2 Md. Abdul Alim 20-07-2010

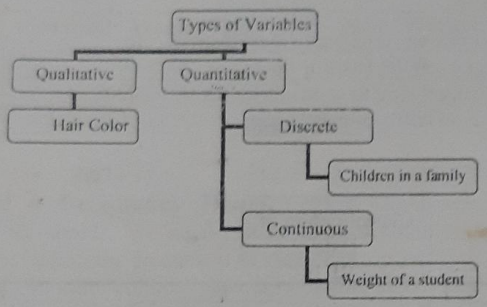
Chapter: Processing of Data

Variable: If we observe a characteristic, we find that it takes on different values in different persons, place or things; we label the characteristic as a variable. We do this for the simple reason that the characteristic is not the same when observed in different possessors of it.

Example

i. The heights of adult males
ii. The weights of pre-school children

Types of Variable



There are two basic types of variables

1. Qualitative and 2. Quantitative

Qualitative Variable or an attribute: When the characteristic being studied is nonnumeric, it is called a qualitative variable or an attribute.

Examples: Qualitative variables are gender, religious, affiliation, type of automobiles owned, state of birth and eye color.

When the data are qualitative, we are usually interested in how many or what proportion fall in each category. For example, what percent of the population has blue eyes? How many Catholics and how many Protestants are there in the United States?

Quantitative Variable: When the variable studied can be reported numerically, the variable is called a quantitative variable.

Examples: Quantitative variables are the balance in your checking account, the ages of company presidents, the life of an automobile battery (such as 42 months) and the number of children in a family.

Quantitative variables are either discrete or continuous.

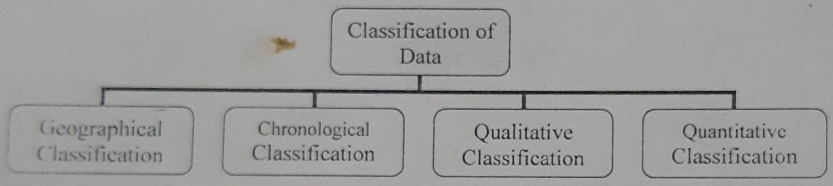
i. **Discrete Variables:** Discrete variables can assume only certain values, and there are usually "gaps" between the values. Examples of discrete variables are the number of bedrooms in a house (1,2,3,4 etc), the number of cars arriving and the number of students in each section course (25 in section A, 42 in section B and 18 in section C).

ii. **Continuous Variables:** Observations of a continuous variable can assume any value within a specific range. Examples of continuous variables are in the air pressure in a tire and the weight of a shipment of tomatoes.

Classification of Data: After collection and editing of data an important step towards processing the data is classification.

Types of Classification: Broadly, the data can be classified on the following four basis:

i. Geographical, i.e. area-wise, e.g., cities, districts, etc
ii. Chronological, i.e. on the basis of time
iii. Qualitative, i.e. according to some attributes
iv. Quantitative, i.e. in terms of magnitudes.



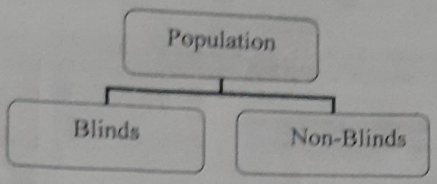
i. **Geographical Classification:** In geographical classification data are classified on the basis of geographical or locational differences between the various items. For example, when we present the production of sugarcane, wheat, rice etc., for various states, this would be called geographical classification.

ii. **Chronological Classification:** When data are observed over a period of time the type of classification is known as chronological classification. For examples, the sales figures of a company are given below:

Year	Sales (Tk. lakhs)
2001	18810
2002	23601
2003	23816
2004	32435
2005	39343

iii. **Qualitative Classification:** In qualitative classification, data are classified on the basis of some attribute or quality such as sex, color of hair, literacy, religion etc. The point to note in this type of classification is that the attribute under study is blindness, we may found out how many persons are blind in a given population.

Md. Mowazzem Hossain
Lecturer, Dept. of Statistics, U



Quantitative Classification: Quantitative classification refers to the classification of data according to some characteristics that can be measured, such as height, weight, income, sales etc. For examples, the workers of a factory may be classified according to wages as follows:

Monthly Wages (Tk.)	No of Workers
2500-2600	50
2600-2700	200
2700-2800	260

Formation of a Frequency Distribution: The process of preparing this type of distribution is very simple. We have just to count the number of times a particular value is repeated which is called the frequency of that class. In order to facilitate counting, prepare a column of "tally". In other column, place all possible values of the variable from the lowest to the highest. Then put a bar (vertical line) opposite the particular value to which it relates.

We finally count the number of bars corresponding to each value of the variable and place it in the column of the frequency.

Example: The number of refrigerators sold on 22 working days by a leading agency house:

23 30 20 26 30 20 23 40 40 26 20 30

23 40 28 26 23 40 28 28 30 30

Frequency distribution of the number of refrigerators sold

No. of Refrigerators	Tally	Frequency (No. of Days)
20	III	3
23	IIII	4
26	III	3
28	III	3
30	~~IIII~~	5
40	IIII	4

Md. Moyazzem Hossain
Lecturer, Dept. of Statistics, IU

The table clearly shows that on 3 days 20 refrigerators were sold each day, on 4 days 23 refrigerators were sold each day etc.

This method of classification helps in condensing the data only where values are largely repeated, otherwise there will be hardly and condensation. In order to make the series more compact so that its characteristics can be easily studied, data may be classified according to class- intervals.

Cumulative Frequency, Relative Frequency and Relative Cumulative Frequency: In some situations, we may be interested, not in the frequencies in various classes, but rather in the frequencies or proportions of observation which are "less than" or "greater than" a given value. This leads to a cumulative frequency distribution. This is derived from a frequency distribution by forming a cumulative frequency column. This column is computed by adding the successive class frequencies from top to bottom. The entry corresponding to the top interval is the frequency of that class., the entry opposite the second interval is the sum of the frequencies in first and second class intervals etc. and so on.

If we divide frequency by N, the total number of observations, we get the relative frequencies. Also, if we divide cumulative frequency by N, the total number of observations, we get the relative cumulative frequencies, which are often expressed in percentage.

Value	f	c. f	Relative f	Relative c. f
0-10	4	4	4/96	4/96
10-20	12	16	12/96	16/96
20-30	24	40	24/96	40/96
30-40	36	76	36/96	76/96
40-50	20	96	20/96	96/96

Classification according to class intervals: This type of classification is most popular in practice. The following technical terms are important when data are classified according to class intervals:

i. **Class limits:** The class limits are the lowest and the highest values that can be included in the class. For example, take the class 20-24. The lowest value of this class is 20 and the highest 40. The two boundaries of a class are known as the lower limit and upper limit of the class. The lower limit of a class is the value below which there can be no value in that class. The upper limit of a class is the value above which no value can belong to that class. Of the class 70-89, 70 is the lower limit and 89 is the upper limit. i.e. in this class there can be no value which is less than 70 or more than 89.

ii. **Class intervals:** The span of a class, that is, the difference between the upper limit and lower limit, is known as class interval. For example, in the class 20-40, the class interval is 20 (i.e. 40 minus 20). The size of the class interval is determined by the number of the classes and the total range in the data.

iii. **Class frequency:** The number of observations corresponding to the particular class is known as the frequency of that class or the class frequency.

Md. Moyazzem Hossain
Lecturer, Dept. of Statistics, JU

iv. **Class mid-point:** It is the value lying half-way between the lower and the upper class limits of a class interval. Mid point of a class is ascertained as follows:

$$\text{Mid point of a class} = \frac{(\text{Upper limit of the class} + \text{Lower limit of the class})}{2}$$

Methods of classifying the data according to class interval: There are two methods of classifying the data according to class intervals namely

a. Exclusive method
b. Inclusive method

a. **Exclusive Method:** When the class intervals are so fixed that the upper limit of one class is the lower limit of the next class it is known as the 'Exclusive' method of classification. The following data are classified on the basis:

Income (Tk.)	No of Employees
1800-1900	50
1900-2000	100
2000-2200	200

It is clear that 'Exclusive method' ensures continuity of data inasmuch as the upper limit of one class is the lower limit of the next class. Thus in the above example, there are 50 persons whose income is between Tk. 1800 and Tk. 1899.99. A person who is getting exactly Tk. 1900 would be included in the class 1900-2000.

b. **Inclusive method:** Under the "Inclusive method' of classification, the upper limit of one class is included in that class itself.

Income (Tk.)	No of Employees
800-899	50
900-999	100
1000-1099	200

In the class 800-899 we include persons whose income is between Tk 800 and Tk. 899. If the income of persons is exactly Tk. 900 he is included in the next class.

Principles of Classification: It is difficult to lay down any hard and fast rules for classifying the data as the type of classification.

- The number of classes should preferably be between 5 and 15. However, there is no rigidity about it. The classes can be more than 15 depending upon the total number of observations in the series and the details required, but they should not be less than five because in that case the classification may not reveal the essential characteristics.

Struges suggested the following formula for determining the approximate number of classes:

$K = 1 + 3.322 \log N$ where, K = The approximate number of classes, N = Total number of observation and Log = The ordinary logarithm to the base of 10.

Md. Moyazzem Hossain
Lecturer, Dept. of Statistics, U

However, the precise number of classes to be used for a given variable depends upon personal judgment and other considerations such as the details required.

- As far as possible one should avoid odd values of class intervals e.g. 3, 7, 11, 26, 39 etc. Preferably, one should have class intervals of either five or multiples of five like 10, 20, 25, 100 etc.

- The starting point, i.e. the lower limit of the first class, should either be zero or 5 or multiple of 5. For example, if the lowest value of the series is 63 and we have taken a class interval of 10, then the first class should be 60-70, instead of 63-75. Similarly, if the lowest value of the series is 76 and the class interval is 5 then the first class should be 75 to 80 rather than 76 to 81.

Example: The profits (in lakhs of Tk's) of 30 Bangladeshi companies for the year 2008-2009 are given below:

18	16	23	37	35	49	63	65	55	29
45	58	57	69	20	22	35	42	37	58
42	48	53	49	65	39	48	67	25	65

Classify the above data taking a suitable class interval.

Solution: Let us determine the suitable class interval with the help of the following formula:

$$i = \frac{Range}{K}$$ where, K = 1+3.322logN and Range = Highest value - Lowest value

We have, N = 30, Highest value = 69, Lowest value = 16

K=1+3.322log30=5.91 ≈ 6, Range = 69-16 = 53

$$i = \frac{Range}{K} = \frac{53}{5.91} = 8.97 \text{ or } 9$$

Since values like 3, 7, 9 etc. should be avoided we will take 10 as the class interval and the first class be 15-25.

Frequency Distribution of the profits

Profit (Tk. lakhs)	Tally	No. of companies
15-25	𝍸	5
25-35	II	2
35-45	𝍸 II	7
45-55	𝍸 I	6
55-65	𝍸	5
65-75	𝍸	5
Total		30

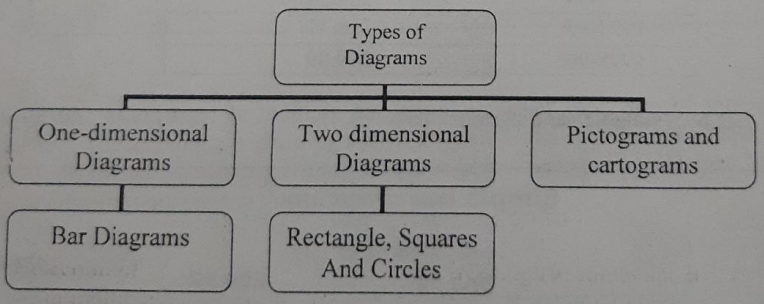
Charting Data: A chart can take the shape of either a diagram or a graph. For the sake of clarity we will discuss them under two separate heads:

- Diagrams
- Graphs

Diagrams: For representing data diagrams are more commonly used than graphs.

General Rules for Constructing Diagrams

1. **Title:** Every diagram must be given a suitable title. The title should convey in as few a words as possible the main idea that the diagram is intended to portray.
2. **Proportion between width and height:** A proper proportion between the height and width of the diagram should be maintained. If either the height or width is too short or too long is proportion, the diagram would give an ugly look.
3. **Selection of appropriate scale:** The scale showing the values should be in even numbers or in multiples of five or ten e.g. 25, 50, 75 or 20, 40, 60. Odd values like 1, 3, 5, and 7 should be avoided.
4. **Footnotes:** In order to clarify certain points about the diagrams footnotes may be given at the bottom of the diagram.
5. **Index:** Index illustration different types of lines or different shades, colors, should be given so that the reader can easily make out the meaning of the diagram.
6. **Neatness and cleanliness:** Diagrams should be absolutely neat and clean
7. **Simplicity:** Diagrams should be as simple as possible so that the reader can understand their meaning clearly.



One- dimensional or Bar Diagrams: Bar diagrams are the most common type of diagrams used in practice. A bar is a thick line whose width is shown merely for attention. They are called one-dimensional because it is only the length of the bar that matters and not the width.

Points to be kept in mind while constructing Bar Diagrams

- The width of the bars should be uniform throughout the diagram.
- The gap between one bar and another bar should be uniform throughout.
- Bars may be either horizontal or vertical. The vertical bars should be preferred because they give better look and also facilitate comparison.
- While constructing the bar diagrams, it is desirable to write the respective figure at the end of each bar so that the reader can know the precise value without looking at the scale.

Types of Bar Diagrams

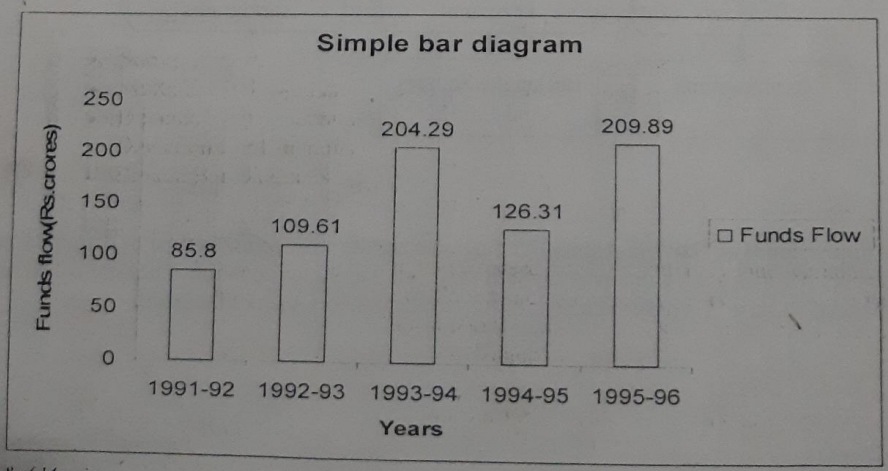
- Simple Bar Diagrams
- Sub-divided Bar Diagrams
- Multiple Bar Diagrams
- Percentage Bar Diagrams
- Deviation Bar Diagrams
- Broken Bar Diagrams

Simple Bar Diagrams: A simple bar diagram is used to represent only one variable. For example the figures of sales, production, population etc, for various years may be shown by means of a simple bar diagram. However, an important limitation of such diagrams is that they can present only one classification or one category of data.

Example: The funds flow of Goodwill India Ltd from 1991-92 to 1995-96 are given below:

Year	Funds Flow (Rs. Crores)
1991-92	85.80
1992-93	109.61
1993-94	204.29
1994-95	126.31
1995-96	209.89

Represent this data by a suitable bar diagram.

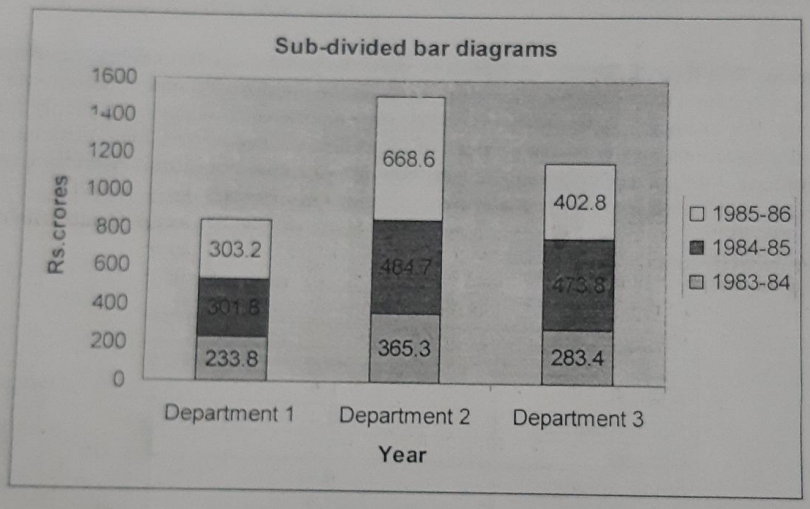


Md. Moyazzem Hossain

Sub-divided Bar Diagrams: These diagrams are used to represent various parts of the total. For example, the number of employees in various departments of a company may be represented by a sub-divided bar diagrams. While constructing such a diagram the various components in each bar should be kept in the same order. To distinguish between the different components, it is useful to use different shades or colors. Sub-divided bar diagrams can be vertical as well as horizontal.

Example: Represent the following data by sub-divided bar diagrams

Year	Department 1	Department 2	Department 3
1983-84	233.8	365.3	283.4
1984-85	301.8	484.7	473.8
1985-86	303.2	668.6	402.8

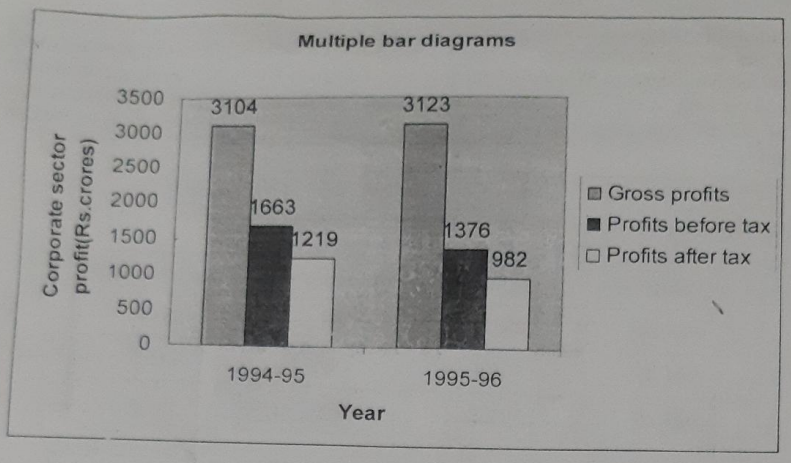


Multiple bar Diagrams: In multiple bar diagram two or more sets of inter-related data are represented. The technique of drawing such a diagram is the same as that of simple bar diagram. The only difference is that since more than one phenomenon is represented, different shades, colors or crossings are used to distinguish between the bars.

Example

	Corporate Sector Profits(Rs.crores)	
	1994-95	1995-96
Gross profits	3104	3123
Profits before tax	1663	1376
Profits after tax	1219	982

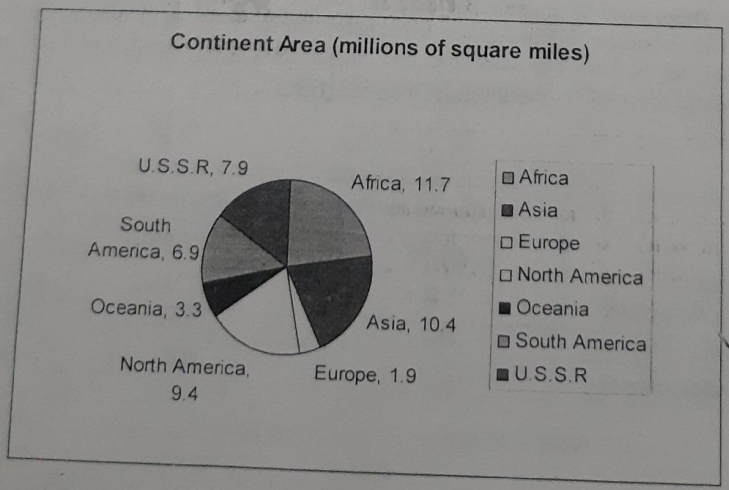
Md. Moyazzem Hossain
Lecturer, Dept. of Statistics, IU



Pie Diagram: This type of diagram enables us to show the portioning of a total into component parts. A very common use of the pie chart is to represent the division of a sum of money into its components. For example, the entire circle or pie, may represent the budget of a family for a month and the sections may represent portions of the budget allotted to rent, food, clothing and so on. Similarly, through a pie diagram we can show how a rupee by a firm is distributed over various heads such as wages, raw materials, administration expenses etc.

Example: Areas of continents of the world

Continent	Area (millions of square miles)
Africa	11.7
Asia	10.4
Europe	1.9
North America	9.4
Oceania	3.3
South America	6.9
U.S.S.R	7.9
Total	51.5



Md. Moyazzem Hossain
Lecturer, Dept. of Statistics, H

The pie diagram is intended to compare the distinct components which together constitute a whole. The whole is represented by a circle of arbitrary radius and the segments of the circle represent the component parts. To construct such a diagram we use the fact "the whole" (51.5 in the above illustration) corresponds to the total number of degrees in the circular arc, namely 360^0. This 360^0 is then proportionately divided among the various components of the whole. Thus the above illustration; the arc of the segment representing

Asia subtends on angle of $73^0 \left(= \frac{360^\circ}{51.5} \times 10.4\right)$ at the centre of the circle.

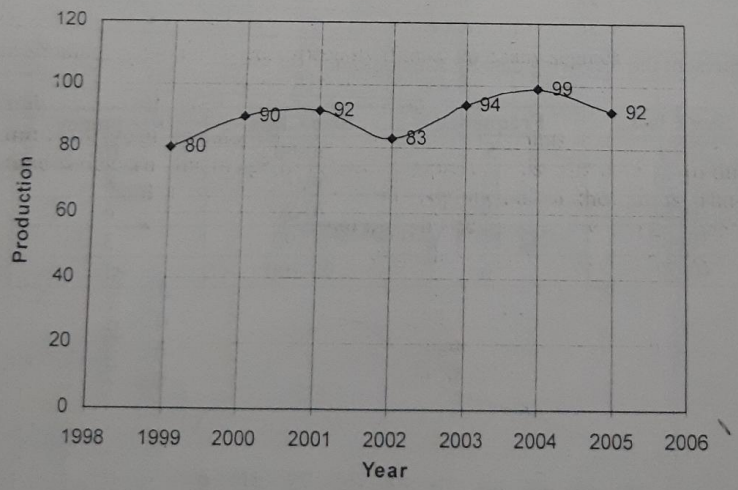
This diagram should be sparingly used, especially if there are many segments.

Line Diagram: If we are given values of a variable at different points 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 connected by straight lines, give the desired diagram. Often smooth curve is drawn through these points. This diagram is alternatively called a line diagram or a time series graph.

Example: The productions (in thousand quintals) of a sugar factory are given below:

Year	Production (in '000 qtl)
1999	80
2000	90
2001	92
2002	83
2003	94
2004	99
2005	92

Line diagrams



Md. Moyazzem Hossain
Lecturer, Dept. of Statistics, U

Graphs of Frequency Distributions: A frequency distribution can be presented graphically in any of the following diagrams:

- Histogram
- Frequency Polygon
- Smoothed frequency curve
- Cumulative frequency curves or 'Ogives'.

Histogram: A histogram is a graphical method for presenting data, where the observations are located on a horizontal axis (usually grouped into intervals) and the frequency of those observations is depicted along the vertical axis.

While constructing histograms the variable (class interval) is always taken on the X-axis and the frequencies depending on it on the Y axis. The distance for each rectangle on the X-axis shall remain the same in case the class intervals are uniform throughout; if they are different the width of the rectangles shall also vary. The Y axis represents the frequencies of each class which constitute the height of its rectangle.

The histogram is most widely used for graphically presentation of a frequency distribution. However, we cannot construct a histogram for distributions with open-end classes.

Histogram and Bar Diagram: First, a histogram is used for representing a frequency distribution only but a bar diagram is never used for representing a frequency distribution. A bar diagram is one-dimensional i.e. only the length of the bar is material and not the width; a histogram is two dimensional, that is in a bar histogram both the length as well as the width are important.

Construction of Histogram when Class-intervals are Equal: When class-intervals are equal, take frequency on the Y axis, the variable on the X-axis and construct adjacent rectangles. In such a case the heights of the rectangles will be proportional to the frequencies.

Example: Represent the following data by a histogram:

Size Class	Frequency	Size Class	Frequency
0-10	5	50-60	10
10-20	11	60-70	8
20-30	19	70-80	6
30-40	21	80-90	3
40-50	16	90-100	1

Md. Moyazzem Hossain
Lecturer, Dept. of Statistics, IU

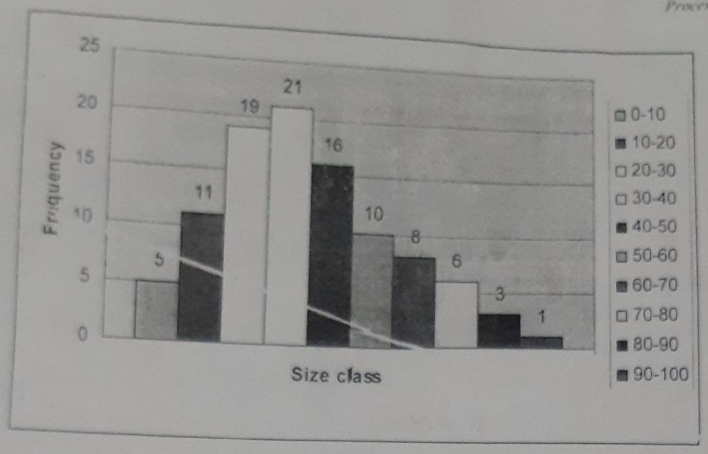
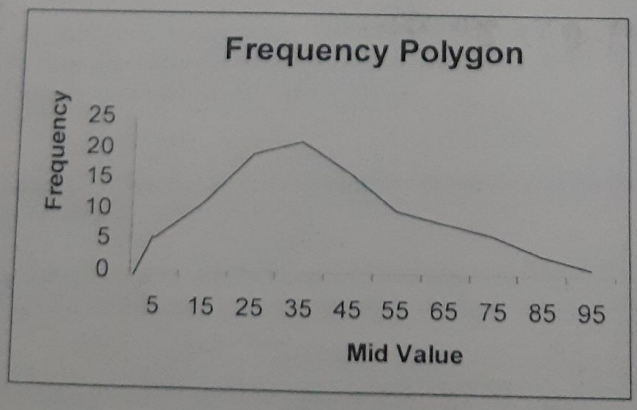


Fig: Histogram when Class intervals are equal

Frequency polygon: In frequency polygon the mid-values of the continuous class intervals are represented along X-axis and the frequencies corresponding to the class intervals are represented along the Y-axis. The class frequencies are plotted against the mid-values of the respective class intervals. These points are then joint by straight lines one after another. The first and the last points are then brought down at each end to the X-axis by joining it to the mid-value of the next out lying interval of zero frequency. The polygon thus obtained is called frequency polygon.

Example: Represent the following data by a histogram:

Size Class	Mid Value	Frequency	Size Class	Mid Value	Frequency
0-10	5	5	50-60	55	10
10-20	15	11	60-70	65	8
20-30	25	19	70-80	75	6
30-40	35	21	80-90	85	3
40-50	45	16	90-100	95	1

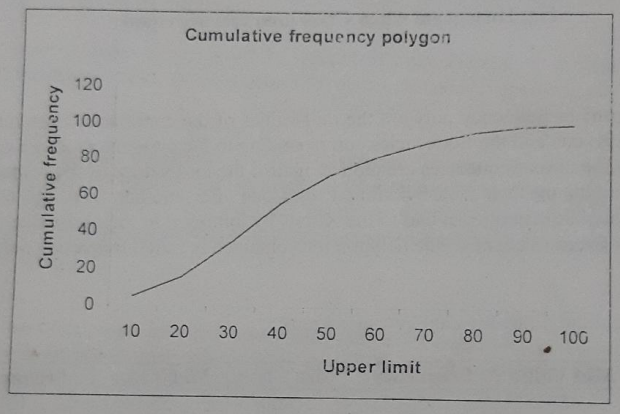


Md. Moyazzem Hossain
Lecturer, Dept. of Statistics, U

Cumulative frequency polygon and Ogive: In cumulative frequency polygon the upper limits of the continuous class intervals are represented in X-axis and the cumulative frequencies are represented to the Y-axis. A free hand curve to smooth a cumulative frequency polygon is called an ogive.

Example: Represent the following data by a histogram:

Size Class	Upper limit	Frequency	Cumulative frequency	Size Class	Upper limit	Frequency	Cumulative frequency
0-10	10	5	5	50-60	60	10	82
10-20	20	11	16	60-70	70	8	90
20-30	30	19	35	70-80	80	6	96
30-40	40	21	56	80-90	90	3	99
40-50	50	16	72	90-100	100	1	100



Md. Movazzem Hossain
Lecturer, Dept. of Statistics, IU