ANALYSING DATA

- Mean
- Median
- Mode

- Range
- Interquartile Range
- 5-number summaries and box plots
- Variance
- Standard Deviation

MEAN

$$\text{mean} = \frac{\text{sum of values}}{\text{number of values}}$$

Population mean

Say we have *N* observations in our population: $x_1, x_2, x_3 \ldots\ldots x_N$

$$\mu = \frac{\sum x}{N}$$

Sample mean

If we have a large population, and our data set is a sample of *n* measurements from the population, our sample data set is
$x_1, x_2, x_3 \ldots\ldots x_n$

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

Calculate the mean of the following sample data: 3, 5, 11, 4, 7

Calculate the mean of the following population data: 1, 1, 3, 9, 11, 8

Calculators with Statistical functions
Your tutor will show you how to use calculator.

Calculation of mean from a frequency table

$$\mu = \frac{\Sigma(f.x)}{N}$$

Calculate the mean of the following data set:
12,13,13,13,13,13,14,14,14,14,14,14,14,15,15,15,15,15,15,15,15,15,16,16,16,16,16,16,17,17,18

x	f	f.x
12		
13		
14		
15		
16		
17		
18		

Exercise

One hundred springs were tested and the expansion length recorded.
The following frequency table summarises the expansion length of the springs (to the nearest cm).
Calculate the mean expansion distance from the data in the following distribution table.

Length (cm)	Frequency	f.x
5	10	
6	20	
7	50	
8	20	

Calculation of mean for data grouped into classes

When the mean is calculated from a frequency table where the data has been grouped into classes, the mean value calculated will only be an estimate. This is because the actual raw data has been obscured by grouping. An approximate value of the mean can be calculated by using midpoints of class intervals to represent all data points falling within the interval.

The following data represents the amount of time that 80 customers spent "browsing" in a bookshop before either leaving the store or making a purchase. Estimate the mean value of time of "customer browsing" the following frequency table:

Time Browsing (minutes)	No. Customers	Midpoint x	(f.x)
2 and under 6	9	4	
6 and under 10	15		
10 and under 14	28		
14 and under 18	21		
18 and under 22	7		

MEDIAN

The median is the midpoint of the observations when they are arranged in ascending order.

The median is the value at the (n+1)/2 position in the data set.

Data sets with an odd number of values.

eg. Consider the following data set: 2, 4, 10, 3, 6

Firstly, arrange the data from smallest to largest: 2, 3, 4, 6, 10
Then determine the middle value.

So the median of the data set is _____

Data sets with an even number of values.

Consider the following data set. 2, 4, 10, 3, 6, 7

Order the data from smallest to largest:
Now, there is no single middle number BUT there is a middle pair :
To find the median we find the mean of the middle pair :
So the median of the data set is _____

Find the median value of each data set

(a) Data set = 1,2,3,9,11 median =

(b) Data set = 1,1,3,11,15 median =

(c) Data set = 15, 10, 16, 9, 18 median =

(d) Data set = 10, 6, 20, 8 median =

Calculation of median from a frequency table when data is ungrouped

Value	Frequency	Cumulative Frequency	Position
12	3		
13	8		
14	19		
15	6		
16	4		

Calculation of median class for grouped data - using frequency table

Determine the median class from the following distribution table

Time (hours)	Frequency	Cumulative Frequency
1.00 - 1.99	5	5
2.00 -2.99	10	15
3.00 - 3.99	30	45
4.00 - 4.99	10	55
5.00 - 5.99	25	80

Median class =

Assuming that the values in the median class are evenly spaced, would you estimate the median value to be close to the lower or upper limit of the median class?

Give a rough estimate (a guess) of the median value:

Calculation of median value for grouped data.

There are two methods estimating the median value when the data is grouped

(1) Graphical and

(2) Numerical

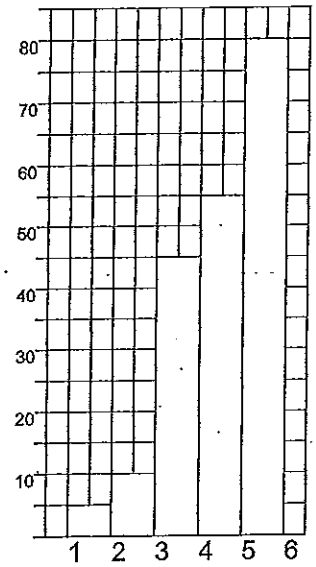
1. Graphical method for calculating median - using cumulative frequency polygon

In order to determine (estimate) the median value, two straight lines need to be superimposed on the cumulative frequency polygon.

Line 1 On the vertical axis, find the cumulative frequency of $\frac{n}{2}$ or $\frac{n+1}{2}$. (when n is large there is very little difference) and draw a horizontal line until it touches the polygon.

Line 2 From the point where Line 1 touches the polygon, draw a vertical line to touch the horizontal axis.

Firstly draw a cumulative frequency polygon (on top of the histogram).
Use the cumulative frequency polygon to estimate the median value of the data.
Note: the cumulative frequency polygon refers to the same data as in previous exercise.



MODE

The mode is the most frequent value among the observations.

Consider the data set

4, 5, 5, 6, 6, 6, 8, 9

The mode here is _____ because this value occurs 3 times and no other value appears more than twice.

That is ____ has the highest _____ among the observation.

Say we have the data set: 1, 4, 5, 5, 5, 6, 8, 9, 9, 9, 12

_____ are modes.

Such a data set is called *bimodal.*

For the data set: 2, 3, 4, 6, 10

there is no mode

Determine the modal value for the following data sets:

(a) Data set = 1,1,3,4,4,4,5,6,6,9,11,11,18 mode =

(b) Data set = 10,11,11,12,12,12,13,14,15,15,15,16,18 mode =

Calculation of modal class for grouped data

Weight (g)	Frequency
10 up to 20	40
20 up to 30	45
30 up to 40	20
40 up to 50	25
50 up to 60	25
	155

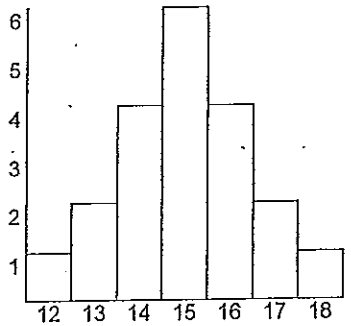
COMPARE THE MEAN, MEDIAN and MODE

The mean, median and mode will be exactly equal whenever the distribution is perfectly symmetric.

For data set A and B:
Determine the mean, median and mode values

Data set A: 12,13,13,14,14,14,14,15,15,15,15,15,15,16,16,16,16,17,17,18

Value (x)	Freq (f)	Freq.xValue (f.x)
12	1	12
13	2	26
14	4	56
15	6	90
16	4	64
17	2	34
18	1	18
	20	**300**



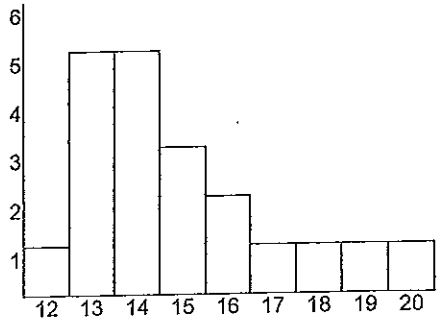
mean =

median =

mode =

Data set B: 12,13,13,13,13,13,14,14,14,14,14,15,15,15,16,16,17,18,19,20

Value (x)	Freq (f)	Freq.xValue (f.x)
12	1	12
13	5	65
14	5	70
15	3	45
16	2	32
17	1	17
18	1	18
19	1	19
20	1	20
	20	298



mean =

median =

mode =

Relative position of mean and median

If a distribution is skewed to the right then _____

If a distribution is skewed to the left then _____

The mean is more sensitive than the median to extreme values.

The median is a more "robust" measure - it is not affected as much by outliers and errors.

Data set A: 3,6,7,8,10	mean =	median =
Data set B: 3,6,7,8,40	mean =	median =
Data set C: 5,5,6,6,8,8,9,10	mean =	median =
Data set D: 5,6,6,8,8,9,10,40	mean =	median =

MEASURING SPREAD OR VARIABILITY

Consider the following three data sets:

Date set 1:	55,55,55,55,55,55,55
Data set 2:	47,51,54,55,56,59,63
Data set 3:	39,47,53,55,57,63,71

In each of the three data sets the mean = and the median =

RANGE

The range of a data set is the difference between the maximum and minimum value of the data set.

That is, the range of the data set is :

For data set 2, 3, 4, 6, 10 the range is:

Calculate the range for following data sets, noting how outliers affect the range.

3,7,8,9,10,10	Range =
3,3,4,4,5,10	Range =
4,5,6,6,8,8,9,10,10,10,10	Range =
4,5,6,6,8,8,9,10,10,10,20	Range =

INTERQUARTILE RANGE

One way to overcome the problem of outliers is to exclude the bottom and top quarters of the data. The range of the remaining middle 50% of data is called the **interquartile range (IQR).**

To calculate the interquartile range, it is necessary to first calculate the **quartiles.**

Lower quartile Q1, divides the bottom ¼ from the top ¾.

Upper quartile Q3, divides the top ¼ from the lower ¾.

The interquartile range is the difference between the upper and lower quartiles.

Interquartile range = upper quartile - lower quartile
IQR = Q3 - Q1

The interquartile range measures the variability of the middle 50% of the data.

It is - easily interpreted, quickly calculated and robust against outliers.

Quartiles

The median divides the whole data set into two halves. It is referred to as the second quartile Q2.

¼ of the values are between the minimum and first quartile (Q_1)

¼ of the values are between the first quartile (Q_1) and second quartile (Q_2)

¼ of the values are between the second quartile (Q_2) and third quartile (Q_3)

¼ of the values are between the third quartile (Q_3) and maximum

eg. A data set has

a minimum value	= 10
a lower quartile (Q_1)	= 18
a second quartile (Q_2)	= 20
an upper quartile (Q_3)	= 25
a maximum value	= 40

(a) What is the interquartile range?

(b) What percentage of values in the data set were between:
(i) 10 and 18? (ii) 18 and 25? (iii) 10 and 20? (iv) 18 and 40?

How to calculate the quartiles for a small data set

To calculate the quartiles, first calculate the median. If the number of data values is odd, include the median in both upper and lower halves. If the number of data values in the set is even, separate the data into two halves.

Lower quartile = median of lower half
Upper quartile = median of upper half

Calculate the quartiles and interquartile range for the following data set

8, 9, 9, 10, 12, 12, 13, 13, 13, 14, 14, 14, 15, 15,15, 15,16,16,16,18,18, 19, 19, 21, 21,22, 22,23,23,24

How to estimate the quartiles for a large data set

The position of the

lower quartile is	0.25 x no. values in data set
second quartile is	0.50 x no. values in the data set
upper quartile is	0.75 x no. values in the data set

For example, if there are 200 values in the data set

Q_1 is at position

Q_2 is at position

Q_3 is at position

Calculate the quartiles and the interquartile range from the following table.

Value	Frequency	Cumulative Frequency
28	20	20
29	15	35
30	17	52
31	23	75
32	17	92
33	40	132
34	28	160
35	25	185
36	15	200

Estimation of the quartiles for large data sets with grouped data

Consider the following data set of 60 values in which the values have been grouped into classes in the frequency table.

Value	Frequency	Cumulative frequency
50 up to 55	10	10
55 up to 60	24	34
60 up to 65	14	48
65 up to 70	6	54
70 up to 75	4	58
75 up to 80	2	60

Calculate the position in the data set of the quartiles, and the class in which the quartiles can be found.

Q_1 at position 15 (.25 x 60 = 15) => 1st quartile is in class 55 up to 60

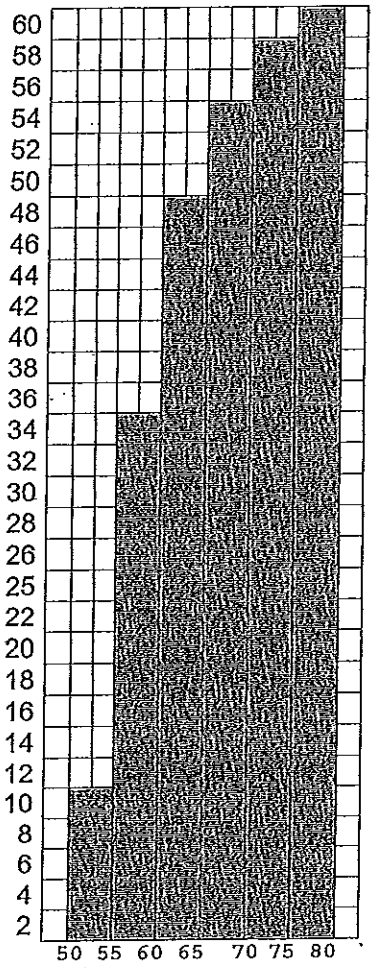
Q_2 at position 30 (.50 x 60 = 30) => 2nd quartile is in class 55 up to 60

Q_3 at position 45 (.75 x 60 = 45) => 3rd quartile is in class 60 up to 65

Formulae may be used to estimate the quartile values rather than the class.

Alternatively the quartiles may be estimated from a cumulative frequency polygon in the same way that the median may be estimated.

Firstly superimpose a cumulative frequency polygon over the histogram.



5-NUMBER SUMMARY and BOX PLOTS

The five-number summary:
minimum value, lower quartile, median, upper quartile, maximum value together with the sample size, is useful for describing the distribution.

The 5-number summary may be graphically displayed in a boxplot.

eg. A data set has

a minimum value	= 10
a lower quartile (Q_1)	= 18
a second quartile (Q_2)	= 20
an upper quartile (Q_3)	= 25
a maximum value	= 40

The 5-number summary may be written simply as: 10, 18, 20, 25, 40

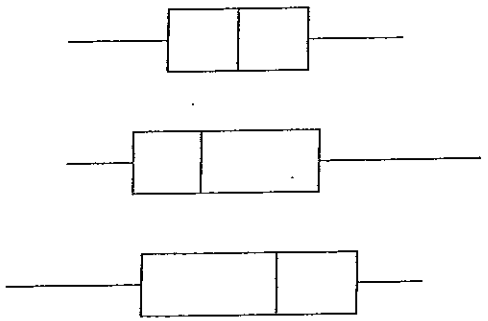
The box plot for this data is:

10 15 20 25 30 35 40

The box plot must be drawn to scale!!!

Assuming there are no outliers, the 5-number summary and box plot readily shows the distribution of values in the data set.

A box plot readily shows the distribution of the data set:



VARIANCE and STANDARD DEVIATION

Variance and standard deviation are measures of the amount of variation of values within a data set.

If a data set was: 5, 5, 5, 5, 5 then variance and standard deviation is zero.

Symbols used to represent variance and standard deviation

	Population data	Sample data
variance	σ^2	s^2
standard deviation	σ	s

Relationship between variance and standard deviation

Standard deviation is the positive square root of the variance.

Standard deviation = $\sqrt{\text{variance}}$

and variance = (standard deviation)^2

Using symbols:

If $\sigma^2 = 36$ then $\sigma =$

if $\sigma = 4$ then $\sigma^2 =$

In words:

If variance is 36, then standard deviation is _____

If standard deviation is 4, then variance is _____

Exercise

(a) If a data set has a standard deviation of 7, calculate the variance.

(b) If a data set has a variance of 10, calculate the standard deviation

Exercise

Which data set (A or B) has the greater amount of variation of values within the data sets:

(a) Data set A has a standard deviation of 18
Data set B has a standard deviation of 26

(b) Data set A has a variance of 6.5
Data set B has a variance of 2.5

(c) Data set A has a variance of 20
Data set B has a standard deviation of 5

Read through this SOLUTIONS

ANALYSING DATA

- Mean
- Median
- Mode

- Range
- Interquartile Range
- 5-number summaries and box plots
- Variance
- Standard Deviation

MEAN

$$\text{mean} = \frac{\text{sum of values}}{\text{number of values}}$$

Population mean

Say we have N observations in our population: $x_1, x_2, x_3 \ldots\ldots x_N$

$$\mu = \frac{\sum x}{N}$$

Sample mean

If we have a large population, and our data set is a sample of n measurements from the population, our sample data set is
$x_1, x_2, x_3 \ldots\ldots x_n$

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

Calculate the mean of the following sample data: 3, 5, 11, 4, 7

= 6 ✓

Calculate the mean of the following population data: 1, 1, 3, 9, 11, 8

5.5 ✓

Calculators with Statistical functions
Your tutor will show you how to use calculator.

Calculation of mean from a frequency table

$$\mu = \frac{\Sigma(f.x)}{N}$$

Calculate the mean of the following data set:
12,13,13,13,13,13,14,14,14,14,14,14,14,15,15,15,15,15,15,15,15,15,16,16,16,16,16,16,17,17,18

x	f	f.x
12	1	12
13	5	65
14	7	98
15	9	135
16	5	80
17	2	34
18	1 ✓	18 ✓
	30 ✓	442

$\therefore$ Mean $= \frac{221}{15}$ ✓

$= 14.7$ (dp)

Exercise

One hundred springs were tested and the expansion length recorded.
The following frequency table summarises the expansion length of the springs (to the nearest cm).
Calculate the mean expansion distance from the data in the following distribution table.

Length (cm)	Frequency	f.x
5	10	50
6	20	120
7	50	350
8	20	160 ✓

Mean = 6.8 ✓

Calculation of mean for data grouped into classes

When the mean is calculated from a frequency table where the data has been grouped into classes, the mean value calculated will only be an estimate. This is because the actual raw data has been obscured by grouping. An approximate value of the mean can be calculated by using midpoints of class intervals to represent all data points falling within the interval.

The following data represents the amount of time that 80 customers spent "browsing" in a bookshop before either leaving the store or making a purchase. Estimate the mean value of time of "customer browsing" the following frequency table:

Time Browsing (minutes)	No. Customers	Midpoint x	(f.x)
2 and under 6	9	4	36
6 and under 10	15	8	120
10 and under 14	28	12	336
14 and under 18	21	16	336
18 and under 22	7	20	140 ✓

$\therefore$ Mean $= \frac{\cancel{12}}{12.1}$ ✓

MEDIAN

The median is the midpoint of the observations when they are arranged in ascending order.

The median is the value at the (n+1)/2 position in the data set.

Data sets with an odd number of values.

eg. Consider the following data set: 2, 4, 10, 3, 6

Firstly, arrange the data from smallest to largest: 2, 3, 4, 6, 10
Then determine the middle value.

So the median of the data set is 4 ✓

Data sets with an even number of values.

Consider the following data set. 2, 4, 10, 3, 6, 7
2 3 4 6 7 10

Order the data from smallest to largest:
Now, there is no single middle number BUT there is a middle pair :
To find the median we find the mean of the middle pair :
So the median of the data set is 5 ✓

Find the median value of each data set

(a) Data set = 1,2,3,9,11 median = 3 ✓

(b) Data set = 1,1,3,11,15 median = 3 ✓

(c) Data set = 15, 10, 16, 9, 18 median = 15 ✓

(d) Data set = 10, 6, 20, 8 median = 9 ✓

Calculation of median from a frequency table when data is ungrouped

Value	Frequency	Cumulative Frequency	Position
12	3	3	3
13	8	11	11
14	19	30	30
15	6	36	36
16	4	40	40 ✓

∴ Median = 14 ✓

Calculation of median class for grouped data - using frequency table

Determine the median class from the following distribution table

Time (hours)	Frequency	Cumulative Frequency
1.00 - 1.99	5	5
2.00 -2.99	10	15
3.00 - 3.99	30	45
4.00 - 4.99	10	55
5.00 - 5.99	25	80

Median class = (3 - 3.99) ✓

Assuming that the values in the median class are evenly spaced, would you estimate the median value to be close to the lower or upper limit of the median class? lower limit ✓

Give a rough estimate (a guess) of the median value: 3.4 ✓

Calculation of median value for grouped data.

There are two methods estimating the median value when the data is grouped

(1) Graphical and

(2) Numerical

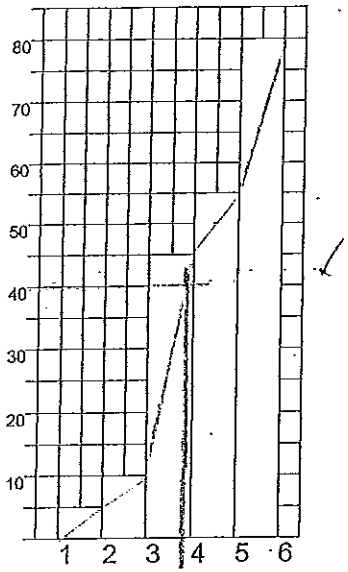
1. Graphical method for calculating median - using cumulative frequency polygon

In order to determine (estimate) the median value, two straight lines need to be superimposed on the cumulative frequency polygon.

Line 1 On the vertical axis, find the cumulative frequency of $\frac{n}{2}$ or $\frac{n+1}{2}$. (when n is large there is very little difference) and draw a horizontal line until it touches the polygon.

Line 2 From the point where Line 1 touches the polygon, draw a vertical line to touch the horizontal axis.

Firstly draw a cumulative frequency polygon (on top of the histogram). Use the cumulative frequency polygon to estimate the median value of the data.
Note: the cumulative frequency polygon refers to the same data as in previous exercise.



3.75✓

MODE

The mode is the most frequent value among the observations.

Consider the data set

4, 5, 5, 6, 6, 6, 8, 9

The mode here is 6 ✓ because this value occurs 3 times and no other value appears more than twice.

That is 6 ✓ has the highest frequency ✓ among the observation.

Say we have the data set: 1, 4, 5, 5, 5, 6, 8, 9, 9, 9, 12

5, 9 ✓ are modes.

Such a data set is called *bimodal*.

For the data set: 2, 3, 4, 6, 10

there is no mode

Determine the modal value for the following data sets:

(a) Data set = 1,1,3,4,4,4,5,6,6,9,11,11,18 mode = 4 ✓

(b) Data set = 10,11,11,12,12,12,13,14,15,15,15,16,18 mode = 12, 15 ✓

Calculation of modal class for grouped data

Weight (g)	Frequency
10 up to 20	40
20 up to 30	45
30 up to 40	20
40 up to 50	25
50 up to 60	25
	155

(20 - 30) ✓

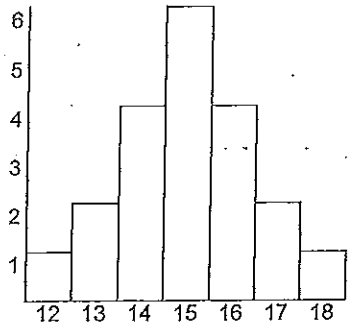
COMPARE THE MEAN, MEDIAN and MODE

The mean, median and mode will be exactly equal whenever the distribution is perfectly symmetric.

For data set A and B:
Determine the mean, median and mode values

Data set A: 12,13,13,14,14,14,14,15,15,15,15,15,15,16,16,16,16,17,17,18

Value (x)	Freq (f)	Freq.xValue (f.x)
12	1	12
13	2	26
14	4	56
15	6	90
16	4	64
17	2	34
18	1	18
	20	**300**



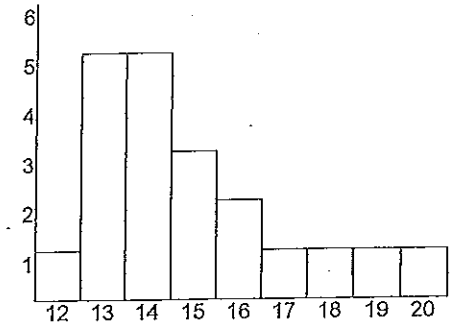
mean = 15 ✓

median = 15 ✓

mode = 15 ✓

Data set B: 12,13,13,13,13,13,14,14,14,14,14,15,15,15,16,16,17,18,19,20

Value (x)	Freq (f)	Freq.xValue (f.x)
12	1	12
13	5 7	65
14	5 12	70
15	3 15	45
16	2 17	32
17	1	17
18	1	18
19	1	19
20	1	20
	20	298



mean = 14.9 ✓

median = 14 ✓

mode = 13, 14 ✓

Relative position of mean and median

If a distribution is skewed to the right then it covers more (positively skewed) larger portions of data are less than or equal to the mean. not bigger nos. OR MEAN > MEDIAN

If a distribution is skewed to the left then it covers smaller nos (negatively skewed) larger portion of data are greater than or equal to the mean OR MEAN < MEDIAN.

The mean is more sensitive than the median to extreme values.

The median is a more "robust" measure - it is not affected as much by outliers and errors.

Data set A: 3,6,7,8,10 mean = ~~34~~ 6.8 ✓ median = 7 ✓

Data set B: 3,6,7,8,40 mean = 12.8 ✓ median = 7 ✓

Data set C: 5,5,6,6,8,8,9,10 mean = 7.125 ✓ median = 7 ✓

Data set D: 5,6,6,8,8,9,10,40 mean = 11.5 ✓ median = 8 ✓

MEASURING SPREAD OR VARIABILITY

Consider the following three data sets:

Date set 1:	55,55,55,55,55,55,55
Data set 2:	47,51,54,55,56,59,63
Data set 3:	39,47,53,55,57,63,71

In each of the three data sets the mean = 1.55 2.55 3.55 ✓ and the median = 1.55 2.55 3.55 ✓

RANGE

The range of a data set is the difference between the maximum and minimum value of the data set.

That is, the range of the data set is :

For data set 2, 3, 4, 6, 10 the range is: 8 ✓

Calculate the range for following data sets, noting how outliers affect the range.

3,7,8,9,10,10 Range = 7 ✓

3,3,4,4,5,10 Range = 7 ✓

4,5,6,6,8,8,9,10,10,10,10 Range = 6 ✓

4,5,6,6,8,8,9,10,10,10,20 Range = 16 ✓

INTERQUARTILE RANGE

One way to overcome the problem of outliers is to exclude the bottom and top quarters of the data. The range of the remaining middle 50% of data is called the **interquartile range (IQR).**

To calculate the interquartile range, it is necessary to first calculate the **quartiles.**

Lower quartile Q1, divides the bottom ¼ from the top ¾.

Upper quartile Q3, divides the top ¼ from the lower ¾.

The interquartile range is the difference between the upper and lower quartiles.

Interquartile range = upper quartile - lower quartile
IQR = Q3 - Q1

The interquartile range measures the variability of the middle 50% of the data.

It is - easily interpreted, quickly calculated and robust against outliers.

Quartiles

The median divides the whole data set into two halves. It is referred to as the second quartile Q2.

¼ of the values are between the minimum and first quartile (Q_1)

¼ of the values are between the first quartile (Q_1) and second quartile (Q_2)

¼ of the values are between the second quartile (Q_2) and third quartile (Q_3)

¼ of the values are between the third quartile (Q_3) and maximum

eg. A data set has

a minimum value	= 10
a lower quartile (Q_1)	= 18
a second quartile (Q_2)	= 20
an upper quartile (Q_3)	= 25
a maximum value	= 40

(a) What is the interquartile range? = 7 ✓

(b) What percentage of values in the data set were between:

(i) 10 and 18? (ii) 18 and 25? (iii) 10 and 20? (iv) 18 and 40?

25% ✓ 50% ✓ 50% ✓ 75% ✓

How to calculate the quartiles for a small data set

To calculate the quartiles, first calculate the median. If the number of data values is odd, include the median in both upper and lower halves. If the number of data values in the set is even, separate the data into two halves.

Lower quartile = median of lower half = 13 ✓
Upper quartile = median of upper half = 19 ✓

Calculate the quartiles and interquartile range for the following data set

8, 9, 9, 10, 12, 12, 13, 13, 13, 14, 14, 14, 15, 15, 15, 15, 16, 16, 16, 18, 18, 19, 19, 21, 21, 22, 22, 23, 23, 24

How to estimate the quartiles for a large data set

The position of the

lower quartile is	0.25 x no. values in data set
second quartile is	0.50 x no. values in the data set
upper quartile is	0.75 x no. values in the data set

For example, if there are 200 values in the data set

Q_1 is at position 50 ✓
Q_2 is at position 100 ✓
Q_3 is at position 150 ✓

Calculate the quartiles and the interquartile range from the following table.

Value	Frequency	Cumulative Frequency
28	20	20
29	15	35
30	17	52
31	23	75
32	17	92
33	40	132
34	28	160
35	25	185
36	15	200

$Q_1 = 30$ ✓
$Q_2 = 33$ ✓
$Q_3 = 34$ ✓

Estimation of the quartiles for large data sets with grouped data

Consider the following data set of 60 values in which the values have been grouped into classes in the frequency table.

Value	Frequency	Cumulative frequency
50 up to 55	10	10
55 up to 60	24	34
60 up to 65	14	48
65 up to 70	6	54
70 up to 75	4	58
75 up to 80	2	60

Calculate the position in the data set of the quartiles, and the class in which the quartiles can be found.

Q_1 at position 15 (.25 x 60 = 15) => 1st quartile is in class 55 up to 60

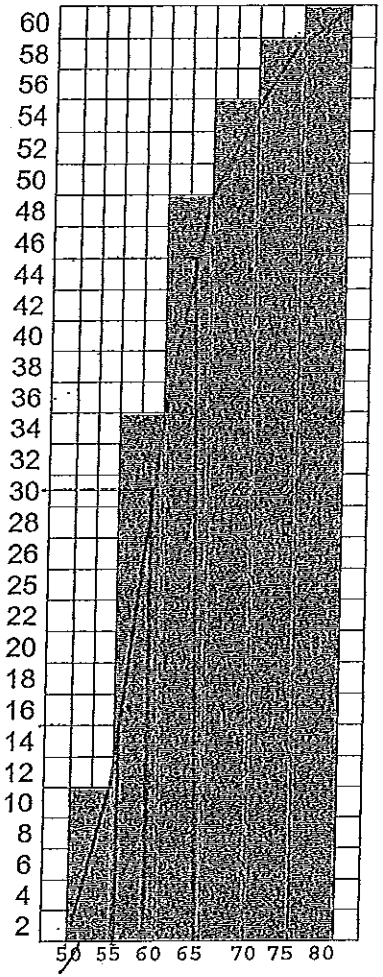
Q_2 at position 30 (.50 x 60 = 30) => 2nd quartile is in class 55 up to 60

Q_3 at position 45 (.75 x 60 = 45) => 3rd quartile is in class 60 up to 65

Formulae may be used to estimate the quartile values rather than the class.

Alternatively the quartiles may be estimated from a cumulative frequency polygon in the same way that the median may be estimated.

Firstly superimpose a cumulative frequency polygon over the histogram.



$Q_1 = 55$ ✓

$Q_2 = 58$ ✓

$Q_3 = 64$ ✓

5-NUMBER SUMMARY and BOX PLOTS

The five-number summary:
minimum value, lower quartile, median, upper quartile, maximum value together with the sample size, is useful for describing the distribution.

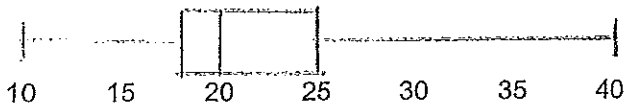
The 5-number summary may be graphically displayed in a boxplot.

eg. A data set has

a minimum value	= 10
a lower quartile (Q_1)	= 18
a second quartile (Q_2)	= 20
an upper quartile (Q_3)	= 25
a maximum value	= 40

The 5-number summary may be written simply as: 10, 18, 20, 25, 40

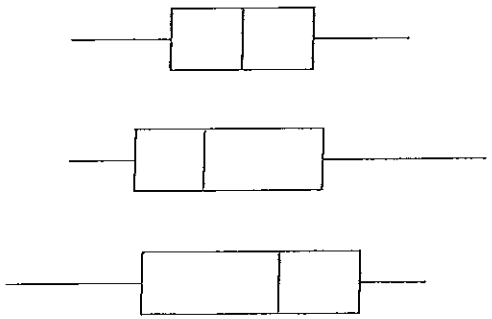
The box plot for this data is:



The box plot must be drawn to scale!!!

Assuming there are no outliers, the 5-number summary and box plot readily shows the distribution of values in the data set.

A box plot readily shows the distribution of the data set:



VARIANCE and STANDARD DEVIATION

Variance and standard deviation are measures of the amount of variation of values within a data set.

If a data set was: 5, 5, 5, 5, 5 then variance and standard deviation is zero.

Symbols used to represent variance and standard deviation

	Population data	Sample data
variance	σ^2	s^2
standard deviation	σ	s

Relationship between variance and standard deviation

Standard deviation is the positive square root of the variance.

Standard deviation = $\sqrt{\text{variance}}$

and variance = (standard deviation)^2

Using symbols:

If $\sigma^2 = 36$ then $\sigma = 6$ ✓

if $\sigma = 4$ then $\sigma^2 = 16$ ✓

In words:

If variance is 36, then standard deviation is 6 ✓

If standard deviation is 4, then variance is 16 ✓

Exercise

(a) If a data set has a standard deviation of 7, calculate the variance.

49 ✓

(b) If a data set has a variance of 10, calculate the standard deviation

$\sqrt{10}$ ✓

Exercise

Which data set (A or B) has the greater amount of variation of values within the data sets:

(a) Data set A has a standard deviation of 18
Data set B has a standard deviation of 26 Set B ✓

(b) Data set A has a variance of 6.5 Set A ✓
Data set B has a variance of 2.5

(c) Data set A has a variance of 20
Data set B has a standard deviation of 5 Set B ✓