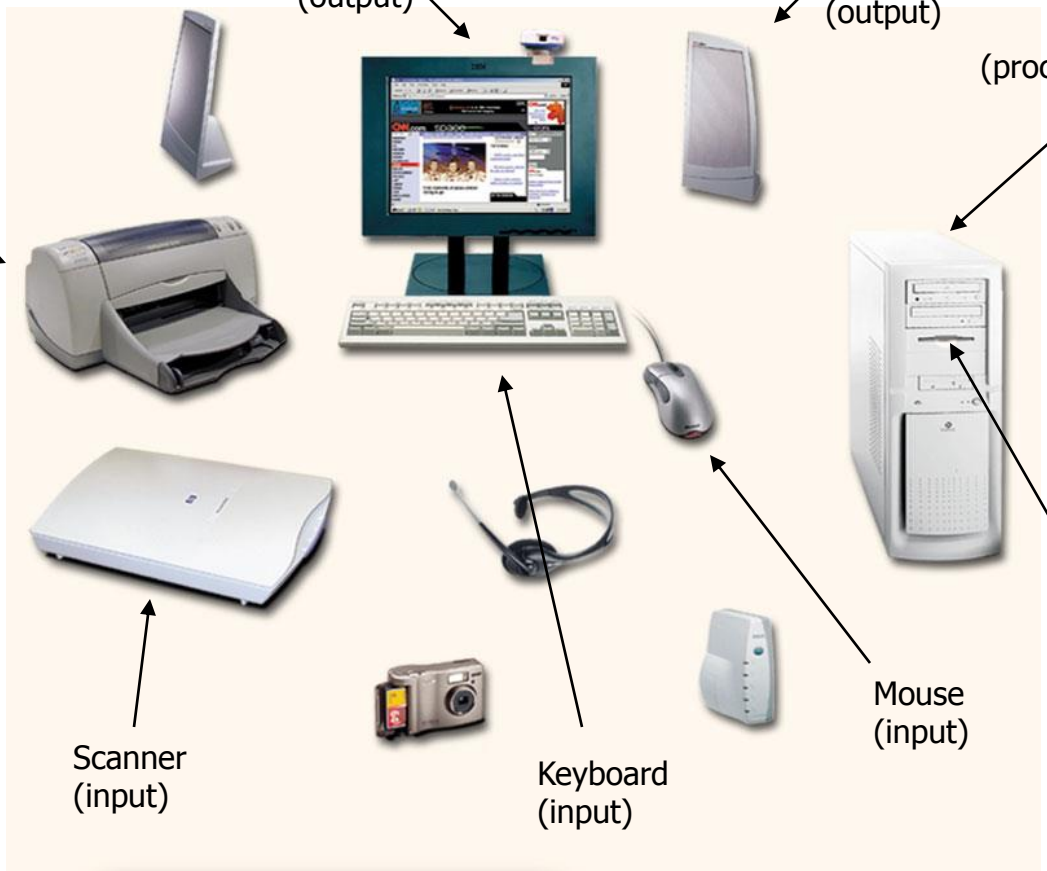
Introduction to Computers

What Is A Computer?

A computer is an electronic device, operating under the control of instructions (software) stored in its own memory unit, that can accept data (input), manipulate data (process), and produce information (output) from the processing. Generally, the term is used to describe a collection of devices that function together as a system.

Devices that comprise a computer system



What Does A Computer Do?

Computers can perform four general operations, which comprise the information processing cycle.

- Input
- Process
- Output
- Storage

Input, Process, Output, & Store data

Input ⇒ **Process** ⇒ **Output**

⇕

Store Data

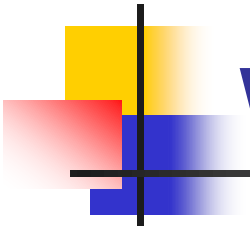
Data and Information

- All computer processing requires **data**, which is a collection of raw facts, figures and symbols, such as numbers, words, images, video and sound, given to the computer during the input phase.
- Computers manipulate data to create information. **Information** is data that is organized, meaningful, and useful.
- During the output Phase, the information that has been created is put into some form, such as a printed report.
- The information can also be put in computer storage for future use.

Why Is A Computer So Powerful?

- The ability to perform the information processing cycle with amazing speed.
- Reliability (low failure rate).
- Accuracy.
- Ability to store huge amounts of data and information.
- Ability to communicate with other computers.

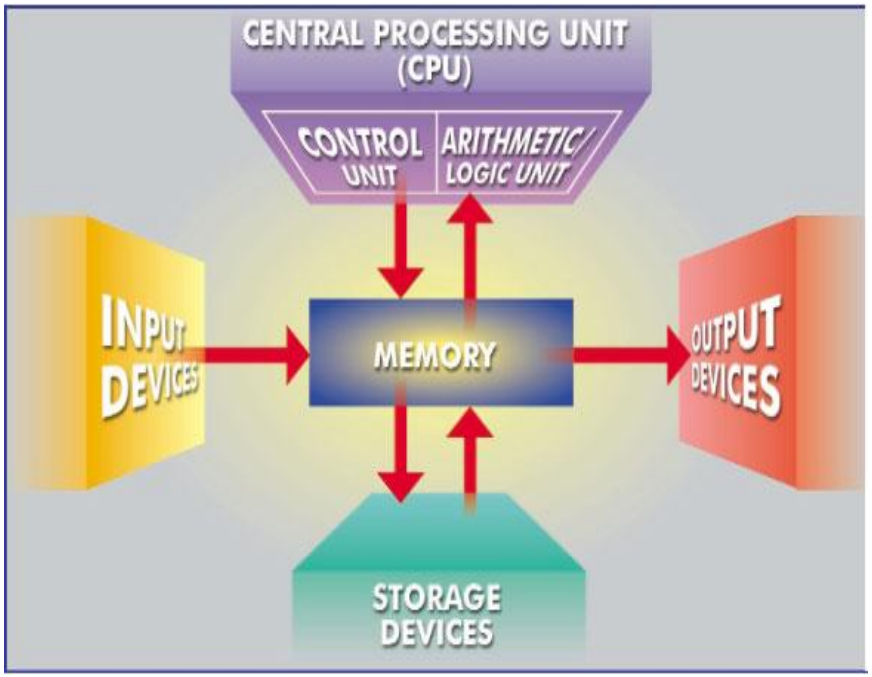
How Does a Computer Know what to do?



- It must be given a detailed list of instructions, called a **compute program** or **software**, that tells it exactly what to do.
- Before processing a specific job, the computer program corresponding to that job must be stored in memory.
- Once the program is stored in memory the compute can start the operation by executing the program instructions one after the other.

What Are The Primary Components Of A Computer ?

- Input devices.
- Central Processing Unit (containing the control unit and the arithmetic/logic unit).
- Memory.
- Output devices.
- Storage devices.



Uses of Computer

PC at Home

Common uses for the computer within the home

- Computer games
- Working from Home
- Banking from Home
- Connecting to the Web

Uses of Computer

Office Applications

Stock Control

Stock control is ideal for automation and in many companies it is now completely computerized. The stock control system keeps track of the number of items in stock and can automatically order replacement items when required.

Accounts / Payroll

In most large organizations the accounts are maintained by a computerized system. Due to the repetitive nature of accounts a computer system is ideally suited to this task and accuracy is guaranteed.

Uses of Computer

Automated Production Systems

Many car factories are almost completely automated and the cars are assembled by computer-controlled robots. This automation is becoming increasingly common throughout industry.

Design Systems

Many products are designed using CAD (Computer Aided Design) programs to produce exact specifications and detailed drawings on the computer before producing models of new products.

Uses of Computer

Computers in Daily Life

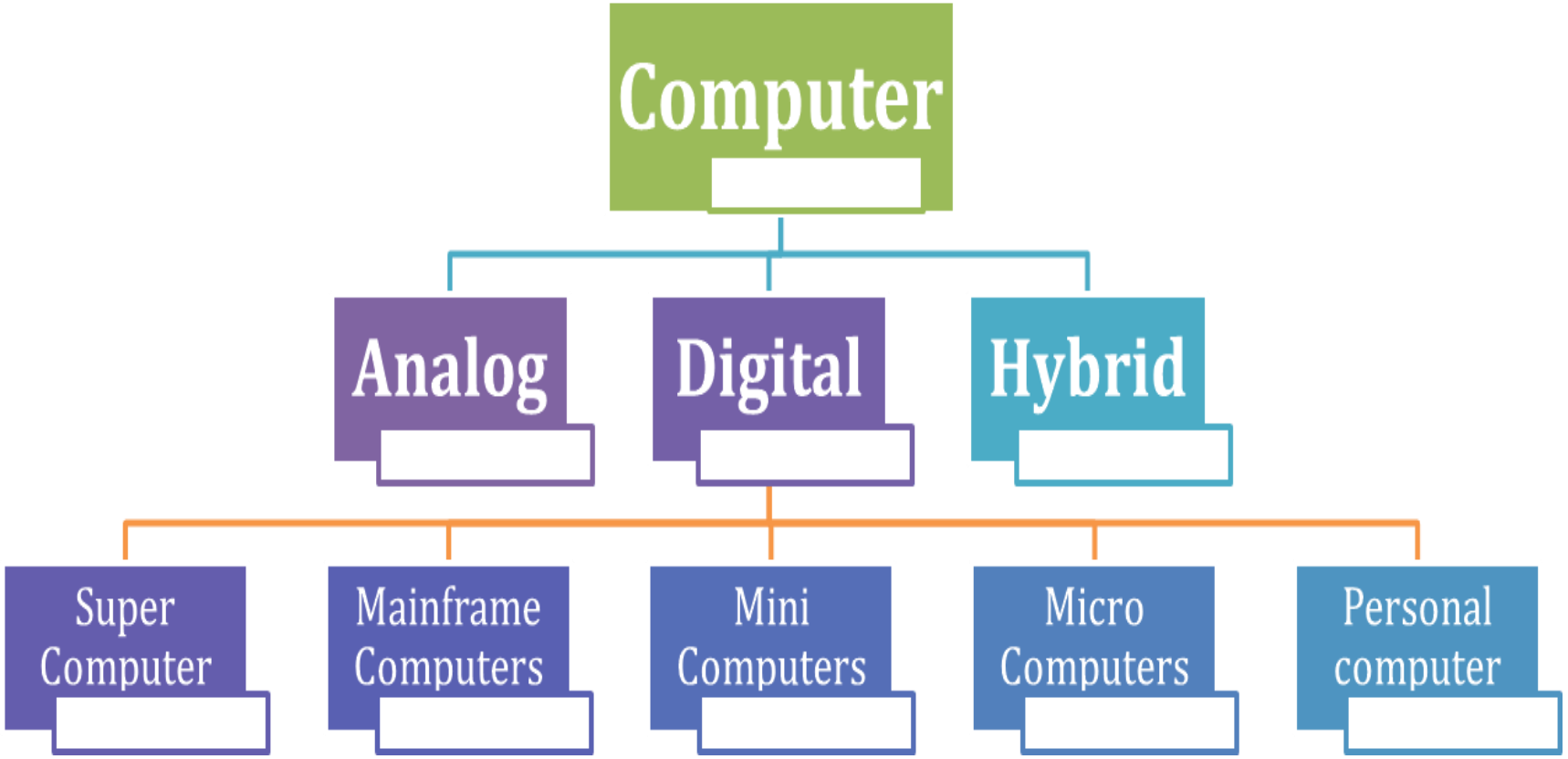
- Accounts
- Games
- Educational
- On-line banking
- Smart ID cards
- Supermarkets
- Working from home (Tele-working)
- Internet

Computer Input Devices

- Keyboard
- Mouse/Trackball
- Joystick
- Light pen
- Pointing Stick
- Touchpad
- Touch screen
- Bar code reader
- Scanner
- Microphone
- Graphics Tablet
- Digital Cameras



Types of Computers



Supercomputer

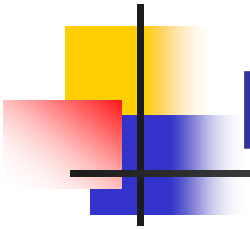
- Fastest and expensive
- Used by applications for molecular chemistry, nuclear research, weather reports, and advanced physics
- Consists of several computers that work in parallel as a single system

Super Computer

- **Advantage**
  - Solve bigger problems
  - Run more problems in shorter time
  - Allows for virtual testing
  - Can be used for R&D
- **Disadvantage**
  - Can be expensive
  - Takes up a lot of space
  - May only be good for specific applications
  - Does not replace physical testing
  - Requires trained staff
- Generate a large amount of heat during operation

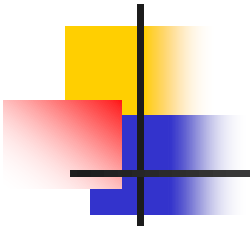
Mainframe

- Known as **enterprise servers**
- Occupies entire rooms or floors
- Used for centralized computing
- Serve distributed users and small servers in a computing network



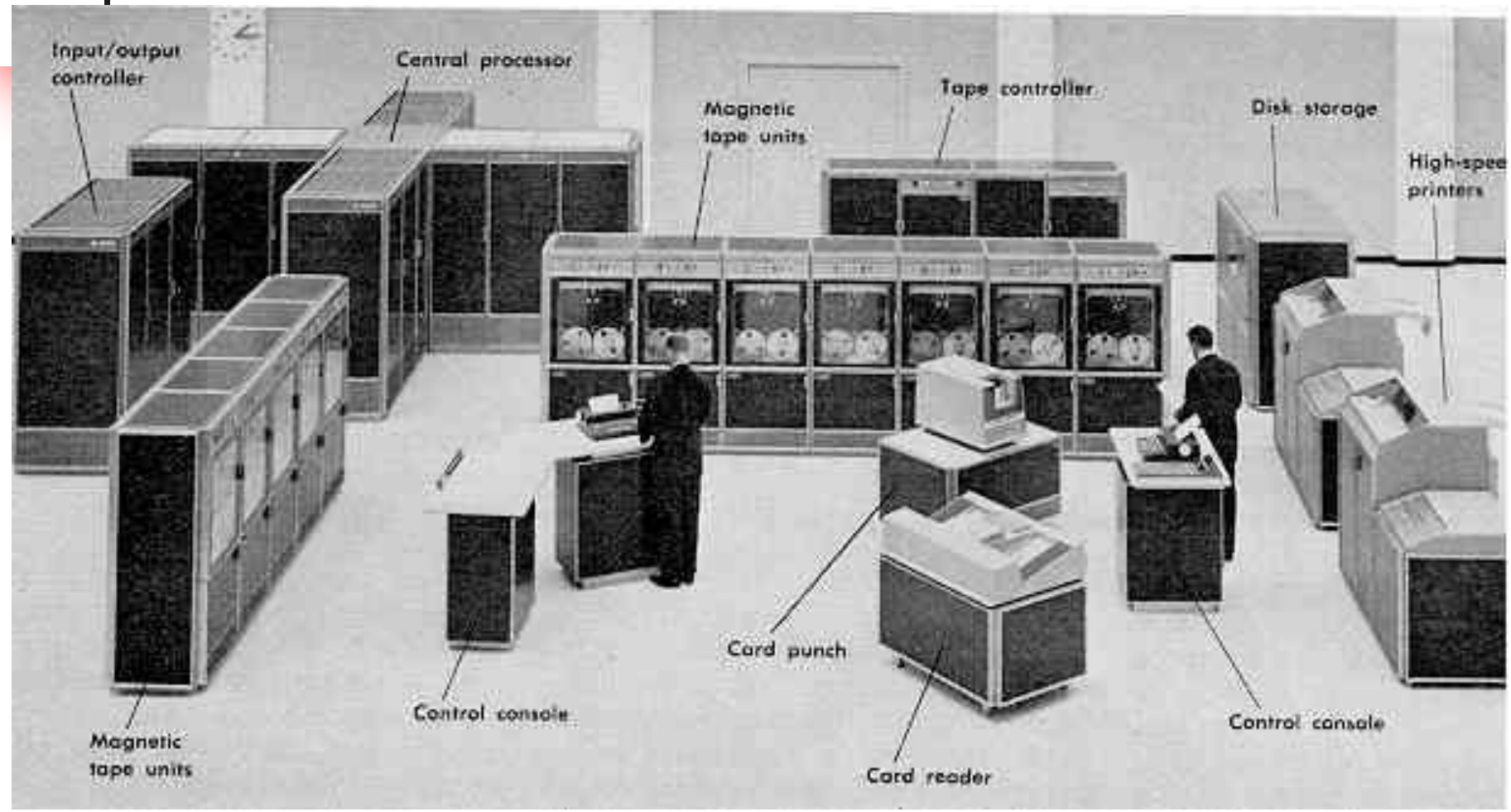
Mainframe

- Large, fast and expensive computer
- Cost millions of dollar
  - e.g. IBM3091, ICL39, etc
- Characteristics:
  - Bigger in size than minicomputers
  - Very expensive
  - Support a few hundred users simultaneously (Multi-Users)
  - Difficult to use
  - More computing power than minicomputers
  - Have to be kept in a special air-conditioned room
  - Used in big business organizations and government departments



Mainframe

- **Advantage**
  - Supports many users and instructions
  - Large memory
- **Disadvantage**
  - Huge size
  - Expensive

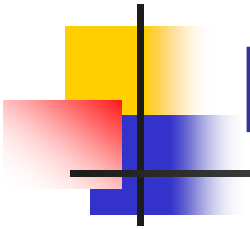

Input/output controller
Central processor
Magnetic tape units
Tape controller
Disk storage
High-spee printers
Card punch
Control console
Magnetic tape units
Card reader
Control console

Areas where mainframes are used

- Airline reservation
- Big banks with hundreds of branches located all over the world
- Big universities with thousands of enrollment
- Natural gas and oil exploration companies
- Space Vehicle control
- Weather forecasting
- Animated Cartoon
- Some mainframes are designed to be extremely fast and called super computers. It is used for space launching, monitoring and controlling.

Minicomputer

- Medium sized computer
- Also called the minis
  - e.g. IBM36, HP9000, etc
- Computing power lies between microcomputer and mainframe computer



MiniComputer

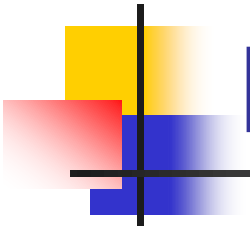
- Characteristics
  - Bigger size than PCs
  - Expensive than PCs
  - Multi-User
  - Difficult to use
  - More computing power than PCs
  - Used by medium sized business organizations, colleges, libraries and banks.

Minicomputer

- **Advantage**
  - Cater to multiple users
  - Lower costs than mainframes
- **Disadvantage**
  - Large
  - Bulky

Uses of Minicomputer

- Control of Automated Teller Machine (ATMs)
- Payroll
- Hospital patients registration
- Inventory Control for supermarket
- Insurance claims processing
- Small bank accounting and customer details tracking



Microcomputer

- Portable PCs
  - Can be moved easily from place to place
  - Weight may varies
  - Small PCs are popular known as laptop
  - Widely used by students, scientist, reporters, etc

Microcomputer

- **Advantages**
  - Small size
  - Low cost
  - Portability
  - Low Computing Power
  - Commonly used for personal applications
- **Disadvantages**
  - Low processing speed

Uses of Microcomputer

- Word Processing
- Home entertainment
- Home banking
- Printing
- Surfing the internet
- etc

Microcomputer Model

Laptop

Notebook

Subnotebook

Palmtop

Computer System

A computer system consists of three primary units:

Input units – accept data

Processor unit – processes data by performing comparisons and calculations

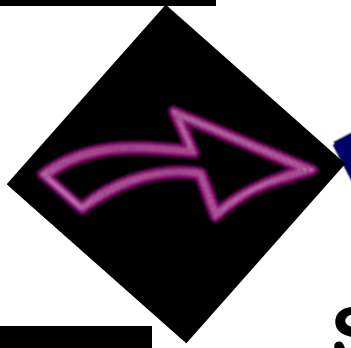
Output units – present the results

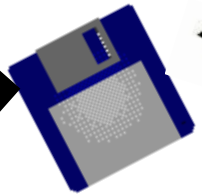
COMPUTER SYSTEM

MONITOR

PROCESSOR

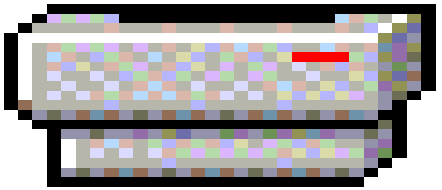
KEYBOARD





Storage devices

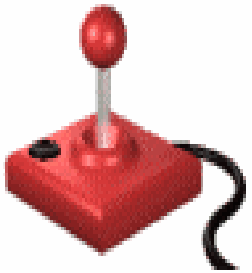
PRINTER



Input Devices

Data are facts, numbers and characters that are entered into the computer via keyboard.

Other types of input devices are mouse, joystick, light pens, scanners, camera, etc.



Processor Unit

Two main parts:

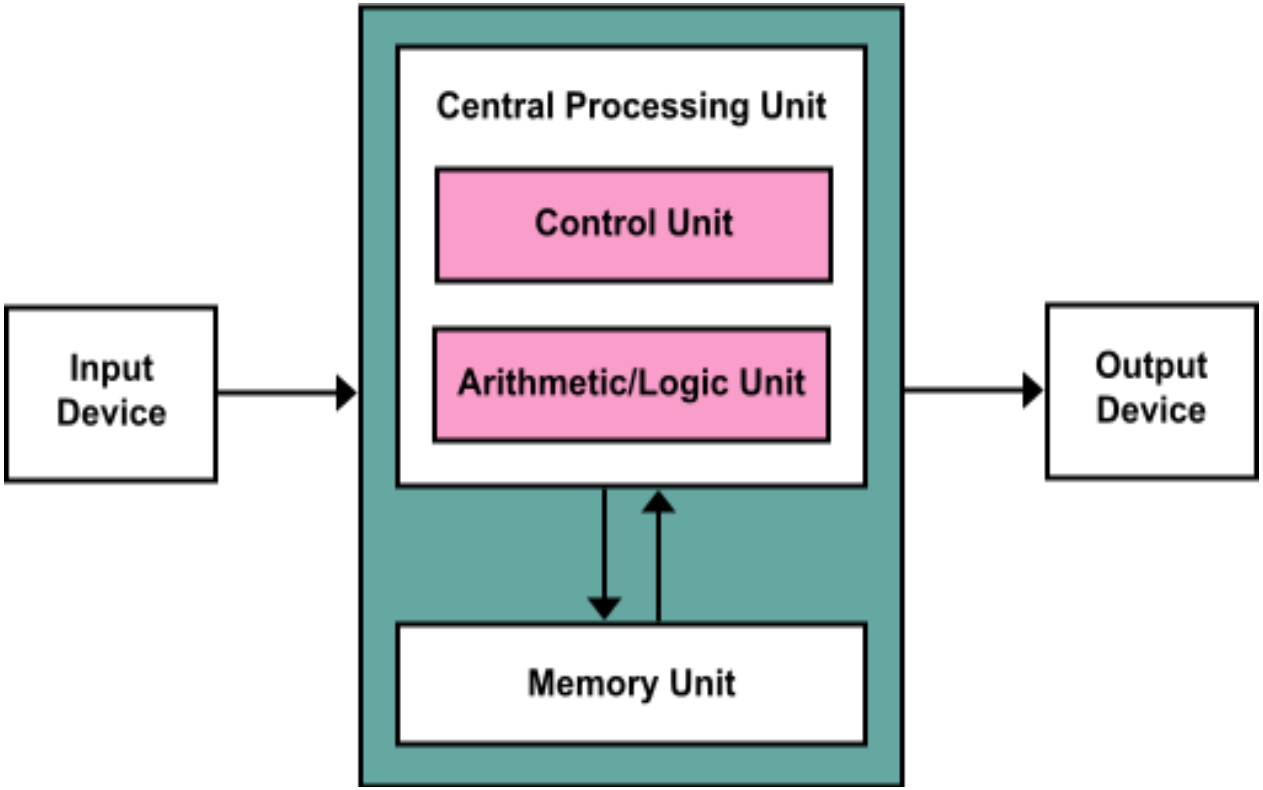
CPU – where the actual processing takes place; and

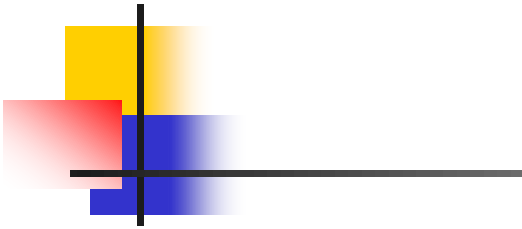
Main memory – where data are stored.

The contents of main memory can be transferred to auxiliary storage devices such as hard disks, floppy diskettes, zip disks, compact disks, or USB flash disk.

Central Processing Unit

- The microprocessor, the brains of the computer. Referred to a CPU or processor
  - Housed on a tiny silicon chip
  - Chip contains millions of switches and pathways that help your computer make important decisions.


Central Processing Unit
Control Unit
Arithmetic/Logic Unit
Input Device
Output Device
Memory Unit



CPU knows which switches to turn on and which to turn off because it receives its instructions from computer programs (software).

- CPU has two primary sections:

  Arithmetic/logic unit

  Control unit

Arithmetic/logic unit (ALU):

- Performs arithmetic computations and logical operations; by combining these two operations the ALU can execute complex tasks.

  - Arithmetic operations include addition, subtractions, multiplication, and division.

  - Logical operations involve comparisons.

Control Unit:

- It is the “boss” and coordinates all of the CPU’s activities.

•Uses programming instructions, it controls the flow of information through the processor by controlling what happens inside the processor.

•We communicate with the computer through programming languages.

Examples: COBOL, C++, HTML, Java Script or VisualBasic.net

Memory

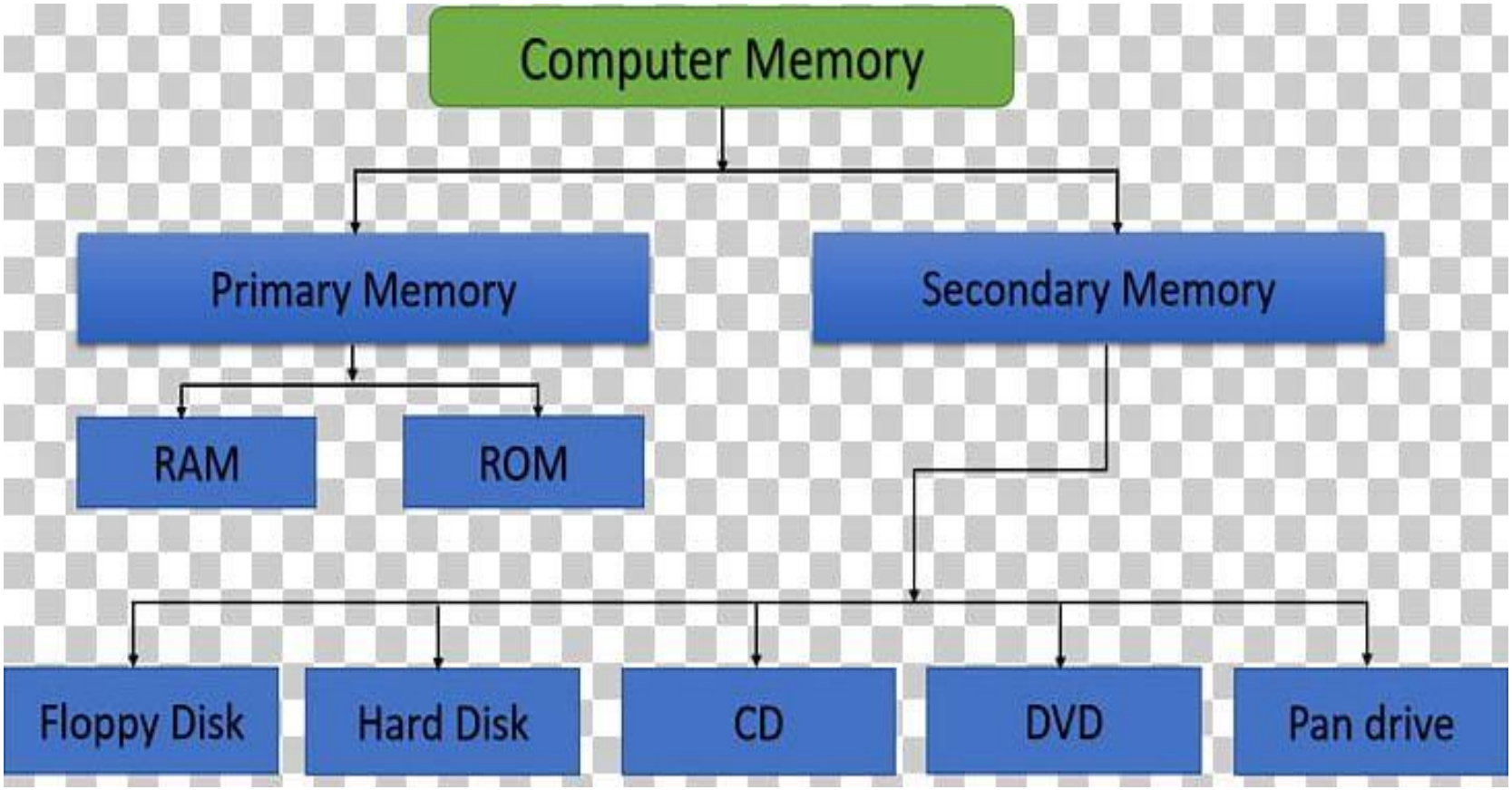
Found on the motherboard

- Short term

  Random Access Memory (RAM)

- Long term

  Read Only Memory (ROM)

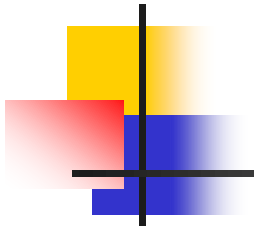

Computer Memory
Primary Memory
Secondary Memory
RAM
ROM
Floppy Disk
Hard Disk
CD
DVD
Pan drive

Random Access Memory (RAM)

Memory on the motherboard that is short term; where data, information, and program instructions are stored temporarily on a RAM chip or a set of RAM chips. Known as the main memory.

This memory is considered volatile.

The computer can read from and write to RAM.



When the computer is turned off or if there is loss of power, what ever is stored in RAM disappears.

“Temporary Memory” – Short Term

Read-Only Memory (ROM)

Memory on the motherboard that is long term; where the specific instructions that are needed for the computer to operate are stored.

This memory is nonvolatile and your computer can only read from a ROM chip.

The instructions remain on the chip regardless if the power is turned on or off.

Most common is the BIOS ROM; where the computer uses instructions contained on this chip to boot or start the system when you turn on your computer.

"Permanent Memory" – Long Term

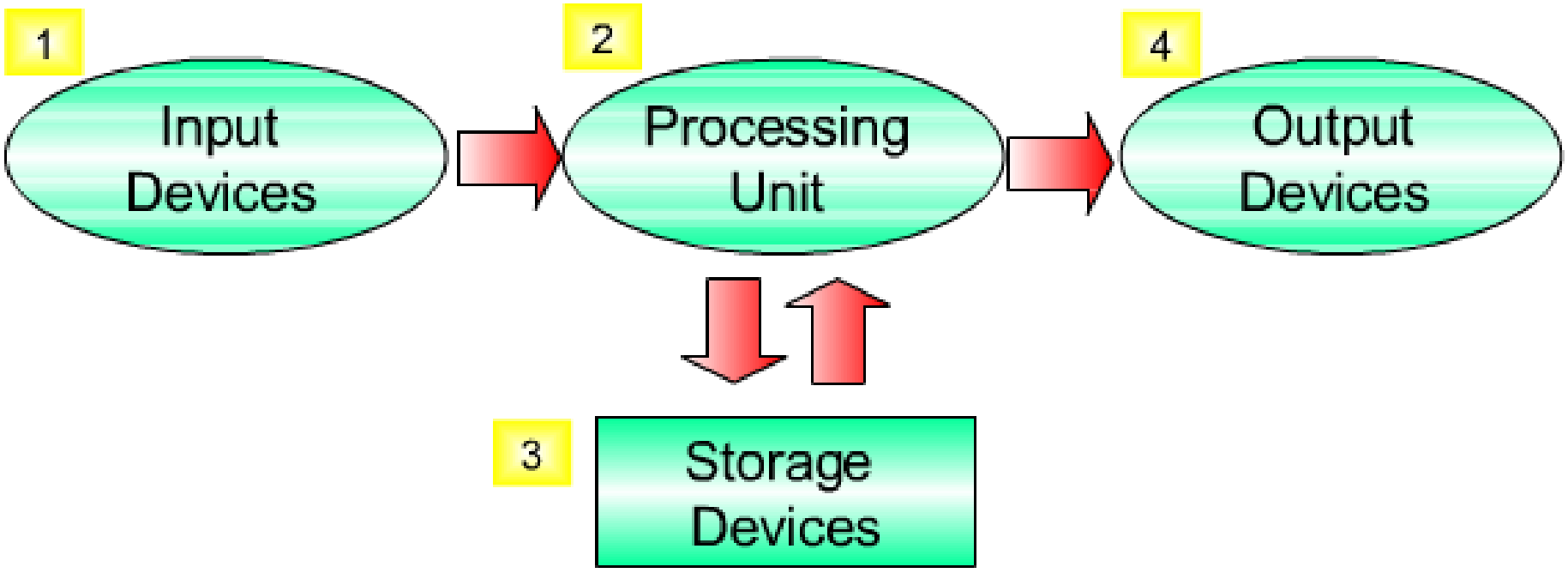
What is Hardware &software?

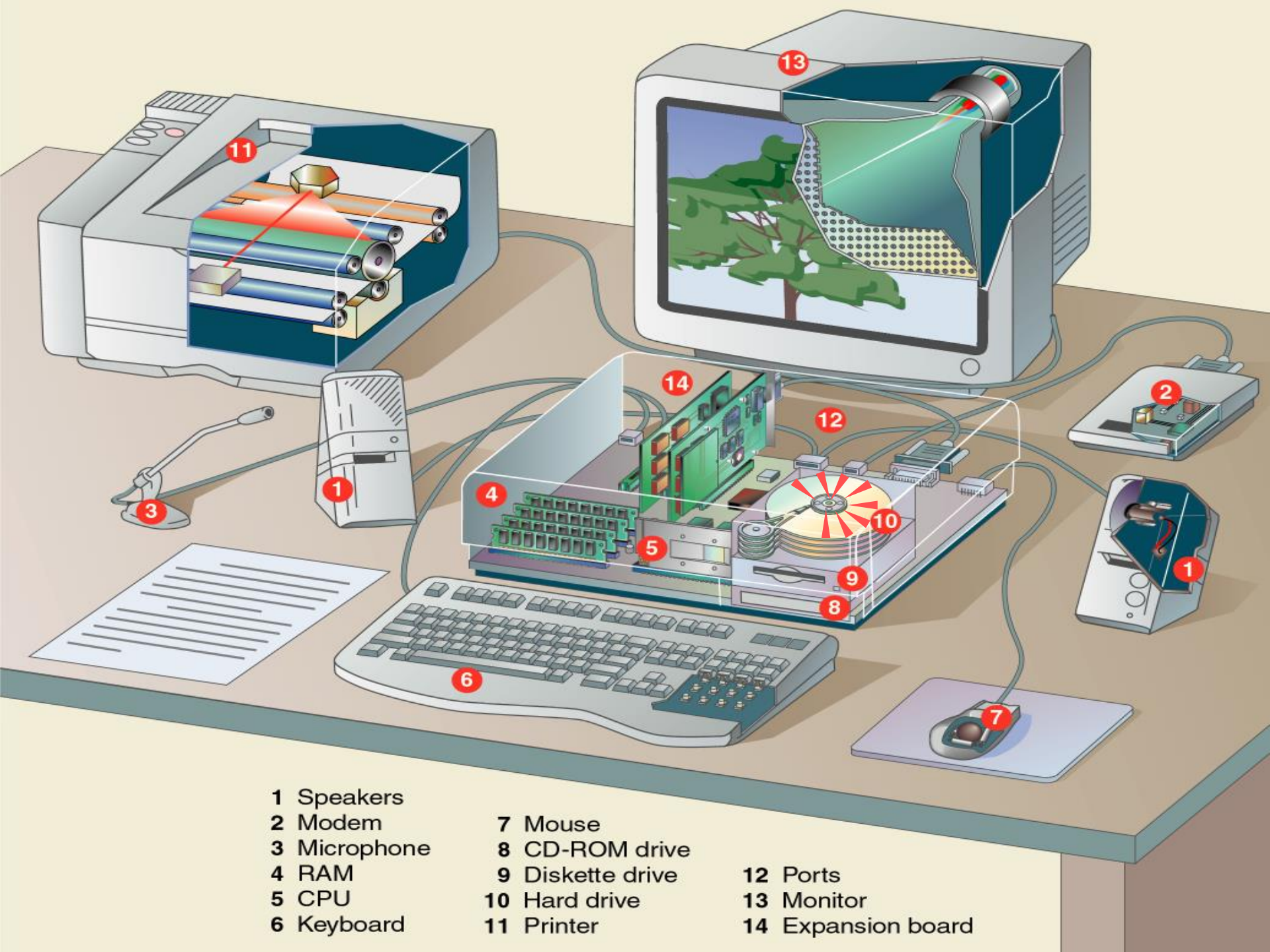
- A **computer** system consists **of** two major elements:
  1. **Hardware**
  2. **Software**.

  - **Computer hardware** is the collection **of** all the **parts** you can physically touch.

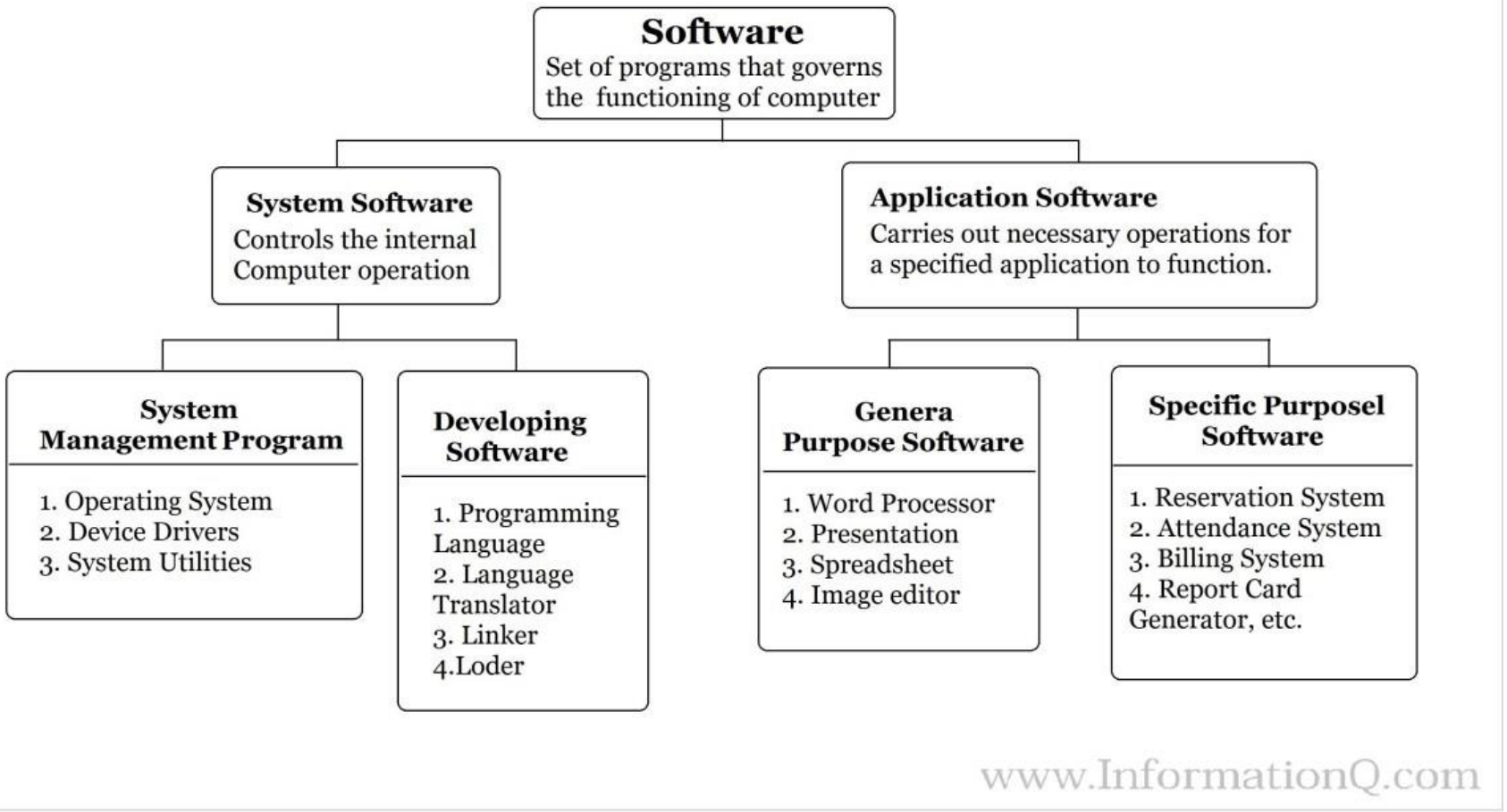
  - **Computer software**, on the other hand, is not something you can touch. **Software** is a set **of** instructions for a **computer** to perform specific operations.

Classification of Hardware




1 Speakers
2 Modem
3 Microphone
4 RAM
5 CPU
6 Keyboard
7 Mouse
8 CD-ROM drive
9 Diskette drive
10 Hard drive
11 Printer
12 Ports
13 Monitor
14 Expansion board

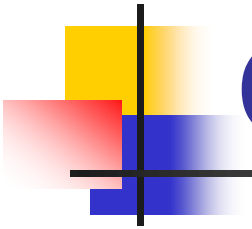
Types of Software



Hardware & Software

COMPARISON CHART

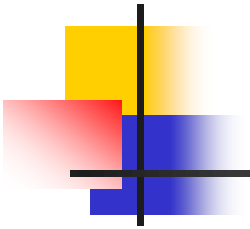
	HARDWARE	SOFTWARE
DEFINITION:	Devices that are required to store and execute the software.	Collection of instructions that enables a user to interact with the Computer.
TYPES:	Input,storage,processing, control and output devices.	System software, Programming software and Application software.
EXAMPLE:	CD-ROM, monitor, printer scanners	Quickbooks, Adobe Acrobat, Microsoft Word.
DEPENDENT:	Hardware starts functioning once software is loaded.	To deliver its set of instructions, Software is installed on hardware.
NATURE:	Hardware is physical in nature.	Software is logical in nature.



Operating System

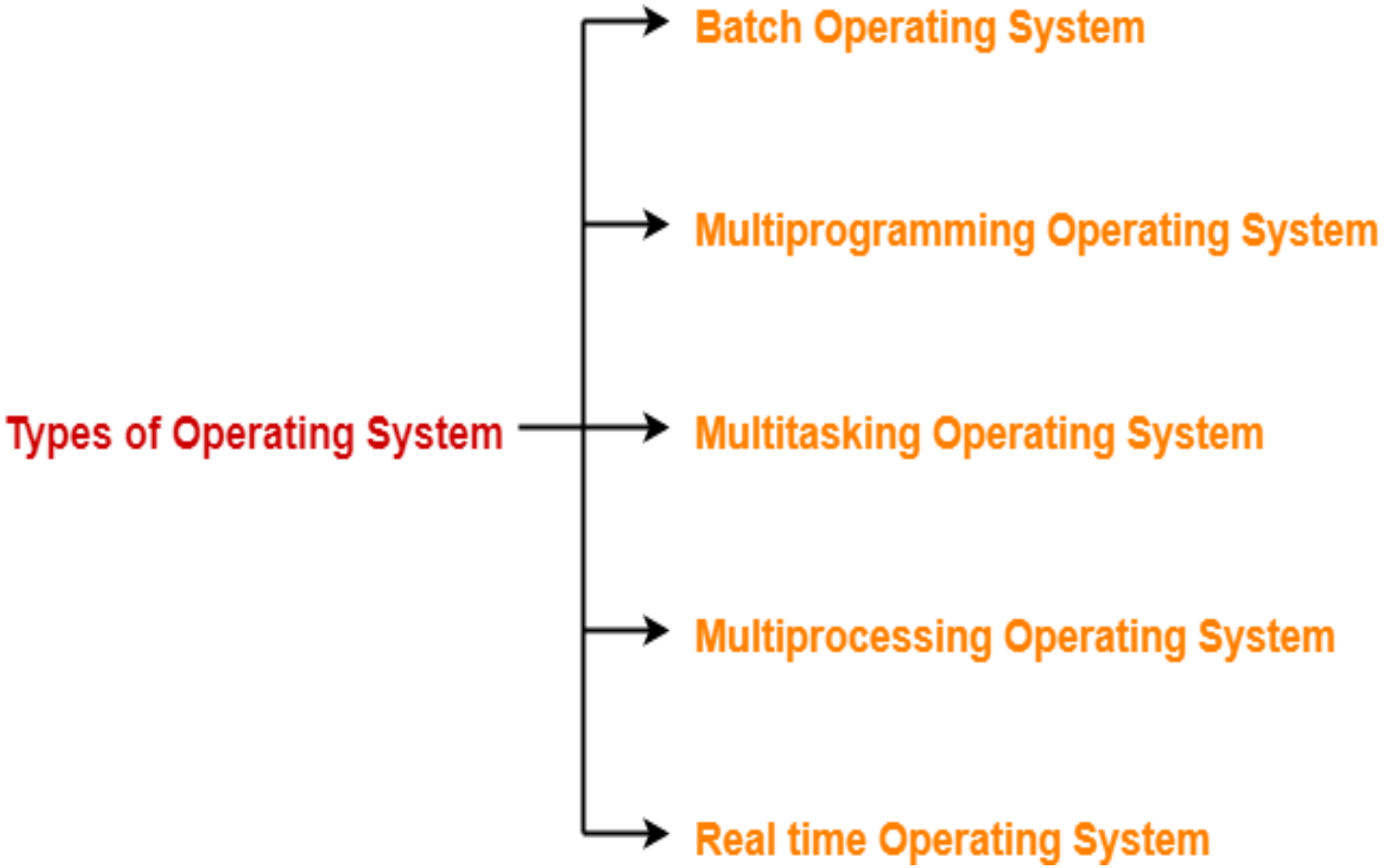
- An *operating system* is a powerful, and usually large, program that controls and manages the hardware and other software on a computer.

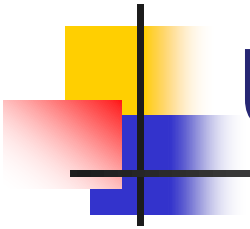
- All computers and computer-like devices require operating systems, including your laptop, tablet, desktop, smartphone, smartwatch, and router.



Some examples

- Microsoft Windows (versions of Microsoft Windows Windows 10, Windows 8, Windows 7, Windows Vista, and Windows XP),
- Apple's macOS (formerly OS X),
- Chrome OS,
- BlackBerry Tablet OS, and
- flavors of the open source operating system Linux.


Types of Operating System
Batch Operating System
Multiprogramming Operating System
Multitasking Operating System
Multiprocessing Operating System
Real time Operating System



User Interface

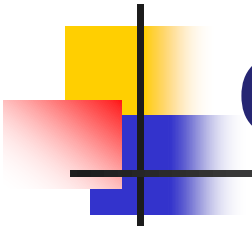
- **User Interface :** User Interface comprises of everything the user can use to interact with the computer. It is basically the means by which the **user and computer system can interact** using input and output devices.

- GUI and CUI are two types of **User Interfaces**.
  - GUI stands for **Graphical User Interface** while
  - CUI stands for **Character User Interface.**

GUI

- **GUI:** GUI stands for **Graphical User Interface**.

- This is a type of user interface where user interacts with the computer **using graphics**.
- Graphics include icons, navigation bars, images etc. Mouse can be used while using this interface to interact with the graphics.
- It is a very **user-friendly** interface and requires no expertise. Eg: Windows has GUI.



CUI

- **CUI:** CUI stands for **Character User Interface**.
- This is a type of user interface where user interacts with computer **using only keyboard**.
- To perform any action **a command is required**. C
- UI is precursor of GUI and was used in most primitive computers. Most modern computers use GUI and not CUI. Eg: MS-DOS has CUI.

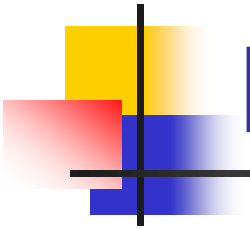
Comparison GUI and CUI

GUI	CUI
Symbols, pictures and pointing commands are use to execute commands	Set of characters and words are used to execute the commands
No need to remember the command. A small practice enables you to use the commands	The syntax and various options are required to be remembered.
General menu structure and commands are used for all the applications.	Different application have their own set of commands.
Number of applications can be opened and executed in different window at the same time	Only one application can run at a time
Minimum use of keyboard	Maximum use of keyboard
Mouse extensively used	Mouse used only in some applications
Easy to operate	Not user friendly hence difficult to operate

GUI VS COMMAND LINE

DOS commands

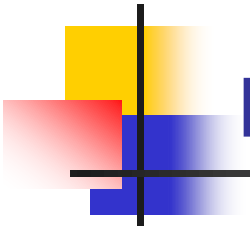
- **ATTRIB**
  Change file attributes. **+** adds an attribute, **-** removes it.
  Attributes are: A=archive; R=read only; S=system; H=hidden.
  ATTRIB -R -A -S -H <VIRUS.EXE>
- **C:** Go to the C: drive. Similarly A: and D: etc.
- **CD** Change directory. CD <DIRECTORYNAME>
- **CLS** Clear the screen. CLS
- **DEL** Delete one or more files in the current directory.
  DEL <VIRUS.EXE>



DBMS

- A **database management** system is important because **it** manages data efficiently and allows users to perform multiple tasks.

- A **database management** system stores, organizes and manages a large amount of information within a single software application.



Benefits of database development

- reduce the amount of time you spend managing data.
- analyse data in a variety of ways.
- promote a disciplined approach to data management.
- turn disparate information into a valuable resource.
- improve the quality and consistency of information.

e- AGRICULTURE and ICT

ICT (Information and Communication Technologies) refers to technologies that provide access to information through telecommunications medium.

Medium of ICT

1. Radio 2. television 3. Cell phone 4. computers 5. satellite technology 6.Internet including email 7. instant messaging, video conferencing and 8. social networking websites which have made it possible for users across the world to communicate with each other to give users quick access to ideas and experiences from a wide range of people, communities and cultures.

➢ Information and Communication Technology (ICT) consists of three main technologies.

They are:

- Computer Technology
- Communication Technology and
- Information Management Technology.

❖These technologies are applied for processing, exchanging and managing data, information and knowledge.

UNIQUE FEATURES OF ICT

- Access to the surprising store-house of information is free
- The information is available immediately round the year and twenty four hours a day,
- Communication can also be interactive
- The information is available from any point on the globe
- The communication is dynamic and ever growing.

Perks of ICT in agriculture

- Access to price information
- Access to agriculture information
- Access to national and international markets
- Increasing production efficiency
- Creating a conducive policy environment

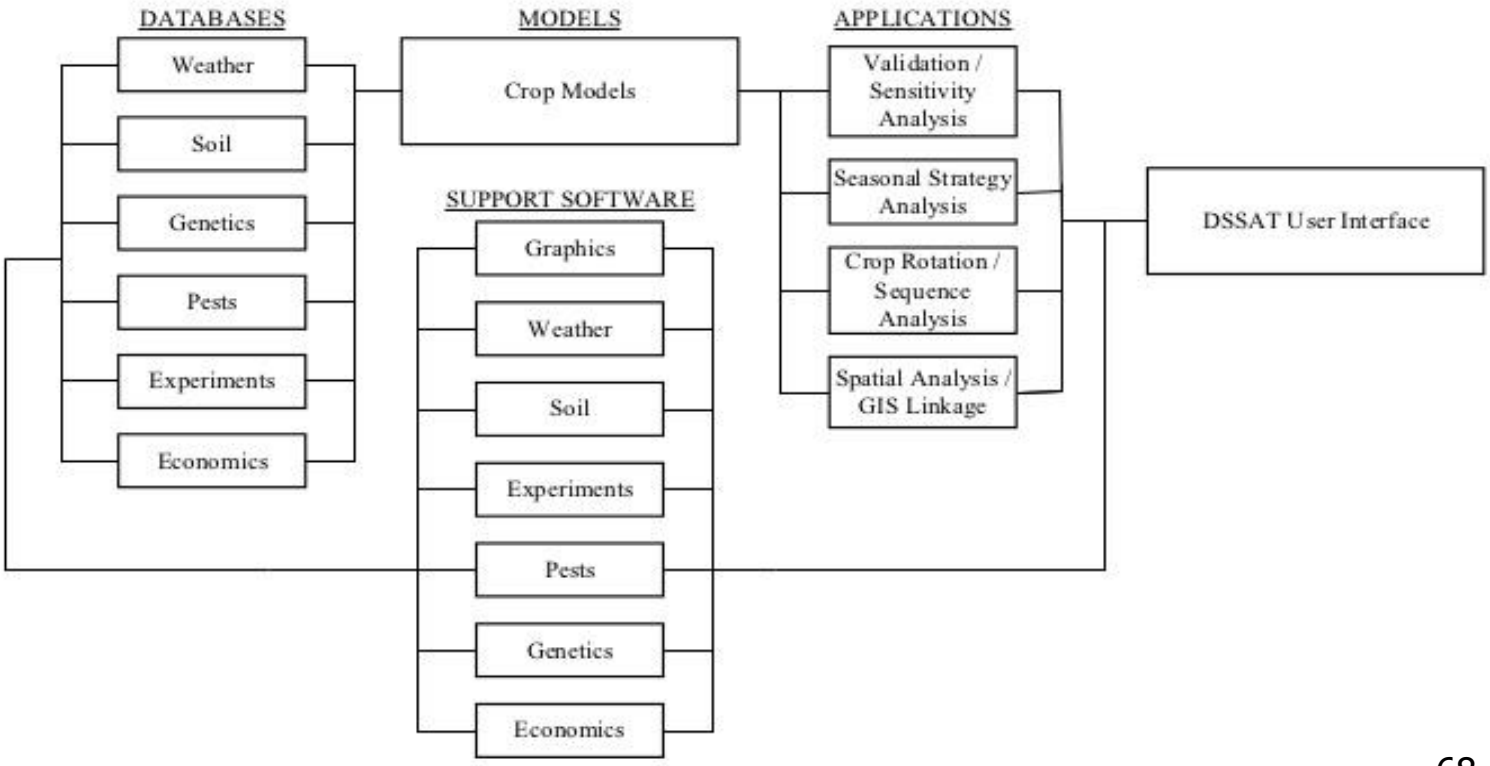
Decision Support System (DSS)

- A decision support system (**DSS**) is a computer-based application that collects, organizes and analyzes business data to facilitate quality business decision-making for management, operations and planning**.**
- **Programmed and Non-programmed Decisions**

DSSAT

- The **Decision Support System for Agrotechnology Transfer** (**DSSAT**) is a set of computer programs for simulating agricultural crop growth. It has been used in over 100 countries by agronomists for evaluating farming methods.
- DSSAT typically requires input parameters related to soil condition,weather, any management practices such as fertilizer use and irrigation, and characteristics of the crop variety being grown.

Components of DSSAT

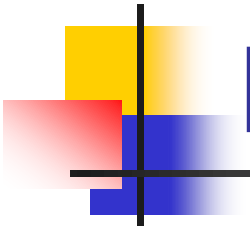


Computer Models In Agriculture

- In addition to field and laboratory experiments, the use of **computer models.**
- It can provide a cost effective and efficient way to analyze.
- Analyze impact of various **agricultural** management practices on crop yields.
- soil and water quality.
- **Crop Simulation Model**

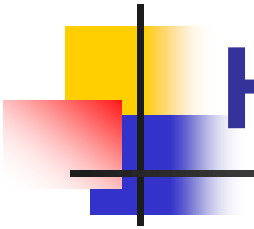
Types of crop simulation models

- **Statistical models**
  - These typically rely on yield information for large areas
- **Mechanistic models**
  - These attempt to use fundamental mechanisms of plant and soil processes to simulate specific outcomes.
- **Functional models**
  - These use simplified closed functional forms to simulate complex processes.



Protocol

- A set of rules or procedures for transmitting data between electronic devices, such as **computers**.
- In order for **computers** to exchange information, there must be a pre-existing agreement as to how the information will be structured and how each side will send and receive it.
- In order for two computers to talk to each other, they must be speaking the same language.
- **TCP**/**IP**, or the Transmission Control Protocol/Internet Protocol, it is a suite of communication protocols
- HTTP, FTP, Telnet, SMTP,



HTTP

- **HTTP** means HyperText Transfer Protocol.
- **HTTP** is the underlying protocol used by the World Wide Web.
- This protocol defines how messages are formatted and transmitted, and what actions Web servers and browsers should take in response to various commands.

Telnet (TELecommunication NETwork)

- **Telnet** is a **protocol** that allows you to connect to remote **computers** (called hosts) over a TCP/IP network (such as the internet).
- Using **telnet** client software on your **computer**, you can make a connection to a **telnet** server (that is, the remote host).
- This network **protocol** that provides a command-line interface to communicate with a device.

FTP (File Transfer **Protocol)**

- File Transfer **Protocol** (**FTP**)
- It is a client/server **protocol** used for transferring files to or exchanging files with a host **computer**.
- **FTP** allows users to access files, programs and other data from the Internet without the need for a user ID or password.
- The File Transfer **Protocol** (**FTP**) is a standard network **protocol** used for the transfer of **computer** files between a client and server on a **computer** network.

SMTP (Simple Mail Transfer **Protocol**)

- Simple Mail Transfer **Protocol**, a **protocol** for sending e-mail messages between servers.
- Most e-mail systems that send mail over the Internet use **SMTP** to send messages from one server to another.
- It only works for outgoing emails.
- When you send an email, the sender, recipients, email body and title heading are separated into sections. SMTP separates the sections using code words.

1.KEYBOARD

Keyboard is the primary input device of the PC. You use the keyboard to enter commands and type text. The keyboards on computers are similar to typewriters. However, a computer has many additional keys. Computer keyboards have not changed a lot since they were introduced. The only changes have been additions to the number of keys in the original keyboards. Present day keyboards have 101 or more keys. Each of these keys performs a different operation.

Types of Keyboards

Type of Keys

- Function Keys
- Alphanumeric Keys
- Numeric Keys
- Navigation Keys
- Punctuation Keys
- Modification Keys
- Special Keys
- Windows Keys

TYPE OF KEYBOARD

1. AT Keyboard
2. P/2 Keyboard
3. USB Keyboard

AT Keyboard

- AT (Advanced Technology) Keyboard

  The 5-pin DIN connector was widely in use in the past. The DIN stands for Deustche Industries Norm. This type of connector is rarely used nowadays.

  (Male connector)

PS/2 Keyboard

- PS/2 (Personal Standard/2)

  IBM introduced the PS/2 port on the back of PCs to connect devices such as the mouse and the keyboard. The PS/2 port supports the 6-pin PS/2 connector, which is very popular in the present. It has replaced most of the 5-pin DIN connectors. The 6-pin PS/2 connector is also known as the 6-pin mini DIN connector

USB Keyboard

- USB (Universal Serial Bus) Keyboard

  The 4-pin USB connector for keyboards is fast replacing the 6-pin PS/2 port as the most widely used connector for keyboards. The USB keyboard is connected to the USB port on the back of the computer and tends to be the fastest.

2.MOUSE

- Douglas Engelbart invented the computer mouse in 1963. It is a device that enables a computer user to move the cursor or pointer to a specific point on the screen. It was given its name because of its appearance and movement, which are similar to those of a mouse. When you move the mouse, the cursor on the screen also moves in the same direction. The mouse is the most used input device after the keyboard. The mouse is particularly important for Graphical User Interface (GUI) applications. You can also use the mouse for drawing objects on the screen by using the mouse as a pencil or paintbrush


Types of Mice

Types of Mouse Devices

There are different types of mouse devices that are widely in use at present. The connected mouse has been replaced to a great extent with cordless and optical mouse devices

- Mechanical Mouse
- Optomechanical Mouse
- Optical Mouse
- Cordless Mouse

Mechanical Mouse

A mechanical mouse is the oldest type of mouse available today. This type of mouse consists of a rubber or metal ball on the underside. The mouse is placed on a mouse pad. As the mouse is moved, the ball rolls in the appropriate direction. This type of mouse has mechanical sensors that detect the direction of movement of the mouse and send signals to the computer to move the pointer accordingly.

Optomechanical Mouse

- The Optomechanical mouse is similar to the mechanical mouse. However, this type of mouse makes use of optical sensors instead of mechanical sensors. The optomechanical mouse is more accurate than the mechanical mouse.

Optical Mouse

- An Optical Mouse uses laser technology to detect the movement of the mouse. It is the most accurate and precise type of mouse. The optical mouse has no moving parts. The optical mouse must be moved along a special mouse pad that has grids on its surface. The grids on the mouse pad provide a frame of reference for the working of the mouse.

Cordless Mouse

- A cordless mouse establishes an infrared or radio link with a transceiver (transmitter/receiver) that is connected to the computer through a cable. Cordless mouse are very convenient and easy to use. Cordless mouse operate on batteries.

Mouse Connectors

- Serial Connector
- PS/2 Connector
- USB Connector

Serial Connector

- The serial mouse connects to the serial port on the computer. This type of connection offers low speed and hence it is being replaced with other connections like the PS/2 and USB ports.

PS/2 Connector

- The PS/2 mouse port on the back of the PC connects to the mouse. This port was originally invented so that other devices like modems could use the serial port. The PS/2 port was also known as the mouse port. This is the most widely used port for connection of a mouse to the computer.

USB Connector

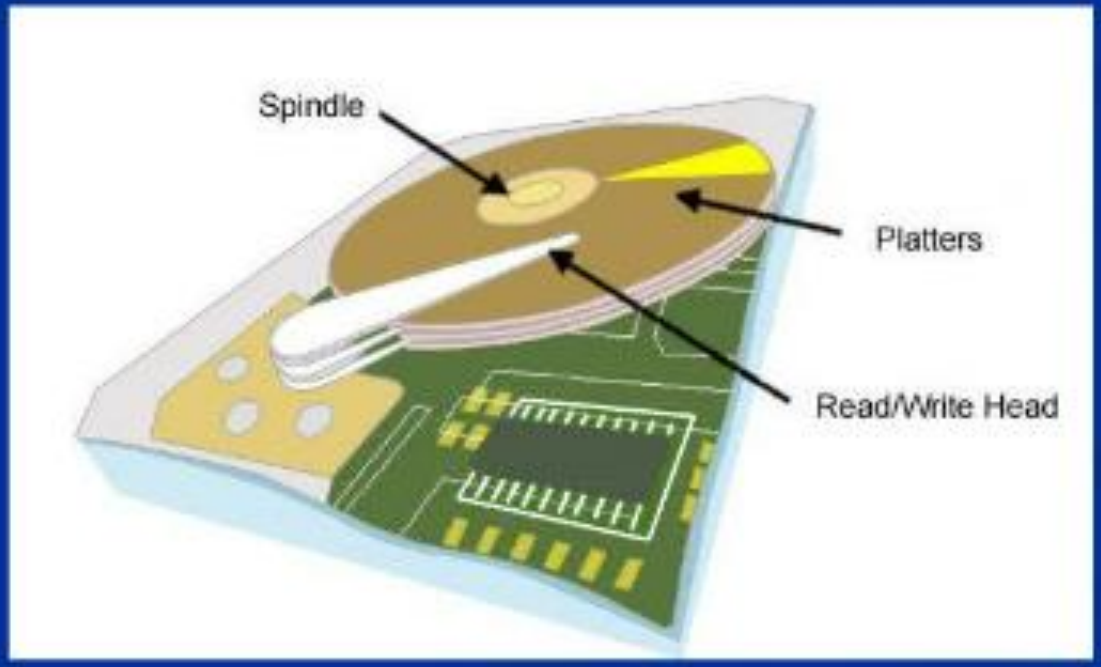
- Some types of mouse devices connect to the computer through the USB port. The USB connection offers good speed.

3. Hard Disk Drive

- A hard disk is the primary and permanent data storage device that is placed in the system. It is similar to a human brain where all the past and present events are stored. It is made up of a magnetic material that helps in storing data for the system by following the magnetic recording techniques. A hard disk stores data from 1 GB to 160 GB or even more depending on the capacity of the hard disk. A hard disk consists of several circular platters and each platter has read/write heads on both the sides of it. The platters are divided into concentric rings, called as tracks, and each track is divided into number of sectors. The read/write heads examines and then records the data in these sectors.

Components of the Hard Disk

- The different components such as the platters, the read / write head and the head actuator form the hard disk. These components are sealed inside the hard disk compartment known as Head Disk Assembly (HAD)



Platters

- The platters are round disks that are made up of metal or glass. The platters of glass are preferred, as the shape of the glass does not change when the hard disk heats up. The platters in the hard disk are stacked over each other. The size of the platter determines the size of the hard disk

Read/Write Head

- The Read/Write head is used to read the data stored on the hard disk and also write the data to the hard disk. While reading, the head converts the data from binary to a magnetic pulse. The magnetic pulse charges the magnetic coating on the platter and stores the data on the disk. While reading data from the disk, the head reads the magnetic data stored on the disk and converts them to binary and sends it to the system. Every platter in the hard disk has two read/write heads one on each side of the platter. While the disk is reading or writing data to the disk the head does not touch the disk. However, when the disk stops spinning the head gently rests on the stationary disk.

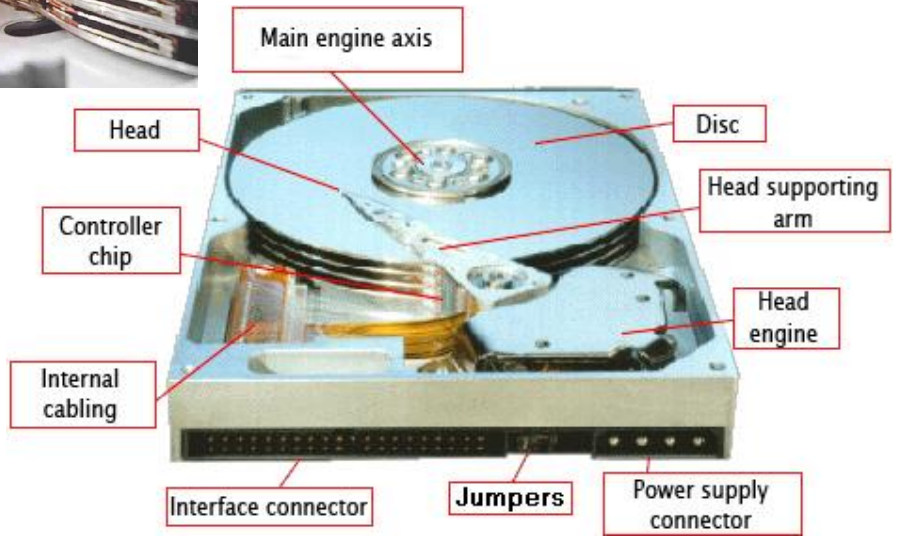
Logic Board

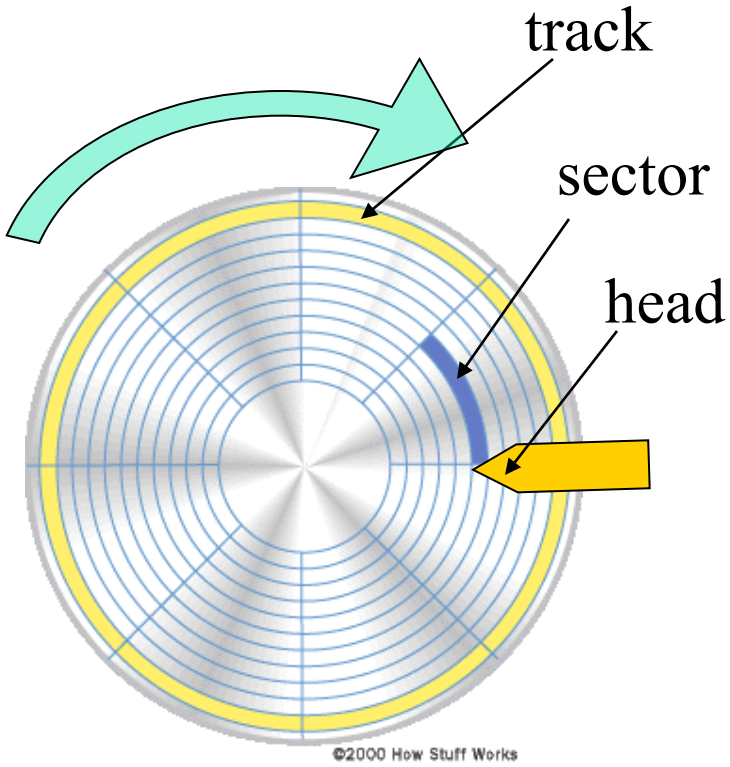
- The Logic Board contains the circuit for controlling the hard disk. All the connectors from the system connect to the logic board. The logic board stores the data on the hard disk platters. Sometimes the logic board of the hard disk may fail due to surges in the power supply. In such a case you can replace the logic board of the hard disk.

Hard Disk Jumpers

- You can also set the hard disk to master, slave or cable select as per the jumper settings displayed on a label affixed on the top cover of the hard disk. If you have two hard disks installed on the system, you must set the jumper of one hard disk to master and the jumper of the other hard disk to slave. This ensures that the correct hard disk responds to the system calls. The system considers the master first and then the slave while assigning drive letters. The system also boots the primary partition from the hard disk set as master.

HD


Main engine axis
Head
Disc
Head supporting arm
Controller chip
Head engine
Internal cabling
Interface connector
Jumpers
Power supply connector


track
sector
head
©2000 How Stuff Works

Hard Disk Inside

Types of Hard Disks

- The different hard disks specify the speed at which the hard disk transfers data and the reliability of the hard disk in storing the data if the hard disk crashes.

Advanced Technology Attachment (ATA)

Serial ATA (SATA)

Small Computer Systems Interface (SCSI)

Advanced Technology Attachment (ATA)

- The ATA transfers data between the hard disk and the system using 16 bits with speeds of upto 100 MB per second. You can connect two hard disks to a single controller on the system. You must set the hard disk to master or slave using the jumpers on the hard disk, so that the system can identify the required hard disk. This hard disk uses a 40-pin connector to connect to the system. And 4 Pins Power Connector.

Serial ATA(SATA)

- The Serial ATA transfers data between the hard disk and the system using only 1 bit at a time with the speed of upto 600 MB per second. You can connect only one Serial ATA hard disk to a single controller on the system, thus this hard disk does not have settings such as master, slave or cable select. The Serial ATA hard disk uses a 7 wire cable to connect to the system. And 15 Pins or 4 Pins Power Connector.

Serial ATA(SATA)

- The Serial ATA transfers data between the hard disk and the system using only 1 bit at a time with the speed of upto 600 MB per second. You can connect only one Serial ATA hard disk to a single controller on the system, thus this hard disk does not have settings such as master, slave or cable select. The Serial ATA hard disk uses a 7 wire cable to connect to the system. And 15 Pins or 4 Pins Power Connector.

SATA Connector

Sata Cables

5.CD-ROM AND CD WRITER DRIVES

Introduction :

- The data stored on the CD can last for many years. The CD can store large amounts of data and is used to distribute software, music, and movies. The CD drive reads the data stored on the CD. The CD-ROM drive can only read data from a CD. Today CD-R and CD-RW drives are available that can read, and also burn data on a CD.

CD Drive

CDs

DVDS

CD-ROM Disc

The 3.5 inch floppy disk stores about 1.44 MB of data. This enabled users to only copy files of small size on the floppy disk. The CD stores more data than the floppy disk. The CD holds about 650 - 700 MB of data. The CD is made out of a clear piece of polycarbonate plastic. The CD has different layers that enable the CD drive to read data from the CD. The layers that make up the CD depend on the type, and the number of times that you can write data on the CD.

Optical Discs

Use laser technology to read and write data on silver platters

Compact Disk (CD) can store 650MB to 800MB of information and data. CD-ROM (Read Only Memory) can only read data from a CD-ROM.

You can store data on a CD only if you have a CD Burner and CD-R (writable) or CD-RW (rewritable) CD.

DVD (Digital Versatile Disk) is the size of a regular CD and can be played in a regular in a DVD movie player.

DVD can store 4.8GB to 8.0GB of information and data. DVD-ROM is readable only (a movie DVD).

You can store data on a DVD only if you have a DVD Burner and DVD+R/DVD-R (writable) or DVD-RW (rewritable) DVD.

- **CD-ROM Drive**

The CD-ROM drive can read data from the CD-ROM disc and the Audio CD. You can use this drive to install software from a CD and play music from an Audio CD.

The speed of the CD drive specifies the amount of data the CD drive reads from the CD in a second. This speed is called the transfer speed of the CD drive. The original speed of the CD drive is 150 KB per second. The speed of the CD drive is specified using a number followed by X. For example, if the speed of the CD drive is specified as 52X, then the CD drives reads data from the CD at a speed of 52X. The number specifying the speed of the CD drive must be multiplied with the original speed of 150 KB per second, to get the speed of the CD drive. For example if the speed of the CD drive is specified as 12X, then the transfer speed is 12 x 150 = 1,800 kbps.

Connectors

- The IDE cable connects the CD drive to the motherboard. The CD drive uses the 40 pin connector. You must ensure that the marking on the IDE cable is on the same side of the power supply cable when inserting it in the CD drive. The CD-ROM drive is also equipped with output connections for sound. You can connect the cable from here to the sound card or the sound output device.


Jumper

DVD ROM AND COMBO DRIVES

- **Introduction :**
- Digital Versatile Disk or Digital Video Disk (DVD) is a form of digital storage. DVD is used to stores music, video, games, and multimedia applications. The DVD is similar to the CD in appearance and structure. The DVD offers higher capacity and better quality as compared to the CD. The DVD storage capacity is seven times that of a CD. To play a DVD, a DVD ROM drive is used. In addition, combo drives are available that play DVDs as well as CDs.

What is a motherboard?

- A **motherboard** is the *central circuit board* making up a personal computer.

A motherboard

Motherboard

- A personal computer is built with the CPU, main memory, and other essential components on the motherboard.
- Other components such as external storage, controllers for video display and sound, and peripheral devices are typically attached to the motherboard via edge connectors and cables.
- In modern computers it is increasingly common to *integrate these "peripherals" into the motherboard.*

PC Subsystems

- **Motherboard** – The main circuit board of a microcomputer

Form Factors of Motherboard

- The form factor of a motherboard refers to its physical shape, layout, and the positioning of the components on it. The form factor of the motherboard determines the type of system case it will fit into. Some motherboards that have the same functionality can be packaged in different form factors. The only major difference between such motherboards will be the form factor. Motherboards are available in different forms. Are as follow:-

Advanced Technology (AT)

Baby AT

Advanced Technology Extended (ATX) Form Factor

Advanced Technology (AT) Form Factor

- The AT form factor is also known as the full-size AT form factor. This form factor matches the original IBM AT motherboard in structure and layout. This type of motherboard is very large and is about 12 inches wide and 13.8 inches deep. The AT form factor is not used much by present-day motherboard manufacturers. This type of motherboard does not fit into most of the popular system cases and its size also makes installation and troubleshooting difficult.

Baby AT Form Factor

- The baby AT form factor was the most popular till recently. This form factor is very much similar to the original IBM XT motherboard structure. The baby AT form factor fits into most of the system cases. A baby AT motherboard is 8.5 inches wide and about 13 inches long. The length of baby AT motherboards is not fixed

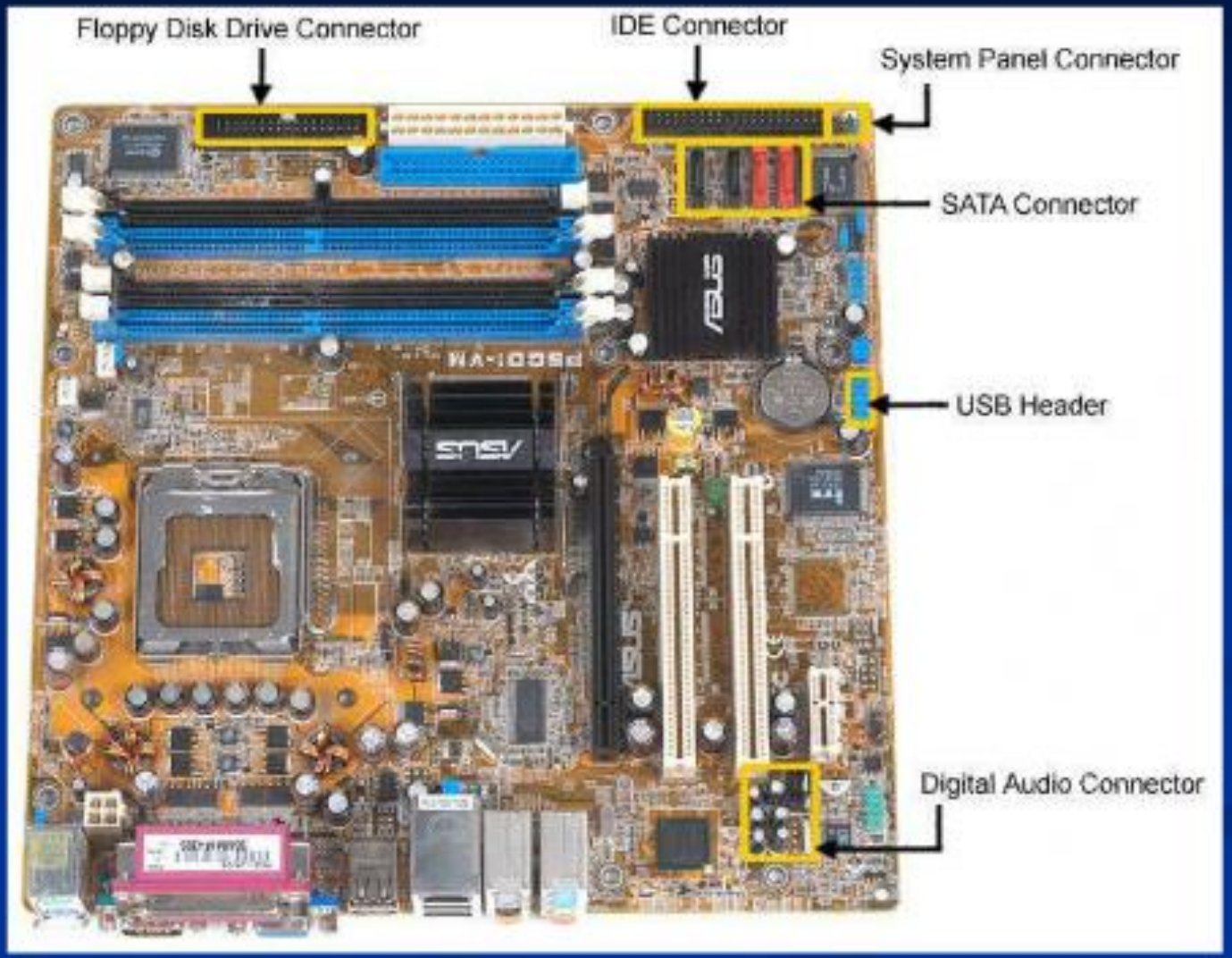
Advanced Technology Extended (ATX) Form Factor

- Intel invented the ATX form factor in the year 1995. The ATX and mini ATX form factors are the most popular ones presently. The ATX form factor has many of the best features of the LPX and AT form factors. The ATX form factor is not compatible with the LPX or AT form factors. As a result, new system cases and power supplies were designed to match this form factor. These cases and power supplies have now become quite common. The ATX

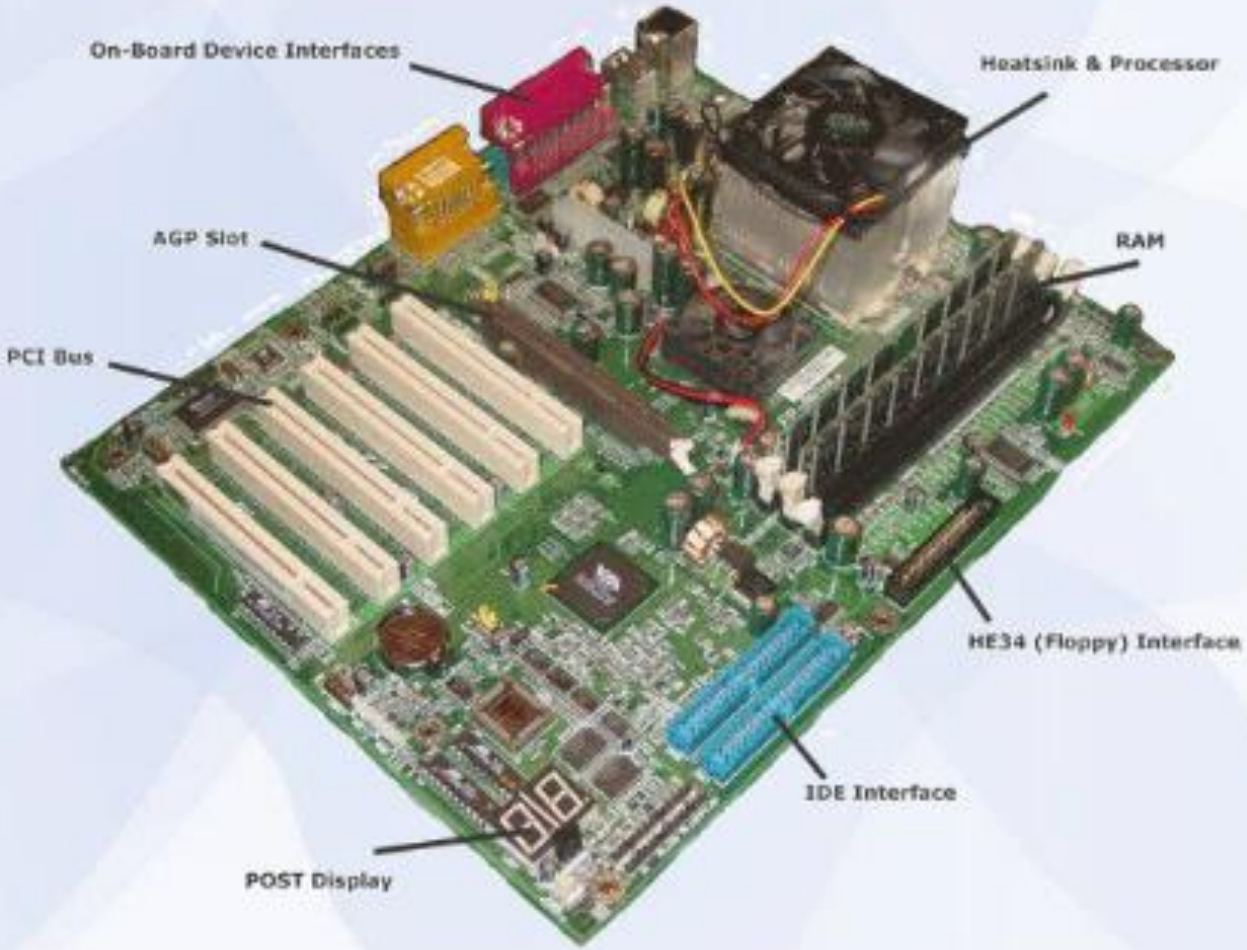
Components of a Motherboard

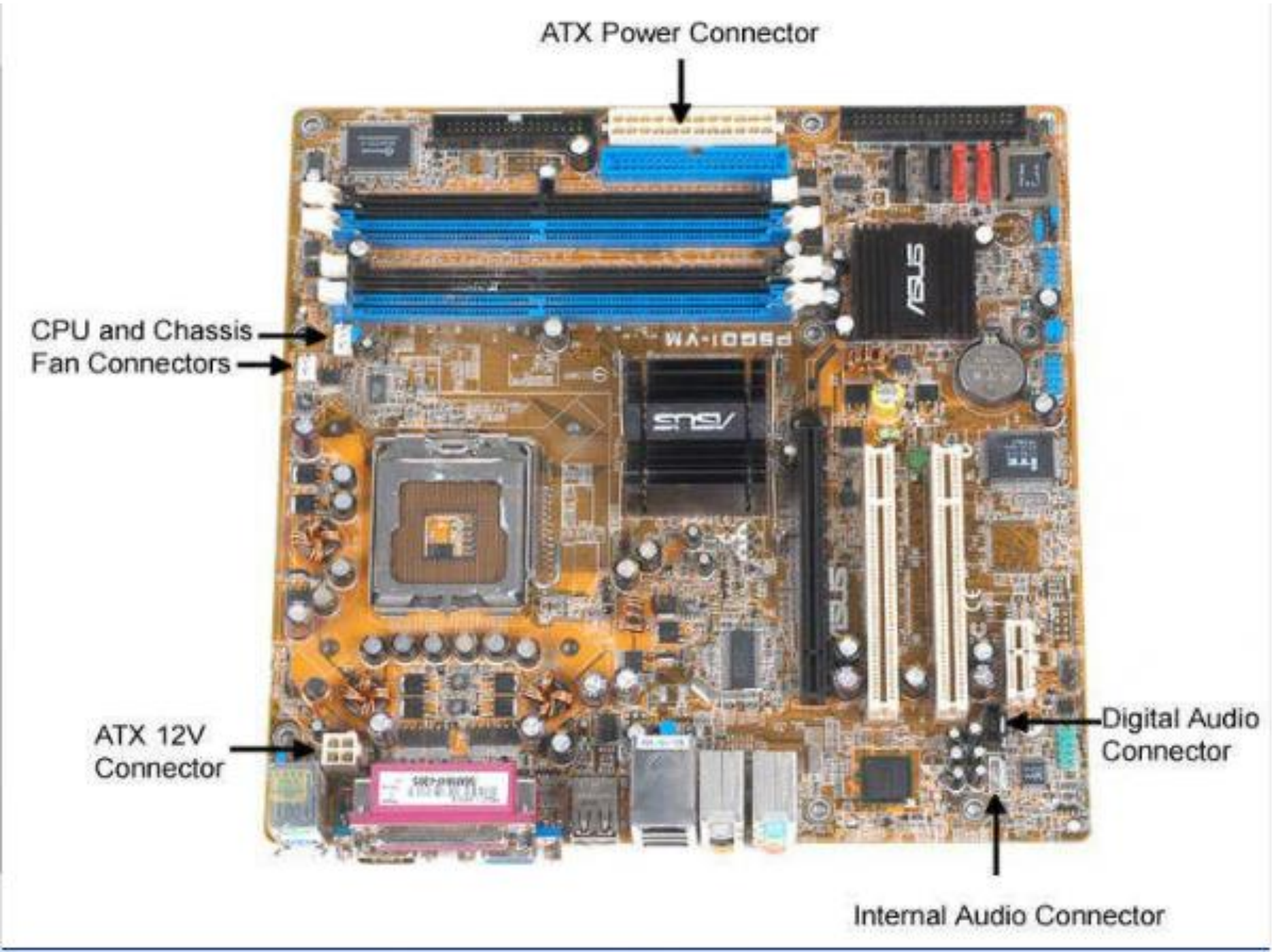
- The motherboard has several connectors, jumpers and expansion slots for connecting various components of the system. You can connect various devices of the system to the connectors and expansion slots of the motherboard. You can configure the motherboard using the jumpers.

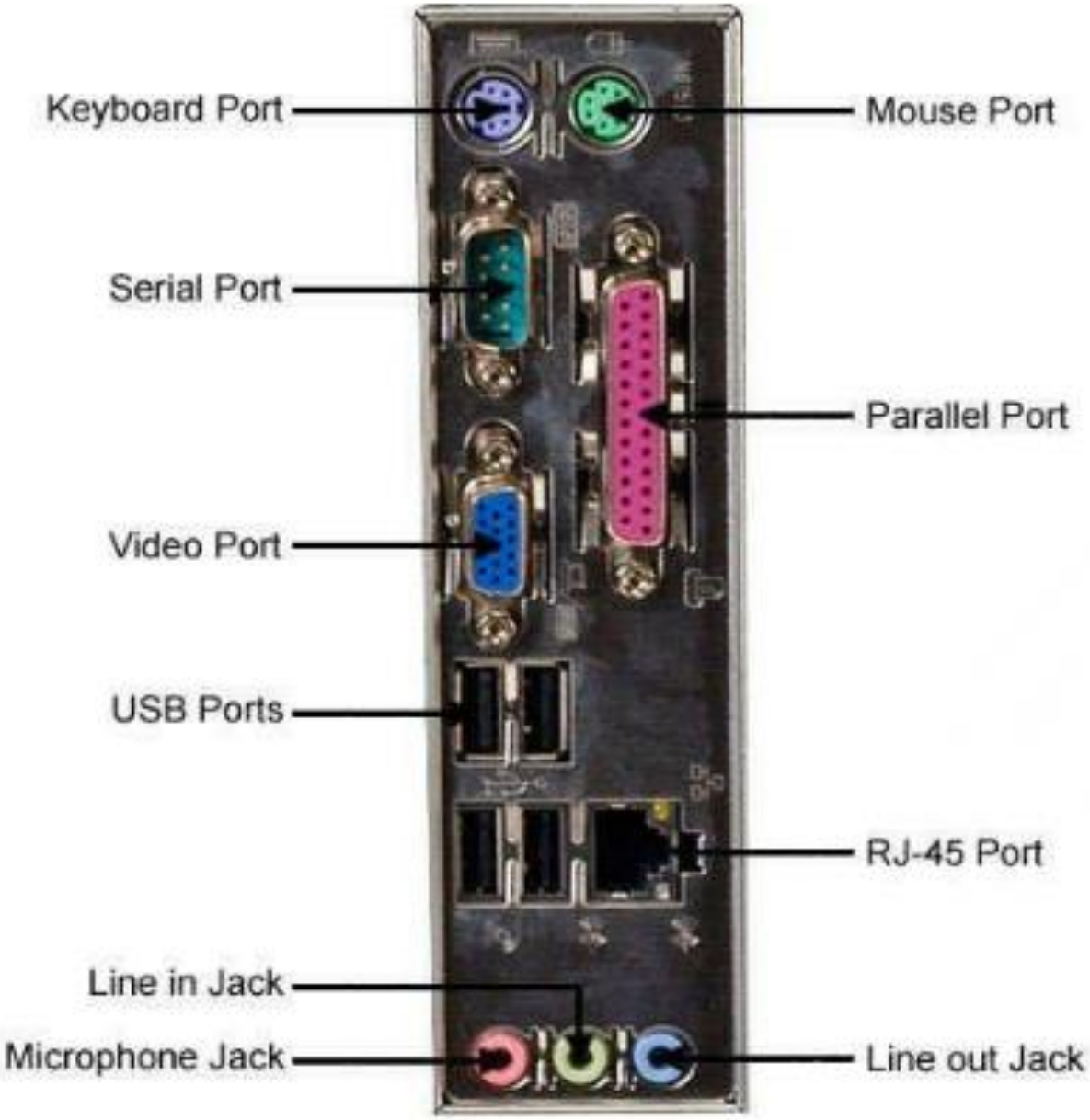
- Connector


Floppy Disk Drive Connector
IDE Connector
System Panel Connector
SATA Connector
USB Header
Digital Audio Connector

Mother Board




ATX Power Connector
CPU and Chassis Fan Connectors
ATX 12V Connector
Digital Audio Connector
Internal Audio Connector


Keyboard Port
Mouse Port
Serial Port
Parallel Port
Video Port
USB Ports
RJ-45 Port
Line in Jack
Microphone Jack
Line out Jack

Expansion Slots

- The expansion slots on a motherboard enable you to connect the expansion cards to the motherboard. The different PCI cards include LAN card, SCSI card and USB card. The cards must comply with PCI specifications. The motherboard has an Accelerated Graphics Port (AGP) slot. You can connect AGP cards that are compatible with your motherboard, to the AGP slot.

Bus Standards

- There are different types of I/O buses that transfer data across components of the system. The different buses have different widths and speeds. The common bus standards are:

Peripheral Component Interconnect (PCI) Local Bus

-

- Is the most popular I/O bus. It has the same speed and width as the VESA local bus. Some PCI buses have a width of 64 bits. It provides much better performance as compared to the VESA local bus. It has separate circuitry that controls it. PCI generally supports 3 or 4 slots. You can connect video cards, SCSI host adapters and network cards to the PCI expansion slot.

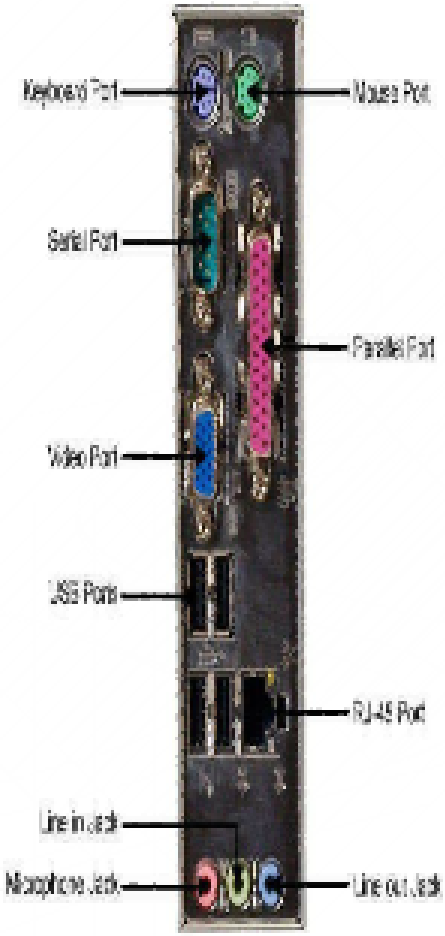
Accelerated Graphics Port (AGP) Bus

- Provides high performance graphics capabilities to a system. This bus forms a dedicated path between the chipset and the graphics subsystem. The AGP bus enables creation of 3D graphics.

I/O PORTS AND DEVICES

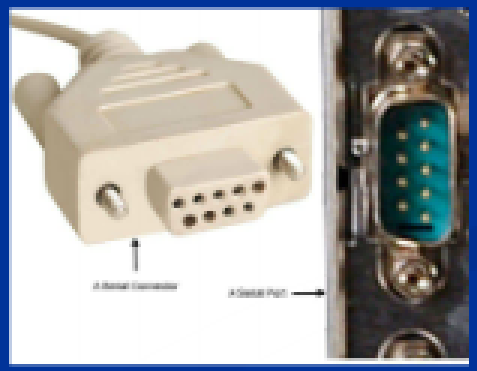
Introduction :

- Input/Output ports (I/O) enable you to connect hardware devices such as the keyboard, mouse, printer and scanners to the system. It is the entry and exit point for data from the system. I/O ports give you freedom in choosing and installing the device because if you have only few ports you can select a device that is available for that port.



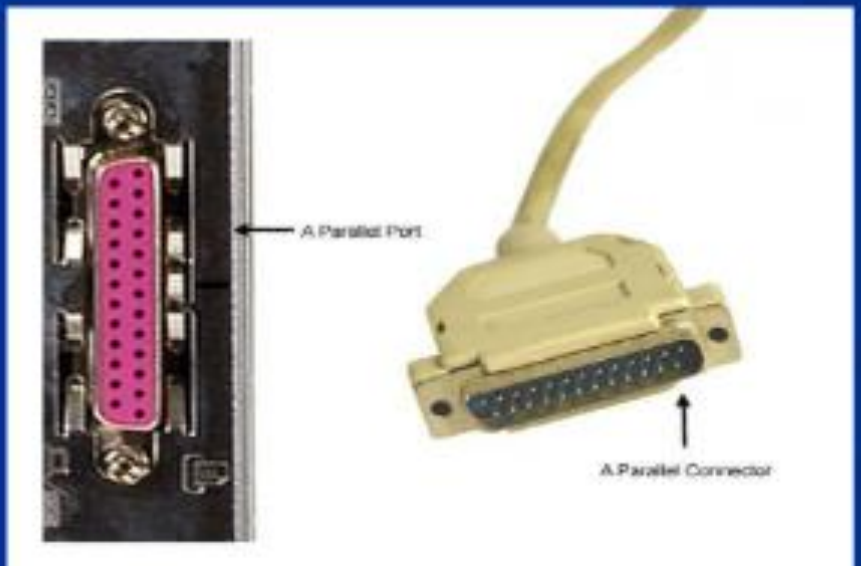
Serial Ports

- A serial port is like a single lane road that sends and receives one bit of data at a time. Thus, the eight bits of data in one byte travel one bit at a time, one behind the other. The serial port connector also known as the Communication or COM port can have 9 or 25 pins. A serial port is used to connect devices such as the mouse and modems to the system.



Parallel Ports

- The parallel port is like an eight lane road that transmits eight bits of data at a time. It is like eight cars moving on a wide road side by side. The parallel port connector has 25 pins to connect devices such as printers, scanners, external hard drives, and tape backup devices.

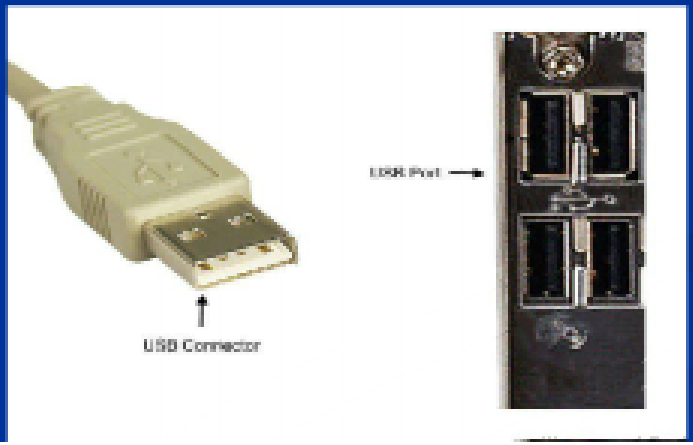


PS/2 Port

- The PS/2 port is used to connect the keyboard and mouse to the system. The ports are available in a color that matches the color of the plug connecting the mouse and keyboard. This port uses 6 pins to connect the device.

Universal Serial Bus (USB) Ports

- The USB port is a rectangular port that is used to connect a variety of devices to the system. The USB port also supplies power to the device such as the web camera, if the device does not use an external power source. To use the USB device you must just plug the device into the USB port, as most USB devices offer Plug-and-Play support. However, you must install the USB driver before using the USB port.



7.MICROPROCESSORS

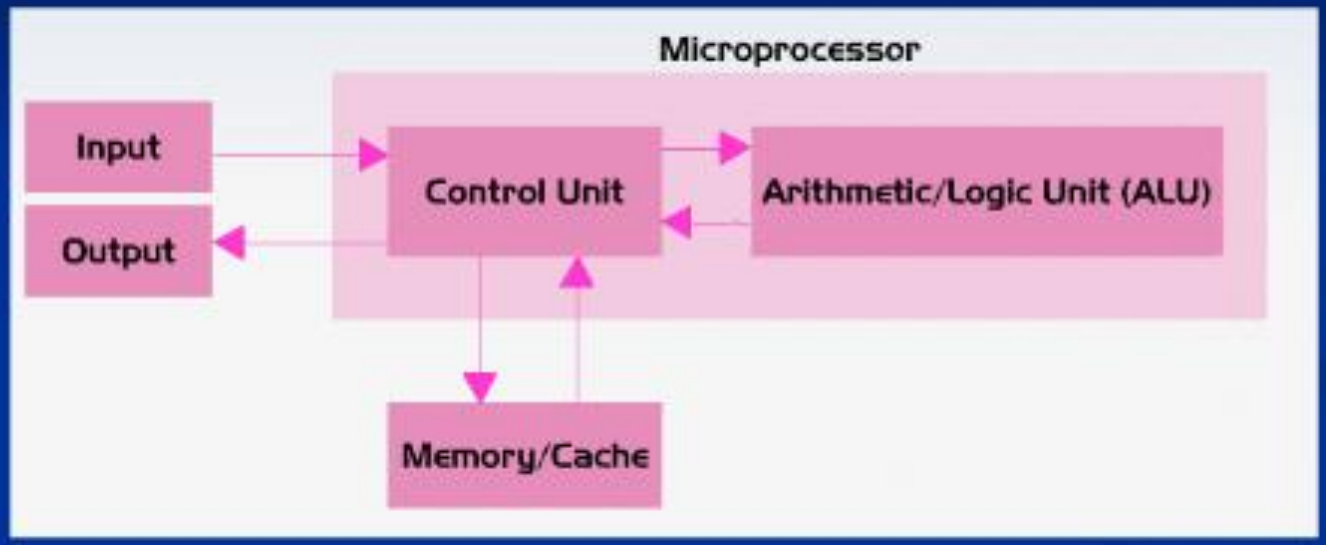
- Computer is capable of performing complex tasks, such as managing the brake system of a car. These tasks are processed by the Central Processing Unit (CPU). The CPU comprises of the microprocessor. The microprocessor accepts input from the user in the form of the data and instructions. It processes the data using the instructions and sends the processed information to the output device. The microprocessor controls the system, therefore it is important to understand it's working. The choice of the microprocessor also depends on computing needs

TYPE

Slot

Socket

Microprocessor


Microprocessor
Input
Output
Control Unit
Arithmetic/Logic Unit (ALU)
Memory/Cache

What are Intel CPU Families for Desktop Computers?

There are 3 Intel CPU families for desktop computers:

1. Core Processor
   http://www.intel.com/products/desktop/processors/index.htm
2. Pentium Processor
   http://www.intel.com/products/desktop/processors/pentium.htm
3. Celeron
   http://www.intel.com/products/desktop/processors/celeron.htm

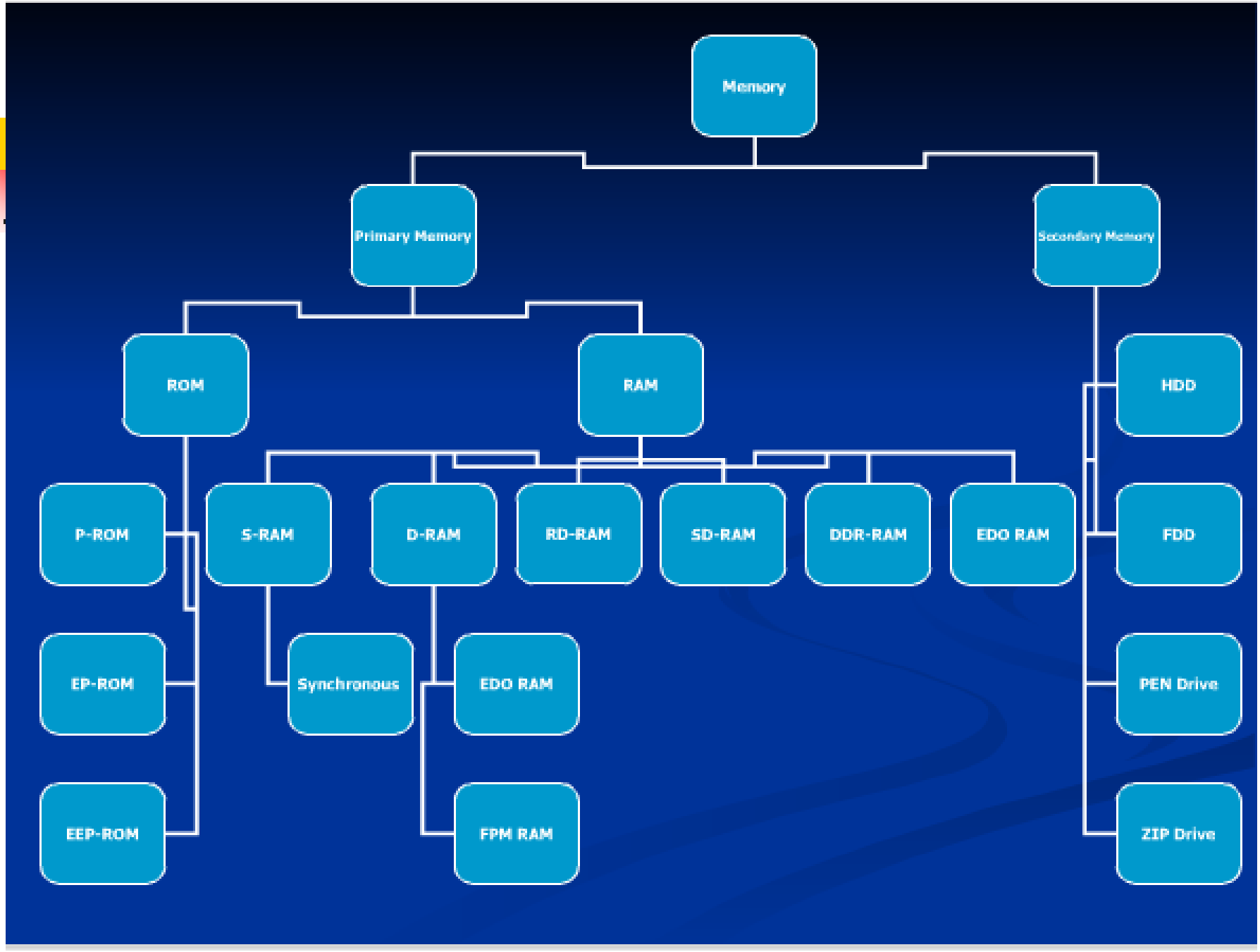
What are AMD CPU Families for Desktop Computers?

There are 3 AMD CPU Families for Desktop Computers:

1. Athlon 64 FX
   http://www.amd.com/us-en/Processors/ProductInformation/0,,30_11
2. Athlon 64 x 2 Dual – Core
   http://www.amd.com/us-en/Processors/ProductInformation/0,,30_11
3. Athlon 64
   http://www.amd.com/us-en/Processors/ProductInformation/0,,30_11

8.Memory

- **Introduction :**
- Memory is one of the functions of the brain that enables to store and remember the past events. Similarly, in computers the term memory refers to a chip that stores data. It also enables us to retrieve the stored data. The processor retrieves the information stored in the memory for processing the data. The storage capacity of a memory depends on the type of the memory package used


Memory
Primary Memory
Secondary Memory
ROM
RAM
HDD
P-ROM
S-RAM
D-RAM
RD-RAM
SD-RAM
DDR-RAM
EDO RAM
FDD
EP-ROM
Synchronous
EDO RAM
PEN Drive
EEP-ROM
FPM RAM
ZIP Drive

Physical Memory

Physical memory comprises of memory chips. Physical memory stores programs and data that the microprocessor requires. Therefore, it enables the microprocessor to access the required programs and data quickly. The different types of physical memory are:

- **Read Only Memory (ROM)**
- **Random Access Memory (RAM)**

Type of ROM

1. P-ROM (Programmable Read-Only Memory)
2. EP-ROM (Erasable Programmable Read-Only Memory)
3. EEP-ROM (Electrically Erasable Programmable Read-Only Memory)

Random Access (RAM)

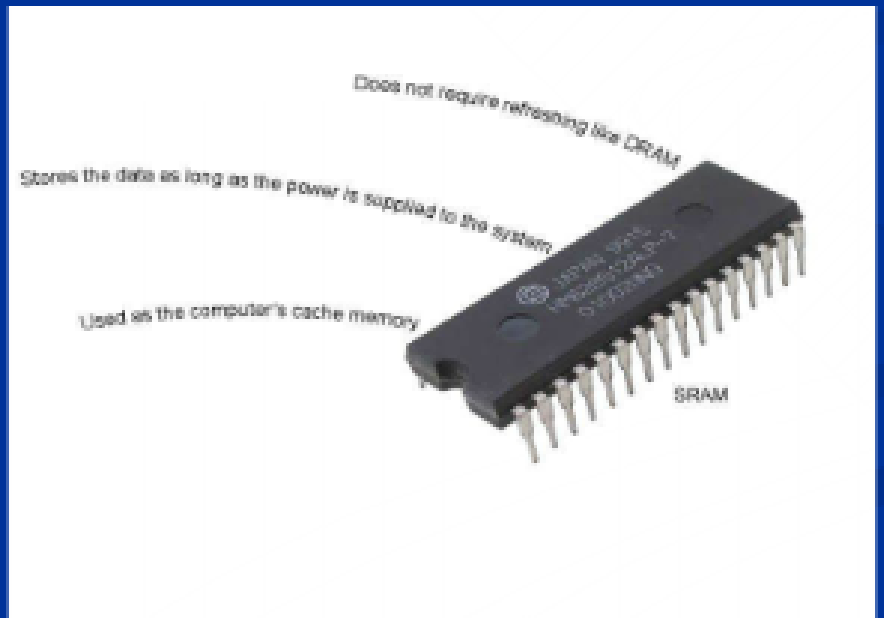
- Random access means any byte can be accessed in any order. The microprocessor can read and write programs and data to the RAM. It is a type of a volatile memory and therefore it is referred to as a temporary data storage area

Type of RAM

1. S-RAM
2. D-RAM
3. DDR-RAM
4. RD-RAM

S-RAM
(Static Random Access Memory)

- This memory is referred to as Static because of the fact that it does not require refreshing like DRAM. SRAM stores the data as long as the power is supplied to the system. This is because SRAM is made up of transistors that do not require refreshing. SRAM is used as the computer's cache memory.



D-RAM
(Dynamic Random Access Memory)

- Today, most PCs use DRAM as the temporary data storage area. DRAM stores data in the memory cells. Each memory cell contains a transistor and a capacitor. Capacitors lose their charge quickly and as a result, they need to be refreshed. Refreshing helps to retain the data in the memory. The life of data in DRAM is very short for about few milliseconds. In order to retain the data, the storage cells need to be refreshed after few milliseconds. Because of the constant refreshing of cells, this memory is referred to as Dynamic RAM

SD-RAM

(Synchronous Dynamic Random Access Memory)

- SDRAM synchronizes the memory speed with the CPU clock speed. The speed of the SDRAM depends on the speed of the CPU bus. It is faster than SRAM, DRAM, EDO DRAM, and VRAM memories. The data transfer speed of SDRAM is measured in nanoseconds and megahertz units. It runs with an average speed of 133 MHz.

DDR-RAM

(Double Data Rate Random Access Memory)

- It is the latest version of SDRAM. DDR is synchronous with the system clock. As a result, the data transfer rate of DDR is faster than SDRAM. It is almost twice the speed of the SDRAM. This memory chip consumes less power. DDR memory supports error correction code and non-parity. The server uses the error correction code, normally called as ECC.

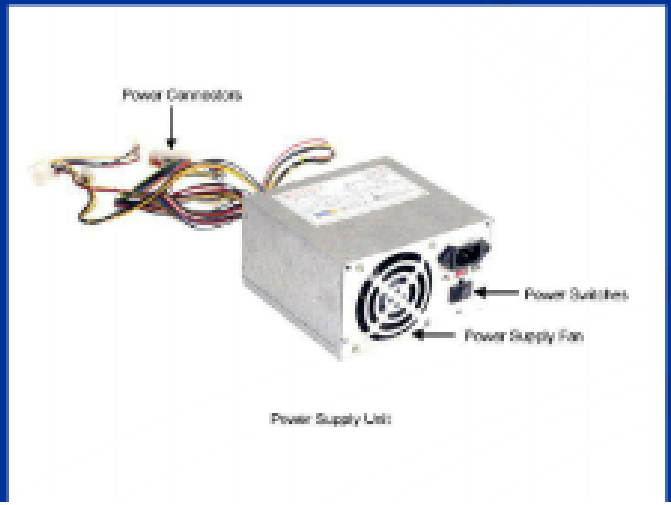
Memory Socket

Flash Memory

- Flash memory is used in the digital camera, cellular phones, LAN switches, PC Cards for notebooks, and video games. This memory performs the action in a flash and as a result, it is termed as the flash memory. Flash memory is a non-volatile memory. It is a type of EEPROM consisting of blocks. EEPROM is erased and reprogrammed in blocks. Flash memory is faster than EEPROM since it is reprogrammed at the block level whereas EEPROM is reprogrammed at the byte level

9. SMPS
(Switch Mode Power Supply)

- The power supply unit supplies power to the different system components such as the motherboard and the device drives. You must protect the system from extreme temperature changes like overheating using cooling devices. You can also use a UPS to protect the system from power fluctuations.

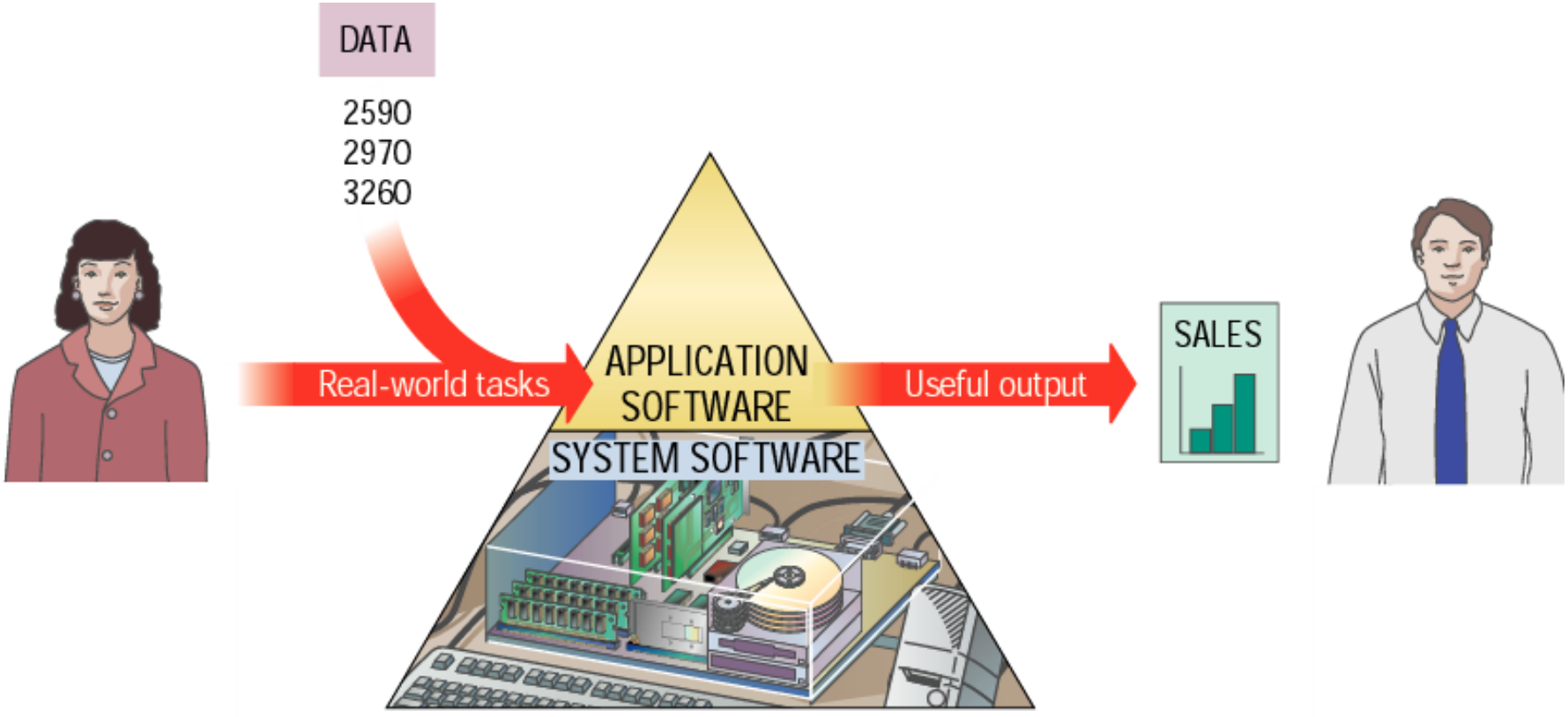


Type of Power Supply

- Personal Computer / Extended Technology (PC/XT)
- Advanced Technology (AT) Form Factor
- Advanced Technology Extended (ATX) Form Factor

Bringing the Machine to Life – What is Software?

- **Software is a set of electronic instructions that tells the computer how to do certain tasks. A set of instructions is often called a program.**
- **When a computer is using a particular program, it is said to be running or executing the program.**
- **The two most common types of programs are system software and application software.**


DATA
2590
2970
3260
Real-world tasks
APPLICATION SOFTWARE
SYSTEM SOFTWARE
Useful output
SALES

Bringing the Machine to Life – System Software

- System software exists primarily for the computer itself, to help the computer perform specific functions.
- One major type of system software is the operating system (OS). All computers require an operating system.
- The OS tells the computer how to interact with the user and its own devices.
- Common operating systems include Windows, the Macintosh OS, OS/2, and UNIX .

Bringing the Machine to Life - Applications

- Application Software consists of programs that tell a computer how to produce information
- Application software tells the computer how to accomplish tasks the user requires, such as creating a document or editing a graphic image.
- Some important kinds of application software are:

Word processing programs
Database management
Graphics programs
Web design tools and browsers
Communications programs
Entertainment and education

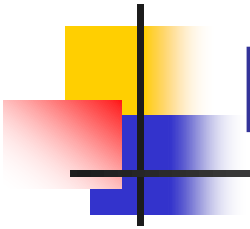
Spreadsheet software
Presentation programs
Networking software
Internet applications
Utilities
Multimedia authoring

OS
(Operating System)

- OS is program that act as an intermediary between a user of a computer and component of the computer Hardware.
- There are two Goal of OS 1) **Primary Goal** to make the computer system convenient to use for the use. 2) **Secondary Goal** is to use the computer Hardware.
- OS controls and co-ordinate the use of the computer Hardware in among the various application program for the various user of the computer.
- OS is one program running at all times on the computer i.e. Kernel with all the other program.

Function Of OS

1. OS provide the method for other programs to communication with the hardware of computer.
2. It create a user interface and to enables user to make changes in the computer.
3. OS must enables user to determine the installed programs and then, use, and shutdown the program of their own choice.
4. OS should enables user to add, move and delete the installed program and data.



Firmware

- Firmware are programs that are permanently written and stored in memory

BIOS

- The BIOS software enables you to control the system and the different hardware components without loading the operating system. The BIOS contains the code required to operate the hardware devices connected to the system such as the keyboard, mouse and the different ports connected to the system. The BIOS chip is of two types.

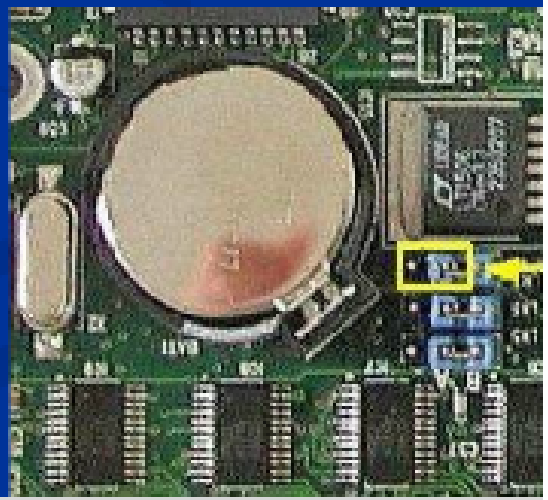
Operating Systems and BIOS

- Hardware in a PC does not know the software and BIOS is the interface between hardware and software.
- BIOS is also called firmware due to its integration with hardware.
- BIOS contains all the code required to control the keyboard, display screen, disk drivers, serial communications and a number of miscellaneous functions.
- BIOS is typically placed in a ROM chip that comes with the computer it is often called a ROM BIOS which ensures that the BIOS will always be available and will not be damaged by disk failures.

CMOS

(Complementary Metal Oxide Semiconductor)

The CMOS is the memory located on the motherboard that stores the BIOS settings. The CMOS has a size of 64 bits. When the system starts the BIOS loads the settings from the CMOS. The CMOS requires a power source to store the settings. It receives power from a battery that is installed on the motherboard. This battery must be replaced when it becomes weak or you may lose the stored BIOS settings. You can clear the BIOS settings stored in the CMOS using the jumpers located on the motherboard or by removing the battery from the motherboard



POST
(Power-On Self-Test)

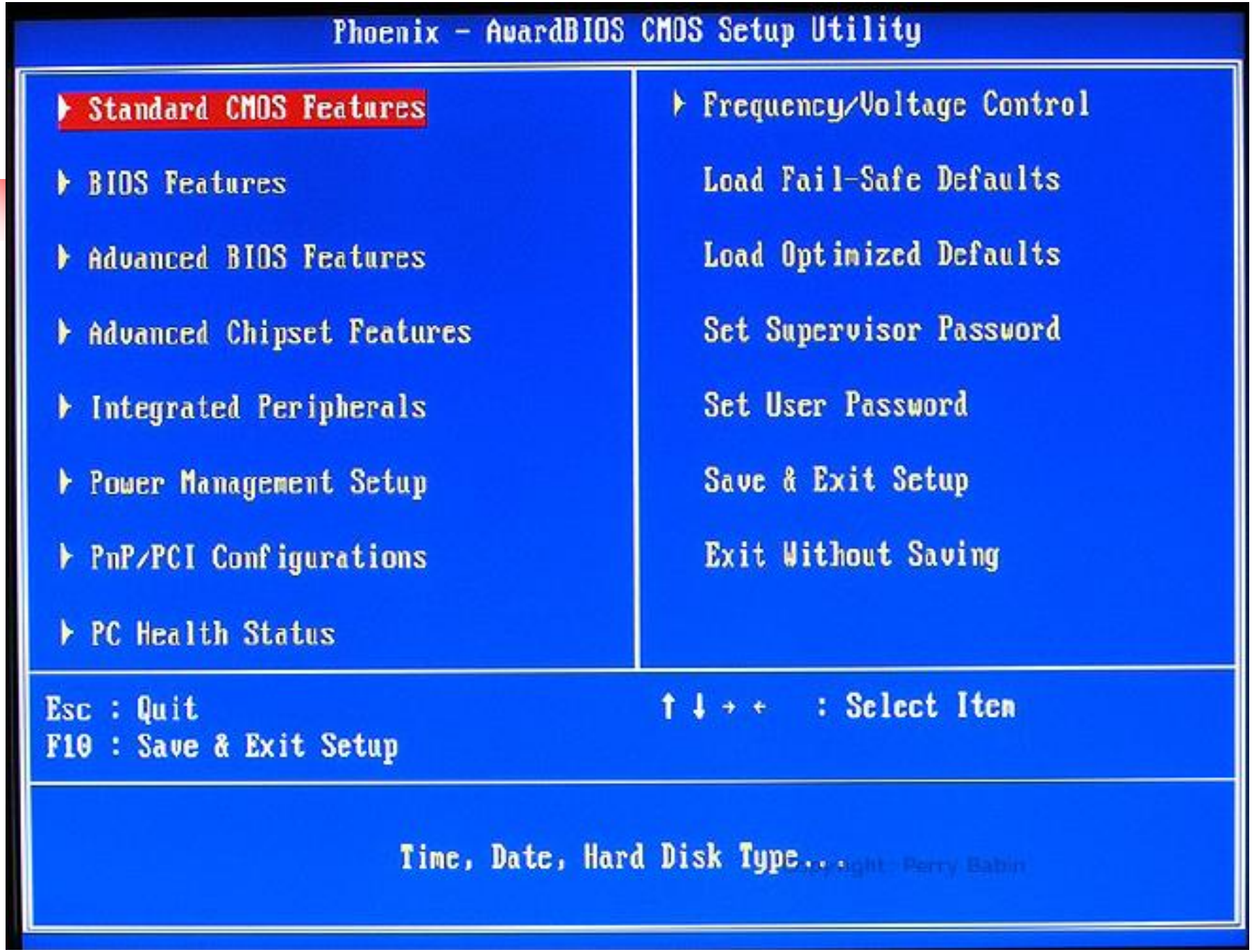
- When the system starts, the BIOS runs the power-on self-test (POST) to test the hardware connected to the system. These are the testing routines that are store in ROM BIOS. When POST delete an error either from the Keyboard, Mouse, Display or Memory or various other component it produces an error warning in the form of Message or Beep sound.

Starting a PC for the First Time

- When we start the PC, the Basic Input Output System (BIOS) runs a test to check if all the peripheral devices, memory and hardware of the PC are working properly. This test is called the Power On Self Test (POST). The PC will boot only if the results of the test are positive. If any of the above components of the PC are not in proper condition, the PC will give some warning signals like beeps. In some cases, if the problem is serious, the PC will not boot at all. The POST happens as soon as the PC starts and before most of the components of the PC start. The errors that cause the system to stop booting are of two types – fatal and non-fatal. In case of fatal errors, the boot process stops immediately. The following are the functions performed by POST:

- Checking the motherboard
  Comparing the system configuration with the PC Configuration Program to find any changes made
  Checking the memory devices and drives
  Checking the system memory
  Starts the display and audio devices

- In case there are no errors in the system configuration or devices, a single beep follows the POST. Then the booting process of the PC starts and the operating system is loaded.


Phoenix - AwardBIOS CMOS Setup Utility
Standard CMOS Features
BIOS Features
Advanced BIOS Features
Advanced Chipset Features
Integrated Peripherals
Power Management Setup
PnP/PCI Configurations
PC Health Status
Frequency/Voltage Control
Load Fail-Safe Defaults
Load Optimized Defaults
Set Supervisor Password
Set User Password
Save & Exit Setup
Exit Without Saving
Esc : Quit
F10 : Save & Exit Setup
↑ ↓ → ← : Select Item
Time, Date, Hard Disk Type...

Computer Software

Computer software is the key to productive use of computers. Software can be categorized into two types:

- Operating system software
- Application software.

Operating System Software

Operating system software tells the computer how to perform the functions of loading, storing and executing an application and how to transfer data.

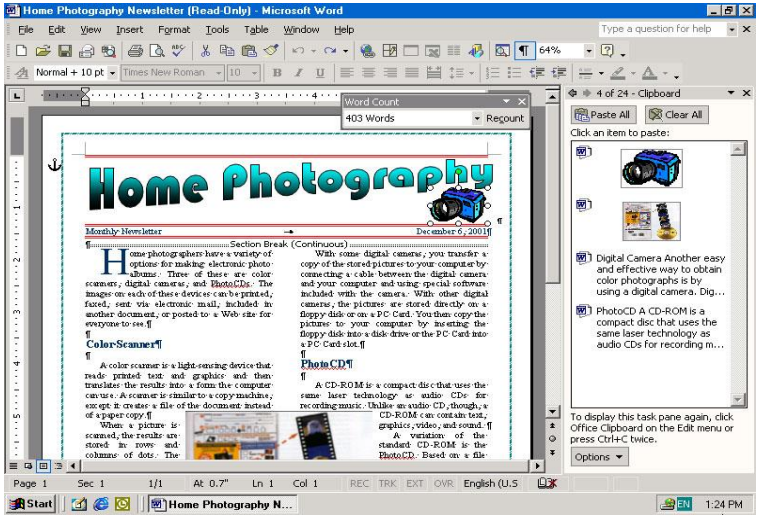
Today, many computers use an operating system that has a graphical user interface (GUI) that provides visual clues such as icon symbols to help the user. Microsoft **Windows 98** is a widely used graphical operating system. **DOS** (Disk Operating System) is an older but still widely used operating system that is text-based.

Application Software

Application Software consists of programs that tell a computer how to produce information. Some of the more commonly used packages are:

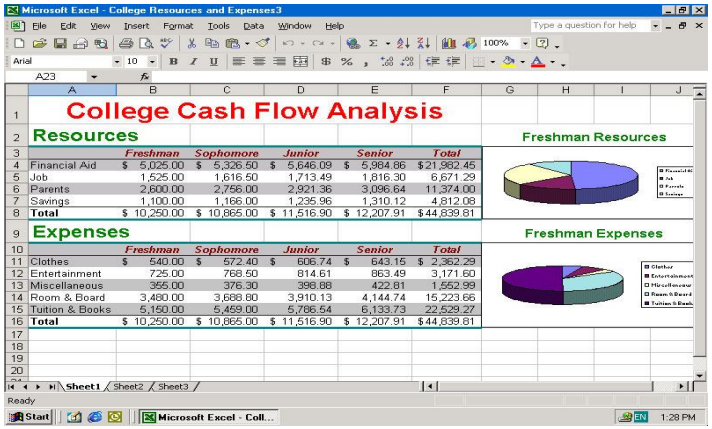
- **Word processing**
- **Electronic spreadsheet**
- **Database**
- **Presentation graphics**

Word Processing



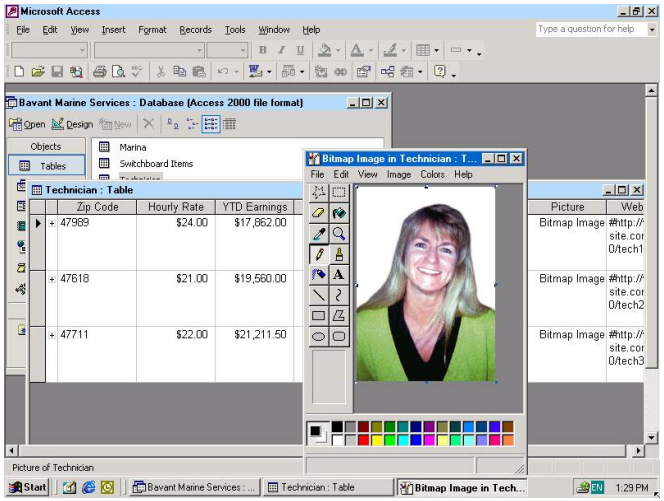
- Word Processing software is used to create and print documents. A key advantage of word processing software is that users easily can make changes in documents.

Electronic Spreadsheets



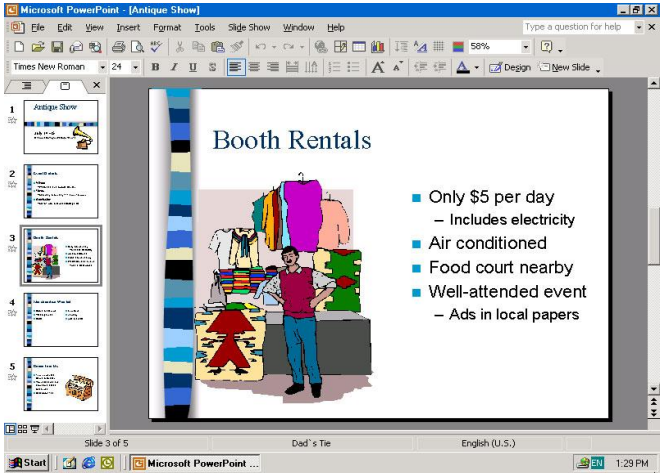
- Electronic spreadsheet software allows the user to add, subtract, and perform user-defined calculations on rows and columns of numbers. These numbers can be changed and the spreadsheet quickly recalculates the new results.

Database Software



- Allows the user to enter, retrieve, and update data in an organized and efficient manner, with flexible inquiry and reporting capabilities.

Presentation Graphics



- Presentation graphic software allows the user to create documents called slides to be used in making the presentations. Using special projection devices, the slides display as they appear on the computer screen.

Course Content

Theory

Introduction to Statistics and its Applications in Agriculture. Graphical Representation of Data, Measures of Central Tendency & Dispersion. Definition of Probability, Addition and Multiplication Theorem (without proof). Simple Problems Based on Probability. Normal Distribution. Definition of Correlation, Scatter Diagram. Karl Pearson's Coefficient of Correlation Linear Regression Equations. Introduction to Test of Significance, One sample & two sample test t for Means, Large sample test (Z test), Chi-Square Test of independence of Attributes in 2 ´2 Contingency Table. Introduction to Analysis of Variance, Principle of experimental designs, Analysis of One Way Classification (CRD and RBD). Introduction to Sampling Methods, Sampling versus Complete Enumeration, Simple Random Sampling with and without replacement, Use of Random Number Tables for selection of Simple Random Sample.

Practical

Graphical Representation of Data. Measures of Central Tendency (Ungrouped data) with Calculation of Quartiles, Deciles & Percentiles. Measures of Central Tendency (Grouped data) with Calculation of Quartiles, Deciles& Percentiles. Measures of Dispersion (Ungrouped Data). Measures of Dispersion (Grouped Data). Moments, Measures of Skewness & Kurtosis (Ungrouped Data).Moments, Measures of Skewness & Kurtosis (Grouped Data). Correlation & Regression Analysis. Application of One Sample t-test. Application of Two Sample Fisher's t-test. Chi-Square test of Goodness of Fit. Chi-Square test of Independence of Attributes for 2 ´2 contingency table. Analysis of Variance One Way Classification. Selection of random sample using Simple Random Sampling.

Text books:

1. A Hand book of Agricultural Statistics by S R S Chandel
2. A text book of Agricultural Statistics by Rangaswamy, R

Reference books

1. Fundamentals of Biostatistics by Irfan Ali Khan and Atiyakhanum
2. Statistics in Applied Science by B. K. Bhattacharya
3. Statistical Methods by for Agricultural Workers by V. G. Panse and P. V. sukhatme
4. Agriculture and Applied Statistics by P. K. Sahu
5. Statistical Procedure for Agricultural Research by K. A. Gomez and a. A. Gomez
6. Theory and Analysis of Sampling survey design by Daroga Singh and F. S. Chaudhary
7. Sampling techniques by Padam Singh
8. Sampling Theory of Survery with Applications by Sukhatme, P. V. , Sukhatme, B. V., Sand Ashok , C.

INTRODUCTION

Statistics, its meaning and definition

Statistics is a branch of applied mathematics and is concerned with observational data.

The word statistics is generally used in two different ways.

(1) When it is used in plural, it means the quantitative data affected to a marked extent by a multiplicity of causes. When we say `collect statistics' it means collect the numerical data which are to be analyzed and interpreted e.g.

(i) Wheat production affected by various causes.

(ii) Collect the data of height of the students of second semester.

(2) When it is used in singular, it means "the science of collecting, classifying and using the data for further statistical treatments". It involve the methods of analysis used in the analysis and interpretation of data and they are known as statistical methods.

Definitions

(1) R. A. Fisher : It is a study of population, variation and the methods for reduction of the data.

(2) A.L.Bowley : It is a science of calculations and averages.

(3) Boddington : It is a science of estimate and probability.

All these definitions are not satisfactory because they cover only a part of the subject.

In general : Statistics may be defined as the science and art of collection, organization, presentation, analysis and interpretation of numerical data. OR Statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting and analyzing the data as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis.

However, it is not used for all these purposes in all field e.g. in administrative and executive department statisticians are interested only in collecting and presentation of data. Such as crop yields, birth and death rates etc. On the other hand a researcher employ the methods which relate to design of experiments and analysis of experimental results.

Biometry : When the principles of statistics are applied on living thing or organisms, the science is called biometry.

Statistical methods : The methods by which statistical data are analyzed are called statistical methods.

Aims of studying statistics

(1) To study the population: - The study of population of any kind is on the basis of sample data.

(2) To understand the nature of variability e.g. Height of plants. The biological phenomena observed under one set of conditions are never duplicated exactly under another set of similar conditions. Therefore, repetition of experiment is necessary to account all the factors causing variation. In biological phenomena were variation is a rule rather than exception it is this function that has wide application.

(3) To express the facts in summary form (the facts that are based on large number of observations). e.g. It is not possible for one to form a precise idea about the income position of the population of India from the records of individuals. However, figure of per capita income can be easily studied from sample data.

(4) To provide correct method(s) for taking sample (sampling).

(5) To provide proper method for comparison of two or more things.

(6) It helps in prediction/ forecasting the yield of a particular crop for a particular year on the basis of the past records.

Limitations of Statistics

Statistics with its wide applications in almost every sphere of human activity, is not without limitations. The following are the important limitations.

1) It does not deal with individual.
2) It deals only with quantitative characters.
3) Statistical results are true only on an average.
4) Statistics can be misused.
5) It does not reveal the entire story.
6) Expert knowledge is must to handle the statistical data.

(1) It does not study individuals

Statistics deals with an aggregate of objects and does not give any specific recognition to the individual items of a series. e.g. the individual figures of agricultural production of any country for a particular year are meaningless unless, to facilitate comparison, similar figures of other countries or of the same country of different years are given. Height of Mr. X is 5'8" does not constitute statistical statement. The average height of an Indian is 5'8".

(2) It deals only with quantitative characters.

Efficiency, honesty, intelligence. These factors can be measure indirectly e.g. efficiency of selling agent can be judged by studying the no. of articles sold by him.

(3) Statistical results are true only on an average.

Average consumption of milk per head in a certain locality is 1/2 liter but it does not give any idea of the shortage of milk faced by the poor. The conclusions obtained statistically are not universally true, they are true only under certain conditions. This is because statistics as a science less exact as compared to natural sciences.

(4) Statistics can be misused.

Because if conclusions are based on incomplete information, statistics can prove anything. There are three types lies : lies, dammed lies and statistics. Statistics are like day of which one can make a God or Devil as he please.

Importance in agricultural research

(1) It helps to understand nature of variability or differences.

(2) To arrive at the meaningful conclusion on the basis of sample study in the field.

(3) Express the data/result of the field experiment in summary form.

(4) Sampling

a) In state Argil. survey for estimation of areas and yield of crops.

b) In price fixation policy of various argil. commodities.

c) In argil extension survey to study the impact of programs.

d) In argil. economics survey to study the demand-supply policy, the growth rate of population and cost of production of various crops.

(5) In argil. meteorology for weather forecasting and to correlate weather parameters with crop production.

Collection of data

Methods of collection data.

1) Measurement : e.g. Length of pod, Height of plant, Area of leaf, Volume of mango fruit etc.

2) Scale in physical science : e.g. Temperature, Humidity, Wind velocity etc.

3) Score or ranks : e.g. Intelligence test by judging candidates interview.

4) Personal contact : e.g. By asking questions to the individual.

5) Questionnaire : e.g. Data collected by mailing a form called questionnaire consisting several questions.

Attributes : Qualitative characteristics of an individual which shows variability is known as attributes.

Variable : The characteristics which show variation or variability are called variables or variates e.g. cabbage yield, wheat yield per hectares of the growers. Variable can of two types,

(i) Qualitative : The characteristics which can not be measured numerically or in terms of magnitude e.g. flower color, nature of surface.

(ii) Quantitative :The characteristics which can be measured in terms of magnitude e.g. yield of crop, height, weight. The quantitative characteristics is of two types.

(a) Discrete : Character which takes only integer values/or whole value. There is a definite gap between two values. e.g. No. of students in a class, No. of bacteria in given area.

(b) Continuous : The quantity which can take any numerical value within a certain range. Height, weight (They are in fraction and there is no definite gap between values).

2. FREQUENCY DISTRIBUTION AND FREQUENCY CURVES

Classification and tabulation of data: The process of reduction of data to a manageable size is called classification OR The process by which the data are arranged in groups or classes according to similarities is known as classification and the process by which the classified data are presented in an orderly manner by being placed in proper rows and columns of a table in order to bring out their essential features or characteristics is known as tabulation.

Objectives of classification:

1. To reduce data in groups/classes according to similarity.
2. To facilitate comparison through statistical analysis.
3. To point out most significant features of the data at a glance.
4. To give importance to a particular item by dropping out the unnecessary elements.
5. To enable a statistical treatment of the material collected.

Types of classification:

1. Geographical: When the classification is made on area basis e.g. district. taluka, city,
2. Chronological: When the classification is made on the basis of time e.g. production of wheat in past 10 years.
3. Qualitative: When classification is made on the basis of some attributes. This classification is further divided into four types.
    (A) Simple classification: Only one attribute is considered e.g. blindness or sex.
    (B) Two way classification: Two attributes are considered e.g. blindness & deafness, colour & shape of flowers.
    (C) Three way classification: Three attributes are considered e.g. sex, education level and residing location.
    (D) Manifold classification: More than three attributes are considered.

e.g.

	Classification		
	One way	Two way	Three way
	Sex	Sex & Marital status	Sex, Marital status & Education level
Population	Male	Married	High
			Medium
			Low
		Unmarried	High
			Medium
			Low
	Female	Married	High
			Medium
			Low
		Unmarried	High
			Medium
			Low

4. Quantitative: When the classification is made in the form of magnitude e.g. cows are classified according to milk yield. This classification is further divided into two types

(A) Discrete classification: Specific value in the range is considered e.g. no. of petal, no. of insects etc.

(B) Continuous classification: Any value in the range of variation is considered e.g. length, width etc.

FREQUENCY DISTRIBUTION

Objectives:

1. To condense the mass of data in such a manner that similarities and dissimilarities can be easily understand.
2. To enable statistical treatment to the data collected.

Frequency: The no. or individual of items occurring in each class is termed as frequency.

Frequency distribution: The manner in which the frequencies are distributed

over the different class is called frequency distribution of the character under study and the table indicating frequency distribution is called frequency table.

Class limit: It is the lowest and highest values of the distribution that can be included in the class e.g 10-20, 20-30 etc. Two boundaries of a class are known as the lower limit and upper limit of a class.

Class interval: The width of a class that is the difference of upper and lower limit of the class is known as class interval.

Class mid point: It is the value lying half way between the lower limit (LL) and upper limit (UL) of a class interval i.e. (LL + UL)/2.

Points while deciding class interval/classes:

1. It should be of uniform width which facilitates the statistical computation.
2. Range of the class should cover the data and should be continuous.
3. It should be convenient to make the mid-point of a class.
4. It should not be over lapping.

Types of frequency distribution:

(1) Discrete frequency distribution

(2) Continuous frequency distribution

Methods of classifying the data according to class interval:

Exclusive method: When the class intervals are so fixed that the upper limit of one class is the lower limit of the next class. This method is known as exclusive method e.g. -10, 10-20. Usually this method is preferred for continuous type of data. The data observed up to 9.99 would be included in 0-10 class while 10 or greater than 10 will be included in 10-20 class.

Inclusive method: In this method of classification, the upper limit of one class is included in that class itself e.g. 100-199, 200-299. The value of 100 and 199 will be included in the class of 100-199. This method is preferred for discrete type of data.

Procedure to form frequency distribution:

Step 1: Find range of the data. Range = Highest value – Lowest value.

Step 2: Fix the number of classes. Number of classes should preferably between 5 to 15 and should not be less than 5 and more than 30.

Approximate no. of classes = K =1 + 3.322 log N (Sturge's rule) where N = no. of observations under study.

Step 3: Fix the class interval = CI = Range/No. of classes or (L-S)/K where L = largest value and S = smallest value

Step 4: Arrange different classes in ascending order of magnitude

Step 5: Pick up the values of observation and make tally mark against respective classes.

Step 6: Find total tally mark of each class which will give the no. of frequencies in the respective classes.

GRAPHICAL REPRESENTATION

Graphical representation is used when we have to represent the data of a frequency distribution and of a time series. It is represented by points which are plotted on a graph paper.

Advantages of graphical representation

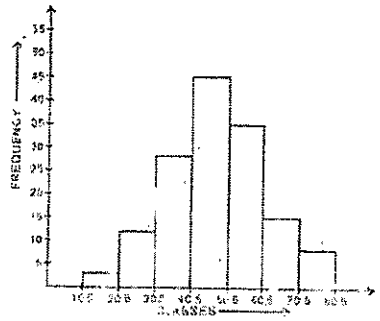
1. Easy to understand and interpret data at a glance.
2. It facilitates comparisons.
3. It gives eye view of complex data.
4. It has long lasting impression.
5. It gives an attractive and interesting view.

Limitations of graphical representation

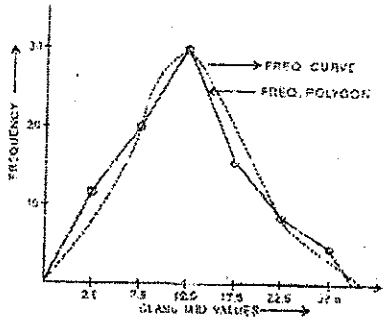
1. It cannot show all those facts which are there in the tables.
2. It shows tendency and fluctuations, actual values are not known.
3. The charts take more time to be drawn than the tables.

Graphs of Frequency Distribution

Histogram: It is a bar diagram which is suitable for frequency distributions with continuous classes. The width of all bars is equal to class interval and heights of the bars are in proportion to the frequencies of the respective classes. In this diagram bars touch each other but one bar never overlaps the other.

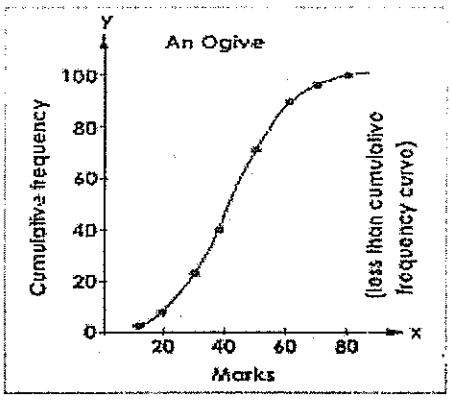


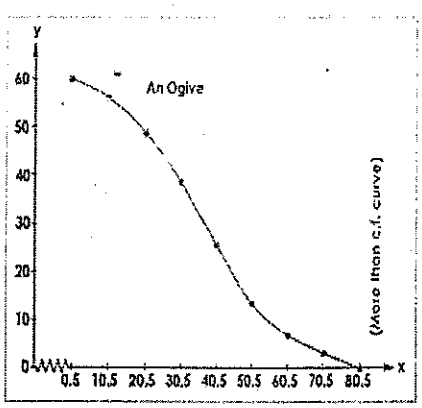
Frequency polygon: When the mid points of the tops of the adjacent bars of a histogram are joined in order by a straight line, then the graph of lines so obtained is called a frequency polygon.



Frequency curve: A frequency curve is a graphical representation of frequencies corresponding to their variates values by a smooth curve. A smoothened frequency polygon represents a frequency curve.

Ogive or Cumulative frequency curve: it is a graph plotted for the variates values and their corresponding cumulative frequencies and joined by a free hand smooth curve. The curve is '*S*' shaped. There are two methods of constructing ogive viz. (i) "less than" method and (ii) "more than" method. In the "less than" method, we start with the upper limit of classes and go on adding the frequencies however, in the "more than" methods, we start with the lower limit of classes. The first method gives a rising curve whereas second method shows a declining curve.





3. MEASURES OF CENTRAL TENDENCY

Different groups of data or statistical series or frequency distributions differ in four characteristics.

1) Central tendency or location
2) Dispersion or variation
3) Skewness or symmetry
4) Kurtosis or peaked ness

Central tendency: Generally it is found that in any distribution, values of the variable tend to cluster around a central value or centrally located observation of the distribution. This characteristic is known as central tendency.

This centrally located value which represents the group of values is termed as the measure of central tendency e.g. an average is called measure of central tendency.

Objectives

1) To get one single value that describes the characteristics of the entire series/group.
2) To compare two or more distributions.

Requisite/Characteristics of ideal measures of central tendency

Since an average is a single value representing a group of values, it is expected that such a value should satisfied the following properties.

1) It should be rigidly defined.
2) It should be based on all the observations.
3) It should be easy to understand or comprehensible, otherwise its use will be limited
4) It should be easy to calculate.
5) It should be amenable to further mathematical treatment.
6) It should be least affected by fluctuation of sampling.
7) It should be least affected by the extreme values.

Different measures of Central tendency

1) Arithmetic mean (A.M.) — Algebraic average
2) Median
3) Mode — } Positional average
4) Geometric mean (GM)
5) Harmonic mean (H.M.)
6) Weighted mean (W.M.) — } Algebraic average

(1) Arithmetic mean or Mean

It is the most common and ideal measure of central tendency. It is defined as the sum of the observed values of the character (or variable) divided by the number of observations considered in obtained sum (total).

Symbolically $\bar{X}$: Sample mean

μ : Population mean

Method of computation

Raw data or ungrouped data

(i) Direct method:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

(ii) Assumed mean method :

$$\bar{X} = A + \frac{\sum_{i=1}^{n} d_i}{n} \qquad d_i = X_i - A \quad (A = \text{Assumed mean})$$

Grouped data

(i) Direct method

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i = n}$$

where, f = freq. of k^{th} class, X = class mid value, k = no. of classes

$$\bar{X} = \frac{8}{4} = 2 \text{ i.e. } 4 - 2$$

If we multiply and divide by 2 then

$$\bar{X} = \frac{32}{4} = 8 \text{ i.e. } 4 \times 2$$

$$\bar{X} = \frac{8}{2} = 2 \text{ i.e. } 4/2$$

Merits and demerits of Arithmetic mean

Merits

1) It is rigidly defined.
2) It is based on all the observations.
3) It is readily comprehensible.
4) It is easy to calculate.
5) Its algebraic (Mathematical) treatment is especially easy and definitely possible.
6) It is also least affected by the fluctuation of sampling.

Demerits

1) It is affected by the extreme values.
2) If there is large variation in the data then A.M. becomes some times meaning less.
3) It is not used to measure rate of growth or rate of speed directly.

Uses: It is most popular and simple estimate and used widely in almost all the fields of studies such as social science, economics, business, agriculture, medical sciences, engineering and such other sciences.

Weighted Mean

When different observations are to be given different weights, arithmetic mean does not prove to be a good measure of central tendency. In such cases weighted mean is to be calculated.

If $X_1, X_2, X_3, \ldots X_n$ are different observation and $W_1, W_2, W_3, \ldots W_n$ are their respective weights then,

$$W.M. = \frac{W_1X_1 + W_2X_2 + \ldots\ldots + W_nX_n}{W_1 + W_2 + \ldots\ldots + W_n} = \frac{\sum W_iX_i}{\sum W_i}$$

(ii) Assumed mean method

$$\bar{X} = A + \frac{\sum f_i d_i}{n}$$ where $d_i = X_i - A$ and A = Assumed mean

(iii) Step deviation method

$$\bar{X} = A + \frac{\sum f_i d_{xi}}{n} \times I$$ where $d_{xi} = (X_i - A)/I$, A = Assumed mean,

I = Class interval, X = class mid value

Properties of Arithmetic mean

1) The algebraic sum of the deviations of a set of values (observed values) from their arithmetic mean is zero.
2) The sum of squares of the deviations of a set of values from their arithmetic mean is always minimum.
3) Amenability of arithmetic mean to further mathematical calculation.

   (a) Let $\bar{X}_1$ be the mean of n_1 observations, $\bar{X}_2$ be the mean of n_2 observations, X_k be the mean of n_k observations then the combined mean of N observations is given by (where $N = n_1 + n_2 + \ldots, + n_k$)

$$\bar{X} = \frac{n_1 X_1 + n_2 X_2 + \ldots\ldots + n_k X_k}{n_1 + n_2 + \ldots\ldots + n_k}$$

   This is also called weighted mean (W.M.)

   (b) Adding or subtracting a constant from each observation of a given series will add or subtract the same constant to the arithmetic mean.

   (c) Multiplying or diving each observation by a constant will multiply or divide the arithmetic mean by the same constant.

Examples:

Let the observations are 2, 3, 4 and 7.

$$\text{Here } \bar{X} = \frac{16}{4} = 4$$

If each value is added and subtracted by 2 then

$$\bar{X} = \frac{24}{4} = 6 \text{ i.e. } 4 + 2$$

Merits and demerits of Weighted mean (W. M.)

W.M. is the A.M., hence merits and demerits are the same as there for the arithmetic mean.

Uses

1) Used when the number of individuals in different classes of grouped widely varying.
2) Used when the importance of all the items in a series is not same.
3) Used when the ratios, percentages or rates e.g. rupees per kilogram, rupees per meter etc. are to be averaged.
4) Weighted mean is particularly used in calculating birth rates, death rates, index numbers, average yield etc.

Geometric Mean

AM gives equal weightage to all the items and has got a tendency towards the higher values. Sometimes it is necessary to get average having a tendency towards the lower values. In such case, geometric mean is helpful. It is defined as the n^{th} root of the product of n items of a series with following relation.

Raw data or ungrouped data

$$GM = \sqrt[n]{X_1, X_2 \ldots X_n} = (X_1, X_2 \ldots X_n)^{1/n} = \frac{\sum_{i=1}^{n} \log X_i}{n}$$

Grouped data

$$GM = \sqrt[n]{X_1^{f_1}, X_2^{f_2} \ldots X_k^{f_k}} = (X_1^{f_1}, X_2^{f_2} \ldots X_k^{f_k})^{1/n}$$

$$= \frac{1}{n}(f_1 \log X_1 + f_2 \log X_2 + \ldots f_k \log X_k) = \frac{1}{n}\sum_{1}^{k} f_i \log X_i$$

Merits and demerits of Geometric mean

Merits

1) It is rigidly defined.
2) It is based on all the observations.
3) It is not much affected by the fluctuation of sampling.
4) It gives less weightage to large items and more to small items.
5) It is suitable for averaging ratios, average rate of change, index nos.

Demerits

1) It is difficult to understand.

2) It cannot be calculated when there are negative values.

3) If any item of the series is zero, it would be also zero.

Harmonic Mean

HM is the reciprocal of the arithmetic mean of the reciprocal of the values of a variable or series.

Raw data or ungrouped data

$$HM = \frac{n}{\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots\ldots + \frac{1}{x_n}\right)} = \frac{n}{\sum \frac{1}{x_i}}$$

Grouped data

$$HM = \frac{\sum f_i}{\left(\frac{f_1}{x_1} + \frac{f_2}{x_2} + \ldots\ldots + \frac{f_k}{x_n}\right)} = \frac{\sum f_i}{\sum \frac{f_i}{x_i}}$$

Merits and demerits of Harmonic mean

Merits

1) It is rigidly defined.

2) It is based on all the observations.

3) It is not much affected by the fluctuation of sampling.

4) It gives greater weightage to smaller values.

5) It is useful in average price, speed, time, distance, quantity etc.

Demerits

1) It is not easy to calculate and understand.

2) It cannot be calculated if any value is zero or negative.

3) It gives large weightage to smaller values.

Uses: Time series data, units purchased per rupee, kilometers covered per hour, problems solved per time.

Relation between AM, GM and HM: AM > GM > HM

Median: The median is the middle most items that divide the series into two equal parts when i^{th} items are arranged in ascending or descending order.

In case of raw data, the median is the $\left(\frac{n+1}{2}\right)^{th}$ term where n is the total no. of observations whereas, in case of grouped data it is given by the formula,

$$\text{Median} = l + \left[\frac{(n/2) - cf}{f} \times CI \right]$$

where, /= lower limit of the median class in which $(n/2)^{th}$ item falls
cf= cumulative frequency of the class preceding the median class
f = frequency of median class
CI= Class interval

Uses: It is useful when the extreme values of the series are either not available or impossible to be obtained or abnormal. When in a group, the individual is denoted by better than half the individual's, median is used. It is also useful when the items are not susceptible to measurement in definite units e.g. intelligence, ability, efficiency etc.

Mode: The value of the variable which occurs most frequently or whose frequency is maximum is known as mode. For grouped data, it is given by

$$Mode = l + \frac{f_1 \text{ or } \Delta_1}{f_1 \text{ or } \Delta_1 + f_2 \text{ or } \Delta_2} \times CI$$

where /= lower limit of modal class

f_1 = difference between frequency of modal class and preceding class

f_2 = difference between frequency of modal class and succeeding class

CI=class interval

Uses: Business forecasting is particularly based on the modal values. Meteorological forecasting is also based on modal values.

Relation between Mean, Median and Mode:

Mode = 3Median – 2Mean

Quartiles:

Quartiles are those three values that divide the total data (already arranged) into four equal parts. Those three values, say Q1, Q2 and Q3 are called the first, second and third quartiles respectively. Q1 is such a value that 25% of the observations are smaller than or equal to and 75% of the observations are large than Q1. Second Quartiles Q2 is the value such that 50% of the observations are less than or equal to and 50% are greater than Q2. Similarly Q3 is the value such that 75% of the observations are less than or equal to and 25% of the observations are greater than it. It has been assumed

in this interpretation that the series is arranged in ascending order of magnitude of the variable.

First, second and third Quartiles are also known as ¼ Fractile, ½ Fractile and ¾ Fractile respectively.

Deciles :

Deciles are those nine values (say $D_1, D_2, \ldots D_9$) that divide the total data (arranged in ascending or descending order) into 10 equal parts.

$D_1, D_2, \ldots D_9$ are also known as $\frac{1}{10}$ Fractile, $\frac{2}{10}$ Fractile , ... $\frac{9}{10}$ Fractile respectively.

Percentiles :

Percentiles are those 99 values that divide the total arranged data in into 100 equal parts. These may be denoted as $P_1, P_2, \ldots P_{99}$

$P_1, P_2, \ldots P_{99}$ are also known as $\frac{1}{100}$ Fractile, $\frac{2}{100}$ Fractile , ... $\frac{99}{100}$ Fractile respectively.

First Percentiles P_1 is a point in a frequency distribution below which one percent of the total measures of score lie.

Tenth Percentiles P_{10} is a point in a frequency distribution below which ten percent of the total measures of score lie.

Similarly P_{99} or median is a point in a frequency distribution below which 50 percent of the total measures of score lie.

Difference between an Average and a Partition Value.

An average is the representative of the whole series while a partition value is the average of a part of the series. For example, the first quartile is the average (and hence the representative) of the first half of the series, while the third quartile is the average of the second half of the series. First decile is the average of the first , $\frac{2}{10}$ or , $\frac{1}{5}^{th}$ part of the series, second decile is the average of the first , $\frac{4}{10}$ *i.e.* $\frac{2}{5}^{th}$ part of the series and so on..

Calculation of partition Values for Individual

as well as Discrete Series

After the data have arrange in ascending order, we can calculate values of partition values by using the following formulae. let N be the number of observation (For discrete series $N=\sum f$)

1. p^{th} quartile Qp = value of $p\left[\frac{N+1}{4}\right]$th item or term. ($p$ = 1,2,3)

   In a particular first quartile Q_1 = value of $1\left[\frac{N+1}{4}\right]$th item

   second quartile Q_2 = value of $2\left[\frac{N+1}{4}\right]$th item

   third quartile Q_3 = value of $3\left[\frac{N+1}{4}\right]$th item

2. p^{th} decile Dp = value of $p\left[\frac{N+1}{10}\right]$th item or term. ($p$ = 1,2,3,.....9)

   In a particular first decile D_1 = value of $1\left[\frac{N+1}{10}\right]$th item

   second decile D_2 = value of $2\left[\frac{N+1}{10}\right]$th item

   .....

   .....

   ninth decile D_9 = value of $9\left[\frac{N+1}{10}\right]$th item

3. p^{th} percentile = value of $p\left[\frac{N+1}{100}\right]$th item or term. ($p$ = 1,2,3,.....99)

   In a particular first percentile P_1 = value of $1\left[\frac{N+1}{100}\right]$th item

   second percentile P_2 = value of $2\left[\frac{N+1}{100}\right]$th item

   .....

   ....

   99^{th} percentile P_{99} = value of $99\left[\frac{N+1}{100}\right]$th item

MEASURES OF DISPERSION

Mean is though an important concept in Statistics it does not give a clear picture as to how the different observations are distributed in a given distribution or the series under study. Consider the following series

Series	Observations	Mean
1	2,3,4,7	4
2	4,4,4,4	4
3	1,1,2,12	4
4	3,4,4,5	4

In the above series, the mean is same i.e. 4 but the spread of the observations about the mean is in different manner. Hence after locating the measures of central tendency, the next point is to find out the center. This can be done by measuring the spread. The spread is also called a Scatter, Variation OR dispersion of the variate values.

Definition

Dispersion may be defined as the extend of the scatterness of observations around a measure of central tendency and a measure of such scatter is called measures of dispersion.

Different measures of dispersion

1) Range.
2) Absolute mean deviation or Absolute deviation (A.M. D.)
3) Standard deviation (S)
4) Variance (S^2)
5) Standard error of mean (Sem.)
6) Coefficient of variation (C.V.%)

Requisite/Characteristics of an ideal measures of dispersion

Measures of dispersion should possess all those characteristics which are considered essential for measures of central tendency viz.

1) It should be based on all observations.
2) It should be readily comprehensible.
3) It should be fairly easily calculated.
4) It should be simple to understand.
5) It should not be affected by sampling fluctuations.
6) It should be amenable to algebraic treatment.

Standard deviation (S)

The standard deviation or "root of mean square deviation" is the most common and efficient estimator used in statistics. It is based on deviation from arithmetic mean and is denoted by S or σ. S = Std. deviation for sample. σ = Std. deviation for population

Definition

"It is a square root of a ratio of sum of square of deviation calculated from arithmetic mean to the total number of observations minus one."

Method of computation

Raw data or ungrouped data

(1) Deviation method

$$S = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n - 1}}$$

(2) Variable square method

$$S = \sqrt{\frac{\Sigma X_i^2 - (\Sigma X_i)^2/n}{n - 1}} = \sqrt{\frac{SS}{df}}$$

Where : X_i = Variate value SS = Sum of square

n = No. of observations. df = Degrees of freedom

(3) Assumed mean method

$$S = \sqrt{\frac{\Sigma d_i^2 - (\Sigma d_i)^2/n}{n - 1}}$$

$d_i = X_i - A$

A = Assumed mean

Grouped data or frequency distribution

(1) Deviation method

$$S = \sqrt{\frac{\Sigma f_i (X - \bar{X})^2}{n - 1}}$$

(2) Variable square method

$$S = \sqrt{\frac{\Sigma f_i X_i^2 - (\Sigma f_i X_i)^2/n}{n - 1}}$$

(3) Assumed mean method

$$S = \sqrt{\frac{\sum fid_i^2 - (\sum fid_i)^2/n}{n-1}}$$

$d_i = (X_i - A)$

A = Assumed mean

f_i= Frequency of i^{th} class

(4) Step deviation method

$$S = \sqrt{\frac{\sum fidx_i^2 - (\sum fidx_i)^2/n}{n-1}} \times I$$

$dx_i = (X_i - A)/I$

Properties of Standard deviation.

(1) Combined standard deviation

Combined standard deviation can be calculated using following formula when two series are given under study. It is symbolically denoted by S_{12}.

$$S_{12} = \frac{N_1S_{12} + N_2S_{22} + N_1d_{12} + N_2d_{22}}{N_1 + N_2}$$

Where, S_{12} = Combined standard deviation ; S_1 = Standard deviation of first group ; S_2 = Standard deviation of second group ; $d_1 = (X_1 - X_{12})$; $d_2 = (X_2 - X_{12})$; N_1 and N_2 are the numbers of observation for series one and two ; X_1 = A.M. for first series ; X_2 = A.M. for second series ; X_{12} = Weighted or combined mean.

(2) The sum of squares of the deviations of items in the series from their arithmetic mean is minimum. This is the reason why standard deviation is always computed from the arithmetic mean.

(3) Addition or subtraction of a constant from the grouped of an observation will not change the value of S.D.

(4) Multiplying or dividing each observation of a given series by a constant value will multiply or divide the std. deviation by the same constant.

Variance

Variance is the square of standard deviation. It is also called the "Mean square deviation" Its being used very extensively in analysis of variance of results from field experiment. Symbolically denoted by

S^2 = Sample variance and σ^2 = Population variance.

Method of computation

Raw data or Ungrouped data

(1) Deviation method

$$S^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X})^2}{(n_1 - 1)}$$

(2) Variable square method

$$S^2 = \frac{\Sigma X_i^2 - (\Sigma X_i)^2 / n}{n - 1} = \frac{S\,S}{df}$$

Where : X_i = Variate value S S = Sum of square
n = No. of observations. df = Degrees of freedom

(3) Assumed mean method

$$S^2 = \frac{\Sigma d_i^2 - (\Sigma d_i)^2/n}{n - 1}$$

$d_i = X_i - A$
A = Assumed mean

Grouped data or frequency distribution

(1) Deviation method

$$S^2 = \frac{\sum_{i=1}^{k} f_i(X_i - \bar{X})^2}{(n-1)}$$

(2) Variable square method

$$S^2 = \frac{\Sigma f_i X_i^2 - (\Sigma f_i X_i)^2/n}{n - 1}$$

(3) Assumed mean method :

$$S^2 = \frac{\Sigma f_i d_i^2 - (\Sigma f_i d_i)^2/n}{n - 1}$$

$d_i = (X_i - A)$
A = Assumed mean
f_i = Frequency of i^{th} class

(4) Step deviation method :

$$S^2 = \frac{\Sigma f_i dx_i^2 - (\Sigma f_i dx_i)^2/n}{n - 1} \times I^2 \qquad dx_i = (X_i - A)/I$$

Properties of variance

1) If $V_{(x)}$ represent the variance of X series and $V_{(y)}$ represent the variance of Y series then $V_{(x+y)} = V_{(x)} + V_{(y)}$

   i.e. $V_{(x+y)} = V_{(x)} + V_{(y)}$ and $V_{(x-y)} = V_{(x)} + V_{(y)}$

2) Multiplying or dividing each observation by a constant will multiply or divide the variance by square of that constant.

   e.g. $V_{(ax)} = a^2 V_{(x)}$.

3) Addition or subtraction a constant from the groups of each observation will not change the value of variance.

Standard error of mean (Sem.)

The standard deviation is the standard error of a single variate where as standard error of mean is the standard deviation of sampling distribution of the sample mean OR it refers to the average magnitude of difference between the sample estimate and population parameter taken over all possible samples from the population.

Definition : It is defined as square root of the ratio of the variance to the total no. of observations in a given set of data.

Symbolically it is written as S_x for sample and σ_x for population.

$S_{\bar{x}} = \frac{S}{\sqrt{n}}$ Where, S = Standard deviation ; n = No. of observation

For statistical analysis work the use of S_x is common. It is also used to provide confidence limit on population mean and for test of significance.

Coefficient of variation (C.V.%)

It is a relative measure of variation and widely used to compare two or more statistical series.

The statistical series may differ from one-another with respect to their mean or standard deviation or both. Some times they may also differ with respect to their units and then their comparison is not possible. To have a comparable idea about the variability present in them, C.V. % is used. It was developed by Karl Pearson".

Definition : "It is a percentage ratio of standard deviation to the arithmetic mean of a given series". It is without unit or unit less.

$$C.V.\ \% = (S / \bar{X}) \times 100$$

The series for which the C.V.% is greater is said to be more variable or we say less consistence, less homogeneous or less stable while the series having lower C.V. % is called more consistence or more homogeneous.

5. PROBABILITY

Statistics concerns itself with inductive reasoning/inference based on the mathematics of probability.

Sampling variation needs support in terms of probability / reliability of inference.

Introduction:

In the study of a population, one can not make any firm statements about the population concerned or its parameters, when only a sample investigation of the population is available for scrutiny. Due to the sampling variation there is doubt about sample investigation and hence it is practice to make statements of less definite nature in terms of probability or chance. The probability or chance for any statement depends on the number of favourable, unfavourable and total possible cases. e.g. in a tossing a coin for getting a head, one has to consider that there are two equally likely cases, head and tail, one is in favour of the statement and the other is against it.

The theory of probability aims to generalize the laws of chance, to discover the regularities in the pattern in which events, depending on chance, repeat themselves. It may be the tossing of a coin, a game of cards or the genetical ratios which may be the object of our investigation.

Jacob Bernoullis, an Italian mathematician was the first to give concept and definition of probability in 1713. The work of Gregor Mendel in Genetics showed that the theory of probability could be applied to biological investigations.

Definition:

Probability is a ratio of the number of "favourable" cases to the total number of equally likely cases. If probability is denoted by P then

$$P = \frac{\text{Number of favourable cases}}{\text{Total number of equally likely cases}}$$

Suppose a coin is tossed, the possible outcomes (events) are head and tail. These are equally likely and mutually exclusive events. The probability (P) of event head is 1/2.

$$\text{i. e.} \quad P(\text{Head}) = \frac{\text{Number of favourable event(s)}}{\text{Total no. of events}} = \frac{1}{2}$$

If n is the number of equally likely and mutually exclusive events for an event A, of which m is the favorable to its occurrence, then the probability of A is the fraction m/n."

$$P(A) = m/n$$

This is "***A PRIORI*** or "***CLASSICAL PROBABILITY***" determined before trials are made actually. The probability is arrived at by examination of the nature of the event rather than from the results of experiment. Here frequencies of the events are known and exact.

The estimation of "a priori" probability is logical. It may mislead, sometimes it fails to answer questions like; what is the probability that male die before the age of 60? What is the probability of rainfall on 15th August? One may reply as 1/2, which is not correct. One has to study here favorable events, the frequency of occurrence of rainfall on 15th August through past records. What is the probability of a student passing in an examination? Here we have two equally likely cases, passing or failing. The probability of passing the examination would be 1/2, then we might be ignoring the facts, the student might be a first class student who has studied well. In such cases the probability of passing the examination would be nearly 1 and not 1/2.

Thus, in many agricultural problems, it may not be possible logically to define equally likely events, which may happen, before trials are made. In such situation, the probability is estimated from a set of observations, which is called "***A POSTERIORI***" or "***EMPIRICAL PROBABILITY***". Such estimate is based on large number of observations.

- "a posteriori" probability is determined after the trial made i.e. the event has already occurred.

Posteriori probability: In this case, the probability is determined after the event has already occurred. The post-facto analysis of the event occurred just to understand the probability and its application to the problem.

Random Experiment: A happening with two or more outcomes is called an

experiment. If the outcomes are associated with uncertainties, the experiment is called random.

Trial and event: An experiment which repeated under essentially identical conditions, possible out comes is called events and experiment is known as trial. (Any out come or results of an experiment is termed as event) e.g. throwing a die is a trial and getting 1 or 2 ... is an event.

Simple event: The occurrence of a single event is known as simple event.

Compound events: The occurrence of two or more in connection with each other, the joint occurrence is called the compound events.

Exhaustive events: All possible out comes of any trial / experiment are known as exhaustive events. e.g: (i) tossing a coin there are two exhaustive events viz. head and tail. (Possibility of the coin standing on an edge being ignored) (ii) throwing of a die, there are six exhaustive events.

Mutually exclusive events: Events are said to be mutually exclusive if happening of any one of them precludes the happening of other OR two events are said to be mutually exclusive when both can not happen simultaneously in a single trial.

Independent events: Events are said to be independent if occurrence of any event is not affected by the occurrence the remaining events e.g. in tossing an unbiased coin event of getting head in the first toss is independent of getting a head in second, third and subsequent throws.

Dependent events: Dependent events are those in which the occurrence or non-occurrence of one event in anyone trial affects the probability of other events in other trial.

Equally likely events: Events are said to be equally likely when one does not occur more often than the others e.g. in throwing a die, all the six faces are equally likely to come.

LAWS OF PROBABILITY

(i) **Law of complementary event** $(\bar{A})$:

Whenever an event A fails to occur, we may say that event "not A" has occurred. The event "not A" is called complement of A and is denoted by $\bar{A}$,

For example,

If "head" is the event A, then "not head" is the complement of A i.e. "not head" means tail. So tail is the complement of the event A, head.

If P(head) = 1/2

then P(not head) = 1/2 i.e. P(tail)

= 1 - P(head)

= 1 - 1/2 = 1/2

Rule: If P is the probability of event A then probability of its complement $\bar{A}$ is

1- P or $P(\bar{A}) = 1 - P(A)$

Proof: Let m are favorable cases to A among n cases. Then remaining $n - m$ cases are favorable to "not A" the complement of A.

P(A) = m/n and P(not A) = (n-m)/n

i.e. $P(\bar{A}) = (n-m)/n = 1 - (m/n) = 1 - P(A)$

Example: A flower is taken out at random from a bag containing 6 red flowers 4 white flowers and 5 yellow flowers. Determine the probability that it is (a) red flower (b) not red flower.

Let R denote red flower. $\bar{R}$ denote "not red" flower, the complement of R.

$P(R) = 6/15$ (because total flowers are 15 and among 15 flowers, 6 are of red colored flowers)

$P(\bar{R}) = (15 - 6)/15$ (because among 15, there are 9 flowers, their colour is not red)

= 1 - 6/15 = 9/15

(ii) Law of addition:

Rule (ii) A: When events are mutually exclusive

If two events A and B are mutually exclusive with probabilities $P1$ and $P2$ respectively, then the probability of occurrence of either of them (A or B) is equal to the sum of the individual probabilities (A and B).

In symbols P(A or B) = P (A) + P(B) = P1 + P2

Proof: If an event A can happen in $m1$ ways and B in $m2$ ways, then the

number of ways in which either event can happen is *m1 + m2*. If the number of possibilities is *n*, then by definition the probability of either the first or the second event happening is

$$P(A \text{ or } B) = \frac{m1 + m2}{n} = \frac{m1}{n} + \frac{m2}{n} = P(A) + P(B) = P1 + P2$$

$$\text{where } P(A) = \frac{m1}{n} = P1 \; ; \; P(B) = \frac{m2}{n} = P2$$

Similarly, in general, P(A or B or C) = P(A) + P(B) + P(C)

If K events are mutually exclusive with individual probabilities *F1, F2, ... ,Pk* then P(anyone among *K* mutually exclusive events) = *P1 + P2 + ...+ Pk.*

Example: If *A* is the event drawing an ace from a pack of cards and *B* is the event drawing a king, then P(ace = A) = 4/52 and P (king = B) = 4/52. The probability of drawing either an ace or a king in a single draw is

P(ace or king) = P(A or B) = P(ace) + P(king)
= P(A) + P(B)
= 4/52 + 4/52 = 8/52

Since both ace and king can not be drawn in a single draw and are thus mutually exclusive events.

From the above explanation, one can point out two facts. They are:

(i) The probability, *P1* of an event lies between zero and one.
(ii) The sum of the probabilities of mutually exclusive events is one.

Rule (ii) B: When events are not mutually exclusive

If *A* and *B* are not mutually exclusive events, then the probability of either of them is equal to the sum of their probabilities less the probability of their simultaneous occurrence.

Symbolically

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$

where *P(AB)* is the probability of joint occurrence of *A* and *B*

Example: This will serve the proof also.

If *A* is the event "drawing an ace" from a pack of cards, and *B* is the event "drawing a spade card"; then *A* and *B* are not mutually exclusive events, since the ace of spade can be drawn. Thus the probability of drawing either

ace or a spade or both is

P(ace or spade) = P(ace) + P(spade) – P(ace of spade)
= 4/52 + 13/52 - 1/52
= 16/52

Similarly we can generalize the rule for more than two events also.

i.e. P(A + B + C) = P(A) + P(B) + P(C) – P(AB) - P(AC) – P(BC) + P(ABC)

(iii) Law of multiplication:

Independent and dependent events: Events are said to be dependent or independent accordingly as the occurrence of one does or does not affect the occurrence of the others. Two events, drawing of a king and queen will be independent if the drawing of the card is replaced after the first draw but if the card after first draw is not replaced and another card is drawn for the second event, the probability of occurrence of the second event will depend on the probability of the occurrence of the first. Hence in the latter case the second event will be dependent on the first.

Rule (iii) A: When events are independent

If *A* and *B* are two independent events, with individual probabilities *P1* and *P2* respectively, then the probability of both happening at a time is the product of their respective probability (*P1.P2*)

i.e. P(AB) = P(A) . P(B)

∴ = Pl . P2

Proof: Let *n1* and *m1* be the possible and favorable numbers of cases for the event *A* and *n2* and *m2* for the event *B* then

P(A) = m1/nl and P(B) = m2/n2

Since two events are independent, we can associate *n2* possible cases for *B* with each of the *n1* possible cases for *A*, so that the total number of possible cases is *n1.n2*.

Similarly the total number of favorable cases for "*A* and *B*" is *m.1 m2*

Thus,

$$P(\text{A and B, both at a time}) = \frac{m1.m2}{n1.n2} = \frac{m1}{n1}.\frac{m2}{n2.} = P1.P2$$

i .e. P(A and B) = P(A).P(B)

Similarly, the probability of occurrence of several independent events is the product of their separate probabilities.

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

Example: One urn contains 6 white flowers and 10 red flowers; second urn contains 8 white flowers and 12 red flowers. One flower is taken out from each of the urn. What is the probability that the flowers drawn are white?

The probability of a white flower from the 1st urn is 6/16 and from 2nd urn is 8/20. Both events are independent & hence required probability is the product 6/16 x 8/20 = 3/20.

Rule (iii) B: When events are dependent

If two events *A* and *B* are dependent then the probability of both happening at a time is given as follows:

$P(AB) = P(A) \cdot P(B/A)$ this is the conditional probability

or $= P(B) \cdot P(A/B)$

Where *P(B/A)* means the probability of second event *B* dependent on the probability of first event *A*.

In above, if P(B/A) = P(B) then *A* and *B* are independent events.

Example: Suppose a box contains 3 white balls and 2 black balls. Let A be the event "first ball drawn is black" and B the event "second ball drawn is black", where the balls are not replaced after being drawn. Here A and B are dependent events.

$P(A) = \frac{2}{3+2} = \frac{2}{5}$ ⟶ Probability of drawing first black ball.

$P(B) = \frac{1}{3+1} = \frac{1}{4} = P(B/A)$ ⟶ the probability of second black ball given the first ball drawn is black

Then P(A.B) = P(both black) = 2/5 . 1/4

= 2/20 = 1/10

Similarly we can generalize the rule for more than one dependent event.

$$P(A.B.C) = P(A).P(B/A).P(C/AB)$$

PROBABILITY DISTRIBUTION

It is also called parent population distribution, theoretical distributions or theoretical frequency distribution. In previous chapter, the probability of the occurrence of a single event is obtained. In scientific research using statistical methodology, it is often required to obtain the probabilities of occurrence of all possible events. A table of the possible values (X_i) which a chance event may assume with a corresponding probability distribution for each value is called a probability distribution for the population. Following table gives the probability distribution of sum of two unbiased dice.

Table : Probability distribution of sum of two dice.

X_i	2	3	4	5	6	7	8	9	10	11	12
f_i	1	2	3	4	5	6	5	4	3	2	1
p_i	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

$$\sum_{i=1}^{k} p_i = 1 \qquad p_i = f(x) = f(x_i)$$

Instead of a table of values such as above, one can represent the outcomes (p_i) by proper mathematical function over a range of X_i. In this chapter we would like to describe three theoretical distribution

(i) Binomial distribution - James Bernoulli (1700)

(ii) Poisson distribution - S.D. Poisson (1857) and

(iii) Normal distribution – De moivre (1733).

The normal distribution is very important in the field of agriculture because this distribution is more befitting to data of various field of agriculture.

Normal distribution

The normal distribution, also called the normal probability distribution, is most useful theoretical distribution for continuous variables. The data of many biological phenomena follow normal distribution. The area under the curve represents the total number of observations. The distribution is represented mathematically by

$$f(X) = \frac{1}{\sqrt{2\pi}\, n}\, e^{-\frac{1}{2}(X-\mu)2 / \sigma 2} \qquad -\infty < X < \infty$$

Where σ = Standard deviation , μ = Mean, e = 2.71828

The quantities μ and σ are parameters of this distribution. The above equation takes following form under the assumption that

$\mu = 0$, $\sigma = 1$ and $(X - \mu) / \sigma = Z$

$$f(Z) = \{1/ (2\pi)^{1/2}\} \; e^{-z2/2}$$

This is standard form of the normal distribution. A variate Z is said to be normally distributed with mean zero and standard deviation unity. It is called normal deviate.

Properties of normal distribution

(1) It is a symmetrical, bell shaped single peaked curve. Its slope grows steeper and steeper as it progress towards the ends. It is asymptotic curve i.e. it approaches closer and closer to the base line but never coincide to the base line.

(2) The shape of the curve at the center towards the x-axis is concave while at end it is convex. The curve changes the shape at the distance of σ from the mean.

(3) There are two parameters viz., μ (mean) σ (standard deviation). The curve can be drawn if we are having the values of both the parameters of population.

When $\mu = 0$ and $\sigma = 1$, then the normal curve is termed as standard normal curve and the variate is called the standard normal variate.

(4) The normal curve is symmetrical about the mean therefore, mean devides the entire area of the curve into two equal parts and hence the mean is also the median. The maximum frequencies are also at the center of the curve and therefore, the mode is also equal to median. Thus mean, mode and median coincide at the center.

(5) If two ordinates at the distance of σ on both the sides of the mean are erected, the area of the curve so cut off is equal to 68.26 percent i.e. about 2/3 of the entire curve.

(6) Similarly if two perpendiculars are erected at the distance of 2σ on both the sides of the mean, the area between these two perpendiculars will be 95.44 percent.

(7) If two ordinates are erected on both the sides of mean at the distance of 3σ, the area so cut off will be 99.74 percent of the entire area of the curve.

(8) The range of the normal distribution is equal to 6 σ. Range = Maximum value - Minimum value = $(\mu + 3\sigma) - (\mu - 3\sigma) = 6\sigma$

(9) The absolute mean deviation about the mean = $0.7999\sigma = 4/5\sigma$

(10) The coefficient of skewness and kurtosis are 0 and 3 respectively.

(11) The quartile deviation = 0.6745 QD = 2/3 σ. This is called probable error of Z.

(12) The odd moment values for standard normal variate will be 0 while the values of the 2^{nd} & 4^{th} moments are 2 and 4 respectively.

Probability Integral

The probability is defined as the ratio of favorable cases to total number of equally likely cases. In the frequency curve, the proportion of area lying under the curve below a given value of the variate is called the probability integral. The proportion of the area lying under the curve beyond a given variate value is also called the probability integral.

Normal probability integral table

This is the table which gives the values of the probability integrals i.e. the proportion of area under the curve to the left of the ordinate at X_i. This is generally represented by $1/2\ (1 + \alpha)$. The variate X_i is the value of the normal deviate Z. This table gives the value of $1/2\ (1 + \alpha)$ for only positive value of normal deviate. The area to the right hand side of the ordinate which is $1/2\ (1 - \alpha)$ can be obtained by subtracting the area to the left hand side at X, from the unit value (one).

How to use this table ?

In crop competition for rice crop at district level 180 farmers participated. The average yield obtained was 51 qtl/ha with standard deviation 5 qtl/ha. Assuming the yield to be normally distributed determine.

1. How many farmers do you expect to obtain.
   (a) 60 qt. or more yield per ha. (b) 45 qt. or less yield per ha.
2. p (50 qt < X < 60 qt)
3. The point that has 99.8% of the are above it.

Solution

1 (a) The first step is to estimate the normal deviation corresponding to 60 qt.

$$Z = \frac{X - \mu}{\sigma} = \frac{60-51}{5} = 1.8$$

Find area under curve from Table 1 for Z = 1.8.

The value at Z = 1.8 is 0.9641 i.e. the area up to Z =1.8 is 0.9641 but we require area for the portion Z >= 1.8.

The are corresponding to >= 1.8 = 1 - Area up to (Z=1.8)
= 1 - 0.9641 = 0.0359
p (X >= 60 qt) = 0.0359

The number of farmers harvesting 60 or more qt./ha can be estimated as under

E = p x N Where, E = Expected numbers (freq.)
p = probability of X
N = Total number of observations

No. of farmers who harvested 60 qt or more yield = 0.0359 x 180
= 7 farmers

(b) Similarly, for X =< 45 qt./ha,

$$Z = \frac{X - \mu}{\sigma} = \frac{45-51}{5} = -1.2$$

The required area under curve is for Z = -1.2 for the area to the left of Z = -1.2. Due to symmetrical form, this can be obtained by estimating area right to the Z = 1.2.

The area Z >= 1.2

= p (Z >= 1.2)

= 1 - p (Z >= 1.2)

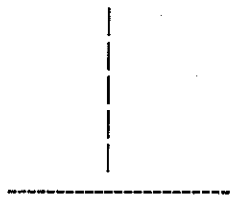
= 1 - 0.8849 = 0.1150

p (X =< 45) = 0.1150

Number of farmers harvesting 45 qt. or less = p x N = 0.1150 x 180 = 21

2. p (50 qt =< X =< 60 qt)

i.e. area between 50 and 60 qt.



$$Z = \frac{50-51}{5} = -0.2$$

$$Z = \frac{60-51}{5} = 1.8$$

The required area = Area up to 60 - Area up to 50

i.e. (Area up to Z = 1.8) - (Area up to Z= - 0.2)
i.e. (Area between Z = -0.2 and Z = 0) + (Area between Z = 0 and Z = 1.8)

Now (Area between Z = -0.2 and Z = 0) is equal to (Area between Z = 0 and Z = 0.2)

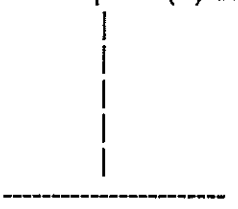
= 0.5793 - 0.5000 = 0.0793

Also p (0 =< Z =< 1.8) = 0.9641 - 0.5000 = 0.4641

p (0.2 =< Z =< 1.8)= 0.0.793- 0.4641 = 0.5434

Thus, p (50 qt =< X =< 60 qt) = 0.5434

3. The point (X) that has 99.8% of the area above it.



Here first find Z corresponding to the area 0.9980 from the table, which is equal to 2.88. The required point is for negative side i.e. for Z = -2.88. To locate X work out $Z = (X-\mu)/\sigma$

i.e. $-2.88 = \frac{X - 51}{5}$ giving X = 36.6 qt., the required point.

6. NORMAL DISTRIBUTION

The normal distribution, also called the normal probability distribution, is most useful theoretical distribution for continuous variables. The data of many biological phenomena follow normal distribution. The area under the curve represents the total number of observations. The distribution is represented mathematically by

$$f(X) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}} \qquad -\infty < X < \infty$$

Where σ = Standard deviation, μ = Mean, e = 2.71828

The quantities μ and σ are parameters of this distribution. The above equation takes following form under the assumption that

$$\mu = 0, \sigma = 1 \text{ and } (X - \mu) / \sigma = Z$$

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

This is standard form of the normal distribution. A variate Z is said to be normally distributed with mean zero and standard deviation unity. It is called normal deviate.

Properties of normal distribution

(1) It is a symmetrical, bell shaped single peaked curve. Its slope grows steeper and steeper as it progress towards the ends. It is asymptotic curve i.e. it approaches closer and closer to the base line but never coincide to the base line.

(2) The shape of the curve at the center towards the x-axis is concave while at end it is convex. The curve changes the shape at the distance of σ from the mean.

(3) There are two parameters viz., μ (mean) σ (standard deviation). The curve can be drawn if we are having the values of both the parameters of population.

When $\mu = 0$ and $\sigma = 1$, then the normal curve is termed as standard normal curve and the variate is called the standard normal variate.

(4) The normal curve is symmetrical about the mean therefore, mean divides the entire area of the curve into two equal parts and hence the mean is also the median. The maximum frequencies are also at the

center of the curve and therefore, the mode is also equal to median. Thus, mean, median and mode coincide at the center.

(5) If two ordinates at the distance of σ on both the sides of the mean are erected, the area of the curve so cut off is equal to 68.26 percent i.e. about 2/3 of the entire curve.

(6) Similarly if two perpendiculars are erected at the distance of 2σ on both the sides of the mean, the area between these two perpendiculars will be 95.44 percent.

(7) If two ordinates are erected on both the sides of mean at the distance of 3σ, the area so cut off will be 99.74 percent of the entire area of the curve.

(8) The range of the normal distribution is equal to 6σ. Range = Maximum value - Minimum value = $(\mu + 3\sigma) - (\mu - 3\sigma) = 6\sigma$

(9) The absolute mean deviation about the mean = $0.7999\ \sigma = 4/5\ \sigma$

(10) The coefficient of skewness and kurtosis are 0 and 3, respectively.

(11) The quartile deviation = 0.6745 QD = $2/3\sigma$. This is called probable error of Z.

(12) The odd moment values for standard normal variate will be 0 while the values of the 2^{nd} & 4^{th} moments are 2 and 4 respectively.

7. CORRELATION ANALYSIS

So far we have studied problems relating to one variable only. In practice we come across a large number of problems involving the use of two or more than two variables.

Univariate population

A population that is characterized by a single variable is termed as univariate population e.g. population of height of students, weight, yield etc.

Bivariate population

When two variables are simultaneously studied in a single population is termed as bivariate population e.g. the height and weight of the students, rainfall and yield, the amount of fertilizer used and the crop yield.

If two quantities vary in such a way that movement in one are accompanied by movements in the other, these quantities are said to be correlated e.g. price of commodities and amount demanded, increase in rainfall up to a point and production of crop. The degree of relationship between the variables under consideration is measured through the correlation analysis.

Correlation

It indicates the association between the two or more variables in a bivariate distribution or an analysis of the covariation of two or more variables is usually called correlation.

Types of correlation

Correlation is described or classified in several different ways. Three of the most important ways of classifying correlation are:

i) Positive or negative

ii) Simple, partial and multiple

iii) Linear and non-linear

Positive and negative correlation

Whether correlation is positive or negative would depend upon the direction of change of the variable. If both the variables are varying in the same direction i.e. if as one variable is increasing the other on an average is also increasing, correlation is said to be positive. If, on the other hand the variable is varying in opposite directions, i.e. as one variable is increasing the

other is decreasing or vice-versa, correlation is said to be negative.

Positive correlation

X: 10 12 15 18 20	X: 80 70 60 40 30
Y: 15 20 22 25 37	Y: 50 45 30 20 10

Negative correlation

X: 20 30 40 60 80	X: 100 90 60 40 30
Y: 40 30 22 15 10	Y: 10 20 30 40 50

Simple, Partial and Multiple correlation

When only two variables are studied it is a problem of simple correlation. When three or more variables are studied it is a problem of either multiple or partial correlation. In multiple correlation, three or more variables are studied simultaneously. In partial correlation, we recognize more than two variables, but consider only two variables to be influencing each other, the effect of other influencing variables being kept constant.

Linear and Non-linear (curvilinear) correlation

If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable than the correlation is said to be linear e.g.

X: 10, 20; 30, 40, 50
Y: 70,140, 210, 280, 350

Correlation would be called non-linear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable.

Methods of studying correlation

There are four major approaches of ascertaining whether two variables are correlated or not:

1. Scatter diagram method
2. Graphic method
3. Algebraic method: Karl Pearson's coefficient of correlation
4. Rank method

1. Scatter diagram method

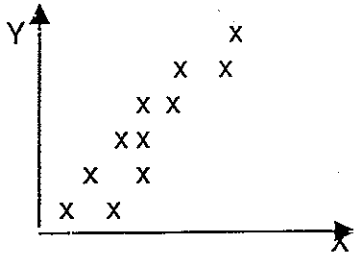
The simplest device for deciding whether two variables are related or not is to prepare a dot chart called scatter diagram. When this method is used

the given data are plotted on a graph paper in the form of dots i.e. for each pair of X and Y values we put a dot and thus obtain as many points as the number of observations. By looking to this scatter of the various points we can form an idea as to whether the variables are related or not. The greater the scatter of the plotted points on the chart, the lesser is the relationship between two variables. The more closely the points come to a straight line, the higher the degree of relationship.

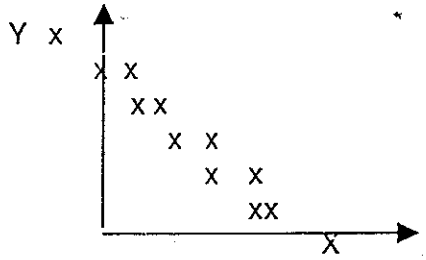
If all the points lie on a straight line falling from the lower left hand corner to the upper right hand corner, correlation is said to be perfectly positive e.g. volume and weight of metal. r = 1

If all the points are lying on a straight line rising from the upper left hand corner to the lower right hand corner of the diagram, correlation is said to be perfectly negative e.g. pressure and volume of gas. r = -1

If the plotted points fall in a narrow band there would be a high degree of correlation between the variables.

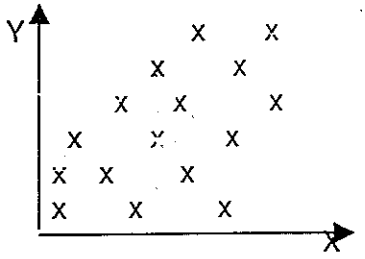


High degree +ve r
Rainfall and yield

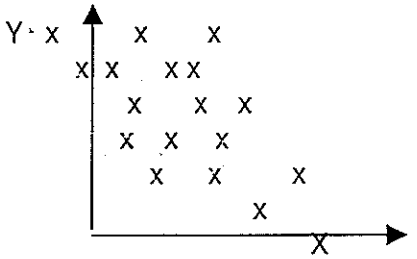


High degree -ve r
Intensity of diseases and yield

If the points are widely scattered over the diagram, it is the indication of very little relationship between the variables.



Imperfect +ve
(Low degree of positive correlation)



Imperfect -ve
(Low degree of negative correlation)

If the points lie on a straight line parallel to the X-axis or in a haphazard manner it shows absence of any relationship between the variables e.g. height of students and marks.

2. Graphic method

When this method is used the individual values of the two variables are plotted on the graph paper. We thus obtain two curves, one for X-variable and another for Y variable. By examining the direction and closeness of the two curves so drawn we can infer whether or not the variables are related. If both the curves drawn on the graph is moving in the same direction (either upper or downward) correlation is said to be positive. On the other hand, if the curves are moving in the opposite directions correlation is said to be negative.

3. Algebraic method (Karl Pearson Coefficient of correlation)

ρ (population) and its estimate as 'r' (sample) indicate Karl Pearson coefficient of correlation.

Definition: It is a measure of intensity of association between two variables in a bivariate population.

Computational formula:

$$\rho = \frac{Cov(XY)}{\sigma_x \sigma_y}$$

$$r = \frac{Cov(XY}{S_x S_y} = \frac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}} = \frac{SP(xy)}{SS_x . SS_y}$$

where, $\sum xy = \sum XY - \frac{(\sum X)(\sum Y)}{n}$

$$\sum x^2 = \sum X^2 - \frac{(\sum X)^2}{n}$$

$$\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

Properties of correlation coefficient:

1. A change in an origin does not affect the value of the correlation coefficient.
2. A change in a scale does not affect the value of correlation coefficient.
3. The value of correlation coefficient lies between -1 to +1.
4. Correlation coefficient is unit free.
5. Geometric mean of two-regression coefficient is equal to correlation coefficient.

Test of significance of correlation coefficient

Comparison of sample 'r' with population value

H_o: $\rho = 0$ (both the variables are not linearly associated)
H_a: $\rho \neq 0$

$t_{(n-2)}$ = r - ρ / SE of r SE of r = $\sqrt{\frac{1-r^2}{n-2}}$

$= \frac{r}{\sqrt{1-r^2}}\sqrt{(n-2)}$ under H_o : $\rho = 0$

If cal. t ≥ table $t_{0.05, (n-2)}$ d.f. H_o: rejected

Rejection of H_o: Means there is an association between two variables under study.

If cal. t < table $t_{0.05\,(n-2)}$ d.f. H_o: accepted

Acceptance of H_o: indicates that there is no association between two variables in the population.

'Z' transformation for test of significance of 'r':

If r' is to be compared with any hypothetical value, 't' distribution is not satisfactory for a test based on t = r-ρ/SE of r under H_o: $\rho = 0$

In order to overcome the difficulty mentioned above, R.A. Fisher has developed a transformation of 'r' to a statistic 'Z', which is approximately normally distributed. This transformation is useful in testing the significance of the difference between the estimated correlation coefficient r_1 and r_2 or the deviation of 'r' from a hypothetical value ρ. It is also helpful in testing homogeneity of several 'r's.

(i) H_o: $\rho = \rho_0$ (comparison of 'r' with hypothetical value of r)
H_a: $\rho \neq \rho_0$

$$Z = \frac{Z_o - Z_h}{1/\sqrt{n-3} = SE\ of\ Z} \qquad Z_o = \frac{1}{2}\log_e \frac{1+r}{1-r} \qquad Z_h = \frac{1}{2}\log_e \frac{1+\rho}{1-\rho}$$

(ii) Comparison of two correlation coefficient of two different populations

H_o: $\rho_1 = \rho_2$
H_a: $\rho_1 \neq \rho_2$

$$Z = \frac{Z_1 - Z_2}{\sqrt{1/(n_1-3) + 1/(n_2-3)} = SE\ of\ (Z_1 - Z_2)}$$

$$Z_1 = \frac{1}{2}\log_e \frac{1+r_1}{1-r_1} \qquad Z_2 = \frac{1}{2}\log_e \frac{1+r_2}{1-r_2}$$

Rank Correlation

The Karl Pearson's method is based on the assumption that the population being studied is normally distributed. When it is known that the population is not normal, or when the shape of the distribution is not known there is a need for a measure of correlation that involves no assumption about the parameters of the population.

This method was developed by Charles Spearman in 1904. This measure is especially useful when quantitative measures for certain factors can not be fixed e.g. (i) correlation between marks obtains in two different subjects by the same group of students. (ii) Correlation of height and weight of the students can be worked out without making exact measurement. We shall first stand the students according to height; the same procedure can be utilized for weight for giving ranks. When there are two or more items are of equal magnitude, their ranks are to be calculated by taking the average of their ranks.

$$R = 1 - \frac{6[\sum d_i^2 + (1/12)\sum (P^3 - p)]}{n(n^2-1)} \quad \text{or} \quad 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Where d_i^2 = square of difference of rank
n = number of pairs
P = number of items where ranks are common

8. REGRESSION ANALYSIS

The word regression was first used by Sir Fransis Galton and he introduces functional relationships between two variables. Many a times it is observed that change in one variable from a bivariate population causes change in the other variable, indicating a cause and effect relationship between the two variables. The former variable is termed as independent variable whereas, the later as dependent variable. Quantity of fertilizer and the crop will have this type of cause and effect relationship, where as quantity of fertilizer could be termed as independent variable and crop yield as dependent variable. The functional relationship between this independent and dependent variable is known as regression relationship.

Definition: Regression is a study of average relationship between two or more variables in terms of original units of the data.

Regression lines

In a scatter diagram if the points are scattered around a line than the relationship between two variables can be considered as linear. The resulting line is termed as regression line or line of best fit. For any pair of two variables that are related with each other linearly a set of two regression lines could be observed and they can be represented by two equations which are called regression equations. Let X and Y are the two variables. Then the two regression lines can be given by the following two equations.

$$Y - \bar{Y} = \beta_{yx} (X - \bar{X}) \quad \ldots\ldots \quad (i)$$

$$X - \bar{X} = \beta_{xy} (Y - \bar{Y}) \quad \ldots\ldots \quad (ii)$$

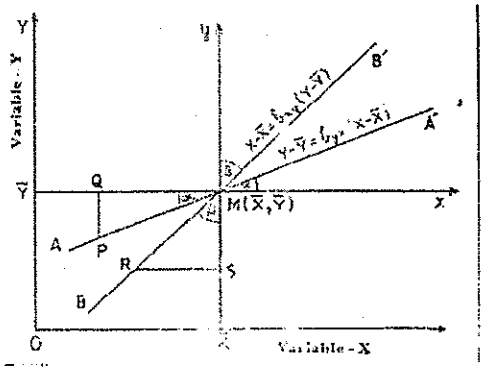
Where,

$\bar{X}$ = Mean of X variable; $\bar{Y}$ = Mean of Y variable

β_{yx} = Reg. coefficient of Y on X ; β_{xy} = Reg. coefficient of X on Y

We may observe that in first regression equation Y is considered as the dependent and X as independent where as in the second it is the reverse case.

These lines have been shown in the following diagram.



$$b_{yx} = \tan \alpha = \frac{PQ}{QM},$$

$$b_{xy} = \tan \beta = \frac{RS}{SM}$$

Fitting of the regression lines

A regression equation which represents a straight line is of the following form.

$$\hat{Y} = \alpha + \beta_{yx} X \quad Y =$$

Here Y is the dependent variable and X is the independent variable. β_{yx} is population regression coefficient of Y on X

$$\text{Intercept } \alpha = \bar{Y} - \beta_{yx} \bar{X}$$

In case of the sample data the estimates of βyx i.e. b_{yx} and the estimate of 'α' as a are obtained and placed in the equation.

$$a = \bar{Y} - b_{yx} \bar{X}$$

In a similar fashion the regression equation of the straight line where X is considered as dependent variable and Y as the independent variable the form of the equation would be

$$\hat{X} = \alpha' + \beta_{xy} Y$$

The estimate of β_{xy} is b_{xy} and α' is $a' = \bar{X} - b_{xy} \bar{Y}$

Regression coefficient

Regression coefficient can be defined as the average increase or decrease in the dependent variable for a unit change in the independent variable or it is the average rate of change in dependent variable with a unit change in independent variable. It is represented by β_{yx} and β_{xy} for the population regression coefficient. In practice they are estimated with the help

of the sample from the bivariate population under consideration and these estimates are generally represented as b_{yx} and b_{xy} respectively.

Method of computation

$$\beta_{yx} = \frac{\sum (X - \mu_x)(Y - \mu_y)}{\sum (X - \mu_x)^2} = \frac{Cov(XY)}{V(X)}$$

$$b_{yx} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{\sum xy}{\sum x^2} = \frac{\sum XY - (\sum X \sum Y)/n}{\sum X^2 - (\sum X)^2/n}$$

Similarly,

$$\beta_{xy} = \frac{\sum (X - \mu_x)(Y - \mu_y)}{\sum (Y - \mu_y)^2} = \frac{Cov(XY)}{V(Y)}$$

$$b_{yx} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (Y - \bar{Y})^2} = \frac{\sum xy}{\sum y^2} = \frac{\sum XY - (\sum X \sum Y)/n}{\sum Y^2 - (\sum Y)^2/n}$$

Test of significance of regression coefficient

1) When our interest is to ascertain whether the effect of the independent variable on the dependent variable is appreciable or not, we employ 't' test.

$H_o : \beta_{yx} = 0$

$H_a : \beta_{yx} = 0$

$$t = \frac{b_{yx}}{SE \text{ of } b_{yx}} \qquad SE \text{ of } b_{yx} = \sqrt{\frac{\sum y^2 - (\sum xy)^2 / \sum x^2}{(n-2)\sum x^2}}$$

Where, n = size of the sample

The calculated *t* value is to be compared with the table *t* value at the desired level of significance with (n-2) d.f. and conclusion is to be drawn.

2) By F test or ANOVA

Source	d.f.	S S	MS= SS/df	Cal.F
Regression	1	$(\Sigma xy)^2/ \Sigma x^2$	V_1	V_1/V_2
Error or Residual	(n-2)	$\Sigma y^2 - (\Sigma xy)^2/\Sigma x^2$	V_2	
Total	(n-1)	Σy^2		

If cal F < table $F_{0.05,(1,n-2)\,d.f.}$ $H_o: \beta_{yx} = 0$ is accepted
If cal F ≥ table $F_{0.05,(1,n-2)\,d.f.}$ $H_o: \beta_{yx} = 0$ is not accepted

Note: $F = t^2$

Properties of regression coefficient

1) Geometric mean between regression coefficients is correlation coefficient i.e. $r = \sqrt{b_{yx} \cdot b_{xy}}$
   a) Arithmetic mean of b_{yx} & b_{xy} is equal to or greater than correlation coefficient i.e. $\frac{b_{yx} + b_{xy}}{2} \geq r$
   b) If one regression coefficient is greater than unity than other regression coefficient must be less than unity.
2) Regression coefficient is independent of origin but not scale
3) Regression coefficient lies between - ∞ to + ∞
4) Regression coefficient has unit of dependant Valve
5) Regression coefficient has one way relationship

Uses of regression

1) To predict the value of Y for a given value of X with the help of regression equation.
2) To know the rate of change in Y for a unit change in X with the help of regression coefficient.

Relations among r, b_{yx}, b_{xy}, S_x and S_y

(i) $r = \sqrt{b_{yx} \cdot b_{xy}}$ (ii) $b_{yx} = r\frac{S_y}{S_x}$ (iii) $b_{xy} = r\frac{S_x}{S_y}$

Differences between Correlation and Regression

	Correlation	Regression
1	It deals with mutual association	It deals with cause and effect relationship
2	It is two way relationship	It is one way relationship
3	Correlation coefficient is unit free	Regression coefficient is in the units of dependent variable
4	Correlation coefficient lies between - 1 to + 1	Regression coefficient lies Letween - ∞ and + ∞
5	For a given value of one variable other variable can not be predicted	For a given value of independent variable the value of the dependent variable can be predicted.

BINOMIAL DISTRIBUTION

This is very important distribution dealing with discrete variable. The binomial distribution has two parameters viz. n and p. In other words, it is completely determined by the values of n and p.

If a coin is tossed once, there are two outcomes, namely, tail or head. The probability of obtaining a head or p=1/2 and probability of obtaining a tail or q=1/2. Thus (q+p)=1. These are terms of binomial (q+p). Similarly, if 2 coins are tossed simultaneously there are four possible outcomes

A	B
T	T
T	H
H	T
H	H

The probability corresponding to these results are

TT	TH	HT	HH
qq	qp	pq	pp
q^2	2pq		p^2

These are the terms of binomial $(q+p)^2$ because

$(q+p)^2 = q^2+2pq+p^2$ Similarly, for three coins

$(q+p)^3 = q^3 + 3q^2p + 3pq^2 + p^3$

In general, $(q+p)^n = q^n+nC_1\, q^{n-1}\, p + nC_2\, q^{n-2}\, p^2 + ... + p^n$

Since by expanding the binomial $(q+p)^n$ we obtained probability of 0,1,2,....n heads the probability distribution is naturally called binomial probability distribution. The general form of distribution is

$$\mathbf{P(r) = nC_r\, q^{n-r}\, p^r}$$

Where P(r) denotes the probability of getting exactly r successes.

Properties of Binomial distribution

(1) The shape of the distribution depends on the values of q and p. If p = q, the shape if it is symmetrical. If p = q the shape of it is, asymmetrical but the asymmetry decreases as n increases.

(2) Arithmetic mean = np

(3) Standard deviation = $\sqrt{npq}$

(4) Variance = npq

(5) Central moments value

(a) First moment $\mu_1 = 0$

(b) Second moment $\mu_2 = npq$

(c) Third moment $\mu_3 = npq\,(q-p)$

(d) Fourth moment $\mu_4 = 3n^2p^2q^2 + npq\,(1-6pq)$

(6) β-coefficients

(a) $\beta_1 = \dfrac{(q-p)^2}{npq}$ (b) $\beta_2 = 3 + \dfrac{1-6pq}{npq}$

Conditions for using Binomial distribution

(1) The outcome or results of each trial in the process are characterized as one of two types of possible outcomes.

(2) The possibility of outcome of any trial does not change and is independence of the results of previous trials.

Use

It is useful in describing an enormous variety of real life events.

POISSON DISTRIBUTION

The Poisson distribution is the limiting form of the binomial probability distributions n become infinitely large and p approaches 0 in such a way that np = m remained constant. Such situation are fairly common. That is to say, a Poisson distribution may be expected in cases were the chance of any individual event being a success is small. e.g. occurrence of comparatively rare event, such as serious floods, percentage infestation of any diseases etc.

Like binomial distribution, the variate of the Poisson distribution is also a discrete one. The probability functions is

$$P_{(x)} = \frac{e^{-m} m^x}{x!}$$

Where $P_{(x)}$ represents the number of successes

m represents the average number of successes ($m = np$)

e is a constant ($e = 2.7183$)

Properties of Poisson distribution

(1) Arithmetic mean = m

(2) Variance = m

(3) Standard deviation = $\sqrt{m}$

(4) Central moment value :

(1) First moment = $\mu_1 = 0$

(2) Second moment = $\mu_2 = m$

(3) Third moment = $\mu_3 = m^2$

(4) Fourth moment = $\mu_4 = m+3m^2$

(5) β Coefficients

$$(1)\ \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{m^2}{m^3} = \frac{1}{m}$$

$$(2)\ \beta_2 = \frac{\mu_4}{\mu_2} = \frac{m+3m^2}{m^2} = 3 + \frac{1}{m}$$

Use

Poisson distribution is used in practice in wide variety of problems where there are infrequently occurring events with respect to time, area, volume or similar unit. For example it is used in quality control statistics to count the number of defects of an item, or in biology to count number of bacteria, insects etc.

STATISTICAL INFERENCE AND TESTING OF HYPOTHESIS

Statistical inference is that branch of statistics, which is concerned with using probability concept to deal with uncertainty in decision making. It refers to the process of selecting and using a sample statistic or estimate to draw inferences about population parameters based on the sample drawn from the population.

The subject of statistics deals with statistical estimation and testing of statistical hypothesis. These are the two important functions for drawing inference about the population parameters. Statistical estimation is the technique of estimating the population parameter values on the basis of information obtained from the sample. Suppose we wish to know the yield of a crop. To know this figure, it is not necessary to harvest entire field of that crop or all the fields of that crop grown in the region. One may collect the sample from the fields by appropriate sampling procedure and on the basis of sample information, one may estimate the average yield of the crop of entire area. The estimate thus obtained is not the final form for drawing valid conclusion regarding population from which the samples are drawn. It needs to be tested by applying an appropriate test or method. Such test is known as the test of significance. Thus, test of significance can be defined as " The statistical procedure for deciding whether the observed difference between sample estimate and population parametric value is significant or not at specified level of significance".

Hypothesis It is the statement specifying the parametric value of a distribution from which the sample/s is/are drawn.

Null Hypothesis It is a hypothesis of no difference between different populations parametric values from which samples are drawn. OR It is the hypothesis of equality of population parametric values from which sample/s is/are drawn.

Procedure for testing a hypothesis

Step I : Set appropriate null hypothesis

Let us consider that there are two methods for preparing a compost.

Method A : standard method and

Method B : new method

Now to test which method is better, the hypothesis can be

1) B is better than A $B > A$
2) A is better than B $A > B$
3) B is not different from A $A = B$

The first two statements indicate a preferential attitude to one or the other of the two methods. Hence it is better to adopt the third statement and make the test. This third statement is called the null hypothesis, which is denoted as Ho: symbolically $H_o : \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$ where μ_1 and μ_2 are the population parametric values.

In the above examples suppose in first method the average nitrogen content is μ_1 and in the second method the average nitrogen is μ_2. Hence $\mu_1 = \mu_2$ can be tested by the appropriate test. As against the null hypothesis the alternative hypothesis should also be set up, which specifies those values, the researcher believes to hold true. Since one is going to accept or reject the null hypothesis one has to set the alternative hypothesis also. It is denoted by $H_a : \mu_1 \neq \mu_2$ or $H_a : \mu_1 < \mu_2$, $\mu_1 > \mu_2$.

Step II : Fix appropriate level of significance

The confidence with which an experimenter reject or accept the null hypothesis depends upon the significance level adopted. It is expressed in percentage such as 5 per cent, 1 per cent etc. When the hypothesis in question, is accepted at 5 per cent level of significance, the experimenter is running the risk that in the repeated cases (experiments/trials), he will be making the wrong decision in about 5 per cent of the cases. By rejecting the hypothesis at the same level, he runs the risk of rejecting a true hypothesis in 5 out of every 100 occasions. Thus, level of significance is defined as :

"It is the maximum probability at which one would like to reject the null hypothesis when it is true. OR The level of significance is the average proportion of incorrect statements made when the null hypothesis is true."

Step III : Set suitable test criterion

To construct a test criterion, one has to select the appropriate probability distribution for the particular test viz. Z, t, F, χ^2 etc.

Step IV : Computation

This step involves the calculations of various statistics from sample data such as mean and standard error of mean.

Step V : Conclusion

After doing the necessary calculations one has to decide whether to accept or reject the null hypothesis at a certain level of significance. Therefore, the computed value of the test criterion is compared with the table value. If the computed value is greater table value, the observed difference is significant and Ho is not accepted. If calculated value is less than or equal to table value the Ho is accepted at a given level of significance. Not acceptance of Ho means the difference between sample estimate and the hypothetical parametric value is a real difference, while acceptance of Ho means the difference between sample estimate and population/hypothetical parametric value can be explained due to chance variation (sampling variation).

Type - I and Type - II errors

While testing the hypothesis one is liable to commit two kinds of errors.

An error of first kind is made by rejecting the true null hypothesis. The probability of committing a type -I error is denoted by α (alpha)

Type II error is committed by accepting the null hypothesis when it is false. The probability of type II error is denoted by β (Beta).

Type I error depends on the level of significance. When 5 per cent level of significance is fixed, we fixed the probability of committing type I error at 5 per cent.

It is possible to control type I error by shifting the level of significance. Type II error increases as the type I error decreases. Therefore, the common practice is to keep the type I error at five percent or one percent fixed and try to decrease type II error by increasing sample size and following refined technique of conducting experiment.

Degrees of freedom

For testing any hypothesis the estimated statistic is compared with table value. The knowledge of degrees of freedom is essential for referring the table value. With $X_1, X_2,.....X_n$ having constant sum, (n - 1) X values can be given freely, but the n^{th} X value will be determined by the condition that the sum of all 'X' is equal to the given constant quantity i.e. one degree of freedom is lost. So in one way classification, number of observations - 1 is called degrees of freedom and in general number of observation minus number of independent constraints or restrictions is called degrees of freedom.

Large sample test : Z - test

It is a large sample test and can be utilized for testing the hypothesis if the following conditions are satisfied.

(1) Data follow normal distribution

(2) Sample size should be large (n > 30) or

(3) The standard deviation of population should be known if sample is not large.

Z test can be defined as " It is the ratio of the difference between the estimated population mean and hypothetical mean to the standard error of mean based on population standard deviation or its estimate from large sample.

One sample Z test

Objective : To test whether the given large random sample has come from the given population with mean μ and variance σ^2 or its estimate S^2.

Procedure :

Step I : Set the null hypothesis : $H_o : \mu = \mu_o$ or $H_o : \mu - \mu_o = 0$

$H_a : \mu \neq \mu_o$ or (two tailed)

$\mu < \mu_o$, $\mu > \mu_o$ (One tailed test)

Where, μ is the population mean from which the random sample has been drawn and μ_o is the mean of the hypothetical population.

Step II : Fix the level of significance

Usually 5 and 1 per cent levels of significance are fixed. If the probability of the observed difference between the sample estimate of μ and the hypothetical mean μ_o is less than 5 per cent, the difference will be considered as significant at 5 per cent level of significance and

if it is less than 1 per cent the difference will be considered highly significant at 1 per cent level of significance. The critical values will be 1.96 (2 tailed) and 1.65 (1 tailed) at 5 per cent level of significance and 2.576 (2 tailed) and 2.33 (1 tailed) at 1 per cent level of significance.

Step III Computation

If (i) $X_1, X_2, X_3, \ldots, X_n$ is the given sample ($n > 30$) or

(ii) class value : $X_1, X_2, \ldots, X_k$ with corresponding frequencies: $f_1, f_2, \ldots, f_k$ (Total = n) of given sample, calculate,

Sample mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i = n}$

(ungrouped data) **(grouped data)**

Variance : $S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)} = \frac{\sum_{i=1}^{k} f_i (X_i - \bar{X})^2}{(n-1)}$

(ungrouped data) **(grouped data)**

Standard error of mean :

$$S_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$ (If population standard deviation is known)

Step IV : Calculate normal deviate

$$Z = \frac{|\bar{X} - \mu_o|}{S_{\bar{X}}}$$

Step - V : Conclusion

If calculated $Z < 1.96$ the difference is non significant at 5% level of significant. H_o: $\mu = \mu_o$ is accepted. Acceptance of hypothesis revealed that the given sample has come from the population having mean μ_c.

If calculated $Z \geq 1.96$ the difference is significant at 5% level of significant. H_o : $\mu = \mu_o$ is rejected. If calculated $Z \geq 2.58$ the difference is highly significant at 1% level of significant. H_o: $\mu = \mu_o$ rejected Rejection of hypothesis indicates that the given sample does not come from the population having mean μ_o. The confidence of rejection being 95 per cent. If the difference is highly significant, the confidence of rejection is 99 per cent.

Two sample Z test

Objective: To test whether two randomly selected samples have come from the same population having mean μ and a standard deviation σ or its estimate S.

Generally there is little interest in comparing sample mean with population mean. A more frequent problem usually met with in agriculture is involved in comparison of 2 samples i.e. means of 2 samples. e.g.

i) We may require to compare variety A with variety B of a crop.

ii) Comparison of two manure for their effect on yield.

iii) Comparison of two rations for milk yield.

iv) Comparison of yield of the same varieties on two different farms etc. need to be tested for their difference in means (for location specificity).

Let, Sample I : $X_1, X_2, \ldots \ldots, X_{n1}$ and

Sample II : $Y_1, Y_2, \ldots \ldots, Y_{n2}$ are random samples drawn from normally distributed populations

OR

Let	Sample - I	Sample –II
Class value :	$X_1, X_2, \ldots, X_{k1}$	$Y_1, Y_2, \ldots, Y_{k2}$
Frequency :	$f_1, f_2, \ldots, f_{k1} = n_1$	$f_1, f_2, \ldots, f_{k2} = n_2$

are two frequency distributions of two samples drawn from a normal populations.

Procedure :

Step I : Set the null hypothesis that both the samples have come from the same population having mean μ and standard deviation σ or its estimate S.

i.e. $H_o : \mu_1 = \mu_2 = \mu$ against $H_a : \mu_1 \neq \mu_2$ or

$\mu_1 > \mu_2$ or $\mu_1 < \mu_2$

Where μ_1 is the population mean from which sample one is drawn and μ_2 is the population mean from which the second sample is drawn.

Step II : Fix the level of significance. Usually 5 per cent and 1 per cent levels of significance are fixed.

Step III : Calculate the following estimates.

Sample-I Sample-II

(i) Mean :

(Ungrouped data)

$$\bar{X} = \frac{\sum_{i=1}^{n_1} X_i}{n_1} \qquad \bar{Y} = \frac{\sum_{i=1}^{n_2} Y_i}{n_2}$$

(Grouped data)

$$\bar{X} = \frac{\sum_{i=1}^{k_1} f_i X_i}{n_1} \qquad \bar{Y} = \frac{\sum_{i=1}^{k_2} f_i Y_i}{n_2}$$

(Ungrouped data)

(ii) Variance :

$$S_x^2 = \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2}{(n_1 - 1)} \qquad S_y^2 = \frac{\sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{(n_2 - 1)}$$

(Grouped data)

$$= \frac{\sum_{i=1}^{k_1} f_i (X_i - \bar{X})^2}{(n_1 - 1)} \qquad = \frac{\sum_{i=1}^{k_2} f_i (Y_i - \bar{Y})^2}{(n_2 - 1)}$$

(iii) Pooled sample variance

$$S_p^2 = \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{n_1 + n_2 - 2} \quad \text{or} \quad S_p^2 = \frac{\sum_{i=1}^{n_1} f_i (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} f_i (Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

(iv) Standard error of mean of differences

$$S_{(\bar{X}-\bar{Y})} = \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Step IV : Calculate normal deviate (Z)

$$Z = \frac{(\bar{X} - \mu_1) - (\bar{Y} - \mu_2)}{S_{(\bar{X}-\bar{Y})}}$$

Step V : Conclusion

If calculated $Z \leq 1.96$ the observed difference is non significant at 5% level of significance and $H_o : \mu_1 = \mu_2$ accepted. Acceptance of $H_o : \mu_1 = \mu_2$ means both the samples have came from the same population

If calculated $Z > 1.96$ the observed difference is significant at 5% level of significant and $H_o : \mu_1 = \mu_2$ rejected at 5 percent level of

significance and if calculated $Z > 2.58$ the observed difference is highly significant 1% level of significance hence $H_o : \mu_1 = \mu_2$ rejected at 1% level of significance. Rejection of $H_o : \mu_1 = \mu_2$ means both the sample are drawn from two different populations.

Small sample or Student's 't' test

When the sample is large and if r is not known, we estimate the same and can be used in Z test. But if 'n' is small error will be more for replacing r by S and under that situation the Z remain no longer normal, but changes to another distribution named "t". The "t" distribution was found out by W.S. Gossett in the name of 'Student' in 1908.

Values of 't' depends on degree of freedom and is always greater than its limiting value of Z for any unit degree of freedom. When d.f. is large t --> Z. Difference between t and Z becomes more and more marked as n become smaller and smaller.

Definition : It is the ratio of the deviation between of sample mean and hypothetical mean to the standard error of mean estimated from the small sample.

Conditions for applying 't' test

1) Data follow normal distribution.
2) The sample is small ($n < 30$) and the standard deviation of the population is estimated from the sample.

Uses

1) Comparing sample mean with hypothetical mean or population mean.
2) Comparing two sample means.
   (a) When the number of observation of both the samples are unequal ($n_1 \neq n_2$)
   (b) When number of observations of both the samples are equal ($n_1 = n_2$)
   (c) When the observations are paired.
3) Comparing the regression coefficient of sample with the hypothetical or population regression coefficient.
4) Comparing the correlation coefficient with the correlation coefficient of population.
5) Comparing two regression coefficients.

Characteristics of "t" distribution

1) It is the exact distribution and not approximate.
2) t value ranges from $-\infty$ to $+\infty$
3) The distribution is symmetrical one.
4) It is flatter than the normal distribution i.e. the area near the tail is large for t distribution compared to normal distribution. Value of coefficient of kurtosis is less than 3.
5) As sample size increases the t distribution approaches to normal distribution.
6) There is need to known the d.f. to obtained the probability value from the table.

One Sample 't' test

Objective : To test whether the given small sample ($n < 30$) has come from the population having mean μ.

Procedure :

If (i) $X_1, X_2, X_3, \ldots, X_n$ is the given sample ($n < 30$) or

(ii) class value : $X_1, X_2, \ldots, X_k$ with corresponding frequencies : $f_1, f_2, \ldots, f_k$ ($\sum f_i = n$) of a given sample,

Step I : Set the null hypothesis : $H_o : \mu = \mu_o$ or $H_o : \mu - \mu_o = 0$

$H_a : \mu \neq \mu_o$ (two tailed test) or

$\mu < \mu_o$, $\mu > \mu_o$ (one tailed test)

Where, μ is the population mean from which the random sample has been drawn and μ_o is the mean of the hypothetical population.

Step II : Fix the level of significance .Usually 5 and 1 per cent levels of significance are fixed.

Step III : Calculate the following estimates.

If (i) $X_1, X_2, X_3, \ldots, X_n$ is the given sample ($n < 30$) or

(ii) class value : $X_1, X_2, \ldots, X_k$ with corresponding frequencies $f_1, f_2, \ldots, f_k$ ($\sum f_i = n$) of a given sample, calculate,

Sample mean $\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$ $= \dfrac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i = n}$

(ungrouped data) **(grouped data)**

Variance : $S^2 = \dfrac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)}$ $= \dfrac{\sum_{i=1}^{k} f_i (X_i - \bar{X})^2}{(n-1)}$

Standard error of mean :

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$ (If population standard deviation is known)

Step IV : Compute the student 't' with (n-1) degree of freedom

$$t = \frac{|\bar{X} - \mu_o|}{s_{\bar{x}}}$$

Step - V : Conclusion

If calculated t < table $t_{0.05,(n-1)\ d.f.}$ observed difference is not significant. H_o: $\mu = \mu_o$ is accepted. Acceptance of $H_o : \mu = \mu_o$ means the given small random sample has come from the hypothetical population having mean μ_o

If calculated t ≥ table $t_{0.05,(n-1)\ d.f.}$ observed difference is significant. $H_o : \mu = \mu_o$ is rejected. If calculated t ≥ table $t_{0.01\ ,\ (n-1)\ d.f.}$ observed difference is highly significant. H_o: $\mu = \mu_o$ is rejected. Rejection of $H_o : \mu = \mu_o$ means the given random sample does not come from the hypothetical population having mean μ_o.

Two sample 't' test (Independent sample)

Objective : To test whether the given two small random samples have come from the same population having mean μ_o.

Let sample - I : $X_1, X_2, \ldots, X_{n1}$

sample - II : $Y_1, Y_2, \ldots, Y_{n2}$ are two random samples drawn from a population.

Procedure:

Step I : Set the null hypothesis that both the samples have come from the same population having mean μ and standard deviation S.

i.e. $H_o : \mu_1 = \mu_2 = \mu$ against $H_a : \mu_1 \neq \mu_2$ or

$\mu_1 > \mu_2$ or $\mu_1 < \mu_2$

Where μ_1 is the population mean from which sample one is drawn and μ_2 is the population mean from which the second sample is drawn.

Step II : Fix the level of significance. Usually 5 per cent and 1 per cent levels of significance are fixed.

Step III : Calculate the following estimates.

Sample - I **Sample - II**

(Ungrouped data)

i) **Mean** $\bar{X} = \frac{\sum_{i=1}^{n_1} X_i}{n_1}$ $\bar{Y} = \frac{\sum_{i=1}^{n_2} Y_i}{n_2}$

(Grouped data)

$\bar{X} = \frac{\sum_{i=1}^{k_1} f_i X_i}{n_1}$ $\bar{Y} = \frac{\sum_{i=1}^{k_2} f_i Y_i}{n_2}$

(Ungrouped data)

ii) Variance

$s_x^2 = \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2}{(n_1 - 1)}$ $s_y^2 = \frac{\sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{(n_2 - 1)}$

(Grouped data)

$= \frac{\sum_{i=1}^{k_1} f_i (X_i - \bar{X})^2}{(n_1 - 1)}$ $= \frac{\sum_{i=1}^{k_2} f_i (Y_i - \bar{Y})^2}{(n_2 - 1)}$

(v) Pooled sample variance

$$s_p^2 = \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{n_1 + n_2 - 2} \quad \text{or} \quad s_p^2 = \frac{\sum_{i=1}^{n_1} f_i (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} f_i (Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

(vi) Standard error of mean of differences

$$s_{(\bar{X}-\bar{Y})} = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Step IV : Calculate student 't' with $n_1 + n_2 - 2$ d.f.

$$t = \frac{(\bar{X} - \mu_1) - (\bar{Y} - \mu_2)}{s_{(\bar{X}-\bar{Y})}}$$

Step V : If cal t < Table $t_{0.05,(n_1+n_2-2)}$ df. difference is non significant at 5% level of significance Ho : $\mu_1 = \mu_2$ accepted. Acceptance of $H_o : \mu_1 = \mu_2$ means both the samples have came from the same population μ

If cal t ≥ table $t_{0.05,(n1+n2-2)}$ d.f. difference is significant at 5 % level of significant $H_o : \mu_1 = \mu_2$ rejected at 5 % level of significance

If cal t ≥ table $t_{0.01,(n1+n2-2)}$ d.f. difference is highly significant 1% level of significance. $H_o : \mu_1 = \mu_2$ rejected at 1% level of significance. rejection of $H_o : \mu_1 = \mu_2$ means both the sample are drawn from two different populations.

Two sample 't' test (Dependent sample) : Paired 't' test

Objective : To test whether the two small related random samples have come from the same population.

Let Sample - I : $X_1, X_2, \ldots, X_n$ and

Sample - II : $Y_1, Y_2, \ldots, Y_n$ be two related sample such that $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are the pairs of related observations.

Procedure :

Step I : Set the null hypothesis : $H_o : \mu_d = 0$; $H_a : \mu_d \neq 0$

Where, μ_d is the average difference between $X_i - Y_i$ in the population.

Step II : Fix the level of significance. Usually 5 per cent and 1 per cent levels of significance are fixed.

Step III : Calculate the following estimates.

(i) $d_i = X_i - Y_i \qquad i = 1, \ldots, n$

(ii) $\bar{d} = \sum d_i / n$

(iii) $$\sum (d_i - \bar{d})^2 = \frac{\sum d_i^2 - (\sum d_i)^2/n}{n-1}$$

(iv) Se. of $$\bar{d} = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n(n-1)}}$$

Step IV : Calculate the student t with n-1 d.f.

$$t = \frac{|\bar{d} - \mu_d|}{S_{\bar{d}}} \qquad (\text{Under } H_o : \mu_d = 0)$$

Step V : Conclusion

If cal. $t <$ table $t_{0.05, (n-1)}$ d.f. Observed difference is non-significant at 5% level of significance and null hypothesis (H_o) is accepted. Acceptance of null hypothesis ($H_o : \mu_d = 0$) means the given two related small samples have come from the same population.

If cal. $t \geq$ table $t_{0.05, (n-1)}$ d.f. Observed difference is significant at 5% level of significance and null hypothesis (H_o) is rejected at 5% level of significant. If cal. $t \geq$ table $t_{0.01, (n-1)}$ d.f. Observed difference is highly significant at 1% level of significance and null hypothesis (H_o) is rejected at 1% level of significance. Rejection of null hypothesis ($H_o : \mu_d = 0$) means the given two related samples does not come from the same population.

Confidence limit

Usually parametric values of various characteristics are not known but their estimates are obtained from the samples. Such estimates are known as point estimates of the corresponding parameters. The reliability of such point estimates varies. Some time the mathematical distribution of such estimates is known. On the basis of the nature of distribution and the null hypothesis for a given probability, interval estimates can be worked out which specify that the parametric value may lie between these two values, with a given probability level of confidence. These two values of such interval estimates are known as confidence limits. Thus in case of test of significance of the differences between the sample mean X and population mean μ, one has to determine with reasonable degree of confidence, the range within which the true mean may lie. The limit of this range is usually expressed as confidence limits and the range of these limits is called confidence interval.

Confidence limits depends on

(1) Size of the sample

(2) Level of significance and

(3) inherent variation exist in the population.

The confidence limit for various test criterions can be worked out as under.

(A) Large sample

(I) One sample

$$\text{Lower limit: } \bar{X} - Z_\alpha . S_{\bar{X}} = I_1$$
$$\text{Upper limit: } \bar{X} + Z_\alpha . S_{\bar{X}} = I_2 \qquad I_1 < \mu_o < I_2$$

(ii) Two sample

$$\text{Lower limit: } \left| \bar{X} - \bar{Y} \right| - Z_\alpha . S_{(\bar{X}-\bar{Y})} = I_1$$
$$\text{Upper limit: } \left| \bar{X} - \bar{Y} \right| + Z_\alpha . S_{(\bar{X}-\bar{Y})} = I_2 \qquad I_1 < \mu < I_2$$

(B) Small sample:

(i) One sample

$$\text{Lower limit: } \bar{X} - t_{\alpha,(n-1)} . S_{\bar{X}} = I_1$$
$$\text{Upper limit: } \bar{X} + t_{\alpha,(n-1)} . S_{\bar{X}} = I_2 \quad I_1 < \mu < I_2$$

(ii) Two sample

$$\text{Lower limit: } \left| \bar{X} - \bar{Y} \right| - t_{\alpha,(n_1+n_2-2)} . S_{(\bar{X}-\bar{Y})} = I_1$$
$$\text{Upper limit: } \left| \bar{X} - \bar{Y} \right| + t_{\alpha,(n_1+n_2-2)} . S_{(\bar{X}-\bar{Y})} = I_2 \quad I_1 < \mu < I_2$$

(iii) Paired observation

$$\text{Lower limit: } |\bar{d}| - t_{\alpha,(n-1)} . S_{\bar{d}} = I_1$$
$$\text{Upper limit: } |\bar{d}| + t_{\alpha,(n-1)} . S_{\bar{d}} = I_2$$

F test

t and Z tests are used for comparing two populations mean. When the population is to be compared with respect to their variances the F test is used.

Definition: It is the ratio of the estimates of greater mean square or variance to smaller variance of two different populations.

$$F = \frac{S_1^2}{S_2^2} \quad ; \quad S_1^2 > S^2_2$$

Compare cal F with table $F_{(n_1-1)}$ and $_{(n_2-1)}$ d.f. and draw the conclusion.

χ^2 - test (Chi-square test)

Chi-square was introduced by Karl Pearson in the year 1899. It is calculated by

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

Where, O_i = Observed frequency of i^{th} class

E_i = Expected frequency of i^{th} class

k = number of classes , i = 1,2,..,k

Definition : "It is the sum of the ratio of the square of deviations obtained between observed and expected frequency to the expected frequency of the respective class of the frequency distribution."

Properties of Chi-Square distribution

1) Chi-square distribution is not exact distribution as "t" distribution.
2) It is not symmetrical distribution but it is positively skewed distribution.
3) The value of its varies from 0 to ∞. When there is a perfect agreement of observed frequency distribution with hypothetical frequency distribution, the value of chi-square will be zero, while the value of its increases as there is a departure from the agreement and will increased up to infinity.
4) If χ^2_1, χ^2_2,... , χ^2_k are chi-square values of different samples with n_1, n_2,...,n_k degrees of freedom respectively, the pooled Chi-square value will be equal to

$$\chi^2_p = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \quad \text{with } n = \Sigma n_i \text{ degrees of freedom.}$$

5) The different central moments are $\mu_2 = 2n$; $\mu_3 = 8n$, $\mu_4 = 48n + 12n^2$.
6) As the number of observation tends to infinity, the chi-square distribution tends to normality.
7) The table chi-square value depends upon degrees of freedom. The table chi-square values can be obtained for 1 to 30 d.f., then it is not available from the table. As the number of degrees of freedom exceeds 30, it is found that $\sqrt{\chi^2}$ will be distributed approximately normal about the mean $\sqrt{2n-1}$ with a unit standard deviation. Therefore, the Z value can be worked out by using the following formula and it should be compared with table Z value at 5 per cent or 1 per cent level of significance.

$$Z = \sqrt{2\chi^2} - \sqrt{2n-1}$$

Conditions for application of Chi-Square

1) Deviations (O_i - E_i) should be normally distributed.
2) Number of observations should be sufficiently large. It should be at least 50.
3) Expected frequency of any cell should not be very small. It should be at least 5 and better if it is 10 .

Uses

1) Testing goodness of fit.
2) Testing the independence of attributes for 2 x 2, 2 x c, r x 2 and r x c contingency table.
3) Testing the agreement of genetic ratio with the observed ratio.
4) Test of homogeneity of the families.
5) Test for the detection of linkage.
6) Testing of homogeneity of various variances (Bartlett's test of homogeneity)
7) Testing the heterogeneity among correlation coefficients.

1) Testing goodness of fit

When chi-square test is used to know whether the given sampling distribution is in agreement with the theoretical or expected frequency distribution the test is known as test of goodness of fit.

Procedure:

Step I : Set the appropriate null hypothesis.

H_o : Given sampling distribution is in the agreement with theoretical or expected frequency distribution.

H_a : Given sampling distribution is not in agreement with theoretical or expected frequency distribution.

Step II : Fix the level of significance.

Step III : Work out expected frequency according to given ratio or expectation.

Step IV Calculate Chi square as

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

Where, O_i = Observed frequency of i^{th} class

E_i = Expected frequency of i^{th} class

k = number of classes , i = 1,2,..,k

Step V Compare cal χ^2 with table value at 5% level of significance and (k-1) degree of freedom.

Step VI If cal χ^2 > table $\chi^2_{0.05,(k-1)d.f.}$ observed difference is significant at 5% level of significance. H_o rejected.

If cal χ^2 < table $\chi^2_{0.05,(k-1)\text{ d.f.}}$ observed difference is not significant at 5% level of significance. H_o accepted.

Step VII Conclusion : Non significance difference indicates that the given sampling distribution is in agreement with theoretical distribution and the fit is good.

Significant difference indicates that the given sampling distribution is not in agreement with theoretical distribution and the fit is poor.

2) Test of Independence

Another common use of the chi square test is in testing independence of classifications.

Independence: The two attributes A and B are said to be independent to each other if the proportion of A's among B's is the same as that in not - B's.

Variable :Any character which varying from individual to individual is termed as variable.

Attribute : Attribute is that which is not capable of being described numerically e.g. sex, blindness, colour, shape.

Contingency table : When the individuals in a sample have two characters or attributes and a frequency distribution is made classifying them according to both so as to show the relation between the characters, the resulted table is termed as contingency table.

Yate's correction

In order to avoid irregularities caused by smaller frequencies in 2 x 2 contingency table, a correction for continuity known as Yate's correction is to be applied. When the product of principal diagonal (ad) is greater than off diagonal (bc) i.e. ad > bc, then 1/2 is to be subtracted from the values of Principle diagonal cell frequencies and 1/2 is to be added to the values of off diagonal so that the marginal total remain unchanged. Similarly bc > ad then 1/2 is to be added and 1/2 is to be subtracted from the values of principle and off diagonal cell frequency respectively.

If ad > bc,

$$\chi^2 = \frac{[(ad - bc) - 1/2(a+b+c+d)]^2 \cdot N}{R_1 R_2 C_1 C_2}$$

$$= \frac{[(ad - bc) - N/2]^2 \cdot N}{R_1 R_2 C_1 C_2}$$

Similarly if bc > ad

$$= \frac{[(ad-bc) + 1/2\ (a+b+c+d)]^2 .\ N}{R_1\ R_2\ C_1\ C_2}$$

$$= \frac{[(ad-bc) + N/2]^2 .\ N}{R_1\ R_2\ C_1\ C_2}$$

In general,

$$\chi^2 = \frac{[(ad-bc) - N/2]^2 .\ N}{R_1\ R_2\ C_1\ C_2}$$

Procedure for test of Independence of attribute in case of 2 x 2 contingency table :

Step I Set the appropriate null hypothesis.

H_o : The given classification of group of individuals independent to each other.

H_a : The given classification of group of individuals is not independent to each other.

Step II Fix the level of significance.

Step III Let group A and B are classified in twc ways, the results of the classification can be set out the following table.

Class	A_1	A_2	Total
B_1	a	b	R_1
B_2	c	d	R_2
Total	C_1	C_2	N

Step IV Calculate Chi square as

$$\chi^2 = \frac{(ad - bc)^2 .\ N}{R_1\ R_2\ C_1\ C_2}$$

Where, a, b, c and d are the observed frequency of the respective cell R_1, R_2, C_1, and C_2 are the rows and column totals. N is the grand total.

Step V Compare calculated χ^2 with table value at 5% level of significance and (r-1)(c-1) degree of freedom.

Step VI If cal χ^2 > table $\chi^2_{0.05,(r-1)(c-1)}$ df. observed difference is significant at 5% level of significance. Ho rejected.

If cal χ^2 < table $\chi^2_{0.05,(r-1)(c-1)}$ df. observed difference is not significant at 5% level of significance. Ho accepted.

Step VII Acceptance of Ho means the two characters are independent to each other Rejection of Ho means the two characters are not independent to each other

Procedure for test of Independence of attribute in case of r x 2 contingency table :

Step I Set the appropriate null hypothesis.

H_o : The given classification of group of individuals independent to each other.

H_a : The given classification of group of individuals is not independent to each other.

Step II Fix the level of significance.

Step III Let a group of individuals is classified in two ways in two column and 'r' rows in the following table.

Class	I	II	Total
1st	a_1	b_1	R_1
2nd	a_2	b_2	R_2
3rd	a_3	b_3	R_3
.	.	.	.
nth	a_n	b_n	R_n
Total	C_1	C_2	N

Step IV Work out expected frequency of each cell as follows.

$$E(a_1) = \frac{R_1C_1}{N} \qquad E(a_2) = \frac{R_2C_1}{N}$$

$$E(b_1) = \frac{R_1C_2}{N} \qquad E(b_2) = \frac{R_2C_2}{N}$$

In general,

$$E(X_{ij}) = \frac{R_iC_j}{N}$$

Step V Calculate Chi square as

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

Where, O_i = Observed frequency of i^{th} class

E_i = Expected frequency of i^{th} class

k = number of classes , i = 1,2,..,k

Step VI Compare calculated χ^2 with table value at 5% level of significance and (r-1)(c-1) degree of freedom.

Step VII If cal χ^2 > table $\chi^2_{0.05,(r-1)(c-1)}$ df. observed difference is significant at 5% level of significance. Ho rejected.

If cal χ^2 < table $\chi^2_{0.05,(r-1)(c-1)}$ df. observed difference is not significant at 5% level of significance. Ho accepted.

Step VIII Acceptance of Ho means the two characters are independent to each other Rejection of Ho means the two characters are not independent to each other

Procedure for test of Independence of attribute in case of 2 x C contingency table :

Step I Set the appropriate null hypothesis.

H_o : The given classification of group of individuals independent to each other.

H_a : The given classification of group of individuals is not independent to each other.

Step II Fix the level of significance.

Step III Let a group of individuals is classified in two ways in two rows and 'C' columns in the following table.

Class	1st	2nd	3rd		nth	Total
B_1	a_1	a_2	a_3	..	a_n	R_1
B2	b_1	b_2	b_3	..	b_n	R_2
Total	C_1	C_2	C_3	..	C_n	N

Step IV Work out expected frequency of each cell as follows.

$$E(a_1) = \frac{R_1C_1}{N} \qquad E(a_2) = \frac{R_2C_1}{N}$$

$$E(b_1) = \frac{R_1C_2}{N} \qquad E(b_2) = \frac{R_2C_2}{N}$$

In general, $$E(X_{ij}) = \frac{R_iC_j}{N}$$

Step V Calculate Chi square as

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

Where, O_i = Observed frequency of i^{th} class
E_i = Expected frequency of i^{th} class
k = number of classes , i = 1,2,..,k

Step VI Compare calculated χ^2 with table value at 5% level of significance and (r-1)(c-1) degree of freedom.

Step VII If cal χ^2 > table $\chi^2_{0.05,(r-1)(c-1)}$ df. observed difference is significant at 5% level of significance. Ho rejected.

If cal χ^2 < table $\chi^2_{0.05,(r-1)(c-1)}$ df. observed difference is not significant at 5% level of significance. Ho accepted.

Step VIII Acceptance of Ho means the two characters are independent to each other Rejection of Ho means the two characters are not independent to each other

10. EXPERIMENTAL DESIGN

Experiment: An experiment is a planned inquiry to obtain new facts or confirm or deny the result of previous experiment. Such inquiry can aid in an administrative decision, such as recommencing a variety, feed, cultural practices, a fertilizer, a pesticide or a fungicide etc.

Experimental Unit: An experimental unit or experimental plot is the unit or material to which a treatment is to be applied OR it is a group of material to which a treatment is to be applied in single trial of the experiment. The experimental unit may be a plot of land, a patient in a hospital, an animal, a group of pigs in a pen, a batch of seeds etc.

Treatment: The treatment is the procedure whose effect is to be measured and compared with other treatments e.g. the varieties, manures, chemicals, methods of seed treatments, nutritional and other factors in case of animal.

Experimental Error: It is a measure of the variation which exists among observations on experimental units treated alike.

There are two major sources of such error.-

a. The inherent variability which exists in the experimental material to which treatments are applied.

b. The variation which results from any lack of uniformity in the physical conduct of the experiment.

How the error can be reduced?

a. Handling the experimental material in such a way that the effects of inherent variability are reduced or error is controlled by

(i) Experimental design,

(ii) Use of concomitant observations i.e. by statistical control of error

(iii) The choice of proper size and shape of the experimental units.

b. Refine the experimental technique of experimental design by

(i) Uniformity in the application of treatments.

(ii) Control be exercised over external influence.

(iii) Suitable and unbiased measures of the effects of the treatments should be made available.

Avoid gross error by proper supervision and scrutiny of data.

PRINCIPLES OF EXPERIMENTAL DESIGN

Experiment is the main tool of agricultural research. The aim of an experimenter is to know whether the given treatment is effective or not, if effective, the magnitude of the effect must also be ascertained. The purpose of conducting an experiment is to know these facts. The main difficulty in conducting an experiment is that in which some extraneous factors confuse the treatment effects. For elimination of these extraneous effects, some basic principles are to be followed in planning an experiment. They are replication randomization and local control.

Replication: When a treatment appears more than once in an experiment, it is said to be replicated. In other words, repetition of treatment in an experiment is known as replication.

Why replications are necessary?

(i) Provide an estimate of experimental error,

(ii) Improve the precision of an experiment by reducing the standard deviation of a treatment mean i.e. standard error,

(iii) Increase the scope of inference of the experiment by selection and appropriate use of quite variable experimental units and

(iv) Effect control of the error variance.

Randomization: The allotment of the treatments to the experimental units by random procedure is known as randomization.

Functions: (i) It assures unbiased estimates of treatment means and differences among them and (ii) It assures unbiased estimates of experimental error and thereby valid test of significance.

Local Control: The random allocation of treatments to experimental units while giving an estimate of treatment differences are free from any systematic influence of environment or bias as well as providing a correct test of significance. It is also desirable to reduce the experimental error as far as practicable without interfering with the statistical requirement of randomness because the lower the experimental error the smaller the real difference between treatments can be detected to be significant.

The reduction of experimental error can be achieved by having more homogeneous adjacent experimental units than those widely separated in an experiment. The principal which provides with greater homogeneity of a group of experimental units to reduce the experimental error is known as the local control. Based on this, various forms of plot arrangements, to suit the requirement of particular problems, experimental designs have been evolved.

Analysis of Variance: It is a mathematical process of partitioning the total sum of squares into various recognized sources of variation.

Assumptions underlying ANOVA

For valid use of ANOVA certain assumption should be satisfied.

1. The population from which each sample mean has been drawn is normally distributed.
2. The treatment effects and error are additive in nature.
3. Errors are normally and independently distributed with mean zero and common variance σ^2.

COMPLETELY RANDOMIZED DESIGN (CRD)

Introduction: This design is useful only when the experimental units are homogenous. Completely randomized design is the simplest of all designs. This may be defined as the one in which the treatments are randomly allotted to the entire experimental area. No effort is made to divide the area into blocks or to confine treatments to any portion of the entire area.

Application: The use of CRD is limited to green house studies, methodological studies and the laboratory experiments where the experimental material is homogeneous. On the other hand this design is seldom used in field experiments. Again the design may be appropriate only in situations where the variation over the entire experimental unit is relatively small.

Layout: The layout refers to the placement of treatments on the experimental site. Suppose there are 3 treatments A, B and C and it is desired to replicate (repeat) them 8 times. The experimental area is divided into 24 equal plots and the treatments are allotted to these plots at random. There is no restriction of assigning the treatments to plots.

Completely Randomized Design with 3 treatments.

B 1	C 9	B 17
A 2	B 10	A 18
A 3	C 12	C 19
A 4	C 12	B 20
A 5	C 13	B 21
A 6	C 14	B 22
A 7	C 15	B 23
A 8	C 16	C 24

Randomization of treatments:

Here, since the number of units is 24, a two digit random number table will be considered and a series of 24 random numbers will be taken excluding those which are greater than 24. Suppose the random numbers chosen are 4, 18, 2, 21, 14, 3, 7, 13, 22, 1, 6, 10, 17, 23, 20, 8, 15, 11, 24, 5, 9, 12, 16 and 19. After this the plots will be serially numbered and the treatment A will be allotted to the plots bearing the serial numbers 4, 18, 2, 21,, 2, 14, 3, 7, and 13; treatment B will be allotted to the plots bearing the serial numbers 22, 1, 6, 10, 17, 23, 20 and 8 and the rest for the treatment C. Alternatively chit box method can also be employed for drawing the random numbers. Here 24 chits are prepared bearing serial numbers 1 to 24. The first 8 chits of numbers drawn will be the experimental plot numbers that will receive treatment A. The remaining chits are drawn and allotted similarly to the treatments B and C.

Statistical Model : $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

Where,Yij = Response of yield from the j^{th} unit receiving the i^{th} treatment

μ = General mean

τ_i = Effect of i^{th} treatment

ε_{ij} = Uncontrolled variation associated with j^{th} unit receiving i^{th} treatments.

Analysis of variance: Completely randomized design with equal replications

Source of variation	DF	Sum of Squares (SS)	M. S. (SS/DF)	Cal. F
Treatment	(t-1)	$\frac{\sum_{i=1}^{t} Y_{i.}^2}{r} - \frac{(\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij})^2}{rt}$	MST	$\frac{MST}{MSE}$
Error	t(r-1)	By subtraction	MSE	
Total	(rt-1)	$\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}^2 - \frac{(\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij})^2}{rt}$		

Analysis of variance: Completely randomized design with unequal replication

Source of variation	DF	Sum of Squares (SS)	M. S. (SS/DF)	Cal. F
Treatment	(t-1)	$\sum_{i=1}^{t} \frac{Y_{i.}^2}{r_i} - \frac{(\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij})^2}{\sum_{i=1}^{t} r_i}$	MST	$\frac{MST}{MSE}$
Error	$\sum_{i=1}^{t} r_i - t$	By subtraction	MSE	
Total	$\sum_{i=1}^{t} r_i - 1$	$\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}^2 - \frac{(\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij})^2}{\sum_{i=1}^{t} r_i}$		

$$SEm = \sqrt{MS_E / r \text{ or } r_0} \qquad \text{or} \qquad SEd = \sqrt{MS_E \left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$$

Where, r = Number of observations for treatments (equal number of observations)

r_o = Harmonic mean of number of observations for different treatments (when unequal number of observation for treatment)

$$r_0 = \frac{\text{No. of Treatments}}{1/r_1 + 1/r_2 + \ldots\ldots + 1/r_t}$$

Where r_1, r_2... are the number of observations for different treatments.

lsd at 5 % if treatment F is significant. = SEd x $t_{0.05,\ ne}$

$$CV\ \% = \frac{\sqrt{MS_{\varepsilon}}}{\bar{Y}_{..}} \times 100$$

Advantages:

1. It is easy to layout the design
2. The statistical analysis is simple
3. There is all flexibility as regards the number of treatments in the experiment and the number of repetitions per treatment.

Disadvantage:

1. If the experimental material is heterogeneous, the error variation is very large and thus the precision will be affected.
2. This design is recommended therefore in glass house studies and in other situations where uniform conditions for the experimental material can be maintained.

RANDOMIZED BLOCK DESIGN (RBD)

Introduction:

In this design the whole experimental material is divided into homogeneous groups, each of which constitutes a single replication. Each of these groups is further divided into a number of experimental units which are equal in all respects. The treatments are applied to these units by any random process. It is important to note that fresh randomization is done in each block. The number of plots in each block is equal to the number of treatments, so that each block is a replicate of each treatment. An important and essential point on which the attention kept is that the experimental errors within each group are to be kept as small as practically possible and the variation from block to block as great as possible. In this way all the treatments which are assigned to one group, experience the same type of environmental effects and are therefore comparable.

Applications:

This experimental design is especially used in the fields of research where the experimental material is expected to be heterogeneous but it is possible to group the homogeneous experimental units in the form of blocks. It's use in agriculture is very common because the fields on which the experiments are to be laid out are generally heterogeneous in nature, but the

fertility gradient of the soil might have a tendency towards a particular direction. It is also possible that adjacent contagious plots forming a block will generally be homogeneous in nature as compared to widely apart experimental plots, forming the experimental material to which the treatments are to be superimposed. Similarly, in case of experiments on trees or animals where it is possible to control the variation in some known characteristics is possible, this design is used.

Layout plan 1 . No. of treatment : 5 (A,B,C,D,E), No. of replications : 4

I

1	2	3	4	5
B	C	E	A	D

II

6	7	8	9	10
D	A	C	E	B

III

11	12	13	14	15
A	C	D	B	E

IV

16	17	18	19	20
B	D	C	A	E

Layout plan 2 : Treatments = 8, No. of replications : 3

I

1	2	3	4
T6	T1	T4	T3
5	6	7	8
T7	T5	T2	T8

II

9	10	11	12
T5	T2	T7	T3
13	14	15	16
T4	T1	T6	T8

III

17	18	19	20
T3	T8	T5	T1
21	22	23	24
T6	T2	T4	T7

Note : Compact near to square block long and narrow plots.

Here the numbers indicate the numbers of experimental units constituting a block or replication and the letters superimposed on the units indicate the randomly allotted treatments.

Randomization of treatments

As indicated earlier, the experimental units forming a block or replication are serially numbered. Randomization of treatments is carried out on these units. Fresh randomization of treatments for different blocks or replications is performed separately as indicated in the above layout plan.

Statistical model:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

Where, Y_{ij} = Response or yield from the j^{th} unit receiving the i^{th} treatment

μ = General mean

τ_i = Effect of i^{th} treatment

β_j = Effect of j^{th} replication

ε_{ij}= Uncontrolled variation associated with i^{th} treatment in j^{th} replication.

Analysis of Variance

Source of variation	DF	Sum of Squares (SS)	M. S. (SS/DF)	Cal. F
Replication	(r-1)	$\frac{\sum_{j=1}^{r} Y_{.j}^2}{t} - \frac{(\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij})^2}{rt}$	MSR	MSR/MSE
Treatment	(t-1)	$\frac{\sum_{i=1}^{t} Y_{i.}^2}{r} - \frac{(\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij})^2}{rt}$	MST	MST/MSE
Error	(t-1)(r-1)	By subtraction	MSE	
Total	(rt-1)	$\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij}^2 - \frac{(\sum_{i=1}^{t}\sum_{j=1}^{r} Y_{ij})^2}{rt}$		

Standard error and critical difference:

The standard error of mean = $S.Em. = \sqrt{MS_{\varepsilon}/r}$

The standard error or the difference between the treatment means based on r replications is estimated by the relation.

$$S.Ed. = \sqrt{2MS_{\varepsilon}/r}$$

where, MSE = Error M.S. r = No. of replications

Critical difference at 5 % level of significance

CD 0.05 = S Em x $\sqrt{2}$ x t0.05,ne or SEd x $t_{0.05,ne}$

(when treatment F is significant)

Advantages:

(1) Any number of treatments can be tried in this design. However, this depends upon the homogeneity of the material within a group of replication.

(2) Any number of replications can be had depending upon the availability of the experimental material.

NON PARAMETRIC TEST

Test of significance viz. Z, t and F are the parametric Tests. The parameters need to be estimated and the hypothesis concerning them needs to be tested. These tests are valid under the assumption that the sample comes from the population having a known distribution.

Many a times it is difficult to specify the distribution easily and therefore it is not possible to make a parametric Test. To handle such data, we need distribution-free statistics. If we do not specify the nature of the parent distribution, then We will not ordinarily deal with parameters. Therefore, these Tests are known as non-parametric test.

Advantages

(1) Since the ranks or the signs of difference are used, they. are easy to learn and apply

(2) For the same reason, they may reduce the work of data collection e.g. In case of insect or disease infestation ranking is easy.

(3) In varietial trials of different locations / years the experimental errors are likely to be heterogeneous and the pooled analysis is complicated. Whereas, the analysis of non parametric procedure is simple.

(4) One does not have to assume specific form of distribution of the population from which the sample has been drawn for the study.

Disadvantages

(1) When the distribution of the population is known and we use non parametric test, we loose some information in ranking or in using only signs of difference of the data.

(2) When null hypothesis is rejected, the non parametric tests are not as effective as parametric tests.

SIGN TEST

This test is used, when observations can be paired. Instead of mean of differences, the median of difference is considered and signs of difference are used in the test.

Procedure

(1) Find the difference of the paired observations and indicate + and - sign.

(2) Count plus and minus signs.

(3) Set the null hypothesis that, differences have median zero, Chi-square test is used under $H_o : p = 0.5$

$$\chi^2 = \frac{(|n_1 - n_2| - 1)^2}{n_1 - n_2}$$

Where n_1 and n_2 are numbers of positive and negative signs of differences. If cal (χ^2 < table (χ^2 value difference is non significant. Hence H_o is accepted. If cal. χ^2 > table χ^2 difference is significant. Hence H_o is rejected.

Example : To test the effect of insecticide the experiment was conducted on paddy crop. The yields of paddy in paired plots are as under.

Pair No.	Yield		Sign test
	Sprayed	Unsprayed	Difference
1	63.3	69.0	-5.7
2	78.1	74.4	+3.7
3	93.0	86.6	+6.4
4	80.0	79.2	+1.5
5	89.0	84.7	+4.3
6	79.9	75.1	+4.8
7	87.3	90.6	-3.3
8	102.4	98.8	+3.6
9	70.7	70.2	+0.5
10	106.1	101.1	+5.0
11	7.4	83.4	+24.0
12	74.0	65.2	+8.8
13	72.6	68.1	+4.5
14	69.5	68.4	+1.1
15	69.5	68.4	+1.1

Step : 1) Find the differences and indicate + and - signs.

2) + signs $n_1 = 12$ and - signs $n_2 = 3$

3) H_o : p = 0.5 (The differences has median zero)

Cal. $\chi^2 = 5.7$ > table value of $\chi^2_{1\ d.f.05} = 3.841$. Result is significant Hypothesis is rejected, means there is a difference in the yield of sprayed and unsprayed plot.

Limitations :

(1) Lots of information of magnitude of differences is wasted in this test.

(2) It is not possible to use this test with less than six pairs and is not much useful and for less than 12 pairs, for 20 or more pairs it becomes more useful.

WILCOXON'S SIGNED RANK TEST

This test is an improvement upon the sign test. Here the magnitudes of differences is considered by ranking the differences.

Procedure

(1) Ranks the difference between paired values from smallest to largest without considering the sign.

(2) Assign average rank if the differences are of equal magnitude.

(3) Assign to the ranks the signs of the original difference.

(4) Compute the sum of either positive or negative ranks which is smaller.

(5) Compare the sum obtained at step 3 with the critical value of Table.

Example : To test the effect of insecticide the experiment was conducted on paddy crop. The yields of paddy in paired plots are as under.

Pair No.	Yield		Difference	Signed Ranks
	Sprayed	Unsprayed		
1	63.3	69.0	-5.7	-12
2	78.1	74.4	+3.7	-7
3	93.0	86.6	+6.4	+13
4	80.0	79.2	+1.5	+4
5	89.0	84.7	+4.3	+8
6	79.9	75.1	+4.8	+10
7	87.3	90.6	-3.3	-5
8	102.4	98.8	+3.6	+6
9	70.7	70.2	+0.5	+1
10	106.1	101.1	+5.0	+11
11	7.4	83.4	+24.0	+15
12	74.0	65.2	+8.8	+14
13	72.6	68.1	+4.5	+9
14	69.5	68.4	+1.1	+1.5
15	69.5	68.4	+1.1	+1.5

Cal. T = 24 < table value for 15 pair = 25 It is significant hence H_o : is rejected. Means there is a difference in the yield of sprayed and unsprayed plot.

Note : 1) Small value of T is significant one.

2) When tie occurs, average value is given to each rank.

	PARAMETRIC TEST	NON PARAMETRIC TEST
1	Parametric test depend on the condition of the parameter of the population	They do not depend upon nature of parent distribution.
2	They are used for the data of at least interval scale	They can be used for the data of nominal and ordinal scale
3	They are most powerful due to strong assumption	They are weak due to weak assumption
4	It compares parameters	It compares distributions
5	They are difficult to learn and apply	They are easy to learn and apply.

TABLE VALUE OF WILCOXON'S SIGNED RANK TEST.

Tabulated value of T are such that smaller values regardless of sign, occurred by chance with stated probability.

Pairs	Probability	
n	0.05	0.01
6	0	-
7	2	-
8	4	0
9	6	2
10	8	3
11	11	5
12	14	7
13	17	10
14	21	13
15	25	16
16	30	20
17	35	23
18	40	28
19	46	32
20	52	38
21	59	43
22	66	49
23	73	55
24	81	61
25	89	68

SAMPLING

Introduction

Our knowledge, our actions and our attitudes are based to a very large extent on samples. This is equally true in everyday life and in scientific research, whenever, a person wants to buy a large quantity of a commodity say wheat, rice, fruits etc. be decided about total lot of just by simply examining a small fraction of it . A doctor's opinion about the state of health is determined by only examining one or two drops of blood. We study the soil about its nutrient status, salinity status etc by using only 5 g. of soil. It has been experienced that the sample survey if planned properly, can give precise information.

A sample is a part of population (total or aggregate). It is to be selected at random for unbiased estimate and can be used as a basis for inferring about the population. Most populations (a crop plant population, groundnut growers, pest population pump-set owners etc) are so large that it is rather next to impossible to contact each of them during specified time, only a fraction of it is investigated and the inference is drawn from it for the population . Consequently many generalizations which arise from ordinary experience are likely to be unwarranted.

The sample has many advantages over a census or complete enumeration of a finite population . If carefully designed, the sample is not only cheaper but may give results which are as accurate or sometimes, more accurate than those of census. The reason is that a census is subjected to biased error than that due to any sampling method employed. The smallness of samples makes possible the use of the more care in the design and execution of each step in the inquiry. In this way sources of error can be investigated and reduced, eliminated or measured and corrected for. The other advantages of sample survey are that it is less time consuming, involves less cost, has greater scope and has greater operational facilities. It is for these reasons that sample surveys are being preferred by the research scientists to complete enumeration.

Terminology

1) Population : A population is the aggregate of individuals or units or objects having atlest one common characteristics.

2) Sample : A sample is a part or fraction of large aggregate (Population) about which some information is required.

3) Sampling : It is the method/process of selection of sample from the population.

4) Sampling Unit : It is the individual element or a group of elements on which observations are to be made.

5) Parameter : Any measurable characteristic of population which is to be worked out by utilizing each and every observation of population parameters help to characterize the population mean μ and σ are the example of the parameter. Parameters are generally unknown and constant.

6) Statistic : Any number estimated from sample (X and S are the statistic)

Type of population

1) Finite population :-

One can count rose plants in a garden i.e. counting the individuals in the population is possible & hence it is finite population

e.g. Rose plants in garden, No. of animals in herd, Fruits on tree etc.

2) Infinite population :-

The rose plants on the earth can not be counted. .Thus it becomes infinite population.

3) Real population :-

This is the population in which the members do exist in reality.

e.g. a heap of food grains, herd of cows etc.

4) Hypothetical population :-

The members of population does not exist in reality.

e.g. Possible throws of a die. Similarly, yield of a new variety to be evolved. Possible results of an experiments etc. are the examples of this type of population.

Advantages of sample study

1) Less expensive.
2) In a limited specific time one can complete the project work and collect the information (i.e. greater speed)
3) With minimum technical persons, one can study the problem
4) More precise and accurate information regarding population (greater accuracy) may be obtained.
5) Sample investigation can be carryout with minimum facilities.
6) Sample investigation is to be done when the unit is supposed to be destroyed while taking observation

Sampling plans/Designs

1) Simple random sampling.
2) Stratified random sampling.
3) Multistage sampling.
4) Cluster sampling.
5) Systematic sampling.
6) Purposive sampling and so on.

(1) Simple random sampling (SRS)

In a sample random sampling, the sample is selected in such a way that every member of the population has an equal and independent chance of being selected in the sample. It implies that selection of a sample from all possible samples that could be chosen is equally likely.

Random selection of units is done using any one of the following methods.

a) Using Tickets, tags etc.

b) Random number tables.

a) Using Tickets, tags etc

To give an example, suppose that an experimenter wishes to draw a random sample of size 10 individuals (Say 10 plants for measuring the heights) from a finite population, say 200 plants. A method of doing this would be assign a number to each member of the population put the individuals into a box, and mix them thoroughly. Draw ten tickets or tags from the box. The number on these ten tickets or tags correspond the plants to be selected.

This method is time and labor consuming and also costly.

b) Use of Random Number table

The procedure (a) can be shortened by the use of a table of random number. Such a table consists of numbers chosen in a fashion similar to drawing numbered tickets or tags out of a box. This table is so made that all numbers 0,1,2.... appear with approximately the same frequency. By combining the numbers in pairs we have two-digit numbers (i.e. from 00 to 99) By using the number three at a time we have three digit numbers from 000 to 999 etc.

The table should be entered in a random manner. Put a pencil aimlessly on a page of the table . The point thus obtained on page is the starting point for selecting the numbers, Record the numbers until the required number of random digits is obtained.

Examples of SRS

1) Impact of T & V system on Socio-economic status of summer groundnut growers in a specific Taluka.

Here population is "Summer groundnut growers registered under T & V system of a given taluka. One can easily prepare the frame for this finite population and select the random sample of summer groundnut growers.

2) Constraint analysis for milk : Productivity in a village

Here population is " milk producers in a given village. The frame for which can be prepared easily and a random sample can be taken to fine out the factors responsible for low productivity.

(2) Stratified Random sampling

In this scheme the population is sub-divided into several groups called stratum ands then samples are drawn independently (at random) from each stratum.

As the sampling variance of the estimate of mean depends on the within strata variation, the stratification of heterogeneous population into homogeneous strata helps in increasing precision of the estimates e.g. While studying average income of the staff members of the Gujarat Agricultural University one has to employ stratified random sampling. Because, simple random sampling may result into under or over estimation of income of the staff member. Suppose only professors are selected for the sample then the average income will be above the true average value, similarly if only helpers are selected in the sample then the results will be on lower side .

When such heterogeneous population is required to be sampled then one has to utilize the stratified random sampling in spite of simple random sampling.

Examples

1) Adoption level of improved agro technology by the cultivators.
2) A survey for nutritional status of school going students.
3) Socio-economic survey for rural and urban people of Valsad district.

(3) Multistage sampling

In this method, the selection is done in stages, called sub sampling. For example in estimating the yield of wheat in a district. Talukas may be considered as primary samplings unit (1st stage), Villages. Within taluka as secondary sampling units (2nd stage) the cultivators within village within taluka as third stage sampling units and so on.

The advantage of this type of sampling is that at the first stage the frame of primary sampling units is required which is easy to have and at the second stage the frame of secondary sampling units is required for the selected primary sampling units only and so on. Moreover , this method allows the use of different selection procedure in different stages. It is because of this consideration that multi-stage sampling is used in most of the large scale surveys.

(4) Systematic sampling

Systematic sampling is slight varying compared to the simple random sampling in which only the first sample units is selected at random and the remaining units are automatically selected in a definite sequence at equal spacing from one another. This technique of drawing samples is usually recommended if the complete and up to date list of the sampling units is available and the units are arranged in some systematic order such alphabetical, Chronological geographical order etc. This requires the sampling units in the population to be ordered in such a way that each item in the population is uniquely identified by its order, for example the name of persons in a telephone directory, the list of voters etc.

Sampling and Non-sampling Errors.

The inaccuracies or errors in any statistical investigation i.e. in the collection, processing, analysis and interpretation of the data may be broadly classified as follows:

1) Sampling errors 2) Non sampling errors.

(1) Sampling Errors

In a sample survey, since only a small portion of population is studies and hence its results are bound to differ from the census results and thus have a certain amount of error. This error would always be there no matter that the sample is drawn at random and that it is highly representative. This error is attributed to fluctuations of sampling and is called sampling error. Sampling error is due to the fact that only a subset of the population (i.e. sample) has been used to estimate the population parameters and draw inferences about the population. Thus sampling error is present only in a sample survey while it is completely absent in census surveys.

Sampling error may be due to following reasons.

(i) Faulty selection of the sample

(ii) Substitution.

(iii) Faulty demarcation of sampling units

(iv) Error due to bias in the estimation method

(v) Variability of the population.

(2) Non Sampling Errors

Non-sampling errors are not attributed to chance and are a consequence of certain factors which are within human control. In other words they are due to certain causes which can be traced and may arise at any stage of the inquiry viz. Planning and execution of the survey and collection, processing and analysis of the data. This error is present in sample and census.

Some of the important factors responsible for non sampling errors are as under.

(i) Faulty planning including vague and faulty definitions of the population of the statistical units to be used, incomplete list of population members.

(ii) Vague and imperfect questionnaire which might result in incomplete or wrong information.

(iii) Defective methods of interviewing and asking questions.

(iv) Vagueness about the type of the data to be collected.

(v) Personal bias of the investigator.

(vi) Lack of trained and qualified investigators and lack of supervisory staff.

(vii) Failure of respondents memory to recall the events or happenings in the past.

(viii) Non response and inadequate or incomplete response.

(ix) Improper coverage.

(x) Compiling errors.

(xi) Publication errors.