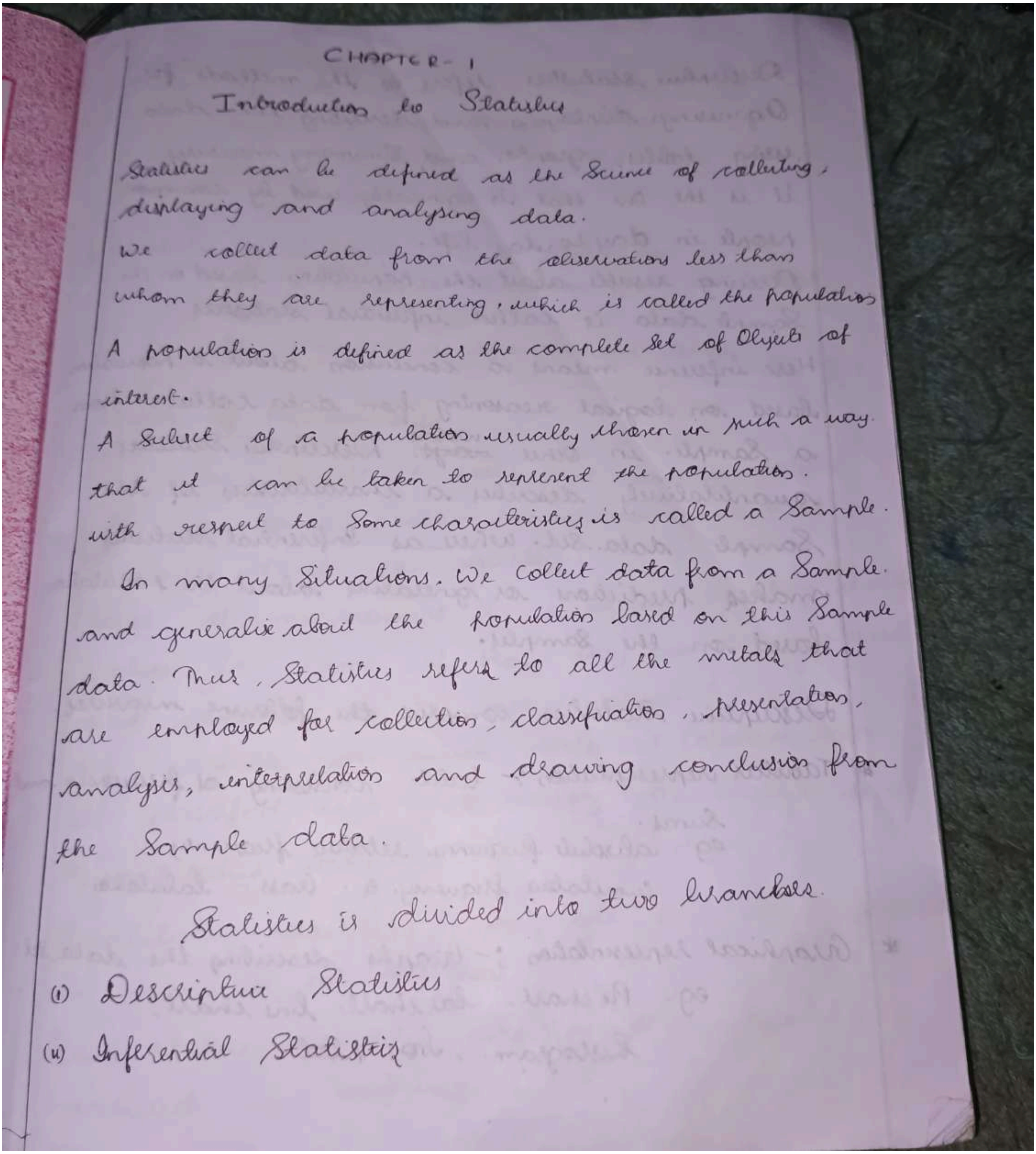


CHAPTER- 1

Introduction to Statistics

Statistics can be defined as the Science of collecting, displaying and analysing data.

We collect data from the observations less than whom they are representing, which is called the population

A population is defined as the complete set of Objects of interest.

A Subset of a population usually chosen in such a way that it can be taken to represent the population. with respect to Some characteristics is called a Sample.

In many Situations. We collect data from a Sample. and generalise about the population based on this Sample data. Thus, Statistics refers to all the metals that are employed for collection, classification, presentation, analysis, interpretation and drawing conclusion from the Sample data.

Statistics is divided into two branches.

(i) Descriptive Statistics

(ii) Inferential Statistics

Descriptive Statistics refers to the methods for Organising, displaying, and describing the data using tables, graphs, and Summary measures. it is the One that is generally used by common people in day to day life.

Deriving results about the population. based on the Sample data is called inferential Statistics

Here inference means a conclusion about a population based on logical reasoning from data collected from a Sample. in Other ~~ways~~ words. Descriptive Statistics. quantitatively describes a characteristics of the Sample data. set. where as Inferential Statistics makes predictions or generalises about the population. based on the Samples.

Descriptive Statistics comprise the following measures.

Tabular representation :- table consisting of frequencies and Sums.

eg:- absolute frequency, relative frequency, cumilative frequency, & Cross - tabulation

raphical representation :- Graphs describing the data set.

eg:- Pie chart, bar chart, line chart, histogram. , box plot.

* Measure of Central tendency :- which describe where a large part of the Sample is located.
  Eg:- Arithematic mean, geometric mean, median, [average] mode, etc.

* Measures of Dispersion :- which describe how much is the spread of the Observations from the Central Value.
  Eg:- Standard deviation, Variance, Range.

Levels of Measurements.

The statistical tools that can be applied for a given problem depend upon the level of measurement of the variables.

Level of measurement is a way of classifing the data values in terms of How they can be measured and compared to the other values.

The variables on which data is Collected can be classified into two namely qualitative and quantitative variables.

Qualitative variables are the Ones which cannot be measured in numbers.

Quantitative variables are the ones which can be measured in numbers.

There are two levels of measurements for qualitative variables, norminial and ordinal. also there are two levels of measurement for Quantitative variable, interval and ratio.

Understanding these four levels of measurement is very important since the level of measurement of a the variable plays a keyrole in deciding which statistical ~~too~~l can apply to a given data.

Nominal Level

Nominal variables are where the categories are just named.
The word Nomine means names. The variables falling into this category have no quantitative value.
In the nominal~~to~~ level only the name matters and order does not matter

eg:- eye colour, Gender.

Ordinal Level.

This level is same as the nominal level, The only difference is that the Order matters.

for eg:- assigning codes to households according to their levels of income. Say 0 to low, 1 to medium, & 2 to high.

Interval level of measurement

in this level of measurement both order and difference between the values of a variable matter but ratio does not matter.

for eg:-
City A has 10°c temperature and City B has 20°c temp. here we can internret that City B is warmer than City B is warmer than city A. But it would be wrong to conclude that city B is twice as city A. also note that zero is an arbitiary number in the interval level. ie it does not mean absence of that variable which means if the temperature of a place is 0°c. then we cannot say that place has no temperature.

Ratio ~~of~~ Level of Measurement

In this level of measurement name, order, difference and ratio all are meaningful.

for eg:-

Suppose a box B_1 has 5 kg weight and box B_2 has 15 kg weight. Then we can say that box B_2 is thrice as heavy as box B_2.

Descriptive Statistics

Quantitative data exhibit certain general characteristics in the following ways.

(1) They show a tendancy to concentrate at certain values usually some where in the centre of the distribution. measures of this tendancy are called measure of central ~~to~~ tendancy or averages.

(2) The data vary about the measure. central tendancy which are called measures of dispersion or variation.

(3) The data in a freauency distribution vary fall into symmetrical or assymmetrical. The measure of direction and degree of symmetrical or assymmetry are known as measure of ~~[illegible]~~ Skewness.

4) The data in a frequency distribution exhibits flatness or peakedness of the frequency curve. This measure are known as measure of kurtosis.

Frequency distribution.

When observations are available on a single variable of a large no of individuals. often it becomes necessary. to Summarise the data as far as possible:

To this end, a tabular representation which shows the distribution of the frequency in different classes may be used. The manner in the class frequency are distributed over the class intervals is called the grouped frequency distribution

eg:- Let us consider the marks in mathematics obtained by 50 students selected at random. from a certain School. the following frequency table is obtained by computing the class intervals with corresponding frequencies.

Class	20-40	40-60	60-80	80-100
frequency	4	20	20	6.

The following points may be kept in mind for Classification.

1) The classes should be clearly defined and Should not leave any ambiguity.

2) the classes should be exhaustive, ie each of the given value should be included in oneof the

(3) The classes should be mutually exclusive and non overlapping.

(4) The class should be of equal width, if the classes are of varying width. the different class frequencies will not be comparable.

(5) indeterminate classes; eg:- open end class like $< a$, or $> b$. should be avoided as far as possible.

(6) The no of classes should neither be too large nor too small. The following formula may be used to determine an approximate no of class.

$K = 1 + 3 \cdot 322 \log_{10} N$.

N is the total frequency.

This equation is also known as sturges formulae

Class limit:

class limit should be chosen in such a way. that the midvalue of the class interval and actual average of the observation in that classinterval are as near to eachother as possible.

Continious frequency distribution -

if we have a continiow variable, it is not passible to arrange the data in the class intervals of above type

for eg:- let us Consider the distribution of age in years if class intervals are ..15-19, 20-24, ... then the person with ages between 19 & 20. years are not taken into consideration. In Such a way that all the person with any fraction of age are included in one group. We may rewrite the above classes. as shown in the following table.

0 - 5	
5 - 10	
10 - 15	
15 - 20.	
20 - 25.	

The frequency distribution with such classes is known as Continous frequency distribution. In the above classes, the upperlimit of each class are excluded from the respective classes and are included in the immediate next class.

Such classes are ~~called~~ known as exclusive classes.

Graphical representation of frequency distribution

it is often useful to represent a frequency distribution by means of a diagram. This representation facilitates the comparision of two or more frequency distributions.

Histogram.

To draw a histogram of a given continous frequency distribution ~~too~~. We first mark the class intervals in the x axis..

On each class interval draw rectangles with height proportional to the frequency of the corresponding class interval. So that the area of the rectangle is ~~so~~ proportional to the frequency of the class. Such diagram of continous rectangles is called

Histog

Remar

- if e contin distr
- To dist co ind

eg

Histograms

Remarks

- If the grouped frequency distribution is not continuous, it should be converted to continuous distribution and then draw the histogram
- To draw the histogram for ungrouped frequency distribution. We have to assume that the frequency corresponding to the value x is spread over the interval ~~$\left(\frac{(x-h)}{2}, \frac{x+h}{2}\right)$~~ $\left(x-\frac{h}{2}, x+\frac{h}{2}\right)$

eg:- construct histogram for the following frequency table.

Mark	No of students
15 - 19	9
20 - 24	11
25 - 29	10
30 - 34	44
35 - 39	45
40 - 44	54
45 - 49	37
50 - 54	26
55 - 59	8
60 - 64	5
65 - 69	1
	250

This frequency distribution is not continuous. So we first convert it into a continuous distribution with exclusive type classes.

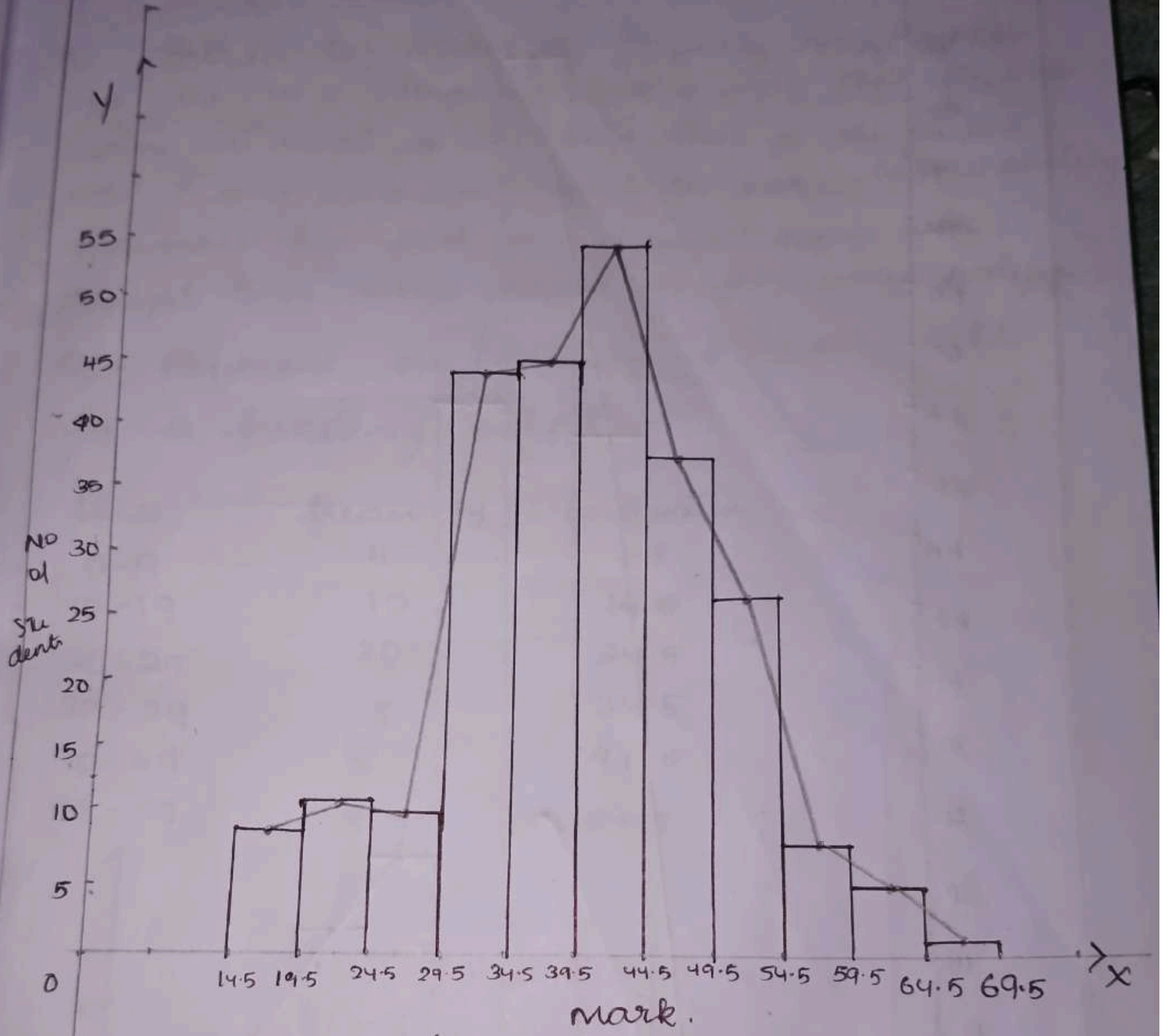
Let d be the gap between the upper limit of any class and lower limit of the succeding class. then the new class boundaries are given by upperclass boundary = upper class unit + $\frac{d}{2}$.

lowerclass boundary = lower class unit − $\frac{d}{2}$.

So in our problem $d = 1$.

∴ the continious frequency distribution is given by

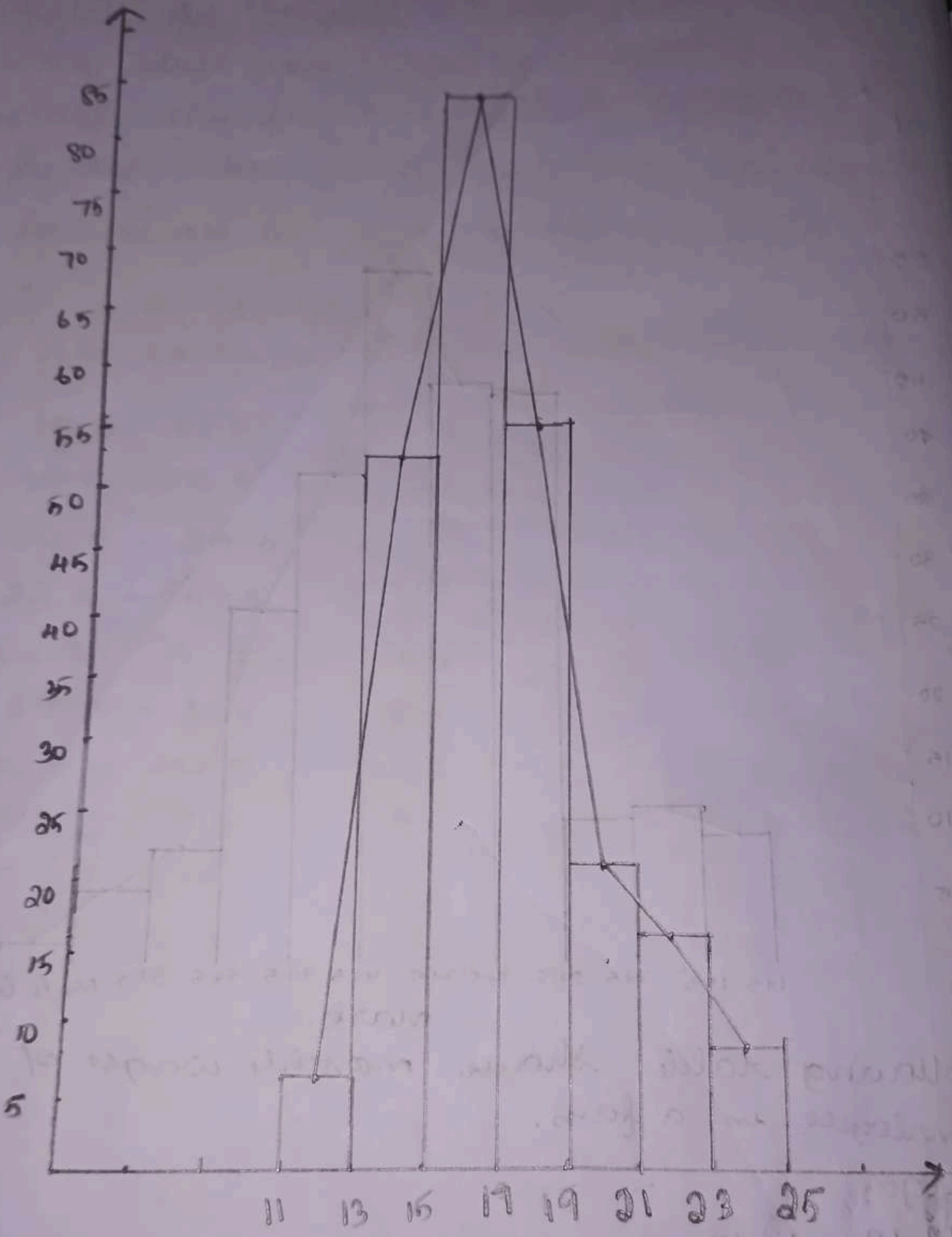
14.5 - 19.5	9
19.5 - 24.5	11
24.5 - 29.5	10
29.5 - 34.5	44
34.5 - 39.5	45
39.5 - 44.5	54
44.5 - 49.5	37
49.5 - 54.5	26
54.5 - 59.5	8
59.5 - 64.5	5
64.5 - 69.5	1



Q) following table shows monthly wages of H.W employees. in a firm.

wages

11-13	13-15	15-17	17-19	19-21	21-23	23-25
6	53	85	56	21	16	8



Frequ

in

valu

and

freq

Stra

eg:

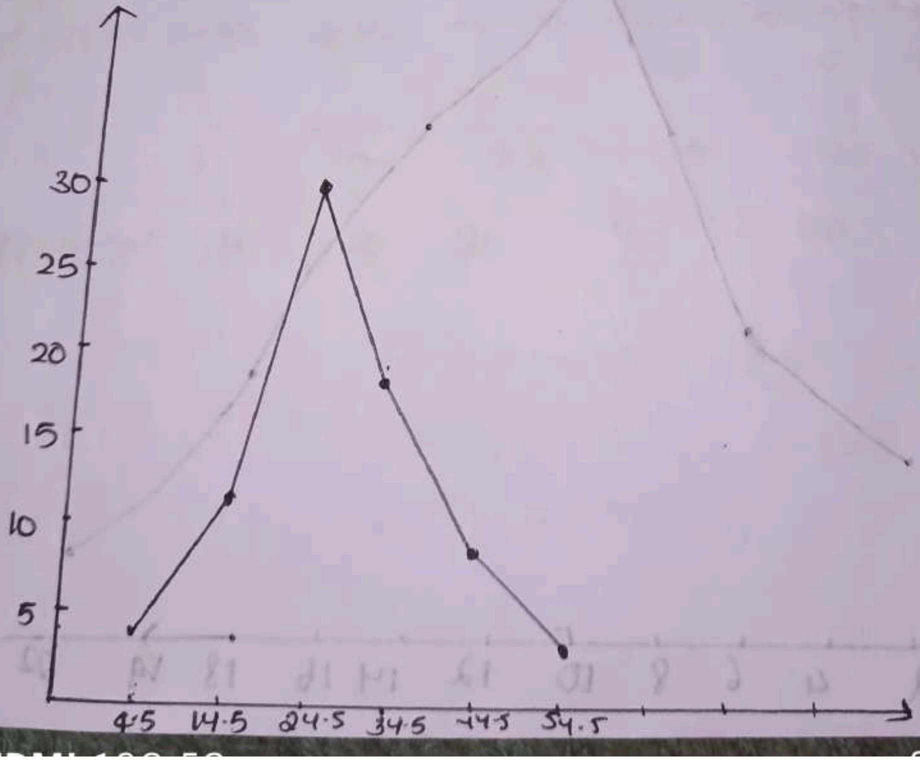
b

Frequency Polygon.

Inorder to draw a frequency polygon plot. in the x-y plain in such a way that the x axis values are eaual to the mid values. of the classes. and Y axis values are eaual to the ~~corre~~ corresponding frequencies. Then joint the adjacent points with Straight lines which results a frequency polygon.

Eg:- Represent the following frequency table by a frequency polygon.

Class	frequency.	mid value.
0-9.	4	4.5
10-19	12	14.5.
20-29.	30.	24.5
30-39.	18	34.5
40-49.	8	44.5
50-59.	2.	54.5.

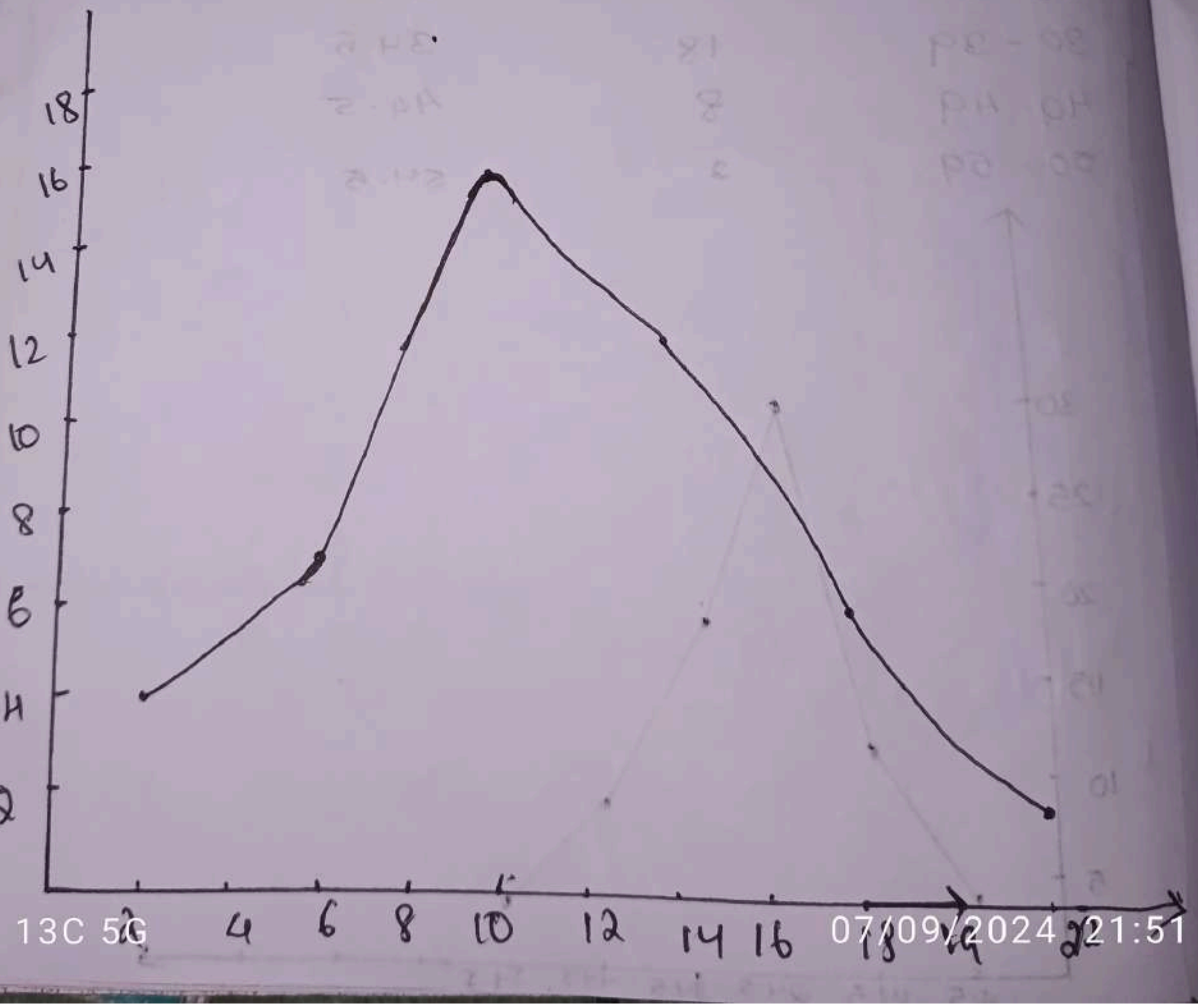


Frequency Curve.

To draw a frequency curve from a frequency table. take the midvalues of classes on the x axis and the frequencies on Y axis.
Mark the points accordingly. next joint these plotted points with a Smooth line, we get a frequency Curve.

Eg:- draw a frequency curve from the following. frequency table.

class.	frequency.	
0-4	4	2.
4-8	7	6.
8-12.	16.	10
12-16	12	14
16-20	6	18
20-24.	2.	22.



Ogives (Cumulative Frequency Curve) .

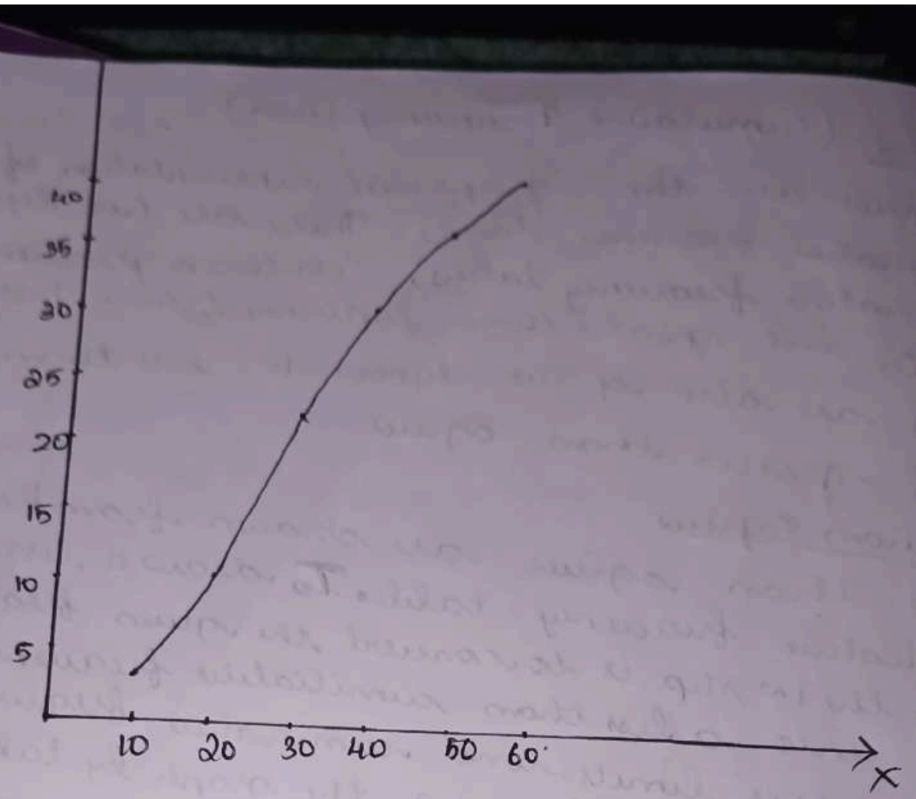
Ogives are. the graphical representation of. Cumiliative frequency tables. There are two types of cumiliative frequency Tables, less than frequency table ~~less th~~ and greater than frequency table. There fore. ogives are also of two types. ie. less than ogives and greater than ogive.

Less than Ogives

Less than ogives are drawn from less than cumiliative frequency table. To draw a less than ogive. the 1st step is to convert the given frequency table into a less than cumiliative frequency table. with upper limits and cumiliative frequencies. Then plot the points on the graph by taking upperlimits on the x axis and C.f on. Y axis. joining these points with a smooth line we get a less than ~~ogl~~ ogive. Less than ogives are upward sloping curves from left to right.

eg:- draw Lessthan ogive curve from following table.

Class.	0-10	10-20	20-30	30-40	40-50	50-60
fre.	3	7	12	8	6	4.
C.F	3	10	22	30	36	40
upperlimit	10	20	30	40	50	60



Greater than Ogive

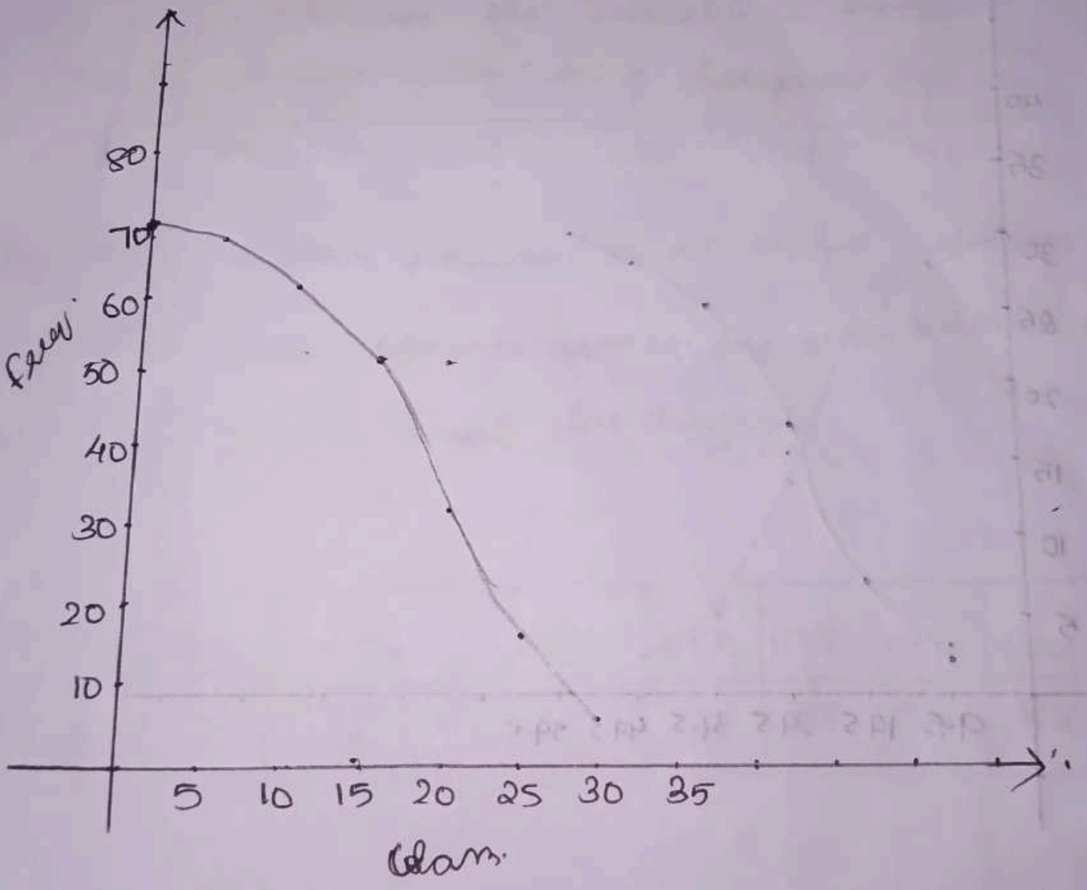
Graph of a Greaterthan Cumilative frequency table is called a greaterthan Ogive.

To draw a greaterthan ogive. marks the points on the graph in such a way that the x axis value are equal to the lower limits and y axis values are respective cumiliative frequency. Joining these points by a smooth line, we get a greater than Ogive. The greater than ogives are downward slopping curves from left to right

Eg:-

represent the following frequency table in a more than ogive.

Class	Freq.	Ls C.F.	lower.	More than cf
0 - 5	2	2	0	70
5-10	5	7	5	68
10 - 15	12.	19	10	63
15 - 20	20.	39	15	51
20 - 25	16.	55	20	31
25 - 30	10	65	25	15
30 - 35	5	70	30.	5

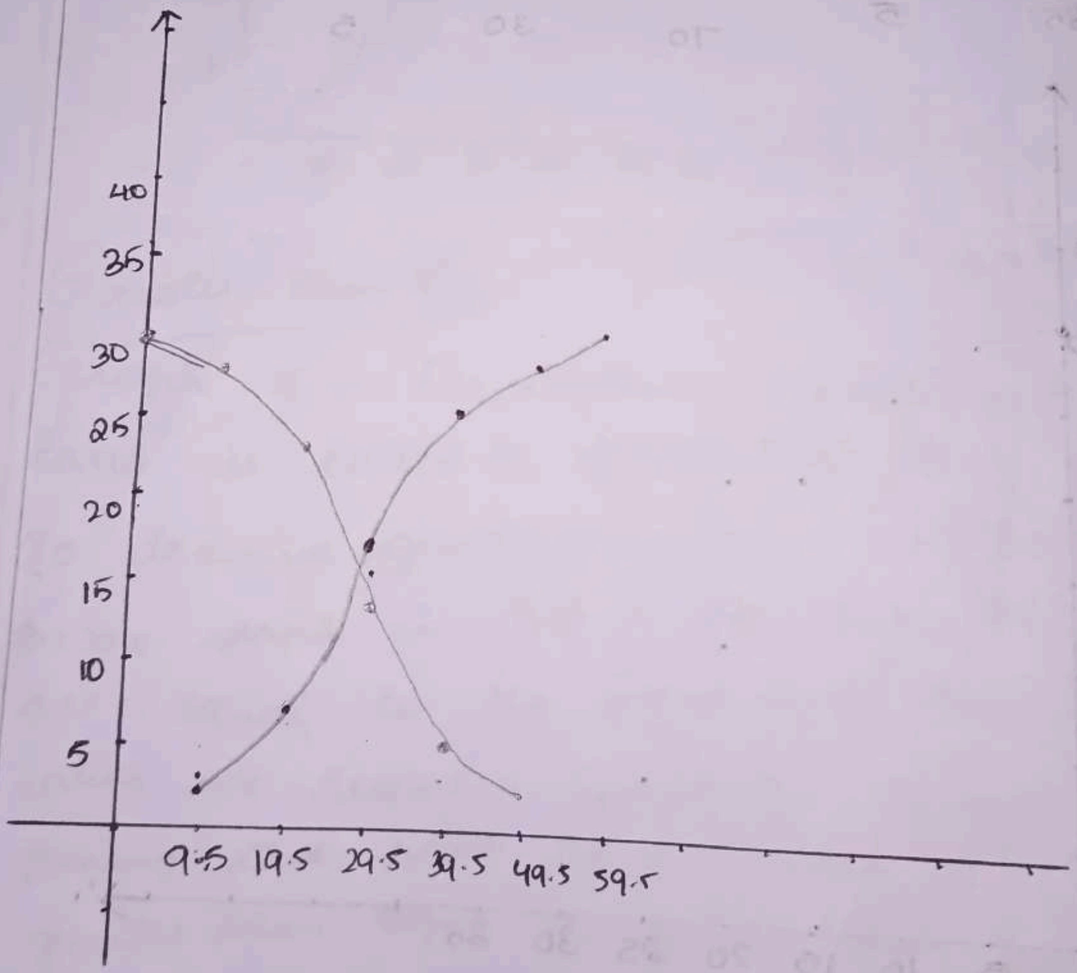


Remark
it is possible to draw a less than ogive and a greater than ogive in the same graph.

Eg:- draw a less than ogive and a greater than ogive on the same graph for the following data.

Class	frequency	Class	L.S C.F.	G. C.F	Low	upper
0-9	2	0-9.5	2	30	0	9.5
10-19	5	9.5-19.5	7	28	9.5	19.5
20-29	10	19.5-29.5	17	23	19.5	29.5
30-39	8	29.5-39.5	25	13	29.5	39.5
40-49	3	39.5-49.5	28	5	39.5	49.5
50-59	2	49.5-59.5	30	2	49.5	59.5

Ans) Class.



Type of diagrams

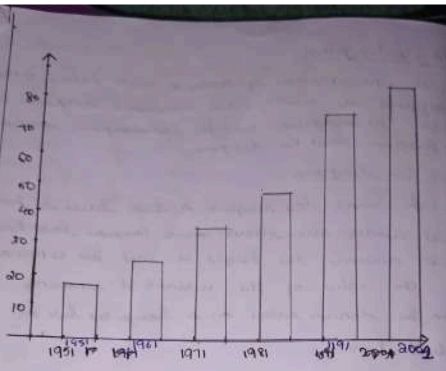
In the presentation of data a wide variety of ~~diagra~~ diagrams are used. Some important diagrams are. Simple bar diagram. multiple bar diagram, component bar diagram and pie diagram.

Simple bar diagram

In Simple bar diagram, a Seer Series of bars of equal width are drawn on a common base line of equal distances the height of each bar is proportional to the value of the variable it represents. it can be drawn either on a horizontal base or vertical base

eg) the urban population of india during the ~~cens.~~ last six census years are given below. Present the data by a Simple bar diagram

Census Year	1951	1961	1971	1981	1991	2001
Population	18	26	35	48	76	82

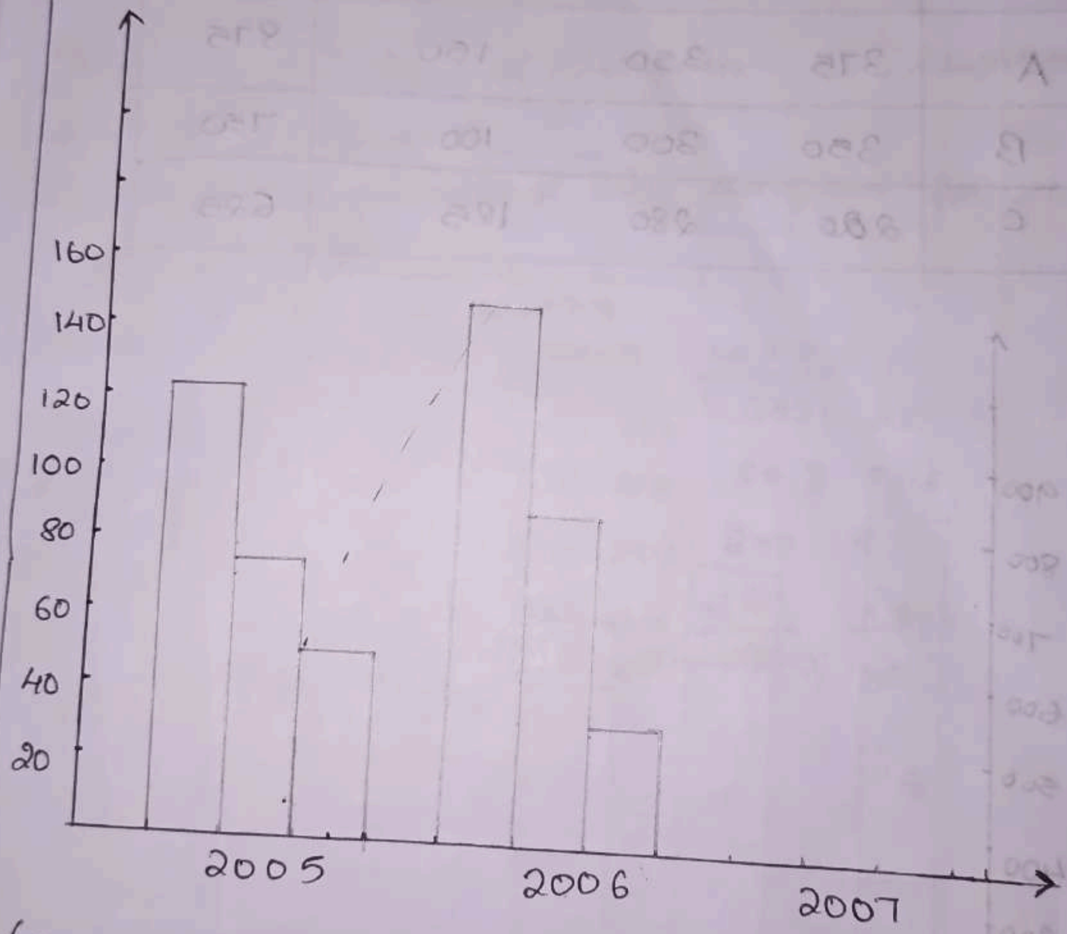


Multipte Bar diagrams

Multiple bar diagrams are used to compare two or more data sets.
in this type of diagram bars representing diff values are drawn side by side and diffrent sets of such bars are shown at equal distances. inorder to distinguish various bars. different colours or shade are used.

eg:- Sales data of 3 companies a, b, c for 3 years are given below. represent the data by means of a multiple bar diagrams.

Year	Sales in years A	B	C
2005	122	75	50
2006	150	90	32
2007	130	81	49



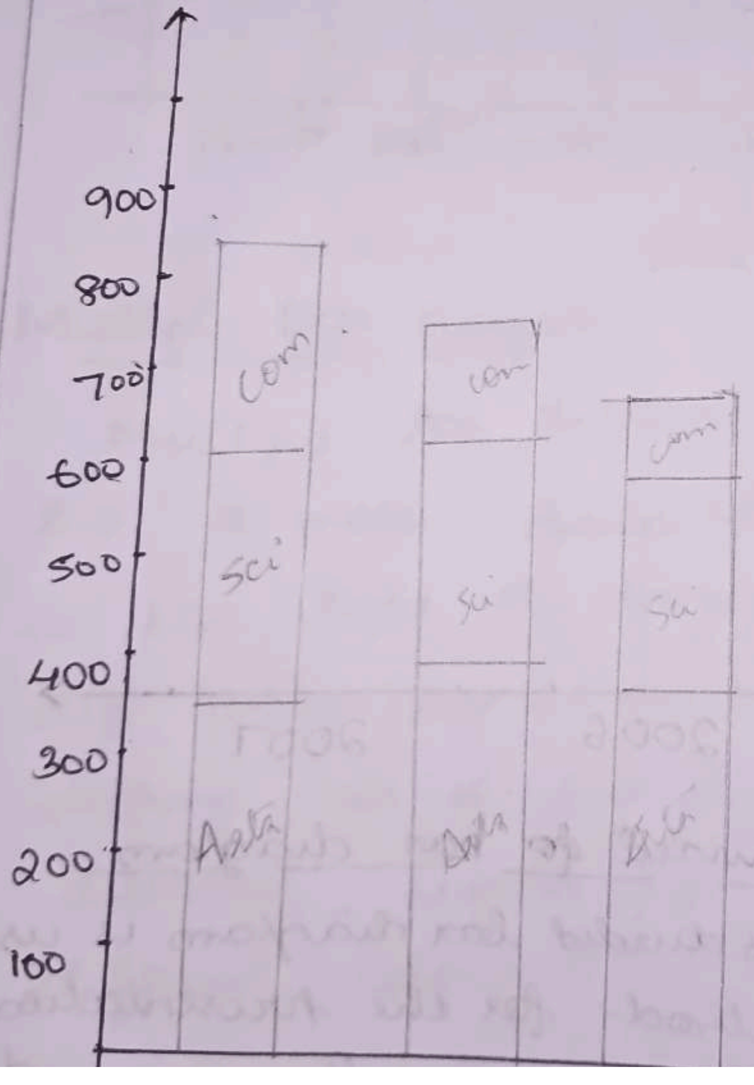
Component or Sub-divided ~~for~~ bar diagram

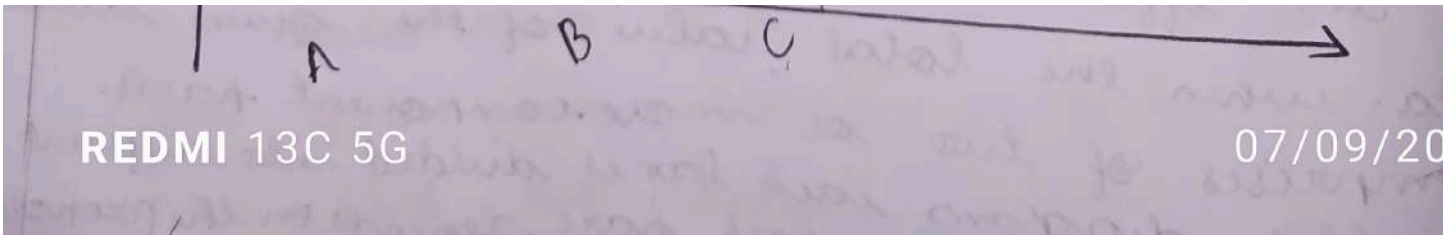
A Component or Sub divided bar diagram is used as an effective method for the presentation of data. when the total value of the given data comprises of two or more component parts. in this diagram each bar is divided into different parts and size of each part depends on the proportional

Share of the component.

eg:- represent the following data by means of component ~~mean st~~ bar diagram.

College.	No of students			Total
	arts	science	commerce	
A	375	350	150	875
B	350	300	100.	750.
C.	220	280	125.	625




A
B
C

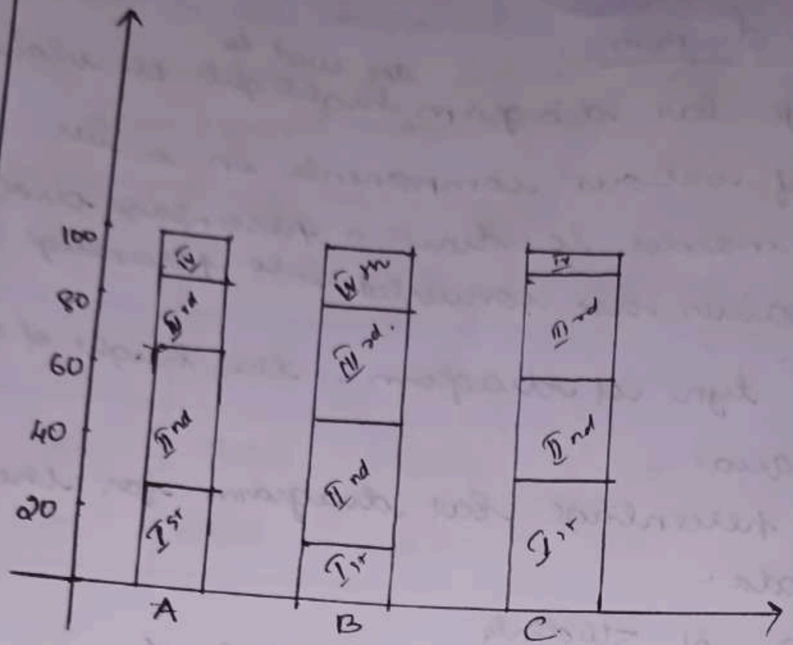
Percentage Bar diagram.

Percentage bar diagram are used to highlights the relative importance of various components in a bar. diagram. inorder to draw a percentage diagram the given values are converted into percentage.

Thus in this type of diagram. the height of all bar will be equal.

eg: Draw a percentage bar diagram for the following data.

Result	NO of students college A	%	college B	%	college C	%
I st	250	29.3	243	15.3	342.	38
II	318	37.2	530	33.5	258	28.6
III	210	24.6	600	37.9	250	27.7
IV.	75	8.85	208	13.15	60	6.666.
	853.	99.95	1581	99.85	900	100.9

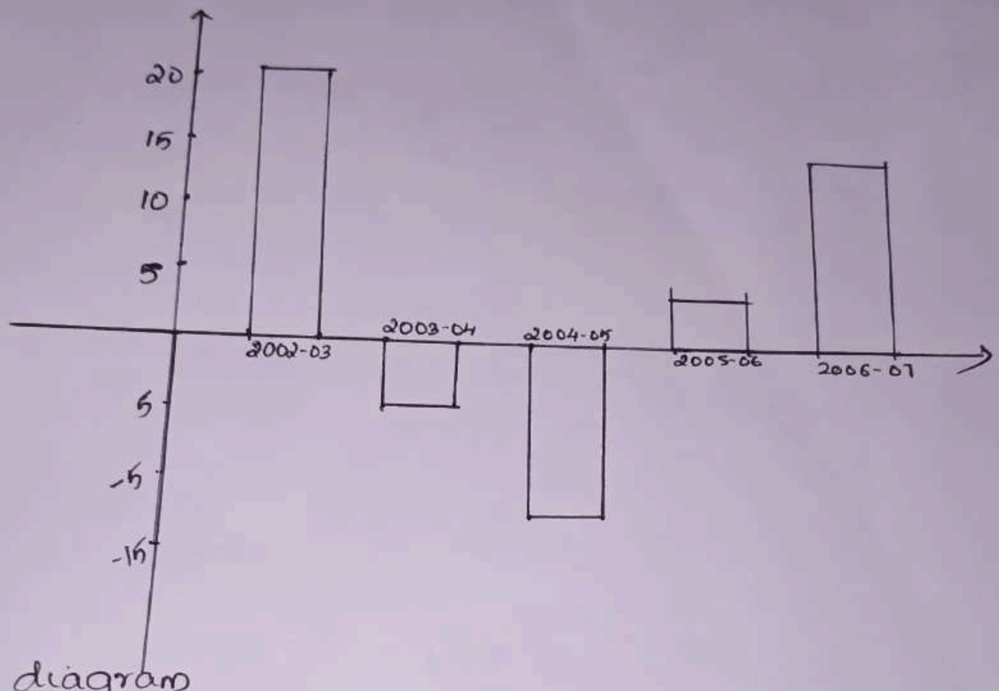


Deviation Bar Diagram:

Values of certain Variables can be positive or negative. in Such cases deviation bar charts can be used to present the data. A deviation chart is a particular type of bar diagram which shows both the +ve & -ve values like other bar diagrams it can be drawn either vertically or horizontally. if bars are drawn vertically, negative values are marked below the base line. and if bars ~~chart~~ are drawn horizontally -ve values are marked in the left side of zero.

Eg:- Represent the following by a deviation chart

Year.	
2002-03	20
2003-04	-5
2004-05	-12.
2005-06	4
2006-07.	14.



Pie diagram

In a pie diagram total value of the given data is shown as the total area of a circle. This circle is then divided into different sectors in such a way that each sector represents a particular component. inorder to design the area of such sectors, central angles proportionate to the corresponding values of the components are to be calculated. This is done by multiplying the ratio of each component to the total value with 360.

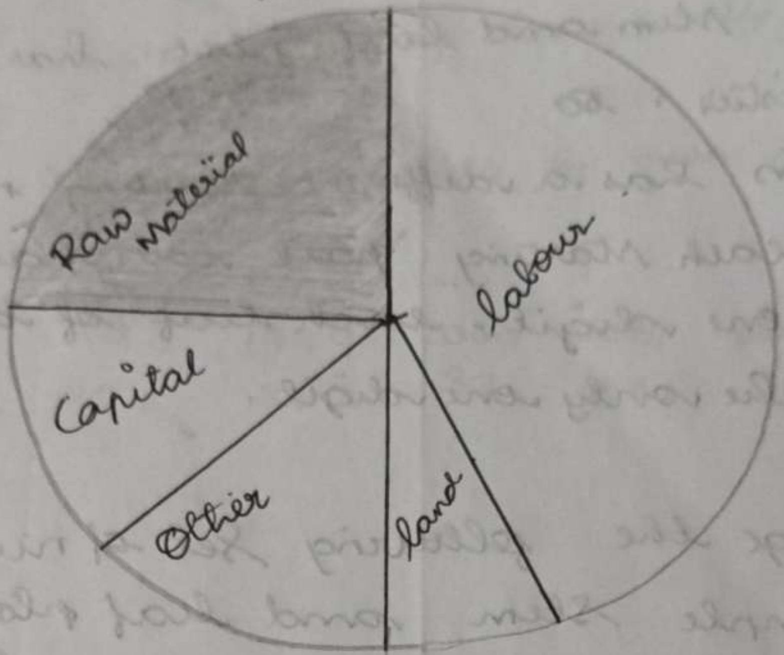
for eg:-
represent the following data using a pie diag

factorinpute	cost in ₹
Land	44
Labour	304
Raw materials	178
Capital	100
Others	94
Total	720

Solution.

The degree of angle representing each component is calculated as follows.

Land	44	$\frac{44}{720} \times 360 = 22$
Labour	304	$\frac{304}{720} \times 360 = 152$
Raw Material	178	$\frac{178}{720} \times 360 = 89$
Capital	100	$\frac{100}{720} \times 360 = 50$
Other	94	$\frac{94}{720} \times 360 = 47$
	720	



Stem And leaf plots.

Stem and leaf plots are techniques that allow rapid and informal exploration of the characterestics of a data ~~d~~ set.

in a typical stem and leaf plot there is a vertical line of numbers called starting parts, and for each starting part. there is a horizontal line of numbers called leaves.

Each complete horizontal line. ~~ie starting part plus~~ (ie starting part plus leaves) is called stem. every number in the data set being displayed has both a starting ~~par~~ ~~poith~~ part and leaf.

The stem width determines which numbers in the data set are recorded on a given stem.

a simple stem and leaf plot has following characterstics · ~~so~~

(1) each stem has a different starting part.

(2) While each starting part can have more than one digit, each leaf of a stem. must be only one digit.

eg-

eg:- Arrange the following set of numbers in a simple stem and leaf plot that has a single digit starting parts & leafs. and stem width of 10.

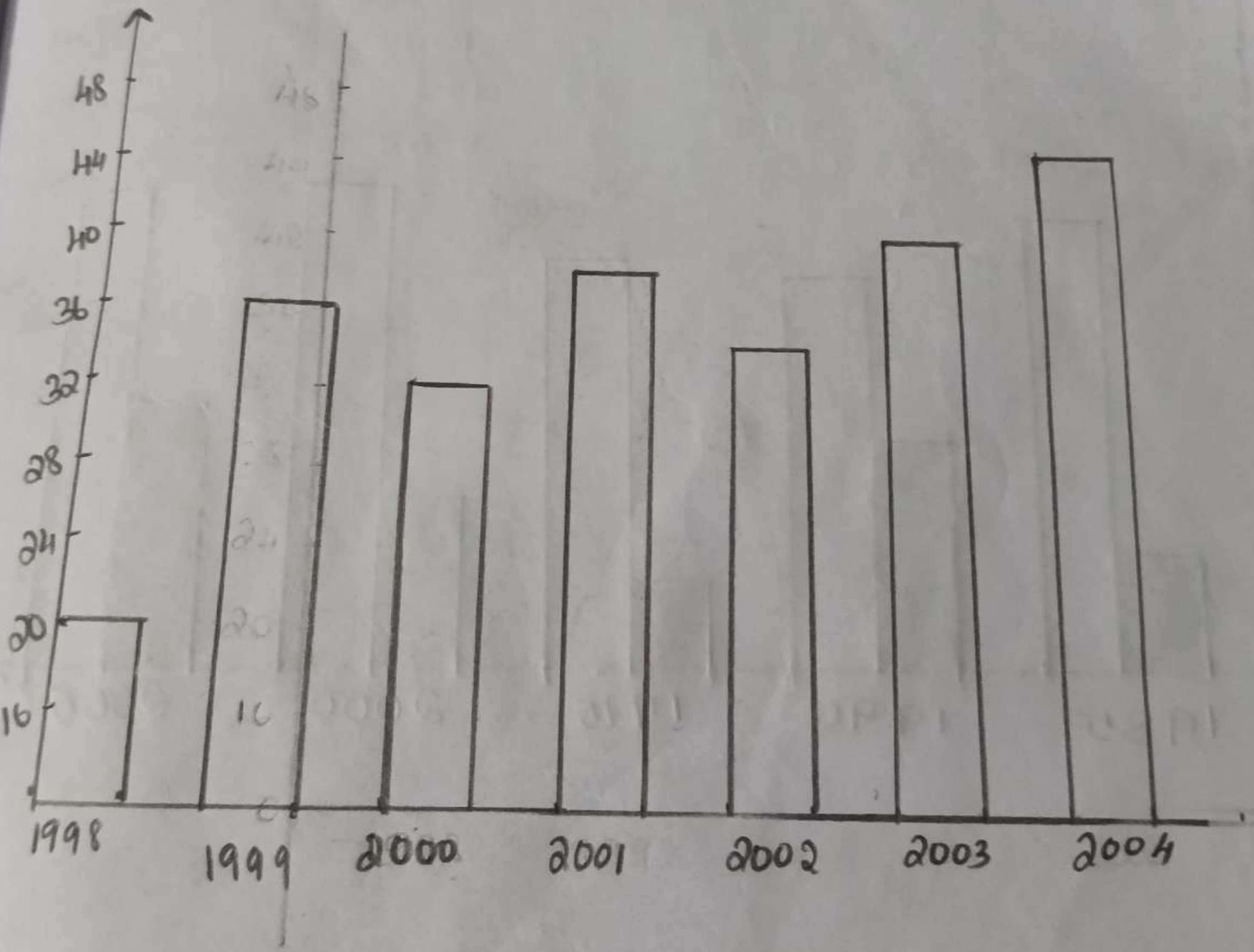
46, 35, 37, 20, 43, 15, 15, 26, 45, 25, 29, 13, 39, 44, 21, 24, 16, 40, 19, 45, 30, 34, 17, 39, 16, 40, 31, 21, 14, 42, 16, 43, 22, 11, 24, 25, 31, 27, 40, 33.

1	5 5 3 6 9 7 6 4 6 1.	(10)
2	0 6 5 9 1 4 1 2 4 5 7	(11)
3	5 7 9 0 4 9 1 3 1 3	(9)
4	6 3 5 4 0 5 0 2 3 0	(10).

Starting parts are the vertical numbers (1, 2, 3, 4) to the left of the vertical line and the leaves are the numbers extending horizontally to the right

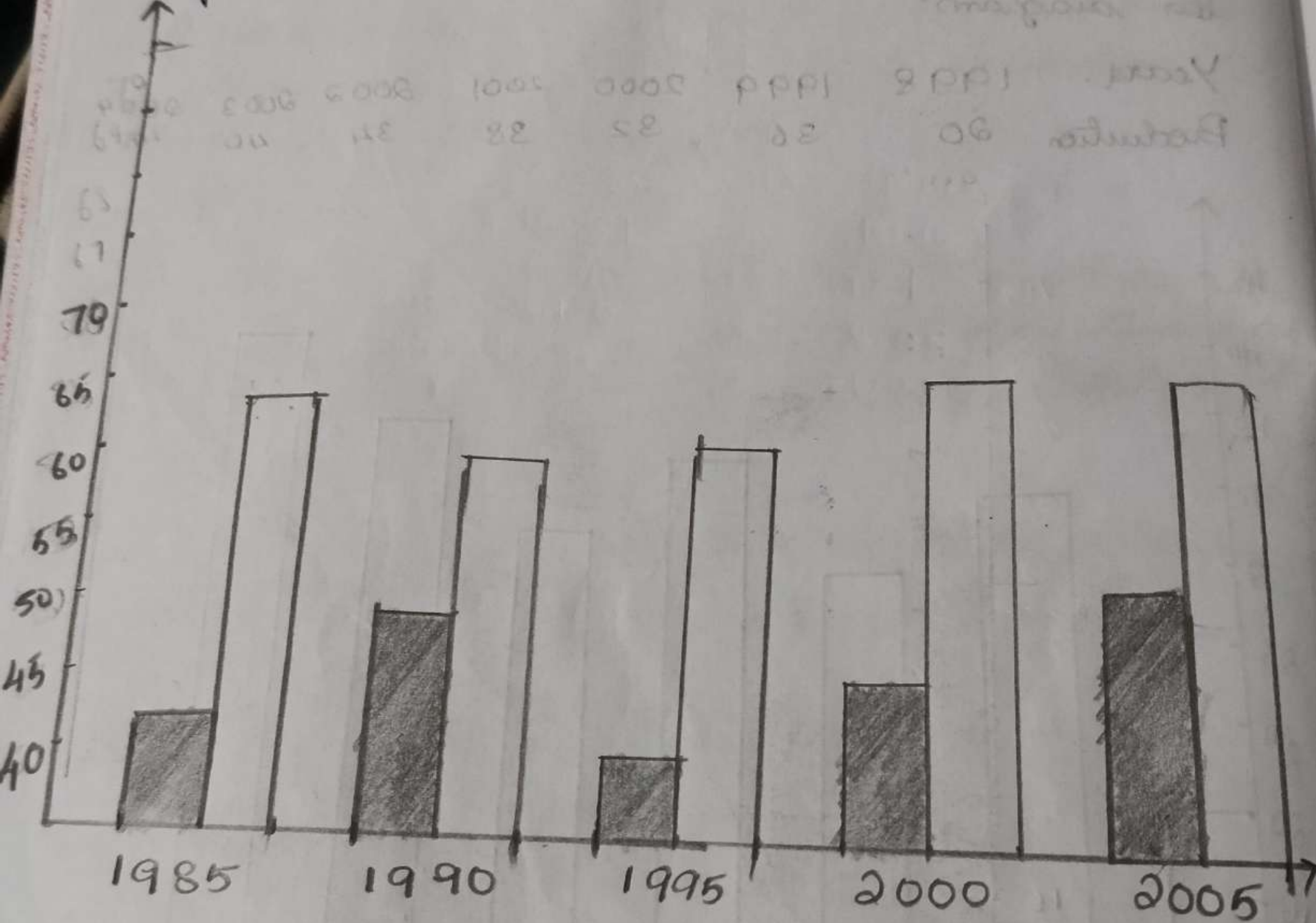
Q). Present the following data by means of a simple bar diagrams.

Years.	1998	1999	2000	2001	2002	2003	2004
Production	20	36	32	38	34	40	45



2 draw a multiple bar diagram for the following data.

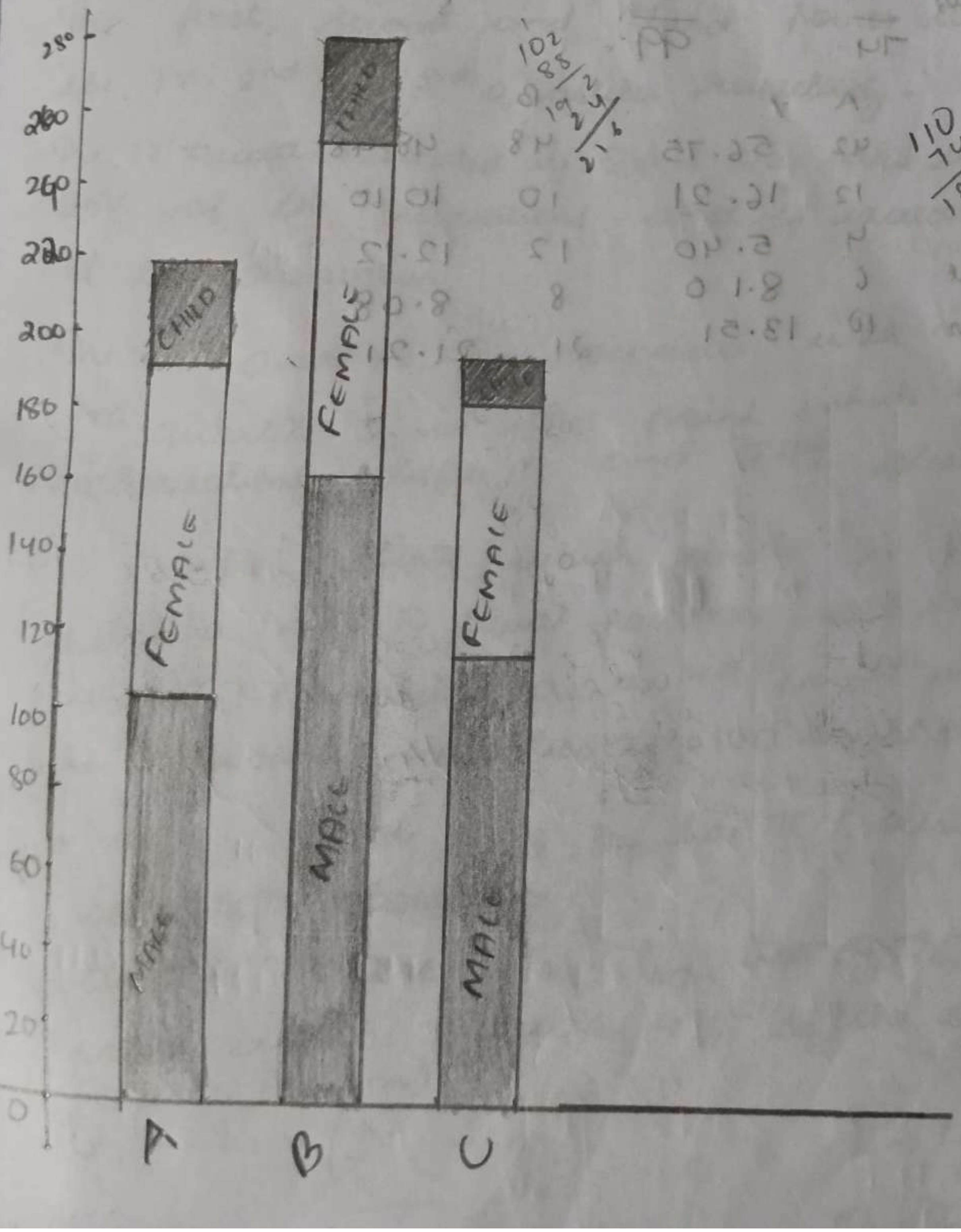
year	1985	1990	1995	2000	2005
export	42	48	40	45	52
import	64	60	62	70	65



70
(10→ 30

3) following data shows the no of workers in 3 factories A, B, C. draw a Component bar diagram.

Factory	No of Workers. Male	female	children	Total
A	102	88	24	214.
B	160	90	30	280.
C	110	74	10	194.



102
88
192
24
216

110
74
184

Partition Values

partition values are the values which divide a frequency distribution into a no of equal parts. the three points which divide the frequency distribution into 4 equal parts are called quartiles. The first, second and third points are known as the 1st, 2nd and 3rd quartiles respectively.

The 1st quartile, denoted by Q_1 is the value which exceed 25% of the observations and is exceeded by 75% of the observation

The 2nd quartile, Q_2 coincides with median the 3rd quartile Q_3 is the point which has 75% observations before it and 25% observations after it.

The 9 points which divide the frequency distribution into 10 equal parts are called deciles. whereas percentiles are the 99 points which divide the frequency distribution into 100 equal parts.

for eg:- The 7th decile, D_7 has 70% observation before i ~~and 30% observation~~.

and 47th percentile. P_{47} ~~has 47% [illegible]~~ is the point which exceeds 47% of the observation

(class & frequency) ↓

$$Q_1 - L_1 + \frac{\left(\frac{N}{4} - m_1\right)}{f_1} \times C.$$

L → Lower limit of Q_1 class.

m_1 → cumulative ~~frequency~~ of just previous Q_1 class.

f_1 → frequency Q_1 class.

c → class width

Graphical Location of Partition Value.

The partition values can be located with the help of a cumilative frequency curve or ogive

The procedure is given as follows.

The procedure is given as follows.

1. first form the less than c.f table. then plot the less than ogive curve. Similiarly to plot the more than ogive curve.

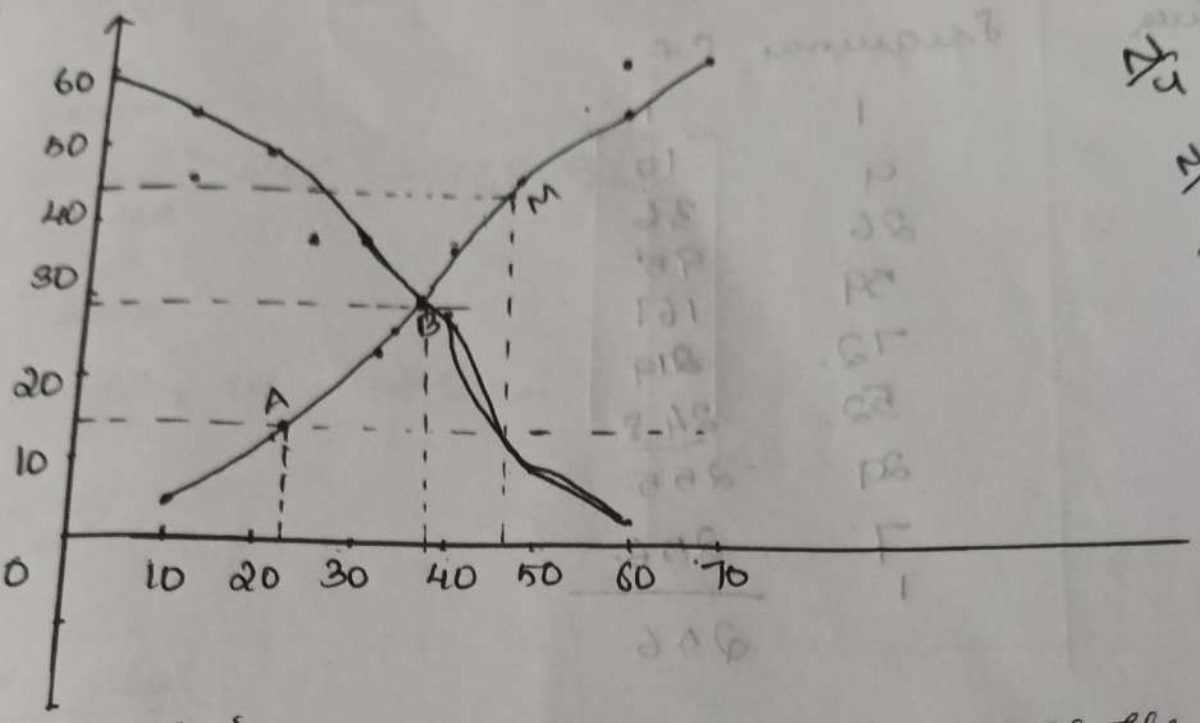
To locate the value of Q_2. mark a point corresponding to $N/2$. where N is the total frequency along Y axis.

at this point draw a line parallel to x axis meeting the ogive at the point A. draw a line ~~from~~ A perpendicular to x axis. meeting it in M. Then ~~absisa~~ abscissa of M gives the value of median.

To locate the values of Q_1 and Q_3. mark the points corresponding to $N/4$. and $3N/4$. and proceed exactly as before.

eg:- locate Q_1, Q_2 & Q_3 for the following data

Mark	No of stud:	L.than		M.than
0-10	4	4		59
10-20	8	12	~~44~~	55
20-30	11	23	~~40~~	47
30-40	15	38	~~32~~	36
40-50	12	50	~~21~~	21
50-60	6	56	~~9~~	9
60-70	3	59	~~3~~	3



$\frac{N}{4} = \frac{59}{4} = 14.$

$\frac{N}{2} = \frac{59}{2} = 29.5$

$\frac{3N}{4} = \frac{3\times59}{4}$

Q) 8 ~~points~~ coins were tossed together and the no of hea
resulting was noted. the experiment is repeate
256 times. and the frequencies that were ~~ob~~
obtained for different values: are shown in the
following tables.

Values	Frequencies..
0	1
1	9
2	26
3	59
4	72
5	52
6	29
7	7
8	1

calculate Q_1, Q_3, D_4, and P_{27}

value	frequency	c.f
0	1	1
1	9	10
2	26	36
3	59	95
4	72	167
5	52	219
6	29	24.8
7	7	255
8	1	256
	256	

$$\frac{N}{4} = \frac{256}{4} = 64.$$

The Cumilative frequency . first greater than 64 is 95.

Q_1 = Value . Corresponding to the frequency 95

= 3.

Q_3 . $\frac{3N}{4} = 192$

$Q_3 = 5$.

$Q_2 = 4$ value corresponding to the $\frac{N}{2}$ th freq

$\underline{\underline{= 4}}$.

to find D_4 :

$$\frac{4}{10}N = \frac{4}{10} \times 256 = 102.4.$$

$P_{27} = \frac{27}{100} \times 256 = 69.12$.

$= 3$.

$P_{66} = \frac{66}{100} \times 256. = 168.96$

$= 5$.

$P_{90} = \frac{90}{100} \times 216 = 230.4$.

$= 6$.

Box - plot.

box plot is a method for demonstrating graphically the spread of numerical data through their quartiles. in addition to the box on a box plot. there can be lines called whiskers. Hence the plot is Sometimes called Box and Whisker plot.

A box plot is a standarised way of displaying the data set based on. the 5 number Summary; minimum, maximum, median, first quartile and. 3rd quartile.

In addition to the minimum and maximum values, another important element can also be used to obtain a box plot is the inter quartile range. $IQR = Q_3 - Q_1$

A box is drawn from Q_1 to Q_3 with a horizontal line drawn inside to denote the median.