Chapter : 2

Data Processing

---

1. Choose the right answer from the four alternatives given below:

(i). The measure of central tendency that does not get affected by extreme values
(a) Mean
(b) Mean and mode
(c) Mode
(d) Median

Answer: (d) Median

(ii). The measure of central tendency always coinciding with the hump of any distribution is:
(a) Median
(b) Median and Mode
(c) Mean
(d) Mode

Answer: (b) Median and Mode

(iii). A scatter plot represents negative correlation if the plotted values run from:
(a) Upper left to lower right
(b) Lower left to upper right
(c) Left to right
(d) Upper right to lower left

Answer: (a) Upper left to lower right

2. Answer the following questions in about 30 words:

(i). Define the mean.

Answer: The mean is the value which is derived by summing all the values and dividing it by the number of observations.

$$\overline{X} = \frac{\text{Sum of observations}}{\text{No. of observations}}$$

(ii). What are the advantages of using mode?

Answer: Mode is the maximum occurrence or frequency at a particular point or value. The biggest advantage of mode is that it is not affected by extreme values. It can be determined even for open ended series.

(iii). What is dispersion?

Answer: The term, 'dispersion', refers to the scattering of scores about the measure of central tendency. It is used to measure the extent to which individual items or numerical data tend to vary or spread about an average value. Thus, to get a better picture of a distribution, we need to use a measure of central tendency and of dispersion or variability.

(iv). Define correlation.

Answer: Correlation is basically a measure of relationship between two or more sets of data. It serves a very useful purpose.

(v). What is perfect correlation?

Answer: Perfect correlation means that there is proportional relationship between two variables. If on doubling x, the value of y also gets doubled, it is perfect positive correlation. On the other hand, if on doubling the variable x, the value of y gets halved, it is called perfect negative correlation.

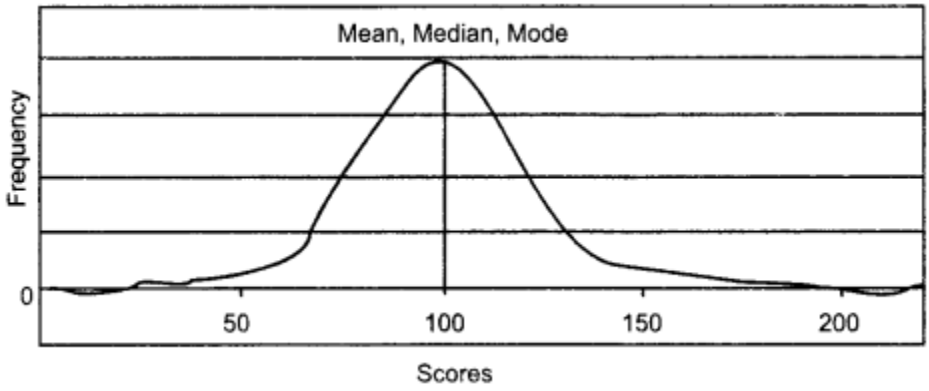
(vi). What is the maximum extent of correlation?

Answer: The value of correlation lies between -1 and +1. Closer it is to zero, weaker is the correlation; closer it is towards ±1, stronger is the correlation. Symbolically $-1 \leq r \leq +1$.

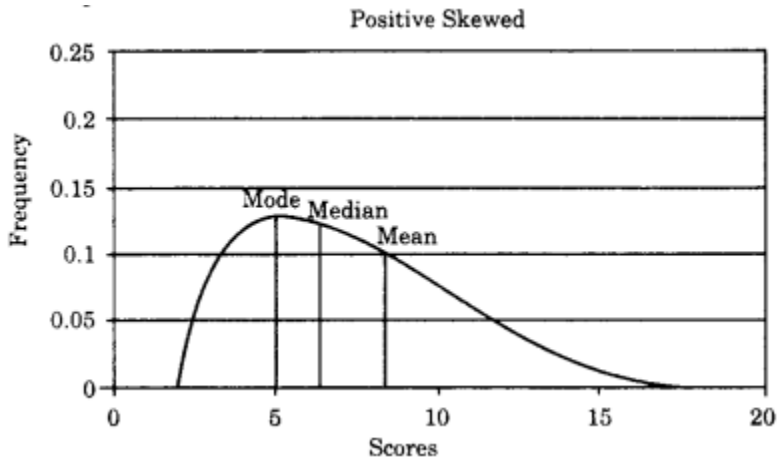
3. Answer the following questions in about 125 words.

(i). Explain relative positions of mean, median and mode in a normal distribution and skewed distribution with the help of diagrams.

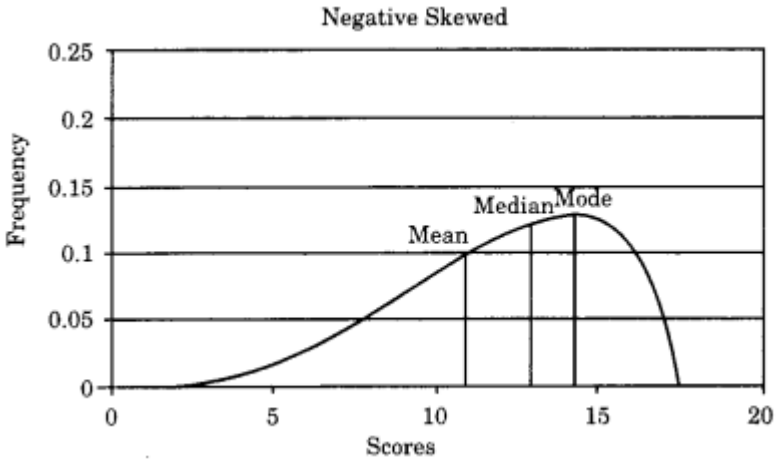
Answer: (a) Normal Curve: In this curve, highest frequency is at the centre and both tails on the left and right hand side are in identical fashion. It is a unimodal curve in which mean, median and mode are equal. It is also known as a bell¬shaped or symmetrical curve. It is shown below:



(b) Positively Skewed Curve: It is a symmetrical curve which has a tail on the right hand side of the graph and frequencies are more for the lower values of the data. These histograms have the curve on the left side of the distribution. If the right tail is longer, the mass of the distribution is concentrated on the left. It has relatively few low values. It is shown below:



(c) Negatively Skewed Curve: It is a symmetrical curve which has a tail on the left hand side of the graph and frequencies are more for the higher values of the data. The left tail is longer, the mass of the distribution is concentrated on the right of the figure. It has relatively few low values. The distribution is said to be left-skewed. It is shown below:



(ii). Comment on the applicability of mean, median and mode (Hint: from their merits and demerits)

Answer:

Mean:

- It is the simplest among all the measures of central tendency.
- It is based on all the items in a series. Hence, it is a representative value of different items.
- It is a value. It has no scope for estimated values.
- It is a stable form of central tendency.
- It can be used for comparison.

Median:

- Median is not affected by the extreme values of the series.
- For finding median, only the middle values and their units are sufficient.
- Median can also be determined through the graphic representation of data.
- Median value is always certain in a series.
- Median value is a real value.

Mode:

- Mode is a very simple measure of central tendency.
- It is less affected by extreme and marginal values.
- Mode is the best representation of the series.

- It can also be determined graphically.

(iii). Explain the process of computing Standard Deviation with the help of an imaginary example.

Answer: Standard deviation (SD) is the most widely used measure of dispersion. It is defined as the square root of the average of squares of deviations. It is always calculated around the mean. The standard deviation is the most stable measure of variability and is used in so many other statistical operations. The Greek character a denotes it.

Steps:

- To obtain SD, deviation of each score from the mean (x) is first squared (x^2).
- It makes all negative signs of deviations positive. It saves SD from the major criticism of mean deviation which uses modulus x. Then, all of the squared deviations are summed -x^2
- (care should be taken that these are not summed first and then squared).
- This sum of the squared deviations (x^2) is divided by the number of cases and then the square root is taken. Therefore, Standard Deviation is defined as the root mean square deviation.

Calculate the standard deviation for the following distribution:

Groups	120-130	130-140	140-150	150-160	160-170	170-180
f	2	4	6	12	10	6

Solution:

The method of obtaining SD for grouped data has been explained in the table below. The initial steps upto column 4, are the same as those we followed in the computation of the mean for grouped data. We begin with deviation value of zero has been assigned to the group. Like wise other deviations are determined. Values in column 4(fx') are obtained by the multiplication of the values in the two previous columns. Values in column 5(fx' 2) are obtained by multiplying the values given in column 3 and 4. Then various columns have been summed.

(1) Group	(2) f	(3) x'	(4) fx'	(5) $fx'2$
120-130	2	–3	–6	18
130-140	4	–2	–8	16
140-150	6	–1	$\frac{-6}{-20}$	6
150-160	12	0	0	0
160-170	10	1	10	10
170-180	6	2	$\frac{12}{22}$	24
	N = 40		$\sum fx' = 2$	$\sum fx'2 = 74$

The following formula is used to calculate the Standard Deviation:

$$SD = t^2 \left| \Sigma fx'^2 - \frac{\Sigma fx'}{N} \right.$$

(iv). Which measures of dispersion is the most unstable statistic and why?

Answer: Range is the most unstable statistic because:

- Range is not based on all the terms. Only extreme items reflect its size. Hence, range cannot be completely representative of the data as all other middle values are ignored.
- Due to the above reason, range is not a reliable measure of dispersion.
- Range does not change even the least even if all other, in between, terms and variables are changed.
- Range is too much affected by fluctuation of sampling. Range changes from sample to sample. As the size of sample increases range increases and vice-versa.
- It does not tell us anything about the variability of other data.
- For open-end intervals, range is indeterminate because lower and appear limits of first and last interval are not given.

(v). Write a detailed note on the degree of correlation.

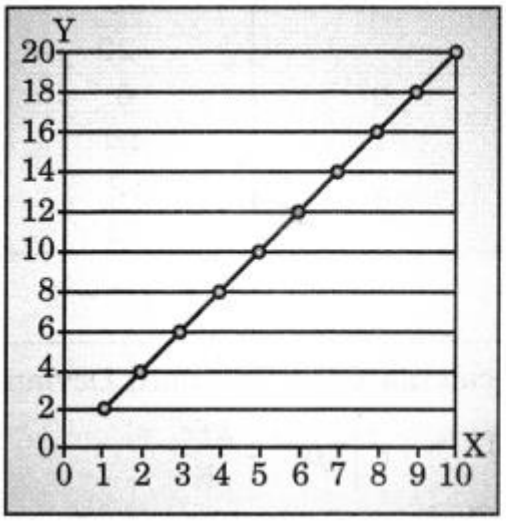
Answer: Through the coefficient of correlation, we can measure the degree or extent of the correlation between the two variables. On the basis of the coefficient of

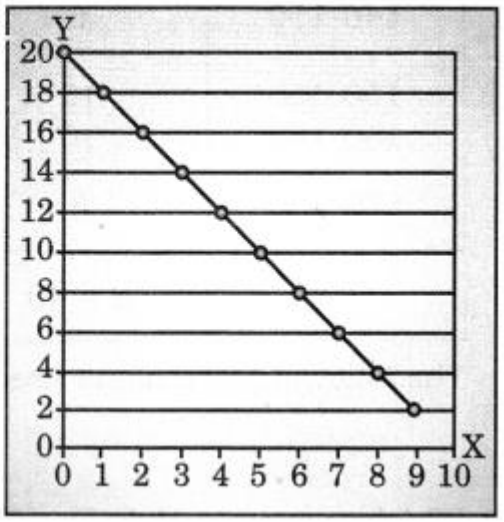
correlation, we can also determine whether the correlation is positive or negative and also its degree or extent.

Perfect correlation: If two variables change in the same direction and in the same proportion, the correlation between the two is perfect positive. According to Karl Pearson, the coefficient of correlation, in this case, is +1. On the other hand, if the variables change in the opposite direction and in the same proportion, the correlation is perfect negative. Its coefficient of correlation is -1. In practice we rarely come across these types of correlations.

Absence of correlation: If two series of two variables exhibit no relations between them or change in variable does not lead to a change in the other variable, then we can firmly say that there is no correlation or absurd correlation between the two variables. In such a case the coefficient of correlation is 0.

Limited degrees of correlation: If two variables are not perfectly correlated or is there a perfect absence of correlation, then we term the correlation as Limited correlation. It may be positive, negative or zero but lies with the limits ± 1.
High degree, moderate degree or low degree are the three categories of this kind of correlation. The following table reveals the effect (or degree) of coefficient or correlation.





(vi). What are various steps for the calculation of rank order correlation?

Answer: Step 1: Rank both sets of data. Give the largest value rank 1, the second largest value rank 2, etc.

Step 2: Calculate the differences in the ranks, d.
Step 3: Work out the squares of the differences (d^2).
Step 4: Calculate the sum of these squared differences, $\sum d2$
Step 5: Spearman's Rank Correlation Coefficient is found by substituting this sum into the following formula:

$$1-\frac{6\times\sum d^2}{n\left(n^2-1\right)}$$

where n is how many pairs of data you have.

Example: Find rank correlation from data given below:

Team	Goals in 1992	Goals in 1993
1	125	109
2	80	76
3	96	101
4	65	77
5	30	27
6	134	142
7	54	76
8	16	12
9	64	80
10	72	93
11	49	82

Team	Goals In 1992	Goals In 1993	R_1	R_2	$R_1 - R_2$	$(R_1 - R_2)^2$
1	125	109	2	2	0	0
2	80	76	4	8.5	–4.5	20.25
3	96	101	3	3	0	0
4	65	77	6	7	–1	1
5	30	27	10	10	0	0
6	134	142	1	1	0	0
7	54	76	8	8.5	-0.5	0.25
8	16	12	11	11	0	0
9	64	80	7	6	1	1
10	72	93	5	4	1	1
11	49	82	9	5	4	16
						39.5

So, for this set of data, the finished equation looks like this:

$$\text{rank} = 1-\frac{6\sum d^2}{n\left(n^2-1\right)}$$

$$1-\frac{6\times 39.5}{11\times\left(11^2-1\right)}$$

$$1-\frac{237}{1320}=-0.17954$$

$$= 0.82046.$$

High Positive Correlation.

Activity

Question 1. Take an imaginary example applicable to geographical analysis and explain direct and indirect methods of calculating mean from ungrouped data.

Answer: From the following data of the marks obtained by 60 students of a class.

Marks	20	30	40	50	60	70
No. of students	8	12	20	10	6	4

Solution-1 (Direct Method):

Calculation of Arithmetic mean:

Marks (x)	**No. of students (f)**	fx
20	8	160
30	12	360
40	20	800
50	10	500
60	6	360
70	4	280
	N = 60	$\sum fx = 2,460$

Here N= total frequency = 60

$$\text{Mean } \bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} \quad \Rightarrow \quad \bar{x} = \frac{2460}{60} = 41$$

Hence,
Average Marks=41

Solution-2 (Short Cut Method):

Calculation of Arithmetic mean:

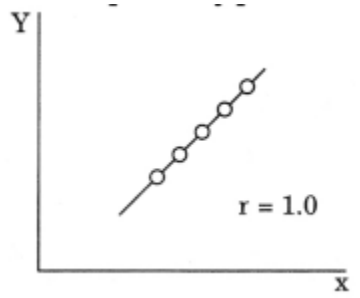
Marks (x)	No. of students (f)	fx	$d = (x - 40)$	d'	Fd'
20	8	160	−20	−2	−16
30	12	360	−10	−1	−12
40	20	800	0	0	0
50	10	500	10	1	10
60	6	360	20	2	12
70	4	280	30	3	12
	N = 60	$\sum fx = 2,460$			$\sum fd' = 60$

$$\text{Mean } \bar{x} = A + \frac{\sum_{i=1}^{n} f_i d_i'}{\sum_{i=1}^{n} f_i} \times i$$

Since $A = 40, \bar{x} = 40 + 10 \times \frac{6}{60} = 40 + 1 = 41$
Hence the average marks = 41.

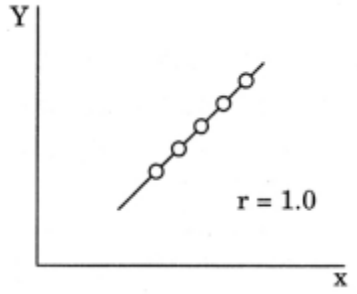
Question 2. Draw scatter plots showing different types of perfect correlations

Answer: Perfect Positive Correlation: If all points lie on a rising straight line the correlation is perfectly positive and r=+1



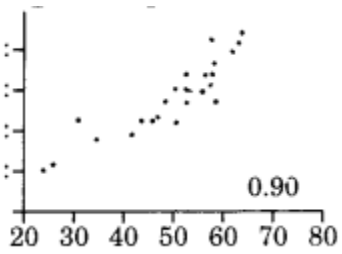
Perfect Negative Correlation: If all points lie on a falling straight line the correlation

is perfectly negative and r = -1.



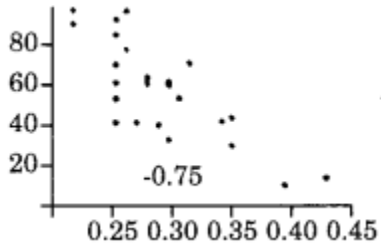
High degree of Positive Correlation:

If the points lie in a narrow strip rising upwards, the correlation is high degree of positive.



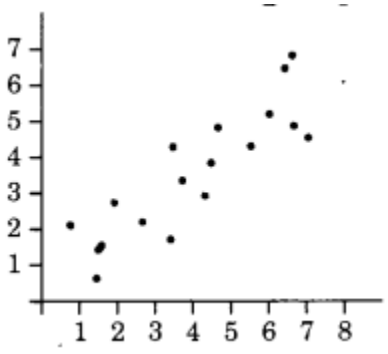
High degree of Negative Correlation:

If the points lie in a narrow strip falling downwards, the correlation is high degree of negative.



Low degree of Positive Correlation:

If the point are spread widely over a broad strip rising upwards, the correlation is low degree positive.



Low degree of Negative Correlation:

If the points are spread widely over a broad strip falling downwards, the correlation is low degree negative.