Introduction to Statistics

Origin of statistics:

statistics has been derived from the latin word "status" or Italian word "statista" or the French word "statistique" and the German word "statistik". Status and statistik means political sad state and statista means activities of political state. At first Professor G. Achenwall used the word statistics in 1749. According to Professor G Achenwall "statistics is the political science of the several countries." PC Mahalanobish name the word of statistics in Bengali word Porisongkkhan.

Definition of statistics:

statistics is a scientific methods for collecting, summerizing, presenting and analyzing sample data as well as drawing valid conclusions about population characteristics and making reasonable decision on the basis of such analysis.

→ Some definition of statistics:

1. statistics can be defined as the collection, presentation and interpretation of numerical data — Croxton and Crowed.

2. statistics are measurement, enumerations or estimates of natural or sociel phenomena systematically arrangement to exhibit their inner relation. ——— conner.

3. The science of statistics is essentially a branch of applied mathematics and can be regarded as a mathematics to applied to observation data. ——— R.A. Fisher.

Types of statistics:

There are two types of statistics. such as

* Descriptive statistics
* Inferential statistics

Descriptive Statistics: Descriptive statistics deals with collection, tabulation, presentation and analysis of data. The study of frequency distribution, measures of central tendancy, measures of dispersion, correlation, regression etc. are included in descriptive statistics.

Inferential Statistics: The descriptive statistics are used for making predictions or decisions relating to unobserved characteristic. The methods of taking decision is known as inferential statistics or statistical inference. The inference is made by sampling, sampling distribution, estimation of parameters and test regarding any hypothesis of parameter.

Population: In statistics, population refers the totality of all the items or individuals having some specific characteristics. For example, all the students of university of Dhaka constitute a population.
A population can be classified into two groups and they are:

(a) Finite Population: A population having a finite numbers of units or individuals or items called a finite population.
For example, The population consisting the students of Dhaka university.

(b) Infinite Population: A population having an infinite number of units or individuals or items in called an infinite population.
For example, The population consisting of all possibl outcomes

Sample: A representative and considerably small part of a population is known as a sample of the population.
For example, a group of 1050 students from 31,000 students of Dhaka University constitute a sample.

Parameters: Any characteristics of population about which inference are to be made is called parameters. Population mean (μ) and population variance (σ^2) are examples of parameters.

Statistic: Any characteristic of sample is usually known an statistic. Sample mean ($\bar{x}$) and sample variance (s^2) are examples of statistic.

Scope and Importance of Statistics:

→ Statistics and planning: Statistics is fundamental into planning in the modern age which is termed as "the age of planning". Almost all over the world the govt. are re-storing to plan for economic development.

→ Statistics and economics: Statistical data and techniques of statistical analysis have to immensely useful involving economical problem. Such as wages, price, time series analysis, demand analysis.

→ Statistics and business: Statistics is an irresponsible tool of production control. Business executive are relying more and more on statistical techniques for studying the much and desire of the valued customer

Parameters: Any characteristics of population about which inference are to be made is called parameters. Population mean (μ) and population variance (σ^2) are examples of parameters.

Statistic: Any characteristic of sample is usually known an statistic. Sample mean ($\bar{x}$) and sample variance (s^2) are examples of statistic.

Scope and Importance of Statistics:

→ Statistics and planning: Statistics is fundamental into planning in the modern age which is termed as "the age of planning". Almost all over the world the govt. are re-storing to plan for economic development.

→ Statistics and economics: Statistical data and techniques of statistical analysis have to immensely useful involving economical problem. Such as wages, price, time series analysis, demand analysis.

→ Statistics and business: Statistics is an irresponsible tool of production control. Business executive are relying more and more on statistical techniques for studying the much and desire of the valued customer

→ Statistics and industry: In industry, statistics is widely used inequality control. In production engineering to find out whether the product is conforming to the specifications or not. statistical tools such as inspection plan, control chart etc.

→ Statistics and mathematics: statistics are intimately related recent advancements in statistical technique are the outcome of wide applications of mathematics.

→ Statistics and modern science: In medical scienc the statistical tools for collection, presentation and analysis of observed facts relating to causes and incidence of dieses and the result of applicatio various drugs and medicine are of great importa nce

→ Statistics, psychology and education: In edudion and psychology statistics has found wide application such as determining or to determ the reliability and validity to a test, feetore anaysis etc.

→ Statistics and war: In war the theory of decision function can be a great assistance to the military and personal to plan "maximum destruction with minimum effort.

Quantitative variable further be classified into two sub-groups.

(a) Discrete Variable: A quantitative variable which possesses isolated or integral value is called discrete variable.
Example: Family size, Population size, number of road accidents per day in Bangladesh.

(b) Continuous Variable: A quantitative variable which takes value within a range or limit is called continuous variable.
Example: Height, Weight etc.

Frequency: Frequency, also called class frequency refers to the number of observations falling within a particular class. It is the number of measurements or counts in a category or class.

Frequency distribution: In statistics, frequency distribution is a graph or a data set organized to show the frequency of occurrence of each possible outcome of a repeatable event observed many times.

→ A frequency distribution in statistics is a representation that displays the number of observations within a given interval.

Frequency distribution can be constructed for both categorical and numerical data.

When numerical data is grouped and organized in a frequency distribution results, it is called grouped frequency distribution.

For ungrouped data, we refer to as ungrouped frequency distribution.

constructing of frequency table using categorical data (qualitative data)

→ The constructing of a frequency table for categorical data with a single categorical variable consists essentially of the following steps:

(a) Choose the category into which the data are to be grouped.

(b) Sort or tally the data into appropriate categories.

(c) count the number of items or measurement falling in each category.

(d) Display the results in a table.

(e) The resulting table represents the desired frequency distribution.

2.2 Example: A market researcher conducted an inventory of 25 firms and categorized them as 'large', 'medium' and 'small' depending on the investment, floor space and number of employees. The categories of the 25 listed firms were as follows:

small	large	small	medium	large
medium	large	small	large	medium
large	small	large	medium	medium
medium	large	large	medium	medium
medium	small	small	medium	medium

present the data in a frequency distribution.

Solution:

Here, the data pertain to three distinct categories of the firm and they are large, medium and small.

count the tallies for each category following the table:

Firm size	Tally	Count
Large	𝍸 III	8
Medium	𝍸 𝍸 I	11
Small	𝍸 I	6

Frequency distribution of the firm size:

Firm size	Number of firms	Percent
large	8	32.0
Medium	11	44.0
Small	6	24.0
Total	25	100.0

$(\frac{8}{25} \times 100)$

Here, the number firms (8, 11, 6) represent the class or frequency for the categories large, medium, small respectively. The count 25 rep is the total frequency.

Since, the data is grouped, it is known as categorical distribution.

Example-2.3: A market research team conducted a servey among 350 boys and 250 girls of about the same age on their preference of cold drink available in the market.

	Types of Drink				
Sex	coca-cola	Sprite	Mountain Dew	7-up	Total
Boys	95	90	85	80	350
Girls	65	50	80	55	250
Total	160	140	165	135	600

(a) of the total children, how many of the girls prefer sprite?

⇒ Here, 50 girls out of 600 like sprite.

$\therefore$ percentage $= \frac{50}{600} \times 100\% = 8.3\%$

(b) How many of the children like mountain dew?

⇒ percentage $= \frac{165}{600} \times 100\% = 27.5\%$

(c) Among the total children, what is the percentage of boys who prefer 7-up?

⇒ percentage $= \frac{80}{600} \times 100\% = 13.3\%$

(d) Among the total children, how many do prefer sprites?

⇒ percentage $= \frac{140}{600} \times 100\% = 23.3\%$

(e) of the total boys, what percentage of them like mountain dew?

⇒ percentage $= \frac{85}{350} \times 100\% = 24.3\%$

(f) of those who prefer coca-cola, what is the percentage of boys?

⇒ percentage $= \frac{95}{160} \times 100\% = 59.4\%$

(9) construct the marginal distribution for boys and girls.

Drink type	No. of children	Percent
coca-cola	160	26.7
sprite	140	23.3
Mountain dew	165	27.5
7-up	135	22.5
Total	600	100

constructing of frequency distribution using numerical data

→ The process of constructing a frequency distribution with numerical or quantitative data is very similar to those for qualitative data, except that now the data have to be grouped into classes of appropriate intervals. The simplest device in doing so is to form an array first.

An array is an ordering of values of the variables in order of their magnitude, usually in ascending order; from smallest to the largest.

4.

50	63	70	75	84
51	65	71	75	85
54	65	72	76	86
56	66	72	77	87
56	67	72	79	88
57	68	73	80	88
59	68	73	81	89
60	69	74	82	93
61	69	74	82	93
62	70	74	83	97

grouped data

ungrouped frequency distribution:

Wage	Frequency
50	1
51	1
54	1
56	2
57	1
⋮	⋮
65	2
⋮	⋮
68	2
69	2
70	2
⋮	⋮
72	3
73	2
74	3
75	2
⋮	⋮
97	1

Various terms of frequency distributions for numerical data:

class: A class is an interval containing observations, each observation being classified into one and only one class.

Frequency: The number of observations or values falling into each group or class is called class frequency. The frequency of a class thus shows how many times a particular value of observation is repeated in that class.

class limits: For numerical data, the frequencies of a particular class are bounded by two values. The smaller value of the class is known as the lower class limit while the larger value is known as the upper class ~~value~~ limit.

class boundary:

class interval: The width (w) or length of the class, formed by two boundary values is known as the class interval or class width. A class interval represents the spread between the class boundaries.

class mark: The class mark is the value that lies in the middle of the class and is obtained by averaging the two class boundaries. The class is also referred to as clas-mid-point or mid-value of the class.

open interval: An open interval is an interval with one of it's limits (in either side) indeterminate. Thus an age of a person recorded as less than 45 years (<45) constitute an open interval.

Steps in constructing grouped frequency distribution

(a) Decide on number of classes and the class width in which the observations are to be grouped

(b) Assign the observations to the appropriately chosen classes. This called tallying.

(c) count the number of observation's falling in each class. These numbers are the frequencies.

(d) Display the results obtained in the above three steps in a table or a chart.

(e) The resulting table is our deserved frequency distribution.

If the smallest value(S) and the largest value(L) in a data set are known, then as a rule of thumb,

the range, $R = L - S$

class width, (W) & Approximate number of classes (K)

$\therefore K = \frac{L-S}{W} = \frac{R}{W}$

Determine, $K = 1 + 3.322 \log N$

$$W = \frac{R}{1 + 3.322 \log_{10} N}$$

This approach has been proposed by Sturge:

Another empirical rule suggested by sturge to determine the number of classes is the "2 to the K rule". This rule suggests that the number of classes should be the smallest whole number K that makes the quantity 2^K greater than the total number of observation's (N) in the data set, that is $2^K \geq N$. Suppose that, $N = 50$ observations. Then, since $2^5 = 32$ which is smaller than N and $2^6 = 64$, which is greater than N. The sturge's rules also dictates us to choose 6 classes, that is $K = 6$

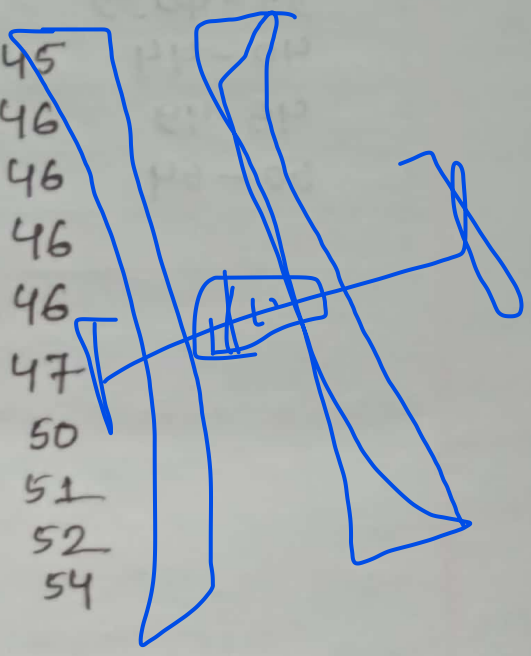
$K = 6$

Constructing frequency distribution using continuous data

Ex-2.5: construct a suitable frequency distribution with class width of appropriate size:

The ages of the 50 workers

25	33	37	42	45
28	34	37	42	46
29	35	37	42	46
30	35	38	43	46
31	35	38	43	46
32	36	38	43	47
32	36	39	44	50
32	36	40	44	51
33	36	41	44	52
33	37	42	45	54



Sol: From giving data,

$N = 50$

number of classes, $K = 1 + 3.322 \log 50 = 6.64$

which is approximately 6 $\therefore K = 6$

$\therefore$ Range, R = upper limit (L) − lower limit (S)

$= (54 - 25) = 29$

$\therefore$ class width, $W = \frac{R}{K} = \frac{29}{6} = 4.83$

which is approximately 5 $\therefore W = 5$

Table: Age Distribution

Age in Years	Tally	Number of Workers
25-29	III	3
30-34	𝍸 IIII	9
35-40 39	𝍸 𝍸 𝍸 I	15
40-44	𝍸 𝍸 II	12
45-49	𝍸 II	7
50-54	IIII	4
		Total: 50

4. Other forms of frequency distribution:

1. Percentage frequency distribution
2. Relative frequency distribution
3. Cumulative distribution

Percentage frequency distribution: A percentage frequency distribution is formed by dividing the number of classes attribute to a category by the total number of cases and multiplying the resulting value by 100. Thus f_i is the frequency of the i-th class of a frequency distribution and N the total frequency then the percentange of i-th class is,

$$P_i = \frac{f_i}{N} \times 100$$

∴ The total frequency in a percent distribution that is $= \sum\left(\frac{f_i}{N}\right) \times 100 = 100$

Relative frequency distribution: Instead of presenting the frequencies in absolute figures, it is sometimes convenient to express the frequencies in related terms. The resulting distribution is then called relative frequency distribution. The relative frequency is simply the fraction or proportion of the total number of items belonging to the class or category. For a data set having a total number of observation, N. The related frequency of i-th class is,

Relative frequency of i-th class $= \frac{f_i}{N}$

The total relative frequency in such a distribution is, $\sum\left(\frac{f_i}{N}\right) = 1.0$

[The relative frequencies are essentially proportions when multiplied by 100, result in percentage frequencies and hence the percentage distribution.]

Cumulative frequency distribution: In statistics, a cumulative frequency distribution is defined as the total of frequencies, that are distributed over different class intervals. There are two type of cumulative frequency distribution and they are:

1. less than type
2. more than type

Less than type frequency distribution: the less than cumulative frequency distribution is obtained by adding successively the frequencies of all the previous classes along with the class.

More than type frequency distribution: The more than cumulative frequency distribution is obtained by determining the total frequencies starting from the highest class to the lowest class.

Table : Frequency distribution

Class boundaries	Absolute frequency	Percentage frequency	Relative frequency	less than type C.F	% less than type C.F	More than type C.F	% more than type C.F
49.5-57.5	6	12.0	0.12	6	12.0	50	100.0
57.5-65.5	7	14.0	0.14	13	26.0	44	88.0
65.5-73.5	14	28.0	0.28	27	54.0	37	74.0
73.5-81.5	10	20.0	0.20	37	74.0	23	46.0
81.5-89.5	10	20.0	0.20	47	94.0	13	26.0
89.5-97.5	3	6.0	0.06	50	100	3	6.0
Total :	50	100	1.0				

percentange frequency for 3rd class $= \frac{f_3}{N} \times 100\% = \frac{14}{50} \times 100\%$

relative frequency $= \frac{f_3}{N} = \frac{14}{50} = 0.28$

Graphical Presentation of Data

→ In addition to presenting a frequency distribution in tabular form, one can present the same through some visual aids. This refers to graphs and diagrams or charts.

As we aware that a frequency distribution can be constructed either from categorical (qualitative) and numerical (quantitive) data, the graphs and diagrams to be constructed will also differ accordingly: categorical or numerical.

Presenting of categorical Data:

→ Bar diagram, stacked bar diagram
→ Pie diagram

Bar diagram: Bar diagram also called bar charts are commonly used to describe categorical data. A bar diagram is a form of presentation in which the frequencies against the categories are represented by rectangles seperated usually along the horizontal axis and drawns as bars of convenient widths. The width of these bars have significance but are taken to make the chart look attractive.

We represent the data by two different bar diagrams. These are,

(a) Vertical bars
(b) Horizontal bars

Ex-2.11: The accompanying table shows the stock position of finished goods in Metric tons as of June 2004 of the Bangladesh chemical industries (BCIC). Represent the data by suitable diagram.

Finished Goods	Quantity (in Metric Ton)
TSP	8916
SSP	18455
Paper	2660
Cement	7048
Sanitary ware	1620
Insulator	3520
Tiles	17335
Total	54928

Sol:

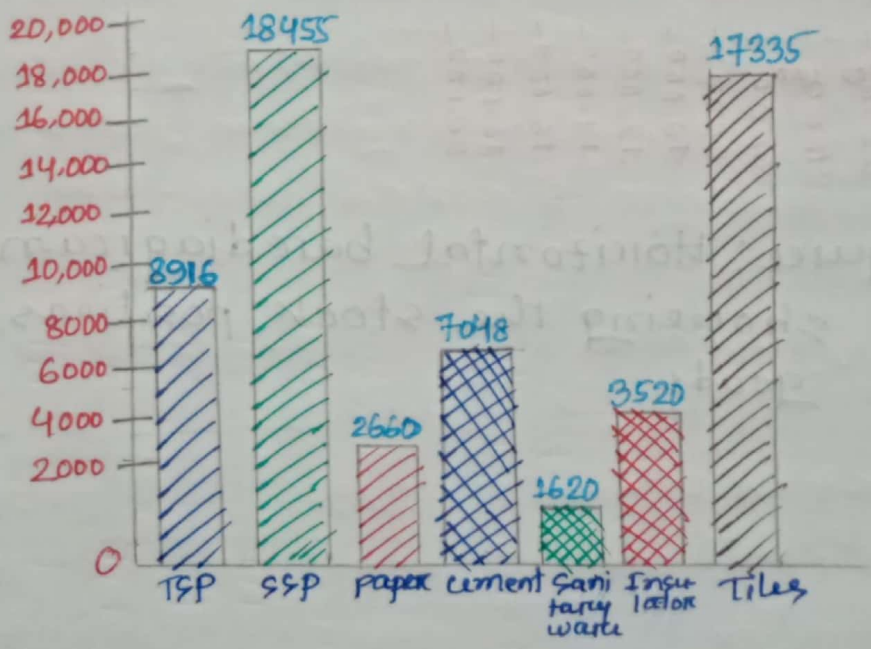


Figure: Vertical bar chart showing the stock positions of goods.

Alternative:

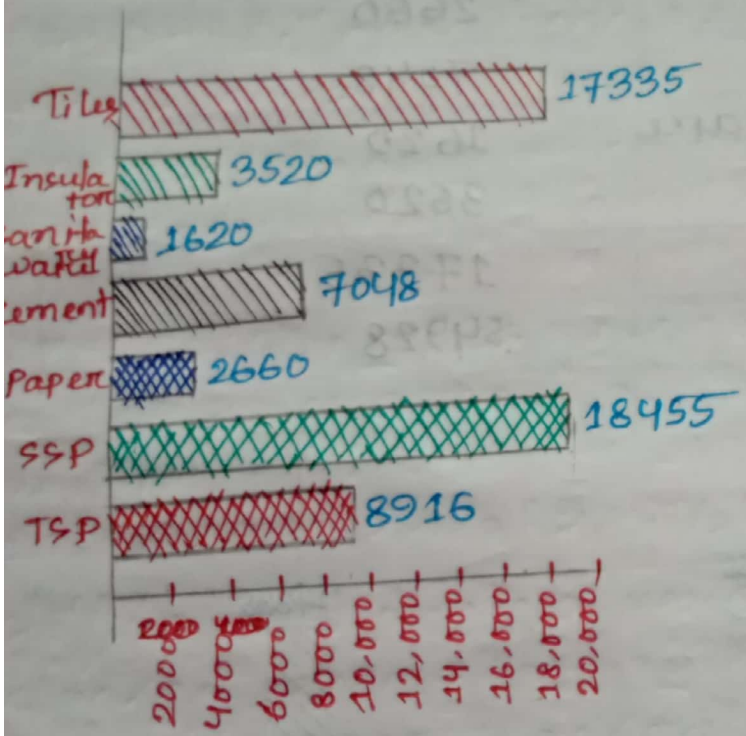


Figure: Horizontal bar diagram showing the stock positions of goods

stacked bar diagram: A stacked bar diagram, variously known as component bar diagram or sub-devided bar diagram, is a side by side diagram that uses bars to show comparisons between categories of data, but with ability to break down and compare parts of a whole. Each bar in the diagram represents a whole and segments in the bar represent different parts or categories of that whole. Stacked bar diagram is a good device to display categorical data. The component parts are variously colored or shaded to make them distinct and attractive.

As with the simple bar diagram, the stacked bar diagram uses rectangular boxes to represent categories of a variable. The variable located on the x-axis is known as the stacked variable. The stacked bar diagram differs from the simple bar diagram in that each rectangular box on the x-axis is made up of smaller individual boxes that we call segments. The segments are stacked on top of one another and could also be called the y-axis variable. The height of each segment represents the number of cases in a category of the stacked and a category of the segment variable. The stacked bar diagram depicts the relationship between the categories of the stacked and segment variable.

There are two types of stacked bar diagram: simple stacked bar and 100 percent stacked bar.

In the construction of 100% stacked bar diagram, we follow the following steps:

(a) convert each component in the data set into percentage value of the corresponding total.

(b) Draw one rectangular for each total, taking equal lengths of 100 units and breadths proportional to the totals.

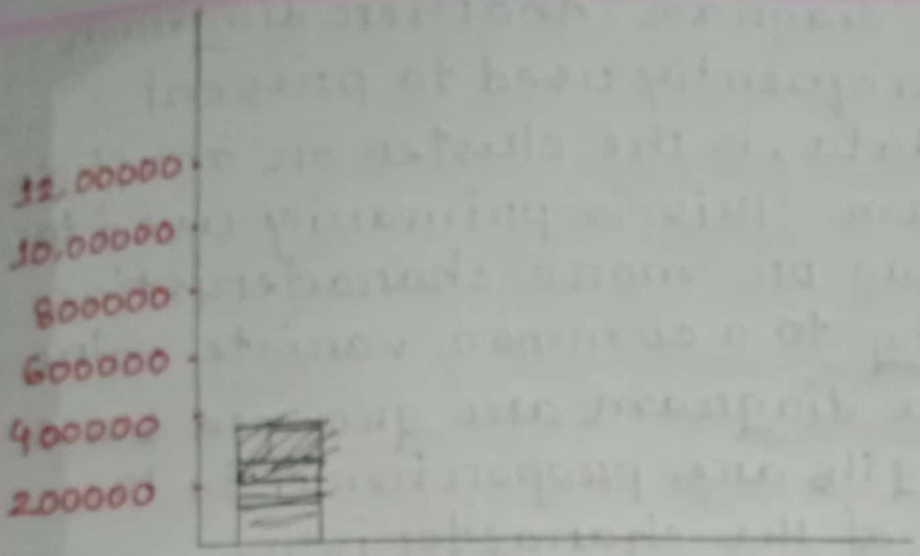
(c) Devide each rectangle so drawen into parts equal

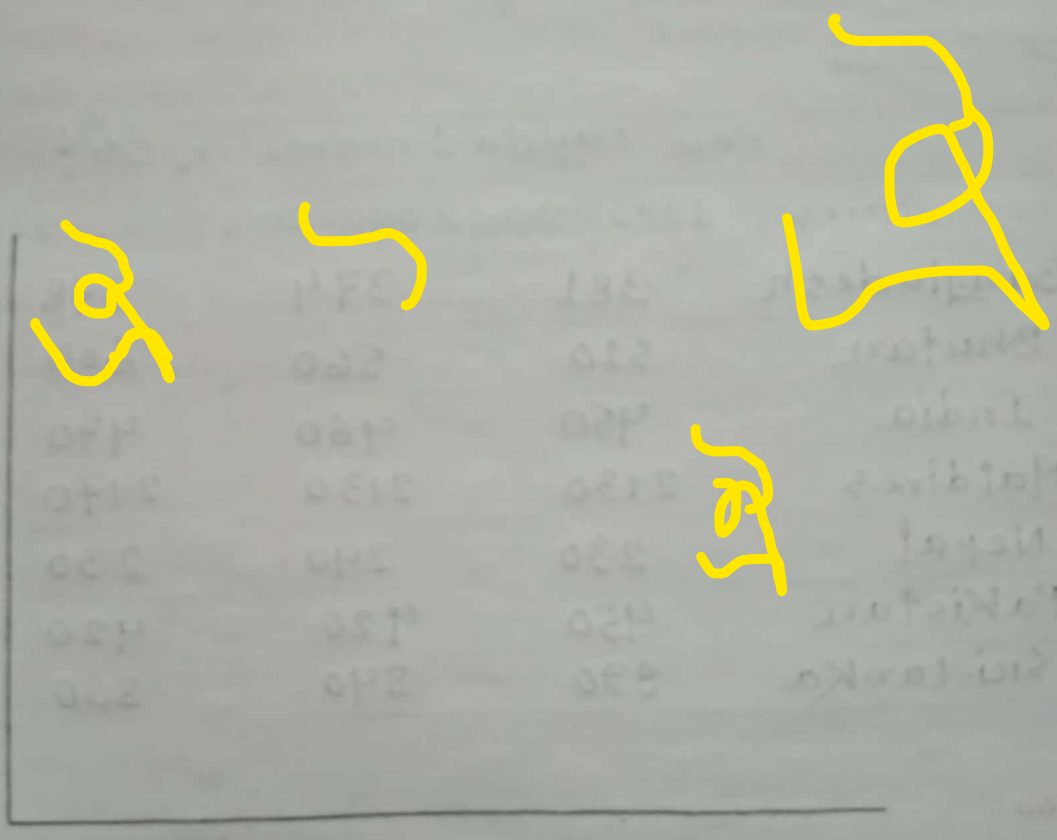
(d) use shading or color to distinguish one component from the others.

Ex - 2.18: The gross revenue expenditures of the government of Bangladesh in million BDT for the financial years 2011-12 and 2012-13 were as follows:

Sol: We first draw a ~~percentage~~ simple stacked bar ~~and~~ diagram followed by a percentange stacked bar.

Heads of revenue expenditure (gross)	% Expenditure in Taka	
	2011-12	2012-13
Wages and salaries	21.9	20.3
commodities & services	4.9	4.5
Transfer	33.7	34.1
Other services	32.5	34.1
Total	100.0	100.0



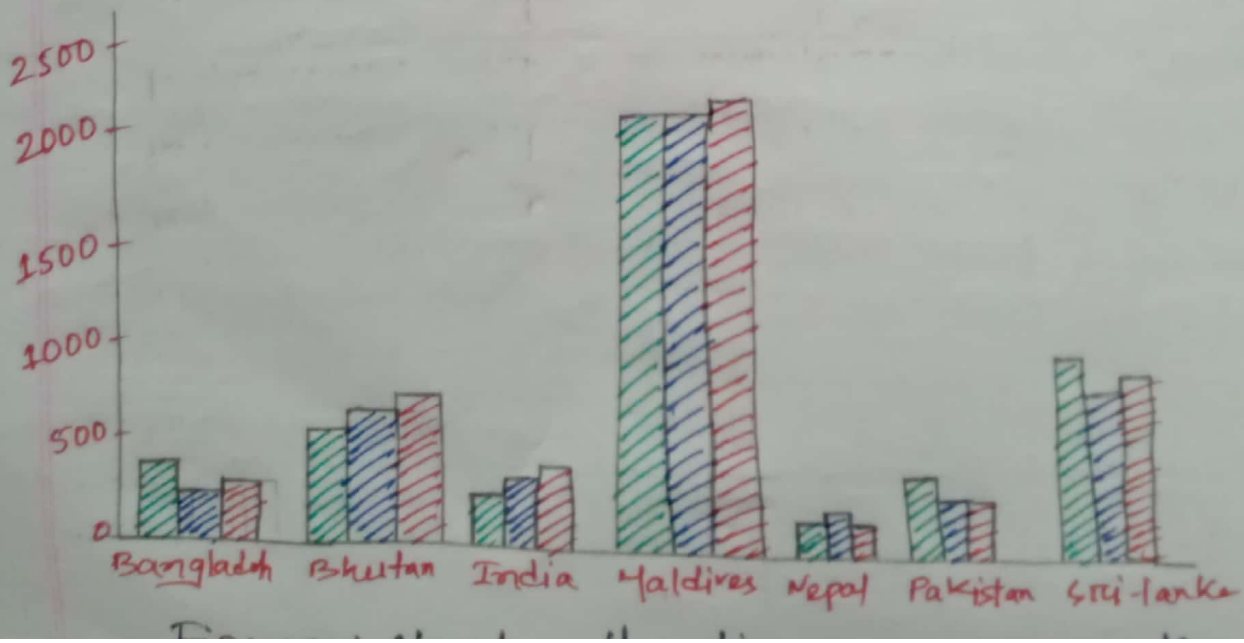


Cluster Bar diagram: Another diagram which is frequently used to present statistical data, is the cluster or multiple bar diagram. This is primarily used to compare two or more characteristics corresponding to a common variate value. cluster bar diagram are grouped bars whoose lengths are proportional to the magnitude of the characteristics.

Ex-2.24:

Per capita Income in US $

countries	1990-2000	2000-2001	2001-2002
Bangladesh	381	374	378
Bhutan	510	560	600
India	450	460	470
Maldives	2130	2130	2170
Nepal	230	240	230
Pakistan	450	420	420
Sri-Lanka	890	840	850



Pie chart: Pie diagram also known as pie chart, is a useful device for presenting categorical data. The pie chart consists of a circle sub devided into sectors whose areas are proportional to the various parts into which the whole quantity is devided. The sectors may be shaded or colored differently to show their individual contributions to the whole. The following steps are involved in constructing a pie chart:

(a) convert the absolute frequencies into relative ~~percentage~~ frequencies for each category of the variable.

(b) Multiply the percentage value by 360° for each category. The resulting values are the angles expressed into degrees.

(c) Draw a circle of appropriate radius.

(d) The resulting figure is the desired pie diagram of the data.

Ex-2.7:

Sectors	Percent of Jobs	Angles
Govt.	4	14.4
Non-govt.	14	50.4
Private	61	219.6
Others	21	75.6
Total	100	360.0

Sol: To draw a pie diagram,

$$\text{Angle} = \frac{14}{100} \times 360^\circ = 50.4^\circ$$

The various parts of the Pie chart drawn may be identified either by angles in degrees or in percentage value.

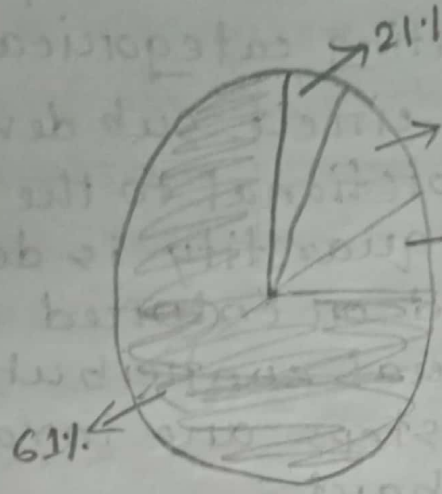


Figure: Pie diagram of this data

Presentation of Continuous Data:

→ Histogram
→ Frequency polygon
→ Ogive

Histogram: The most common form of graphical presentation of a frequency distribution is the frequency histogram. A frequency histogram is constructed by placing the class boundaries on the horizontal axis of a graph and the frequencies on the vertical axis. Each class is shown on the graph by drawing a rectangle whose base is the class boundary and whose height is the corresponding frequency for the class.

Ex-2.39:

House rent (% of income in taka)	Frequency (no. of families)
4.5-9.5	8
9.5 - 14.5	29
14.5 - 19.5	27
19.5 - 24.5	12
24.5 - 29.5	4
Total	80

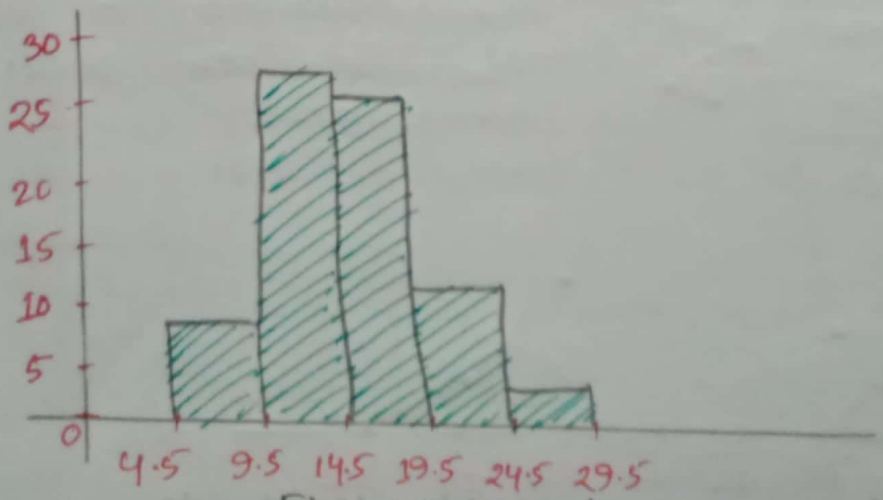


Figure: Histogram with equal class wid

Frequency Polygon: A frequency polygon provides an alternative to a histogram as a way of graphically presenting a distribution of a continuous variable. The presentation involves placing the mid-values on the horizontal axis and the frequencies on the vertical axis. Instead of using rectangles, as with the histogram, we find the class mid-points on the horizontal axis and then plot points directly above the class mid-points at a height corresponding to the frequency of the class.

Ex-2.41:

Data for constructing Frequency polygon

Class boundary	Mid-Values	Frequency
-0.5-4.5	2	0
4.5-9.5	7	8
9.5-14.5	12	29
14.5-19.5	17	27
19.5-24.5	22	12
24.5-29.5	27	4
29.5-34.5	32	0
Total		80

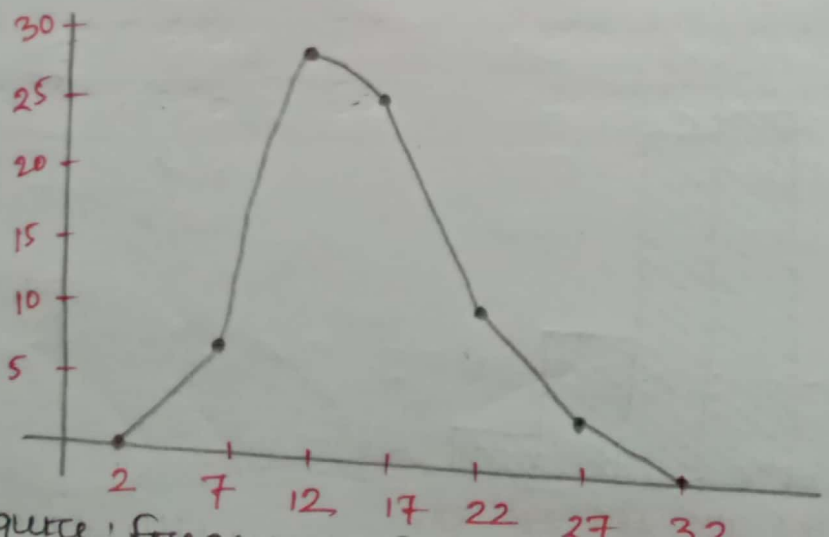


Figure: F...

Cumulative frequency polygon: A graph of the cumulative frequency distribution or cumulative relative frequency distribution is called an ogive. Two types of ogive can be constructed:

(a) more than type ogive
(b) less than type ogive

To construct a less than type ogive, follow the steps below:

(a) Put the upper class limits (precisely the upper boundaries) on the horizontal axis and cumulative frequency on the vertical ~~bar~~ axis.

(b) Plot a point directly above each upper class limit at a ~~high~~ height corresponding to the cumulative frequency at that upper class limit.

(c) Plot one additional point above the lower class limit for the first class at a height ~~corresponding~~ of zero.

(d) connect these points by straight lines.

To construct a more than type ogive, the steps are as follows:

(a) Put the lower class limits on the horizontal axis

(b) Plot a point against each lower class limit at a height corresponding to the cumulative frequ at that lower class limit.

(c) Plot an additional point above the upper class limit for the terminal class at a height of zero frequency.

Ex - 2.43:

longevity (in month)	Number of bulbs
1.45 - 1.95	4
1.95 - 2.45	2
2.45 - 2.95	8
2.95 - 3.45	30
3.45 - 3.95	20
3.95 - 4.45	10
4.45 - 4.95	6
Total	80

Sol:

less than type		more than type	
Inflation rate (%)	cumulative frequency	Inflation rate (%)	cumulative frequency
less than 1.45	0	more than 1.45	80
less than 1.95	4	more than 1.95	76
less than 2.45	6	more than 2.45	74
less than 3.45 2.95	14	more than 2.95	66
less than 3.95 3.45	44	more than 3.45	36
less than 4.45 3.95	64	more than 3.95	16
less than 4.45	74	more than 4.45	6
less than 4.95	80	more than 4.95	0

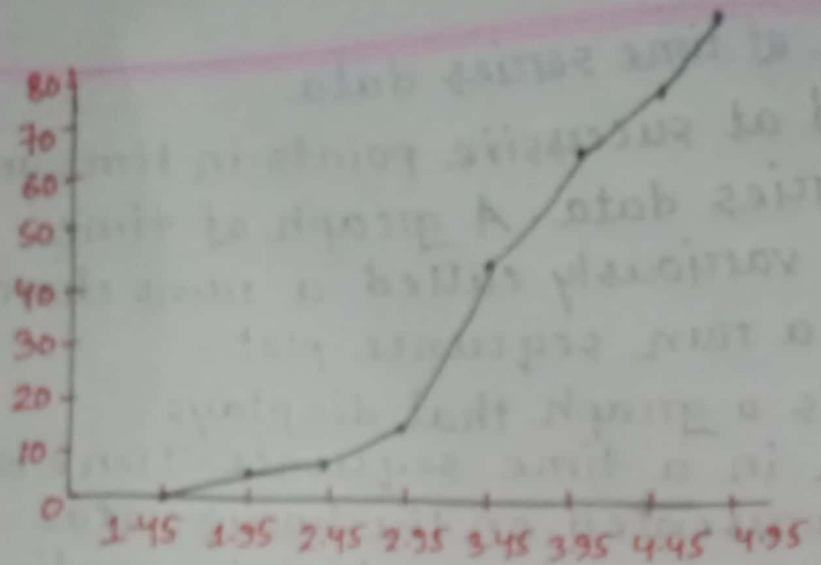


Figure: less than type ogive

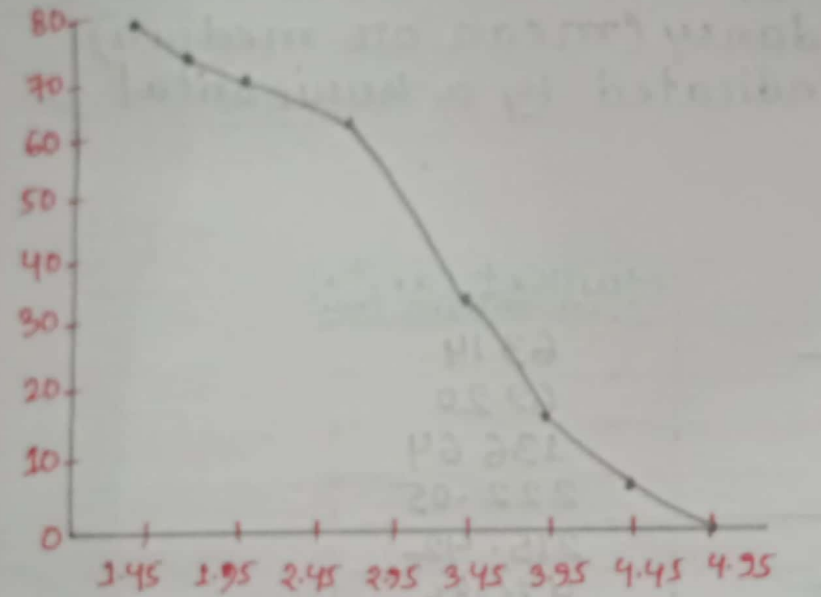
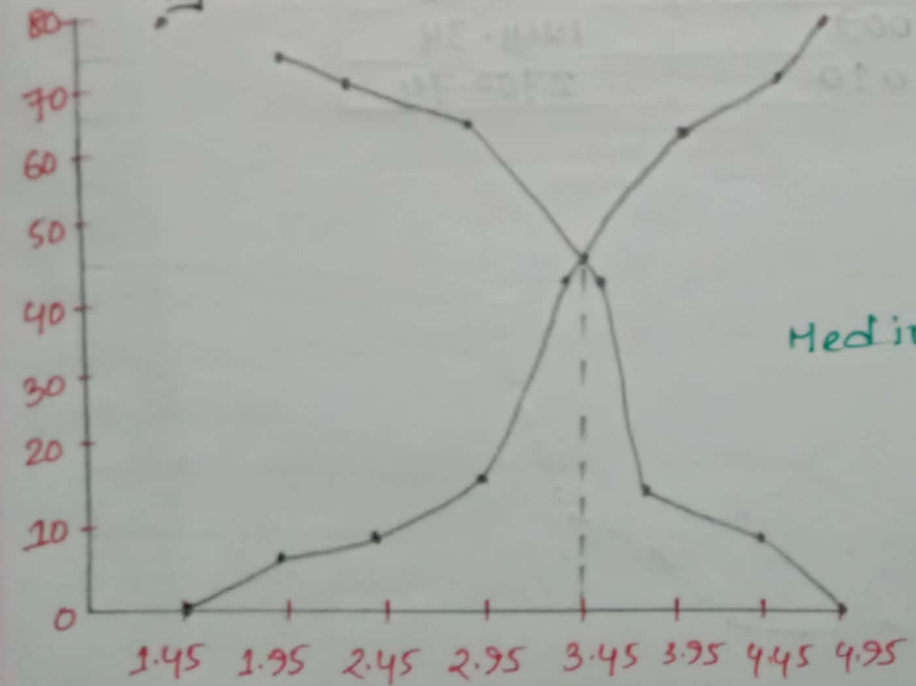


Figure: More than type ogive



Medium = 3.45

Presentation of time series data

Data measured at successive points in time are called time series data. A graph of time series data is variously called a runs chart, line chart or a run sequence plot.

A run chart is a graph that displays observed data in a time sequence. Time is generally represented on the horizontal (x) axis and the property under observation on the vertical (y) axis. often, some measure of central tendancy (mean or median) of the data is indicated by a horizontal reference line.

Ex: 2.49

Year	Market capital in billion taka
2002	63.14
2003	69.20
2004	136.64
2005	222.05
2006	215.42
2007	475.86
2008	931.03
2009	1241.34
2010	2700.74