DATA HANDLING (1)

LESSON 41

Learning Outcomes and Assessment Standards

Learning Outcome 4: Data handling and probability
Assessment Standard AS 1(a)
Calculate and represent measures of central tendency and dispersion in univariate numerical data by:

- five number summary
- box and whisker diagrams
- ogives
- variance and standard deviation.

Overview

Overview

In this lesson you will:

- Revise the terms mean, median, mode and quartiles
- Learn about five number summaries
- Learn about box and whisker plots.

Lesson

Lesson

Revision of Grade 10 concepts

The mean

The mean of a set of data is defined as $\bar{x}$ (x bar).

$$\bar{x} = \frac{\text{sum of the values}}{\text{number of the values}} = \frac{\Sigma x}{n}$$

Example

Example

Calculate the mean of the following data:

12, 13, 13, 15, 16, 16, 16, 16, 16, 17, 17, 18, 18, 18, 18

$$\bar{x} = \frac{\Sigma x}{n} = 12+13+13+15+16+16+16+16+16+17+17+18+18+18+18 = \frac{239}{15} = 15{,}9$$

The mode

The mode is the most commonly occurring observation.

Example

Example

Class A	1	1	1	2	4	5	7	9	10							
Class B	1	1	1	2	4	5	5	5	5	7	7	8	8	8	9	10

For Class A, the mode is 1. For Class B, the mode is 5.

Consider the following set of marks for a Class C:

Class C	0	1	1	2	2	2	3	4	4	4	5	5	6	7	9	10

There are two modes in this set of data: 2 and 4 (they appear the same number of times and are the most frequently occurring marks. The data is said to be **bimodal**.

Quartiles

Quartiles are measures of dispersion (or spread) around the median, which is a better

measure of central tendency. The median divides the data into two halves. The quartiles further subdivide the data into quarters.

There are therefore three quartiles:

The lower quartile (Q_1):	This is the median of the lower half of the values. We also call this the 25th percentile.
The median (Q_2):	The value that divides the data into halves. We also call this the 50th percentile.
The upper quartile(Q_3):	This is the median of the upper half of the values. We also call this the 75th percentile.

Useful formulae to determine the position of the quartiles are:

The lower quartile (q_1): $\frac{1}{4}(n+1)$

The median (q_2)::

$\frac{1}{2}(N+1)$

The upper quartile(q_2): $\frac{3}{4}(n+1)$

Example

Example 1 (Odd number of values)

Note:
If the number of values (n) in the data set is odd, the median will always be part of the data set.

To find the median we use $\left(\frac{n+1}{2}\right)$

The lower and upper quartiles will be part of the data set if $\frac{1}{4}(n+1)$ and $\frac{3}{4}(n+1)$ work out to be whole numbers.

The lower and upper quartiles will not be part of the data set if $\frac{1}{4}(n+1)$ and $\frac{3}{4}(n+1)$ do not work out to be whole numbers.

(a) Consider the following set of marks obtained on a class test out of 10 marks. The number of marks is odd.

2	2	**3**	4	5	**5**	6	7	**7**	8	9
		Lower quartile Q_1			Median Q_2			Upper quartile Q_3		

The position of $Q_2 = \frac{1}{2}(11+1) = 6$.

The **median** of the data is 5 (the 6th value).

The position of $Q_1 = \frac{1}{4}(11+1) = 3$

The **lower quartile** of the data is 3 (the 3rd value). It is a part of the data set.

The position of $Q_3 = \frac{3}{4}(11+1) = 9$

The **upper quartile** of the data is 7 (the 9th value). It is part of the data set.

(b) Consider the following set of 13 marks obtained on a class test out of 10 marks:

2	3	4	5	5	5	6	7	7	8	9	10	10

The position of $Q_2 = \frac{1}{2}(13+1) = 7$th position.

The **median** of the data is the 7th value:

$Q_2 = 6$

The position of $Q_1 = \frac{1}{4}(13 + 1) = 3$, 5th position (In the middle of point 3 and 4)

The **lower quartile** of the data is the average between the 3rd and 4th value:

$Q_1 = \frac{4 + 5}{2} = 4{,}5$ (not part of the data set)

The position of $Q_3 = \frac{3}{4}(13 + 1) = 10$, 5th position

The **upper quartile** of the data is the average between the 10th and 11th value:

$Q_3 = \frac{8 + 9}{2} = 8{,}5$ (not part of the data set)

2	3	4	4,5	5	5	5	6	7	7	8	8,5	9	10	10
			Q_1				Q_2				Q_3			

Example

Example 2 (Even number of values)

Note:
If n is even, the median will not be part of the data set.

If n is even and $\frac{n}{2}$ is even, the lower and upper quartiles will not be values in the data set. Round off the position values up or down to the nearest whole number.

If n is even and $\frac{n}{2}$ is odd, the lower and upper quartiles will be values in the data set.

(a) Consider the following set of 12 marks obtained by a class on a class test out of 100 marks. The number of marks is even.

20	32	43	54	55	61	73	78	89	90	91	98

The position of $Q_2 = \frac{1}{2}(12 + 1) = 6{,}5$ (average of the 6th and 7th value).

The **median** of the data is $Q_2 = \frac{61 + 73}{2} = 67$

Since n is even and since $\frac{n}{2} = \frac{12}{2} = 6$ which is even, the lower and upper quartiles will not be values in the data set.

The position of $q_1 = \frac{1}{4}(12 + 1) = 3{,}25$ (average of the 3rd and 4th value).

The **lower quartile** of the data is $q_1 = \frac{43 + 54}{2} = 48{,}5$

The position of $q_3 = \frac{3}{4}(12 + 1) = 9{,}75$ (average of the 9th and 10th value).

The **upper quartile** of the data is $q_3 = \frac{89 + 90}{2} = 89{,}5$

20	32	43	48,5	54	55	61	67	73	78	89	89,5	90	91	98
			Lower quartile 48,5				Median 67				Upper quartile 89,5			

(b) Consider the following set of 10 marks obtained by a class on a class test out of 150 marks. The number of marks is even.

12	60	95	105	120	125	130	135	140	142

The position of $Q_2 = \frac{1}{2}(10 + 1) = 5{,}5$th position (average of the 5th and 6th value).

The **median** of the data is $\frac{120 + 125}{2} = 122{,}5$

Since n is even and since $\frac{n}{2} = \frac{10}{2} = 5$ which is odd, the lower and upper quartiles will be values in the data set.

The position of $Q_2 = \frac{1}{4}(10 + 1) = 2{,}75$ (Round up to the 3rd value)

The **lower quartile** of the data is 95

The position of $q_3 = \frac{3}{4}(10 + 1) = 8{,}25$ (round down to the 8th value).

The **upper quartile** of the data is 135

12	60	95	105	120	122,5	125	130	135	140	142

Interquartile range (iqr)

The difference between the lower and upper quartile is called the Interquartile Range. It is a better measure of dispersion than the range because it is not affected by extreme values. It is based on the middle half of the data. It indicates how densely the data in the middle is spread around the median. Consider the previous example.

12	60	95	105	120	122,5	125	130	135	140	142

The Interquartile Range (IQR)= $Q_3 - Q_1 = 135 - 95 = 40$

Semi-interquartile range

The semi-interquartile range is half of the interquartile range.

The semi-IQR for the previous example is $\frac{Q3 - Q_1}{2} = \frac{135 - 95}{2} = \frac{40}{2} = 20$.

Activity

Activity 1

1. For each set of data, determine the quartiles:

A	2	3	5	7	9	10	11	13	15	16	17	18	19	21	22	23	25	32	
B	2	3	5	7	9	10	11	13	15	16	17	18	19	21	22	23			
C	2	3	5	7	9	10	11	13	15	16	17	18	19	21	22	23	25	32	34
D	2	3	5	7	9	10	11	13	15	16	17	18	19	21	22	23	25		

2. Class results for a test out of 30 are recorded in the table below.

10A	16	12	16	11	14	15	22	16	17	15	26	23	16	22	16	17	24	19		
10B	20	19	14	10	14	9	8	13	14	30	27	23	24	28	17	29	20	16	14	18
10C	5	20	14	12	7	2	12	21	14	26	14	14	12	14	21	24	14	14		

   (a) Calculate the mean for each class.

   (b) Calculate the mode for each class.

   (c) Calculate the median for each class.

   (d) Calculate the range for each class.

   (e) Calculate the lower quartile for each class.

   (f) Calculate the upper quartile for each class.

   (g) Calculate the interquartile range for each class.

   (h) Calculate the semi-interquartile range for each class.

3. A teacher has recorded the test marks of forty grade 10 learners. The test was out of 10. Draw a frequency table and then calculate the mean, median and mode for this data.

1	9	10	4	7	4	4	10

7	3	9	3	8	9	3	7
7	3	9	4	5	8	6	6
1	3	10	2	2	7	8	7
7	2	7	6	2	8	7	6

Lesson

Lesson

Five number summaries and box and whisker plots

Five number summaries and box and whisker plots help us to represent and analyse the spread of data about the median.

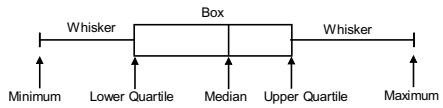
Five number summaries

The five number summary uses the following measures of dispersion:

- Minimum: the smallest value in the data
- Lower quartile: the median of the lower half of the values
- Median: the value that divided the data into halves
- Upper quartile: the median of the upper half of the values
- Maximum: the largest value in the data

Box and whisker plots

A box and whisker plot is a graphical representation of the five number summary.



Note:

- Half of the values lie between the minimum value and the median.
- Half of the values lie between the median and the maximum value.
- One quarter of the values lies between the minimum value and the lower quartile.
- One quarter of the values lies between the lower quartile and the median.
- One quarter of the values lies between the median and the upper quartile.
- One quarter of the values lies between the upper quartile and the maximum value.
- Half of the values lie between the lower quartile and upper quartile.

Example

Example

Consider the following set of marks for a class test (out of 10) for three classes.

CLASS A	1	1	2	2	3	3	4	4	4	6	7	8	8	9	10	10	10
CLASS B	1	2	4	4	4	4	5	7	8	8	8	8	9	9	9	10	10
CLASS C	1	2	3	3	3	4	5	5	5	6	6	7	7	7	8	9	10

For each class above, create a five number summary and hence a box and whisker plot.

CLASS A	1	1	2	2	2,5	3	3	4	4	4	6	7	8	8	8,5	9	10	10	10

Five number summary

Minimum: 1

Lower quartile (Q_1): Position of $Q_1 = \frac{1}{4}(17 + 1) = 4{,}5$

(average of 4th and 5th value)

$\therefore Q_1 = \frac{2 + 3}{2} = 2{,}5$

Median (M or Q_2): Position of $Q_2 = \frac{1}{2}(17 + 1) = 9$

(9th value)

$\therefore Q_2 = 4$

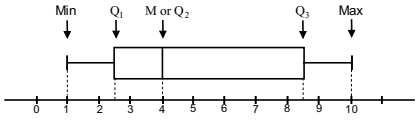
Upper quartile (Q_3): Position of $Q_3 = \frac{3}{4}(17 + 1) = 13{,}5$

(average of 13th and 14th value)

$\therefore Q_3 = \frac{8 + 9}{2} = 8{,}5$

Maximum: 10

Box and whisker plot



CLASS B	1	2	4	4	4	4	4	5	7	8	8	8	8	9	9	9	9	10	10

Five number summary

Minimum: 1

Lower quartile (Q_1): Position of $Q_1 = \frac{1}{4}(17 + 1) = 4{,}5$

(average of 4th and 5th value)

$\therefore Q_1 = \frac{4 + 4}{2} = 4$

Median (M or Q_2): Position of $Q_2 = \frac{1}{2}(17 + 1) = 9$

(9th value)

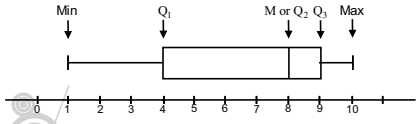
$\therefore Q_2 = 8$

Upper quartile (Q_3): Position of $Q_3 = \frac{3}{4}(17 + 1) = 13{,}5$

(average of 13th and 14th value)

$\therefore Q_3 = \frac{9 + 9}{2} = 9$

Maximum: 10



box and whisker plot

CLASS C	1	2	3	3	3	3	4	5	5	5	6	6	7	7	7	7	8	9	10

Five number summary

Minimum: 1

Lower quartile (Q_1): Position of $Q_1 = \frac{1}{4}(17 + 1) = 4{,}5$

(average of 4th and 5th value)

$\therefore Q_1 = \frac{3 + 3}{2} = 3$

Median (M or Q_2): Position of $Q_2 = \frac{1}{2}(17 + 1) = 9$

(9th value)

$\therefore Q_2 = 5$

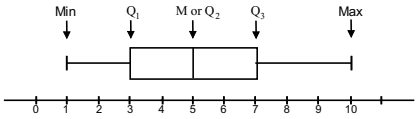
Upper quartile (Q_3): Position of $Q_3 = \frac{3}{4}(17 + 1) = 13{,}5$

(average of 13th and 14th value)

$\therefore Q_3 = \frac{7 + 7}{2} = 7$

Maximum: 10

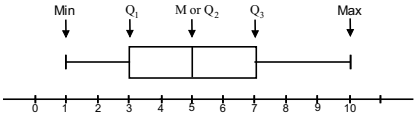
Box and whisker plot



Symmetrical and skewed data

- Symmetrical data set (relative to the median)

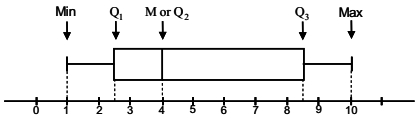
If the data to the left of the median balances with the data on the right, then the data is **symmetrical about the median**.



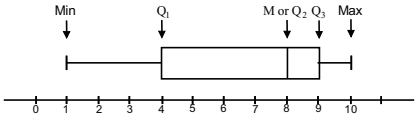
Consider, for example, CLASS C.

- Skewed data (relative to the median)

If the data is clustered predominantly to the right of the median, the data is said to be **skewed to the right.** Consider, for example, CLASS A.



If the data is clustered predominantly to the left of the median, the data is said to be **skewed to the left.** Consider, for example, CLASS B.



Activity 2

Activity

The number of points scored by four Formula One racing drivers over a number of races are given below:

A	1	1	1	2	6	6	8	8	8	8	10	10	10
B	1	2	6	8	8	8	8	8	8	10	10	10	–
C	1	1	2	2	4	4	6	6	8	8	10	–	–
D	2	2	2	4	4	6	6	8	8	10	10	10	–

(A) Calculate the mean for each of the drivers.

(B) List the five number summary for each driver.

(C) Draw a box and whisker plot for each driver.

(d) Discuss each driver's distribution of scores in terms of the spread about the median.

(e) Compare the performance results for each driver by using the information obtained above.