^{n}x_i\bar{y}+\sum_{i=1}^{n}\bar{x}\bar{y}\\ &= \sum_{i=1}^{n}x_iy_i-\bar{x}\sum_{i=1}^{n}y_i-\bar{y}\sum_{i=1}^{n}x_i+n\bar{x}\bar{y}\\ &= \sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}-n\bar{y}\bar{x}+n\bar{x}\bar{y}\\ &= \sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}\\ &= \sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i\right)/n\end{aligned}$$

Using these alternative forms, equation (1) becomes

$$r=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{\left[\left\{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\right\}\left\{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2\right\}\right]}} \qquad (2)$$

Table 6.4 Life expectancy and number of people per physician.

Country	Life expectancy (years)	People per physician (thousands)
Burma	47	9.6
Canada	72	0.7
China	50	3.4
Haiti	44	13.4
Malaysia	57	4.7
Madagascar	41	10.0
United Kingdom	71	0.9
Pakistan	47	6.1
Turkey	55	2.3
Venezuala	64	1.2
Zambia	43	16.1

It can be shown that r can only take values between + 1 and – 1. For exact direct linear correlation, i.e. when the points on the scatter diagram lie exactly on a straight line of positive slope, $r = + 1$: for exact inverse linear correlation, $r = - 1$. If there is no correlation r is zero. It is important to remember that r only measures *linear* correlation. For the scatter diagram shown in Figure 7 r would be close to zero: there is obviously correlation between X and Y but the relationship is not linear. Conversely for the data shown in Table 6.4 and illustrated in figure 8 r takes the value – 0.83 even though the relationship is evidently not linear. In fact we should not expect a linear relationship since, however few physicians there are, the life expectancy can never be negative!

Example 6.1

Calculate r for the data in Table 6.4.

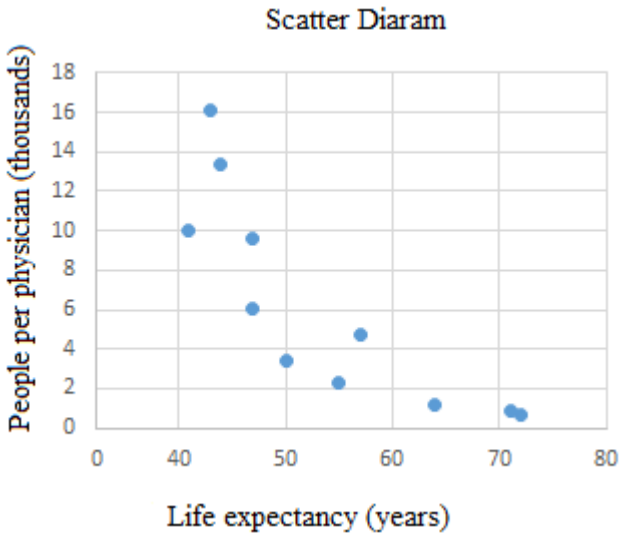


Figure 6.8 Scatter diagram for the data in Table 6.4

Denoting the marks in paper 1 and paper 2 by *X* and *Y* respectively, the calculation is shown in Table 6.5. Substitution in equation (2) gives = 0.825

This rather tedious calculation can sometimes be simplified by using coded values. If we use coded values u_i and v_i for x_i and y_i such that

$$x_i = A + Bu_i \qquad \Rightarrow \bar{x} = A + B\bar{u}$$

$$y_i = C + Dv_i \qquad \Rightarrow \bar{y} = C + D\bar{v}$$

Table 6.5 Calculation of r for data in Table 6.1

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
42	31	1764	961	1302
84	83	7056	6889	6972
50	42	2500	1764	2100
42	60	1764	3600	2520
33	28	1089	784	924
50	63	2500	3969	3150
69	59	4761	3481	4071
81	92	6561	8464	7452
50	73	2500	5329	3650
35	40	1225	1600	1400
$\sum x = 536$	$\sum y = 571$	$\sum x^2 = 31\,720$	$\sum y^2 = 36841$	$\sum xy = 33541$

(then from equation (1)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}$$

$$r = \frac{\sum_{i=1}^{n}(Bu_i - b\bar{u})(Dv_i - D\bar{v}0}{\sqrt{\left[\sum_{i=1}^{n}(Bu_i B\bar{u})^2 \sum_{i=1}^{n}(Dv_i - D\bar{v})^2\right]}}$$

$$r = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\left[\sum_{i=1}^{n}(u_i - \bar{u})^2 \sum_{i=1}^{n}(v_i - \bar{v})\right]}}$$

which using the identity proved above (see equation(3) becomes

$$r = \frac{n\sum_{i=1}^{n} u_i v_i - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i}{\sqrt{\left[\left\{n\sum_{i=1}^{n} u_i^2 - \left(\sum_{i=1}^{n} u_i\right)^2\right\}\left\{n\sum_{i=1}^{n} v_i^2 - \left(\sum_{i=1}^{n} v_i\right)^2\right\}\right]}} \qquad (3)$$

Modern calculators simplify the computation of r, many of them being able to calculate it directly when the values of X and Y are fed in. However, it is wise to draw a scatter diagram first to whether or not the calculation of r is appropriate. (What r is calculated the values of x_i^2, y_i^2 and $x_i y_i$ should not be rounded off too much as this may result in a value of $|r|$ greater than one!)

6.1 UNIT ACTIVITY

(1) Calculate the correlation coefficient for the data in (a) Table 2, (b) Table 3.

(2) The following experiment was performed at the Much more Crops Institute. Carnations were grown on standard plots inside and outside glasshouses. For each plot the amount of a certain vital, but toxic, chemical was measured together with the average number of blooms per plant.

The results were as in table 6.6 and 6.7.

Calculate the product moment correction coefficient for each set of data.

Table 6.6 Glasshouse crop

Amount of chemical in standard plot x (micrograms)	3	4	6	7	8	10	12	15	16
Average number of blooms per plant for this plot y	3.2	2.9	3.7	2.2	1.8	2.3	1.7	0.8	0.3

(You are given that $\sum v = 18.9$. $\sum v^2 = 49.33$: other sums should be calculated.)

Table 6.7 Outdoor crop

Amount of chemical in standard plot x (micrograms)	3	5	6	10	11	12	14	15
Average number of								

blooms per plant y								
for this plot	4.0	4.2	3.6	2.3	2.5	1.9	1.3.	1.1

(You are given that $\sum y = 20.9$, $\sum y^2 = 64.65$: other sums should be calculated.)

(3) (a) Pupils transferring from primary to secondary education are given two verbal reasoning tests. The score of twelve pupils (quoted relative to the mean score for each test) are in Table 6.8. Show that the product moment correlation coefficient of the test scores is approximately 0.871.

Table 6.8 Two verbal reasoning test

Pupil	A	B	C	D	E	F	G	H	J	K	L	M
Test 1	- 15	- 14	-10	- 5	- 4	- 1	2	5	10	13	17	20
Test 2	- 10	- 7	- 4	- 6	-1	-6	- 2	5	9	0	10	12

(b) Corresponding values of the variable X and Y are given in Table 6.9.

These values give a product moment correlation coefficient of 0.845

Table 6.9 Corresponding values of the variable *X* and *Y*

X	-15	- 14	-10	- 5	-1	4	10	17	20
Y	- 10	- 7	0	5	-8	10	11	10	9

Compare and contrast, without further calculation, the relationship between the test scores in (a) and the relationship between X and Y.

Some fallacies in the interpretation of *r*.

(i) Correlation should not be confused with causation. Figure for the years 1955-65 show high correlation between the number of television licences and the number of road accidents in this country but no one would suppose that one of these factors *causes* the other. Such a spurious relationship can result when each of the two variables depends on a third variable, in this case the 'standard of living', which produces a simultaneous change in both of them. It is particularly common if the variables are measured over a period of time.

(ii) Even if it appears that there is a causal relationship between two variables, the correlation coefficient does not help us to establish which variable depends on which.

(iii) As already mentioned a non-significant value of r only implies an absence of linear correlation.

(iv) A linear relationship established over a particular range of values of X and Y should not be assumed to hold outside that range by extrapolation. For example, there is a strong direct correlation between the age and weight of a baby, but to extend this linear relationship would imply that older children and adults continued to grow at this same rate!

6.5.3 Spearman's rank correlation coefficient, r_s

A manufacturer of margarine has been experimenting with different recipes and wishes to test public Reaction to them. Table 6.12 shows how two tasters rank eight varieties in order of preference. Hardly surprisingly the testers do not show the same order of preference but the manufacturer would like to know whether there is any degree of correlation between their rankings. In this case you do not have two continuous variates which are normally distributed but

Table 6.10 Public Reaction to taste

Variety	Rankings by Taster 1	Taster 2
A	5	7
B	4	3
C	2	1
D	1	2
E	6	6
F	7	8
G	3	4
H	8	5

two discrete varieties in the form of the rankings. Using these rankings we can calculate the correlation coefficient using formula (2). However, since the values of X (and Y) are the first n integers, this formula can be considerably simplified. If there are n ranks then

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} i = \frac{1}{2} n(n+1)$$

$$\sum_{i=1}^{n} x^2{}_i = \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} i^2 = \frac{1}{6} n(n+1)(2n+1)$$

Using these expressions the denominator of formula (2) becomes

$$n * \frac{1}{6} n(n+1)(2n+1) - \left[\frac{1}{2} n(n+1) \right]^2$$

$$= \frac{1}{12} n^2 (n^2 - 1) \quad \text{(after some manipulation)}$$

The numerator of (equation 1) can also be simplified by letting d_i be the difference between ranks for the ith item. Then $d_i = x_i - y_i$

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (x_i - y_i)^2 = \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i^2$$

Rearranging

$$\sum_{i=1}^{n} x_i y_i = \frac{1}{2} \left[\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} d_i^2 \right]$$

$$= \frac{1}{2} \left[2 * \frac{1}{6} n(n+1)(2n+1) - \sum_{i=1}^{n} d_i^2 \right]$$

Using this expression for $\sum x_i y_i$ and substituting for $\sum x_i$ and $\sum y_i$ as before, the numerator of equation (3.) becomes

$$= \frac{1}{2} n \left[2 * \frac{1}{6} n(n+1)(2n+1) - \sum_{i=1}^{n} d_i^2 \right] - \left\{ \frac{1}{2} n(n+1) \right\}^2$$

$$= \frac{1}{12} n^2 (n^2 - 1) - \frac{1}{2} n \sum_{i=1}^{n} d_i^2$$

Combining the expression obtained for the numerator and denominator we obtain for the rank correlation coefficient, r s,

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \qquad (4)$$

This correlation coefficient is due to Spearing (1904) and is usually called after him. Obviously when the rankings are identical the d_i s are all zero and $r_s = 1$.

Example 6.2 Calculate the Spearman rank correlation coefficient for the data in Table 6.10 (reproduced in Table 6.11).

Table 6.11 Public Reaction to taste

Variety	Taster 1, x_i	Taster 2, y_i	d_i	d_i^2
A	5	7	- 2	4
B	4	3	+1	1
C	2	1	+1	1
D	1	2	- 1	1
E	6	6	0	0
F	7	6	- 1	1
G	3	8	- 1	1
H	8	5	+3	9
			0	18

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6*18}{8(8^2-1)} = 0.79$$

As for *r*, r_s can vary between – 1 and +. The critical values of r_s at the 5% level, taking the null hypothesis that there is no correlation, are given in Table 6.12. As the number of

Table 12 Public Reaction to taste

Number of ranks	Critical value of $\|r_s\|$
5	1.00
6	0.89
7	0.79
8	0.74
9	0.70
10	0.65
20	0.45
25	0.40
50	0.28

ranks increase the cortical values appropriate for *r* can be used. Using this table the value of r_s = 0.79 obtained in the preceding example is significant at the 5% level since the critical region is $|r_s| > 0.74$ indicating there is a degree of correlation between the rankings of the two tasters.

In the example given here the variate measure by the tasters, i.e. how much they liked the margarine, was qualitative and the correlation could *only* be measured by computing r_s. Rank correlation is also used for quantitative variables when *neither* of them is normally distributed. One problem that may arise in this case is that two more values may be equal. What ranking value should they be given? Consider the following arranged in increasing order:

4.3, 5.1, 5.1, 6.3, 7.5, 8.6, 8.6, 8.6

The second and third values are 'tied' and are both given an average ranking of 2½. Similarly the 6th, 7th and 8th values are given the average ranking of 7. The complete sequence of rankings is

therefore: 1, 2½, 2½, 4, 5, 7, 7, 7. Note that the sequence has the same total as 1, 2, 3, 4, 5, 6, 7, 8.

Example 6.3

Two judges of an international beauty competition award the marks given in Table 6.13.

Calculate the Spearman rank correlation coefficient.

Do the judges agree on the order in which they place the candidates?

Table 6.13 international beauty competition

	Competitor						
	Zambia	France	Malawi	Germany	Botswana	USA	Canada
Judge A	5.8	5.5	5.9	4.9	5.6	5.6	5.0
Judge B	5.5	5.4	5.6	5.3	5.7	5.7	5.7

Judges *A's* marks, placed in descending order, are

Mark	5.9	5.9	5.8	5.6	5.5	5.0	4.9
Rank	1½	1½	3	4	5	6	7

with the rankings below.

For judge *B*, marks and their ranks are

Mark	5.8	5.7	5.7	5.7	5.5	5.4	5.3
Rank	1	3	3	3	5	6	7

The calculation of r_s is shown in Table 6.14.

Our null hypothesis is that there is no correlation between the judges' rankings. Using Table 6.13 the critical region for $n = 7$ is $|r_s| > 0.79$. The value obtained is not significant at the 5% level.

We retain H_o :

Table 6.14 Results of a beauty competition

Competitor	Judges A's rank	Judge B's rank	di	d_i^2
Zambia	3	5	-2	4
France	5	6	-1	1
USSR	1 ½	1	½	¼
Zimbabwe	7	7	0	0
Germany	1 ½	3	-1½	2¼
RSA	4	3	+ 1	1
Canada	6	3	+3	9
			0	17½

Using equation (1),

$r_s = 1 - \frac{6 \times 17\frac{1}{2}}{7(7^2 - 1)} = 0.6875$

there is no evidence that the judges agree on the order in which they place the competitors. (Strictly speaking, the formula for r_s should be modified for tied ranks since its derivation used the fact that the ranks were the first n natural numbers. However, the correction involved is small.)

6.2 UNIT ACTIVITY

(1) The Organisers of a flower completion intend to base their order of merit on the opinions of judge X and Judge Y. The marks given by the two judges are set out in Table 6.15.
(a) Carry out a rank correlation test to determine whether the two judges' opinions are consistent.
(b) Determine an order of merit of the competitors.

(2) In a village flower show the six competitors in the finals had their entries assessed by three judges. The placing's were as in Table 6.16
Calculate a rank correlation coefficient between (a) the first and second judges' placing's, and (b) the first and third judges' placing's. Comment on your results.

Table 6.15 Village flower competition

Competitor	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Judge X marks										
(max. 100)	48	50	55	51	51	47	48	46	52	50
Judge Y marks										
(max. 100)	18	19	29	22	26	14	22	11	24	17

Table 6.16

Competitor	*P*	*Q*	*R*	*S*	*T*	*U*
1st Judge	3	5	6	1	4	2
2nd Judge	4	3	5	2	6	1
3rd Judge	2	6	4	1	3	5

Table 17

x	29	81	60	88	91	91	86	99	73	42
y	74	86	63	74	70	63	41	81	29	56

Table 18 Results of *X* and *Y* for judges at a beauty contest

Competitor	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
Judge X	4	9	2	5 =	3	10	5 =	7	8

Judge Y	6	10	2	8	1	9	7	4	5

(3) Explain briefly how a scatter diagram can show whether there is direct or indirect correlation between two quantities. Give one example each of data where (a) direct, (b) indirect correlation might be expected. Table 6.17 shows some values of a variable x and the corresponding values of a second variable y. Calculate a coefficient of rank correlation between x and y and comment briefly on the result you obtain.

(4) (a) X and Y were judges at a beauty contest in which there were ten competitors. Their rankings are shown in Table 6.18.
Calculate a coefficient of rank correlation between these two sets of ranks and comment briefly on your result.

(b) Illustrate by means of two scatter diagrams rank correlation coefficients of 0 and - 1 between two variable X and Y.

6.3 UNIT ACTIVITY

(5) In an experiment the values of two variables, X and Y, were measured. Ten such experiments were performed with the following result for (X, Y):

(6, 13), (12, 2,) (9, 12), (5, 15), (2, 17), (12, 5), (8, 10), (3, 13), (11, 12), (7, 11).

Illustrate these data by means of a scatter diagram and comment briefly on the type of correlation that this shows. Calculate a coefficient of rank correlation between the ten values of X and Y.

(6) In each of the following section, which purport to be extracts from reports, the second sentence is an inference from the statement made in the first sentence. State whether the inferences are valid or invalid, and give the reasons for your decisions.

(a) The amount of fertiliser applied was varied from plot to plot and the correlation coefficient between the yield of corn from a plot and the amount of fertiliser applied was found to be 0.02. This shows that there was no relation between yield and amount of fertiliser applied.

(b) Inspection of the police reports of car accidents in the town during 1975 revealed that, when the number of accidents involving drivers of a particular age was

correlated with that age, the correlation coefficient was – 0.72. We conclude that older drivers are less likely to have an accident than younger drivers.

(c) The correlation coefficient between the sugar content *s* of the peas and the length of time *t* they have been in the greengrocer's shop was negative. It follows that s decreases with increasing *t*.

(d) The correlation coefficient of percentage of children over sixteen at school and the Gross National Product, over the years for which we have records, is 0.91, which is a high correlation. It is obvious that if many more children can be persuaded to stay on at school after sixteen then the Gross National Product will increase substantially.

(7) (a) Explain the use of (i) the product moment correlation coefficient, (ii) the coefficient of correlation by ranks.

Give two examples of the appropriate application of each.

(b) At the final judging of a 'Cow of the Year Show', two judges gave the descending orders of merit of ten cows as *EAHJBIFCGD* and *EHJAFICBDG*. Find the rank correlation coefficient between these two orders. Discuss the significance of the result.

(8) Define the product moment correlation coefficient and explain how you would interpret values of 0 and 1. Guess the value of the correlation coefficient between the variables in the following situation and interpret your estimate.

(a) The height of water and the volume of road traffic at London Bridge, if high tide is at 7 a.m. The interval between successive high tides is about 12 hours.

(b) The marks in paper I and total marks in a two-paper examination.

(9) Explain clearly what is meant by the statistical term 'correlation'.

Vegboost Industries, a small chemical firm specialising in garden fertilisers, set up an experiment to study the relationship between a new fertiliser compound and the yield from tomato plants. Eight similar plants were selected and treated regularly throughout their life with *x* grams of fertilizer diluted in a standard volume of water. The yield *y*, in kilograms, of good tomatoes was measure for each plant. Table 6.19 summarises the results.

(a) Calculate the product moment correlation coefficient for these data.

(b) Calculate Spearman's rank correlation coefficient for these data.

(c) Is there any evidence of a relationship between these variables? Justify your answer. (No formal test is required.)

Table 6.19 Fertiliser compound and the yield from tomato plants

Plant	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Amount of fertilizer *x* (g)	1.2	1.8	3.1	4.9	5.7	7.1	8.6	9.8
Yield *y* (kg)	4.5	5.9	7.0	7.8	7.2	6.8	4.5	2.7

(10) Explain how you would calculate the product moment correlation coefficient from a sample of n pairs of values x_i, y_i (i = 1, 2, . . . , n).

The number of eggs laid, x, by a certain species of bird is either 1 or 2, and this is related to the age of the bird, y, which is also 1 or 2. The frequencies of the four possible combinations of values x, y are recorded for a total of n birds, as shown in Table 17.28. Show that the product moment correlation coefficient between x and y is $(ad - bc)/ \{(a + b)(c + d)(a + c)(b + d)\}^{1/2}$.

Table 6.20 Number of eggs laid, *x*, by a certain species of bird

	Numbers of eggs laid (x)		
	1	2	Total
Age y (years) 1	*a*	*b*	*a* + *b*
2	*c*	*d*	*c* + *d*

Total	$a + c$	$b + d$	n

(11) What is meant by the statistical terms 'independent' and uncorrelated', and how are they related?

At the University of Zambia, students for the degree in chemical engineering take two examinations. They have a part I examination at the end of the second year of their studies, and a part II examination at the end of the third year. Degree classification

Table 6.21 Marks of a Chemical Engineering Examination

		Part I marks					
		40-49	50-59	60-69	70-79	80-89	90-99
Part II marks	40-49	3	5	4	0	0	0
	50-59	3	6	6	2	0	0
	60-69	0	5	9	5	2	0
	70-79	0	0	5	10	8	1
	80-89	0	0	0	5	6	5
	90-99	0	0	0	2	4	4

depends upon an average of those marks. Table 21 gives the part I and part II marks for a particular set of 100 students. Use the coding method to calculate the product moment correlation coefficient for these data. (Assumed means of 64.5 for part I marks and 74.5 for part II marks should be used.)

6.0 UNIT SUMMARY

The following formulas are very useful in the calculation of correlation coefficient r.

1. $$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}$$

2. $$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[\left\{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right\}\left\{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right\}\right]}}$$

This rather tedious calculation can sometimes be simplified by using coded values. If we use coded values u_i and v_i for x_i and y_i such that

3. $$r = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\left[\sum_{i=1}^{n}(u_i - \bar{u})^2 \sum_{i=1}^{n}(v_i - \bar{v})\right]}}$$

4. $$r = \frac{n\sum_{i=1}^{n} u_i v_i - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i}{\sqrt{\left[\left\{n\sum_{i=1}^{n} u_i^2 - \left(\sum_{i=1}^{n} u_i\right)^2\right\}\left\{n\sum_{i=1}^{n} v_i^2 - \left(\sum_{i=1}^{n} v_i\right)^2\right\}\right]}}$$

5. Spearman's rank correlation coefficient, r_s

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Correlation coefficient is used to find the relationship between to variable X and Y.

UNIT 7 THE ANALYSIS OF VARIANCE

7.1 Unit Introduction

Welcome to Unit 6 in which you will learn how to conduct analysis of variance. In this Unit you are going to learn how to design experiments. You will learn how to allocate units in block designing and carry out the analysis for each design.

7.2 Unit Aim

The aim of this Unit is teach you on how to design an experiment and analysis.

7.3 Unit Objectives

By the end of the unit you should be able to:

- Perform **analysis of variance** by hand
- Appropriately interpret results of **analysis of variance** tests.
- Distinguish between one and two factor **analysis of variance** tests.
- Identify the appropriate hypothesis testing procedure based on type of outcome variable and number of samples

Terminology

Total SS Total sum of squares

SST Sum of squares of Treatments

SSC – Sum of squares of columns

SSR – Sum of squares of rows

SSE – sum of squares of errors

MSC- Mean squares of columns

MSR – means squares of errors

MST-Mean squares of Treatments

F- Distribution

7.4 Unit Time required

You need 20 hours for this unit

7.5 Unit Topics

7.5.1 The Analysis of Variance

The methodology for the analysis of experiments involving several independent variables can best be explained in terms of the linear probabilistic models. Although elementary and unified, this approach is not susceptible to the condensation necessary for inclusion in an elementary text. Instead, we shall attempt an intuitive discussion using a procedure known as the *analysis of variance*. Actually, the two approaches are connected and the analysis of variance can easily be explained in a general way in terms of the linear model.

As the name implies, the analysis-of-variance procedure attempts to analyze the variation of a response and to assign portions of this variation to each of a set of independent variables. The reasoning is that response variables vary only because of variation in a set of unknown independent variables. Since the experimenter will rarely, if ever, include all the variables affecting the response in his experiment, random variation in the response is observed even though all independent variables considered are held constant. The objective of the analysis of variance is to locate important independent variables in a study and to determine how they interact and affect the response.

The rationale underlying the analysis of variance can be indicated best with a symbolic discussion. The actual analysis of variance-that is, "how to do it" it can be illustrated with an example.

You will recall that the variability of a set of n measurements is proportional to the sum of squares of deviations, $\sum_{i=1}^{n}(y-\bar{y})^2$, and that this quantity is used to calculate the sample variance. The analysis of variance partitions the sum of squares of deviations, called the total sum of squares of deviations, into parts, each of which is attributed to one of the independent variables in the experiment, plus a remainder that is associated in Figure 1 for three independent variables. If a multivariate linear model were written for the response, as suggested

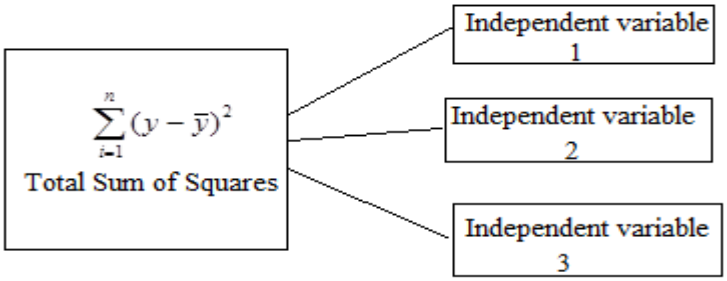


Figure 7. 1 The Partitioning of the total Sum of Squares of Deviations

The portion of the total sum of squares of deviations assigned to error would be labelled SSE.

For the cases that we consider, and when the independent variables are unrelated to the response, it can be shown that each of the pieces of the total sum of squares of deviations, divided by an appropriate constant, provides an independent and unbiased estimator of σ^2, the variance of the experimental error. When a variable is highly related to the response, its portion (called the "sum of squares" for that variable) will be inflated. This condition can be detected by comparing the estimate of σ^2 for a particular independent variable with that obtained from SSE using a F test. If the estimate for the independent variable is significantly larger, the F test will reject a hypothesis of "no effect for the independent variable: and produce evidence to indicate a relation to the response.

The mechanism involve in an analysis of variance can best be illustrated by considering a familiar example, say the comparison of means for the unpaired experiment for the special case where $n_1 = n_2$. This experiment, formerly analysed by the use of Student's t, will now be approached from another point of view. The total variation of the response measurements about their mean for the two samples is

$$Total\,SS = \sum_{i=1}^{n}\sum_{j=1}^{n}(y_{i,j}-\bar{y})^2$$

Where $y_{i,j}$ denotes the jth observation in the ith sample and y is where y_{ij} denotes the jth observation in the ith sample and $\bar{y}$ is the mean of all $2n_1 = n$ observations. This quantity can be partitioned into two parts. That is,

$$Total\,SS = \sum_{i=1}^{2}\sum_{j=1}^{n_1}(y_{i,j} - \bar{y})^2$$

$$= n_1\sum_{i=1}^{2}(y_i - \bar{y})^2 + \sum_{i=1}^{2}\sum_{j=1}^{n_1}(y_{ij} - \bar{y}_i)^2$$

(proof deferred), where y_1 is the average of the observations in the ith sample i = 1, 2, and the second quantity on the right-hand side of the equality is simply the pooled sum of squares of deviations used to calculate s^2 . (Recall that we assume equal population variances for this statistical method). We will denote this quantity as SSE.

Note that we have partitioned the "total sum of squares" of deviations into two parts. One part, SSE, can be used to obtain a pooled estimator of σ^2 . the other part,

$$n_1\sum_{i=1}^{2}(\bar{y}_i - \bar{y})^2 = \frac{n_1}{2}(\bar{y}_1 - \bar{y}_2)^2 = SST$$

which we will call the sum of squares for treatments, will increase as $(\bar{y}_1 - \bar{y}_2)$ increases. Hence, the larger SST, the weight of evidence to indicate a difference in $(\mu_1 - \mu_2)$

How large is "large?" When will SST be large enough to indicate a real difference between μ_1 and μ_2 ?

$$s^2 = MSE = \frac{SSE}{n_1 + n_2 - 2} = \frac{SSE}{2n_1 - 2}, \text{ (since } n_1 = n_2)$$

provides an unbiased estimator of σ^2 . (MSE denotes mean square for error.) Also, when the null hypothesis is true (that is, $\mu_1 = \mu_2$), SST divided by an appropriate number of degrees of freedom yields a second unbiased estimator for …^{2}, which we will denote as MST. For this example, the number of degrees of freedom for MST is equal to 1. When the null hypothesis is true, MSE and

MST (the mean square fir treatments) will estimate the same quantity and should be "roughly" for the same magnitude. When the null hypothesis is false and $\mu_1 \ldots \mu_2$, MST will tend to be larger than MSE. You can show that E(MST) $= \sigma^2$. when $\mu_1 = \mu_2$.

The preceding discussion, together with a review of the variance ration, suggests the use of $\frac{MST}{MSE}$ as a test statistic to test the hypothesis, $\mu_1 = \mu_2$, against the alternative, $\mu_1 \neq \mu_2$. Indeed, when both populations are normally distributed, it can be shown that MST and MSE are independent and that $F = \frac{MST}{MSE}$ follows the F probability distribution. Disagreement with the null hypothesis is indicated by a large value of F, and hence the rejection region for a given α will be $F \geq F_\alpha$

The analysis-of-variance test results in a one-tailed F test, the degrees of freedom for F will be those associated with MST and MSE, which we will denote as v_1 and v_2, respectively. Although we have not indicated, in general, how one determines v_1 and v_2, $v_1 = 1$ and $v_2 = (2n_1 - 2)$ for the two sample experiment described.

Proof of the independence of MST and MST will not be considered here because it is a special case of a; more general and more complicated theorem that applies to many different analyses of variance.

Example 7.1: The coded values for the measure of elasticity in plastic, prepared by two different processes, for samples of six drawn randomly from each of the two processes, are as follows:

Table 7.1 Measure of elasticity in plastic

A	B
6.1	9.1
7.1	8.2
7.8	8.6
6.9	6.9
7.6	7.5
8.2	7.9

Do the data present sufficient evidence to indicate a difference in mean elasticity for the two processes?

Solution: Although the student's t could be used as the test statistic for this example, we shall use our analysis-of-variance F test since it is more general and can be used to compare more than two means.

The three desired sums of squares of deviations are

$$\text{Total SS} = \sum_{i=1}^{2}\sum_{j=1}^{6}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{2}\sum_{j=1}^{6}y_{ij}^2 - \frac{(\sum_{i=1}^{2}\sum_{j=1}^{6}y_{ij})^2}{12}$$

$$= 711.35 - \frac{(91.9)^2}{12} = 7.5492$$

$$SST = n_1\sum_{i=1}^{2}(\bar{y}_1-\bar{y})^2 = 6\sum(\bar{y}_1-\bar{y})^2 = 1.6875$$

$$SSE = \sum_{i=1}^{2}\sum_{j=1}^{6}(y_{ij}-\bar{y}_i)^2 = 5.8617$$

(You may verify that SSE is the pooled sum of squares of the deviations for the two samples. Also, note that Total SS = SST + SSE.) The mean squares for treatment and error are, respectively,

$$\text{MST} = \frac{SST}{1} = 1.6875$$

$$\text{MSE} = \frac{SSE}{2n_1 - 2} = \frac{5.8617}{10} = 0.58617$$

To test the hypothesis, $\mu_1 = \mu_2$, we compute the test statistic

$$\text{F} = \frac{MST}{MSE} = \frac{1.6875}{0.58617}$$

The critical value of the F statistic for $\alpha = .05$ is 4.96. Although the mean square for treatments is almost three times as large as the mean square for error, it is not large enough to reject the null hypothesis. Consequently, there is not sufficient evidence to indicate a difference between μ_1 and μ_2.

As noted, the purpose of the preceding example was to illustrate the computations involved in a simple analysis of variance. The F test for comparing two means is equivalent to a student's test because an F statistic with one degree of freedom in the numerator is equal to t^2. You can easily verify that the square of $t_{.025} = 2.228$ (used for the two-tailed test with $\alpha = .05$ and $v = 10$ degrees of freedom) is equal to $F_{.05} = 4.96$. Similarly, the value of the t statistic for Example7.1 would equal the square root of the computed F = 2.88.

7.5.2 A Comparison of More Than Two Means

An analysis of variance to detect a difference in a set of more than two population means is a simple generalization of the analysis of variance of Section 2.5.2. The random selection of independent samples from *p* populations is known as a *completely randomized* experimental *design.*

Assume that independent random samples have been drawn from *p* normal populations with means $\mu_1, \mu_2, \ldots \mu_p$, respectively, and variance σ^2. All populations are assumed to possess equal variances. To be completely general, we shall allow the sample size to be unequal and let n_i, i = 1,

2, …, p, be the number in the sample drawn from the ith population. The total number of observations in the experiment will be $n = n_1 + \ldots + n_p$.

Let y_{ij} denote the measured response on the jth experimental unit in the ith sample and let T_i and $\bar{T}_i$ represent the total and mean, respectively for the observations in the ith sample. (The modification in the symbols for sample totals and averages will simplify the computing formulae for the sums of squares.) Then, as in the analysis of variance involving two means,

Total SS = SST + SSE

$$Total\,SS = \sum_{i=1}^{n}\sum_{j=1}^{n_1}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{p}\sum_{i=1}^{n_1} y_{ij}^2 - CM$$

$$CM = \frac{(total\,of\,all\,observations)^2}{n}$$

$$= \frac{\left(\sum_{i=1}^{p}\sum_{j=1}^{n_i} y_{ij}\right)^2}{n} = n\bar{y}^2$$

(the symbol CM denotes "correction for the mean"),

$$SST = \sum_{i=1}^{p} n_i(\bar{T}_i - \bar{y})^2 = \sum_{i=1}^{p}\frac{T_i^2}{n_i} - CM$$

SSE = Total SS - SST

Although the easy way to compute SSE is by subtraction as shown above, it is interesting to note that SSE is the pooled sum of squares for all p samples and is equal to

$$SSE = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{T}_i)^2$$

The unbiased estimator of σ^2 based on $(n_1 + n_2 + \ldots + n_p - p)$ degrees of freedom is

$$s^2 = MSE = \frac{MSE}{n_1 + n_2 + \ldots + n_p - p}$$

The mean square for treatments will possess (p-1) degrees of freedom, that is, one less than the number of means, and $MST = \frac{SST}{p-1}$

To test the null hypothesis, $H_o : \mu_1 = \mu_2 = ... = \mu_p$ against the alternative that at least one of the equalities does not hold, MST is compared with MSE using the F statistic based upon $v_1 = (p - 1)$ and $v_2 = \left(\sum_{i=1}^{p} n_i - p\right)(n - p)$ degrees of freedom. The null hypothesis will be rejected if

$$F = \frac{MST}{MSE} > F_\alpha$$

where F_α is the critical value of F for probability of a type I error, α.

Intuitively, the greater the difference between the observed treatment means, $\overline{T}_1, \overline{T}_2, ..., \overline{T}_p$, the greater the will be the evidence to indicate a difference between their corresponding

Population means. It can be seen from the above expression that SST = 0 when all the observed treatment means are identical, because then $\overline{T}_1 = \overline{T}_2 = ... = \overline{T}_p = \bar{y}$ and the deviations appearing in SST, $(\overline{T}_i - \bar{y})$, i= 1, 2, ……p, will equal zero. As the treatment means get farther apart, the deviations, $(\overline{T}_i - \bar{y})$, will increase in absolute value and SST will increase in magnitude. Consequently, the larger the value of SST will increase in magnitude. Consequently, the larger the value of SST, the greater will be the weight of evidence favouring a rejection of the null hypothesis. This same line of reasoning will apply to the F tests employed in the analyses of variances for all designed experiments.

The assumptions underlying the analysis-of-variance F tests should receive particular attention. The samples are assumed to have been randomly selected from the p populations in an independent manner. The populations are assumed to be normally distributed with equal variances, σ^2, and means, $\mu_1, \mu_2, ..., \mu_p$. Moderate departures from these assumptions will not seriously affect the properties of the test. This is particularly true of the normality assumption.

Example 7.2: Four groups of students were subjected to different teaching techniques tested at the end of a specified period of time. As a result of dropouts from the experimental groups (due to sickness, transfer, and so on), the number of students varied from group to group. ; do the data shown present sufficient evidence to indicate a difference in the mean achievement for the four teaching techniques?

Table 7.2 Different teaching techniques

	1	2	3	4
	65	75	59	94
	87	69	78	89
	73	83	67	80
	79	81	62	88
	81	72	83	
	69	79	76	
		90		
T_i	454	549	425	351
n_i	6	7	6	4
$\bar{T}_i$	75.67	78.43	70.83	87.75

Solution:

$$CM = \frac{\left(\sum_{i=1}^{4}\sum_{j=1}^{n_i} y_{ij}\right)^2}{n} = \frac{(1779)^2}{23} = 137{,}601.8$$

$$Total\,SS = \sum_{i=1}^{4}\sum_{j=1}^{n_i} y_{ij}^2 - CM$$

$$= 139{,}511 - 137{,}601.8 = 1909.2$$

$$SST = \sum_{i=1}^{4}\frac{T_i^2}{n_i} - CM$$

$$= 138{,}214.4 - 137{,}601.8 = 712.6$$

SSE = Total SS –SST =1196.6

The mean squares for treatment and error are

$$\text{MST} = \frac{SST}{p-1} = \frac{712.6}{3} = 237.5$$

$$\text{MSE} = \frac{SSE}{n_1 + n_2 + \ldots + n_p - p} = \frac{SSE}{n-p} = \frac{1196.6}{19} = 63.0$$

The test statistic for testing the hypothesis, $\mu_1 = \mu_2 = \mu_3 = \mu_4$, is

$$F = \frac{MST}{MSE} = \frac{237.5}{63.0} = 3.77$$

Where $\nu_1 = p - 1 = 3$, $\nu_2 = \sum_{i=1}^{p} n_i - 4 = 19$

The critical value of F for α = .05 is $F_{.05} = 3.13$. Since the computed value of F exceeds $F_{.05}$, you reject the null hypothesis and conclude that the evidence is sufficient to indicate a difference in mean achievement for the four teaching procedures.

You may feel that the above conclusion could have been made on the basis of visual observation of the treatment means. It is not difficult to construct a set of data that will lead the "visual" decision maker to erroneous results.

7.5.3 Proof of Additivity of the Sums of Squares and E(MST) for a Completely Randomized Design

The proof that Total SS = SST + SSE

For the completely randomized design is presented in this section for the benefit of the interested reader. It may be omitted without loss of continuity.

The proof utilizes elementary results on summations that appear in the exercises for unit 1, and the device of adding and subtracting T_I within the expression for the Total SS. Thus

$$\text{Total SS} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{T}_i + \bar{T}_i - \bar{y})^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{n_i}[(y_{ij} - \bar{T}_i) + (\bar{T}_i - \bar{y})]^2$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{n_i}[(y_{ij} - \bar{T}_i)^2 + 2(y_{ij} - \bar{T}_i)(\bar{T}_i - \bar{y}) + (\bar{T}_i - \bar{y})^2]$$

Summing first over j, we obtain

$$\text{Total SS} = \sum_{i=1}^{p}\left[\sum_{j=1}^{n_i}(y_{ij} - \bar{T}_i)^2 + 2(\bar{T}_i - \bar{y})\sum_{i=1}^{n_i}(y_{ij} - \bar{T}_i) + n_i(\bar{T}_i - \bar{y})^2\right]$$

where

$$\sum_{i=1}^{n_i}(y_{ij} - \bar{T}_{ij}) = T_i - n_i\bar{T}_i = T_i - T_i = 0$$

Consequently, the middle term in the expression for the Total SS is equal to zero.

Then, summing over i, we obtain

$$\text{Total SS} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_{ij} - \bar{T}_i)^2 + \sum_{i=1}^{p} n_i(\bar{T}_i - \bar{y})^2,$$

= SSE + SST.

The first expression is SSE, the pooled sum of squares of deviations of the sample measurements about their respective means. The second is the formula for SST.

Proof of the additive of the analysis-of-variance sums of squares for other experimental designs can be obtained in a similar manner. The procedure is tedious.

You will now derive the expected value of MST for a completely randomized design. Assume that Y_{ij}, the jth sampled value from the ith population, has $E(Y_{ij}) = 1, \ldots, p, j = 1, \ldots, n_i$. since T_i is the average of n_i independent random variables, Y_{ij}, $j = 1, \ldots, n_i$, it follows that $E(T_i) =$, μ_i and $V(T_i) = \frac{\sigma^2}{n_i}$. Similarly, Y is given by

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij} ,$$

and hence

$$E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i}\mu_i = \frac{1}{n}\sum_{i=1}^{p} n_i\mu_i$$

We will denote $E(\bar{Y})$ by $\bar{\mu}$

$$V(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^{p}\sum_{j=1}^{n_i} V(Y_{ij}) = \frac{\sigma^2}{n}$$

Then

$$E(MST) = \frac{1}{p-1}E\left[\sum_{i=1}^{p} n_i(\bar{T}_i - \bar{Y})^2\right]$$

$$= \frac{1}{p-1}E\left[\sum_{i=1}^{p} n_i(\bar{T}^2{}_i - 2\bar{T}_i\bar{Y} + \bar{Y}^2)\right]$$

$$= \frac{1}{p-1}E\left[\sum_{i=1}^{p} n_i\bar{T}_i^2 - n\bar{Y}^2\right]$$

$$= \frac{1}{p-1}\left[\sum_{i=1}^{p} n_i E(\bar{T}_i^2) - nE(Y^2)\right]$$

$$= \frac{1}{p-1}\left[\sum_{i=1}^{n} n_i\left(\frac{\sigma^2}{n_i} + \mu_i^2\right) - n\left(\frac{\sigma^2}{n} + \bar{\mu}^2\right)\right]$$

On noting that, for random variable U, $E(U^2) = V(U) - [E(U)]^2$. It then follows that

$$E(MST) = \frac{1}{p-1}\left(\sum_{i=1}^{p}\sigma^2 + \sum_{i=1}^{p} n_i\mu_i^2 - \sigma^2 - n\bar{\mu}_i^2\right)$$

$$= \frac{1}{p-1}\left[\sigma^2(p-1) + \sum_{i=1}^{p} n_i\mu_i^2 - n\mu_i^2\right]$$

$$= \sigma^2 + \frac{1}{p-1}\sum_{i=1} n_i(\mu_i - \bar{\mu})^2$$

Under $H_o : \mu_1 = \mu_2 = ... = \mu_p$, $\mu_i = \bar{\mu}$, $i = 1,2,..,p$, and hence E(MST) = σ^2 Thus MST/MSE is a ratio of unbiased estimators of σ^2 when H_o is true.

7.5.4 An Analysis-of-Variance Table for a Completely Randomized Design

The calculations of the analysis of variance are usually displayed in an analysis-of-variance (ANOVA or AOV) table. The table shown in Table 7.1.

Table 7.3 ANOVA table for a Complete Randomized Design

Source	d.f.	SS	MS	F
Treatments	p - 1	SST	$MST = \frac{SST}{P-1}$	$\frac{MST}{MSE}$
Error	n – p	SSE	$MSE = \frac{SSE}{n-p}$	
Total	n – 1	$\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2$		

The first column shows the source of each sum of squares of deviations; the second column gives the respective degrees of freedom; the third and fourth columns give the corresponding sums of squares and mean squares, respectively. A calculated value of F, comparing MST and MSE, is usually shown in the fifth column. Note that the degrees of freedom and sums of squares add to their respective totals.

The ANOVA table for Example 7.2, shown in Table 7.3, gives a compact presentation of the appropriate computed quantities for the analysis of variance.

Table 7.4 ANOVA Table for Example 7.2

Source	d.f.	SS	MS	F
Treatments	3	712.6	237.5	3.77
Error	19	1196.6	63.0	
Total	22	1909.2		

7.5.5 Estimation for the Completely Randomized Design

Confidence intervals for a single treatment mean and the difference between a pair of treatment means. The confidence interval for the mean of treatment i or the difference between treatments i and j are, respectively,

$$\bar{T}_i \pm \frac{t_{\alpha/2} S}{\sqrt{n_i}}$$

And $(\bar{T}_i - \bar{T}_j) \pm t_{\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$,

Where $s = \sqrt{s^2} = \sqrt{MSE} = \sqrt{\frac{SSE}{n_1 + n_2 + ... + n_p - p}}$

And $t_{\alpha/2}$ is based upon (n – p) degrees of freedom

Note that the confidence intervals stated above are appropriate for single treatment means or a comparison of a pair of means selected prior to observation of the data. The stated confidence coefficients are based on random sampling. If one were to look at the data and always compare the largest and smallest sample means, the assumption of randomness would be disturbed. Certainly the difference between the largest and smallest sample means is expected to be larger than for a pair selected at random.

Example 7.3: Find a 95 percent confidence interval for the mean score for teaching technique 1,

Solution: The 95 percent confidence interval for the mean score is

$$\bar{T}_1 \pm \frac{t_{0.025} s}{\sqrt{6}}$$

or

$$75.67 \pm \frac{(2.093)(7.94)}{\sqrt{6}}$$

$$75.67 \pm 6.78$$

Example 7.**4**: Find a 95 percent confidence interval for the difference in mean score for teaching techniques 1 and 4, Example 2

Solution: The 95 percent confidence interval is

$$(\bar{T}_1 - \bar{T}_4) \pm (2.093)(7.94)\sqrt{1/6 + 1/4}$$

-12.08 ± 10.74

Hence the 95 percent confidence interval for $(\mu_1 - \mu_4)$ is -22.82 to - 1.34. This suggests that

$\mu_4 > \mu_1$

7.5.6 The Analysis of Variance for a Randomized Block Design

The method for constructing a randomized block design was presented in Section 7.5.5.

The randomized block design implies the presence of two qualitative independent variables, "blocks" and treatments." Consequently, the total sum of squares of deviations of the response measurements about their mean may be partitioned into three parts, the sums of squares for blocks, treatments, and error.

Denote the total and average of all observations in block I as B_i as and $\overline{B}_i$, respectively. Similarly, let T_j and $\overline{T}_j$ represent the total and the average for all observations receiving treatment j. Then, for a randomized block design involving b blocks and p treatments,

Total SS = SSB + SST + SSE

$$= \sum_{i=1}^{b}\sum_{j=1}^{p}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{b}\sum_{j=1}^{p} y_{ij}^2 - CM$$

$$SSB = p\sum_{i=1}^{b}(\overline{B}_i - \bar{y})^2 = \frac{\sum_{i=1}^{b} B_i^2}{p} - CM$$

$$SST = b\sum_{j=1}^{p}(\overline{T}_j - \bar{y})^2 = \frac{\sum_{j=1}^{p} T_j^2}{b} - CM$$

SSE = Total SS –SSB -SST

In the formulas, $\bar{y} = (average of all n = bp observations)$

$$= \frac{\sum_{i=1}^{b}\sum_{j=1}^{p} y_{ij}}{n}$$

$$CM = \frac{(total\ of\ all\ observations)^2}{n} \text{ and}$$

$$= \frac{\left(\sum_{i=1}^{b}\sum_{j=1}^{p} y_{ij}\right)^2}{n}$$

The analysis of variance for the randomized block design is presented in table 7.4. The degrees of freedom associated with each sum of squares is

Table 7.5 ANOVA Table for a Randomized Block Design

Source	d.f.	SS	MS
Blocks	b - 1	SSB	$\frac{SSB}{b-1}$
Treatments	p - 1	SST	$\frac{SST}{p-1}$
Error	n - b – p +1	SSE	MSE
Total	n - 1	Total SS	

shown in the second column. Mean squares are calculated by dividing the sums of squares by their respective degrees of freedom.

To test the null hypothesis "there is no difference in treatment means," you use the *F* statistic,

$$F = \frac{MST}{MSE}$$

and reject if $F > F_{\alpha}$ based on $\nu_1 = b-1$ and $\nu_2 = n-b-p+1$ degrees of freedom.

Blocking not only reduces the experimental error, it also provides an opportunity to see whether evidence exists to indicate a difference in the mean response for blocks. Under the null hypothesis that there is no difference in mean response for blocks, MSB provides an unbiased estimator for σ^2 based on (b-1) degrees of freedom. Where real differences exist among block means, MSB will tend to be inflated in comparison with MSE and

$$F = \frac{MSB}{MSE}$$

provides a test statistic. As in the test for treatments, the rejection region for the rest will be

$$F > F_{\alpha}$$

Based on $\nu_1 = b - 1$ and $\nu_2 = n - b - p + 1$ degrees of freedom.

Example 7.**5** A stimulus-response experiment involving three treatments was laid out in a randomized block design using four subjects. The response was the

Table 7.6 Allocating of units in blocks

	Subjects			
	1	2	3	4
	① 1.7	③ 2.1	① .1	② 2.2
	③ 2.3	① 1.5	② 2.3	① 0.6
	② 3.4	② 2.6	③ 0.8	③ 1.6

length of time to reaction measured in seconds. The data, arranged in blocks, are shown in Table 7.6. The treatment number is circled and shown above each observation. Do the data present sufficient evidence to indicate a difference in the mean response for stimuli (treatments)? Subjects?

Solution: The sums of squares for the analysis of variance are shown individually below and jointly in Table 7.5 Thus

$$CM = \frac{(total)^2}{n} = \frac{(21.2)^2}{12} = 37.45$$

$$\text{Total SS} = \sum_{i=1}^{4}\sum_{j=1}^{3}(y_{ij} - \bar{y})^2 = i = \sum_{i=1}^{3}\sum_{j=1}^{3} y_{ij}^2 - CM$$

$$= 46.86 - 37.45 = 9.41,$$

$$SSB = \frac{\sum_{i=1}^{4} B_i^2}{3} - CM = 40.93 - 37.45 = 3.48$$

$$SSE = \frac{\sum_{j=1}^{3} T_j^2}{4} - CM = 42.93 - 37.45 = 5.48$$

SSE = Total SS – SSB – SST

= 9.41 – 3.48 – 5.48 = 0.45

Table 7.7 ANOVA Table

Source	d.f.	SS	MS	F
Blocks	3	3.48	1.160	15.47
Treatments	2	5.48	2.740	36.53
Error	6	.45	.075	
Total	11	9.41		

You use the ratio of mean-square treatment to mean-square error to test a hypothesis of no difference in the expected response for treatments. Thus

$$F = \frac{MST}{MSE} = \frac{2.74}{0.075} = 36.53$$

The critical value of the F statistic (a = .05) for $v_1 = 2$ and $v_2 = 6$ degree of freedom is $F_{.05} = 5.14$. Since the computed value F exceeds the critical value, there is sufficient evidence to reject the null hypothesis and conclude that real difference do exist among the expected responses for the three stimuli. Table 7.7 shows the ANOVA table for RBD.

A similar test may be conducted for the null hypothesis that no difference exists in the mean response for subjects. Rejection of this hypothesis would imply that subject-to-subject variability does exist, and that blocking is desirable. The computed value of F based on $v_1 = 3$ and $v_2 = 6$ degrees of freedom is

$$F = \frac{MST}{MSE} = \frac{1.16}{0.075} = 15.47$$

Since this value of F exceeds the corresponding tabulated critical value, $F_{.05} = 4.76$, we reject the null hypothesis and conclude that a real difference exists in the expected response for the group of subjects.

7.5.7 Estimation for the Randomized Block Design

The confidence interval for the difference between a pair of means is exactly the same as for the completely randomized design, Section 7.5.6. It is

$$(T_i - T_j) \pm t_{\alpha/2} s \sqrt{\frac{2}{b}}$$

where $n_i = n_j = b$, the number of observations contained in a treatment mean, and s = MSE. The difference between the confidence intervals for the completely randomised block design and the randomized block designs is that s, appearing in the expression above, will tend to be smaller than for the completely randomized design.

Similarly, one may construct a $(1-\alpha)$ confidence interval for the difference between a pair of block means.; each block contains p observations corresponding to the p treatments. Therefore, the confidence interval is

$$(B_i - B_j) \pm t_{\alpha/2} s \sqrt{\frac{2}{p}}$$

Example 7.6 Construct a 95 percent confidence interval for the difference between treatments 1 and 2, Example 7.5.

Solution: The confidence interval for the difference in mean response for a pair of treatments is

$$(T_i - T_j) \pm t_{0.025} S \sqrt{\frac{2}{b}}$$

where for our example $t_{.025}$ is based upon six degrees of freedom. For treatments 1 and 2 we have

$$(0.98 - 2.63) \pm (2.447)(0.27)\sqrt{\frac{2}{4}}$$

$$-1.65 \pm 0.47$$

7.5.8 The Analysis of Variance for a Latin-Square Design

The method for constructing a Latin-square design for comparing p treatments is presented in Section 7.5.5. The purpose of the design is to remove unwanted variation as might occur in the mechanized application of icing to cakes on a conveyor belt. Variation in the thickness of icing could occur across the belt due to the variation in pressure at the applicator nozzles. Similarly the thickness of icing could vary somewhat along the length of the belt due to variations in the consistency of the icing supplied to the machine. Now suppose that we wish to compare three different types of cake mixes, A, B, and C, that result in different porosities which affect absorption of the icing into the cakes. Then the thickness of the resulting icing, y, could be compared for the three treatments (mixes) by employing a 3 x 3 Latin-square design. Each mix would appear in each column (across the conveyor belt) and in each row as one proceeds down the belt. The design configuration is shown in Figure 7.2.

Table 7.8 Latin square Design

			columns	
		(Position across the belt		
		1	2	3
	Rows	B	A	C
(Positions down the		C	B	A
belts)		A	C	B
		Conveyor Belt		

Figure 7.2 A 3 x 3 Latin-Square Design

The three independent variables in a Latin-square design are "rows", "columns," and treatments. All are qualitative variables, although the treatments could be levels of a single quantitative factor or combinations of levels for two or more factors. Thus the total variation in an analysis of variance can be partitioned into four parts, one each corresponding to the variation in rows, columns, treatments, and experimental error. The analysis-of –variance table for a p x p Latin-square design is shown in Table 7.6. As for previous designs, the four sums of squares add to the total sum of squares of deviations. Table 7.9 gives an ANOVA table for a Latin square design.

Table 7.9 ANOVA Table for a Latin-Square Design

Source	d.f.	SS	MS
Rows	$p-1$	SSR	$\frac{SSR}{P-1}$
Columns	$p-1$	SSC	$\frac{SSC}{P-1}$
Treatments	$p-1$	SST	$\frac{SST}{P-1}$
Error	$n-3p+2$	SSE	MSE
Total	$n-1$	Total SS	

The formulas for computing the total sums of squares, SSR, SSC, and SST, are identical to the corresponding formulas given for the randomized block design. Let y_{ij} denote an observation in row i and column j. Then

$$Total\ SS = \sum_{i=1}^{p}\sum_{j=1}^{p}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} y_{ij}^2 - CM$$

Where $CM = \dfrac{(total\ of\ all\ observations)^2}{n}$

$$= \frac{\left(\sum_{i=1}^{p}\sum_{j=1}^{p} y_{ij}\right)^2}{n}$$

Similarly, let R_i and $\bar{R}_i$, $\bar{C}_i$ C_j, and T_k and $\bar{T}_i$ represent the total and average for all observations in row i, column j, and treatment k, respectively. Then the sums of squares for rows, columns, and treatments are

$$SSR = P\sum_{i=1}^{p}(\bar{R}_i - \bar{y})^2 = \frac{\sum_{i=1}^{p} R_i^2}{p} - CM$$

$$SSC = P\sum_{i=1}^{p}(\bar{C}_j - \bar{y})^2 = \frac{\sum_{i=1}^{p} C_j^2}{p} - CM \quad \text{and}$$

$$SST = P\sum_{i=1}^{p}(\bar{T}_k - \bar{y})^2 = \frac{\sum_{i=1}^{p} T_{ik}^2}{p} - CM$$

The sums of squares for error, SSE, can be obtained by subtraction. Thus

Total SS = SSR + SST + SSE

Hence

SSE = Total SS – SSR – SSC - SST

The mean squares corresponding to rows, columns, and treatments can be obtained by dividing the respective sum of squares by (p – 1). Thus

$$F = \frac{SST}{P-1}$$

The hypothesis “no difference in mean response for treatments” is tested using the F statistic,

$$F = \frac{SST}{MSE}$$

Similarly, the F statistic can be used to test a hypothesis of no difference between rows (or columns) by using the ratio of MSR (or MSC) to MSE. We will illustrate with an example.

Example 7.7 An experiment was conducted to investigate the difference in mean time to assemble four different electronic devices, 1, 2, 3, and 4. Two sources of unwanted variation affect the response – the variation between people and the effect of fatigue if a person assembles a series of the devices over time. Consequently, four assemblers were selected and each assembled all four of the devices in the Latin-square design of Table 7. (The observed responses, in

Table 7. 10 Latin Square Design

Rows (Position in Assembly sequence)	Columns (Assemblers) 1	2	3	4	Total
1	③ 44	① 41	② 30	④ 40	155
2	② 41	③ 42	④ 49	① 49	181
3	① 59	④ 41	③ 59	② 34	193
4	④ 58	② 37	① 53	③ 59	207
Total	202	161	191	182	736

minutes, are shown in the cells. Circled numbers in the design above indicates the treatments employed.) Do the data provide sufficient evidence to indicate a difference in mean time to assemble the four devices? A difference in the meantime to assemble for people? Is there evidence of a fatigue factor (a difference in mean response for positions in the assembly sequence).

Solution: The totals for rows, columns, and treatments are as shown in Table 8

Table 7.11 Latin Square Design revised

	i			
	1	2	3	4
R_i	155	181	193	207
C_i	202	161	191	182
T_i	202	142	204	188

And

$$CM = \frac{(736)^2}{16} = \frac{541{,}696}{16} = 33{,}856.0$$

Then

$$\text{Total SS} = \sum_{i=1}^{4}\sum_{j=1}^{4} y_{ij}^2 - CM$$

$$= 35{,}186.0 - 33{,}856.0 = 1330.0$$

$$SSR = \frac{\sum_{i=1}^{4} R_i^2}{4} - CM$$

$$= \frac{(155)^2 + (181)^2 + (193)^2 + (207)^2}{4} - CM$$

$$= 34{,}221.0 - 33{,}856.0 = 365.0$$

$$SSC = \frac{\sum_{i=1}^{4} C_j^2}{4} - CM$$

$$= \frac{(202)^2 + (161)^2 + (191)^2 + (182)^2}{4} - CM$$

$$= 34{,}082.5 - 33{,}856.0 = 226.5,$$

$$SST = \frac{\sum_{i=1}^{4} T_j^2}{4} - CM$$

$$\text{SST} = \frac{(202)^2 + (142)^2 + (204)^2 + (188)^2}{4} - CM$$

$$= 34{,}482.0 - 33{,}856.0 = 626.0.$$

Finally,

SSE = Total SS – SSR – SSC – SST

= 1330.0 – 365.0 – 226.5 – 626.0

= 112.5.

The analysis-of variance table for this example is shown as Table 9. Note that the mean squares were obtained by dividing the sums of squares by their, respective degrees of freedom.

Table 7.12 ANOVA table

Source	d.f.	SS	MS	F
Rows	3	365.0	121.67	6.49
Columns	3	226.5	75.50	4.03
Treatments	3	626.0	208.67	11.13
Error	6	112.5	18.75	
Total	15	1330.0		

All the computed F statistics are based on $v_1 = 3$ and $v_2 = 6$ degrees of freedom. The corresponding tabulated critical value is $F_{3,6} = 4.76$ (a = .05). A comparison of the computed F statistics with the tabulated value indicates that the computed F's for both rows and treatments exceed the critical value. Thus the data provide sufficient evidence to indicate a difference in the meantime to assemble the four devices. The data also show a difference in mean time to assemble for rows or, equivalently, positions in the sequences of assembly. It appears that fatigue, boredom, or some other factor increases the mean time to assembly as the length of employment increases.

7.5.8 Estimation for the Latin-Square Design

The $(1 - \alpha)$ confidence interval for the difference between a pair of treatment means is obtained in the same manner as for the completely randomized and the randomized block designs. Since each treatment mean will contain p observations, $n_1 = n_2 = p$ and the standard deviation of the difference between a pair of means is

$$\sigma_{(\bar{T}_i - \bar{T}_j)} = \sigma\sqrt{\frac{1}{p} + \frac{1}{p}}$$

The $(1 - \alpha)$ confidence interval is

$$(\bar{T}_i - \bar{T}_j) \pm t_{\alpha/2} S\sqrt{\frac{2}{p}}$$

Corresponding confidence intervals for the difference between a pair of row or column means are respectively,

$$(\bar{R}_i - \bar{R}_j) \pm t_{\alpha/2} S\sqrt{\frac{2}{p}}$$

$$(\bar{C}_i - \bar{C}_j) \pm t_{\alpha/2} S\sqrt{\frac{2}{p}}$$

Example 7.8: Refer to Example 7.7. Estimate the difference in mean response between treatments 1 and 2 using a 95 percent confidence interval.

Solution: From Table 7.12 observe that $s^2 = MSE = 18.75$ is based on six degrees of freedom. Consequently, the tabulated value, $t_{a/2}$, with six degrees of freedom is $t_{.025} = 2.447$. Then the 95 percent confidence interval for the difference between the treatment means, $(\mu_1 - \mu_2)$, is

$$(\bar{T}_i - \bar{T}_j) \pm t_{\alpha/2} S\sqrt{\frac{2}{p}}$$

$$(50.5 - 35.5) \pm (2.447)(4.33)\sqrt{\frac{2}{4}}$$

$$15 \pm 7.49$$

Thus we estimate the difference in mean time to assemble the two devices to be between 7.51 and 22.49 minutes.

7.5.9 Selecting the Sample Size

Selecting the sample size for the completely randomized or the randomized block design and the Latin-square design is an extension of the procedures of Section 7.5.7. We confine our attention to the case of equal sample sizes, $n_1 = n_2 = \ldots = n_p$, for the treatments of the completely randomized design. The number of observations per treatment is equal to b for the randomized block design and for a b x b Latin-square design. Thus the problem is to select n_1 or b for these three designs so as to purchase a specifies quantity of information.

The selection of sample size follows a similar procedure for all three designs; we will outline a general method. First the experimenter must decide on the parameter (or parameters) of major interest. Usually, he will wish to compare a pair of treatment means. Second, he must specify a bound on the error of estimation that he is willing to tolerate. Once determined, he need only select n_i (the number of observations in a treatment mean for a randomized block or Latin-square design) that will reduce the half-width of the confidence interval for the parameter so that it is less than or equal to the specified bound on the error of estimation. It should be emphasized that the sample-size solution will always be an approximation, since α is unknown until the sample is acquired. The best available value will be used for s in order to produce an approximate solution. We will illustrate the procedure with an example.

Example 7.9: A completely randomized design is to be conducted to compare teaching techniques in classes of equal size. Estimation of the difference in mean response on an achievement test is desired correct to within 30 test-score points, with probability equal to .95. It is expected that the test scores for a given teaching technique will possess a range approximately equal to 240. Find

the approximate member of observations required for each sample in order to acquire the specified information.

Solution: The confidence interval for the difference between a pair of treatment means is

$$(\bar{T}_i - \bar{T}_j) \pm t_{\alpha/2} S \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

Therefore, we will wish to select n_i and n_j so that

$$t_{\alpha/2} s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \leq 30$$

value of α is unknown and s is a random variable. However, an approximate solution for $n_i = n_j$ can be obtained by guessing s to be roughly equal to one-fourth of the range. Thus $s = 240/4 = 60$. The value of $t_{a/2}$ will be based upon $(n_1 + n_2 + \ldots + n_5 - 5)$ degrees of freedom, and for even moderate values of n_i, $t_{.025}$ will approximately equal 2. Then

$$t_{0.025} s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = (2)(60)\sqrt{\frac{2}{n_i}} \leq 30$$

$n_i = 32$, i= 1, 2,…,5.

Example 7.10: An experiment is to be conducted to compare the toxic effect of three chemicals on the skin of rats. The resistance to the chemicals was expected to vary substantially from rat to rat. Therefore, all three chemicals were to be tested on each rat, thereby blocking out rat-to-rat variability.

The standard deviation of the experimental error was unknown, but prior experimentation involving several applications of a given chemical on the same rat suggested a range of response measurements equal to 5 units.

Find a value for b such that the error of estimating the difference between a pair of treatment means is less than one unit, with probability equal to .95.

Solution: A very approximate value for s would be one-fourth the range, or s = 1.25. Then we wish to select b so that

$$t_{0.025}s\sqrt{\frac{1}{b}+\frac{1}{b}} = t_{0.025}s\sqrt{\frac{2}{b}} \leq 1$$

Since $t_{0.025}$ will depend upon the degrees of freedom associated with s^2, which will be (n – b – p + 1), we will guess $t_{0.025} \approx 2$. Then

$$(2)(1.25)\sqrt{\frac{2}{b}} \leq 1$$

Or b ≈ 13, i=1,2,3.

Approximately 13 rats will be required to obtain the desired information.

The degrees of freedom associated with s^2 will be 24, based on this solution. Therefore, the guessed value of t would seem to be adequate for this approximate solution.

The sample-size solutions for Examples 7.9 and 7.10 are very approximate, and are intended to provide only a rough estimate of approximate size and consequent cost of the experiment. The experimenter will obtain information on α as the data are being collected and can recalculate a better approximation to n as he proceeds.

Selecting the sample size for Latin-square designs follows essentially the same procedure as for the completely randomized and the randomized block designs. The only difference is that one must decide on the number of p x p Latin squares that will be needed to acquire the desired information. We omit discussion of this topic because we have not shown how to conduct an analysis of variance for more than one p x p Latin square. The reader interested in this topic should consult the references at the end of the unit.

7.0 UNIT ACTIVITY

1. State the assumptions underlying the analysis of variance of a completely randomized design.

2. Refer to Example 7.2. Calculate SSE by pooling the sums of squares of deviations within each of the four samples, and compare with the value obtained by subtraction. Note that this is an extension of the pooling procedure used in the two-sample case discussed in Section 7.5.2.

3 To compare the strengths of concrete produced by four experimental mixes, three specimens were prepared from each type of mix. Each of the 12 specimens was subjected to increasing compressive loads until breakdown. The following compressive loads in tons per square inch were attained at breakdown. Specimen numbers 1-12 are indicated in parentheses for identification purposes.

Table 7.13 Completely Randomized Design

Mix A	Mix B	Mix C	Mix D
(1) 2.30	(2) 2.20	(3) 2.15	(4) 2.25
(5) 2.20	(6) 2.10	(7) 2.15	(8) 2.15
(9) 2.25	(10) 2.20	(11) 2.20	(12) 2.25

Assuming that the requirements for a completely randomized design are met, analyze the data. State whether there is statistical support at the $\alpha = .05$ level of significance for the conclusion that the four types of concrete differ in average strength.

4 A clinical psychologist wished to compare three methods for reducing hostility levels in university students. A certain psychological test (HLT) was used to measure the degree of hostility. High scores on this test were taken to indicate great hostility. Eleven students obtaining high and nearly equal scores were used in the experiment. Five were selected at random from among the 11 problem cases and treated by method *A*. three were taken at

random from the remaining 6 students and treated by method *B*. the other 3 students were treated by method C. All treatments continued throughout a semester. Each student was given the HLT test again at the end of the semester, with the following results:

Table 7.14 Records of hostility levels

Method A	*Method B*	*Method C*
73	54	79
83	74	95
76	71	87
68		
80		

(a) Perform an analysis of variance for this experiment.

(b) Do the data provide sufficient evidence to indicate a difference in mean student response for the three methods after treatment?

5. A study was initiated to investigate the effect of two drugs, administered simultaneously, in reducing human blood pressure. It was decided to utilize three levels of each drug and to include all nine combinations in the experiment. Nine high-blood-pressure patients were selected for the experiment, and one was randomly assigned to each of the drug combinations. The response observed was a drop in blood pressure over a fixed interval of time.

(a) Is this a randomized block design?

(b) Suppose that two patients were assigned to each of the nine drug combinations. What type of experimental design is this?

6. A dealer has in stock three cars (car A, car B, and car C) of the same make and model. Wishing to compare these cars in gas consumption, a customer arranged to test each car with each of three brands of gasoline (brand A, brand B, brand C). In each trial, a gallon of gasoline was added to an empty tank and the car was driven without stopping until it ran

out of gasoline. The following table shows the number of miles covered in each of the nine trials.

7.15 Distance (miles) Brand of

Gasoline	Car A	Car B	Car C
A	22.4	17.0	19.2
B	20.8	19.4	20.2
C	21.5	18.7	21.2

(a) Should the customer conclude that the three cars differ in gas mileage? Test at the $\alpha = .05$ level.

(b) Do the data indicate that the brand of gasoline affects gas mileage?

7. A portion of a questionnaire was constructed to enable judges to evaluate a certain aspect of observed classroom teaching. Four films portraying teaching performances, and differing markedly in the teaching characteristic under study, were viewed by each of eight judges. The order of viewing the four films was assigned in a random manner to each judge. The data obtained are as follows:

Judges.

Table 7.16 Evaluation of classroom teaching

Films	1	2	3	4	5	6	7	8
1	9	10	7	5	12	7	8	6
2	4	9	3	0	6	8	2	4
3	12	16	10	9	11	10	10	14
4	9	11	7	8	12	7	7	8

(a) Give the type of design employed for this experiment and justify your diagnosis.

(b) How many degrees of freedom are available for estimating α^2? Perform an analysis of variance on the data.

(c) Do the data provide sufficient evidence to indicate that the mean questionnaire score varies from film to film? Test using $\alpha = .05$.

(d) Suppose that the data did provide sufficient evidence to indicate differences among the mean questionnaire scores for the four films. Would this imply that the

questionnaire was able to detect a difference in the teaching characteristic exhibited in the four films?

8. A completely randomized design was conducted to compare the effect of five stimuli on reaction time. Twenty-seven people were employed in the experiment, which was conducted using a completely randomized design. Regardless of the results of the analysis of variance, it is desired to compare stimuli *A* and *D*. the reaction times (in seconds) were as follows:

Table 7.17 Effect of five stimuli on reaction time

	Stimulus				
	A	*B*	*C*	*D*	*E*
	.8	.7	1.2	1.0	.6
	.6	.8	1.0	.9	.4
	.6	.5	.9	.9	.4
	.5	.5	1.2	1.1	.7
		.6	1.3	.7	.3
		.9	.8		
		.7			
Total	2.5	4.7	6.4	4.6	2.4
Mean	.625	.643	1.067	.920	.48

(a) Conduct an analysis of variance and test for a difference in mean reaction time due to the five stimuli.

(b) Compare stimuli A and D to see if there is a difference in mean reaction time.

9. An experiment was conducted to determine the effect of three methods of soil preparation on the first-year growth of slash pine seedlings. Four locations (state forest lands) were selected and each location was divided into three plots. Since it was felt that soil fertility within a location was more homogeneous than between locations, a randomized block design was employed using locations as blocks. The methods of soil preparation were A (no preparation), B (light fertilization), and C (burning). Each soil preparation was randomly applied to a plot within each location. On each plot the same number of

seedlings were planted, and the observation recorded was the average first-year growth (in centimeters) of the seedlings on each plot.

(a) Conduct an analysis of variance. Do the data provide sufficient evidence to indicate a difference in the mean growth for the three soil preparations?

(b) Is there evidence to indicate a difference in mean growth for the four locations?

(c) Use a 90 percent confidence interval to estimate the difference in mean growth for methods *A* and *B*.

Table 7.19 Effect of three methods of soil preparation

Soil Preparation	Location 1	2	3	4
A	11	13	16	10
B	15	17	20	12
C	10	15	13	10

10. Give the analysis of variance for the following 3 x 3 Latin-square design.

Table 7.20 Latin Square Design

Rows	*Columns* 1	2	3
1	B	A	C
	12	7	17
2	C	B	A
	10	7	4
3	A	C	B
	2	8	12

Find a 95 percent confidence interval for the difference in mean response for treatments A and C assuming that this is a preplanned comparison.

11. The following measurements are of the thickness of cake icing for the Latin-square design discussed in Section 7.5.2. The only difference between this exercise and the text discussion is that five (not three) mixes, A, B, C, D, and E, were employed in a 5 x 5 Latin-square design. Thickness measurements are given in hundredths of an inch.

Table 21 The thickness of cake icing

	Columns				
Rows	1	2	3	4	5
1	*D*	*E*	*C*	*A*	*B*
	9	18	6	8	11
2	*B*	*D*	*E*	*C*	*A*
	12	17	10	4	5
3	*A*	*B*	*D*	*E*	*C*
	6	16	10	9	4
4	*C*	*A*	*B*	*D*	*E*
	4	13	11	8	13
5	*E*	*C*	*A*	*B*	*D*
	14	11	7	10	15

Do the data provide sufficient evidence to indicate a difference in mean thickness of icing for the cake mixtures? Estimate the difference in mean thickness for mixtures *A* and *B*.

12. An experiment was conducted to investigate the toxic effect of three chemicals, *A, B,* and *C*, on the skin of rats. One-inch squares of skin were treated with the chemicals and then scored from 0 to 10, depending on the degree of irritation. Three adjacent 1-inch squares were marked on the backs of eight rats, and each of the three chemicals was applied to each rat. The experiment was blocked on rats to eliminate the variation in skin sensitivity from rat to rat. The data are as follows:

Table 7.22 Investigate the toxic effect of three chemicals, *A, B,* and *C*

1	2	3	4	5	6	7	8
B	A	A	C	B	C	C	B
5	9	6	6	8	5	5	7

A	C	B	B	C	A	B	A
6	4	9	8	8	5	7	6
C	*B*	*C*	*A*	*A*	*B*	*A*	*C*
3	9	3	5	7	7	6	7

(a) Do the data provide sufficient evidence to indicate a difference in the toxic effect of the three chemicals?

(b) Estimate the difference in mean score for chemicals A and B using a 95 percent confidence interval.

13. Consider the following one-way classification consisting of three treatments, *A, B,* and *C*, where the number of observations per treatment varies from treatment to treatment.

Table 7.23 classification of three treatments, *A, B,* and *C*,

A	*B*	*C*
24.2	24.5	26.0
27.5	22.7	
25.9		
24.7		

Do the data present sufficient evidence to indicate a difference between treatments?

7.0 UNIT SUMMARY

The completely randomized, the randomized block, and the Latin-square designs are illustrations of experiments involving one, two and three qualitative independent variables, respectively. The analysis of variance partitions the total sum of squares of deviations of the response measurements about their mean into portions associated with each independent variable and the experimental error. The former may be compared with the sum of squares for error, using mean squares and the F statistics, to see whether the mean squares for the independent variable are unusually large and thereby indicative of an effect on the response.

In this unit you have presented a very brief introduction to the analysis of variance and its associated subject, the design of experiments. Experiments can be designed to investigate the effect of many quantitative and qualitative variables on a response. These may be variables of primary interest to the experimenter as well as nuisance variables, such as blocks, which we attempt to separate from the experimental error. These experiments are subject to an analysis of variance when properly designed. A more extensive coverage of the basic concepts of experimental design and the analysis of experiments will be found in the references.

UNIT 8 LINEAR REGRESSION

8.1 Unit Introduction

Welcome to Unit 8 in which you will learn methods of linear regression analysis. In this unit you will learn methods Statistical Data Analysis technique. You will further learn that linear regression is used to determine the extent to which there is a linear relationship between a dependent variable and one or more independent variables. You will also learn that you can use this analysis to make predictions of future values.

8.2 Unit Aim

The aim of this Unit is to teach you how to conduct linear regression analysis.

8.3 Unit Objectives

By the end of the unit you should be able to:

- Estimate the relationship between a set of independent variables (regressors) and some dependent variable (outcome).
- Predict" the value of the dependent variable based upon the values of one or more independent variables.
- Predict a continuous dependent variable from a number of independent variables. If the dependent variable is dichotomous, then logistic regression should be used

Terminology

$\hat{a}$ **:** Intercept estimate

$\hat{b}$ **:** Slope estimates

8.4 Unit Time required

You need 20 hours for this unit

8.5 Unit Topics

8.5.1 Analysing the results of an experiment

Figure 8.1 illustrates the apparatus used in a simple experiment which may people will have performed for themselves. When weights are added to the scale pan, the spring stretches. Table 8.1 shows the result obtained when different loads were applied in a random order, and the corresponding scatter diagram is shown in Figure 2. In this experiment the values of *Y*, the length of the spring in cm, were measured for preselected values of *X*, the load in newtons, and for this reason X is called the independent variable and *Y* the dependent variable.

Figure 8.2 suggests that there is a linear relationship between *X* and *Y*: how should we draw a straight line to represent this relationship? Drawing a line by eye is obviously not satisfactory, since the result will be subjective, but as a method it can lead us to a mathematical method. Figure 3 shows schematically the points on a scatter diagram and a possible straight line which relates them. In this experiment the values of *X* are extremely accurately known. The measurement of the length of the spring is less accurate. If the experiment were repeated (using the same loads) the values of *Y* for a given value of *X* would show random variation. It is this variation which we may reasonably assume causes the points on the scatter diagram to deviate from a straight line. For example, in Figure 3 (x_i, y_i) deviates by e_1 and (x_2, y_2) by e_2 (in the opposite direction) from the straight line which has been drawn. We can represent the relationship between any pair of values (x_i y_i) by

$$y_i = \alpha + \beta x_i + e_i \qquad (1)$$

Where $\alpha + \beta x_i$ represents the linear relationship between y_i and x_i and e_i is a random error. β is called the coefficient of regression of Y on X. We shall assume that the e_i's are independent of x_i and normally distributed with mean 0 and s.d. $\sigma_{Y/X}$

Table 8.1 Results for extension of a spring

x_i, load (in newtons)	y_i, length of spring (cm)
0.1	10.7
0.2	11.3
0.3	12.0
0.4	12.4
0.5	13.0
0.6	13.7
0.7	14.5
0.8	15.1
0.9	15.6
1.0	16.0

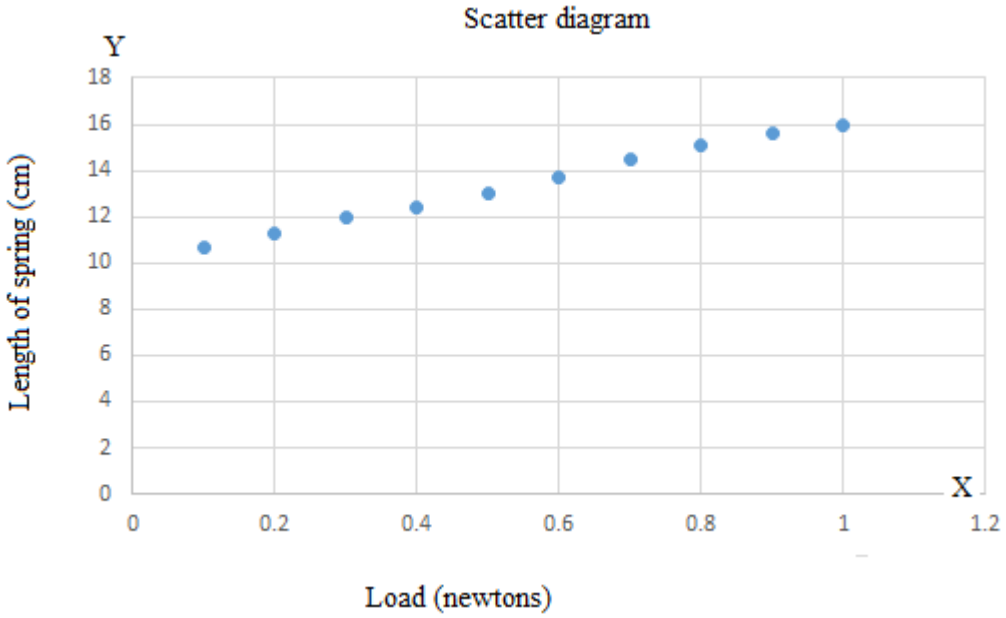


Figure 8.1. Scatter diagram for the data in Table 8.1

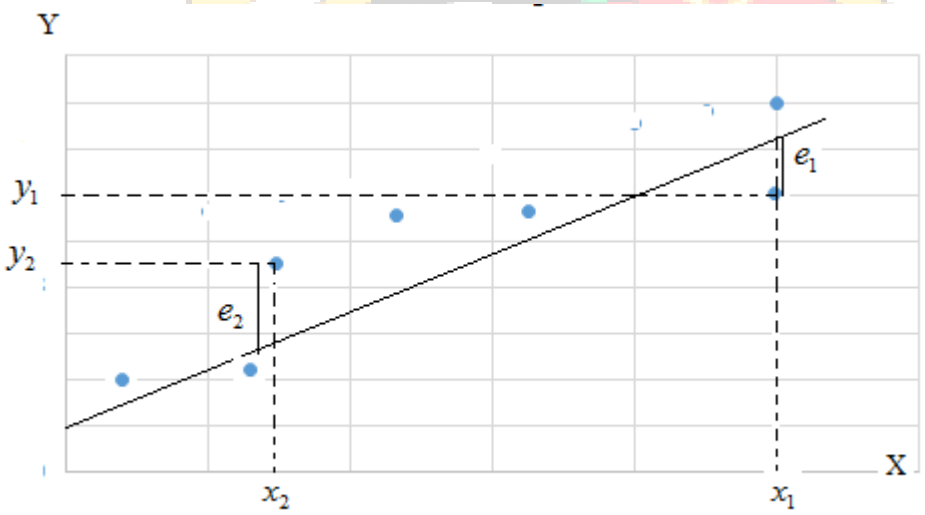


Figure 8.2. Schematic diagram showing the deviation of the measurements from a possible straight line relating *X* and *Y*.

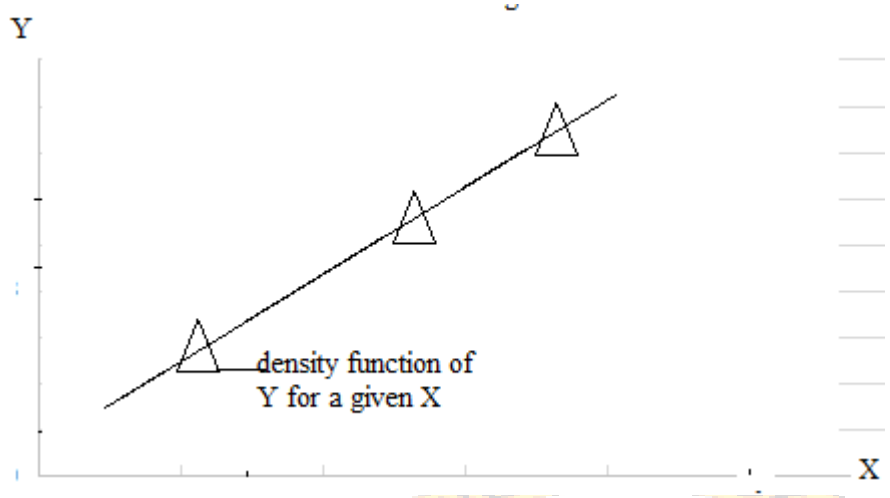


Figure 8.3. Diagram showing the distribution of repeated values of *Y* for a given value of *X*

Distribution of repeated measurements of y_i for each x_i and the appropriate line relating x_i and y_i which Passes through the mean values of the y_i distributions.

8.5.2 The method of least squares

The parameters α and β are estimated by the **method of least squares**, so called because estimates, $\hat{a}$ and $\hat{b}$, of α and β respectively are chosen so as to minimise $\sum_{i=1}^{n} e_i^2$ (n is the number of points on the scatter diagram.) The squares of the deviates are used for the same reasons as the squares of the deviations are used in the calculation of standard deviation.

Rearranging equation (1) $e_i = y_i - \alpha - \beta x_i$ $\sum_{i=1}^{n} e_i^2 = \sum (y_i - \alpha - \beta x_i)^2$.

This expression can be varied by varying α and β, $\hat{a}$ and $\hat{b}$, the estimates of α and β, are chosen so that it is minimised. The approximate values of $\hat{a}$ and $\hat{b}$ are found by putting the partial derivatives of $\sum (y_i - \hat{a} - \hat{b}x_i)^2$ with respect to $\hat{a}$ and to $\hat{b}$ equal to zero:

$$\frac{\partial}{\partial \hat{a}} \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2 = \sum_{i=1}^{n} -2(y_i - \hat{a} - \hat{b}x_i) = 0$$

$$\frac{\partial}{\partial \hat{b}} \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2 = \sum_{i=1}^{n} -2x_i(y_i - \hat{a} - \hat{b}x_i) = 0$$

This gives two simultaneous equation in $\hat{a}$ and $\hat{b}$. These equations can be rewritten as follows:

$$\sum_{i=}^{n} -2(y_i - \hat{a} - \hat{b}x_i) = 0$$

Gives

$$\sum_{i=1}^{n} y_i - n\hat{a} - \hat{b}\sum_{i=1}^{n} x_i = 0 \qquad (2)$$

and

$$\sum_{i=1}^{n} -2x_i(y_i - \hat{a} - \hat{b}x_i) = 0$$

gives

$$\sum_{i=1}^{n} x_i y_i - \hat{a}\sum_{i=1}^{n} x_i - \hat{b}\sum_{i=1}^{n} x_i^2 = 0 \qquad (3)$$

(2) and (3) are known as the normal equations. They can be solved simultaneously.

Subtracting (3) multiplied by n from (2) multiplied by $\sum_{i=1}^{n} x_i$ gives:

$$\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - \hat{b}\left(\sum_{i=1}^{n} x_i\right)^2 - n\sum_{i=1}^{n} x_i y_i + n\hat{b}\sum_{i=1}^{n} x_i^2 = 0$$

Rearranging

$$\hat{b} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (4)$$

From (2) a, the estimated value of a, is given by

$$\hat{a} = \left(\sum_{i=1}^{n} y_i - \hat{b}\sum_{i=1}^{n} x_i\right)/n$$

giving $\hat{a} = \bar{y} - \hat{b}\bar{x}$

The relationship between x and y is given by

$$y = \hat{a} + \hat{b}x_i \qquad (5)$$

and this is known as the regression line of Y on X, An alternative form is given by substituting the expression (5) for a in equation (5): $y = \bar{y} - \hat{b}\bar{x} + \hat{b}x$.

Giving $y - \bar{y} = \hat{b}(x - \bar{x})$ (6)

From (6) you can see that the regression line passes through (x, y), a fact which is useful when drawing it on a scatter diagram.

Table 8.2 Strength of the string

x_i	y_i	$x_i y_i$	x_i^2
0.1	10.7	1.07	0.01
0.2	11.3	2.26	0.04
0.3	12.0	3.60	0.09
0.4	12.4	4.96	0.16
0.5	13.0	6.50	0.25
0.6	13.7	8.22	0.36
0.7	14.5	10.15	0.49
0.8	15.1	12.15	0.64
0.9	15.6	14.04	0.81
1.0	16.0	14.04	1.00
5.5	134.3	78.88	3.85

Example 8.1

Calculate the regression line of Y on X for the data given in Table 8.1 (reproduced in Table 8.2), and use it to predict the length for a load of 0.65 N. n = 10

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5.5}{10} = 0.55$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{134.3}{10} = 13.43$$

$$\hat{b} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{b} = \frac{10 * 78.88 - 5.5 * 134.3}{10 * 3.85 - 5.5^2} = 6.08$$

The regression line of Y on X is

$$y - \bar{y} = \hat{b}(x - \bar{x}) \quad (6)$$

$$y - 13.43 = 6.08(x - 0.55)$$

$$y = 6.08x + 10.1$$

To predict the length for a load of 0.65 N we substitute this value of x into the equation of the regression line.

$$y = 6.08 * 0.65 + 10.1$$
$$= 14.1$$

8.5.3 Coded values

The calculation in the previous section can be greatly simplified using coded values as in Section 7.5.2. With $x_i = A + Bu_i y_i = C + Dv_i$

$$\hat{b} = \frac{D}{B}\left\{\frac{n\sum_{i=1}^{n} u_i v_i - \sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i}{n\sum_{i=1}^{n} u_i^2 - \left(\sum_{i=1}^{n} u_i\right)}\right\} \qquad (7)$$

(Calculation may often be further simplified by choosing a value of A which makes $\sum_{i=1}^{n} u_i = 0$

Example 8.2

Calculate the regression line of Y on X for the data in Table 1 using coded values.

Table 8.3 Strength of the string

x_i	y_i	u_i	v_i	$u_i v_i$	u_i^2
0.1	10.7	- 5	- 2.3	11.5	25
0.2	11.3	- 4	- 1.7	6.8	16
0.3	12.0	- 3	- 1.0	3.0	9
0.4	12.4	-2	-0.6	1.2	4
0.5	13.0	-1	0	0	1
0.6	13.7	0	0.7	0	0
0.7	14.5	1	1.5	1.5	1
0.8	15.1	2	2.1	4.2	4
0.9	15.6	3	2.6	7.8	9
1.0	16.0	4	3.0	12.0	16
		-5	4.3	48.0	85

Solution

Taking A = 0.6, C = 13.0, B = 0.1 , D = 1 you have table 8.3. Using equation (7)

$$\hat{b} = \frac{1}{0.1} * \frac{10*48.0-(-5)*4.3}{10*85-(-5)^2} = 6.08$$

$$\bar{x} = A + B\bar{u}$$

$$\bar{x} = 0.6 + 0.1*\left(\frac{-5}{10}\right)$$

$$= 0.55$$

$$\bar{y} = C + D\bar{v}$$

$$\bar{y} = 13 + 1*\frac{4.3}{10} = 13.43$$

And the regression line is:

$$y = 6.08x + 10.1$$

8.1 UNIT ACTIVITY

(1) A scientist, working in an agricultural research, believes there is a relationship between the hardness of the shells of eggs laid by chickens and the amount of a certain food supplement put into the diet of the chickens. He selects ten chickens of the same breed and collects the data of Table 4.

(Hardness is measured on a 0 – 10 scale, 10 being the hardest. There are no units attached).

(a) Calculate the equation of the regression line of y and x.

(b) Calculate the product moment correlation coefficient.

(c) Do you b8elieve that his linear model will continue to be appropriate no matter how large or small x becomes?

Just your reply

Table 8.4 Measurements of hardness of egg shells

Chicken	A	B	C	D	F	G	H	I	J
Amount of food supplement x(g)	7.0	9.8	11.6	17.5	7.6	8.2	12.4	15.5	9.5	19.5

Hardness of shells	1.2	2.1	3.4	6.1	1.3	1.7	3.4	6.2	2.1	7.1

Table 8.5 Consumptions of petrol in Lusaka

Year (x)	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972
Consumption (y) (millions of gallons)	32.5	37.1	35.5	37.7	41.5	46.4	44.8	45.8	53.9	62.0

(2) The figures in Table 8.5 give the wine consumption in Lusaka in millions of gallons (y) for the years 1963 to 1972 (x).

Draw a scatter diagram to show these data.

Determine the last squares estimate of the regression line of y and x, showing all your working. Draw this line on your scatter diagram and use it to estimate the consumption for 1973.

Comment on the appropriateness of a linear regression model in this case, given also that the actual wine consumption in 1973 was 78.3 million gallons.

(3) The body and heart masses of fourteen ten-month-old male mice are given in Table 8.6.

(a) Draw a scatter diagram of these data.

(b) Calculate the equation of the regression line of y and x and draw this line on the scatter diagram.

(c) Calculate the product moment coefficient of correction

(4) A regression line, y = a + bx, is to be fitted to a set of data points (x,y). The data are coded to X = (x – c1)d1, Y = (y – x2)/d2, and in terms of these the regression line is Y = a T BX. Find a and b in term s of A and B and the coding constants.

In such a problem the values of the independent variable are equally paced. Explain, with reference to the appropriate least squares equations, how this simplifies the fitting of the line.

Table 8.6 Measurements of body mass

Body mass (x)

(grams)27	30	37	38	32	36	32	32	38	42	36	44	33	38	
Heart mass(y) (milligrams)														
118	136	156	150	140	155	157	114	144	156	149	170	131	160	

Table 8.7 shows the mass of a certain animal at weekly intervals.

Age (weeks)	11	12	13	14	15	16
Mass (g)	357	382	404	423	440	451

State whether it would be more appropriate to fit a regression line of age on mass or mass on age to these data, and justify your choice. What value does the line of your choice give for the mass at 17 Weeks? Comment on the validity of your estimate. (You are advised to plot rough graph.)

8.5.4 Estimating a value of $\sigma^2_{Y/X}$

For a value of xi the value of Y predicated by the regression line is given by $y_i' = \hat{a} + \hat{b}x_i$

The difference between y_i and y_i' known as a residual, gives the values of e_i for that point.

Each value of $y_i' - y_i$ gives a value of e_i from a distribution which is $N(0, \sigma^2_{Y/X})$. The unbiased estimate $\hat{s}_{Y/X}$ of $\sigma_{Y/X}$ is given by

$$\hat{s}_{Y/X} = \sqrt{\left(\frac{\sum_{i=1}^{n}(y_i' - y_i)^2}{n-2}\right)} \qquad (8)$$

The divisor is (n – 2) since two degrees of freedom are lost because there are two constraints: the vlues of $\hat{a}$ and $\hat{b}$ are calculated from the values of X and Y.

Example 8.3

Obtain an estimate of $\sigma^2_{Y/X}$ for the data in Table 1.

$$\hat{s}_{Y/X} = \sqrt{\left(\frac{\sum_{i=1}^{n} (y_i' - y_i)^2}{n-2} \right)}$$

$$\hat{s}_{Y/X} = \frac{0.11784}{8} = 0.121$$

In section 8.5.2 the regression line of Y and X was calculated as y = 6.08x + 10.1. Using this equation, the predicted values for Y are given in Table 8.8.

From equation (8)

Table 8.8 Strength of string

x_i	y_i	y_i'	$(y_i' - y_i)^2$
0.1	10.7	10.708	0.000 06
0.2	11.3	11.316	0.000 26
0.3	12.0	11.924	0.005 78
0.4	12.4	12.532	0.017 42
0.5	13.0	13.140	0.019 60
0.6	13.7	13.748	0.002 30
0.7	14.5	14.356	0.020 74
0.8	15.1	14.964	0.018 50
0.9	15.6	15.572	0.000 78
1.0	16.0	16.180	0.032 40
			0.117 84

8.5.5 Confidence limits for β

How accurate is the estimate, b, which we have made of β ? This section will show how the confidence limits for β can be found in terms of $\sigma_{Y/X}$ the s.d. of the random errors, e_y. To recapitulate, we assume the e_i 's are randomly distributed with a Normal distribution mean 0.s.d. $\sigma_{Y/X}$ thus by definition, $E(e_i) = 0$, (9)

$\text{var}(e_i) = \sigma^2{}_{Y/X}$ (10)

We will now calculate the expected value and variance of $\hat{b}$ in terms of the e_i's

From equation

$$\hat{b} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (11)$$

Substituting $y_i = \beta x + \alpha + e_i$ (equation (1) the numerator becomes

$$n\sum_{i=1}^{n} x_i(\beta x_i + \alpha + e_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} (\beta x_i + \alpha + e_i)$$

$$= n\beta\sum_{i=1}^{n} x_i^2 + n\alpha\sum_{i=1}^{n} x_i + n\sum_{i=1}^{n} x_i e_i - \beta\left(\sum_{i=1}^{n} x_i\right)^2 - n\alpha\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} e_i$$

$$= \beta\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] + n\sum_{i=1}^{n} x_i e_i - n\sum_{i=1}^{n} \bar{x} e_i$$

$$= \beta\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right] + n\sum_{i=1}^{n} (x_i - \bar{x}) e_i$$

Dividing by the denominator $n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2$ in the expression for $\hat{b}$, we have

$$\hat{b} = \beta + \frac{n\sum_{i=1}^{n} (x_i - \bar{x}) e_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \beta + \frac{n\sum_{i=1}^{n} (x_i - \bar{x}) e_i}{n\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \beta + \frac{\sum_{i=1}^{n} (x_i - \bar{x}) e_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

In calculating the expected value and variance of $\hat{b}$, the term in x_i is constant for each e_i since the values of x_i are predetermined.

This gives

$$E(\hat{b}) = E\left[\beta + \frac{\sum_{i=1}^{n}(x_i - \bar{x})e_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

$$= \beta + \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(e_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \beta$$

Since $E(e_i) = 0$

Thus $\hat{b}$ gives an unbiased estimate of β. Also

$$\text{Var}(\hat{b}) = \text{var}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})e_i}{i=1\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

Since β is constant

Using the relationship var(cy)= c^2 var(y) where c is a constant

$$\text{var}(\hat{b}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ var}(e_i)}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$$

$$\text{var}(\hat{b}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma^2_{Y/X}}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$$

$$\text{var}(\hat{b}) = \frac{\sigma^2_{Y/X}}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]} \qquad (12)$$

Assuming the sampling distribution of $\hat{b}$ is Normal, the 95% confidence limits of β are given by

$$\beta = \hat{b} \pm \frac{1.96\sigma_{Y/X}}{\sqrt{\left\{\sum(x_i - \bar{x})^2\right\}}} \qquad (13)$$

If $\sigma_{Y/X}$ is estimated from the samples as then the 95% confidence limits of β are

$$\beta = \hat{b} \pm \frac{t_{n-2,5\%}\hat{s}_{Y/X}}{\sqrt{\left\{\sum(x_i - \bar{x})^2\right\}}} \qquad (14)$$

Where the t-distribution is used because the variance is estimated from the sample with n – 2 degrees of freedom.

Example 8.4

Calculate the 95% confidence limits for the estimate of β for the data in Table 8.1.

In equation (12) we cannot substitute $\sigma_{Y/X}$ but only the estimate of it, since $\hat{s}_{Y/X}$, which are found in Section 8.5.4 to be 0.2.1 $\sum_{i=1}^{n}(x_i - \bar{x})^2$. Can be found most simply using the identity

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2 / n$$

$$= 3.85 - \frac{5.5^2}{10} = 0.825 \quad \text{(Using the values obtained in Example 8.2)}$$

Which gives our estimate of the variance of $\hat{b}$ as

$$\sqrt{\text{var}(\hat{b})} = \frac{\hat{s}_{Y/X}}{\sqrt{\left\{\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2 / n\right\}}} = \frac{0.121}{\sqrt{0.825}}$$

The 95% confidence limits of β are, from equation (14),

$$\beta = \hat{b} \pm t_{n-2,5\%} \sqrt{[\text{var}(\hat{b})]}$$
$$= 6.08 \pm 2.31 * 0.133$$
$$= 6.08 \pm 0.31$$

8.5.6 Confidence limits for α

From equation (5) $\hat{a} = \bar{y} - \hat{b}\bar{x}$

$$= \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{b}\bar{x}$$

Thus

$$E(\hat{a}) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) - E(\hat{b}\bar{x}) \text{ Since}$$

$$y_i = \alpha + \beta x_i + e_i$$

$$E(y_i) = \alpha + \beta x_i + E(e_i)$$

$$= \alpha + \beta x_i$$

Giving

$$E(\hat{a}) = \frac{1}{n}\sum_{i=1}^{n} (\alpha + \beta x_i) - \bar{x}E(\hat{b})$$

$$= \alpha + \beta\bar{x} - \beta\bar{x}$$

$$= \alpha$$

Showing that $\hat{a}$ is an unbiased estimate of α. Also

$$\text{var}(\hat{a}) = \text{var}\left\{\sum_{i=1}^{n} y_i / n - \hat{b}\bar{x}\right\}$$

It can be shown that the covariance of $\bar{y}$ and $\hat{b}$ is zero so that

$$\text{var}(\hat{a}) = \text{var}\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) + \text{var}(\hat{b}\bar{x})$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(y_i) + \bar{x}^2\ \text{var}(\hat{b})$$

Since $y_i = \alpha + \beta x_i + e_i$

Var(y_i) $= \text{var}(e_i) = \sigma^2_{Y/X}$ and equation(12) gives a value for $\text{var}(\hat{b})$ substituting, you have

$$\text{var}(\hat{a}) = \frac{1}{n^2} * n\sigma^2_{Y/X} + \frac{\bar{x}^2\sigma^2_{Y/X}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\text{var}(\hat{a}) = \sigma^2_{Y/X}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \qquad (15)$$

Example 8.5

Find the s,d, and 95% confidence limits for α for the data in table 8.1

Solution

You do not have an exact value for $\sigma_{Y/X}$, but only its estimate, found in section 8.5.4 to be 0.121. The summation $\sum_{i=1}^{n}(x_i - \bar{x})^2$ was found in example 8.5 to be 0.825 and $\bar{x} = 0.55$

substituting these values in (equation 15) gives

$$\text{var}(\hat{a}) = 0.121^2\left[\frac{1}{10} + \frac{0.55^2}{0.825}\right]$$

$$\sqrt{\text{var}(\hat{a})} = 0.0827$$

The 95% confidence limits of α are;

$$\alpha = \hat{a} \pm t_{n-2,5\%}\sqrt{[\text{var}(\hat{a})]}$$

Where the t distribution is use because var($\hat{a}$) is estimated from the sample. Substituting

$$\alpha = 10.1 \pm 2.31 * 0.0827$$
$$= 10.1 \pm 0.2$$

8.5.7 Confidence limits of predicted values

In example 8.5 the regression line was used to predict the true value of Y, y_o, for a value of X, x_o. To do this you can use equation (6). $y_o - \bar{y} = \hat{b}(x_o - \bar{x})$

Which gives $y_o = \hat{b}(x_o - \bar{x}) + \bar{y}$

Since the covariance of $\bar{y}$ and $\hat{b}$ is zero, gives

$\text{Var}(y_o) = (x_o - \bar{x})^2 \text{ var}(\hat{b}) + \text{var}(\bar{y})$

From equation (15)

$$\text{Var}(\hat{b}) = \frac{\sigma^2_{Y/X}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

And

$$\text{Var}(\bar{y}) = \frac{\sigma^2_{Y/X}}{n}$$

Giving

$$\text{var}(y_o) = \sigma^2_{Y/X}\left[\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \qquad (16)$$

Returning to example in section 8.5.2, the regression line of Y on X was

$y = 6.08x + 10.1$

When x = 0.65, $y = 6.08 * 0.65 + 10.1 = 14.10$

In this example 8.5 you found $\hat{s}_{Y/X} = 0.121$ and in example 8.5 using equation (16) gives your estimate of the variance of y_o as

$$\text{var}(y_o) = 0.121^2\left[\frac{1}{n} + \frac{(0.65-0.55)^2}{0.825}\right]$$

$$s.d(y_o) = 0.0405$$

And 95% confidence limits of y_o are $y_o = 14.10 \pm 2.31 * 0.0405 = 14.10 \pm 0.09$

Whereas before t_{n-2} has been used since the s.d. was estimated from the sample.

Study of equation (16) shows that the confidence limits for y_o will have the smallest range when $x_o = \bar{x}$, and the range increases towards the ends of the regression line.

Example 8.6

It is known that the true response Y in a certain chemical experiment is a linear function of the operating temperature X. However, the experimental determinations of Y are subject

Table 8.9 Measurements of temperature

Temperature(X)	30	40	50
Observed response(Y)	14	10	7
	12	11	6

To random errors, so that when an experiment is performed at temperature x_i the observe response y_i in such that $y_i = \alpha + \beta x_i + e_i$

Where $\alpha + \beta x_i$ is the true response and e_i is the error. Table 8.9 gives the observed responses in six experiments two at each of the three temperature.

Use the data to obtain the least estimate of the linear relationship connecting X and Y (You are given that $\sum_{i=1}^{n} x_i y_i = 2270$)

The errors e_i are independent and normally distributed with zero mean and unit standard deviation. Calculated 90% confidence limits for

(a) The value of α (b) the value of β (c) the true value of Y, when X is 50.

Solution

Using the six pairs of values in the table you have

$$\sum_{i=1}^{n} x_i y_i = 2270,\ \sum_{i=1}^{n} x_i^2 = 10000,\ \sum_{i=1}^{n} x_i = 240,\ \sum_{i=1}^{n} y_i = 60$$

From equation (4)

$$\hat{b} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{6*2270 - 240*60}{6*10000 - 240^2}$$

$$= -0.325$$

From equation (6)

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$= \frac{60}{6} - (-0.325)*\frac{240}{6}$$

$$= 23$$

From equation (6),

$$\text{var}(\hat{b}) = \frac{\sigma^2_{Y/X}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\text{var}(\hat{b}) = \frac{\sigma^2_{Y/X}}{\left\{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n\right\}}$$

You are given $\sigma^2_{Y/X} = 1$ so

$$\text{var}(\hat{b}) = \frac{1^2}{\left\{10000 - \frac{240^2}{6}\right\}} = \frac{1}{400}$$

$$s.d.(\hat{b}) = 0.05$$

From equation (15)

$$\text{var}(\hat{a}) = \sigma^2_{Y/X}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

$$= 1^2\left[\frac{1}{6} + \frac{40^2}{400}\right]$$

s.d. $(\hat{a}) = 2.04$

(a) Using the value of $\hat{a}$ and its variance obtained above the 90% confidence limits of α are

$\alpha = 23 \pm 1.64 * 2.04 = 23 \pm 3.35$

(Where z is used since $\sigma_{Y/X}$ is known).

i. Using the value of $\hat{b}$ and its variance obtained above , the 90% confidence limits of β are

$\beta = -0.325 \pm 1.64 * 0.05 = -0.325 \pm 0.082$

ii. Using the regression line of Y on X, the value of Y, y_o, when $x_o = 50$ is given by

$$\text{var}(y_o) = \sigma^2_{Y/X}\left[\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

$$= 1^2\left[\frac{1}{6} + \frac{(50-40)^2}{400}\right]$$

$$\sqrt{[\text{var}(y_o)]} = 0.645$$

The 90% confidence limits of the predicted value are

$y_o = 6.75 \pm 1.64 * 0.645$

$= 6.8 \pm 1.1$

8.2 UNIT ACTIVITY

1. Two variable X and Y of interest in an experiment are known to be linearly related, but the coefficients in this relationship are not known. A series of 15 experiments was conducted in which three determinations of y were made for each of the five values of x. The x- values are accurate but the determinations of the y- values are subject to independent random errors that are normally distributed with mean zero and standard deviation 1.25.The observed values in the

series of experiments are shown in table 8.10. The following equations were calculated from the table

Table 8.10 Two variable X and Y

x	1	2	3	4	5
	19	17	12	9	3
y	18	16	11	10	3
	21	17	13	7	4

$\sum x = 45,\ \sum y = 180,\ \sum x^2 = 165,\ \sum xy = 420$

(a) Calculate the equation of the least square estimate of the linear relationship between x and y.

(b) Estimate the true value of y when x – 4. Determine the standard error of this estimate. Explain why this estimate of y when x = 4 is preferable to that obtained from averaging the three observed values of y when x = 4.

(c)Calculate a 90% confidence interval for the true value of y when x = 4.

2. In an experiment to find the young modulus for a brass wire the eleven pairs of values x

Table 8.11 Measurements of young modulus for a brass wire

X	1	1.5	2	2.5	3	3.5	3	2.5	2	1.5
y	-1.1	-0.6	0	0.4	0.9	1.5	1.0	0.6	0.1	-0.5

(suspended mass kg) and y (length of wire in mm-7000mm) given in table 11 are obtained.

The equation connecting x and y is assumed to take the form y = c + kx

Obtain the least squares estimate values for c and k and 95% confidence limits for k.

$\sum x^2 = 57.25, \sum y^2 = 7.22,\ \sum xy = 10.0$

3. The size (z) of an organism was measured at different times 9x giving the data shown in the first two rows of table 8.12. The third row of the table gives the values of $y = \log_{10}^{z}$ to two decimal places.

Table 8.12 The size of an organism

Time(x)	1	3	5	7	9
Size(z)	1.48	3.02	5.37	9.55	17.38
$y = \log_{10}^{z}$	0.17	0.48	0.73	0.98	1.24

Without carrying out any nontrivial calculations show that the assumption of a linear relationship between y and x is more realistic than between z and x.

The following quantities were calculated from the data:

$$\sum x = 25,\ \sum x^2 = 165,\ \sum y = 3.6,\ \sum xy = 23.28$$

Suppose that $y = \alpha + \beta x$ and that for given values of x the observed values of y are subject to independent errors which are normally distributed with mean zero and standard deviation 0.02.

(a) Calculate the least squares estimates of α and β

(b) Obtain 95% confidence limits for the value of y when x = 10. Hence find 95% confidence limits for the size of the organism at time x = 10.

4. In an investigation of the effect of duration of training (x) on performance time (y) for a certain repetitive job, the following observations were obtained from 26 trainees;

$$\sum x = 104,\ \sum y = 208,\ \sum (x-\bar{x})^2 = 56,\ (x-\bar{x})(y-\bar{y}) = -56,\ \sum (y-\bar{y})^2 = 62$$

Plot the linear relationship of y to x which best fits the observation. Calculate 95% confidence limits for the value predicted for y from this relationship when x = 6. Sketch the form of these confidence limits for varying values of x.

Calculate the value predicted for y when x = 12: What does this value suggest to you concerning the form of the relationship between y and x?

8.0 UNIT SUMMARY

Regression line of Y on X, $y = \alpha + \beta x$

Model: $y_i = \alpha + \beta x_i + e_i$ where e_i is $N(0, \sigma^2_{Y/X})$

$$\hat{b}(least\ \ square\, estimate\, of\ \beta) = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{a}\,(least\ squaresestimateof\ a) = \bar{y} - \hat{b}\bar{x}$$

$$\hat{s}_{Y/X}\,(unbiased\, estimate\, of\ \sigma_{Y/X}) = \sqrt{\left(\frac{\sum_{i=1}^{n}(y_i' - y_i)^2}{n-2}\right)}$$

Where $y_i' = -y_i$ (the residuals) $= \hat{a} + \hat{b}x_i - y_i$

9.0 MODULE SUMMARY

The module MAT 2602 is an introduction to statistics but this will give you a very good start in the field of statistics. As defined in unit 1, were statistics is a science that deals with the methods of data: Collections, Compilation, Presentation, analysis, and Interpretation of results, Conclusion and above all making decisions based on the study of the data. What you will learning in this module will be up to conclusion about a particular problem. Unit 1 gives you basic ideas about data in general in which you will learn that data can be discrete or continuous and how you can calculate the statistics. In this unit you will learn how to present different types of data in a pictorial why to make non-mathematician audience understand what the data all about. In unit 2 you will learn sampling procedures and how to estimate parameters from samples. In this unit you will learn how to estimate into the population and make a distinction between biased and unbiased estimates which will lead

you to: Confidence interval and significance test in unit 3 and 4 respectively. Unit 5 will give you methods of test of goodness which is similar to the previous unit. Unit 6 will give you procedures of how to conduct a correlation analysis in which you will learn how to calculate the correlation coefficient between two variables, say X and Y. You will learn that a relationship can either be shown as a scatter diagram or calculate the correlation coefficient. In the previous units under tests of hypotheses you learnt how to conduct tests between a value or two values. In unit 7 under the analysis of variance you will learn how to find differences between means of more than two variables. Lastly, in unit 8 you will learn how to conduct a linear regression analysis. In this unit you will learn how to estimate a least regression line and how to predict the future. This is a basis for advanced statistical analysis.

10.0 SYLLABUS: MAT 2602 - INTRODUCTION TO STATISTICS

Pre-requisites: MAT 1100-Foundation Mathematics

1. **Introduction**

   What is statistics, elements of statistics; Sampling Methods

   (Data collection); Data representation: graphical techniques, frequency distribution.

2. **Estimation and sampling distributions**

   Point estimates for mean, difference of means, proportions, difference of proportions and their sampling distributions; Confidence intervals for mean, difference of means, proportion, difference of proportions, variance and ratio of variances.

3. **Statistical hypothesis testing**

   Definition of statistical hypothesis, type I and type II errors; Tests of hypothesis about the mean, proportion, the difference of means and proportions for large and small samples; Tests of hypothesis about variance and ratio of variances; Tests of independence and goodness of fit test.

4. **Analysis of variance**

   Introduction to linear model and experimental design; Completely randomized design (CRD), balanced and unbalanced; Randomized block design (RBD), unreplicated and replicated.

5. **Linear regression and correlation analysis**

   Introduction; Simple linear regression model, least squares, ANOVA table, t-test, and F-tests, confidence intervals for the intercept and slope; Correlation analysis, coefficient of correlation.

Prescribed textbook

1. Statistics. Mclave, J.T.and Dietrich, F.H. (1979). Delloen Publishing Company.
2. Introduction to Statistics; (2016)., David Lane, Rice University, Publisher: Saylor Foundation
3. Introductory Statistics: (2017); Douglas S. Shafer, University of North Carolina. ISBN 13: 9781453344873

Recommended text book

1. Introduction to statistics. Walpole, R.E. 1984, Macmillan.

11.0 Answers to Unit Activities

1.1 Unit Activities

1. (a) discrete (b) cont. (c) discrete (d) cont. (d) cont.

1.2 Unit Activities

1. 4.47; 4; 1;2; 4; median

2. (b) P 65.9 , q 71.3

3. (b)4.475, 4.525,4.500,4.458,,4.450, 4.550, 4.525, 4.242, 4.117,, 3.958, 3.942, 3.950,3.892, 3.867, 3.883, 3.767, 3.742, 3.650,4.025, 3.975, 3.958, 3.992,3.917,3.933,4.017 (h)

4. (a)k62, (b) k 74 72%(taking upper limits as; k35.885, k39.995, et) true class limits; 0, 34.995, 35.995, 39.995, etc . Corresponding frequencies; 0.1,0.2,0.4, 0.6, 0.7, 0.7, 1.3, 0.9, 0.8, 0.4. mean = 68

2.1 Unit Activities

1. 2, 1

$\overline{X}$	1	5/3	7/3	3
$p(\overline{X})$	1/8	3/8	3/8	1/8

2.

2. (a) $\mu = 0$ (b) $\sigma^2 = 1.2$

2.2 Unit Activities

1. $\frac{r_1}{n_1}$, $\sqrt{\left\{\frac{p(1-p)}{n_1}\right\}}$; $\frac{1}{2}\left[p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right]^{\frac{1}{2}}$; $\frac{1}{3} < \frac{n_1}{n_2} < 3$

3. $\hat{\mu}_1$ is more efficient; $\text{var}(\hat{\mu}_1)$; $\text{var}(\hat{\mu}_2) = 3$; 5

5. k = 0.5

2.3 Unit Activities

1. (a) 0.46.2 (b) (i) 0. 3821 (ii) 0.1587

2. 9a) (ii) (64.4 – 67.6)kg (b) (i) K773 – 817 (c) 51.48 – 51.72

3. N(54,64), 0.0228, c = 16 ; N(5,144), 0.662

2.4 Unit Activities

1. (a) 4, 14.4

(b) (i) $\bar{x}$ $6\frac{2}{3}$, $5\frac{1}{3}$, $4\frac{2}{3}$, $3\frac{1}{2}$, 2, $1\frac{1}{3}$

P($\bar{x}$) 0.1 0.2 0.2 0.3 0.1 0.1

(iii) 2.4

2. (a) 2μ, $\sigma\sqrt{2}$ (b) 0 , $\sigma\sqrt{2}$ (c) μ, $\sigma/\sqrt{2}$; 0.71 , 0.92

3. $v = 1.645/\sqrt{n}$, $n \geq 215$

4. (a) $E(X) = 2\vartheta$, var(x) = ϑ^2/n (b) $N(\mu, 1/n)$; $n \geq 97$

5. 7, 5.83; $N(7n, 5.83n)$

6. $\frac{n_1}{n_2}$

7. $P(R = r) = \binom{r-1}{a-1} p^a (1-p)^{r-a}$, $r \geq a$

3.1 Unit Activities

1. (a) (3.114 -3.166)m (b) (3.109 – 3.171)m

2. (15.03 – 15.39)g

3.2 Unit Activities

1. (a) (15.7 – 18.68)h (b) (15.44 – 18.96)h

2. 47.86 – 48.14

3. 105.5h, 78.1h ; (87.7 – 125.3)h

4. 11.41mm, 0.120mm; (11.39 – 11.43)mm

3.3 Unit Activities

1. (0.9714 -0.9886)cm

2. (23.1 – 25.3)g

3. N(51.81, 54.18)g

4. (6.754 – 8.182)m

3.4 Unit Activities

1. 0.34 – 0.62

2. (a) 0.023 – 0.227, (b) 27 – 27.3

3. (a) 34% (b) 29.4% - 38.6% (c) 2155

4. 0.162 s.e = 0.024; 0.114 – 0.210; 1400; 0.054 – 0.112

3.5 Unit Activities

1. (a) 0.866 (b) (i) 2.00mg (ii) -0.01 to 1.01mg

2. (a) (50.0 – 54.6)kg (b) (1.80 – 2. 44)kg

3. (a) (0.252 – 0.262)kg (b) 246

4. (a) 2.04 – 3.96 (b) 25.2% - 34.8% (c) retired unemployed

5. (a) $c = \bar{x}$ (b) (i) 456, 108 (ii) 449.4-462.6 (iii) 457, 89

6. 10000, 7240 – 16200

7. (b) (7.3 – 9.3)%

8. (a) 26,500 (b) 1068

9. Yes

10. 4, 2.47 – 4.96; 4.1, 3.90 – 4.10

4.1 Unit Activities

1. $H_o : p = \frac{1}{4}$, $H_1 : p > \frac{1}{4}$

2. 0.0766

3. 0.1178

4. 0.1348

4.2 Unit Activities

1. z = 1.71

2. z = 2.83

3. 3mm, 0.02m; z =1.8

4. z= 1.79

5. mean = 4.6, z = 7.7

6. z=-1.50

7. z = 3

4.3 Unit Activities

1. 26.75yr 42.22yr2; z = 1.85

2. z = 4.44

3. z = 5.3

4. z = -1.47

5. z = -2.2

6. z= -4.68

4.4 Unit Activities

1. $t_6 = -0.298$

2. 8.6, $t_5 = 3.54$

3. $t_4 = 2.21$

4. $t_7 = 1.67$

4.5 Unit Activities

1. z = 4.8; 65.13 – 67.33 68.83 – 70.65

2. z = 1.44

3. z = 2.86

4. z = 4.6

5. z = -2.15

4.6 Unit Activities

1. 0.34 ; $t_9 = 1.16$

2. 0.0005; yes; $t_{24} = 5.17$

4.7 Unit Activities

1. no, z = 1.40

2. yes, z = 2.08

3. yes ,z = 2.59

4. yes , z = 4.06

4.8 Unit Activities

1. yes ; z = 2.57

2. yes ; z = 2.99

4.9 Unit Activities

1. 0.082

2. z = -2.16

4.10 Unit Activities

1. r = 3 to 9 ; 0.182

2. 21.81 – 26.19 ; 0.05 ;0.855

3. n = 35 , v = 3.028

4. (a) 2%, 0.377

(b) $\bar{x} > 54.66$, (i) 0.749 (ii) 0.0918

(a) 1 in 40 line between 5 and 6 defectives, 1 in 100 line between 8 and 9 defectives

(b) (i) 0.364 , (ii) 0.0681 (iii) 0.754

3. (a) $t_9 = 1.96$, (b) 12.9 ± 13.04

3. 0.0064

4. -1.01 -0.21

5. $t_9 = 2.33$

6. yes, z = 4; 0.2 – 0.6

7 $a = \frac{1}{3}; \frac{4}{9}$

8. 0.109, ; yes, z = 1.83

9. z = 4

10. yes; $t_{12} = 0.429$

11. 0.30 < p< 0.38; 0.309; z = =3.03

12. 11.642 , 1.892; no $z \approx t_{99} = 1.88$; $12 + 4.66/\sqrt{n}$

13. (a) yes, z= 2.4 (b) yes ; z = 2.95

14. yes, $t_{120} = 2.1$

15. 3.62, 0.29; $t_9 = 4.5$

5.1 Unit Activities

1. yes, $X^2 = 20.2, 9df$

2. 12,24,36,48,60,72,60,48,36,24,12; A, $X^2 = 14.5$, B , $X^2 = 2.12$

5.2 Unit Activities

1. reject null hypothesis, $X^2 = 9.6,\ 2df$

2. no, $X^2 = 0.23,\ 1df$

3 retain null hypothesis , $X^2 = 3.2,\ 7df$

4. 3; 2.5, 7.5, 11.2, 11.2, 8.4, 5.0, 2.5, 1.1, 0.6,; $X^2 = 0.28, 3df$

5.3 Unit Activities

1. no , $X^2 = 2.8, 1df$

2. 21.0 10.0 7.0

   15.5 7.4 5.2

   41.5 19.7 13.8

   No evidence, $X^2 = 7.9, 4df$

6.1 Unit Activities

1. (a) -0.98 (b) -0.03

2. -0.91, -0.98

6.2 Unit Activities

1. (a) r = 0.88 (b) C,I,E,D,B,G,J,A,F,H

2. (a) r = 0.66 , (b) r = 0.54

3. r = 0.28

4. (a) r = 0.75

6.3 Unit Activities

1. r = -0.86

2. (a) invalid (b) invalid (c) valid (d) invalid

3. (b) r = 0.87

5.(a) -0.37 (b) -0.27 (c) no evidence

6. r = 0.97

7. r = 0.78

7.0 Unit Activities

3 SSE =0.020; F = 2.0; do not reject the null hypothesis

4 (a) $2.20 < \mu_A < 2.30$

(b) $-0.01 < (\mu_A - \mu_B < 0.18$

5 (c) SSE = 498.67 , yes

9 (a) no, it is completely RD (b) CRD with n treatment and two treatment

10(c) SSE = 4.99 (d) no, no

11(e) RBD, (f) 21 (g) SSE = 58.91; F = 3.07 ; YES; (h) no

12(a) SSE = 0.571; SST = 1.212;F = 11.7; YES; (b) t =--2.73 ; v = 7; yes

13

source	df	SS	MS	F
BLOCKS	3	0.140	0.047	6.62
TREATMENT	4	0.787	0.197	27.7
ERRORS	12	0.085	0.0071	
TOTAL	19	1.012		

F = 27.7 , REJECT THE NULL HYPOTHESIS

14

source	df	SS	MS	F
TREATMENT	2	38.00	19.00	10.05
BLOCKS	3	61.67	20.56	10.88
ERRORS	6	11.33	1.89	
TOTAL	11			

i. yes; F = 10.05

ii. yes ; F = 10.88

iii. 3.50 ± 1.89

15

source	df	SS	MS	F
ROWS	2	46.89	23.45	11.11
COLUMNS	2	22.89	11.45	5.43
TREATMENTS	2	91.56	45.78	21.70
ERRORS	2	4.22	2.11	
TOTAL	8	165.56		

7.33 ± 5.11

18. F= 1.535 no evidence of a difference between treatment

8.1 Unit Activities

1. (a) $y = 0.486x - 2.397$ (b) 0.997 (c) no

2. y = 2.79x – 5451 ; $53.7x10^6$

3. (b) y = 2.75x +48.35 (c) 0.79

4. y = 18.9x + 154 ; 476g

8.2 Unit Activities

1. (a) y = -4x +24 (b) 8, 0.395 (c) 7.4 – 8.6

2. k = 0.9948, c = -1.998; 0.92 – 1.07mm

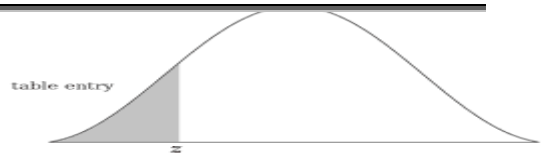
3. (a) $\hat{a} = 0.06$, $\hat{b} = 0.132$ (b) 1.344 – 1.416; 22.1 – 26.1

4. y = -x + 12 ; 5.67 – 6.33; 0 ,no linear

5. (a) 18.8 – 27.2

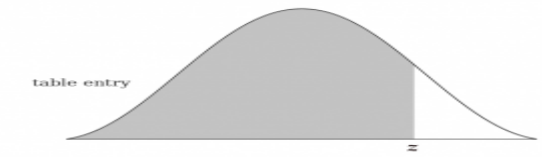
(b) (-0.428) – (-0.222)

12.1 Table A1 Z-Table



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
-0.1	.46017	.45620	.45224	.44828	.44433	.44034	.43640	.43251	.42858	.42465
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-1	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08692	.08534	.08379	.08226
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-2	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-3	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-4	.00003	.00003	.00003	.00003	.00003	.00003	.00002	.00002	.00002	.00002

1.1 – Negative Z Table



Use the positive Z score table below to find values on the right of the mean as can be seen in the graph alongside. Corresponding values which are greater than the mean are marked with a positive score in the z-table and represent the area under the bell curve to the left of z.

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
+0.1	.53983	.54380	.54776	.55172	.55567	.55966	.56360	.56749	.57142	.57535
+0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
+0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
+0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
+0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
+0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
+0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
+0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
+0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
+1	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
+1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
+1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
+1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91308	.91466	.91621	.91774
+1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
+1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
+1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
+1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
+1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
+1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
+2	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
+2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
+2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
+2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
+2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
+2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
+2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
+2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
+2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
+2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
+3	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
+3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
+3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
+3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
+3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
+3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
+3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
+3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
+3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
+3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997
+4	.99997	.99997	.99997	.99997	.99997	.99997	.99998	.99998	.99998	.99998

1.2 – Positive Z Table

12.2 Table A3 Random Numbers

1000 Random Numbers

46 65137 89808 87910 21015 08217 80808 79191 79323 61832 59344 50445 69982
97 77449 73675 73906 50231 58945 13486 26566 00967 05591 01990 00727 84468
44 78926 96293 24899 40709 57141 10013 63645 09751 19181 41396 41981 82285
25 87148 04928 99092 17839 26631 98905 04880 62315 32680 87327 44912 51325
32 44692 71845 79687 32010 16245 33677 89725 40301 01165 74548 05222 04476
56 19137 94311 74888 58615 99810 92844 82632 50348 04714 62825 43615 79631
37 03061 94590 17898 98043 76155 79923 73055 24097 48407 36923 05793 39680
16 14042 19949 66919 75685 69107 23429 75448 18190 62320 06608 62459 72513
32 94658 42653 05156 41932 27496 76878 29262 14266 01533 06647 09317 91022
28 12487 59967 40811 72174 25148 18785 98952 18967 46122 96803 67069 33096
94 60107 94157 07087 66803 60892 52665 73970 63932 01168 61844 81420 61626
19 55096 92490 62188 51207 58556 07036 68769 05803 96746 11534 23072 64503
46 40033 27229 89316 75780 66969 35411 22804 26221 87936 80640 87728 54618
14 82496 95642 96675 15909 59958 16307 00525 31824 54816 28926 96091 69860
54 75062 19994 17468 71988 36352 53240 91388 54256 02463 41159 93864 73122
36 35310 47550 09898 69405 19978 74724 09845 50155 60957 27105 51751 52293
73 05819 50576 57168 58553 59665 34276 95349 01163 85374 21528 65455 20264
13 54165 26337 67481 14820 29241 10412 12353 13920 55794 84194 55678 14926
97 72928 97100 74040 38599 27534 06646 50310 20067 77294 99870 68124 99169
57 79685 10130 35049 55279 78372 51494 23674 76245 79953 78320 73497 60884
52 17014 37692 33711 15252 30584 98473 71898 64094 18848 54310 67498 12962
25 52495 77052 78228 93907 57997 26098 59434 90254 74058 87606 70832 93293
94 78704 50476 01804 91636 59439 37884 75736 86367 50661 62974 29201 64767
11 13220 05229 62642 33829 58739 62347 35258 09073 78734 90179 13309 15276
16 98620 13577 43945 70877 04725 53003 77451 40013 13357 03987 18858 57710
50 02484 72786 53917 78170 71225 88366 66770 85826 34607 64163 66552 29960
54 17390 05121 25759 56099 57067 71954 92666 78155 47562 55473 02094 94916
58 60914 57848 17676 30054 86079 32281 69595 73639 82301 87683 96187 40469
32 75636 03232 12871 50427 07089 22864 37008 53920 93901 95558 44534 79509
38 29134 38539 19547 92608 93926 22217 59359 25637 62441 38864 67839 49928
77 13986 86035 70906 56987 45417 33856 26100 20432 92998 02367 11804 14615
38 47769 92986 25152 68955 85391 86298 51686 94944 15652 95832 47132 61462
12 73063 31656 65488 78903 62676 33057 75569 30815 85480 54024 69527 03359
40 57444 99312 00118 62222 51099 52649 62202 34686 82560 36710 99692 78197
53 93492 18220 62244 19301 49490 48591 01038 38196 11970 28655 83330 48866
21 98784 35644 10069 75532 81676 49396 06531 97987 03789 39625 91143 03256
55 15899 23850 26797 81306 96485 08582 18134 95699 13094 04385 18836 54260
05 12414 80105 20631 12012 67745 04664 00299 37973 19964 82701 87112 39766
13 80909 26673 24773 68041 94286 41157 42480 56449 29477 59338 41672 22667
18 38960 10821 97428 74785 99933 46577 59130 07768 71711 44063 53221 90851
51 29146 04814 21797 82633 33517 74440 89702 73571 01743 71640 17462 57211
34 19045 05917 06085 66977 68864 03260 98650 60498 99653 01836 55097 59637
26 01235 36397 83494 28207 04804 36555 75426 00190 70519 19415 11831 61018
28 93295 43518 12227 91713 65163 02840 34483 49288 44418 34145 58365 49247

31 37802 99547 75489 07887 98771 55629 15275 89432 76428 80789 90404 37213
76 02335 01749 81480 52119 79456 73439 82255 27262 46628 19115 28650 62808
02 79472 72390 63406 24799 06045 59786 24875 94420 60096 33887 18806 88921
51 86791 58735 79469 51318 02366 43184 57590 47546 46910 34871 31576 03444
54 66869 33050 12871 82067 17249 77150 19851 68280 62148 56833 62662 19266
31 87286 92600 71817 92436 24520 40200 76113 63264 85888 59539 32768 74280
42 40619 59652 89510 82816 23008 35848 15829 64138 23208 27753 87598 31374
38 44446 81471 24151 78598 36542 92611 53450 27542 04187 71243 97853 78441
50 41346 30322 58946 30172 13949 64004 28432 35117 83902 24772 59237 28370
09 98897 03471 02096 12929 29057 28542 05226 93504 42433 48107 92454 19052
45 09676 23184 86805 27259 80695 89742 73286 34344 84856 26014 43206 13379
38 91262 36726 71276 98554 92690 76119 47065 14276 41249 68258 82196 99239
20 86233 16363 38711 77203 13871 41189 66016 14942 66692 09605 05128 89525
30 62168 59361 97316 38597 18066 63467 94592 42983 41263 09820 32705 22126
11 11643 87985 80291 83128 04103 46469 22247 30038 70038 53670 10630 59934
01 52238 49501 45841 36185 95811 18710 63433 32749 83097 50919 09606 90838
43 26531 16174 77169 48160 42016 33707 57604 32756 35518 36333 25336 27730
75 02099 85271 82477 94442 16548 69974 69213 36700 07378 69678 78523 08500
40 71521 79056 94969 75423

12.3 Table A4 T-Distribution Table (One Tail and Two-Tails)

df	a = 0.1	0.05	0.025	0.01	0.005	0.001	0.0005

∞	t_a = 1.282	1.645	1.960	2.326	2.576	3.091	3.291
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	1.886	2.920	4.303	6.965	9.925	22.328	31.600
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883

12.4 Table A6 Upper-tail critical values of chi-square distribution with *v* degrees of freedom

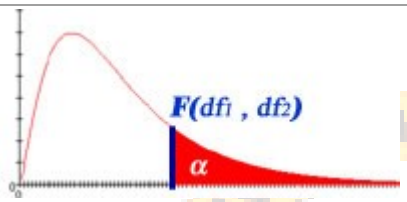
	Probability less than the critical value				
v	0.90	0.95	0.975	0.99	0.999
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266
4	7.779	9.488	11.143	13.277	18.467
5	9.236	11.070	12.833	15.086	20.515
6	10.645	12.592	14.449	16.812	22.458
7	12.017	14.067	16.013	18.475	24.322
8	13.362	15.507	17.535	20.090	26.125
9	14.684	16.919	19.023	21.666	27.877
10	15.987	18.307	20.483	23.209	29.588
11	17.275	19.675	21.920	24.725	31.264

12	18.549	21.026	23.337	26.217	32.910
13	19.812	22.362	24.736	27.688	34.528
14	21.064	23.685	26.119	29.141	36.123
15	22.307	24.996	27.488	30.578	37.697
16	23.542	26.296	28.845	32.000	39.252
17	24.769	27.587	30.191	33.409	40.790
18	25.989	28.869	31.526	34.805	42.312
19	27.204	30.144	32.852	36.191	43.820
20	28.412	31.410	34.170	37.566	45.315
21	29.615	32.671	35.479	38.932	46.797
22	30.813	33.924	36.781	40.289	48.268
23	32.007	35.172	38.076	41.638	49.728
24	33.196	36.415	39.364	42.980	51.179
25	34.382	37.652	40.646	44.314	52.620
26	35.563	38.885	41.923	45.642	54.052
27	36.741	40.113	43.195	46.963	55.476
28	37.916	41.337	44.461	48.278	56.892
29	39.087	42.557	45.722	49.588	58.301
30	40.256	43.773	46.979	50.892	59.703
31	41.422	44.985	48.232	52.191	61.098
32	42.585	46.194	49.480	53.486	62.487
33	43.745	47.400	50.725	54.776	63.870
34	44.903	48.602	51.966	56.061	65.247
35	46.059	49.802	53.203	57.342	66.619
36	47.212	50.998	54.437	58.619	67.985
37	48.363	52.192	55.668	59.893	69.347
38	49.513	53.384	56.896	61.162	70.703
39	50.660	54.572	58.120	62.428	72.055
40	51.805	55.758	59.342	63.691	73.402
41	52.949	56.942	60.561	64.950	74.745
42	54.090	58.124	61.777	66.206	76.084
43	55.230	59.304	62.990	67.459	77.419
44	56.369	60.481	64.201	68.710	78.750
45	57.505	61.656	65.410	69.957	80.077
46	58.641	62.830	66.617	71.201	81.400
47	59.774	64.001	67.821	72.443	82.720
48	60.907	65.171	69.023	73.683	84.037
49	62.038	66.339	70.222	74.919	85.351
50	63.167	67.505	71.420	76.154	86.661
51	64.295	68.669	72.616	77.386	87.968
52	65.422	69.832	73.810	78.616	89.272
53	66.548	70.993	75.002	79.843	90.573
54	67.673	72.153	76.192	81.069	91.872
55	68.796	73.311	77.380	82.292	93.168

56	69.919	74.468	78.567	83.513	94.461
57	71.040	75.624	79.752	84.733	95.751
58	72.160	76.778	80.936	85.950	97.039
59	73.279	77.931	82.117	87.166	98.324
60	74.397	79.082	83.298	88.379	99.607
61	75.514	80.232	84.476	89.591	100.888
62	76.630	81.381	85.654	90.802	102.166
63	77.745	82.529	86.830	92.010	103.442
64	78.860	83.675	88.004	93.217	104.716
65	79.973	84.821	89.177	94.422	105.988
66	81.085	85.965	90.349	95.626	107.258
67	82.197	87.108	91.519	96.828	108.526
68	83.308	88.250	92.689	98.028	109.791
69	84.418	89.391	93.856	99.228	111.055
70	85.527	90.531	95.023	100.425	112.317
71	86.635	91.670	96.189	101.621	113.577
72	87.743	92.808	97.353	102.816	114.835
73	88.850	93.945	98.516	104.010	116.092
74	89.956	95.081	99.678	105.202	117.346
75	91.061	96.217	100.839	106.393	118.599
76	92.166	97.351	101.999	107.583	119.850
77	93.270	98.484	103.158	108.771	121.100
78	94.374	99.617	104.316	109.958	122.348
79	95.476	100.749	105.473	111.144	123.594
80	96.578	101.879	106.629	112.329	124.839
81	97.680	103.010	107.783	113.512	126.083
82	98.780	104.139	108.937	114.695	127.324
83	99.880	105.267	110.090	115.876	128.565
84	100.980	106.395	111.242	117.057	129.804
85	102.079	107.522	112.393	118.236	131.041
86	103.177	108.648	113.544	119.414	132.277
87	104.275	109.773	114.693	120.591	133.512
88	105.372	110.898	115.841	121.767	134.746
89	106.469	112.022	116.989	122.942	135.978
90	107.565	113.145	118.136	124.116	137.208

Table A7 F Distribution Tables

F Table for $\alpha = 0.05$



/	df_1=1	2	3	4	5	6	7	8	9	10	
df_2= 1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817	24
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959	1
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644	
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600	
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022	

15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437

The F distribution is a right-skewed distribution used most commonly in Analysis of Variance. When referencing the F distribution, the **numerator degrees of freedom are always given first**, as switching the order of degrees of freedom changes the distribution (e.g., $F_{(10,12)}$ does not equal $F_{(12,10)}$). For the four F tables below, the rows represent denominator degrees of freedom and the columns represent numerator degrees of freedom. The right tail area is given in the name of the table. For example, to determine the .05 critical value for an F distribution with 10 and 12 degrees of freedom, look in the 10 column (numerator) and 12 row (denominator) of the F Table for alpha=.05. $F_{(.05,\ 10,\ 12)} = 2.7534$. You can use the Java Applet or the HTML5/JavaScript Webapp interactive F-Distribution calculators to obtain more accurate measures of probability or critical values.