Universiti Malaysia PAHANG
Engineering • Technology • Creativity

DUM 2413 STATISTICS & PROBABILITY

CHAPTER 2
DESCRIPTIVE STATISTICS

PREPARED BY:
DR. CHUAN ZUN LIANG; DR. NORATIKAH ABU; DR. SITI ZANARIAH SATARI
FACULTY OF INDUSTRIAL SCIENCES & TECHNOLOGY
chuanzl@ump.edu.my; atikahabu@ump.edu.my; zanariah@ump.edu.my

CONTENT

EXPECTED OUTCOMES

- Able to organise and represent qualitative and quantitative data using an appropriate analysis tool
- Able to differentiate between the grouped and ungrouped data
- Able to summarise the data using non-graphical and graphical exploratory data analysis tools
- Able to apply Chebyshev's Theorem in applications

2.3
SUMMARY STATISTICS (DATA DESCRIPTION)

2.4
EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an approach using statistical tools to analyse the data sets in order to summarise or describe their important characteristics.

Exploratory Data Analysis (EDA)

- **Non-graphical**
  - Measures of central tendency
  - Measures of position
  - Measure of variation
- **Graphical**
  - Histogram
  - Stem and Leaf Plot
  - Boxplots
  - Time series graph

Measures of central tendency, histogram, stem and leaf plot, and box plots can be used to identify the shape of distribution.

MEASURES OF CENTRAL TENDENCY (UNGROUPED DATA)

MEASURES OF CENTRAL TENDENCY

MEAN

Population:

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

N = Population size

Sample:

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

n = Sample size

MEDIAN

If n is odd:

$$\text{Median} = x_{\left(\frac{n}{2}\right)}$$

If n is even:

$$\text{Median} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}\right)+1}}{2}$$

MODE

The mode is the value which has the highest frequency in a data set.

MIDRANGE

$$\text{Midrange} = \frac{x_{\min} + x_{\max}}{2}$$

where

$x_{\min}$ = lowest value (minimum)

$x_{\max}$ = highest value (maximum)

EXAMPLE 2.7

A sample of 10 students in UMP showed the following credit hours taken during the first year of this program.

17 18 18 18 19 20 21 21 22 24

Compute the mean, median, mode, and midrange.

SOLUTION

Mean = 19.8

Median =19.5

Mode =18.0

Midrange =20.5

MEASURES OF CENTRAL TENDENCY
(GROUPED DATA)

MEASURES OF CENTRAL TENDENCY

MEAN

$$\text{Mean} = \frac{\sum f_i x_i}{\sum f_i}$$

MEDIAN

$$\text{Median} = L + \left(\frac{\left({}^{n}\!/_{2} \right) - f_L}{f_{\text{median}}} \right) * C$$

MODE

$$\text{Mode} = L_{\text{mode}} + \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right) * C_{\text{mode}}$$

- x_i — The midpoint of the ith class
- f_i — The corresponding frequency
- L_{median} — Lower boundary of the median class
- f_L — Cumulative frequency until point L
- f_{median} — Frequency of the class median
- C — Size of median class
- L_{mode} — Lower boundary of the modal class
- C_{mode} — Size of modal class
- λ_1 — Difference between the frequency of the modal class and the previous class
- λ_2 — Difference between the frequency of the modal class and the next class

EXAMPLE 2.8

Calculate the mean, mode and median of the following data.

Height (cm)	Frequency	Midpoint	Cumulative Frequency
$120 \le x < 125$	1	122.5	1
$125 \le x < 130$	3	127.5	4
$130 \le x < 135$	6	132.5	10
$135 \le x < 140$	12	137.5	22
$140 \le x < 145$	17	142.5	39
$145 \le x < 150$	18	147.5	57
$150 \le x < 155$	15	152.5	72
$155 \le x < 160$	5	157.5	77
$160 \le x < 165$	2	162.5	79
$165 \le x < 170$	1	167.5	80

SOLUTION

$$\text{Median} = \left(\frac{80}{2}\right)\text{th} = 40\text{th}$$

Class boundary=145-150

Mode:

Class boundary=145-150

EXAMPLE 2.8-CONTINUE

$$\text{Mean} = \frac{\sum f_i x_i}{\sum f_i} = 144.9375$$

$$\text{Median} = L + \left(\frac{\left(n/2 \right) - f_L}{f_m} \right) * C$$

$$L = 145;\ f_L = 39;\ f_m = 18;\ C = 150 - 145 = 5$$

$$= 145 + \left(\frac{\left(80/2 \right) - 39}{18} \right) * 5$$

=145.2778

$$\text{Mode} = L_{\text{mode}} + \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right) * C$$

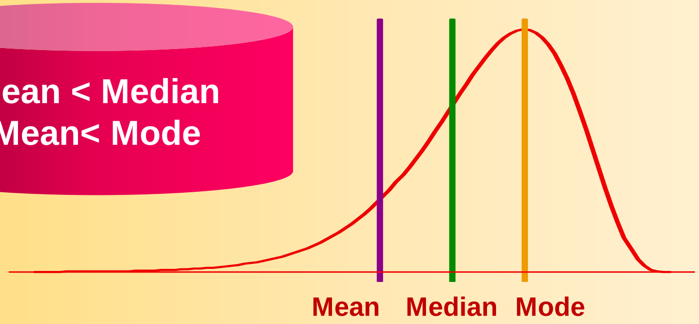
$$L_{\text{mode}} = 145;\ \lambda_1 = 18 - 17 = 1;\ \lambda_2 = 18 - 15 = 3$$

$$= 145 + \left(\frac{1}{1+3} \right) * 5$$

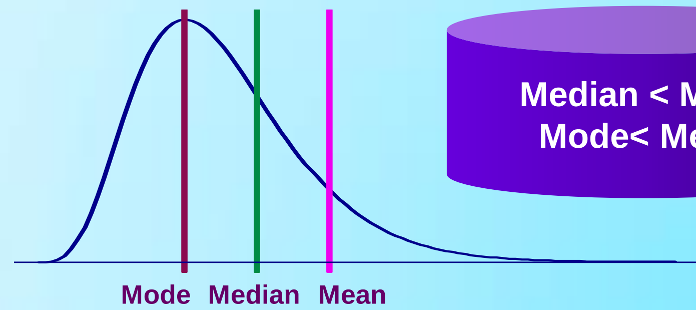
=146.2500

IDENTIFY THE SHAPE OF DISTRIBUTION USING MEASURES OF CENTRAL TENDENCY

Mean < Median
Mean< Mode



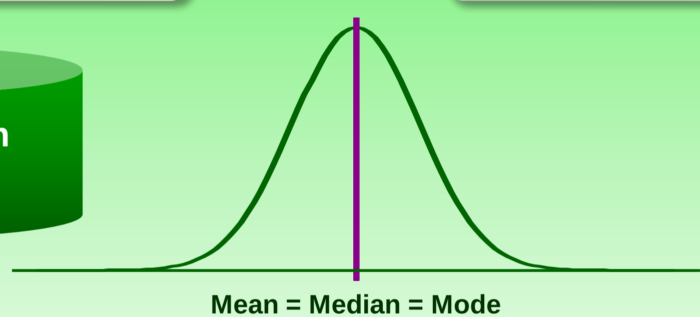
LEFT-SKEWED DISTRIBUTION
Mean < Median < Mode



Median < Mean
Mode< Mean

RIGHT-SKEWED DISTRIBUTION
Mode < Median < Mean

Mean = Median
Mean=Mode



SYMMETRICAL DISTRIBUTION
Mean = Median = Mode

EXAMPLE 2.9

Determine the type of distribution of the following data

(i)

11.6	12.6	12.7	12.8	13.3	13.3	13.6	13.7	13.8	11.4

(ii) Mean=Mode=Median=1

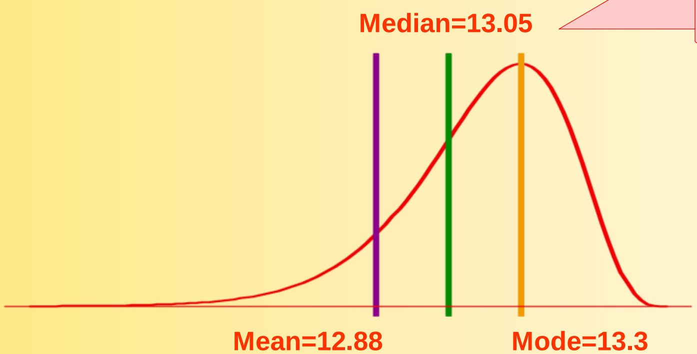
(iii) Mean=25, Mode=13, Median=17

(iv) Mean=5, Mode=73, Median=17

EXAMPLE 2.9-CONTINUE

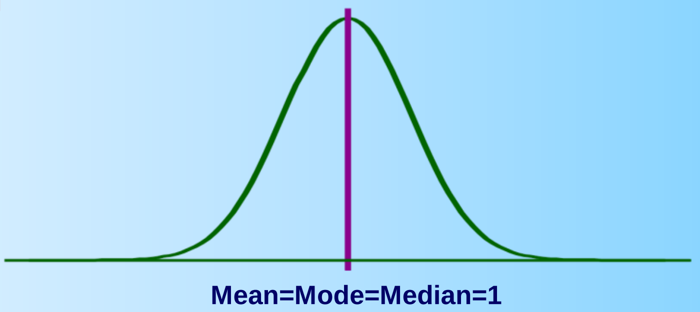
SOLUTION

(i)



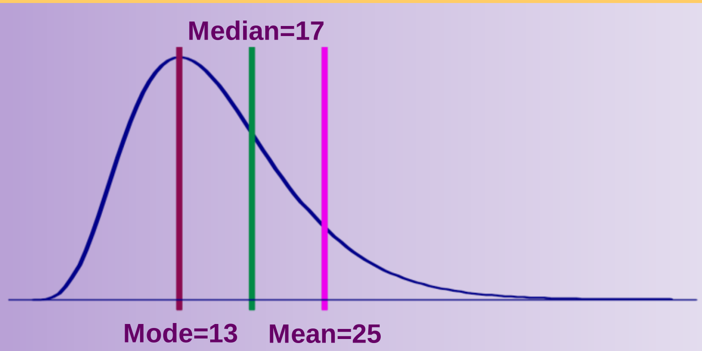
Left-skewed Distribution
Reason: Mean < Median < Mode

(ii)



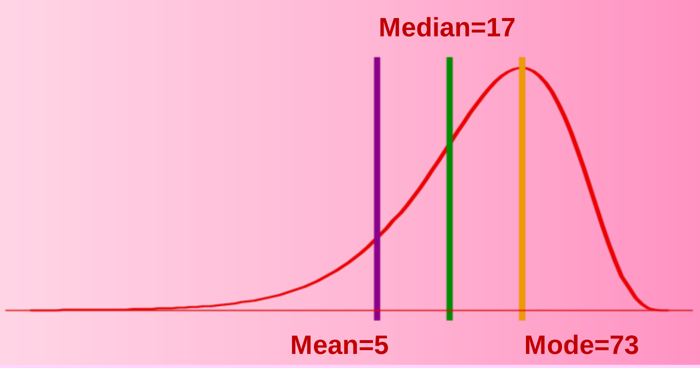
Symmetrical Distribution
Reason: Mean = Median = Mode=1

(iii)



Right-skewed Distribution
Reason: Mode < Median < Mean

(iv)



Left-skewed Distribution
Reason: Mean < Median < Mode

EXAMPLE 2.10

The table shows the speed of the tracks passing through a hilling road.

Speed	Frequency	Class Boundary	Midpoint	Cumulative frequency
56-58	4	55.5-58.5	57	4
59-61	12	58.5-61.5	60	16
62-64	28	61.5-64.5	63	44
65-67	58	64.5-67.5	66	102
68-70	44	67.5-70.5	69	146
71-73	18	70.5-73.5	72	164
74-76	10	73.5-76.5	75	174

Find the mean, mode and median. Hence, identify the shape of distribution based on measures on central tendency.

SOLUTION

$$\text{Median} = \left(\frac{174}{2}\right)\text{th=87th}$$

Class boundary=64.5-67.5

Mode:

Class boundary=64.5-67.5

EXAMPLE 2.10-CONTINUE

$$\text{Mean} = \frac{\sum f_i x_i}{\sum f_i} = 66.7931$$

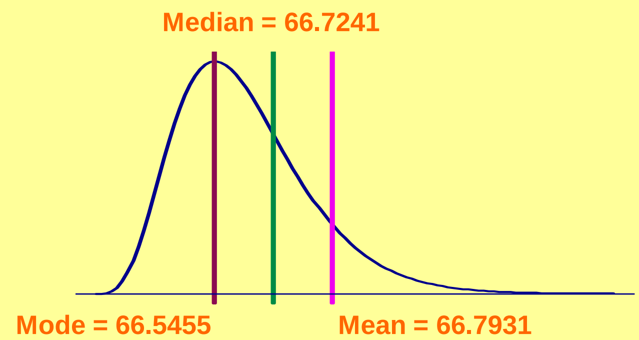
$$\text{Median} = L + \left(\frac{\left({n}/{2} \right) - f_L}{f_m} \right) * C$$

$$L = 64.5;\ f_L = 44;\ f_m = 58;\ C = 67.5 - 64.5 = 3$$

$$= 64.5 + \left(\frac{\left({174}/{2} \right) - 44}{58} \right) * 3$$

$$= 66.7241$$

Mode < Median < Mean
Right-skewed Distribution



$$\text{Mode} = L_{\text{mode}} + \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right) * C$$

$$L_{\text{mode}} = 64.5;\ \lambda_1 = 58 - 28 = 30;\ \lambda_2 = 58 - 44 = 14$$

$$= 64.5 + \left(\frac{30}{30 + 14} \right) * 3$$

$$= 66.5455$$

MEASURES OF VARIATION (UNGROUPED DATA)

MEASURES OF VARIATION

RANGE

$$\text{Range} = x_{\max} - x_{\min}$$

where

$x_{\min}$ = lowest value (minimum)

$x_{\max}$ = highest value (maximum)

VARIANCE

Population

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)}{N}$$

Sample

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})}{n-1}$$

STANDARD DEVIATION

Population

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)}{N}}$$

Sample

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})}{n-1}}$$

NOTE:

- Variance, (σ^2/s^2) AND Standard Deviation (σ/s), can be used to determine the spread and consistency of the data. For example, $\sigma_A > \sigma_B \;/\; s_A > s_B$, this means sample A is more dispersed/variable compares to sample B.

EXAMPLE 2.11

Note:
Similar Terms
- Less consistent
- More spread
- More variable
- Large variation
- Less precise
- Less stable

Which of the following set of sample data is more dispersed.

(i)

A	4.2	6.7	7.3	7.5	8.0	8.5	8.7	8.8	9.2	9.3
B	9.6	9.7	9.8	9.9	10.1	10.2	11.0	11.0	11.0	11.1

(ii)

Method A	79	73	78	76	80	75	82	70	77
Method B	80	85	78	79	75	73	70	60	65

SOLUTION

(i)

Since $s_A = 1.5296 > s_B = 0.6150$, therefore data A is more variable compared to B.

(ii)

Since $s_A = 3.7642 < s_B = 7.8493$, therefore data of Method B is more variable compared to Method A.

MEASURES OF VARIATION (GROUPED DATA)

MEASURES OF VARIATON

VARIANCE

POPULATION

$$\sigma^2 = \frac{\sum f_i (x_i - \mu)^2}{\sum f_i} = \frac{\sum f_i x_i^2}{\sum f_i} - \left(\frac{\sum f_i x_i}{\sum f_i}\right)^2$$

SAMPLE

$$s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i - 1} = \frac{\sum f_i x_i^2}{\sum f_i - 1} - \left(\frac{\sum f_i x_i}{\sum f_i - 1}\right)^2$$

STANDARD DEVIATION

POPULATION

$$\sigma = \sqrt{\frac{\sum f_i (x_i - \mu)^2}{\sum f_i}} = \sqrt{\frac{\sum f_i x_i^2}{\sum f_i} - \left(\frac{\sum f_i x_i}{\sum f_i}\right)^2}$$

SAMPLE

$$s = \sqrt{\frac{\sum f_i (x_i - \bar{x})^2}{\sum f_i - 1}} = \sqrt{\frac{\sum f_i x_i^2}{\sum f_i - 1} - \left(\frac{\sum f_i x_i}{\sum f_i - 1}\right)^2}$$

x_i **: The midpoint of the ith class;** f_i **: The corresponding frequency**

EXAMPLE 2.12

The table below shows the lifetime (hours) of 112 light bulbs. Find the sample mean, standard deviation and variance of the lifetime of these light bulbs.

Lifetime (hours)	Number of bulbs
800 ≤ x < 1000	5
1000 ≤ x < 1200	17
1200 ≤ x < 1400	26
1400 ≤ x < 1600	38
1600 ≤ x < 1800	13
1800 ≤ x < 2000	8
2000 ≤ x < 2200	5

SOLUTION

Lifetime (hours)	Midpoint (x_i)	Number of bulbs (f_i)
800 ≤ x < 1000	900	5
1000 ≤ x < 1200	1100	17
1200 ≤ x < 1400	1300	26
1400 ≤ x < 1600	1500	38
1600 ≤ x < 1800	1700	13
1800 ≤ x < 2000	1900	8
2000 ≤ x < 2200	2100	5

$$\bar{x} = 1444.6429$$
$$s = 281.8342$$
$$s^2 = 79430.5019$$

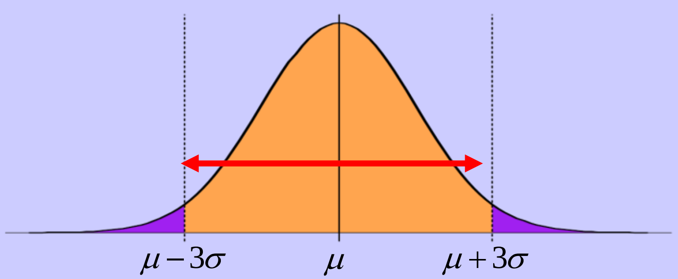
THE EMPIRICAL RULE AND CHEBYSHEV'S THEOREM WITH A BELL-SHAPED DISTRIBUTION

CHEBYSHEV'S THEOREM

The proportion of any distribution that lies within k standard deviation of the mean is at least $1-\dfrac{1}{k^2}$, where is any positive number greater than 1.

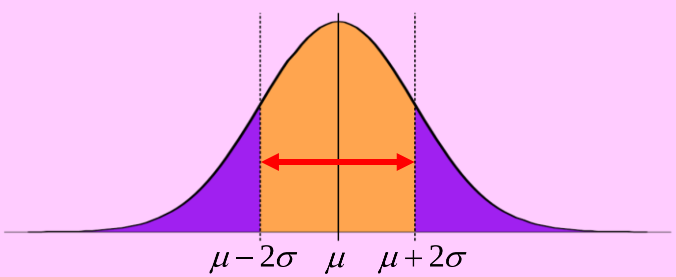
1821-1894

EXAMPLE: $k = 3$



Empirical Rule	Chebyshev's Theorem
0.9973 (Approximately 99.73% of the data)	0.8889 (At least 88.89% of the data)

EXAMPLE: $k = 2$

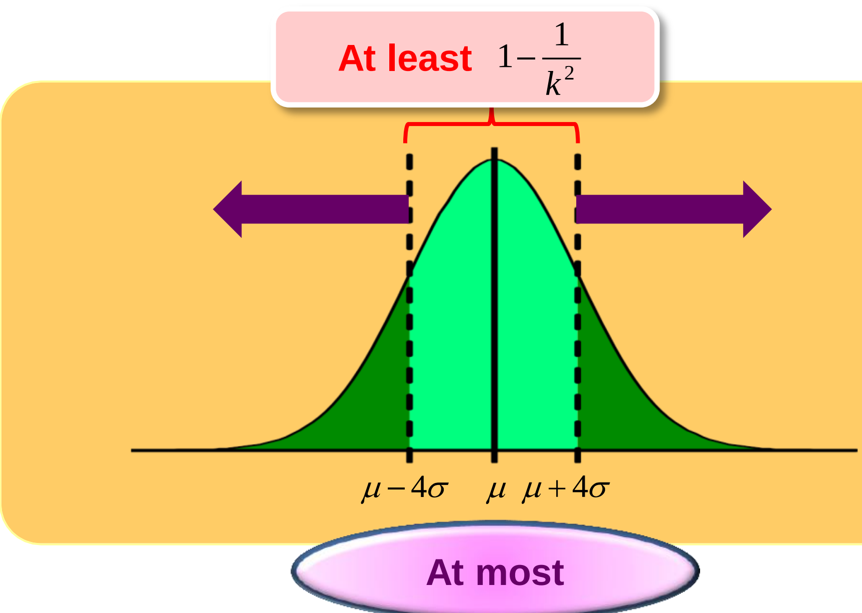


Empirical Rule	Chebyshev's Theorem
0.9545 (Approximately 95.45% of the data)	0.7500 (At least 75% of the data)

EXAMPLE 2.13

Chebyshev's theorem stated the proportion of any distribution that lies within k standard deviation of the mean. For instance, when k=2, it can interpret as "at least 75% of the data fall within 2 standard deviation of the mean. This also equivalent to state that "at most, 25% will be more than 2 standard deviations away from the mean." At most, what percentage of a distribution will be 4 or more standard deviations from the mean?

SOLUTION



In general, we know that the *total area under curve* is equal to *1*.

Therefore,
the percentage of a distribution will *be 4 or more standard deviation from mean*:

$$=1-\left(1-\frac{1}{k^2}\right)$$

$$=1-\left(1-\frac{1}{4^2}\right)$$

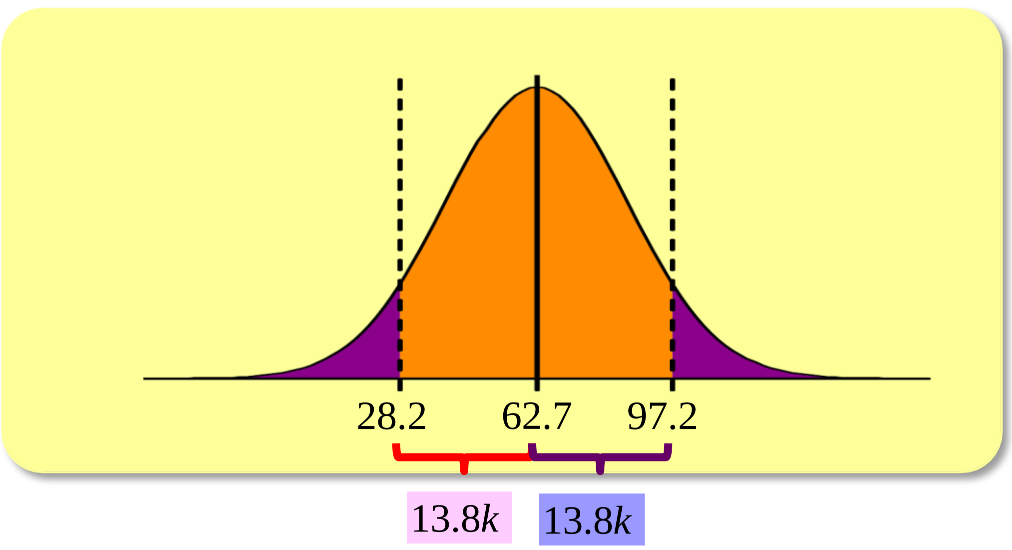
$$=0.0625$$

At most 6.25% of a distribution will be 4 or more standard deviations from the mean.

EXAMPLE 2.14

A lecturer conducted an analysis regarding the students' performance in the subject of DUM 2413 Statistics & Probability. The analysis results showed that the average marks of this subject is 62.7%, with a standard deviation of 13.8%. According to Chebyshev's theorem, at least what percent of the students' performance in the subject of DUM 2413 Statistics & Probabilityis between 28.2% and 97.2%.

SOLUTION



$$13.8k = 97.2 - 62.7$$

$$k = 2.5$$

Chebyshev's Theorem

$$= 1 - \frac{1}{k^2}$$

$$= 0.84$$

Therefore, at least 84% of student' performance in subject of DUM 2413 Statistics & Probability is between 28.2% and 97.2%.

MEASURES OF POSITION (UNGROUPED DATA)

MEASURES OF POSITION

QUARTILES

Split data into **4** equal parts

$$Q_i = x_c = x_{\frac{in}{4}}$$

DECILES

Split data into **10** equal parts

$$D_i = x_c = x_{\frac{in}{10}}$$

PERCENTILES

Split data into **100** equal parts

$$P_i = x_c = x_{\frac{in}{100}}$$

If c is not a whole number, round it up to the next whole number.

If c a whole number, then use $Q_i, D_i, P_i \approx \dfrac{x_c + x_{c+1}}{2}$.

EXAMPLE 2.15

A manufacturing company has 550 operators. A random sample of 11 operators is randomly selected and the numbers of sick leave (in days) last year for these operators are recorded as shown below.

3	6	1	4	4	2	5	1	7	2	4

(i) Calculate the first, second and third quartile.
(Note: second quartile equivalent to median)

1	1	2	2	3	4	4	4	5	6	7

$$Q_1 = x_{\frac{1(11)}{4}} = x_{2.75} \approx x_3 = 2; Q_2 = x_{\frac{2(11)}{4}} = x_{5.5} \approx x_6 = 4; Q_3 = x_{\frac{3(11)}{4}} = x_{8.25} \approx x_9 = 5$$

(ii) Calculate the 25%, 50% and 75% percentile.

$$P_{25} = x_{\frac{25(11)}{100}} = x_{2.75} \approx x_3 = 2; P_{50} = x_{\frac{50(11)}{100}} = x_{5.5} \approx x_6 = 4; Q_{75} = x_{\frac{75(11)}{100}} = x_{8.25} \approx x_9 = 5$$

EXAMPLE 2.16

1. Given

9	2	1	4	3	7	5	4	6

1	2	3	4	4	5	6	7	9

(i) Find the value correspond to 4th deciles.
(ii) Find the value correspond to 3rd quartiles.

(i) $D_4 = x_{\frac{4(9)}{10}} = x_{3.6} \approx x_4 = 4$

(ii) $Q_4 = x_{\frac{3(9)}{4}} = x_{6.75} \approx x_7 = 6$

2. Given

9	22	11	14	13	3	7	15	18	16

(i) Find the value correspond to 20th percentiles.
(ii) Find the value correspond to 7th deciles.

3	7	9	11	13	14	15	16	18	22

(i) $P_{20} = x_{\frac{20(10)}{100}} = x_2 \approx \frac{x_2 + x_3}{2} = 8$

b) $D_7 = x_{\frac{7(10)}{10}} = x_7 \approx \frac{x_7 + x_8}{2} = 15.5$

MEASURES OF POSITION
(GROUPED DATA)

MEASURES OF POSITION

QUARTILES

$$Q_i = L_i + \left(\frac{\left(in/4 \right) - f_L}{f_i} \right) * C; \; i = 1, 2, 3$$

DECILES

$$D_i = L_i + \left(\frac{\left(in/10 \right) - f_L}{f_i} \right) * C; \; i = 1, 2, \ldots, 10$$

PENCENTILES

$$P_i = L_i + \left(\frac{\left(in/100 \right) - f_L}{f_i} \right) * C; \; i = 1, 2, \ldots, 99$$

$$n = \sum f_i$$

- L_i: Lower boundary of the class Q_i lies
- f_L: Cumulative frequency until point L_i
- f_i: Frequency of the class where Q_i lies
- C: Size of the class where Q_i lies

EXAMPLE 2.17

The frequency distribution depicted the times taken for 70 workers to complete a single challenging task assigned by their manager.

Time (min)	Number of workers
20 ≤ x < 25	10
25 ≤ x < 30	8
30 ≤ x < 35	9
35 ≤ x < 40	18
40 ≤ x < 45	21
45 ≤ x < 50	4

Determine Q_1, D_3 **and** P_{75}.

SOLUTION

Time (min)	Number of workers	Cumulative Frequency
20 ≤ x < 25	10	10
25 ≤ x < 30	8	18
30 ≤ x < 35	9	27
35 ≤ x < 40	18	45
40 ≤ x < 45	21	66
45 ≤ x < 50	4	70

EXAMPLE 2.17-CONTINUE

$$Q_i = L_i + \left(\frac{\left({}^{in}\!/_{4} \right) - f_L}{f_i} \right) * C$$

$L_i = 25;\ n = 70;\ f_L = 10;\ f_i = 8;\ C = 30 - 25 = 5$

$$Q_1 = 25 + \left(\frac{{}^{(1*70)}\!/_{4} - 10}{8} \right) * 5$$

$$= 29.6875$$

$$Q_1 = \left(\frac{1}{4} * 70 \right) \text{th} = 17.5\text{th} \approx 18\text{th}$$

Class boundary = 25-30

$$D_i = L_i + \left(\frac{\left({}^{in}\!/_{10} \right) - f_L}{f_i} \right) * C$$

$L_i = 30;\ n = 70;\ f_L = 18;\ f_i = 9;\ C = 35 - 30 = 5$

$$D_3 = 30 + \left(\frac{{}^{(3*70)}\!/_{10} - 18}{9} \right) * 5$$

$$= 31.6667$$

$$D_3 = \left(\frac{3}{10} * 70 \right) \text{th} = 21\text{th}$$

Class boundary $= 30 - 35$

$$D_i = L_i + \left(\frac{\left({}^{in}\!/_{10} \right) - f_L}{f_i} \right) * C$$

$L_i = 40;\ n = 70;\ f_L = 45;\ f_i = 21;\ C = 45 - 40 = 5$

$$P_{75} = 40 + \left(\frac{{}^{(75*70)}\!/_{100} - 45}{21} \right) * 5$$

$$= 41.7857$$

$$P_{75} = \left(\frac{75}{100} * 70 \right) \text{th} = 52.5\text{th} \approx 53\text{th}$$

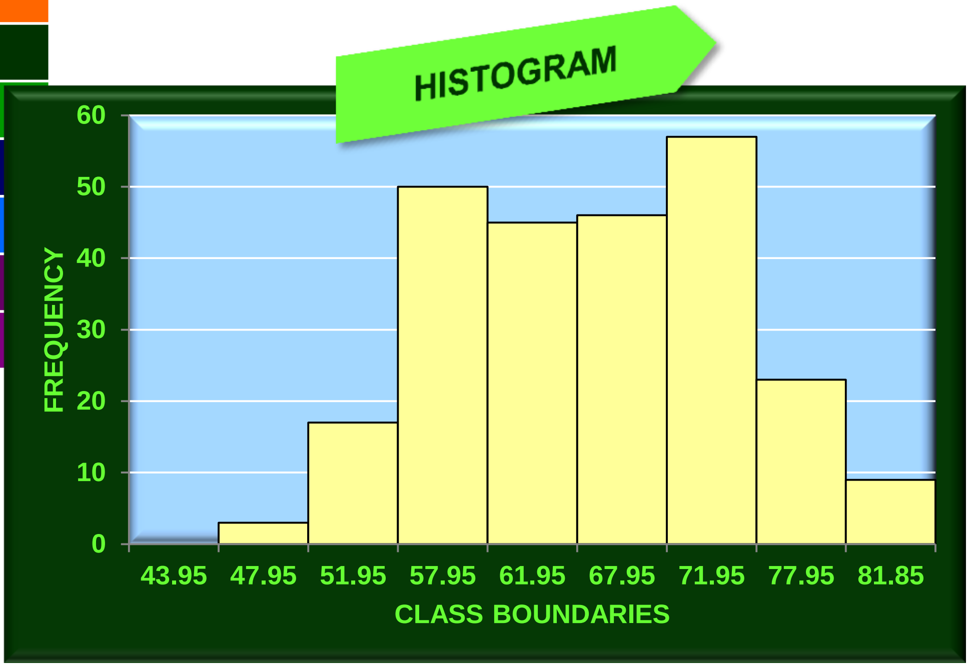
Class boundary $= 40 - 50$

HISTOGRAM

Histogram is a **bar graph** that **represents a frequency distribution** of a quantitative variable.

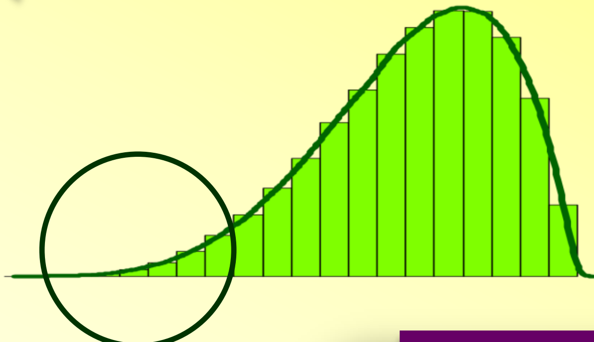
EXAMPLE: The weight of 250 sacks of durian (in kg)

Weight (kg)	Number of sacks of durian
44.0-47.9	3
48.0-51.9	17
52.0-57.9	50
58.0-61.9	45
62.0-67.9	46
68.0-71.9	57
72.0-77.9	23
78.0-81.9	9

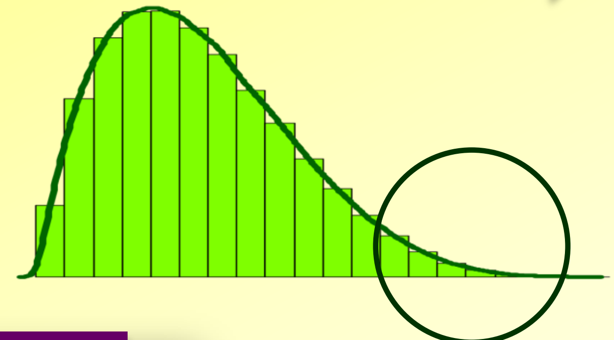


HISTOGRAM
(IDENTIFY THE SHAPE OF DISTRIBUTION)-THREE IMPORTANT SHAPES

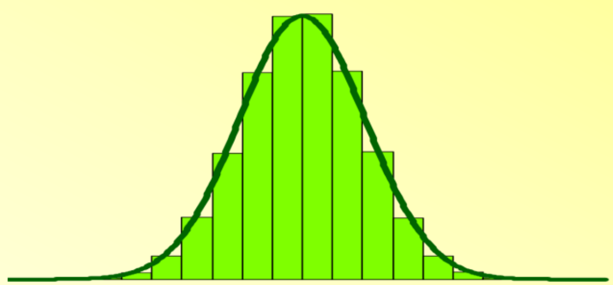
Left-skewed Distribution



Right-skewed Distribution



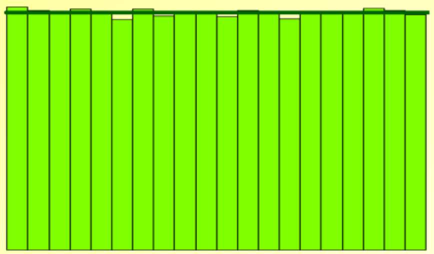
Symmetrical Distribution



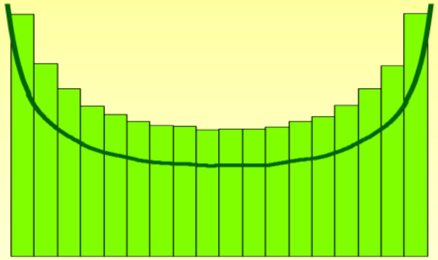
Identify the direction of skewed based on the "TAIL"

HISTOGRAM
(IDENTIFY THE SHAPE OF DISTRIBUTION)-THREE IMPORTANT SHAPES

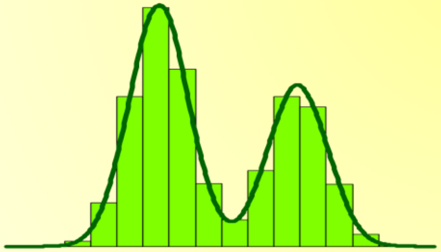
UNIFORM



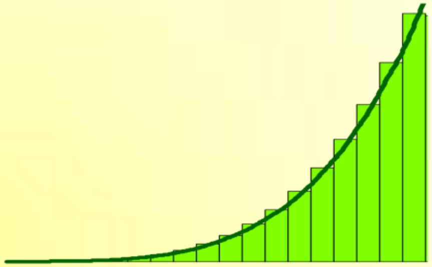
U-SHAPE



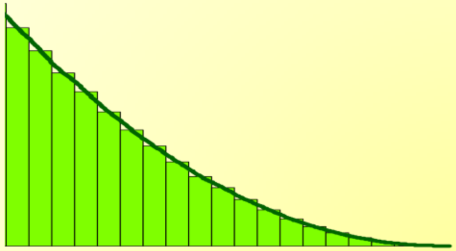
BIMODAL



J-SHAPE



REVERSE J



EXAMPLE 2.18

The traffic police observed the speeds of 55 cars passing through an accident crime scene in a village using radar device.

27	23	22	38	43	24	35	26	28	18	20
25	23	22	52	31	30	41	45	29	27	43
29	28	27	25	29	28	24	37	28	29	18
26	33	25	27	25	34	32	36	22	32	33
21	23	24	18	48	23	16	38	26	21	23

(i) Classify these data into a grouped frequency distribution using class boundaries 12-18, 18-24, …, 48-54.

(ii) Find the class width.

(iii) For the class 24-30, find the class midpoint, the lower and upper class boundaries.

(iv) Construct a frequency histogram of these data. Then, identify the shape of distribution.

EXAMPLE 2.18-CONTINUE

SOLUTION

(i)

Class limits	Frequency
12-18	1
18-24	14
24-30	22
30-36	8
36-42	5
42-48	3
48-54	2

(ii) Class width=18-12=6

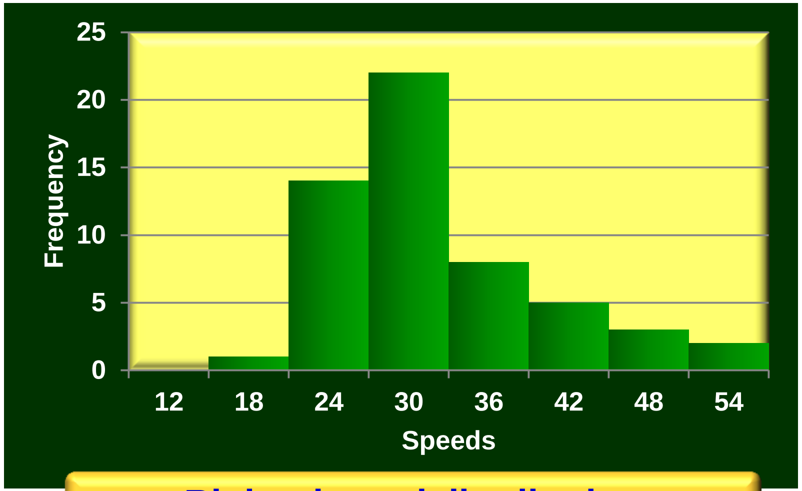
(iii)

$$\text{Midpoint}=\frac{24+30}{2}=27$$

Lower class boundaries = 24

Upper class boundaries = 30

(iv)



Right-skewed distribution

STEM AND LEAF PLOT

A stem and leaf plot displays the data of a sample using the actual digits that make up the data values.

EXAMPLE:
The response times of 30 integrated circuits (in picoseconds)

4.6	4.0	3.7	4.1	4.1	5.6
4.5	6.0	6.0	3.4	3.4	4.6
3.7	4.2	4.6	4.7	4.1	3.7
3.4	3.3	3.7	4.1	4.5	4.6
4.4	4.8	4.3	4.4	5.1	3.9

Stem	Leaf	Key: 3|0 means 3.0
3	3 4 4 4 7 7 7 7 9	
4	0 1 1 1 1 2 3 4 4 5 5 6 6 6 6	
5	1 6	
6	0 0	

Stem and leaf plot

EXAMPLE:
The heat rates for two different groups

GROUP 1

115	115	117	126	127	127	128	128	129	129
129	129	130	134	134	136	136	140	142	144

GROUP 2

125	125	126	134	136	138	138	142	143	146
146	147	148	148	153	155	155	157	162	164

Group 1	Stem	Group 2
7 5 5	11	
9 9 9 9 8 8 7 7 6	12	5 5 6
6 6 4 4 0	13	4 6 8 8
4 2 0	14	2 3 6 6 7 8 8
	15	3 5 5 7
Key: 11|5 means 115	16	2 4

Back-to-back stem and leaf plot

EXAMPLE 2.19

The following data show the 22 final examination marks for DUM 2413 Statistics & Probability course.

44	52	70	75	53	44	52	66	57	79	83
68	94	66	59	45	69	48	53	80	95	44

Construct the stem-and-leaf plot for the data. Then, identify the distribution.

SOLUTION

Key: 4|4 means 44

Stem	Leaf					
4	4	4	4	5	8	
5	2	2	3	3	7	9
6	6	6	8	9		
7	0	5	9			
8	0	3				
9	4	5				

SHAPE OF DISTRIBUTION: RIGHT-SKEWED DISTRIBUTION

EXAMPLE 2.20

The data shown represents the sample of percentage of unemployment in a particular country according to gender. Construct a back-to-back (mixture) stem and leaf plot. Then, compare the distribution of the two groups.

Females	4.9	5.0	5.3	5.5	5.6	5.6	5.8	6.1	6.3	6.6	6.7	7.1	7.4	7.6	6.9
Males	2.1	2.3	2.3	2.7	3.0	3.3	3.3	3.6	3.7	3.9	4.2	4.2	4.4	4.5	5.6

SOLUTION

Females	Stem	Males
	2	1 3 3 7
	3	0 3 3 6 7 9
9	4	2 2 4 5
8 6 6 5 3 0	5	6
9 7 6 3 1	6	
6 4 1	7	

Key: 2|1 means 21

RIGHT-SKEWED DISTRIBUTION

RIGHT-SKEWED DISTRIBUTION

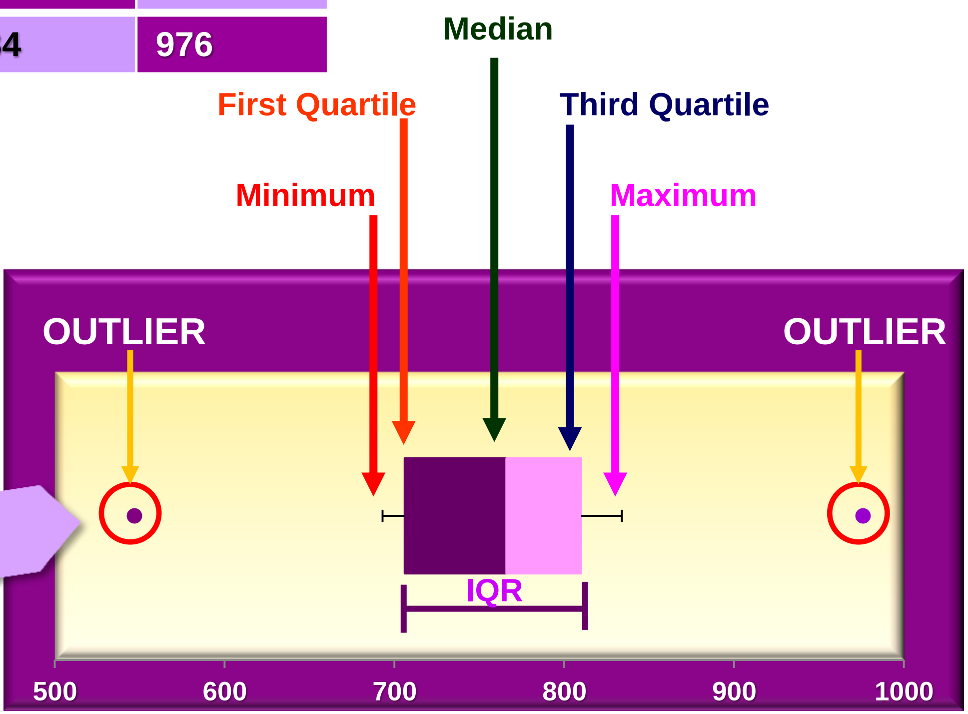
BOXPLOT

A **boxplot** is a **graphic representation** of the **5-number summaries** (**minimum. first quartile, median (second quartile), third quartile and maximum**).

EXAMPLE: FICO credit rating scores

547	693	698	714	751	753
779	789	802	818	834	976

Boxplot



PROCEDURE FOR CONSTRUCTING A BOXPLOT

STEP 1

Arrange the data in ascending order.

STEP 2

Find the 1st quartile, Q_1, 2nd quartile (median), Q_2, and 3rd quartile, Q_3.

STEP 3

Find the outliers(extreme values→ extremely low/extremely high).

$$x < Q_1 - 1.5(Q_3 - Q_1) \text{ or } x > Q_3 + 1.5(Q_3 - Q_1)$$

$IQR\,(Interquartile\ Range)$

*Note: The larger value of IQR, the larger of variability

STEP 4

Draw a scale for the data on the x-axis.

STEP 5

Locate the minimum value, 1st quartile, 2nd quartile (median), 3rd quartile, the maximum value, and outliers on the scale.

STEP 6

Draw a box around 1st quartile and 3rd quartile, draw a vertical line through the median, and connect the upper and lower value.

EXAMPLE 2.21

Plot a box-plot for the following data. Then describe the shape of distribution.

a.	3.2	5.9	4.3	6.9	4.5	8.0	4.7	8.9	5.7	11.9
b.	5.8	9.7	6.7	13.4	6.8	14.7	7.2	16.4	8.2	28.1

SOLUTION

STEP 1

Arrange the data.

3.2	4.3	4.5	4.7	5.7	5.9	6.9	8.0	8.9	11.9

STEP 2

Find the 1st quartile, 2nd quartile (median) and 3rd quartile.

$$Q_1 = x_{c=\frac{1(10)}{4}=2.5} \Rightarrow x_3 = 4.5$$

$$Q_2 = x_{c=\frac{2(10)}{4}=5} \Rightarrow \frac{x_5 + x_6}{2} = \frac{5.7+5.9}{2} = 5.8$$

$$Q_3 = x_{c=\frac{3(10)}{4}=7.5} \Rightarrow x_8 = 8.0$$

EXAMPLE 2.21-CONTINUE

STEP 3

Find the outliers.

$$\text{Lower Limit}: Q_1 - 1.5(Q_3 - Q_1) = 4.5 - 1.5(8.0 - 4.5) = -0.75$$

$$\text{Upper Limit}: Q_3 - 1.5(Q_3 - Q_1) = 8.0 - 1.5(8.0 - 4.5) = 13.25$$

Since the smallest value is 3.2 and the largest value is 11.9.
Therefore, no outlier exists in this data set.

STEP 4

Draw a scale for the data on the x-axis.

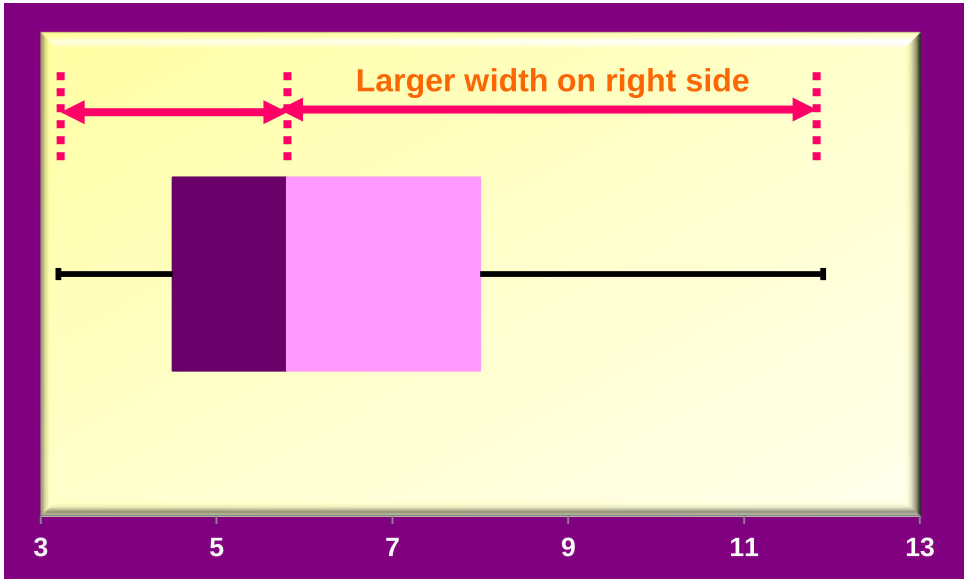
STEP 5

Locate the minimum value, 1st quartile, 2nd quartile (median), 3rd quartile, the maximum value, and outliers on the scale.

STEP 6

Draw a box around 1st quartile and 3rd quartile, draw a vertical line through the median, and connect the upper and lower value.

EXAMPLE 2.21-CONTINUE



SHAPE OF DISTRIBUTION: RIGHT-SKEWED DISTRIBUTION

EXAMPLE 2.21-CONTINUE

Plot a box-plot for the following data. Then describe the shape of distribution.

a.	3.2	5.9	4.3	6.9	4.5	8.0	4.7	8.9	5.7	11.9
b.	5.8	9.7	6.7	13.4	6.8	14.7	7.2	16.4	8.2	28.1

SOLUTION

STEP 1

Arrange the data.

5.8	6.7	6.8	7.2	8.2	9.7	13.4	14.7	16.4	28.1

STEP 2

Find the 1st quartile, 2nd quartile (median) and 3rd quartile.

$$Q_1 = x_{c=\frac{1(10)}{4}=2.5} \Rightarrow x_3 = 6.8$$

$$Q_2 = x_{c=\frac{2(10)}{4}=5} \Rightarrow \frac{x_5 + x_6}{2} = \frac{8.2+9.7}{2} = 8.95$$

$$Q_3 = x_{c=\frac{3(10)}{4}=7.5} \Rightarrow x_8 = 14.7$$

EXAMPLE 2.21-CONTINUE

STEP 3

Find the outliers.

$$\text{Lower Limit}: Q_1 - 1.5(Q_3 - Q_1) = 6.8 - 1.5(14.7 - 6.8) = -5.05$$

$$\text{Upper Limit}: Q_3 - 1.5(Q_3 - Q_1) = 14.7 - 1.5(14.7 - 6.8) = 26.55$$

Since the smallest value is 5.8 and the largest value is 28.1.
Therefore, outlier in this data set is 28.1 $(x = 28.1) > (\text{Upper limit} = 26.55)$.

STEP 4

Draw a scale for the data on the x-axis.

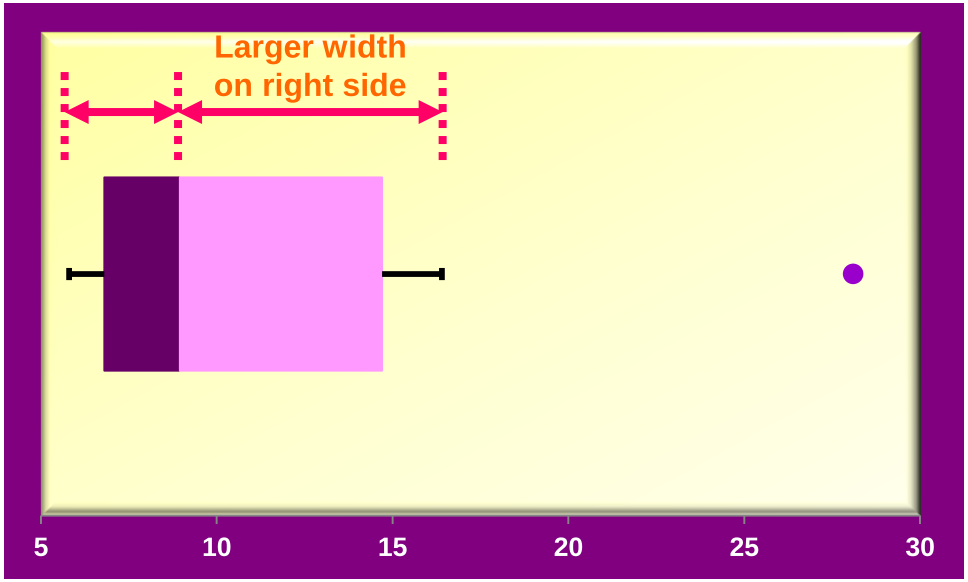
STEP 5

Locate the minimum value, 1st quartile, 2nd quartile (median), 3rd quartile, the maximum value, and outliers on the scale.

STEP 6

Draw a box around 1st quartile and 3rd quartile, draw a vertical line through the median, and connect the upper and lower value.

EXAMPLE 2.21-CONTINUE



SHAPE OF DISTRIBUTION: RIGHT-SKEWED DISTRIBUTION

EXAMPLE 2.22

Two sample of ten spring made out of the steel rods supplied by two different companies were compared. The measurement of flexibility (in N/m) for each spring was recorded as follows.

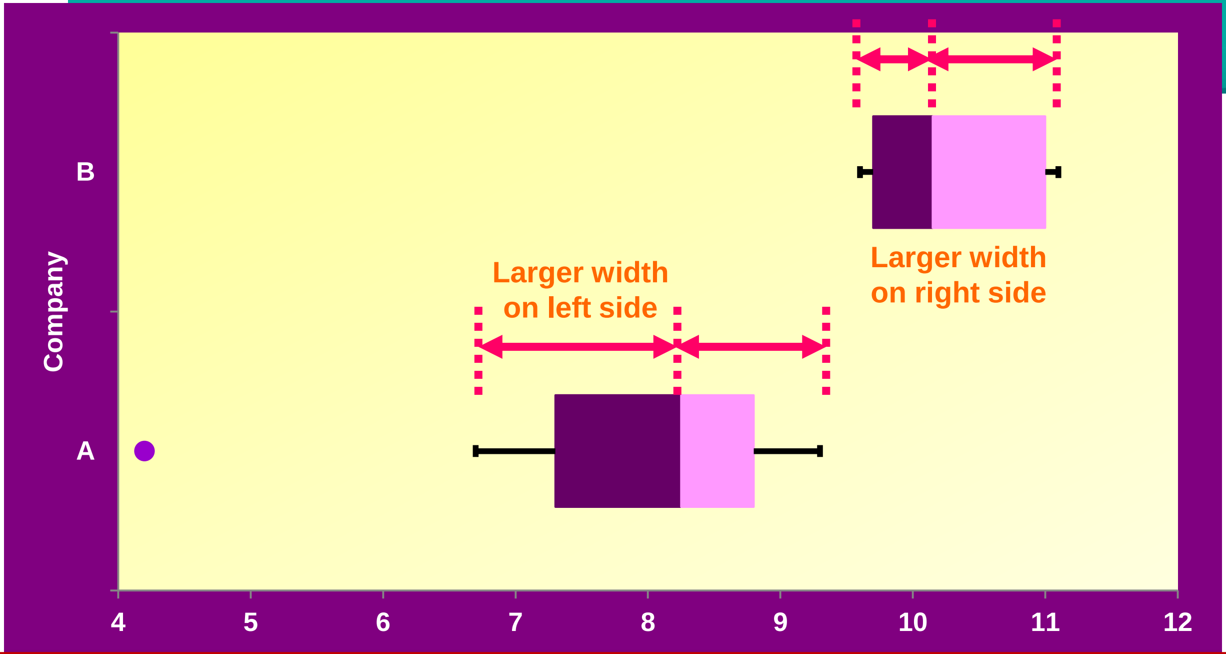
Company A	4.2	6.7	7.3	7.5	8.0	8.5	8.7	8.8	9.2	9.3
Company B	9.6	9.7	9.8	9.9	10.1	10.2	11.0	11.0	11.0	11.1

Compare the distributions, average and variation of both data using boxplots.

SOLUTION

Company	A	B
Minimum	6.7	9.6
1st Quartile	7.3	9.8
2nd Quartile	8.25	10.15
3rd Quartile	8.8	11.0
Maximum	9.3	11.1
Outlier *Upper Limit* *Lower Limit*	1st observation: 4.2 *5.05* *11.05*	No outlier *8.00* *12.8*

EXAMPLE 2.22-CONTINUE



SHAPE OF DISTRIBUTION:
Company A: ***Left-skewed distribution***; Company B: ***Right-skewed distribution***

AVERAGE:
Data of Company B has higher average compared to Company A. This is due to
$$\left(\text{Median}_B = 10.15\right) > \left(\text{Median}_A = 8.25\right)$$

VARIATION:
Data of Company A is more variable compared to Company B. This is due to
$$\left(IQR_A = 8.8 - 7.3 = 1.5\right) > \left(IQR_B = 11.0 - 9.8 = 1.2\right)$$

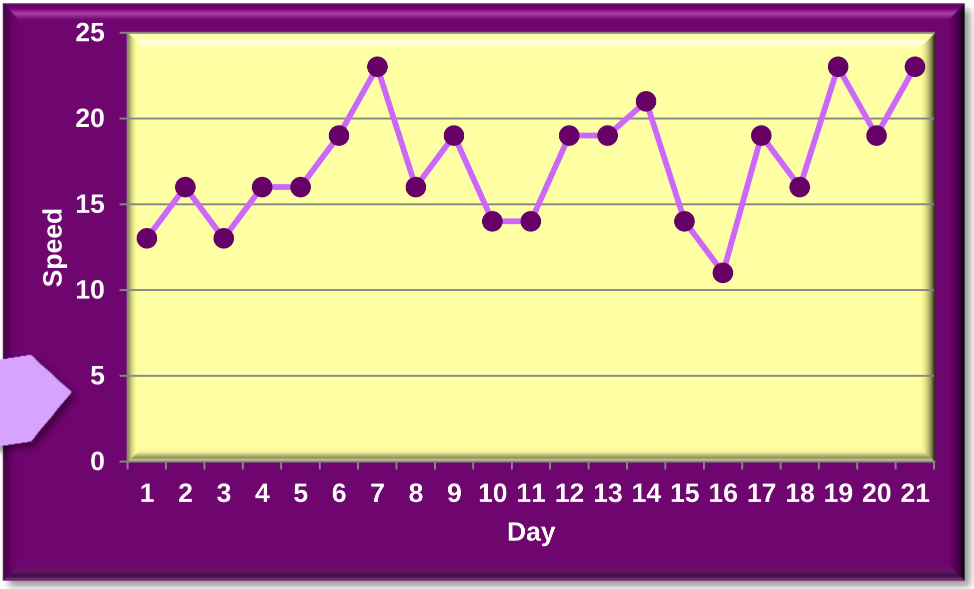
***NOTE: The median is more robust in measure the average of the skewed data compare to mean. Thus we always use the median to measure the average of skew data.**

TIME-SERIES GRAPH

A **time-series** graph is a graph of time-series data, which are **quantitative data** that have been **collected at different points in time (yearly, monthly, quarterly, weekly, etc.)**.

EXAMPLE: The daily wind speed (km/h) in Sepang Malaysia

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Speed	13	16	13	16	16	19	23	16	19	14	14	19	19	21	14	11	19	16	23	19	23



TIME-SERIES GRAPH

EXAMPLE 2.23

Table below shows Malaysia's exports (RM million) of tyres from 2002-2011.

Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Exports	164.01	191.84	164.43	321.63	292.64	243.89	245.02	428.20	301.73	413.65

Construct a time-series graph for the data above.

SOLUTION

THANK YOU

END OF CHAPTER 2 (PART 2)