CHAPTER 7

Measures of Dispersion for Grouped Data

What will you learn?

- Dispersion
- Measures of Dispersion

Why study this chapter?

Statistical analysis such as measure of dispersion is widely applied in various fields, including medicine, agriculture, finance, social science and many more. The career fields that apply statistical analysis include biometrics, actuarial science and financial analysis that use big data to obtain statistical values, and hence represent the data in statistical graphs.

Do you know?

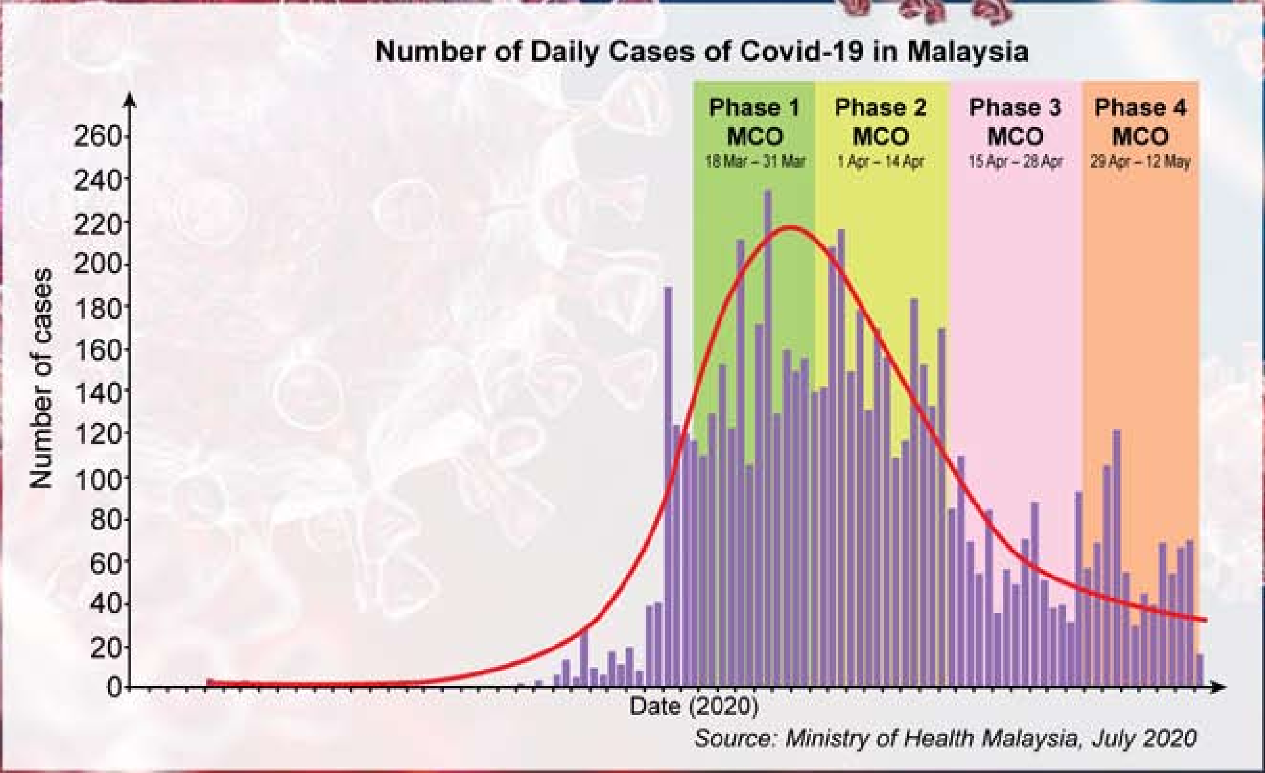
William Playfair (1759-1823) was a Scottish economist who used various common statistical graphs in his book, The Commercial and Political Atlas, published in 1786.

For more information:

bit.do/DoYouKnowChap7

WORD BANK

grouped data	*data terkumpul*
histogram	*histogram*
cumulative histogram	*histogram longgokan*
cumulative frequency	*kekerapan longgokan*
quartile	*kuartil*
ogive	*ogif*
statistical investigation	*penyiasatan statistik*
percentile	*persentil*
frequency polygon	*poligon kekerapan*



Source: Ministry of Health Malaysia, July 2020

The outbreak of the Covid-19 pandemic in early 2020 has forced Malaysian to adjust to a new normal. The swift and efficient action taken by the authorities in tackling the pandemic has helped Malaysia to control the increasing number of patients infected by the virus. Malaysia is successful in flattening the curve of the number of daily infected cases by issuing the Movement Control Order (MCO). In your opinion, how will the shape of the graph be if MCO has not been implemented?

7.1 Dispersion

How to construct histogram and frequency polygon?

Learning Standard

Construct histogram and frequency polygon for a set of grouped data

In Form 4, you have learnt about the ways to interpret the dispersion of ungrouped data based on the stem-and-leaf plots and dot plots. We can observe the dispersion for a grouped data by constructing histogram and frequency polygon. Prior to that, you need to know the class interval, lower limit, upper limit, midpoint, lower boundary, upper boundary and cumulative frequency that can be obtained from a frequency table.

Info Bulletin

Class interval is the range of a division of data.

MIND MOBILISATION 1 Group

Aim: To recognise the lower limit, upper limit, midpoint, lower boundary and upper boundary of a set of data.

Steps:

The data shows the amount of daily pocket money in RM, received by 20 pupils on a particular day.

8	10	4	7	1
5	2	8	11	4
5	7	15	3	4
14	12	7	11	9

1. Identify the smallest data and the largest data.
2. By referring to the data, group the data into 3, 4, 5 or 6 parts in sequence. For example, a group of three uniform parts means 1 – 5, 6 – 10 and 11 – 15.
3. By using the tally method, choose and insert the data according to the parts of the group.
4. Based on each part of the data, determine
   (a) the lower limit (the smallest value in a part of the data) and the upper limit (the largest value in a part of the data),
   (b) the midpoint of each part of the data,
   (c) (i) the middle value between the lower limit of a part and the upper limit of the part before it,
   (ii) the middle value between the upper limit of a part and the lower limit of the part after it.
5. Complete the frequency table with the results of steps 3, 4(a), 4(b), 4(c)(i) and 4(c)(ii) as shown below.

Pocket money (RM)	Frequency	Step 4(a)		Step 4(b)	Step 4(c)	
		Lower limit	Upper limit	Midpoint	(i)	(ii)

Discussion:

Discuss and write down the definition to determine the lower limit, upper limit, midpoint, lower boundary and upper boundary of a set of data.

The results of Mind Mobilisation 1 show that;

Size of class interval

$$= \left(\frac{\text{Largest data value} - \text{Smallest data value}}{\text{Number of classes}}\right)$$

Lower limit is the smallest value and upper limit is the largest value of a class.

$$\text{Midpoint} = \left(\frac{\text{Lower limit} + \text{Upper limit}}{2}\right)$$

Lower boundary

$$= \left(\frac{\text{Upper limit of the class before it} + \text{Lower limit of the class}}{2}\right)$$

Upper boundary

$$= \left(\frac{\text{Upper limit of the class} + \text{Lower limit of the class after it}}{2}\right)$$

Example 1

The data on the right shows the heights, to the nearest cm, of a group of Form 5 pupils.

153	168	163	157
158	161	165	162
145	150	158	156
166	163	152	155
158	173	148	164

(a) Determine the class intervals for the data, if the number of classes required is 6.

(b) Construct a frequency table based on the information in (a). Hence, complete the frequency table with the lower limit, upper limit, midpoint, lower boundary and upper boundary.

i-Technology

Scan the QR code or visit bit.do/WSChap7i to explore ways to organise raw data in frequency table by using spreadsheet.

Solution:

(a) The largest data is 173 and the smallest data is 145.
If the number of classes is 6, then the size of each class interval

$$= \frac{173 - 145}{6}$$
$$= 4.7 \approx 5$$

Size of class interval $= \left(\frac{\text{Largest data value} - \text{Smallest data value}}{\text{Number of classes}}\right)$

Therefore, the class intervals are 145 – 149, 150 – 154, 155 – 159, 160 – 164, 165 – 169 and 170 – 174.

(b)

Height (cm)	Frequency	Lower limit	Upper limit	Midpoint	Lower boundary	Upper boundary
145 – 149	2	145	149	147	144.5	149.5
150 – 154	3	150	154	152	149.5	154.5
155 – 159	6	155	159	157	154.5	159.5
160 – 164	5	160	164	162	159.5	164.5
165 – 169	3	165	169	167	164.5	169.5
170 – 174	1	170	174	172	169.5	174.5

For a grouped data in uniform class intervals, the size of class interval can be calculated using two methods.

Method 1: The difference between the lower limits or the upper limits of two consecutive classes.

Size of class interval of the first two classes

= 150 – 145 Lower limit of class 145 – 149

= 5 Lower limit of class 150 – 154

or

= 154 – 149 Upper limit of class 145 – 149

= 5 Upper limit of class 150 – 154

Method 2: The difference between the upper boundary and the lower boundary of a class interval.

Size of class interval of the first class

= 149.5 – 144.5 Lower boundary of class 145 – 149

= 5 Upper boundary of class 145 – 149

Smart TIPS

To determine the size of class interval, avoid using lower and upper limits of a class. For example, for class interval 145 –149, the size of class interval
= 149 – 145
= 4 (Not true)

Info Bulletin

In Example 1, class 150 – 154 is actually inclusive of the values from 149.5 to 154.5 because the data is a continuous data. The lower boundary 149.5 and the upper boundary 154.5 are used to separate the classes so that there are no gaps between 149 cm and 150 cm, also 154 cm and 155 cm.

The cumulative frequency of a data can also be obtained from a frequency table. The **cumulative frequency** of a class interval is the sum of the frequency of the class and the total frequency of the classes before it. This gives an ascending cumulative frequency.

Example 2

Construct a cumulative frequency table from the frequency table below.

Age	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59
Frequency	4	5	8	7	3

Solution:

Age	Frequency	Cumulative frequency
10 – 19	4	4
20 – 29	5	9
30 – 39	8	17
40 – 49	7	24
50 – 59	3	27

This value of 17 means there are 17 people aged 39 years old and below

Info Bulletin

- Continuous data is a data measured on a continuous scale. For example, the time taken by pupils to buy food at the canteen, and the pupils' heights.
- Discrete data is a data involving counting. For example, the number of pupils in Mathematics Club.

Histogram

Histogram is a graphical representation in which the data is grouped into ranges by using contiguous bars. The height of the bar in histogram represents the frequency of a class. Steps for constructing a histogram:

Find the lower boundary and upper boundary of each class interval. → Choose an appropriate scale on the vertical axis. Represent the frequencies on the vertical axis and the class boundaries on the horizontal axis. → Draw bars that represent each class where the width is equal to the size of the class and the height is proportionate to the frequency.

Frequency polygon

A frequency polygon is a graph that displays a grouped data by using straight lines that connect midpoints of the classes which lie at the upper end of each bar in a histogram. Steps for constructing a frequency polygon:

Info Bulletin

Histogram and frequency polygon can only be constructed by using continuous data.

Mark the midpoints of each class on top of each bar. → Mark the midpoints before the first class and after the last class with zero frequency. → Draw straight lines by connecting the adjacent midpoints.

Example 3

The frequency table below shows the speed of cars in km h^{-1}, recorded by a speed trap camera along a highway in a certain duration. Represent the data with a histogram and frequency polygon by using a scale of 2 cm to 10 km h^{-1} on the horizontal axis and 2 cm to 10 cars on the vertical axis.

Speed (km h^{-1})	70 – 79	80 – 89	90 – 99	100 – 109	110 – 119	120 – 129
Number of cars	5	10	20	30	25	10

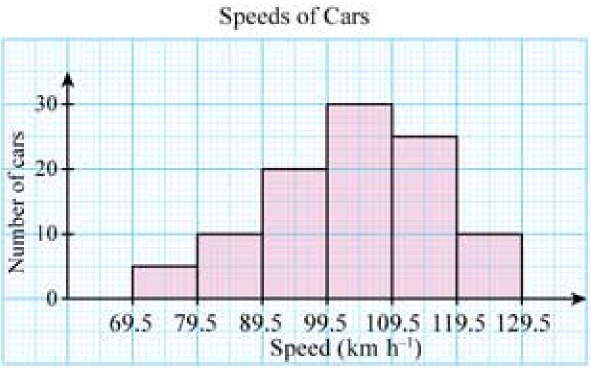
Solution:

Speed (km h^{-1})	Number of cars	Midpoint	Lower boundary	Upper boundary
70 – 79	5	74.5	69.5	79.5
80 – 89	10	84.5	79.5	89.5
90 – 99	20	94.5	89.5	99.5
100 – 109	30	104.5	99.5	109.5
110 – 119	25	114.5	109.5	119.5
120 – 129	10	124.5	119.5	129.5

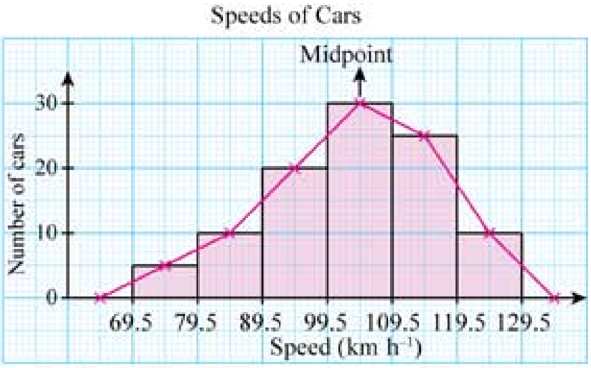
Critical Mind

By using the frequency polygon, explain the speed of cars of more than 90 km h^{-1}.

Histogram:



Frequency polygon:



The frequency polygon can also be constructed without constructing a histogram. Steps for constructing a frequency polygon from a frequency table:

Add one class interval before the first class and after the last class with zero frequency. → Find the midpoint of each class interval. → Choose an appropriate scale on the vertical axis. Represent the frequencies on the vertical axis and the midpoints on the horizontal axis. → Mark the midpoint with the corresponding frequency. → Connect each midpoint with a straight line.

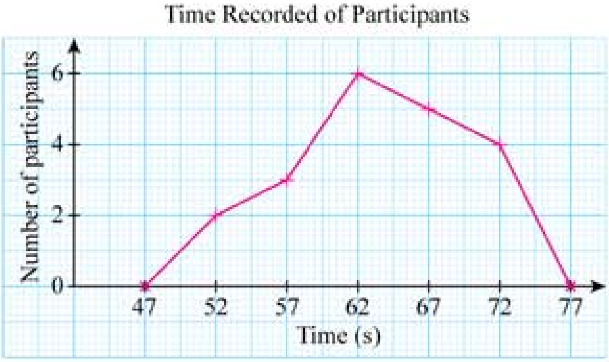
Example 4

The frequency table below shows the time in seconds, recorded by 20 participants in a qualifying round of a swimming competition. Represent the data with a frequency polygon by using a scale of 2 cm to 5 seconds on the horizontal axis and 2 cm to 2 participants on the vertical axis.

Time recorded (s)	50 – 54	55 – 59	60 – 64	65 – 69	70 – 74
Number of participants	2	3	6	5	4

Solution:

Time recorded (s)	Number of participants	Midpoint
45 – 49	0	47
50 – 54	2	52
55 – 59	3	57
60 – 64	6	62
65 – 69	5	67
70 – 74	4	72
75 – 79	0	77



Add a class interval with zero frequency before the first class and after the last class

Self Practice 7.1a

1. The data below shows the time taken by 50 pupils to go to school from their houses. The time recorded is in the nearest minute.

6	15	32	16	18	31	38	20	17	32
18	8	25	35	13	24	14	8	8	25
16	25	30	10	18	14	14	10	25	30
23	30	12	18	6	23	1	15	30	12
40	15	5	14	22	49	12	19	33	25

Construct a frequency table such that there are 5 classes. Then, state the lower limit, upper limit, midpoint, lower boundary and upper boundary of each class interval.

2. The frequency table below shows the masses in kg, of new-born babies in a hospital in a month. State the midpoint, lower limit, upper limit, lower boundary, upper boundary and cumulative frequency of the data.

Mass (kg)	2.0 – 2.4	2.5 – 2.9	3.0 – 3.4	3.5 – 3.9	4.0 – 4.4
Number of babies	9	15	24	20	10

3. The frequency table below shows the number of hours of sleep per day of a group of workers in a factory. By using a scale of 2 cm to 1 hour on the horizontal axis and 2 cm to 20 workers on the vertical axis, construct a histogram and frequency polygon on the same graph to represent the data.

Number of hours of sleep per day	4.05–5.04	5.05–6.04	6.05–7.04	7.05–8.04	8.05–9.04	9.05–10.04	10.05–11.04
Number of workers	2	4	22	64	90	14	2

4. The frequency table below shows the height in m, of sugar cane plants or also known as *Saccharum officinarum* taken from a plantation. Represent the data with a frequency polygon by using a scale of 2 cm to 1 m on the horizontal axis and 2 cm to 10 sugar cane plants on the vertical axis.

Height (m)	1.0 – 1.9	2.0 – 2.9	3.0 – 3.9	4.0 – 4.9	5.0 – 5.9	6.0 – 6.9
Number of sugar cane plants	25	33	46	50	44	36

How to compare and interpret the dispersions based on histogram and frequency polygon?

Learning Standard

Compare and interpret the dispersions of two or more sets of grouped data based on histogram and frequency polygon, hence make conclusion.

Distribution shapes of data

When describing a grouped data, it is important to be able to recognise the shapes of the distribution. The distribution shapes can be identified through a histogram or frequency polygon.

MIND MOBILISATION 2 Group

Aim: To explore the possible shapes of a distribution.

Steps:

1. Divide the class into groups.
2. Open the worksheet by scanning the QR code. Each group is given the worksheet.
3. In the group, classify the distribution shapes into two categories, symmetrical or skewed.

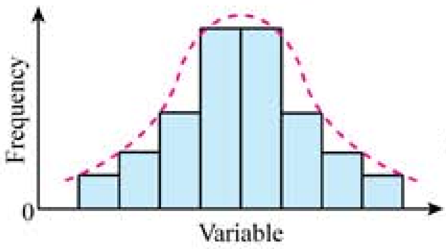
Discussion:

Can you differentiate between symmetrical and skewed shapes?

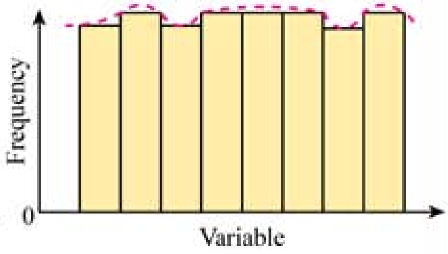
Scan the QR code or visit bit.do/WSChap7ii to obtain the worksheet.

The results of Mind Mobilisation 2 show that a distribution is symmetric if the shape and size of the distribution are almost the same when divided into two parts, left and right. The shape of distribution is skewed if one tail of the histogram is longer than the other tail.

Symmetric Histogram

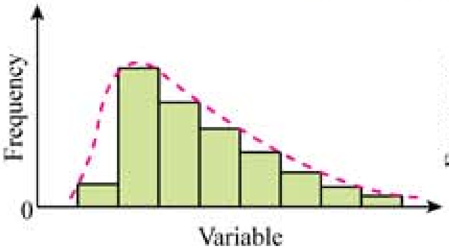


Bell-shaped

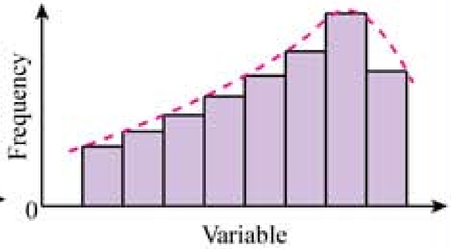


Uniform-shaped

Skewed Histogram



Right-skewed

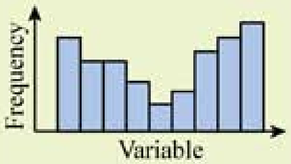


Left-skewed

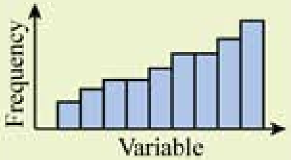
Info Bulletin

Other distribution shapes:

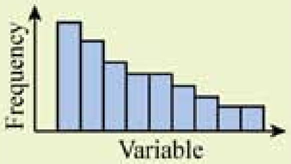
(i) U-shaped



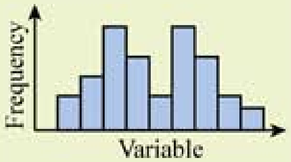
(ii) J-shaped



(iii) Reverse J-shaped

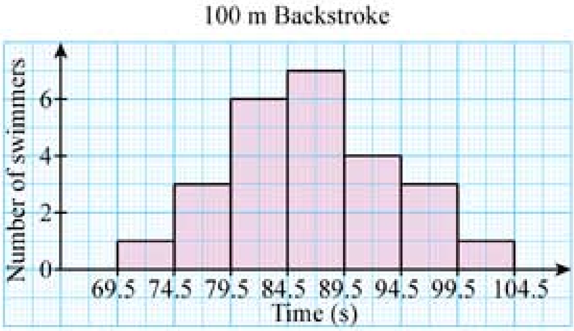


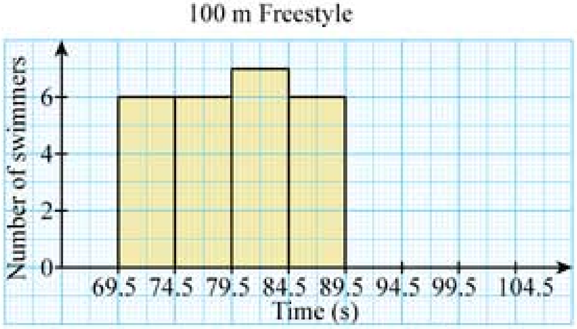
(iv) Bimodal



Example 5

The diagram below shows two histograms representing the time taken by 25 swimmers to complete two different events.





(a) State the distribution shape of the histogram for the two events.

(b) Which event has a wider dispersion of the time taken? Give your reason.

(c) Between backstroke and freestyle, in which event did the swimmers perform better?

Smart TIPS

Distributions are most often not perfectly shaped. Therefore, it is necessary to identify an overall pattern.

Solution:

(a) The histogram for the 100 m backstroke shows a bell-shaped distribution and for 100 m freestyle shows a uniform distribution.

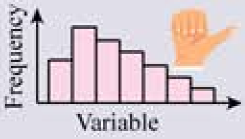
(b) The 100 m backstroke event has a wider dispersion because the difference of the time recorded is larger, that is 30 seconds (102 s – 72 s).

(c) 100 m freestyle. This is because most of the swimmers recorded a better time.

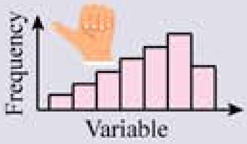
Smart TiPS

To determine distribution shape using hands:

(i) Skew to the right



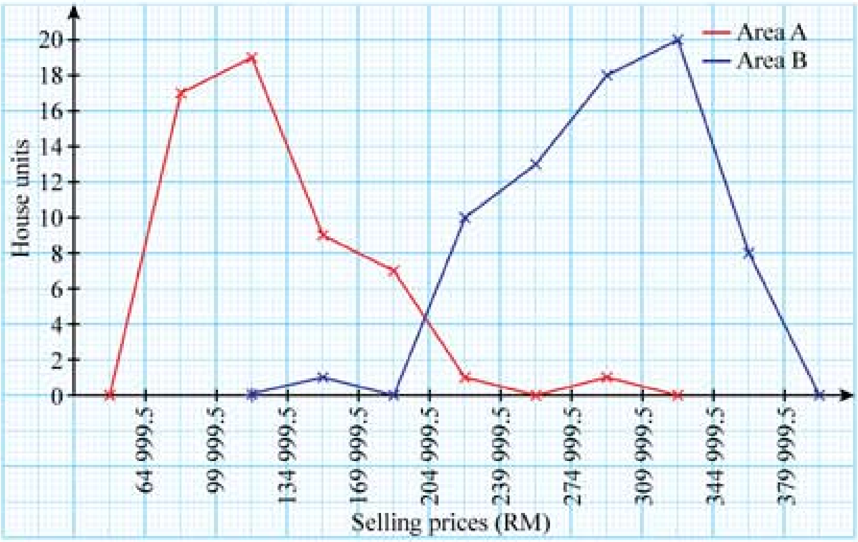
(ii) Skew to the left



Example 6

The frequency polygon below shows the selling prices of the houses that were sold in two different areas in the last six months.

Selling Prices of Houses in Area A and Area B



(a) State the distribution shapes in the two areas.

(b) Compare the dispersions of the house prices in the two areas.

(c) In your opinion, which area represents an urban area and which area represents a rural area?

Solution:

(a) The distribution shape of the selling prices in area A is skewed to the right whereas in area B is skewed to the left.

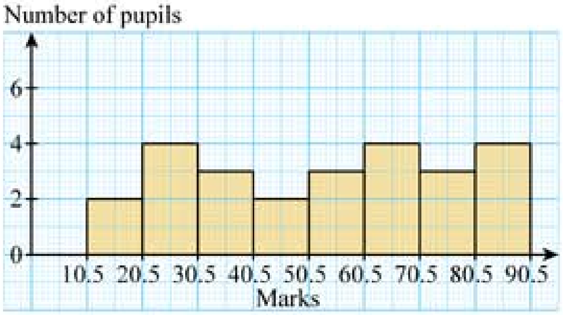
(b) The dispersions of the selling prices in area A and area B are approximately the same even though their distribution shapes are different.

(c) Area A represents a rural area because most of the selling prices are lower whereas Area B represents an urban area because most of the selling prices are higher.

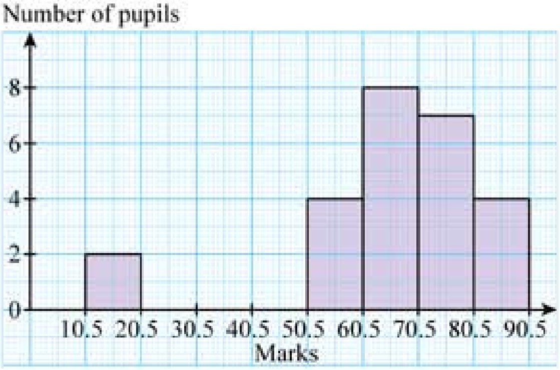
Self Practice 7.1b

1. The diagram below shows two histograms of Mathematics test marks obtained by two groups, Arif and Bestari.

Mathematics Test Marks of Arif Group



Mathematics Test Marks of Bestari Group



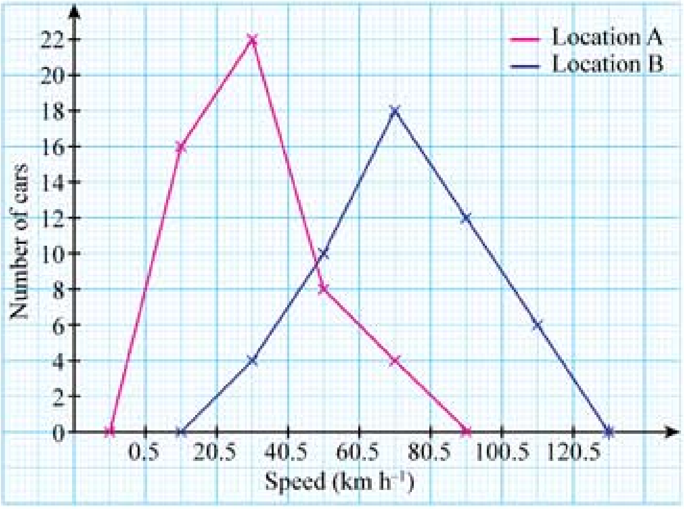
(a) State the distribution shape of the histogram for the two groups.

(b) Compare the dispersions of test marks between the two groups.

(c) Which group shows better results? Give your reason.

2. The diagram below shows the survey results of the traffic flow in two different locations. Each location records the speeds of 50 cars.

Speeds of Cars



(a) State the distribution shapes in both locations.

(b) Compare the dispersions of the car speeds in both locations.

(c) In your opinion, which location is a highway and which location is a housing area?

How to construct an ogive for a set of grouped data?

Learning Standard

Construct an ogive for a set of grouped data and determine the quartiles.

Besides histogram and frequency polygon, a frequency distribution can also be displayed by drawing a cumulative frequency graph, also known as an ogive. When the cumulative frequencies of a data are plotted and connected, it will produce an S-shaped curve. Ogives are useful for determining the quartiles and the percentiles. We will learn how to use an ogive for this purpose in the next section.

Steps for constructing an ogive:

Add one class before the first class with zero frequency. Find the upper boundary and the cumulative frequency for each class. → Choose an appropriate scale on the vertical axis to represent the cumulative frequencies and the horizontal axis to represent the upper boundaries. → Plot the cumulative frequency with the corresponding upper boundary. → Draw a smooth curve passing through all the points.

Quartile

For a grouped data with number of data N, the quartiles can be determined from the ogive. Q_1, Q_2 and Q_3 are the values that correspond to the cumulative frequency $\frac{N}{4}$, $\frac{N}{2}$ and $\frac{3N}{4}$ respectively.

Example 7

The frequency table on the right shows the salt content of 60 types of food.

(a) Construct an ogive to represent the data.

(b) From your ogive, determine

(i) the first quartile

(ii) the median

(iii) the third quartile

Salt content (mg)	Frequency
100 – 149	4
150 – 199	11
200 – 249	15
250 – 299	21
300 – 349	8
350 – 399	1

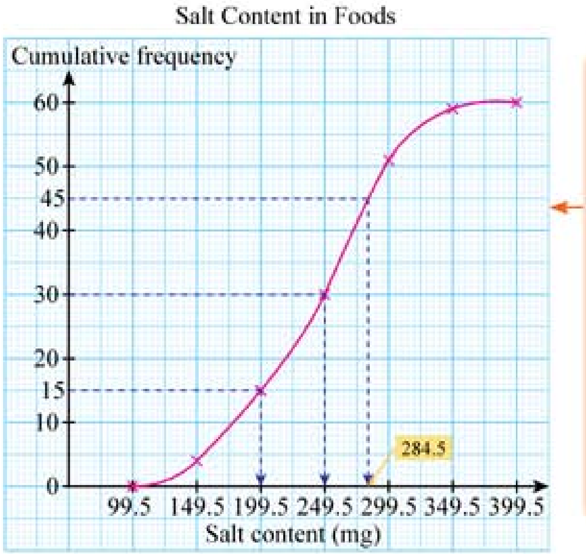
Solution:

(a)

Salt content (mg)	Frequency	Upper boundary	Cumulative frequency
50 – 99	0	99.5	0
100 – 149	4	149.5	4
150 – 199	11	199.5	15
200 – 249	15	249.5	30
250 – 299	21	299.5	51
300 – 349	8	349.5	59
350 – 399	1	399.5	60

MEMORY BOX

- Quartiles are values that divide a set of data into four equal parts. Each set of data has three quartiles, which are Q_1, Q_2 (median) and Q_3.
- The first quartile Q_1, also known as the lower quartile, is the middle value of the lower half of the data before the median or a quartile that contains 25% of the data.
- The second quartile, Q_2, also known as median is the middle value of a set of data.
- The third quartile, Q_3, also known as the upper quartile, is the middle value of the upper half of the data after the median or a quartile that contains 75% of the data.



Steps to determine the quartiles:

1. Number of data, $N = 60$, therefore $\frac{N}{4} = 15$, $\frac{N}{2} = 30$ and $\frac{3N}{4} = 45$.
2. Draw a horizontal line from the axis of cumulative frequency at 15 until it intersects the ogive.
3. From the intersection point in step 2, draw the vertical line down until it meets the axis of salt content at the horizontal axis.
4. The value of the salt content obtained is the value of Q_1.
5. Repeat steps 2 to 4 for the values of 30 and 45 to obtain the values of Q_2 and Q_3.

(b) $\frac{1}{4} \times 60 = 15$

From the graph, the first quartile, $Q_1 = 199.5$ mg

The salt content of 15 types of food are less or equal to 199.5 mg

$\frac{1}{2} \times 60 = 30$

From the graph, the median, $Q_2 = 249.5$ mg

The salt content of 30 types of food are less or equal to 249.5 mg

$\frac{3}{4} \times 60 = 45$

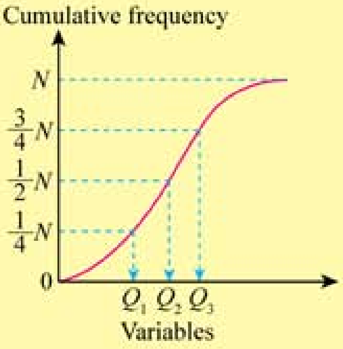
From the graph, the third quartile, $Q_3 = 284.5$ mg

The salt content of 45 types of food are less or equal to 284.5 mg

Info Bulletin

The average salt intake per day among Malaysians is 7.9 g (1.6 teaspoons). This is above the level recommended by the World Health Organization (WHO), which is less than 5 g (one teaspoon) per day.

From Example 7, the first quartile, median and third quartile of a grouped data can be determined by using an ogive.



The first quartile position, Q_1
$= \frac{1}{4} \times$ total frequency, N

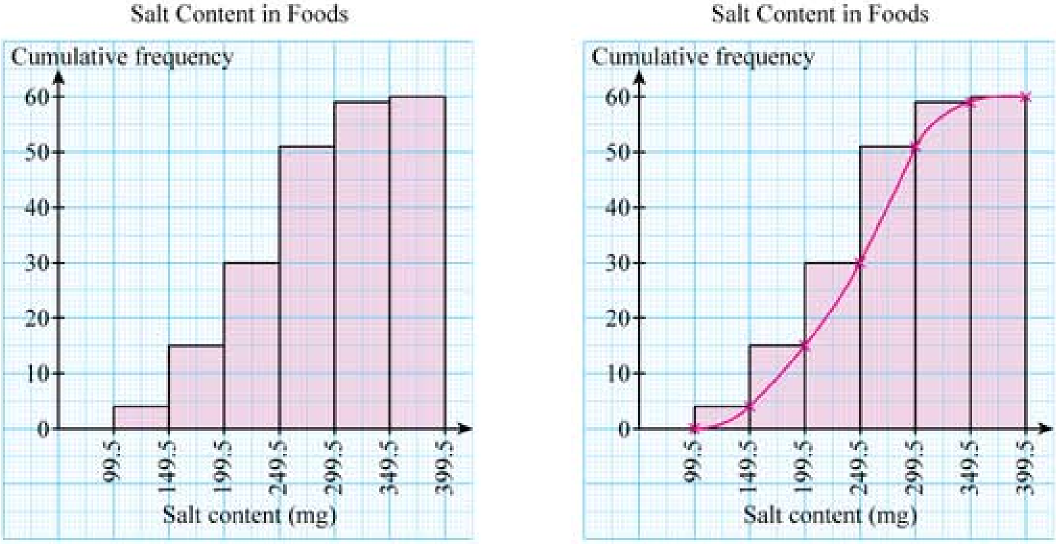
The median position, Q_2
$= \frac{1}{2} \times$ total frequency, N

The third quartile position, Q_3
$= \frac{3}{4} \times$ total frequency, N

Application & Career

A financial manager needs to be an expert in the features of market capital that involve financial assets such as stocks and bonds. Statistical method can be used to analyse the features of market capital through the stocks and bonds distributions.

Cumulative histogram and ogive can be constructed using cumulative frequency table. Cumulative histogram is constructed just like histogram, but the vertical axis is represented by cumulative frequency. By referring to Example 7, the cumulative histogram and the related ogive are as shown below.



How is the construction of ogive related to the construction of cumulative histogram?

Percentile

We can analyse a large data more easily and effectively when we divide the data into small parts which is known as percentile. A **percentile** is a value that divides a set of data into 100 equal parts and is represented by P_1, P_2, P_3, …, P_{99}.

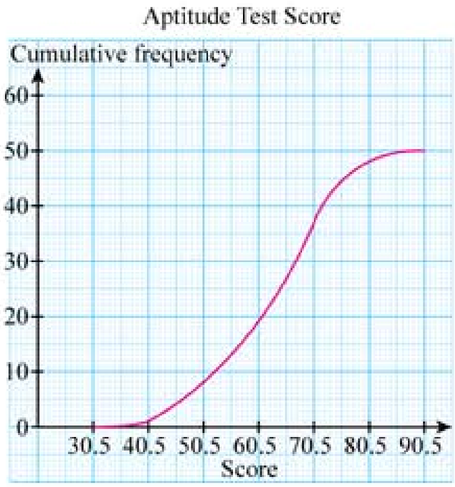
Example 8

The ogive on the right shows the scores of an aptitude test obtained from candidates who are applying for a post in a company.

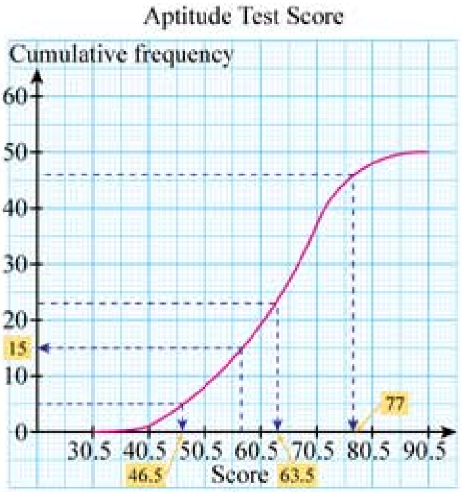
(a) Based on the ogive, find
 (i) the 10th percentile, P_{10}
 (ii) the 46th percentile, P_{46}

(b) Only those candidates who obtained 92nd percentile and above will be called for an interview. What is the minimum score required in order to be called for an interview?

(c) What is the percentage of the candidates who obtained a score of 57 and below?



Solution:



(a) (i) 10% of the total frequency $= \frac{10}{100} \times 50$

$= 5$

From the ogive, $P_{10} = 46.5$

(ii) 46% of the total frequency $= \frac{46}{100} \times 50$

$= 23$

From the ogive, $P_{46} = 63.5$

(b) 92% of the total frequency $= \frac{92}{100} \times 50$

$= 46$

$P_{92} = 77$. Therefore, only candidates with a minimum score of 77 will be called for an interview.

(c) From the ogive,

$\frac{15}{50} \times 100 = 30\%$

Therefore, 30% of the candidates obtained a score of 57 and below.

Info Bulletin

25th percentile is also known as first quartile, 50th percentile as median and 75th percentile as third quartile.

Critical Mind

What is the difference between quartile and percentile?

Self Practice 7.1c

1. The frequency table on the right shows the marks of 100 pupils in an examination.
   (a) Construct an ogive to represent the data.
   (b) From your ogive, determine
      (i) the first quartile
      (ii) the median
      (iii) the third quartile

Marks	Number of pupils
11 – 20	2
21 – 30	13
31 – 40	25
41 – 50	25
51 – 60	19
61 – 70	10
71 – 80	4
81 – 90	2

2. The frequency table on the right shows the length of the soles of 40 pupils.
   (a) Construct an ogive to represent the data.
   (b) Based on the ogive, find
      (i) the 20th percentile, P_{20}
      (ii) the 55th percentile, P_{55}
      (iii) the 85th percentile, P_{85}
   (c) What is the percentage of the pupils having a sole length of 24.6 cm and below?

Length of soles (cm)	Number of pupils
21.0 – 21.9	1
22.0 – 22.9	4
23.0 – 23.9	10
24.0 – 24.9	18
25.0 – 25.9	5
26.0 – 26.9	2

7.2 Measures of Dispersion

How to determine range, interquartile range, variance and standard deviation for grouped data?

In Form 4, you have learnt ways to determine range, interquartile range, variance and standard deviation as a measure to describe dispersion for ungrouped data. In this section, we shall proceed to the measures of dispersion for grouped data.

Learning Standard

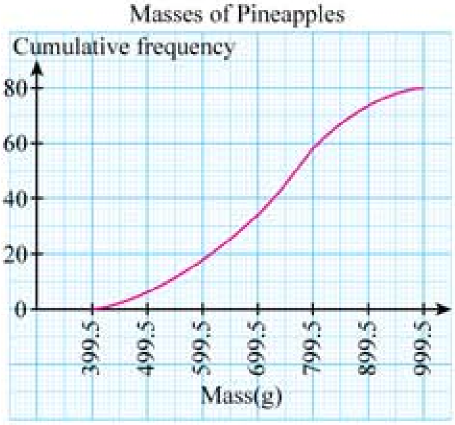
Determine range, interquartile range, variance and standard deviation as a measure to describe dispersion for grouped data.

Range and Interquartile Range

Example 9

Pak Hamidi had recorded the mass of pineapples that he harvested from his farm. The following frequency table and ogive show the data that he obtained. Determine the range and interquartile range for the data.

Mass (g)	Number of pineapples
400 – 499	6
500 – 599	12
600 – 699	16
700 – 799	24
800 – 899	14
900 – 999	8



MEMORY BOX

Interquartile range $= Q_3 - Q_1$

Smart TIPS

Interquartile range of a set of grouped data can be determined from ogive by finding Q_1 and Q_3 first.

Solution:

Range = midpoint of the highest class – midpoint of the lowest class

$= \frac{900 + 999}{2} - \frac{400 + 499}{2}$

$= 949.5 - 449.5$

$= 500$ g

Difference between the heaviest pineapple and the lightest pineapple is 500 g.

From the ogive,

the position of Q_1:

$\frac{1}{4} \times 80 = 20$

$Q_1 = 614.5$

the position of Q_3:

$\frac{3}{4} \times 80 = 60$

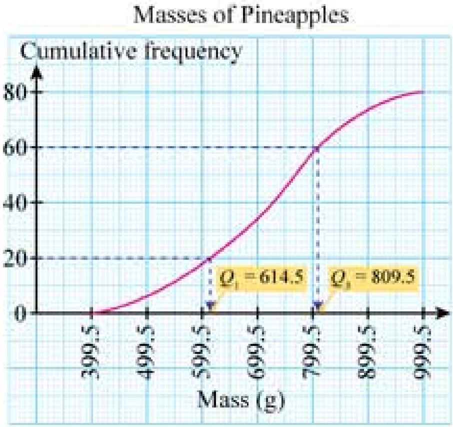
$Q_3 = 809.5$

Therefore, the interquartile range

$= 809.5 - 614.5$

$= 195$ g

Difference between the heaviest pineapple and the lightest pineapple that lies in the middle 50% of the distribution is 195 g.



Variance and Standard Deviation

Variance and standard deviation for a grouped data can be obtained using the following formulae.

Variance, σ^2

$$= \frac{\sum fx^2}{\sum f} - \bar{x}^2$$

Standard deviation, σ

$$= \sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2}$$

where
x = midpoint of the class interval
f = frequency
$\bar{x}$ = mean of the data

MEMORY BOX

- Variance is the average of the square of the difference between each data and the mean.
- Standard deviation is a measure of dispersion relative to its mean, which is measured in the same unit of the original data.

Example 10

The frequency table below shows the volumes of water to the nearest litres, used daily by a group of families in a housing area. Calculate the variance and standard deviation of the data.

Volume of water (ℓ)	150 – 159	160 – 169	170 – 179	180 – 189	190 – 199	200 – 209
Number of families	8	12	15	24	20	16

Solution:

Volume of water (ℓ)	Frequency, f	Midpoint, x	fx	x^2	fx^2
150 – 159	8	154.5	1 236	23 870.25	190 962
160 – 169	12	164.5	1 974	27 060.25	324 723
170 – 179	15	174.5	2 617.5	30 450.25	456 753.75
180 – 189	24	184.5	4 428	34 040.25	816 966
190 – 199	20	194.5	3 890	37 830.25	756 605
200 – 209	16	204.5	3 272	41 820.25	669 124
	$\sum f = 95$		$\sum fx = 17\ 417.5$		$\sum fx^2 = 3\ 215\ 133.75$

$$\text{Mean, } \bar{x} = \frac{\sum fx}{\sum f} = \frac{17\ 417.5}{95} = 183.34\ \ell$$

$$\text{Variance, } \sigma^2 = \frac{\sum fx^2}{\sum f} - \bar{x}^2 = \frac{3\ 215\ 133.75}{95} - \left(\frac{17\ 417.5}{95}\right)^2 = 229.1856 = 229.19\ \ell^2 \text{ (correct to 2 decimal places)}$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2} = \sqrt{229.1855956} = 15.1389 = 15.14\ \ell \text{ (correct to 2 decimal places)}$$

Checking Answer

1. Press MODE MODE
   Display SD REG BASE
   1 2 3
   Choose 1
2. Enter (midpoint), press SHIFT , (frequency) M+, and repeat for the subsequent values.
3. Press AC SHIFT 2
   Display $\bar{x}$ $x\sigma n$ $x\sigma n{-}1$
   1 2 3
   Press 1 for mean:
   Display 183.3421053
   Press 2 for standard deviation:
   Display 15.13887696

Self Practice 7.2a

1. The frequency table below shows the electricity bills of apartment units for a certain month.

Electricity bill (RM)	30 – 49	50 – 69	70 – 89	90 – 109	110 – 129
Number of apartment units	4	9	11	15	13

Construct an ogive for the data and hence, calculate the range and interquartile range. Explain the meaning of the range and interquartile range obtained.

2. Calculate the variance and standard deviation of each of the following data. Give your answer correct to two decimal places.

(a)

Time (minutes)	1 – 2	3 – 4	5 – 6	7 – 8	9 – 10	11 – 12
Frequency	15	20	28	35	30	24

(b)

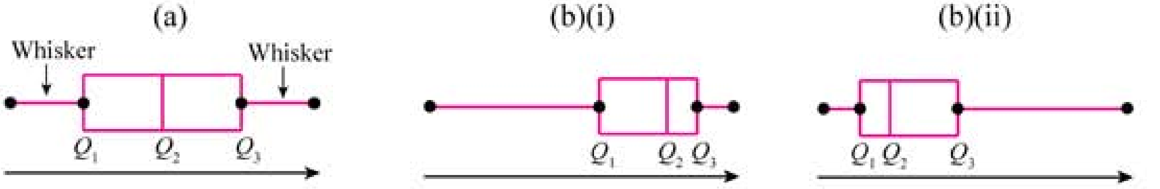
Distance (m)	11 – 20	21 – 30	31 – 40	41 – 50	51 – 60	61 – 70	71 – 80
Frequency	5	8	13	20	22	21	11

How to construct and interpret a box plot for a set of grouped data?

Learning Standard

Construct and interpret a box plot for a set of grouped data.

You have learnt that a box plot is a method to display a group of numerical data graphically based on the five number summary of data. They are the minimum value, first quartile, median, third quartile and maximum value. Similar to the histrogram and frequency polygon, the shape of a distribution can also be identified through the box plot.



Symmetric distribution **Left-skewed distribution** **Right-skewed distribution**

(a) The median lies in the middle of the box and the whiskers are about the same length on both sides of the box.
(b) The median cuts the box into two different sizes.
  (i) If the left side of the box is longer, then the data distribution is left-skewed.
  (ii) If the right side of the box is longer, then the data distribution is right-skewed.

Left whisker and right whisker represent the score outside of the median. If the box is divided into the same size but the left whisker is longer than the right whisker, then the data distribution is left-skewed, and vice versa.

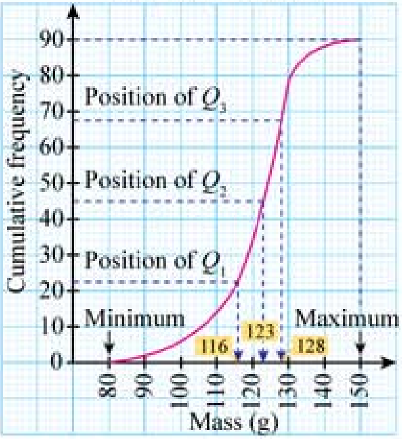
Example 11

The ogive on the right shows the masses in g, of 90 starfruits.

(a) Construct a box plot based on the ogive.

(b) Hence, state the distribution shape of the data.

Masses of Starfruits

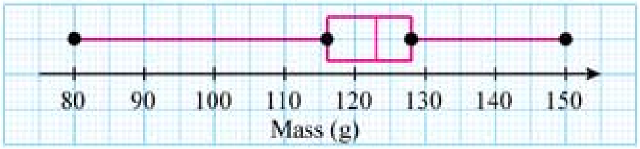


Solution:

(a) From the ogive:

- Minimum value = 80
- Maximum value = 150
- Position of Q_1: $\frac{1}{4} \times 90 = 22.5$
  $Q_1 = 116$
- Position of Q_2: $\frac{1}{2} \times 90 = 45$
  $Q_2 = 123$
- Position of Q_3: $\frac{3}{4} \times 90 = 67.5$
  $Q_3 = 128$

Box plot:

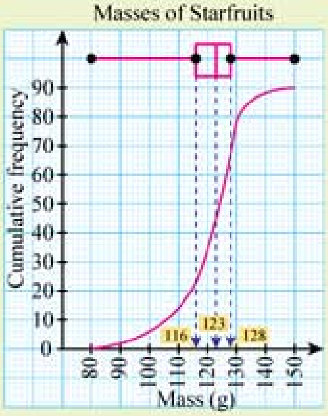


(b) The distribution of the data is skewed to the left because the left side of the box plot is longer than the right side of the box plot.

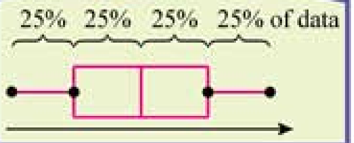
Info Bulletin

Ogive and box plot on the same graph:

Masses of Starfruits



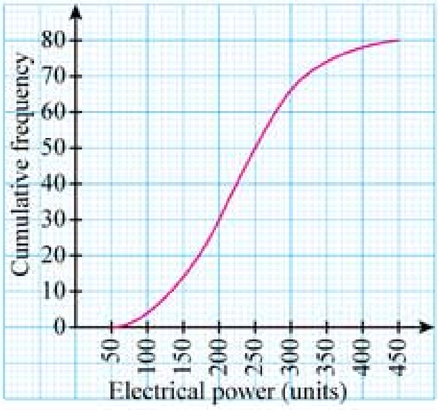
Info Bulletin



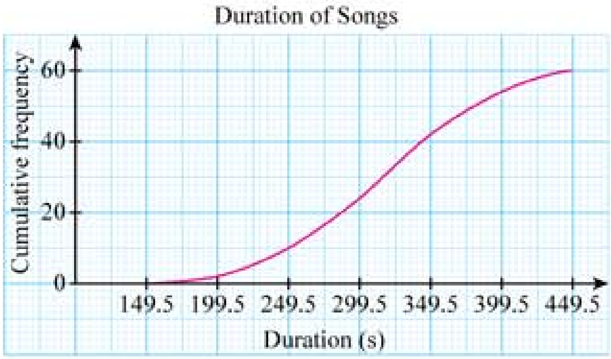
Self Practice 7.2b

1. The ogive on the right shows the number of units of electrical power, consumed by 80 households in a particular month.
   (a) Construct a box plot based on the ogive.
   (b) Hence, state the distribution shape of the data.

Units of Power Consumed



2. The ogive on the right shows the duration in seconds, of 60 songs aired by a radio station at a certain time.
   (a) Construct a box plot based on the ogive.
   (b) Hence, state the distribution shape of the data.



How to compare and interpret two or more sets of grouped data based on measures of dispersion?

Learning Standard

Compare and interpret two or more sets of grouped data, based on measures of dispersion hence make conclusion.

Example 12

A botanist sowed 40 samples of hibiscus seeds using two different hybrids, A and B. The diameters of both hybrids are measured under close guard to develop an extra-large hibiscus. The following frequency table shows the diameters of petals for hybrid A and B.

Diameter (cm)	13.0 – 13.4	13.5 – 13.9	14.0 – 14.4	14.5 – 14.9	15.0 – 15.4
Hybrid A	4	8	9	10	9
Hybrid B	9	10	8	6	7

Based on the mean and standard deviation, determine which hybrid produces larger and more consistent petals. Justify your answer.

Solution:

For hibiscus of hybrid A,

Diameter (cm)	Frequency, f	Midpoint, x	fx	x^2	fx^2
13.0 – 13.4	4	13.2	52.8	174.24	696.96
13.5 – 13.9	8	13.7	109.6	187.69	1 501.52
14.0 – 14.4	9	14.2	127.8	201.64	1 814.76
14.5 – 14.9	10	14.7	147	216.09	2 160.9
15.0 – 15.4	9	15.2	136.8	231.04	2 079.36
	$\sum f = 40$		$\sum fx = 574$		$\sum fx^2 = 8\ 253.5$

$$\text{Mean, } \bar{x} = \frac{574}{40}$$
$$= 14.35 \text{ cm}$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{8\ 253.5}{40} - 14.35^2}$$
$$= \sqrt{0.415}$$
$$= 0.64 \text{ cm}$$

For hibiscus of hybrid B,

Diameter (cm)	Frequency, f	Midpoint, x	fx	x^2	fx^2
13.0 – 13.4	9	13.2	118.8	174.24	1 568.16
13.5 – 13.9	10	13.7	137	187.69	1 876.9
14.0 – 14.4	8	14.2	113.6	201.64	1 613.12
14.5 – 14.9	6	14.7	88.2	216.09	1 296.54
15.0 – 15.4	7	15.2	106.4	231.04	1 617.28
	$\sum f = 40$		$\sum fx = 564$		$\sum fx^2 = 7\ 972$

Mean, $\bar{x}$

$= \dfrac{564}{40}$

$= 14.1$ cm

Standard deviation, σ

$= \sqrt{\dfrac{7\ 972}{40} - 14.1^2}$

$= \sqrt{0.49}$

$= 0.7$ cm

Hybrid A produces larger petals because the mean is larger than hybrid B (14.35 cm > 14.1 cm) and the smaller standard deviation (0.64 cm < 0.7 cm) shows that the diameter of the petals is more consistent.

My Malaysia

Tunku Abdul Rahman Putra Al-Haj declared the hibiscus as The National Flower in 1960. The five petals of the flower represent the five principles of *Rukun Negara*.

Self Practice 7.2c

1. A ball manufacturing factory needs to regulate the internal air pressure in psi, of the produced ball before being marketed. The frequency table below shows the internal air pressures of 50 ball samples taken from machine P and machine Q.

Air pressure (psi)	8.0 – 8.9	9.0 – 9.9	10.0 – 10.9	11.0 – 11.9	12.0 – 12.9	13.0 – 13.9
Machine P	7	11	13	12	5	2
Machine Q	1	3	5	20	18	3

The factory specified that the internal air pressure of a ball should be between 11.3 psi to 11.7 psi. Which machine shows better performance in terms of air pressure accuracy?

2. The frequency table below shows the lifespans in years, of brand X and brand Y batteries.

Lifespan (years)	0 – 0.9	1.0 – 1.9	2.0 – 2.9	3.0 – 3.9	4.0 – 4.9
Brand X battery	4	10	17	20	9
Brand Y battery	10	21	15	8	6

By using suitable measures, determine which brand of battery is better and lasts longer.

How to solve problems involving measures of dispersion for grouped data?

Learning Standard

Solve problems involving measures of dispersion for grouped data.

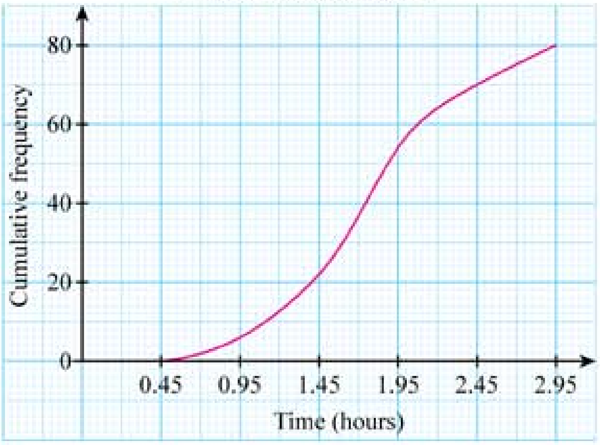
Example 13

A survey on the duration of time in hours, spent by customers to buy goods in a supermarket is carried out. The results of the survey are shown in the ogive on the right.

(a) Construct a frequency table for the time taken by the customers to buy goods in the supermarket using the classes 0.5 – 0.9, 1.0 – 1.4, 1.5 – 1.9, 2.0 – 2.4 and 2.5 – 2.9.

(b) Hence, estimate the mean and standard deviation of the data.

Time Spent by Customers



Solution:

Understanding the problem

Determine the mean and standard deviation from the ogive.

Devising a strategy

(a) Construct the frequency table from the ogive.

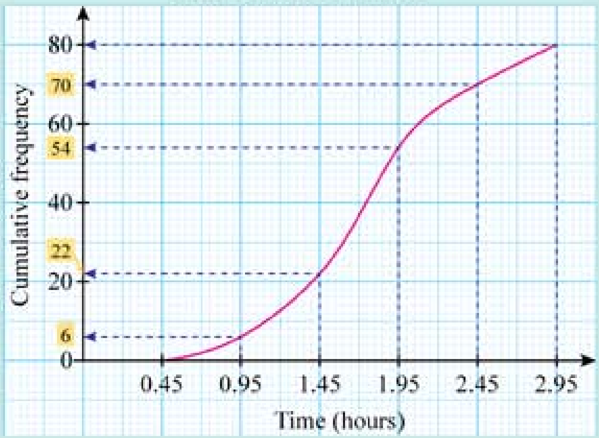
(b) Calculate the mean and standard deviation using formula.

Implementing the strategy

(a)

Time (hours)	Number of customers	
0.5 – 0.9	6	6 – 0 = 6
1.0 – 1.4	16	22 – 6 = 16
1.5 – 1.9	32	54 – 22 = 32
2.0 – 2.4	16	70 – 54 = 16
2.5 – 2.9	10	80 – 70 = 10

Time Spent by Customers



(b)

Time (hours)	Frequency, f	Midpoint, x	fx	x^2	fx^2
0.5 – 0.9	6	0.7	4.2	0.49	2.94
1.0 – 1.4	16	1.2	19.2	1.44	23.04
1.5 – 1.9	32	1.7	54.4	2.89	92.48
2.0 – 2.4	16	2.2	35.2	4.84	77.44
2.5 – 2.9	10	2.7	27	7.29	72.9
	$\sum f = 80$		$\sum fx = 140$		$\sum fx^2 = 268.8$

$$\text{Mean, } \bar{x} = \frac{140}{80} = 1.75 \text{ hours}$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{268.8}{80} - 1.75^2} = \sqrt{0.2975} = 0.55 \text{ hours}$$

Making a conclusion

(a)

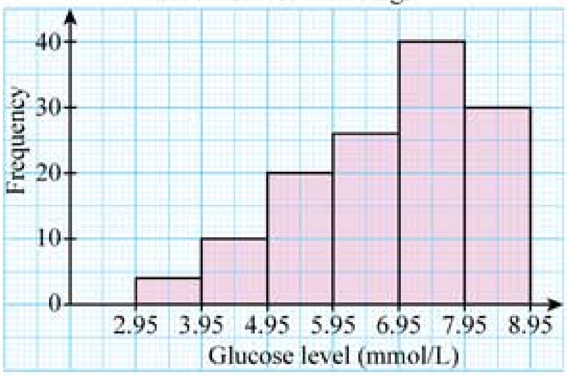
Time (hours)	Number of customers
0.5 – 0.9	6
1.0 – 1.4	16
1.5 – 1.9	32
2.0 – 2.4	16
2.5 – 2.9	10

(b) Mean, $\bar{x}$ = 1.75 hours
Standard deviation, σ = 0.55 hours

Self Practice 7.2d

1. The histogram on the right shows the blood glucose readings taken from a group of patients in a clinic.
   (a) Based on the histogram, is the distribution symmetrical? Give your reason.
   (b) Calculate the mean and standard deviation of the blood glucose readings.
   (c) Compare the standard deviation of the blood glucose readings between 6.0 mmol/L and 8.9 mmol/L with the standard deviation in (b). Justify your answer.

Blood Glucose Readings



2. The table below shows the results of statistical analysis for the price in RM, of 10 kg of rice in supermarket P and supermarket Q. The number of data taken is 20 rice bags from each supermarkets respectively.

Supermarket	Mean	Standard deviation	Minimum value	First quartile	Median	Third quartile	Maximum value
P	32	5.62	26	30	32	34	40
Q	32	4.05	26	32	34	34	40

(a) State the mean and range of the price of the rice in both supermarkets.

(b) The price distribution of the rice in which market is more symmetric? Explain your answer.

(c) Discuss about the median and interquartile range of the two data.

How to design and conduct a mini-project involving statistical investigations?

Learning Standard

Design and conduct a mini-project involving statistical investigations based on measures of central tendency and measures of dispersion and hence interpret and communicate research findings.

PROJECT

The statistical data collected by the National Health and Morbidity Survey (NHMS) shows that more Malaysians are becoming obese with the rate of one in two adults suffering from overweight. Other surveys also found that overweight and obesity among school pupils make up 30% of the population. Obesity increases the chances of developing health problems such as diabetes, heart disease and stroke.

Your school has decided to launch "Activate Your Life" campaign with the aim to create awareness among pupils about obesity and motivate them to adopt healthy lifestyle. Your mathematics teacher wishes to display the health level of pupils on the school bulletin board according to their gender.

Title: Health Level of Pupils

Material: Measuring tape, weighing scale

Procedure:

1. Each group will investigate the health level of pupils based on their gender by using the Body Mass Index (BMI). Divide the class into five groups where each group conducts survey to Form 1, 2, 3, 4 and 5 pupils. Fix the equal number of respondents according to gender from each form.

Smart TIPS

You can carry out the research outside the teaching and learning (PdP) session.

2. Each group is asked to prepare a project report as part of the learning in the class. The report needs to cover the following aspects:

(a) **Survey**
Generate suitable questions such as gender, height, mass and the number of hours spent in sport activities by the respondents within a week as part of the data collection process.

(b) **Data collection method**
Choose a collection method to obtain the data. Choose your respondents randomly.

(c) **Data organisation method**
Construct a frequency table to organise your data. Choose an appropriate class interval for each data.

(d) **Graphical representations**
Present your data using histogram, frequency polygon and other suitable representations.

Smart TiPS

You can organise the data and construct the graphical representation using electronic spreadsheet.

(e) **Data analysis**

(i) Calculate the suitable measures of central tendency and measures of dispersion for each of your data.

(ii) Calculate the Body Mass Index (BMI) of each pupil using the following formula.

$$\text{BMI} = \frac{\text{Mass (kg)}}{\text{Height (m)} \times \text{Height (m)}}$$

(iii) The table below shows the BMI according to age for a teenage boy.

Age	Underweight	Normal	Overweight	Obese
13	$\leqslant 14.8$	14.9 – 20.8	20.9 – 24.8	> 24.8
14	$\leqslant 15.4$	15.5 – 21.8	21.9 – 25.9	> 25.9
15	$\leqslant 15.9$	16.0 – 22.7	22.8 – 27.0	> 27.0
16	$\leqslant 16.4$	16.5 – 23.5	23.6 – 27.9	> 27.9
17	$\leqslant 16.8$	16.9 – 24.3	24.4 – 28.6	> 28.6

The table below shows the BMI according to age for a teenage girl.

Age	Underweight	Normal	Overweight	Obese
13	$\leqslant 14.8$	14.9 – 21.8	21.9 – 26.2	> 26.2
14	$\leqslant 15.3$	15.4 – 22.7	22.8 – 27.3	> 27.3
15	$\leqslant 15.8$	15.9 – 22.5	22.6 – 28.2	> 28.2
16	$\leqslant 16.1$	16.2 – 24.1	24.2 – 28.9	> 28.9
17	$\leqslant 16.3$	16.4 – 24.8	24.9 – 29.3	> 29.3

Source: World Health Organization (WHO), 2007

Based on the BMI table and the data collected, determine the percentage of pupils that are under the category 'Underweight' and 'Obese' by constructing an ogive.

(f) **Description and conclusion**

(i) Interpret your findings of the research. Make conclusions about the physical state of pupils according to their gender for each form.

(ii) Suggest follow-up actions that need to be taken by the pupils who are underweight, overweight and obese.

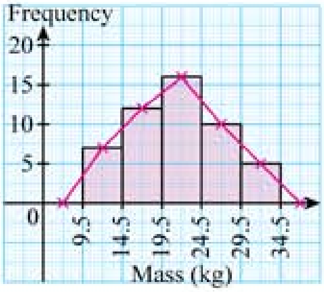
3. Write down your findings of the research on the cardboards and paste it on your school bulletin board.

Summary Arena

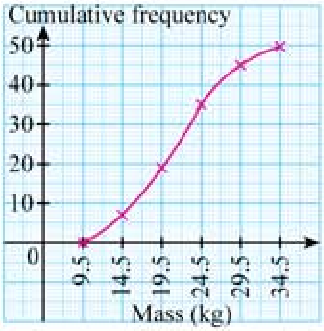
MEASURES OF DISPERSION FOR GROUPED DATA

Data Representation

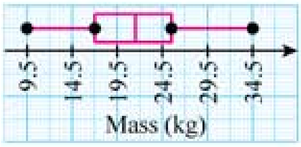
Histogram and frequency polygon:



Ogive:



Box plot:



Measures of Dispersion

Range
= Midpoint of the highest class − Midpoint of the lowest class

Interquartile range
$= Q_3 - Q_1$
(The values of Q_3 and Q_1 are determined from ogive)

Percentile

Variance, $\sigma^2 = \dfrac{\sum fx^2}{\sum f} - \bar{x}^2$

Standard deviation, σ
$= \sqrt{\dfrac{\sum fx^2}{\sum f} - \bar{x}^2}$

Reflection

At the end of this chapter, I can

	🙂	😵
construct histogram and frequency polygon for a set of grouped data.		
compare and interpret the dispersions of two or more sets of grouped data based on histogram and frequency polygon, hence make conclusion.		
construct an ogive for a set of grouped data and determine the quartiles.		
determine range, interquartile range, variance and standard deviation as a measure to describe dispersion for grouped data		
construct and interpret a box plot for a set of grouped data.		
compare and interpret two or more sets of grouped data, based on measures of dispersion hence make conclusion.		
solve problems involving measures of dispersion for grouped data.		
design and conduct a mini-project involving statistical investigations based on measures of central tendency and measures of dispersion and hence interpret and communicate research findings.		

MINI PROJECT

You are required to investigate the population distribution of Malaysia, Indonesia and Singapore from 1990 to 2019. You can obtain the population data by scanning the QR code on the right.

Then, organise the data in frequency table using the appropriate class interval. Construct a suitable data representation to see the data distribution.

For the data of each country, obtain the values of the measures of central tendency and the measures of dispersion. By using the value of suitable measure, compare the populations in these three countries from the aspect of total population and population dispersion. Make a conclusion for the population distribution and relate it with the population density in each country.

Scan the QR code or visit bit.do/MPChap7 to obtain the population data.

Extensive Practice

Scan the QR code or visit bit.do/QuizE07 for interactive quiz

UNDERSTAND

1. For each of the following class intervals, determine the lower limit, upper limit, midpoint, lower boundary and upper boundary.

(a)

Length (m)	10 – 14	15 – 19	20 – 24	25 – 29	30 – 34

(b)

Volume (cm^3)	25.0 – 25.9	26.0 – 26.9	27.0 – 27.9	28.0 – 28.9	29.0 – 29.9

(c)

Mass (g)	0 – 0.24	0.25 – 0.49	0.50 – 0.74	0.75 – 0.99

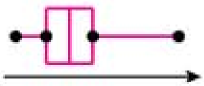
2. Identify whether the following descriptions is for a histogram, a frequency polygon or an ogive.

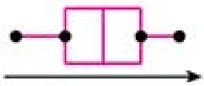
(a) Graph that represents cumulative frequency of classes in a frequency distribution.

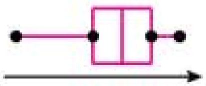
(b) Displays data using side by side bars. The height of the bar is used to represent the class frequency.

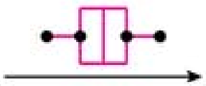
(c) Displays data using straight lines that connect the midpoints of the class interval. Frequency is represented by the height of these midpoints.

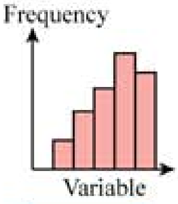
3. The box plot below shows the distribution shapes of data. Match the following histograms with the corresponding box plot in the space provided.

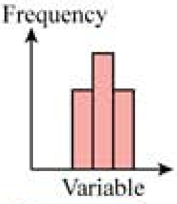

A

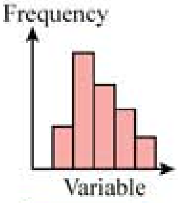

B

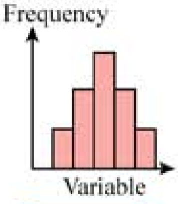

C


D

(a)


(b)


(c)


(d)


MASTERY

4. The data on the right shows the heights in cm, of 30 pupils in Form 5.

146	163	156
152	174	156
178	151	148
166	154	150
164	157	171
168	159	170
163	157	161
167	162	157
166	160	155
168	158	162

(a) Organise the data by completing the frequency table below. Then, draw a histogram of the data using a suitable scale.

Height (cm)	Lower boundary	Upper boundary	Tally	Frequency
145 – 149				
150 – 154				

(b) Construct a new frequency table by rearranging the class intervals beginning with 145 – 148 cm, 149 – 152 cm, 153 – 156 cm and so forth. Hence, construct a histogram to display the data.

(c) Compare the distribution shapes of the two histograms. In your opinion, what is the conclusion that can be made from this comparison?

5. The frequency table below shows the time spent watching television in a week by 30 families.

Time (hours)	2 – 4	5 – 7	8 – 10	11 – 13	14 – 16	17 – 19	20 – 22
Number of families	8	9	6	4	2	0	1

(a) On the same graph, construct a histogram and frequency polygon of the data using a suitable scale.

(b) Comment on the distribution shape of the data displayed.

6. The frequency table below shows the Mathematics test marks of a group of pupils.

Marks	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 – 99
Number of pupils	4	8	12	10	9	7

Draw an ogive of the data and calculate

(a) the range,

(b) the interquartile range,

(c) the 40th percentile and 80th percentile for the test marks.

CHALLENGE

7. Khuzairi is a dairy cow raiser. He rears 130 dairy cows in his farm. The frequency table below shows the volumes in litres, of milk produced by his cows in a certain week.

Volume of milk (litres)	5 – 10	11 – 16	17 – 22	23 – 28	29 – 34	35 – 40
Number of dairy cows	15	28	37	26	18	6

(a) Construct a cumulative histogram of the data.

(b) On the same graph in (a), construct an ogive. Hence, estimate the interquartile range of the distribution.

8. The frequency table below shows the blood pressure readings taken from a group of patients before and after taking a dose of a type of medicine in lowering blood pressure.

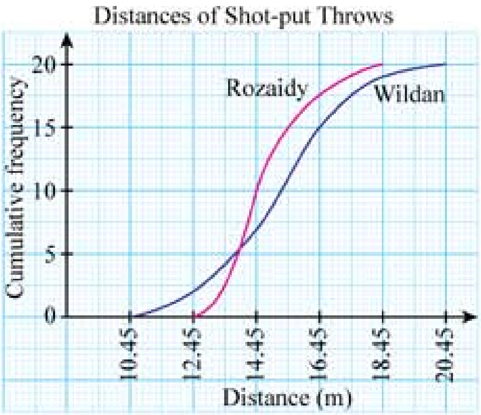
Systolic blood pressure (mmHg)	120 − 134	135 − 149	150 − 164	165 − 179
Before	4	7	8	6
After	9	8	7	1

Calculate the mean and standard deviation of the data. Is the medicine effective in lowering down the blood pressure of the group of patients after taking a dose of the medicine? Justify your answer.

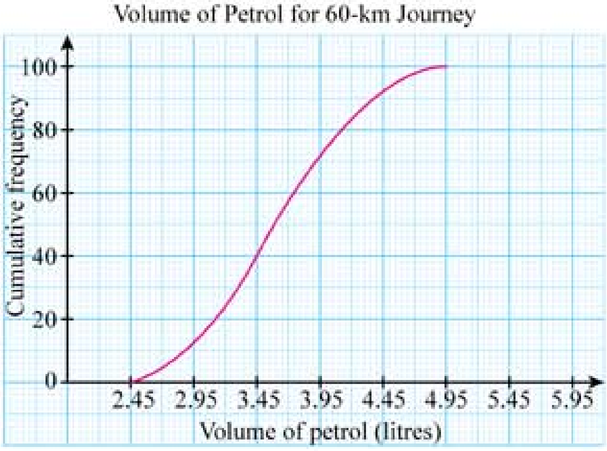
9. The ogive on the right shows the distances of shot-put throws obtained by Rozaidy and Wildan in a training session.

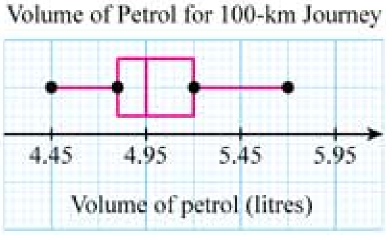
(a) Calculate the percentage of the throwing distances that exceeds 15.45 m obtained by Rozaidy and Wildan.

(b) Based on the median and the third quartile of both performances, determine who perform better during the training session.



10. Volumes of petrol consumed by 100 cars were recorded. The ogive shows the volume of petrol consumed for a 60-km journey and the box plot shows the volume of petrol consumed for a 100-km journey.





(a) Redraw the ogive for the 60-km journey. On the same graph, draw an ogive for the volume of petrol consumed for the 100-km journey.

(b) If a car uses 3.7 litres petrol for the 60-km journey, calculate the volume of petrol consumed for the 100-km journey. Justify your answer.

EXPLORING MATHEMATICS

Instructions:

(i) Do this activity in a small group.

(ii) Each group answer the question in the activity worksheet (scan the QR code).

(iii) After completing the worksheet, each group needs to construct a mind map that summarises the distribution shape and the suitable measures to describe the data.

(iv) Present the outcome of your group. The best presentation will be displayed at the Mathematics corner in your class.

Scan the QR code or visit bit.do/EMChap7 to perform this activity.