1.1 ELEMENTS OF STRUCTURED DATA AND FURTHER READING

- Structured data is the data which conforms to a data model, it has well-defined structure and it follows a consistent order and can be easily accessed and used by a person or a computer program.
- Structured data is generally stored in well-defined schemas such as Databases.
- It is generally tabular with column and rows that define its attributes.
- Data conforms to a data model and has easily identifiable structure.

1.1.1 Example : (Characteristics)

(i) Data is well-organised : Definition, Format and Meaning of Data is explicitly known.

(ii) Data is kept in fixed fields within a record or file.

(iii) Similar entities are grouped together to form relations or classes.

(iv) Entities in the same group have same attributes.

(v) Easy to access and query.

(vi) Data elements are addressable, it is efficient to analyse and process.

1.1.2 Sources of Structured Data

(i) SQL Databases

(ii) Spreadsheets such as Excel

(iii) OLTP systems

(iv) Online forms

(v) Sensors such as GPS or RFID tags.

(vi) Network and web server logs

(vii) Medical devices.

1.1.3 Advantages of Structured Data

(i) Structured data have a well-defined structure and it helps in accessing of data.

(ii) Data can be indexed based on text string as well as attributes. This makes search operation hassle-free.

(iii) Data mining is easy, i.e. knowledge can be easily extracted from data.

(iv) Due to its well structured form, operations such as updating and deleting is quite easy.

(v) Business intelligence operations such as Data ware housing can be easily undertaken.

(vi) Easily scalable if there is an increment of data.

(vii) Ensuring security to data is easy.

(viii) Structured data accounts for only about 25% of data but because of its high degree of organisation and performance make it the foundation of Big Data.

1.2 RECTANGULAR DATA

- Rectangular data are the staple of statistical and machine learning models.
- Rectangular data are **multivariate cross-sectional data** (i.e., not time-series or repeated measure).
- Here, each column is a variable (feature), and each row is a case or record.
- Multivariate cross-sectional data (i.e. not time-series or repeated measure) are indicated by rectangular data; Each column of it is a variable (feature) and each row is a record.

1.2.1 Procedure of Representation of Rectangular Data

(i) First procedure is to map rectangle data on to a higher-dimensional point data and use point-based data structure procedures such as the **grid-file**, **PR quadtree**, **point quadtree** and k-d-tree.

Procedure mapping of the rectangular data to a four-dimensional point can be represented as : x and y coordinates of one corner and the width and height.

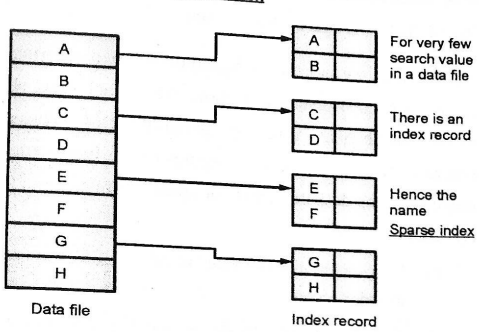

Fig. 1.3.3

Examples to be added

Data warehouses Relational database	Data takes non-relational databases.
Several predetermined formats	A huge array of formats

1.4 NON-RECTANGULAR DATA STRUCTURE

Spatial data structures, which are used in **geolocation analytics**, are very complex and they are quite different from rectangular data structures.

In the object representation, the focus of the data is an object and its spatial coordinates.

Examples of unstructured data are :

(i) Rich media. Media and entertainment data, surveillance data, geo-spatial data, audio, weather data.

(ii) Document collections, Invoices, records, emails, productivity applications.

(iii) Internet Of Things (IOT), Sensor data, ticket data.

(iv) Analytics, Machine Learning, Artificial intelligence (AI).

1.4.1 Structured V/S Unstructured Data

Structured data	Unstructured data
Structured data stands for information that is highly organised, factual and to the point	Unstructured data doesn't have any predefined structure to it and comes in all its diversity of forms
Quantitative	Qualitative

1.5 MEASURES OF CENTRAL TENDENCY

- One of the most important aspects of describing a distribution is the central value around which the observations are distributed.
- Any arithmetical measure which gives the centre or central value of a set of observations is known as a measure of central tendency or measure of location.

1.6 REVIEW OF BASIC RESULTS IN THE THEORY OF STATISTICS

(i) Arithmetic mean

If f_1, f_2, f_n are frequencies of the variates x_1, x_2, ..., x_n then,

$$M = \text{Arithmetic Mean (A.M.)} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Short-cut method of finding mean :

Let, $x' = \frac{x - x_o}{h}$; where x_o is assumed mean and h is length of the class interval

Then $M = x_o + hA$

Where, $A = \frac{\sum f \cdot x'}{\sum f}$

(ii) Standard Derivation

Let f_i (i = 1 to n) be the frequencies of x_i (i = 1 to n). Let M be the mean. If σ is standard deviation (S.D.) then,

$$\sigma^2 = \frac{\sum_{i=1}^{n} [(x_i - M)^2 \cdot f_i]}{\sum_{i=1}^{n} f_i}$$

Short-cut method to find σ :

Let, $x' = \frac{x - x_o}{h}$

Then $\sigma^2 = h^2 \left[\frac{\sum f_i x_i'^2}{\sum f_i} - \left(\frac{\sum f_i x_i'}{\sum f_i} \right)^2 \right]$

Note that, σ^2 is also called as Variance and denoted by Var.(x).

(iii) r^{th} moments about mean

Let, (x_i, f_i) be the given frequency distribution, then the r^{th} moment about mean M is given by,

$$\mu_r = \frac{\sum f_i (x_i - M)^2}{\sum f_i}$$

If r = 1, then $\mu_1 = 0$

$$\because \mu_1 = \frac{\sum f_i (x_i - M)}{\sum f_i}$$

$$= \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i M}{\sum f_i}$$

$$= \frac{\sum f_i x_i}{\sum f_i} - M \frac{\sum f_i}{\sum f_i}$$

$$= M - M \cdot 1 = 0$$

and for, r = 2

$$\mu_2 = \text{Variance} = \frac{\sum f_i (x_i - M)^2}{\sum f_i}$$

$$= \sigma^2$$

= Square of standard deviation

(iv) Definition : r^{th} moment about origin are given by,

$$\mu_r' = \frac{\sum f_i x_i^r}{\sum f_i}$$

if, r = 1, $\mu_1' = \frac{\sum f_i x_i}{\sum f_i}$ = Mean M,

Again $\mu_2' = \frac{\sum f_i x_i^2}{\sum f_i}$

$$\mu_3' = \frac{\sum f_i x_i^3}{\sum f_i}$$

$$\mu_4' = \frac{\sum f_i x_i^4}{\sum f_i}$$

(v) Relation between moments about mean (μ_r) and moments about origin (μ_r')

(i) $\mu_2 = \mu_2' - (\mu_1')^2$

(ii) $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$

(iii) $\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$

Also note that,

If $x' = \frac{x - x_0}{h}$

Then, $\mu_r' = h^r \frac{\sum f_i x_i'^r}{\sum f_i}$

(Vi) Kari pearson's coefficients of kurtosis

(i) β_1 = Measure of skewness $= \frac{\mu_3^2}{\mu_2^3}$

(ii) β_2 = Measure of flatness of single humped distribution

$$= \frac{\mu_4}{\mu_2^2}$$

Note : For normal distribution, $\beta_2 = 3$. If $\beta_2 > 3$, then distribution is peaked sharply than the normal cure and is known as lepto-kurtic.

If $\beta_2 < 3$, the distribution is flat compared to normal curve and is known as plato-kurtic.

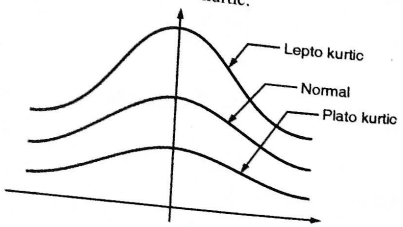

Fig. 1.6.1

(vii) The expected value of x : (mean value of x) :

If X is a random variable then the expected value of X is denoted by E (X) and means the value, on average, that X takes.

Definition : If $X = x_i$, i = 1 to n is a discrete random variable with frequencies f_i, i = 1 to n, then

$$E(X) = \sum_{i=1}^{n} x_i f(x_i)$$

Note : The expected value of X is also called as the Mean value of X and is also denoted by M.
i.e. M = E (X)

Properties

(i) The r^{th} moment about origin is also written as,

$$\mu_r' = E(x^r) = \sum_{i=1}^{n} \left(x_i^r\right) f_i$$

Clearly, $E(x) = \mu_1' = \sum_{i=1}^{n} x_i f_i$

$$E(x^2) = \mu_2' = \sum_{i=1}^{n} x_i^2 f_i$$

$$E(x^3) = \mu_3' = \sum_{i=1}^{n} x_i^3 f_i \text{ and}$$

$$E(x^4) = \mu_4' = \sum_{i=1}^{n} x_i^4 f_i \text{ and so on.}$$

(ii) Moment about the mean $\bar{x}$ are defined as,

$$\mu_r = E\left[x_i - \bar{x}\right]^r = \sum_{i=1}^{n} \left(x_i - \bar{x}\right) f_i$$

and is called as r^{th} moment about mean $\bar{x}$.

Clearly, $\mu_1 = 0$

$$\mu_2 = E\left(x_i - \bar{x}\right)^2 = \mu_2' - \mu_1'^2$$

$$\mu_3 = E\left(x_i - \bar{x}\right)^3$$

$$= \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

and $\mu_4 = E\left(x_i - \bar{x}\right)^4$

$$= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

Examples

Ex. 1.6.1 : For the following distribution, find :

(i) Arithmetic mean

(ii) Standard derivation

(iii) First 4 moments about the mean

(iv) β_1 and β_2.

x	2	2.5	3	3.5	4	4.5	5
f	5	38	65	92	70	40	10

Let $x' = \frac{x - 3.5}{0.5}$

$x_o = 3.5$

$h = 0.5$

x	f	x′	fx′	fx′²	fx′³	fx′⁴
2	5	−3	−15	45	−135	405
2.5	38	−2	−76	152	−304	608
3	65	−1	−65	65	−65	65
3.5	92	0	0	0	0	0
4	70	1	70	70	70	70
4.5	40	2	80	160	320	640
5	10	3	30	90	270	810
Total	**320**		**24**	**582**	**156**	**2598**

(i) Arithmetic mean : Using result (i) in section 10.3 we have,

$$A = \frac{\sum fx'}{\sum r} = \frac{24}{320} = 0.075$$

and arithmetic mean = $x_o + hA$

$$= 3.5 + (0.5)(0.075) = -3.538$$

(ii) Standard deviation : Using the result (ii) in section 10.3, we have

$$\sigma^2 = h^2 \left\{ \frac{\sum fx'^2}{\sum f} - \left(\frac{\sum fx'}{\sum r} \right)^2 \right\}$$

$$= (0.5)^2 \left\{ \frac{582}{320} - \left(\frac{24}{320} \right)^2 \right\} = 0.453$$

$$\sigma = 0.673$$

(iii) Moments about the mean M

When assumed mean is $A = x_o = 3.5$ and using (iii) section – 10.3, we have,

$$\mu_1' = h \frac{\sum fx'}{\sum f} = 0.5 \left(\frac{24}{320} \right) = 0.0375$$

$$\mu_2' = h^2 \frac{\sum fx'^2}{\sum f} = (0.5)^2 \left(\frac{582}{320} \right) = 0.4546$$

$$\mu_3' = h^3 \frac{\sum fx'^3}{\sum f}$$

$$= (0.5)^3 \left(\frac{156}{320} \right)$$

$$= 0.0609$$

$$\mu_4' = h^4 \frac{\sum fx'^4}{\sum f}$$

$$= (0.5)^4 \left(\frac{2598}{320} \right)$$

$$= 0.5074$$

Using result, (iv) in section 10.3, we have for moments about the mean M.

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_2'^2 = (0.4546) - (0.0375)^2 = 0.0453$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$= (0.0609) - 3(0.4546)(0.0375) + 2(0.0375)^2$$

$$= 0.0600$$

$$\mu_4 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^4$$

$$= (0.5074) - 4(0.0609)(0.0375)$$

$$+ 6(0.0375)^2 (0.4546) - 3(0.0375)^4$$

$$= 3.2385$$

(iv) By definition of β_1 and β_2 we get,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.0600)^2}{(0.0453)^3}$$

$$= 32.22$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3.2385}{(0.0453)^2}$$

$$= 1578.21$$

Since $\beta_1 > 3$, the distribution is lepto-curtic i.e. it is peaked up sharply than the normal distribution.

UEx. 1.6.2 (Ref. - Dec. 98)
From the following frequency distribution compute the standard deviation of 100 students.

Table P. 1.6.2

Mass in kg.	No. of students
60-62	5
63-65	18
66-68	42
69-71	27
72-74	8

Soln. : We construct the table,

Let $\bar{x} = 67$ be assumed mean

and $h = 3$

= class width

Let $u = \frac{x - 67}{3}$

Table P. 1.6.2(a)

Class of masses	Midpoint of Classes x	Number of students f	$u = \frac{x-67}{3}$	fu	fu^2
60-62	61	5	– 2	– 10	20
63-65	64	18	– 1	– 18	18
66-68	67	42	0	0	0
69-71	70	27	1	27	27
72-74	73	8	2	16	32
Total ($\sum$)		100	0	15	97

We have $\sum f \cdot u = 15$, $\sum fu^2 = 97$

$h = 3$ and $N = \sum f = 100$

By definition, $\sigma = h\sqrt{\frac{1}{N}\sum fu^2 - \left(\frac{1}{N}\sum f_n\right)^2}$

$= 3\sqrt{\frac{1}{100}(97) - \left(\frac{15}{100}\right)^2}$

$\therefore \sigma = 2.92$

1.7 THE MEDIAN

- The median of a set of n observations $x_1, x_2, \ldots x_n$ is the middle value when the observations are arranged in an array according to their order of magnitude.
- If n is odd, the middle value which is the $\left(\frac{n}{2}+1\right)^{th}$ in the ascending order of magnitude is unique and is the median.
- If n is even, there are two middle values and the average of three values is the median.
- For example, the median of the set 2, 3, 5, 6, 7 is 5 and that of the set – 3, – 1, 0, 1, 2, 3 is $\frac{0+1}{2} = 0.5$.
- The median is the value which divides the set of observations into two equal halves, such that 50% of the observations lie below the median and 50% above the median.

1.7.1 The Median of Grouped Data

In case of grouped data, median lies in a class interval. To find its value, we use the formula,

$$M = l_1 + \frac{l_2 - l_1}{f_1}(m - C)$$

M = Median

l_1 = The lower limit of the class in which median lies

l_2 = The upper limit of the class in which median lies

f_1 = The frequency of the class in which the median lies

m = Middle item and

C = Cumulative frequency of the group preceding the median group

Let us take an example.

Ex. 1.7.1 : Find the median of the following distribution :

Table P. 1.7.1

Class interval Rs.	Frequencies
1-3	6
3-5	53
5-7	85
7-9	56
9-11	21
11-13	16
13-15	4
15-17	4
Total	245

☑ **Soln. :**

Here N = Total frequency = 245

$\therefore \frac{N}{2} = 122.5$

We prepare the table with cumulative frequency.

Table P. 1.7.1(a)

Class-interval	Frequency	Cumulative frequency
1-3	6	6
3-5	53	59
5-7	85	144
7-9	56	200
9-11	21	221
11-13	16	237
13-15	4	241
15-17	4	245

Now, median = The value of $\frac{N}{2}$ i.e.

122.5 items, which lies in 5 – 7 group,

Applying the formula,

$$M = l_1 + \frac{l_2 - l_1}{f_1}(m - C) \text{ we get}$$

$$M = 5 + \frac{7-5}{85}(122.5 - 59) = 6.5$$

1.8 MODE

The word mode comes from the French word '*la mode*' (which means the fashion). The mode is the observation which occurs most frequently in a set.

1.8.1 Calculation of Mode

1. The computation of mode from a set of observations is simple enough. It is the observations which occurs most frequently. For example the mode of 3, 2, 3, – 1, 0, – 3, 2, 5 is 3. The mode may not exist for certain sets such as set – 1, 0, 1, 5. Also note that there may be more than one mode.
2. For grouped data mode is determined as follows :

$$M = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) h$$

Where M is mode,

L_1 = Lower boundary of the modal class.

Δ_1 = Excess of the modal class frequency over the frequency of the class to its left and

Δ_2 = Excess of the modal class frequency over the frequency of the class to its right and

h = Size of the class-intervals.

Ex. 1.8.1 : Calculate the mode for the given data :

Table P. 1.8.1

Class-interval (cm)	Frequency
145-146	2
147-148	5
149-150	8
151-152	15
153-154	9
155-156	6
157-158	4
159-160	1
Total	50

☑ **Soln. :** First we find the modal class i.e. the interval which corresponds to the maximum frequency. Here the modal class is 151-152. Class midpoint is 151.5. Its lower boundary is 150.5 and upper boundary is 152.5. Again h = 2, $\Delta_1 = 15 - 8 = 7$, $\Delta_2 = 15 - 9 = 6$.

$$\therefore M = 150.5 + \left(\frac{7}{7+6}\right) \times 2 = 150.577$$

1.9 GEOMETRIC AND HARMONIC MEAN

There are two other averages, the geometric mean and the harmonic mean which are sometimes used.

(1) The Geometric Mean (G.M.) of a set of observations is such that its logarithm is equal to the Arithmetic mean of the logarithms of the values of the observations. This

is given by the same formula even if the observations occur with certain frequencies.

Consider the set of values 2, 3, 5, 6 which occur with frequencies 10, 16, 24, 10 respectively. If x is the Geometric Mean then,

$$\log x = \frac{10 \log 2 + 16 \log 3 + 24 \log 5 + 10 \log 6}{10 + 16 + 24 + 10}$$

$$= 0.5867$$

$$\therefore \quad x = 3.86$$

(2) The harmonic mean (H.M.) of a set of observations is such that its reciprocal is the arithmetic mean (A.M.) of the reciprocals of the values of the observations. Consider the above set of values. The harmonic mean y of the set of values is given by,

$$\frac{1}{y} = \frac{10 \times \frac{1}{2} + 16 \times \frac{1}{3} + 24 \times \frac{1}{5} + 10 \times \frac{1}{6}}{60}$$

$$= 0.28$$

$$\therefore \quad y = 3.57$$

Note :

1. The Geometric mean can be found only if the values assumed by the observations are positive.
2. It can be shown that A.M. ≥ G.M. ≥ H.M.

1.9.1 Other Measures of Location, Quartiles, Deciles and Percentiles

- We have seen that the median of a set of measurements is the value which divides the set into two equal halves, each containing 50% of the measurements.
- In the same way, some other measures of location can be considered. We define the three quartiles, Q_1, Q_2 and Q_3.
- They are such that when the measurements are arranged in increasing order, they divide the set of measurements into four equal parts, the first quartile Q_1 contains the 25% of the measurement, the second quartile Q_2 contains 50% of the measurements and the third quartile Q_3 contains 75% of the measurements.
- Actually the second quartile Q_2 is the median.
- Similarly we define the deciles. The first decile D_1 contains 10% of the measurements, the second decile D_2 contains 20% of the measurements and so on.
- The fifth decile is the median.
- In the same manner, we define percentiles. The 99 percentiles P_1, ... P_{99} divide the set of measurements into 100 equal parts.
- The first percentile P_1 contains 1% of the measurements, the second percentile P_2 contains 2% of the measurements and so on, the 12^{th} percentile contains 12% of the measurements.
- The 50^{th} percentile is therefore the median. The method of finding out the quartiles, deciles and percentiles is basically the same as that of finding the median.
- The median divides the set of observations into two equal values, each containing 50% of the measurements, the 3^{rd} decile divides the set into two parts, the first part being 30% of the set and the other containing 70% of the observations.

Ex. 1.9.1 : The distribution of fortnightly wages of 280 employees of an undertaking is as given.

Find the first quartile, the median and the third quartile, find D_4, P_{66}, P_{10} and P_{90}.

Table P. 1.9.1

Fortnightly wages (Rs.)	Frequency
Less than 200	12
200-400	16
400-600	38
600-800	78
800-1000	80
1000-1200	35
1200-1400	14
Above 1400	7
Total	**280**

☑ **Soln.:**

First we prepare cumulative frequency table.

Table P. 1.9.1(a)

Wages (Rs.)	Frequency	Cumulative Frequency
Less than 200	12	12
200-400	16	28
400-600	38	66
600-800	78	144
800-1000	80	224
1000-1200	35	259
1200-1400	14	273
Above 1400	7	280

▶ **Step I :** To find quartiles

The observation for the first quartile Q_1 corresponds to $\frac{280}{4} = 70^{th}$ observations, which lies in the interval 600-800, with lower class boundary 600. This interval contains 78 observations and the interval preceding this contains 66 observations. Hence,

$$Q_1 = l_1 + \frac{l_2 - l_1}{f_t} (m - C)$$

where l_1 = Lower limit of the class in which Q_1 lies

l_2 = The upper limit of HM class in which Q_2 lies

f_1 = Positive frequency of the class

$m = \frac{N}{4}$

C = Cumulative frequency of the group preceding the Q_1 class.

$$\therefore \quad Q_1 = 600 + \frac{200}{78} \left(\frac{280}{4} - 66 \right)$$

$$= 610.25 \text{ Rs.}$$

The median which is the second quartile Q_2, is given by,

$$Q_2 = 600 + \frac{200}{78} \left(\frac{280}{2} - 66 \right)$$

$$= 600 + 189.74 = 789.74 \text{ Rs.}$$

The third quartile Q_3 is given by,

$$Q_3 = 800 + \frac{200}{80} \left[280 \times \frac{3}{4} - 144 \right]$$

$$= 965 \text{ Rs.}$$

▶ **Step II :** The observation for 4^{th} decile corresponds to the $\frac{280 \times 4}{10} = 112^{th}$ observation, which lies in the interval 600 – 800. Hence,

$$D_4 = 600 + \frac{200}{78} (112 - 66) = 717.95 \text{ Rs.}$$

▶ **Step III :** The observation for 66^{th} percentile corresponds to $280 \times \frac{66}{100} = 184^{th}$ observation which lies in the interval 800 – 1000. Thus the 66^{th} percentile P_{66} is given by,

$$P_{66} = 800 + \frac{200}{80} (184.8 - 144) = 902 \text{ Rs.}$$

In the same way,

$$P_{10} = 400 + 0 = 400$$

$$\text{and} \quad P_{90} = 1000 + \frac{200}{35} \left[\frac{280 \times 90}{100} - 220 \right]$$

$$= 1182.96 \text{ Rs.}$$

1.10 THE RANGE

The range of set of numbers is the difference between the largest and the smallest items of the set. The range is a very crude measure. It does not tell us about the distribution of the values of the set relative to the average.

1.10.1 The Semi-Interquartile Range

This is a more refined form of range. It is defined by,

$$Q = \frac{Q_3 - Q_1}{2}$$

And is called as semi-interquartile range.

Q_1 = First quartile

Q_3 = Third quartile

1.11 THE MEAN DEVIATION

Consider a set of observations, $x_1, x_2, ..., x_n$. The mean deviation (or average deviation) is defined by,

$$M.D. = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

where, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

is the arithmetic mean and $|x_i - \bar{x}|$ is the absolute value of the deviation.

Ex. 1.11.1 : Find the mean deviation of the set of measurement 1, 3, 8.

Soln. : Here the arithmetic mean,

$$\bar{x} = \frac{1+3+8}{3} = 4$$

$$\therefore M.D. = \frac{|1-4| + |3-4| + |8-4|}{3}$$

$$= 2.67$$

1.11.1 Mean Deviation for Grouped Data

Let $x_1, x_2, ..., x_n$ occur with the corresponding frequencies $f_1, f_2, ..., f_n$, then

$$M.D. = \frac{\sum_{i=1}^{n} f_i |x_i - \bar{x}|}{\sum_{i=1}^{n} f_i}$$

where, $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$

Note that the above formula is also applicable in the case of a frequency distribution whose class intervals have mid-points $x_1, x_2, ..., x_n$ and the classes have frequencies $f_1, f_2, ... f_n$.

Ex. 1.11.2 : Calculate the mean deviation from the mean of the following distribution **Table P. 1.11.2**

Marks	Number of students
0-10	5
20-20	8
20-30	15
30-40	16
40-50	6
Total	**50**

Soln. : We first calculate mean then find mean deviation.

Table P. 1.11.2(a)

Mid value	$u = \frac{x-25}{10}$	f	fu	$x - \bar{x}$	$f\|x - \bar{x}\|$
5	−2	5	−10	−22	110
15	−1	8	−8	−12	96
25	0	15	0	−2	30
35	1	16	16	8	128
45	2	6	12	18	108
Total		**50**	**10**		**472**

Here, Mean $= 25 + \frac{10}{50} \times 10 = 27$ marks and

Mean deviation $= \frac{\sum f_i |x_i - \bar{x}|}{\sum f_i}$

$= \frac{472}{50} = 9.44$ marks

Ex. 1.11.3 : The mean annual salary paid to all employees of a company was Rs. 5000. The mean annual salaries paid to male and female employees were Rs. 5200 and Rs. 4200 respectively. Determine the percentage of males and females employed by the company.

Soln. : Let n_1 and n_2 represent percentage of males and females respectively.

Then $n_1 + n_2 = 100$ ...(1)

Now, mean annual salary of males $= \bar{x}_1 = 5200$ Rs.

Mean annual salary of females $= \bar{x}_2 = 4200$ Rs.

Mean annual salary of all employees $= \bar{x} = 5000$ Rs.

Now, $\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$ (Note the formula)

$\therefore 5000 = \frac{n_1\, 5200 + n_2\, 4200}{100}$

$\therefore 26\, n_1 + 21\, n_2 = 2500$ ...(2)

From Equations (1) and (2),

$n_1 = 80, n_2 = 20.$

$\therefore$ Percentages are 80 and 20.

Ex. 1.11.4 : The first of the two samples has 100 items with mean 15 and standard deviation 3. If the whole group has 250 items with mean 15.6 and standard deviation $\sqrt{13.44}$, find the standard deviation of the second group.

Soln. : We have,

$n_1 = 100, \quad \bar{x}_1 = 15, \quad \sigma_1 = 3$

$n = n_1 + n_2 = 250$

$\bar{x} = 15.6$

$\sigma = \sqrt{13.44}$

$\therefore n_2 = 250 - 100 = 150$

We use the formula,

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

$$\therefore 15.6 = \frac{100\,(15) + 150\, \bar{x}_2}{250}$$

$$\therefore \bar{x}_2 = 16$$

The variance of the combined group σ^2 is given by the formula,

$$\sigma^2 = \frac{n_1 \sigma_1^2 + n_2 \sigma_2^2}{n_1 + n_2} + \frac{n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}$$

(Note the formula)

Where $d_1 = \bar{x}_1 - \bar{x} = 15 - 15.6 = -0.6$

and $d_2 = \bar{x}_2 - \bar{x} = 16 - 15.6 = 0.4$

$\because \sigma = \sqrt{13.44}$

$$\therefore 13.44 = \frac{100\,(9) + 150\, \sigma_2^2}{250} + \frac{100\,(-0.6)^2 + 250\,(0.4)^2}{250}$$

$$\therefore 250 \times 13.44 = 900 + 36 + 40 + 150\, \sigma_2^2$$

$$\therefore \sigma_2 = 4$$

Ex. 1.11.5 : Calculate the first four moments of the following distribution about mean and hence find β_1 and β_2.

x	0	1	2	3	4	5	6	7	8
f	1	8	28	56	70	56	28	8	1

Soln. : We first calculate the first four moments about the point x = 4.

We prepare the table.

Table P. 1.11.5

x	f	d = x − 4	fd	fd^2	fd^3	fd^4
0	1	−4	−4	16	−64	256
1	8	−3	−24	72	−216	648
2	28	−2	−56	112	−224	448
3	56	−1	−56	56	−56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256
Σ	N = 256	–	0	512	0	2816

Recall that,

$\mu_r' = \frac{1}{N}\sum f \cdot d^r$ (Get acquainted with the notation)

where $d = x - 4$

Now, $\mu_1' = \frac{1}{N}\sum fd = 0$

$\mu_2' = \frac{1}{N}\sum fd^2 = \frac{512}{256} = 2$

$\mu_3' = \frac{1}{N}\sum fd^3 = 0$

$\mu_4' = \frac{1}{N}\sum fd^4 = \frac{2816}{256} = 11$

Moments about mean are $\mu_1 = 0$

$\mu_2 = \mu_2' - \mu_1'^2 = 2 - 0 = 2$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$

$= 0 - 0 + 0 = 0$

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$

$= 11 - 0 + 0 - 0 = 11$

Now, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0}{4} = 0$

And $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{11}{4} = 2.75$

1.12 ROBUST ESTIMATES

Robust estimates are statistics with good performance for data drawn from a wide range of **probability distributions**, especially those distributions which are not normal.

Robust statistical methods are used for many common problems, such as estimating **location, scale** and regression parameters.

One motivation is to produce **statistical** methods that are not affected by outliers.

Another motivation is to provide methods with good performance when there are small departures from a **parametric distribution.**

For example, robust methods gives good result for mixture of two normal distributions with different standard deviations.

1.12.1 Robust Parametric Statistics

Robust parametric statistics can proceed in two ways.

(i) by designing estimators so that a pre-selected behaviour of the influence function is achieved.

(ii) by replacing estimators that are optimal under the assumption of a normal distribution with estimators that are optimal for other distributions.

For example, using **t-distribution** with low degrees of freedom (i.e. high kurtosis, degrees of freedom between 4 and 6) or with a mixture of two or more distribution.

1.12.2 Robust Estimates For The Problems

Robust estimates have been studied for the following problems.

(i) estimating **location parameters**

(ii) estimating **scale parameters**

(iii) estimating **regression coefficients**

(iv) estimation of **model-states** in models expressed in **state-space** form.

Example :

(1) The mean Is not a robust measure of central tendency. If the given data set is the values {1, 3, 5, 8, 9}, then if we add another data point with the value – 2000 or + 2000 to the data, the resulting mean will be different to the mean of the original data.

Similarly, if we replace one of the values with a data point of valve – 2000 or + 2000 then the resulting mean will be very different to the mean of the original data.

(2) The median is a robust measure of central tendency. Considering the same dataset {1, 3, 5, 8, 9}, if we add another datapoint of value – 2000 or + 2000, then the median will change slightly, but it will be similar to the median of the original data.

Also if we replace one of the values with a datapoint of valve – 2000 or + 2000, then the resulting median will still be similar to the median of the original data.

(3) The median **absolute deviation** and **interquartile range** are **robust measures** of statistical dispersion, but the **standard deviation** and **range** are not.

(4) **Trimmed estimators** and **Winsorised estimators** are general methods to make statistics more robust.

L-estimators are a general class of simple statistics, but it is robust. And **M-estimators** are a general class of robust statistics. In this case solution is preferred even though it involves lot of calculations.

1.12.3 Measures of Robustness

The main tools used to describe robustness are : (i) The **breakdown point**, (ii) **The influence function** and (iii) The **sensitivity curve**.

(i) The breakdown point

- The breakdown point of an **estimator** is the proportion of incorrect observations (e.g. arbitrarily large observations).
- An estimator can handle before giving an incorrect (e.g. arbitrarily large) result. Usually the asymptotic (infinite sample) limit is quoted as the breakdown point. But the finite-sample breakdown point may be more useful.

  For example, consider n independent random variables $(X_1, X_2, \ldots, X_n)$ and the corresponding realisations $x_1, x_2, \ldots, x_n$, then we have $\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$, to estimate mean. Such as estimator has a breakdown point of 0 (or finite sample breakdown point of $\frac{1}{n}$). It is because we can make $\bar{x}$ arbitrarily large by changing any of $x_1, x_2, \ldots, x_n$.
- The higher the breakdown point of an estimator, the more robust it is.
- A breakdown point cannot exceed 50% because if more than half the of the observations are contaminated, it is not possible to distinguish between the underlying distribution. Therefore, the maximum breakdown point of 0.5. There are estimators which achieve such a breakdown point.
- Statistics with high breakdown points are called as **resistant statistics.**

(ii) Influence Function

- The influence function is also termed as the empirical influence function. It is a measure of the dependence of the estimator on the value of any one of the points in the sample.
- It relies on calculating the estimator again with a different sample. Hence it is called as a **model-free measure**.
- In mathematical terms, on influence function is defined as a vector in the space of the estimator, which is in turn defined for a sample which is a subset of the population.

1. (Ω, A, p) is a probability space.
2. (X, Σ) is a measurable **space** (state space)
3. θ is a parameter space of dimension. $P \in N^*$.
4. $(\Gamma, S) = (R, B)$

The empirical influence function is defined as follows :

Let $n \in N^*$ and

$X_1, \ldots, X_n : (\Omega, A) \to (X, \Sigma)$ are i.i.d.

And $(x_1, \ldots, x_n)$ is a sample from these variables.

$T_n : (X^n, \Sigma^n) \to (\Gamma, S)$ is an estimator. Let $i \in (1, \ldots, n)$. The empirical influence function (E I F_i) at observation i is defined by

$$E\,I\,F_i : x \in X \to n.\,(T_n(x_1, \ldots, x_{i-1}, x, x_{i+1}))$$

Here, we are replacing the i-th value in the sample by an arbitrary value and looking at the output of the estimator.

Thus EIF is defined as the effect, scaled by (n + 1) instead of n, on the estimator of adding the point x to the sample.

(iii) Sensitivity Curve

- Instead of relying on data, we can use the distribution of the random variables.

- Here we are going to study, what happens to an estimator when we change the distribution of the data slightly :- it assumes a **distribution**, and measures sensitivity to change in this distribution.
- Here, the empirical influence assumes a **sample** set, and measures sensitivity to change in the samples.

1.13 ESTIMATES BASED ON PERCENTILES

- In a given dataset, the P^{th} percentile is a value Such that at least P percent of the values or less than or at least (100 – P) percent of the values are to be Considered.

  For example, to find 70^{th} percentile, we sort the given data. We begin with the smallest data and then proceed 70 percent of the way to the largest value.

  We note that the **median** is the same thing as the 50^{th} percentile.
- A common measure of variability is the difference between 25^{th} percentile and 75^{th} percentile. It is termed as Interquartile Range (or IQR).
- Consider an example { 5, 3, 9, 7, 2, 1, 3, 6}. We sort the elements and get {1, 2, 3, 3, 5, 6, 7, 9} Now, the 25^{th} percentile is 2.5 and 75^{th} percentile is 6.5

  $\therefore$ interquartile range = 6.5 – 2.5 = 4.

1.13.1 Percentile for Even Data

- If we have data containing even number of elements (neven).

Here, we take any value between the order statistics x_i and $x_i + 1$, where is satisfies :

$$100 \cdot \frac{i}{n} \leq p \leq 100\left(\frac{i+1}{n}\right)$$

Also, the percentile for the weighted average :

Percentile (P) = $(1 - w)\, x_i + w x_{(i+1)}$ For some weight w between o and 1.

1.13.2 Exploring the Data Distribution

The estimates that we have covered so far indicate that the data can be expressed in a single number to describe its location or variability. It also tells us the nature of its distribution.

We mention below the main terms used for exploring the distribution.

(i) **Boxplot :** Tukey has introduce Boxplat. It visualises the distribution of data quickly.

(ii) **Frequency Table :** It is the count of numeric data values in a set of intervals (bins).

(iii) **Histogram :** A plot of the frequency table with the given intervals on x-axis and the corresponding count on y-axis.

(iv) **Density Plot :** It is a smoothed version of the histogram.

1.13.3 Percentiles and Boxplots

- Another way of describeing data is with a box plot.
- To construct the boxplot, we proceed as follows:

  (i) Draw a rectangular box whose bottom is the lower quartile (ie.25^{th} percentile) and whose top is the upper quartile (75^{th} percentile)

  (ii) Draw a horizontal line segment inside the box to represent the median.

  (iii) Extend horizontal line segments from each end of the box out to the most extreme observations.

Box plots can be either vertically drawn or horizontally drawn.

Remark

If you score in the 75^{th} percentile, then 75% of population score lower than you.

Ex. 1.13.1 : Suppose the test score where : 22, 34, 68, 75, 79, 79, 81, 83, 84, 87, 90, 92, 96 and 99. If your score was 75, in what percentile did you score?

Soln. :

There were 14 scores reported and there were 4 scores at or below yours.

Now, $\frac{4}{14} \times 100\% = 29$

$\therefore$ You scored in the 29^{th} percentile.

Ex. 1.13.2 : Find Q_1, Q_2 and Q_3 for the following data set, and draw a box-and-whisker plot {2, 6, 7, 8, 8, 11, 12, 13, 14, 15, 22, 23}.

Soln. :

Step I : There are 12 data points arranged: The middle two are 11 and 12 so the median i.e. Q_2 is 11.5.

The 'lower half' of the data set is the set {2, 6, 7, 8, 8, 11} The median of this lower half is 7.5; i.e. $Q_1 = 7.5$.

The 'upper half' of the data set is the set { 12, 13, 14, 15, 22, 23}. The median here is 14.5; which is Q_3 $\therefore Q_3 = 14.5$.

Step (II) : A box and whisker plot displays the values Q_1, Q_2 and Q_3 along with the extreme values of the data set {2 and 23, in this problem}.

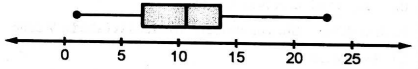

Fig. P.1.13.2

A box and whisker plot shows a "box" with left edge at Q_1, right edge at Q_3, the 'middle' of the box at Q_2 (the median)

The maximum and minimum as 'whiskers'.

We observe that the plot divides the data into 4 equal parts.

The left whisker represent the bottom 25% of the data, the left half of the box represent the second 25% the right half of the box represent the third 25% and the right whisker represent the top 25%.

1.13.4 Outliers

If a data value is very far away from the quartiles (i e. either much less than Q_1 or much greater than Q_3), it is called as an **outlier.**

Instead of being shown using the whiskers of the box-and whisker plot, outliers are generally shown as separately plotted points.

The standard definition for an outlier is a number which is less than Q_1 or greater than Q_3 by more than 1.5 times the **interquartile range** (i.e., $IQR = Q_3 - Q_1$).

Thus an outlier is any number

(i) less than $Q_1 - (1.5 \times IQR)$ or

(ii) greater than $Q_3 + (1.5 \times IQR)$.

Ex. 1.13.3 : Find Q_1, Q_2 and Q_3 for the following data set. Identify any outliers and draw a box and whisker plot.

{ 5, 40, 42, 46, 48, 49, 50, 50, 52, 53, 55, 56, 58, 75, 102}

Soln. :

Step (I) : There are totally 15 values of arranged in increasing order.

Hence, Q_2 is the 8th data point.

$\therefore Q_2 = 50$

Q_1 is 4^{th} data point, $\therefore Q_1 = 46$ and

Q_3 is 12^{th} data point, $Q_3 = 56$.

Now, the interquartile range $IQR = Q_3 - Q_1$

$= 56 - 46$

$= 10$

Step (II) : To find, if any, the values less than $Q_1 - (1.5 \times IQR)$. or greater than $Q_3 + (1.5 \times IQR)$.

Now, $Q_1 - (1.5 \times IQR)$.

$= 46 - (1.5 \times 10)$

$= 46 - 15$

$= 31$

and $Q_3 + (1.5 \times IQR)$

$= 56 + 15 = 71$

Since 5 is less than 31 and 75 and 102 are greater than 71, hence there are 3 outliers.

we exhibit the box and whisker plot

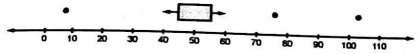

Fig. P. 1.13.3

1.13.5 Density Plots and Estimates

A density plot is a representation of the distribution of a numeric variable. It uses a kernel density estimate to show the probability density function of the variable.

It is the smoothed version of the histogram and is used in the same concept.

Interpretation of Density Curve

(1) If a density curve is a left skewed, them the mean is less than the median

(2) If a density curve is a right skewed, then the mean is greater than the median.

(3) If a density curve has no skew, then the mean is equal to the median.

(4) The peaks of density **plot help display where values are concentrated over the interval.**

An advantages of density plot over Histograms is that they are better at determining the distribution shape because they are not affected by the number of bins.

(5) Difference between density plot and histogram is as follows : Histogram provides the **visual** representation of data distribution. By using the histogram we can represent a large amount of data, and its frequency.

Density plot is the continuous and smoothened version of the histogram estimated from the data. It is estimated through **Kernel Density Estimation.**

(6) Data density in terms of sampling is how many items of particular information set are examined. Data density affect accuracy. Higher densities mean more accuracy.

(7) The y-axis in a density plot is the probability density function for the kernel density estimation. But we have to be careful to specify that it is a probability density and not a probability. The probability density is the probability per unit on x-axis.

1.14 INTRODUCTION TO MEASURES OF CENTRAL TENDENCY

- Statistics is the science that studies the collection and interpretation of numerical data.
- An example of an investigation is say, survey of Indian Births, 1990. The aim was to improve the survival rate and the care of the babies soon after birth by collecting the data on new born babies. The aim is to gain information about, a population, namely 'all Indian births'.

Definitions

A population is the collection of items under discussion. It may be finite, as in this example, or infinite, it may be real or hypothetical.

A large amount of information was collected about each baby, its mother and the circumstances of the birth. Data recorded included the birth-weight and sex of the baby. We call these classified records as *Attributes.*

Each of the quantities or attributes recorded is called a Variate.

Definition (1)

A Variate is any quantity or attribute whose value varies from one unit of investigation to another.

Definition (2)

An observation is the value taken by a variate for a particular unit of investigation.

Variates differ in nature, and the methods of analysis of a variate depend on its nature.

And we can distinguish between quantitative variates (like the birth-weight of the baby, etc.) and qualitative variates (such as the sex of the baby etc.)

Definition (3)

A quantitative variate is a variate where values are numerical.

Definition (4)

A qualitative variate or attribute is a variate whose values are not numerical.

Qualitative variates can also be divided into two types :

(i) They may be continuous, if they can take any value we specify in some range or

(ii) Discrete if their values change by steps or jumps.

Definition (5)

A continuous variate is a variate which may take all values within a given range.

Definition (6)

A discrete variate is a variate whose values change by steps.

The choice of which variates to record is important in any investigation. Once choice is made, the information can be summarized by the frequency distribution of the possible 'values'.

Definition (7)

The frequency distribution of a (discrete) variate is the set of possible values of the variate, together with the associated frequencies.

Definition (8)

The frequency distribution of a (continuous) variate is the set of class-intervals for the variate, together with the associated class-frequencies.

If we classify the whole population according to birth-weights, then instead of looking at the frequency of each variate, we first group the values into intervals, which is the sub-division of the total range of possible values of the variate.

In this example, the variate may be classified as 1-500, 500-1000, 4501-5000, 5001-5500 grams.

Definition (9)

A class-interval is a sub-division of the total range of values which a (continuous) variate may take.

Definition (10)

The class-frequency is the number of observations of the variate which fall in a given interval.

Definition (11)

Cumulative frequency distribution is the sum of all observations which are less than the upper boundary of a given class interval : or this number is the sum of the frequencies upto and including that class to which the upper class boundary corresponds.

- For example, consider the heights of 50 students. We prepare the Table 1.14.1.

Table 1.14.1 : Cumulative frequency (more than) Table

Class (cm) interval	Frequency	Cumulative frequency more than
145-146	2	2
147-148	5	7
149-150	8	15
151-152	15	30
153-154	9	39
155-156	6	45
157-158	4	49
159-160	1	50
Total	**50**	

Table 1.14.2 : Cumulative frequency (less than) Table

Class (cm) interval	Frequency	Cumulative frequency more than
145-146	2	50
147-148	5	48
149-150	8	43
151-152	15	35
153-154	9	20
155-156	6	11
157-158	4	5
159-160	1	1
Total	**50**	

Definition (12)

Points to note while constructing the Tables,

1. Make the table self-explanatory provide a title, a brief description of a source of the data, State in what units the figures are expressed, label rows and columns where appropriate.
2. Keep the table as simple as possible.
3. Distinguish between zero values and missing observations.
4. Make alternations clearly.
5. Give the calculations a logical pattern on the sheet.

1.15 HISTOGRAM

- Histogram is commonly used device for charting **continuous** frequency distribution.
- It consists in erecting a series of adjacent vertical rectangles on the section of the horizontal axis (X - axis), with bases (sections) equal to the width of the corresponding class intervals and heights are so taken that the areas of rectangles are equals to the frequencies of the corresponding classes.

1.15.1 Construction of Histogram

The variate values are taken along X-axis and the frequencies along Y-axis.

Case (I) Histogram with equal classes

- If classes are of equal magnitude, each class interval is drawn on X-axis by a section which is equal to the magnitude of the class interval.
- On each class interval erect a rectangle with the height proportional to the corresponding frequency of the class.
- The series of adjacent rectangles (one for each class) so formed gives the histogram of the frequency distribution and its area represents the total frequency of the distribution.

Case (ii) : Histogram with unequal classes

- If the classes are not uniform, then the different classes are represented on X-axis by sections which are equal to the magnitude of the corresponding classes and the heights of the corresponding rectangles are to be adjusted so that the area of the rectangle is equal to the frequency of the corresponding class.
- This can be done by taking the height of each rectangle equal to the corresponding frequency density of each class, where.

$$\text{Frequency density of a class} = \frac{\text{Frequency of the class}}{\text{Magnitude of the class}}$$

1.15.2 Example

Ex. 1.15.1 : Represent the adjoining distribution of marks of 100 students in the examination by a histogram.

Marks	Obtained	Number of students
Less than	10	4
Less than	20	6
Less than	30	24
Less than	40	46
Less than	50	67
Less than	60	86
Less than	70	96
Less than	80	99
Less than	90	100

Soln. :

First we convert the given cumulative frequency distribution into the frequency distribution of marks :

Marks	Number of students
0-10	4
10-20	6 – 4 = 2
20-30	24 – 6 = 18
30-40	46 – 24 = 22
40-50	67 – 46 = 21
50-60	86 – 67 = 19
60-70	96 – 86 = 10
70-80	99 – 96 = 3
80-90	100 – 99 = 1

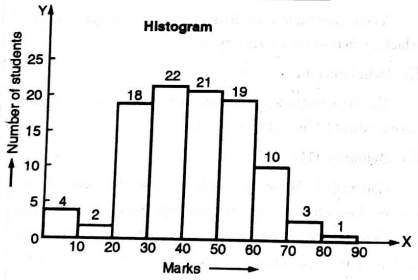

Fig. P. 1.15.1

Ex. 1.15.2 : Represent the following data by means of a histogram :

Weekly wages (100 Rs.)	10-15	15-20	20-25	25-30	30-40	40-60	60-80
Number of workers	7	19	27	15	12	12	8

Soln. :

- Here the class-intervals are of unequal magnitude, the corresponding frequencies have to be adjusted to obtain the 'frequency-density', so that area of the rectangle is equal to the class frequency.
- We note that the first four classes are of magnitude 5, the class 30-40 is of magnitude 10 and the last two classes 40-60 and 60-80 are of magnitude 20.
- Since 5 is the minimum class interval, the frequency of the class 30-40 is divided by 2 and the frequencies of classes 40-60 and 60-80 are to be divided by 4 as shown :

Weekly wages (100Rs.)	Number of workers	Magnitude of class	Height of rectangle
10-15	7	5	7
15-20	19	5	19
20-25	27	5	27
25-30	15	5	15
30-40	12	10	$\frac{12}{2} = 6$
40-60	12	20	$\frac{12}{4} = 3$
60-40	8	20	$\frac{8}{4} = 2$

Histogram

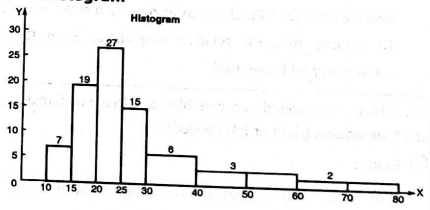

Fig. P. 1.15.2

1.15.3 Exercise

- Describe briefly the construction of histogram of a frequency distribution.

Prepare a histogram from the following data :

Class :	0-6	6-12	12-18	18-24	24-30	30-36
f :	4	8	15	20	12	6

1.16 PIE-DIAGRAM

Just as rectangles are used to represent the total magnitude and its various components, the circle may be divided into various sections or segments representing certain proportion or percentage of the various component parts to the total. Such a sub-divided circle diagram is known as **an angular or pie diagram.**

1.16.1 Steps for construction of Pie-Diagram

(i) Express each of the component values as a percentage of the respective total.

(ii) Since the angle at the centre of the circle is 360°, the total magnitude of the various components is taken to be equal to 360°.

The degrees represented by the various component parts of a given magnitude can be obtained as follows :

Degree of any component part

$$= \frac{\text{component value}}{\text{Total value}} \times 360°$$

(iii) Pie-diagram is also known as **circular diagram**

1.16.2 Example

Draw a pie-diagram to represent the following data of proposed expenditure by a state-govt. for the year 1947-98.

Items	Agriculture and rural development	Industries and urban development	Health and education	Miscellaneous
Proposed expenditure (in million Rs.)	4200	1,500	1000	500

☑ **Soln. :**

☞ **Calculation of Pie-chart**

Items	Proposed expenditure	Angle at the centre
(1)	(2)	(3) $\frac{(2)}{7200} \times 360°$
Agriculture and Rural development	4,200	$\frac{4200}{7200} \times 360° = 210°$
Industries and urban development	1500	$\frac{1500}{7200} \times 360° = 75°$
Health and education	1000	$\frac{1000}{7200} \times 360° = 50°$
Miscellaneous	500	$\frac{500}{7200} \times 360° = 25°$
Total	7200	360°

☞ **PIE-diagram representing proposed expenditure**

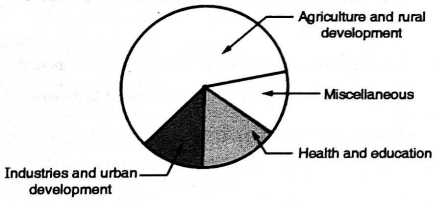

Fig. 1.16.1

1.16.3 Exploring Binary and Categorical Data

- Categorical variables take on a fixed and limited number of possible values. For example grades, gender blood group type etc.
- Also, in the case of categorical variables, the logical order is not the same as categorical data, e.g. 'one', 'two', 'three'.
- But the sorting of these variables uses logical order for example, gender is a categorical variable and has categories male and female. And there is no ordering to assign categories

1.16.4 Two or more Variables

- We have seen that mean, variance, standard deviation deal with one single variable (called as univariate analysis).
- Correlation analysis compares two variables and is called as **bivariate analysis.**

  Now we consider an additional estimate and plots and at more than two variables and it is called as **multivariate analysis.**
- To study multivariate analysis, we come across following terms:

(1) **Contingency table :** In statistics, a contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables.

They are used in survey research, business intelligence, engineering and scientific research.

A crucial problem of **multivariate statistics** is finding the direct-dependence structure of the variables contained in the contingency table.

If some of the **conditional** independence are revealed, then the storage of the data can be done in a better way. To achieve this, the relative frequencies from the contingency table are used.

Ex. 1.16.1 : Considered two variables, **sex** (male or female) and **handedness** (right or left handed).

☑ **Soln. :**

We select 100 individuals randomly from a large population to study of sex differences in handedness.

We from the table (contingency table).

Handedness / Sex	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

The number of the males, females, and right - and left - handed individuals are called marginal totals.

We observe that the population of men who are right-handed is about the same as the proportion of women who are right-handed.

The above table is 2 × 2 contingency table.

There may be more than two variables but higher-order contingency table are difficult to represent.

1.16.5 Hexagonal Binning

Hexagonal binning is a plot of two numeric variables with the records binned into hexagons.

Hexagonal binning is another way to manage the problem of having too many points that start to overlap.

Hexagonal binning plot density, rather than points.

Points are binning into gridded hexagons with distribution (the number of points per hexagon) is displayed using either colour or the area of the hexagons.

Ex. 6.16.2 : Convert binary number 1101010 into hexadecimal number.

☑ **Soln. :**

(i) Converting the given number into decimal number.

$$(1101010)_2 = 1\times2^6+1\times2^5+0\times2^4+1\times2^3+0\times2^2+1\times2^1+0\times2^0$$
$$= 64+32+0+8+0+2+0$$
$$= 106$$

(ii) Hexadecimal number :

$$(106)_{10} = 6\times16^1+10\times16^0$$
$$= (6A)_{16}$$

1.16.6 Contour Plot

- A contour plot is a curve along which the function of two variables, has a constant value.
- It is a plan section of the three dimensional graph of the function f (x, y) parallel to the x , y plane.
- A Contour line joints points of equal elevation (height) above a given level.
- The contour interval of a contour map is the difference in elevation between successive Contour lines.
- The most common form is the rectangular counter plot, which is shaped like a rectangle.

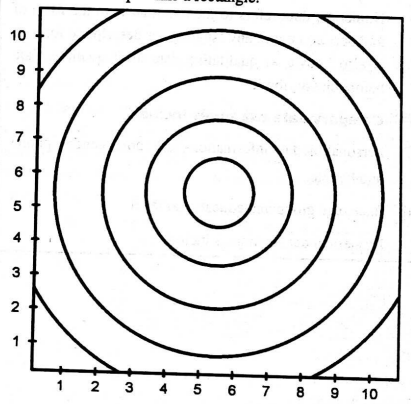

Fig. : Ex. 6.16.2

Rectangular contract plot showing symmetrical surface with a Central peak.

1.16.7 Two Category Variables

- If two category variables are independent, then the value of the variable does not change the probability distribution of the other.
- If two category variables are related, then the distribution of one depends on the other.

1.16.8 Chi-Square Test of Independence

- This taste is used to determine if category variables are independent or if they are in fact related to one another.
- This test measures the difference in the observed conditioner distribution of one variable across levels of the other and compares it to the distribution of that variable.

1.16.9 Categorical and numerical data

- Category data refers to a data type that can be stored and identified based on the names or labels given to them.
- Numerical data refers to the data that is in the form of numbers and not in any language or descriptive form. It is also known as qualitative data as it qualifies data before classifying it.

Category data examples include

(i) Personal biodata information- full name, gender, phone number etc.

(ii) Race, age group and educational level.

Numerical data examples include

(i) (GPA, calculator, interval sale, It is always collected in number form, although there are other type of data that can appear in number form.

(ii) The number of people that attended the movie theatre over the course of a month.

1.16.10 Multiple Variables

- It is possible to have experiments in which there can be multiple variables.
- There may be more than one dependent variable and/or independent variable.
- This is especially the case if one is conducting an experiment with multiple stages or sets of procedures.
- A function is called multivariable if its input is made up of multiple numbers.
- If the output of a function consists of multiple numbers, it can also be called multivariable, but these ones are also commonly called vector-valued functions.

Chapter Ends...

□□□

Module 2

CHAPTER 2 Data and Sampling Distributions

Syllabus :

2.1 Random Sampling and Sample Bias, Bias, Random Selection, Size Versus Quality, Sample Mean Versus Population Mean, Selection Bias, Regression to the Mean, Sampling Distribution of a Statistic, Central Limit Theorem, Standard Error, The Bootstrap, Resampling Versus Bootstrapping.

2.2 Confidence Intervals, Normal Distribution, Standard Normal and QQ-Plots, Long-Tailed Distributions, Student's t-Distribution, Binomial Distribution, Chi-Square Distribution, F-Distribution, Poisson and Related Distributions, Poisson Distributions, Exponential Distribution, Estimating the Failure Rate, Weibull Distribution.