Single Variable Data Anayisis

Lesson 1

9G Mean, median and mode REVISION

The discipline of statistics involves collecting and summarising data. It also involves drawing conclusions and making predictions, which is why many of the decisions we make today are based on statistical analysis. The type and amount of product stocked on supermarket shelves, for example, is determined by the sales statistics and other measures such as average cost and price range.

- The two most commonly used measures of location are the **mean** and the **median**. These are also called '**measures of centre**' or '**measures of central tendency**'. Mean and median can only be applied to numerical data.

- The **mean** is sometimes called the '**arithmetic mean**' or the '**average**'. The formula used for calculating the mean, $\bar{x}$, is:

$$\bar{x} = \frac{\text{sum of data values}}{\text{number of data values}}$$

For example, in the following data set

5	7	2	5	1

the mean is $\frac{5+7+2+5+1}{5} = 20 \div 5 = 4$

- The **median** divides an ordered data set into two sets, each of which contain the same number of data values. It is often called the 'middle value'. The median is found by firstly ensuring the data values are in ascending order, then selecting the 'middle' value.

If the number of values is odd, simply choose the one in the middle.

2 2 3 4 6 9 9

median = 4

If the number of values is even, find the average of the two in the middle.

2 2 3 4 7 9 9 9

Median $= (4 + 7) \div 2 = 5.5$

The Mode

- The mode is the number with more occurrences in a data set or the number with the higher frequency.

- For example, in the following set of numbers: 2 5 2 6 3 1, the Mode will be 2 because it is the number that occurs the most in this particular set of numbers
- In this example: 2, 1, 4, 1, 5, 4. There is two modes 1 and 4 with 2 occurrences each.
- In this example 2 4 1 8 9 10. There is no mode because all scores have the same number of occurrences.

Example 11 Finding measures of centre

For the given data sets, find:

i the mean ii the median iii the mode

a 5 2 4 10 6 1 2 9 6

b 17 13 26 15 9 10

SOLUTION

a i $\text{Mean} = \dfrac{5+2+4+10+6+1+2+9+6}{9}$

$= 5$

ii 1 2 2 4 (5) 6 6 9 10

Median = 5

iii Mode = 2 and 6

EXPLANATION

Find the sum of all the numbers and divide by the number of values.

First, order the data. The median is the middle value.

The data set is bimodal since there are two numbers with the highest frequency.

3 For the given data sets, find the:

i mean ii median iii mode

a 7 2 3 8 5 9 8

b 6 13 5 4 16 10 3 5 10

c 12 9 2 5 8 7 2 3

d 10 17 5 16 4 14

e 3.5 2.1 4.0 8.3 2.1

f 0.7 3 2.9 10.4 6 7.2 1.3 8.5

g 6 0 −3 8 2 −3 9 5

h 3 −7 2 3 −2 −3 4

4 These data sets include an outlier. Write down the outlier, then calculate the mean and the median. Include the outlier in your calculations.

a 5 7 7 8 12 33

b 1.3 1.1 1.0 1.7 1.5 1.6 −1.1 1.5

c −58 −60 −59 −4 −64

5 Decide if the following data sets are bimodal.

a 2 7 9 5 6 2 8 7 4

b 1 6 2 3 3 1 5 4 1 9

c 10 15 12 11 18 13 9 16 17

d 23 25 26 23 19 24 28 26 27

Range

Tha range of a set of data is the difference between the highest and the lowest scors of the set.

Example 1) Find the range of 14, 21, 9, 32, 27, 15, 12, 30

Step 1) Put the data into an ordered list.

This gives us: 9, 12, 14, 15, 21, 27, 30, 32

Step 2) Check the number of data points in both lists is the same.

Both lists have 8 data points.

Step 3) The range is the difference or gap between the largest and smallest numbers.

Answer: the range is 32-9=23.

Find the range in each of the sets of data below by ordering the numbers and then finding the difference between the highest and lowest.

Data		Ordered list	Range
1)	45, 23, 17, 20, 27, 11	11, 17, 20, 23, 27, 45	45-11=34
2)	6, 23, 12, 19, 2, 7		
3)	19, 26, 23, 35, 29, 21		
4)	5, 2, 0, 8, 11, 16, 10		
5)	62, 78, 56, 61, 59, 83		
6)	73, 87, 65, 92, 89, 91		
7)	18, 6, 23, 2, 15, 21, 7		

Lesson 2.
Stem and Leaf Plots

Stem-and-leaf plots (or stem plots) are commonly used to display a single data set or two related data sets. They help to show how the data is distributed like a histogram but retain all the individual data elements so no detail is lost. The median and mode can be easily read from a stem-and-leaf plot because all the data sits in order.

https://www.youtube.com/watch?v=MUCvUgGfzdo

Example 13 Constructing and using a stem-and-leaf plot

Consider the following set of data.

0.3	2.5	4.1	3.7	2.0	3.3	4.8	3.3	4.6	0.1	4.1	7.5	1.4	2.4
5.7	2.3	3.4	3.0	2.3	4.1	6.3	1.0	5.8	4.4	0.1	6.8	5.2	1.0

a Organise the data into an ordered stem-and-leaf plot.
b Find the median.
c Find the mode.
d Describe the data as symmetrical or skewed.

SOLUTION

a

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

3 | 4 means 3.4

b $\text{Median} = \frac{3.3 + 3.4}{2}$
$= 3.35$

c Mode is 4.1.

d Data is approximately symmetrical.

EXPLANATION

The minimum is 0.1 and the maximum is 7.5 so stems range from 0 to 7. Place leaves in order from smallest to largest. Since some numbers appear more than once, e.g. 0.1, their leaf (1) appears the same number of times.

There are 28 data values. The median is the average of the two middle values (the 14th and 15th values).

The most common value is 4.1.

The distribution of numbers is approximately symmetrical about the stem containing the median.

https://www.youtube.com/watch?v=pfujiA5Mk_U

3 For each of the following data sets:

i organise the data into an ordered stem-and-leaf plot

ii find the median

iii find the mode

iv describe the data as symmetrical or skewed

a 41 33 28 24 19 32 54 35 26 28 19 23 32 26 28

b 31 33 23 35 15 23 48 50 35 42 45 15 21 45
51 31 34 23 42 50 26 30 45 37 39

c 34.5 34.9 33.7 34.5 35.8 33.8 34.3 35.2 37.0 34.7
35.2 34.4 35.5 36.5 36.1 33.3 35.4 32.0 36.3 34.8

d 167 159 159 193 161 164 167 157 158 175 177 185
177 202 185 187 159 189 167 159 173 198 200

4 The number of vacant rooms in a motel each week over a 20-week period is shown below.

12 8 11 10 21 12 6 11 12 16 14 22 5 15 20 6 17 8 14 9

a Draw a stem-and-leaf plot of this data.

b In how many weeks were there fewer than 12 vacant rooms?

c Find the median number of vacant rooms.

Example 14 Constructing back-to-back stem-and-leaf plots

A shop owner has two jeans shops. The daily sales in each shop over a 16-day period are monitored and recorded as follows.

Shop A

3 12 12 13 14 14 15 15 21 22 24 24 24 26 27 28

Shop B

4 6 6 7 7 8 9 9 10 12 13 14 14 16 17 27

a Draw a back-to-back stem-and-leaf plot with an interval of 10.

b Compare and comment on differences between the sales made by the two shops.

SOLUTION

a

Shop A		Shop B
3	0	4 6 6 7 7 8 9 9
5 5 4 4 3 2 2	1	0 2 3 4 4 6 7
8 7 6 4 4 4 2 1	2	7

1 | 3 means 13

b Shop A has the highest number of daily sales. Its sales are generally between 12 and 28, with one day of very low sales of 3. Shop B sales are generally between 4 and 17 with only one high sale day of 27.

EXPLANATION

The data for each shop is already ordered. Stems are in intervals of 10. Record leaf digits for Shop A on the left and for Shop B on the right.

Look at both sides of the plot for the similarities and differences.

https://www.youtube.com/watch?v=_84JyTh7tZo

5 For each of the following sets of data:

i draw a back-to-back stem-and-leaf plot

ii compare and comment on the difference between the two data sets

a

Set A:	46	32	40	43	45	47	53	54	40	54	33	48	39	43
Set B:	48	49	31	40	43	47	48	41	49	51	44	46	53	44

b

Set A:	1	43	24	26	48	50	2	2	36	11	16	37	41	3	36		
	6	8	9	10	17	22	10	11	17	29	30	35	4	23	23		
Set B:	9	18	19	19	20	21	23	24	27	28	31	37	37	38	39	39	39
	40	41	41	43	44	44	45	47	50	50	51	53	53	54	54	55	56

c

Set A:	0.7	0.8	1.4	8.8	9.1	2.6	3.2	0.3	1.7	1.9	2.5	4.1	4.3	3.3	3.4
	3.6	3.9	3.9	4.7	1.6	0.4	5.3	5.7	2.1	2.3	1.9	5.2	6.1	6.2	8.3
Set B:	0.1	0.9	0.6	1.3	0.9	0.1	0.3	2.5	0.6	3.4	4.8	5.2	8.8	4.7	5.3
	2.6	1.5	1.8	3.9	1.9	0.1	0.2	1.2	3.3	2.1	4.3	5.7	6.1	6.2	8.3

9J Measures of spread: range and interquartile range

The mean, median and mode are three numbers that help define the centre of a data set; however, it is also important to describe the spread. Two teams of swimmers from different countries, for example, might have similar mean race times but the spread of race times for each team could be very different.

Example 16 Finding the range and quartiles for an odd number of data values

The following data values are the results for a school mathematics test.

67 96 62 85 73 56 79 19 76 23 68 89 81

a List the data in order, from smallest to largest.

b Find the range.

c Find the:

i median (Q_2)

ii lower quartile (Q_1)

iii upper quartile (Q_3)

iv interquartile range (IQR)

https://www.youtube.com/watch?v=K3wsOqIqA6k

SOLUTION

a 19 23 56 62 67 68 73 76 79 81 85 89 96

Q_1 Q_2 Q_3

b Range $= 96 - 19$
$= 77$

c **i** $Q_2 = 73$

ii $Q_1 = \frac{56 + 62}{2} = 59$

iii $Q_3 = \frac{81 + 85}{2} = 83$

iv $IQR = 83 - 59 = 24$

EXPLANATION

Order the data.

Range = maximum value − minimum value

The median is 73. Since there is an odd number of values, exclude the number 73 before finding Q_1 and Q_3.

Q_1 (the middle value of the lower half) is halfway between 56 and 62.

Q_3 (the middle value of the upper half) is halfway between 81 and 85.

$IQR = Q_3 - Q_1$

3 For each of the following sets of data:

- **i** list the set of data in order, from smallest to largest
- **ii** find the range
- **iii** find the median (Q_2)
- **iv** find the lower quartile (Q_1)
- **v** find the upper quartile (Q_3)
- **vi** find the interquartile range (IQR)

a 5 7 3 6 2 1 9 7 11 9 0 8 5

b 38 36 21 18 27 41 29 35 37 30 30 21 26

c 180 316 197 176 346 219 183 253 228

d 256 163 28 520 854 23 367 64 43 787 12 343 76 3 28

e 1.8 1.9 1.3 1.2 2.1 1.2 0.9 1.7 0.8

f 10 35 0.1 2.3 23 12 0.02

Example 17 Finding quartiles for an even number of data values

Here is a set of measurements, collected by measuring the lengths, in metres, of 10 long-jump attempts.

6.7 9.2 8.3 5.1 7.9 8.4 9.0 8.2 8.8 7.1

a List the data in order, from smallest to largest.

b Find the range.

c Find the:

i median (Q_2)

ii lower quartile (Q_1)

iii upper quartile (Q_3)

iv interquartile range (IQR)

d Interpret the IQR.

SOLUTION

a 5.1 6.7 7.1 7.9 8.2 8.3 8.4 8.8 9.0 9.2

Q_1 Q_2 Q_3

b Range $= 9.2 - 5.1$
$= 4.1$

c i $Q_2 = \frac{8.2 + 8.3}{2}$
$= 8.25\,\text{m}$

ii $Q_1 = 7.1\,\text{m}$

iii $Q_3 = 8.8\,\text{m}$

iv IQR $= 8.8 - 7.1$
$= 1.7\,\text{m}$

d The middle 50% of jumps differed by less than 1.7 m.

EXPLANATION

Order the data to locate Q_1, Q_2 and Q_3.

Range = highest value − lowest value

Q_2 is halfway between 8.2 and 8.3.

The middle value of the lower half is 7.1.
The middle value of the upper half is 8.8.
IQR is the difference between Q_1 and Q_3.

The IQR is the range of the middle 50% of the data.

https://www.youtube.com/watch?v=4Nd7-iUTDOI

4 The running time, in minutes, of 16 movies at the cinema were as follows:

123 110 98 120 102 132 112 140 120 139 42 96 152 115 119 128

a Find the range.

b Find the:

i median (Q_2)

ii lower quartile (Q_1)

iii upper quartile (Q_3)

iv interquartile range (IQR)

c Interpret the IQR.

Lesson 4

9K Box plots

A box plot is a commonly used graph for a data set showing the maximum and minimum values, the median and the upper and lower quartiles. Box plots are often used to show how a data set is distributed and how two sets compare. Box plots are used, for example, to compare a school's examination performance against the performance of all schools in a state. They are also used to show medical test results before and after treatment.

https://www.youtube.com/watch?v=fJZv9YeQ-qQ

Example 18 Interpreting a box plot

This box plot summarises the price of all the books in a book shop.

a State the minimum and maximum book prices.
b Find the range of the book prices.
c State the median book price.
d Find the interquartile range.
e Fifty per cent of the books are priced below what amount?
f Twenty five per cent of the books are priced above what amount?
g If there were 1000 books in the store, how many would be priced below $15?

SOLUTION

a Minimum book price = $5
Maximum book price = $50

b Range = $50 − $5
= $45

c Median book price = $20

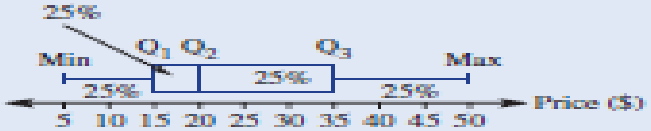
d Interquartile range = $35 − $15
= $20

e $20

f $35

g 0.25 × 1000 = 250
250 books would be below $15

EXPLANATION



Range = maximum price − minimum price

The median is Q_2.

Interquartile range = $Q_3 - Q_1$

50% of books are below Q_2.

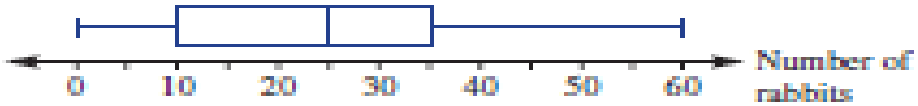
The top 25% of books are above Q_3.

25% of books are priced below $15.

4 This box plot summarises the length of babies born in a particular week in a hospital.

a State the minimum and maximum baby lengths.
b Find the range of the length of the babies.
c State the median baby length.
d Find the interquartile range.
e Fifty per cent of the baby lengths are below what amount?
f Twenty-five per cent of the babies are born longer than what amount?
g If there were 80 babies born in the week, how many would be expected to be less than 45 cm in length?

5 This box plot summarises the number of rabbits spotted per day in a paddock over a 100-day period.



a State the minimum and maximum number of rabbits spotted.
b Find the range of the number of rabbits spotted.
c State the median number of rabbits spotted.
d Find the interquartile range.
e Seventy-five per cent of the days the number of rabbits spotted is below what amount?
f Fifty per cent of the days the number of rabbits spotted is more than what amount?
g How many days was the number of spotted rabbits less than 10?

Lesson 5

Example 19 Constructing a box plot

Consider the following set of data representing 11 scores resulting from rolling two dice and adding their scores.

7 10 7 12 8 9 6 6 5 4 8

a Find the:
- I minimum value
- II maximum value
- III median
- IV lower quartile
- V upper quartile

b Draw a box plot to represent the data.

SOLUTION

a

Q_1 Q_2 Q_3

4 5 (6) 6 7 (7) 8 8 (9) 10 12

- I Min.value = 4
- II Max.value = 12
- III $Q_2 = 7$
- IV $Q_1 = 6$
- V $Q_3 = 9$

b Box plot: Rolling two dice

EXPLANATION

Order the data.

Determine the minimum and maximum value.

The median is the middle value.

Disregard the middle value, then locate the lowerquartile and upper quartile.

Draw a scaled horizontal axis.

Place the box plot above the axis marking in the five key statistics from part a.

6 For each of the sets of data below:

- **i** state the minimum value
- **ii** state the maximum value
- **iii** find the median (Q_2)
- **iv** find the lower quartile (Q_1)
- **v** find the upper quartile (Q_3)
- **vi** draw a box plot to represent the data

a 2 2 3 3 4 6 7 7 7 8 8 8 8 9 11 11 13 13 13

b 43 21 65 45 34 42 40 28 56 50 10 43 70 37 61 54 88 19

c 435 353 643 244 674 364 249 933 523 255 734

d 0.5 0.7 0.1 0.2 0.9 0.5 1.0 0.6 0.3 0.4 0.8 1.1 1.2 0.8 1.3 0.4 0.5

PROBLEM-SOLVING AND REASONING	7, 8, 10	7, 8, 10–12	8, 9, 11–13

7 The following set of data describes the number of cars parked in a street on 18 given days.

14 26 39 46 13 30 5 46 37 26 39 8 8 9 17 48 29 27

a Represent the data as a box plot.

b On what percentage of days were the number of cars parked on the street between:

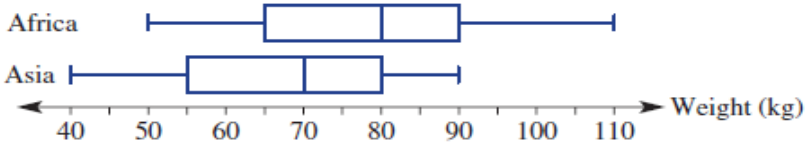
i 5 and 48?

ii 13 and 39?

iii 5 and 39?

iv 39 and 48?

8 The weights of two samples of adult leopards from Africa and Asia are summarised in these box plots.



a Which leopard population sample has the highest minimum weight?

b What is the difference between the ranges for both population samples?

c Is the IQR the same for both leopard samples? If so, what is it?

d What percentage of leopards have a weight less than 80 kg for:

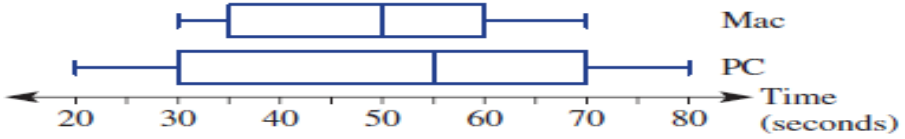
i African leopards?

ii Asian leopards?

e A leopard has a weight of 90 kg. Is it likely to be an Asian or African leopard?

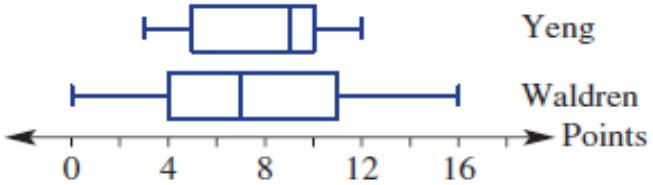
Lesson 6

9 The time that it takes for a sample of computers to start up is summarised in these box plots.



a What type of computer has the lowest median?
b What percentage of Mac computers loaded in less than 1 minute?
c What percentage of PC computers took longer than 55 seconds to load?
d What do you notice about the range for Mac computers and the IQR for PC computers? What does this mean?

10 The number of points per game for two basketball players over a season is summarised in these box plots.



a Which player has the highest maximum?
b Which player has the highest median?
c Which player has the smallest IQR?
d Which player is a more consistent basketball scorer? Give reasons.
e Which player most likely scored the greatest number of points? Give reasons.

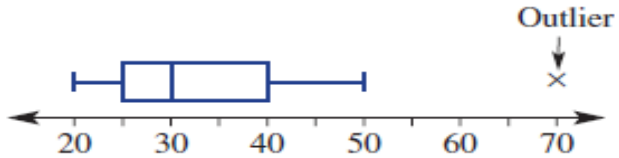
11 Give an example of a small data set that has the following.

- **a** Maximum = upper quartile
- **b** Median = lower quartile

12 Does the median always sit exactly in the middle of a box on a box plot? Explain.

13 Could the mean of a data set be greater than Q_3? Explain.

14 Outliers on box plots are shown as a separate point.



The life in months of a particular kind of battery used in a special type of high-powered calculator is shown in this data set.

3 3 3 4 4 5 6 6 6 7 8 8 9 17

- **a** Use all the values to calculate Q_1, Q_2 and Q_3 for the data set.
- **b** Do any of the values appear to be outliers?
- **c** Not including the outlier, what is the highest value?
- **d** Draw a box plot for the data using a cross (×) to show the outlier.
- **e** Can you give a logical reason for the outlier?

Lesson 7

6 The median of the data in this stem-and-leaf plot is:

A 74 B 71 C 86
D 65 E 70

Stem	Leaf
5	358
6	147
7	02479
8	266

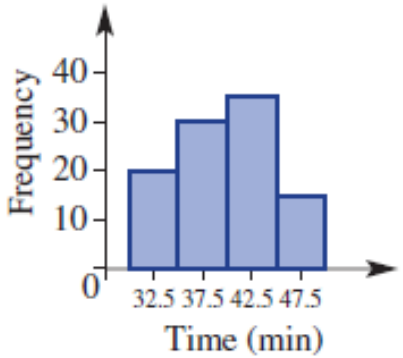
7 | 4 means 74

7 If Jacob achieved scores of 12, 9, 7 and 12 on his last four language vocabulary tests, what score must he get on the fifth test to have a mean of 11?

A 16 B 14 C 11
D 13 E 15

8 This frequency histogram shows the times of finishers in a fun run. The percentage of competitors that finished in better than 40 minutes was:

A 55% B 85% C 50%
D 62.5% E 60%

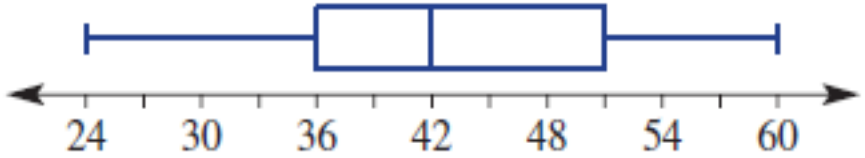


9 The interquartile range of the set of ordered data below is:

1.1 2.3 2.4 2.8 3.1 3.4 3.6 3.8 3.8 4.1 4.5 4.7 4.9

A 2–5 **B** 1.7 **C** 3.8 **D** 1.3 **E** 2.1

10 Choose the *incorrect* statement about the box plot below.



A The range is 36.

B Fifty per cent of values are between 36 and 51.

C The median is 42.

D Twenty-five per cent of values are below 36.

E The interquartile range is 20.