Chapter 1

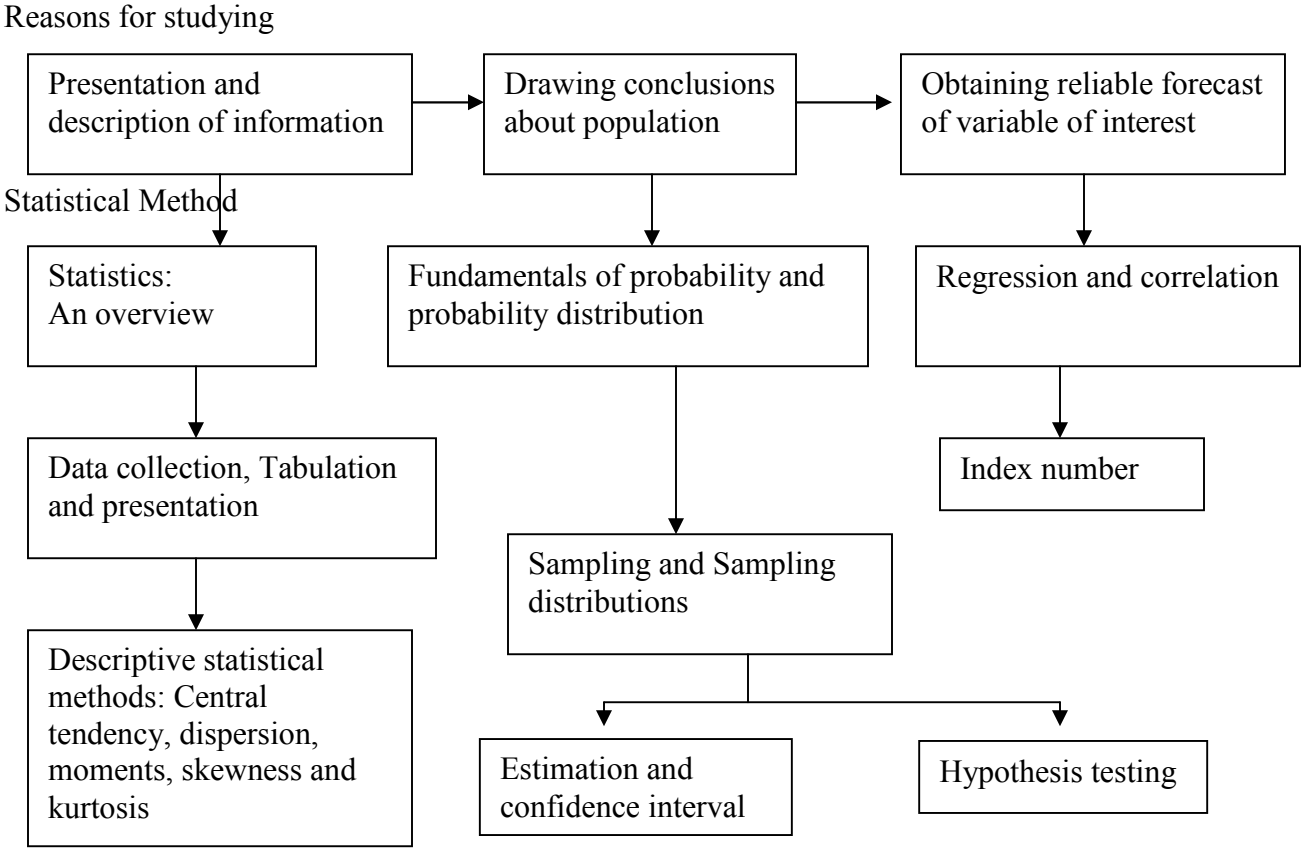
Definition, Importance and Uses of Statistics

Definition of Statistics: Statistics may be defined as a science which deals with collection, organization, classification, presentation, summarization, analysis and interpretation of statistical data in any field of inquiry.

Reasons for learning Statistics: In the context of today's competitive business environment it is important for decision-maker to develop the ability to exact meaningful information from raw data to make better decision. It is possible only through the careful analysis of data guided by statistical thinking. The data analysis shows the variable and its causes in any phenomenon that leads to better decisions about it that produces data. So, the learning of statistics enables the decision-maker to understand how to:

- present and describe information (data) so as to improve decisions
- draw conclusions about the large population based upon information obtained from samples.
- seek out relationship between pairs of variables to improve processes
- obtain reliable forecasts of justified variables of interest

Flow chart of reason of studying Statistics:



Types of Statistical Method:

(i) **Descriptive statistics:** It includes statistical methods involving the collection, presentation and characterization of a set of data. The methods are: graphic, measures of central tendency, measure of dispersion, skewness and kurtosis.

(ii) **Inferential statistics:** It includes statistical methods for estimating the characteristics of a population (parameter) or making decisions concerning a population on the basis of sample results. It can be categorized as parametric and nonparametric. In parametric statistics parent population is considered normal and measured scale is interval and ratio. But in nonparametric statistics there is no assumption of normality and the measured scale is nominal and ordinal.

Importance of Statistics in CSE and IIT: Computer and information technology, in general have had a fundamental effect on most business and service organization. Over the last decades personal computer (PC) has revolutionized both the areas to which statistical techniques are applied. PC facilities such as spreadsheets or common statistical packages have now made such analysis readily available to any business decision-maker. Computer helps in processing and maintaining past records of operations involving payroll calculations, inventory management, railway/airline reservation etc.

Limitation of statistics: Although Statistics has it's application in almost all sciences-social, physical, and natural- it has it's own limitations as well, which restrict it's scope and utility.

a) **Statistics does not study qualitative phenomena**: Since Statistics deals with numerical data, it cannot be applied in studying those problems which can be stated and expressed quantitatively. Qualitative characteristics such as honesty, poverty, welfare, beauty or health cannot be measured directly quantitatively. However, these subjective concepts can be related in an indirect manner to numerical data after assigning particular scores.
b) **Statistics does not study individuals**: Statistics always deals with aggregated data that Statistics always helps for taking decision about a population based on sample information.
c) **Statistics can be misused**: Statistics are liable to be misused. For proper use of statistics one should have enough skill, knowledge, experience to draw accurate and sensible conclusion. Further, valid results cannot be drawn from the use of statistics unless one has a proper understanding of the subject to which it is applied.

Variable: A characteristics which varies individual to individual is called variable. As for example, height, weight, income, number of courses, expenditure etc.

Dr. Mohd. MuziburRahmann
Professor, Department of Statistics, JU

Dummy Variable: The variable which takes only two values either 0 or 1. As for example, yes or no, present or absent etc.

Primary Data: The data which collected from the respondent directly by different survey using questionnaire or schedule is known as primary data.

Secondary Data: The data collected from different published journals or books or other sources is known as secondary data.

Classification of Variable: Variables can be classified in several ways. One method of classification refers to the type and amount of information contained in the data. Data are either categorical or numerical. Another method is to classify data by levels of measurement, giving either qualitative or quantitative.

Categorical (Qualitative) Variables: Categorical variables produce response that belongs to groups or categories. For example, response to yes/no questions are categorical. "Do you own a mobile phone?" and did you ever visit Hiroshima, Japan? Are limited to yes or no answer? Other examples of categorical variables include questions on gender, marital status, and your major in college. Sometime categorical variables include a range of choices such the instructor in this course was an effective teacher (1: strongly disagree 2: slightly disagrees, 3: neither agree nor disagree, 4: slightly agree, 5: strongly agree).

Numerical (Quantitative) Variables: Numerical variables include both discrete and continues variable.

Discrete Variable: A discrete variable may (but does not necessary) have a finite number of values. However, the most common type of discrete that we will encounter produces a response that comes from a counting process. As for example, the number of students enrolled in a class, the number of university credits earned by a student at end of a particular semester and the number of insurance claims filed following a particular hurricane in any particular state.

Continuous Variable: A continuous variable may take any value within a given range of real numbers and usually arises from a measurement (not a counting) process. As for example continuous numerical variables include height, weight, time, distance, temperature etc. Someone might say that he is 6 feet (or 72 inches) tall but his height could be actually be 72.1 inches, 71.8 inches or some other similar number, depending on the accuracy of the instrument used to measure height. Other examples of continuous numerical variables include the weight of cereal boxes, the time to run a race, and the distance between two cities etc.

Measurement and Scaling Concept:

Measurement: Measurement is the assignment of numbers or other symbols to characteristics of objects according to certain pre-specified rules. Note that what we measure is not the object, but some characteristic of it. Thus, we do not measure objects-

Dr. Mohd. MuziburRahmann
Professor, Department of Statistics, JU

only their perceptions, attitudes, preferences, or other relevant characteristics. In research, numbers are usually assigned for one of two reasons. First, numbers permit statistical analysis of the resulting data. Second, numbers facilitate the communication of measurement rules and results. In case of qualitative data there is no measurable to the "difference" in numbers. For one basketball player is assigned the number "20" and another player has the number "10" we can not conclude that the first player is twice as good as the second player. However, with qualitative data, there is a measurable meaning to the difference in numbers. When one student scores 90 on examination and another student scores 45, the difference is measurable and meaningful.

Scale: A Scale may be defined as any series of items that are arranged progressively according to value or magnitude into which an item can be placed according to its quantification. It can be defined as a continuous spectrum or series of categories.

Purpose of Scaling: The purpose of scaling is to represent usual quantitatively an items, a person's or an event's place in the spectrum. The type of scale determines what numerical and statistical operations can be used in analyzing measurements.

Types of Scale: There are four types of scales, such as: (i) Nominal scale (ii) Ordinal scale (iii) Interval scale (iv) Ratio scale.

Nominal Scale: It is a measurement scale of simplest type in which the number or letters assigned to objects serve only labels or tags for identifying and classifying objects with a strict one-to-one correspondence between the numbers and the objects.

Example: In business research if we give the coding of males as 1 and females as 2. These two numbers are nothing but levels.

Ordinal Scale: It is type of scale that arranges objects or alternatives according to their magnitudes in an ordered relationship. When the respondents are ordered, ordinal values are assigned. Thus it is possible to determine whether an object has more of less of a characteristic than some other object.

Example: In business research if we ask to rate companies as excellent, good, fair, poor we know excellent is higher.

Dr. Mohd. MuziburRahmann
Professor, Department of Statistics, JU

Interval Scale: It is another type of scale that not only arranges objects according to their magnitudes but also distinguishes their ordered arrangement in units of equal intervals. An interval scale contains all the information of an ordinal scale, but it also allows you to compare the differences between objects. The difference between any two scale values is identical to the difference between any other two adjacent values of an interval scale. There is a constant or equal interval between scale values.

Example: The classic example is Fahrenheit temperature scale. If a temperature is 80^0 it cannot said that is twice as hot as 40^0. The reason is far that 0^0 does not represent the lack of temperature, but a relative point on the Fahrenheit scale.

Ratio Scale: A ratio scale possesses all the properties of the nominal, ordinal and interval scales and in addition, an absolute zero point. It possesses an absolute zero. Thus, in ratio scales we can identify or classify objects, rank the objects and compare intervals or differences. It is also meaningful to compute ratios of scale values.

Example: Money and weight are ratio because they posses an absolute zero and interval properties.

Mathematical and Statistical Analysis of Scales

Scale	Basic Characteristics	Common Examples	Marketing Examples	Numerical Operation	Permissible Statistics	
					Descriptive	Inferential
Nominal	• Numbers identify and classify objects	• Social Security numbers, numbering of football players	• Brand numbers, store types, sex classification	• Counting	• Percentages, mode	• Chi-square, binomial test
Ordinal	• Numbers indicate the relative positions of the objects but not the magnitude of differences between them	• Quality rankings, rankings of teams in a tournament	• Preference rankings, market position, social class	• Rank ordering	• Percentile, median	• Rank-order correlation, Friedman ANOVA

Dr. Mohd. MuziburRahmann
Professor, Department of Statistics, JU

Interval	• Differences between objects can be compared; zero point is arbitrary	• Temperature (Fahrenheit, Centigrade), IQ score	• Attitudes, opinions, index numbers	• Arithmetic operations on intervals between numbers	• Range, mean, standard deviation	• Product-moment correlations, t-tests, ANOVA, regression, factor analysis
Ratio	• Zero point is fixed; ratios of scale values can be computed	• Length, weight	• Age, income, costs, sales, market shares	• Arithmetic operations on actual quantities	• Geometric mean, harmonic mean	• Coefficient of variation

Frequency distribution:

A frequency distribution is a table used to organize data. A frequency distribution divides observations in the data set into conveniently established, numerically ordered classes (groups or categories). The number of observations in each class is referred to as frequency.

Constructing a Frequency Distribution:

If the variation within the data set is not so wide, then it is wide to construct ungrouped frequency distribution for summarizing data. If the number of observations obtained gets large, the method discussed above to summarize data become difficult and time consuming. Thus to further summarizing the data into group frequency distribution tables, the following steps should be taken:

i) select an appropriate number of non-overlapping class intervals
ii) determine the width of the class interval
iii) determine class limits (or boundaries) for each class interval to avoid overlapping.

For constructing the frequency distribution, we need the following steps:

Step 1: **Determine range**: From the given data set, find out the lowest value and the highest value. Then range is the difference between highest value and lowest value.

Step 2: Determine the number of class: If K determine the number of classes and N the total number of observations, then the value K will be the smallest exponent of number 2, that is $2^k \geq N$.
Another way to find the value of K by using Sturge's rule is given by: $K = 1 + 3.322 \log_{10}\mathrm{N}$, where $\log_{10}N$ is the logarithm (base 10) of total number.

Step 3: **Determine the width or the interval of classes**: For constructing the frequency distribution determine the suitable class interval *i,*

Dr. Mohd. MuziburRahmann
Professor, Department of Statistics, JU

$$i=\frac{\text{Range}}{K}=\frac{Range}{1+3.322\ \log_{10}N}$$

Both *K* and *i* should be rounded upward, possible to the next longest integer.

Step 4: Determine the class limits (boundaries): The limits of each class interval should be clearly defined so that each observation (element) of the data set belongs to one and only one class. The class interval must be inclusive and non-overlapping such as 20-29, 30-39, etc. Sometimes we also need exclusive types of class, where upper limit of each classes are excluded from the each class (such as 20-30, 30-40, 40-50 etc.)

Step 5: Mid-point of class interval: The class mid-point is the point halfway between the boundaries of each class. That means, it is the average of upper limit and of lower limit of each classes.

Step 6.Tally marks: Now each and every observations of the data set are matched with the respective classes and put a tally for every observation, after completing the whole data set, the tallies of every class are added and put it on corresponding classes. This is known as frequencies.

Step 7: Cumulative frequency (less than): If you add the frequencies of each classes with next class from the top in a cumulative form then it is known as cumulative frequencies less than. If do the same thing from the bottom then it is known as cumulative frequencies more than. If we divide the frequencies of each class by the total frequencies then it is known as relative frequency.

Example 1: Following data shows the total time (in hours) work by 30 machinists. Construct a frequency distribution.

90	88	90	89	90	84	86	90	84	89	93	84	90	94	91
94	93	93	92	92	85	88	86	91	87	94	89	85	90	95

Solution: Here the variations among the data set are not vary wide, so we construct a ungroup frequency distribution as follows:

Table 1: Frequency distribution of total time hours work by 30 machinists.

Working hours	84	85	86	87	88	89	90	91	92	93	94	95
No. of Employee	3	2	2	1	2	3	6	2	2	3	3	1

Dr. Mohd. MuziburRahmann
Professor, Department of Statistics, JU

Example 2: Following data shows the weekly overtime (in hours) of 50 employees in a reputed fashion design company. Construct a frequency distribution by taking suitable class interval.

22	77	79	82	65	50	65	73	60	33	75	66	65	30	63	41	55
65	67	62	45	49	75	59	55	54	51	28	39	25	50	48	68	55
81	35	65	65	79	61	45	53	81	49	37	57	78	27	87	77	

Solution: Here, $2^6 \geq 50$, so the value of number of classes is 6. And the range of the data is 22-87=65. Therefore the width of the class inter is $10.83 \cong 11$. On the other hand, we know that, suitable class interval $i = \frac{Range}{1+3.322log_{10}50} = \frac{65}{1+3.322\times1.6989}$=$9.7837 \cong 10$

Here the nearest value of the width is 10 (we select it multiple of 5).

Overtime (in hours)	No. of employee
20-30	4
30-40	5
40-50	6
50-60	10
60-70	13
70-80	8
80-90	4

Example: The management of a factory wants to know per month working pattern of workers of their factory. In this connection, a survey was conducted on randomly selected 48 workers of the factory. Following data give the number of hours work per month of the 48 workers of the factory.

140	165	103	110	130	144	133	204	175	156	187	195
162	161	167	184	151	149	157	124	87	71	79	155
164	40	94	113	108	146	122	87	69	164	116	203
121	128	149	148	30	93	114	104	150	62	143	42

Construct a frequency distribution by using suitable class interval.

Describing Data: Graphical:

Once we carefully define a problem, we will need to collect data for making decision. Often the number of observations collected is so large that the actual findings of the study are unclear. For this reason, it is necessary to summarize data in such a way that a clear and accurate picture emerges. Unfortunately, there is no single method or way to describe data. Rather, the appropriate line is typically problem-specific, depending on two factors,

Dr. Mohd. MuziburRahmann
Professor, Department of Statistics, JU

the type of data and the purpose of the study. Tables and graphs help us to gain a better understanding of data and provide visual support for improved decision making.

Use of graphs: Following are the uses of graphs:

(i) It is helpful in explaining the main features of a set of data
(ii) It is often valuable in suggesting an appropriate method of analysis and in explaining the conclusions founded upon the analysis.
(iii) It can sometimes pinpoint gross errors in statistical records.

Basic principle of graphs:

(i) A graph should be clear and simple; a complicated graph defeats its own purpose.
(ii) A graph should be completely self explanatory.
(iii) The origin, the vertical and the horizontal scales should be so chosen that a graph does not convey a false impression about the nature of the data.

Limitation of graphs:

(i) They may be misleading, unless drawn and studied with care.
(ii) The conclusions drawn from the graphs should normally be regarded as tentative and therefore, the graphs are no substitute for more critical statistical analysis.

Types of diagrams:

(i) bar diagram, (ii) pie diagram, (iii) histogram, (iv) frequency polygon, (v) line diagram, (vi) ogive curve (vii) scatter diagram.

Bar diagram and pie diagram are mainly used for representing qualitative data. The former is also frequently used for depicting numerical values of a given item over a period of time. Histogram, frequency polygon and cumulative frequency polygon are used to represent frequency distributions. Line diagram is widely used to study the changes in the values of a variable with the passage of time. Scatter diagram is very useful in studying the interrelationship of two variables. .

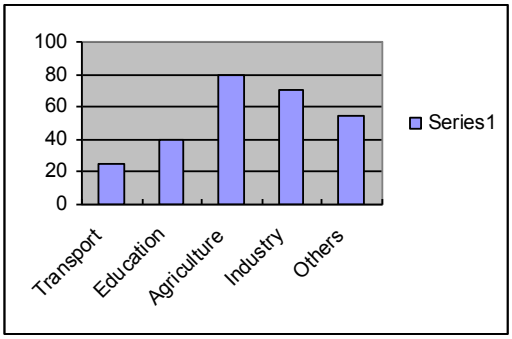
Bar Diagram: This diagram is drawn by constructing a series of blocks of equal widths but the heights of the blocks or rectangles is proportional to the values corresponding to different time period or categories. Following Table shows the distribution of the expenditure budget (in core taka) of different sector of country in the year 2012 as follows:

Sector of Expenditure	Transport	Education	Agriculture	Industry	Others
Expenditure (in core	25	40	80	70	55

Dr. Mohd. MuziburRahmann
Professor, Department of Statistics, JU

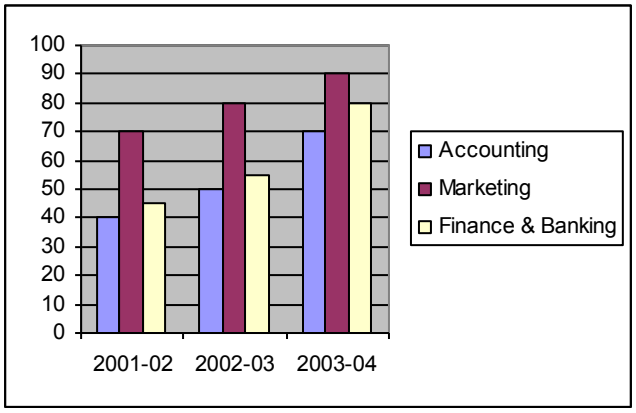
taka)					

Now if we put the categories (sector) in the x-axis and the expenditure in y-axis then the diagram will be a bar diagram where the width of the bars are equal but the heights are proportional to the expenditure of the sectors.



An interesting and useful extension to the simple bar chart can be used when components of individual categories are also of interest. As for example, following Table shows the number of students enrolled in three business majors for three different years of BUP.

Subject	2001-02	2002-03	2003-04
Accounting	40	50	70
Marketing	70	80	90
Finance & Banking	45	55	80



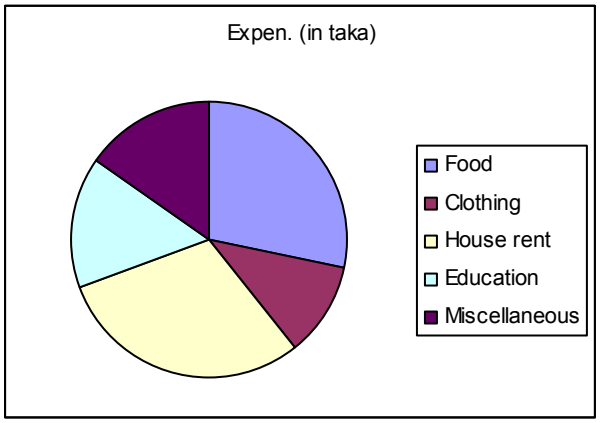
This information can be shown in a bar chart by breaking down the total number of students for each year so that the three components are distinguished by differences called components or bar chart. This graph allows us at make visuals comparisons of totals and individual components. In this example it appears that the increase in enrollment between 2001 and 2004 was almost uniform over the three majors.

Pie diagram: Pie charts are also used to describe categorical data. If we want to draw attention to the proportion of frequencies in each category, then we will probably use a

Dr. Mohd. MuziburRahmann
Professor, Department of Statistics, JU

pie chart to depict the division of a whole into its constituent parts. The circle represents the total, and the segments cut from its center depict shares of that total. Following Table shows the distribution of monthly expenditure of the students of BUP.

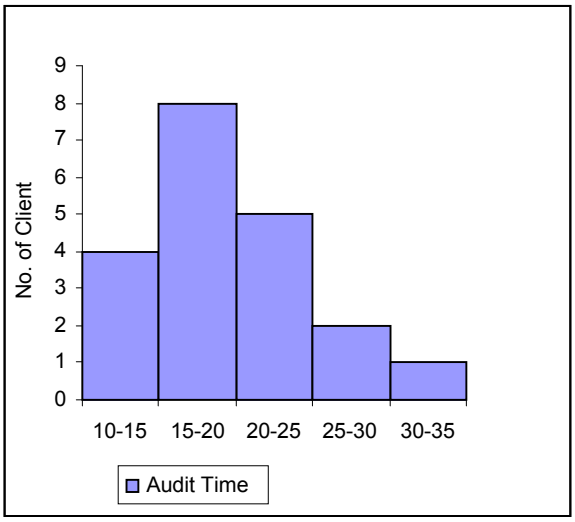
Item of Expenditure	Expenditure (in taka)
Food	6500
Clothing	2500
House rent	7000
Education	3500
Miscellaneous	3500



Histogram: It is a graphical method for representing a frequency distribution. To construct this diagram the horizontal axis (x-axis) is divided into segments corresponding to the class boundaries of the frequency distribution. On each segment a rectangle with area proportional to the frequency in the class is erected. The set of adjacent rectangles so constructed, constitutes a histogram. Following distribution shows the audit time of 20 clients by exclusive method:

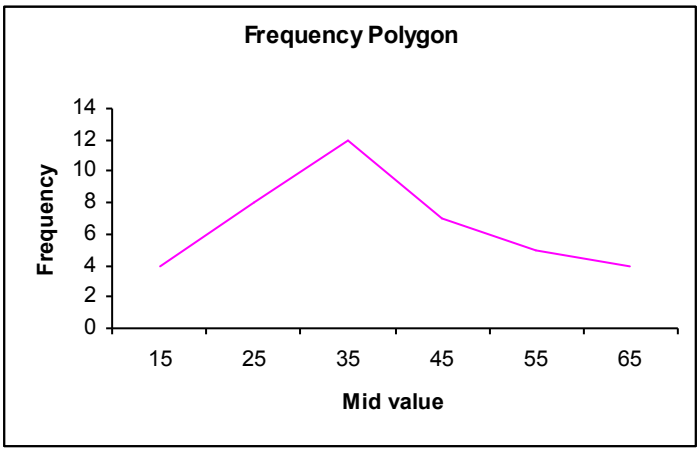
Audit time (in hours)	10-15	15-20	20-25	25-30	30-35
Number of clients	4	8	5	2	1

Dr. Mohd. MuziburRahmann
Professor, Department of Statistics, JU



Frequency Polygon: It is a diagram used to represent a frequency distribution. The mid-values of class intervals are plotted along the x-axis and corresponding frequencies are plotted along the y-axis. These later points are then joined by straight lines. This forming with the x-axis a polygon called frequency polygon. The frequency polygon should be brought down-at each end to the x-axis by joining it to mid value (on the base line) of the next outlying interval.

Audit time (in hours)	10-20	20-30	30-40	40-50	50-60	60-70
Number of clients	4	8	12	7	5	4

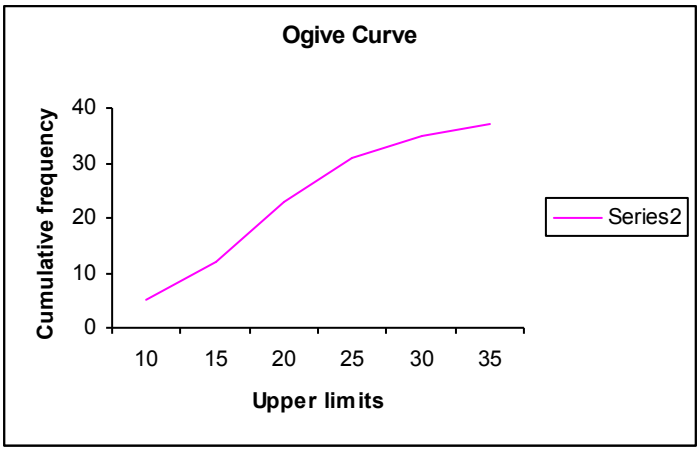


Ogive (less than): In the X axis we plot upper limit of the class and in Y axis we plot cumulative frequency less than.

Class interval	5-10	10-15	15-20	20-25	25-30	30-35

Dr. Mohd. MuziburRahmann
Professor, Department of Statistics, JU

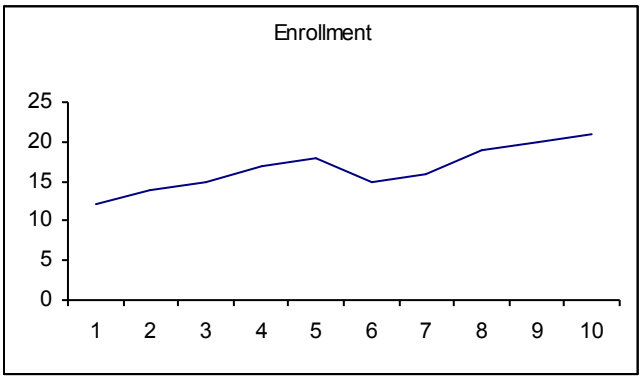
Frequency	5	7	11	8	4	2
Cumulative frequency	5	12	23	31	35	37



Graphs to Describe Time-series Data:

Line diagram: If we are given the values of a variable at different point of time, the set of values is known as a time series. The line diagram is used to represent this type of data. In this diagram time is represented along the x-axis and the variable is plotted along the y-axis. Thus we get a point for each time period and successive points, when it connected by straight line, gives the desired diagram.

Year of enrollment	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
No. of student (‘00)	12	14	15	17	18	15	16	19	20	21



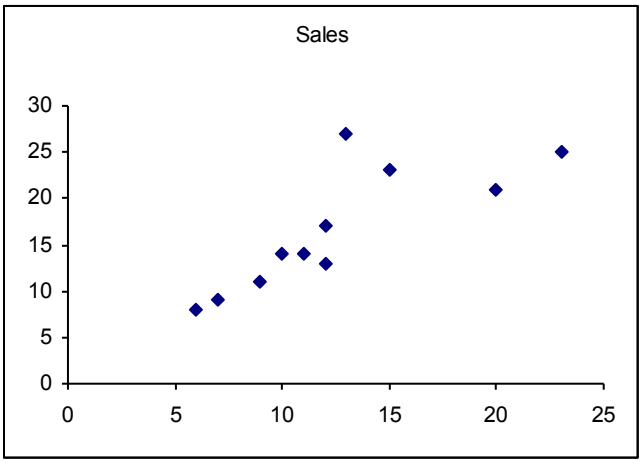
The situations in which the line diagram is particularly useful are:

Dr. Mohd. MuziburRahmann
Professor, Department of Statistics, JU

a) When the emphasis is on the movement of a variable rather then on it's actual magnitude.
b) When several series are compared on the same chart.
c) When estimates or forecasts of a variable are to be obtained or displayed graphically.

Scatter diagram: Sometimes the data consist of pair values of two related variables, and the statistical problem is to investigate the inter-relationship between the variables. The pairs of values of such related variable are: height and weight, income and expenditure, price and consumption etc. When the given pair of values is plotted on ordinary graph paper, we get a scatter diagram. If the dotted points form an upward trend on the graph paper then the relationship between the variable is positive. If it forms a down ward trend on the graph paper then the relationship between two variables is negative.

Expense of Ad.	10	12	15	20	23	9	6	7	11	12	13
Sales (in lac)	14	17	23	21	25	11	8	9	14	13	27



Stem-and-leaf display: Stem and leaf display is another form of presentation of quantitative data. It allows us to condense data, but still retain the individuality of the data. This presentation shows the range, concentration, presence of outlier, if any, and distribution of the data set at a glance.

The stem of an observation is the leading digit or digits and the leaf of an observation is the trailing digit. All the values in the stem are listed in order in a column, a vertical line is drawn beside them and then all the corresponding leaf values are recorded for each stem in row, to right of vertical line.

Steps for construction of stem and leaf plot or display:

a) Divide each observation into two parts: the stem and the leaf.
b) List the leaf in a column, with a vertical line to their right.
c) For each observation, record the leaf portion in the same raw as its corresponding stem.
d) Order the leaves from lowest to highest in each stem.

Dr. Mohd. MuziburRahmann
Professor, Department of Statistics, JU

e) Mention the leaf unit to understand the actual observation.

Example: The prices (in taka) of 20 different brand of walking shoes are given below:

45	70	70	55	75	73	70	65	68	60	74	80	83	58	68	85	90	64	75	82

Construct a stem and leaf plot to display the distribution of the data.

Solution: The stem and leaf display of the data is follows:

Stem	**Leaf**
4	5
5	5 8
6	5 8 0 8 4
7	0 0 5 3 0 4 5
8	0 3 5 2
9	0

Now we arrange the digits of each leaf in ascending order we get:

Stem	**Leaf**
4	5
5	5 8
6	0 4 5 8 8
7	0 0 0 3 4 5 5
8	0 2 3 5
9	0

From the display it is seen that lowest price of walking shoe is 45 and highest is 90. And the most common price is 70.

Dr. Mohd. MuziburRahmann
Professor, Department of Statistics, JU