A computer system is an integrated form of different components that work together to give a desirable result. It has different components and each works for a specific purpose; however, they generate a common result as required by the user.

Components of Computer System

The following basic components of a computer system are

- Hardware
- Software
- Humanware
- Firmware
- Bridgeware

Hardware

The physical components collectively form the hardware of a computer system. Hardware comprises the equipment that helps in the working system of the computer.

Following are the different types of hardware components (which have specific functions) −

- **Monitor** − It displays (visual) the result.
- **CPU** − It is the Central Processing Unit that controls the computer's functions and transmits data.
- **Motherboard** − It is mainly accountable to establish communication between components and transmission of information.
- **RAM** − It is the Random Access Memory and responsible for the storage of programs that are currently running and also stores data temporarily.
- **Hard Disk Drive** − It is a permanent memory storage device.
- **Floppy Disk Drive** − It is hardly being used in recent times.
- **Optical disks** − It is a device that also store data. For example, CD, DVD, etc.

Humanware

Humanware refers to the persons who design, program, and operate a computer installation.

There are numerous categories of jobs, but the three principal positions required in a large computer installation are

1. **System analyst:**

- study information and processing requirements.
- defines the applications problem,
- determines systems specifications,
- recommends hardware and software changes,
- designs information processing procedures.

2. **Programmer:**

- code or prepare programs based on the specifications made by the systems analyst.

3. **computer operator.**

- performs a series of well defined tasks that will keep the computer operating at maximum efficiency.

Firmware

- Firmware is a software program on the hardware device, which perform functions like basic input/output tasks and offers necessary instructions for the device to communicate with other computing devices.
- It is a set of instructions programmed that is permanently etched into a hardware device like video cards, BIOS,, keyboards , or hard drives
- written to a hardware device's non-volatile memory such as flash read-only memory (ROM) , can be erased and rewritten.
- Generally, these are booting programs which help in the starting of a computer. Such programs cannot be erased or overwritten.

Bridgeware

- Any software or hardware that eases the transition from use of one computer system to use of another not entirely compatible one. Bridgeware is normally supplied by a computer manufacturer when a new range of machines does not offer complete upward compatibility from some previous range.
- These necessary because different computers are made by different manufacturers.

Input & Output Device

The following table categorically lists down the input and output device –

Input Device	**Output Device**	**Input Device**	**Output Device**
Mouse	Monitor	Microphone	Speaker
Keyboard	Printer	Camera	Earphone
Scanner	Projector	Trackball	Monitor
Touchpad	Plotter	Joystick	Monitor

Software

The hardware components can only function when software components are added to the computer system. Software is a program that performs different commands given by a user. Software is an intangible part of hardware and controls the sequence of operations.

Types of Software

Depending on the basic features and functionality, software can be categorized as −

- Operating Systems (OS)
- Application Software (AS)
- E-accessibility Software

Let us now discuss the software components in brief.

Operating System

This software helps to load the basic program automatically as soon as the computer is started. Following are the major types of operating system −

Operating Software	**Examples**
Microsoft Windows	XP, Vista, etc.
Mac OS X	Panther, Cheetah, Snow leopard, etc.
Linux	Debian, Ubuntu, Fedora, Knoppix, etc.

Application Software

The software, which can be used on an installed operating system, is known as application software. Following are the significant examples of application software −

Application Software	**Examples**
Office programs	Microsoft Office, OpenOffice, LibreOffice, etc.
Web browser	Internet Explorer, Mozilla Firefox, Google Chrome, Opera, Safari, etc.
Antivirus Program	Norton, McAfee, Quick Heal, Avira, Kaspersky, etc.

E-accessibility Software

The E-accessibility software components additional facilities to users such as −

- Voice recognition software
- Screen reader
- Magnifying tool
- On-screen keyboard
- Video games
- Learning software, etc.

Programming Languages

The computer system is simply a machine and hence it cannot perform any work; therefore, in order to make it functional different languages are developed, which are known as programming languages or simply computer languages.

Following are the major categories of Programming Languages −

- Machine Language
- Assembly Language
- High Level Language

Let us discuss the programming languages in brief.

Machine Language or Code

This is the language that is written for the computer hardware. Such language is effected directly by the central processing unit (CPU) of a computer system.

Assembly Language

It is a language of an encoding of machine code that makes simpler and readable.

High Level Language

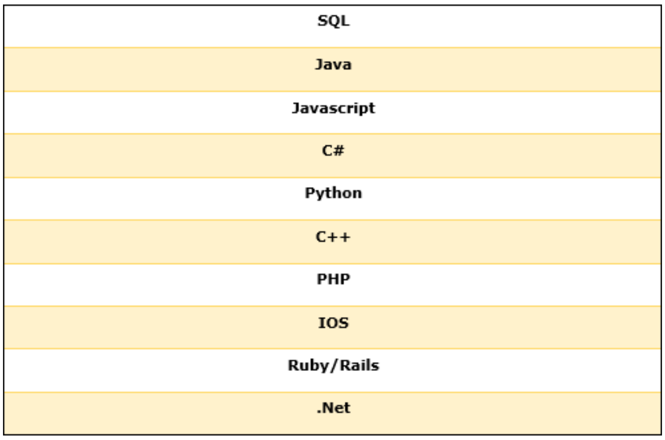
The high level language is simple and easy to understand and it is similar to English language. For example, COBOL, FORTRAN, BASIC, C, C+, Python, etc.

High-level languages are very important, as they help in developing complex software and they have the following advantages −

- Unlike assembly language or machine language, users do not need to learn the high-level language in order to work with it.
- High-level languages are similar to natural languages, therefore, easy to learn and understand.
- High-level language is designed in such a way that it detects the errors immediately.
- High-level language is easy to maintain and it can be easily modified.
- High-level language makes development faster.
- High-level language is comparatively cheaper to develop.
- High-level language is easier to document.

Although a high-level language has many benefits, yet it also has a drawback. It has poor control on machine/hardware.

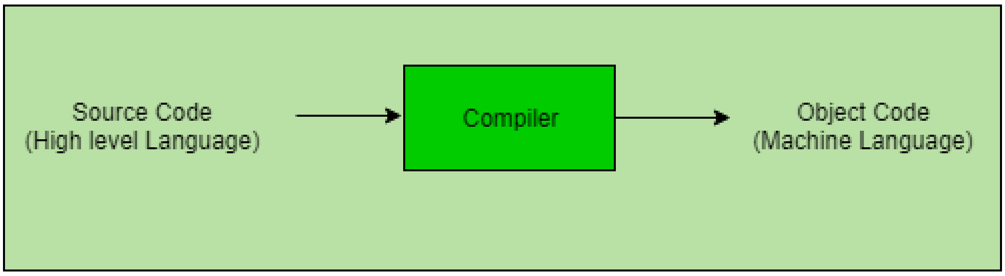
The following table lists down the frequently used languages −



SQL
Java
Javascript
C#
Python
C++
PHP
IOS
Ruby/Rails
.Net

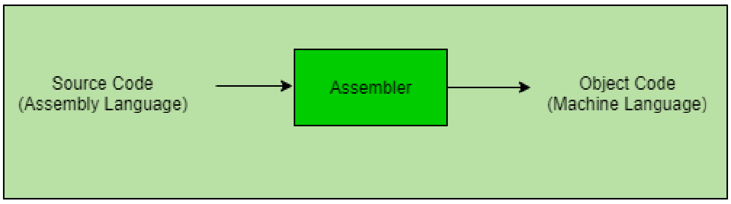
<u>Compiler</u>

The language processor that reads the complete source program written in high-level language as a whole in one go and translates it into an equivalent program in machine language is called a <u>Compiler</u>. Example: C, C++, C#, Java.



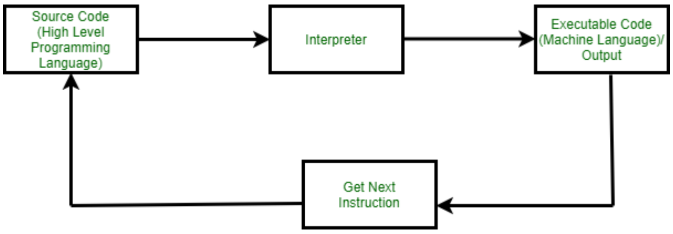
Assembler:

The Assembler is used to translate the program written in Assembly language into machine code.



Interpreter:

The translation of a single statement of the source program into machine code is done by a language processor and executes immediately before moving on to the next line is called an interpreter.



Hardware and Software

The following table highlights the points that differentiate a hardware from a software.

Hardware	Software
It is the physical component of a computer system.	It is the programming language that makes hardware functional.
It has the permanent shape and structure, which cannot be modified.	It can be modified and reused, as it has no permanent shape and structure.
The external agents such as dust, mouse, insects, humidity, heat, etc. can affect the hardware (as it is tangible).	The external agents such as dust, mouse, insects, humidity, heat, etc. cannot affect (as it is not tangible).
It works with binary code (i.e., 1's to 0's) .	It functions with the help of high level language like COBOL, BASIC, JAVA, etc.

It takes in only machine language, i.e., lower level language.	It takes in higher level language, easily readable by a human being.
It is not affected by the computer bug or virus.	It is affected by the computer bug or virus.
It cannot be transferred from one place to other electronically.	It can transfer from one place to other electronically.
Duplicate copy of hardware cannot be created.	A user can create copies of a software as many as he wishes.

Operating System

An operating system is the fundamental basis of all other application programs. Operating system is an intermediary between the users and the hardware.

Operating system controls and coordinates the use of hardware among application programs.

The major services of an operating system are −

- Memory management
- Disk access
- Creating user interface
- Managing the different programs operating parallel
- Likewise, it controls and manage the hardware's working

Applications of Operating System

Following are the major *applications* of an operating system −

- An operating system is accountable for the formation and deletion of files and directories.
- An operating system manages the process of deletion, suspension, resumption, and synchronization.
- An operating system manages memory space by allocation and de-allocation.
- An operating system stores, organizes, and names and protects the existing files.

Types of Operating System

Following are the major types of operating system −

- Disk Operating System (DOS)
- Windows Operating System
- Unix Operating System

Let us now discuss each operating system in detail.

Disk Operating System

MS-DOS is one of the oldest and widely used operating system. DOS is a set of computer programs, the major functions of which are file management, allocation of system resources, providing essential features to control hardware devices.

DOS commands can be typed in either upper case or lower case.

Features of DOS

Following are the significant features of DOS −

- It is a single user system.
- It controls program.
- It is machine independence.
- It manages (computer) files.
- It manages input and output system.
- It manages (computer) memory.
- It provides command processing facilities.
- It operates with Assembler.

Types of DOS Commands

Following are the major types of DOS Command −

- **Internal Commands** − Commands such as DEL, COPY, TYPE, etc. are the internal commands that remain stored in computer memory.
- **External Commands** − Commands like FORMAT, DISKCOPY, etc. are the external commands and remain stored on the disk.

Windows Operating System

The operating system window is the extension of the disk operating system.

It is the most popular and simplest operating system; it can be used by any person who can read and understand basic English, as it does not require any special training.

However, the Windows Operating System requires DOS to run the various application programs initially. Because of this reason, DOS should be installed into the memory and then window can be executed.

Elements of Windows OS

Following are the significant element of **W**indows **O**perating **S**ystem (WOS) −

- Graphical User Interface
- Icons (pictures, documents, application, program icons, etc.)
- Taskbar

- Start button
- Windows explorer
- Mouse button
- Hardware compatibility
- Software compatibility
- Help, etc.

Versions of Windows Operating System

Following are the different versions of Windows Operating System −

Version	Year	Version	Year
Window 1.01	1985	Windows XP Professional x64	2005
Windows NT 3.1	1993	Windows Vista	2007
Windows 95	1995	Windows 7	2009
Windows 98	1998	Windows 8	2012
Windows 2000	2000	Windows 10	2015
Windows ME	2000	Windows Server 2016	2016
Windows XP	2001		

Unix Operating System

The Unix Operating System is the earliest operating system developed in 1970s. Let us consider the following points relating to the Unix Operating System −

- It is an operating system that has multitasking features.
- It has multiuser computer operating systems.
- It runs practically on every sort of hardware and provides stimulus to the open source movement.
- It has comparative complex functionality and hence an untrained user cannot use it; only the one who has taken training can use this system.
- Another drawback of this system is, it does not give notice or warn about the consequences of a user's action (whether user's action is right or wrong).

Types of Computers:

ll the computers that are developed are not alike rather they have different designs and features. Some computers have very high capacity as well as working speed; however, some are slow. Depending upon the requirements, computers are being developed.

Types of Computer

Depending upon the internal structure and subsequent features and applicability, computer system is categorized as follows −

Mainframe Computer

It is high capacity and costly computer. It is largely used by big organizations where many people can use it simultaneously.

Super Computer

This category of computer is the fastest and also very expensive. A typical supercomputer can solve up to ten trillion individual calculations per second.

Workstation Computer

The computer of this category is a high-end and expensive one. It is exclusively made for complex work purpose.

Personal Computer (PC)

It is a low capacity computer developed for single users.

Apple Macintosh (Mac)

It is a sort of personal computer manufactured by Apple company.

Laptop computer (notebook)

It is a handy computer that can be easily carried anywhere

Tablet and Smartphone

Modern technology has advanced further. It has helped develop computers that are pocket-friendly. Tablets and smartphones are the best examples of such computer.

Computer - Input Devices

The input devices enable users to **send signals** to the computer in order for it to complete a certain task.

The **Central Processing Unit (CPU)** then receives the signal and transmits it to the output devices.

Some **examples of Input devices are**:

- Keyboard
- Mouse
- Joy Stick
- Light pen

- Track Ball
- Scanner
- Graphic Tablet
- Microphone
- Magnetic Ink Card Reader(MICR)
- Optical Character Reader(OCR)
- Bar Code Reader
- Optical Mark Reader(OMR)

Keyboard

Keyboard is the most common and very popular input device which helps to input data to the computer. The layout of the keyboard is like that of traditional typewriter, although there are some additional keys provided for performing additional functions.

Keyboards are of two sizes 84 keys or 101/102 keys, but now keyboards with 104 keys or 108 keys are also available for Windows and Internet.

The keys on the keyboard are as follows –

Typing Keys

These keys include the letter keys (A-Z) and digit keys (09) which generally give the same layout as that of typewriters.

Numeric Keypad

It is used to enter the numeric data or cursor movement. Generally, it consists of a set of 17 keys that are laid out in the same configuration used by most adding machines and calculators.

Function Keys

The twelve function keys are present on the keyboard which are arranged in a row at the top of the keyboard. Each function key has a unique meaning and is used for some specific purpose.

Control keys

These keys provide cursor and screen control. It includes four directional arrow keys. Control keys also include Home, End, Insert, Delete, Page Up, Page Down, Control(Ctrl), Alternate(Alt), Escape(Esc)

Special Purpose Keys

Keyboard also contains some special purpose keys such as Enter, Shift, Caps Lock, Num Lock, Space bar, Tab, and Print Screen.

Mouse

Mouse is the most popular pointing device. It is a very famous cursor-control device having a small palm size box with a round ball at its base, which senses the movement of the mouse and sends corresponding signals to the CPU when the mouse buttons are pressed.

Generally, it has two buttons called the left and the right button and a wheel is present between the buttons. A mouse can be used to control the position of the cursor on the screen, but it cannot be used to enter text into the computer.

Advantages

- Easy to use
- Not very expensive
- Moves the cursor faster than the arrow keys of the keyboard.

Joystick

Joystick is also a pointing device, which is used to move the cursor position on a monitor screen. It is a stick having a spherical ball at its both lower and upper ends. The lower spherical ball moves in a socket. The joystick can be moved in all four directions.

The function of the joystick is similar to that of a mouse. It is mainly used in Computer Aided Designing (CAD) and playing computer games.

Light Pen

Light pen is a pointing device similar to a pen. It is used to select a displayed menu item or draw pictures on the monitor screen. It consists of a photocell and an optical system placed in a small tube.

When the tip of a light pen is moved over the monitor screen and the pen button is pressed, its photocell sensing element detects the screen location and sends the corresponding signal to the CPU.

Track Ball

Track ball is an input device that is mostly used in notebook or laptop computer, instead of a mouse. This is a ball which is half inserted and by moving fingers on the ball, the pointer can be moved.

Since the whole device is not moved, a track ball requires less space than a mouse. A track ball comes in various shapes like a ball, a button, or a square.

Scanner

Scanner is an input device, which works more like a photocopy machine. It is used when some information is available on paper and it is to be transferred to the hard disk of the computer for further manipulation.

Scanner captures images from the source which are then converted into a digital form that can be stored on the disk. These images can be edited before they are printed.

Digitizer

Digitizer is an input device which converts analog information into digital form. Digitizer can convert a signal from the television or camera into a series of numbers that could be stored in a computer. They can be used by the computer to create a picture of whatever the camera had been pointed at.

Digitizer is also known as Tablet or Graphics Tablet as it converts graphics and pictorial data into binary inputs. A graphic tablet as digitizer is used for fine works of drawing and image manipulation applications.

Microphone

Microphone is an input device to input sound that is then stored in a digital form.

he microphone is used for various applications such as adding sound to a multimedia presentation or for mixing music.

Magnetic Ink Card Reader (MICR)

MICR input device is generally used in banks as there are large number of cheques to be processed every day. The bank's code number and cheque number are printed on the cheques with a special type of ink that contains particles of magnetic material that are machine readable.

This reading process is called Magnetic Ink Character Recognition (MICR). The main advantages of MICR is that it is fast and less error prone.

Optical Character Reader (OCR)

OCR is an input device used to read a printed text.

OCR scans the text optically, character by character, converts them into a machine readable code, and stores the text on the system memory.

Bar Code Readers

Bar Code Reader is a device used for reading bar coded data (data in the form of light and dark lines). Bar coded data is generally used in labelling goods, numbering the books, etc. It may be a handheld scanner or may be embedded in a stationary scanner.
stationary scanner.

Bar Code Reader scans a bar code image, converts it into an alphanumeric value, which is then fed to the computer that the bar code reader is connected to.

Optical Mark Reader (OMR)

OMR is a special type of optical scanner used to recognize the type of mark made by pen or pencil. It is used where one out of a few alternatives is to be selected and marked.

it is specially used for checking the answer sheets of examinations having multiple choice questions.

Computer - Output Devices

Following are some of the important output devices used in a computer.

- Monitors
- Graphic Plotter
- Printer

Monitors

Monitors, commonly called as **Visual Display Unit** (VDU), are the main output device of a computer. It forms images from tiny dots, called pixels that are arranged in a rectangular form. The sharpness of the image depends upon the number of pixels.

There are two kinds of viewing screens used for monitors.

- Cathode-Ray Tube (CRT)
- Flat-Panel Display

Cathode-Ray Tube (CRT) Monitor

The CRT display is made up of small picture elements called pixels. The smaller the pixels, the better the image clarity or resolution. It takes more than one illuminated pixel to form a whole character, such as the letter ‘e’ in the word help.

A finite number of characters can be displayed on a screen at once. The screen can be divided into a series of character boxes - fixed location on the screen where a standard character can be placed. Most screens are capable of displaying 80 characters of data horizontally and 25 lines vertically.

There are some disadvantages of CRT −

- Large in Size
- High power consumption

Flat-Panel Display Monitor

The flat-panel display refers to a class of video devices that have reduced volume, weight and power requirement in comparison to the CRT. You can hang them on walls or wear them on your wrists. Current uses of flat-panel displays include calculators, video games, monitors, laptop computer, and graphics display.

The flat-panel display is divided into two categories −

- **Emissive Displays** − Emissive displays are devices that convert electrical energy into light. For example, plasma panel and LED (Light-Emitting Diodes).
- **Non-Emissive Displays** − Non-emissive displays use optical effects to convert sunlight or light from some other source into graphics patterns. For example, LCD (Liquid-Crystal Device).

Printers

Printer is an output device, which is used to print information on paper.

There are two types of printers −

- Impact Printers
- Non-Impact Printers

Impact Printers

Impact printers print the characters by striking them on the ribbon, which is then pressed on the paper.

Characteristics of Impact Printers are the following −

- Very low consumable costs
- Very noisy
- Useful for bulk printing due to low cost
- There is physical contact with the paper to produce an image

These printers are of two types −

- Character printers
- Line printers

Character Printers

Character printers are the printers which print one character at a time.

These are further divided into two types:

- Dot Matrix Printer(DMP)
- Daisy Wheel

Dot Matrix Printer

In the market, one of the most popular printers is Dot Matrix Printer. These printers are popular because of their ease of printing and economical price. Each character printed is in the form of pattern of dots and head consists of a Matrix of Pins of size (5*7, 7*9, 9*7 or 9*9) which come out to form a character which is why it is called Dot Matrix Printer.

Advantages

- Inexpensive
- Widely Used
- Other language characters can be printed

Disadvantages

- Slow Speed
- Poor Quality

Daisy Wheel

Head is lying on a wheel and pins corresponding to characters are like petals of Daisy (flower) which is why it is called Daisy Wheel Printer. These printers are generally used for word-processing in offices that require a few letters to be sent here and there with very nice quality.

Advantages

- More reliable than DMP
- Better quality
- Fonts of character can be easily changed

Disadvantages

- Slower than DMP
- Noisy
- More expensive than DMP

Line Printers

Line printers are the printers which print one line at a time.

Non-impact Printers

Non-impact printers print the characters without using the ribbon. These printers print a complete page at a time, thus they are also called as Page Printers.

These printers are of two types −

- Laser Printers
- Inkjet Printers

Characteristics of Non-impact Printers

- Faster than impact printers
- They are not noisy

- High quality
- Supports many fonts and different character size

Laser Printers

These are non-impact page printers. They use laser lights to produce the dots needed to form the characters to be printed on a page.

Advantages

- Very high speed
- Very high quality output
- Good graphics quality
- Supports many fonts and different character size

Disadvantages

- Expensive
- Cannot be used to produce multiple copies of a document in a single printing

Inkjet Printers

Inkjet printers are non-impact character printers based on a relatively new technology. They print characters by spraying small drops of ink onto paper. Inkjet printers produce high quality output with presentable features.

They make less noise because no hammering is done and these have many styles of printing modes available. Color printing is also possible. Some models of Inkjet printers can produce multiple copies of printing also.

Advantages

- High quality printing
- More reliable

Disadvantages

- Expensive as the cost per page is high
- Slow as compared to laser printer

Role of computers in bioinformatics

Bioinformatics, a rapidly evolving discipline, is the application of computational tools and techniques to the management and analysis of biological data. Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. It was Paulien Hogeweg who invented the term Bioinformatics in 1979 to study the processes of information technology into biological systems. The science of bioinformatics actually develops algorithms and biological software for computers to analyze and record the data related to biology, for example the data of genes, proteins, drug ingredients and metabolic pathways. The term bioinformatics is relatively new, and as defined here, it encroaches on such terms as "computational biology" and others. The use of computers in biology research predates the term bioinformatics by many years. As biological data is always in raw form and there is a need for a certain storage house in which the data can be stored, organized and manipulated. Biological software and databases provide the scientists this opportunity so that the data can be extracted from these databases easily and can be used by the scientists. However, the applications of computer science in bio-informatics are broadly classified as:

1. **Data mining** Computers are able to control various types of equipment and can pre-program biological experiments, saving a lot of manpower for the implementation of experiments. Data mining is done by cleaning and integrating data and then analyzing and predicting trends. In modern biological research, computer data mining techniques, such as neural networks and support vector machines, are used to predict the secondary structure of the amino acids by coding the 20 amino acids that make up the protein and then selecting a window of odd length and using the constructed neural network model or vector machine model. The data mining can also be used for microarray data analysis, protein homology study, etc.

2. **Structural simulation** Modeling software can be used on the computer to save cost and time for building structural models, and we can directly input conditions and data for the computer to calculate the results, which facilitates the testing of structural hypotheses. This computer 3D modeling makes the structure of genes and proteins more visual, which is very helpful for the study of biological information.
3. **High performance computers in Genomics research** Using genomics, large-scale gene sequencing is done to obtain information about genes and genomes, however, it is now common to use a method consisting of a large number of random short sequences spliced by sequences. The determination of all short sequences is performed by a highly automated sequencing machine, which is responsible for converting the physicochemical signals about sequence information (i.e., the arrangement of base pairs A, T, C, and G in DNA) obtained in biological sequencing into digital information that can be processed by a computer, which is simply analyzed by a general computer and gives preliminary results, equivalent to raw experimental data. After obtaining the raw data for gene sequencing, the process of data processing and analysis involving huge amounts of data follows. All of this places high demands on the high-performance computing environment, and with the increasing maturity of gene sequencing technology, a country's advantage in genomics research is largely dependent on the high-speed computing power it can provide
4. **Development of software for molecular biology tools** software tools are used by biologists in molecular biology research, for, such as sequence splicing software, gene discovery software, gene structure analysis software, and gene function prediction software, which are all essential computational and analytical tools for biologists.

   Currently, homology comparison is the main method for gene function prediction. The common algorithms for homology comparison are Smith-Waterman algorithm, BLAST algorithm and FASTA algorithm.

   Data mining technology is an extremely fast growing research field in computer science at present, which integrates machine learning, statistical analysis and database technology to provide services for decision-oriented use of data in databases. The essence of data mining is knowledge discovery, including rule generation, classification, clustering, sequence analysis, etc. Data mining techniques have great potential for gene finding and gene function prediction software design and development.

   Artificial neural network technology is another branch of computer science that is developing fast. Artificial neural network is an information processing system that simulates the structure and characteristics of human brain neurons and the cognitive function of human brain with highly nonlinear dynamics. ANN technology can be used for bioinformatics also. The authors argue that the key to the use of ANN techniques for bioinformatics research lies in the selection of the appropriate model from among the many types of artificial neural networks currently available, and in making some necessary improvements to the chosen model, according to the characteristics of biology.

Nowadays, computer application technology is developing rapidly, and now the research in the field of artificial intelligence is advancing by leaps and bounds with remarkable results. In the future, it may allow bioinformatics to be combined with machine learning, so that computers can automatically summarize and analyze, reason and summarize, and derive the correct conclusions when analyzing biological data, which will make the research on the interpretation of biological data information twice as fast with half the effort, and the efficiency will be greatly improved. Bioinformatics will be the basis for guiding the direction of treatment in the future. Using computer data models it is possible to design vaccines, disease diagnostics, and drug treatments for the structure of gene molecules or proteins.

Mean, Median, Mode, and Range

Mean, median, and mode are three kinds of "averages". There are many "averages" in statistics, but these are three most common.

The Mean, Median and Mode are the three measures of central tendency. A **measure of central tendency** describes a set of data by identifying the central position in the data set as a single value.

Choosing the best measure of central tendency depends on the type of data we have

Mean is the arithmetic average of a data set. This is found by adding the numbers in a data set and dividing by the number of observations in the data set.

The median is the middle number in a data set when the numbers are listed in either ascending or descending order.

The mode is the value that occurs the most often in a data set and the range is the difference between the highest and lowest values in a data set.

$$\text{Mean} = \frac{\text{Sum of all observations}}{\text{Number of observations}}$$

The Mean

$$\bar{\mathbf{X}} = \frac{\sum \mathbf{X}}{\mathrm{N}}$$

Here,

∑ represents the summation

X represents observations

N represents the number of observations.

Types of Data

Data can be present in **raw form** or **tabular form**.

Let's find the mean in both cases.

1. Raw Data

Let $x_1, x_2, x_3 \ldots\ldots x_n$ be n observations.

We can find the arithmetic mean using the mean formula.

$$\textbf{Mean, } \bar{x} = \frac{x_1 + x_2 + \ldots x_n}{n}$$

Example 1

Find the mean of the following distribution:

If the heights of 5 people are 142 cm, 150 cm, 149 cm, 156 cm and 153 cm.

Find the mean height.

Mean height

$$= \bar{x} = \frac{142 + 150 + 149 + 156 + 153}{5} = \frac{750}{5} = 150$$

$$\boxed{\textbf{Mean, } \bar{x} = 150 \textbf{ cm}}$$

2. Frequency Distribution (Tabular) Form

When the data is present in tabular form, we use the following formula:

$$\textbf{Mean} = \bar{X} = \frac{X1f1 + X2f2 + X3f3 + \ldots\ldots Xnfn}{f1 + f2 + f3 \ldots\ldots fn}$$

Consider the following example.

Example 1

Find the mean of the following distribution:

X	4	6	9	10	15
f	5	10	10	7	8

Solution

Calculation table for arithmetic mean:

xi	fi	xifi
4	5	20
6	10	60
9	10	90
10	7	70
15	8	120
	∑fi = 40	∑xifi = 360

$$\therefore \text{Mean} = \bar{x} = \frac{\sum xifi}{\sum fi} = \frac{360}{40} = 9$$

Example 2

Here is an example where the data is in the form of class intervals.

The following table indicates the data on the number of patients visiting a hospital in a month.

Find the average number of patients visiting the hospital in a day.

Number of patients	**Number of days visiting hospital**
0-10	2
10-20	6
20-30	9
30-40	4
40-50	5
50-60	7

Solution

In this case, we find the class mark (also called as mid-point of a class) for each class.

Note: Class mark $=\frac{\text{lower limit + upper limit}}{2}$

Let x_1, x_2, x_3 ……x_n be the class marks of the respective classes.

Hence, we get the following table

Class mark (x_i)	**frequency (f_i)**	**$x_i f_i$**
5	2	10
15	6	90
25	9	225
35	7	245
45	4	180
55	2	110
Total	**∑fi=30**	**∑fixi=860**

$$\therefore \text{Mean} = \overline{x} = \frac{\sum xifi}{\sum fi} = \frac{860}{30} = 28.67$$

The Median

If the total number of observations (n) is an odd number, then the formula is given below:

$$\textbf{Median} = \frac{(n+1)}{2}^{\text{th}} \textbf{ observation}$$

Example 1

Let's consider the data: 56, 67, 54, 34, 78, 43, 23.

What is the median?

Solution

Arranging in ascending order, we get: 23, 34, 43, 54, 56, 67, 78.

Here, n (no.of observations) = 7

So,

$$\frac{7+1}{2} = 4 \quad \therefore \textbf{Median = 4th observation}$$

If the total number of the observations (n) is an even number, then the formula is given below:

$$\textbf{Median}=\frac{\left(\frac{n}{2}\right)^{th}\textbf{ observation}+\left(\frac{n}{2}+1\right)^{th}\textbf{ observation}}{2}$$

Example 2

Let's consider the data: 50, 67, 24, 34, 78, 43.

What is the median?

Solution

Arranging in ascending order, we get: 24, 34, 43, 50, 67, 78.

Here, n (no.of observations) = 6

$\frac{6}{2} = 3 = 3$

Using the median formula,

$$\text{Median} =\frac{3^{rd}\text{obs.}+4^{th}\text{obs.}}{2}=\frac{43+50}{2}$$

Case 2: Grouped Data

When the data is continuous and in the form of a frequency distribution, the median is found as shown below:

Step 1: Find the median class.

Let n = total number of observations i.e. $\sum$**fi**

Note: Median Class is the class where n/2 lies.

Step 2: Use the following formula to find the median.

$$\text{Median} = \left[\, L+ \frac{\frac{n}{2}-c}{f} \times H \,\right]$$

where,

l=lower limit of median class

c= cumulative frequency of the class preceding the median class

f=frequency of the median class

h=class size

Let's consider the following example to understand this better.

Example 1

Find the median marks for the following distribution:

Classes	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**
Frequency	**2**	**12**	**22**	**8**	**6**

Solution

We need to calculate the cumulative frequencies to find the median.

Calculation table:

Classes	**Number of students**	**Cumulative frequency**

0-10	**2**	**2**
10-20	**12**	**2 + 12 = 14**
20-30	**22**	**14 + 22 = 36**
30-40	**8**	**36 + 8 = 44**
40-50	**6**	**44 + 6 = 50**

N=50

$\frac{N}{2}=\frac{50}{2}=25$

Median Class =20−30

l=20

f=22

c.f=14, before the median group

h=10

Using Median formula:

$$\text{Median} = [\, L + \frac{\frac{n}{2} - c}{f} \times H \,]$$

$$= 20 + \frac{25 - 14}{22} \times 10$$

$$= 20 + \frac{11}{22} \times 10$$

$$= 20 + 5 = 25$$

The Mode

The mode is the most frequently occurring observation or value.

Case 1: Ungrouped Data

For ungrouped data, we just need to identify the observation which occurs maximum times.

Mode = Observation with maximum frequency

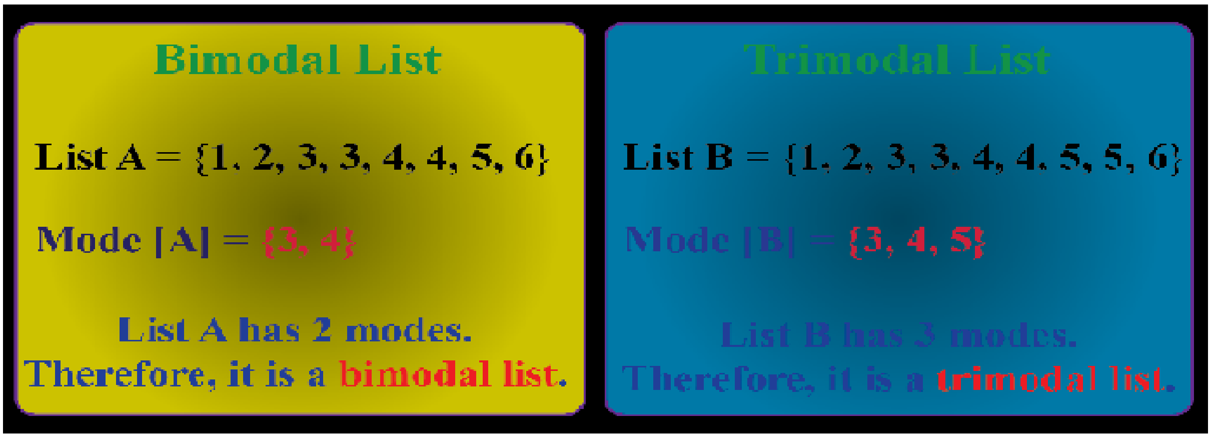
For example in the data: 6, 8, 9, 3, 4, 6, 7, 6, 3 the value 6 appears the most number of times. Thus, mode = 6.

An easy way to remember mode is: **M**ost **O**ften **D**ata **E**ntered.

Note: A data may have no mode, 1 mode or more than 1 mode.

Depending upon the number of modes the data has, it can be called unimodal, bimodal, trimodal or multimodal.

The example discussed above has only 1 mode, so it is unimodal.



Bimodal List

List A = {1, 2, 3, 3, 4, 4, 5, 6}

Mode [A] = {3, 4}

List A has 2 modes.
Therefore, it is a bimodal list.

Trimodal List

List B = {1, 2, 3, 3, 4, 4, 5, 5, 6}

Mode [B] = {3, 4, 5}

List B has 3 modes.
Therefore, it is a trimodal list.

Case 2: Grouped Data

When the data is continuous, the mode can be found using the following steps:

Step 1: Find modal class i.e. the class with maximum frequency.

Step 2: Find mode using the following formula:

$$\text{Mode}=\left[L+\frac{fm - f1}{2fm- f1-f2} \text{x}\, H \right]$$

where, l= lower limit of modal class,

fm= frequency of modal class,

f1= frequency of class preceding modal class,

f2= frequency of class succeeding modal class,

h= class width

Consider the following example to understand the formula.

Example 1

Find the mode of the given data:

Marks Obtained	0-20	20-40	40-60	60-80	80-100
Number of students	5	10	12	6	3

Solution

The highest frequency == 12, so the modal class is 40-60.

l= lower limit of modal class = 40

fm= frequency of modal class =12

f1= frequency of class preceding modal class = 10

f2= frequency of class succeeding modal class = 6

h= class width == 20

Using the mode formula,

$$\text{Mode}=\left[L+\frac{fm - f1}{2fm- f1-f2} \text{x}\, H \right]$$

$$= 40+ \left[\frac{12 - 10}{2\text{x}12\text{-}10\text{-}6}\right] \text{ x } 20$$

$$= 40 + \left[\frac{2}{8}\right] \text{x } 20$$

= 45

Relation Between Mean, Median and Mode

The three measures of central values i.e. mean, median and mode are closely connected by the following relations (called an **empirical relationship**).

2Mean + Mode =3Median

For instance, if we are asked to calculate the mean, median and mode of a continuous grouped data, then we can calculate mean and median using the formulae as discussed in the previous sections and then find mode using the empirical relation.

For example, we have a data whose mode == 65 and median == 61.6.

Then, we can find the mean using the above relation.

2Mean+Mode=3 Median

$\therefore 2\text{Mean}=3\times 61.6-65$

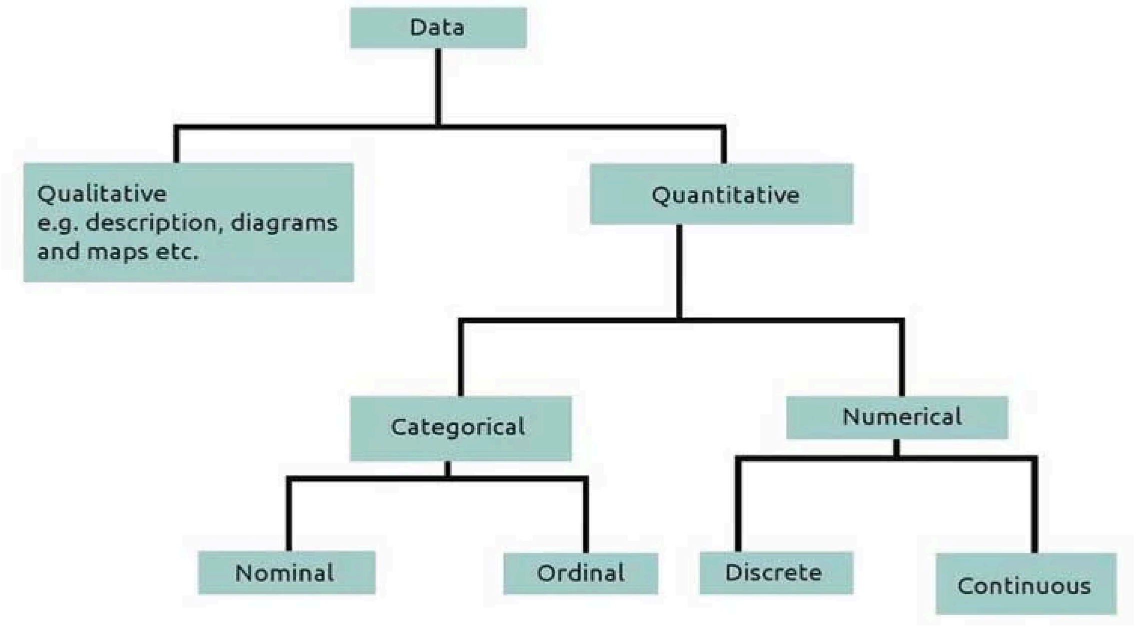
$\therefore 2\text{Mean}=119.8$

$\Rightarrow \text{Mean}=119.8/2 \Rightarrow \text{Mean}=59.9$

DATA AND ITS TYPES

- Data is a set of values of subjects with respect to qualitative or quantitative variables.
- Data is raw, unorganized facts that need to be processed. Data can be something simple and seemingly random and useless until it is organized.
- When data is processed, organized, structured or presented in a given context so as to make it useful, it is called information.
- Information, necessary for research activities are achieved in different forms.
- The main forms of the information available are:

1. Primary data
2. Secondary data
3. Cross-sectional data
4. Categorical data
5. Time series data
6. Spatial data
7. Ordered data



Primary Data

- Primary data is an original and unique data, which is directly collected by the researcher from a source according to his requirements.
- It is the data collected by the investigator himself or herself for a specific purpose.
- Data gathered by finding out first-hand the attitudes of a community towards health services, ascertaining the health needs of a community, evaluating a social program, determining the job satisfaction of the employees of an organization, and ascertaining the quality of service provided by a worker are the examples of primary data.

Secondary Data

- Secondary data refers to the data which has already been collected for a certain purpose and documented somewhere else.
- Data collected by someone else for some other purpose (but being utilized by the investigator for another purpose) is secondary data.
- Gathering information with the use of census data to obtain information on the age-sex structure of a population, the use of hospital records to find out the morbidity and mortality patterns of a community,

Cross-Sectional Data

- Cross-sectional data is a type of data collected by observing many subjects (such as individuals, firms, countries, or regions) at the same point of time, or without regard to differences in time.
- It is the data for a single time point or single space point.

Categorical Data

- Categorical variables represent types of data which may be divided into groups. Examples of categorical variables are race, sex, age group, and educational level.
- The data, which cannot be measured numerically, is called as the categorical data. Categorical data is qualitative in nature.
- The categorical data is also known as attributes.

Example of categorical data: Intelligence, Beauty, Literacy, Unemployment

Time-Series Data

- Time series data occurs wherever the same measurements are recorded on a regular basis.
- Quantities that represent or trace the values taken by a variable over a period such as a month, quarter, or year.
- The values of different phenomenon such as temperature, weight, population, etc. can be recorded over a different period of time.
- The values of the variable remain increasing or decreasing or constant.
- The data according to time periods is called time-series data. e.g. population in a different time period.

Spatial Data

- Also known as geospatial data or geographic information it is the data or information that identifies the geographic location of features and boundaries on Earth, such as natural or constructed features, oceans, and more.
- Spatial data is usually stored as coordinates and topology and is data that can be mapped.

Ordered Data

- Data according to ordered categories is called as ordered data.
- Ordered data is similar to a categorical variable except that there is a clear ordering of the variables.
- For example for category economic status ordered data may be, low, medium and high.

What Is a Sample?

A sample refers to a smaller, manageable version of a larger group. It is a subset containing the characteristics of a larger population. Samples are used in statistical testing when population sizes are too large for the test to include all possible members or observations. A sample should represent the population as a whole and not reflect any bias toward a specific attribute.

Samples are used in a variety of settings where research is conducted. Scientists, marketers, government agencies, economists, and research groups are among those who use samples for their studies and measurements.

Measures of variability: coefficient of variation, variance, standard deviation

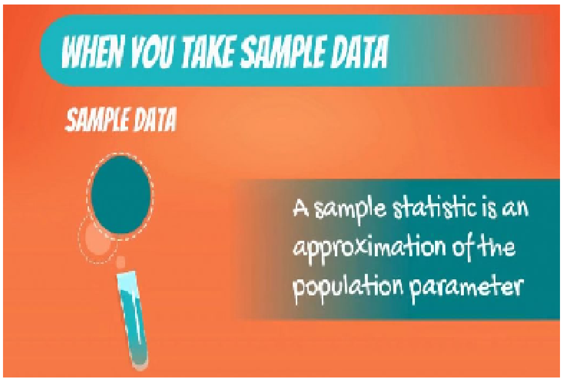
A population is the collection of all items of interest to our study and is usually denoted with an uppercase N. The numbers we've obtained when using a population are called parameters. A sample is a subset of the population and is denoted with a lowercase n, and the numbers we've obtained when working with a sample are called statistics.

In the field of statistics, we typically use different formulas when working with population data and sample data.

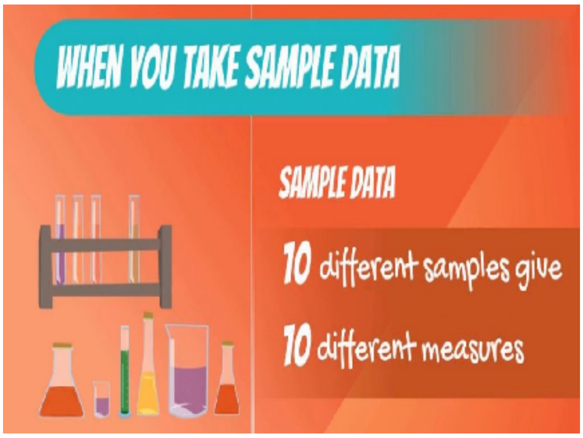
Sample Formulas vs Population Formulas

When we have the whole population, each data point is known to you are 100% sure of the measures we are calculating.

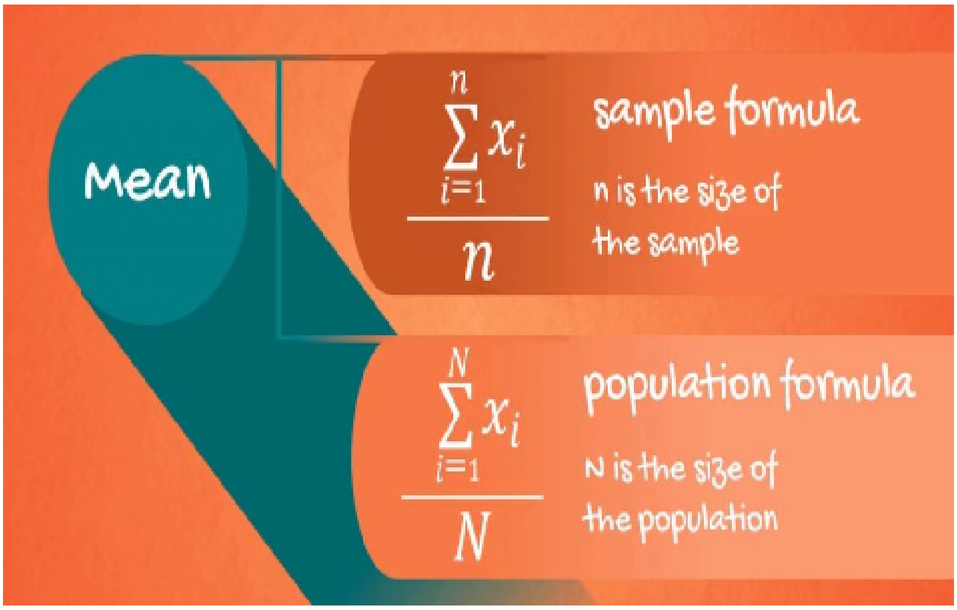
When we take a sample of this population and compute a sample statistic, it is interpreted as an approximation of the population parameter.

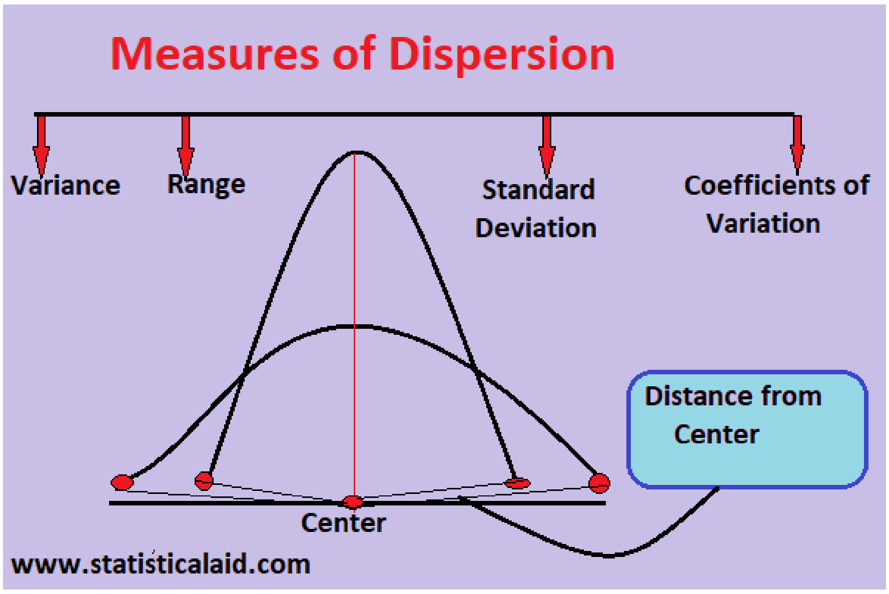

WHEN YOU TAKE SAMPLE DATA
SAMPLE DATA
A sample statistic is an approximation of the population parameter

Moreover, if we extract 10 different samples from the same population, we will get 10 different measures.



The **sample mean** is the average of the sample data points, while the **population mean** is the average of the population data points.

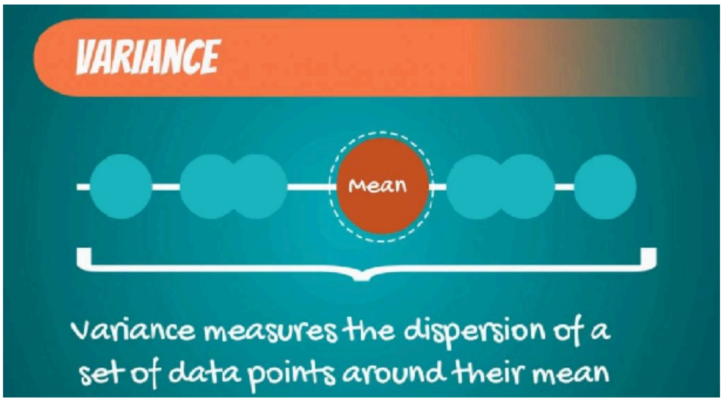

Mean
$\frac{\sum_{i=1}^{n} x_i}{n}$
sample formula
n is the size of the sample
$\frac{\sum_{i=1}^{N} x_i}{N}$
population formula
N is the size of the population



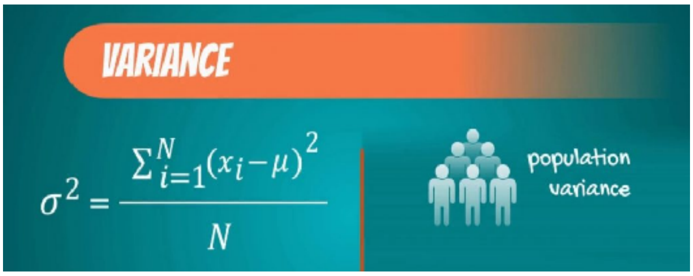
Dispersion is a statistical measure that indicates how the observations are spread out or scattered on each side of the center. If the value of the dispersion is small, it indicates the high uniformity of the observations. The absence of dispersion in the data indicates the perfect uniformity. So, this situation arises when all the observations are identical.

Variance Formula: Sample Variance and Population Variance

Variance measures the dispersion of a set of data points around their **mean** value.


VARIANCE
Mean
variance measures the dispersion of a set of data points around their mean

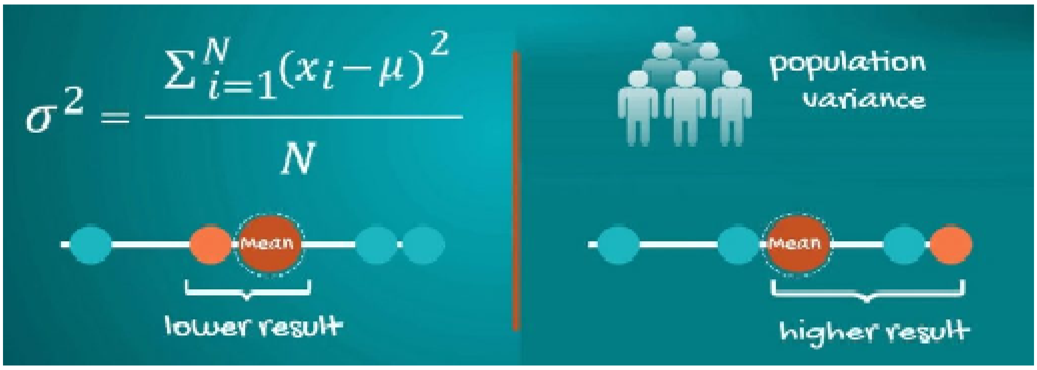
Population variance, denoted by *sigma* squared, is equal to the sum of squared differences between the observed values and the **population mean**, divided by the total number of observations.



$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

The main part of the formula is its *numerator*, so that's what we want to comprehend.

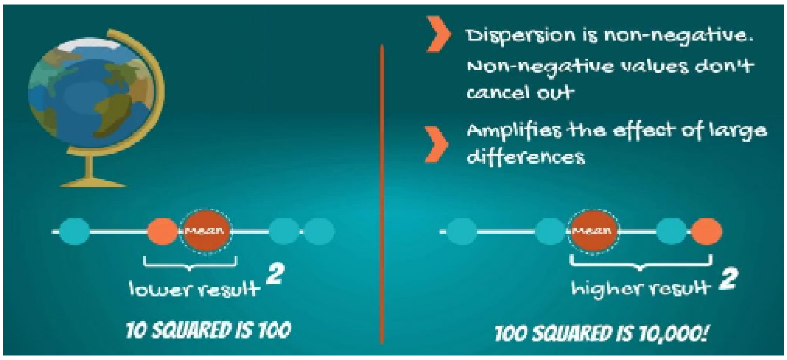
The sum of differences between the observations and the **mean**, squared. So, this means that the closer a number is to the **mean**, the lower the result we obtain will be. And the further away from the **mean** it lies, the larger this difference.



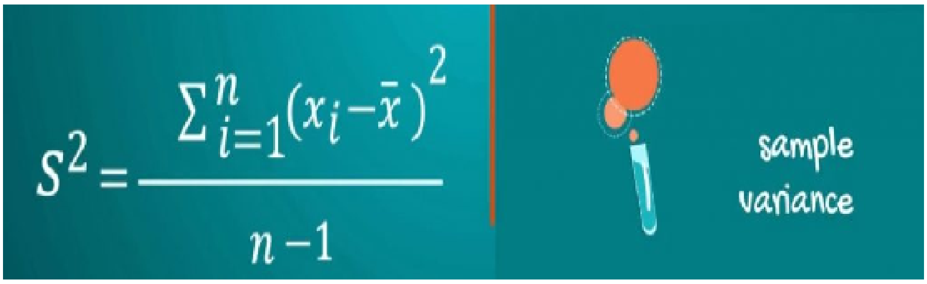
$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

Squaring the differences has two main purposes.

1. First, by squaring the numbers, we always get non-negative computations. Without going too deep into the mathematics of it, it is intuitive that dispersion cannot be negative. Dispersion is about distance and *distance cannot be negative*.
2. Second, squaring amplifies the effect of large differences. For example, if the **mean** is 0 and you have an observation of 100, the squared spread is 10,000!


Dispersion is non-negative.
Non-negative values don't cancel out
Amplifies the effect of large differences
Mean
lower result 2
10 SQUARED IS 100
Mean
higher result 2
100 SQUARED IS 10,000!

Sample variance, on the other hand, is denoted by s squared and is equal to the sum of squared differences between observed **sample** values and the **sample mean**, divided by the number of sample observations minus 1.

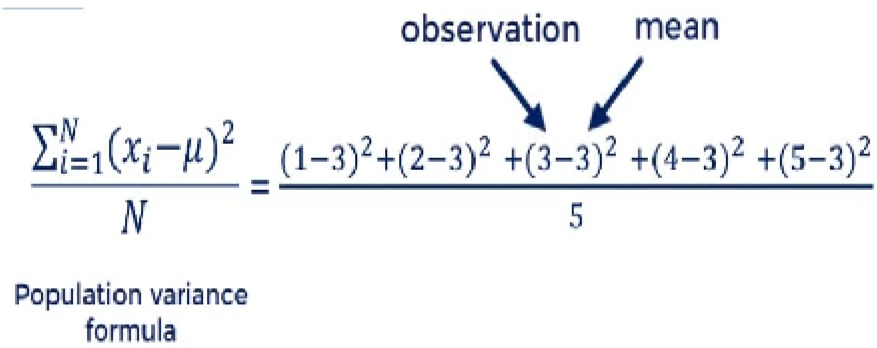


Putting the Population Formula to Use

Alright, enough dry theory. It is time for a practical example. We have a population of five observations – 1, 2, 3, 4 and 5. Let's find its **variance**.

We start by calculating the **mean**: (1 + 2 + 3 + 4 + 5) / 5= 3.

Then we apply the formula which we just discussed: $((1 – 3)^2 + (2 – 3)^2 + (3 – 3)^2 + (4 – 3)^2 + (5 – 3)^2) / 5$.


observation
mean
$\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} = \frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5}$
Population variance formula

the **population variance** of the data set is 2.

Calculating the Sample Variance

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{4}$$

sample variance of 2.5

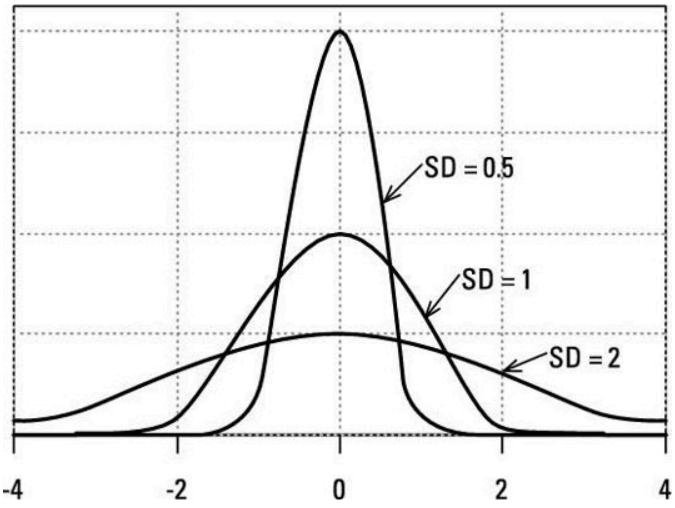
In the first case, we knew the population. That is, we had all the data and we calculated the **variance**. In the second case, we were told that 1, 2, 3, 4 and 5 was a sample, drawn from a bigger population.

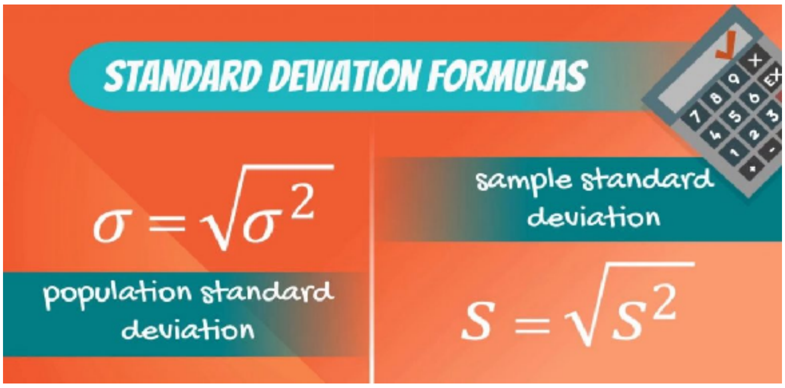
Standard Deviation Formula: Sample Standard Deviation and Population Standard Deviation

While **variance** is a common measure of data dispersion, in most cases the figure you will obtain is pretty large. Moreover, it is hard to compare because the unit of measurement is squared. The easy fix is to calculate its square root and obtain a statistic known as **standard deviation**.

In most analyses, **standard deviation** is much more meaningful than **variance**.

Similar to the **variance** there is also **population** and **sample standard deviation**. The formulas are: the square root of the **population variance** and square root of the **sample variance** respectively.




STANDARD DEVIATION FORMULAS
$\sigma = \sqrt{\sigma^2}$
population standard deviation
sample standard deviation
$s = \sqrt{s^2}$

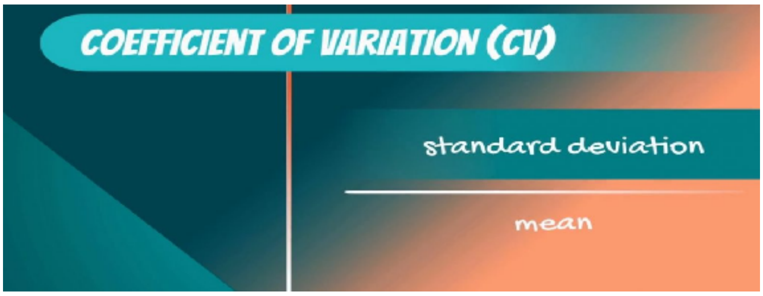
For an IQ example (84, 84, 89, 91, 110, 114, and 116) where the mean is 98.3, you calculate the SD as follows:

$$SD = \sqrt{\frac{(84-98.3)^2 + (84-98.3)^2 + \ldots + (116-98.3)^2}{7-1}} = 14.4$$

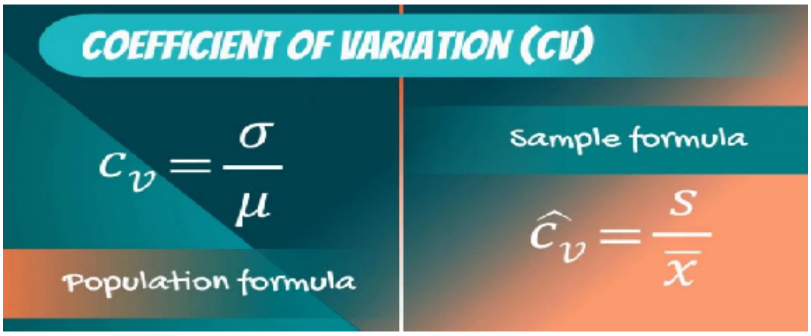
Standard deviations are very sensitive to extreme values (outliers) in the data. For example, if the highest value in the IQ dataset had been 150 instead of 116, the SD would have gone up from 14.4 to 23.9.

The Coefficient of Variation (CV)

The last measure which we will introduce is the **coefficient of variation**. It is equal to the **standard deviation**, divided by the **mean**.

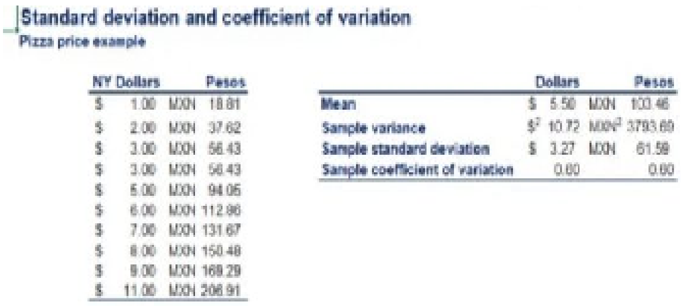


Another name for the term is **relative standard deviation**. This is an easy way to remember its formula – it is simply the **standard deviation** relative to the **mean**.


COEFFICIENT OF VARIATION (CV)
$c_v = \frac{\sigma}{\mu}$
Population formula
Sample formula
$\hat{c}_v = \frac{s}{\bar{x}}$

Why We Need the Coefficient of Variation

So, **standard deviation** is the most common measure of variability for a single data set. But why do we need yet another measure such as the **coefficient of variation**? Well, comparing the **standard deviations** of two different data sets is meaningless, but comparing **coefficients of variation** is not.



Standard deviation and coefficient of variation

Pizza price example

NY Dollars			Pesos
$	1.00	MXN	18.81
$	2.00	MXN	37.62
$	3.00	MXN	56.43
$	3.00	MXN	56.43
$	5.00	MXN	94.05
$	6.00	MXN	112.86
$	7.00	MXN	131.67
$	8.00	MXN	150.48
$	9.00	MXN	169.29
$	11.00	MXN	206.91

		Dollars		Pesos
Mean	$	5.50	MXN	103.46
Sample variance	$^2	10.72	MXN^2	3793.69
Sample standard deviation	$	3.27	MXN	61.59
Sample coefficient of variation		0.60		0.60

$$CV = \frac{s}{\bar{x}}$$

Statistics is a process to convert data into a set of equations that can help us solve problems. Using statistics, we can analyze data in different fields to monitor changing patterns, then use this analysis to draw conclusions and make forecasts.