Definition

And, what is the meaning of sample?

Sample is a part of population being studied or observed.

# Sample and Population in Daily Life

   Suppose we want to know the average weight of Junior High School students in Subang. Considering limited time, labor and cost, we cannot measure all the students. Thus, we should pick a random sample consisting of all students from a number of public or private schools.

   In this case :

i.   all Junior High School students in Subang constitute the *population,*

ii. Students whose heights are measured constitute the *sample.*

The Sampling Process

POPULATION (N)
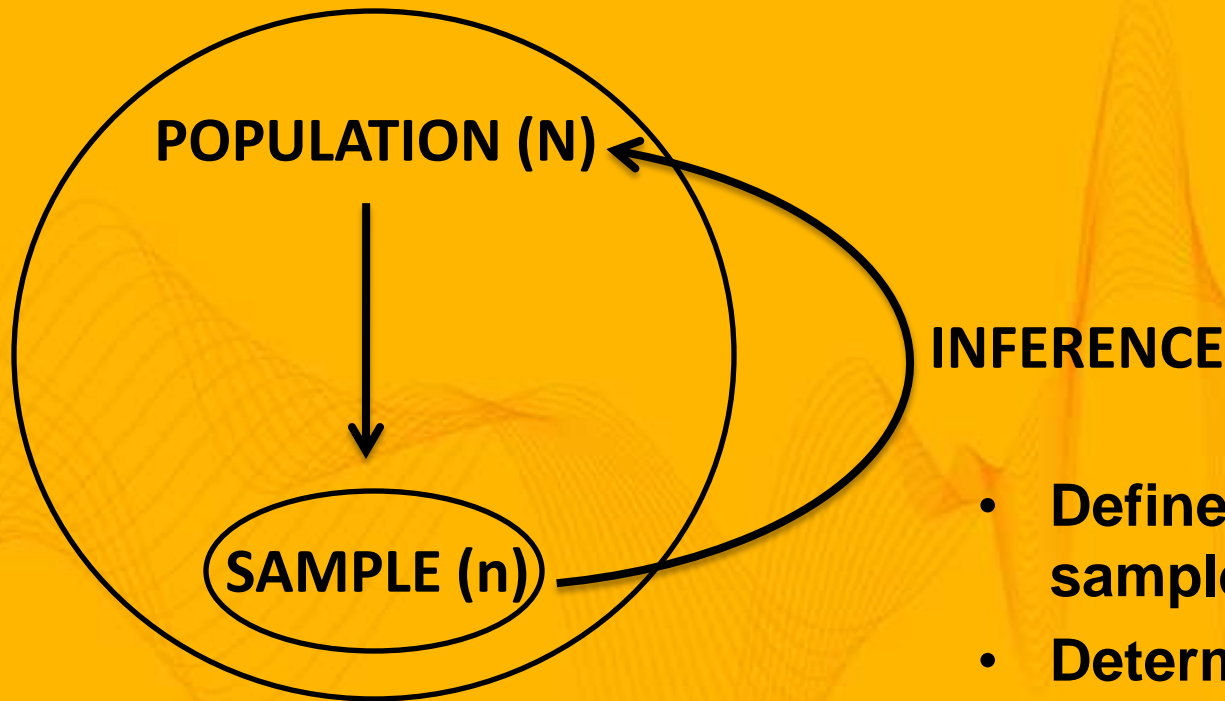
INFERENCE

SAMPLE (n)

- **Define population (N) to be sampled**

- **Determine sample (n) from population (N) to be observed**

- **Determine an inference of population based on sample**

Example 2

A research has been carried out to determine the pollution level in a lake.

  a. What is the population?

  b. How to collect the sample?

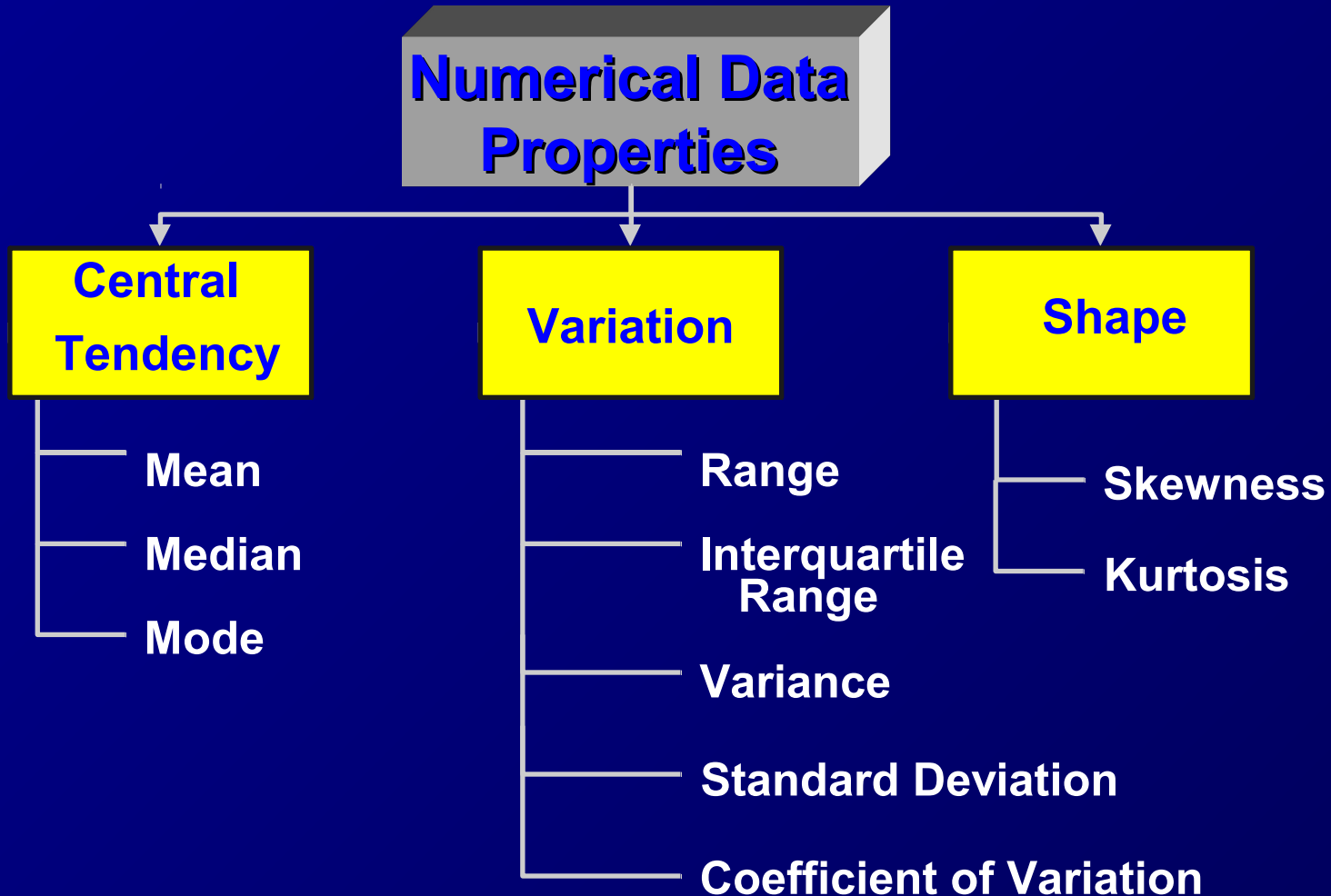*Answer:*

  a. The population is a lake.

  b. A research only observe some quantities of water taken randomly as the sample of the lake (population) to determine the pollution level in a lake.

Descriptive Statistics Numerically Summarizing Data

Central tendency

- **Introduction:**

  Given a set of data, one invariably wishes to find a value about which the observations tend to cluster. The three most common values are the mean, the median, and the mode. They are known as measures of central tendency- the tendency of a set of data to center around certain numerical values.

The Arithmetic Mean))

This is what people usually have in mind when they say

"average"

The Arithmetic Mean

- **May be considered the balance point, in a distribution of observations.**

- **Computed by summing all the observations in the sample and dividing the sum by the number of observations.**

The **sample arithmetic mean**, is computed using sample data.

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum\limits_{i=1}^{n} x_i}{n}$$

The sample mean is a statistic

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- $\overline{x}$ (pronounced "x bar"), representing the sample mean; x_1 is the first and x_i the *i*th in a series of observations.

- The symbol $\sum$ is the Greek letter sigma and denotes "the sum of."

- Thus $\sum_{i=1}^{n}$ indicates that the sum as to begin with *i* = 1 and increment by one up to and including the last observation *n*.

Example

- Consider 7 observations: 4.2, 4.3, 4.7, 4.8, 5.0, 5.1, 9.0.

- By definition

  = (4.2+ 4.3+ 4.7+ 4.8+ 5.0+ 5.1+ 9.0)/7 = 5.3

The population arithmetic Mean

$$\mu = \frac{\sum\limits_{i=1}^{N} x_i}{N} = \frac{\text{Sum of the values of all observations in population}}{\text{Total number of observations in population}}$$

- The symbol for the mean of a population is the Greek letter mu, or μ.
- The population mean is a *parameter*.

Weighted Mean

The weighted mean of a set of numbers   , with corresponding weights

        , is computed from the following formula:

Example: Al-Quds Hospital at Gaza pays its hourly employees $16.50, $19.00, or $25.00 per day. There are 26 daily employees, 14 of which are paid at the $16.50 rate, 10 at the $19.00 rate, and 2 at the $25.00 rate. What is the mean hourly rate paid the 26 employees?

Mean of Grouped Data

In a grouped distribution, we use the middle point   of each interval as *x* value.

Example: find the mean of the age for the following data

Interval (age)	Middle point ( )	Frequency ( )
1-3	2	18
4-6	5	27
7-9	8	34
10-12	11	22
13-15	14	13
Total		**114**

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{(2\times18)+(5\times27)+(8\times34)+(11\times22)+(14\times13)}{18+27+34+22+13} = \frac{867}{114} = 7.61 \quad year$$

Advantages of the mean:

- **It is a measure that can be calculated and is unique.**

- **It is useful for performing statistical procedures such as comparing the means from several data sets.**

Disadvantages of the mean:

- It is affected by extreme values.

$= (4.2+ 4.3+ 4.7+ 4.8+ 5.0+ 5.1+ 9.0)/7 = 5.3$

$= (4.2+ 4.3+ 4.7+ 4.8+ 5.0+ 5.1)/6 = 4.7$

- It would be more representative to calculate the mean without including such an extreme value.

The Median))

- The *median* of a variable is the numerical value that lies in the middle of the data when arranged in ascending order. That is, half the data is below the median and half the data is above the median.

Steps in computing the Median of a data set

1. Arrange the data in ascending order.
2. Determine the number of observation *n.*
3. Determine the observation in the middle of the data set.

   ▮ If the number of observations is odd, then the median is the data value that is exactly in the middle of the data set. That is, it is the observation that lies in the (n +1)/2 position.

▮ **Example**

Find the median of the data set consisting of the observations 7, 4, 3, 5, 6, 8, 10.

Solution: First, we arrange the data set in ascending order

$$3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 10.$$

Since the number of observations is odd, then median = (7+ 1)/2 = 4th number in the ordered list, namely **6**.

Steps in computing the Median of a data set

- If the number of observations is even, then the median is the arithmetic mean of the two middle observations in the data set. That is, it is the arithmetic mean of the data values that lie in the $n/2$ and $(n/2)+1$ position.

- **Example**

Suppose we have the observations 7, 4, 3, 5, 6, 8, 10, 1. Find the median of this data set.

Solution: First, we arrange the data set in ascending order

1 3 4 **5 6** 7 8 10.

Since the number of the observations $n = 8$, then by Definition the median is the average of the 4th ($n/2 = 8/2 = 4$th) and the 5th i.e. Median = $(5+6)/2 = 5.5$

Advantage of the median over the mean:

■ It may be determined even if the values of all observations are not known.

$$3 \quad 4 \quad 5 \quad 6 \quad x_1 \quad x_2 \quad x_3$$

■ Extreme values in data set do not affect the median as strongly as they do the mean.

Example

Consider 5 physicians who practice in Gaza Strip are sampled and asked how much an office visit costs. Suppose we get the answers: 7.5, 7.5, 8.0, 8.0, and 28.0 JD. The mean charge for the sample of five doctors is

$$\bar{x} = \frac{7.5 + 7.5 + 8.0 + 8.0 + 28.0}{5} = \frac{59.0}{5} = \text{JD } 11.8$$

While the median is 8.0. This value is easily seen to be more representative of the values than was the sample mean, JD 11.8 which was affected by the extreme value of 28.0.

Median of grouped data

In a grouped distribution, the following steps are followed:

Step 1: Form the cumulative frequency (F)
Step 2: Find the value of   where
Step 3: Find F value that the first exceeds   , which identifies the median class M.
Step 4: Calculate the median using the following formula

where;

-         lower bound of the median class

-         cumulative frequency of class immediately prior to the median class

-         actual frequency of median class

-         median class width.

Median of grouped data

Example: Estimate the median for the Age in the following data set

Age	20-25	25-30	30-35	35-40	40-45	45-50
frequency	2	14	29	43	33	9

Solution: *Step* 1

Age	(f)	(F)
20-25	2	2
25-30	14	16
30-35	29	45
35-40	43	88
40-45	33	121
45-50	9	130

Step **2:**   =130/2 = 65   *Step* **3:** Median class is 35-40
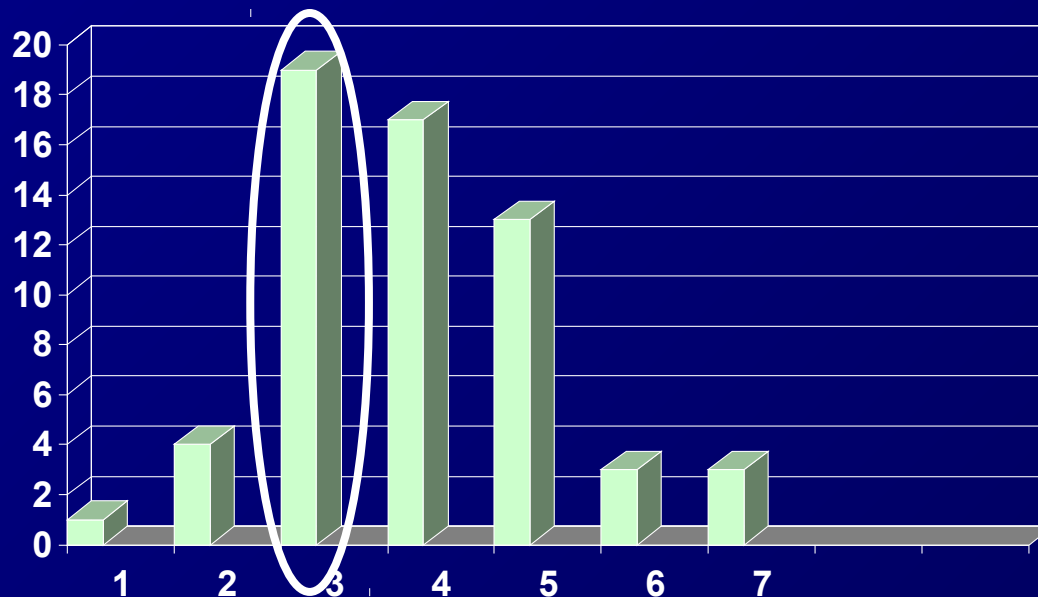
Step **4:**   =35;   =45;   =5.

years

The Mode))

■ The mode is the observation that occurs most frequently. i.e., is repeated most often in the data set.

For a given sample N=16:
33 35 36 37 38 38 38 39 39 39 39 40 40 41 41 45
The mode = 39

■ It corresponds to the highest point on the frequency distribution.

Example

Find the mode of the data set in The Table

Quantity of glucose (mg%) in
blood of 25 students

70	88	95	101	106
79	93	96	101	107
83	93	97	103	108
86	93	97	103	112
87	95	98	106	115

Solution:

■ First we arrange this data set in the ascending order

70	88	95	101	106
79	**93**	96	101	107
83	**93**	97	103	108
86	**93**	97	103	112
87	95	98	106	115

This data set contains 25 numbers. We see that, the value of 93 is repeated most often. Therefore, the mode of the data set is 93.

■ **Multimodal distribution**: A data set may have several modes. In this case it is called multimodal distribution.

■ **Example** The data set has two modes: 1 and 4. This distribution is called bimodal distribution.

0	2	6	9
0	4	6	10
1	4	7	11
1	4	8	11
1	5	9	12
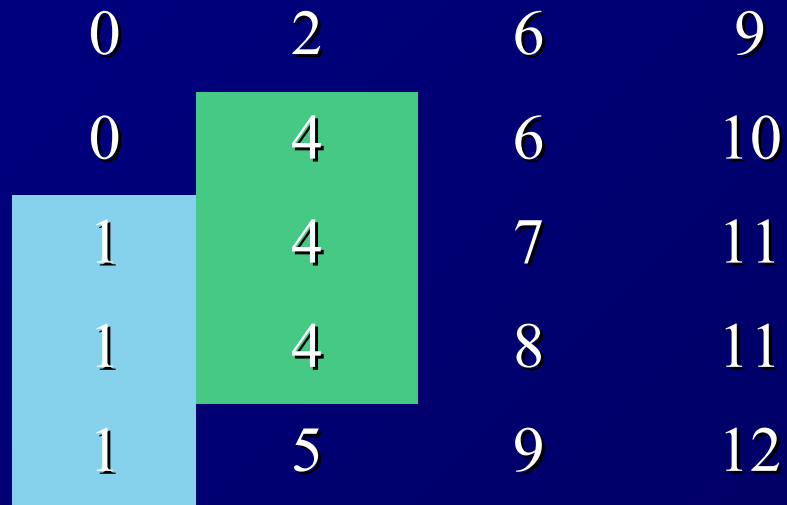
Advantage of the mode

■ Like the median, the mode is **not** affected by extreme values.

For a given sample N=16:

33 35 36 37 38 38 38 39 39 39 39 40 40 41 41

The mode = 39

■ Easily determined for categorical data

Mode of grouped data

In a grouped distribution, the following steps are followed:

Step 1: Determine the model class (class with the largest frequency).

Step 2: Calculate   = Difference between the largest frequency and frequency immediately preceding it.

Step 3: Calculate   = Difference between the largest frequency and the frequency immediately following it.

Step 4: Obtain the mode using the following formula

-     = Lower bound of the modal class
-     = Model class width
-     and   are described in *Step* 2 and *Step* 3.

Mode of grouped data

Example: Estimate the mode for the Age in the following data set

Age	20-25	25-30	30-35	35-40	40-45	45-50
frequency	2	14	29	43	33	9

Solution:
Step 1*:*

Age	Number (*f*)
20-25	2
25-30	14
30-35	29
35-40	43
40-45	33
45-50	9

Step 2: = 43-29=14

Step 3: = 43-33=10

Step 4: =35; = 40-35 =5

years

Disadvantages of the mode:

■ Too often, there is no modal value because the data set contains no values that occur more than once. Other times, every value is the mode because every value occurs the same number of times. Clearly, the mode is a useless measure in these cases.

For a given sample N=16:

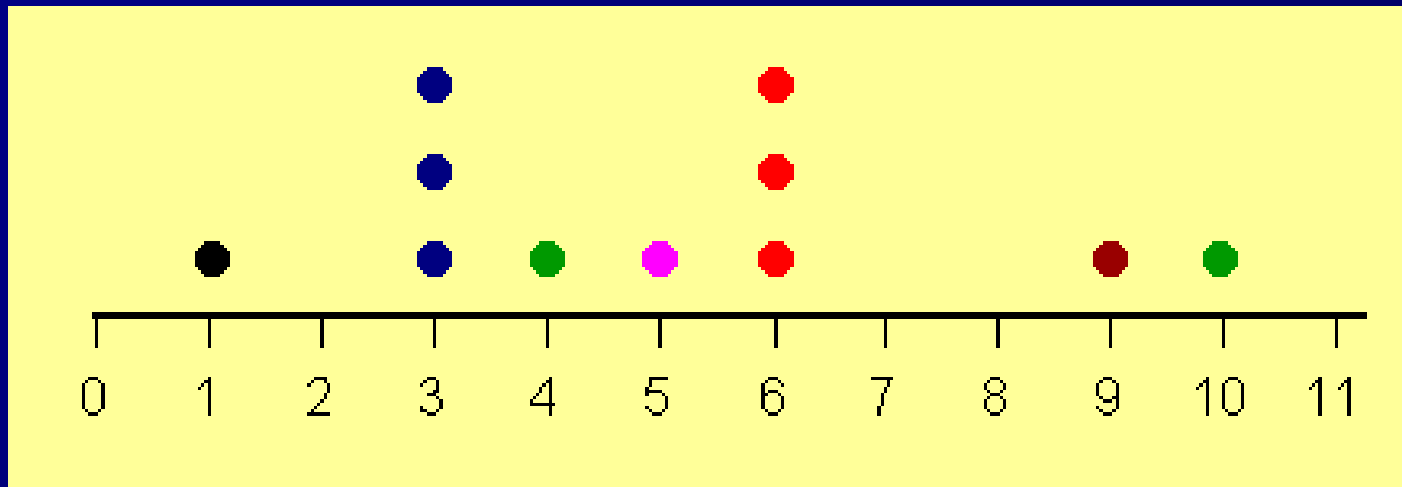33 33 34 34 35 35 36 36 37 37 38 38 39 39 40 40

No unique mode
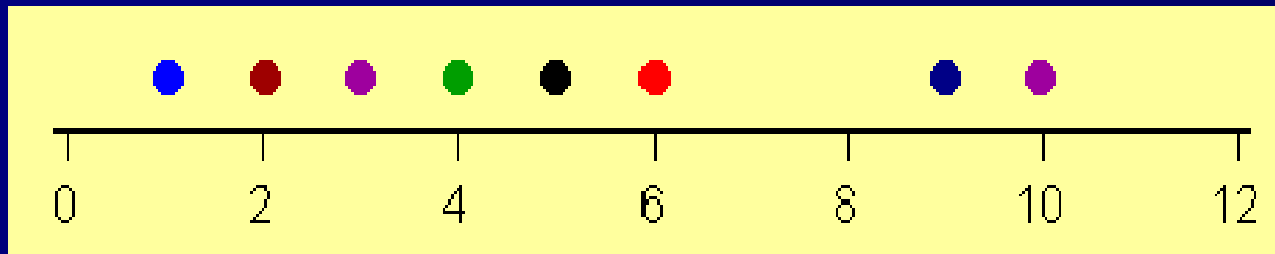
Disadvantages of the mode:

- When data sets contain two, three, or many modes, they are difficult to interpret and compare.

For a given sample N=16:
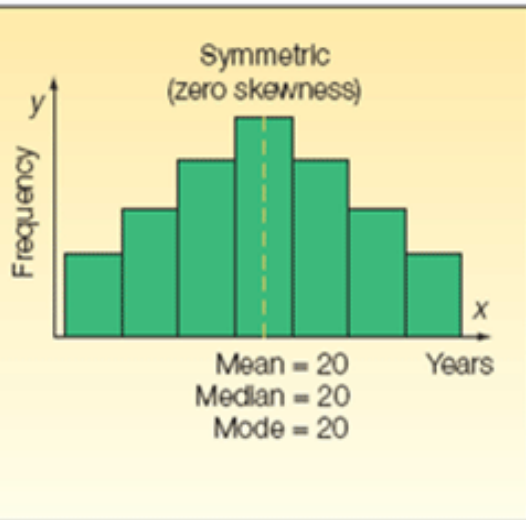
34 34 35 35 35 35 36 37 38 38 39 39 39 39 40 40

The modes = 35 and 39

The Shape of Distributions

- **Distributions can be either <u>symmetrical</u> or <u>skewed</u>, depending on whether there are more frequencies at one end of the distribution than the other.**
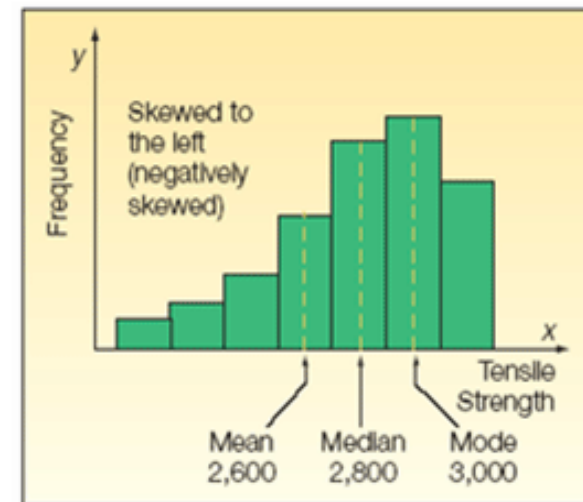
Relative Positions of the Mean, Median and the Mode

Selecting an Appropriate Measure of Central Tendency

■ There are two general criteria for choosing between the measures of central tendency

1. Scale of measurement
   – Nominal scale data, you can only use the Mode
   – Ordinal scale data, you can only use Median or Mode; Median is more informative
   – Interval or ratio scale data, you can use any one of the three.

1. Shape of the distribution
   – Mean is more informative, if you don't have a skewed distribution
   – If you have skewed distribution, you use the median in place of mean.

Measure of Central Tendency	Computation	Interpretation	When to Use
Mean	Population Mean: $\mu = \dfrac{\sum x_i}{N}$ Sample Mean: $\bar{x} = \dfrac{\sum x_i}{n}$	Center of Gravity	When data are quantitative and the frequency distribution is roughly symmetric
Median	Arrange data in ascending order and divide the data set in half	Divides the bottom 50% of the data from top 50% of data	When the data are quantitative and the frequency distribution is skewed left or skewed right
Mode		Most frequent observation	When most frequent observation is desired measure of central tendency or the data are qualitative

Measures of Variation (dispersion)

Measures of Variation (dispersion)

- Just as measures of central tendency locate the "center" of a relative frequency distribution, measures of variation measure its "spread". When the variation is small, this means that the values are close together (but not the same).

To understand Measures of Variation consider the following two examples:

Example 2

Example 1

Think of the difference between an exam with an average mark of 65 in which scores ranged from (62 to 66) and an exam with an average score of 65 in which scores ranged from (30 to 90).

| Night and Day Temperatures (^{o}C) | |
Country A	Country B
22	17
36	40
23	16
35	42
20	20
34	35
Average 28.3	28.3

Two frequency distributions with equal means but different amounts of variation.

Population 1

Population 2

Mean

Measures of variability

- Three statistics to measure variability
  - **Range**
  - **Variance**
  - **Interquartile range**

Range

■ The range is defined as the difference in value between the highest (maxi-mum) and lowest (minimum) observation:

$$\text{Range} = x_{max} - x_{min}$$

■ The range can be computed quickly, but it is **not very useful** since is considers only the extremes and **does not take into consideration the bulk of the observations**.

Variance

- The Variance is a measure which uses the mean as a point of reference.

- The Variance is less when all value are close to the mean while it is more when the values are spread out from the mean.

Population variance

- The **population variance** of the population of the observations *x* is defined the formula

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

σ^2 (sigma squared) = population variance

x_i = the item or observation

μ = population mean

N = total number of observations in the population.

Population variance

The **population variance** of a variable is the sum of squared deviations about the population mean divided by the number of observations in the population, N.

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

That is it is the arithmetic mean of the sum of the squared deviations about the population mean.

The **standard deviation** of a population

- The **standard deviation** of a population is equal to the square root of the variance

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

Since most populations are large, the computation of σ^2 and σ are rarely performed. In practice, the population variance (or standard deviation) is usually estimated by taking a sample from the population and using s^2 and s as a estimate of σ^2 and σ respectively.

The sample variance

■ **The sample variance of the sample of the observations is defined the formula**

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1} \quad \textbf{OR} \quad s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

where:

s^2 =sample variance

$\overline{x}$ = sample mean

n = total number of observations in the sample

Standard deviation of the sample

- The standard deviation of the sample is

$$s = \sqrt{s^2}$$

- It could be also determined from the equations:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad \text{OR} \quad s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}}$$

Remark: In the denominator of the formula for s^2 we use n-1 instead n because statisticians proved that if s^2 is defined as above then s^2 is an unbiased estimate of the variance of the population from which the sample was selected (i.e. the expected value of s^2 is equal to the population variance).

Note: Whenever a statistic consistently overestimates or underestimates a parameter, it is called **biased**. To obtain an **unbiased** estimate of the population variance, we divide the sum of the squared deviations about the mean by n - 1.

Example

A pediatric registrar in a district general hospital is investigating the amount of lead in the urine of children from a nearby housing estate. In a particular street there are 15 children whose ages range from 1 year to under 16, and in a preliminary study the registrar has found the amounts given in the Table below of urinary lead (μmol/24hr),

Urinary concentration of lead in 15 children from housing estate (μmol/24hr)

0.6, 2.6, 0.1, 1.1, 0.4, 2.0, 0.8, 1.3, 1.2, 1.5, 3.2, 1.7, 1.9, 1.9, 2.2

What is the variance and standard deviation?

Note: When using the variance formula, do not round until the last computation. Use as many decimals as allowed by your calculator in order to avoid round off errors.

Solution

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

$$s^2 = \frac{9.96}{14} = 0.7114 \quad (\mu\text{mol/24hr})$$

$$s = \sqrt{s^2}$$

One can apply the following equation as an alternative

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

Calculation of standard deviation				
	(1) Lead concentration x	(2) Differences from mean	(3) Differences squared	(4) Observations in col. (1) squared
	0.1	-1.4	1.96	0.01
	0.4	-1.1	1.21	0.16
	0.6	-0.9	0.81	0.36
	0.8	-0.7	0.49	0.64
	1.1	-0.4	0.16	1.21
	1.2	-0.3	0.09	1.44
	1.3	-0.2	0.04	1.69
	1.5	0	0	2.25
	1.7	0.2	0.04	2.89
	1.9	0.4	0.16	3.61
	1.9	0.4	0.16	3.61
	2.0	0.5	0.25	4.00
	2.2	0.7	0.49	4.84
	2.6	1.1	1.21	6.76
	3.2	1.7	2.89	10.24
Total	=22.5	= 0	=9.96	= 43.71
n= 15, = 1.5				

Coefficient of Variation

- One important application of the mean and the standard deviation is the coefficient of variation. It is defined as the ratio of the standard deviation to the value of the mean, expressed as a percentage.

$$cv = \text{Coefficient of variation} = \frac{\text{Standard deviation}}{\bar{x}} \times 100\%$$

- **Since both standard deviation and the mean are expressed in same units, therefore *cv* is unitless *or* dimensionless.**
- **Therefore, it is possible to use it to compare the relative variation of even unrelated quantities. It also useful in comparing the variability among different variables that vary in magnitude of the values (elephant weight versus mouse weight)**

Example

- Suppose that each day laboratory technician *A* completes 40 analyses with a standard deviation of 5. Technician *B* completes 160 analyses per day with a standard deviation of 15. Which employee shows less variability?

- At first glance, it appears that technician *B* has three times more variation in the output rate than technician *A*. But *B* completes analyses at a rate 4 times faster than *A*. Taking all this information into account, we compute the coefficient of variation for both technicians:

  For technician *A*: *cv*=5/40 x 100% = 12.5%
  For technician *B*: *cv*=15/160 x 100% = 9.4%.

- So, we find that, technician *B* who has more absolute variation in output than technician *A*, has less relative variation.

Means and standard deviations from grouped data

- More often than not, data are presented in *grouped* form. That is, the data are in part summarized and grouped in a frequency table.

Formulas for calculating the mean and the standard deviation for grouped data:

$$\overline{x} = \frac{\sum\limits_{i=1}^{k} f_i x_i}{\sum\limits_{i=1}^{k} f_i}$$

$$s = \sqrt{\frac{\sum\limits_{i=1}^{k} f_i x_i^2 - \frac{\left(\sum\limits_{i=1}^{k} f_i x_i\right)^2}{n}}{n-1}}$$

where $\overline{x}$ = mean of the data set,

s = standard deviation of the data set

x_i = midpoint of the ith class,

f_i = frequency of the ith class,

k = number of classes,

n = total number of observations in the data set.

Example

Given below are the frequency distributions for the heights (in centimeters) of a sample of 100 student in the Islamic University, find the approximate value for the standard deviation for students.

Frequency of heights of a sample of 100 students in the Islamic University

Class interval	x_i	x_i^2	fi	fx_i	fx_i^2
150-154	152	23,104	9	1,368	207,936
155-159	157	24,649	22	3,454	542,278
160-164	162	26,244	31	5,022	813,564
165-169	167	27,889	24	4,008	669,336
170-174	172	29,584	13	2,236	384,592
175-179	177	31,329	1	177	31,329
Total			100	16,265	2,649,035

Class interval	x_i	x_i^2	f_i	fx_i	fx_i^2
150-154	152	23,104	9	1,368	207,936
155-159	157	24,649	22	3,454	542,278
160-164	162	26,244	31	5,022	813,564
165-169	167	27,889	24	4,008	669,336
170-174	172	29,584	13	2,236	384,592
175-179	177	31,329	1	177	31,329
Total			100	16,265	2,649,035

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{16,265}{100} = 162.65 cm$$

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i x_i^2 - \frac{\left(\sum_{i=1}^{k} f_i x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{2,649,035 - 2,645,502.25}{99}}$$

$$= \sqrt{35.68}$$

Note that there is some difference between results from computations ungrouped and grouped data. The size of the discrepancy depends on width of the class interval and on the number of observations within an interval. With short class intervals and large samples, the discrepancy is negligible.

MEASURES OF POSITION: Percentiles, Deciles, and Quartiles

- In cases where our data distribution are heavily skewed or even bimodal, we often get a better summary of the distribution by utilizing relative position of data rather than exact value.

- *Measures of position* are used to describe the location of a particular observation in relation to the rest of the data set.

- Recall that the median is an average computed by using relative position of the data. If we are told that 71 is the median score on a biology test, we know that after the data have been ordered, 50% of the data fall at or below the median value of 71. The median is an example of a *percentile*; in fact, it is the 50th percentile. The general definition of the Pth percentile follows.

Percentiles

■ **Percentiles** are values that divide the ranked data set into 100 equal parts. These values, denoted by *P1*, *P2*, …., *P99*, are such that 1% of the data falls below *P1*, 2% falls below *P2*, …., and 99% falls below *P99*.

	1%	1%	1%	1%	1%		1%	1%	1%
Lowest	1st	2nd	3rd	4th	5th		98th	99th	Highest
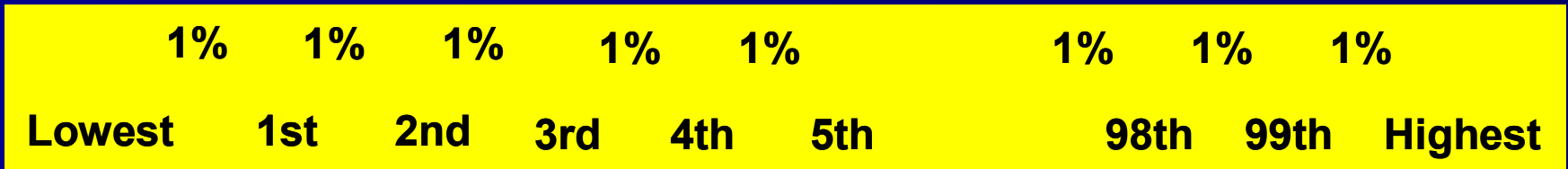
Deciles

■ **Deciles** are values that divide the ranked data set into 10 equal parts. These values, denoted $D1$, $D2$, ..., $D9$, are such that 10% of the data falls below $D1$, 20% falls below $D2$, ..., and 90% falls below $D9$.

10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
Lowest 1st	2nd	3rd	4th	5th	6th	7th	8th	9th	Highest

Quartiles

- **Quartiles** are values that divide the ranked data set into 4 equal parts. These values are denoted by $Q1$, $Q2$, and $Q3$ are such that 25% of the data falls below $Q1$, 50% falls below $Q2$, and 75% falls below $Q3$.

	25%	25%	25%	25%	
Lowest		Q1	Q2	Q3	Highest
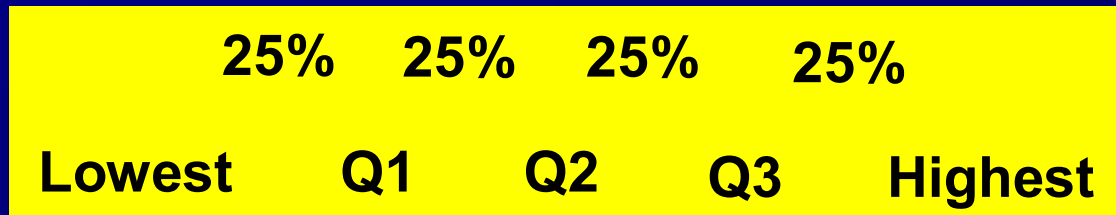
Percentiles, deciles and quartiles

■ All the quartiles and deciles are percentiles. For example, the 7th decile is the 70th percentile and the 1st quartile is the 25th percentile. Consequently, deciles and quartiles are often stated as percentiles.

25%	25%	25%	25%
Lowest	Q1	Q2	Q3	Highest

Median
50th percentile
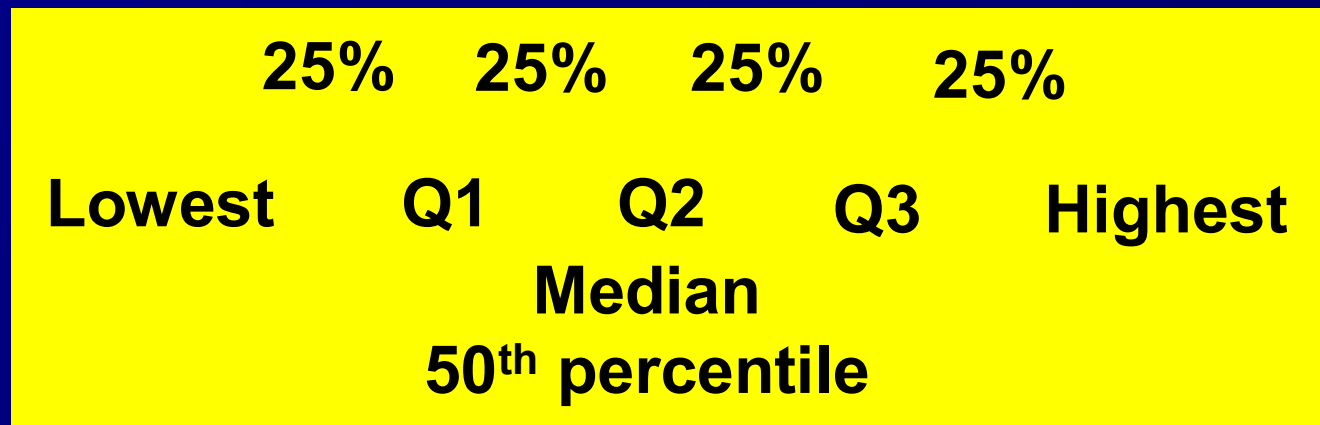
The 50th percentile, 5th decile, and 2nd quartile of a distribution are all the same and correspond to the median

Finding the Percentile that Corresponds to a Data Value

Step 1: Arrange the data in ascending order.

Step 2: Use the following formula to determine the percentile of the score, *x*:

$$\text{Percentile of x} = \frac{\text{Number of data values less than x}}{\text{Total number of values}} \cdot 100$$

This percent is then rounded to the nearest whole number (integer) to give the percentile for observation *x*.

Finding the Percentile that Corresponds to a Data Value (5.5 and 10)

The table contains the ranked aortic diameters measured in centimeters for **45** patients. Notice that the data in the Table are already ranked. Raw data need to be ranked prior to finding measures of position.

3.0	5.0	6.2	7.6	9.4
3.3	5.2	6.3	7.6	9.5
3.5	**5.5**	6.4	7.7	9.5
3.5	5.5	6.6	7.8	**10.0**
3.6	5.5	6.6	7.8	10.5
4.0	5.8	6.8	8.5	10.8
4.0	5.8	6.8	8.5	10.9
4.2	5.9	6.8	8.8	11.0
4.6	6.0	7.0	8.8	11.0

Example 1 The number of observations less than 5.5 is 11 .

$$\frac{11}{45}.100 = 24.4\%$$

This percent rounds to 24. The diameter 5.5 is the 24th percentile and we express this as P24 = 5.5.

Example 2 The number of observations less than 10.0 is 39. Thus $\frac{39}{45}.100 = 86.7\% \approx 87$ **we write P87 = 10.0**

Computing the *p*th Percentile

The *pth percentile* of a data set is a value such that *p* percent of the observations less than this value and (100 - p) percent of the observations are more than this value .

The *pth percentile* for a ranked data set consisting of *n* observations is found by a two-step procedure.

- The first step is to compute index $i = \dfrac{(p)(n)}{100}$.

- If *i* is not an integer, , round up to the next highest integer. Locate the *i*th value of the data set written in ascending order. This number represents the *p*th percentile.

- If *i* is an integer, the pth percentile is the average of the observations in positions *i* and *i* + 1 in the ranked data set.

EXAMPLE

To find the tenth percentile for the data of the Table,

compute $i = \dfrac{(10)(45)}{100} = 4.5$.

The next integer greater than 4.5 is 5. The observation in the fifth position in the Table is 3.6.

Therefore, P10 = 3.6.

Note that at least 10% of the data in the Table are 3.6 or less (the actual amount is 11.1%) and at least 90% of the data are 3.6 or more (the actual amount is 91.1%).

For very large data sets, the percentage of observations equal to or less than P10 will be very close to 10% and the percentage of observations equal to or greater than P10 will be very close to 90%.

3.0	5.0	6.2	7.6	9.4
3.3	5.2	6.3	7.6	9.5
3.5	5.5	6.4	7.7	9.5
3.5	5.5	6.6	7.8	10.0
3.6	5.5	6.6	7.8	10.5
4.0	5.8	6.8	8.5	10.8
4.0	5.8	6.8	8.5	10.9
4.2	5.9	6.8	8.8	11.0
4.6	6.0	7.0	8.8	11.0

EXAMPLE

To find the fortieth percentile for the data in the Table,

compute $i = \dfrac{(40)(45)}{100}$ = 18.

The fortieth percentile is the average of the observations in the 18th and 19th positions in the ranked data set.

The observation in the 18th position is 6.0 and the observation in the 19th position is 6.2.

Therefore P40 $= \dfrac{(6.0)(6.2)}{2}$ = 6.1.

Note that 40% of the data in the Table are 6.1 or less and that 60% of the observations are 6.1 or more.

3.0	5.0	6.2	7.6	9.4
3.3	5.2	6.3	7.6	9.5
3.5	5.5	6.4	7.7	9.5
3.5	5.5	6.6	7.8	10.0
3.6	5.5	6.6	7.8	10.5
4.0	5.8	6.8	8.5	10.8
4.0	5.8	6.8	8.5	10.9
4.2	5.9	6.8	8.8	11.0
4.6	6.0	7.0	8.8	11.0

Procedure to compute quartiles

- Order the data from smallest to largest.
- Find the median. This is the second quartile.
- The first quartile *Q1* is then the median of the lower half of the data; that is, it is the median of the data falling below the *Q2* position (and not including *Q2*).
- The third quartile *Q3* is the median of the upper half of the data; that is, it is the median of the data falling above the *Q3* position (and not including *Q2*).

Lower Quartile (*Q1*)   Upper Quartile (*Q3*)

6, 8, 2, 7, 4, 5, 8

Lower half   median (*Q2*)   Upper half
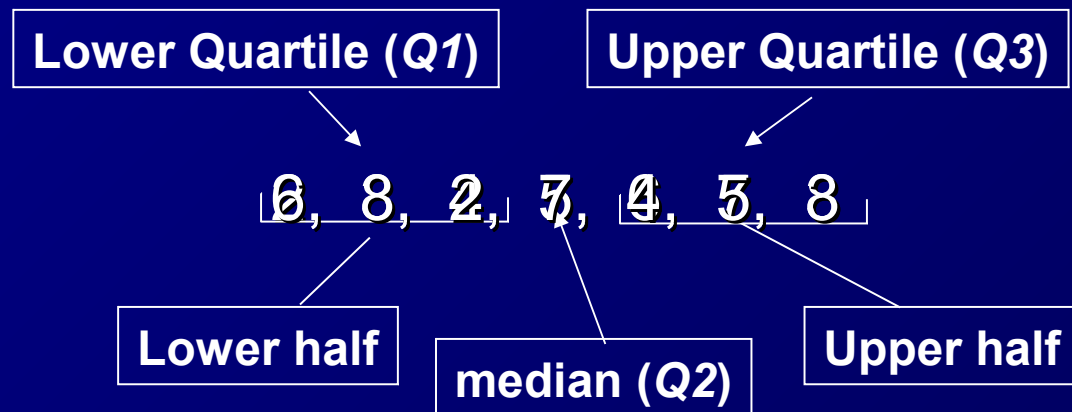
Example 2 ….Even number

- Find the median, and upper and lower quartiles of this set: 22, 19, 27, 32, 38, 25, 32, 26

- First step, order the data:

  19, 22, 25, 26, 27, 32, 32, 38

- So, there are eight numbers, the median is the average of the fourth and fifth numbers.

Median = (26+27)/2 = 26.5

- The lower quartile is the median of the first four numbers,

Lower Quartile = (22+25)/2 = 23.5

- and the upper quartile is the median of the last four numbers.

Upper Quartile = (32+32)/2 = 32

Interquartile Range (*IQR*)

- The interquartile range tells us the spread of the middle half of the data.

Interquartile range =Upper Quartile - Lower Quartile

- Or,

$$IQR = Q3 - Q1$$

25%	25%	25%	25%

Q1   Q2   Q3
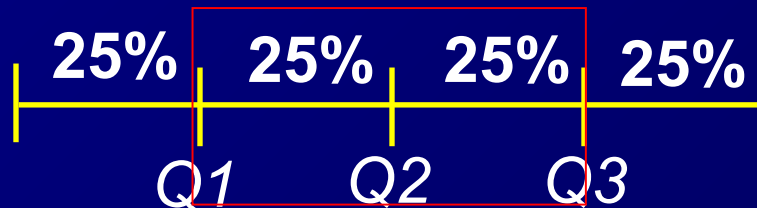
Outliers

■ An **outlier** is a number that is so far above the data set or below most of the data set as to be *considered abnormal and therefore of questionable accuracy*.

Outliers may be from
■ data collection errors,
■ data entry errors,
■ or simply valid but unusual data values.

Regardless of the reason, it is important to identify the outliers in the data set and examine outliers carefully to determine if they are an error.

■ An outlier is **defined** to be any data point that is 1.5 *IQR*s below the lower quartile or above the upper quartile.

Outliers

Example

28, 55, 57, 58, 61, 61, 63, 65, 83

- *UQ* = (65+63)/2 = 64
- *LQ* = (55+57)/2 = 56
- *IQR* = 64 – 56 = 8
- So any number below LQ – 1.5(*IQR*) = 56 – 1.5(8) = 44 or any number above *UQ* + 1.5(*IQR*) = 64 + 1.5(8) = 78 is an outlier.
- Therefore the outliers of this data set are 28 & 83.

Box-and –Whisker Plots

- The quartiles together with the low and high data values give us a very useful five number summary of the data and their spread. These Five-number summary include;

Lowest value, *Q1*, median, *Q3*, and highest value.

- These five numbers can be used to create sketch of the data called a *box-and-Whisker plot*. Box-and-Whisker plots provide another useful technique for describing data.

Q1   Q2   Q3

(lowest value)   (median)   (highest value)

To make Box-and-Whisker plot

1. Draw a vertical scale to include the lowest and highest data values.
2. To the right of the scale draw a box from *Q1* to *Q3*.
3. Include a solid line through the box at the median level.
4. Draw solid lines, called **whiskers**, from *Q1* to the lowest value and from *Q3* to the highest value.
5. Any outliers are marked with an asterisk (*).

Highest value

Q_3

Median

Q_1

Lowest value
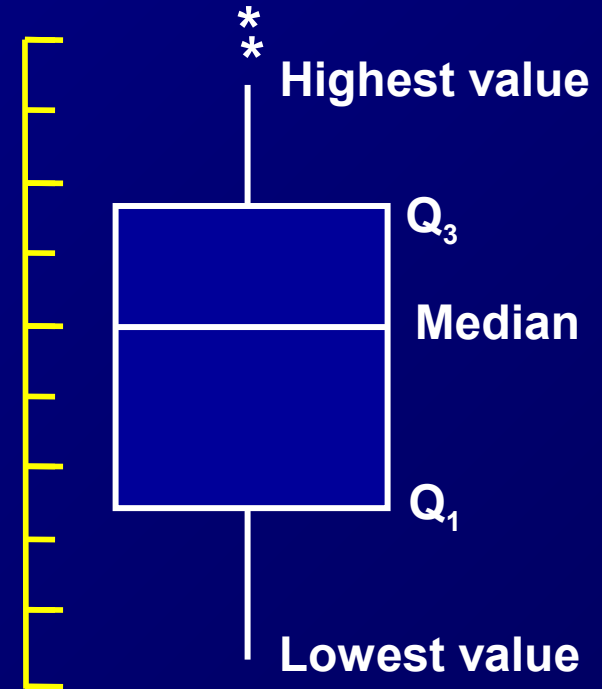
Construct a
Box-and-Whisker Plot:

12	15	16	16	17	18	22	22
23	24	25	30	32	33	33	34
41	45	51					

median = 24
lower quartile = 17
upper quartile = 33
minimum value = 12
maximum value = 51
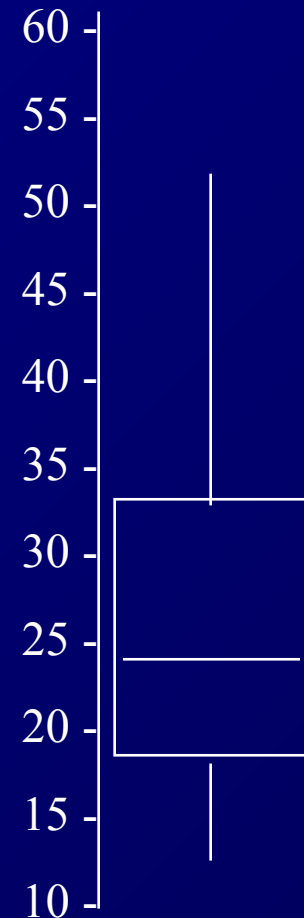
Distribution Shape Based Upon Boxplot

1. Symmetric

If the median is near the center of the box and each of the horizontal lines are approximately equal length, then the distribution is roughly symmetric.
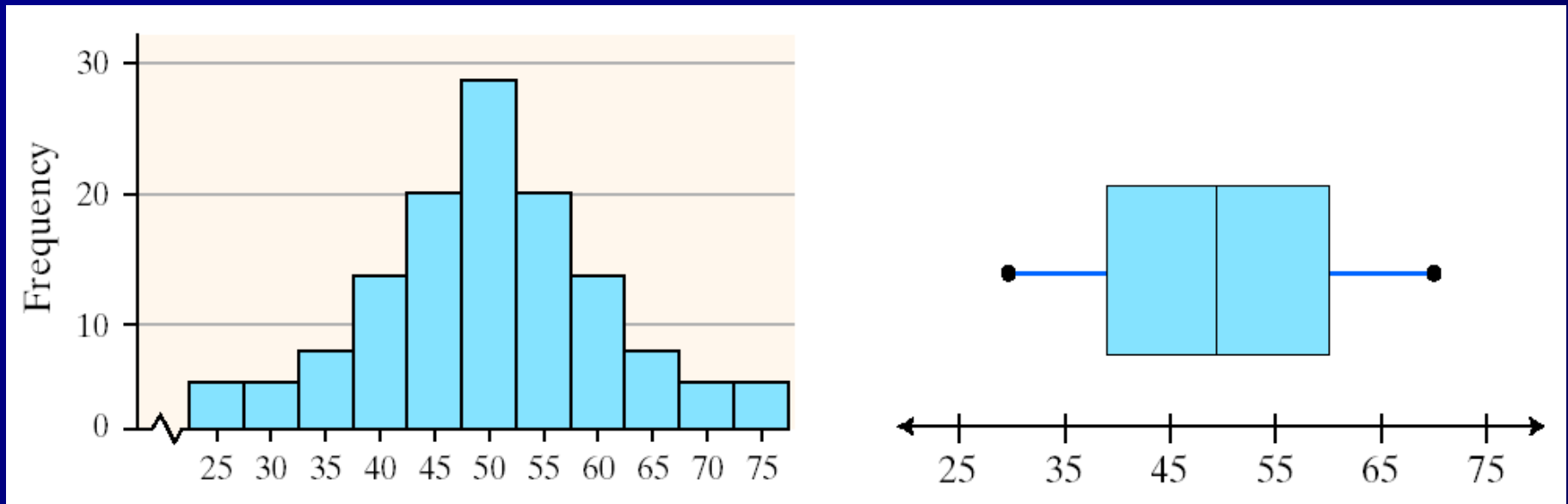
2. Skewed Right

If the median is left of the center of the box and/or the right line is substantially longer than the left line, the distribution is right skewed.
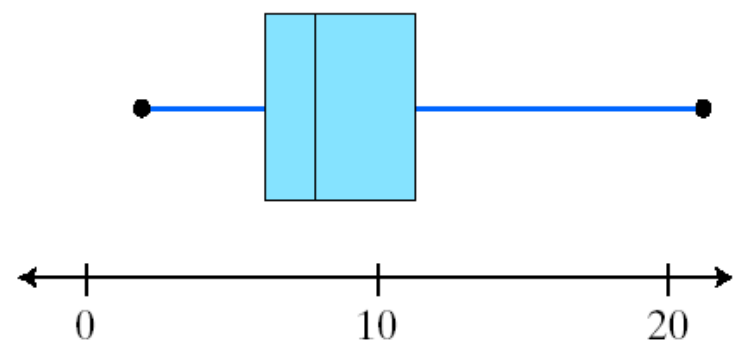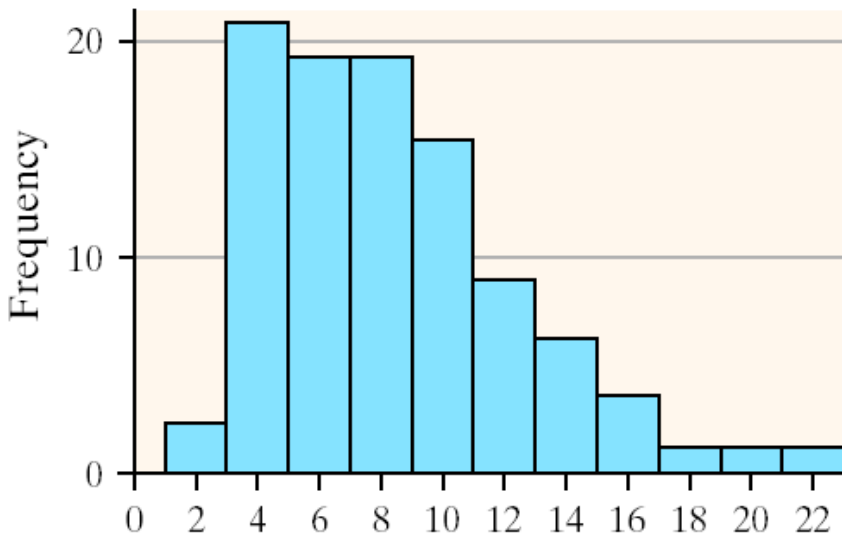
3. Skewed Left

If the median is right of the center of the box and/or the left line is substantially longer than the right line, the distribution is left skewed
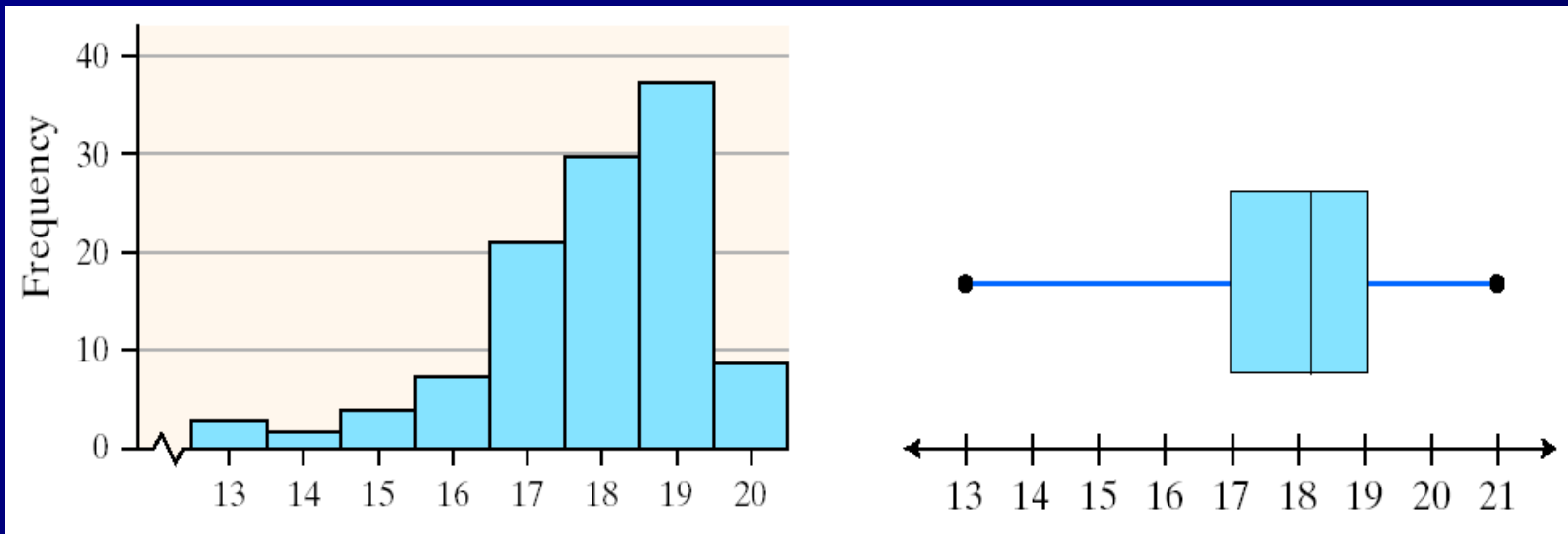
Meaning Of Graphical Representation Of Data

- A picture is said to be more effective than words for describing a particular thing.

- A graphic representation is the geometrical image of a set of data .

- It is a mathematical picture.

- It enables us to think about a statistical problem in visual terms.

- It is an effective and economic device for the presentation , understanding and interpretation of the collected data.

IMPOTANCE OF GRAPHICAL REPRESENTATION

- **It is used to make the data understandable to common man.**

- **It helps in easy and quick understanding of data.**

- **Data displayed by graphical representation can be memorised for a long time.**

- **Can be compared at a glance.**

TYPES OF GRAPHICAL REPRESENTATION

Ungrouped Data	Grouped Data
Line Graph	Histogram
Bar Graph	Frequency Polygon
Pie Diagram Or Circle Graph	Frequency Curve

Line graph:

line graphs are simple Mathematical graphs that are drawn on the graph paper by plotting the data connecting one variable on the horizontal X- axis and other variable of data on the vertical Y-axis.

People in a Store
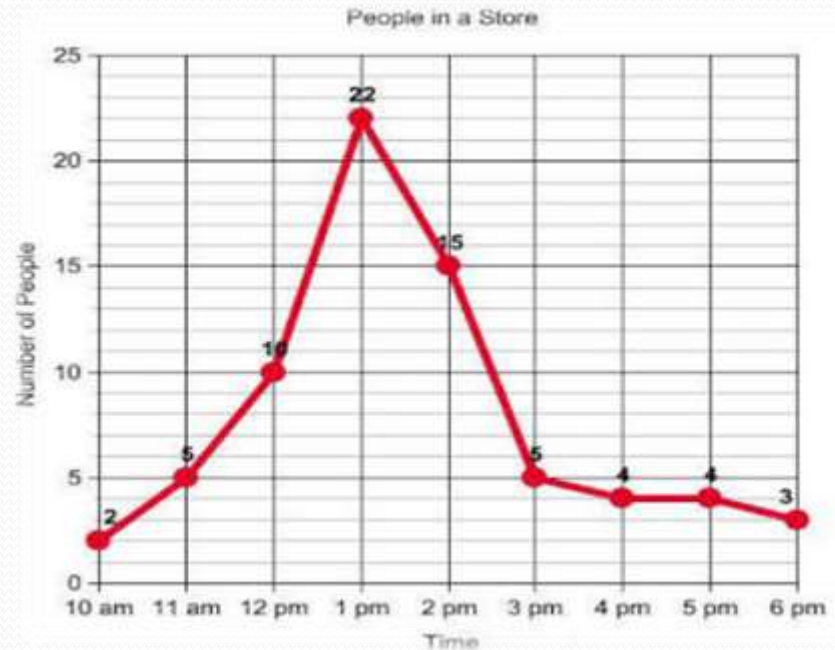
EXAMPLE:

Time	10 am	11 am	12 pm	1 pm	2 pm	3 pm	4 pm	5 pm	6 pm
No of People	2	6	10	22	15	5	4	4	3

Bar graph:

➢ In bar graphs data is represented by bars.

➢ The bars can be made in any direction i.e. vertical or horizontal.

➢ The bars are taken of equal weight and start from a common horizontal or vertical line and their length indicates the corresponding values of statistical data.

When do we use bar diagram ?

- When the data are given in whole numbers.

- When the data are to be compared easily.

How To Make A Bar Graph ?

Months	Jan	Feb	Mar	Apr	May	June	Jul	Aug
No. of buses manufactured	600	800	1000	1200	1400	1600	1800	1800
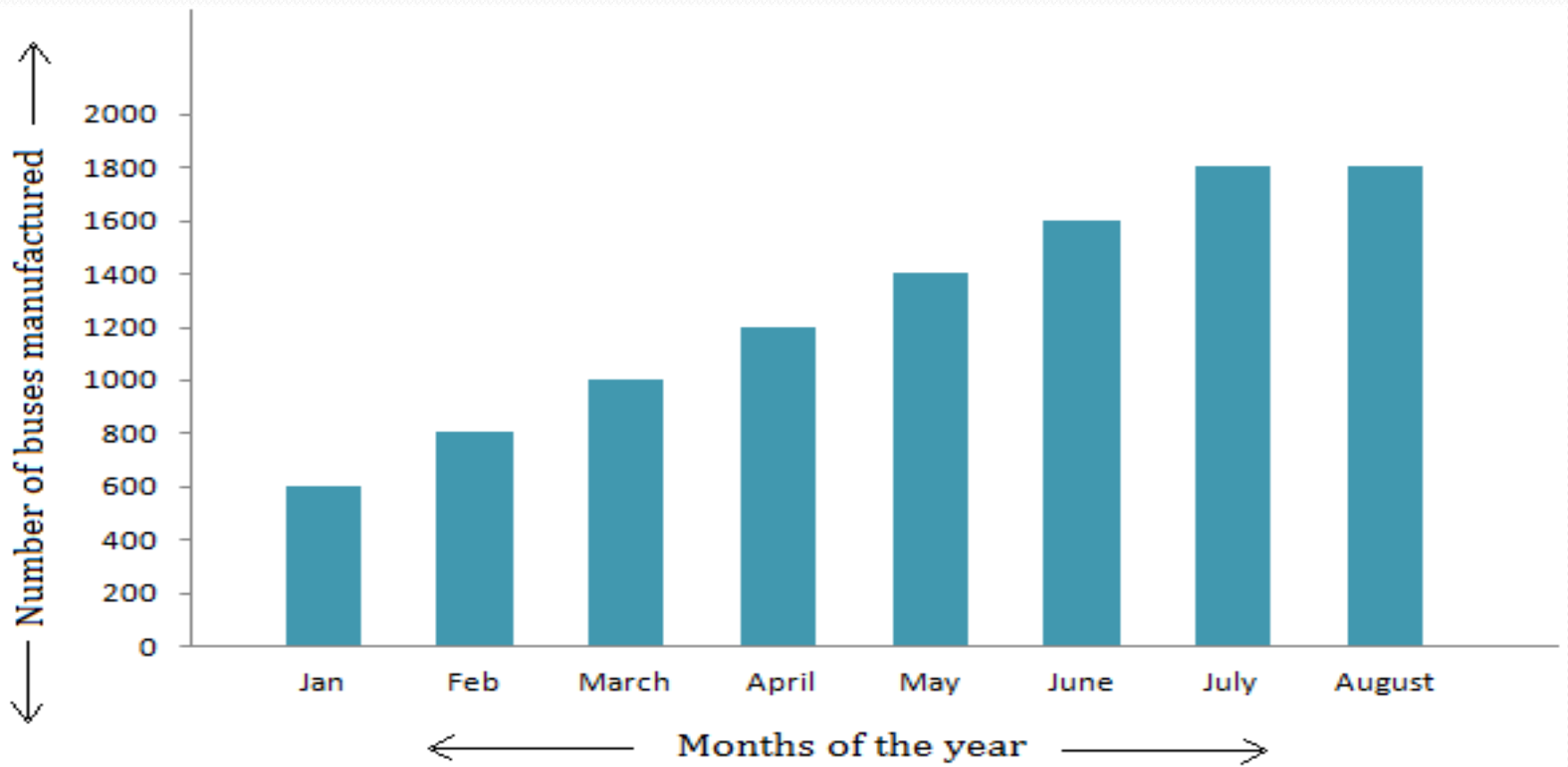
Pie diagram:

➢ It is a circle in which different components are represented through the sections or portions of a circle.

➢ To draw a pie diagram, first the value of each category is expressed as a percentage of the total and then the angle 360^0 is divided in the same percentages.

➢ Then at the centre of a circle these angle are drawn simultaneously starting from a particular radius.

➢ In this way we get a set of sectorial areas proportional to the values of the items.

When do we use pie diagram?

- When the data are given in percentage.

- When different aspect of a variable are to be displayed.

- When the data are to be compared normally.

HOW TO MAKE A PIE DIAGRAM ?

EXAMPLE:

(Table: the result of class 10 of a school)

Marks Division	First	Second	Third	Failures
% of student	20%	56%	20%	4%

Marks Division	% of student	Approx . Angle in degree
First	20%	$\left(\dfrac{360}{100} \times 20\right)^0 = 72^0$
Second	56%	$\left(\dfrac{360}{100} \times 56\right)^0 = 201.6^0$
Third	20%	$\left(\dfrac{360}{100} \times 20\right)^0 = 72^0$
Failures	4%	$\left(\dfrac{360}{100} \times 4\right)^0 = 14.4^0$

■ second div.   ■ first div.

■ failure   ■ third div.
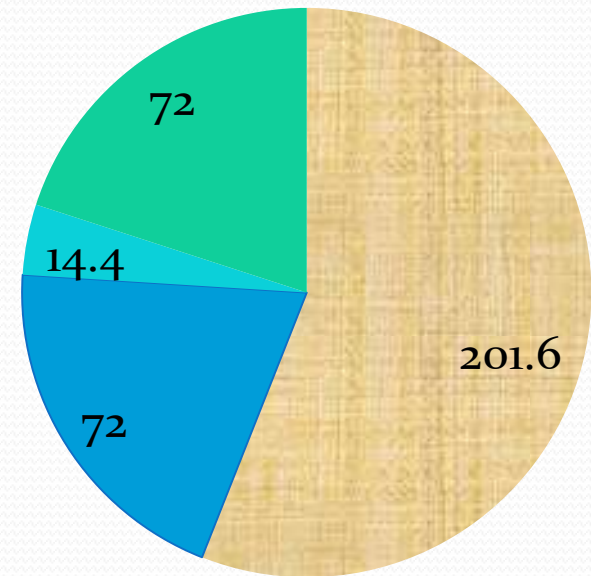
HISTOGRAM:

- A histogram is essentially a bar graph of a frequency distribution.

- It can be constructed for equal as well as unequal class intervals.

- Area of any rectangle of a histogram is proportional to the frequency of that class.

When do we use histogram ?

❑ When data are given in the form of frequencies.

❑ When class interval has to be displayed by a diagram.

❑ When we need to calculate the **Mode** of a distribution graphically.

How to make Histogram ?

Histogram for equal class width:

Class Interval (Height in cm)	Freq.
155-160	3
160-165	2
165-170	9
170-175	7
175-180	10
180-185	5
185-190	5
190-195	1
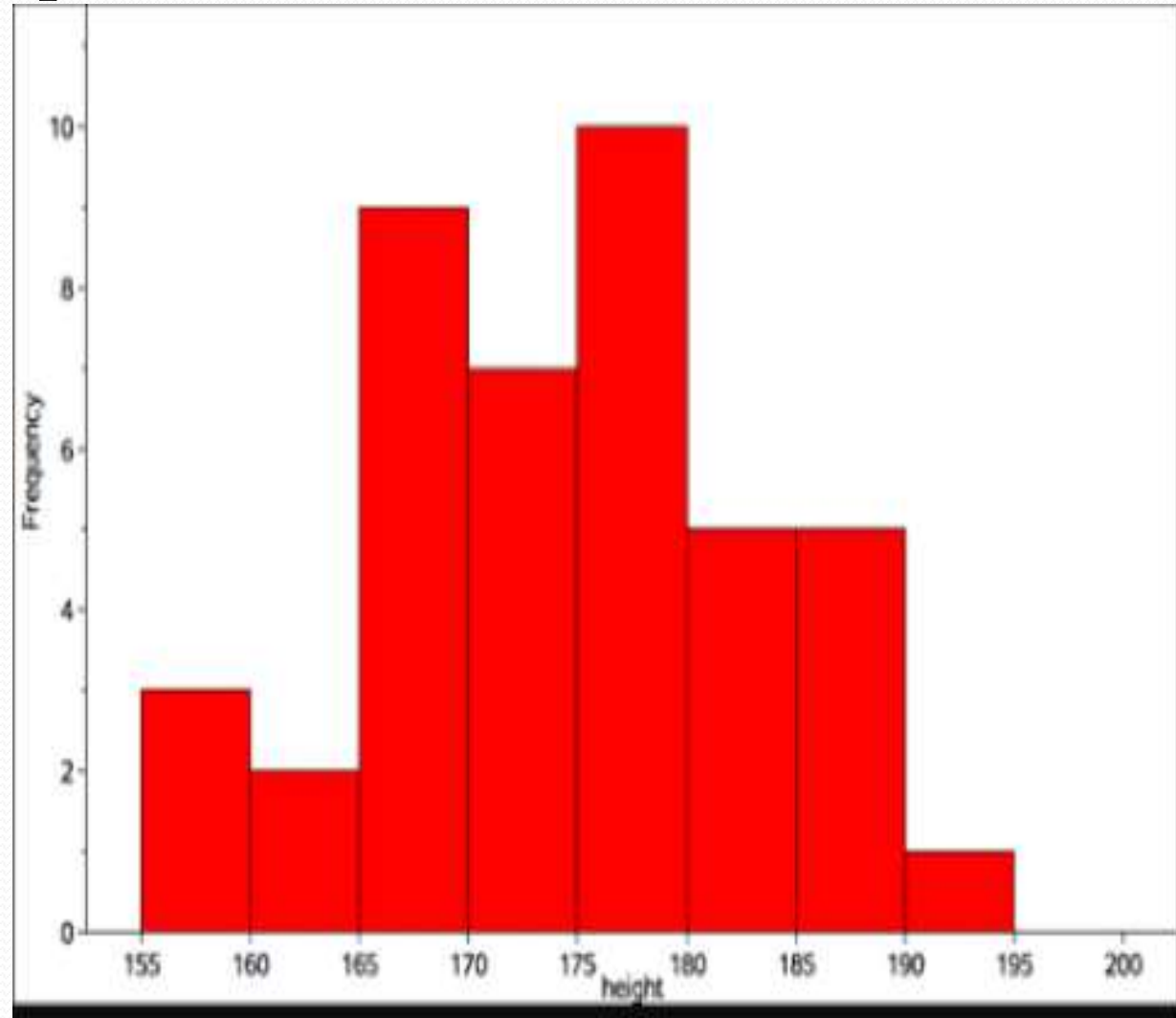
Histogram for unequal class width:

Class boundary	Frequency	Class Width	Frequency Density
0-10	8	10	$\frac{8}{10} = 0.8$
10-15	6	5	$\frac{6}{5} = 1.2$
15-20	12	5	$\frac{12}{5} = 2.4$
20-24	14	4	$\frac{14}{4} = 3.5$
24-35	7	11	$\frac{7}{11} = 0.64$
35-40	3	5	$\frac{4}{7}$
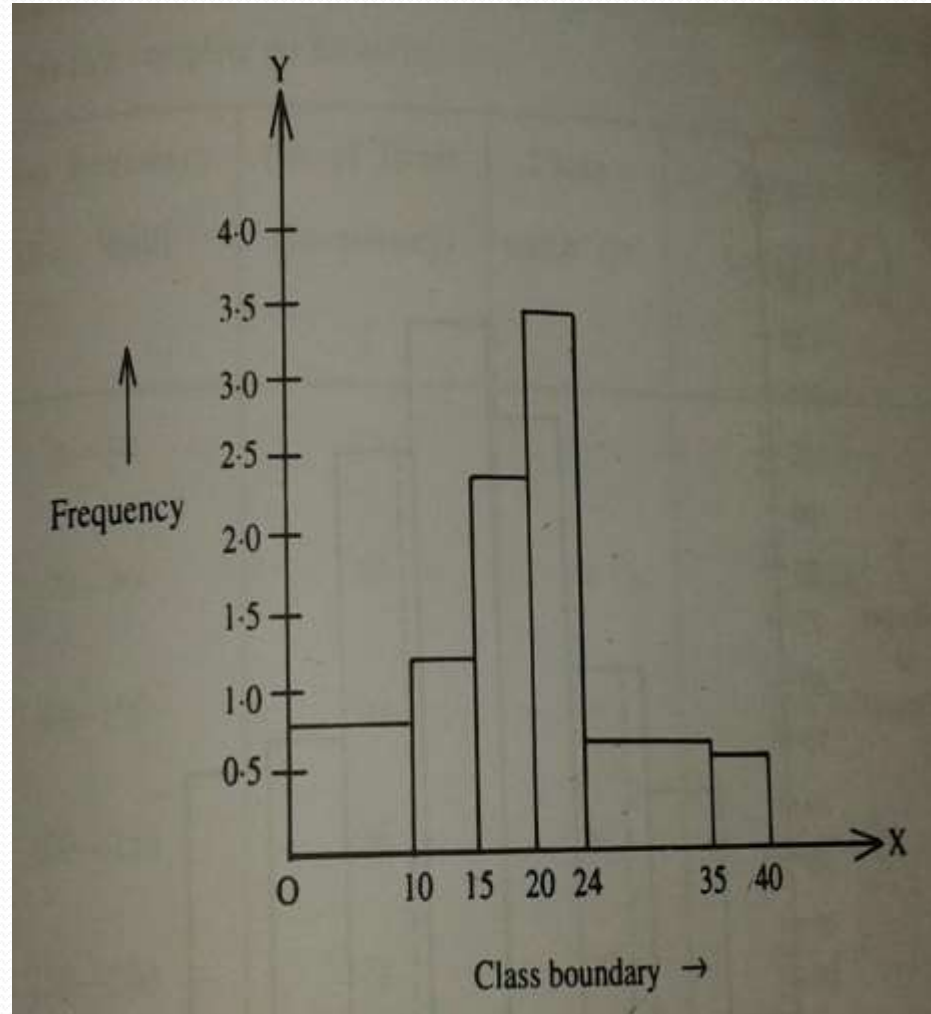
Calculation of **MODE** through Histogram
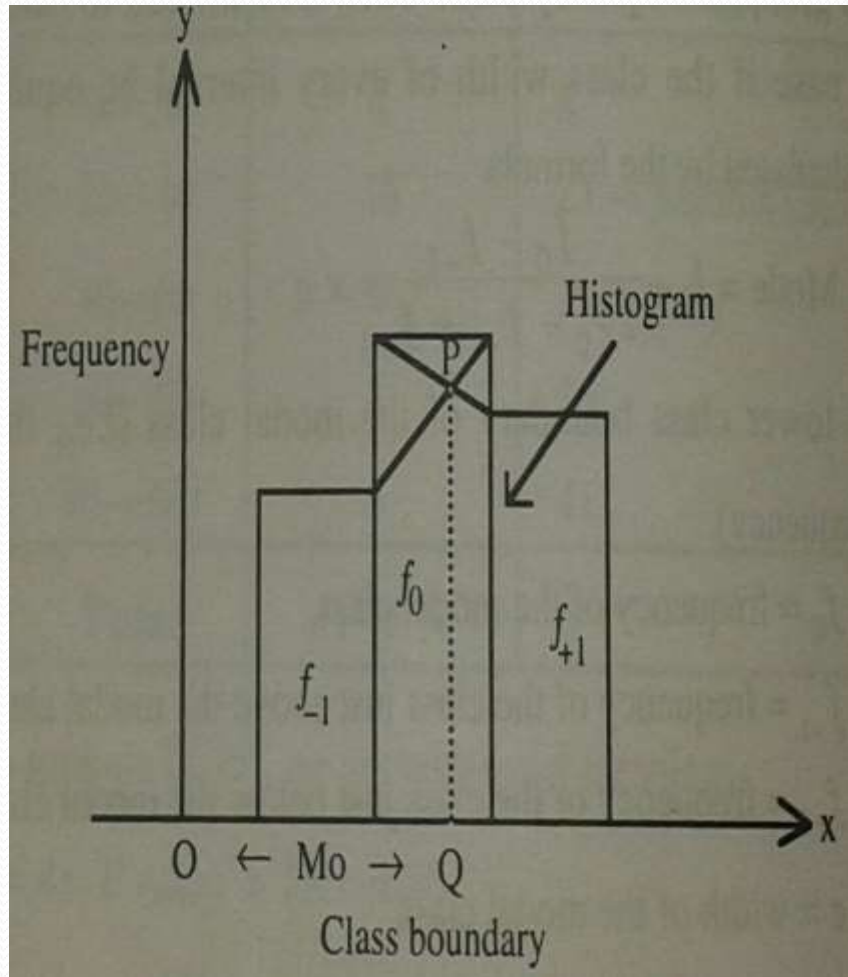
Mode = OQ
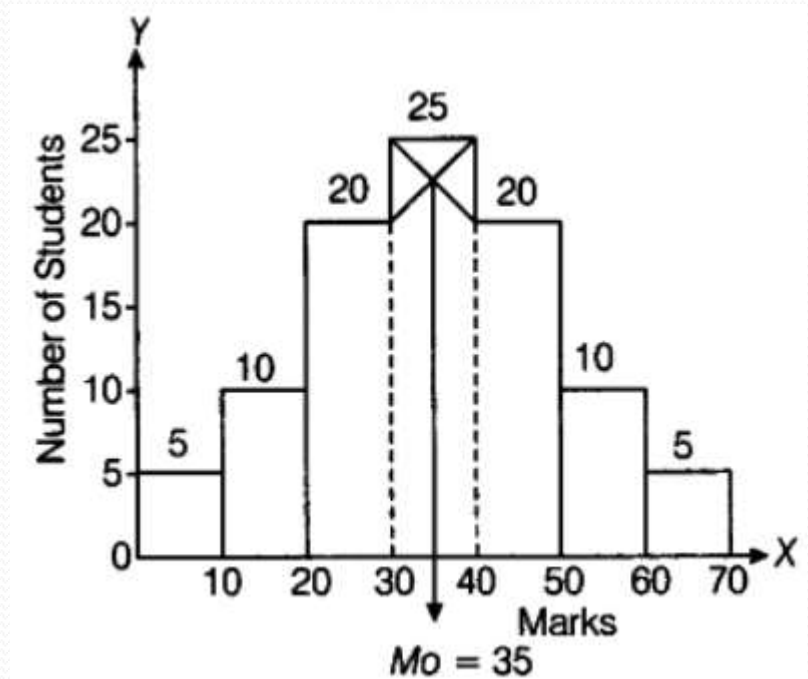
Mode = 35

CALCULATION 0F MODE

- (40,20) and (30,25)
- (30,20) and (40,25)

1st straight line : $\dfrac{y-20}{x-40}$

$= \dfrac{25-20}{30-40}$

$\Rightarrow y - 20 = \dfrac{5}{10} * (x - 40)$

.... (1)

2nd Straight line :

$$\dfrac{y-20}{x-30} = \dfrac{25-20}{40-30}$$

$$\Rightarrow y - 20 = -\dfrac{5}{10} * (x - 30)$$

............(2)

Solving equations (1) and (2) , we get

$$x = 35$$

Mo = 35
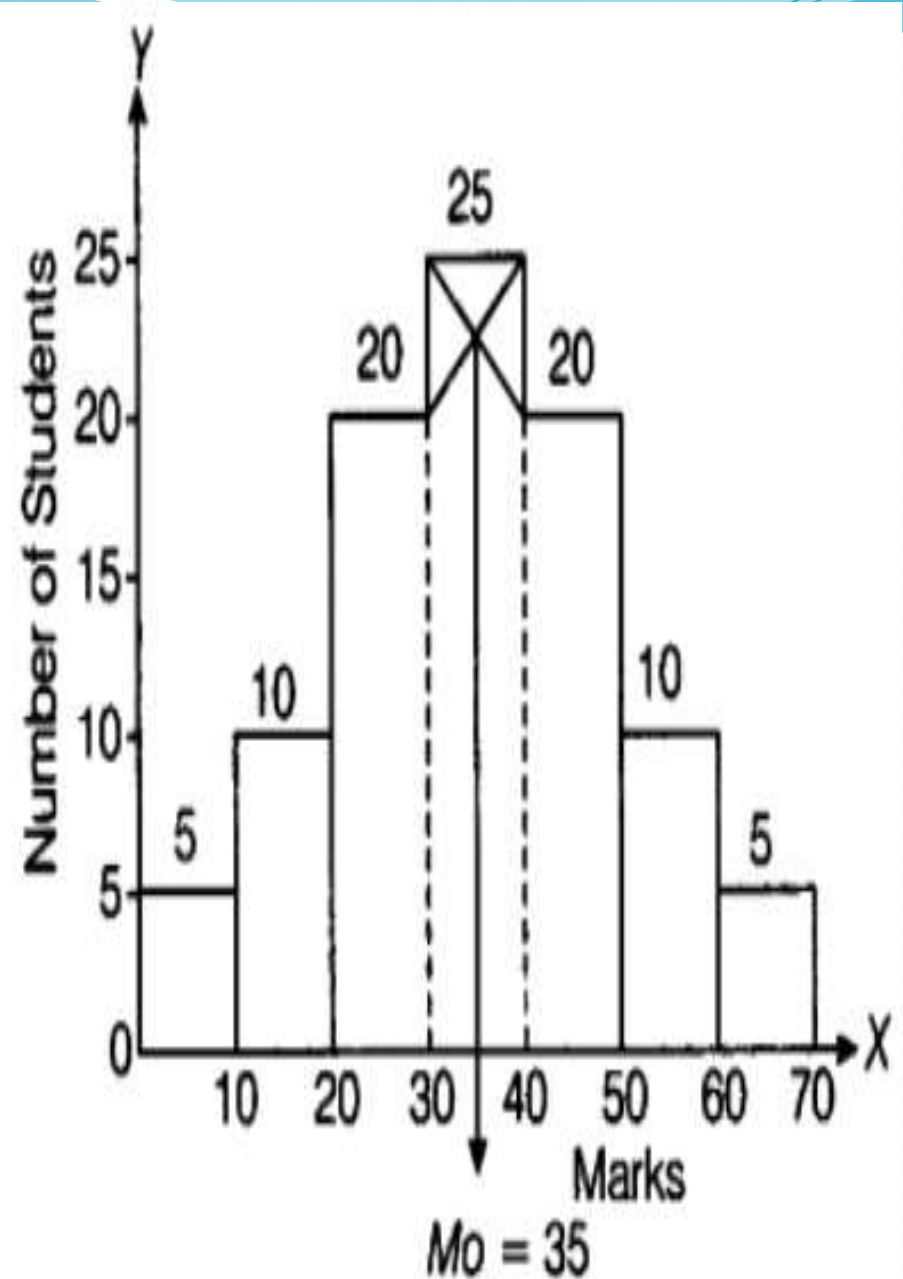
FREQUENCY PLOGON:

- A frequency polygon is essentially a line graph .

- We can get it from a histogram, if the mid points of the upper bases

   of the rectangles are connected by straight lines.

- But it is not essential to plot a histogram first to draw it.

- We can construct it directly from a given frequency distribution.

When do we use Frequency polygon?

o When data are given in the form of frequencies.

o When two or more groups have to be displayed in one

diagram.

o When two or more groups are to be compared.

How to draw frequency polygon?

Height in Cm (class interval)	Mid value	frequency
150-154	152	10
154-158	156	15
158-162	160	20
162-166	164	12
166-170	168	8
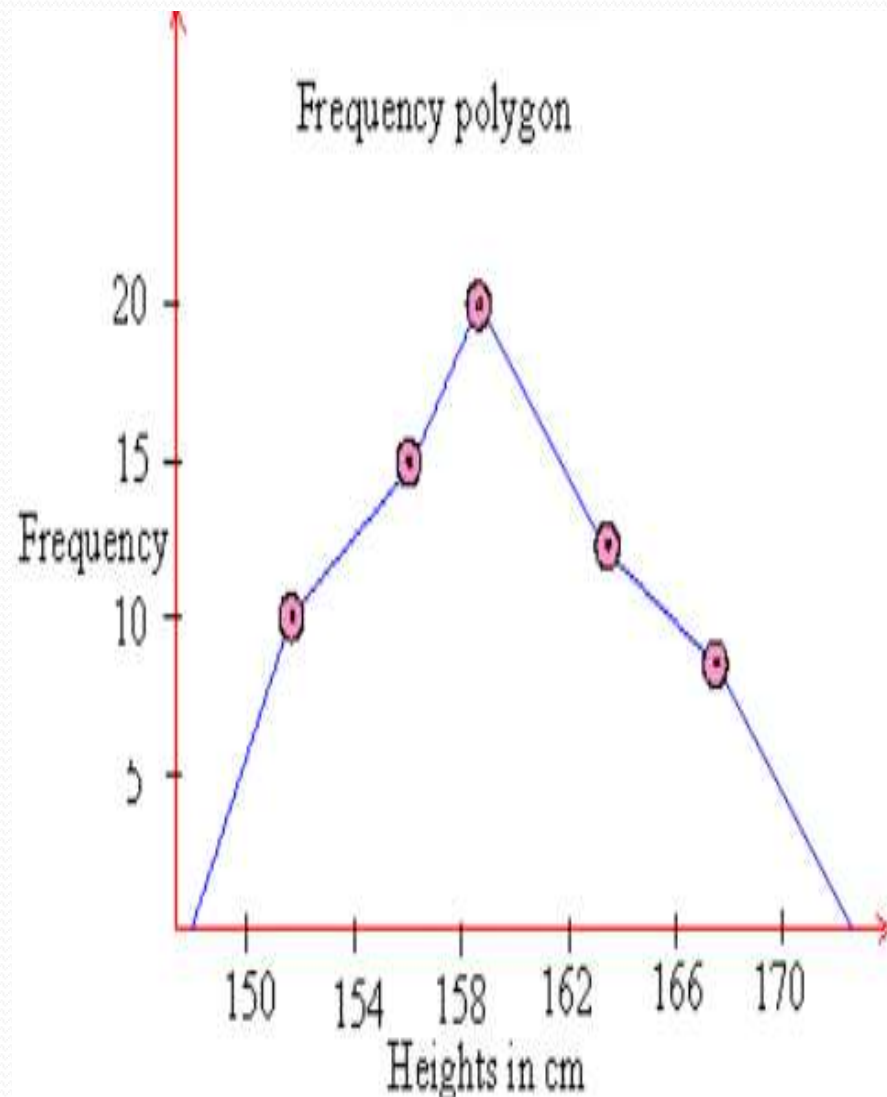
Two or more groups can be compared through Frequency Polygon

Cases of disease X, Y, Z, by month, 2012
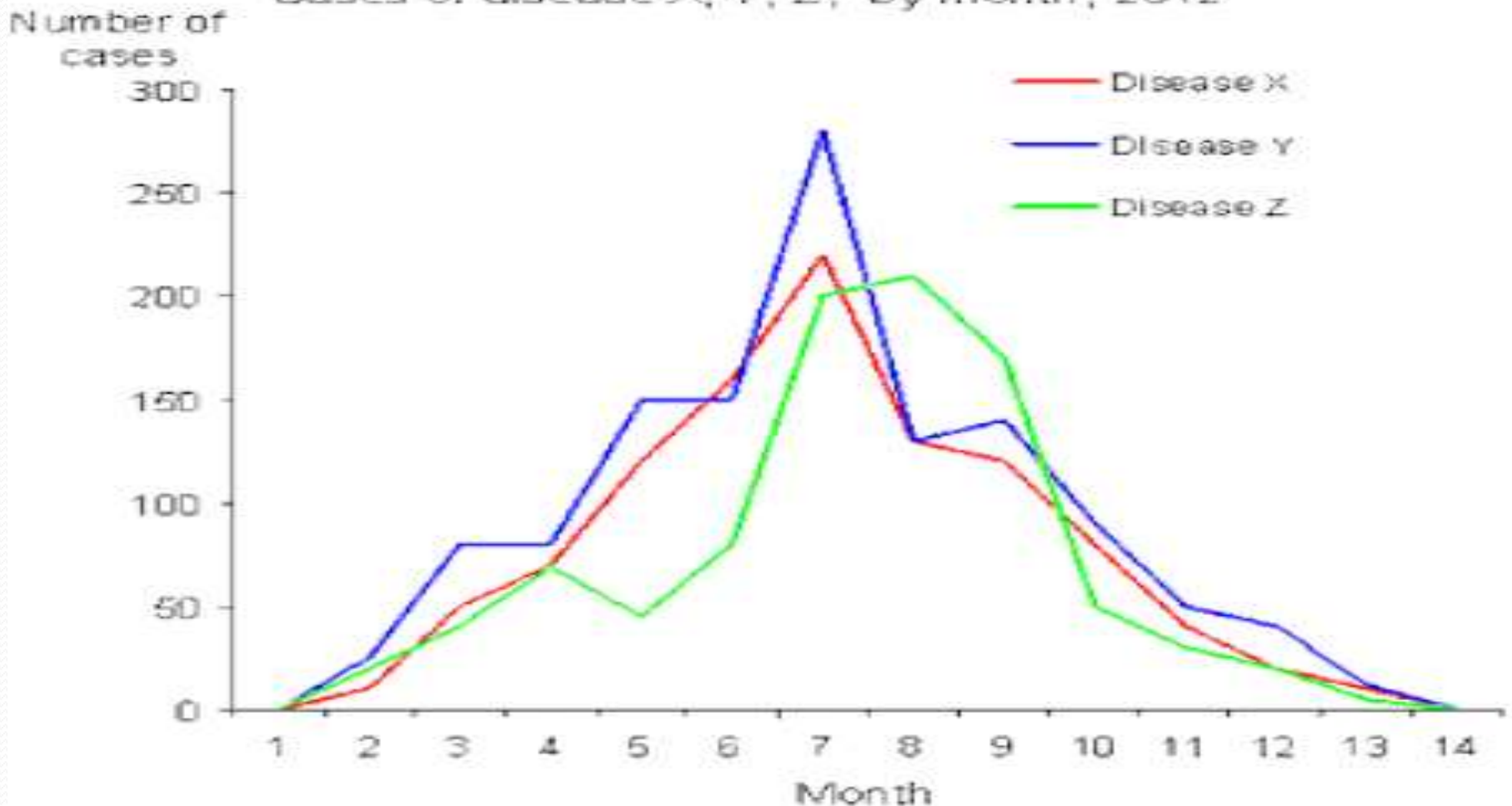
FREQUENCY CURVE:

- Frequency curve is another type of graphical representation of data.

- When then top points of a frequency polygon are joined not by straight lines but by curved ones.

- Frequency polygon is drawn using scale while while Frequency curve is drawn using free hand.

When do we use frequency curve ?

When the number of class intervals are very large i.e.,width of the class intervals are very small and the total number of sample values be increased indefinitely.

FREQUENCY POLYGON
V/S
FREQUENCY CURVE
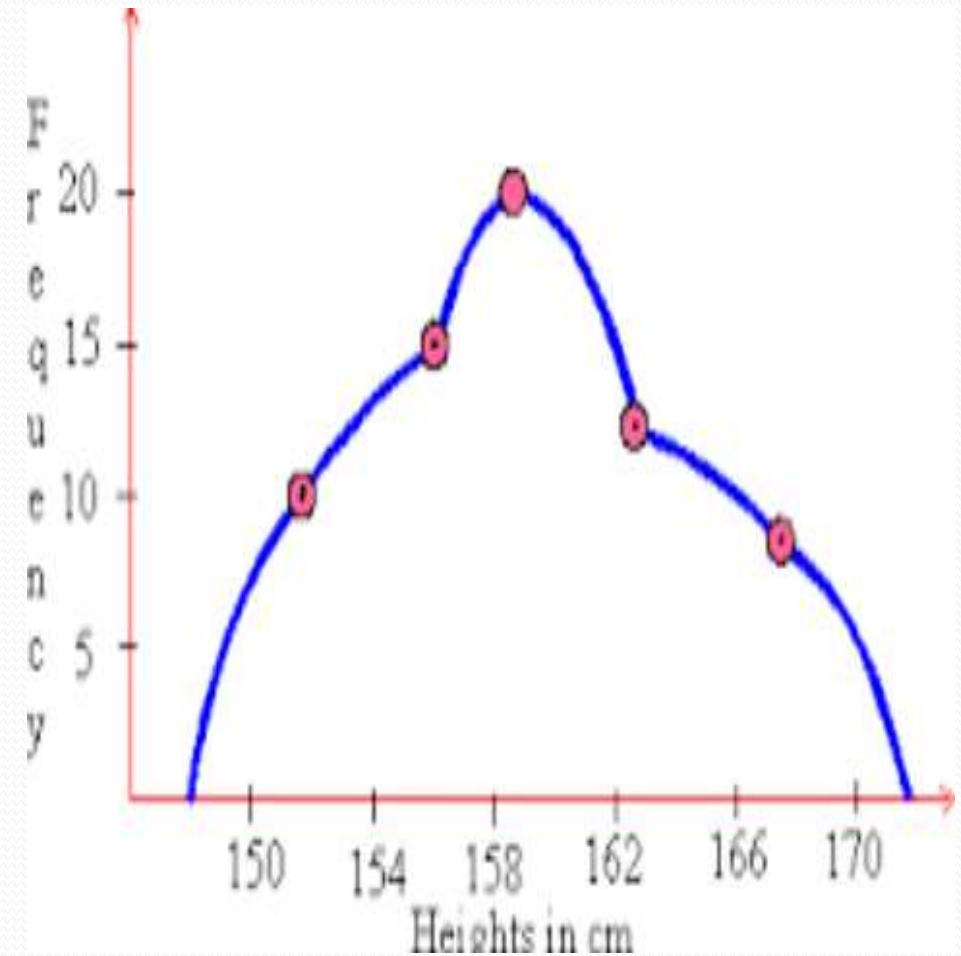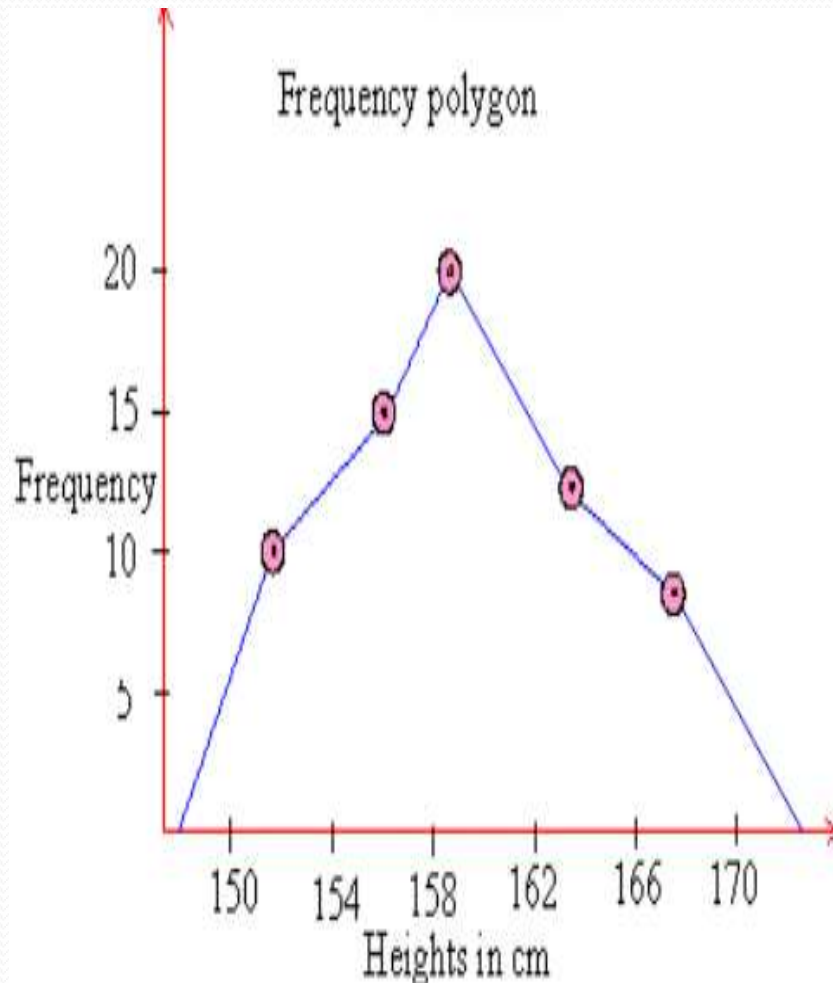
CONCLUSION

So we can conclude that statistical data may be presented in a more attractive form with the help of some graphic aids i.e., pictures and diagrams which carries a lot of communication power and the task of understand and interpretation of data becomes simple, accurate and practicable.

What is Hypothesis?

▶ Hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent variables to some dependent variable.

▶ A hypothesis states what we are looking for and it is a proportion which can be put to a test to determine its validity

e.g.

Students who receive counseling will show a greater increase in creativity than students not receiving counseling

Characteristics of Hypothesis

▶ Clear and precise.

▶ Capable of being tested.

▶ Stated relationship between variables.

▶ limited in scope and must be specific.

▶ Stated as far as possible in most simple terms so that the same is easily understand by all concerned. But one must remember that simplicity of hypothesis has nothing to do with its significance.

▶ Consistent with most known facts.

▶ Responsive to testing with in a reasonable time. One can't spend a life time collecting data to test it.

▶ Explain what it claims to explain; it should have empirical reference.

Null Hypothesis

▶ It is an assertion that we hold as true unless we have sufficient statistical evidence to conclude otherwise.

▶ Null Hypothesis is denoted by H_0

▶ If a population mean is equal to hypothesised mean then Null Hypothesis can be written as

$$H_0 : \mu = \mu_0$$

Alternative Hypothesis

▶ The Alternative hypothesis is negation of null hypothesis and is denoted by H_a

If Null is given as $$H_0: \mu = \mu_0$$

Then alternative Hypothesis can be written as

$$H_a: \mu \neq \mu_0$$

$$H_a: \mu > \mu_0$$

$$H_a: \mu < \mu_0$$

Level of significance and confidence

▶ Significance means the percentage risk to reject a null hypothesis when it is true and it is denoted by α. Generally taken as 1%, 5%, 10%

▶ $(1 - \alpha)$ is the confidence interval in which the null hypothesis will exist when it is true.

Risk of rejecting a Null Hypothesis when it is true

Designation	Risk α	Confidence $1 - \alpha$	Description
Supercritical	0.001 0.1%	0.999 99.9%	More than $100 million (Large loss of life, e.g. nuclear disaster
Critical	0.01 1%	0.99 99%	Less than $100 million (A few lives lost)
Important	0.05 5%	0.95 95%	Less than $100 thousand (No lives lost, injuries occur)
Moderate	0.10 10%	0.90 90%	Less than $500 (No injuries occur)

Type I and Type II Error

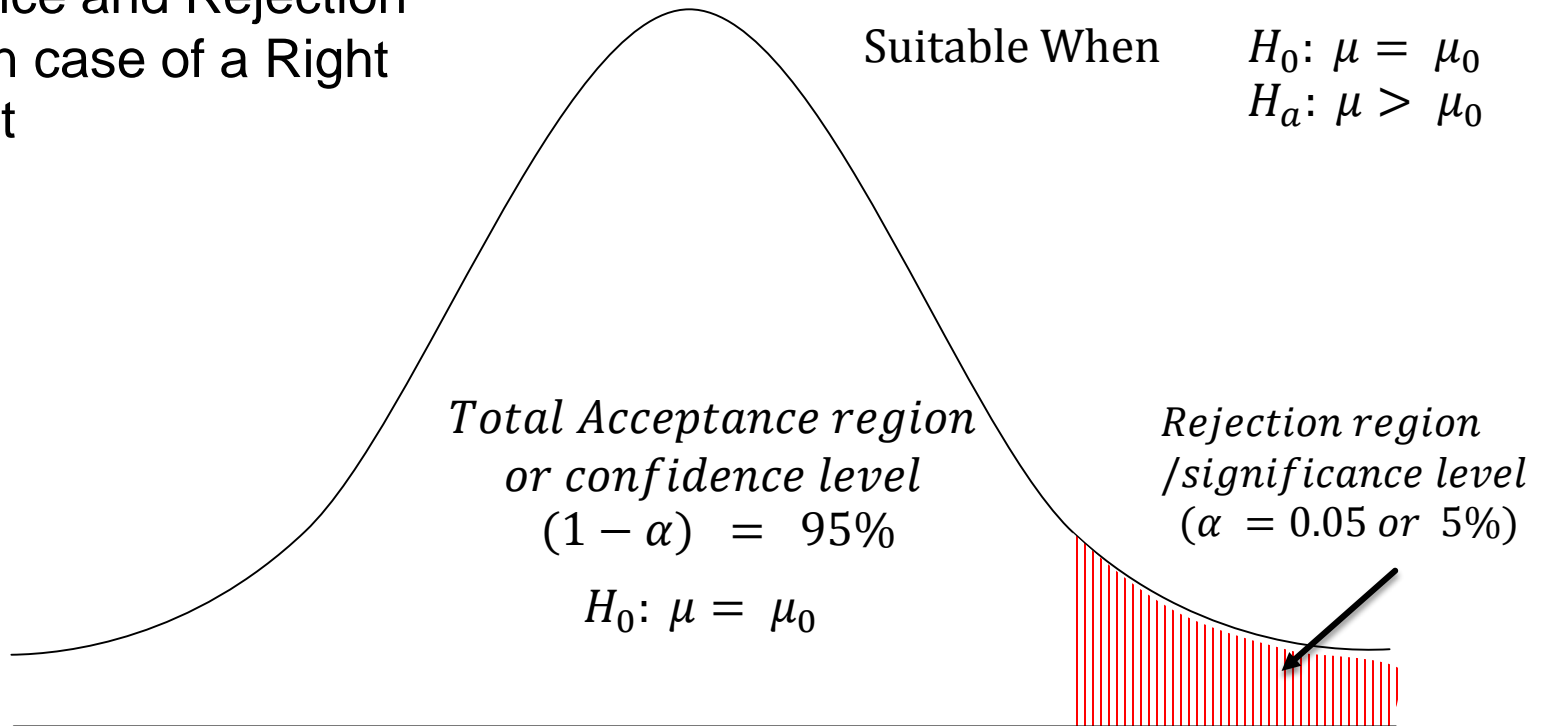| Situation	Decision	
	Accept Null	Reject Null
Null is true	Correct	Type I error ($\alpha\ error$)
Null is false	Type II error ($\beta\ error$)	Correct

Two tailed test @ 5% Significance level

Acceptance and Rejection regions in case of a Two tailed test

Suitable When

$H_0: \mu = \mu_0$
$H_a: \mu \neq \mu_0$

Rejection region /significance level $(\alpha = 0.025 \ or \ 2.5\%)$

Total Acceptance region or confidence level $(1 - \alpha) = 95\%$

$H_0: \mu = \mu_0$

Rejection region /significance level $(\alpha = 0.025 \ or \ 2.5\%)$
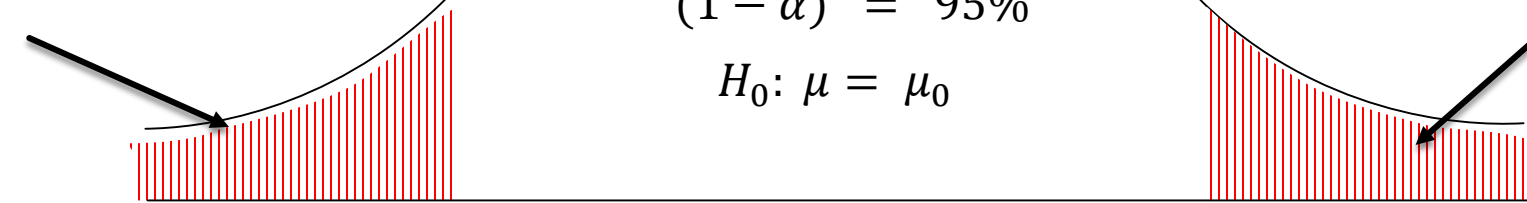
Left tailed test @ 5% Significance level

Acceptance and Rejection regions in case of a left tailed test

Suitable When

$H_0: \mu = \mu_0$
$H_a: \mu < \mu_0$

Rejection region
/significance level
$(\alpha = 0.05 \ or \ 5\%)$

Total Acceptance region
or confidence level
$(1 - \alpha) = 95\%$

$H_0: \mu = \mu_0$
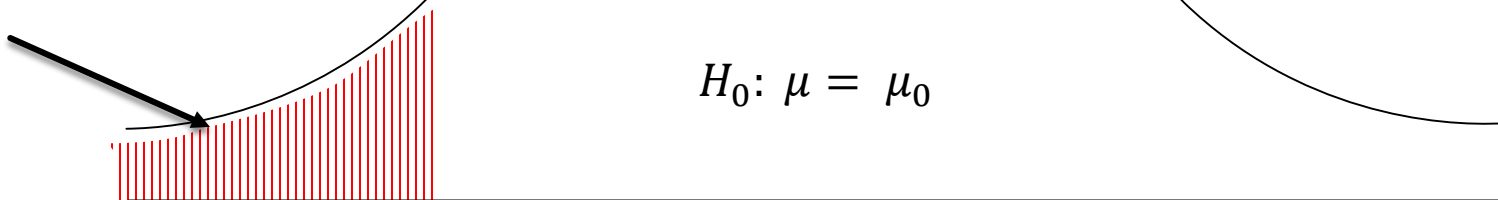
Right tailed test @ 5% Significance level

Acceptance and Rejection regions in case of a Right tailed test

Suitable When $H_0: \mu = \mu_0$
$H_a: \mu > \mu_0$

Total Acceptance region or confidence level
$(1 - \alpha) = 95\%$

$H_0: \mu = \mu_0$

Rejection region /significance level
$(\alpha = 0.05\ or\ 5\%)$

Procedure for Hypothesis Testing

```
State the null          State a              Decide a test         Calculate the
(Ho)and alternate  →  significance level;  →  statistics; z-test, t-  →  value of test
(Ha) Hypothesis       1%, 5%, 10% etc.       test, F-test.           statistics
```

```
Calculate the p-        Compare              P-value >
value at given     →    the p-value     →    Calculated       →     Accept Ho
significance level      with                 value
from the table          calculated
                        value
                          ↓
                        P-value <
                        Calculated       →     Reject Ho
                        value
```
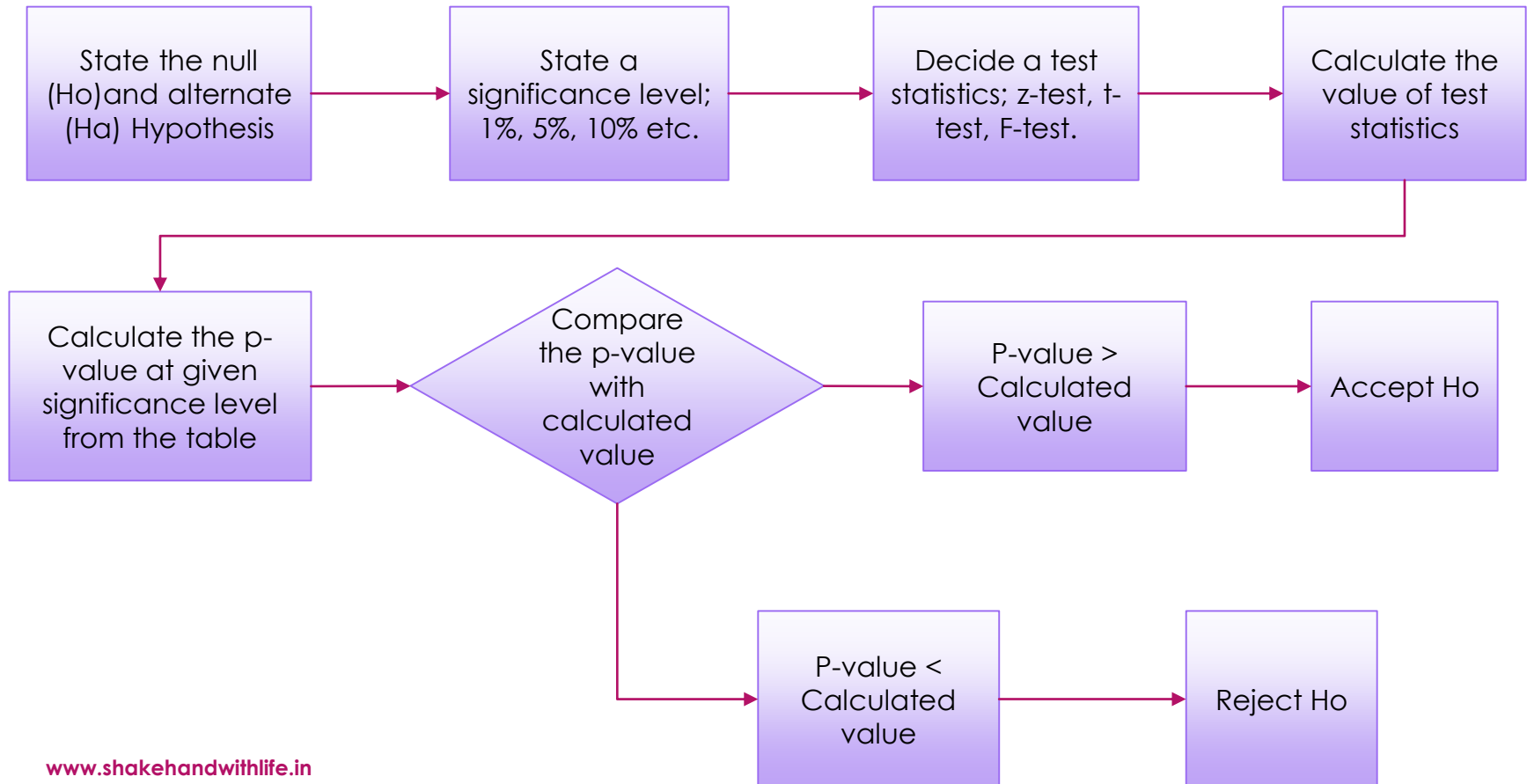
Hypothesis Testing of Means

Z-TEST AND T-TEST

Z-Test for testing means

Test Condition

- Population normal and infinite
- Sample size large or small,
- Population variance is known
- Ha may be one-sided or two sided

Test Statistics

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p / \sqrt{n}}$$

Z-Test for testing means

Test Condition

▶ Population normal and finite,

▶ Sample size large or small,

▶ Population variance is known

▶ Ha may be one-sided or two sided

Test Statistics

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p/\sqrt{n} \times \left[\sqrt{(N-n)/(N-1)}\right]}$$

Z-Test for testing means

Test Condition

▶ Population is infinite and may not be normal,

▶ Sample size is large,

▶ Population variance is unknown

▶ Ha may be one-sided or two sided

Test Statistics

$$Z = \frac{\bar{X} - \mu_{H_0}}{\sigma_S / \sqrt{n}}$$

Z-Test for testing means

Test Condition

- Population is finite and may not be normal,

- Sample size is large,

- Population variance is unknown

- Ha may be one-sided or two sided

Test Statistics

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_s / \sqrt{n} \times \left[\sqrt{(N-n)/(N-1)}\right]}$$

T-Test for testing means

Test Condition

- Population is infinite and normal,

- Sample size is small,

- Population variance is unknown

- Ha may be one-sided or two sided

Test Statistics

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma_s / \sqrt{n}}$$

$$with\ d.f. = n - 1$$

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

T-Test for testing means

Test Condition

▶ Population is finite and normal,

▶ Sample size is small,

▶ Population variance is unknown

▶ Ha may be one-sided or two sided

Test Statistics

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma_s/\sqrt{n} \times \left[\sqrt{(N-n)/(N-1)}\right]}$$

$$with\ d.f. = n - 1$$

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

Hypothesis testing for difference between means

Z-TEST, T-TEST

Z-Test for testing difference between means

Test Condition

- Populations are normal

- Samples happen to be large,

- Population variances are known

- Ha may be one-sided or two sided

Test Statistics

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_{p1}^2}{n_1} + \dfrac{\sigma_{p2}^2}{n_2}}}$$

Z-Test for testing difference between means

Test Condition

- Populations are normal

- Samples happen to be large,

- Presumed to have been drawn from the same population

- Population variances are known

- Ha may be one-sided or two sided

Test Statistics

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

T-Test for testing difference between means

Test Condition

- Samples happen to be small,
- Presumed to have been drawn from the same population
- Population variances are unknown but assumed to be equal
- Ha may be one-sided or two sided

Test Statistics

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)\sigma_{s1}^2 + (n_2 - 1)\sigma_{s2}^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$with\ d.f. = (n_1 + n_2 - 2)$$

Hypothesis Testing for comparing two related samples

PAIRED T-TEST

Paired T-Test for comparing two related samples

Test Condition

▶ Samples happens to be small

▶ Variances of the two populations need not be equal

▶ Populations are normal

▶ Ha may be one sided or two sided

Test Statistics

$$t = \frac{\bar{D} - 0}{\sigma_{diff.}/\sqrt{n}}$$

$$with\ (n-1)\ d.f.$$

$\bar{D}$ = Mean of differences

$\sigma_{diff.}$ = Standard deviation of differences

$n = Number\ of\ matched\ pairs$

Hypothesis Testing of proportions

Z-TEST

Z-test for testing of proportions

Test Condition

▶ Use in case of qualitative data

▶ Sampling distribution may take the form of binomial probability distribution

▶ Ha may be one sided or two sided

▶ $Mean = n.p$

▶ $Standard\ deviation = \sqrt{n.p.q}$

Test statistics

$$z = \frac{\hat{p} - p}{\sqrt{\dfrac{p.q}{n}}}$$

$$\hat{p} = proportion\ of\ sucess$$

Hypothesis Testing for difference between proportions

Z-TEST

Z-test for testing difference between proportions

Test Condition

▶ Sample drawn from two different populations

▶ Test confirm, whether the difference between the proportion of success is significant

▶ Ha may be one sided or two sided

Test statistics

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1 \hat{q}_1}{n_1} + \dfrac{\hat{p}_2 \hat{q}_2}{n_2}}}$$

$\hat{p}_1$ = proportion of success in sample one

$\hat{p}_2$ = proportion of success in sample two

Hypothesis testing of equality of variances of two normal populations

F-TEST

F-Test for testing equality of variances of two normal populations

Test conditions

- The populations are normal
- Samples have been drawn randomly
- Observations are independent; and
- There is no measurement error
- Ha may be one sided or two sided

Test statistics

$$F = \frac{\sigma_{s1}^2}{\sigma_{s2}^2}$$

$with\ (n_1 - 1)\ and\ (n_2 - 1)\ \mathrm{d.f.}$

σ_{s1}^2 is the sample estimate for σ_{p1}^2

σ_{s2}^2 is the sample estimate for σ_{p2}^2

Limitations of the test of Hypothesis

▶ Testing of hypothesis is not decision making itself; but help for decision making

▶ Test does not explain the reasons as why the difference exist, it only indicate that the difference is due to fluctuations of sampling or because of other reasons but the tests do not tell about the reason causing the difference.

▶ Tests are based on the probabilities and as such cannot be expressed with full certainty.

▶ Statistical inferences based on the significance tests cannot be said to be entirely correct evidences concerning the truth of the hypothesis.

Parametric	Non-Parametric
Parametric analysis to test group means	Nonparametric analysis to test group medians
Information about population is completely known	No Information about the population is available
Specific assumptions are made regarding the population	No assumptions are made regarding population
Applicable only for variable	Applicable to both variable and attributes
Samples are independent	Not necessarily the samples are Independent

Parametric	**Non-Parametric**
Assumed normal distributions	No Assumed Shape / distribution
Handles Interval data or Ratio data	Handles Ordinal data, Nominal (or Interval or Ratio), ranked data
Results can be significantly affected by outliers	Results cannot be seriously affected by outliers
Perform well when the spread of each group is different, might not provide valid results if groups have a same spread	Perform well when the spread of each group is same, might not provide valid results if groups have a different spread
Have more statistical power	It is not so powerful like parametric test

Parametric test for Means	**Non-Parametric test for Medians**
1-sample t-test	1-sample Sign, 1-sample Wilcoxon
2-sample t-test	Mann-Whitney test
One-Way ANOVA	Kruskal-Wallis, Mood's median test
Factorial DOE with one factor and one blocking variable	Friedman test

Parametric Tests

Perform well with skewed and non-normal distributions:

This may be a surprise but parametric tests can perform well with continuous data that are non-normal if you satisfy these sample size guidelines.

Parametric analyses	Sample size guidelines for non-normal data
1-sample t test	Greater than 20
2-sample t test	Each group should be greater than 15
One-Way ANOVA	If you have 2-9 groups, each group should be greater than 15. If you have 10-12 groups, each group should be greater than 20.

Parametric or Non-Parametric Determination

Type of Data		

Categorical	Metric	
Non-Parametric Tests	Are the Data approximately normally distributed?	

	NO	YES
	Non-Parametric Tests	Are the variances of populations equal?

		NO	YES
		Non-Parametric Tests	Parametric Tests

Conclusive Thoughts

	Parametric	Non-parametric
Assumed distribution	Normal	Any
Assumed variance	Homogeneous	Any
Typical data	Ratio or Interval	Ordinal or Nominal
Data set relationships	Independent	Any
Usual central measure	Mean	Median
Benefits	Can draw more conclusions	Simplicity; Less affected by outliers
Tests		
Choosing	Choosing parametric test	Choosing a non-parametric test
Correlation test	Pearson	Spearman
Independent measures, 2 groups	Independent-measures t-test	Mann-Whitney test
Independent measures, >2 groups	One-way, independent-measures ANOVA	Kruskal-Wallis test
Repeated measures, 2 conditions	Matched-pair t-test	Wilcoxon test
Repeated measures, >2 conditions	One-way, repeated measures ANOVA	Friedman's test

Introduction to Computer

Chapter 1: Introduction to Computer

- Introduction
  - What is a Computer ?
  - History
  - The Shapes of Computers Today
- Computer Systems
  - Hardware
  - Software
  - Data
  - Users
  - Networks

What is a Computer ?

- There are many points of view to be discussed about the definition, e.g.,
    - Electronics ?
    - Digital ?
    - Programmable ?
    - Manipulate data ?
    - Automated calculation ?
- In this course, we define a computer as an **electronic device** used to process **data** according to a list of **instructions**.

History ~ Ancient calculators

Antikythera mechanism is known as **the 1st mechanical calculator/computer** used for astronomical calculation.

2700-2300 BC   150-100 BC   14th century

A sumerian abacus was capable to add and subtract by counting

With appropriate **procedures**, a **chinese abacus** could calculate multiplication, division, square root, and cube root.

History ~ Mechanical calculators

John Napier invented **Napier's bone**. (He also discovered **logarithms** in 1614)

Pascaline - A mechanical calculator invented by **Blaise Pascal**.

1617   1622   1643   1694
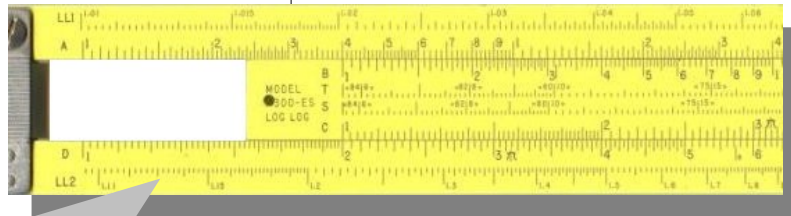
Invented by **William Oughtred**, can calculate log, exp, trigonometry.
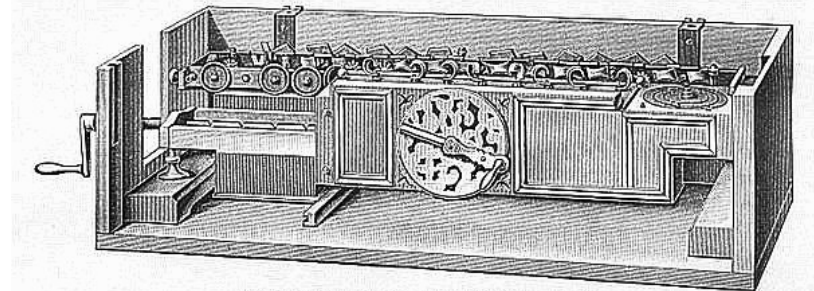
2. Rechenmaschine von Leibniz (1673, Hannover).

Stepped Reckoner, invented by **G. W. Leibniz**, can compute + - x /

History ~ Programmable machines

Charles Babbage attempted to build the **Analytical Engine**, a general-purpose computer, controlled by a list of instruction.

1801   1837   1887

Joseph Marie Jacquard "programmable" loom

Herman Hollerith developed a **punched card tabulating machine**, capable to sort over 200 cards per minute. He founded TMC, merged with CTR which renamed IBM.
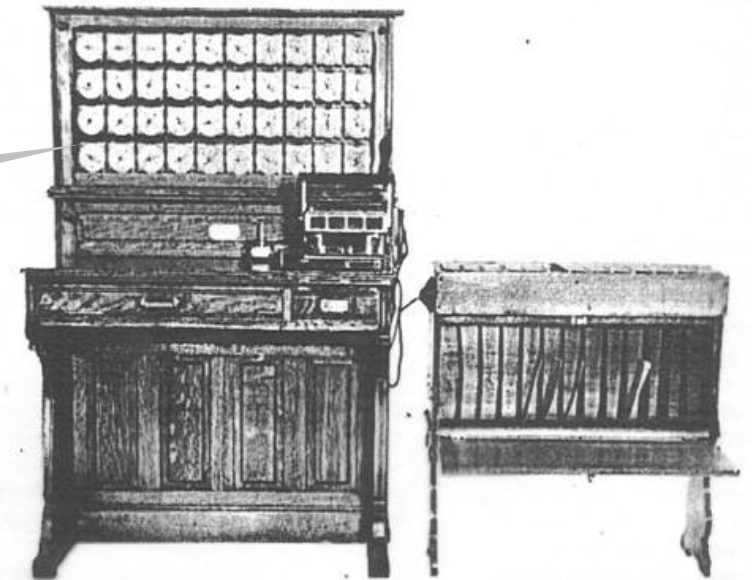
Photo courtesy of IBM

History ~ 1st Gen. (Vacuum Tubes)
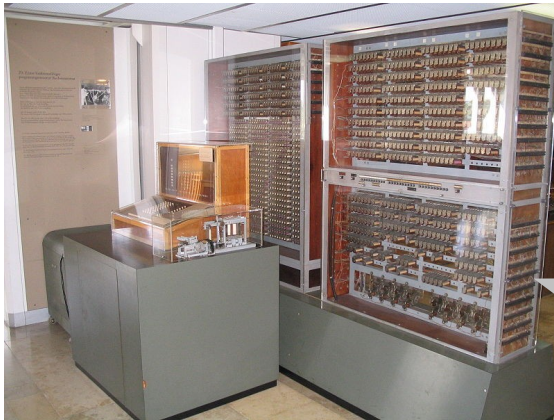
Konrad Zuse's Z3 – the 1st **programmable** (punched film) **turing-complete** digital computer, used relay switches.

1941
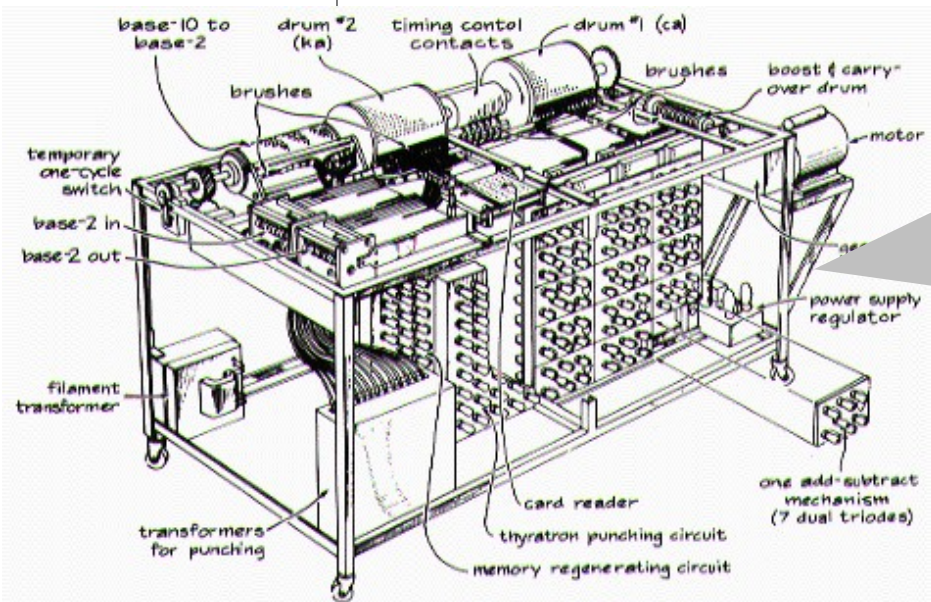
Atanasoff-Berry Computer - the 1st **electronic** (vacuum tube) digital computer. It was not programmable, and not turing-complete.

(cont'd.)

ENIAC – the 1st **all-electronic turing-complete programmable** (wiring, then punched card) computer. It weighted 30 tons, took 63 sq.m. contained 17,468 vacuum tubes, and consumed 150 kW. Performance ~ 300 operations per sec.

1944
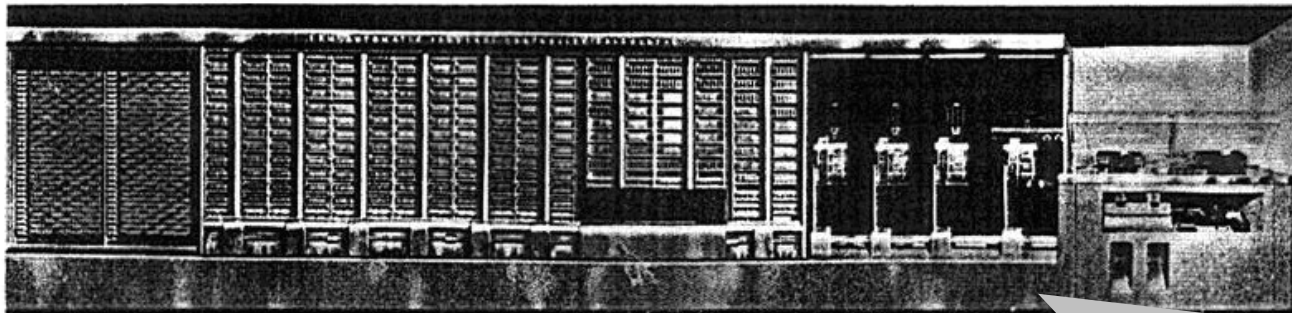
Havard Mark I (IBM ASCC) – the 1st large-scale **automatic** digital computer, used relays, can be programmed by punched paper tape, contained 72 storage registers.

History ~ 2ⁿᵈ Gen. (Transistors)

Bell Lab invented the **transistor** – function like vacuum tubes but smaller, lower power consumption, more reliable.

1947   1951

UNIVAC I – the 1ˢᵗ **commercial** computer. Original priced at US$ 159,000 then rose to US$ 1,500,000. Totally 46 systems installed.

(cont'd.)

(cont'd.)

Mercury Delay Line Memory used in UNIVAC I
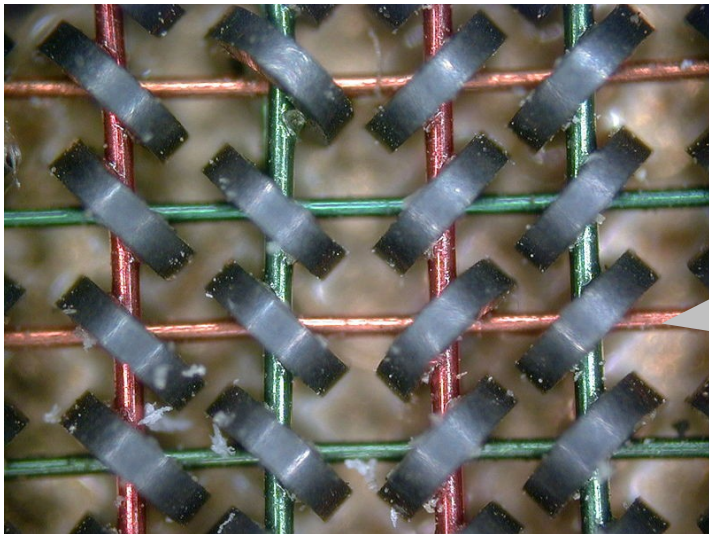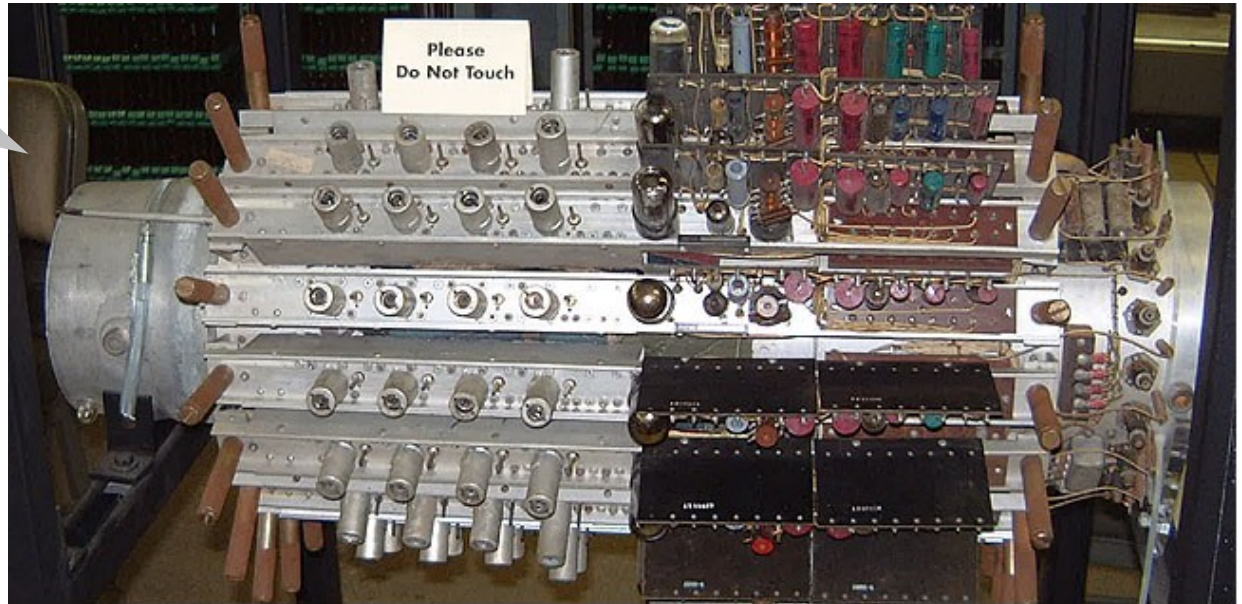
Magnetic Core Memory used in later models of UNIVAC

History ~ 3ʳᵈ Gen. (Integrated Circuits)

Jack Kilby invented the **Miniaturized Electronic Circuit**

DEC PDP-8 – started from US$ 16,000, it is the first successful **minicomputer**.

1958   1964   1965

IBM introduced **System/360** – a highly configurable, highly backward compatible, **mainframe** computer system.

History ~ 4th Gen. (Microprocessors)

Apple I – the 1st PC of Apple, with the price tag of US$ 666.66

1971   1975   1976

Intel 4004 – The 1st commercial microprocessor

MITS Altaire 8800 – the 1st microcomputer, based on intel 8080, sold as mail-ordered kit.

(cont'd.)

Apple II – The beginning of PC era. It's the 1st highly successful mass-produced PC.

1977   1981

IBM PC – Because of the name of IBM, business adopted using a PC for the office work.

(cont'd.)

Apple Macintosh – the 1ˢᵗ commercially successful computer that uses a GUI.

1982   1984   1985

Compaq Portable – The first 100% compatible IBM PC.

Microsoft Windows – GUI for IBM PC & Compatible.

(cont'd.)

Tim Berners Lee invented WWW.

Deep Blue defeated the world #1 G. Kasparov.

1989 **1990** **1991** **1997**

Linux – a free/open source alternative OS originally written by **Linus Torvalds**.
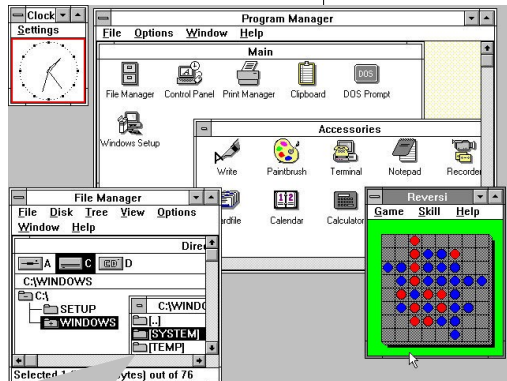
Microsoft Windows 3.0 - *de facto* GUI for PC.

(cont'd.)

AMD Athlon 64 X2 – the 1st **64-bit dual-core** processor for PCs
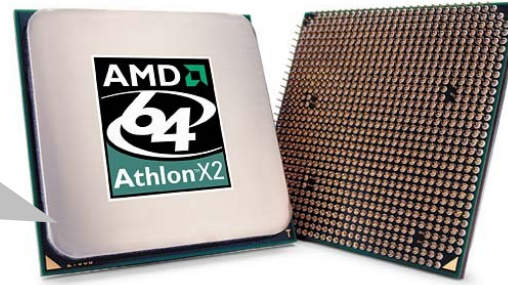
2001   **2005**   **2007**
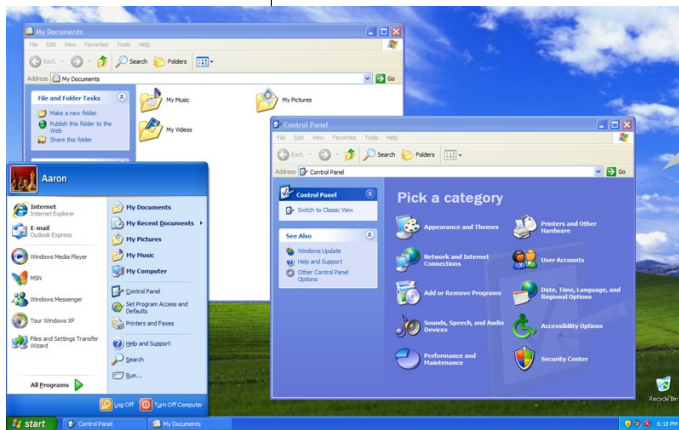
Windows XP released.

Windows Vista released

Moore's Law

US$ per GFLOPS
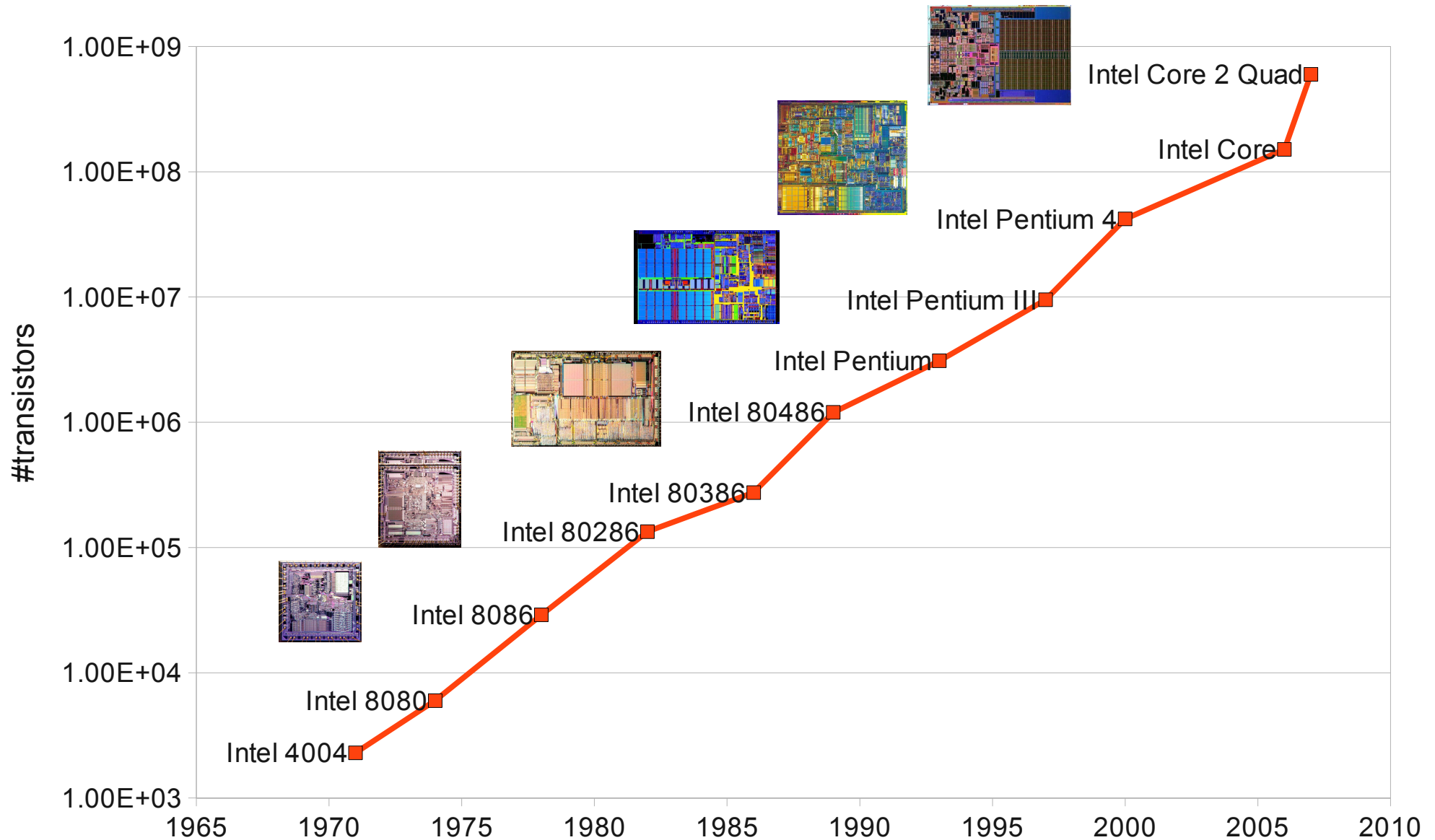
Chart plotting US$ (vertical axis, logarithmic, from 1.00E-03 to 1.00E+13) versus year (horizontal axis, 1960 to 2020). Labeled data points: IBM 1620, Cray 1, Cray XMP/48, Beowulf 2x16xPentium Pro, Bunyip, KLAT2, KASY0, ATI X1900, Ambric AM2045, Sony PS3. Annotations: "Gordon Bell Prize 'Less than US$ 1 per MFLOPS'" and "US$ 0.2 per MFLOPS".
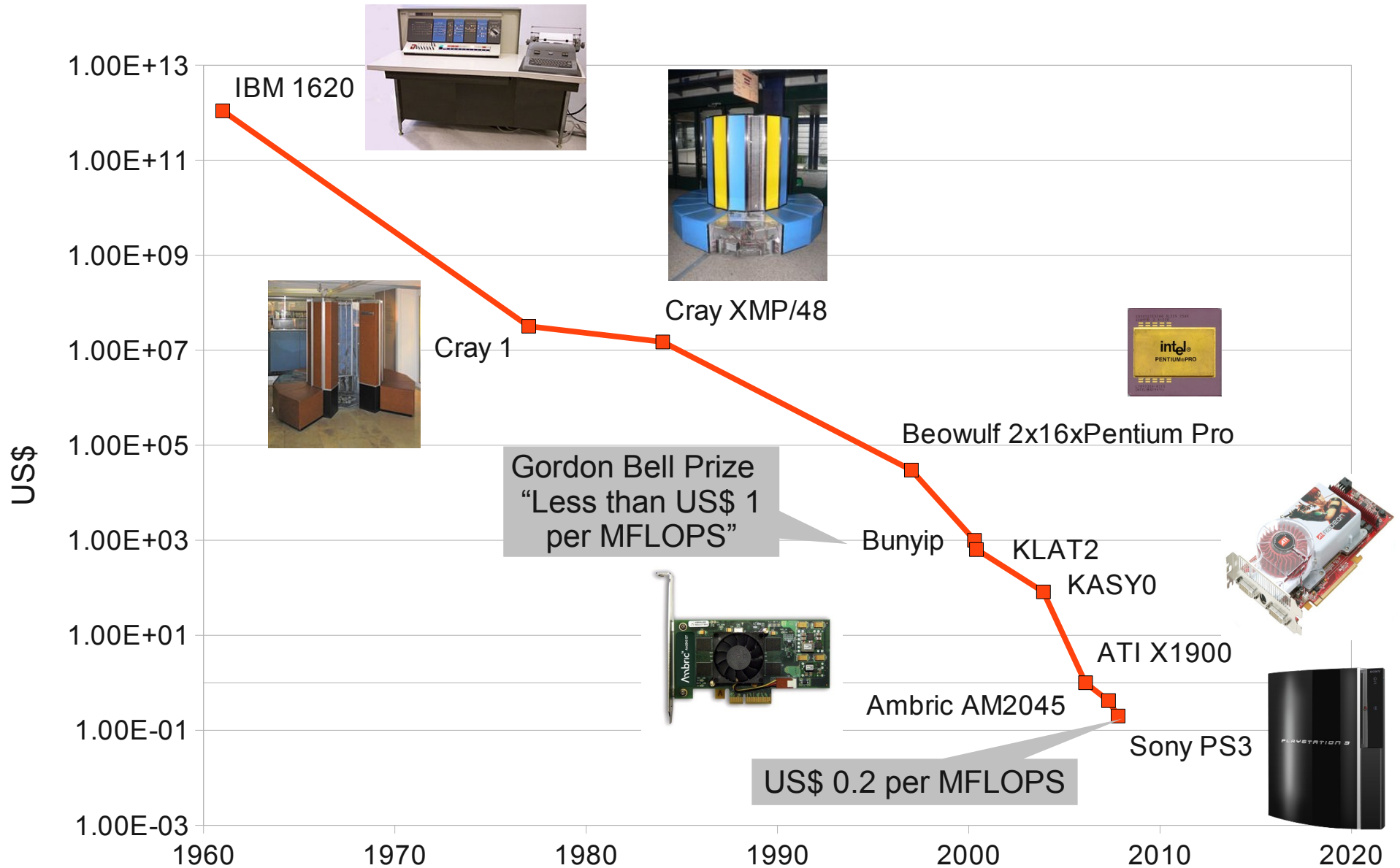
The Shapes of Computers Today

- Although the capabilities and type of computer have changed quickly. There are the terms describing:
    - Supercomputers
    - Mainframes
    - Minicomputers
    - Microcomputers
- All these types of computers can be connected together to form networks of computers, but each individual computer, whether it is on a network or not, falls into one of these four categories.

Supercomputers

- Supercomputers are **the most powerful** computers. They are used to process huge amounts of data, model of complex processes and simulate the processes.
  - Nuclear fission
  - Air pollution
  - Weather forecast
  - Astrophysics
  - Fluid dynamics
  - Genetic
  - Chess
  - Breaking ciphers

Mainframes

- Mainframe computer is the **largest** type computer in use.
  - Large memory, storage, I/O.
- They are used where many people in a large organization need frequent access to the same information which is organized into one or more huge databases.
  - Transactions
  - Accounting
  - ERP

Minicomputers

- The capabilities of a minicomputer lies somewhere between those of mainframes and those of microcomputer.
  - But they can handle more I/O and/or more terminals.
- Obsoleted by microcomputer.

Microcomputers

- The least powerful, but most widely used.
- The term microcomputer and personal computer are interchangeable.
  - PCs are intented to be operated by end users.
  - Size, price, capabilities are right for individuals.
- Fastest growing ~ microprocessors, memory chips, and storage devices keep making gains in speed and capacity, while physical size and price remain stable or in some cases are reduced.

Microcomputer – Desktop

Microcomputer – Notebook

Microcomputer – Tablet

Microcomputer – Handheld

Phone + Computer

INTRODUCTION TO PROGRAMMING
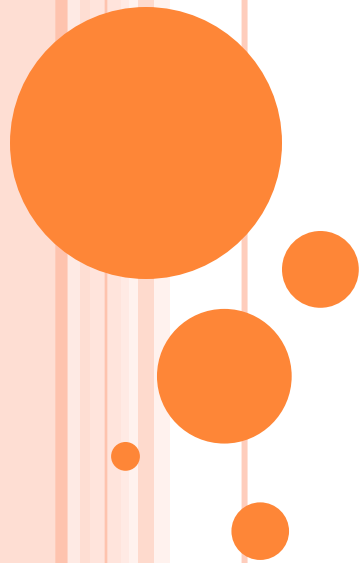
Algorithm

- It is a list of instructions specifying a precise description of a step by step process that terminates after a finite number of steps for solving an algorithm problem producing the correct answer in the end.

- It is a recipe for solving problems.

- A finite set of an instruction that specifies a sequence of operation to be carried out in order to solve a specific problem.

- An unambiguous procedure specifying a finite number of steps to be taken.

METHODS OF SPECIFYING ALGORITHM

- **Pseudocode** - specifies the steps of algorithm using essentially natural language of superimposed control structure.

- **Flowchart** - a traditional graphical tool with standardized symbols. Show the sequence of steps in an algorithm.

PROPERTIES OF ALGORITHM

- **Finiteness** - there is an exact number of steps to be taken and has an end.

- **Absence of Ambiguity** - means that every instruction is precisely described and clearly specified.

- **Sequence of Execution** - instructions are performed from top to bottom.

- **Input and Output** - defined the unknowns of the problem is specified and with the expected outcome.

- **Effectiveness** - the solution prescribed is guaranteed to give a correct answer and that the specified process is faithfully carried out.

- **Scope Definition** - applies to a specific problem or class of problem.

Steps in Program Development

1. **State the problem clearly**- a problem cannot be solved correctly unless it is being understood.

2. **Plan and Write the Logical Order of Instructions** - the computer follows the direction exactly at the given sequence.

3. **Code the Program** - write the programming statements in the desired language.

4. **Enter the program into the computer** - key in or type the statement into the computer.

5. **Run and Debug the program** - check if you have the desired output; if not, trace the possible error.
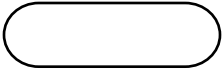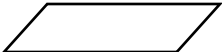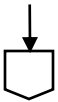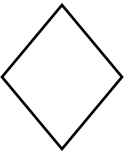
Flowcharting Guidelines
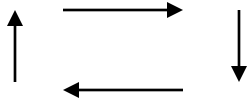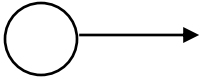
1. The flowchart should flow from top to bottom

2. If the chart becomes complex, utilize connecting blocks

3. Avoid intersecting flow lines

4. Use meaningful description in the symbol

Flowcharting Symbols

SYMBOL	NAME	DESCRIPTION
⬭	**TERMINAL**	Defines the starting and ending point of a flowchart.
⬡	**INITIALIZATION**	The preparation or initialization of memory space for data processing.
▱	**INPUT/OUTPUT**	The inputting of data for processing, and printing out of processed data.
▭	**PROCESS**	Manipulation of data(assignments and mathematical computations)

Flowcharting Symbols

SYMBOL	NAME	DESCRIPTION
	FLOW LINES	Defines logical sequence of the program. Its points to the ext symbol to be performed
	ON-PAGE CONNECTOR	Connects to the flowchart to avoid spaghetti connection on the same page
	OFF-PAGE CONNECTOR	Connects the flowchart on different page to avoid spaghetti connection
	DECISION	Process conditions using relational operators. Used for trapping and filtering data.

Sample Exercises

Sample 1: Write a program that calculates the sum of two input numbers and display the result.

Sample 2: Write a program to calculate the area of a circle and display the result. Use the formula: $A=\pi r^2$ where Pi is approximately equal to 3.1416.

Sample 3: Write a program that computes the average of three input quizzes, and then display the result.

Sample 4: Write a program that converts the input Fahrenheit degree into its Celsius degree equivalent. Use the formula: $C= (5/9)*F-32$.

Sample 5: Create a program to compute the volume of a sphere. Use the formula: $V= (4/3)* \pi r^3$ where is pi equal to 3.1416 approximately. The r^3 is the radius. Display result.

Sample 6: Write a program that converts the input Celsius degree into its equivalent Fahrenheit degree. Use the formula: $F= (9/5) * C+32$.

Introduction to Windows

What is Windows?

Windows

- Windows is an operating system which makes the computer system work.
- Without Windows, you cannot operate the computer such as the system unit, monitor and the keyboard.

What are the

BASIC ELEMENTS OF WINDOWS XP?

BASIC ELEMENTS OF WINDOWS 7?

Basic Elements of
WINDOWS XP

Desktop

- The desktop is the entire screen on Windows XP. It is called a desktop because it looks and functions very much like a neat desk or working table.

Icons

- Icons are small pictures found on the desktop. These are symbols representing programs, applications, or files. Each icon is a shortcut to an item, file, or program inside the computer.

Start Button

- It is a button found on the lower left part of the desktop. It is used to start a program or open a document.

Taskbar

- It is a rectangular bar found at the lower part of the desktop. It contains the Start button and shows what programs or documents are currently open.

Quick Launch Bar

- This provides an easy way to launch a program with just one click.

System Tray

- In the System Tray, this is where you can find icons belonging to some of the programs that are currently running.

Wallpaper

- This is the background design of the desktop. You can customize the wallpaper according to your preference.

Basic Elements of WINDOWS 7

Icons

Desktop

Wallpaper

Page 1 of 1   Search gadgets

Some gadgets access the Internet. For more information, see the privacy statement online.

Calendar   Clock   CPU Meter   Currency   Feed Headlines

Picture Puzzle   Slide Show   TuneUp Utilities   Weather   Windows Media...

Show details   Get more gadgets online

Notification Area

10:53 AM
8/19/2013

Start Button   **Taskbar**

Desktop

- The desktop is the main screen area that you'll see.

- Contains mini-programs called Gadgets.
- Clock, currency, slideshow, picture puzzle, weather

Icons

- Icons are small pictures found on the desktop. These are symbols representing programs, applications, or files. Each icon is a shortcut to an item, file, or program inside the computer.

Start button

- It is the round button found on the lower part of the desktop which you can use to open programs and folders.

Taskbar

- It is found at the bottom of your screen. It also contains the start button, the middle section and notification area.

Notification Area

- The Notification Area informs you of the status of the running program such as anti-virus, printing, computer updates, and time.

Wallpaper

- This is the picture or design used as a background of the desktop.

- In Windows 7, there are new wallpaper designs and desktop slideshow, which display a rotating series of pictures.

What is MS?

- Microsoft is a company/corporation in USA.
- Bill Gates is the owner of Microsoft Company.

What is Microsoft Office?

- Microsoft office is used for special purpose office work such as:
- Documentation
- Work Sheet
- Presentation
- Data base.

MS-OFFICE HAS FOUR PAKAGES

MS-WORD

MS-EXCEL

MS-
POWERPOINT

MS-ACCESS

FIRST PACKAGE

MS-WORD
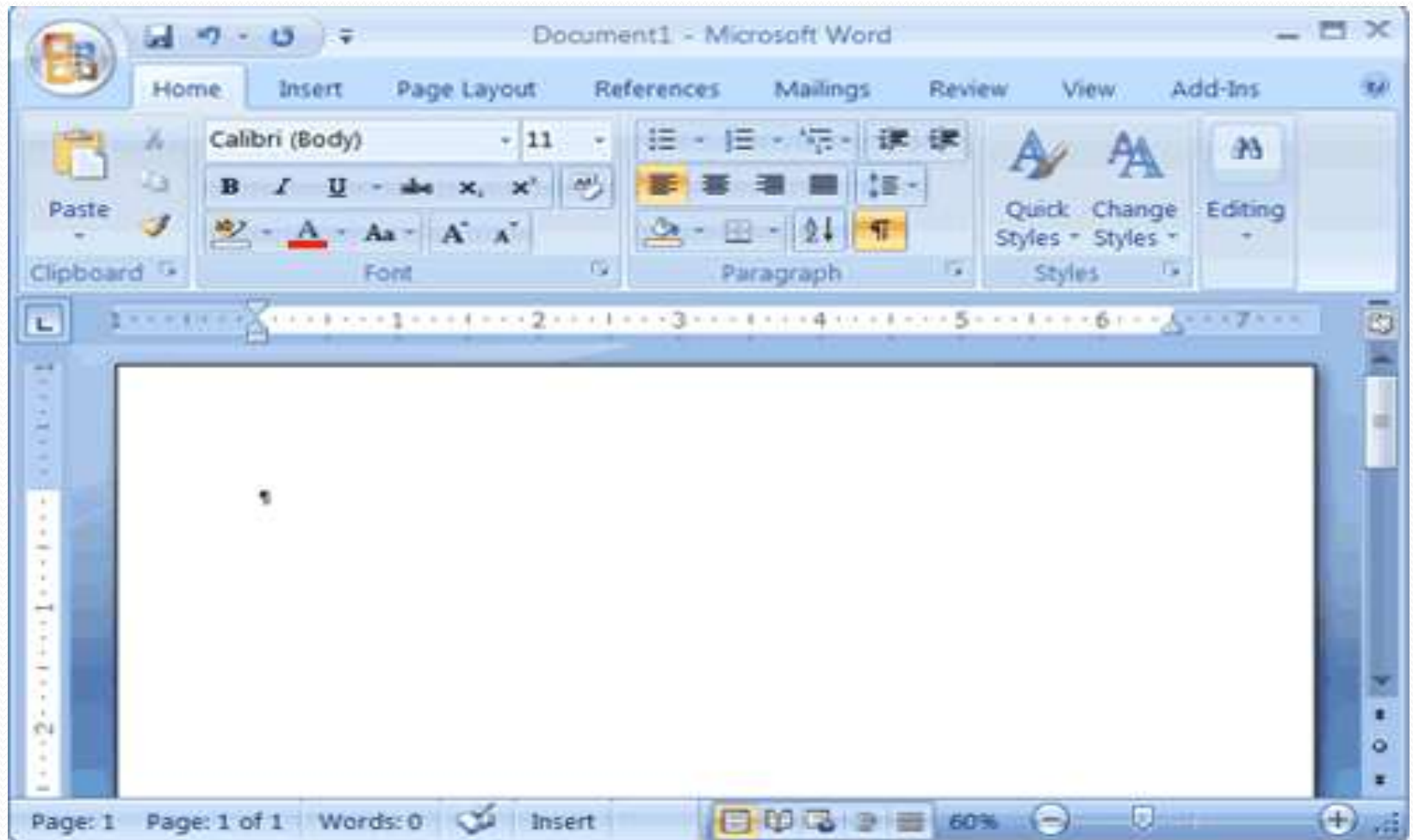
What is MS-WORD ?

- Microsoft Word is a word processing software pakage.
- You can use it to type letters ,reports,and other documents.
- The four main operations of a word processing pakage are:
- ➢ Defining the form of the document
- ➢ Entering a document from a keyboard
- ➢ Editing (modifying) the document
- ➢ Printing the document.

Components of Microsoft Word
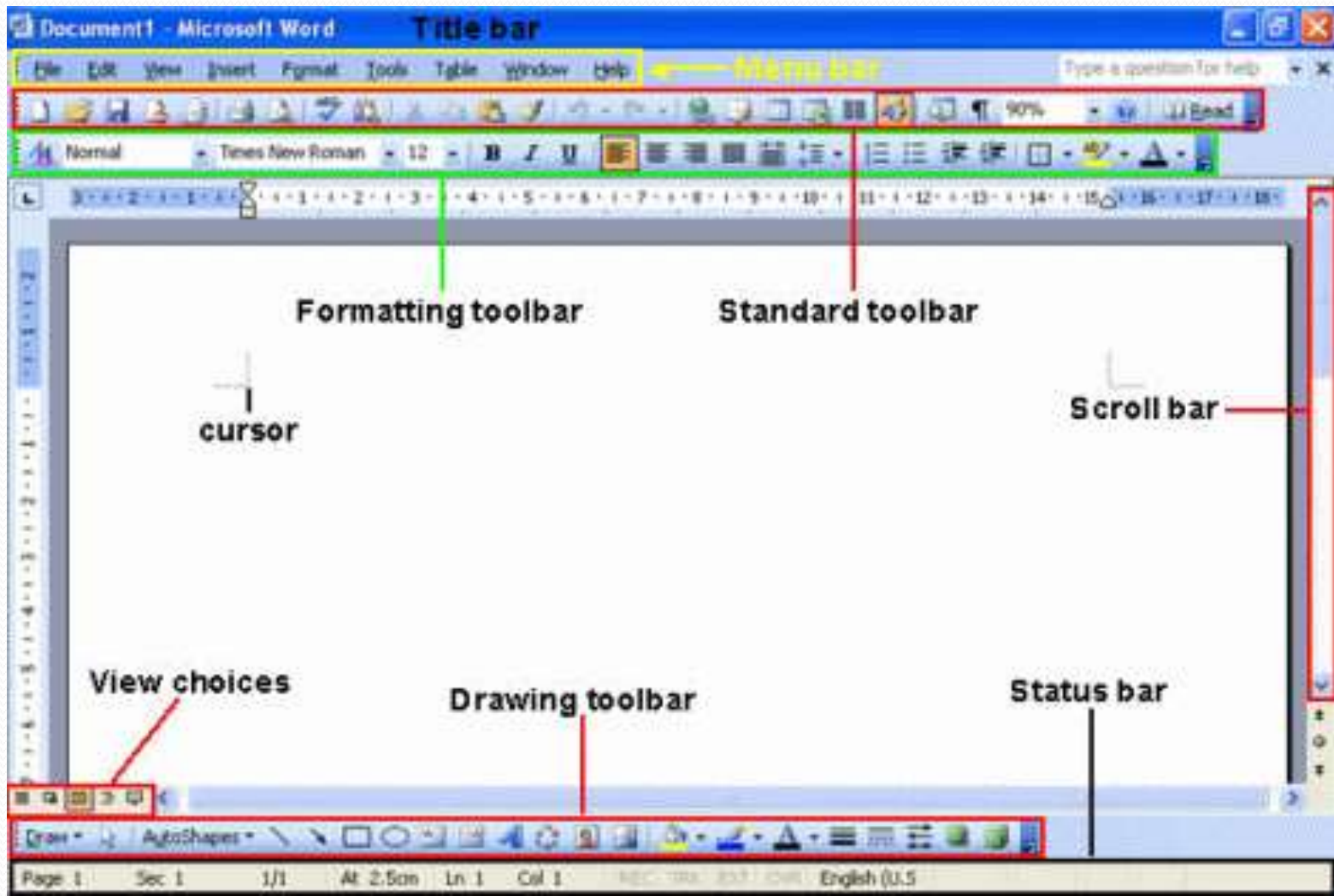
The Ribbon that displays various commands and features of all the tabs
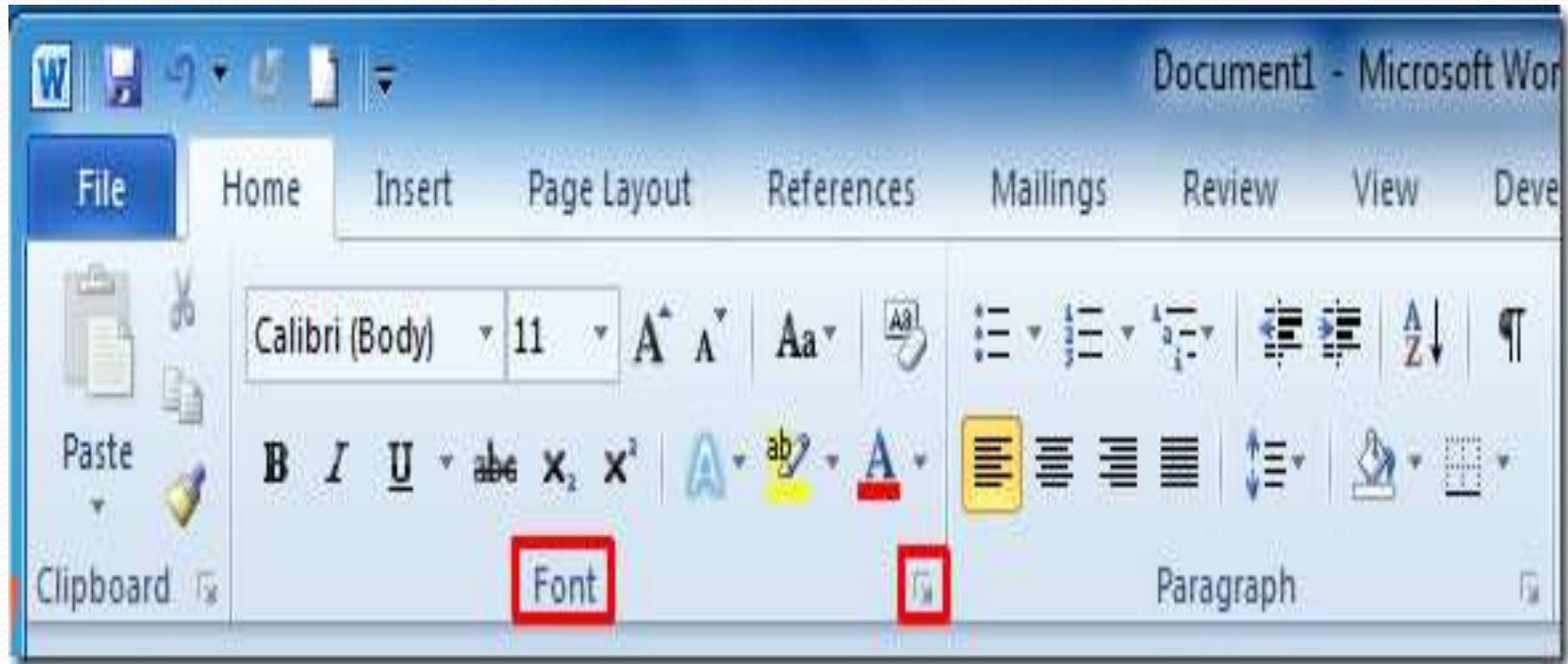
Features of Microsoft Word

➢ **Creating Document.**

➢ **Editing document.**

➢ **Graphics.**

➢ **Word Art.**

➢ **Printing Document.**

Proofing Word Document

- Spelling Checker
- Grammar checker
- Thesaurus
- Auto correct.

Formatting Word Document

- The process to change the appearance of a document is called the document formating you can format a single character, word, lines, paragraph or whole document.

- The document is formated to make it more attractive and beautiful.

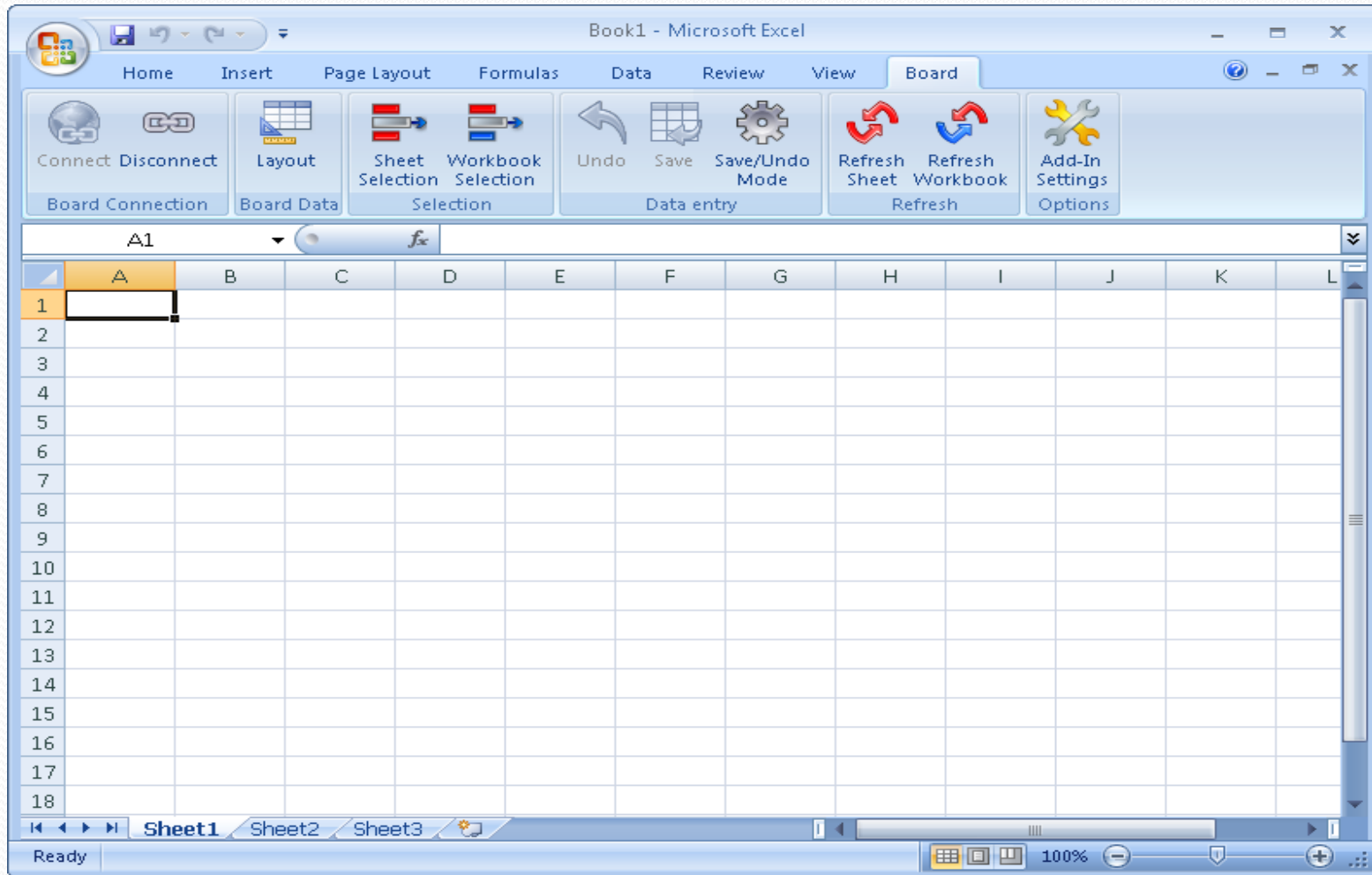- The commands used to format the document are selected from the Home tab.

Creating Tables

- Microsoft word provides a tool called table. It is used to organize information of a rows and columns .

- A table is made of series of rows and columns.

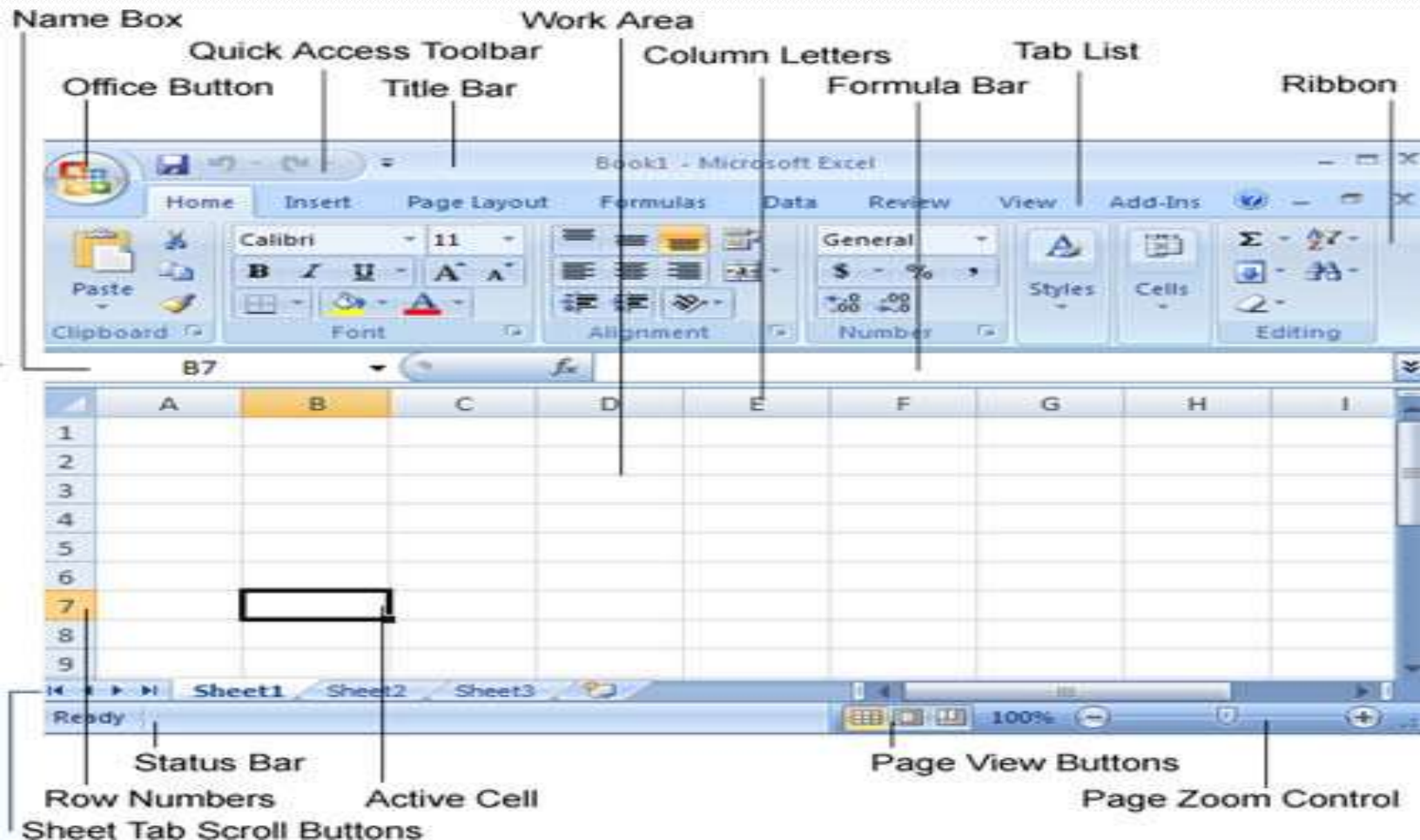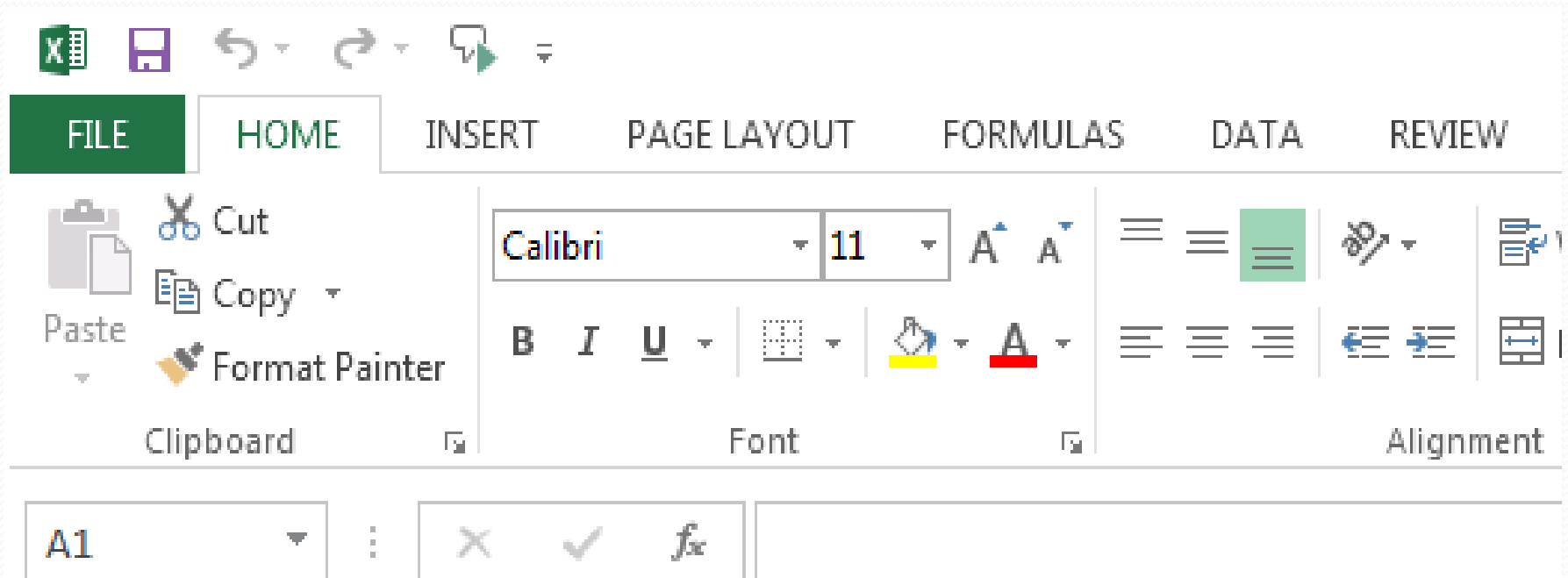- The intersection of row and column is called cell.

MS-EXCEL

Components of MS-Excel

# The Ribbon displays various commands and features of all the tabs.

MS-Excel

- **Excel is an Microsoft Application that is mainly used for calculations and mathematical works.**

a) It is a spreadsheet application in which we can add sheets as per our requirements. In a single sheet, it consists of rows and columns and cells, where every cell has different address.

b) Sum, product, subtraction, division and many mathematical, logical functions are available within it.

c) Other features include tables, charts, clip art and more.

d) It is basically used for payroll, accounts, mathematical, and for other business purposes.

Features of MS-Excel

- **Hyperlink.** We can link one file to another file or page.
- **Clip art.** We can add images and also audio and video clips.
- **Charts.** With charts, we can clearly show a product(s) evaluation to a client. For example, you can display a chart showing which product is selling more or less by month, week, and so forth.
- **Tables.** Tables are created with different fields (e.g. name, age, address, roll number, and so forth). You can add a table to fill these values.
- **Functions.** There are both mathematical functions (add, subtract, divide, multiply), and logical ones (average, sum, mod, product).
- **Images and backgrounds.** You can incorporate images and backgrounds into each sheet.
- **Macros.** Macros are used for recording events for future use.
- **Database:** With the data feature, you can add any database from other sources to it.
- **Sorting and filtering.** We can sort and/or filter our data so that anything redundant or repetitive can be removed more easily.
- **Data validations.** This tool can help you consolidate your data.

- **Grouping.** The grouping feature helps you both to group your data and ungroup it so that you have subtotals and so forth.
- **Page layout.** Themes, colors, sheets, margins, size, backgrounds, breaks, print, titles, sheets height, width, scaling, grids, headings, views, bring to front of font or back alignment, and many more are available for you to lay out your page.

Arithmetic calculation

Graphs And Charts
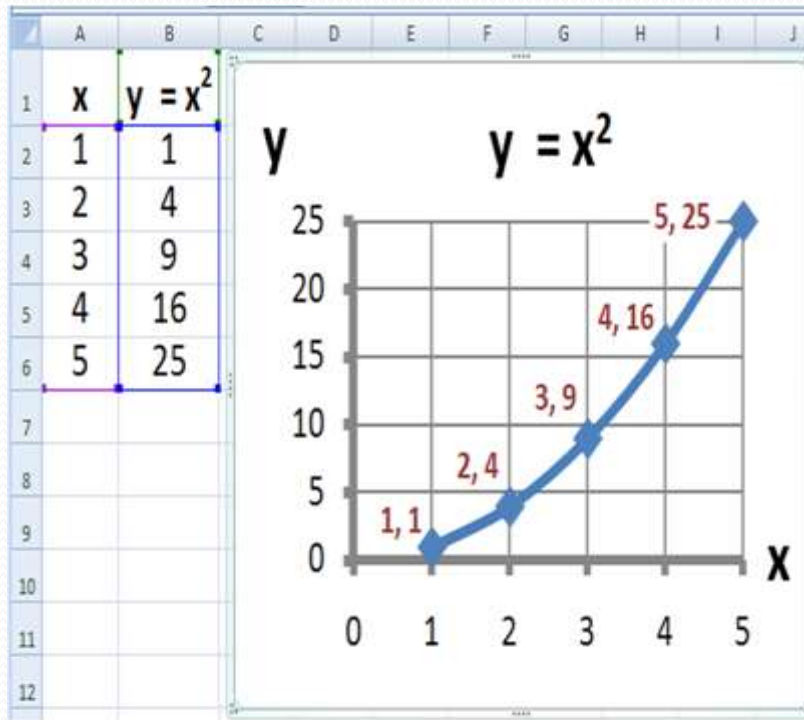
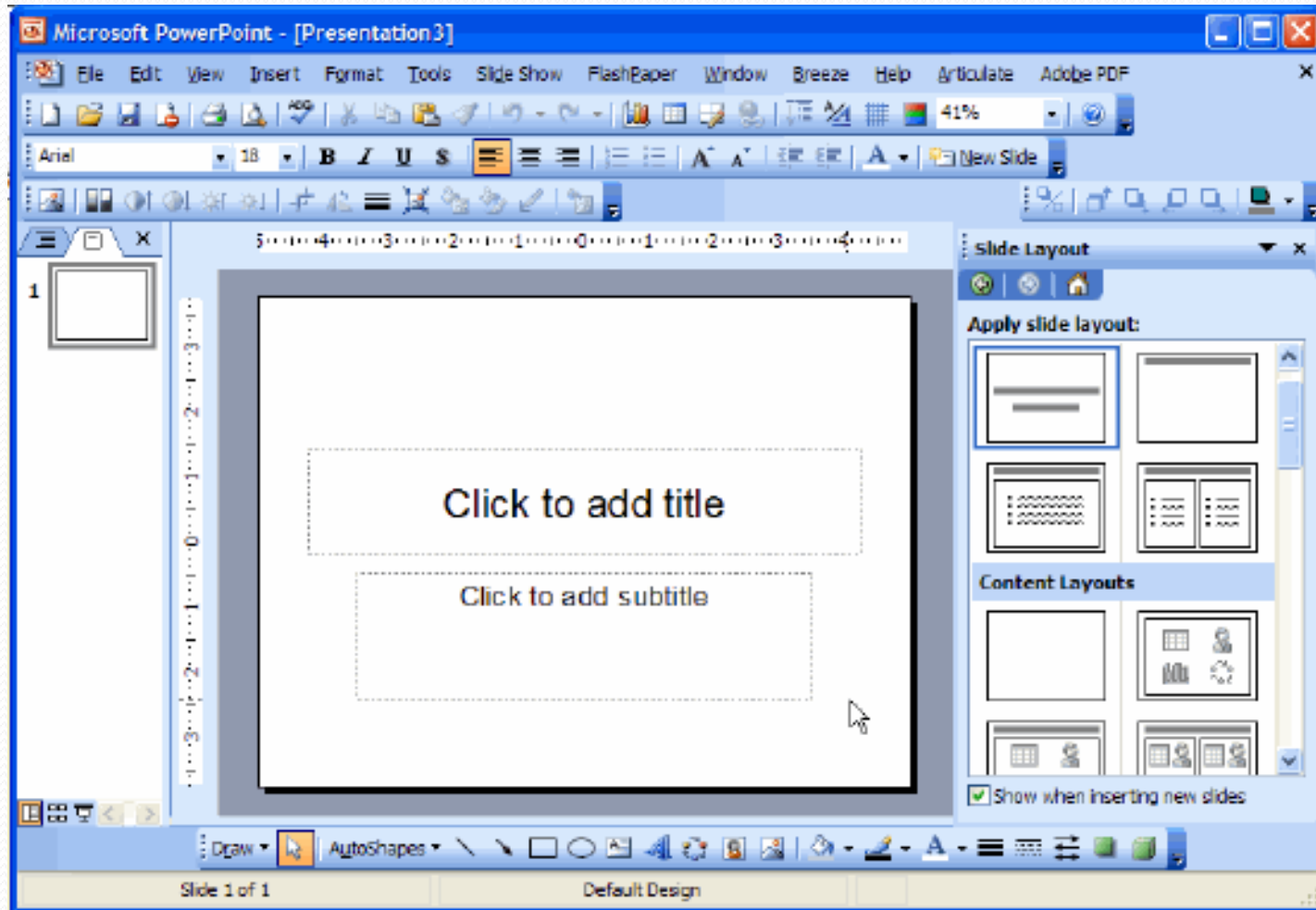# Some Functions of MS-Excel

- Average
- Count
- Pi
- Sum
- Product
- Max
- If

MS-Power point

What is Power Point ?

- **PowerPoint** is computer software created by Microsoft which allows the user to create slides with recordings, narrations, transitions and other features in order to present information. An example of**PowerPoint** is presentation software made by Microsoft.

Features of MS-Power Point

- Animatation
- Auto shapes
- Editing presentation
- Spell checking in presentation
- Hiding and Un-hiding slides
- Running presentation
- Slide transition
- Saving presentation
- Printing presentation

What is MS-ACCESS ?

- MS ACCESS is a tool which used for create database and it is also a application software.

Features of MS-ACCESS

- Database
- Record
- Field
- Table
- Form
- Report
- Primary key

Conclusion

- Simply
- Ms word used for to write documents
- Ms excel used for to make spread sheets
- Ms PowerPoint for presentations
- Ms access for data base management purpose