Session 6 - 7

Descriptive Statistics - Data Summarization

Prof. Jigar M. Shah

Session 6 - 7

Descriptive Statistics - Data Summarization

- Summary Statistics
- Central Tendency
  - Meaning
  - Measures of Central Tendency
- Dispersion
  - Meaning
  - Measures of Dispersion
- Skewness
  - Meaning
  - Measures of Skewness
- Kurtosis
  - Meaning
  - Measures of Kurtosis

Descriptive Statistics - Data Summarization

Summary Statistics

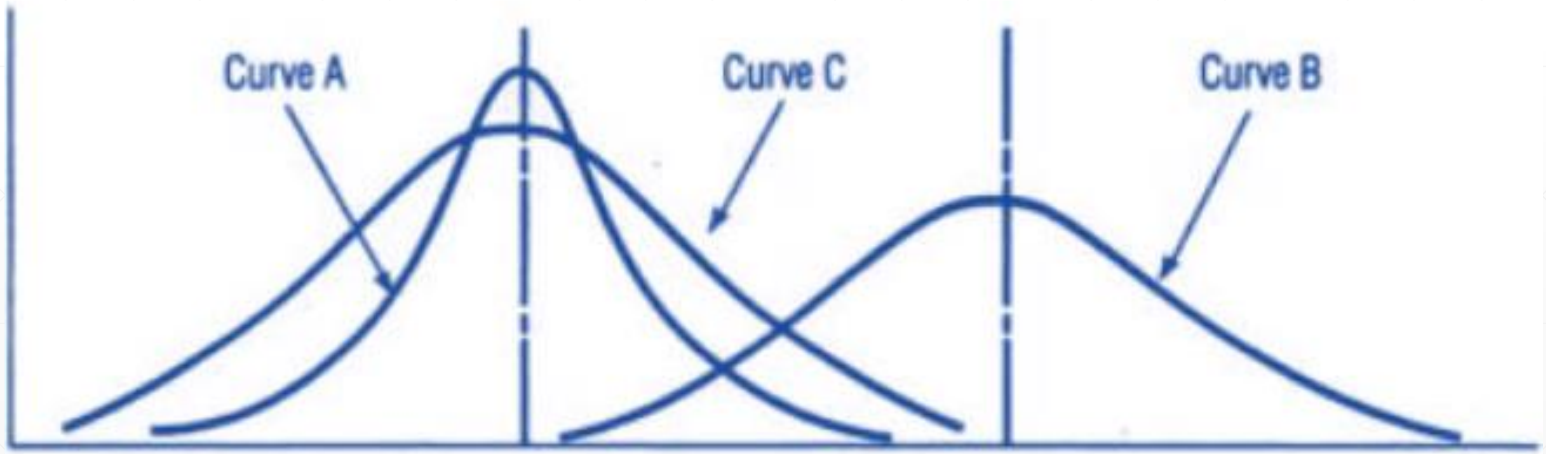
- Tabular summaries & graphical visualizations of data describe the data by illustrating the trends & patterns in the data
- Summary Statistics - numeric descriptors that provide more exact description of data by way of single numbers to describe the characteristics of the data
- Important characteristics of data of interest include:
  - Central Tendency
  - Dispersion
  - Skewness
  - Kurtosis

Provide numeric measures of location, dispersion & shape of distribution (compared to trends & patterns that are provided by frequency distributions)

Descriptive Statistics - Data Summarization

Central Tendency

- Meaning
  - Middle point of a distribution
  - Characteristic that describes the location of the central portion of the distribution



Comparison of central location of three curves

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Objectives of an Ideal Measure of Central Tendency	Requisites of an Ideal Measure of Central Tendency
• To condense data in a single value • To facilitate comparisons between data sets	• It should be rigidly defined • It should be readily comprehensible and easy to calculate • It should be based on all the observations • It should be suitable for further mathematical treatment • It should have sampling stability • It should not be affected much by extreme values

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Some Measures of Central Tendency		
• Mean	• Arithmetic Mean	• Simple Arithmetic Mean
		• Weighted Arithmetic Mean
	• Geometric Mean	
	• Harmonic Mean	• (Simple) Harmonic Mean
		• Weighted Harmonic Mean
• Median		
• Mode		

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Arithmetic Mean - Simple Arithmetic Mean	
For Raw (Ungrouped) Data	Simple arithmetic mean for a set of observations is their sum divided by the number of observations For a sample of n observations, $x_1, x_2, x_3, \ldots, x_n$, the simple arithmetic mean $\bar{x}$ is given by $\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ where i = 1, 2, 3, …, n
For Grouped Data with Single-Value Groups	In case of grouped data with single-value groups having frequency distribution $x_i \mid f_i$, i = 1, 2, 3, …, m, where f_i is the frequency of the variable x_i m is the no. of distinct values of the variable x, its simple arithmetic mean is given by $\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \ldots + f_mx_m}{f_1 + f_2 + f_3 + \ldots + f_m} = \frac{\sum_{i=1}^{m} f_ix_i}{\sum_{i=1}^{m} f_i} = \frac{1}{n}\sum_{i=1}^{m} f_ix_i$ since $\sum_{i=1}^{m} f_i$ = n, the total no. of observations

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Arithmetic Mean - Simple Arithmetic Mean

For Grouped Data with Class Intervals

In case of grouped data with class intervals, its arithmetic mean is computed in the same manner as that for single value groups, with the exception that the value of x_i is taken as the mid-point or the class-mark of the corresponding class interval

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \ldots + f_mx_m}{f_1 + f_2 + f_3 + \ldots + f_m} = \frac{\sum_{i=1}^{m} f_ix_i}{\sum_{i=1}^{m} f_i} = \frac{1}{n}\sum_{i=1}^{m} f_ix_i$$

since $\sum_{i=1}^{m} f_i = n$, the total no. of observations

where m is the no. of distinct values of the variable x,

x_i is the mid-point or class mark of the i^{th} class with i = 1, 2, 3, …, m

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Arithmetic Mean - Simple Arithmetic Mean

Property 1

Algebraic sum of the deviations of a set of values from their mean is zero

If $x_i \mid f_i$, i = 1, 2, 3, …, m, is the frequency distribution, then

$\sum_{i=1}^{m} f_i(x_i - \bar{x}) = 0$ where m is the no. of distinct values of the variable x,

f_i is the frequency of the variable x_i, i = 1, 2, 3, …, m

$\bar{x}$ is the arithmetic mean of the distribution

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Arithmetic Mean - Simple Arithmetic Mean

Property 2

The sum of the squares of the deviations of a set of values is minimum when taken about mean

If $x_i \mid f_i$, i = 1, 2, 3, ..., m, is the frequency distribution, and $z = \sum_{i=1}^{m} f_i (x_i - A)^2$, then

z is minimum when $A = \bar{x}$

where

m is the no. of distinct values of the variable x,

f_i is the frequency of the variable x_i, i = 1, 2, 3, ..., m

$\bar{x}$ is the arithmetic mean of the distribution

A is any arbitrary value

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Arithmetic Mean - Simple Arithmetic Mean

Property 3 — Mean of the combined series

If $\overline{x_i}$, i = 1, 2, 3, ..., k are the means of k series of sizes n_i, i = 1, 2, 3, ..., k, respectively, then the mean $\bar{x}$ of the combined series obtained by combining each series is given by

$$\bar{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2} + n_3\overline{x_3} + \ldots + n_k\overline{x_k}}{n_1 + n_2 + n_3 + \ldots + n_m} = \frac{\sum_{i=1}^{k} n_i\overline{x_i}}{\sum_{i=1}^{k} n_i}$$

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency
  - Arithmetic Mean
    - Simple arithmetic mean gives equal weightage to all observations (weight $\frac{1}{n}$ if there are n observations), but in cases where some observations may be more important than others, for the average to be representative of the distribution, proper weights must be assigned to observations based on their relative importance in the distribution

Arithmetic Mean - Weighted Arithmetic Mean	
Weighted Arithmetic Mean	Arithmetic mean computed by assigning different weights to different observations For a sample of n observations, $x_1, x_2, x_3, \ldots, x_n$, with weights $w_1, w_2, w_3, \ldots, w_n$, respectively, weighted arithmetic mean $\bar{x}$ is given by $\bar{x} = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n}{w_1 + w_2 + w_3 + \ldots + w_n} = \frac{\sum_{i=1}^{n} w_ix_i}{\sum_{i=1}^{n} w_i}$ where i = 1, 2, 3, …, n

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Arithmetic Mean

Merits	Demerits
• It is rigidly defined • It is easy to understand & easy to calculate • It is based upon all observations • It is amenable to algebraic treatment • Of all the averages, arithmetic mean is affected least by fluctuations of sampling i.e., it has sampling stability	• It cannot be determined by inspection nor can it be located graphically • It cannot be used with qualitative characteristics which cannot be measured quantitatively • It cannot be obtained even if a single observation is missing or illegible • It is affected very much by extreme values • It may lead to wrong conclusions if details of the data from which it is computed are not given • It cannot be calculated if the extreme classes are open • It cannot be calculated even if a single observation is missing • It is usually not a suitable measure of location in extremely asymmetrical (skewed) distribution

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency
  - Arithmetic Mean

Graduate	Monthlty Starting Salary (INR)
1	63500
2	79000
3	75000
4	59500
5	75500
6	73500
7	74500
8	66000
9	54500
10	72500
11	77500
12	66000

No. of TVs	No. of Households
0	1
1	16
2	14
3	12
4	3
5	2
6	2

Weight (kg)	No. of Males
55 - 65	4
65 - 75	11
75 - 85	15
85 - 95	4
95 - 105	2
105 - 115	0
115 - 125	0
125 - 135	1
135 - 145	0

Month	Average Daily Metro Ridership
Jan'18	45892
Feb'18	43085
Mar'18	48150
Apr'18	42495
May'18	40867
Jun'18	43095
Jul'18	49174
Aug'18	50782
Sep'18	47295
Oct'18	44986
Nov'18	46871
Dec'18	48159

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency
  - Arithmetic Mean
    - For the following sample of five purchases of raw material over the past three months, find the average cost of raw material per kg.

Purchase	Cost per kg. (Rs.)	Quantity (kg.)
1	300	1200
2	340	500
3	280	2750
4	290	1000
5	325	800

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Geometric Mean	
Geometric mean of a series of n observations is the n[th] root of their product	
For Raw (Ungrouped) Data	For a sample of n observations, $x_1, x_2, x_3, \ldots, x_n$, the geometric mean G is given by $G = \sqrt[n]{x_1 \times x_2 \times x_3 \times \ldots \times x_n} = (x_1 \times x_2 \times x_3 \times \ldots \times x_n)^{\frac{1}{n}}$ where i = 1, 2, 3, …, n $\log G = \frac{1}{n}(\log x_1 + \log x_2 + \log x_3 + \ldots + \log x_n) = \frac{1}{n}\sum_{i=1}^{n} \log x_i$ $G = \text{antilog}\left(\frac{1}{n}\sum_{i=1}^{n} \log x_i\right)$

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Geometric Mean

For Grouped Data with Single-Value Groups

In case of grouped data with single-value groups having frequency distribution $x_i \mid f_i$, i = 1, 2, 3, ..., m,

where f_i is the frequency of the variable x_i m is the no. of distinct values of the variable x,

its geometric mean is given by

$$G = \sqrt[n]{{x_1}^{f_1} \times {x_2}^{f_2} \times {x_3}^{f_3} \times \ \ldots \times {x_m}^{f_m}} = \left({x_1}^{f_1} \times {x_2}^{f_2} \times {x_3}^{f_3} \times \ \ldots \times {x_m}^{f_m}\right)^{\frac{1}{n}}$$ where $\sum_{i=1}^{m} f_i = n$, total no. of observations

$$\log G = \frac{1}{n}(f_1 \log x_1 + f_2 \log x_2 + f_3 \log x_3 + \ \ldots + f_m \log x_m) = \frac{1}{n}\sum_{i=1}^{m} f_i \log x_i$$

$$G = \text{antilog}\left(\frac{1}{n}\sum_{i=1}^{n} f_i \log x_i\right)$$

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Geometric Mean

For Grouped Data with Class Intervals

In case of grouped data with class intervals, its geometric mean is computed in the same manner as that for single value groups, with the exception that the value of x_i is taken as the mid-point or the class-mark of the corresponding class interval

$$G = \sqrt[n]{x_1^{f_1} \times x_2^{f_2} \times x_3^{f_3} \times \ \ldots \times x_m^{f_m}} = \left(x_1^{f_1} \times x_2^{f_2} \times x_3^{f_3} \times \ \ldots \times x_m^{f_m}\right)^{\frac{1}{n}}$$ where $\sum_{i=1}^{m} f_i = n$, total no. of observations

$$\log G = \frac{1}{n}\left(f_1 \log x_1 + f_2 \log x_2 + f_3 \log x_3 + \ \ldots + f_m \log x_m\right) = \frac{1}{n}\sum_{i=1}^{m} f_i \log x_i$$

$$G = \text{antilog}\left(\frac{1}{n}\sum_{i=1}^{n} f_i \log x_i\right)$$

where m is the no. of distinct values of the variable x

x_i is the mid-point or class mark of the i^{th} class with i = 1, 2, 3, ..., m

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Geometric Mean	
Property	Mean of the combined series If n_1 & n_2 are the sizes of two series with G_1 & G_2 as their geometric means, respectively, then the geometric mean G of the combined series is given by $G = \text{antilog}\left(\frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}\right)$

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Geometric Mean

Merits	Demerits	Uses
• It is rigidly defined • It is based upon all observations • It is suitable for further mathematical treatment • It is not affected much by fluctuations of sampling • It gives comparatively more weight to small items	• Because of its abstract mathematical character, it is not easy to understand and calculate for a non-mathematics person • If any one of the observations is 0, geometric mean becomes 0 regardless of the magnitude of the other items • If any one of the observations is negative, geometric mean becomes imaginary regardless of the magnitude of the other items	• To find the rate of population growth & the rate of interest • In the construction of index numbers

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency
  - Geometric Mean
    - Find the geometric mean of i. 2 & 18 ii. 10, 51.2 & 8 iii. (1/2), (1/4), (1/5), (9/72) & (7/4)
    - Stock price of a company increased by 50% in one year, by 20% in the second year, and 90% in the third year. What is the average annual price increase in percentage?

Value	Frequency
1	5
2	9
3	12
4	17
5	14
6	10
7	6

Marks	No. of Students
0-9	12
10-19	18
20-29	27
30-39	20
40-49	17
50-59	6

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Harmonic Mean	
Harmonic Mean of a series of n observations is the reciprocal of the arithmetic mean of the reciprocals of the given values	
For Raw (Ungrouped) Data	For a sample of n observations, $x_1, x_2, x_3, \ldots, x_n$, the harmonic mean H is given by $H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$ where i = 1, 2, 3, …, n
For Grouped Data with Single-Value Groups	In case of grouped data with single-value groups having frequency distribution $x_i \mid f_i$, i = 1, 2, 3, …, m, where f_i is the frequency of the variable x_i m is the no. of distinct values of the variable x its harmonic mean is given by $H = \frac{1}{\frac{1}{n}\sum_{i=1}^{m}\frac{f_i}{x_i}}$ where $\sum_{i=1}^{m} f_i$ = n, total no. of observations

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Harmonic Mean

For Grouped Data with Class Intervals

In case of grouped data with class intervals, its harmonic mean is computed in the same manner as that for single value groups, with the exception that the value of x_i is taken as the mid-point or the class-mark of the corresponding class interval

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{m}\frac{f_i}{x_i}}$$

where m is the no. of distinct values of the variable x

x_i is the mid-point or class mark of the i^{th} class with i = 1, 2, 3, …, m

$\sum_{i=1}^{m} f_i$ = n, total no. of observations

If equal distances are travelled per unit of time with varying speeds, S_1, S_2, S_3, …, S_n, then the average speed is given by the harmonic mean of the speeds S_1, S_2, S_3, …, S_n, i.e., Average Speed $= \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{S_i}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{S_i}}$

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Harmonic Mean - Weighted Harmonic Mean

If instead of fixed (constant) distance being travelled with varying speeds, varying distances S_1, S_2, S_3, ..., S_n, are travelled with varying speeds S_1, S_2, S_3, ..., S_n, respectively, then the average speed is given by the weighted harmonic mean of the speeds S_1, S_2, S_3, ..., S_n, with the weights being the corresponding distances being travelled i.e.

$$\text{Average Speed} = \frac{D_1 + D_2 + D_3 + \ldots + D_n}{\frac{D_1}{S_1} + \frac{D_2}{S_2} + \frac{D_3}{S_3} + \ldots + \frac{D_n}{S_n}} = \frac{\sum_{i=1}^{n} D_i}{\sum_{i=1}^{n} \frac{D_i}{S_i}}$$

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Harmonic Mean

Merits	Demerits
• It is rigidly defined • It is based upon all observations • It is suitable for further mathematical treatment • It is not affected much by fluctuations of sampling • It gives greater importance to small items	• It is not easily understood • It is difficult to compute

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency
  - Harmonic Mean
    - Find the harmonic Mean of 4, 8, 12, 16, 3, 5, 7 and 9
    - A cyclist pedals from his house to his college at a speed of 10 km/hr & back from college to his house at a speed of 15 km/hr. Find the average speed.
    - A trip entailed travelling 900 km by train at an average speed of 60 km/hr, 3000 km by boat at an average speed of 25 km/hr, 400 km by plane at an average speed of 350 km/hr and finally 15 km by taxi at an average speed of 25 km/hr. What is the average speed for the entire distance?

Value	Frequency
1	5
2	9
3	12
4	17
5	14
6	10
7	6

Marks	No. of Students
0-9	12
10-19	18
20-29	27
30-39	20
40-49	17
50-59	6

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Relationship between Arithmetic Mean (AM), Geometric Mean (GM) & Harmonic Mean (HM)	
$AM \times HM = GM^2$	$AM \geq GM \geq HM$

  - The harmonic mean of two numbers is 3, and their arithmetic mean is 4. Find the two numbers and their geometric mean.
  - The geometric mean of two numbers is 8, and their harmonic mean is 6.4. Find the two numbers and their arithmetic mean
  - The geometric mean of two numbers is 25, and their arithmetic mean is 65. Find the two numbers and their harmonic mean.

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Median

- Median is a single value from the data set that measures the central item in the data
- It is the middlemost or most central item in the set of numbers
- Half of the items lie above this point, and the other half lie below it

For Raw (Ungrouped) Data	• The median is the value in the middle when the data are arranged in ascending order (smallest value to largest value) • With an odd number of observations, the median is the value of middle observation • With an even number of observations, the median is the average of the values of the two middle observations

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Median	
For Grouped Data with Single-Value Groups	In case of grouped data with single-value groups having frequency distribution $x_i \mid f_i$, i = 1, 2, 3, ..., m, where f_i is the frequency of the variable x_i m is the no. of distinct values of the variable x the median is the value of x for which the cumulative frequency is just greater than $\frac{n}{2}$ where $n = \sum_{i=1}^{m} f_i$
For Grouped Data with Class Intervals	In case of grouped data with class intervals formed by cut-point grouping, the class corresponding to the cumulative frequency just greater than $\frac{n}{2}$ is called the median class, and the median is given by $\text{Median} = l + \frac{w}{f}\left(\frac{n}{2} - c\right)$ where l is lower cut-point of the median class w is width of the median class f is frequency of the median class c is cumulative frequency of the class preceding the median class

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Median

Merits	Demerits	Uses
• It is rigidly defined • It is easily understood & easy to calculate. In some cases it can be located by merely inspection • It is not affected by extreme values • It can be calculated for distributions with open-ended classes	• For even no. of observations, it cannot be determined exactly (estimated as arithmetic mean of middle two observations) • It is not based on all observations • It is not amenable to algebraic treatment • As compared with mean, it is affected much by sampling fluctuations	• Only average that can be used with qualitative data that cannot be measured quantitatively but can be arranged in ascending or descending order of magnitude • In finding typical values of wages, income distribution, etc.

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency
  - Median
    - Find the median age of employees for the following set of sample data showing the ages of ten employees in an organization.

47	25	46	35	52	45	23	34	54	25

    - Find the median customer service rating for the following set of sample data showing the customer service ratings on a scale of 1-100 provided by 15 customers of a restaurant.

95	85	90	99	80	68	95	77	80	60	88	95	80	95	75

    - Obtain the median for the following frequency distribution.

Value	1	2	3	4	5	6	7	8	9
Frequency	8	10	11	16	20	25	15	9	6

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency
  - Median
    - Find the median of the following wage distribution.

Wages (Rs./hr)	No. of Labourers
20 - less than 30	3
30 - less than 40	5
40 - less than 50	20
50 - less than 60	10
60 - less than 70	5

- For the following sample data of marks of students, compute the median marks.

Marks	No. of students
0 – 9	12
10 – 19	18
20 – 29	27
30 – 39	20
40 – 49	17
50 – 59	6

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Mode	
For Raw (Ungrouped) Data	• Mode is that value of the variable that occurs most frequently in a set of observations • If no value occurs more than once, then the data set has no mode • If two values have the same maximum frequency, then the data set has two modes • If more than two values have the same maximum frequency, then the data set has multiple modes
For Grouped Data with Single-Value Groups	In case of grouped data with single-value groups having frequency distribution $x_i \mid f_i$, $i = 1, 2, 3, \ldots, m$, where f_i is the frequency of the variable x_i m is the no. of distinct values of the variable x, the mode is the value x_i having the highest frequency f_i

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Mode

For Grouped Data with Class Intervals

In case of grouped data with class intervals formed by cut-point grouping, the class corresponding to the maximum frequency is called the modal class, and the mode is given by

$$\text{Mode} = l + \frac{w(f_1 - f_0)}{(f_1 - f_0) - (f_2 - f_1)} = l + \frac{w(f_1 - f_0)}{2f_1 - f_0 - f_2}$$

where
- l is lower cut-point of the modal class
- w is width of the modal class
- f_1 is frequency of the modal class
- f_0 is frequency of the class preceding the modal class
- f_2 is frequency of the class succeeding the modal class

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Mode

Merits	Demerits	Uses
• It is readily comprehensible & easy to calculate • It can be readily located by mere inspection in some cases • It is not at all affected by extreme values • It can be conveniently located even if the frequency distribution has open-ended classes or unequal magnitude provided the modal class and its preceding & succeeding classes are of the same magnitude	• It is ill defined, i.e., it is not always possible to find a clearly defined mode (in cases of bimodal & multimodal distributions) • It is not based on all observations • It is not capable for further mathematical treatment • As compared with mean, it is more affected by fluctuations in sampling	• In finding typical values of a variable (height, weight, size)

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency
  - Mode
    - Find the mode if the batch sizes for a sample of five batches in a coaching center are 46, 54, 42, 46, and 32.

Graduate	Monthlty Starting Salary (INR)
1	63500
2	79000
3	75000
4	59500
5	75500
6	73500
7	74500
8	66000
9	54500
10	72500
11	77500
12	66000

Value	1	2	3	4	5	6	7	8	9
Frequency	8	10	11	16	20	25	15	9	6

Wages (Rs./hr)	No. of Labourers
20 - less than 30	3
30 - less than 40	5
40 - less than 50	20
50 - less than 60	10
60 - less than 70	5

Marks	No. of students
0 – 9	12
10 – 19	18
20 – 29	27
30 – 39	20
40 – 49	17
50 – 59	6

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

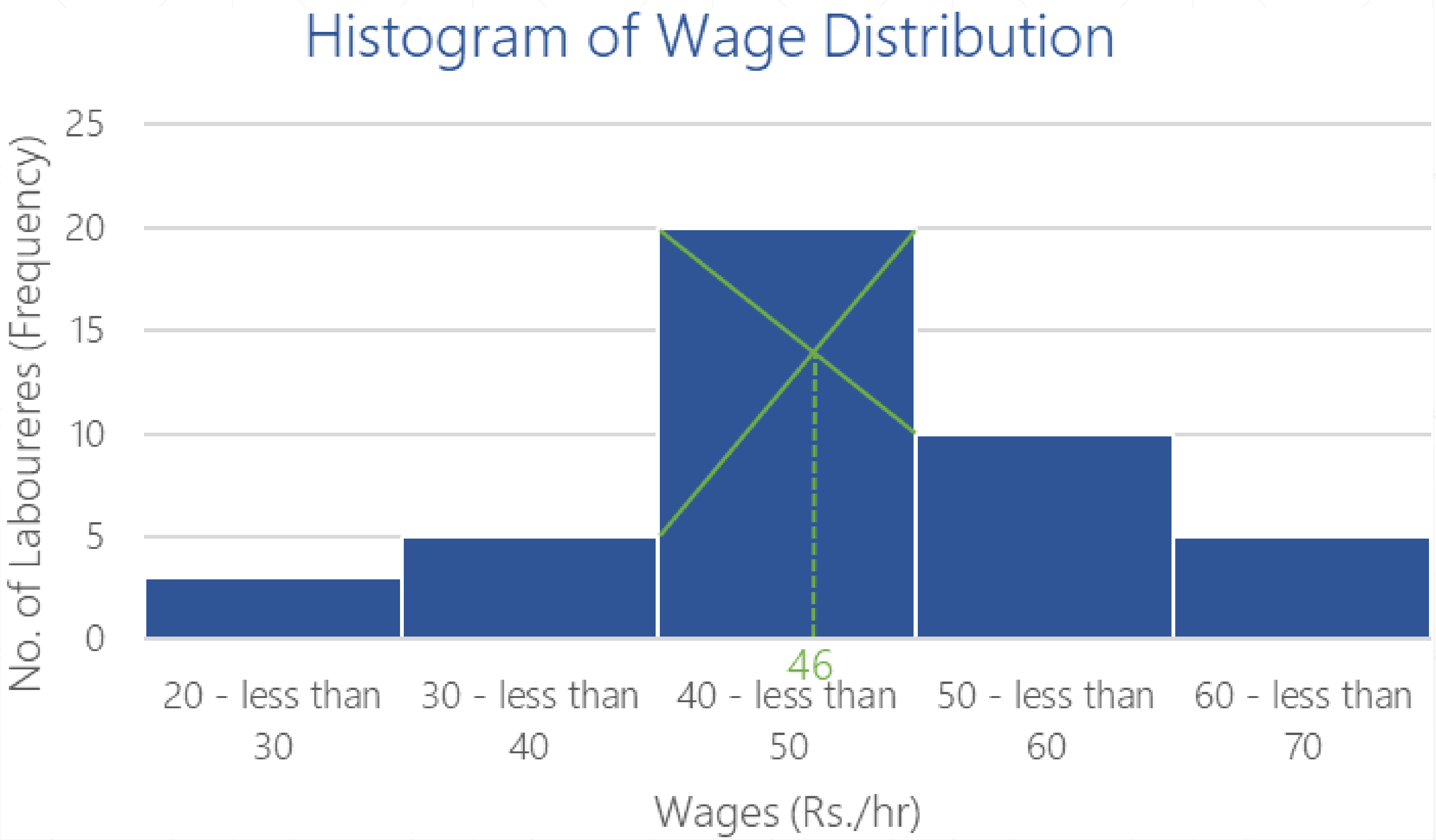
Mode	
For Grouped Data with Class Intervals	Mode can also be located graphically using histograms 1. Construct a histogram for the given data 2. The highest vertical bar in histogram has the highest frequency representing the modal class 3. Draw a straight line from the right corner of the vertical bar of the modal class to the right corner of the vertical bar of the class preceding the modal class 4. Draw a straight line, from the left corner of the vertical bar of the modal class with the left corner of the vertical bar of the class succeeding the modal class 5. From the point of intersection of the two lines drawn above, draw a line parallel to the vertical axis till it meets the horizontal axis at a point 6. The abscissa (x-coordinate) of above the point on the horizontal axis gives the value of the mode

Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency
  - Mode

Wages (Rs./hr)	No. of Labourers
20 - less than 30	3
30 - less than 40	5
40 - less than 50	20
50 - less than 60	10
60 - less than 70	5



Descriptive Statistics - Data Summarization

Central Tendency

- Measures of Central Tendency

Measure	Computation	Scale of Data	Characteristics	Suitability
Mean	Sum of values divided by no. of values	Interval Ratio	• Numerical center of data • Sum of deviations from mean is 0 • Sensitive to extreme values	• Most appropriate measure for data with interval & ratio scale if it is not highly skewed
Median	Middle value of data sorted in ascending order	Ordinal Interval Ratio	• Not sensitive to extreme values • Computed only from center values • Does not use information from all the data	• Most appropriate measure for data with interval & ratio scale if it is highly skewed
Mode	Value/s that occur/s the most frequently in the data	Nominal Ordinal Interval Ratio	• May not reflect the center • May not exist • Might have multiple modes	• Most appropriate measure for data with nominal scale • Results in loss of power in terms of information that could be gained from the data if used with ordinal, interval & ratio scales

Descriptive Statistics - Data Summarization

Dispersion

- Meaning
  - Averages or measures of central tendency provide an idea about the concentration of the observations about the central part of the frequency distribution, but not a complete picture of the distribution

Series 1	7	9	11	8	10
Series 2	15	12	6	3	9
Series 3	9	13	17	1	5

- Each series has the same no. of observations , 5, and the same mean, 9
- Given just the mean and the no. of observations, it is not possible to identify the series being referred to

  - Measures of central tendency are inadequate in providing a complete picture of the distribution, and hence must be supported by some other measures, like dispersion

Descriptive Statistics - Data Summarization

Dispersion

- Meaning
  - For the following daily production output over a period of five days for two different manufacturing facilities of a company,

Plant A	15 units	25 units	35 units	20 units	30 units
Plant B	23 units	26 units	25 units	24 units	27 units

the following production report is submitted by the two plant managers to the company vice president

	Mean	Median
Plant A	25 units	25 units
Plant B	25 units	25 units

Descriptive Statistics - Data Summarization

Dispersion

- Meaning

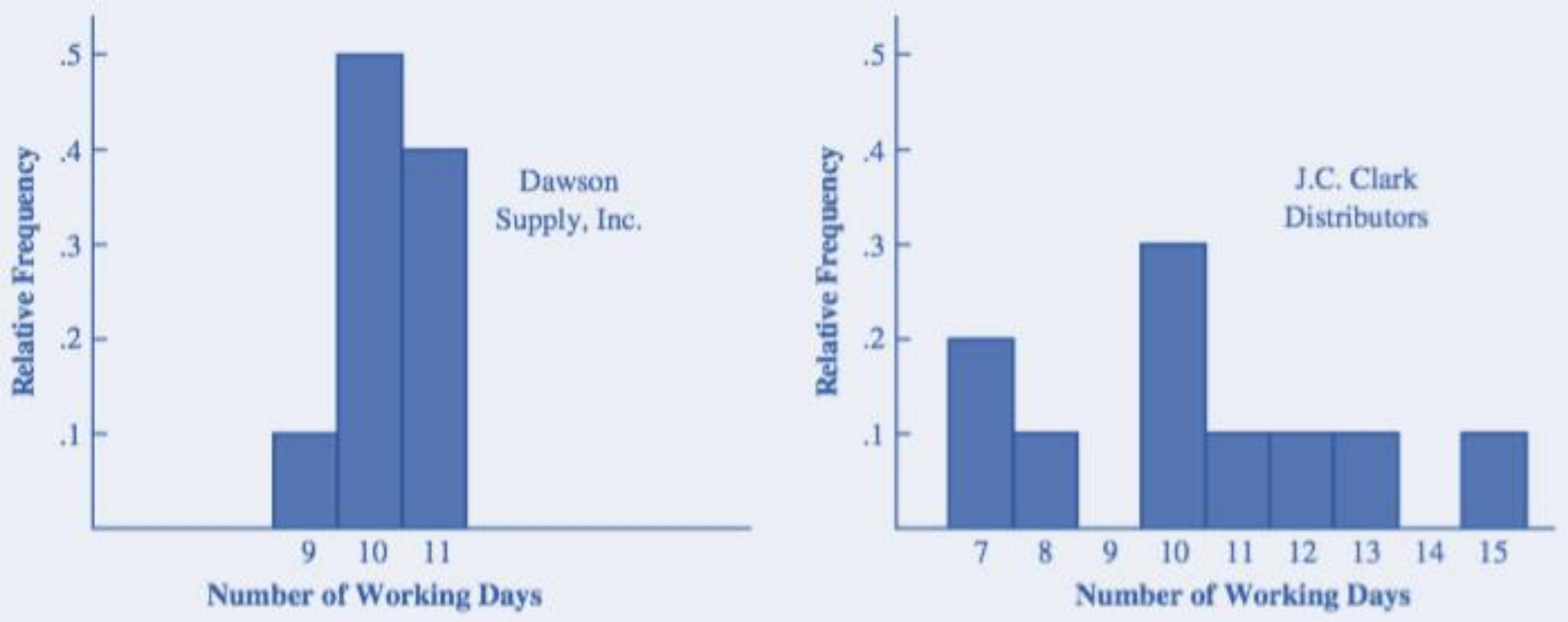
Conclusions Based on the Summary Report Only	Closer Look at Data Suggests
• Average production is the same at both plants • At both plants, the output is at or more than 25 units half the time and at or fewer than 25 units half the time • Because the mean and median are equal, the distribution of production output at the two plants is symmetrical • Based on these statistics, there is no reason to believe that the two plants are different in terms of their production output	• Big difference between the two plants in terms of production variation from day to day • Actually, Plant B is more stable producing almost the same quantity every day • Production in Plant A varies considerable with some low-production days, and some high-production days

- Looking at only measures of the data's central location can be misleading

Descriptive Statistics - Data Summarization

Dispersion

- Meaning



Number of working days required to fill orders from two suppliers

Descriptive Statistics - Data Summarization

Dispersion

- Meaning
  - To fully describe a set of data, a measure of variation / spread / dispersion is also required in addition to the measure of the central tendency
  - Dispersion - scatter
  - Dispersion provides an idea about the heterogeneity or homogeneity of the distribution
  - A more homogeneous series is less dispersed / scattered
  - A more heterogeneous series is more dispersed / scattered

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

<table>
<tr><th>Requisites of an Ideal Measure of Central Tendency</th><th colspan="2">Types of Measures</th></tr>
<tr><td rowspan="2">It should be rigidly defined
It should be easy to calculate & easy to understand
It should be based on all observations
It should be amenable to further mathematical treatment
It should be affected as little as possible by fluctuations of sampling</td><td>Absolute Measures</td><td>Measures of dispersion indicate the amount of variation in a set of values, in terms of units of observations</td></tr>
<tr><td>Relative Measures</td><td>Measures of dispersion are free from units of measurements of observations and are used to compare the variation in two or more sets, which are having different units of measurements of observations</td></tr>
</table>

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Some Measures of Dispersion	
Absolute Measures	Relative Measures
• Range	• Coefficient of Range or Coefficient of Dispersion
• Quartile Deviation	• Coefficient of Quartile Deviation or Quartile Coefficient of Dispersion
• Mean Absolute Deviation	• Coefficient of Mean Deviation or Mean Deviation of Dispersion
• Standard Deviation	• Coefficient of Standard Deviation or Standard Coefficient of Dispersion
• Variance	• Coefficient of Variation (special case of Standard Coefficient of Dispersion)

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Range

- An absolute measure of dispersion, the range of a data set is the difference between the maximum (largest) & the minimum (smallest) observations

Range = Largest Value – Smallest Value

- In case of grouped data with class intervals, the range is the difference between the upper limit of the highest class and the lower limit of the lowest class

Merits	Demerits
• Easy & quick to compute	• Considers only the largest & smallest values, hence very sensitive to extreme values in the data set • Computed from only two values of the data set irrespective of the no. values in the sample or population

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion
  - Range

Series 1	30	40	40	40	40	40	50

Series 2	1330	1335	1340	1340	1340	1345	1350

Coefficient of Range

- A relative measure of dispersion based on the value of range

$$\text{Coefficient of Range} = \frac{\text{Largest Value} - \text{Smallest Value}}{\text{Largest Value} + \text{Smallest Value}}$$

- Since it is a ratio, it is dimensionless, & hence can be used to compare the dispersions of different sets

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion
  - Coefficient of Range
    - For the following frequency distribution, compute the range & coefficient of range.

Value (x)	1	2	3	4	5	6	7
Frequency (f)	5	9	12	17	14	10	6

    - For the following sample data of marks of students, compute the range & coefficient of range.

Marks (x)	10-19	20-29	30-39	40-49	50-59
No. of Students (f)	18	27	20	17	6

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Mean Absolute Deviation

- Mean Absolute Deviation (MAD) - mean of the absolute differences between each data value and the average
- Gives the mean amount of spread relative to the average
- Better measure of dispersion than range, since it is based on all observations
- Does not consider the polarity (sign), so it creates artificiality & hence, cannot be used for further mathematical treatment
- Deviation can be measured from any of the measures of central tendency (usually the mean, median or mode)

For Raw (Ungrouped) Data	For a sample of n observations, $x_1, x_2, x_3, \ldots, x_n$, having arithmetic mean, $\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ where i = 1, 2, 3, …, n, the Mean Absolute Deviation = $\frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \bar{x}\rvert$

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Mean Absolute Deviation	
For Grouped Data with Single-Value Groups	In case of grouped data with single-value groups having frequency distribution $x_i \mid f_i$, i = 1, 2, 3, …, m, where f_i is the frequency of the variable x_i m is the no. of distinct values of the variable x, arithmetic mean, $\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \ldots + f_mx_m}{f_1 + f_2 + f_3 + \ldots + f_m} = \frac{\sum_{i=1}^{m} f_ix_i}{\sum_{i=1}^{m} f_i} = \frac{1}{n}\sum_{i=1}^{m} f_ix_i$ since $\sum_{i=1}^{m} f_i = n$, the total number of observations, Mean Absolute Deviation $= \frac{1}{n}\sum_{i=1}^{n} f_i\lvert x_i - \bar{x}\rvert$
For Grouped Data with Class Intervals	In case of grouped data with class intervals, its mean absolute deviation is computed in the same manner as that for single value groups, with the exception that the value of x_i is taken as the mid-point or the class-mark of the corresponding class interval

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Mean Absolute Deviation

- For mean absolute deviation from median, instead of arithmetic mean, the median is used in the above set of formulae
- Similarly, the mean absolute deviation from any given value can be computed by substituting that value for the value of the arithmetic mean used in the above set of formulae

- The batch size for a sample of five batches in a coaching center are 46, 54, 42, 46, and 32. Find the mean absolute deviation of the batch size.
- Find the MAD of the monthly salary data shown in the adjacent table.

Graduate	Monthlty Starting Salary (INR)
1	63500
2	79000
3	75000
4	59500
5	75500
6	73500
7	74500
8	66000
9	54500
10	72500
11	77500
12	66000

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion
  - Mean Absolute Deviation
    - Find the mean absolute deviation for the following data.

Value (x)	1	2	3	4	5	6	7
Frequency (f)	5	9	12	17	14	10	6

    - For the following sample data of marks obtained by students, compute the arithmetic mean & the mean absolute deviation.

Marks (x)	10-19	20-29	30-39	40-49	50-59
No. of Students (f)	18	27	20	17	6

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Coefficient of Mean Absolute Deviation

- A relative measure of dispersion based on the value of mean absolute deviation and the value of the measure of central tendency used to calculate the mean absolute deviation

$$\text{Coefficient of Mean Deviation} = \frac{\text{Mean Absolute Deviation}}{\text{Mean}}$$

Since it is a ratio, it is dimensionless & can be used to compare the dispersions of different sets

- If instead of the mean, the median is used to calculate the mean absolute deviation, then the

$$\text{Coefficient of Mean Deviation from Median} = \frac{\text{Mean Absolute Deviation from Median}}{\text{Median}}$$

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion
  - Coefficient of Mean Absolute Deviation
    - Following are the number of hours a machine worked for the last 9 weeks: 47, 63, 75, 39, 10, 60, 96, 32, 28. Find the:
      - Mean absolute deviation from the mean
      - Coefficient of mean deviation from mean
      - Mean absolute deviation from the median
      - Coefficient of mean deviation from median
    - Following are the observations showing the age of 50 employees working in a wholesale center. Find the

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion
  - Coefficient of Mean Absolute Deviation
    - Following are the number of hours a machine worked for the last 9 weeks: 47, 63, 75, 39, 10, 60, 96, 32, 28. Find the:
      - Mean absolute deviation from the mean
      - Coefficient of mean deviation from mean
      - Mean absolute deviation from the median
      - Coefficient of mean deviation from median
    - Following are observations showing the age of 50 employees working in a wholesale center. Find the:
      - Mean absolute deviation from the mean
      - Coefficient of mean deviation from mean
      - Mean absolute deviation from the median
      - Coefficient of mean deviation from median deviation

Age	No. of Employees
40-44	4
45-49	7
50-54	14
55-59	11
60-64	8
65-69	6

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Standard Deviation

- Standard Deviation - the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean
- It is the measure of spread most commonly used in statistical practice when the mean is used to calculate central tendency i.e., it measures spread around the mean
- Because of its close links with the mean, standard deviation can be greatly affected if the mean gives a poor measure of central tendency
- Standard deviation is also influenced by outliers, one value could contribute largely to the results of the standard deviation, and in that sense, the standard deviation is a good indicator of the presence of outliers
- Standard deviation is a very useful measure of spread for symmetrical distributions with no outliers

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Standard Deviation

- Standard deviation is useful when comparing spread of two separate data sets that have approximately same mean
- The data set with the smaller standard deviation has a narrower spread of measurements around the mean and therefore usually has comparatively fewer high or low values
- An item selected at random from a data set whose standard deviation is low has a better chance of being close to the mean than an item from a data set whose standard deviation is higher

For Raw (Ungrouped) Data	For a sample of n observations, $x_1, x_2, x_3, \ldots, x_n$, having arithmetic mean, $\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ where i = 1, 2, 3, …, n, the Standard Deviation of the Sample, s = $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Standard Deviation	
For Grouped Data with Single-Value Groups	In case of grouped data with single-value groups having frequency distribution $x_i \mid f_i$, i = 1, 2, 3, …, m, where f_i is the frequency of the variable x_i m is the no. of distinct values of the variable x, arithmetic mean, $\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \ldots + f_mx_m}{f_1 + f_2 + f_3 + \ldots + f_m} = \frac{\sum_{i=1}^{m} f_ix_i}{\sum_{i=1}^{m} f_i} = \frac{1}{n}\sum_{i=1}^{m} f_ix_i$ since $\sum_{i=1}^{m} f_i = n$, the total number of observations, Standard Deviation of the Sample, s = $\sqrt{\frac{1}{n-1}\sum_{i=1}^{m} f_i(x_i - \bar{x})^2}$
For Grouped Data with Class Intervals	In case of grouped data with class intervals, its standard deviation is computed in the same manner as that for single value groups, with the exception that the value of x_i is taken as the mid-point or the class-mark of the corresponding class interval

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Standard Deviation	
For Raw (Ungrouped) Population Data	For an entire population of N observations, $x_1, x_2, x_3, \ldots, x_N$, having arithmetic mean, $\mu = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$ where i = 1, 2, 3, …, N, the Standard Deviation of the Population, $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Standard Deviation	
For Grouped Population Data with Single-Value Groups	In case of grouped population data with single-value groups having frequency distribution $x_i \mid f_i$, i = 1, 2, 3, …, m, where f_i is the frequency of the variable x_i m is the no. of distinct values of the variable x, arithmetic mean, $\mu = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \ldots + f_mx_m}{f_1 + f_2 + f_3 + \ldots + f_m} = \frac{\sum_{i=1}^{m} f_ix_i}{\sum_{i=1}^{m} f_i} = \frac{1}{N}\sum_{i=1}^{m} f_ix_i$ since $\sum_{i=1}^{m} f_i = N$, the total number of observations, Standard Deviation of the Population, $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{m} f_i(x_i - \mu)^2}$
For Grouped Population Data with Class Intervals	In case of grouped population data with class intervals, its standard deviation is computed in the same manner as that for single value groups, with the exception that the value of x_i is taken as the mid-point or the class-mark of the corresponding class interval

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Standard Deviation		
Properties	Merits	Demerits
• Standard deviation is only used to measure spread or dispersion around the mean of a data set • Standard deviation is never negative • For data with approximately the same mean, the greater the spread, the greater the standard deviation • If all values of a data set are the same, Standard deviation is 0 • Standard deviation overcomes the drawback of ignoring the polarity (sign) associated the mean deviation (mean absolute deviation)	• Standard deviation is suitable for further mathematical treatment • Of all the measures, standard deviation is affected least by fluctuations of sampling • It is regarded the best & most powerful measure of dispersion	• It is not readily comprehensible • It provides greater weight to extreme values, and is sensitive to outliers. A single outlier can raise the standard deviation and in turn, distort the picture of spread

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion
  - Standard Deviation
    - Weights (in gm) of eight eggs laid by a hen are 60, 56, 61, 68, 51, 53, 69, 54. Calculate the standard deviation of the weight of the eggs.
    - Thirty farmers were asked how many farm workers they hire during a typical harvest season. Their responses are summarized in the adjacent table. Calculate standard deviation of no. of workers hired.
    - 220 students were asked the number of hours per week they spent watching television. With this information, calculate the mean and standard deviation of hours spent watching television by the 220 students.

Workers	Frequency
0	1
1	1
2	2
3	3
4	6
5	5
6	4
7	3
8	3
9	2

Hours	No. of Students
10-14	2
15-19	12
20-24	23
25-29	60
30-34	77
35-39	38
40-44	8

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Coefficient of Standard Deviation

- Coefficient of Standard Deviation (also known as the Standard Coefficient of Dispersion) - a relative measure of dispersion based on the value of the standard deviation and the mean

$$\text{Coefficient of Standard Deviation (for population)} = \frac{\text{Standard Deviation}}{\text{Mean}} = \frac{\sigma}{\mu}$$

$$\text{Coefficient of Standard Deviation (for sample)} = \frac{\text{Standard Deviation}}{\text{Mean}} = \frac{s}{\bar{x}}$$

Since it is a ratio, it is dimensionless & can be used to compare the dispersions of different sets

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Coefficient of Variation

- Coefficient of Variation - a special case of the standard coefficient of dispersion, is a relative measure of dispersion based on the value of the standard deviation and the mean, and is expressed as a percentage

$$\text{Coefficient of Variation (for population)} = \left(\frac{\text{Standard Deviation}}{\text{Mean}} \times 100\right)\% = \left(\frac{\sigma}{\mu} \times 100\right)\%$$

$$\text{Coefficient of Variation (for sample)} = \left(\frac{\text{Standard Deviation}}{\text{Mean}} \times 100\right)\% = \left(\frac{s}{\bar{x}} \times 100\right)\%$$

It is generally used to compare the dispersions of different sets

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Variance

- Variance - the average of the squared deviations of the given values from their arithmetic mean i.e., the square of the standard variation
- The unit of measure for the variance is the squared of the unit of measure used for the variable

For Raw (Ungrouped) Data	For a sample of n observations, $x_1, x_2, x_3, \ldots, x_n$, having arithmetic mean, $\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ where i = 1, 2, 3, …, n, the Variance of the Sample, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Variance	
For Grouped Data with Single-Value Groups	In case of grouped data with single-value groups having frequency distribution $x_i \mid f_i$, i = 1, 2, 3, ..., m, where f_i is the frequency of the variable x_i m is the no. of distinct values of the variable x, arithmetic mean, $\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \ldots + f_mx_m}{f_1 + f_2 + f_3 + \ldots + f_m} = \frac{\sum_{i=1}^{m} f_ix_i}{\sum_{i=1}^{m} f_i} = \frac{1}{n}\sum_{i=1}^{m} f_ix_i$ since $\sum_{i=1}^{m} f_i = n$, the total number of observations, Variance of the Sample, $s^2 = \frac{1}{n-1}\sum_{i=1}^{m} f_i(x_i - \bar{x})^2$
For Grouped Data with Class Intervals	In case of grouped data with class intervals, its variance is computed in the same manner as that for single value groups, with the exception that the value of x_i is taken as the mid-point or the class-mark of the corresponding class interval

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Variance	
For Raw (Ungrouped) Population Data	For an entire population of N observations, $x_1, x_2, x_3, \ldots, x_N$, having arithmetic mean, $\mu = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$ where i = 1, 2, 3, …, N, the Variance of the Population, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Variance	
For Grouped Population Data with Single-Value Groups	In case of grouped population data with single-value groups having frequency distribution $x_i \mid f_i$, i = 1, 2, 3, …, m, where f_i is the frequency of the variable x_i m is the no. of distinct values of the variable x, arithmetic mean, $\mu = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \ldots + f_mx_m}{f_1 + f_2 + f_3 + \ldots + f_m} = \frac{\sum_{i=1}^{m} f_ix_i}{\sum_{i=1}^{m} f_i} = \frac{1}{N}\sum_{i=1}^{m} f_ix_i$ since $\sum_{i=1}^{m} f_i$ = N, the total number of observations, Variance of the Population, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{m} f_i(x_i - \mu)^2$
For Grouped Population Data with Class Intervals	In case of grouped population data with class intervals, its variance is computed in the same manner as that for single value groups, with the exception that the value of x_i is taken as the mid-point or the class-mark of the corresponding class interval

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion
  - Variance
    - Weights (in gm) of eight eggs laid by a hen are 60, 56, 61, 68, 51, 53, 69, 54. Calculate the variance of the weight of the eggs.
    - Thirty farmers were asked how many farm workers they hire during a typical harvest season. Their responses are summarized in the adjacent table. Calculate the variance of no. of workers hired.
    - 220 students were asked the number of hours per week they spent watching television. With this information, calculate the mean and variance of hours spent watching television by the 220 students.

Workers	Frequency
0	1
1	1
2	2
3	3
4	6
5	5
6	4
7	3
8	3
9	2

Hours	No. of Students
10-14	2
15-19	12
20-24	23
25-29	60
30-34	77
35-39	38
40-44	8

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Variance - Alternate Formula	
For Raw (Ungrouped) Data	For a sample of n observations, $x_1, x_2, x_3, \ldots, x_n$, having arithmetic mean, $\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ where i = 1, 2, 3, …, n, the Variance of the Sample, $s^2 = \frac{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2}{n-1}$

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Variance - Alternate Formula	
For Grouped Data with Single-Value Groups	In case of grouped data with single-value groups having frequency distribution $x_i \mid f_i$, i = 1, 2, 3, …, m, where f_i is the frequency of the variable x_i m is the no. of distinct values of the variable x, arithmetic mean, $\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \ldots + f_mx_m}{f_1 + f_2 + f_3 + \ldots + f_m} = \frac{\sum_{i=1}^{m} f_ix_i}{\sum_{i=1}^{m} f_i} = \frac{1}{n}\sum_{i=1}^{m} f_ix_i$ since $\sum_{i=1}^{m} f_i = n$, the total number of observations, Variance of the Sample, $s^2 = \frac{\sum_{i=1}^{m} f_i{x_i}^2 - n\bar{x}^2}{n-1}$
For Grouped Data with Class Intervals	In case of grouped data with class intervals, its variance is computed in the same manner as that for single value groups, with the exception that the value of x_i is taken as the mid-point or the class-mark of the corresponding class interval

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Variance - Alternate Formula	
For Raw (Ungrouped) Population Data	For an entire population of N observations, $x_1, x_2, x_3, \ldots, x_N$, having arithmetic mean, $\mu = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$ where i = 1, 2, 3, ..., N, the Variance of the Population, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} {x_i}^2 - \mu^2 = \frac{\sum_{i=1}^{N} {x_i}^2 - N\mu^2}{N}$

Descriptive Statistics - Data Summarization

Dispersion

- Measures of Dispersion

Variance - Alternate Formula	
For Grouped Population Data with Single-Value Groups	In case of grouped population data with single-value groups having frequency distribution $x_i \mid f_i$, i = 1, 2, 3, …, m, where f_i is the frequency of the variable x_i m is the no. of distinct values of the variable x, arithmetic mean, $\mu = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \ldots + f_mx_m}{f_1 + f_2 + f_3 + \ldots + f_m} = \frac{\sum_{i=1}^{m} f_ix_i}{\sum_{i=1}^{m} f_i} = \frac{1}{N}\sum_{i=1}^{m} f_ix_i$ since $\sum_{i=1}^{m} f_i = N$, the total number of observations, Variance of the Population, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{m} f_i{x_i}^2 - \mu^2 = \frac{\sum_{i=1}^{m} f_i{x_i}^2 - N\mu^2}{N}$
For Grouped Population Data with Class Intervals	In case of grouped population data with class intervals, its variance is computed in the same manner as that for single value groups, with the exception that the value of x_i is taken as the mid-point or the class-mark of the corresponding class interval

Descriptive Statistics - Data Summarization

Dispersion

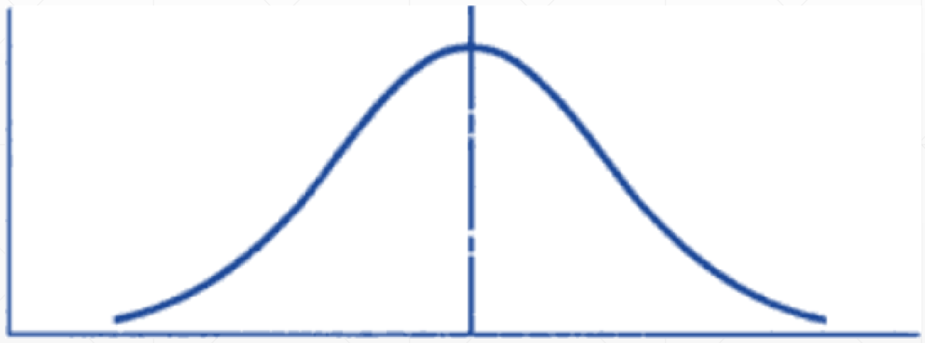
- Measures of Dispersion
  - Variance
    - For a group of 200 students, the mean and standard deviation of scores was found to be 40 and 15, respectively. Later on it was discovered that the scores 43 and 35 were misread as 34 and 53, respectively. Find the corrected mean & standard deviation corresponding to the corrected scores.

Descriptive Statistics - Data Summarization

Skewness

- Meaning
  - Skewness - lack of symmetry
  - While the average measures the central tendency of a distribution and the dispersion measures the scatter of the distribution, the skewness measures the shape of the distribution in terms of its symmetry
  - Skewness indicates the degree of distortion from symmetrical bell curve
  - Symmetrical Curve - a vertical line drawn from the center of the curve to the horizontal axis divides the area of the curve into two equal parts, each being the mirror image of the other (also known as Bell Curve or Normal Curve or Normal Distribution)

Symmetrical Curve
(Bell Curve / Normal Distribution)

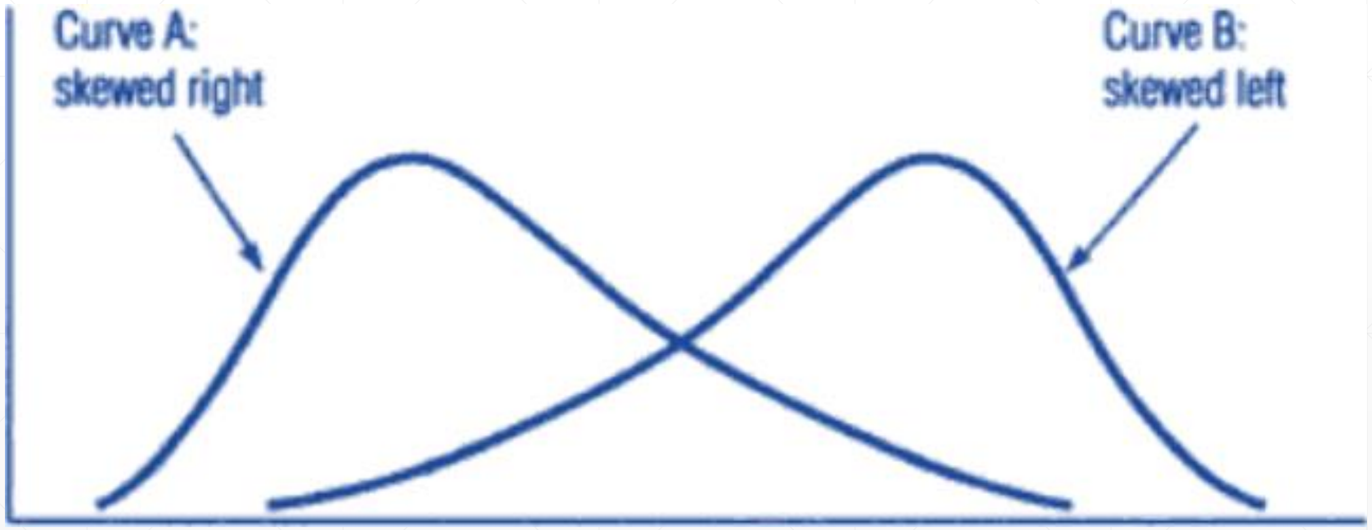


Descriptive Statistics - Data Summarization

Skewness

- Meaning
  - Skewed Curves - curves where values in their frequency distributions are not equally distributed and are concentrated at either the lower end or the higher end
  - Right (Positively) Skewed Curves - curves which tail off (extend) toward the higher end of the measuring scale
  - Left (Negatively) Skewed Curves - curves which tail off (extend) toward the lower end of the measuring scale

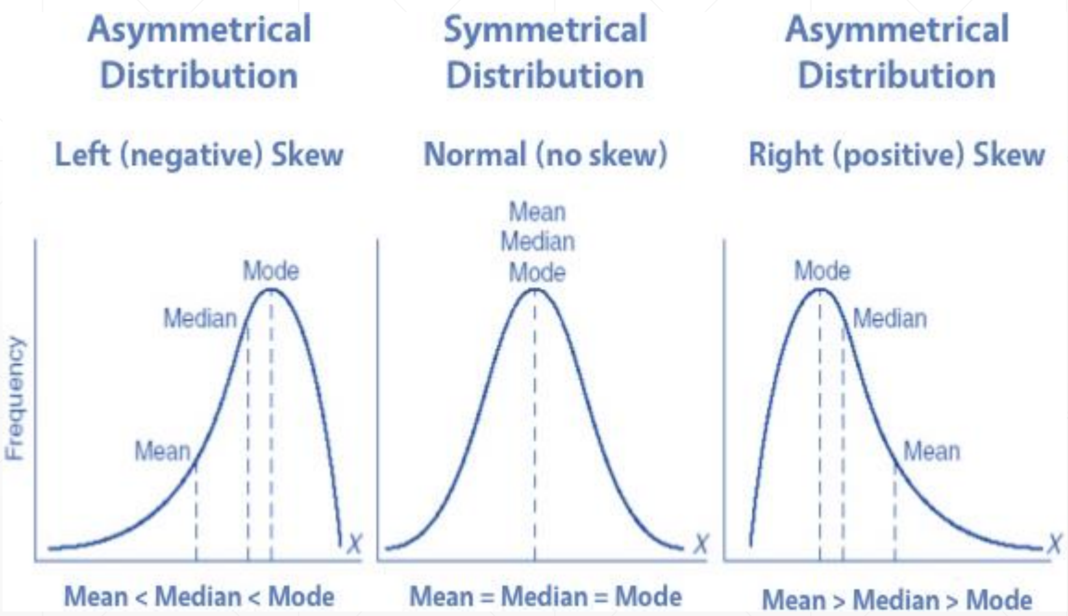
Skewed Curves



Descriptive Statistics - Data Summarization

Skewness

- Meaning
  - When the frequency distribution of the variable is symmetrical, the mean, median & mode are all equal
  - For asymmetrical frequency distribution, the mean, median & mode will be different
    - For left-skewed (negative skewed) distribution, the mean will be less than the median that will be less than the mode i.e., Mean < Median < Mode
    - For left-skewed (negative skewed) distribution, the mean will be less than the median that will be less than the mode i.e., Mean > Median > Mode



Descriptive Statistics - Data Summarization

Skewness

- Measures of Skewness

Absolute Measures of Skewness				
S_k	=	Mean – Median	=	$M - M_d$
S_k	=	Mean – Mode	=	$M - M_o$
S_k	=	$(Q_3 - M_d) - (M_d - Q_1)$	where Q_3 is the third quartile and Q_1 is the first quartile	

- The units for these absolute measures of skewness are the same as that used to measure the data
- To compare different distributions (or data sets), relative measures of skewness, which are pure numbers independent of the units of measurement of data, are used

Descriptive Statistics - Data Summarization

Skewness

- Measures of Skewness

Relative Measures of Skewness
Karl Pearson's Coefficient of Skewness
$S_k = \dfrac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \dfrac{M - M_o}{\sigma}$
• If the mode is ill-defined, then $M_o = 3M_d - 2M$ for a moderately asymmetrical curve, and hence,
$S_k = \dfrac{3(M - M_d)}{\sigma}$
• The limits for Karl Pearson's Coefficient of Skewness are ± 3

Descriptive Statistics - Data Summarization

Skewness

- Measures of Skewness
  - Karl Pearson's Coefficient of Skewness
    - Compute the Karl Pearson's Coefficient of Skewness for the following data.

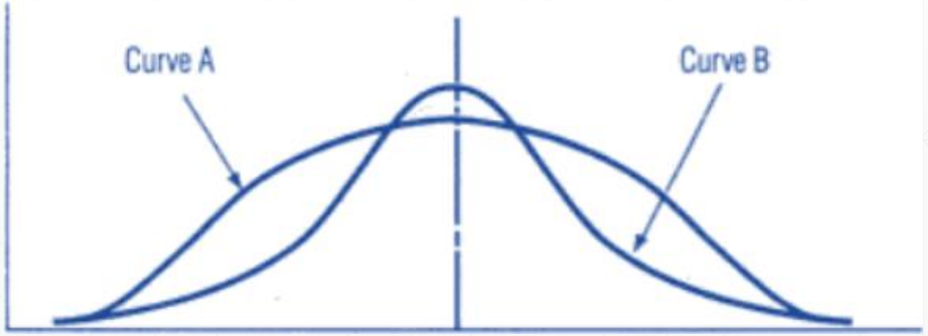
Height (inches)	No. of People
58	10
59	18
60	30
61	42
62	35
63	28
64	16
65	8

Descriptive Statistics - Data Summarization

Kurtosis

- Meaning
  - Kurtosis - measure related to the peaked-ness of the distribution
  - While the average measures the central tendency of a distribution, the dispersion measures the scatter of the distribution, the skewness measures the shape of the distribution in terms of its symmetry, the kurtosis measures the peaked-ness (convexity of curve) of the distribution
  - Together all these measures (average, dispersion, skewness & kurtosis) provide an appropriate summarization of the distribution

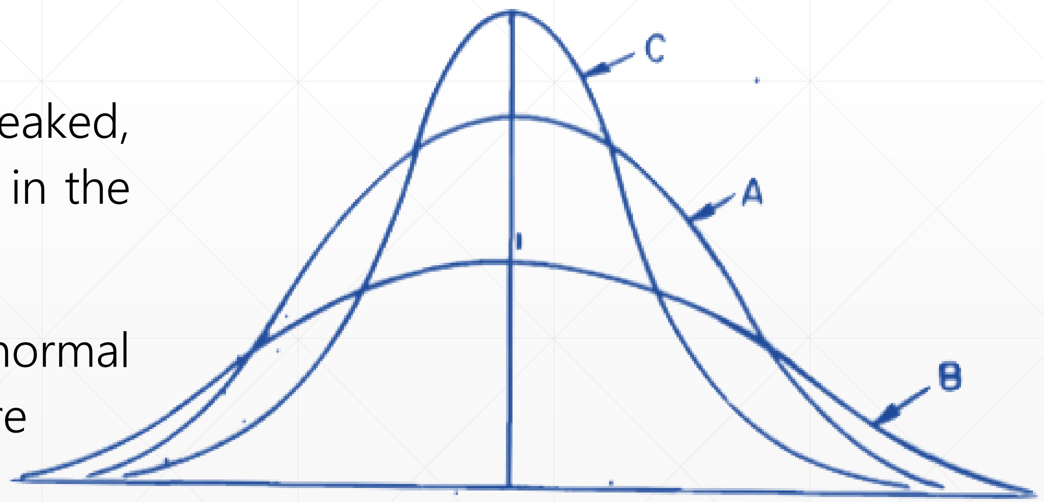
Two Symmetrical Curves with Same Central Location & Dispersion, but Different Kurtosis



Descriptive Statistics - Data Summarization

Kurtosis

- Meaning
  - Different distributions can be classified into three categories depending upon the shape of their peak
  - Platykurtic Curve - a curve flatter than the normal curve ($\beta_2 < 3$, $i.e.$, $\gamma_2 < 0$), Curve B in the adjacent figure
  - Mesokurtic Curve - a curve which is neither flat nor peaked, called the normal curve ($\beta_2 = 3$, $i.e.$, $\gamma_2 = 0$), Curve A in the adjacent figure
  - Leptokurtic Curve - a curve more peaked than the normal curve ($\beta_2 > 3$, $i.e.$, $\gamma_2 > 0$), Curve C in the adjacent figure



Descriptive Statistics - Data Summarization

Kurtosis

- Measures of Kurtosis

Karl Pearson's Coefficient of Kurtosis

$\beta_2 = \dfrac{\mu_4}{\mu_2^{\,2}}$ and $\gamma_2 = \beta_2 - 3$ where μ_4 is the fourth central moment & μ_2 is the second central moment

Thank You

Prof. Jigar M. Shah