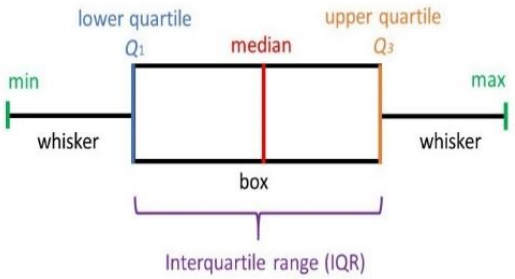
Week 2/3
Term 3
2023

HAWKER COLLEGE
Engage | Inspire | Achieve

MA2
Univariate data

Goals



This fortnight we are going to:

- describe the distribution of a numerical dataset, it's shape (symmetric versus positively or negatively skewed), location and spread, and outliers, and interpret this information in the context of the data
- determine the mean and standard deviation of a dataset and use these statistics as measures of location and spread of a data distribution, being aware of their limitations

Theoretical Components

Resources:
PDF file: Week 2/3 Notes and Exercises

Knowledge Checklist

- Describing distributions of numerical data
- Measures of central tendency
  - Mean
  - Median
  - Mode
- Measures of spread
  - Range
  - IQR
  - Standard deviation
- Effect of extreme values
- Box plots

Order

1. Read through the notes and examples
2. Work through the exercises
3. Complete the investigation at the end of the booklet.
4. Complete the reflection at the end of the booklet
5. Come and see your teacher and make sure you are up to date.

Practical Components

Work through the exercises and show the completed tasks to your teacher.

Be sure to ask for help as you need for the successful completion of all tasks.

Remember to regularly check Google Classroom for messages.

Investigation

Complete the task at the end of the booklet and submit your work for checking. ☺

QFO

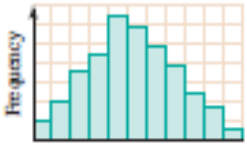
MATHEMATICAL APPLICATIONS 2

WEEK 2/3 NOTES & EXERCISES

DESCRIBING DISTRIBUTIONS OF NUMERICAL DATA

Symmetric distributions

The data shown in the histogram below can be described as *symmetric*.



There is a single peak and the data trail off on both sides of this peak in roughly the same fashion.

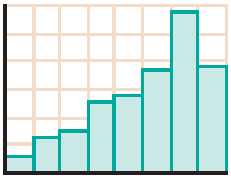
Similarly, in the stem plot below, the distribution of the data could be described as symmetric.

Stem	Leaf
0	7
1	2 3
2	2 4 5 7 9
3	0 2 3 6 8 8
4	4 7 8 9 9
5	2 7 8
6	1 3

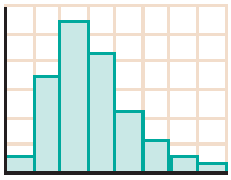
The single peak for these data occurs at the stem 3. On either side of the peak, the number of observations reduces in approximately matching fashion.

Skewed distributions

Each of the histograms below show examples of skewed distributions.



Negatively skewed distribution

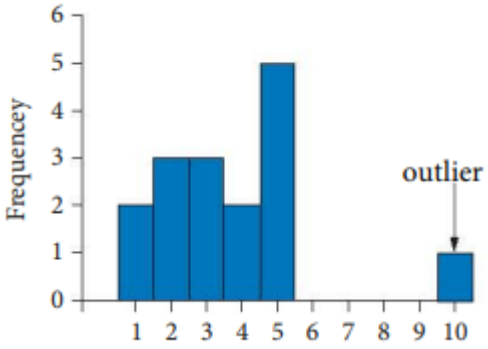


Positively skewed distribution

The figure on the left shows data which is *negatively skewed*. The data in this case peaks to the right and trails off to the left. The figure on the right shows *positively skewed* data. The data in this case peaks to the left and trails off to the right.

In a stem and leaf plot, data is positively skewed if it peaks to the top and trails off below, and negatively skewed if it peaks to the bottom and trails off above.

Outliers are data values that stand out from the main body of data.



EXERCISE 1

1. For each of the following stem plots, describe the shape of the distribution of the data and comment on the existence of any outliers.

a)

Stem	Leaf
0	1 3
1	2 4 7
2	3 4 4 7 8
3	2 5 7 9 9 9 9
4	1 3 6 7
5	0 4
6	4 7
7	1

Key: 1|2 = 12

b)

Stem	Leaf
1	3
2	6
3	3 8
4	2 6 8 8 9
5	4 7 7 7 8 9 9
6	0 2 2 4 5

Key: 2|6 = 2.6

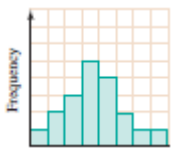
c)

Stem	Leaf
2	3 5 5 6 7 8 9 9
3	0 2 2 3 4 6 6 7 8 8
4	2 2 4 5 6 6 6 7 9
5	0 3 3 5 6
6	2 4
7	5 9
8	2
9	7
10	
11	
12	
13	5

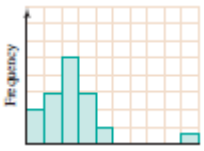
Key: 10|4 = 104

2. For each of the following histograms, describe the shape of the distribution of the data and comment on the existence of any outliers.

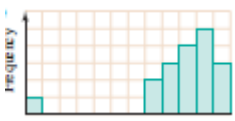
a)



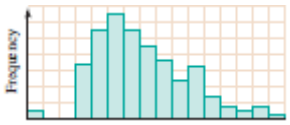
b)



c)



3. [Multiple Choice] The distribution of the data shown in this histogram could be described as:



A negatively skewed

B negatively skewed with one outlier

C positively skewed

D positively skewed with one outlier

E symmetric.

4. The number of hours of exercise completed each week by a group of employees at a company is shown on the stem plot below.

Stem	Leaf
0	0 0 0 0 1 1
0	2 2 2 2 3 3 3
0	4 4 5
0	6 7
0	8
1	
1	
1	4
1	
1	9

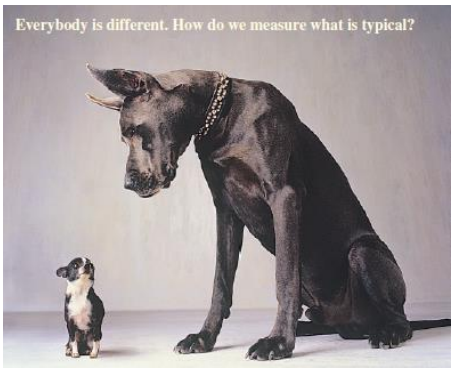
Key: 0|1 = 1 hr

a) Describe the shape of the distribution of these data and comment on the existence of any outliers.

b) What does this tell us about the number of hours of exercise completed weekly by the employees in this company?

MEASURES OF CENTRAL TENDENCY – MEAN, MEDIAN, MODE

One of the main tasks of the statistician is to summarise large volumes of data. It is useful to be able to find one score which is typical of a whole set of data, or a few figures which can describe its distribution.



After displaying data using a histogram or stem plot, we can make even more sense of the data by calculating what are called *summary statistics*. Summary statistics are used because they give us an idea about:

1. where the centre of the distribution is
2. how the distribution is spread out.

Finding the mean, median and mode are three different methods of arriving at a score which is typical or central to a set of data. Mean, median and mode are often called *measures of central tendency*.

- The **mode** is the most common or frequent score.
- The **median** is the middle score when the scores are placed in order from smallest to largest.
- The **mean** is calculated by adding all the scores and dividing by the number of scores. This is what most people call the 'average'.

The mean

The most commonly used measure is the *mean*.

To find the mean all the scores are added together, then the total is divided by the number of scores. The symbol $\bar{x}$ is commonly used to denote the mean. The operation of finding the mean could be written as a formula:

$$\bar{x} = \frac{\sum x_i}{n}.$$

The $\sum$ symbol is the Greek capital letter sigma. In maths it is used to signify 'the sum of'. So, this formula could be read as: 'The mean equals the sum of all the scores (x) divided by the number of scores (n)'.

Example

The following data give the number of pets kept in each of 10 different households:

3, 5, 4, 4, 2, 3, 0, 1, 4, 5

We find the mean as follows:

$$\text{Mean} = \bar{x} = \frac{\sum x_i}{n} = \frac{3+5+4+4+2+3+0+1+4+5}{10} = 3.1$$

The median

The *median* is the midpoint of a set of data. Half the data values lie below the median and half the data values lie above the median.

The position of the median is the $\frac{n+1}{2}$th data value, where n is the total number of data values.

Example

Consider the set of data: 2 5 6 8 11 12 15. These data are in ordered form (that is, from lowest to highest). There are 7 observations. The median in this case is the middle or fourth score; that is, 8.

Example

Consider the set of data: 1 3 5 6 7 8 8 9 10 12. These data are in ordered form also; however, in this case there is an even number of scores, that is, there are 10 scores. The median in this case lies halfway between the 5th score (7) and the 6th score (8). So the median is 7.5. (Alternatively, median equals $\frac{7+8}{2}$ = 7.5)

A stem plot provides a quick way of locating a median since the data in a stem plot are already ordered.

Example

Consider the stem plot below which contains 22 observations. What is the median?

Stem	Leaf
2	3 3
2	5 7 9
3	1 3 3 4 4
3	5 8 9 9
4	0 2 2
4	6 8 8 8 9

Key: 3|4 = 34

There are 22 scores. The median has 11 scores above and 11 scores below. Thus it is half way between the 11th and 12th score. The 11th score is 35 and the 12th score is 38. $\frac{35+38}{2}$ gives 36.5. Therefore, the median (half-way point) of this set of scores is 36.5.

The mode

The *mode* of a group of scores is the score that occurs most often. That is, it is the score with the highest frequency.

In the case of the household pets data, a score of 4 occurs three times, making it the most frequently occurring score. It is therefore the mode of the data.

In some cases there will be two or more scores which occur equally 'most often'. In such cases all of them are modes (- do not attempt to average them). However, if every data value occurs the same number of times then there is no mode.

Example

The following data give the number of hours spent on homework by 8 students:

2, 2, 3, 0, 1, 1, 5, 1

a) Mean $\text{Mean} = \frac{2+2+3+0+1+1+5+1}{8} = 1.875$

b) Arranging in order: 0, 1, 1, 1, 2, 2, 3, 5 gives a median of 1.5

c) The mode is 1 (there are three 1s)

If the data have been presented as a frequency table, then our techniques have to be adapted a little in order to find the mean, median and mode.

Example

The following data show the number of cinema visits made by each of 20 students.

Number of visits	0	1	2	3	4
Frequency	6	7	4	2	1

This data can be arranged in a frequency table.

When calculating the mean of data which is represented in a frequency table the following formula applies.

$$\text{Mean, } \bar{x} = \frac{\sum fx}{\sum f}$$

The use of this formula requires an extra column to be added to the frequency table, $f \times x$.

No of visits	Freq	$f \times x$	Cumul freq
0	6	0	6
1	7	7	13
2	4	8	17
3	2	6	19
4	1	4	20
Total		25	

The column $f \times x$ allows for easy calculation of the sum of the scores, in this case, 25.

The cumulative frequency (cumul freq) column allows us to find the median score.

a) To find the mean of the data, divide the total of all scores by the number of students.

Mean = 25 ÷ 20 = 1.25

b) The scores are already arranged in order on the frequency distribution table. The cumulative frequency column tells us that the 6th score was a 0. Then all the scores up to the 13th score were all the number 1. We are interested in only the 10th and 11th scores which both must be 1. So the median is 1.

c) The mode is the score with the highest frequency. This is easy to find when the data are presented as a frequency distribution table. For the cinema data the score of 1 occurred 7 times and no other score occurred as often. The mode is 1.

The effect of extreme values

An *extreme value* (or *outlier*) is a score which is considerably different from the rest of the data in the set. The presence of extreme values may affect the representative nature of any statistics calculated.

[We can make use of the '**Q1 – 1.5 x IQR' and 'Q3 + 1.5 x IQR**' criteria for formally identifying possible outliers – covered later].

Example

Consider again the number of household pets data that were introduced earlier in this section: 3, 5, 4, 4, 2, 3, 0, 1, 4, 5.

We have already determined the mean, median and mode:

Mean = 3.1 Median = 3.5 Mode = 4.

Let's include an extra household in this survey — the household of dear old Mrs. Kindheart and her 99 cats! How would this extra piece of data affect the mean, median and mode?

There would be 11 pieces of data: 3, 5, 4, 4, 2, 3, 0, 1, 4, 5, 99.

We would find that:

Mean = 11.82 Median = 4 Mode = 4. (Check these calculations for yourself)

The inclusion of an extreme value has dramatically increased the mean of the data, marginally increased the value of the median and left the mode unchanged. The important point to learn from this illustration is that when a set of data includes extreme values then the mean may not be truly representative of typical scores in the set. In the case of the household pets data none of the houses had a number of pets like 11.82. Under such circumstances the median or mode may be a more useful way of relating the notion of a typical score.

EXERCISE 2

1. Find the mean, median and mode of each of the following sets of data.

	Mean	Median	Mode
6, 8, 4, 7, 6, 7, 6			
25, 27, 29, 27, 26, 28, 29, 28, 27, 28			
5.6, 5.2, 5.4, 5.3, 5.8, 5.4, 5.3, 5.4			

2. For each of the following frequency distributions, find:

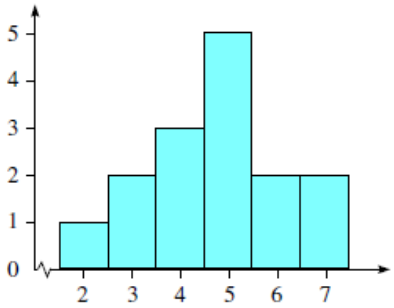
(i) the number of scores represented

(ii) the mean (to 1 decimal place)

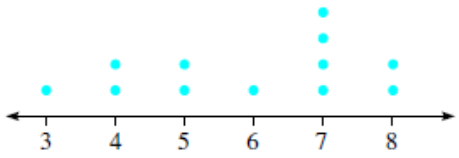
(iii) the median

(iv) the mode

a)



b)



c)

Key: 4|7 means 47%

Stem	Leaf					
4	4	7	7	8		
5	2	6	8	9	9	
6	1	3	5	5	7	8
7	0	2	3	4	5	
8	3	7	8			
9	2	8				

3. Students were surveyed on the number of compact discs (CDs) they had purchased in the last 6 months, with the results shown.

No. of CDs	No. of students
0	6
1	7
2	8
3	10
4	9
5	5
6	5

Add two columns to this table to help you answer the following questions.

a) How many students were included in the survey?

b) What was the modal (mode) number of CDs purchased?

c) Calculate the mean number of CDs purchased.

d) What was the median number of CDs purchased?

e) Construct a histogram of the results and comment on the distribution.

4. Meg's Matches has 'average contents 50' printed on each box. A quality controller counted the contents of a sample of 160 matchboxes. The results are displayed in the frequency table below.

Number of matches (x)	Frequency (f)
48	10
49	45
50	52
51	39
52	9
53	5

a) Calculate the average number of matches for the sample (to 1 dec. place).

b) Is the claim 'average contents 50' justified? Give a reason for your answer.

5. The police used radar to check the speeds of motor vehicles driving in a 40 km/h zone outside of a local primary school one morning. They recorded the results in the classed frequency table below.

Speed (km/h)	Midpoint (x)	Frequency (f)	$f \times x$
36-40		64	
41-45		36	
46-50		18	
51-55		15	
56-60		11	
61-65		5	
	$\sum =$		

$$\text{Mean, } \bar{x} = \frac{\sum fx}{\sum f}$$

a) How many motor vehicles had their speeds checked?

b) Calculate the mean speed of the vehicles (using the class midpoint as x), correct to 2 decimal places. Explain why this answer can only be considered as an estimate.

6. Ten houses were sold this week in Darwin for the following prices.

$376 000	$1 200 000	$270 000	$308 000	$372 000
$409 000	$387 000	$582 000	$460 000	$238 000

a) Calculate the mean house price.

b) Calculate the median house price.

c) Which measure is higher, the mean or the median?

d) Which measure of centre is more appropriate for describing the average house price? Explain your reasoning.

7. Mark and Steve's batting scores for six innings of cricket are shown below.

Mark:	45	48	53	38	32	40	51
Steve:	23	57	6	125	65	5	37

a) Calculate the mean score for each player (to 1 decimal place).

b) Which player is better if you use the mean?

c) Find the median score for each player.

d) Which player is better if you use the median?

e) Which player would you rather have in your team? Justify your answer.

RANGE AND INTERQUARTILE RANGE

The mean, median and mode are measures of centre for a data set. There are three summary statistics that are measures of spread: the **range**, the **interquartile range** (IQR) and the **standard deviation**.

The range

The range is the simplest measure of spread. It is also the easiest of this group of summary statistics to calculate.

The *range* of a set of data is the difference between the highest and lowest values in that set. It is usually not too difficult to locate the highest and lowest values in a set of data.

Example

The ages of the patients who attended the casualty department of an inner suburban hospital on one afternoon are shown below.

14 3 27 42 19 17 73 60 62 21 23 2 5 58 33 19 81 59 25 17 69

The highest score is 81 and the lowest score is 2.

The range is 81 – 2 = 79.

While the range gives us some idea about the spread of the data it is not terribly informative since it gives us no indication of how the data are distributed between the highest and lowest values.

The interquartile range

We have seen that the median divides a set of data in half. Similarly, quartiles divide a set of data in quarters. The symbols used to refer to these quartiles are $Q1$, $Q2$ and $Q3$. The middle quartile, $Q2$, is the median.

The interquartile range IQR = $Q3 - Q1$

The interquartile range gives us the range of the middle 50% of values in our set of data.

There are four steps to locating $Q1$ and $Q3$.

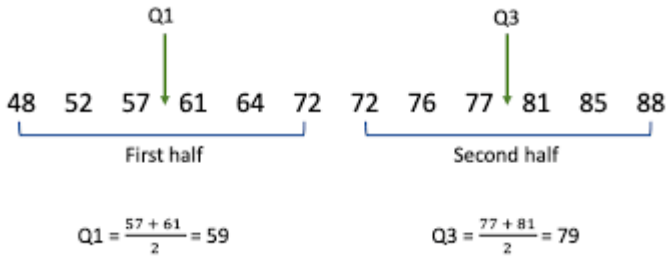
Step 1. Write down the data in ordered form from lowest to highest.

Step 2. Locate the median; that is, locate $Q2$.

Step 3. Now consider just the lower half of the set of data. Find the middle score. This score is $Q1$.

Step 4. Now consider just the upper half of the set of data. Find the middle score. This score is $Q3$.

Case 1 even number of observations

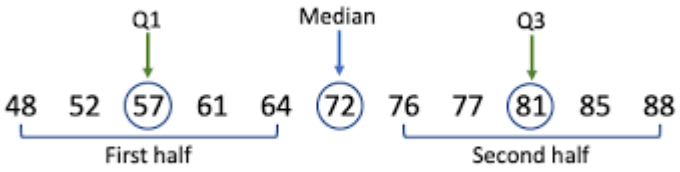


The data is already ordered. The median is between the 6th and 7th data points. The median is 72.

Consider the lower half of the set. The middle score is between the 3rd and 4th data values. Q1 is 59.

Consider the upper half of the set. The middle score is between the 3rd and 4th data values. Q3 is 79.

Case 2 odd number of observations



The data is already ordered. The median is 72.

Consider the lower half of the set. The middle score is 57, so $Q1 = 51$.

Consider the upper half of the set. The middle score is 81, so $Q3 = 81$.

The interquartile range (IQR) is 81 – 51 = 30

The middle 50% of the data values have a range of 30.

Example

The ages of the patients who attended the casualty department of an inner suburban hospital on one afternoon are shown below.

14 3 27 42 19 17 73 60 62 21 23 2 5 58 33 19 81 59 25 17 69

Find the interquartile range of these data.

Order the data.

2 3 5 14 17 17 19 19 21 23 25 27 33 42 58 59 60 62 69 73 81

Find the median.

The median is 25 since ten scores lie below it and ten lie above it.

Find the middle score of the lower half of the data.

For the scores 2 3 5 14 17 17 19 19 21 23, the middle score is 17.

So, *Q*1 is 17.

Find the middle score of the upper half of the data.

For the scores 27 33 42 58 59 60 62 69 73 81, the middle score is 59.5.

So, *Q*3 is 59.5.

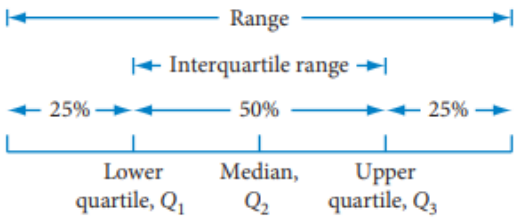
Calculate the interquartile range.

IQR: *Q*3 - *Q*1 = 59.5 - 17 = 42.5

remember

1. The median is the midpoint of a set of data. Half the data are less than or equal to the median.
2. When there are n observations in a set of ordered data, the median can be located at the $\left(\frac{n+1}{2}\right)$th position.
3. The interquartile range IQR $= Q_3 - Q_1$.
4. The interquartile range gives us the range of the middle 50% of values in our set of data.
5. There are four steps to locating Q_1 and Q_3.
   - Step 1. Write down the set of data in ordered form from lowest to highest.
   - Step 2. Locate the median, that is, locate Q_2.
   - Step 3. Now consider just the lower half of the set of data. Find the middle score. This score is Q_1.
   - Step 4. Now consider just the upper half of the set of data. Find the middle score. This score is Q_3.
6. The range of a set of data is the difference between the highest and lowest values in that set.

The **range** represents the total spread of scores but it is not a good measure of spread if there are outliers. The **interquartile range** is not affected by outliers, because it measures the range of half of the data.



Outliers

The IQR can be used to determine outliers in a set of data. If a data entry is 1 and a half times the IQR away from Q1 or Q3, it can be considered an outlier.

'Q1 – 1.5 x IQR' and 'Q3 + 1.5 x IQR' An outlier will lie outside this interval.

Example

Consider the data set:

1, 2, 4, 5, 6, 8, 18

We have Q1 = 2, Q2 = 5, Q3 = 8.

IQR = Q3 – Q1 = 6

1.5 * IQR = 9

8 + 9 = 17. Therefore, any value above 17 is an outlier.

2 – 9 = -7. Therefore, any value below -7 is an outlier.

We can conclude that 18 is an outlier.

EXERCISE 3

1. Write down the median and the range of the sets of data shown in the following stem and leaf plots.

a)

Stem	Leaf
0	7
1	2 3
2	2 4 5 7 9
3	0 2 3 6 8 8
4	4 7 8 9 9
5	2 7 8
6	1 3

The key for each stem plot is 3|4 = 34

b)

Stem	Leaf
0	0 0 1 1
0	2 2 3 3
0	4 4 5 5 5 5 5 5 5 5
0	6 6 6 6 7
0	8 8 8 9
1	0 0 1
1	3 3
1	5 5
1	7
1	

2. For each of the following sets of data find the median, the interquartile range and the range.

a) 16 19 12 11 8 6 7 15 26 32 32 18 15 43 51 31 29 23 45 23

b) 1.2 6.1 2.3 3.7 4.1 5.4 2.4 3.7 1.5 5.2 3.7 3.8 6.1 6.3 2.4 7.1 3.6 4.9 1.2

3. Find the median and inter-quartile range of the following sets of data.

a) On the 9th of August, the number of cars that stopped at the drive-in area at a McBurger restaurant during each hour (from 7.00 am until 10.00 pm) is shown below.

14 18 8 9 12 24 25 15 18 25 24 21 25 24 14

b) At the nearby Kenny's Fried Chicken restaurant on the same day, the number of cars stopping during each hour that *it* was open is shown below.

7 9 13 16 19 12 11 18 20 19 21 20 18 10

c) Which appears to be the busiest take-away? Explain.

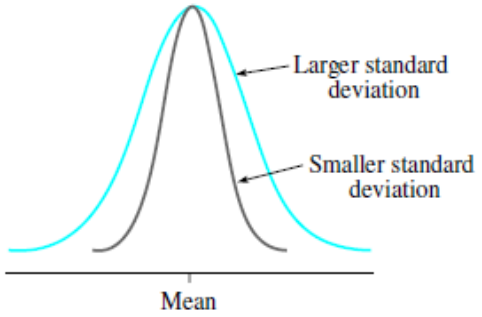
4. Find the range, Q1, Q2, Q3, IQR and comment on any outliers of the following set of data.

a) 10, 12, 15, 15, 16, 18, 19, 20, 21, 21, 23, 25, 27, 28, 30

b) 18, 21, 25, 30, 35, 37, 40, 41, 55, 63, 71, 72, 80, 104, 125, 215

STANDARD DEVIATION

The standard deviation gives us a measure of how data are spread around the mean. It is a sophisticated measure that gives us a lot of information about the spread of the scores, in particular how far the scores are from the mean.



These two distributions have the same mean but different standard deviations.

The larger the standard deviation, the more spread out the scores are from the mean.

A large standard deviation indicates that the scores are more spread out, while a small standard deviation indicates scores that are closer together, or more consistent.

Standard deviation is found using the formula:

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

To find the standard deviation we use the following method.

Data: 8, 10, 11, 12, 12, 13

Mean $= \frac{\sum x}{n} = \frac{66}{6} = 11$

We then set up a table:

Score	Deviation $(x-\bar{x})$	$(x-\bar{x})^2$
8	-3	9
10	-1	1
11	0	0
12	1	1
12	1	1
13	2	4
	Sum = 0	Sum = 16

Standard deviation is found from the formula:

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{16}{5}} = 1.8$$

Example

Consider the two sets of data below. Note: Both these sets are in order.

Set 1: 3, 4, 5, 7, 7, 7, 8, 10, 12

Set 2: 1, 2, 4, 7, 7, 7, 9, 12, 15

Set 1

Mode = 7

Median = 7

Mean = $\frac{3+4+5+7+7+7+8+10+12}{9}$ = 7

Spread

Range = 12 – 3 = 9

IQR = 9 – 4.5 = 4.5

Score	Deviation $(x-\bar{x})$	$(x-\bar{x})^2$
3	-4	16
4	-3	9
5	-2	4
7	0	0
7	0	0
7	0	0
8	1	1
10	3	9
12	5	25
		$\sum = 64$

s = $\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ = $\sqrt{\frac{64}{9}}$ = 2.7

Set 2

Mode = 7

Median = 7

Mean = $\frac{1+1+4+7+7+7+9+12+15}{9}$ = 7

Spread

Range = 15 – 1 = 14

IQR = 10.5 – 3 = 7.5

Score	Deviation $(x-\bar{x})$	$(x-\bar{x})^2$
1	-6	36
1	-6	36
4	-3	9
7	0	0
7	0	0
7	0	0
9	2	4
12	5	25
15	8	64
		$\sum = 174$

s = $\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ = $\sqrt{\frac{174}{9}}$ = 4.4

So, while both sets of scores have identical mean, median and mode the range, IQR and standard deviation of Set 2 are larger verifying that Set 2 has a bigger spread.

EXERCISE 4

1. Using the technique shown above compare the two sets of data below.

Set 1: 17, 20, 19, 25, 29, 27, 28, 25, 18, 15

Set 2: 12, 14, 15, 12, 14, 19, 17, 15, 18, 20

Arrange the numbers in order first.

Calculate the mean. Then, calculate the standard deviation. $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

SET 1

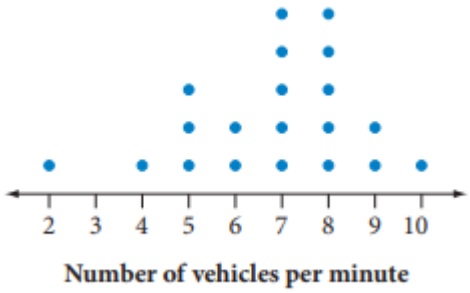
Score	Deviation $(x-\bar{x})$	$(x-\bar{x})^2$

SET 2

Score	Deviation $(x-\bar{x})$	$(x-\bar{x})^2$

2. Use an online standard deviation calculator (such as https://www.calculatorsoup.com/calculators/statistics/standard-deviation-calculator.php), table or spreadsheet to answer the following:

a) find the mean and *sample* standard deviation of the following data set.



b) How many scores were within one standard deviation of the mean?

3.

The cricket selectors are trying to choose between two pairs of indoor cricket batsmen for the state team. Both pairs have an average (mean) of 33 runs. Two sets of results for the pairs of batsmen (in runs) are:

Pair A	34	30	36	35	29	34
Pair B	41	26	37	35	25	34

The standard deviation for pair A is 2.58 runs, while the standard deviation for pair B is 5.74 runs.

Which pair is more consistent?

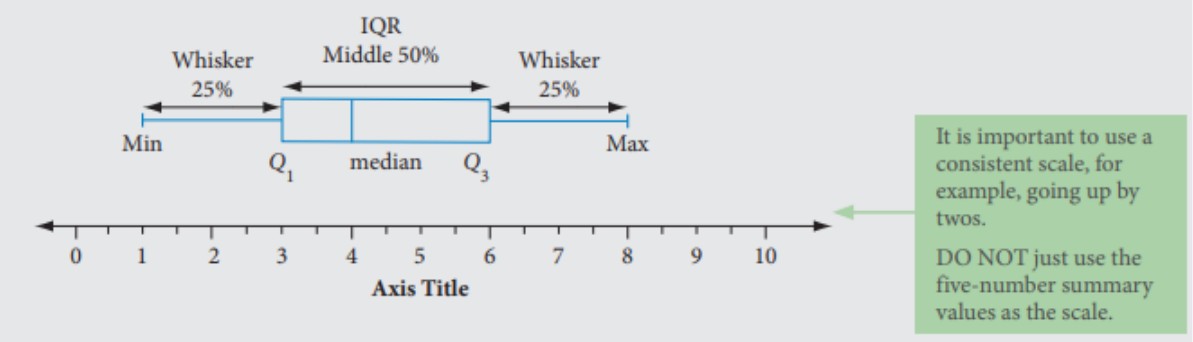
BOXPLOTS

The summary statistics that we looked at in the previous section can be illustrated very neatly in a special diagram known as a *boxplot* (or *box-and-whisker* diagram). The diagram is made up of a box with straight lines (whiskers) extending from opposite sides of the box.

A boxplot displays the minimum and maximum values of the data together with the quartiles and is drawn with a scale. The length of the box gives us the interquartile range. A boxplot gives us a very clear visual display of how the data are spread out.

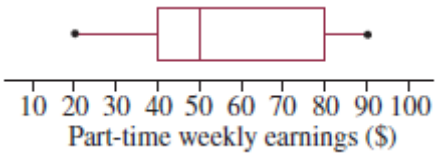
To construct a boxplot you need the five-number summary which consists of the:

Minimum, Q1, median, Q3, maximum



Example

The boxplot below shows the distribution of the part-time weekly earnings of a group of Year-11 students. Write down the range, the median and the interquartile range for these data.



Range = Maximum value - Minimum value.

The minimum value is 20 and the maximum value is 90.

Range = 90 - 20 = 70

The median is located at the bar inside the box.

Median = 50

The ends of the box are at 40 and 80.

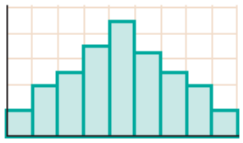
IQR = *Q*3 - *Q*1 *Q*1 = 40 and *Q*3 = 80

IQR = 80 - 40 = 40

Previously, we noted three general types of shape for histograms and stem plots: symmetric, negatively skewed and positively skewed. It is useful to compare the corresponding boxplots of distributions with such shapes.

In the figure below a symmetric distribution is represented in the histogram and in the boxplot.

The characteristics of this boxplot are that the whiskers are about the same length and the median is located about halfway along the box.

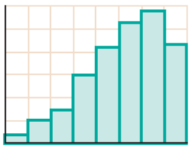


Symmetric histogram

Symmetric boxplot

The figure below shows a negatively skewed distribution. In such a distribution, the data peak to the right on the histogram and trail off to the left.

In corresponding fashion on the boxplot, the bunching of the data to the right means that the left-hand whisker is longer and the right-hand whisker is shorter; that is, the lower 25% of data are sparse and spread out whereas the top 25% of data are bunched up. The median occurs further towards the right end of the box.

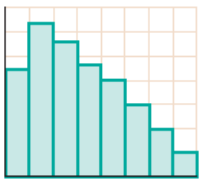


Negatively skewed histogram

Negatively skewed boxplot

In the figure below we have a positively skewed distribution. In such a distribution, the data peak to the left on the histogram and trail off to the right.

In corresponding fashion on the boxplot, the bunching of the data to the left means that the left-hand whisker is shorter and the right-hand whisker is longer; that is, the upper 25% of data are sparse and spread out whereas the lower 25% of data are bunched up. The median occurs further towards the left end of the box.

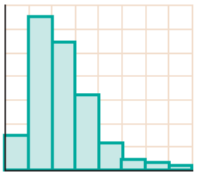


Positively skewed histogram

Positively skewed boxplot

Example

Explain whether or not the histogram and the boxplot shown below could represent the same data.



The histogram shows a distribution which is positively skewed.
The boxplot shows a distribution which is approximately symmetric.

The histogram and the boxplot could not represent the same data since the histogram shows a distribution that is positively skewed and the boxplot shows a distribution that is approximately symmetric.

Example

The results out of 20 of oral tests in a Year-12 Indonesian class are shown. Display these data using a boxplot.

15 12 17 8 13 18 14 16 17 13 11 12

Rank the scores:

8 11 12 12 13 13 14 15 16 17 17 18

Find *Q*2, the median score is 13.5

The lower half of the scores are 8 11 12 12 13 13.

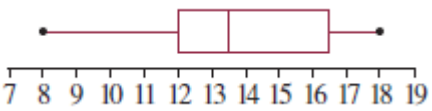
So, *Q*1 is 12

The upper half of the scores are 14 15 16 17 17 18.

So, *Q*3 is 16.5

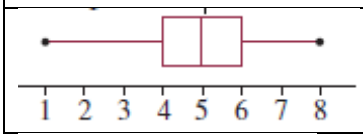
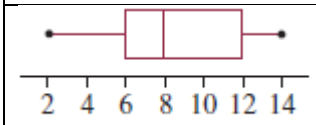
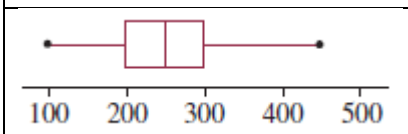
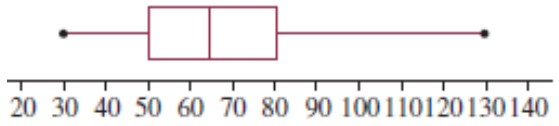
The lowest score is 8. The highest score is 18.

Using these 5 key scores, draw the boxplot

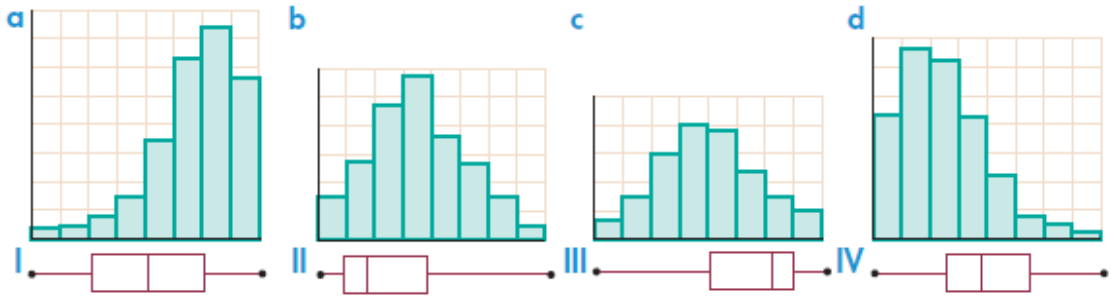


EXERCISE 5

1. For the boxplots shown, write down the range, the interquartile range and the median of the distributions which each one represents.

	Range	Interquartile range	Median
			
			
			
			

2. Each of the histograms shown below is labelled with a letter and each of the boxplots is labelled with a number. Match each histogram with a boxplot which could show the same distribution.



3. For each of the following sets of data construct a boxplot.

a) 3 5 6 8 8 9 12 14 17 18

b) 11 13 15 15 16 18 20 21 22 21 18 19 20 16 18 20

4. The maximum daily temperatures (in degrees C) for the month of October in Melbourne are:

18 26 28 23 16 19 21 27 31 23 24 26 21 18 26

27 23 21 24 20 19 25 27 32 29 21 16 19 23 25

27

Represent this data in a boxplot. *You may use a spreadsheet to help you sort the data and find the five number summary.*

WEEK 2/3 INVESTIGATION

Find sets of integers that satisfy the following:

show working out and reasoning where necessary

1. Three numbers with mean 3 and mode 2

_____ _____ _____

2. Three numbers with mean 7 and mode 10

_____ _____ _____

3. Three numbers with mean 8, median 10 and range 8

_____ _____ _____

4. Four numbers with mean 7.5, mode 6 and median 7

_____ _____ _____ _____

5. Four numbers with mean 6, median 6.5 and range 11

_____ _____ _____ _____

6. Five numbers with mean 4, mode 3 and range 9

_____ _____ _____ _____ _____

7. Five numbers with mean 4, mode 2 and range 6

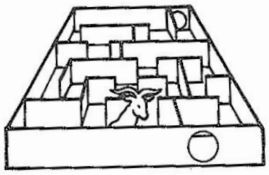
_____ _____ _____ _____ _____

8. Five numbers with mean 7, mode 7 and range 10

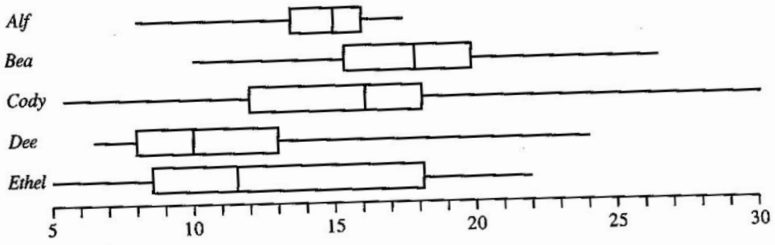
_____ _____ _____ _____ _____

Task 2

David's dilemma



Five mice—Alf, Bea, Cody, Dee and Ethel—are timed travelling through a maze. Each mouse is timed on 100 different occasions. Only one mouse is in the maze at any time. Here are box and whisker plots showing the times taken by the five mice to complete the maze.



Taking all of the available information into account, rank the five mice from fastest to slowest.

Rank the five mice from fastest to slowest. Give reasons to justify your ranking.

MARKING RUBRIC

CRITERIA	EXPECTATIONS	POSS	MULT	GIVEN	TOTAL
Practical	Student completes practical work of the brief to an acceptable standard set by the teacher.	**2**	**3**		**/6**
Investigation	Student completes the investigation of the brief to an acceptable standard set by the teacher.	**2**	**2**		**/4**
Reasoning and communications	Student responses are accurate and appropriate in presentation of mathematical ideas in different contexts, with clear and logical working out shown.	**4**	**-**		**/4**
Concepts and techniques	Student submitted work selects and applies appropriate mathematical modelling and problem solving techniques to solve practical problems, and demonstrates proficiency in the use of mathematical facts, techniques and formulae.	**4**	**-**		**/4**
	Submission Guidelines				
Timeliness	Student submits the exercises and portfolio task by the set deadline. See scoring guidelines for specific details.	**2**	**-**		**/2**
		FINAL			**/20**

Student Reflection:
How did you go with this week's work?

What was interesting?

What did you find easy?

What do you need to work on?