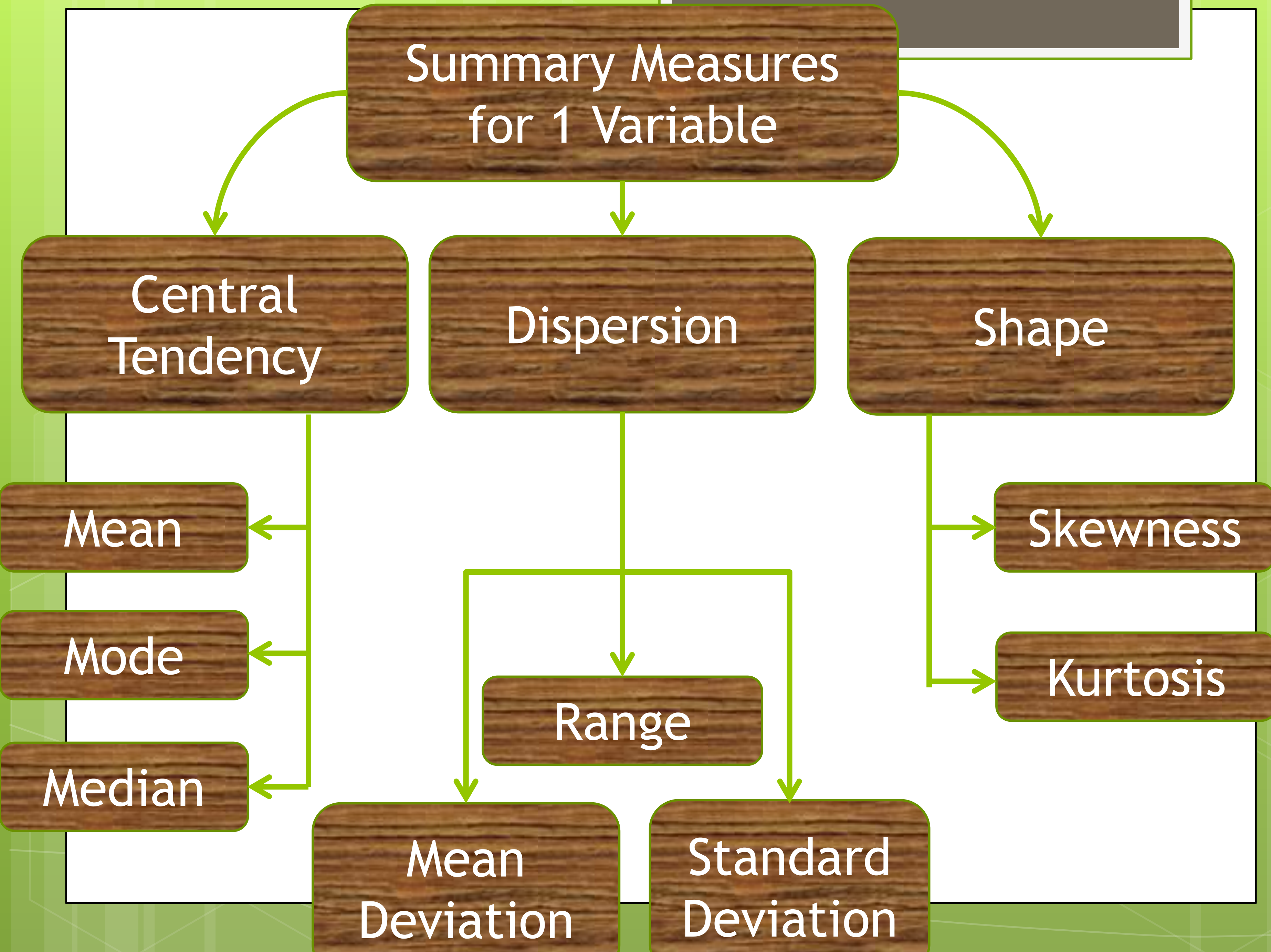
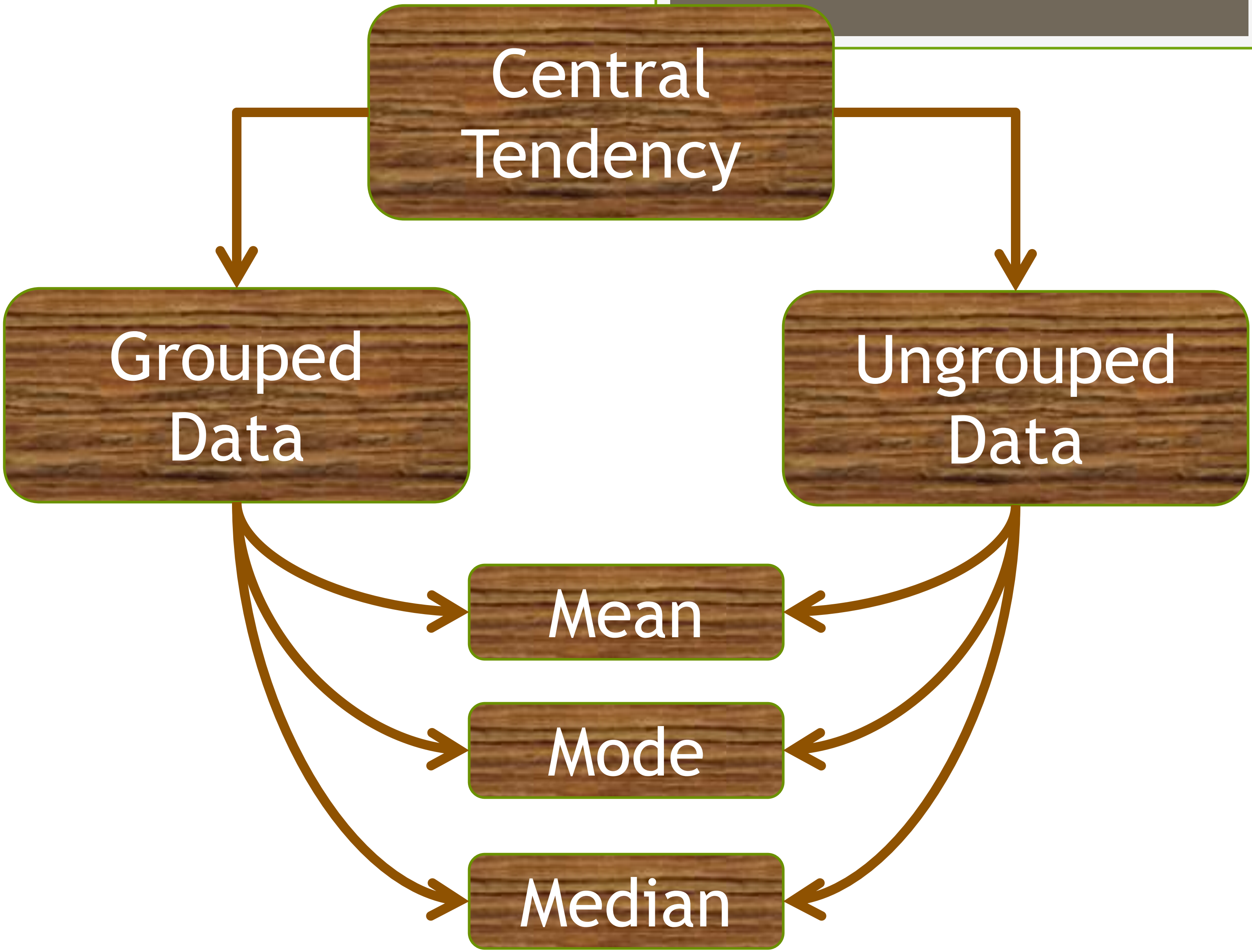
Descriptive Statistics

Jamilusmani


Summary Measures for 1 Variable
Central Tendency
Dispersion
Shape
Mean
Mode
Median
Range
Mean Deviation
Standard Deviation
Skewness
Kurtosis


Central Tendency
Grouped Data
Ungrouped Data
Mean
Mode
Median

Mean

- The most widely used measure
- It is a computed value therefore it is affected by all the values
- It is possible that Mean is not the part of data which it represents
- Its value is affected by extreme value, so in case of skewness in data, mean is not a good measure of central tendency
- It can not be computed from an open-ended distribution

Mean for Ungrouped Data

- Mean
- Population Mean = $\mu = \frac{\sum X}{N}$
- Sample Mean = X-bar $= \frac{\sum X}{n}$
- Find the Mean for 25, 30, 40, 45

	X
	25
	30
	40
	45
Total	140

Average	140/4
Average	35

Mode

- The most repeated value
- Used for Nominal Data
- It is not widely used
- It is not affected by extreme value
- There can be no mode or more than one mode

Mode for Ungrouped Data

- In the class there are 20 boys and 35 girls, what is the mode
  - ___________(20, 35, Boy, Girl, None of them)
- Find the mode for the data set given below
- 2, 5, 7, 4, 2, 8, 7, 12, 5, 2, 8, 10
  - 2
- Is it Possible for a data to have more than one mode
  - Yes

Median

- It is easy to define and easy to understand
  - It is the Middle Value of the arranged data
- It is affected by the number of observations, but not by the value of observation
- Extreme value does not affect it
- It is used when data is skewed
- It may be computed for an open-ended distribution

Median

- If Odd Number of Data Points are there
  Median Position is the $\frac{n+1}{2}$
- If Even Number of Data Points are there
  Median Position is the $\frac{n}{2}$ and $\frac{n+2}{2}$
  Median is the average of these two

Median for the Ungrouped Data

- Compute Median for these values
  - 2, 4, 3, 2, 8, 6, 9, 7, 6

Step 1: Arrange the Data 2, 2, 3, 4, 6, 6, 7, 8, 9

Step 2: Find whether Odd or Even Number of Observations

9 observations, so it is odd number of observations

Step 3: Find the Median Position and Value

Median position = $\frac{n+1}{2} = \frac{9+1}{2} = 5^{th}$ Which is 6

Median for the Ungrouped Data

- Compute Median for these values
  - 28, 18, 76, 56, 34, 25, 30, 45

Step 1: Arrange the Data 18, 25, 28, 30, 34, 45, 56, 76

Step 2: Find whether Odd or Even Number of Observations

8 observations, so it is Even number of observations

Step 3: Find the Median Position and Value

Median positions = $\frac{n}{2}$ and $\frac{n+2}{2}$ = $\frac{8}{2}, \frac{10}{2}$ = 4^{th} and 5^{th}

Median = average of 4^{th} and 5^{th} Values = (30+34)/2 = 32

Mean for the Grouped Data

$$\text{Mean} = \frac{\sum fXm}{\sum f}$$

Xm is the midpoint of Class Interval

Class Interval	Frequency	Xm	F Xm
1----5	2	3	6
6----10	4	8	32
11---15	6	13	78
16---20	5	18	90
21---25	3	23	69
26---30	2	28	56
Total	22		331

$$\text{Mean} = \frac{331}{22} = 15.04$$

Mode for the grouped Data

$$Mode = l + h\left\{\frac{fm - f1}{2fm - f1 - f2}\right\}$$

Modal Group: fm
Group with Highest Frequency

Class Interval	Class Boundary	Frequency
1----5	0.5 ----5.5	2
6----10	5.5 ----10.5	4
11---15	10.5 ----15.5	6
16---20	15.5 ----20.5	5
21---25	20.5 ----25.5	3
26---30	25.5 ----30.5	2
Total		22

Preceding Modal Group: f1

Proceeding Modal Group: f2

$$Mode = 10.5 + 5\left\{\frac{6 - 4}{2(6) - 4 - 5}\right\}$$

l is the lower class boundary for modal group
h is the height of Modal Class = 15.5 – 10.5 = 5

$$Mode = 13.83$$

Median For the Grouped Data

$$Median = l + \frac{h}{f}\left\{\frac{N}{2} - C\right\}$$

Now Finding the Values for Formula

Find Cumulative Frequency

Class Interval	Class Boundary	Frequency	Cumulative Freq < Upper CB
1----5	0.5 ----5.5	2	2
6----10	5.5 ----10.5	4	6
11---15	10.5 ----15.5	6	12
16---20	15.5 ----20.5	5	17
21---25	20.5 ----25.5	3	20
26---30	25.5 ----30.5	2	22
Total		22	

Find Median Group
Group which is containing the middle value
22/2 =11 is the middle value

Where is the 11th Value

7th to 12th are there in group Shown in Red

l = 10.5

f = 6

C = 6

h = 15.5 - 10.5

$$Median = 10.5 + \frac{5}{6}\left\{\frac{22}{2} - 6\right\} = 14.66$$

In the Same Way Quartile, Decile, Percentile Can Be Calculated

$$Qi = l + \frac{h}{f}\left\{\frac{iN}{4} - C\right\}$$

Only Quartile is Being Explained

i = 1, 2 and 3

$$Q1 = l + \frac{h}{f}\left\{\frac{N}{4} - C\right\}$$

$$Q2 = l + \frac{h}{f}\left\{\frac{2N}{4} - C\right\} = \text{Median}$$

$$Q3 = l + \frac{h}{f}\left\{\frac{3N}{4} - C\right\}$$

Quartiles

CB	F	C
0.5-----10.5	2	2
10.5----20.5	4	6
20.5----30.5	3	9
30.5----40.5	6	15
40.5----50.5	5	20
50.5----60.5	7	27
60.5----70.5	4	31
70.5----80.5	7	38
80.5----90.5	4	42
90.5---100.5	6	48
100.5---110.5	5	53
110.5----120.5	4	57
120.5----130.5	3	60
Total	60	

Find Quartile Positions through N

Q1 is at N/4
which is 15

$$Q1 = l + \frac{h}{f}\left\{\frac{N}{4} - C\right\}$$

$$Q1 = 30.5 + \frac{10}{6}\left\{\frac{60}{4} - 9\right\}$$

$$Q1 = 40.5$$

Q2 is at N/2
which is 30

$$Q2 = l + \frac{h}{f}\left\{\frac{N}{2} - C\right\}$$

$$Q2 = 60.5 + \frac{10}{4}\left\{\frac{60}{2} - 27\right\}$$

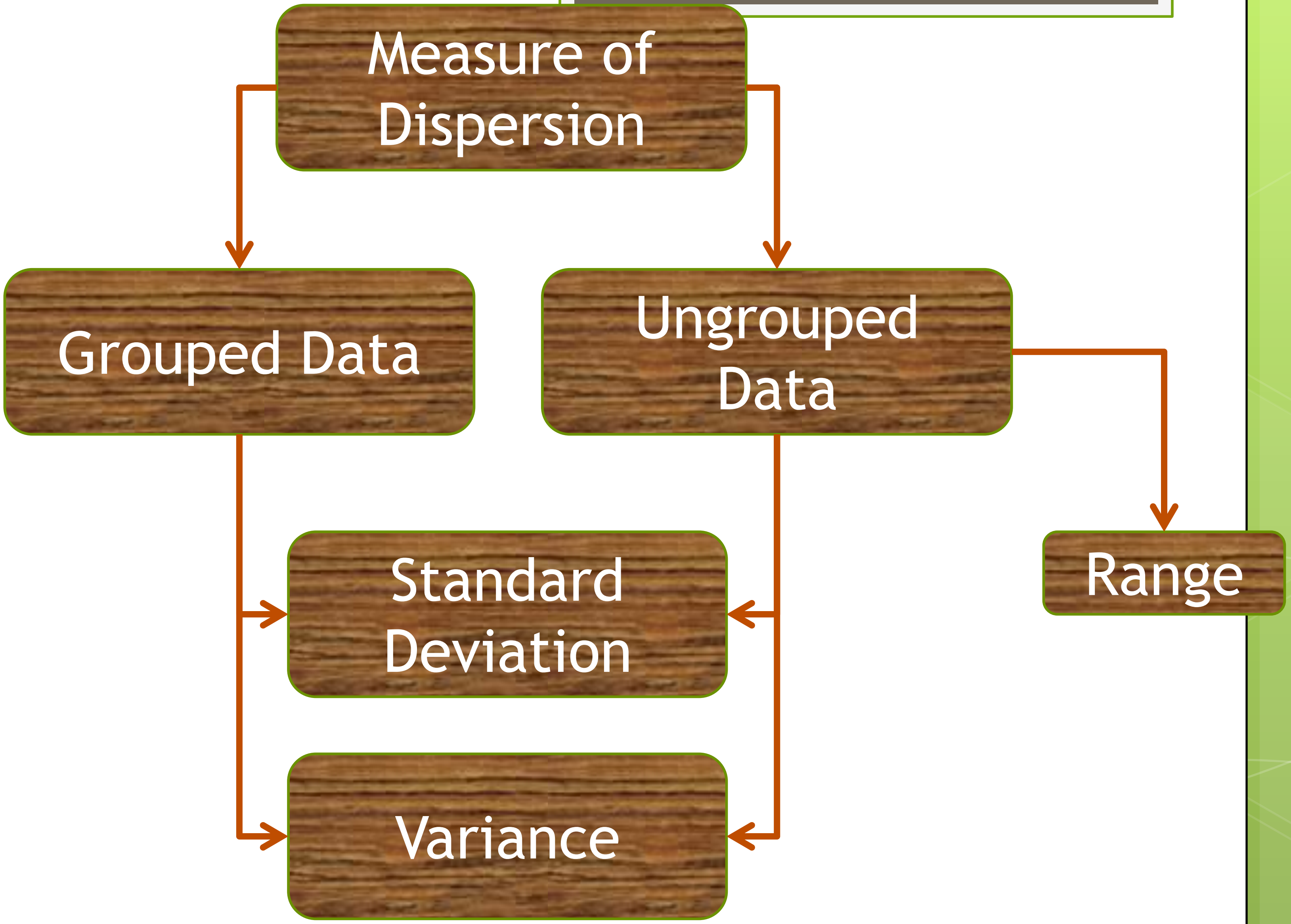
$$Q2 = 67.5$$

Q2 is at 3N/4
which is 45

$$Q3 = l + \frac{h}{f}\left\{\frac{3N}{4} - C\right\}$$

$$Q3 = 90.5 + \frac{10}{6}\left\{\frac{180}{4} - 42\right\}$$

$$Q3 = 95.5$$


Measure of Dispersion
Grouped Data
Ungrouped Data
Standard Deviation
Variance
Range

Range

- Easy to Measure and Compute
- It Emphasizes only the extreme values so it gives a very distorted picture

Range

- Find the Range for the given data
- Data Set: 25, 30, 40, 45
- Range = Max – Min = 45 – 25 = 20

Standard Deviation

- It is the most frequently used measure of dispersion
- It is a computed measure whose value is affected by every value
- Its value may be distorted by extreme values
- It can not be computed from an open-ended distribution

Standard Deviation and Variance for Ungrouped Data

Find the Standard Deviation for the given data
Data Set: 25, 30, 40, 45

Defining Formula

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

$$S = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Computing Formula

$$\sigma = \sqrt{\frac{\sum x^2}{N} - \mu^2}$$

$$S = \sqrt{\frac{\sum x^2}{n-1} - \frac{n\bar{x}^2}{n-1}}$$

Calculating Standard Deviation Through Defining Formula

	X	$(x-\mu)^2$	$(x-\mu)^2$
	25	$(25-35)^2$	100
	30	$(30 - 35)^2$	25
	40	$(40 - 35)^2$	25
	45	$(45 - 35)^2$	100
Total	140	250	250

Average	140/4
Average	35

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

$$\sigma = \sqrt{\frac{250}{4}}$$

σ = 7.905

S = 9.128

Calculate Standard Deviation Through Computing Formula

	X	X^2
	25	625
	30	900
	40	1600
	45	2025
Total	140	5150

Average	140/4
Average	35

$$\sigma = \sqrt{\frac{\sum x^2}{N} - \mu^2}$$

$$\sigma = \sqrt{\frac{5150}{4} - 35^2}$$

$$\sigma = 7.90564$$

$$S = \sqrt{\frac{\sum x^2}{n-1} - \frac{nx^2}{n-1}} = \sqrt{\frac{5150}{4-1} - \frac{4\,(35)^2}{4-1}} = 9.128$$

Standard Deviation For Grouped Data

Defining Formula

$$\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{N}} \qquad S = \sqrt{\frac{\sum f(x-x)^2}{n-1}}$$

Computing Formula

$$\sigma = \sqrt{\frac{\sum fx^2}{N} - \mu^2} \qquad S = \sqrt{\frac{\sum fx^2}{n-1} - \frac{(\sum f\, x)^2}{n(n-1)}}$$

Standard Deviation For the Grouped Data

Class Interval	f	Xm	F Xm	F Xm	Xm^2	fXm^2
1----5	2	3	6	6	9	18
6----10	4	8	32	32	64	256
11---15	6	13	78	78	169	1014
16---20	5	18	90	90	329	1620
21---25	3	23	69	69	529	1587
26---30	2	28	56	56	784	1568
Total	22		331	331		6063

$$\sigma = \sqrt{\frac{\sum f x^2}{N} - \mu^2} = \sqrt{\frac{6063}{22} - (15.04)^2} = 7.023 \qquad \mu = \frac{\sum f x}{\sum f}$$

$$S = \sqrt{\frac{\sum f x^2}{n-1} - \frac{(\sum f\, x)^2}{n(n-1)}} = \sqrt{\frac{6063}{22-1} - \frac{(331)^2}{22(22-1)}} = 7.181$$

Variance

- Variance = σ^2

Coefficient of Variation

- If means are same and Standard Deviation are different then the data with less standard deviation is more stable
- But if means are different and Standard deviation are different than how will compare the deviation?
- Suppose

$\bar{X}_1 = 50$
$\sigma_1 = 10$

$\bar{X}_2 = 70$
$\sigma_2 = 15$

CV1= 10/50
= 0.2
= 20%

CV2= 15/70
= 0.214
= 21.4

Quartiles

$Q1 = 40.5$

$Q2 = 67.5$

$Q3 = 95.5$

- Already discussed along Median
- Inter Quartile Range
  - $IQR = Q_3 - Q_1$
  - Quartile Deviation = $\frac{Q_3 - Q_1}{2}$
  - Find IQR and Quartile Deviation

Quartile Deviation

- It is very much similar to Range
- It is the Range of 50% middle value
- It is used for skewed data set
- It may be computed in open-ended distribution

Outliers

Observation which fall well outside the overall pattern of the data

Detecting Outliers

Lower Limit = Q_1 – 1.5 IQR

Upper Limit = Q_3 + 1.5 IQR

Values Lying outside the lower and upper limits are outliers

Reason for Outliers & Decisions

Measurement or any other Error → Remove Outlier

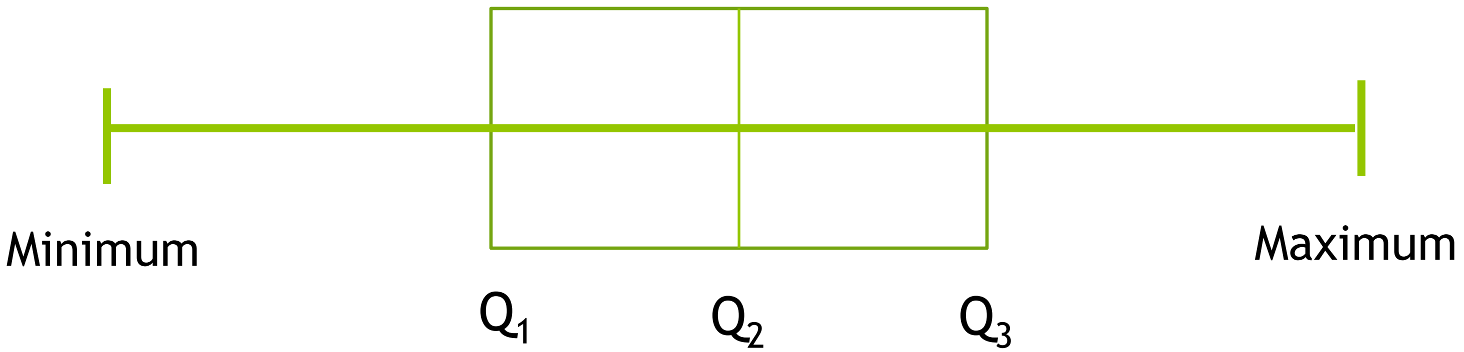
There is no error → Decision is difficult

Five Number Summary

- Outer Limits
  - Minimum
  - Maximum
- Inner Distribution
  - Q_1
  - Q_2
  - Q_3

Box Plot is a visual summary which is produced with the help of 5 Number Summary

Box and Whiskers Display



Box Plot

Box Plot

Box Plot whose outer range is made of Min & Max

Modified Box Plot

Box Plot showing outliers (Adjacent Values) show the limits of MBP

Construction of Box Plot

Box Plot

1. Determine five Number summary
2. Mark these values on X-axis
3. Connect the quartile to form box and then connect the box with minimum & Maximum

Modified Box Plot

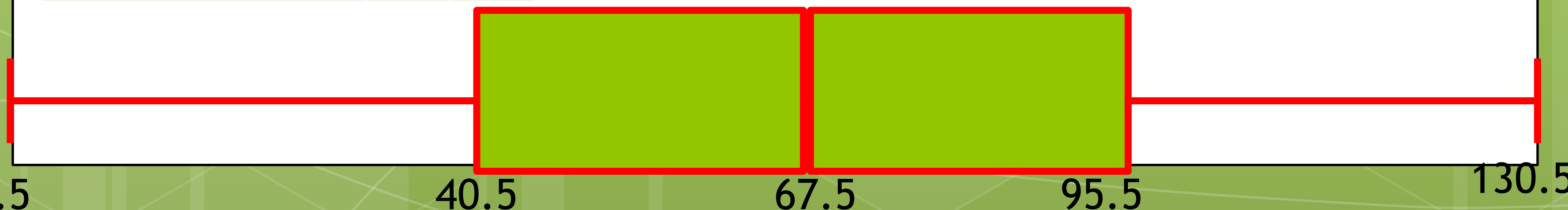
1. Determine the Quartiles
2. Determine the adjacent values & potential outliers
3. Mark the values on X-axis
4. Connect the values
5. Plot each potential outlier with a asterisk

Box Plot

CB	F	C
0.5-----10.5	2	2
10.5----20.5	4	6
20.5----30.5	3	9
30.5----40.5	6	15
40.5----50.5	5	20
50.5----60.5	7	27
60.5----70.5	4	31
70.5----80.5	7	38
80.5----90.5	4	42
90.5---100.5	6	48
100.5---110.5	5	53
110.5----120.5	4	57
120.5----130.5	3	60
Total	60	

$Q1 = 40.5$ $Q2 = 67.5$ $Q3 = 95.5$

$\text{Min} = 0.5$ $\text{Max} = 130.5$



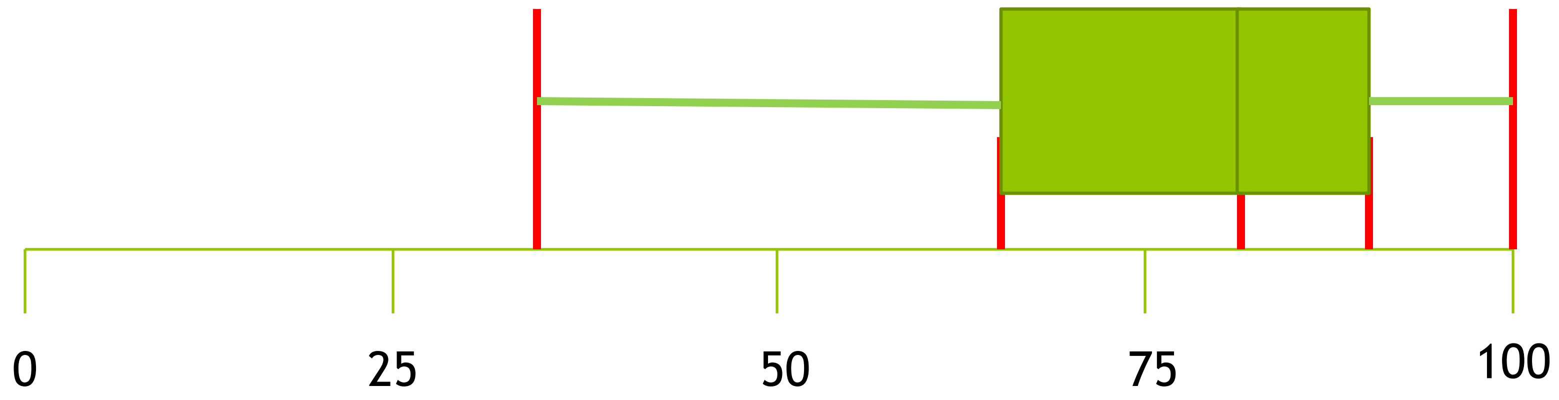
3.83

- Data
- 88, 85, 90, 81, 67, 82, 63, 96, 64, 39, 89, 100, 76, 75, 90, 70, 86, 34, 84, 96
- Arranged Data
- 34, 39, 63, 64, 67, 70, 75, 76, 81, 82, 84, 85, 86, 88, 89, 90, 90, 96, 96, 100

Min = 34 Max = 100 n = 20

Position Q_1 = 5.25	Position Q_2 = 10.5	Position Q_3 = 15.75
Q_1 = 75% of 5^{th} Value & 25% of 6^{th} Value	Q_2 = Avg(10^{th}, 11^{th})	Q_3 = 75% of 16^{th} and 25% of 15^{th} value
Q_1 = 67.375	Q_2 = 83	Q_2 = 89.75

Box Plot



Modified Box Plot

$IQR = Q_3 - Q_1$

$IQR = 89.75 - 67.37$

$IQR = 22.38$

$UV = Q_3 + 1.5\ IQR$

$UV = 89.75 + 1.5\ (22.38)$

$UV = 123.2$

$LV = Q_1 - 1.5\ IQR$

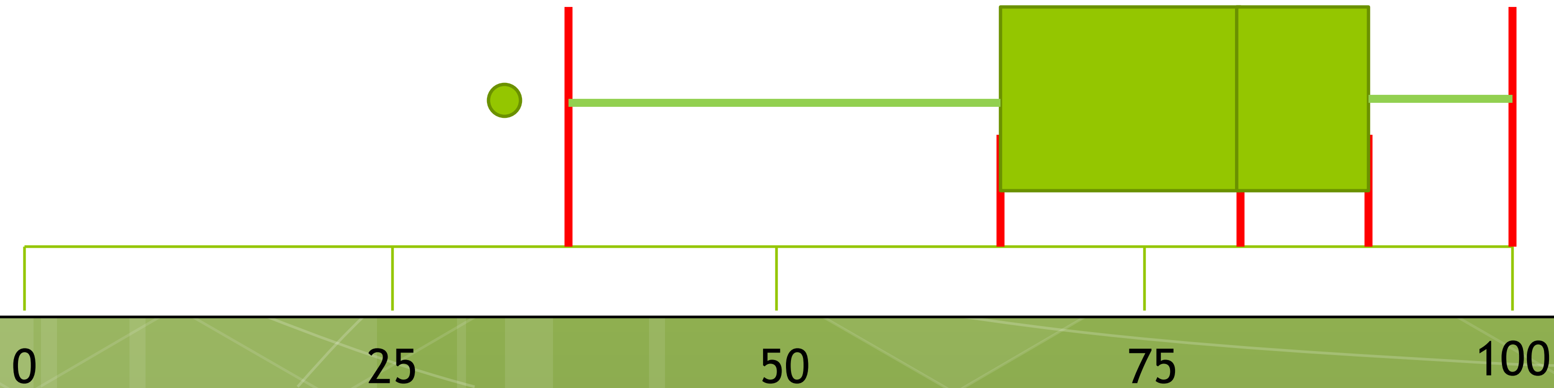
$LV = 67.37 - 1.5\ (22.38)$

$LV = 34$

Lower Adjacent Value = 39

Upper Adjacent Value = 100

Outlier = 34



Skewness

- Measures asymmetry of data
  - Positive or right skewed: Longer right tail
  - Negative or left skewed: Longer left tail

Let $x_1, x_2, \ldots x_n$ be n observations. Then,

$$\text{Skewness} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

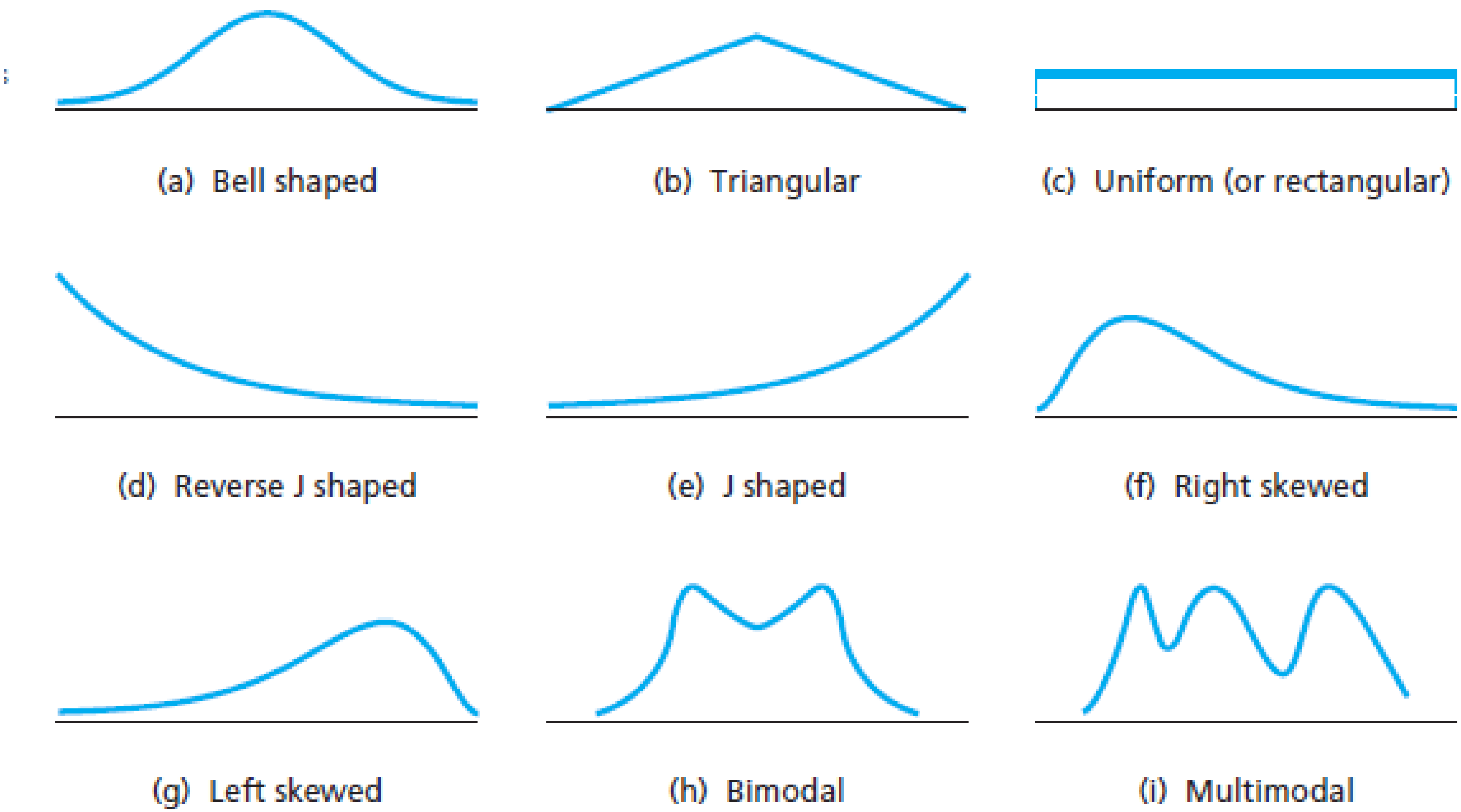
Kurtosis

- Measures peakedness of the distribution of data. The kurtosis of normal distribution is 0.

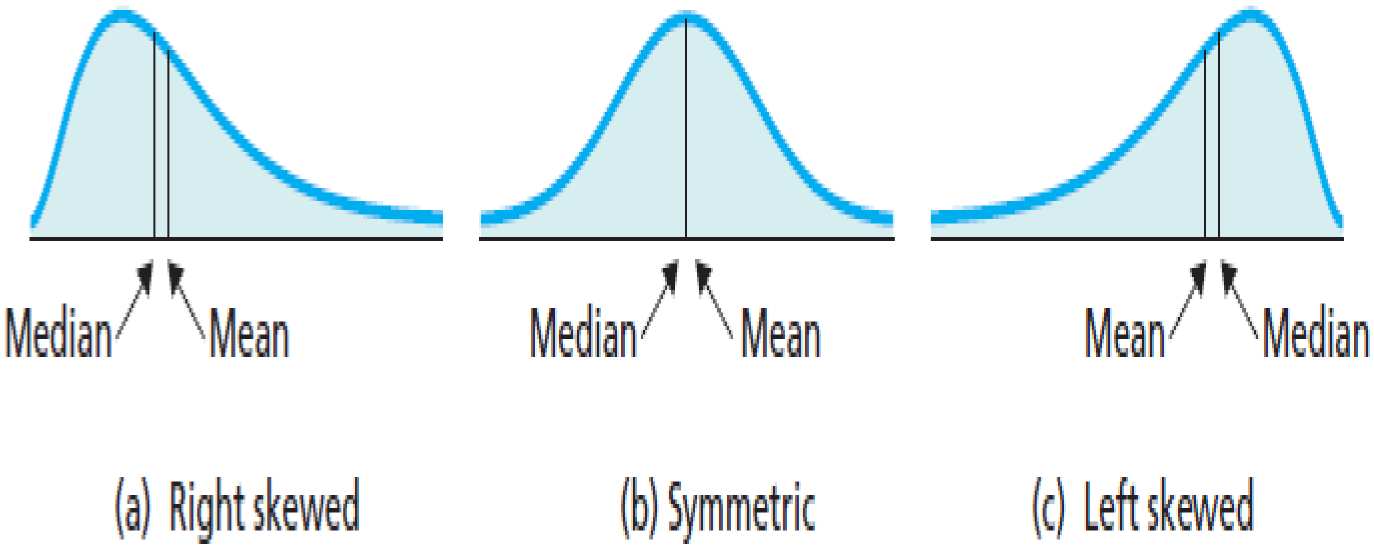
Let $x_1, x_2, \ldots x_n$ be n observations. Then,

$$\text{Kurtosis} = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$

Common distribution shape



Comparison of Mean,Median,Mode



Skewness with Quartile & BoxPlot