

Statistical Computing

STA 231

UNIVERSITY OF BRADY LIBRARY

Copyright © 2002, Revised in 2016 by Distance Learning Centre, University of Ibadan, Ibadan.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval System, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN 978-021

General Editor: Prof. Bayo Okunade

Error! Use the Home tab to apply Guide Publishing Institute to the text that you want to appear here.

University of Ibadan,
Nigeria

Telex: 31128NG

Tel: +234 (80775935727)

E-mail: ssu@dlc.ui.edu.ng

Website: www.dlc.ui.edu.ng

Vice-Chancellor's Message

The Distance Learning Centre is building on a solid tradition of over two decades of service in the provision of External Studies Programme and now Distance Learning Education in Nigeria and beyond. The Distance Learning mode to which we are committed is providing access to many deserving Nigerians in having access to higher education especially those who by the nature of their engagement do not have the luxury of full time education. Recently, it is contributing in no small measure to providing places for teeming Nigerian youths who for one reason or the other could not get admission into the conventional universities.

These course materials have been written by writers specially trained in ODL course delivery. The writers have made great efforts to provide up to date information, knowledge and skills in the different disciplines and ensure that the materials are user-friendly.

In addition to provision of course materials in print and e-format, a lot of Information Technology input has also gone into the deployment of course materials. Most of them can be downloaded from the DLC website and are available in audio format which you can also download into your mobile phones, IPod, MP3 among other devices to allow you listen to the audio study sessions. Some of the study session materials have been scripted and are being broadcast on the university's Diamond Radio FM 101.1, while others have been delivered and captured in audio-visual format in a classroom environment for use by our students. Detailed information on availability and access is available on the website. We will continue in our efforts to provide and review course materials for our courses.

However, for you to take advantage of these formats, you will need to improve on your I.T. skills and develop requisite distance learning Culture. It is well known that, for efficient and effective provision of Distance learning education, availability of appropriate and relevant course materials is a *sine qua non*. So also, is the availability of multiple plat form for the convenience of our students. It is in fulfilment of this, that series of course materials are being written to enable our students study at their own pace and convenience.

It is our hope that you will put these course materials to the best use.



Prof. Abel Idowu Olayinka
Vice-Chancellor

Foreword

As part of its vision of providing education for “Liberty and Development” for Nigerians and the International Community, the University of Ibadan, Distance Learning Centre has recently embarked on a vigorous repositioning agenda which aimed at embracing a holistic and all encompassing approach to the delivery of its Open Distance Learning (ODL) programmes. Thus we are committed to global best practices in distance learning provision. Apart from providing an efficient administrative and academic support for our students, we are committed to providing educational resource materials for the use of our students. We are convinced that, without an up-to-date, learner-friendly and distance learning compliant course materials, there cannot be any basis to lay claim to being a provider of distance learning education. Indeed, availability of appropriate course materials in multiple formats is the hub of any distance learning provision worldwide.

In view of the above, we are vigorously pursuing as a matter of priority, the provision of credible, learner-friendly and interactive course materials for all our courses. We commissioned the authoring of, and review of course materials to teams of experts and their outputs were subjected to rigorous peer review to ensure standard. The approach not only emphasizes cognitive knowledge, but also skills and humane values which are at the core of education, even in an ICT age.

The development of the materials which is on-going also had input from experienced editors and illustrators who have ensured that they are accurate, current and learner-friendly. They are specially written with distance learners in mind. This is very important because, distance learning involves non-residential students who can often feel isolated from the community of learners.

It is important to note that, for a distance learner to excel there is the need to source and read relevant materials apart from this course material. Therefore, adequate supplementary reading materials as well as other information sources are suggested in the course materials.

Apart from the responsibility for you to read this course material with others, you are also advised to seek assistance from your course facilitators especially academic advisors during your study even before the interactive session which is by design for revision. Your academic advisors will assist you using convenient technology including Google Hang Out, You Tube, Talk Fusion, etc. but you have to take advantage of these. It is also going to be of immense advantage if you complete assignments as at when due so as to have necessary feedbacks as a guide.

The implication of the above is that, a distance learner has a responsibility to develop requisite distance learning culture which includes diligent and disciplined self-study, seeking available administrative and academic support and acquisition of basic information technology skills. This is why you are encouraged to develop your computer

skills by availing yourself the opportunity of training that the Centre's provide and put these into use.

In conclusion, it is envisaged that the course materials would also be useful for the regular students of tertiary institutions in Nigeria who are faced with a dearth of high quality textbooks. We are therefore, delighted to present these titles to both our distance learning students and the university's regular students. We are confident that the materials will be an invaluable resource to all.

We would like to thank all our authors, reviewers and production staff for the high quality of work.

Best wishes.



Professor Bayo Okunade

Director

UNIVERSITY OF IBADAN LIBRARY

Course Development Team

Content Authoring	Udomboso, Christopher Godwin
Content Editor	Prof. Remi Raji-Oyelade
Production Editor	Ogundele Olumuyiwa Caleb
Learning Design/Assessment Authoring	Folajimi Olambo Fakoya
Managing Editor	Ogunmefun Oladele Abiodun
General Editor	Prof. Bayo Okunade

UNIVERSITY OF IBADAN LIBRARY

Table of Contents

Study Session 1: General Computing	12
Introduction.....	12
Learning Outcomes from Study Session 1.....	12
1.1 Brief History of the Computer	12
In-Text Question	14
In-Text Answer	14
1.1.1 Definition of Computer.....	14
1.1.2 Generations of Computers	15
1.2 Types of Computers.....	16
In-Text Question	18
In-Text Answer	18
1.2.1 Components of the Computer	19
In-Text Question	20
In-Text Answer	20
Input	21
1.2.2 Introduction to Operating System.....	22
1.3 Implications of Computerization	22
Summary	24
Self-Assessment Questions (SAQs) for study session 1	24
Study Session 2: Computer in Statistics Profession	26
Introduction.....	26
Learning Outcomes from Study Session 2.....	26
2.1 The Growth of Statistics and Computational Devices	26
In-Text Question	29
In-Text Answer	29
2.2 Programing Techniques for Statistics	29
Summary for 2	30
Self-Assessment Questions (SAQs) for study session 2.....	31
Study Session 3: Representation of Data in the Computer	32
Introduction.....	32
Learning Outcomes from Study Session 3.....	32

3.1 Binary Representation.....	32
3.1.1 Converting from Decimal to Binary	34
3.1.2 Converting from Binary to Decimal	34
3.2 Simple Computer Arithmetic.....	35
Summary for 3	40
Self-Assessment Questions (SAQs) for study session 3.....	40
Study Session 4: Introduction to Microsoft Excel	42
Introduction.....	42
Learning Outcomes from Study Session 4.....	42
4.1 Spreadsheet	42
4.1.1 Simple Operations.....	43
In-Text Question	43
In-Text Answer	43
4.1.2 Selecting a Cell or Multiple Cells.....	44
4.2 Types of Sheet in MS Excel.....	44
In-Text Question	45
In-Text Answer	45
Summary for 4	46
Self-Assessment Questions (SAQs) for study session 4.....	46
Study Session 5: Data Entry in Microsoft Excel	48
Introduction.....	48
Learning Outcomes from Study Session 5.....	48
5.1 Types of Data.....	48
In-Text Question	50
In-Text Answer	50
5.2 Data Entry Techniques.....	50
5.2.1 Using Autofill	51
In-Text Question	52
In-Text Answer	52
5.3 Creating Simple Trends and Forecasts	53
Summary for 5	54
Self-Assessment Questions (SAQs) for study session 5.....	54
Study Session 6: Data Analysis using Microsoft Excel.....	56
Introduction.....	56

Learning Outcomes from Study Session 6.....	56
6.1 General Rule on Computation in MS Excel	56
In-Text Question	58
In-Text Answer	58
Summary for 6	59
Self-Assessment Questions (SAQs) for study session 6.....	60
Study Session 7: Using the Function and Chart Wizards	62
Introduction.....	62
Learning Outcomes from Study Session 7.....	62
7.1 Using Functions	62
In-Text Question	64
In-Text Answer	64
7.1.1 Statistical Function.....	67
In-Text Question	69
In-Text Answer	69
7.1.2 Using Charts.....	70
Summary for 7	71
Self-Assessment Questions (SAQs) for study session 7.....	71
Study Session 8: Algorithm and Flow Chart	74
Introduction.....	74
Learning Outcomes from Study Session 8.....	74
8.1 Algorithm.....	75
In-Text Question	76
In-Text Answer	76
8.2 Flowchart	76
8.2.1 Flowchart Symbols	77
Summary for 8	80
Self-Assessment Questions (SAQs) for study session 8.....	80
Study Session 9: Review of the BASIC Programming Language.....	82
Introduction.....	82
Learning Outcomes from Study Session 9.....	82
9.1 The BASIC Program.....	83
Summary for 9	90
Self-Assessment Questions (SAQs) for study session 9.....	90

Study Session 10: Descriptive Statistics	92
Introduction.....	92
Learning Outcomes from Study Session 10.....	92
10.1 Rounding of Numerical Data	93
10.1.1 Error	95
In-Text Question	95
In-Text Answer	96
10.1.2 Ratios and Percentages	97
10.1.3 The Median	104
In-Text Question	106
In-Text Answer	106
Summary for 10	117
Self-Assessment Questions (SAQs) for study session 10.....	117
Study Session 11: Probability Theory.....	120
Introduction.....	120
Learning Outcomes from Study Session 11.....	120
11.1 Set Theory.....	121
In-Text Question	122
In-Text Answer	122
11.2 Mutually Exclusive Events	130
Summary for 11	136
Self-Assessment Questions (SAQs) for study session 11	136
Study Session 12: Probability Distribution Functions	138
Introduction.....	138
Learning Outcomes from Study Session 12.....	138
12.1 Probability Function (The Discreet Case).....	139
Summary for 12	156
Self-Assessment Questions (SAQs) for study session 12.....	156
Study Session 13: Correlation and Linear Regression.....	159
Introduction.....	159
Learning Outcomes from Study Session 13.....	159
13.1 The Theory of Correlation	160
13.2 Coefficient of Correlation	161
In-Text Question	165

In-Text Answer	165
Summary for 13	169
Self-Assessment Questions (SAQs) for study session 13	169
Study Session 14: Elementary Time Series Analysis	171
Introduction.....	171
Learning Outcomes from Study Session 14.....	171
14.1 Components of a Time Series	171
In-Text Question	172
In-Text Answer	172
14.1.1 Models of Time Series	172
Summary of 14.....	176
Self-Assessment Questions (SAQs) for study session 14.....	177
Study Session 15: Statistical Tests and Confidence Intervals	180
Introduction.....	180
Learning Outcomes from Study Session 15.....	180
15.1 Point and Interval Estimates	181
15.2 Hypothesis Testing.....	190
Summary of 15.....	192
Self-Assessment Questions (SAQs) for study session 15.....	192
Study Session 16: Introduction to MATLAB	194
Introduction.....	194
Learning Outcomes from Study Session 16.....	194
16.1 MATLAB.....	195
Summary of 16.....	217
Self-Assessment Questions (SAQs) for study session 16.....	217

Study Session 1: General Computing

Introduction

The Computer has become an indispensable part of the human life which has literally phased out many jobs which were hitherto done by several men. The relevance of computer in almost every human activity cannot be overemphasized.

It is also important to know how the computer came to be, the generations, types, configuration, and the likes. The statistics profession has become, in the last 50 to 60 years, a major player in the computing world.

Learning Outcomes from Study Session 1

At the end of this study session, you should be able to:

- 1.1 Concept of a modern computer;
- 1.2 Types of computers;
- 1.3 Identify the implications of computers in every profession.

1.1 Brief History of the Computer

Data and information are so important to human beings because they are needed for decision making. The human brain is equipped to file and retrieve data and information. Billions of such activities are carried out every day.

While data are raw facts that are almost useless, it has to be processed before it can make any meaning. These processed data is called information. The brain performs numerous processes everyday.

Due to its limitations (for example, stress due to overload of data and information), it is not sufficient to rely only on it (the brain). This brought about the art of recording. Recording is writing or drawing 'something' on a physical object so as to reduce load on the brain and enhance memory.

Prehistoric cave dwellers painted pictures on walls of caves, while the ancient Egyptians wrote on a crude form of paper called papyrus. The Sumerians, around 3000B.C, developed a device for representing numbers by use of stones in a box, which the Chinese, about 1000 B.C, improved upon by tying stones on strings in a wooden frame.

The device was called Abacus, so named after the Chinese name for box, baccus. Abacus remained a powerful mathematical tool especially for business for several centuries. The Europeans also tried by trying to simplify complex tasks into simple ones.

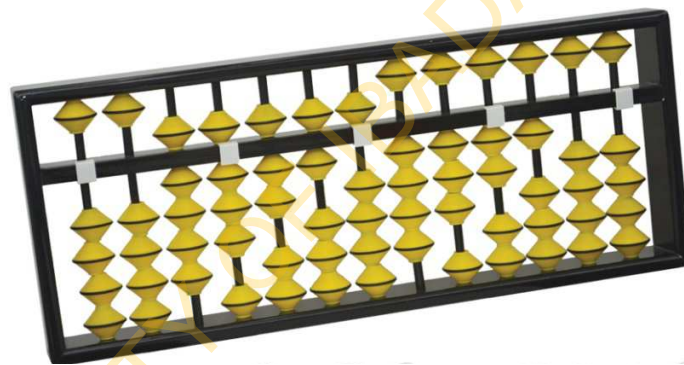


Figure1.1: Abacus Computer.

Source: <http://mmebsabacus.blogspot.com.ng/2015/04/what-exactly-is-abacus.html>

During the industrial revolution in 1804, a Frenchman named Josef Marie Jacquard developed a device that could control the operations of the weaving loom. He used cards with holes punched into them at appropriate location, which could program a textile machine (or weaving loom) to weave specific patterns with specific texture using specific colours. His machine was a mechanical device called Jacquard weaving loom.

His machine had three units. These are input unit, processing unit and output unit. In 1842, an English mathematician named Charles Babbage started work on a calculating device

called Difference Engine, which he did not complete before abandoning it to work on a general – purpose digital calculating machine called Analytical Engine.

In-Text Question

.....remained a powerful mathematical tool especially for business for several centuries.

In-Text Answer

Abacus

He adapted the idea of punched cards to this machine. Working closely with him was a woman named Augusta Ada Byron, nicknamed ‘Lady Lovelace’. She wrote programs that made Babbage’s machine to work. This machine had all the units of the Jacquard weaving loom, i.e. input unit, processing unit and the output unit. The major difference was that Babbage’s machine used electricity, thus making it an electronic device.

This machine marked the beginning of modern electronic computer, and is therefore called the ‘father of modern computer’. At this point we can now define a computer.

1.1.1 Definition of Computer

A computer is an electronic device that accepts data from the input unit, processes the data in the processing unit according to the instruction given, and produces a result in response to format specification through an output unit.



Figure 1.2: Computer

Source: <https://ucpcentralmn.org/computers/>

1.1.2 Generations of Computers

The term 'generation' was applied to different types of computers in order to delineate the major technological developments in hardware and software. So far there are four distinct generations, with the fifth under construction. In brief, they are outlined below:

1. First generation
2. Second generation
3. Third generation
4. Fourth generation
5. Fifth generation

1. First generation (1944 – 1958) – The technology used was vacuum tubes. They used punched cards and magnetic tapes and were slow and large, producing a tremendous amount of heat, and running one program at a time. Examples are ENIAC and UNIVAC1.

2. **Second generation** (1959 – 1963) – The technology used was transistors and some other solid – state devices. These were much smaller than the vacuum tubes, and made computers to be smaller, more reliable and significantly faster.
3. **Third generation** (1964 – 1970) – Now the Integrated Circuit (IC) replaced the transistorized circuitry as the technology improved. The use of magnetic disc became widespread, and computers began to support the capabilities such as multiprogramming and sharing. Operating systems and application software became increased and rapidly produced, and the sizes of computers became much more reduced.
4. **Fourth generation** (1971 – Now) – Improvement in technology caused the replacement of the Integrated circuit by the Large Scale Integrated Circuit (LSIC). The computers of this era had a much larger capacity to support main memory, and the use of keyboard as an input device began to be popular.
5. **Fifth generation** (Now and in the future) – this is still under construction. However what constitutes the fifth generation computers has not been well defined. Nevertheless, everyone agrees that there must be a great improvement over the LSIC technology.

1.2 Types of Computers

Computers can be categorized according to their (a) capabilities and (b) primary functions.

- a. Types of computers according to capability
 - i. **Super computer** – this is about 50,000 times faster than a microcomputer and can handle large amount of scientific computation. It is maintained in a special room or environment.



Figure 1.3: Super Computer

Source: <http://www.networkworld.com/article/2848788/data-center/the-10-mightiest-supercomputers-on-the-planet.html>

- ii. **Mainframe computer** – this can support the processing requirements of hundreds and often thousands of users and computer professionals. It is large and also maintained a controlled environment.
- iii. **Minicomputer** – also called midsize or low-end mainframe computer, its function is similar to the main frame computer, and can support 2 to about 50 users and computer professionals.
- iv. **Microcomputer** – this is the most common computer seen in almost every office, home and everywhere computer is used. It is also known as a personal computer (P.C) and comes in a variety of forms such as diary, notebooks, laptop and desktop computers.



Figure 1.4: Microcomputer

Source: <https://www.quora.com/What-is-Micro-computer-super-computer-mainframe-and-Minicomputer>

In-Text Question

First Generation technology used was

In-Text Answer

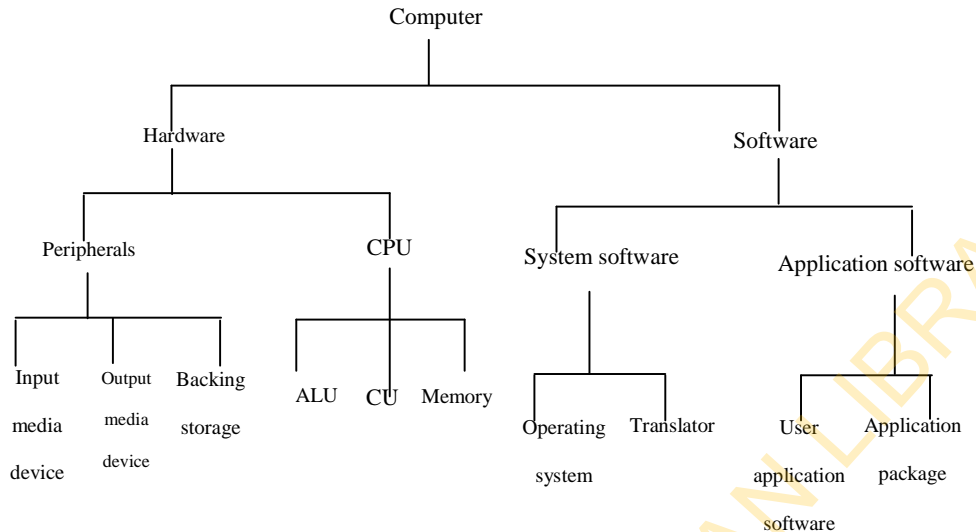
Vacuum tubes.

b. Types of computers according to their primary function

- i. **Digital computers** – these operate on data of discrete forms and perform mathematical computations on them. It is most suitable for business and statistical analyses.
- ii. **Analog computers** – these operate on data in the form of continuous variable quantities. It is most suitable for engineering purposes and other physical sciences.
- iii. **Hybrid computers** – these combine the features of both the digital and analog computers.

1.2.1 Components of the Computer

The computer is made up of the hardware and software components



The computer system is made up of two parts, namely the hardware and the software.

A. The hardware comprises of the

- i. Peripheral devices
- ii. Central Processing Unit (CPU)

1. The peripheral devices are made up of the

- i. Input media devices, like the keyboard, mouse, light pen, scanner, joystick, etc.
- ii. Output media devices, like the monitor, printer, graph plotter, speaker, etc.
- iii. Backing storage which is used in storing data and information.
- iv. The monitor goes by different names such as the visual display unit (VDU), display, cathode ray tube (CRT), and so on.

2. The Central Processing Unit (CPU) also known as the Central Processing Zone (CPZ), is comprised of the Arithmetic/Logical Unit (ALU), the Control Unit (CU) and the memory, also known as the main or primary memory.

The ALU is where the computer processes arithmetic and logical operations. The operators for these operations include:

- i. Arithmetic operators - $+$ $-$ \times \div
- ii. Logical operators - $>$ $<$ \geq \leq

The CU is the heart of the computer. It is where everything that goes on in the computer is controlled. It is also called the supervisor. The memory stores data and information even while the file processing is in progress. There are two main parts of this memory.

These include the Read Only Memory (ROM), and the Random Access Memory (RAM). The ROM is a non-volatile or permanent memory, while the RAM is a volatile or temporary memory. Information stored in the ROM cannot be lost even when the power is turned off. Information in the RAM is lost when power is turned off. In order not to lose information we have to *save*.

In-Text Question

..... comprised of the Arithmetic/Logical Unit (ALU)

In-Text Answer

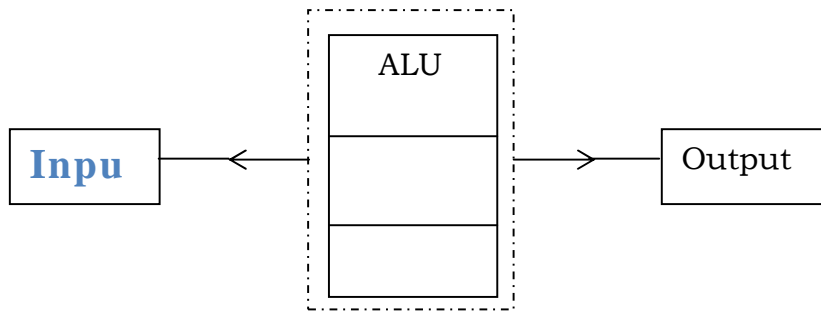
Central Processing Unit (CPU)

The path a data takes through the computer for it to turn to information is:

- Input of data through the input media device
- Processing of data through the CPU
- Output of information through the output media device

This is illustrated below;

The computer is made up of the hardware and software



CPU or CPZ

- B. The software is a set of instructions written for the computer to carry out its task. There are two types of software, namely the system software, and the application software.
1. The system software is the set of instructions written for the computer to manage itself. There are two of them, which are the operating system (OS), and the compiler and interpreter.
 - a. *The operating system* is a set of system program that acts as an interphase between the user and the machine.
 - b. *The translator commonly called compiler* is a set of system program that translates user language to machine readable language. The Interpreter is peculiar to only the BASIC programming language.
 2. The application software is a set of instruction or program written purposely for the user to perform his/her task. This is also of two types, namely the user application software and the application package.
 - a. *The User Application Software*: This is the program written for private use.
 - b. *The Application Package*: This program is designed for public use. For example, MS Word, MS Excel, SPSS, E-views, AutoCAD, Peachtree Accounting, etc.

Application software comes in different forms. These are

- i. Low Level Language e.g. Assembly Language
- ii. High Level Language e.g. BASIC, FORTRAN, COBOL, C++, JAVA

iii. DOS-based Applications e.g. Lotus 1-2-3, WordPerfect, Dbase 3+

iv. Windows-based Application e.g. MS Word, MS Access, Corel Draw.

1.2.2 Introduction to Operating System

The operating system (OS) is system software that acts as an interphase between the machine and the user. It is also called control program or system program. There are two types of OS.

1. Single – Use OS

2. Multi – User OS

A single-user OS accepts commands from one user at a time and perform a single task. Examples include MSDOS, IBMDOS, PCDOS, and all versions of WINDOWS.

A multi-user OS accepts commands from different users working at different local terminals simultaneously. Examples include UNIX, XENIX, LAN, WAN, MOS.

1.3 Implications of Computerization

Computers have essentially revolutionized data and information processing. They have also changed some industries and actually created others. Many people focus on the freedom from routine and boring activities that computers give.

The phasing out of many jobs by the introduction of the computer has necessitated people rushing to train on the use of the computer. Today most employees are computer literate, at least to some extent. In fact, computer literacy is fast becoming a major requirement for employment no matter the certificate or degree one has.

Computer are used in virtually every field of human endeavour such as in business, government, legal profession, medicine, education, industry, entertainment and sports, agriculture, and the home. (Please make a research into how the computer fits in into these fields).

The following are a few examples of jobs that computers create;

- i. Software/Firmware Engineer
- ii. Decision support Analyst
- iii. Programmer Analyst
- iv. MIS Manager
- v. Desktop Publishing/Graphic Artist
- vi. Word processor
- vii. Application Programmer
- viii. Telecommunications Engineer
- ix. Data Processing Specialist
- x. Computer Training Specialist
- xi. Data Processing Position
- xii. Computer Technician/Engineer
- xiii. Data Entry clerk
- xiv. Systems Programmer
- xv. Computer Marketing Professional
- xvi. System Analyst

Summary

In this study session you have learnt about:

- Data and information processing have been of utmost importance to humans long before the advent of computers.
- Through the ages, human beings have made efforts to invent devices that could give them ease in the processing of data, and storage of information. This led to the abacus and other calculating devices.
- The first computer was developed by Charles Babbage in 1842, and Augusta Ada Byron wrote the program that this computer used.
- A computer is composed of the input unit, processing unit and the output unit.
- Since 1944 we have witnessed four distinct generations of computer. Computers are classified according to their capabilities and primary functions, and are made up of the hardware and software.
- The use of computers in society cannot be overemphasized. It has virtually taken over many jobs

Self-Assessment Questions (SAQs) for study session 1

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

SAQ

1. Differentiate among the generations of computers.
2. Differentiate between the operating system and compiler.
3. Write on the effects of computers on the society.

References

Goal 2000 Computer Networks, Module 1: **Using Micro-Computers**. MAGNA Computer School, Ibadan: Nigeria.

Introduction to Micro computing and Operating System - A handbook of Computer Appreciation. MAGNA Computer School, Ibadan: Nigeria.

Study Session 2: Computer in Statistics Profession

Introduction

In the previous study session, you noted that computers have taken over most jobs. There is hardly any profession that has not applied the computer to its job. Statistics is no different.

Due to the enormous data that statisticians, and majority of other profession that uses data and information, encounter, you cannot underestimate the volume of computations being performed every day.

Without the use of computation devices (which include the computer) data processing can become very boring. Even with ordinary calculators it is boring. Thus, the need for statisticians and users of statistical data to be well acquainted with the use of computers cannot be overemphasized.

Learning Outcomes from Study Session 2

At the end of this study session, you should be able to:

- 2.1 The Growth of Computational devices in Statistic
- 2.2 Identify packages that are used to solve statistics problems.

2.1 The Growth of Statistics and Computational Devices

The study of statistics started with an Englishman called John Graunt in the eighteenth century. He was a commoner and a storekeeper, who was interested in reviewing a weekly

church publication issued by the local parish clerk that listed the number of births, christenings, and deaths in each parish.

These so called Bills of Mortality also listed the causes of deaths. He organized this data in the forms we call descriptive statistics, which was published as NATURAL AND POLITICAL OBSERVATION MADE UPON THE BILL OF MORTALITY.

Shortly thereafter, he was elected as a member of the Royal Society. Statistics, which comes from the Italian word standing for state, deals with data collected by a nation just for records purpose. The history of data collection dates back to between 3000B.C. and 4000B.C.

Then there was no interest in this collection more than just to be acquainted with the total population of people or things. This form of data collection is very much common in the bible, especially the Old Testament.



Figure 2.1: John Graunt

Source: https://en.wikipedia.org/wiki/John_Graunt

From the days of Graunt, interest shifted from population to such statistics as the averages, percentages and proportions. This started the development of statistics as a unique area of study and professionalism. Before long it began to find application in many fields much as agriculture and the social sciences, especially economics. Today statistics is a unique feature of the 20th century. In the 19th century, computation depended on calculating machines. The profound statistician, Karl Pearson (the first head of department of statistics in the world, in the University of London), wrote in a letter in 1894: "I want to purchase a Brunsviga calculating machine before anything else, and am making inquiries about it. I think it would make moment – calculating easy". In the early 1900s, routine statistical analysis involved mostly tabulating data and calculating averages and index numbers. As such, the machines of those times were made to add and subtract, print out totals, sub-totals, and individual items if required.

The first modern and automatic electronic computers were introduced in the 1940s, precisely the ENIAC. The advantage of the electronic computer is the speed at which computations are carried out. However, statisticians were slow in integrating the use of computers for their use due to the fact that, then, they were more concerned with much smaller sets of data compared to the large data the computer can handle.

They felt their calculators are enough for the small task. Apart from that, most computer programs then were written by mathematicians and could not do exactly what the statistician desired. Besides, statisticians were unwilling to write programs, at least for their use.

To crown their unwillingness to use computers, statisticians want to be in direct contact with the data and the processes of analysis, which makes them to be very familiar with the result for proper decision making. With the computer, direct input with the data is most possible, and most statisticians do not know what the computer does with their data.

The arrival of microcomputers has had positive effect on statistical analysis. These effects include its computing power, which is enormously greater than that of the desk machines, in that they can be programmed using a high-level language, and the more recent windows

application programs. With the microcomputer, one can keep the analysis under his control, that is, keep close to the original data while analysis is going on.

Recently, (about three decades after the introduction of electronic computers), the technology of calculators improved such that we now have pocket-size hand-held electronic calculators which could go a long way in assisting in statistical computations.

Unlike the old electro-mechanical desk machines, they have a number of built-in programs, ranging from a mean-and-variance routine to (in more powerful versions) routines for regression and validity probability densities.

In-Text Question

The study of statistics startedin the eighteenth century

In-Text Answer

John Graunt

2.2 Programing Techniques for Statistics

In the early days of computer programming, machine codes were the only familiar software. However, development has grown from these codes to the high-level languages like BASIC, FORTRAN, COBOL and PASCAL. Some of these high-level programs were not so user-friendly, though more user friendly than the machine codes.

This led to the development of more user-friendly programmes (due to the evolvment of windows) like the SPSS (Statistical Package for the Social Sciences), developed in the USA, and GENSTAT (a General Statistical Program) and GLIM, both developed in Britain. The most user-friendly program that could analyze statistical problems is the Microsoft Excel. This is a software for most forms of computations, developed by the Microsoft Corporation, USA. It is easy to learn.

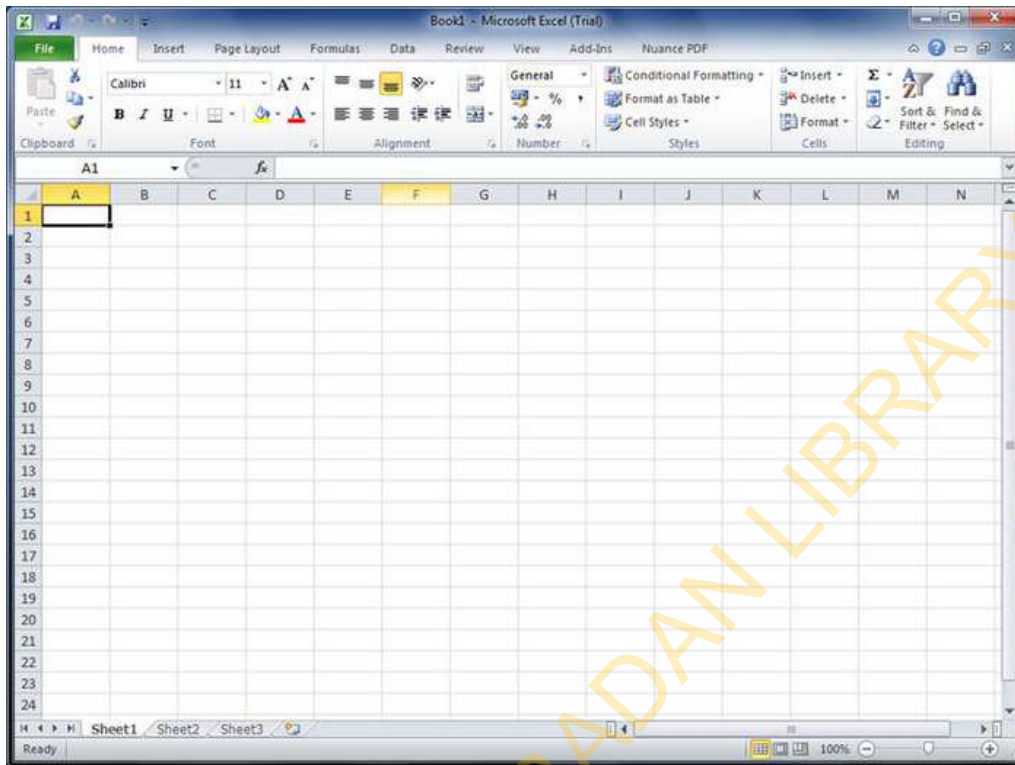


Figure 2.2 : Microsoft Excel

Other statistical packages include the E-views (Econometric Views) for econometric and time series analyses, SAS (Statistical Application Software), STATA, statistical software for almost all statistical theory, and a host of many others. The effect of computers is much more noticeable in advanced statistics, such as the Multiple Regression.

The whole field of multivariate analysis has been opened up through the advances in computers, but the techniques developed there are based on more complicated mathematics. In this book, all our computations shall be based on MS Excel and BASIC.

Summary for 2

Even though statistics have been in use for a very long time, its scope had not gone beyond knowing the population of a country. However, its significance began to unfold in the 18th century when John Graunt began to collect demographic data recorded in the church.

Early statisticians had no interest in the use of the computer, but with the development of statistical tools, interest began to shift towards the device. Today most statistical analyses are computed through the use of computer devices.

Self-Assessment Questions (SAQs) for study session 2

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

Reference

Cooke D., Craven A. H. and Clarke G. M.: Basic Statistical Computing. Second Edition.
Edward Arnold. A division of Hodder and Stoughton.

Study Session 3: Representation of Data in the Computer

Introduction

Computers do not understand the human code. Before processing can be done there must be leverage between the user and the computer. That is, the computer must be able to understand what you are saying.

You may not be able to understand the computer language, but you speak to the computer in your own language. There is a device in the computer that converts your language to that which the computer is able to understand. This device is called the COMPILER. In BASIC programming language, it is called the INTERPRETER.

Learning Outcomes from Study Session 3

At the end of this study session, you should be able to:

- 1.1 Explain the Binary representation
- 3.2 Simple Computer Arithmetic

3.1 Binary Representation

Computers use the Binary code for data or character representation. This numbering system has only two digits, 0 and 1. Each is referred to as binary digit (or bit). Each bit is used to denote the presence and absence of electrical pulse or signal in the computer circuitry.

Most numbering systems are called positional, because the physical location or position of digit within the number affects the value. For instance, 64 and 46 have same digits, but their values are different because of the position of the digits. The same with 52 and 25. Let us

look at the following decimal value of position as well as their corresponding exponential value of position.

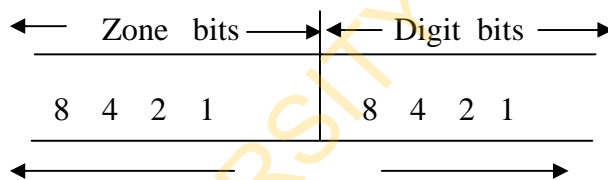
Decimal value of position	1000	100	10	1
Exponential value of position	10^3	10^2	10^1	10^0

Decimal value of position	16	8	4	2	1
Exponential value of position	2^4	2^3	2^2	2^1	2^0

The binary numbering system has a base of 2.

Character Representation

Each storage position in the main memory is referred to as a byte. Since each byte contains four bits that represent the decimal numbers 8, 4, 2, and 1, then this makes it possible to represent the decimal digits 0 to 15. However, for the alphabets and special characters an 8-bit code is usually used. This is divided into two, viz, four *zone* bits and four *digit* bits. This is represented below:



One Byte

The Extended Binary Coded Decimal Interchange code (EBCDIC) is the most commonly used computer code for character representation. The four zone bits are used to indicate codes for letters, unsigned numbers, positive numbers, negative numbers, and special characters. Other computer codes for character representation include:

Binary Coded Decimal (BCD) and American Standard Code for Information Interchange (ASCII). There is an 8-bit ASCII and also a 7-bit ASCII.

3.1.1 Converting from Decimal to Binary

Using the positional numbering system, you can determine what combination of these positional values equals the decimal value.

Example 1: Convert 13_{ten} to binary.

Solution: Write out the positional values and place the digit 1 under the values that add up to 13. Put 0 elsewhere.

$$\begin{array}{cccc} 8 & 4 & 2 & 1 \\ 1 & 1 & 0 & 1 \end{array}$$

So that $13_{\text{ten}} = 1101_{\text{two}}$.

Using the remainder theorem, divide the decimal value by 2 and write down the remainders. Then rearrange the remainders from the bottom to the top.

$$\begin{array}{r|l} 2 & 13 \\ 2 & 6 \text{ r } 1 \\ 2 & 3 \text{ r } 0 \\ 2 & 1 \text{ r } 1 \\ 2 & 0 \text{ r } 1 \end{array} = 1101_{\text{two}}$$

3.1.2 Converting from Binary to Decimal

It can also use the positional numbering system to determine this.

Examples 2: Convert 1111_{two} to decimal.

Solution: Write out the positional codes and write under the digits of the binary code. Multiply each code with the corresponding value and add.

8	4	2	1	
			-----	$1 \times 1 = 1$
		-----	-----	$1 \times 2 = 2$
	-----	-----	-----	$1 \times 4 = 4$
-----	-----	-----	-----	$1 \times 8 = 8$
				15

$\therefore 1111_{\text{two}} = 15_{\text{ten}}$

3.2 Simple Computer Arithmetic

1. *Addition:* Addition of binary numbers is the same with decimal numbers. The highest value for any addition is 1. Anything above 1 is taken as a 'carry over' to the next addition.

Example 3: $10_{\text{two}} + 11_{\text{two}}$, and compute its equivalence in decimal.

Binary	Decimal
10	2
+ 11	+ 3
-----	-----
101_{two}	5_{ten}

Example 4: $1011_{\text{two}} + 1110_{\text{two}}$

Binary	Decimal
1011	11
+ <u>1110</u>	+ <u>14</u>
1100 ¹ _{two}	25

2. *Subtraction*: While binary subtraction is done in much the same way as the decimal subtraction, computers perform this task using *complement addition*. A complement of a number is the value which must be added to it to get its number base. For example, in the decimal system, the complement of 7 is 3, and that of 6 is 4. Using this to perform a subtraction, say 7-4, obtain the complement of 4 and add to 7. Then discard the carry, and leave the rightmost digit. That is, the complement of 4 is 6, so that 7 + 6 = 13. We discard 1 and write 3, this gives the same result as 7 - 4 = 3. The same method is also applied to the binary system.

The reason for this is that unique property of binary numbers allows the determination of the 2's complement to be very simple, a fact that has important implications for simplifying computer circuit design. The digit being subtracted is called the 'subtrahend' while the digit from which the subtrahend is taken from is called the 'minuend'.

Example 5: Perform 7 - 4 in binary with and without complement addition

- a. Without complement addition.

Decimal	Binary
7	0111
- <u>4</u>	- <u>0100</u>
3	0011

b. With complement addition

Decimal	Binary
7	0111
<u>+ 6</u>	<u>+ 1011</u>
* 3	1100
	<u>11</u>
	0011

3. *Multiplication:* This is done by ‘a shift-left and add’ operations as it is in the case of decimal. For example, 12 multiplied by 12 in decimal is

$$\begin{array}{r}
 12 \\
 \times 12 \\
 \hline
 24 \\
 12 \\
 \hline
 144
 \end{array}$$

Example 6: Multiply 12 x 5 in binary

Decimal	Binary
12	1100
<u>x 5</u>	<u>x 0101</u>
60	1100
	+ 0000
	1100
	37

0000

0111100

$60_{\text{ten}} = 111100_{\text{two}}$ verify this.

Example 7: What is 12^2 in binary

```
      1100
    x 1100
    -----
      0000
    + 0000
    -----
     1100
    -----
    10010000
```

$10010000_{\text{two}} = 144_{\text{ten}}$ verify this.

4. *Division:* This is the opposite of multiplication. Instead of shifting the multiplier left and adding the intermediate result, division is performed by shifting the divisor right and subtracting it from the quotient initially and then from each intermediate remainder until zeros are obtained. The subtraction would actually be performed using the 2's complement addition method.

Example (without complement addition)

$25 \div 5 = 5$ $25_{\text{ten}} = 11001_{\text{two}}$ and $5_{\text{ten}} = 101_{\text{two}}$

$$\begin{array}{r}
11001 \\
- \underline{101} \\
\hline
\cancel{10100} \\
- \underline{101} \\
\hline
01111 \\
- \underline{101} \\
\hline
01010 \\
- \underline{101} \\
\hline
00101 \\
- \underline{101} \\
\hline
00000
\end{array}$$

Count the number of subtraction and convert it to decimal. In this example, the number of subsequent subtraction is 5. so $5_{\text{ten}} = 101_{\text{two}}$. Therefore $25_{\text{ten}} \div 5_{\text{ten}} \Rightarrow 11001_{\text{two}} \div 101_{\text{two}} = 101_{\text{two}}$

Optal and Hexadecimal

Other useful number systems in computer are the octal and hexadecimal number systems. The octal number system uses 8 symbols (0 -7) as its digits. Note that each of these octal digits can be represented by 3 bits. The hexadecimal number system on the other hand, uses 16 symbols (0-9 and A – F). This A- F represents 10 - 15. Each of these hexadecimal digits can be represented by 4 bits.

Summary for 3

- The only way the computer can decode the data and information entered into it is for it to convert it (data and information) to machine code.
- This conversion is brought about by the compiler. The codes used by the computer to represent data are the digits 0 and 1, which represent the presence and absence of electrical pulse or signal in the computer circuitry.
- These codes are called binary digits. Each storage position is called a byte, containing 4-bits. The alphabets and special characters are represented by 8-bits comprising 4-zone bits and 4-digit bits.
- Apart from the binary digits, other useful numbers in the computer are the octal and hexadecimal.

Self-Assessment Questions (SAQs) for study session 3

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

1. Convert the following to binary: (a) 0.1 (b) $\frac{1}{2}$ (c) 8^3 (d) F
2. Obtain the following using binary codes:
(a) $\text{LOG}_5 125$ (b) $3^2 + 2^3$ (c) 0.5×2 (d) $50 \div 10$
3. Given a base of 9 and a height of 5, what goes on in the compiler when the area of a parallelogram is computed? (Show your workings).
4. A trapezium has its height as 6, and the parallel sides as 4 and 7.5. Obtain its area in binary and convert your result to decimal.

5. Convert the following to decimal: (a) 1100110_{two} (b) 0010010_{two} (c) 1111111_{two} (d) 000011_{two}

Reference

Ojo S. O. (1991): *Introduction to Computer Science*. Revised Edition. Department of Computer Science, University of Ibadan, Nigeria.

UNIVERSITY OF IBADAN LIBRARY

Study Session 4: Introduction to Microsoft Excel

Introduction

Analysis of data on computer is fast taking the shine off electronic calculators and adding machines. One of the commonest and easiest to learn software for analysis is the Microsoft Excel. MS Excel is a powerful tool from the Microsoft Corporation. It is an interesting package for introduction to data analysis.

Learning Outcomes from Study Session 4

At the end of this study session, you should be able to:

- 1.1 Explain the Spreadsheet
- 1.2 Types of Sheet in MS Excel.

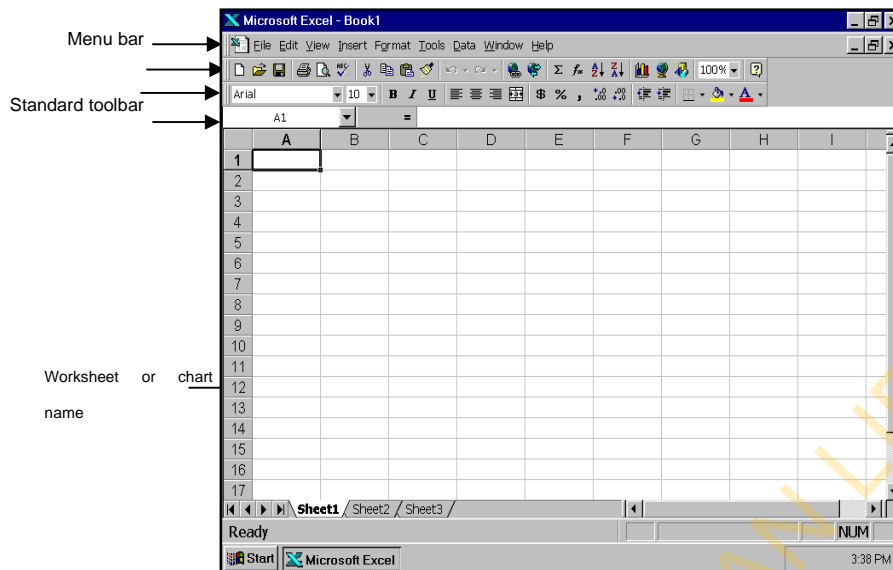
4.1 Spreadsheet

Microsoft Excel is a powerful electronic spreadsheet for Microsoft Windows on which data are entered and calculations performed. A spreadsheet is a sheet that has rows and columns in which data are entered into for the purpose of mathematical, statistical, logical, operation, and so on.

An MS Excel spreadsheet is made up of grid lines (vertical and horizontal) that form cells. A cell is an intersection of four grid lines, two of which are vertical and the other two horizontal.

In other words, it is the intersection of a row and a column. There are about 256 columns and 65,536 rows. The columns are labeled A to iv, while the rows are labeled numerically. Every cell has a name which is always corresponding to the column and the row. For example a cell that is made up of column D and row 8 is called cell D8.

Diagram of MS Excel spreadsheet



4.1.1 Simple Operations

To go to any cell, simply type the cell name in the *Name Box*. The Name box is directly under the toolbars. To execute the command to go to a named cell after typing the cell name (in the name box), press the ENTER key. Every time you type into any cell, the content is displayed in the *Formula Bar*. The Formula Bar is adjacent to the Name Box.

When you want to change the content of a cell, simply move your cell pointer to that cell and begin typing the new content. The former content will automatically be deleted. If you want to add to the content of a cell, take your cell pointer to the cell, and click your mouse on the formula Bar.

In-Text Question

.....is made up of grid lines that form cells.

In-Text Answer

MS Excel spreadsheet

Use the arrow or direction) keys to move to the place you want to make an additional insertion, are start typing. Every time you enter into a cell, press the ENTER or arrow keys.

4.1.2 Selecting a Cell or Multiple Cells

1. To select a cell, simply click on that cell or use the arrow keys to move to the cell you want to select.
2. To select *adjacent cells*, click the first cell and drag your mouse to the last cell that you want to include in the selection. Using the keyboard, take the cell pointer to the first cell in the selection, hold down the SHIFT key and use the arrow keys to include every cell for the selection.
3. To select *non-adjacent cells*, hold down the control key and click the cells you want to include into the selection.

4.2 Types of Sheet in MS Excel

There are several workbooks in the MS Excel, and in each workbook are several sheets. Each workbook and sheet runs into thousands respectively. You can therefore imagine how many sheets are available in the MS Excel. Check it.

There are different types of sheets in a workbook. These include

1. Worksheet
2. Chart sheet
3. Visual Basic module sheet
4. Dialog sheet
5. Macro sheet
6. International macro sheets

However, not all of these may be available in your MS Excel package. The commonest sheets of the lot are the worksheet and chart sheet. To insert a worksheet, from the INSERT menu choose worksheet. You can as well press SHIFT + F11 on the keyboards.

To insert a chart sheet, from the INSERT menu choose chart sheet, and then *As New Sheet*. The chart wizard appears. Then follow the instruction on the screen. You can as well press F11 on the keyboard.

Sheets can be renamed, copied or moved within the workbook, or to another workbook. You can also hide sheets within the workbook.

In-Text Question

To select a cell, simply click on that cell or use the arrow keys to move to the cell you want to select. True or False

In-Text Answer

True.

The worksheet is where calculations are performed, whereas the chart sheet is where charts are displayed. However, charts can also be displayed on the worksheet that contains the data used to draw the chart. Any alteration on the data also alters the chart. Deleting the data also deleted the chart.

Sizing Cells

Cells can be enlarged or reduced. Three methods would be presented here.

Method 1:

- i. Take the cell pointer to the cell you want to enlarge or reduce.
- ii. Click the FORMAT menu.
- iii. Point to COLUMN and click 'width'
- iv. Enter a size value that could accommodate the entry in the cell.

(However you must be aware of the standard size of a cell which is normally 8.43 pixels).

- v. Press ENTER or click OK.

Method 2:

- i. Select a cell you want to resize
- ii. Click the FORMAT menu
- iii. Point to COLUMN and click 'Auto fit selection'.

Method 3:

- i. Go to the column label corresponding to the cell you want to resize.
- ii. Place your cell pointer on the right grid line of the column.
- iii. Hold down the mouse button and drag to the desired size, then release.

Note: Whatever resizing done to any cell affects every cell in that column.

Summary for 4

In this study session,

- You have been introduced to MS Excel as electronic spreadsheet software with simple operation of how to move to a named cell, change or add to the entry of the cell.
- You have also learnt the different types of sheets in a single workbook, and how to select a cell or multiple cells. Also learnt is how to size a cell to accommodate its entry.

Self-Assessment Questions (SAQs) for study session 4

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

1. Define a Microsoft Excel.
2. How many cells are in a single worksheet?
3. How do you select non-adjacent cells of a worksheet?
4. What are the kinds of computations that could be performed with MS Excel?
5. Mention five types of charts that could be drawn with MS Excel.

Reference

Goal 2000 Computer Networks, Module 4: Using Microsoft Excel. MAGNA Computer School, Ibadan, Nigeria.

UNIVERSITY OF IBADAN LIBRARY

Study Session 5: Data Entry in Microsoft Excel

Introduction

The simplest task in MS Excel is entering data into the cells. This is the first step before any analysis could be carried out. There are different types of data that MS Excel can accept. However, there are rules governing entry of data. There is no data that operations cannot be performed upon.

Learning Outcomes from Study Session 5

At the end of this study session, you should be able to:

- 5.1 Explain the types of data to be entered into an MS Excel worksheet;
- 5.2 Explain the techniques of entering data into the cells;
- 5.3 Create simple forecasts and trends.

5.1 Types of Data

You can enter two types of data into a worksheet – these include

1. Constant values
2. Formulas

A constant value is data that you type into a cell directly. It comes in various forms. These are numeric, alphabetic, alphanumeric, date, time, currency, logical, percentages, fractions, and scientific notations. Values are constant and do not change unless you select the cell and edit the value yourself.

On the other hand, a formula is a sequence of constant values, cell references, names, functions, or operators that gives rise to a new value from existing values. In MS Excel, formulas always begin with an equal sign, '='. In cases where cell references are used to

create a formula, a value that is produced as the result of the formula can change when other values in the worksheet change.

Numbers

The combination of numeric characters, 0 through 9, and any special character such as + - (), / & % make up Numbers.

When entering numbers,

- i. You can include commas, for example 1,000,000
- ii. A single period is treated as a decimal point.
- iii. Plus signs entered before numbers are ignored.
- iv. Precede negative numbers with a minus sign or enclose them within parentheses (or brackets).

MS Excel uses General number format as default, except otherwise stated. When it can, it automatically assigns the correct number format to your entry. For instance, when you enter a dollar sign before a number, MS Excel automatically converts your entry into a currency format. When entered, numbers in the cell align to the right.

Dates and Times

If you want to display the time using the 12-hour clock, type am or pm. For example 5:00p.m, 1.00a.m. You can type 'a' or 'p' instead of am or pm. However, you must include a space between the time and the letter. Unless you type am or pm, MS Excel automatically displays time using the 24-hour clock. For example: 20:00, 23.00, and 13.00.

You can type a date and time in the same cell. There must be a space between them. In entering date, you can use either a slash (/) or a hyphen (-). There are several standard formats for displaying date and time; however, MS Excel stores all dates as serial numbers and all times as decimal fractions.

In-Text Question

What is constant value?

In-Text Answer

A constant value is data that you type into a cell directly.

MS Excel sees dates and times as numbers. This makes it possible to perform different arithmetic operations on them. For example, the difference between 5th of December, 2005 and 3rd of May 2005 is written.

$$= "5/12/05" - "3/5/05"$$

and the result that would be displayed is 175.

5.2 Data Entry Techniques

To enter data into a cell

- i. Select the cell you want to enter data.
- ii. Type the data.
- iii. Press ENTER.

To type same entry into several cells at once

- i. Select the range of cells you want to enter data. The selected cells can be adjacent or non-adjacent.
- ii. Type the data
- iii. Press CTRL + ENTER

To enter numbers with fixed decimal places

- i. From the tools menu, choose options

- ii. Select the Edit tab
- iii. Select the fixed decimal check box, and then select the number of decimal places in the places box.
- iv. Choose the OK button
- v. Begin entering numbers into cells without typing the period for the decimal places.

5.2.1 Using Autofill

The AutoFill feature is used to create a series of incremental or fixed values on a worksheet by dragging the fill handle with the mouse. For example, you can copy the value from one cell into five cells below it. In this instance, Auto fill works in the same way as the Fill commands on the Edit menu.

You can also drag the fill handle to increment a series or you can use the series command (Edit menu, Fill submenu).

For instance, if you type January and February in consecutive columns, and then drag the fill handle to the right, MS Excel fills March, April, May and so on into the selected cells.

Series are useful when creating table row or column headings on a worksheet, or anytime you need to enter a series of incremental numbers, dates, or time periods. Series can be created in any direction. AutoFill can only fill a range of adjacent cells.

To copy by dragging the fill handle.

- i. Select the cell containing the data you want to copy.
- ii. Drag the fill handle across the cells you want to fill and then release the mouse button.

Any existing values or formulas in the cell you fill will be replaced.

To create a series increment:

- i. Type into the first 2 or 3 cells of the data in order.
- ii. Select the two or three cells that contain the data you have typed and which you want to create a series.
- iii. Drag the fill handle across the cells you want to fill and then release the mouse button.

In-Text Question

What is AutoFill?

In-Text Answer

The AutoFill feature is used to create a series of incremental or fixed values on a worksheet by dragging the fill handle with the mouse.

To copy using the Fill Right and Fill Down commands

- i. Select the cell or cells you want to copy and the adjacent cells you want to fill.
- ii. To copy the selection's first column into the adjacent cells to the right, choose Fill from the Edit menu, and then choose Right.

Keyboard command = CTRL + R

To copy the selection's first row into the adjacent cells below, choose Fill from the Edit menu, and then choose down.

Keyboard command = CTRL + D

To copy the selection's last column into the adjacent cells to the left, hold down SHIFT, choose FILL from the Edit menu, and then choose Left.

Keyboard command = SHFT + FILL Right button (Edit category)

To copy the selections last row into the adjacent cells above, hold down SHIFT, choose FILL from the Edit, and then choose up.

Keyboard command = SHIFT + Fill Down button (Edit category)

5.3 Creating Simple Trends and Forecasts

Trend means behaviour. You can observe the behaviour of your data, e.g. sales. The pattern of sales of a company or business enterprise could be studied. Also a future value could also be predicted for any data. In this section, we are going to see how to create simple trends and forecasts given a data.

To create a linear or Growth Trend series using the Auto Fill shortcut menu:

- i. Select the cell range containing the values on which you want to base you trend.
- ii. Hold down the right mouse button and drag the fill handle the direction you want to fill.

(The AutoFill shortcut menu is displayed)

- iii. Choose linear trend or Growth Trend.

To create Linear or Growth Trend series using the series command

- i. Select the cell range containing the values on which you want to base your trend.
- ii. From the Edit menu, choose Fill, and then choose series.
- iii. Under series in, select the Rows option button or the Column option button, depending on your selection.
- iv. Under type, select the linear option button to produce a linear growth trend, or select the Growth option button to create an exponential growth trend.
- v. Select the trend check box.
- vi. Choose the OK button.

Summary for 5

Entering data into MS Excel is a simple process.

- Two types of data can be typed into a worksheet cell. These are constant values and formulas. Any data typed directly into a cell is called a constant value, and they come in different forms.
- On the other hand, a sequence of constant values, cell references and so on is a formula. Data can be entered into a cell or range of cells at a time. Also you can use AutoFill command to copy and create series.
- With this command also, trends could be observed, and future values could be forecast given existing values.

Self-Assessment Questions (SAQs) for study session 5

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

1. Enter the following data into an MS Excel worksheet

Income	Expenditure
₦	₦
25,000	25,000
15,000	13,000
30,000	25,000
45,000	44,000

35,000 36,000

20,000 17,000

2. What is the & value of the following

i. = “13/1/02” - “29/9/01”

ii. = “4:40pm” - “3:15am”

3. The following are sales data from ESCRAVOS LTD for the first quarter of the year 1999.

MONTH	SALE
	(₦)
January	251,050
February	270,500
March	269,150

Forecast the sales of the company by the tenth month, and obtain the predicted total sales.

Reference

Goal 2000 Computer Networks, Module 4: *Using Microsoft Excel*. MAGNA Computer School, Ibadan, Nigeria.

Study Session 6: Data Analysis using Microsoft Excel

Introduction

No data is useful when it is not processed. The processing of data is also known as analysis. Analysis of data provides information for the purpose of making decisions. The Microsoft Excel is mainly for data analysis and there are various analyses that could be performed.

Learning Outcomes from Study Session 6

At the end of this study session, you should be able to:

6.1 Explain the general rule on Computation in MS Excel

6.1 General Rule on Computation in MS Excel

The development of the Microsoft Excel is similar to that of LOTUS 123. Being a software for data analysis, the rules of mathematics must be followed. MS Excel makes use of arithmetic and logical operators in returning the output involving two or more figures. The operator are +, -, x and \div (for arithmetic), and <, >, \leq \geq (for logical). The rule, BODMAS, is also important in the order MS Excel performs its calculations.

However, to return the output of any formula or expression, while Lotus 123 uses '@' before writing the expression, MS Excel uses '='. Any expression that is not preceded by the "=" sign will be regarded as a cell entry. Expression with the '=' sign is called a formula. To correct a formula entry or expression click the cell that contains the formula or expression, and go to the Formula bar and correct. To execute a formula, press the ENTER key.

Simple Computations using Raw Figures

In this section you shall learn how to calculate simple mathematical expressions. We shall start with simple arithmetic operations.

1. Addition of 2 and 2 is written

$$=2+2$$

2. Subtraction of 6 from 10 is written

$$=10-6 \quad \text{or} \quad =(-6)+10$$

3. Multiplication of 3 by 4 is written

$$=3*4 \quad \text{or} \quad =4*3$$

4. Division of 25 by 5 is written

$$=25/5 \quad \text{or} \quad =25*(1/5)$$

The above are the basic usage of the arithmetic operators. Now we shall go ahead to look at a situation where we have more than one expression in a formula.

1. $\frac{1}{3}$ of 9 is written

$$= (1/3)*9$$

2. $\frac{2}{3} + \frac{4}{5}$ is written

$$= (2/3)*(4/5)$$

3. $2 - \frac{1}{2} + 1.5$ is written

$$=2-(1/2)+1.5$$

Just as in normal mathematics, MS Excel will follow the rule of BODMAS, and solve the part in bracket, followed by the addition before the subtraction.

4. $3\frac{1}{4} \div \frac{9}{11}$ is written

$$= (3*(1/4)) / (9/11)$$

The easiest way to know how to write an MS Excel expression for any mathematical expression is to follow the logical procedures for solving the problem (assuming there is no computer).

In-Text Question

Using computation operation, Multiplication of 3 by 4 is written

In-Text Answer

=3*4 or =4*3

Simple Computations Using Cell References

A reference identifies a cell or group of cells on a worksheet. References tell MS Excel which cells to look into to find the values you want to use in a formula. With references, you can use data contained in different parts of a worksheet in one formula and use the value of one cell in several formulas.

You can also refer to cells in other sheets in a workbook, to other workbooks, and to data in other applications. References to cells in other workbooks are called *external references*. References to data in other application are called *remote references*.

We are going to use the examples in the previous section to work in this section, using only cell references.

1. Assume the first digit 2 is stored in A3, and the second digit 2 is stored in cell A5, then the addition is written

=A3+A5

2. Assume the digit 6 is stored in cell A2, and the digit 10 is in cell B5, then the subtraction is

=B5-A2

Assume a negative digit 6 is in cell A2, then the second method of the subtraction (i.e. addition) is =A2+B5.

3. In the example on multiplication, put 3 in AJ20, and 4 in K65, then we have

$$=AJ20*K65 \quad \text{or} \quad =K65*AJ20$$

In the fourth example, suppose 25 is stored in cell BK 2 and 5 in CF 25, then we write

$$=BK2/CF25$$

Using the second example, assume 25 remained as stored, and the digit 1 stored in B25, and 5 in J49, then the same result is obtained from

$$=BK2*(B25/J49)$$

We go ahead to look at the other examples having more than one expression,

1. Put 1 in A23, 3 in C5, and 9 in A12, then the expression is written

$$=(A23/C5)*A12$$

2. Let 2 be in B3, 3 in B4, 4 in A13 and 5 in J1, then example 2 is written

$$=(B3/B4)*(A13/J1)$$

3. Store 2 in cell F14, $\frac{1}{2}$ written as 0.5 in cell D7, and 1.5 in cell A1 then examples 3 is written

$$=F4-D7+A1$$

4. The last example could have 3 in cell A3, 1 in cell B1, 4 in cell D8, 9 in cell F29 and 11 in cell M16, and then we write.

$$=(A3*(B1/D8))/(F29/M16)$$

Note: It is preferred that cell references be used to write formulas. The reason is because whenever any change is made in any part of the original data or some results that leads to the final result, then subsequent results changes automatically, also affecting the final result.

Summary for 6

- Problem solving with MS Excel is not different from the usual mathematics you are familiar with.
- MS Excel performs calculations by use of arithmetic and logical operation, and also follows the rule of BODMAS.
- To execute an expression or formula, an equal sign '=' must proceed the expression or formula.

- It is best to use cell references for expressions or formulas whenever the references or formulas are with respect to a data already stored in the workbook.

Self-Assessment Questions (SAQs) for study session 6

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

Post -Test

1. Write an MS Excel expression for calculating
 - i. $\frac{1}{2}bh$, where $b=5$, $h=9$
 - ii. $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$, where $a=1$, $b=-3$, $c=2$
 - iii. $2^3 \times 3^2$
2. Assume the following data 2, 1, 5, 2, 7, 6, 3, 10, 9, 3. Let these data be stored in cells B2 to B11 respectively.

Write an MS Excel expression for obtaining

- i. the sum (using the arithmetic operation '+') and store cell D2.
 - ii. the average and store in cell D3.
 - iii. the subtractions of (ii) from each of the data points in B2 to B11 and store in cells C2 to C11.
 - iv. what have you just calculated?
 - v. what do you expect the sum of (iii) above to be?
3. Assume the data in (2) above is stored in cells B2 to K2 respectively.

Write an MS Excel expression for obtaining

- i. the sum as 2(i) above and store in cell E3.
- ii. the average and store in cell E4.

iii. the subtractions of (ii) from each of the data points in B2 to K2, and store in cells C5 to L5.

Reference

Goal 2000 Computer Networks, Module 4: *Using Microsoft Excel*. MAGNA Computer School, Ibadan Nigeria.

UNIVERSITY OF IBADAN LIBRARY

Study Session 7: Using the Function and Chart Wizards

Introduction

The ultimate in the use of the Microsoft Excel is analysis of data. In this study session, you are going to discover the different types of calculations that MS Excel can perform. MS Excel can perform a wide range of analysis, which makes it a ready tool for almost all kinds of professions. There are also wide ranges of charts that can be drawn.

Learning Outcomes from Study Session 7

At the end of this study session, you should be able to:

7.1 Discuss the use of Functions

Pre -Test

1. Mention two functions an MS Excel can perform.
2. What is the action of each of the following sub-functions and under which function can each be found?
 - i. ABS
 - ii. KURT
 - iii. AVERAGE
 - iv. SINH
 - v. IF
 - vi. MOD
 - vii. STDEV
 - viii. VAR
 - ix. EXP
 - x. AND
3. Mention three charts you know.
4. Differentiate between a bar chart and pie chart.

7.1 Using Functions

You use the MS Excel built-in functions to perform standard worksheet and macro sheet calculations. The values that you give a function to perform operations on are called arguments. The values that the functions return are called results. You use functions by entering them into formulas on your worksheet.

The sequence of characters you use in a function is called the syntax. All functions have the same basic syntax. If you do not follow this syntax, MS Excel displays a message indicating that there is an error in the formula.

Brackets tell MS Excel where the arguments begin and end. Remember to include both brackets, with no spaces preceding or following. You specify arguments within these brackets. Arguments can be numbers, text, logical values, arrays, error values, or references. The argument you designate must produce a valid value for that argument.

Arguments can also be constants or formulas. The formulas themselves can contain other functions. When an argument to a function is itself a function, it is said to be nested. In MS Excel, you can nest up to seven levels of functions in a formula.

To insert a built-in function

1. Select the cell where you want to enter the formula
2. From the insert menu, choose function.
3. Select the function category
4. Select the function name
5. Choose the next button
6. Enter values for the arguments
7. Choose the finish button

We are going to consider some categories under three functions, namely

1. Mathematical and Trigonometrical
2. Statistical
3. Logical

(Note: the student is expected to obtain, study and understand all categories)

In-Text Question

To insert a built-in Function, which one of the following is not correct.

- A. Select the function category
- B. Select the function name
- C. Choose the next button
- D. None of the above.

In-Text Answer

D.

Mathematical and Trigonometrical Function

There are fifty categories under this function. We shall consider just ten.

- i. ABS – returns the absolute value of a number.

The syntax is

=ABS(number). Example: =ABS(5), =ABS(-3), =ABS(A7)

- ii. SIN – returns the sine of a number.

The syntax is

=SIN(number). Example: =SIN(30), =SIN(-60), =SIN(J25)

- iii. COMBIN - returns the combination of a number given a chosen number

The syntax is

= COMBIN(number, number_chosen)

Example: =COMBIN(5,2), =COMBIN(B10, A1)

- iv. LOG – returns the logarithm of a number given a base

The syntax is

=LOG(number, base)

Example: =LOG(25, 2), =LOG(36,6)

- v. LOGIO – returning the logarithm of a number given base 10

The syntax is

=LOG10(number)

Example: =LOGIO(100), =LOG10(1)

- vi. LN – returns the natural logarithm of a number

The syntax is

=LN(number)

Example: =LN(2.513), =LN(1.045)

- vii. FACT – returns the factorial of a number

The syntax is

=FACT(number)

Example: =FACT(5), =FACT(3)

- viii. MMULT – returns the matrix multiplication of two matrices or arrays.

The syntax is

=MMULT(array1, array2)

Example: MMULT(A2: B3, C4:D5)

- ix. SUM – returns the sum of several values

The syntax is

=SUM (number 1, number 2,...)

Example: SUM (A2:A10), =SUM(B1: D1)

=SUM(1, 2, 1, 3)

- x. MOD – returns the modulo of a number given a division

The syntax is

=MOD (number, divisor)

Example: = MOD (5,2), =MOD(23,3)

UNIVERSITY OF IBADAN LIBRARY

7.1.1 Statistical Function

There are eighty categories under this function. We shall consider just ten

- i. AVERAGE – returns the arithmetic mean of several values

The syntax is

=AVERAGE (number 1, number 2, ...)

Example: =AVERAGE(1,2,1,3), =AVERAGE(B3:B7)

- ii. GEO MEAN – returns the geometric mean of several values.

The syntax is

= GEOMEAN(number 1, number 2,...)

Example: =GEOMEAN(1, 2, 1, 3), =GEOMEAN(B2: B7)

- iii. HARMEAN - returns the harmonic mean of several values.

The syntax is

=HARMEAN(number 1, number 2,...)

Example: =HARMEAN(1, 2, 1, 3), =HARMAEN(B2: B7)

- iv. CORREL – returns the correlation between two arrays

The syntax is

=CORREL(array 1, array 2)

Example: =CORREL(A1: A5, B1 : B5)

- iv. VAR – returning the variance of several values

The syntax is

=VAR(number 1, number 2,...)

Example: =VAR(1, 2, 1, 3), =VAR(B2: B7)

vi. PERCENTILE – returns the k percentile in a given array

The syntax is

=PERCENTILE(array, K)

Example: =PERCENTILE(I15 : I25, 25)

Example: =PERCENTILE(M7 : K17, 90)

vii. COUNT – returns the total number of entry in an array

The syntax is

=COUNT(value 1, value 2, ...)

Example: =COUNT(1, 2, 1, 3), =HARMAEN(B2: B7)

viii. TTEST – returns the value of student – t, test of 2 arrays.

The syntax is

=TTEST(array 1, array 2, tails, type)

Example: =TTEST(A2 : A5, B2 : B5, 1, 2)

ix. CONFIDENCE – returns the confidence interval of addition.

The syntax is

=CONFIDENCE(alpha, standard_dev, size)

Example: =CONFIDENCE(0.05, 1.05, 5)

- x. BINOMDIST – returns the probability of binomial distribution

The syntax is

=BINOMDIST(number_s, trials, probability_s, cumulative)

Example: =BINOMDIST(5,3, 0.5, 1)

In-Text Question

List four statistical function in MS Excel.

In-Text Answer

- GEOMEAN
- HARMEAN
- COUNT
- PERCENTILE

Logical Function

There are six categories under this function. We shall consider all of them.

- i. AND – returns true if all arguments are fine, otherwise false if any is false.

The syntax is

=AND(logical 1, logical 2, ...)

Example: =AND(B2 >1, B5 < 1)

- ii. IF – returns true or false given splid condition.

The syntax is

=IF(logical-test, value-if-true, value-if-false)

Example: =IF(B2 <1, B5 >1)

- iii. NOT - returns true if any is true and false if all are false.

The syntax is

=NOT(logical)

Example: =NOT(TRUE), =NOT(FALSE)

- iv. OR – returns true if any is true and false if all are false.

The syntax is

=OR(logical 1, logical 2)

Example: =OR(B2 < 1, B5 >1)

- v. FALSE – returns the logical value False.

The syntax is

=FALSE()

- vi. TRUE - returns the logical value TRUE

The syntax is

=TRUE()

7.1.2 Using Charts

Charts are pictorial representation of a given data. There are different types of charts such as bar chart, pie chart, line graph, column chart, and so on. Charts can be inserted in the worksheet that contain the data being used for displaying the chart, or on a new sheet called the chart sheet. To display a chart,

- i. Select the range cells containing the data to be used for the chart.
- ii. On the Insert menu, click chart

iii. Select the type of chart you want to use

iv. Follow the instruction on the screen

Instead of first selecting the range of cells, you could go straight to the chart wizard and select the chart you want. And on the space or spaces provided for the cells range, enter or drag the array into it. Then follow the instruction on the screen.

Summary for 7

You can use the built in formula in MS Excel to perform calculation.

- However, you provide MS Excel with the arguments, placed inside the brackets. When the syntax is not correct, MS Excel flashes an error message.
- There are different functions in MS Excel, each having categories.
- The most common and most used are the mathematical/trigonometrical functions, as well as the statistical functions

Self-Assessment Questions (SAQs) for study session 7

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

Post -Test

1. Write an MS Excel formula to obtain
 - i. the arrangement of 5 items taken 2 at a time
 - ii. the selection of 5 items taken 2 at a time

- iii. the medium of 1, 2, 1, 1, 3, 1, 5, 2
 - iv. the modulo of 13 divided 3
 - v. the standard deviation of (iii) above
 - vi. the correlation of (hypothetical) data in cells B1 to B5 and D1 to D5.
 - vii. empty cells in cells A15 to K 15
 - viii. the truth value of cells B3 less than 3 and C4 greater than 2.
 - ix. the slope of a regression line
 - x. the kurtosis of a group of data
2. A researcher wishes to know whether a student's IQ depends on his age, and thus obtained information on five randomly selected students from five departments.

Age (Y)	Y ₁	Y ₂	Y ₃	Y ₄	Y ₅
Performance (X)	X ₁	X ₂	X ₃	X ₄	X ₅

Assuming he uses the Microsoft Excel package, and stores Age variables in cells A2 to A6, and performance variables in cells B2 to B6, while the headings, Age and Performance occupy cells A1 and B1 respectively. Help him in writing expressions to

- i. Store the squares of performance variables in cells C2 to C6.
- ii. Store the product of Age and Performance variables in cells D2 to D6.
- iii. Store the sum and mean of Age variables in cells A8 and A9 respectively.
- iv. Store the sum and mean of performance variables in cells B8 and B9 respectively.
- v. Store the sum of the squares of Performance variables in cell C8.
- vi. Store the sum of the products of Age and Performance variables in cell D8.
- vii. Store the number of data entered, of either variable, in cell F9.

3. Display with MS Excel a compound bar chart for the following data

INCOME	EXPENDITURE
25,000	25,500
15,000	13,000
30,000	25,000
45,000	44,000
35,000	36,000
20,000	17,000

Reference

The Microsoft Excel Software: *Excel 2003*

UNIVERSITY OF IBADAN LIBRARY

Study Session 8: Algorithm and Flow Chart

Introduction

Everything has a way of approach. The tendency is to find an easy means of solving a problem. Most problems are solved starting with an algorithm, and then a flowchart. A problem solved with algorithm and flowchart is almost solved.

Learning Outcomes from Study Session 8

At the end of this study session, you should be able to:

- 3.1 Write an algorithm to solve problems
- 3.2 Algorithm with a flowchart.

Pre-Test

1. Write an algorithm for the preparation of “Eba”
2. Write an algorithm to sum five digits
3. Draw a flowchart for the preparation of “Eba”
4. Draw a flowchart to sum five digits

8.1 Algorithm

An algorithm is a step-by-step method of solving a problem. It does not use any special code other than the languages we speak. Algorithm is a description on how to approach a situation. For example, I could offer help to someone going to UI from Lagos by way of describing how to get to UI. In doing this, what do I tell him?

1. Get to the motor garage and board a vehicle going to Iwo Road in Ibadan.
2. From Iwo Road, board a vehicle going to Agbowo – Ojoo route.
3. Drop at Agbowo junction on the express, and enter another vehicle going to UI.
4. Drop at the gate and enter the campus to the transport unit, and enter a vehicle going to your destination on campus.

There is no general rule in writing an algorithm. Sometimes an algorithm can be written in form of a real program. As long as it provides an easy way to (draw flowchart and) writing the actual program, an algorithm is accepted.

Example 1: Write an algorithm to compute $n!$

- i. Write down the digit n
- ii. Subtract 1 from n and multiply the result with n
- iii. Subtract 1 from the result of the subtraction in (ii) above and multiply the value with the multiplication result in (ii)
- iv. Continue the process until the result of the subtraction becomes 1.

The above algorithm tells you that $n!$ is

$$n \times (n-1) \times (n-2) \times \dots \times (n-n+1)$$

Assume $n = 5$, then $5! = 5 \times 4 \times 3 \times 2 \times 1$

Example 2: Write an algorithm to calculate the area of a triangle given the base and the height.

Solution

1. Multiply the base with the height
2. Divide the result by 2

In-Text Question

What is Algorithm?

In-Text Answer

An algorithm is a step-by-step method of solving a problem

8.2 Flowchart

Just as the algorithm, flowcharts are graphic methods of solving a problem, by a sequence of operations. A graphic is a two-dimensional pictorial format. They serve as a means of communication, telling how an operation should be performed. The name flowchart comes about from the use of charts to display the orderly passing of control from one operation to the next in an explicit sequence.

Prior to the advent of computers the name 'flowchart' was used by systems analyst to designate a means of describing the flow of documents carrying data in an organization. Nowadays we use it in describing the operations of the computer. Flowchart comes in a variety of names such as logic chart, run diagram, process chart, flow diagram, procedure chart, block diagram, system chart, and logic diagram, amidst many others.

Flowchart is adaptable to any kind of program or operation. Like in algorithm, the normal language (e.g. English Language) that you speak is used to describe the processing of an

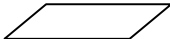
operation of a flowchart. It is a foundation for writing programs, and is most useful in a team work as all members of the team understand its language regardless of their training in programming language.

8.2.1 Flowchart Symbols

The basic flowchart symbols include the following

- i. Start/Stop  (A flattened ellipse)

This is used to indicate the beginning and ending of a process.

- ii. Input/Output  (A parallelogram)

This is used for input and output processes. You input data, and output results.

- iii. Processing  (A rectangle or square)

This is used in arithmetic and/or logical processes in an operation.

- iv. Connector  (A small circle)

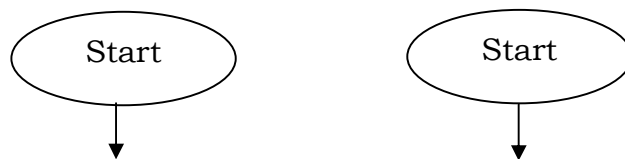
This is used in connecting two sequences of flowcharts that are broken due to insufficient space.

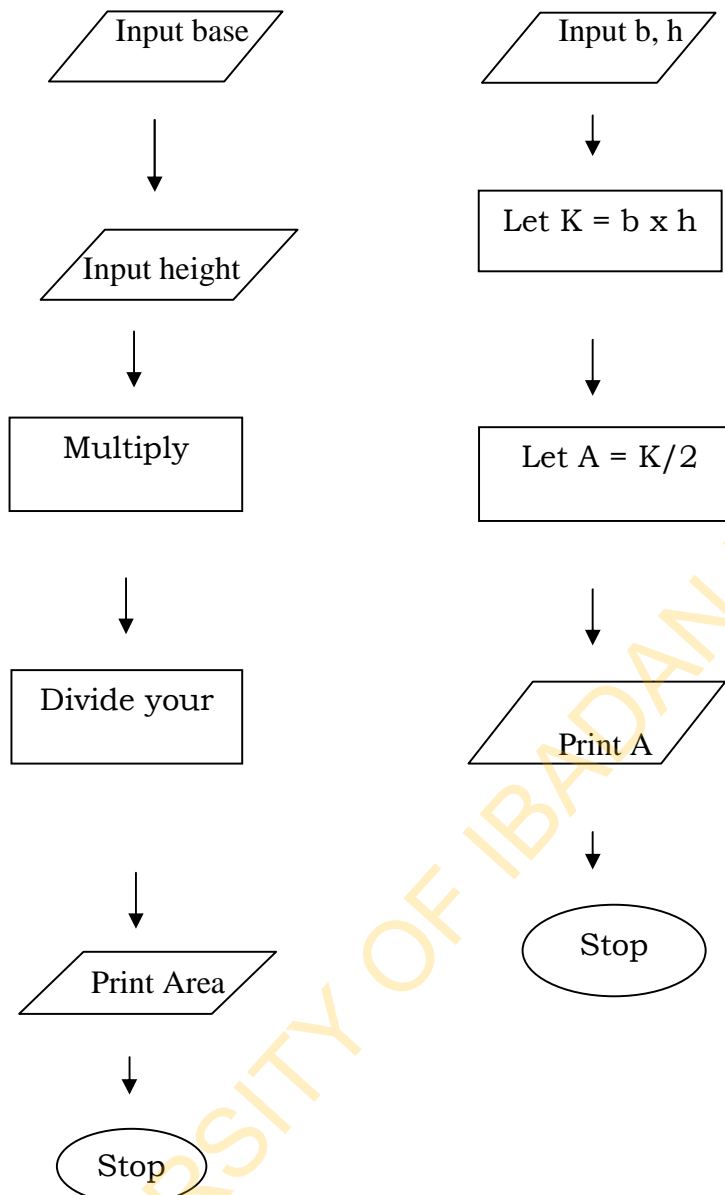
- v. Decision  (A kite)

This is used in setting a logical rule.

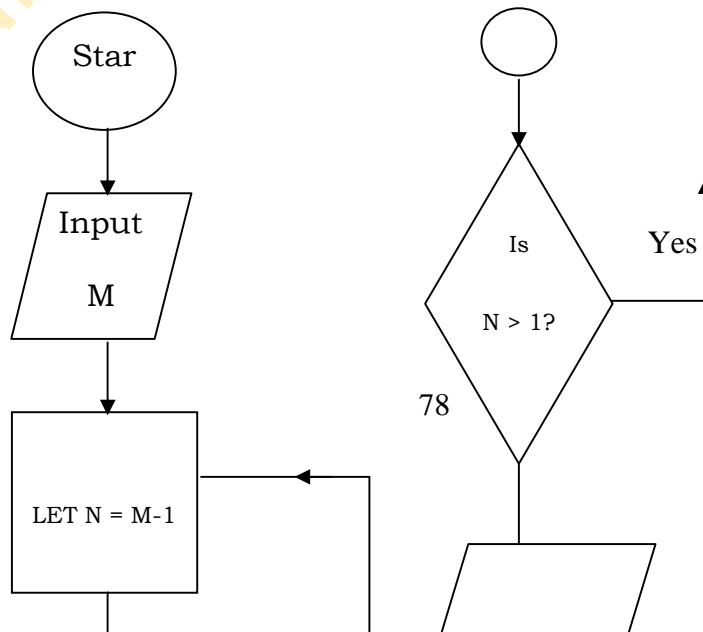
To make this charts meaningful, arrows, are used to join one chart with the other. This is where the flow comes from, and hence, the name “flowchart”.

Example 3: We shall draw a flowchart for the calculation of the area of a triangle in example 3.



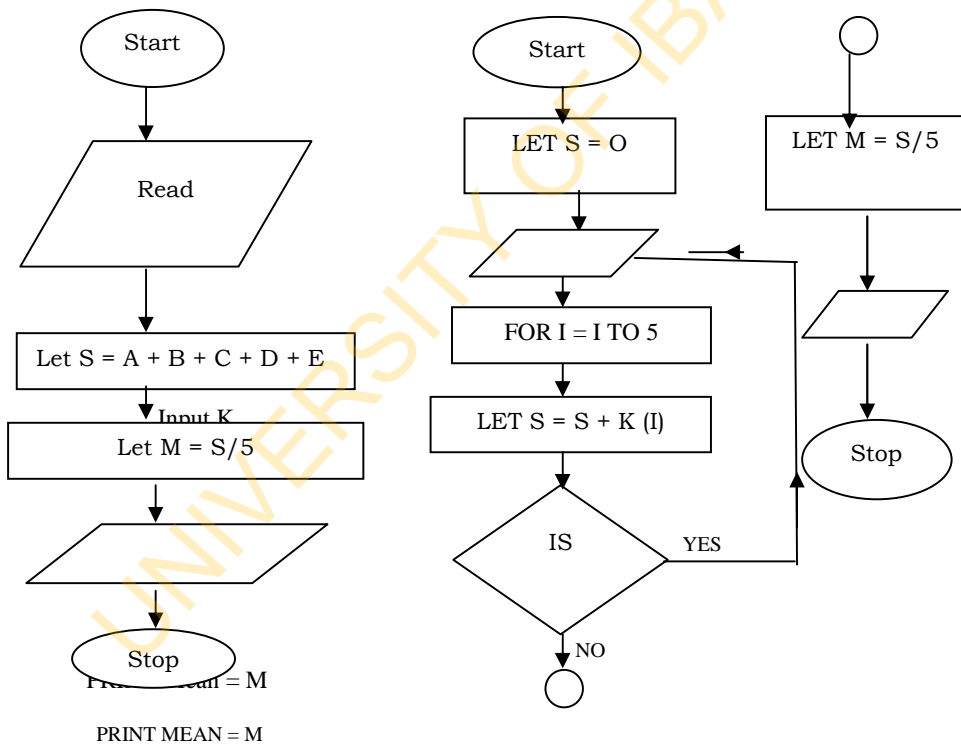


Example 4: Draw a flowchart for example 1.



No

Example 5: Draw a flowchart to calculate the mean of a set of 5 digits.



Summary for 8

- In solving a problem, the easiest way of approach is by the use of algorithm and flowchart.
- An algorithm is a step-by-step easy method of describing the solution to an event or process. It has no code other than the language you speak.
- On the other hand, a flowchart is a graphic method of operation of a process. It uses charts (a two-dimensional pictorial diagram) in describing an operation.
- The language used in a flowchart is not different from that used in the algorithm.
- Flowchart is easy to be understood especially as it is very useful in team work, where all members of the team need not be programming experts.

Self-Assessment Questions (SAQs) for study session 8

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

Post-Test

1. Write an algorithm and draw a flowchart to solve the following:
 - i. the area of a circle
 - ii. the area of a triangle given the base, height and an angle
 - iii. the quadratic equation using its formula
 - iv. the standard deviation of a set of 5 data
 - v. the geometric and harmonic mean of 5 digits
 - vi. the union of two sets given that the events are independent
 - vii. the sum of two probabilities given that the events are not independent.

2. Differentiate between an algorithm and flowchart and write down their uses.

References

- Cooke D., Craven A. H. and Clarke G. M.: *Basic Statistical Computing*. Second Edition. Edward Arnold. A division of Hodder and Stoughton.
- Ojo S. O. (1991): *Introduction to Computer Science*. Revised Edition. Department of Computer Science, University of Ibadan, Nigeria.

Study Session 9: Review of the BASIC Programming Language

Introduction

The BASIC programming language is an acronym for BEGINNERS ALL-PURPOSE SYMBOLIC INSTRUCTIONAL CODE. It is called the starter's code because it is a rudimentary lesson for the learning of programming language. It is easy to learn.

Instead of being phased out, as most early languages, it rather undergoes modification. That is the reason for the different versions we have today like the QBASIC, GWBASIC and visual BASIC.

The version we shall be reviewing here is the QBASIC, which shall, alongside the MS Excel, be the main software for our analyses here (especially as we get to the aspect of statistical analyses from the next chapter).

Learning Outcomes from Study Session 9

At the end of this study session, you should be able to:

- 9.1 identify a BASIC program;

Pre-Test

1. What are the type of instructions a BASIC program can accept?
2. What is a variable?
3. Return the output of the following program:

10 Read A, B, C, D

20 M = A + B + C + D

```
30    PRINT M
40    DATA 3.1, 2.7, 4.0, 1.5
50    END
```

9.1 The BASIC Program

The BASIC program is essentially made up of a set of instructions logically sequenced together. These instructions are common to all other programming languages, and they include;

- i. Input – this instruction reads data into the computer, which is considered variable because its constants are not known during program writing.
- ii. Processing – this instruction carries out arithmetic and logical operations on data.
- iii. Output – this instruction produces information from the data that has been processed.

To run a BASIC program, a BASIC interpreter is needed.

Fundamental Rules for Basic Programs

- i. Each instruction is written as a separate statement.
- ii. Every statement must appear on a separate line.
- iii. A statement cannot exceed one line in length.
- iv. Each statement must begin with a positive integer
- v. Successive statement must have increasing statement numbers.
- vi. Each statement number must be followed by a BASIC keyword, indicating the type of instruction to be carried out.

- vii. Blank spaces may be inserted as desired to improve readability.
- viii. Statements are executed in statement number sequence unless a deliberate “jump” is indicated.
- ix. On most computers, programs must end with an END statement.
- x. Numeric variable names can be represented by a letter or a letter followed by a digit.
- xi. Arithmetic operation are coded with the following symbols in a LET Statement: + (Addition), - (subtraction), *(multiplication), /(Division), ** or ^ (Exponentiation).
- xii. Numeric constant can be used in arithmetic statements
- xiii. To branch to a different place in a program, we use GOTO statement.

Fundamental Concepts of the BASIC Language

1. Character set- these are alphabet (A to Z), numeric (0 to 9) are special characters like ! @, \$, #, %, &, +, >, ?, {, and as on.
2. Constants - these are exact data elements supplied to BASIC which could be a string or numeric constant. String constants are enclosed in quotation marks. A numeric constant can be integer, fixed points, and/or floating point.

Integer constants are whole numbers with or without a prefix + or – sign. Fixed point constants are positive or negative real (decimal) numbers.

Floating point constants are positives or negative number represented in exponential form.

Fixed point and floating point numbers can be either simple or double precision numbers. Single precision constants has seven or fewer digit with an exponentiation form using E, or a trailing exclamation mark (!), while a double precision constant has eight or more digits with an experimental form using D, or a trailing numbers sign (#).

Examples:

Single Precision

Double Precision

4815.0

2033411.2013

1.52!

2208.0#

45.7

3105112590

-2.11E-05

-1.23056D-07

3. Variables: A variable is a name that represents a number or string. It can be called a storage compartment in a program where a value is placed. A variable can be a numeric variable or string variable depending on the type of variable stored. Ideally, a numeric variable is a letter, or letter followed by an integer, while a string variable is a letter followed by a dollar sign.

Writing Simple Programs

The reader is expected to consult relevant BASIC programming books as this text is just a review. You are not expected to learn BASIC here, but its applications. However, listed below are BASIC statements:

REM	INPUT	READ-DATA	PRINT	LET
GOTO	IF-THEN	FOR-NEXT	ON-GO-TO	STOP
DIM	GOSUB- RETURN	END	RANDOMIZE	RESTORE
STEP	TO	TROFF	TRON	DOWHILE
CLS	ELSE	ENDIF	CALL	CLOSE
DEF	DO-	FUNCTION-	ONxGOSUB	OPEN file FOR

	LOOPUNTIL	ENDFUNCTION		
OUTPUT AS	PRINT USING	SELECT CASE	SUB-ENDING	WRITE

The arithmetic operators used by BASIC have been explained earlier. These are

Addition; +

Subtraction: –

Multiplication; *

Division: /

Exponentiation: ** or ^

While the relational operators include

Equal to: =

Less than: <

Greater than: >

Less than or equal to: <=

Greater than or equal to : >=

Not equal to: <>

Relational operators are used in logical or conditional statements.

Look at these simple programs:

Example 1:

Write a program that computes the area of a rectangle

```
10 REM This program computes the area of a rectangle
20 INPUT L, B
30 LET A = L * B
40 PRINT "Area is "; A
50 END
```

Example 2:

Assuming data is given for the length and breadth, say, 25 and 16 respectively. Write another simple program that can solve this problem.

```
10 REM this program computes the area of a rectangle
20 READ A, B
25 LET A = L * B
30 PRINT "Area is"; A
35 DATA 25, 16
40 END
```

The difference between these two programs is that while the first can solve any given data, the second is restricted to only two given data, i.e. 25 and 16. The INPUT statement is shown here to be more flexible than the READ –DATA statement.

Example 3:

The graduating diploma students of the University of Ibadan are to be determined whether they are eligible to apply for direct entry. Given that the pass mark is 60, write a simple BASIC program that determines this.

```
10 REM this program prints names of graduating diploma students and
15 REM determines whether they are eligible to apply for direct entry
20 INPUT N$, S
30 IF S >= 60.0 then 40
35 IF S < 60.0 then 50
40 Print N$, "Eligible to apply for direct entry"
50 Print N$, "Not Eligible to apply for direct entry"
55 GO TO 20
60 END
```

Example 4:

Write a program to compute the factorial of a number N.

```
10 REM Program computes the factorial of a number N
20 LET F = 1
30 INPUT N
40 FOR I = 1 TO N
50 LET F = F * X(I)
60 NEXT I
70 PRINT "N! is"; F
80 END
```

Example 5:

The next program prints the sine and cosine of an angle and calculates its tangent.

```
5  REM program prints the sine and cosine of an angle, and
10 REM Calculates the tangent
20 INPUT K
30 LET A = SIN (K)
35 LET B = COS (K)
40 LET C = A / B
50 PRINT K
55 PRINT "Sine of K is "; A; "Cosine of K is "; B
60 PRINT "Tangent of K is"; C
70 END
```

Example 6:

Program that computes and prints group average score

```
5  REM program computers are prints group average score
10 FOR I = 1 To 10
15 LET F = 0
20 FOR J = 1 TO 30
25 INPUT K
30 LET F = F + K
35 NEXT J
40 LET A = F / 30
```

```
45 PRINT "Group"; I; "Average is"; A
50 NEXT I
60 END
```

Summary for 9

- BASIC programming language is a starter's code meaning BEGINNERS ALL-PURPOSE SYMBOLIC INSTRUCTIONAL CODE.
- BASIC is easy to learn and is not phased out, but rather undergoes modifications.
- It is made up of a set of instructions (just like other programming languages) logically sequenced together.
- These are the input, processing and output instructions.
- BASIC has fundamental rules that are needed for it to work (a good grasp of these rules is needed).

Self-Assessment Questions (SAQs) for study session 9

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

Post Test

1. Write a program that computes the standard deviation of a named student.
2. Write a BASIC program that computes the simple interest of a named client.

3. Differentiate between the floating and fixed point constants.
4. Differentiate between the single and double precision constants.
5. Write a BASIC program that uses (a) INPUT (b) READ-DATA statements to solve the same problem.
6. Write out a
 - i. symbol you may use for a string variable.
 - ii. symbol you may use for a numeric variable

References

- Cooke D., Craven A. H. and Clarke G. M. O: *Basic Statistical Computing*. Second Edition. Edward Arnold. A division of Hodder and Stoughton.
- Ojo S. O. (1991): *Introduction to Computer Science*. Revised Edition. Department of Computer Science, University of Ibadan, Nigeria.

Study Session 10: Descriptive Statistics

Introduction

In the early days of statistics, not much was done with gathered data other than using the raw data itself to describe certain phenomenon in the form of charts and simple calculation like the averages, ratios, proportions, percentages and index numbers.

Because these simple result and charts describes or summarizes the data and the object it is representing, they are rightly called descriptive statistics. However, over the years it has undergone some expansions to include frequency distribution, measures of central tendencies (location, petition and dispersion, and measure of skewness and kurtosis).

Learning Outcomes from Study Session 10

At the end of this study session, you should be able to:

10.1 Explain the Rounding of Numerical Data

Pre-Test

1. Mention three charts you are familiar with in descriptive statistics, and sketch them.
2. Differentiate between the mean, median and mode.
3. Distinguish between grouped and ungrouped data.
4. What is a deviation?
5. List the deviations you are familiar with.

10.1 Rounding of Numerical Data

Statistical data comes from counting or measurement. Measurement does not always give accurate figures, so an approximate or round form is often appreciated. Hence, rounding is simply approximating numbers in such a way as to replace the affected digits by zero.

This makes the number clearer and understandable. For example, in giving the number of accident casualties, one may not be exact with the actual figure. Let us say, in actual fact 23 people really died in an accident, a round figure of 20 (say; not fewer than 20') may be given.

Look at these two results:

1. 4,034,713 candidates registered for University entrance examination in 1974.
2. 4,000,000 candidates registered for University entrance examination in 1974.

We could round to either of the following ways.

- i. Specific units, e.g. nearest 100, or 1000.
- ii. Specific significant figures, e.g. 3 significant figures.
- iii. Specific decimal places, e.g. 2 decimal places

Example 1: a. Round 213, 530 to the nearest

- i. 100
 - ii. 1000
 - iii. 4 significant figures
 - iv. 3 significant figures
- b. Round 213.3780 to the nearest 2 decimal places.
- c. Round 115.3051 to the nearest 3 decimal places.
- d. Round 115.3051 to the nearest 5 significant figures.

Solution a i. 213,500, a ii. 214,000, a iii. 213,500, a iv. 214,000

b. 213.38

c. 115.305

d. 115.31

Using the MS Excel, to round up a digit, the syntax is Round

Use this command to perform the exercises in example 1.

UNIVERSITY OF IBADAN LIBRARY

10.1.1 Error

In rounding of numerical data, we are bound to commit error. This is because the rounded figure is an approximate. Error is defined as the difference between the actual figure and the rounded figure. That is,

Error = Actual figure – rounded figure.

However, since it is possible to obtain a negative value which we are not interested in, we thus seek to obtain the absolute value, which ignores the negative sign. So we talk of **absolute error** instead of just 'error'. Therefore, when we talk about error, you should have it in mind that we are referring to absolute error.

Assume we define the actual figure as A, rounded figure as B, and absolute error as E, then.

1. Absolute error, $E = A - B$
2. Relative error, $R = E / B$
3. Percentage error, $P = E / B \times 100$ or $\frac{100E}{B}$

Using the MS Excel, we define the above (note that only the syntax is presented):

1. Absolute error E, =ABS(A-B)
2. Relative error R, =E/B
3. Percentage error P = (E/B)*100 or =(100*X)/B or =PERCENT(E/B) or =PERCENT(R)

In-Text Question

What is Error?

In-Text Answer

Error is defined as the difference between the actual figure and the rounded figure.

The Basic program for the errors defined above are given below:

1. Absolute error

```
10  REM Program calculates (Absolute) Error
15  INPUT A, B
20  LET E = A - B
25  PRINT 'Error is "; E
```

2. Relative Error

```
10  REM program calculates Relatives Error
20  INPUT A, B
30  LET E = A - B
35  LET R = E / B
40  PRINT "Relative Error is "; R
```

3. Percentage Error

```
10  REM Program calculates percentage Error
15  INPUT A, B
25  LET E = A - B
35  LET R= E / B
```



```

40   LET P = R * 100
50   PRINT "Percentage Error is"; P
60   END

```

Note: Write a single program that computes and prints these three errors.

10.1.2 Ratios and Percentages

Ratios are fractions or decimals that express variations in data, irrespective of actual or absolute sizes of the data. On the other hand, percentages are ratios per 100. If in a class of 50 students, 32 are male, the ratio and percentage of female students in the population is

$\frac{18}{50} = \frac{9}{25}$ or 0.36 and $\frac{18}{50} \times 100$ or $0.36 \times 100 = 36\%$ respectively.

Using MS Excel, since ratio is a fraction, we use the syntax for calculating a fraction (as in the case of reaction error).

Example 2: Consider the given data.

Item	Expenditure (₦)
Food	9,000
Clothing	2,000
Housing	3,500
Recreation	1,500
Savings	5,000
Others	4,000

- i. Use MS Excel to obtain the ratio and percentage of each item
- ii. Write a BASIC program to obtain the ratio and percentage of each item.
- iii. For (i) above, what would your expressions look like if the data are stored in cells A1 to A7 and B1 to B7 for each heading respectively. Let the ratios occupy cells C1 to C7 and the percentage cells D1 to D7.

Solution

i.

	Item	Expenditure	Ration	Percentage
	Food	9,000	=9000/sum()	=(9000/sum())*100
	Clothing	2,000	=2000/sum()	=(2000/sum())*100
	Housing	3,500	=3500/sum()	=(3500/sum())*100
	Recreation	1,500	=1500/sum()	=(1500/sum())*100
	Saving	5,000	=5000/sum()	=(5000/sum())*100
	Others	41,000	=4000/sum()	=(4000/sum())*100
		=sum(9000+ ... +4,000)		

ii.

10 REM Program computes and prints Ratios and Percentages of

15 REM The Expenditures of six items

20 READ A, B, C, D, E, F

30 LET S= A + B + C + D + E + F

40 FOR I = 1 TO 6

45 INPUT N\$, X

50 LET R = X / S

55 LET P= R * 100

```
60 PRINT " "; N$
65 PRINT "Ratio is"; R; "Percentage is"; P
70 NEXT I
80 END
90 DATA 900, 3500, 1500, 5000, 4000
```

Another method is given below

```
10 REM Program computer and prints Ratios and Percentages of
20 REM The Expenditures of six items
30 LET S = 0
40 FOR I = 1 TO 6
45 INPUT X
50 LET S = S + X
60 NEXT I
70 FOR J = 1 TO 6
75 INPUT N$, X
80 LET R = X / S
85 LET P = R * 100
90 PRINT " "; N$; "Ratio is"; R; "Percentage is"; P
100 NEXT J
110 END
```

This second program is more flexible than the first in that it can accept any data, unlike the first program that is restricted to the given data.

iii.

Item	Expenditure (₹)	Ratio	Percentage
Food	900	=B2/B8	=C2*100 or =PERCENT(C2)
Clothing	2000	=B3/B8	=C3*100 or =PERCENT(C3)
Housing	3500	=B4/B8	=C4*100 or =PERCENT(C4)
Recreation	1500	=B5/B8	=C5*100 or =PERCENT(C5)
Saving	5000	=B6/B8	=C6*100 or =PERCENT(C6)
Others	400	=B7/B8	=C7*100 or =PERCENT(C7)
	=SUM(B2:B7)		

Measures of Location

The Arithmetic mean (generally called the mean) is the average of all members of a data.

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

In Excel, the syntax is

$$= \text{AVERAGE}(X_1, X_2, \dots, X_n)$$

or
$$= (X_1 + X_2 + \dots + X_n) / n$$

or
$$= (X_1 + X_2 + \dots + X_n) * 0.5$$

or
$$= (\text{Sum}(X_1, X_2, \dots, X_n)) / n$$

or
$$= (\text{Sum}(X_1, X_2, \dots, X_n)) * 0.5$$

Using cell references, we write

= AVERAGE (cell 1: cell N)

or = (Sum(cell 1:cell N))/2

or = (Sum(cell 1:cell N))*0.5

Programming in BASIC, we write

```
10 REM mean of a set of data
20 INPUT N
30 S = 0
40 FOR I= 1 TO N
50 INPUT X
60 S = S + X
70 NEXT I
80 M = S/N
90 PRINT "Mean ="; M
100 END
```

A second method is using the READ-DATA statement when large data is to be computed. This is to avoid mistakes when typing in the data when prompted.

```
5 REM computation of mean six observations
10 S = 0
15 FOR I = 1 TO 6
20 READ X
```

```

25     S = S + X
30     NEXT I
35     M = S / 6
40     PRINT "Mean="; M
45     END
50     DATA 3.63, 5.71, 8.02, 6.62, 4.14, 1.91

```

Yet another way of writing the above is

```

10     REM program computes mean
20     READ N
30     S = 0
40     FOR I = 1 TO N
50         READ X
60         S = S + X
70     NEXT I
80     M = S / N
90     PRINT "Number of observation ="; N
100    PRINT "Mean ="; M
110    END
120    DATA 6
130    DATA 3.63, 5.71, 8.02, 6.62, 4.14, 1.91

```

We could also write MS Excel expression as well as BASIC program for geometric and harmonic mean.

The geometric mean is given as

$$\sqrt[n]{\prod_{i=1}^n X_i} \quad \text{or} \quad \frac{1}{n} \sum_{i=1}^n \ln X_i$$

MS Excel returns the geometric mean for its syntax;

$$= \text{GEOMEAN}(\text{cell 1:cell N})$$

or $= \text{GEOMEAN}(X_1, X_2, \dots, X_N)$

Also, we can first find the product:

$$= \text{PRODUCT}(X_1, X_2, \dots, X_N)$$

or $= (X_1 * X_2 * \dots * X_N)$

Then we obtain the n^{th} root

$$= (\text{PRODUCT}(X_1, X_2, \dots, X_N))^{**}(1/N)$$

or $= \text{PRODUCT}(X_1, X_2, \dots, X_N)^{(1/N)}$

or $= (X_1 * X_2 * \dots * X_N)^{**}(1/N)$

or $= (X_1 * X_2 * \dots * X_N)^{(1/N)}$

Using the last BASIC programme for the arithmetic mean, we can write for geometric mean by modification.

10 REM Geometric mean

```

20  READ N
30  S = 1
40  FOR I= 1 TO N
50  READ X
60  S = S * X
70  NEXT I
80  K = 1 / N
90  M = S * K
100 PRINT "Number of observation ="; N
110 PRINT "Geometric mean="; M
120 END
130 DATA 6
140 DATA 3-63, 5.71, 8.02, 6.62, 4.14, 1.91

```

10.1.3 The Median

This is the middle number in a given set of numbers or of a frequency distribution when arranged in order of magnitude.

For an ungrouped data,

When odd, median, $\tilde{x} = x_m$ where x_m is the middle number.

When even, median, $\tilde{x} = \frac{x_m + x_n}{2}$, if $m \neq n$ and x_m, x_n are the middle numbers.

For a grouped data,

$$\text{Median, } \tilde{x} = L_1 + \left(\frac{\frac{N}{2} - \sum f_i}{f_{med}} \right) C$$

Where the symbols are as defined in the appropriate course.

The expression for median in MS excel is given as

$$= \text{MEDIAN}(\text{Cell 1:Cell N})$$

or $= \text{MEDIAN}(X_1, X_2, \dots, X_n)$

The Mode

This is the number that occurs most in a distribution.

For an ungrouped data

The mode is obtained by counting the value of the numbers separately for a grouped data

$$\text{Mode, } \hat{x} = L_1 + \left(\frac{D_1}{D_1 + D_2} \right) C$$

Where the symbols are as defined in the appropriate course.

The MS Excel command is

$$= \text{MODE}(\text{Cell 1: Cell N})$$

or $= \text{MODE}(X_1, X_2, \dots, X_n)$

Measures of Partition

The other name for the measures of partition is also 'Quantiles', which are a useful class of descriptive statistics. They are observations that divide, in specified proportions, the total frequency of a set of observations. The commonest of quantities is the 'median'. The various measures of partition used are:

1. The **quartiles** which divide the total frequency into quarters (or four parts)
2. The **deciles** which divide the total frequency into ten parts.
3. The **percentile** which divide the total frequency into one hundred parts.

In-Text Question

What is Mode?

In-Text Answer

This is the number that occurs most in a distribution.

For the MS Excel, we write only for quartiles and percentiles.

The Quartile

= QUARTILE (array, quart).

That is

= QUARTILE(Cell 1:Cell N, quart)

= QUARTILE (X₁, X₂, ..., X_n, quart)

The percentile

= PERCENTILE(array, K).

That is

= PERCENTILE(X₁, X₂, ..., X_n, K)

The general BASIC Program for any quartile is given below. Note that Q stands for the quantile, and P for the part of quantile. E.g. P=1, Q=2 result in the median, P= 25, Q= 100 results in the 25th percentile P=7, Q=10 results in the 7th decile, and P= 1, Q=4 results in the 1st quartile.

```
5   REM Quartiles of a set of data
15  REM INPUT: Data A (.), number of data N,
30  REM      P, Q for P-th Q-tile (P<Q)
45  REM OUTPUT: QO= P-th Q-tile
50  REM REQUIRES: Shellsort at 250
60  REM ARRAYS: A(N)
70  REM VARIABLES – Integer: J,K
85  GOSUB 250: REM Shellsort
90  J = INT(P * N / Q) + 1
105 K = N - INT((Q - P) * N / Q)
120 QO = (A(J) + A(K)) / 2
130 RETURN
```

The Shellsort program is not available in this book for now

Measure of Dispersion

This is simply defined as the disparity in a distribution. That is, how far an observation is from a particular (often central) observation. There are several of them which we are going to consider.

The Range: This is the difference between the maximum and minimum observation of a set of data. That is

$$\text{Range} = X_{\text{MAX}} - X_{\text{MIN}}$$

The procedure in MS Excel is by

- i. Obtaining the maximum value of the data
- ii. Obtaining the minimum value of the data
- iii. Obtaining their difference.

The **maximum** value is the result of

$$= \text{MAX}(X_1, X_2, \dots, X_n)$$

or $= \text{MAX}(\text{Cell 1:Cell N})$

The **minimum** value is the result of

$$= \text{MIN}(X_1, X_2, \dots, X_n)$$

or $= \text{MIN}(\text{Cell 1:Cell N})$

So that the range is the result

$$= \text{MAX}() - \text{MIN}()$$

The Quartile Deviation (MS excel only)

1. Compute the third quartile
 $= \text{QUARTILE}(\text{array}, 3)$
2. Compute the first quartile

$$= \text{QUARTILE}(\text{array}, 1)$$

3. Compute their difference and divide by 2 that is,

$$= (\text{QUARTILE}(\text{array}, 3) - \text{QUARTILE}(\text{array}, 1)) / 2 \text{ or}$$

$$= (\text{QUARTILE}(\text{array}, 3) - \text{QUARTILE}(\text{array}, 1)) * 0.5$$

The quartile deviation is also called the semi-interquartile range.

The Semi- Interpercentile range (Ms Excel only)

1. Compute the 90th percentile

$$= \text{PERCENTILE}(\text{array}, 90)$$

2. Compute the 10th percentile

$$= \text{PERCENTILE}(\text{array}, 10)$$

3. Compute the difference and divide by 2 just like in the quartile deviation.

Mean Deviation

This is the mean of sum of each of the absolute difference between an observation in a data set and the mean. This is given mathematically as

$$\frac{\sum |x_i - \bar{x}|}{n}$$

In MS Excel, the principle is to obtain the mean, and subsequently the deviation, from the mean, of each observation.

The syntax is (after the mean is obtained);

$$= ((X_1 - \text{AVERAGE}(X_1, X_2, \dots, X_n)) + \dots + (X_n - \text{AVERAGE}(X_1, X_2, \dots, X_n))) / 2, \text{ or}$$

$$= ((\text{cell 1} - \text{AVERAGE}(\text{cell 1: cell N})) + \dots + (\text{cell N} - \text{AVERAGE}(\text{cell 1: cell N}))) / 2$$

You can use the sum command as well.

The BASIC program for this is given below;

```
10  REM Program Computes mean Deviation
20  INPUT N
30  S = 0
40  D = 0
50  FOR I = 1 TO N
60  INPUT X
70  S = S + X
80  NEXT I
90  M = S / N
100 PRINT "Mean="; M
110 FOR J = 1 TO N
120 INPUT X
130 Y = X - M
140 Y = ABS(Y)
150 D = D + Y
160 NEXT J
170 P = D / N
180 PRINT "Mean Deviation="; P
190 END
```

Variance and Standard Deviation

The variance can also be called the 'mean squared deviation' while the standard deviation is also known as the 'root mean squared deviation'. They are so called because they derive their formulas from the mean deviation. Instead of the absolute deviations, we square the deviations, and sum them, and obtain the mean of the result. In the case of standard deviation we obtain the square root of the last result.

Using the formula for the population, we have that the variance σ^2 is given as

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N} = \frac{\sum_{i=1}^N X_i^2 - \left(\sum_{i=1}^N X_i\right)^2 / N}{N} \\ &= \frac{\sum_{i=1}^N X_i^2}{N} - \left(\frac{\sum_{i=1}^N X_i}{N}\right)^2\end{aligned}$$

The population standard deviation σ , is given as

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}} = \sqrt{\frac{\sum_{i=1}^N X_i^2 - \left(\sum_{i=1}^N X_i\right)^2 / N}{N}} \\ &= \sqrt{\frac{\sum_{i=1}^N X_i^2}{N} - \left(\frac{\sum_{i=1}^N X_i}{N}\right)^2} = \sqrt{\frac{\sum_{i=1}^N X_i^2}{N} - \left(\frac{\sum_{i=1}^N X_i}{N}\right)^2}\end{aligned}$$

We can approach the computation of the variance and standard deviation in MS Excel in two ways. First is by using the defined function. The second is by direct formula approach.

For the **variance**

1. By defined function approach

$$=VARIANCE(X_1, X_2, \dots, X_N)$$

or

$$=VARIANCE(\text{cell 1:cell N})$$

2. By direct formula approach follow the steps below

- i. Obtain the mean of data
- ii. Obtain the individual deviations.
- iii. Square the deviation in (ii)
- iv. Sum the squared deviations
- v. Obtain the mean of the squared deviation in (iv)

For the **standard deviation**;

1. By defined function approach

$$=STDEV(X_1, X_2, \dots, X_N), \text{ or}$$

$$=STDEV(\text{cell 1:cell N})$$

2. By direct formula approach, follow the steps for variance and then obtain the square root of the result for variance. You can use

$$=SQRT(\text{result of variance}), \text{ or}$$

$$=(\text{result of variance})^{*0.5}, \text{ or}$$

$$=(\text{result of variance})^{^0.5}$$

We shall write a single BASIC Program which shall calculate the population variance and population standard deviations.

```
10  REM Program computes Variance and Standard Deviations
20  INPUT N
30  S = 0
40  FOR I = 1 TO N
50  INPUT X
60  S = S + X
70  NEXT I
80  M=S/N
90  PRINT "Mean="; M
100 D = 0
110 FOR J = 1 TO N
120 INPUT X
130 Y = X - M
140 Y = Y ^ 2
150 D = D + Y
160 NEXT J
170 V = D / N
180 T = V ^ 0.5
190 PRINT "Variance="; V; "STANDARD DEVIATION=", T
200 END
```

You can learn how to write for the sample variance and standard deviations respectively.

Coefficients of Variation

This reduces the various measures into ordinary numbers so that comparison can then be made. We are going to consider three types.

1. Coefficient of Quartile Deviation = Quartile Deviation/Median. This is defined as

$$\frac{Q_3 - Q_1}{2} \bigg/ \frac{Q_3 + Q_1}{2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

2. Coefficient of Mean Deviation = Mean Deviation/Arithmetic Mean. This is defined

$$\text{as } \frac{\sum |x - \bar{x}|}{n} \bigg/ \frac{\sum x}{n} = \frac{\sum |x - \bar{x}|}{n\bar{x}}$$

3. Coefficient of Standard Deviation = Standard Deviation/Arithmetic Mean this is defined as

$$\frac{\sigma}{\bar{x}}$$

In most cases, the accepted form of the coefficient of standard deviation is $\frac{\sigma}{\bar{x}} \times 100$

This gives the percentage of the former result.

Note: You are expected to learn how to use MS Excel, and as well write a BASIC program to calculate these variations.

Diagrams, Charts and Graphs

There are numerous diagrams, charts and graphs in statistics that are very useful. In this text, we shall refer to them as graphs on a general term. Graphs are important in that they present a graphical (or pictorial) view of a distribution. A graph tells us the nature of a set of observations even before a summary statistics is obtained from them. The various graphs you are expected to review in this course include:

1. Bar chart

2. Column chart
3. Pie Chart
4. Histogram
5. Frequency polygon
6. Line graph
7. Cumulative frequency curve (Ogive)

Consult the MS Excel in drawing any graph. The graph commands are located in the Chart Wizard under the Insert menu.

Skewness and Kurtosis

Skewness is the degree at which a distribution departs from symmetry. Positive skewness means the distribution has a longer tail to the right than to the left of the central maximum.

The coefficient of skewness is defined as

$$\frac{\text{mean} - \text{mode}}{\text{standard deviation}}$$

But mode = mean – 3 (mean - median), so that

$$\text{coefficient of skewness} \equiv \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

For a univariate data X_1, X_2, \dots, X_N , skewness is defined as

$$\frac{\sum_{i=1}^N (X_i - \bar{X})^3}{(N-1)s^3} \quad \text{where } s \text{ is the standard deviation}$$

The MS Excel command for skewness is

$$=\text{SKEW}(X_1, X_2, \dots, X_n)$$

or =SKEW(cell 1:cell N)

Kurtosis, on the other hand, shows the highest peak of a set of random variables. That is, the highest point on the graph.

For a univariate data X_1, X_2, \dots, X_N , kurtosis is defined as

$$\frac{\sum_{i=1}^N (X_i - \bar{X})^4}{(N-1)s^4} \quad \text{where } s \text{ is the standard deviation}$$

The kurtosis of the standard normal distribution is 3, so therefore excess kurtosis is defined as

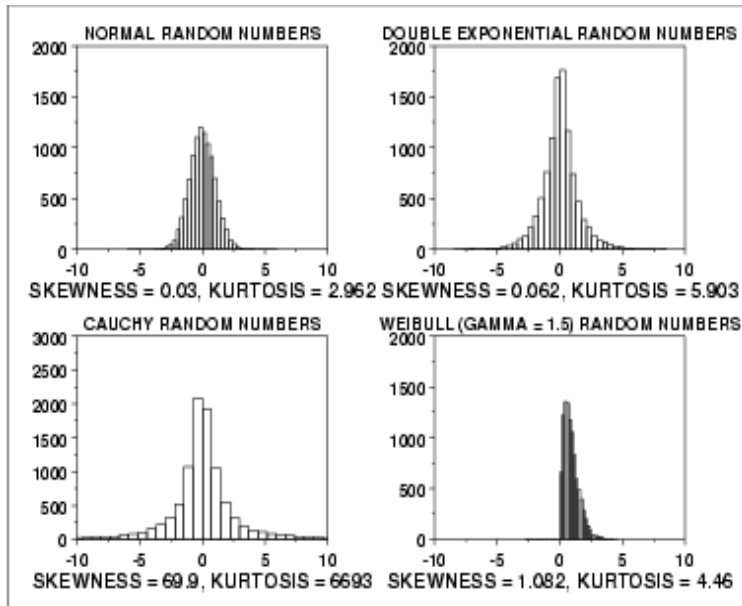
$$\frac{\sum_{i=1}^N (X_i - \bar{X})^4}{(N-1)s^4} - 3 \quad \text{which results gives zero.}$$

The MS Excel command for kurtosis is

=KURT(X_1, X_2, \dots, X_n)

or =KURT(Cell 1:Cell N)

See graphic examples below;



Summary for 10

- Descriptive statistics is all about summarizing a data into quantities that could easily be used for comparison, as well as hypothesis testing.
- There are various statistics ranging from the mean, ratio, and measures of central tendencies to such ones like the skewness and kurtosis of a distribution.

Self-Assessment Questions (SAQs) for study session 10

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

Post-Test

1. Write a BASIC program that computes the harmonic mean.
2. Write a BASIC program that computes the mean deviation.

- 3 (i) Write a stepwise MS Excel procedure (ii) Write a BASIC program that computes (i) the coefficient of mean deviation, and (ii) the coefficient of standard deviation.
4. The life of 100 electric lamps was recorded in hours, to the nearest hour as in the following table (after deducting 600 hours from each observed value):

90	101	122	84	60	99	115	142	126	116
128	105	93	91	88	106	107	91	101	113
140	62	76	138	114	103	95	92	99	63
98	87	103	82	99	92	114	124	81	85
94	105	107	73	117	107	96	97	119	89
112	133	105	91	94	116	145	92	124	108
79	80	54	96	69	85	125	104	80	101
89	102	110	96	97	109	121	77	90	114
85	124	136	126	88	92	128	56	100	71
102	96	108	98	110	82	94	76	64	95

Use MS Excel to:

- i. Obtain the mean, median, mode, variance, standard deviation the quartile deviations, the deciles deviations, the skewness and kurtosis of the distribution.
- ii. Obtain the geometric and harmonic means, and compare your result with the arithmetic means in (i)
- iii. Draw its histogram and frequency polygon.

(State your steps).

References

- Adamu S. O. and Johnson T. L. (1985): *Statistics for Beginners, Book 1*. Second Edition. Lagos: KOLA Publishers Limited, Nigeria.
- Cooke D., Craven A. H. and Clarke G. M. O: *Basic Statistical Computing*. Second Edition. Edward Arnold. A division of Hodder and Stoughton.
- Omotosho Y. (1990): *College and University Text Statistics*. Ibadan: NPS Educational Publishers Limited, Nigeria.

Study Session 11: Probability Theory

Introduction

Have you ever made a statement and somebody asks you 'are you sure?' There is always an air of uncertainty over every statement or event. In verifying the possibility of an event, you make statements about its probability. Probability is the level of certainty of an event. It is the verification of the certainty of an event as compared to a surer event. The probability of a sure event is 1, while that of an impossible event is 0. It therefore means the probability of any event lies between 0 and 1.

Learning Outcomes from Study Session 11

At the end of this study session, you should be able to:

11.1 Explain the Set Theory.

11.2 calculate permutation and combination of any number.

Pre-Test

What command in MS Excel returns the number of observations in a set or an array?

Write and return the result of the expression, =COMBINE (5,2).

11.1 Set Theory

This is a collection of well-defined objects. Each element of a set must be related to others. There must each be characteristic of the phenomenon that brings them together. An example is the set of numbers on the faces of a fair die; 1, 2, 3, 4, 5, 6, or the set of outcome in the throwing of a fair coin; H, T.

In each case, the set can be written as $S = (1, 2, 3, 4, 5, 6)$ and $S = (H, T)$.

The *sample space* of an experiment is the set of all possible outcomes in an experiment, also called universal set U. In the casting of a fair die above, the sample space is $S = (1, 2, 3, 4, 5, 6)$ while that of throwing a fair coin is $S = (H, T)$.

The *sample point* is a member of the sample space. The sample point is also called an *element*. Like in our example above, 4 is a sample point or element in the throwing of a fair coin.

An *Event* is possible outcome associated with an experiment. Thus, (1, 3, 5) is an event of odd numbers occurring in the experiment of casting of fair die.

The *Number* of elements in a set is obtained by counting. This is calculated using MS Excel using the COUNT function.

The syntax is

$$=COUNT(X_1, X_2, \dots, X_n)$$

Suppose $S = (a, b, c, d)$, $n(S) = 4$ which we can obtain on MS Excel by using the COUNT function = COUNT (Cell 1:Cell 4) which returns 4.

Probability of Equally Likely Events

$$P(E) = \frac{n(E)}{n(S)}$$

This is defined as

$n(E)$ is the number of member in the event.

$n(S)$ is the number of members in the sample space.

We compute $n(E)$ as =COUNT(E) and $n(S)$ as =COUNT(S), so that the $P(E)$ is written as =COUNT(E)/COUNT(S)

In-Text Question

Which of the following MS Excel function is correct?

- A. COUNT
- B. COUT
- C. COUNTS
- D. NONE OF THE ABOVE

In-Text Answer

A.COUNT

Example 1: In the throwing of 2 coins, the sample space is given as $S = (HH, HT, TH, TT)$.

The event of at least a tail is given as $E = (HT, TH, TT)$. Therefore,

$n(S)$ is written as =COUNT(S) which returns 4, and

$n(E)$ is written as =COUNT(E) which return 3.

So that $P(E) = \frac{n(E)}{n(S)}$ written as =COUNT(E)/COUNT(S) which returns 0.75.

Example 2: In the last example, the event of just a tail is $E = (HT, TH)$. The probability of the event will then be $=\text{COUNT}(E)/\text{COUNT}(S)$ which returns 0.5.

Example 3: In the table below

Mark scored	10	15	20	25	30
No of students	7	18	7	3	5

We want to find the probability that a student scores (a) at most 15 marks (b) 20 and 30 marks inclusive.

The total number of students = $n(S) = \text{COUNT}(S) = 40$

$n(E) = \text{COUNT}(E) = 25$.

Therefore $P(E) = \text{COUNT}(E)/\text{COUNT}(S) = 0.625$.

b. $n(E) = \text{COUNT}(E) = 15$

Therefore, $P(E) = \text{COUNT}(E)/\text{COUNT}(S) = 0.375$.

In writing a BASIC program for this example, we want to exclude the marks scored.

10 REM Calculation of probabilities

20 READ A, B, C, D, E

30 N1 = A + B

40 N2 = C + D + E

```

50      S = A + B + C + D + D
60      P1 = N1 / S
70      P2 = N2 / S
80      PRINT "Prob (student scored ≤ 15 marks) ="; P1
90      PRINT "Prob (student scores 20 ≤ X ≤ 30) ="; P2
100     END
110     DATA 7,18,7, 3, 5

```

Union and Intersection of Sets

The union of two sets A and B denoted by $A \cup B$ consists of all members of the sets A and B irrespective of an element appearing or not appearing in both sets. $A \cup B$ reads “ A union B ” can also be read A or B meaning too, all members in set A or, in set B .

For example, if $A = (5, 7, 9)$ and $B = (1, 7, 5, 10)$. Then $A \cup B = (1, 5, 7, 9, 10)$.

On the other hand, the intersection of two sets A and B denoted by $A \cap B$ is all the members of set A which are also found in set B . In the last example,

$$A \cap B = (5, 7)$$

The BASIC program for *union* and *intersection* of sets can be written in the format below:

Union of sets

```

10  REM Program prints union of two sets
20  REM INPUT: Data X(.) and Y(.), number of data N, M
30  REM OUTPUT: Union of sets X(.), Y(.)
40  REM variables – Integer I, J
50  INPUT N

```

```

60  FOR I = 1 TO N
70  READ X(I)
80  NEXT I
90  INPUT M
100 FOR J = 1 TO M
110 READ Y(J)
120 NEXT J
130 PRINT "Number of observations for the X's =" ; N
140 PRINT "Number of observations for the Y's =" ; M
150 PRINT "Union of the X's and Y's =" ; X(I), Y(J)
160 END
170 DATA X (1), X (2), ..., X(N)
180 DATA Y(1), Y(2), ..., Y(M)

```

Intersection of sets

```

10  REM Program prints intersection of sets
20  REM INPUT: Data X(.) and Y(.), number of data N,M
30  REM OUTPUT: Intersection of sets X(.) = Y(.)
40  REM variables – Integer I, J
50  INPUT N, M
60  FOR I = 1 TO N
70  READ X(I)
80  FOR J = 1 TO M
90  READ Y(J)
100 IF Y(J) = X(I) GO TO 160
120 NEXT J
130 NEXT I
140 PRINT "Number of observations for the X's =" ; N
150 PRINT "Number of observations for the Y's =", M
160 PRINT "Intersection of the X's and Y's =", Y(J)

```

```

170  END
180  DATA X(1), X(2), ... , X(N)
190  DATA Y(1), Y(2), ... , Y(M)

```

We write a simple BASIC program to return the number of elements in a set.

```

10  REM Program to count the number of elements in a set
15  INPUT N
20  FOR I = 1 TO N
30  READ X(I)
40  X(I) = I
50  IF I >= N GO TO 70
60  NEXT I
70  PRINT "Number of elements in the set ="; I
80  END
90  DATA X(1), X(2), ... , X(N)

```

With respect to example 1, we want to write a simple BASIC program to compute a probability.

```

10  REM Program calculates probability
20  REM INPUT: Date X(.) and Y(.), number of data N, M
30  REM OUTPUT: Probability
40  REM variables – Integer I, J
                – Real P
45  INPUT N, M
50  FOR I = 1 TO N
60  READ X(I)

70  X(I) = I

```

```

80  NEXT I
90  FOR J = 1 TO M
100 READ Y(J)
110  Y(J) = J
120  NEXT J
130  P = I / J
140  PRINT "Probability of .... ="; P
150  END
160  DATA X(1), X(2),..., X(N)
170  DATA Y(1), Y(2),..., Y(M)

```

A simple approach is given below:

```

10  REM Program calculates probability
20  INPUT N, M
30  P = N / M
40  PRINT "Probability of ...="; P
50  END

```

The difference between these last two programs is that the first counts the number of elements in the two sets before computing the probabilities. But the second inputs directly the number of elements in the two sets and computes the probability.

Addition Law of Probability

This is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The simple process in any application we wish to use is to

1. Calculate the probability of set A
2. Calculate the probability of set B
3. Calculate the probability of the intersection of sets A and B
4. Calculate the probability of the union of sets A and B by obtaining the result of the linear combination on the RHS.

In MS Excel, the procedure for computing $P(A)$, $P(B)$ and $P(A \cap B)$ is preceded by writing the command to count the elements in the universal sets, S , and the elements in sets A and B respectively, so that

$$P(A) = \text{COUNT}(A) / \text{COUNT}(S)$$

$$P(B) = \text{COUNT}(B) / \text{COUNT}(S)$$

and $P(A \cap B) = \text{COUNT}(A \cap B) / \text{COUNT}(S)$

Therefore, $P(A \cup B) = (\text{COUNT}(A) / \text{COUNT}(S)) + (\text{COUNT}(B) / \text{COUNT}(S)) -$

$$(\text{COUNT}(A \cap B) / \text{COUNT}(S))$$

For instance, consider the spreadsheet below

	B	C	D	E
--	---	---	---	---

	S	A	B	$A \cap B$
1	11	12	11	13
2	12	13	13	15
3	13	15	15	
4	14		16	
5	15			
6	16			
7				
8				
9	$n(S)$	$n(A)$	$n(B)$	$n(A \cap B)$
10		$P(A)$	$P(B)$	$P(A \cap B)$
11	$P(A \cup B)$			

You are required to store the results of the computations in the cells indicated.

Cell B9 is for $n(S)$, \therefore B9 =COUNT(S)

Cell C9 is for $n(A)$, \therefore C9 =COUNT(A)

Cell D9 is for $n(B)$, \therefore D9 =COUNT(B)

Cell E9 is for $n(A \cap B)$, $\therefore E9 = \text{COUNT}(A \cap B)$

Cell C10 is for $P(A)$, $\therefore C10 = C9/B9$

Cell D10 is for $P(B)$, $\therefore D10 = D9/B9$

Cell E10 is for $P(A \cap B)$, $\therefore E10 = E9/B9$

Finally,

Cell B11 is for $P(A \cup B)$, $\therefore B11 = C10 + D10 - E10$

11.2 Mutually Exclusive Events

It is possible that there is no intersection between A and B , that is the two events cannot occur simultaneously:

In such case $P(A \cap B) = 0$. This results in the addition law of probability becoming

$$P(A \cup B) = P(A) + P(B)$$

Independent Events

Here, the events A and B are independent of each other. That is the two events can exist simultaneously. The occurrence of one cannot prevent the occurrence of the other.

Therefore $P(A \cap B) = P(A)P(B)$, also called multiplication law of probability applicable to independent events. Thus the addition law of probability becomes

$$P(A \cup B) = P(A) + P(B) - P(A)P(B)$$

Conditional Probability

Two events cannot be both independent and mutually exclusive. It has to be one of the two. But it is not a general rule. One event occurring can be based on the occurrence of the other. An example is that a student is given a registration form on the condition that he has registered. Let the event of a student registering be R , and the event that he is given a registration form be F , so the conditional probability that a student is given a registration form given that he has registered is written $P(F/R)$. When this happens, the multiplication and addition laws are respectively written as

$$P(A \cap B) = P(A) P(B/A)$$

$$P(A \cup B) = P(A) + P(B) - P(A) P(B/A)$$

The Factorial Notation

The factorial of any number n is the product of n and all its immediate lower digits until 1. It is written as

$$n! = n(n-1)(n-2)(n-3)...4.3.2.1$$

In MS Excel, the factorial is calculated using the defined function FACT.

The syntax is =FACT(number)

Example 4: the factorial of 5 is calculated thus, =FACT(5) which returns 120.

The alternative is to multiply the digits 5,4,3,2 and 1.

The BASIC program is

```

10  REM Factorial of N
20  INPUT N, M
30  F = 1
40  FOR I = 1 TO N
50  READ X(I)
60  F = F * X(I)
70  NEXT I
80  PRINT "N factorial ="; F
90  END
100 DATA X(1), X(2),... X(N)

```

Permutation

This is an arrangement of a number of objects in a definite order. The factorial of N is an example of the arrangement of N objects in N ways if all the objects are to be taken at one time. If R of the N distinct objects are to be taken at a time, then the number of arrangement is known as permutation, denoted by

$${}^N P_R = \frac{N!}{(N-R)!}$$

So the permutation of 5 objects taken 3 at a time is given as

$${}^5 P_3 = \frac{5!}{(5-3)!} = \frac{5!}{2!} = 60$$

In MS Excel, the permutation of N objects taken R at a time is written as

$$=PERMUT(N, R)$$

E.g. 5P_3 =PERMUT(5, 3)

Its BASIC programming can be given as

```
10  REM permutation of N objects taken R at a time
20  INPUT N
25  INPUT R
30  F = 1
40  K = I
50  FOR I = 1 TO N
60  READ X(I)
70  F = F * X(I)
80  NEXT I
90  FOR J = 1 TO (N-M)
100 READ Y(J)
110 K = K * Y(J)
120 NEXT J
130 P = F / K
140 PRINT "N Permutation R ="; P
150 END
160 DATA 1, 2, 3, ..., N
170 DATA 1, 2 ..., (N-R)
10  REM Combination of N objects taking R
20  INPUT N
30  INPUT R
40  F = 1
50  K = 1
60  S = 1
70  M = N-R
```

```

80  FOR I = 1 TO N
90  READ A(I)
100 F = F*A(I)
110 NEXT I
120 FOR J = 1 TO R
130 READ B(J)
140 K = K*B(J)
150 NEXT J
160 FOR L = 1 TO M
170 READ C (L)
180 S = S*C(L)
190 NEXT L
200  P1 = K*S
210  P = F/P1
220  PRINT "N Combination R = ;"P
230  END
240  DATA A(1), A(2), ..., (A(N)
250  DATA B(1), B(2),..., B(M)
260  DATA C(1), C(2),..., C(M)

```

Combination

This is the selection of N objects taken R at a time. In this case order is not of any interest.

This is written as

$${}^N C_R = \frac{{}^N P_R}{R!} = \frac{N!}{R!(N-R)!}$$

Example: ${}^5 C_2 = {}^5 C_3 = \frac{5!}{2!3!} = 10$

In MS Excel, it is written as

$$= \text{COMBIN}(N, R)$$

E.g. ${}^5C_2 = \text{COMBIN}(5,2)$, and ${}^5C_3 = \text{COMBINE}(5,3)$. Both returns 10.

The BASIC format may be written as

```
10      REM Combination
20      INPUT N, M
30      F = 1
40      R = I
50      S = I
60      FOR I = 1 TO N
70          READ A(I)
80          F = F * A(I)
90      NEXT I
100     FOR J = 1 TO M
110     READ B(J)
120     R = R * B(J)
130     NEXT J
140     FOR K = 1 TO N - M
150     READ C(K)
160     S = S * C(K)
170     NEXT K
180     P1 = F
190     P2 = R * S
200     P3 = P1 / P2
```

```
210 PRINT "F Combination R ="; P3
220 END
230 DATA A(1), A(2), ..., (A(N))
240 DATA B(1), B(2), ..., B(M)
250 DATA C(1), C(2), ..., C(M)
```

Summary for 11

Probability is a game of chance. Probability of any event lies between 0 and 1. In computing the probability of any event, the procedure for counting the number of elements in the sets is very important. Permutation and combination are integral parts in the computation of many probabilities. While permutation deals with arrangement with regard to order, combination deals with selection with no regard to order.

Self-Assessment Questions (SAQs) for study session 11

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

Post-Test

1. Write a BASIC program that computes a combinatoric.

2. Given that A and B are two independent events with respective probabilities 0.3 and 0.45. Write a BASIC program to compute (a) $P(A \cap B)$ (b) $P(A \cup B)$ (c) $P(A/B)$.
3. Consider the universal set, $S = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$. Two events, A and B , are drawn from S . $A = (2, 4, 5, 6, 7)$, $B = (6, 1, 9)$. Use the MS Excel to compute (a) $P(A)$ (b) $P(B)$ (c) $P(A \cup B)$ (State clearly the steps needed to arrive at the results)

References

- Adamu S. O. and Johnson T. L. (1985): *Statistics for Beginners, Book 1*. Second Edition. Lagos: KOLA Publishers Limited, Nigeria.
- Cooke D., Craven A. H. and Clarke G. M.: *Basic Statistical Computing*. Second Edition. Edward Arnold. A division of Hodder and Stoughton.
- Omotosho Y. (1990): *College and University Text Statistics*. Ibadan: NPS Educational Publishers Limited, Nigeria.
- Ross S. (1994): *A First Course in Probability*. Fourth Edition. Prentice Hall, Englewood Cliffs, NJ 07632.

Study Session 12: Probability Distribution Functions

Introduction

The sum of probabilities of different events whose elements are members of a particular universal set always equals to one. The function that gives rise to the probability of each event is known as the ‘probability distribution function’.

In the case of the discrete random variables, the associated function is known as the probability mass function (p.m.f), while that of the continuous random variables is known as the ‘probability density function (p.d.f)’. Their sum (in case of p.m.f) and integral (in case of p.d.f) is called the ‘cumulative distribution function (c.d.f)’.

Learning Outcomes from Study Session 12

At the end of this study session, you should be able to:

12.1 Define a probability function;

1. compute the probabilities of a given function;
2. write a program for probability functions; and
3. draw a graph of probability functions and their c.d.f.

Pre-Test

1. Define a probability distribution.
2. Write out the formulas of three p.m.f and two p.d.f that you know.

12.1 Probability Function (The Discrete Case)

A probability function of X (a discrete random variable), $p(x)$, is a mathematical function which assigns positive values to the X 's in such a way that their sum is equal to 1. This makes a probability function look like a cumulative frequency distribution.

The difference is that while probability function is a rational number, in most cases, values that make a cumulative frequency distribution are real digits.

Consider the tossing of a fair coin once. The probabilities are given as

Outcome (X)	Probability $p(x)$
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$
Total	1

The function that gives rise to $\frac{1}{6}$ in each case is known as a probability function, while the total is known as the cumulative distribution function.

Therefore, let us denote the probability function as f , and the cumulative distribution function as F . Then for the discrete case, the probability that X takes the particular value x is given as

$$\Pr(X = x) = f(x)$$

And the probability that X is less than or equal to some particular value, say b , is given as

$$F(b) = \Pr(X \leq b)$$

For example

$$F(b) = \sum_{x=0}^b f(x) \quad (b \leq n) \quad x = 0, 1, 2, \dots, n$$

The second line is for the particular case when x runs from 0 to n .

Also for the *continuous case*, the probability that X takes a value in the interval (a, b) is given as

$$\Pr(a \leq X \leq b) = \int_a^b f(x) dx$$

And the probability that X is less than or equal to some particular value, say b , is given as

$$F(b) = \Pr(X \leq b) = \int_{-\infty}^b f(x) dx$$

Mean and Variance of a Probability Function

Having known the values of X and their probabilities, we now set forth to obtain their mean and variances.

Consider the table below:

X	$p(x)$	$xp(x)$
x_1	$p(x_1)$	$x_1p(x_1)$
x_2	$p(x_2)$	$x_2p(x_2)$

x_3	$p(x_3)$	$x_3 p(x_3)$
\vdots	\vdots	\vdots
x_b	$p(x_b)$	$x_b p(x_b)$
Total	1	

For the discrete case

The mean of a probability function is given as

$$\begin{aligned}\mu &= x_1 p(x_1) + x_2 p(x_2) + x_3 p(x_3) + \dots + x_b p(x_b) \\ &= \sum_{x=1}^b x_i p(x_i)\end{aligned}$$

and the variance given as

$$\begin{aligned}\sigma^2 &= (x_1 - \mu)^2 p(x_1) + (x_2 - \mu)^2 p(x_2) + (x_3 - \mu)^2 p(x_3) + \dots + (x_b - \mu)^2 p(x_b) \\ &= \sum_{x=1}^b (x_i - \mu)^2 p(x_i)\end{aligned}$$

For the continuous case

The mean is given as

$$\mu = \int_{-\infty}^b x_i f(x_i) dx$$

and the variance is given as

$$\sigma^2 = \int_{-\infty}^b (x_i - \mu)^2 f(x_i) dx$$

Example 1

Use the MS Excel to compute the mean and variances of the outcomes of tossing two unbiased coins.

UNIVERSITY OF IBADAN LIBRARY

Solution

The possible outcome will include

				T	H			
				x	x			
HH	—	2H	—	OT	—	0	2	
HT	—	IH	—	1T	—	1	OR	1
TH	—	IH	—	1T				
TT	—	OH	—	2T	2	—	0	

We can use either outcomes (H or T) for our computation.

Store the results of outcome in cells, say D4 to D6, and their probabilities in cells, say E4 to E6.

D4	=	0	E4	=	0.25
D5	=	1	E5	=	0.5
D6	=	2	E6	=	0.25

Observe that the total of their probabilities is 1

Multiply the adjacent cells D_*E_ and store the result in F4 to F6

F4	=	D4*E4
F5	=	D5*E5
F6	=	D6*E6

Sum up the cells F4 to F6 and store the result in F7

$$F7 = \text{Sum}(F4:F6)$$

F7 is the mean of the outcomes of the throwing the two unbiased coins.

For the variance, obtain the deviations of each outcome and store in cells say G4 to G6

$$G4 = D4 - F7$$

$$G5 = D5 - F7$$

$$G6 = D6 - F7$$

Square these new results and store in say H4 to H6.

$$H4 = G4^2$$

$$H5 = G5^2$$

$$H6 = G6^2$$

Multiply each of the above with the probabilities and store in cells, say I4 to I6

$$I4 = H4 * E4$$

$$I5 = H5 * E5$$

$$I6 = H6 * E6$$

Sum up these products and store in cell, say I7

$$I7 = \text{Sum}(I4:I6)$$

I7 now is the variance of the outcomes of the throwing of two unbiased coins.

Discreet Random Variables

The Bernoulli and Binomial distributions

We are going to consider these two distributions because they are related, in that, the Bernoulli is the probability of x successes in a single trial, whereas the binomial is the probability of x successes in n trials (in both cases the probability of success is p , while that of failure is $q = 1 - p$).

The Bernoulli function is given as

$$f(x) = \binom{1}{x} p^x q^{1-x}, \quad x = 0, 1$$
$$= p^x q^{1-x}, \quad \text{since } \binom{1}{0} = \binom{1}{1} = 1$$

while that of the Binomial function is given as

$$f(x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \dots, n$$

In MS excel, to compute a binomial probability, the syntax is given as

=BINOMDIST(number_s, trials, probability_s, cumulative)

This is a direct computation. However, we can use a step wise computation (having known our p) by

1. computing the combinatory part
2. computing the p^x
3. computing the q^{n-x}
4. computing the product of the three results above.

For the Bernoulli distribution, the BASIC program below can be used compute the probabilities.

```
10    REM Bernoulli probability
20    INPUT X, P
30    READ A
40    Q = 1 - P
50    P1 = P ^ X
60    W = A - X
```

```

70     Q1 = Q ^ W
80     B = P1 * Q1
90     PRINT "Bernoulli Probability ="; B
100    END
110    DATA 1

```

The Binomial distribution BASIC program is similar to the Bernoulli. The only difference is that the combinatory part is also computed.

```

10     REM This program computes Binomial probability
20     REM combination
30     INPUT N, P, M
40     F = 1
50     R = 1
60     S = 1
70     FOR I = 1 TO N
80         READ A (I)
90         F = F * A(I)
100    NEXT I
110    FOR J = 1 TO M
120        READ B(J)
130        R = R * B(J)
140    NEXT J
150    FOR K = 1 TO N-M

```

```

160    READ C(K)
170    S = S * C(K)
180    NEXT K
190    P1 = F
200    P2 = R * S
210    P3 = P1 / P2
220    REM Binomial probability
230    Q = 1 - P
240    X = M
250    P4 = P ^ X
260    W = N - X
270    Q1 = Q^W
280    B1 = P3 * P4 * Q1
290    PRINT "Binomial Probability =" ; B1
300    END
310    DATA A(1), A(2), ..., A(N)
320    DATA B(1), B(2),..., B(M)
330    DATA C(1), C(2),... C(M)

```

The Poisson distribution

The p.m.f is given as

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \dots$$

The MS Excel syntax is given as

=POISSON(n, mean, cumul)

A do it by yourself method is to obtain the value of

1. λ^x (having being given λ)
2. The exponent of negative λ
3. The product of (1) and (2) above
4. The factorial of x
5. The quotient of the value in (3) and (4)

The BASIC program is given below:

```
10 REM Poisson Probability
20 REM Factorial of X
30 INPUT N, L
40 F = 1
50 FOR I = 1 TO N
```

```

60    READ A(I)
70    F = F * A(I)
80    NEXT I
90    X = N
100   B = L * X
110   C = EXP (L)
120   D = F * C
130   E = B / D
140   PRINT "Poisson Probability ="; E
150   END
160   DATA A(1), A(2), ..., A(N)

```

The Geometric distribution

The print is given as

$$\begin{aligned}
 f(x) &= p(1-p)^{1-x} \\
 &= pq^{1-x} \quad x = 1, 2, \dots
 \end{aligned}$$

The simple BASIC program for this distribution is

```

10    REM Geometric distribution
20    INPUT X, P
30    Q = 1 - P
40    W = 1 - x

```

```

50     Q1 = Q ^ W
60     G = P * Q1
70     PRINT "Geometric Probability ="; G
80     END

```

Continuous Random Variables

Most of the common continuous random variables have probability density functions that look complicated. However, their evaluation is straight forward. One line of program will serve to calculate each of these functions for a particular value of x (or b).

The Exponential distribution

The p.d.f is given as

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

and c.d.f

$$F(x) = 1 - e^{-\lambda x}$$

Look at this

```

10     REM Exponential distribution
20     INPUT X, L
30     DEF FNP = L * EXP (-L * X)
40     DEF FNC = 1 - EXP (-L * X)
50     PRINT "Probability density function ="; FNP
60     PRINT "Cumulative distribution function ="; FNC

```

70 END

The Normal distribution

The Normal distribution has p.d.f given as

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty$$

The MS Excel format for calculating the probability is given as

=NORMDIST(x, mean, std_dev, cum)

The syntax assumes that the mean and standard deviation is known.

You can also use a stepwise procedure.

The BASIC program for this function is given as follows:

```
10      REM The Normal Probability
20      INPUT X, U, R
30      READ P
40      D = 2 * P * R
50      D1 = SQR (D)
60      F = 1 / D1
70      Z = (X - U) / R
80      Z = (Z ^ 2) * 0.5
90      N = G * EXP (-Z1)
100     PRINT "Normal probability ="; N
```

110 END

120 DATA 3.1428571

Cumulative Function of Continuous Random Variables

Evaluating the cumulative distribution function may be odd and cumbersome since there is not always an equivalent explicit expression such as exists for the probability density function. So we resort to the computation of integrals. And the most appropriate approach is to use numerical integration of the probability density functions. We shall use the Simpson's Rule.

Simpson's rule for approximating the integral $\int_a^b f(x)dx$ is

$$\frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + 2f_4 + \dots + 4f_{n-3} + 2f_{n-2} + 4f_{n-1} + f_n)$$

which may be written as

$$\frac{h}{3}(S^{(1)} + 2S^{(2)} + 4S^{(4)})$$

where $S^{(i)}$, with $i = 1, 2$ or 4 , is the sum of the function values

which have coefficient i .

$$S^{(1)} = f_0 + f_n$$

which means, $S^{(2)} = f_2 + f_4 + \dots + f_{n-4} + f_{n-2}$

$$S^{(4)} = f_1 + f_3 + \dots + f_{n-3} + f_{n-1}$$

The BASIC programme for numerical integration by Simpson's Rule is given below.


```

610  REM Simpson's Rule
620  INPUT U, L
630  A = 1.0 E-6
640  N = 2: H = 0.5 * (U - L)
650  S1 = FNF (L) + FNF (u)
660  S2 = 0
670  S4 = FNF (0.5 * (L + U))
680  S = H * (S1 + 4 * S4)
690  W = S: N = N + N : H = H / 2
700  S2 = S2 + S4
710  S4 = 0 : I = 1
720  S4 = S4 + FNF (L + I * H)
730  I = I + 2
740  IF I <= N THEN 120
750  S = H * (S1 + 2 * S2 + 4 * S4)
760  IF ABS (S - W) > A THEN 90
770  S = S / 3

```

The Normal distribution (cumulative function)

We shall apply the Simpson's Rule in evaluating the cumulative distribution function of the normal distribution. The c.d.f is given as

$$F(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

The above function assumes standard normal distribution with mean 0, and variance 1.

To use Simpson's Rule we rewrite the integral in a symmetrical form about zero. That is

$$\begin{aligned} F(z) &= \left(\int_{-\infty}^0 + \int_0^z \right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \\ &= 0.5 + \int_0^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \end{aligned}$$

```

10      REM Normal probability (integration)
20      DEF FNF (x) = 0.39894228 * EXP (-0.5 * X ^ 2)
30      L = 0.0: u = z
40      GOSUB 610: REM Integration, Simpson's Rule
50      P = S + 0.5
60      RETURN

```

This program takes into account that the Simpson's Rule has been included in the program somewhere along the line. So the GOSUB – RETURN statement recalls it.

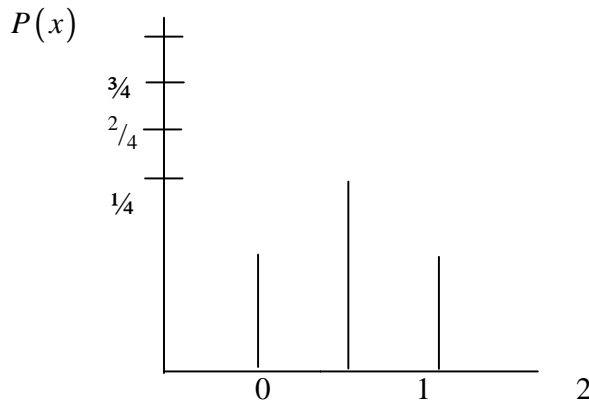
Chart or Graph of Random Variables

In plotting the graph of random variables, we make use of the probability of each event. We can also draw the graph of their cumulative distribution function.

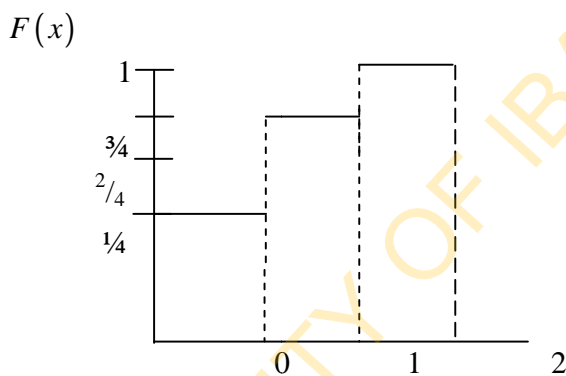
Example 2

The problem that deals with the throwing of two good and unbiased coins once could be used to illustrate this purpose.

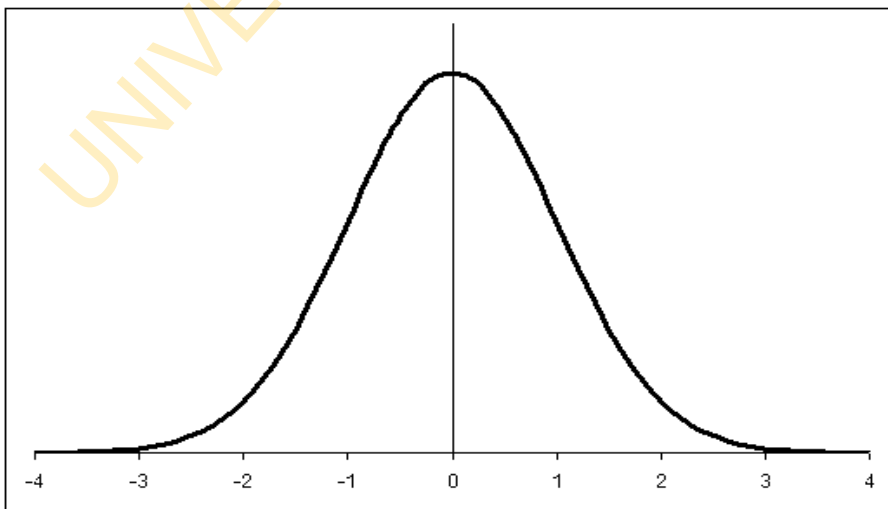
We found out that the graph of the probability distribution is as below:



While that of the cumulative functions is as below:



The broken lines show that it is a discrete function. An example of a cumulative function curve of a continuous random variable is that of the normal which is as given below:



Therefore, in plotting a graph of a probability mass function or probability density function, evaluate the various probability and plot against the values (of x) given. And in plotting the graph of cumulative distribution function, sum up the various probabilities and plot their points as the summation progresses.

The MS Excel chart wizard is a good tool in plotting these graphs.

Summary for 12

- Probabilities are events under uncertainty.
- The sum of all probabilities of events from the same universal set equals 1.
- The function giving rise to the various probabilities is known as the probabilities mass function (p.m.f) for the discrete case, and the probability density function (p.d.f) for the continuous case.
- Their sum or integral is known as the cumulative distribution function. In evaluating the cumulative distribution function of the continuous random variables (using BASIC), the numerical integration approach by Simpson's Rule is very useful.
- To draw the graph of a p.m.f or p.d.f and c.d.f, evaluate the various probabilities.

Self-Assessment Questions (SAQs) for study session 12

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

Post-Test

- 1 i. Define a probability mass function (p.m.f)
- ii. Define a probability density function (p.d.f)

iii. Define a cumulative distribution function (c.d.f)

2. If 40% of cocoa seeds brought by a produce buyer are defective, write a BASIC program to compute the probability that out of 5 cocoa seeds selected at random, (a) exactly 3 are defective (b) at most 2 are defective.

5% of the inhabitants of a village who were attacked by cholera died. Write a BASIC program to compute the probability that out of a sample of 60 cholera patients selected at random in the village (i) exactly 2 (ii) at least 2 (iii) less than 2 will die.

The mean life span of a number of cancer patients treated by one specialist is 72 years. If the life span of the patients are normally distributed with standard deviation 3.5, use the MS Excel to compute the probability that any cancer patient treated by the specialist will

live for less than 69 years

live for more than 76 years

live for between the ages of 74 and 80 years

(Write out the steps you need to take)

Given that n is large, and p is the probability of success, write a BASIC program that computes the standardized Z – value of the binomial distribution.

Write (i) an MS Excel expression (ii) a BASIC program

to evaluate $\sqrt{(3)}$.

References

Adamu S. O. and Johnson T. L. (1985): Statistics for Beginners, Book 1. Second Edition. Lagos: KOLA Publishers Limited, Nigeria.

Cooke D., Craven A. H. and Clarke G. M.: Basic Statistical Computing. Second Edition. Edward Arnold. A division of Hodder and Stoughton.

Omotosho Y. (1990): College and University Text Statistics. Ibadan: NPS Educational Publishers Limited, Nigeria.

UNIVERSITY OF IBADAN LIBRARY

Study Session 13: Correlation and Linear Regression

Introduction

You often find variables that are related, especially in economics. For example, it is known that income and expenditure have a relationship.

It may not be good enough to examine how some variables affect others that they are related with, but we can examine the degree of their association.

The degree of association of related variables is called Correlation, while that of effect of one variable upon the other(s) is called Regression. We shall approach this chapter, and subsequent ones using the Microsoft Excel.

Learning Outcomes from Study Session 13

At the end of this study session, you should be able to:

13.1 Explain the Theory of Correlation

Pre Test

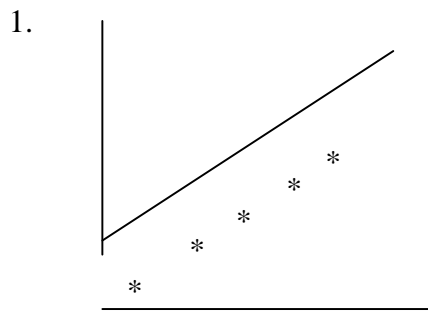
1. Mention five pairs of variables that are related with each other.
2. Enumerate the types of correlation you know
3. List the types of correlation coefficients you are familiar with.
4. What do you understand by the term 'Regression'?

13.1 The Theory of Correlation

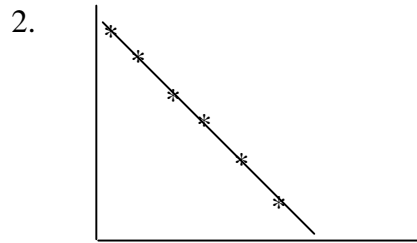
Correlation is simply the degree of association between two variables. To obtain a meaningful correlation, the variables must be related. There are different types of correlation which are described below.

1. Perfect and positive correlation - This is the case where all points of the scatter diagram fall on a straight line which slopes upwards from left to right, or downwards from right to left.
2. Perfect and negative correlation - This is the case where all points of the scatter diagram fall on a straight-line which slopes upwards from right to left, or downwards from left to right.
3. Positive or direct correlation - This is the case where one variable increases as the other increases in a left-right direction (i.e. the line slopes upwards from left to right).
4. Negative or inverse correlation - This is the case where one variable increases as the other decreases, or one variable decreases as the other increases. The line slopes downward in a left- right direction, or upwards in a right-left direction.
5. Null correlation-This is the case where there is no definite pattern in the direction of the two variables.

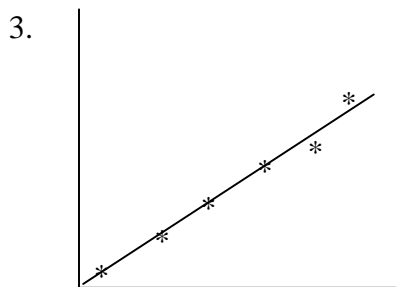
The different correlation could be illustrated with a diagram



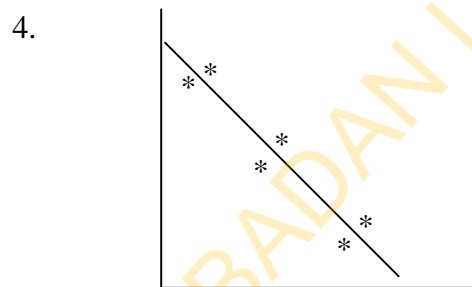
Perfect and Negative
Correlation



Perfect and positive Correlation



Positive or direct
Correlation



Perfect or Inverse
Correlation



Null correlation

13.2 Coefficient of Correlation

There are different types of correlation coefficient, of which two are very paramount. These two are:

1. The spearman's rank correlation coefficient.
2. The Karl Pearson's Product Moment Correlation Coefficient.

The most popular, and often used, of the two is the Product Moment Correlation Coefficient.

The spearman's rank correlation coefficient is given as

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

The Karl Pearson's Product moment correlation coefficient is given as

$$\begin{aligned} \rho &= \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} \\ &= \frac{\sum XY - \sum X \sum Y}{\sqrt{[\sum X^2 - (\sum X)^2][\sum Y^2 - (\sum Y)^2]}} \end{aligned}$$

ρ is the population correlation coefficient. Most often we compute the sample correlation coefficient because the population variance is often not known. We rather use the estimated value given as $\hat{\rho}$, or simply r .

Our interest here is not to show how these coefficients of correlation are derived, but its application in MS Excel. The syntax for this expression is given as:

=Correl(array 1, array 2)

Assume your arrays are in cell B9 to K9, and B10 to K10, then we can obtain the correlation coefficient thus:

= CORREL (B9: K9, B10: K10)

The result that will be obtained will lie between -1 and + 1 inclusive. + 1 shows perfect and positive relationship, while -1 shows perfect and negative relationship. 0 shows null relationship.

It should be noted here that MS Excel uses the Karl Pearson's Products moment correlation coefficient to compute the relationship. You can also work it out yourself by using the formula.

Example 1

Determine the correlation coefficient between the given data

X	X_1	X_2	X_3	X_4	X_5
Y	Y_1	Y_2	Y_3	Y_4	Y_5

Solution:

- i. Enter the variables in an MS Excel spreadsheet

	D	E	F	G	h
1	X	Y	X^2	Y^2	XY
2	X_1	Y_1	X_1^2	Y_1^2	X_1Y_1
3	X_2	Y_2	X_2^2	Y_2^2	X_2Y_2
4	X_3	Y_3	X_3^2	Y_3^2	X_3Y_3

5	X_4	Y_4	X_4^2	Y_4^2	X_4Y_4
6	X_5	Y_5	X_5^2	Y_5^2	X_5Y_5
7	$\sum_{i=1}^5 X_i$	$\sum_{i=1}^5 Y_i$	$\sum_{i=1}^5 X_i^2$	$\sum_{i=1}^5 Y_i^2$	$\sum_{i=1}^5 X_iY_i$
8	\bar{X}	\bar{Y}			

ii. Using the defined function, we write

$$= \text{CORREL}(D2:D6, E2:E6)$$

However, if you want to use the formula, you can go ahead to obtain the other results in columns F, G and H, as well as the results in D7 to H7, as well as D8 and E8, so that we could use either of the two formula.

$$\sum XY = H7, \quad \sum X = D7, \quad \sum Y = E7, \quad \sum X^2 = F7, \quad \sum Y^2 = G7$$

Therefore, the coefficient of correlation is written as

$$= (H7 - (D7 * E7)) / (((F7 - (D7^2)) * (G7 - (E7^2)))^{0.5})$$

This looks a little bit clumsy. You can simplify it by solving the arguments one by one, storing them in different cells, and further using the new results in the new cells.

Another way of approaching this problem is to use the defined function for variance and covariance. That is, for the numerator (store in cell, say F9);

$$F9 = \text{COVAR}(D2:D6, E2:E6)$$

And for the denominator, F or the X variables (store in cell, say G9);

$$G9 = \text{VAR}(D2:D6)$$

And for the Y variables (store in cell say, H9);

$$H9 = \text{VAR}(E2:E6)$$

Find the product of the resulting variances and store in all cell,

$$F10 = G9 * H9.$$

And obtain the square not of F10, store in cell G10, say,

$$G10 = \text{SQR}(F10)$$

This finally gives the correlation as

$$= F9 / G10$$

Another easier way, after obtaining the covariance, instead of the variances, we look for the standard deviations of the variables. That is, $G9 = \text{STDEV}(D2:D6)$

$$\text{and } H9 = \text{STDEV}(E2:E6)$$

so that the correlation becomes

$$= F9 / (G9 * H9).$$

In-Text Question

List the different type of coefficient of correlation.

In-Text Answer

Spearman's Rank correlation coefficient

Karl Pearman's Product Moment Correlation coefficient

The Theory of Linear Regression

Let us start this section by mentioning the equation of a straight line given by

$$y = mx + c, \quad \text{where}$$

y = dependent variable

m = slope (gradient) of the line

x = independent variable

c = intercept on the y -axis

This equation shows the relationship between the variables x and y . According to the definition, we see that the result of y depends on what x is with the fact that there is a gradient between them, and an interruption. There are cases where the intercept does not exist. That is, the line of relationship passes through the origin. In this case, $c = 0$ so that $y = mx$.

But in regression theory, the intercept is common because it is uncommon to have a relationship that will pass through the origin. The equation $y = mx + c$ is called the regression line, though often written in the form $y = a + bx$, where a is the interrupt, b the slope, and x and y are as defined.

The regression line is the best line that could be drawn to fit into a scatter diagram. This is because not all the points of a scatter diagram of a bivariate data (x, y) very often lie on a straight line. Listed are some of the methods used to fit a regression line.

1. The free hand method
2. The grand mean method
3. The semi-average method
4. The least squares method

The one we shall be interested in for the purpose of this course is the least squares method. (You should familiarize yourself with the other methods.) The regression line obtained here is unique.

The name ‘least squares’ is so called because we determine the slope and intercept by minimizing the error sum of squares. The actual regression line is $y = a + bx + e$, where e is the error or variation between the variable y and its estimated value \hat{y} . That is

$$e = y - \hat{y}, \quad \text{so that}$$

$$e^2 = (y - \hat{y})^2, \text{ and}$$

$$\sum e^2 = \sum (y - \hat{y})^2$$

The last expression is called the error sum of squares.

On minimizing the variation, we obtain the normal equations

$$\sum y = b \sum x + na \quad \text{--- (i)}$$

$$\sum xy = b \sum x^2 + a \sum x \quad \text{--- (ii)}$$

On solving the two simultaneous equation we obtain the estimated values for a and b (properly written as \hat{a} and \hat{b}).

$$\hat{b} = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

$$\hat{a} = \frac{1}{n} (\sum y - b \sum x) = \bar{y} - b\bar{x}$$

The estimated regression line becomes

$$\hat{y} = \hat{a} + \hat{b}x$$

In example 1, using the formula to find \hat{a} and \hat{b} , we proceed as follows

For $\hat{b} = ((5*H7)-(D7* E7))/(5*F7)-(D7^2))$

However, if the number of observations in variables is stored in a cell, say, C8, we can replace 5 by the cell address thus.

$$=((C8* H7)-(D7*E7))/((C8*F7)-(D7^2))$$

We can store this result for \hat{b} in a cell, say D9, so that

for $\hat{a} =E8-(D9*D8)$

Whatever value we obtain will be substituted in the estimated regression line.

However, in recent MS Excel packages, we can obtain a regression equation directly from the Tools menu in Data Analysis. Familiarize yourself with this new innovation in MS Excel and observe the results displayed.

You can also use defined functions for covariance and variance to compute \hat{b} (just like we did in the case of correlation).

Coefficient of Determination (R^2)

This is the amount of variation in the independent variable that explains the variation in the dependant variable, and is mathematically given as

$$R^2 = \frac{\text{explained variation squared}}{\text{total variation squared}}$$
$$= \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

You can go further to obtain this result using the MS Excel by applying the formula. This result is as well given in the Regression analysis in the Data Analysis under the Tools menu.

Summary for 13

- Related variables often have effect on each other, but normally are associated. But the degree of association is what is to be determined.
- Correlation is the degree of association; while the effect one variable has on other(s) is known as regression.
- You also learnt that to determine the extent an explained variable has on the unexplained, you determine the (the determination coefficient).

Self-Assessment Questions (SAQs) for study session 13

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

Post -Test

Use MS Excel (stating your procedure) to obtain

1. The Product moment correlation coefficient between

a. Botany (X) 3 6 4 6 4 7 5 5 4 7

Zoology (Y) 4 6 5 7 4 7 6 6 5 8

b. Reading (X) 4 6 5 5 4 6 6 6

Arithmetic (Y) 3 3 3 3 2 3 2 2

c. Agriculture (Y) 4.4 4.3 12.4 13.4 24.6 38.8 18.5

42.0 20.5

Health (X) 12.0 14.1 19.9 18.3 29.1 62.4 83.5
85.0 95.5

2. The Spearman's Rank correlation coefficient of the data in 1a - 1c.
3. The Regression equation of Y on X , and X on Y if the data in 1a - 1c, and also obtain their coefficients of determination, R^2 .

References

- Adamu S. O. and Johnson T. L. (1985): *Statistics for Beginners, Book 1*. Second Edition. Lagos: KOLA Publishers Limited, Nigeria.
- Cooke D., Craven A. H. and Clarke G. M. O: *Basic Statistical Computing*. Second Edition. Edward Arnold. A division of Hodder and Stoughton.
- Omotosho Y. (1990): *College and University Text Statistics*. Ibadan: NPS Educational Publishers Limited, Nigeria.

Study Session 14: Elementary Time Series Analysis

Introduction

A time series could be defined as data obtained overtime, that is, at equal intervals of time. This could be weekly, fortnightly, monthly, quarterly or yearly. However, a fortnight interval is not common since some months break into five weeks as against the traditional four weeks. Examples of time series are annual financial reports, monthly sales of particular good, quarterly products of export/import of products, or weekly collection of rainfall in a particular month.

Learning Outcomes from Study Session 14

At the end of this study session, you should be able to:

14.1 Components of a Time Series

1. observe the trend of a time series data; and
2. make simple forecast.

Pre-Test

1. What do you understand by variation in values?
2. What are the causes of variation in values?
3. How many models has a time series? Name them
4. What do you understand by trend?
5. How do you estimate trend?

14.1 Components of a Time Series

A time series data is never smooth. It is subject to variation which could be attributed to climatic, social, economic or accidental factors. These factors make up features of the four components of time series.

They are:

- i. Trend – This is the path a time series follow over a long period of time. It is best illustrated by the graph of the time series data.
- ii. Seasonal variation – This is as a result of climatic and social factors. It is usually (but not perfectly) regular, especially regular with seasons and climatic changes. The graph is up and down in movement.
- iii. Cyclical variation - This is caused by economic factor. It is usual to experience periods of booms, recesses and recovery. The graph is oscillatory in nature.
- iv. Irregular variation- This variation is accidental in nature. This is due to factors that do not occur on a regular basis, or may not be expected to occur, such as war, flood, drought, industrial actions, fire disasters, elections, special occasions, and so on. The graph is erratic in nature.

In-Text Question

The four component of Time Series are;

In-Text Answer

- Trend
- Seasonal variation
- Cyclical variation
- Irregular variation

14.1.1 Models of Time Series

There are two types of models that are most appropriate for associating the components of time series. These are

i. Additive model,
$$Y_t = T_t + S_t + C_t + I_{t,t}$$

ii. Multiplicative model,
$$Y_t = T_t S_t C_t I_{t,t}$$

where in both cases,

Y_t = Observed data

T_t = Trend values

S_t = Seasonal variation

C_t = Cyclical variation

I_t = Irregular variation

Estimating Trend

The methods that may be applied are:

- i. The free-hand method
- ii. The semi-average method
- iii. The least squares method
- iv. The moving average method

We shall consider the 'popular' least squares method, and the moving average method.

The least squares method is as described in the previous lesson under the linear regression theory. However, instead of the y in the estimated regression line, we may write T , which stands for the Trend. That is, in the normal regression equation, we write

$$y = a + bx + e$$

which can be written as

$$T = a + bx + e$$

and the estimated line $\hat{y} = \hat{a} + \hat{b}x$ becomes

$$T = \hat{a} + \hat{b}x$$

so that every calculated value of \hat{y} now becomes the trend. It should be noted that while the original y values may not be smoothly increasing, the trend of necessity is. This is because the trend is a linear combination of the x values, with the slope and intercept. However, the trend of a time series may not be a sufficient tool for prediction.

A typical time series data look like what we have below:

Year	Q	X	Y
1974	1	1	5.8
	2	2	2.1
	3	3	6.8
	4	4	51.1
1980	1	5	7.5
	2	6	1.2
	3	7	2.1
	4	8	34.8

1981	1	9	14.6
	2	10	1.5
	3	11	1.0
	4	12	40.6
1982	1	13	18.7
	2	14	21.5
	3	15	11.2
	4	16	35.1

The linear regression equation of this time series is given as

$$y_t = a + bx_t + e_t$$

and the estimated equation is

$$T = \hat{y}_t = \hat{a} + \hat{b}x_t$$

If you solve this, it results in

$$T = 8.02 + 0.94x$$

Confirm that the trend is as given below:

9.0, 9.9, 10.8, 11.8, 12.7, 13.7, 14.6, 15.5, 16.5, 17.4, 18.4, 19.3, 20.2, 21.2, 22.1, 23.1

Compare these results with the y values and comment.

The computer procedure is as described for the simple linear regression in the last session

The Moving Average Method

The procedure for computing the moving average depends on whether the desired moving average is to be odd or even. A good example of an odd moving average is the 3-point moving average, which is preceded by computing a 3-point moving totals.

The 3- point moving average is the desired trend. Using the data given earlier, we find out that the trend is given as

4.9, 20.0, 21.8, 19.9, 3.6, 12.7, 17.2, 17.0, 5.7, 14.4, 20.1, 26.9, 17.7, 22.6

An example of an even moving average is the 4-point moving average, which is preceded by the computation of the 4-point moving totals and the 2-4 point moving totals. The resulting results of the 4- point moving average is known as the trend. Confirm the results given below:

16.7, 16.8, 16.1, 13.4, 12.3, 13.2, 13.1, 13.7, 14.9, 18.0, 21.7, 22.3,

Compare these three trends and comment.

The procedure for computer analysis is very simple as it is all of addition and division. You are expected to carry out the computer analysis, stating the steps you have taken.

Built-In Trend Command

There is a built-in trend command as explained in lecture five. This is carried out in the Edit menu. Carry out this built-in solution and compare your results as well.

Summary of 14

- Time series data is that collected at equal interval of time.
- It is affected by certain factors which are summarized into four features, viz; trend, seasonal variation, cyclical variation, and irregular variation.

- The model of a time series could be additive or multiplicative. In estimating a trend, most used methods are the least squares and the moving average.

Self-Assessment Questions (SAQs) for study session 14

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

Post-Test

1. State the essential components of a time series data and give a mathematical relationship between each data unit and the various components. The data below shows the total domestic production of cements in Nigeria during the period 1974 – 1977. Figures are in thousand metric tons. Use the computer to perform the computation of the trend.

Year	Quarter			
	I	II	III	IV
1974	279	308	328	311
1975	471	381	225	311
1976	312	294	339	329

1977	305	309	315	332
------	-----	-----	-----	-----

Source: National Bureau of Statistics

2. Consider the time series data below

Year	Output	'000	Year	Output	'000
1950	30		1963	62	
1951	29		1964	69	
1952	31		1965	76	
1953	32		1966	71	
1954	33		1967	74	
1955	37		1968	81	
1956	41		1969	82	
1957	43		1970	85	
1958	41		1971	91	
1959	49		1972	93	
1960	55		1973	96	
1961	58		1974	103	
1962	60		1975	108	

Use the computer to

- i. Plot the time series data

- ii. Calculate a five-year moving average
- iii. Plot the moving average graph on the same graph as (i).

Reference

Omotosho Y. (1990): *College and University Text Statistics*. Ibadan: NPS Educational Publishers Limited, Nigeria.

UNIVERSITY OF IBADAN LIBRARY

Study Session 15: Statistical Tests and Confidence Intervals

Introduction

Making inference about a population is what is of utmost concern to the statistician. If a conclusion cannot be drawn over a given data, then no work is done. In making an inference about a population, you need to set up a hypothesis, always known as the null hypothesis, and then you make suitable assumptions about the sample data. Next is to obtain a suitable statistic, say t -statistic or F -statistic

This statistic must justify our assumptions, as well as having knowledge of the sampling distribution. You then calculated the value of the statistic of interest from our data, and compare the result with the tabulated value of the test statistic.

In finding a confidence interval for a population parameter, for example the mean, you need an estimate of the parameter, and its sampling distribution. In most cases the assumption holding for the sampling distribution is the normal or t distribution.

Learning Outcomes from Study Session 15

At the end of this study session, you should be able to:

15.1 Point and Interval Estimates

15.2 Explain the Hypothesis Testing.

Pre- Test

1. What in a hypothesis?
2. Define a Standard Normal Variate.

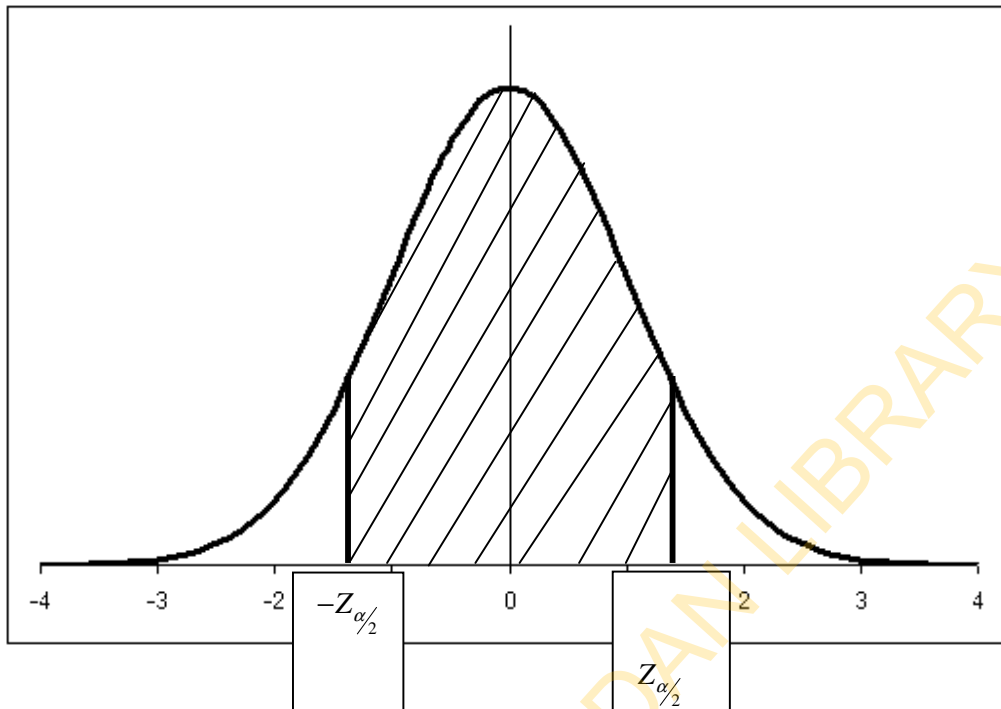
3. Define a t -test statistic.

15.1 Point and Interval Estimates

In study session ten, you were able to obtain estimates such as the mean. A population parameter estimated as a single number is known as a point estimate. On the other hand, if it lies between two numbers it is known as interval estimate. For example, $\mu = 3.5$ is a point estimate, and $2.5 \leq \mu \leq 4.5$ is an interval estimate of the population mean.

Confidence Interval and Limits for the Population Mean (μ)

1. **Large sample case-** The definition of the confidence interval for the mean in large samples is approached using the normal distribution. We know that μ and σ are respectively the population mean and standard deviation. The critical value is defined as $\pm Z_{\alpha/2}$, where Z is the standard normal variate. The acceptance region for μ must be within the bounds of the critical values inclusive. This is illustrated in the graph below.



The shaded region is the acceptance region. Setting up a confidence interval for μ , from the figure above, mathematically μ must be such that

$$-Z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}$$

which results

$$\bar{x} - Z_{\alpha/2} \sigma/\sqrt{n} \leq \mu \leq \bar{x} + Z_{\alpha/2} \sigma/\sqrt{n}$$

This is the confidence interval

From here we obtain the confidence limits as

$$\bar{x} \pm Z_{\alpha/2} \sigma/\sqrt{n}$$

A test could be 1-tailed or 2-tailed.

For 1-tailed, the 95% confidence interval and limits are given as

$$\mu \leq \bar{x} + 1.65\sigma/\sqrt{n} \text{ and } \bar{x} + 1.65\sigma/\sqrt{n} \text{ respectively for upper tails,}$$

and $\mu \geq \bar{x} - 1.65\sigma/\sqrt{n}$ and $\bar{x} - 1.65\sigma/\sqrt{n}$ respectively for lower tail.

For 2-tailed, the 95% confidence interval and limits are given as:

$$\bar{x} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{x} + 1.96\sigma/\sqrt{n} \text{ and } \bar{x} \pm 1.96\sigma/\sqrt{n} \text{ respectively.}$$

The standard normal variate is defined as

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

In MS Excel, the syntax for Z is given as

$$=STANDARDIZE (X, \text{mean}, \text{std_dev})$$

The form of Z used here is

$$Z = \frac{\bar{x} - \mu}{\sigma}$$

(Note, however, that σ/\sqrt{n} is the standard error) You could obtain the same result in both cases using a stepwise format. That is,

for $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$, having given the population and sample mean, as well as the standard

deviation and total sample number, we could write a general syntax as

$$Z = (\bar{X} - \mu) / (\sigma / \text{SQRT}(n))$$

Note that SQR(n) must have been solved earlier.

$$\text{OR } Z = (\bar{X} - \mu) / (\sigma / (n^{0.5}))$$

Since this is a syntax, it would be best to represent the parameters and statistics above by real letters. That is, let $X = \bar{X}$, $U = \mu$ and $R = \sigma$, so that

$$Z = (X - U) / (R / (N^{0.5}))$$

The second form of Z , that is $Z = \frac{\bar{x} - \mu}{\sigma}$, could also be written as

$$Z = (X - U) / R$$

Here is a BASIC program for $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

```
10  REM Standard Normal Variate
20  INPUT N, X, U, R
30  Z = (X - U) / (R / SQR(N))
35  PRINT "Standard normal Test"
40  PRINT "Mean of null hypothesis="; U
50  PRINT "sample mean="; X
60  PRINT "Standard error of the mean="; R/SQR(N)
70  Print "Z- statistic ="; Z
```

The standard error given as $R/SQR(N)$, where R denotes the standard deviation could also be written as $SQR(V/N)$, where V denotes the variance.

The confidence interval is not strenuous as well. For 2-tailed, the simple program is as follows;

```
10 REM Confidence interval for 2-tailed
20 INPUT N, X, R
30 READ Z
40 N1 = SQR(N)
50 I1 = X - (Z * (R / N1))
60 I2 = X + (Z * (R / N1))
70 PRINT "Upper Confidence Limit="; I2
80 PRINT "Lower Confidence Limit="; I1
90 PRINT "95% Confidence Interval for mean is"; I1; "<= U <="; I2
100 END
110 DATA 1.96
```

And for a 1-tailed tests,

```
10 REM Confidence Interval for 1-tailed
20 INPUT N, X, V
30 READ Z
40 S = SQR (V / N)
50 I1 = X + (Z * S)
60 I2 = X - Z * S
70 PRINT "Upper Tail Confidence Limit ="; I1
```

```

80 PRINT "Lower Tail Confidence Limit="; I2
90 PRINT "95% Upper Tail Confidence Interval is U <="; I1
100 PRINT "95% Lower Tail Confidence Interval is U >="; I2
110 END
120 DATA 1.65

```

While it is assured that there is not much difference between the two programs, the only difference that may be noticed is that 'R' is exchanged for 'V' in the second program. R stands for the 'standard deviation', and V stands for 'variance'.

2. **Small Sample Case-** The appropriate distribution used here is the *student-t* distribution, which become an alternative for the Z. Z is useful only when the population standard deviation is known. The parameter, mentioned is largely unknown. We rather make do with the estimate. This results in the new variate (*t*), given below as :

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

Where *s* is the sample standard deviation, and *n-1* is the degree of freedom, *n* being the sample size.

The MS Excel format for the *t-test* is

= TTEST(array1, array 2, tails, type)

The *t*- distribution has the syntax

=TDIST(x, deg_freedom, tails)

For the population mean μ , to all fall within the acceptance region, it must be such that

$$t_{\alpha/2} \leq \frac{|\bar{x} - \mu|}{s/\sqrt{n-1}}$$

This means that

$$-t_{\alpha/2} \leq \frac{|\bar{x} - \mu|}{s/\sqrt{n-1}} \leq t_{\alpha/2}$$

Therefore, for a 2-tailed, the confidence interval is

$$\bar{x} - t_{\alpha/2} s/\sqrt{n-1} \leq \mu \leq \bar{x} + t_{\alpha/2} s/\sqrt{n-1}$$

having confidence limits $\bar{x} \pm t_{\alpha/2} s/\sqrt{n-1}$

And for a 1-tailed, the upper and lower confidence intervals are

$$\mu \leq \bar{x} + t_{\alpha/2} s/\sqrt{n-1} \text{ and } \mu \geq \bar{x} - t_{\alpha/2} s/\sqrt{n-1}$$

respectively, having confidence limits

$\bar{x} + t_{\alpha/2} s/\sqrt{n-1}$ for the upper tail, and $\bar{x} - t_{\alpha/2} s/\sqrt{n-1}$ for the lower tail.

The value $t_{\alpha/2}$ and $-t_{\alpha/2}$ could be read from the table of t -distribution under $n-1$ degrees of freedom.

In MS Excel, confidence interval could be calculated directly using the syntax.

=CONFIDENCE(alpha, standard_dev, size).

However, you are encouraged to apply a do-it-yourself method as well so that you could compare results with the built-in facilities of the MS Excel.

Unpaired Sample- Difference of means

Assume we have two sample x and y with sizes n_1 and n_2 respectively. The t -test for these two samples from normal population is given as

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where \bar{x} and \bar{y} are means of each sample, with pooled variance S^2

$$S^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The degrees of freedom of the t distribution defined above is $n_1 + n_2 - 2$.

The BASIC program for this t - statistic involving two sample is

```

10 REM t- statistic for unpaired samples
20 INPUT X, Y, N1, N2, V1, V2
30 K = N1 + N2 - 2
40 S1 = (N1 - 1) * V1 + (N2 - 1) * V2
50 S2 = S1 / K
60 S1 = SQR (S2 / N1 + S2 / N2)
70 T = (X - Y) / S1
80 PRINT "t-statistic for unpaired samples ="; T
90 END

```

Example

The mean weight and standard deviation of 150 tins of vegetable cooking oil produced by an oil producing factory showed 3.8 litres and 0.25 litres respectively. Write a BASIC program

to compute (a) 2- tailed (b) 1- tailed (upper) (c) 1- tailed (lower) 95% confidence limits and intervals for the mean.

Solution: Since n is large (i.e. 150), normal distribution is applicable.

a.

```
10 REM Confidence limits and Intervals for 2- tailed
20 READ A, B, C, D
30 N1 = SQR ( A)
40 I1 = B - (D * (C / N1))
50 I2 = B + (D * (C / N1))
60 PRINT "Upper Confidence Limit ="; I2
70 PRINT "Lower Confidence Limit ="; I1
80 PRINT "95% Confidence Interval for Mean ="; I1; "<= U <="; I2
90 END
100 DATA 150, 3.8, 0.25, 1.96
```

b.

```
10 REM Confidence Limit and Interval for 1- tailed (upper)
20 READ A, B, C, D
30 N = SQR (A)
40 I = B + (D * (C / N))
50 PRINT "Upper tail confidence limit ="; I
60 PRINT "95% upper tail confidence Interval is U <="; I
```

```

70 END

80 DATA 150, 3.8, 0.25, 1.65

10 REM Confidence Limit and Interval for 1- tailed (lower)

20 READ A, B,C, D

30 N = SQR (A)

40 I = B - (D * (C / N))

50 PRINT "Lower tail Confidence Limit =" ; I

60 PRINT "95% Lower tail Confidence Interval is U >=" ; I

70 END

80 DATA 150, 3.8, 0.25, 1.65

```

You are encouraged to write a single program for (a), (b) and (c), as well as use the MS Excel to compute these results.

15.2 Hypothesis Testing

A statistical hypothesis is an assumption made on a population in the process of taking certain decision(s) on the population. The assumption made with the aim of rendering a statistical hypothesis insignificant is called 'null hypothesis'.

It assumes no difference in certain condition (or parameter of interest). Symbolically, it denoted by H_0 . Of necessity, another assumption is made to counter the null hypothesis, which is called the alternative hypothesis, denoted by H_1 or H_A . For example

H_0 : the 2 students are equally brilliant

H_1 : the 2 students are not equally brilliant.

In statistics, hypotheses are better stated using real values. For example:

$$H_0: \mu = 2$$

$$H_1: \mu > 2$$

In stating a hypothesis, errors might be committed (due to our 'imperfections'). There are two types of error. These are

1. **Type 1 error**, which is committed if a hypothetical value is taken to fall within the rejection region when it is supposed to fall within the acceptance region.
2. **Type 2 error**, which is committed if a hypothetical value is taken to fall within the acceptance region when it ought to fall within the rejection region.

In general, the following is the procedure for carrying out a hypothesis testing (our test-statistic of interest are the normal and t);

1. Make statement about your hypotheses, H_0 and H_1 .
2. Obtain the value for mean and standard deviation of the given population, or their estimates.
3. Compute the test statistic, Z or t , of the hypothetical value.
4. Determine the critical value(s) that corresponds to the given significance level and hence the acceptance region.
5. Compare your results in (3) and (4) and either accept or reject your H_0 as the case may be. If the computed test statistic is less than the critical value (i.e. table value), accept H_0 . Otherwise reject it.

Summary of 15

- No statistical analysis is computed if an inference cannot be drawn on the entire population given the sample data.
- In any sample or population, it is always important to know the mean and the standard deviation, in order to be able to carry out tests.
- However, the population standard deviation is largely unknown, so we limit ourselves to the estimate, that is, sample standard deviation.
- To carry out a test, a hypothesis, known as the null hypothesis must first be conjectured, which serves as springboard for the test. Accept or reject if your computed test- statistic falls within or outside the region of acceptance, respectively

Self-Assessment Questions (SAQs) for study session 15

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

Post -Test

1. Write a BASIC program to compute the 95% confidence limits and confidence intervals of both 2-tailed and 1- tailed (upper and lower).
- 2 a. Write stepwise the MS Excel format. b. write a BASIC program to carry out a test at 5% level to find whether or not a claim of an average return of ₦5550 by a sales representative could be regarded as the mean for all months if in 5 years he made an average monthly returns of ₦5550 with standard deviation of ₦250.
3. Rewrite all the BASIC programs in this lecture and assess yourself.

References

- Adamu S. O. and Johnson T. L. (1985): *Statistics for Beginners, Book 1*. Second Edition. Lagos: KOLA Publishers Limited, Nigeria.
- Cooke D., Craven A. H. and Clarke G. M.: *Basic Statistical Computing*. Second Edition. Edward Arnold. A division of Hodder and Stoughton.
- Omotosho Y. (1990): *College and University Text Statistics*. Ibadan: NPS Educational Publishers Limited, Nigeria.

Study Session 16: Introduction to MATLAB

Introduction

The name MATLAB stands for MATrix Laboratory. It is a high-performance language for technical computing. It integrates *computation*, *visualization*, and *programming* environment.

It is a modern programming language environment with sophisticated *data structures*, built-in editing and *debugging tools*, and supports *object-oriented programming*. These factors make MATLAB an excellent tool for teaching and research.

MATLAB has many advantages compared to conventional computer languages (e.g., C, FORTRAN) for solving technical problems. It is an interactive system whose basic data element is an *array* that does not require dimensioning. Specific applications are collected in packages referred to as *toolbox*.

There are toolboxes for signal processing, symbolic computation, control theory, simulation, optimization, and several other fields of applied science and engineering.

Learning Outcomes from Study Session 16

At the end of this study session, you should be able to:

1. Recognize MATLAB as scientific analytic software.
2. Understand the basic features of MATLAB.
3. Use MATLAB as a calculator.

Pre-Test

1. What is the meaning of MATLAB?
2. Differentiate between *who* and *whos* commands.
3. What are the hierarchies of arithmetic operations as used in MATLAB?

16.1 MATLAB

After logging into your account, you can enter MATLAB by double-clicking on the MATLAB shortcut *icon* on your Windows desktop. When you start MATLAB, a special window called the MATLAB desktop appears. The desktop is a window that contains *other* windows. The major tools within or accessible from the desktop are:

The Command Window

The Command History

The Workspace

The Current Directory

The Help Browser

You can customize the arrangement of tools and documents to suit your needs. Excellent graphics facilities are available, and the pictures can be inserted into LATEX and Word documents.

MATLAB can be used in a number of different ways or modes; as an advanced calculator in the calculator mode, in a high level programming language mode and as a subroutine called from a C-program.

To execute a command in MATLAB, type on the command prompt sign, “>>” when it is activated in the command window.

Command Line Help

Help is available from the command line prompt.

```
>> help command
```

To obtain help on *elementary math functions*, for instance, type

```
>> help elfun
```

This gives rather a lot of information so, in order to see the information one screenful at a time, first issue the command *more on*, i.e.,

```
>> more on
```

```
>> help elfun
```

Hit any key to progress to the next page of information.

Another way to get help is to use the *lookfor* command. The *lookfor* command differs from the help command. The help command searches for an exact function name match, while the *lookfor* command searches the quick summary information in each function for a match. For example

```
>> lookfor inverse
```

```
>> help sqrt
```

The doc function opens the on-line version of the help manual. This is very helpful for more complex commands.

```
>> doc plot
```

Use *lookfor* to find functions by keywords. The general form is

```
>> lookfor functionname
```

Managing the Workspace

The contents of the workspace persist between the executions of separate commands. Therefore, it is possible for the results of one problem to have an effect on the next one. To avoid this possibility, it is a good idea to issue a clear command at the start of each new independent calculation.

```
>> clear
```

The command *clear* or *clear all* removes all variables from the workspace. This frees up system memory. In order to display a list of the variables currently in the memory, type

```
>> who
```

while *whos* will give more details which include size, space allocation, and class of the variables.

To keep a record, issuing the command

```
>> diary mysession
```

This will cause all subsequent text that appears on the screen to be saved to the file *mysession* located in the directory in which MATLAB was invoked. You may use any legal filename except the names on and off. The record may be terminated by

```
>> diary off
```

The file *mysession* may be edited with your favourite editor (the MATLAB editor, emacs, or even MS Word) to remove any mistakes.

If you wish to quit MATLAB midway through a calculation so as to continue at a later stage type

```
>> save thissession
```

This will save the current values of all variables to a file called *thissession.mat*. This file cannot be edited. When you next startup MATLAB, type

```
>> load thissession
```

and the computation can be resumed where you left off. A list of variables used in the current session may be seen with the *whos* command, for example

```
>> whos
```

Output:

Name	Size	Elements	Bytes	Density	Complex
ans	1 by 1	1	8	Full	No
v	1 by 3	3	24	Full	No
v1	1 by 2	2	16	Full	No
v2	1 by 2	2	16	Full	No
v3	1 by 3	3	24	Full	No
v4	1 by 3	3	24	Full	No
x	1 by 1	1	8	Full	No
y	1 by 1	1	8	Full	No

Grand total is 16 elements using 128 bytes

Copying to and from Word and other applications

There are many situations where one wants to copy the output resulting from a MATLAB command (or commands) into a Windows application such as *MS Word* or into a Unix file editor such as *Emacs*.

Copying material is made possible on the Windows operating system by using the Windows clipboard. Also, pictures can be exported to files in a number of alternative formats such as

encapsulated postscript format or in jpeg format. MATLAB is so frequently used as an analysis tool that many manufacturers of measurement systems and software find it convenient to provide interfaces to MATLAB which make it possible, for instance, to import measured data directly into a *.mat MATLAB file.

Quitting MATLAB

To end your MATLAB session, type *quit* in the command window, or select *File - Exit MATLAB* in the desktop main menu.

Creating MATLAB Variables

MATLAB variables are created with an assignment statement. The syntax of variable assignment is

variable name = a value (or an expression)

For example,

```
>> x = expression
```

where *expression* is a combination of numerical values, mathematical operators, variables, and function calls. On other words, *expression* can involve:

manual entry

built-in functions

user-defined functions

Variables

```
>> 3-2^4
```

```
ans =
```

```
-13
```

```
>> ans*5
```

```
ans =
```

```
-65
```

The result of the first calculation is labeled *ans* by MATLAB and is used in the second calculation where its value is changed.

We can use our own names to store numbers:

```
>> x = 3-2^4
```

```
x =
```

```
-13
```

```
>> y = x*5
```

```
y =
```

```
-65
```

so that *x* has the value -13 and *y* = -65 . These can be used in subsequent calculations. These are examples of assignment statements: values are assigned to variables. Each variable must be assigned a value before it may be used on the right of an assignment statement.

Variable Names

Legal names consist of any combination of letters and digits, starting with a letter.

These are allowable: X, Y, NetCost, Left2Pay, x3, X3, z25, c5, and so on

These are not allowable: Net-Cost, 2pay, %x, @sign

It is important to use names that reflect the values they represent.

Special names: you should avoid using

eps = $2.2204e-16 = 2^{-54}$ (The largest number such that $1 + \text{eps}$ is indistinguishable from 1),

and

pi = $3.14159... = \pi$

If you wish to do arithmetic with complex numbers, both i and j have the value $\sqrt{-1}$ unless you change them

```
>> i,j, i=3
ans = 0 + 1.0000i
ans = 0 + 1.0000i
i = 3
```

Overwriting Variable

Once a variable has been created, it can be reassigned. In addition, if you do not wish to see the intermediate results, you can suppress the numerical output by putting a semicolon (;) at the end of the line. Then the sequence of commands looks like this:

```
>> t = 5;
>> t = t+1
t =
6
```

Entering Multiple Variables per Line

It is possible to enter multiple variables per line. Use commas (,) or semicolons (;) to enter more than one statement at once. Commas (,) allow multiple statements per line without suppressing output.

```
>> a=7; b=cos(a), c=cosh(a)
b =
0.6570
c =
548.3170
```

Error Messages

If we enter an expression incorrectly, MATLAB will return an error message. For example, in the following, we left out the multiplication sign, *, in the following expression

```
>> x = 10;
```

```
>> 5x
```

```
??? 5x
```

```
||
```

Error: Unexpected MATLAB expression.

Basic Arithmetic Operators

Symbol Operation Example

+	Addition	$2 + 3$
-	Subtraction	$2 - 3$
*	Multiplication	$2 * 3$
/	Division	$2/3$

MATLAB as a Calculator

The basic arithmetic operators are + - * / ^ and these are used in conjunction with brackets: (). The symbol ^ is used to get exponents (powers): $2^4=16$. You should type in commands shown following the prompt: >>.

```
>> 2 + 3/4*5
```

```
ans =
```

```
5.7500
```

```
>>
```

Is this calculation $2 + 3/(4*5)$ or $2 + (3/4)*5$?

MATLAB works according to the priorities:

1. quantities in brackets,
2. powers $2 + 3^2$) $2 + 9 = 11$,
3. * /, working left to right ($3*4/5=12/5$),
4. + -, working left to right ($3+4-5=7-5$),

Thus, the earlier calculation was for $2 + (3/4)*5$ by priority 3.

Hierarchy of arithmetic operations

Precedence Mathematical operations

- First** The contents of all parentheses are evaluated first, starting from the innermost parentheses and working outward.
- Second** All exponentials are evaluated, working from left to right
- Third** All multiplications and divisions are evaluated, working from left to right
- Fourth** All additions and subtractions are evaluated, starting from left to right

Now, consider another example:

$$\frac{1}{2 + 3^2} + \frac{4}{5} \times \frac{6}{7}$$

In MATLAB, it becomes

```
>> 1/(2+3^2)+4/5*6/7
```

```
ans =
```

```
0.7766
```

or, if parentheses are missing,

```
>> 1/2+3^2+4/5*6/7
```

```
ans =
```

```
10.1857
```

So here what we get: two different results. Therefore, we want to emphasize the importance of precedence rule in order to avoid ambiguity.

Controlling the Hierarchy of Operations or Precedence

Let's consider the previous arithmetic operation, but now we will include *parentheses*. For example, $1 + 2 \times 3$ will become $(1 + 2) \times 3$

```
>> (1+2)*3
```

```
ans =
```

```
9
```

and, from previous example

```
>> 1+2*3
```

```
ans =
```

```
7
```

MATLAB arithmetic operators obey the same *precedence* rules as those in most computer programs. For operators of *equal* precedence, evaluation is from *left to right*.

Mathematical Functions

Some commonly used functions, where variables x and y can be numbers, vectors, or matrices.

Elementary Functions

cos(x)	Cosine	abs(x)	Absolute value
sin(x)	Sine	sign(x)	Signum function
tan(x)	Tangent	max(x)	Maximum value
acos(x)	Arc cosine	min(x)	Minimum value
asin(x)	Arc sine	ceil(x)	Round towards $+\infty$
atan(x)	Arc tangent	floor(x)	Round towards $-\infty$
exp(x)	Exponential	round(x)	Round to nearest integer
sqrt(x)	Square root	rem(x)	Remainder after division
log(x)	Natural logarithm	angle(x)	Phase angle

log10(x) Common logarithm conj(x) Complex conjugate

Predefined constant values

pi The π number, $\pi = 3.14159 \dots$

i,j The imaginary unit $i, \sqrt{-1}$

Inf The infinity, ∞

NaN Not a number

Examples

We illustrate here some typical examples which related to the elementary functions previously defined.

As a first example, the value of the expression $y = e^{-a} \sin(x) + 10\sqrt{y}$, for $a = 5, x = 2$, and $y = 8$ is computed by

```
>> a = 5; x = 2; y = 8;
>> y = exp(-a)*sin(x)+10*sqrt(y)
y =
28.2904
```

The subsequent examples are

```
>> log(142)
ans =
4.9558
>> log10(142)
ans =
2.1523
```

Note the difference between the natural logarithm $\log(x)$ and the decimal logarithm (base10) $\log_{10}(x)$.

To calculate $\sin(\pi/4)$ and e^{10} , we enter the following commands in MATLAB,

```
>> sin(pi/4)
ans =
0.7071
>> exp(10)
ans =
2.2026e+004
```

Suppressing Output

One often does not want to see the result of intermediate calculations. We can terminate the assignment statement or expression with semi-colon

```
>> x=-13; y = 5*x, z = x^2+y
y =
-65
z =
104
>>
```

the value of x is hidden. Note also we can place several statements on one line, separated by commas or semi-colons.

Arrays

An array is an ordered collection of numbers. Each number can be accessed directly using an index (much like the subscript notation used in mathematics to refer to an element of a vector or a matrix). An array is entered into MATLAB as a list of numbers e.g.

```
x = [1,3,5,7,9]
```

or instead of using commas we can use blanks

```
x = [1 3 5 7 9]
```

This creates a variable called *x* which has elements *x*(1), *x*(2), ..., *x*(5). Each such element is a real number and can be used directly. To access the third element we type:

```
z = x(3)
```

This code would result in the variable *z* assuming a value of 5.

Arrays become useful when one wants to perform the same action on every element of an array. For example, supposed we have an array describing the flows of a set of four components in a mixture. We can find out the total flow by simply adding up the individual flows. One way would obviously be to write a statement of the form

```
Totalflow = x(1) + x(2) + x(3) + x(4);
```

Though, luckily, MATLAB has a built in function *sum* which can be used in this instance.

We write

```
Totalflow = sum(x);
```

Dot Division of Arrays (./)

There is no mathematical definition for the division of one vector by another. However, in MATLAB, the operator ./ is defined to give element by element division – it is therefore only defined for vectors of the same size and type.

```
>> a = 1:5, b = 6:10, a./b
```

```
a =
```

```
    1    2    3    4    5
```

```
b =
```

```
    6    7    8    9   10
```

```
ans =
```

```
0.1667    0.2857    0.3750    0.4444    0.5000
```

```
>> a./a
```

```
ans =
```

```
    1    1    1    1    1
```

```
>> c = -2:2, a./c
```

```
c =
```

```
   -2   -1    0    1    2
```

```
Warning: Divide by zero
```

```
ans =
```

```
-0.5000   -2.0000    Inf    4.0000    2.5000
```

The previous calculation required division by 0 - notice the *Inf*, denoting infinity, in the answer.

```
>> a.*b -24, ans./c
```

```
ans =
```

```
   -18   -10    0    12    26
```

```
Warning: Divide by zero
```



```
ans =
```

```
9      10     NaN    12     13
```

Here we are warned about 0/0 - giving a NaN (Not a Number).

Example

Estimate the limit

$$\lim_{x \rightarrow \infty} \frac{\sin \pi x}{x}$$

The idea is to observe the behaviour of the ratio $\frac{\sin \pi x}{x}$ for a sequence of values of x that approach zero. Suppose that we choose the sequence defined by the column vector

```
>> x = [0.1; 0.01; 0.001; 0.0001]
```

then

```
>> sin(pi*x)./x
```

```
ans =
```

```
3.0902  
3.1411  
3.1416  
3.1416
```

which suggests that the values approach π . To get a better impression, we subtract the value of π from each entry in the output and, to display more decimal places, we change the format

```
>> format long
```

```
>> ans -pi
```

```
ans =
```

```
-0.05142270984032  
-0.00051674577696
```

```
-0.00000516771023
```

```
-0.00000005167713
```

Can you explain the pattern revealed in these numbers?

We also need to use `./` to compute a scalar divided by a vector:

```
>> 1/x
```

```
??? Error using ==> /
```

```
Matrix dimensions must agree.
```

```
>> 1./x
```

```
ans =
```

```
10    100   1000  10000
```

so `1./x` works, but `1/x` does not.

Dot Power of Arrays (`.^`)

To square each of the elements of a vector we could, for example, do `u.*u`. However, a neater way is to use the `.^` operator:

```
>> u.^2
```

```
ans =
```

```
100   121   144
```

```
>> u.*u
```

```
ans =
```

```
100   121   144
```

```
>> u.^4
```

```
ans =  
    10000    14641    20736
```

```
>> v.^2
```

```
ans =  
    400  
    441  
    484
```

```
>> u.*w.^(-2)
```

```
ans =  
    2.5000   -11.0000    1.3333
```

Recall that powers (\wedge in this case) are done first, before any other arithmetic operation.

Other Array arithmetic

Addition and subtraction

Matrix addition and subtraction are in fact carried out on an element-by-element basis and so there is no need for separate array addition and subtraction operations.

Multiplication

How can we square all the elements in matrix A? If we write a program that calculates A^2 (this is equivalent to carrying out the matrix operation $A * A$)

then we might write:

```
A = [1 2; 3 4];
```

```
B = A^2; % or B = A * A
```

however, this gives the output

B =

```
7    10
15   22
```

So we need to specify that we want to carry out the operation on an element-by-element basis, hence:

```
A = [1 2; 3 4];
```

```
C = A.^2; % or C = A .* A
```

C =

```
1    4
9   16
```

In general, the operation $A(i,j) .* B(i,j)$ will result in a matrix with elements $a_{ij}b_{ij}$.

Division

In a similar way to multiplication if we want to carry out division on an element-by-element basis then we used the array operations. Thus, the operation $A(i,j) ./ B(i,j)$ will result in a matrix with elements a_{ij}/b_{ij} . Hence,

```
A = [1 2; 3 4];
```

```
C = [1 4; 9 16];
```

```
D = C./A
```

D =

```
1    2
3    4
```

Note: For both array multiplication and division, unless either variable is a scalar, then both matrices being operated on must be the same size. If you don't understand the rules of matrix

and array arithmetic then it may help to refer to your maths notes or a standard maths text book.

Matrices

Thus far we have only introduced one dimensional arrays, however many engineering and mathematical calculations require the use of matrices, which are two dimensional arrays. These are entered in a similar way

```
>> A = [1 2;3 4]
```

```
A =
```

```
    1    2
    3    4
```

We can determine the size of any matrix, using the *size* function

```
>> size(A)
```

where the output is

```
ans =
```

```
    2    2
```

This tells us that *A* is a matrix with two rows and two columns. Actually, all variable types in MATLAB (scalar variables, and row and column arrays) are in fact matrices of different sizes. So if we determine the size of our row vector, *x*

```
>> size(x)
```

we are told

```
ans =
```

```
    1    5
```

So x is a matrix 1 row by 5 columns.

Matrix arithmetic operations in MATLAB

When earlier we considered the multiplication and addition of simple (scalar) variables, we used normal mathematical operators. However, for calculations with matrix variables we will see that there are two types of arithmetic operations:

Matrix arithmetic: which is based on the rules of standard linear algebra. The standard operators are used (+, -, *, /, \, ^).

Array arithmetic: which is carried out element-by-element. To distinguish from matrix arithmetic the standard operators are preceded by a full-stop (.*, ./, .^). The results of the operations are generally quite different, and therefore, it is essential to determine which option you require before writing your program.

Addition and subtraction

Under the standard rules of matrix arithmetic, normal addition and subtraction are carried out on an element-by-element basis. For example

```
>> A = [1 2;3 4] ;  
>> B = [2 4;3 5];  
>> C = A + B  
>> D = A - B
```

gives the output

```
C =  
     3     6  
     6     9  
D =  
    -1    -4
```

0 -1

and of course the operation $C - A$ gives B.

Multiplication

Matrix multiplication is illustrated by the following program:

```
A = [1 2; 3 4];
```

```
B = A * A
```

```
C = B *2
```

The output of the program is

```
B =
```

```
7    10
```

```
15   22
```

```
C =
```

```
14   20
```

```
30   44
```

Notice that the multiplications are carried out under the standard rules of matrix arithmetic and thus the result of the first operation is not simply the square of all the elements but for this example is calculated as below:

```
B = (1*1+2*3) (1*2+2*4)
```

```
(1*3+4*3) (3*2+4*4)
```

Division

In MATLAB both left and right division are possible. So if we wish to divide two scalars then two results are possible:

```
>>2\3
```

```
ans =
```

1.5

>>2/3

ans =

0.6667

Of course we can also use matrix division to reverse a multiplication e.g.

A = [1 2; 3 4];

C = [14 20; 30 40];

D = C/2

E = D/A

this program gives

D =

7 10

15 22

E =

1 2

3 4

In general if A is a square matrix then we can say that

$A \setminus B = \text{inv}(A) * B$

$A / B = B * \text{inv}(A)$

Solving linear equations

Perhaps the most useful application of matrix division is in the solution of linear equations.

In fact in MATLAB a system of n linear equations in n unknowns can be solved easily using left division (\). (The maths for over or under specified systems of equations is much more complicated and thus we won't consider it.).

Consider a system of two simultaneous linear equations containing two unknowns:

$$2x_1 + 5x_2 = 3$$

$$4x_1 + 5x_2 = 11$$

This can be written in matrix form

$$A \cdot X = B$$

where

$$A = \begin{bmatrix} 2 & 5 \\ 4 & 5 \end{bmatrix} \quad X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad B = \begin{bmatrix} 3 \\ 11 \end{bmatrix}$$

Summary of 16

- MATLAB is an acronym for MATrix LABoratory.
- It is a high-performance language for technical computing as well as a modern programming language environment with sophisticated *data structures*, built-in editing and *debugging tools*, and supports *object-oriented programming*.
- These factors make it an excellent tool for teaching and research. It has many advantages compared to conventional computer languages (e.g., C, FORTRAN) for solving technical problems. The basic features are interactive.
- It can also be used as a calculator. It has found relevant usage in all areas of science and technology.

Self-Assessment Questions (SAQs) for study session 16

Now that you have completed this study session, you can assess how well you have achieved its Learning outcomes by answering the following questions. Write your answers in your study

Diary and discuss them with your Tutor at the next study Support Meeting. You can check your Define School answers with the Notes on the Self-Assessment questions at the end of this Module.

Post – Test

1. In each case find the value of the expression in MATLAB and explain precisely the order in which the calculation was performed.

i) -2^3+9

ii) $2/3*3$

iii) $3*2/3$

iv) $3*4-5^2*2-3$

v) $(2/3^2*5)*(3-4^3)^2$ vi) $3*(3*4-2*5^2-3)$

2. Solve for x_1 and x_2 in the following simultaneous equation:

$$2x_1 + 5x_2 = 3$$

$$4x_1 + 5x_2 = 11$$

References

1. Griffiths D. F. (2005): An Introduction to MATLAB. *A Publication of the Department of Mathematics, University of Dundee, Scotland, 3rd Ed.*
2. Houcque D. (2005): Introduction to MATLAB for Engineering Students. *A Publication of Northwestern University, Illinois, USA.*