(x) at least two of A B C occur

Elements of Statistics :~

Frequency Distribution :~

Let X be a population r.v. which is discrete and a sample of size 'n' be taken. If $x_1, x_2 \cdots x_n$ are the distinct values of X in the sample, when x_i occurs f_i times $i = 1, 2 \cdots k$, f_i is called the frequency of x_i. ~~Thus the~~

clearly $f_1 + f_2 + \cdots f_k = n$.= Size of the sample.

x	Frequency
x_1	f_1
x_2	f_2
⋮	⋮
x_k	f_k

Ex – Class of 20 students scored the following marks in an exam.
0, 10, 12, 5, 6, 8, 12, 15, 10, 18, 20, 15
18, 11, 25, 28, 32, 30, 42, 48

Frequency Table :

Marks x_i	Number of students f_i
	1
0	1
5	1
6	1
8	2
10	1
11	2
12	2
15	2
18	1
20	1
25	1
28	1
32	1
42	1
48	
	20

~~For~~ X cont^d : We devide the range of X into some intervals (usually of equal length). called class interval.

The limits within which a class interval lies are called class limits.

Class boundaries : The true values of the end points of the class intervals.

Total frequency : The sum of all class frequency.

Midpoint or class mark :

$$\frac{\text{upper limit of the class} + \text{lower limit of the class}}{2}$$

Class length : The difference between the upper & the lower boundary of a class.

Class limit : The limits with in which a class interval lies are called class limits.

class limit	Class boundary	class limit	Class boundary
10–14	9.5 – 14.5	10 – 14.9	9.95 – 14.95
15–19	14.5 – 19.5	15–19.9	14.95 – 19.95

open end class : The lowest class lacking a lower limit or the highest class lacking an upper limit are both side to be open-end classes.

→ The income of 850 people in a factory in a certain year is given as Under.

Problem

Income	Frequency
Under 2000	250
2000 – 3000	100
3000 – 4000	150
4000 – 5000	100
5000 – 6000	70
6000 – 7000	50
7000 and outer (open end class)	130
	850

Cumulative Frequency :- The 'less than' cumulative frequency (cf) of a class is the ~~Frequency~~ total frequency of all values 'less than' the upper boundary of the class. Similarly, the 'more than' cumulative frequency (cf) of a class is the total frequency of all values which are greater than the lower boundary of the class.

Frequency		Less than cf		More than cf	
Rs (x)	No of workers	Less than x (or upper boundary)	cf	More than x (or lower boundary)	cf
50.5 – 60.5	5	60.5	5	50.5	40
60.5 – 70.5	7	70.5	12	60.5	35
70.5 – 80.5	8	80.5	20	70.5	28
80.5 – 90.5	10	90.5	30	80.5	20
90.5 – 100.5	10	100.5	40	90.5	10
Total	40				

Scanned by CamScanner

<u>relative frequency</u> . if f_i is the frequency of a class & its length, and N the total frequency. then

relative frequency of the class = f_i/N

frequency density of the class = f_i/l .

Graphical representation of frequency distribution

Frequency polygon :

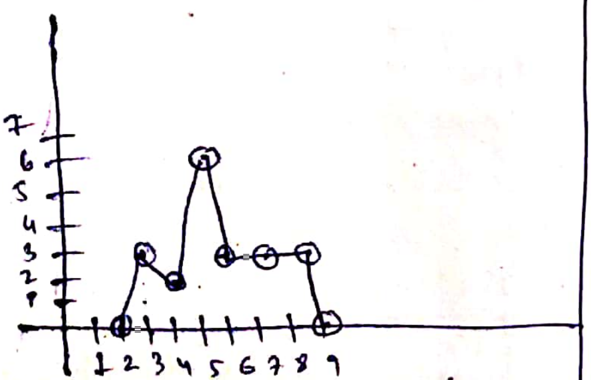
EX → The number of petals Counted for 20 flowers of a certain species yields the following observation.

8, 8, 4, 5, 3, 7, 5, 6, 3, 4
5, 6, 7, 5, 5, 3, 5, 6, 7, 8

The frequency Table →

No of petals	Frequency
2	0
3	3
4	2
5	6
6	3
7	3
8	3
9	0
Total	20

frequency polygon :

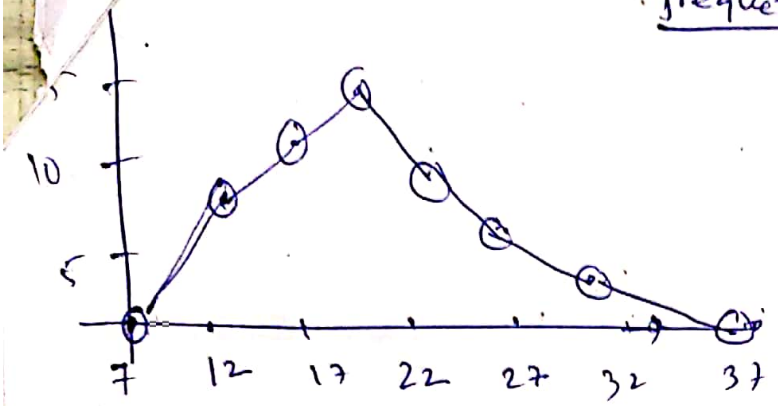


EX-2 : Wages in Rs — 10-14 15-19 20-24 25-29 30-34
No of worker — 8 12 15 10 5

Class interval	Mid point	Frequency
5-9	7	0
10-14	12	8
15-19	17	12
20-24	22	15
25-29	27	10
30-34	32	5
35-39	37	0

1.

frequency polygon.



Histogram :- ~~Dra~~

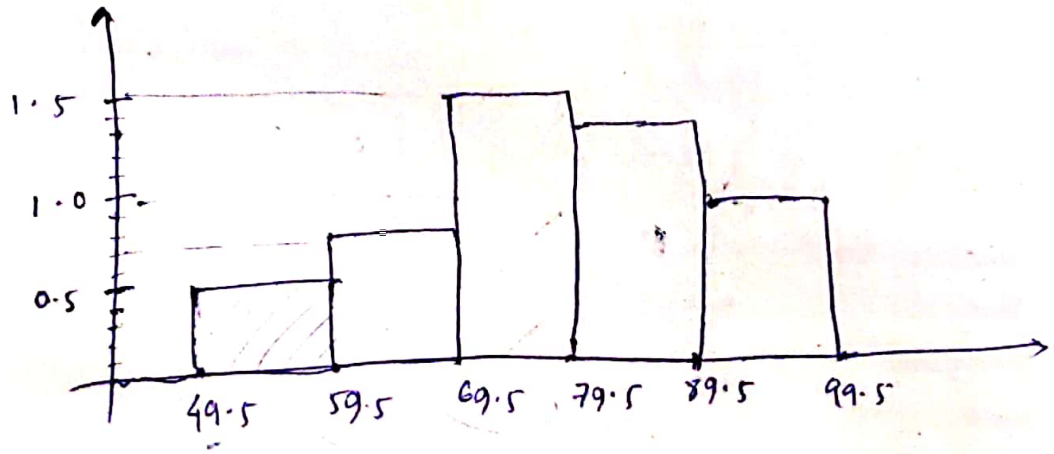
Ex-1: Draw a histogram of ~~a~~ the following data.

Rs	50–59	60–69	70–79	80–89	90–99
No of workers	5	7	15	13	10

As:

Frequency Table

class interval	class boundary	f	Frequency density
50–59	49.5 – 59.5	5	0.5
60–69	59.5 – 69.5	7	0.7
70–79	69.5 – 79.5	15	1.5
80–89	79.5 – 89.5	13	1.3
90–99	89.5 – 99.5	10	1.0

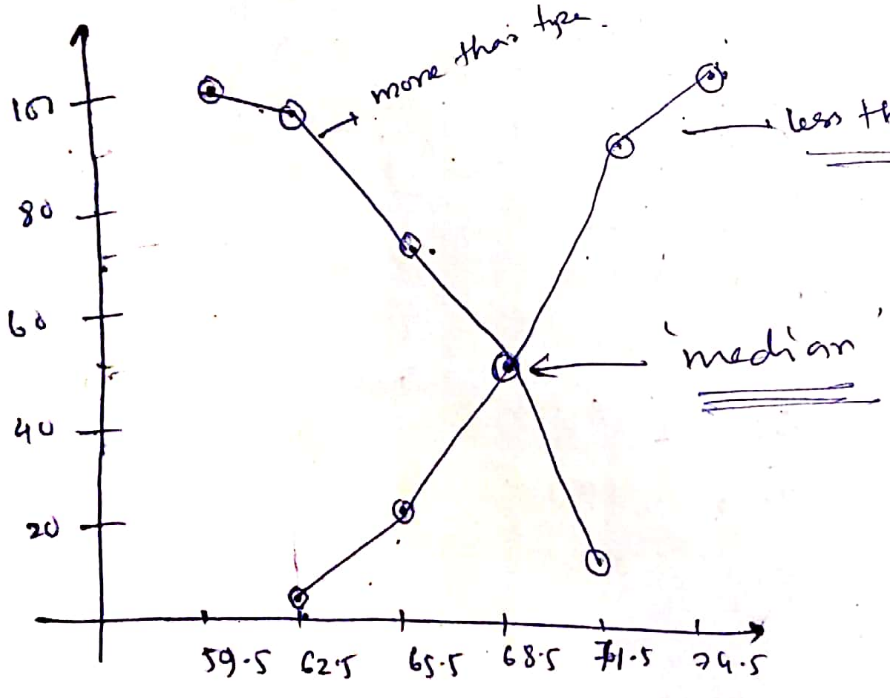


(x) at least two of A B C occur $(A\cap B)\cup(B\cap C)\cup(A$ ... $)^c)$

Ogive :- Ex :

Weight (kg)	60 – 62	63 – 65	66 – 68	69 – 71	72
No. of Persons :	5	20	25	40	10

Frequency		Less than 'cf'		More than 'cf'	
class boundary	f	less than x	cf	More than x	cf
59.5 – 62.5	5	62.5	5	59.5	100
62.5 – 65.5	20	65.5	25	62.5	95
65.5 – 68.5	25	68.5	50	65.5	75
68.5 – 71.5	40	71.5	90	68.5	50
71.5 – 74.5	10	74.5	100	71.5	10



A. Measures of location or measures of Central Tendency:

I. Arithmetic mean or simply the mean
II. Geometric mean
III. Harmonic Mean
IV. Median and quartile
V. Mode

Arithmetic Mean: The arithmetic mean (A.M.) or the mean of a set of numbers $x_1, x_2 \ldots x_n$, denoted by $\bar{x}$, is defined as

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

If the numbers $x_1, x_2 \ldots x_n$ occur $f_1, f_2 \ldots f_n$ times respectively (i.e. with respective frequencies $f_1, f_2 \ldots f_n$). The arithmetic mean is given by

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$$

Where $N = \sum_{i=1}^{N} f_i$

change of origin and scale: For simplification of numerical calculation, we often change the origin and diminish the Scale as stated in the following theorem.

Thm: If $x_1, x_2 \ldots x_n$ are a set of n-numbers with respective frequencies $f_1, f_2 \ldots f_n$ and if

$$u_i = \frac{x_i - x_0}{h}, \quad i = 1, 2 \ldots n$$

then $\bar{x} = x_0 + h\bar{u}$. and $\sum_{i=1}^{n} f_i = N$.

where h and x_0 are Constants

Proof: clearly, $x_i = x_0 + hu_i$. So,

$$\frac{1}{N} \sum_{i=1}^{n} f_i x_i = x_0 \frac{1}{N} \sum_{i=1}^{n} f_i + h \frac{1}{N} \cdot \sum_{i=1}^{n} f_i u_i$$

$$\Rightarrow \bar{x} = x_0 + h\bar{u}$$

Since $\sum_{i=1}^{n} f_i = N$.

Ex:

class interval	class mark(x_i)	f_i	u_i	$f_i u_i$
20 - 29	24.5	2	-3	-6
30 - 39	34.5	4	-2	-8
40 - 49	44.5	8	-1	-8
50 - 59	54.5	27	0	0
60 - 69	64.5	18	1	18
70 - 79	74.5	15	2	30
80 - 89	84.5	6	3	18
		$N = 80$		44

Here, we have taken, $x_0 = 54.5$,

$$\& \quad u_i = \frac{x_i - x_0}{h} = \frac{x_i - 54.5}{10}$$

$$\therefore \sum f_i u_i = 44$$

$$\therefore \bar{u} = \frac{44}{80}$$

$$\therefore \bar{x} = x_0 + h\bar{u} = 54.5 + 10 \cdot \frac{44}{80} = 54.5 + 5.5 = 60.$$

Geometric mean (GM) The geometric mean of a set of N +ve numbers $x_1, x_2 \cdots x_n$ denoted by x_g, is defined by.

$$x_g = (x_1 x_2 \cdots x_n)^{1/n}$$

or equivalently, $(x_g)^n = x_1 x_2 \cdots x_n$.

If the +ve numbers $x_1, x_2 \cdots x_n$ occur with respective frequencies $f_1, f_2 \cdots f_n$ such that $\sum_{i=1}^{n} f_i = N$

then

$$x_g = \left(x_1^{f_1} x_2^{f_2} \cdots x_n^{f_n}\right)^{1/N}$$

or

$$\boxed{\log(x_g) = \frac{1}{N} \sum_{i=1}^{n} f_i \log(x_i)}$$

Harmonic mean (H.M)

The Harmonic mean (H.M) of a set of N +ve numbers $x_1, x_2 \cdots x_N$ denoted by x_h is defined as

$$x_h = \frac{N}{(1/x_1) + (1/x_2) + \cdots (1/x_N)} = \frac{N}{\sum_{i=1}^{N} (1/x_i)}$$

If the +ve numbers $x_1, x_2 \cdots x_n$ occur with resptive frequencies $f_1, f_2 \cdots f_n$ such that $\sum_{i=1}^{n} f_i = N$ then

$$HM = \frac{N}{(f_1/x_1) + (f_2/x_2) + \cdots (f_n/x_n)} = \frac{N}{\sum_{i=1}^{n} f_i/x_i}$$

Thm: $AM \geq GM \geq HM$.

Median :- The median of a set of numbers arranged in the order of their magnitudes, i.e in an array, is the middle value or the arithmatic mean of the two middle values.

So if N is odd, median $= \frac{N+1}{2}$ th item

if N is even, media $= \frac{1}{2}\left(\frac{N}{2} \text{ th item} + \left(\frac{N}{2}+1\right) \text{th item}\right)$

[V] For a grouped frequency distribution: the median is obtained by the formula.

$$\text{median} = L_1 + \frac{(N/2) - C}{f} \cdot i$$

where L_1 = lower class boundary of the median class. (i.e the class containing median)

N = Total frequency

C = Sum of the frequencies of all classes lower than the median class. (i.e less than cumulative frequency of the class preceding the median class).

f = frequency of the median class

i = width of the median class.

Quantile: If the data is arranged in order of magnitude then quantile are the three values which divide the data into four equal parts. The three quartiles are denoted by Q_1, Q_2, and Q_3, where Q_1 is the lower or first quartile, Q_2, the 2nd quartile is also the median and Q_3 is the upper ~~third~~ or third quartile.

\# mid range $= \frac{1}{2}(x_1 + x_n)$ x_1 is smallest, x_n largest value.

$$\text{Range} = x_n - x_1$$

$$\text{Inter quantile range} = Q_3 - Q_1.$$

For a grouped frequency distribution quatiles are obtained as follows Q_k = Value of $\left(k \frac{N}{4}\right)$ th item, $k = 1, 2, 3$

After the quantile class is determined, the quantile Q_K is then obtained by the method of interpolation as:

$$Q_K = L_1 + \frac{k(N/4) - C}{f} i \quad , \quad k = 1, 2, 3.$$

where
L_1 = lower class boundary of the quartile class
N = Total frequency
C = 'less than' Cumulative frequency of the class preceding the quartile class.
f = frequency of the quartile class
i = width of the quartile class.

For ungroup (discrete) frequency distribution.

$$Q_K = \text{Value of } \left(k \frac{N+1}{4}\right)\text{th item}, \quad k = 1, 2, 3.$$

Mode :~ Mode is the value of the sample which occurs with the highest frequency.

For samples taken from a continuous population, where the data is grouped into classes, the mode is obtained by the method of interpolation,

i.e. $$\text{Mode} = L_1 + \frac{d_1}{d_1 + d_2} \times i$$

Where L_1 = lower class boundary of the modal class (i.e. the class with the highest frequency)
f = frequency of the modal class
f_1 = " of the class immediately preceding the modal class
f_2 = frequency of the class immediately following the modal class.

i = width of the modal class.

$d_1 = f - f_1$

$d_2 = f - f_2$

###

Measures of Dispersion :~ It indicates the extend to which the values are scattered away from the center (mean).

There are three measures of ~~dis~~ dispersion in general.

I. Standard deviation !

let $x_1, x_2 \cdots x_N$ be a set of N real numbers. The standard deviation S.D of these numbers, denoted by S_x is defined as, $S_x = \left[\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2\right]^{1/2}$.

here the quantity. S_x^2 is called the variance of the given numbers.

.. When some of the values are repeated i.e if $x_1, x_2 \cdots x_n$ occur $f_1, f_2 \cdots f_n$ times such that $\sum_{i=1}^{n} f_i = N$. then

$$S_x = \left[\frac{1}{N} \sum_{i=1}^{n} f_i (x_i - \bar{x})^2\right]^{1/2}.$$

For a sample drawn from a grouped frequency distribution (conts), we will take the mid point as the representative of the class.

II Change of origin and scale : let the numbers $x_1, x_2 \cdots x_n$ occurs with frequencies $f_1, f_2 \cdots f_n$ respectively and $u_i = \frac{x_i - x_0}{h}$, $i = 1, 2 \cdots n$. & x_0 & h are constant.

Then $S_x^2 = h^2 S_u^2$ or $S_x = |h| S_u$.

Proof : Do it your self.

II Mean absolute deviation : The mean absolute deviation (M.A.D), also called the mean deviation of a sample $\{x_1, x_2 \cdots x_N\}$ of size N is defined as

$$M.A.D = \frac{1}{N}\sum_{i=1}^{N}|x_i - \bar{x}|,$$ where $\bar{x}$ is the mean.

If $x_1, x_2 \cdots x_n$ occur with corresponding frequencies $f_1, f_2 \cdots f_n$, then

$$M.A.D = \frac{1}{N}\sum_{i=1}^{N} f_i |x_i - \bar{x}|, \quad \text{when } \sum_{i=1}^{N} f_i = N$$

For a grouped frequency distribution, the mid point of a class is taken as the respresentative member of the class.

Bivariate Data: $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$

[I] **Covariance:** The co-variance of a bivariate data $\{(x_i, y_i) : i = 1, 2 \cdots N\}$ denoted by S_{xy} or $Cov(x, y)$ is defined as

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}).$$

where $\bar{x}$ = mean of $\{x_i : i = 1, 2 \cdots N\}$
$\bar{y}$ = mean of $\{y_i : i = 1, 2 \cdots N\}$

Th^m: $S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i y_i) - \bar{x}\bar{y}$

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\frac{1}{N}\sum_{i=1}^{N} x_i + \bar{y}\frac{1}{N}\sum_{i=1}^{N} y_i + \bar{x}\bar{y}$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y} - \cancel{\bar{x}\bar{y}} + \cancel{\bar{x}\bar{y}}$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

[II] **Correlation Co-efficient:** -

$$r = r_{xy} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y} = \frac{S_{xy}}{S_x S_y} = \frac{Cov(x, y)}{S_x S_y}$$

* <u>change of origin and scale</u> :

$u_i = \frac{x_i - x_0}{h}$ & $v_i = \frac{y_i - y_0}{k}$, $i = 1, 2 \ldots N$

$r_{xy} = r_{uv}$ if h, k are same sign

$r_{xy} = -r_{uv}$ if h, k are opposit sign.

<u>Th^m</u> : $-1 \leq r_{xy} \leq 1$

<u>Regression lines</u> :-

The regression line of y on x is

$$y - \bar{y} = b_{yx} (x - \bar{x})$$

when $b_{yx} = r \frac{S_y}{S_x} = \frac{Cov(x,y)}{S_x^2}$

Similarly, the regression line of x on y is

$$x - \bar{x} = b_{xy} (y - \bar{y})$$, when $b_{xy} = r \frac{S_x}{S_y} = \frac{Cov(x,y)}{S_y^2}$

<u>Relationship between Mean, Mode & Median</u> :

$$\boxed{(\text{Mean} - \text{Mode}) = 3\,(\text{Mean} - \text{Median})}$$