Introduction to Statistics

Statistics is a branch of mathematics that is concerned with the planning, collection, organisation, analysis, reporting of data and the interpretation of results. The aim of this module is to give you an introduction to statistics. This module gives you an overview of different methods of displaying and organising data, calculating measures of central tendency, calculating measures of spread and comparing like data. Correlation and Regression using Excel will inform the understanding of linear relationships.

Collecting Data

When data is collected from every member of the group, a census is held. The group in this instance is called the population.

When only selected members of the population contribute to the data, this is referred to as a sample of the population and a survey takes place. Mostly, it is impractical and too expensive to obtain data from a population so a sample is selected. Information from a sample is often used to predict information about the population. This is the process used to predict the outcomes of elections.

The selection of a sampling method is very important. For example it may be important to collect data from different geographical locations, different ages or different socioeconomic backgrounds. For quality control in industry, systematic sampling, taking (say) every tenth item to check quality could be used. It takes careful planning to conduct investigations with little or no bias. Bias is an unfair preference towards one group which may lead to a distortion of the statistical results.

The researcher must also decide on sample size. If the sample size is too small, then the validity of the finding could be in doubt. If the sample size is too large, the cost of obtaining the data may be prohibitively high. It is possible to obtain an appropriate sample size using statistical processes that considers the accuracy of results required with the number to survey to give that accuracy.

Statistical data can be of different types, the type of the data may determine the statistical processes that can be undertaken. Data can be classified as either Categorical or Numerical. Categorical data is usually in word form. Numerical data is usually in number form, however, some data in number form such as postcodes could be considered Categorical data as no statistical analysis would be performed; for example; the average postcode.

There are two types of numerical data: discrete and continuous. Most of this module is about numerical data.

Discrete data is numerical data where the values are in set amounts, for example, the number of people in a classroom. There could be 25, 26, 27 people in a classroom but not any amounts in between such as 25.72 people.

Continuous data is numerical data where the values can be any value, for example, the height of people. The height of people can be measured, in theory, to any precision. The only limitation is the limitation of the measuring instrument, but in theory the measurement can be to any precision.

Let's consider some examples:

(a) Temperature – this is continuous because it can be measured to any precision.

(b) Number of students at a lecture – this is discrete because only a whole number of students could attend.

(c) The number of children in a family – this is discrete because only a whole number of children is possible.

(d) Age – this is continuous as age can be measured to any precision. For convenience, age in years may be recorded – this is still continuous because age can be measured to any precision. The decision of discrete vs continuous should be based on what is possible rather than what is chosen for convenience.

(e) Weight – this is continuous as weight can be measured to any precision.

(f) Shoe size – this is discrete because shoe size has set values of half sizes and none in between.

Module contents

Introduction

- **Organising Data - Tables**
- **Graphs**
- **Measures of Central Tendancy**
- **Methods of Spread**
- **Comparing Like Data**
- **Correlation and Regression**

Answers to activity questions

Outcomes

- Display data appropriately using charts and graphs.
- Organise data using tables.
- Calculate descriptive statistics for sets of data.
- Calculate correlation coefficient and equations of regression lines using Excel.

Check your skills

This module covers the following concepts, if you can successfully answer these questions, you do not need to do this module. Check your answers from the answer section at the end of the module.

1. For the data (right) about the Average Daily Hours of Sunshine, calculate the mean, mode and median.
2. For the same data, calculate the range, interquartile range and standard deviation.
3. For the same data, use a 5 number summary to draw a box and whisker plot.
4. Use Excel to do a scatterplot of Hours of Sunshine vs Latitude. Latitude is the independent variable
5. Use Excel to determine the correlation and the regression equation.

	Data	
Location	Latitude °S	Av. Daily Hours of Sunshine
Darwin	12.42	6.9
Brisbane	27.48	7.4
Perth	31.93	8.8
Sydney	33.86	6.8
Adelaide	34.93	7.0
Canberra	35.3	7.2
Melbourne	37.81	6
Hobart	42.89	5.9
Macquarie Island	54.5	2.3

Topic 1: Organising Data - Tables

The quantity of data collected will determine how it is organised.

If the heights of 25 students are collected, then no table will be required. These 25 **individual values** can be analysed without the need to organise the data.

If the grade point average of all the students at this university were to be analysed, then the data would need to the organised into an appropriate table before any statistical analysis could take place. The structure of the table will vary depending on the type of data (discrete or continuous) and the variation between lowest and highest values. With spreadsheet computer programs such as Excel, the need to organise data using the traditional table structures is not as important as in the past.

The first table to be considered is a Frequency Distribution Table. To assist in the development of ideas in this module, some key examples will be used.

Ungrouped Frequency Distribution Tables

The data below is the number of matches in each box for 50 boxes, it is discrete data. This number of values is too many to have as individual values so organising the values into a table will assist when analysing the data. The average number of matches per box is given as 50. The number of matches per box typically varies between 46 and 55 matches. The raw data is given below:

49	50	50	49	50	48	50	53	48	53
51	47	54	48	52	50	47	55	49	55
52	50	51	51	50	50	53	49	52	48
54	49	46	50	50	49	50	50	50	54
46	51	51	47	48	50	52	50	51	50

When the data is initially entered into a table, tally marks are used to record each piece of data. Tally marks are a small vertical line. After four tally marks, the fifth is put through the four to make grouping of 5.

When recording the occurrence of each value, data should be systematically entered. It is during this process that many mistakes are made – you have been warned! Watch the video below to see how to do this.

Video ‘Entering Data into a frequency distribution table’

After the data is entered into the table it will have the appearance below. The word frequency is the number of occurrences of that value or how *frequently* it occurs.

Number of Matches	Tally	Frequency
46	II	2
47	III	3
48	𝍸	5
49	𝍸 I	6
50	𝍸 𝍸 𝍸 I	16
51	𝍸 I	6
52	IIII	4
53	III	3
54	III	3
55	II	2
Total		50

After this initial table is constructed, the table has extra columns added to help with the summary data required.

The Tally is not usually repeated.

Commonly added columns are Relative Frequency, %Relative Frequency and Cumulative Frequency.

The **Relative Frequency** is the proportion of the data that has that value. This can be expressed as a decimal or fraction. It is calculated by taking the frequency for each score and dividing by the total number of scores. Find a total for this column.

The **%Relative Frequency** is the Relative Frequency made into a percentage. This is achieved by multiplying the Relative Frequency by 100. Find a total for this column.

The **Cumulative Frequency** is a running total of the frequency. The cumulative frequency is found on any line of the table by adding the frequency for that line to the total of frequencies for the previous lines. Do not obtain a total for the cumulative frequencies as it is already a (running) total. The cumulative frequency column is used extensively for finding the median and the quartiles (covered later).

Number of Matches	Frequency	Relative Frequency	% Relative Frequency	Cumulative Frequency
46	2	$\frac{2}{50} = 0.04$	0.04 x 100=4%	2
47	3	0.06	6%	2+3=5
48	5	0.1	10%	5+5=10
49	6	0.12	12%	10+6=16
50	16	0.32	32%	16+16=32
51	6	0.12	12%	32+6=38
52	4	0.08	8%	38+4=42
53	3	0.06	6%	42+3=45
54	3	0.06	6%	48
55	2	0.04	4%	50
Total	50	1.00	100%	

The accuracy of the calculations can be checked by:

1. The total of the Relative Frequencies should be 1
2. The total of the %Relative Frequencies should be 100%
3. The last Cumulative Frequency should be equal to the total of the frequency column.

From this table, answer to questions expressed in certain ways can be obtained.

For example:

1. How many boxes contained 51 matches? The frequency for this was 6, so there were 6 boxes that contained 51 matches.

2. For what percentage of boxes contained 50 matches? 32% of the values were 50.

3. How many boxes of matches contained 50 or less?
   The cumulative frequency for 50 is the total of the frequencies including the frequency for 50. This is 32. So 32 boxes contained 50 or less matches.

4. What proportion of boxes contained 49, 50 or 51 matches?
   This is found by adding up the relative frequencies for those numbers of matches. So the proportion is 0.12 + 0.32 + 0.12 = 0.56, this can also be expressed as a percentage, 56%.

5. How many matches were in the 27^{th} box?
   The frequency distribution table has ordered the data. For the number of boxes containing 50 matches, the cumulative frequency is 32. The previous cumulative frequency is 16. This means that the 17^{th} (the one after the 16^{th}) box through to the 32^{nd} box contains 50 matches. As the 27^{th} box is between the 16^{th} and the 32^{nd} box, the 27^{th} box will contain 50 matches.

Grouped Frequency Distribution Tables

When the difference between the lowest and highest scores is larger, groups may have to be used. Consider the following table for data for amount spent by seventy children at a recent show.

This data is continuous.

Forming **groups** for continuous data is similar to those below, that is, in the form of:

$$Lower\ amount \quad \text{but less than } higher\ amount$$

or

$$lower\ amount \leq x < higher\ amount$$

When deciding on groups it is important that:

(i) Each piece of data will be placed in **one group only.**
(ii) There should be enough groups to **include all values.**
(iii) There should be between **5 and 15 groups**. Use practical divisions based on the data.
(iv) Each group should cover the **same number of values**.
In the table below; the upper boundaries subtract lower boundaries give exactly the same value, for example, in the first group $10 - $0 = $10, the second group $20 - $10 = $10 and so on.
Although the upper boundary is always worded 'but less than n.n', in a practical sense it is suggested that the upper boundary is so close to n.n that it is taken as n.n.

A Grouped Frequency Distribution Table has extra column added.

The Group Midpoint is the numerical middle value, found by:

$$GroupMidpoint = \frac{(lower\ amount + upper\ amount)}{2}$$

This value will be used in future topics.

The Grouped Frequency Distribution Table below was derived from a list of 70 values less than $100. The groups were formed and the frequency of each group recorded in the frequency column. The next four columns were calculated from the groups or their frequencies.

Amount Spent$	Frequency	Group Midpoint	Relative Frequency	% Relative Frequency	Cumulative Frequency
\$0 but less than \$10	2	5	$\frac{2}{70} = 0.029$	0.04 x 100=4%	2
\$10 but less than \$20	3	15	0.043	4.3%	2+3=5
\$20 but less than \$30	5	25	0.071	7.1%	5+5=10
\$30 but less than \$40	4	35	0.057	5.7%	14
\$40 but less than \$50	2	45	0.029	2.9%	16
\$50 but less than \$60	8	55	0.114	11.4%	24
\$60 but less than \$70	14	65	0.2	20%	38
\$70 but less than \$80	18	75	0.257	25.7%	56
\$80 but less than \$90	12	85	0.171	17.1%	68
\$90 but less than \$100	2	95	0.029	2.9%	70
Total	70		1.00	100	

When a table like this is given, there is no **original data**! The original values are lost into the groups. This is the big **disadvantage** of Grouped Frequency Distribution Tables. The only assumption that can be made about the original values is that they are **evenly spread** throughout the group.

For example; the 18 values in the group *'\$70 but less than \$80'* are assumed to be equally spaced out between \$70 and \$80. The higher the number of children surveyed, the more likely this is to be true. This idea is used when calculating some statistical measures in future topics.

If the data is **discrete,** the construction of the table is a little different.

The addition of a column labelled Group Boundaries is required for the construction of a frequency histogram and polygon (a graph).

The data below is number of mp3 players sold per day during a 60 day sale.

Number Sold	Group Boundaries	Frequency	Class Midpoint	Relative Frequency	% Relative Frequency	Cumulative Frequency
1 - 10	0.5 - 10.5	6	5.5	0.1	10%	6
11 - 20	10.5 - 20.5	10	15.5	0.167	16.7%	16
21 - 30	20.5 - 30.5	15	25.5	0.25	25%	31
31 - 40	30.5 - 40.5	12	35.5	0.2	20%	43
41 - 50	40.5 - 50.5	9	45.5	0.15	15%	52
51 - 60	50.5 - 60.5	6	55.5	0.1	10%	58
61 - 70	60.5 - 70.5	2	65.5	0.033	3.33%	60
Total		60		1.00	100	

From this table, answer to questions expressed in certain ways can be obtained.

1. On how many days were 15 mp3 players sold?
   Because the original data is lost it is not possible to determine this.

2. On how many days were 11 - 20 mp3 players sold?
   This occurred on 10 days.

3. On how many days were 40 or less mp3 players sold?
   The cumulative frequency for 31 - 40 is the total of the frequencies including the frequency for 31 - 40. This is 43. So on 43 days 40 or less mp3 players were sold.

4. On what proportion of days were 21 – 60 players sold?
   This is found by adding up the relative frequencies for those numbers of players. So the proportion is 0.25 + 0.2 + 0.15+ 0.1 = 0.7, this can also be expressed as a percentage, 70%.

5. How many players sold per day are represented by the 27^{th} score (when put in number sold order)?
   The frequency distribution table has ordered the data. For the group 21 - 30, the cumulative frequency is 31. The previous group's cumulative frequency is 16. This means that the 17^{th} (the one after the 16^{th}) box through to the 31^{st} box between 21 – 30 days. As the 27^{th} score is between the 16^{th} and the 31^{st} score, the 27^{th} score will be between 21 – 30 players sold. A more detailed look at this will occur later in this module.

Stem and Leaf Plots

The ages of the 40 people can be displayed in a stem and leaf plot. The raw data for the ages is:

18	22	41	19	30	31	27	20
32	27	31	25	35	24	19	40
35	32	44	37	17	20	45	32
23	27	34	19	47	33	24	41
39	26	29	30	44	24	28	32

If the numbers are entered from the table above starting with the first row, working from left to right, the Stem and Leaf will be:

Stem	Leaf
1	8 9 9 7 9
2	2 7 0 7 5 4 0 3 7 4 6 9 4 8
3	0 1 2 1 5 5 2 7 2 4 3 9 0 2
4	1 0 4 5 7 1 4
	3|2 means 32 years old

When constructing a Stem and Leaf Plot, make sure that the numbers are equally spaced. The length of each leaf can give some general information about the data. Each table should have a statement that gives place value to the data. Once obtaining this Stem and Leaf plot, it is now useful to order each of the leaves (leafs!).

The Ordered Stem and Leaf Plot is:

Stem	Leaf
1	7 8 9 9 9
2	0 0 2 3 4 4 4 5 6 7 7 7 8 9
3	0 0 1 1 2 2 2 2 3 4 5 5 7 9
4	0 1 1 4 4 5 7
	3|2 means 32 years old

Note: Stem And Leaf Plots are presented with leaves ordered – this is convention.

The advantages of a Stem and Leaf plot over a frequency distribution table are:

(i) The data is grouped and the **original values are retained.**

(ii) The data is listed in **numerical order** from the lowest value (top row on left) to the highest value (bottom row on right). This will be very useful later on when calculating 5 number summaries.

If there is concern about the number of groups, that is, not enough groups, then each group can be subdivided into two groups. There are various ways this is done, but the system used here is to replace the existing 2 stem (values from 20 - 29) with the stems: 2 (values 20 - 24) and 2* (values 25 - 29).

Now the data is spread over 7 groups instead of 4.

Stem	Leaf
1*	7 8 9 9 9
2	0 0 2 3 4 4 4
2*	5 6 7 7 7 8 9
3	0 0 1 1 2 2 2 2 3 4
3*	5 5 7 9
4	0 1 1 4 4
4*	5 7
	3|2 means 32 years old

 Video ‘Organising Data using Tables’

Activity

1. The table of data below represents the speed of 40 cars passing a school at 9am on a school day.

Car Speed Data (in km/hr)

12	41	44	45	28	40	32	62	46	25
31	35	31	20	59	27	49	19	58	38
22	50	46	14	33	48	25	32	52	69
40	52	57	27	61	42	39	64	52	27

(a) Enter the data into a **grouped frequency distribution table**. Include a cumulative frequency column. Make the first 'group 10 to but less than 20'.
(b) Enter the data into Stem and Leaf Plot.
(c) What percentage of cars were doing 40km/hr or more?

2. The number of students attending a class (maximum 25) for 30 lessons is given in the table below:

Students Attending Class

25	24	24	25	24	23
25	24	23	24	25	25
24	25	20	23	25	24
22	24	25	24	23	21
25	23	24	25	24	22

(a) Is the data discrete or continuous?
(b) Enter the data into a frequency distribution table (groups not required). Include a relative frequency and % relative frequency column.
(c) What proportion of lessons contained 22 students?
(d) What percentage of lessons were fully attended?

3. The systolic blood pressures in mmHg (this is the higher value of the two blood pressure figures) of 30 patients are given in the table below.

Systolic Blood Pressures of 35 patients at a Cardiac Clinic

122	175	114	92	128	155
138	115	88	134	141	146
112	124	107	121	118	145
126	188	134	110	122	139
133	149	120	102	95	109
144	127	143	161	137	

(a) Enter the data into a Stem and Leaf Plot. Use the key: 11|4 means 114.
(b) If hypertension (high blood pressure) is defined by a systolic blood pressure 140 or above, what percentage of this group are suffering hypertension?

4. The Shot Put distances thrown by 27 world champion shot putters are given in the table below. The unit is metres (m).

22.25	20.19	21.39	21.25	21.19	22.07
21.72	20.37	20.45	23.09	21.19	21.22
21.07	21.55	23.12	20.91	22.58	21.97
20.22	20.38	22.37	22.19	21.70	20.54
22.67	21.58	21.72			

(a) Enter the data into a Frequency Distribution Table. Your FDT must have at least 5 groups.
(b) How many have thrown less than 22m? (Use cumulative frequency to answer this)
(c) What percentage threw 21 to but less than 22m? (Use a % Relative Frequency Column).

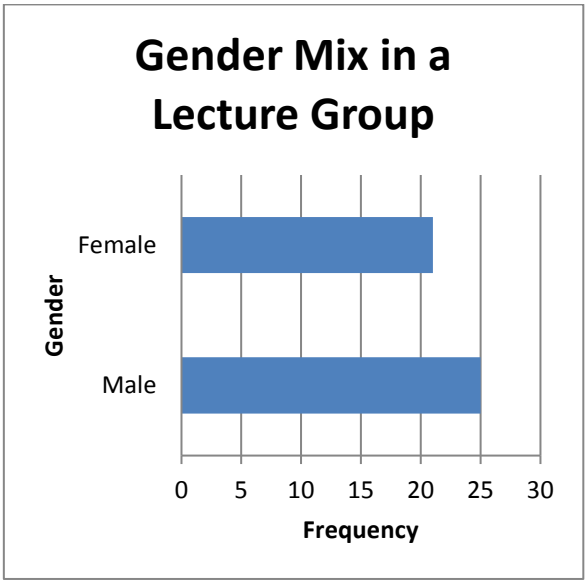
Topic 2: Graphs

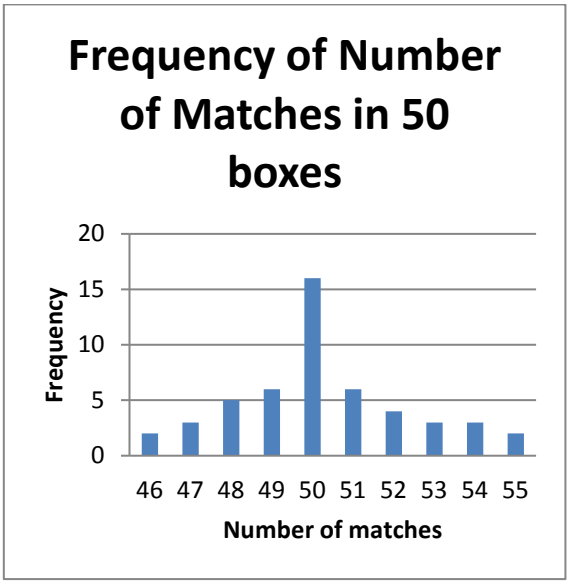
The graphs you use to present information depend upon the nature and the type of data.

Bar / Column Graphs

Categorical and discrete data can be displayed effectively in bar or column graphs.

Below are two graphs; a bar graph for the gender breakdown in a Lecture Group and a column graph for the number of matches in 50 boxes (covered previously). These were drawn using Excel. Notice the equal spacing and thickness of the bars. The spacing of bars reflects the nature of discrete data.



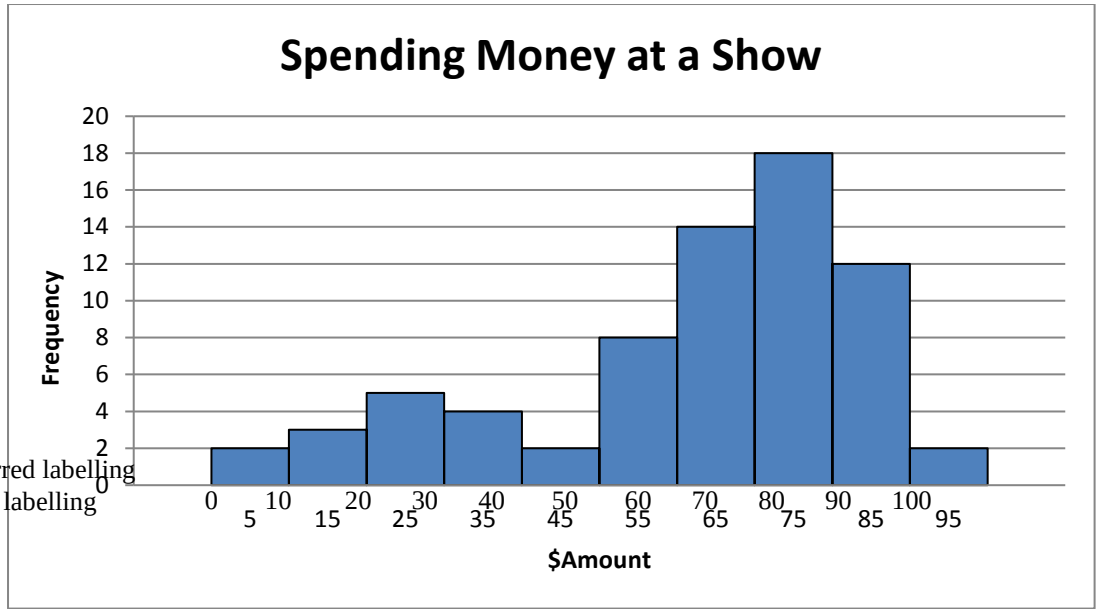


Histograms

Numerical data can be used to construct a histogram. A histogram is basically a column graph with the bars joined together. This reflects the nature of continuous data, however, discrete data can also be graphed as a histogram

The graph below was drawn using Excel. Frequency is shown on the vertical axis. The horizontal axis shows the amount spent at the show and is labelled with the class mark of each group. This is not ideal. The preferred way to label the axis is to show the boundaries for each group on the bar, this is also shown below.

This type of graph is called a 'Frequency Histogram'.



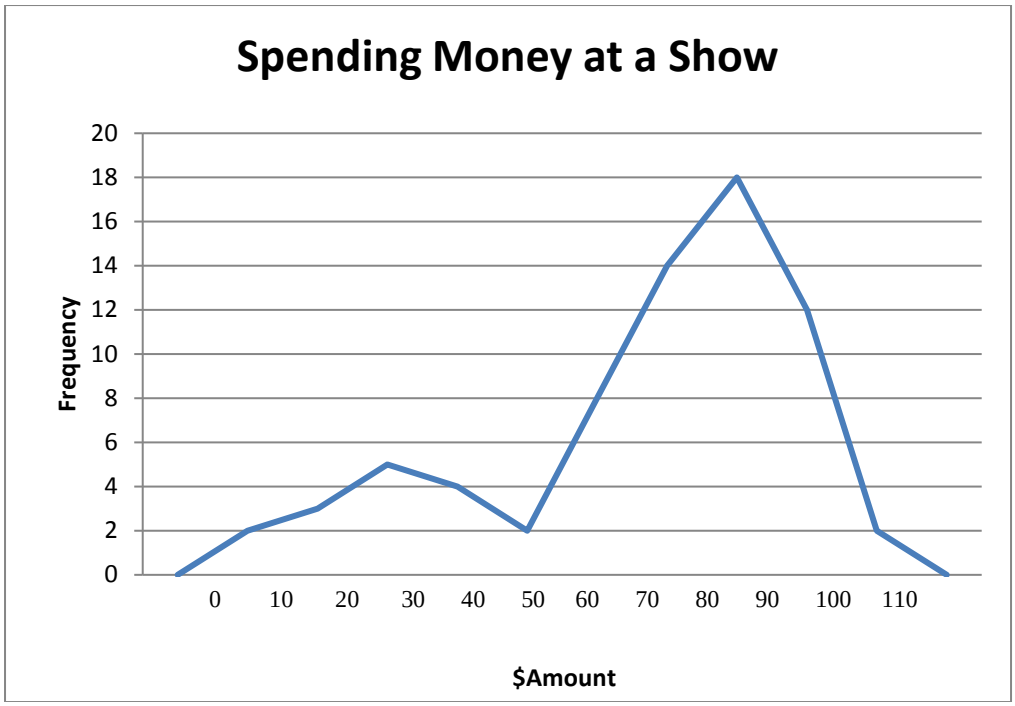
Preferred labelling
Excel labelling

A line called the frequency polygon can be drawn on the frequency histogram.

From this it is possible to comment on the shape of the distribution.

The line is drawn from the centre of each column to the centre of the next. It should start from the horizontal axis from the centre of an imaginary group below and extend back to the horizontal axis to the centre of an imaginary group above.

The polygon can be drawn with or without the histogram present. The polygon is drawn below.

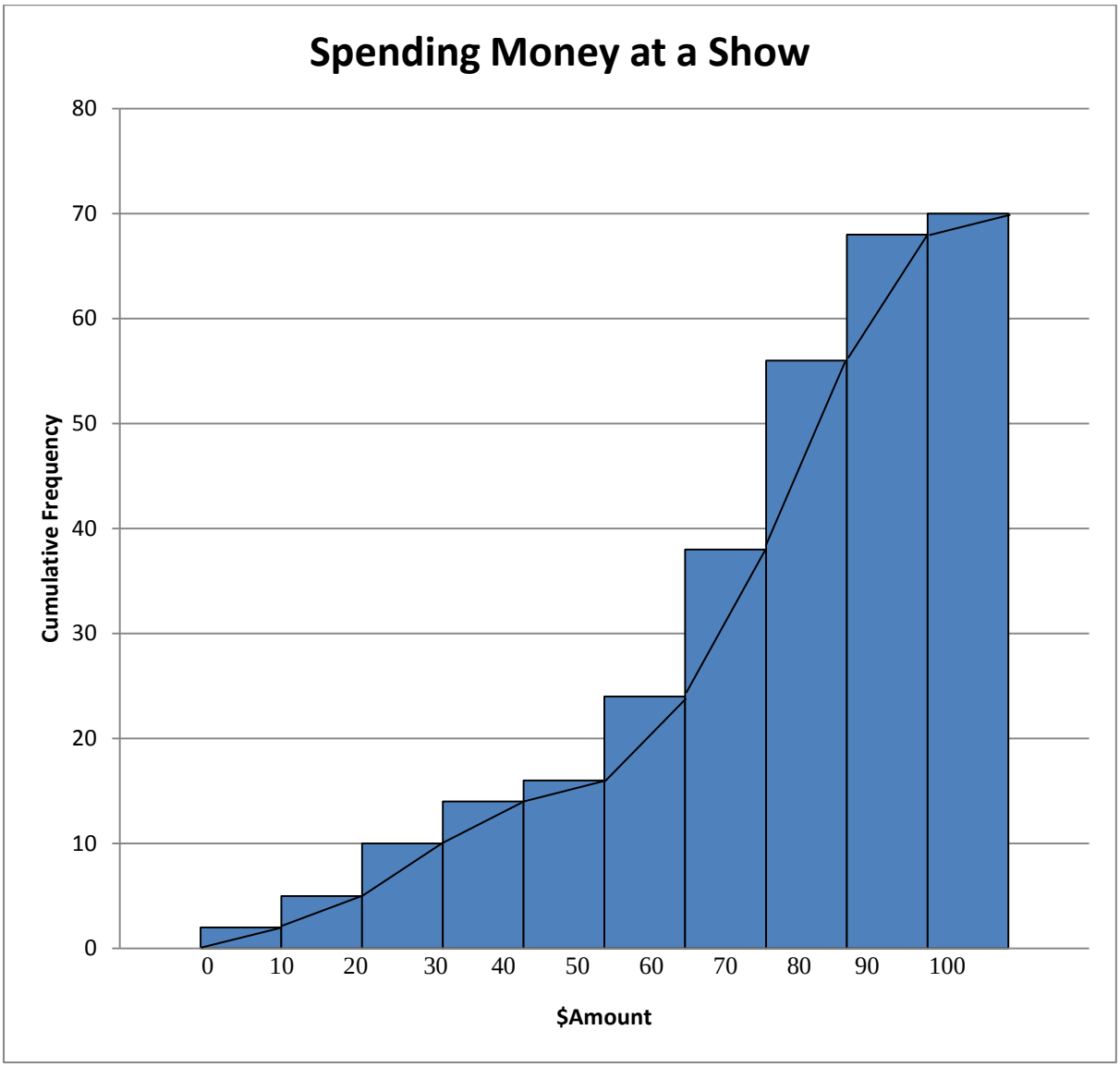


Another type of graph is based on the Cumulative Frequency Histogram.

A line is draw on this called an Ogive.

The Ogive can be drawn with or without the histogram.

The Ogive is drawn from the previous cumulative frequency on the left of the column to the current cumulative frequency on the right of the column.



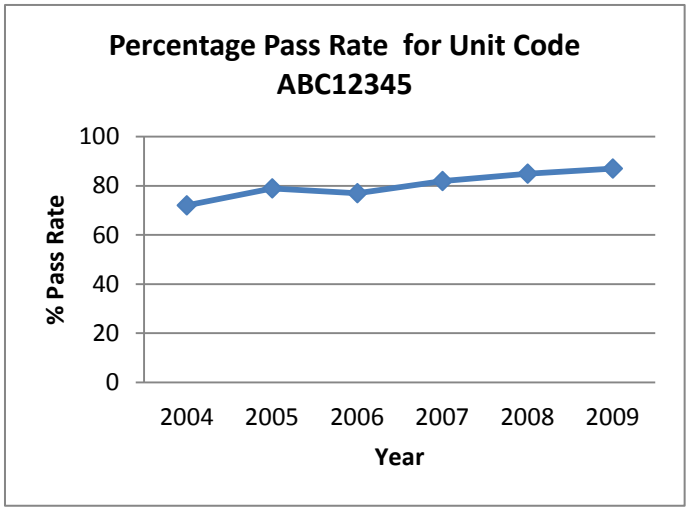
Line Graphs

Another important type of graph is the Line Graph.

A line graph shows change of a variable (usually) over a period of time.

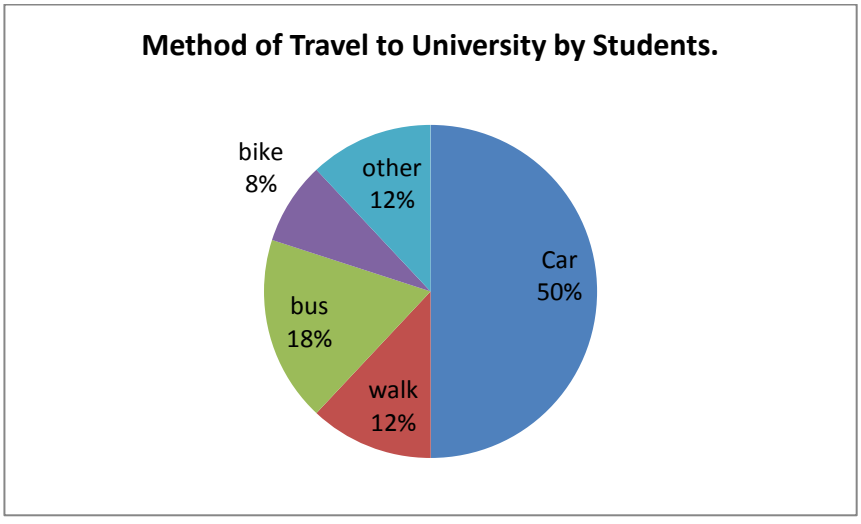
Because none of the data collected for our scenario is suitable for a line graph, a simple example has been made up.

Year	% Pass Rate
2004	72
2005	79
2006	77
2007	82
2008	85
2009	87



Pie or Sector Graphs

The next graph is very useful when you are trying to show how a whole set of data is divided up into components. For example: If you collected data about how students travel to university, the whole group can be divided up into those that travel by bus, car, bike, walk etc. A pie graph is good at showing the whole in its individual parts.



Visually, it is easy to say that most students travel to uni by car. With the percentages given, it is also possible to quantify information from the graph.

For example: If there are 2000 students on campus today, approximately how many travelled by bike? Number travelling by bike =12% of 2000 = 0.12 x 2000 = 240 students.

A variation of the pie graph is the percentage bar graph. Ideally a bar of length 100mm (or 200mm, 300mm etc.) is divided by in the percentages given.

Car 50%	Bus 18%	Walk 12%	Bike 8%	Other 12%

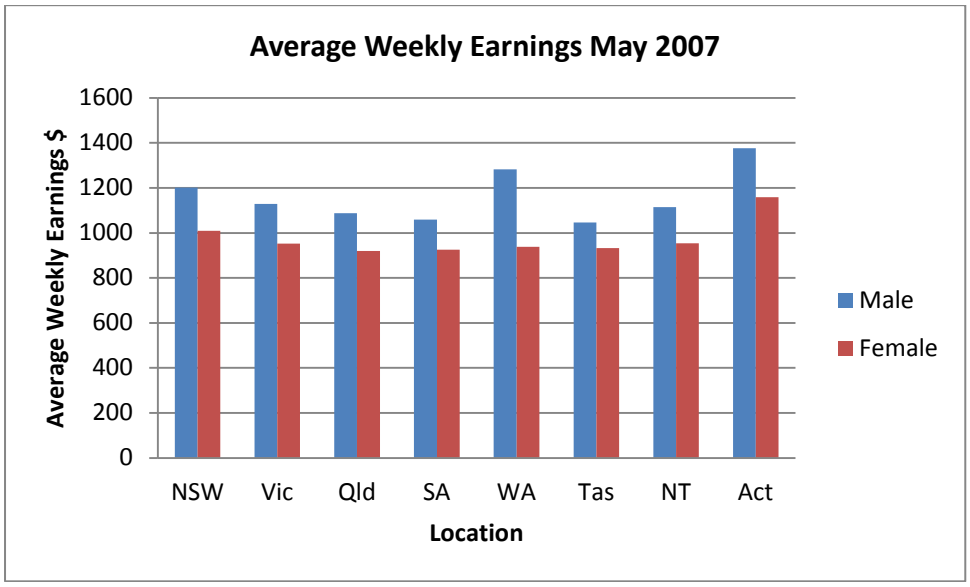
Composite Bar / Column Graph

Another graph to consider is a composite bar graph.

This graph can be drawn either vertically or horizontally.

It can be used to compare two or more sets of similar data.

The graph below compares average weekly earnings in May 2007 for males and females working full time in different states and territories.



Data sourced from:
http://www.abs.gov.au/ausstats/abs@.nsf/2f762f95845417aeca25706c00834efa/0796FAC7CDEEA671CA2573D20010F2DD?opendocument

Video ‘Presenting Data in Graphs’

Activity

1. The height of a plant is measured every Friday morning for twelve weeks. Which type of graph would be best to show the growth of the plant?
2. A 'sound and vision' shop sells CDs, DVDs, console games and computer games. Which type of graph would best display the relative sales of the different items sold?
3. A student wishes to compare the number of motorcycle fatalities between the states. To get a more accurate picture, the data is collected for the years 2007, 2008 and 2009. Which type of graph allows all this data to be presented in one graph?
4. The number of students attending a class (maximum 25) for 30 lessons is given in the table below:

Number attending	Tally	Frequency
20	I	1
21	I	1
22	II	2
23	$\cancel{\text{IIII}}$	5
24	$\cancel{\text{IIII}}$ $\cancel{\text{IIII}}$ I	11
25	$\cancel{\text{IIII}}$ $\cancel{\text{IIII}}$	10
	Total	30

Construct a column graph to represent this data.

5. The Shot Put distances thrown by 27 world champion shot putters are given in the Frequency Distribution Table below. The unit is metres (m).

Distance	Tally	Frequency	Cumulative Frequency	% Relative Frequency
20 to but less than 20.5	$\cancel{\text{IIII}}$	5	5	$\frac{5}{27}\times 100 = 18.5\%$
20.5 to but less than 21	II	2	7	$\frac{2}{27}\times 100 = 7.4\%$
21 to but less than 21.5	$\cancel{\text{IIII}}$ I	6	13	$\frac{6}{27}\times 100 = 22.2\%$
21.5 to but less than 22	$\cancel{\text{IIII}}$ I	6	19	$\frac{6}{27}\times 100 = 22.2\%$
22 to but less than 22.5	IIII	4	23	$\frac{4}{27}\times 100 = 14.8\%$
22.5 to but less than 23	II	2	25	$\frac{2}{27}\times 100 = 7.4\%$
23 to but less than 23.5	II	2	27	$\frac{2}{27}\times 100 = 7.4\%$
	Total	27		100%

(a) Construct a frequency histogram for this information. As a second step, put a frequency polygon on the histogram.

(b) Construct a cumulative frequency histogram and then add an Ogive to the histogram.

Topic 3: Measures of Central Tendency

With any set of numerical data, there is always a temptation to summarise the data to a single value that attempts to be a typical value for the data. There are three commonly used measures to describe the 'central' or 'typical' value of the data, the Mean, Median and Mode.

Mean

The mean is commonly referred to as the average.

The mean is the most common measure used. It is found by adding up all the values and dividing by the number of values. Because of this, **every value** contributes to the mean.

For a set of values written as $X_1, X_2, X_3, X_4, X_5, \ldots\ldots X_n$, the **sample mean** is calculated using the equation:

$$\overline{X} = \frac{\text{sum of the values}}{\text{number of values in the sample}} = \frac{\sum X}{n}$$

where $\overline{X}$ is the mean of the n values in the sample.

The symbol Σ is shorthand for "the sum of".

The equation for the **population mean** uses some different pronumerals.

$$\mu = \frac{\text{sum of the values}}{\text{number of values in the population}} = \frac{\sum X}{N}$$

where μ is the mean of the N values in the population.

Because the calculation of the sample and population means is the same (except for the pronumerals used in the equation), calculators have just one key to cover both.

Finding the mean of Individual Values

This data is the test results of a sample 15 students on a maths test (in order).

23, 45, 50, 54, 55, 57, 59, 59, 59, 61, 63, 75, 75, 81, 90.

$$\overline{X} = \frac{\sum X}{n}$$

$$\overline{X} = \frac{23+45+50+.....+75+81+90}{15}$$

$$\overline{X} = 60.4$$

The mean can easily be calculated on a calculator, especially a scientific calculator with STATS mode (covered later).

Finding the mean from a Frequency Distribution Table

Let's recall the data about the number of matches in 50 boxes which was organised in a FDT with no grouping. The table has a new column added headed $f \times x$. This column is the **score** multiplied by the **frequency**.

Number of Matches *(X)*	Frequency *(f)*	$f \times X$
46	2	2 x 46 = 92
47	3	3 x 47 = 141
48	5	240
49	6	294
50	16	800
51	6	306
52	4	208
53	3	159
54	3	162
55	2	110
Total	$\Sigma f = 50$	$\Sigma fX = 2512$

92 represents the total of all the 46s in the data.

141 represents the total of all the 47s in the data.

2512 is the total of all the values in the data.

The mean number of matches per box is given by the equation:

$$\overline{X} = \frac{\text{sum of the values}}{\text{number of values in the sample}}$$

$$\overline{X} = \frac{\Sigma fX}{\Sigma f}$$

$$\overline{X} = \frac{2512}{50}$$

$$\overline{X} = 50.24$$

The mean number per box is 50.24 matches.

Now let's look at the example about the money spent by children at the show. This was organised using a Grouped FDT.

Finding the mean is performed in a similar manner to above except the Group Midpoint is used to represent the group.

Amount Spent$	Frequency *(f)*	Group Midpoint *(X)*	$f \times X$
$0 but less than $10	2	5	2 x 5 = 10
$10 but less than $20	3	15	3 x 15 = 45
$20 but less than $30	5	25	125
$30 but less than $40	4	35	140
$40 but less than $50	2	45	90
$50 but less than $60	8	55	440
$60 but less than $70	14	65	910
$70 but less than $80	18	75	1350
$80 but less than $90	12	85	1020
$90 but less than $100	2	95	190
Total	$\Sigma f = 70$		$\Sigma fX = 4320$

The mean amount spent is given by the equation:

$$\overline{X} = \frac{\text{sum of the values}}{\text{number of values in the sample}}$$

$$\overline{X} = \frac{\Sigma fX}{\Sigma f}$$

$$\overline{X} = \frac{4320}{70}$$

$$\overline{X} = \$61.71 \text{ (to the nearest cent)}$$

Remember this method assumes that the values in each group are evenly spread. This assumption is not always true so the figure obtained for the mean using this method can slightly different to the mean if the original values where used.

Finding the mean using a Calculator

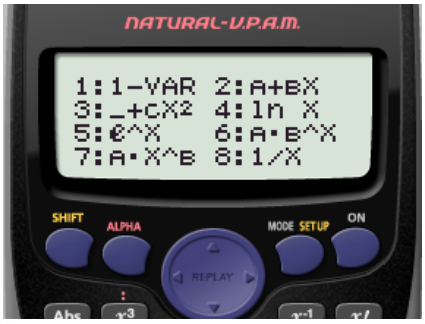
A scientific calculator is usually capable of performing this operation when STATS mode is used.

The instructions below apply to the Casio fx-82AU and the Sharp EL531 only. If you have a different scientific calculator to this, you will need to read the user's guide for your calculator.

There are 3 steps to using a scientific calculator in STATS mode:

1. Getting the calculator into STATS mode.
2. Entering the values (data).

3. Obtaining the statistical measure required.

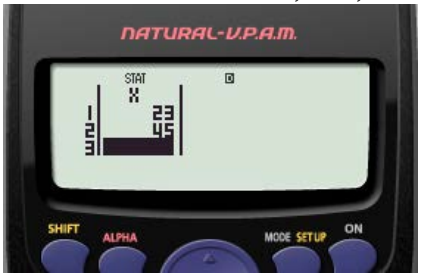
On a Casio fx-82AU:	On a Sharp EL531
To get this calculator into STATS mode	To get this calculator into STATS mode
Press *Mode* . The display will be:  Press 2 for *STAT*  Press 1 for *1-VAR* The calculator should display a small ***STAT*** on the top line of the display.	Press *Mode* . Press 1 for STAT Press 0 for *Standard Deviation* (+ more) The calculator should display ***Stat 0*** in the top left of the display.

Now your calculator is in STATs mode, the next stage is to **enter the data**. There are slight differences depending on how the data is organised.

Finding the Mean of Individual Values

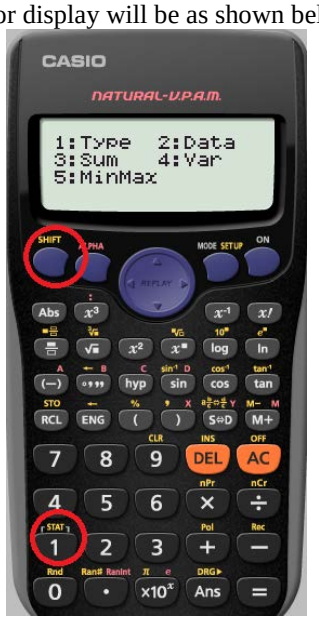
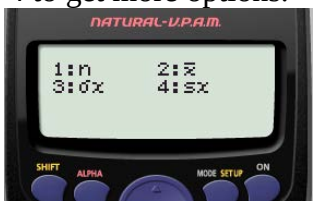
This data is the test results of a sample 15 students on a maths test (in order).

23, 45, 50, 54, 55, 57, 59, 59, 59, 61, 63, 75, 75, 81, 90.

On a Casio fx-82AU:	On a Sharp EL531
Each number is entered into the list, 23 =, 45 =:  	Each number from the list is entered as: 23 then press *M +* the calculator display will show DATA SET= 1 45 then press *M +* the calculator display will show DATA SET= 2 . . . . .

If you make a mistake, use the big round blue 'Replay' button to navigate back to the position of the incorrect value and enter the correct value.	90 then press $M+$ the calculator display will show DATA SET= 15 The display DATA SET= 15 informs you that 15 numbers have been entered into the STATs memory. If you make a mistake and need to start again, the STATs memory is cleared by pressing $2ndF$ CA

Now the data is entered, the calculated value of the sample and population standard deviations can be obtained.

On a Casio fx-82AU:	On a Sharp EL531
When all the values have been entered, press AC This is important. Now press $SHIFT$ 1. This is the same as $SHIFT$ $STAT$. The calculator display will be as shown below.  Now press 4 to get more options.  Press 2 = – for the mean Press 3 = – for the population standard dev. Press 4 = – for the sample standard dev. Always press = to get the value required.	To obtain the value of the population standard deviation; press RCL σx the symbol σ is used to represent the population standard deviation. To obtain the value of the sample standard deviation; press RCL sx the symbol s is used to represent the sample standard deviation. (Think of RCL as recall) Other values are found by; To obtain the value of the mean; press RCL $\bar{x}$ the symbol $\bar{x}$ means the mean. To obtain the number of values entered; press RCL n the symbol n means the number of numbers entered. To obtain the sum of the values entered; press RCL Σx the symbol Σx means the sum of the numbers entered. To obtain the sum of the squares of the values entered; press RCL Σx^2 the symbol Σx^2 means the sum of the squares of the numbers entered.

Do the calculation yourself, the values obtained should be:

Mean = 60.4

Finding the Mean from a Frequency Distribution Table

Let's recall the data about the number of matches in 50 boxes which was organised in a FDT with no grouping.

Number of Matches *(X)*	Frequency *(f)*
46	2
47	3
48	5
49	6
50	16
51	6
52	4
53	3
54	3
55	2
Total	$\Sigma f = 50$

It is worth recalling that this table is informing us that there are 2 occurrences of the value 46, 3 occurrences of 47, 5 occurrences of 48, ….
Instead of entering the value of 50, 16 times a modified entry process is used. The process for finding the standard deviation is exactly the same.

Remember to clear STATs memory first.

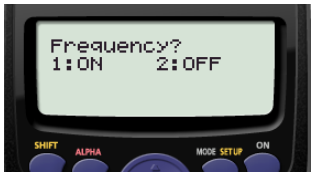
With a Frequency Distribution Table, the data is entered into the **Sharp EL531** calculator in the format: *score, frequency* .

On the **Casio fx-82AU**, the table the values are entered into needs to be changed into a frequency like the example above. To do this:

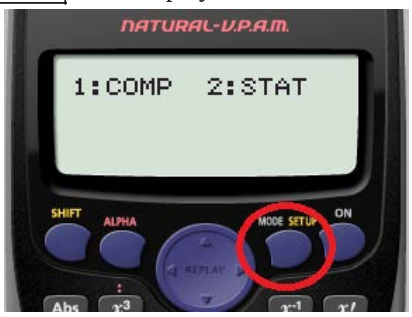
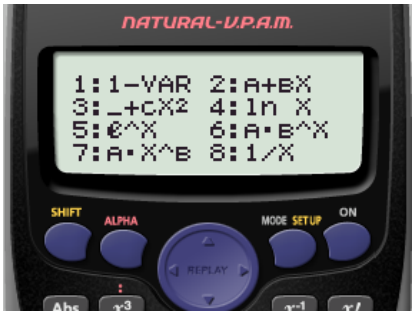
Press *SHIFT* *MODE* . Eight options will be displayed.

Press the big blue key down  .
This will take you to another screen with 5 options.

Press 3: STAT The screen will then display



Press 1: ON This will add a frequency column

On a Casio fx-82AU:	On a Sharp EL531
Get the calculator into STATS mode: Press [*Mode*] . The display will be:  Press [2] for *STAT*  Press [1] for *1-VAR* This time the calculator will display an X column and a FREQ column. Enter a value followed by =. Navigate to the frequency column using the big blue REPLAY key.  	Each row from the table is entered as: [46] [,] [2] [*M +*] the calculator display will show [*DATA SET=* *1*] [47] [,] [3] [*M +*] the calculator display will show [*DATA SET=* *2*] The display [*DATA SET=* *2*] informs you that 2 **rows** have been entered into the STATs memory. [48] [,] [5] [*M +*] the calculator display will show [*DATA SET=* *3*] . . . . [55] [,] [2] [*M +*] the calculator display will show [*DATA SET=* *15*] If you make a mistake and need to start again, the STATs memory is cleared by pressing [*2ndF*] [*CA*]

Obtaining values for the mean is exactly the same process as in the previous section.

Do the calculation yourself, the values obtained should be:

Mean = 50.24

If the data is grouped, the process is almost identical except that the group midpoint is used.

Amount Spent$	Group Midpoint *(X)*	Frequency *(f)*
$0 but less than $10	5	2
$10 but less than $20	15	3
$20 but less than $30	25	5
$30 but less than $40	35	4
$40 but less than $50	45	2
$50 but less than $60	55	8
$60 but less than $70	65	14
$70 but less than $80	75	18
$80 but less than $90	85	12
$90 but less than $100	95	2
Total		$\Sigma f = 70$

The data is entered into the calculator as Group Midpoint, Frequency or Group Midpoint; Frequency depending on the calculator used.

Median

The median is the middle value of the data if the values are arranged in order.

As a measure of centre, it is a value based on its position; it is not influenced by the size of the values that are above or below it.

Therefore it is quite different to the mean because it is based on position and for the mean, every value contributes.

Finding the median of Individual Values

This data is the test results of a sample 15 students on a maths test (in order).

23, 45, 50, 54, 55, 57, 59, 59, 59, 61, 63, 75, 75, 81, 90.

When finding the median for a set of individual values, the first step is to order the values. Usually this is done in ascending order, but descending order gives the same result.

The **position** of the median can always be found by using a simple expression:

$$\text{The position of the median is } \frac{n+1}{2}.$$

As there are 15 values, the position of the median is:

$$\frac{n+1}{2} = \frac{15+1}{2} = \frac{16}{2} = 8$$

The 8^{th} value is the median.

The median is the 8th value, which is 59. The median is the 8th value, so there are 7 values before it and 7 values after it.

23, 45, 50, 54, 55, 57, 59	59	59, 61, 63, 75, 75, 81, 90
7 values	median	7 values

= 15 values

The median cannot be calculated on a scientific calculator.

Finding the median from a Frequency Distribution Table

Let's recall the data about the number of matches in 50 boxes which was organised in a FDT with no grouping.

To find the median, the cumulative frequency column is required.

Number of Matches *(X)*	Frequency *(f)*	Cumulative Frequency
46	2	2
47	3	5
48	5	10
49	6	16
50	16	32
51	6	38
52	4	42
53	3	45
54	3	48
55	2	50
Total	$\Sigma f = 50$	

The 17th through to the 32nd values are in this group.

This means that the 25th and 26th values are both 50.

The median is 50.

As there are 50 values, the position of the median is: $\frac{n+1}{2} = \frac{50+1}{2} = \frac{51}{2} = 25.5$

In this case the 25th and 26th values are required. The median is the average of the 25th and 26th values. As the 25th and 26th values are both 50, the median is 50.

Now let's look at the example using a Grouped FDT. Finding the median can be performed in one of two ways: interpolation method or graphical method.

(a) Interpolation Method (Amount Spent at a Show Data)

Amount Spent$	Frequency *(f)*	Group Midpoint *(X)*	Cumulative Frequency
$0 but less than $10	2	5	2
$10 but less than $20	3	15	5
$20 but less than $30	5	25	10
$30 but less than $40	4	35	14
$40 but less than $50	2	45	16
$50 but less than $60	8	55	24
$60 but less than $70	14	65	38
$70 but [illegible]	18	[illegible]	[illegible]
$80 bu[illegible]	2	[illegible]	[illegible]
$90 – $100	2	95	70
Total	$\Sigma f = 70$		

This means that the 24th score is $60

This means that the 38th score is $70

The group width =70-60 =10

There are 14 values in this group.

As there are 70 values, the position of the median is: $\frac{n+1}{2} = \frac{70+1}{2} = \frac{71}{2} = 35.5$

In this case the 35th and 36th values are required. Both of these values are located in the group '$60 but less than $70'. This means that the median is greater than 60 but less than 70. This means that the median is (35.5 – 24 =) $^{11.5}/_{14}$ of the way through the '$60 but less than $70' group. The median is:

$$Median = 60 + \frac{(35.5 - 24)}{14} \times 10 \ (Group\ width)$$

$$Median = 68.2$$

Or expressed more generally:

$$Median = L_m + \frac{\left(\frac{n+1}{2}\right) - cf_{m-1}}{f_m} \times Group\ width$$

where L_m *is the lower limit of the median group*

f_m *is the frequency of the median group*

cf_{m-1} *is the cumulative frequency of the previous group*

(b) Graphical method (using the Ogive)

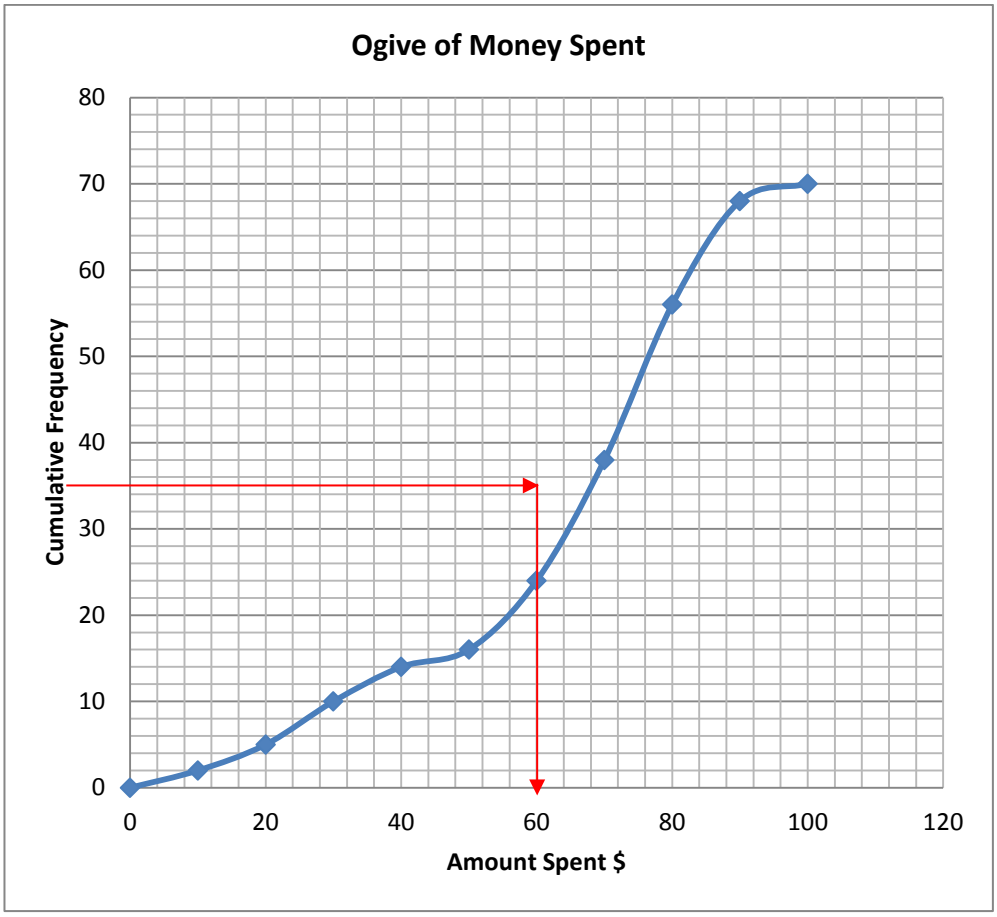
There are 70 values in the table. This is shown on the vertical axis (Cumulative Frequency).

The **first step** is to come half way up the vertical axis.

This is to 35.

The **second step**; from this position, move horizontally across the graph until the cumulative frequency Ogive is found.

The **third step** is to move down (vertically) to the horizontal axis and read the value.



From the graph, the median is approximately 68.

Remember this method assumes that the values in each group are evenly spread. This assumption is not always true, so the figure obtained for the median using this method can slightly different to the actual median if the original values where used.

Mode

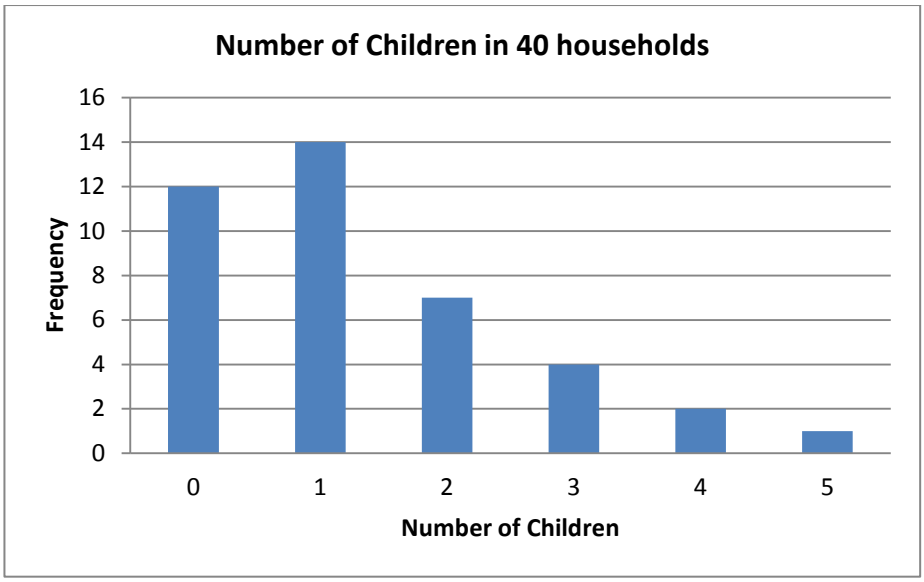
The Mode is the value that occurs the most often.

The mode may not exist because values only occur once.

The mode could even be quite different to the mean and median, it may be much higher or much lower depending on the meaning of the data.

There may also be more than one mode. If there are two modes, the data is said to be ‘bimodal’.

Finding the mode from Graphs



From the graph, it can be observed that '1 child per household' has the highest frequency. This is the mode of these values.

Finding the mode of Individual Values

This data is the test results of a sample 15 students on a maths test (in order).

23, 45, 50, 54, 55, 57, 59, 59, 59, 61, 63, 75, 75, 81, 90.

When inspecting this data it can be seen that the 59 occurs 3 times.

This is the value that occurs the most frequently, that is, it has the highest frequency.

The mode for this data is 59.

Finding the mode from a Frequency Distribution Table

Let’s recall the data about the number of matches in 50 boxes which was organised in a FDT with no grouping.

To find the mode, the frequency column is inspected. The highest frequency of 16 is related to a box containing 50 matches. The mode of this data is 50 matches. Take care to give the mode as the score (50 matches) not the frequency (16).

Number of Matches *(X)*	Frequency *(f)*
46	2
47	3
48	5
49	6
50	16
51	6
52	4
53	3
54	3
55	2
Total	$\Sigma f = 50$

The mode for this data is 50 matches as it has the highest frequency (occurs the most often)

When data is grouped, it is more relevant to refer to the modal group.
From the scenario, ‘Amount Spent at the Show’, the group with the highest frequency is '$70 but less than $80' which has a frequency of 18.

It can be stated that the modal group is '$70 but less than $80'.

Amount Spent$	Frequency *(f)*	Group Midpoint *(X)*
$0 but less than $10	2	5
$10 but less than $20	3	15
$20 but less than $30	5	25
$30 but less than $40	4	35
$40 but less than $50	2	45
$50 but less than $60	8	55
$60 but less than $70	14	65
$70 but less than $80	18	75
$80 but less than $90	12	85
$90 but less than $100	2	95
Total	$\Sigma f = 70$	

The modal group for this data is ‘$70 but less than $80’ as it has the highest frequency (occurs the most often)

Either the group ‘$70 but less than $80’ or the Group Midpoint ‘$75’ can be given as the mode.

Which measure of centre should be used and when?

Every value from the data is used to calculate the mean. As long as the data is spread out without excessive variation, the mean is usually used. A lecturer would use the mean to find a measure of centre because the marks would be usually spread out from (say) 30 to 100%. In this data the low or high values are not excessively different to a measure of centre.

In real estate, the median is often used as a measure of centre because there are some excessively high values that are quite different to the bulk of the market. Real estate sales generally consist of many properties at the cheaper part of the market and fewer properties at the expensive part of the market. If the mean was to be used, the fewer, more expensive properties would have a huge effect on the mean. Because the median is a measure of centre based on location, variation in the price of high value properties and the number of high value properties will have no effect on the median but would have a significant effect on the mean.

In the clothing industry, the mode is often used as a measure of centre. If a shop sells mostly size 12 clothing, then size 12 is the mode of the sizes sold. The mean is of little use because it could be a value that is not even a possible size. The median is a better measure but fails to reflect the nature of the sales environment.

 Video ‘Measures of Central Tendency’

Activity

1. The lifetime, in hours, of a sample of 15 light bulbs is: 351, 429, 885, 509, 317, 753, 827, 737, 487, 726, 395, 773, 926, 688, 485.

   Calculate the mean, mode and median of the values. Do not organise into a table.

2. The number of children in a 10 families is: 1, 5, 2, 2, 2, 3, 1, 4, 3, 2. Calculate the mean, mode and median of the values. Do not organise into a table.

3. For the Car Speed Data, calculate the mean, mode and median of the values after organising the data in Frequency Distribution Table (The FDT can be found in the answers to the Topic 'Graphs').

Car Speed Data (in km/hr)

12	41	44	45	28	40	32	62	46	25
31	35	31	20	59	27	49	19	58	38
22	50	46	14	33	48	25	32	52	69
40	52	57	27	61	42	39	64	52	27

4. The number of students attending a class (maximum 25) for 30 lessons is given in the table below:

Students Attending Class

25	24	24	25	24	23
25	24	23	24	25	25
24	25	20	23	25	24
22	24	25	24	23	21
25	23	24	25	24	22

   (a) Calculate the mean, mode and median using a Frequency Distribution Table. (The FDT can be found in the answers to the Topic 'Graphs').

   (b) Discuss the appropriateness of each measure of central tendency as a typical value.

5. The Shot Put distances thrown by 27 world champion shot putters are given in the table below. The unit is metres (m).

22.25	20.19	21.39	21.25	21.19	22.07
21.72	20.37	20.45	23.09	21.19	21.22
21.07	21.55	23.12	20.91	22.58	21.97
20.22	20.38	22.37	22.19	21.70	20.54
22.67	21.58	21.72			

   Calculate the mean, mode and median using a grouped Frequency Distribution Table. (The FDT can be found in the answers to the Topic 'Graphs')

Topic 4: Measures of Spread

Consider the two sets of values below. They both have a mean and median of 50, but they are quite different.

Set A: 20, 35, 50, 65, 80 Set B: 40, 45, 50, 55, 60

The reason the sets are different is because of the spread of the values; the values in Set A are more spread out than those in Set B.

Range

In statistics, not only is a measure of centre important but a measure of spread is also important. The most basic measure of spread is the range.

Range = Highest Value - Lowest Value

For the data above,

Set A	Set B
Range = Highest Value - Lowest Value Range = 80 – 20 Range = 60	Range = Highest Value - Lowest Value Range = 60 – 40 Range = 20

From this it is possible to say that Set A is has a higher spread than Set B. The problem with the range is that it uses the lowest and highest values. In any set of data, the lowest or highest score could be an odd or unusual value.

Using odd or unusual values to measure spread will not produce a result that properly reflects the data. If you consider students sitting an examination, a student not feeling well sits an exam and records an unusually low score. This score is not typical of the group and so the range is much larger than it should be for the group.

Scores that are atypical of the group are called outliers. There are ways of identifying outliers, but these will not be covered in this unit.

Finding the range of Individual Values

This data is the test results of a sample 15 students on a maths test (in order).

23, 45, 50, 54, 55, 57, 59, 59, 59, 61, 63, 75, 75, 81, 90.

Once the data is ordered, the lowest and highest values are easy to locate.

Range = Highest Value - Lowest Value
Range = 90 – 23
Range = 67

Finding the range from a Frequency Distribution Table

Let's recall the data about the number of matches in 50 boxes which was organised in a FDT with no grouping. To find the range, the 'Number of Matches' column is used.

Number of Matches *(X)*	Frequency *(f)*
46	2
47	3
48	5
49	6
50	16
51	6
52	4
53	3
54	3
55	2
Total	$\Sigma f = 50$

The lowest score obtained was 46. The highest score obtained was 55.
Range = 55 – 46 = 9

When data is grouped it is more difficult to determine the highest and lowest values so an assumption must be made.

Amount Spent$	Frequency *(f)*	Group Midpoint *(X)*
$0 but less than $10	2	5
$10 but less than $20	3	15
$20 but less than $30	5	25
$30 but less than $40	4	35
$40 but less than $50	2	45
$50 but less than $60	8	55
$60 but less than $70	14	65
$70 but less than $80	18	75
$80 but less than $90	12	85

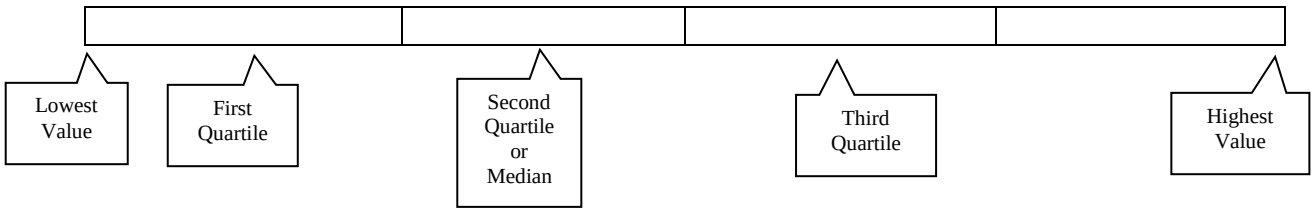
\$90 but less than \$100	2	95
Total	$\Sigma f = 70$	

The only assumption (because the original values are not present) that can be made here is that the lowest value is \$0 and the highest value is \$100.
The range is 100 – 0 = 100

Interquartile Range

The next measure of spread is the Interquartile Range (IQR). It is still a range but it is between the first and third quartiles. The first and third quartiles are values that are ¼ and ¾ of the way through the ordered data. Another way to think about the IQR is to consider it the range of the middle 50% of the data.

Diagrammatically, it can be represented as below:



Inter-Quartile Range = Third Quartile - First Quartile

Or

$$IQR = Q_3 - Q_1$$

The five values shown in this diagram are commonly referred to as a 'five number summary'. A five number summary is very useful for calculating the range, IQR and for drawing a box and whisker diagram (covered later).

Finding the interquartile range of Individual Values

This data is the test results of a sample 15 students on a maths test (in order).

23, 45, 50, 54, 55, 57, 59, 59, 59, 61, 63, 75, 75, 81, 90.

Once again the data must be ordered. Previously a simple equation was used to locate the median. This is similar.

The first quartile is the $\frac{n+1}{4}^{th}$ *score.* *The third quartile is the* $\frac{3(n+1)}{4}^{th}$ *score.*

For the data above:

First Quartile (Q_1)	Third Quartile (Q_3)
$\frac{n+1}{4}$ $=\frac{15+1}{4}$ $=\frac{16}{4}$ $=4^{th}$ value	$\frac{3(n+1)}{4}$ $=\frac{3\times16}{4}$ $=\frac{48}{4}$ $=12^{th}$ value

23, 45, 50, 54, 55, 57, 59, 59, 59, 61, 63, 75, 75, 81, 90.

4^{th} value

12^{th} value

$$IQR = Q_3 - Q_1$$
$$IQR = 75 - 54$$
$$IQR = 21$$

The advantage of using the IQR is that the first and third quartiles are stable values; they are not outliers or 'odd' values.

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

Finding the interquartile range from a Frequency Distribution Table

Let's recall the data about the number of matches in 50 boxes which was organised in a FDT with no grouping. To find the IQR, the 'Cumulative Frequency' column is used.

Number of Matches *(X)*	Frequency *(f)*	Cumulative Frequency
46	2	2
47	3	5
48	5	10
49	6	16
50	16	32
51	6	38
52	4	42
53	3	45
54	3	48
55	2	50
Total	$\Sigma f = 50$	

The 11^{th} through to the 16^{th} values are 49 matches, this includes the 12^{th} and 13^{th} values.

The first quartile is 49.

The 33^{rd} through to the 38^{th} values are 51 matches.

The 38^{th} value is 51 matches.

The 39^{th} through to the 42^{nd} values are 52 matches.

The 39^{th} value is 52 matches.

First Quartile *(Q_1)*	Third Quartile *(Q_3)*

$\frac{n+1}{4}$ $=\frac{50+1}{4}$ $=\frac{51}{4}$ $=12.75^{th}$ value	$\frac{3(n+1)}{4}$ $=\frac{3\times 51}{4}$ $=\frac{153}{4}$ $=38.25^{th}$ value

To find the 12.75th value, means finding the 12th and 13th values. In this case the 12th and 13th values are both 49, so the First Quartile is 49.

To find the 38.25th value, means finding the 38th and 39th values. In this case the 38th value is 51 and the 39th value is 52, so the Third Quartile is 51+ 0.25 x Difference between the 51st and 52nd values.

This is equal to 51+ 0.25 x 1 = 51.25

$$IQR = Q_3 - Q_1$$
$$IQR = 51.25 - 49$$
$$IQR = 2.25$$

When data is grouped, the cumulative frequency is used to obtain the quartiles. There are two methods to find the quartiles; the interpolation method and the graphical method.

(a) Interpolation Method

Amount Spent$	Frequency *(f)*	Group Midpoint *(X)*	Cumulative Frequency
$0 but less than $10	2	5	2
$10 but less than $20	3	15	5
$20 but less than $30	5	25	10
$30 but less than $40	4	35	14
$40 but less than $50	2	45	16
$50 but less than $60	8	55	24
$60 but less than $70	14	65	38
$70 but less than $80	18	75	56
$80 but less than $90	12	85	68
$90 but less than $100	2	95	70
Total	$\Sigma f = 70$		

The 17th through to the 24th values are found in the group '$50 but less than $60'

The 17.75th value is in this group.

The 39th through to the 56th values are found in the group '$70 but less than $80'

The 53.25th value is in this group.

First Quartile (Q_1)	Third Quartile (Q_3)
$\frac{n+1}{4}$ $=\frac{70+1}{4}$ $=\frac{71}{4}$ $=17.75^{th}$ value	$\frac{3(n+1)}{4}$ $=\frac{3\times 71}{4}$ $=\frac{213}{4}$ $=53.25^{th}$ value

The first quartile is found in the group '\$50 but less than \$60'.

In this group there are 8 values (the 17th to the 24th) and the class width is 10.

This means the First Quartile is $\frac{17.75-CF_{PreviousGroup}}{GroupFrequency}=\frac{17.75-16}{8}=\frac{1.75}{8}$ of the way through the group '\$50 but less than \$60'.

The third quartile is found in the group '\$70 but less than \$80'.

In this group there are 18 values (the 39th to the 56th) and the class width is 10.

This means that the Third Quartile is $\frac{53.25-CF_{PreviousGroup}}{GroupFrequency}=\frac{53.25-38}{18}=\frac{15.25}{8}$ of the way through the group '\$70 but less than \$80'.

$$\textit{The first quartile } (Q_1)=50+\frac{1.75}{8}\times 10 \qquad \textit{The third quartile } (Q_3)=70+\frac{15.24}{18}\times 10$$
$$=52.19 \qquad =78.47$$

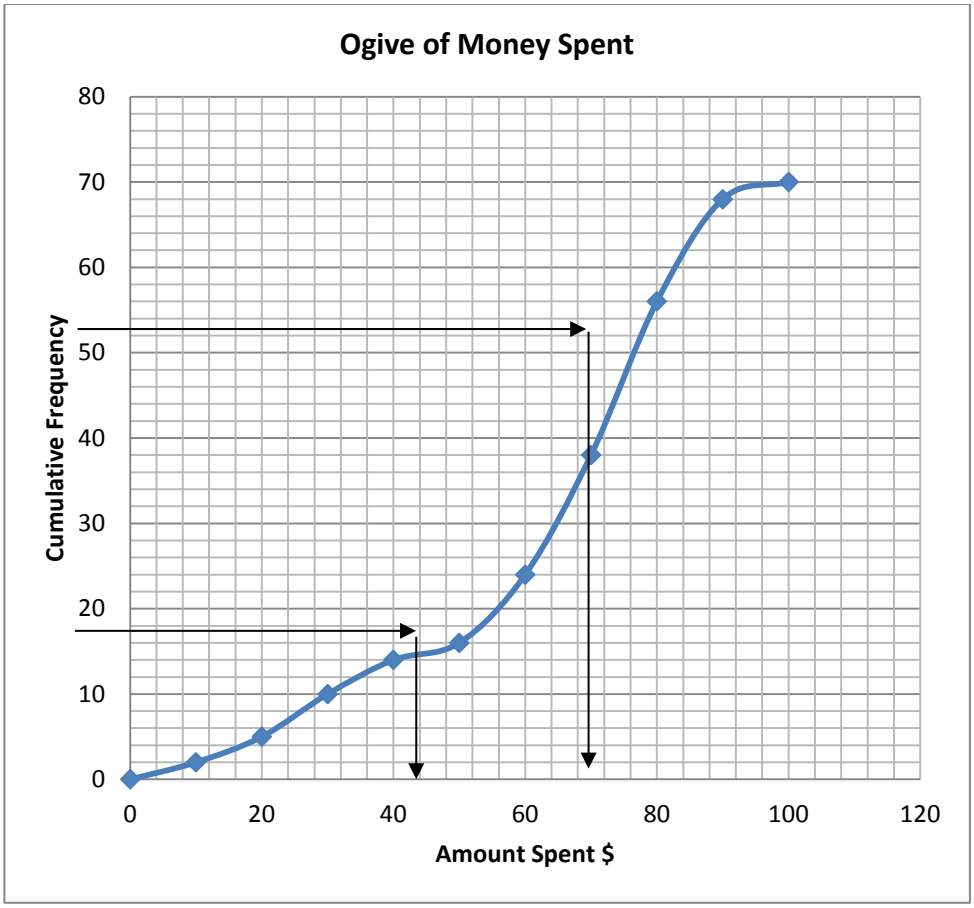
$$IQR=Q_3-Q_1$$
$$IQR=78.47-52.19$$
$$IQR=26.28$$

(b) Graphical Method (using the Ogive)

There are 70 values in the table. This is shown on the vertical axis.

To find the first quartile, come one quarter of the way up the vertical axis. This is to 17.5 (One quarter of 70). From this position, move horizontally across the graph until the cumulative frequency line is found, then move down (vertically) to the horizontal axis and read the value.

To find the third quartile, come three quarters of the way up the vertical axis. This is to 52.5 (three quarters of 70). From this position, move horizontally across the graph until the cumulative frequency line is found, then move down (vertically) to the horizontal axis and read the value.



The graph value for the first quartile is (about) 52 and the third quartile is (about) 78, giving an IQR of 78 – 52 = 26.

Standard Deviation

The third measure of spread is the Standard Deviation. The key word in this measure is deviation. The word deviation in this context is how much each value deviates (differs) from the mean.

For example: consider the sample data set 2, 4, 6, 8.

Data	Mean	Deviation from Mean $\left(X-\bar{X}\right)$	Square the deviations (this makes the deviations positive) $\left(X-\bar{X}\right)^2$
2	$\bar{X}=\frac{\sum X}{n}$	2 - 5 = -3	9
4	$=\frac{2+4+6+8}{4}$	4 - 5 = -1	1
6	$=\frac{20}{4}$	6 - 5 = 1	1
8	$=5$	8 - 5 = 3	9
			$\sum\left(X-\bar{X}\right)^2=20$

The standard deviation of a **sample** is calculated using

$$S=\sqrt{\frac{\sum\left(X-\bar{X}\right)^2}{n-1}}=\sqrt{\frac{20}{4-1}}=\sqrt{6.6666\overline{6}}=2.58$$

If the set of data is a **population**, then the standard deviation has a slightly different equation:

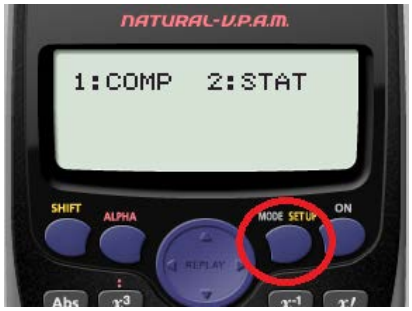
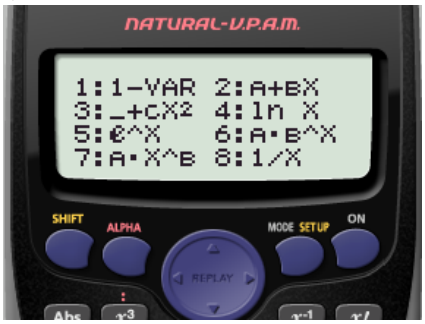
$$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}} = \sqrt{\frac{20}{4}} = \sqrt{5} = 2.24$$

A scientific calculator is usually capable of performing this operation when STATS mode is used.

The instructions below apply to the Casio fx-82AU and the Sharp EL531 only. If you have a different scientific calculator to this, you will need to read the user's guide for your calculator.

There are 3 steps to using a scientific calculator in STATS mode:

1. Getting the calculator into STATS mode.
2. Entering the values (data).
3. Obtaining the statistical measure required.

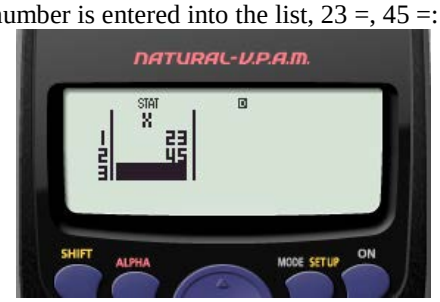
On a Casio fx-82AU:	On a Sharp EL531
To get this calculator into STATS mode	To get this calculator into STATS mode
Press *Mode*. The display will be:  Press 2 for *STAT*  Press 1 for *1-VAR* The calculator should display a small ***STAT*** on the top line of the display.	Press *Mode*. Press 1 for STAT Press 0 for *Standard Deviation* (+ more) The calculator should display ***Stat 0*** in the top left of the display.

Now your calculator is in STATs mode, the next stage is to **enter the data**. There are slight differences depending on how the data is organised.

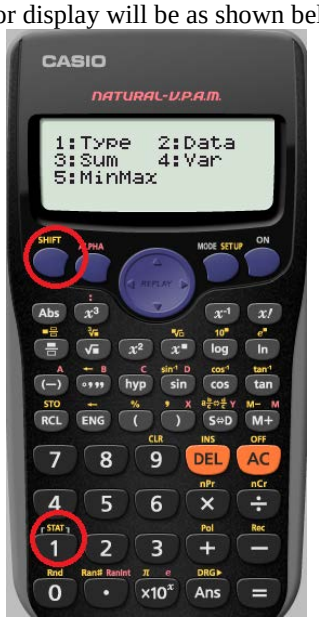
Finding the Standard Deviation of Individual Values

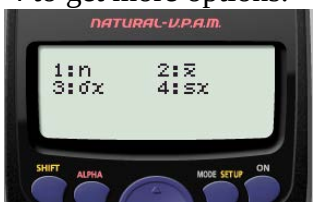
This data is the test results of a sample 15 students on a maths test (in order).

23, 45, 50, 54, 55, 57, 59, 59, 59, 61, 63, 75, 75, 81, 90.

On a Casio fx-82AU:	On a Sharp EL531
Each number is entered into the list, 23 =, 45 =:  If you make a mistake, use the big round blue 'Replay' button to navigate back to the position of the incorrect value and enter the correct value.	Each number from the list is entered as: 23 then press $M+$ the calculator display will show DATA SET= 1 45 then press $M+$ the calculator display will show DATA SET= 2 . . . . . 90 then press $M+$ the calculator display will show DATA SET= 15 The display DATA SET= 15 informs you that 15 numbers have been entered into the STATs memory. If you make a mistake and need to start again, the STATs memory is cleared by pressing $2ndF$ CA

Now the data is entered, the calculated value of the sample and population standard deviations can be obtained.

On a Casio fx-82AU:	On a Sharp EL531
When all the values have been entered, press AC This is important. Now press $SHIFT$ 1. This is the same as $SHIFT$ $STAT$. The calculator display will be as shown below.  	To obtain the value of the population standard deviation; press RCL σx the symbol σ is used to represent the population standard deviation. To obtain the value of the sample standard deviation; press RCL sx the symbol s is used to represent the sample standard deviation. (Think of RCL as recall) Other values are found by; To obtain the value of the mean; press RCL $\bar{x}$ the symbol $\bar{x}$ means the mean. To obtain the number of values entered; press RCL n the symbol n means the number of numbers entered. To obtain the sum of the values entered; press

Now press 4 to get more options.  Press 2 = – for the mean Press 3 = – for the population standard dev. Press 4 = – for the sample standard dev. Always press = to get the value required.	$\boxed{RCL}\boxed{\Sigma x}$ the symbol Σx means the sum of the numbers entered. To obtain the sum of the squares of the values entered; press $\boxed{RCL}\boxed{\Sigma x^2}$ the symbol Σx^2 means the sum of the squares of the numbers entered.

Do the calculation yourself, the values obtained should be:

Population Standard Deviation =15.4177387 rounded gives 15.4

Sample Standard Deviation = 15.95887572 rounded gives 16.0

Finding the Standard Deviation from a Frequency Distribution Table

Let's recall the data about the number of matches in 50 boxes which was organised in a FDT with no grouping.

Number of Matches *(X)*	Frequency *(f)*
46	2
47	3
48	5
49	6
50	16
51	6
52	4
53	3
54	3
55	2
Total	$\Sigma f = 50$

It is worth recalling that this table is informing us that there are 2 occurrences of the value 46, 3 occurrences of 47, 5 occurrences of 48, ….
Instead of entering the value of 50, 16 times a modified entry process is used. The process for finding the standard deviation is exactly the same.

Remember to clear STATs memory first.

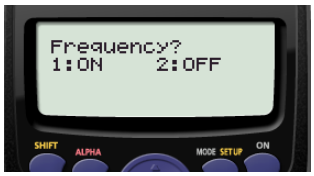
With a Frequency Distribution Table, the data is entered into the Sharp EL531 calculator in the format:
score, *frequency* .

On the Casio fx-82AU, the table the values are entered into needs to be changed into a frequency like the example above. To do this:

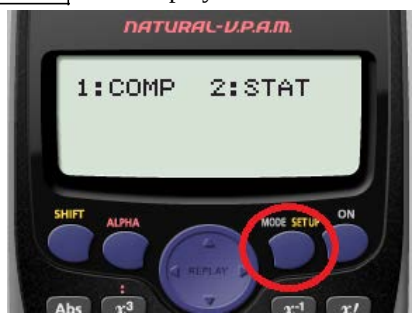
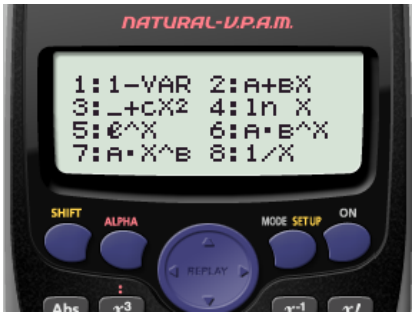
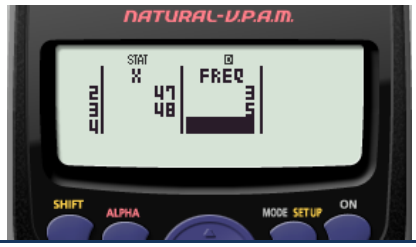
Press SHIFT MODE . Eight options will be displayed.

Press the big blue key down  .
This will take you to another screen with 5 options.

Press 3: STAT The screen will then display



Press 1: ON This will add a frequency column

On a Casio fx-82AU:	On a Sharp EL531
Get the calculator into STATS mode: Press Mode . The display will be:  Press 2 for *STAT*  Press 1 for *1-VAR* This time the calculator will display an X column and a FREQ column. Enter a value followed by =. Navigate to the frequency column using the big blue REPLAY key. 	Each row from the table is entered as: 46 , 2 M+ the calculator display will show DATA SET= 1 47 , 3 M+ the calculator display will show DATA SET= 2 The display DATA SET= 2 informs you that 2 **rows** have been entered into the STATs memory. 48 , 5 M+ the calculator display will show DATA SET= 3 . . . . 55 , 2 M+ the calculator display will show DATA SET= 15 If you make a mistake and need to start again, the STATs memory is cleared by pressing 2ndF CA

Obtaining values for the population and sample standard deviation is exactly the same process as in the previous section.

Do the calculation yourself, the values obtained should be:

Population Standard Deviation =2.140554106 rounded gives 2.14

Sample Standard Deviation = 2.162387192 rounded gives 2.16

If the data is grouped, the process is almost identical except that the group midpoint is used.

Amount Spent$	Group Midpoint *(X)*	Frequency *(f)*
$0 but less than $10	5	2
$10 but less than $20	15	3
$20 but less than $30	25	5
$30 but less than $40	35	4
$40 but less than $50	45	2
$50 but less than $60	55	8
$60 but less than $70	65	14
$70 but less than $80	75	18
$80 but less than $90	85	12
$90 but less than $100	95	2
Total		$\Sigma f = 70$

The data is entered into the calculator as Group Midpoint, Frequency or Group Midpoint; Frequency depending on the calculator used.

Video ‘Measures of Spread’

Activity

1. The lifetime, in hours, of a sample of 15 light bulbs is: 351, 429, 885, 509, 317, 753, 827, 737, 487, 726, 395, 773, 926, 688, 485.

   Calculate the range, inter-quartile range and standard deviation of the values. Do not organise into a table.

2. The number of children in a 10 families is: 1, 5, 2, 2, 2, 3, 1, 4, 3, 2. Calculate the range, inter-quartile range and standard deviation of the values. Do not organise into a table.

3. For the Car Speed Data, calculate the range, inter-quartile range and standard deviation of the values after organising the data in Frequency Distribution Table (This can be found in the answers to the Topic 'Graphs').

Car Speed Data (in km/hr)

12	41	44	45	28	40	32	62	46	25
31	35	31	20	59	27	49	19	58	38
22	50	46	14	33	48	25	32	52	69
40	52	57	27	61	42	39	64	52	27

4. The number of students attending a class (maximum 25) for 30 lessons is given in the table below:

Students Attending Class

25	24	24	25	24	23
25	24	23	24	25	25
24	25	20	23	25	24
22	24	25	24	23	21
25	23	24	25	24	22

   Calculate the range, inter-quartile range and standard deviation using a Frequency Distribution Table. (This can be found in the answers to the Topic 'Graphs').

5. The Shot Put distances thrown by 27 world champion shot putters are given in the table below. The unit is metres (m).

22.25	20.19	21.39	21.25	21.19	22.07
21.72	20.37	20.45	23.09	21.19	21.22
21.07	21.55	23.12	20.91	22.58	21.97
20.22	20.38	22.37	22.19	21.70	20.54
22.67	21.58	21.72			

   Calculate the range, inter-quartile range and standard deviation using a grouped Frequency Distribution Table. (This can be found in the answers to the Topic 'Graphs')

Topic 5: Comparing like data

The results of two classes of maths students are compared. This is an example where like data (exam results) exists in two different settings (classes). It is possible to compare these. To do this we will use Box and Whisker diagrams.

Before continuing, it is important to make a point about the limitations of this process. When samples of data are collected there is always random variation between samples. For example; an environmental scientist collected tadpoles from two different ponds each day for a week. The number of tadpoles in the first pond was lower than the second pond. It was suspected that the first pond receives water runoff from an industrial area which is polluted. The lower number of tadpoles in the first pond could be explained by one of two explanations: (a) chance variation or (b) the first pond was polluted. Perhaps if the difference is small, the variation is chance variation. If the difference is large, the variation is due to pollution in the first pond. What is a small or large variation?

In this module it is not possible to separate chance variation from variation due to some effect. In most first year university statistics courses the topic 'Hypothesis Testing' is covered. Hypothesis testing is a process that gives some certainty to situation outlined above.

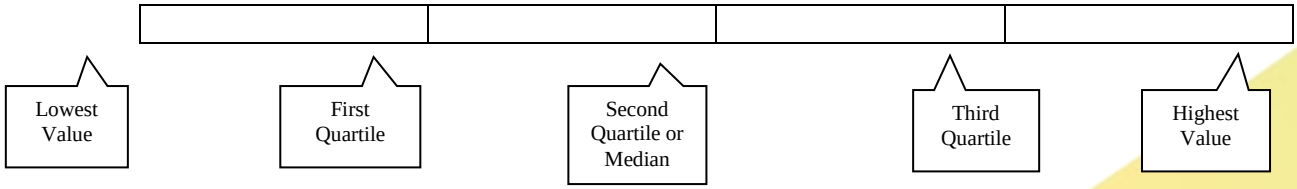
Box and Whisker Diagrams

In this section some new data is required. Two brands of batteries were tested on a common battery operated toy. A sample of twenty batteries of each brand was used. Each battery was used to operate the toy until it became non-operational. The time taken for this to occur (to the nearest hour) was recorded.

The data collect is listed in the table below:

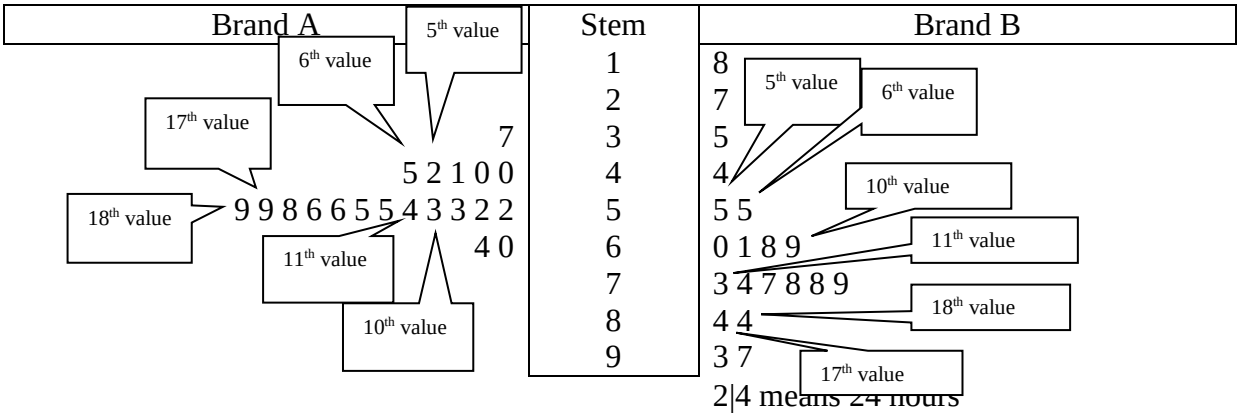
Brand A										Brand B									
53	40	37	41	40	56	56	60	50	52	27	18	60	97	35	79	55	73	44	68
58	55	53	52	54	59	59	64	45	42	77	84	93	84	61	78	69	74	55	78

For each brand a 5 number summary is determined. This was mentioned in the section on 'Interquartile Range'. Diagrammatically, it is represented as below:



A five number summary is necessary for drawing a box and whisker diagram.

One way to organise this data is to use a back to back stem and leaf plot.
This method of organising the data works well here because this data is easily put into a stem and leaf plot (see section 2 Tables) and some trends about the data may be seen from the plot.



The 5 number summary for the Brand A was:

Lowest Value = 37	First Quartile is the $\frac{n+1}{4}$ $=\frac{20+1}{4}$ $=\frac{21}{4}$ $=5.25^{th}$ value. This means that the first quartile is a quarter of the way between the 5th and 6th values. First Quartile = 42 + 0.25 x (45-42) = 42.75	Median is the $\frac{n+1}{2}$ $=\frac{20+1}{2}$ $=\frac{21}{2}$ $=10.5^{th}$ value Median = (53+54)÷2=53.5	Third Quartile is the $\frac{3(n+1)}{4}$ $=\frac{3\times 21}{4}$ $=17.75^{th}$ value This means that the third quartile is three quarters of the way between the 17th and 18th values. Third Quartile = 59+0.75x(59-59) =59	Highest Value = 64
Lowest Value = 37	First Quartile *(Q_1)* = 42.75	Median = 53.5	Third Quartile *(Q_3)* = 59	Highest Value = 64

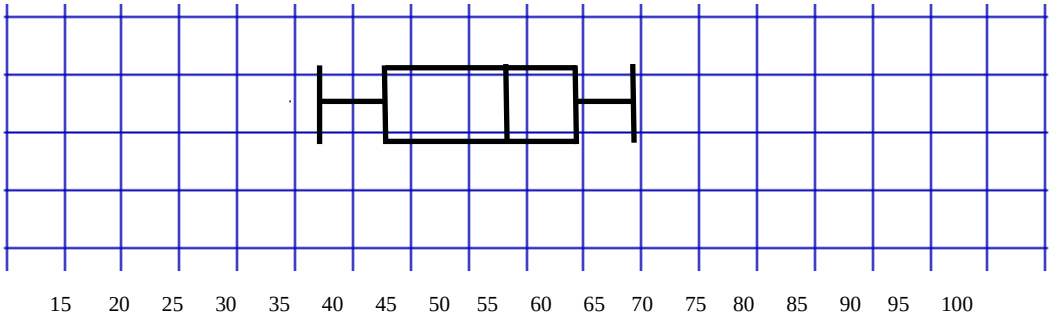
To draw a box and whisker plot:

1) Draw short vertical lines at the 5 values from the 5 number summary.

2) Draw a box from the first quartile to the third quartile.

3) Connect the first quartile to the lowest value with a whisker. Likewise the third quartile to the highest value. These should be located in the middle between the top and bottom of the box.

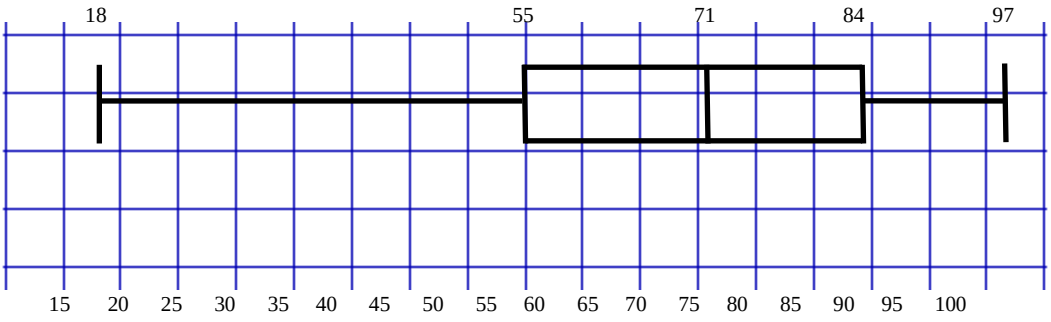
37 42.75 53.5 59 64



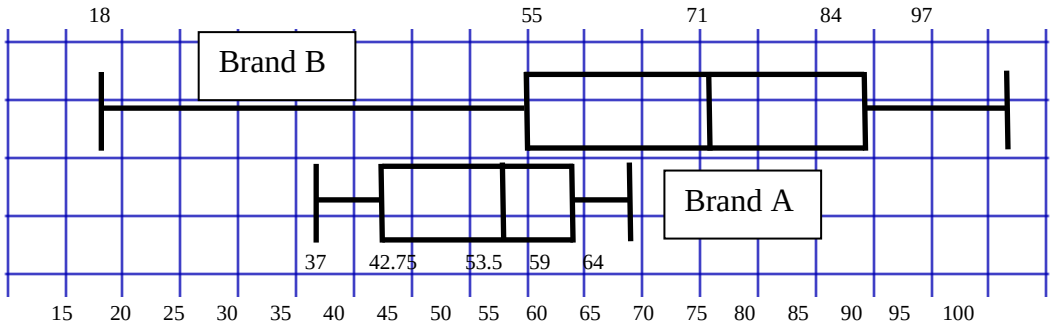
The 5 number summary for the Brand B was:

Lowest Value = 18	First Quartile is the $\frac{n+1}{4}$ $=\frac{20+1}{4}$ $=\frac{21}{4}$ $=5.25^{th}$ value This means that the first quartile is a quarter of the way between the 5th and 6th values. First Quartile = 55+ 0.25 x (55-55) = 55	Median is the $\frac{n+1}{2}$ $=\frac{20+1}{2}$ $=\frac{21}{2}$ $=10.5^{th}$ value Median = (69+73)÷2=71	Third Quartile is the $\frac{3(n+1)}{4}$ $=\frac{3\times 21}{4}$ $=17.75^{th}$ value This means that the third quartile is three quarters of the way between the 17th and 18th values. Third Quartile = 84+0.75x(84-84) =84	Highest Value = 97
Lowest Value = 18	First Quartile *(Q_1)* = 55	Median = 71	Third Quartile *(Q_3)* = 84	Highest Value = 97

Drawing the box and whisker plot:



Now to do the comparing the Box and Whisker Plots should be arranged side by side.



Comparing:

1. Based on the lengths of the overall plots and the lengths of the boxes, Brand B batteries are more variable in lifetime than Brand A.

2. Based on the first quartiles, median and third quartiles, it is possible to state that Brand B will mostly last longer than Brand A. The lowest value for Brand B is lower than the lowest for Brand A. This is not a major consideration as lowest or highest values may be 'odd' values. Because the lower quartile, median and upper quartiles are free of 'odd' values, they are the most reliable values to base generalisations on.

Video 'Comparing Data'

Activity

1. The lifetime, in hours, of a sample of 15 light bulbs is: 351, 429, 885, 509, 317, 753, 827, 737, 487, 726, 395, 773, 926, 688, 485.

   Construct a box and whisker plot for this data.

2. The Car Speed Data was collected from a school zone just prior to an advertising campaign. The data collected was:

Car Speed Data (in km/hr)

12	41	44	45	28	40	32	62	46	25
31	35	31	20	59	27	49	19	58	38
22	50	46	14	33	48	25	32	52	69
40	52	57	27	61	42	39	64	52	27

After an advertising campaign to make motorists more aware of speeding in school zones, the following data was collected.

22	45	42	27	27	40	45	26	75	25
31	48	30	16	28	42	19	55	46	38
34	50	14	59	33	42	30	32	36	18
38	41	50	14	31	39	25	46	39	27

Draw side by side box and whisker plot to determine if the campaign was successful.

3. The heart rate of a group of athletes was compared to the heart rate of a group of office workers after climbing a set of stairs. The data is presented below:

Athletes

96	111	88	79	101	91
104	106	121	93	103	96
85	117	126	97	83	93
112	106	110	116	91	88

Office Workers

114	107	94	113	113	97
118	103	121	127	117	131
145	132	118	108	145	126
100	138	120			

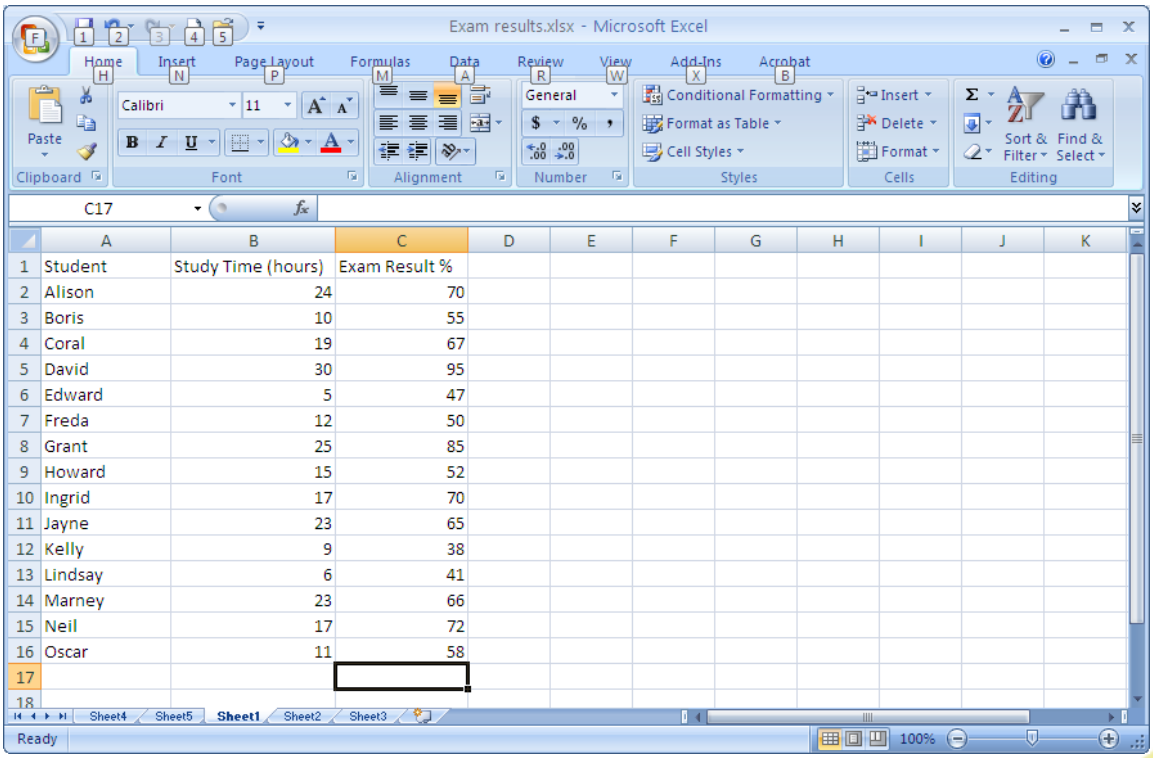
Draw side by side box and whisker plots and make generalisations about the two groups.

Topic 6: Correlation and Regression

In this section, the relationship between two variables is being considered.
Correlation is an attempt to measure the strength of the relationship between the two variables. For example: a retail store may vary the amount they spend on advertising and then measure the revenue obtained. There seems to be a natural link between the amount spent on advertising and the revenue obtained from sales. Calculating the correlation coefficient will give information about the strength and nature of the linear relationship.

Once the correlation coefficient suggests a reasonable linear relationship exists, the linear relationship can be described by an equation. The process of finding the "theoretical" or "ideal" linear equation is called Regression.

The scenario used as an example is relevant to university study. Many studies have shown that time on task is an important indicator to the overall success of students at university. In this example the success of students on an examination (as a percentage) will be one variable and the hours spent studying will be the other variable. The data will be entered directly into Excel. The spreadsheet is shown below:



	A	B	C	D	E	F	G	H	I	J	K
1	Student	Study Time (hours)	Exam Result %								
2	Alison	24	70								
3	Boris	10	55								
4	Coral	19	67								
5	David	30	95								
6	Edward	5	47								
7	Freda	12	50								
8	Grant	25	85								
9	Howard	15	52								
10	Ingrid	17	70								
11	Jayne	23	65								
12	Kelly	9	38								
13	Lindsay	6	41								
14	Marney	23	66								
15	Neil	17	72								
16	Oscar	11	58								
17											
18											

The next step is to draw the graph of the data. For the scenario, the dependent variable is Exam Result% because the underlying relationship could be that Exam Result% depends upon the Study Time. Study Time is the independent variable. When two variables are graphed together, the graph is called a scatter plot or scatter diagram.
To graph the data from above

>Highlight the data in the table

>Click on the Insert tab

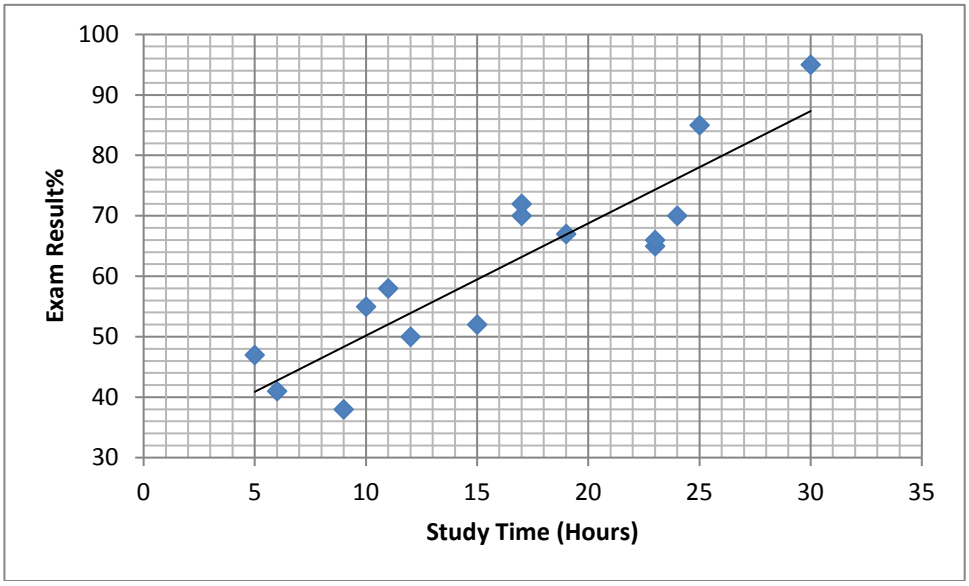
> In the Charts Area click on Scatter

> Choose the option that plots points (not lines)

The **correlation** can now be considered.

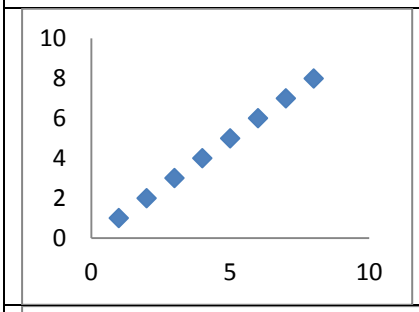
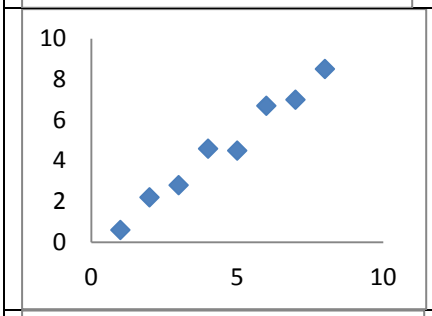
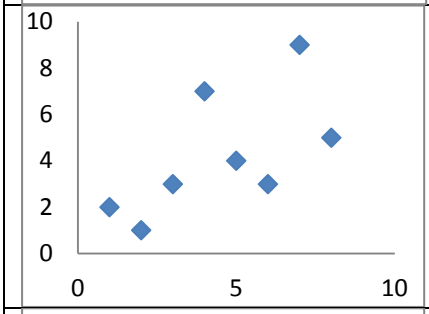
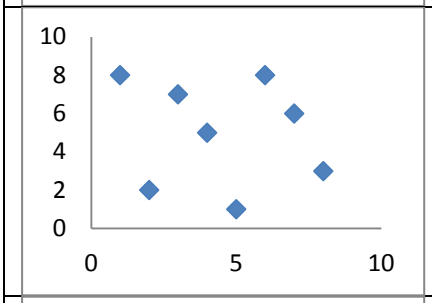
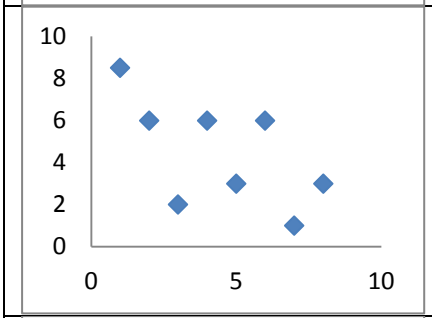
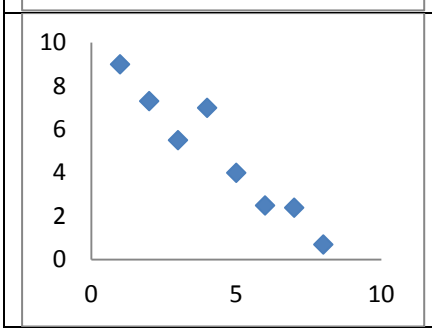
Strong correlation is indicated by the points making a straight line. The straight line may have a positive slope, indicating positive correlation or a negative slope, indicating negative correlation. If the points form a line with slope close to zero or seem to be randomly arranged, then there is no correlation.
For positive correlation, it is possible to say 'when the value of one variable increases, the value of the other variable will also increase'.

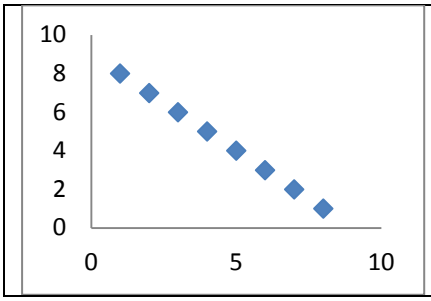
For negative correlation, it is possible to say 'when the value of one variable increases, the value of the other variable will decrease'.



The appearance of the graph suggests positive correlation. Generally speaking this means that as the Study Time increases there is an increase in the Exam Result.

Correlation has a number associated with it called the correlation coefficient. The correlation coefficient ranges from -1 to 1. A correlation coefficient of -1 represents strong negative correlation, 0 represents no correlation, 1 represents strong positive correlation. The following table shows the link between the appearance of graphs, the correlation, possible correlation values and the link between the two variables.

Graph	Correlation description	Correlation coefficient	Link between variables
	Strong Positive	0.8 to 1	The greater the value of the x variable, the greater the value of the y variable.
	Quite Strong Positive	0.6 to 0.8	There is some evidence to suggest that the greater the value of the x variable, the greater the value of the y variable.
	Weak Positive	0.4 to 0.6	There is little evidence to suggest that the greater the value of the x variable, the greater the value of the y variable.
	No relationship	-0.4 to 0.4	There is no evidence to support a linear relationship.
	Weak Negative	-0.6 to -0.4	There is little evidence to suggest that the greater the value of the x variable, the lower the value of the y variable.
	Quite Strong Negative	-0.8 to -0.6	There is some evidence to suggest that the greater the value of the x variable, the lower the value of the y variable.

	Strong Negative	-1 to -0.8	The greater the value of the x variable, the lower the value of the y variable.

Looking at the table above and comparing to the graph for our scenario, it is clear that there is some correlation. The wording 'Quite Strong Correlation' is almost suitable, so a term to suggest slightly lower correlation could be 'Reasonable Correlation'. The correlation coefficient could be estimated to be about 0.7; overall the correlation could be described as Reasonable Positive Correlation.

Excel can calculate the correlation coefficient using a formula.

The data must be present in a table. Excel can give a graph for the visual assessment of the correlation and the CORREL function will give the correlation coefficient.

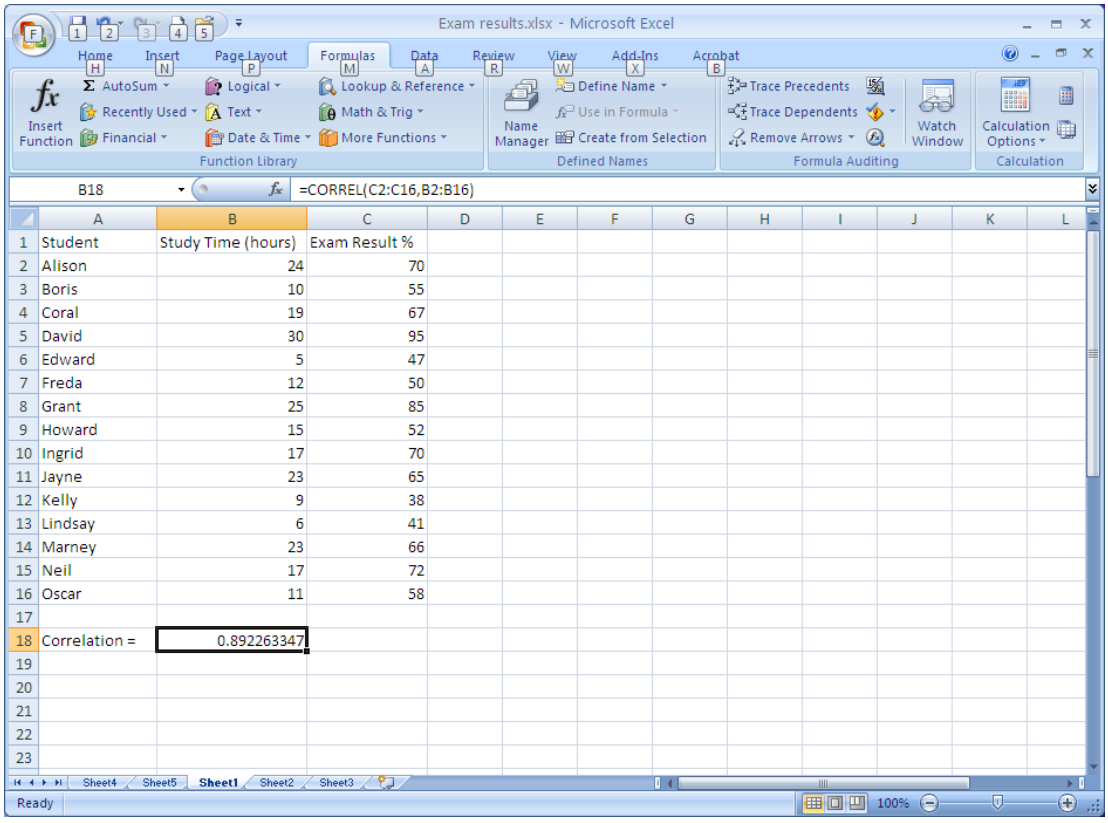
In the cell (say) A18 write the word 'correlation='

In the cell B18, click on the tab Formulas

> click on More Functions
> click on Statistical
Come down the list to CORREL and click.

Enter array 1 by going to the sheet and highlighting the Exam Result values with no heading.

Enter array 2 by going to the sheet and highlighting the Study Time value with no heading. The spreadsheet should look like the spreadsheet below.



	A	B	C
1	Student	Study Time (hours)	Exam Result %
2	Alison	24	70
3	Boris	10	55
4	Coral	19	67
5	David	30	95
6	Edward	5	47
7	Freda	12	50
8	Grant	25	85
9	Howard	15	52
10	Ingrid	17	70
11	Jayne	23	65
12	Kelly	9	38
13	Lindsay	6	41
14	Marney	23	66
15	Neil	17	72
16	Oscar	11	58
17			
18	Correlation =	0.892263347	
19			
20			
21			
22			
23			

The correlation coefficient was 0.892; there is quite strong positive correlation meaning that there is evidence to suggest that the greater the amount of Study Time, the higher the amount of Exam Result will be.
The correlation coefficient confirms the relationship between the variables is stronger than the graph suggests.

As the correlation is strong it is worthwhile determining the **regression equation**.

In the graph above, Excel has plotted the regression line. The spreadsheet can be modified to include the slope and y-intercept of the regression line. From this, the equation can be obtained. (You may need to brush up on Equations of Straight Lines from the Linear Relationships module.)

In cell B20, write the words “Slope =” and in cell B21 write the words “Y-intercept =” The slope is found by doing:
Click on the cell C20, click on the Formulas tab
> click on More Functions
> click on Statistical
Come down the list to SLOPE and click.

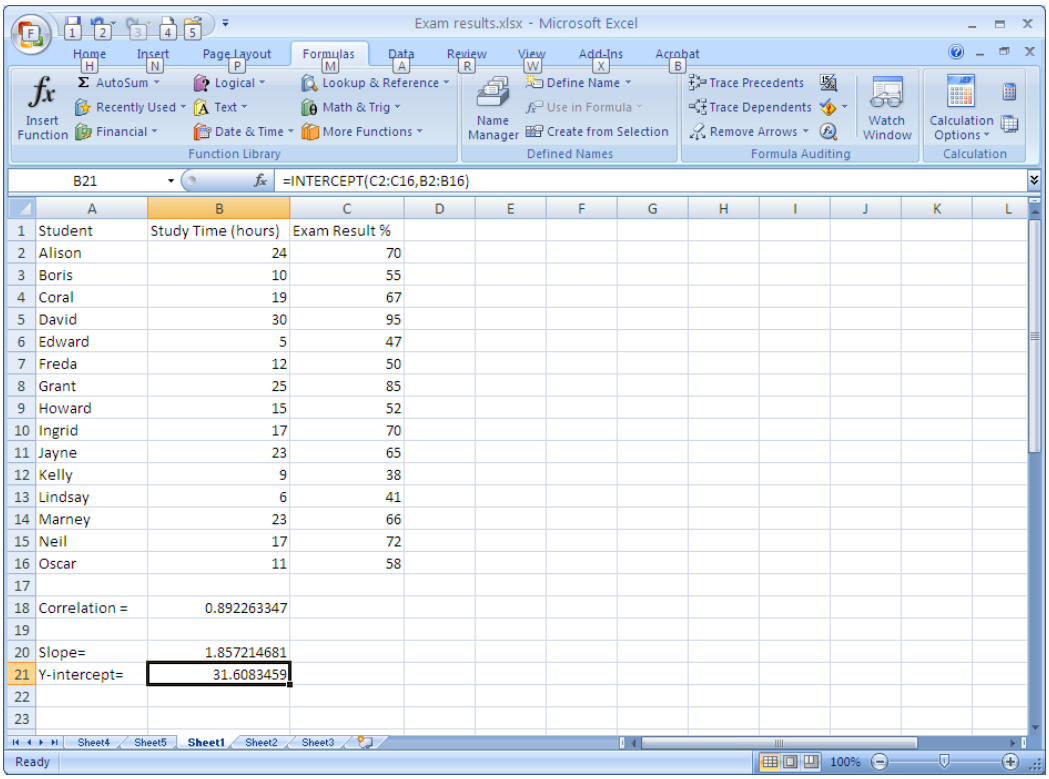
Enter y values (Exam Result%) by going to the sheet and highlighting the exam result values with no heading.

Enter x values (Study Time) by going to the sheet and highlighting the study time values with no heading.

The y-intercept is found by doing:
Click on the cell C21, click on the Formulas tab
> click on More Functions
> click on Statistical
Come down the list to INTERCEPT and click.

Enter y values (Exam Result%) by going to the sheet and highlighting the exam result values with no heading.

Enter x values (Study Time) by going to the sheet and highlighting the study time values with no heading.



	A	B	C
1	Student	Study Time (hours)	Exam Result %
2	Alison	24	70
3	Boris	10	55
4	Coral	19	67
5	David	30	95
6	Edward	5	47
7	Freda	12	50
8	Grant	25	85
9	Howard	15	52
10	Ingrid	17	70
11	Jayne	23	65
12	Kelly	9	38
13	Lindsay	6	41
14	Marney	23	66
15	Neil	17	72
16	Oscar	11	58
17			
18	Correlation =	0.892263347	
19			
20	Slope=	1.857214681	
21	Y-intercept=	31.6083459	
22			
23			

The equation of the line is:

$$R = 1.86T + 31.6$$

where R is the exam result % and T is study time in hours

Note1: In Business text books, the equation may be written as $R = 31.6 + 1.86T$

Note2: In some university units, students are required to correlation coefficient and regression equation is calculated using formulas. Scientific calculators and computer spreadsheets use these formulas in these calculations.

Video ‘Correlation & Regression’

The purpose of describing the relationship between the variables as an equation is to predict values.

For example: If a student studies for 24 hours, what exam result would be expected?

$$R = 1.86T + 31.6$$
$$R = 1.86 \times 24 + 31.6$$
$$R = 76.24$$

The prediction for the exam result is about 76%.

Because of the least-squares method used to calculate regression lines, the value of the y variable (exam result) can be calculated from the value of the x variable (study time) but not the other way round. This means that calculating the hours of study required to get a result of (say) 90% should not be attempted. For

this to take place, a new regression equation would have to be recalculated giving study time as the dependent (y) variable and exam result as the independent (x) variable.

In summary,

(a) The strength of the relationship between two variables is called correlation.

(b) If the graph indicates strong positive or negative correlation supported by the correlation coefficient then determining the regression is worthwhile.

(c) The regression equation can be used to predict the dependent variable from a given independent variable.

(d) Prediction should only be within the range of independent values that were used to calculate the regression equation. In this example the regression equation was based on independent variable values from 5 to 30 hours. Predicting within the range of independent values is called 'Interpolation'.

(e) Prediction outside the range of independent values is called 'Extrapolation'. Dependent values found by extrapolation should be considered unreliable and this practice should be avoided.

(f) The regression equation only applies to this exam with a cohort of student similar to those used to form the equation.

Activity

1. The table below contains information about different types of milks. The information given is the grams of fat and the energy of the food in kilojoules.

Milk (1 cup)	Amount of Fat (g)	Energy (kJ)
Skim	0	336
1% Fat	2.4g	420
2% Fat	4.7g	504
Whole	8g	630
Breast milk	10.7g	735
Goat milk	10g	706
Sheep milk	17g	1113

   (a) Enter the data into Excel and produce the scatterplot. The hypothesis is 'the Energy in the milk depends on the amount of Fat'. Energy is the dependent variable. Comment on the correlation by viewing your scatterplot.

   (b) Modify your spreadsheet to calculate the correlation coefficient. Comment on this.

   (c) Modify your spreadsheet to calculate the slope and y-intercept of the regression line.

2. A large company is making note of the amount spent on advertising each month and then comparing this with the amount of monthly sales. The data is given in the table below:

Month	Advertising ($ x1000)	Sales ($ x1000)
Jan	2	77
Feb	2.4	79
Mar	6	105
Apr	3.3	110
May	1.5	85
Jun	3	95
Jul	3	110
Aug	3.6	124
Sept	4	136
Oct	2.7	118
Nov	5	127
Dec	6.5	154

   (a) Enter the data into Excel and produce the scatterplot. Comment on the correlation by viewing your scatterplot.

   (b) Modify your spreadsheet to calculate the correlation coefficient. Comment on this.

   (c) Modify your spreadsheet to calculate the slope and y-intercept of the regression line.

3. Information about climate is given below for towns on roughly the same latitude but varying longitudes. The longitude is not given but distance inland (east) of the coastal location of Yeppoon is

given. The purpose of the question is to determine if there is a correlation between the climate of a location and its distance from the sea. (Distance Inland is the independent variable)

Location	Distance inland (km)	January Av. Temp Max°C	July Av. Temp Min °C	Annual Rainfall
Yeppoon	0	29.3	11.8	885
Rockhampton	26	31.9	9.5	796
Walterhall	36	31.4	8	815
Blackwater	193	34.1	6	542
Emerald	266	34.2	6.9	640
Springsure	276	34	6.2	682
Blackall	549	36	6.9	529
Barcaldine	568	35	7.9	500
Longreach	674	37.3	6.8	434
Bedourie	1131	38.4	7.7	262

Data from http://www.bom.gov.au/

Using Excel, find the correlation coefficient of the Distance Inland vs the three climate data given. Calculate the regression line as appropriate and make comments about what other information could be applicable in this situation

Answers to activity questions

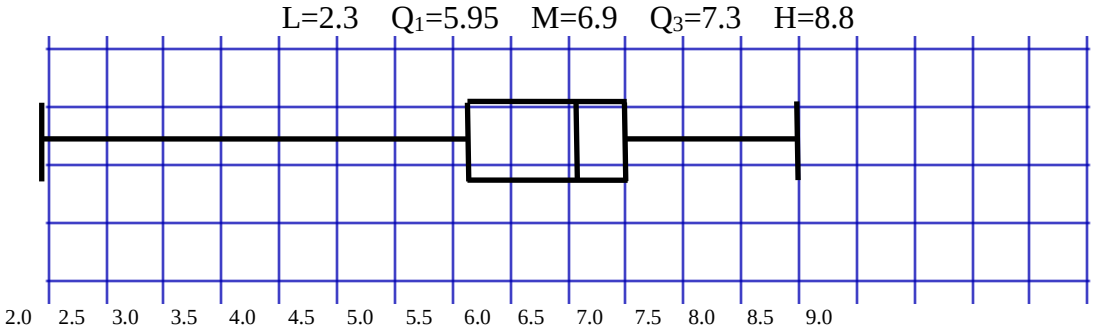
Check your skills

1. For the data (right) about the Average Daily Hours of Sunshine, calculate the mean, mode and median.

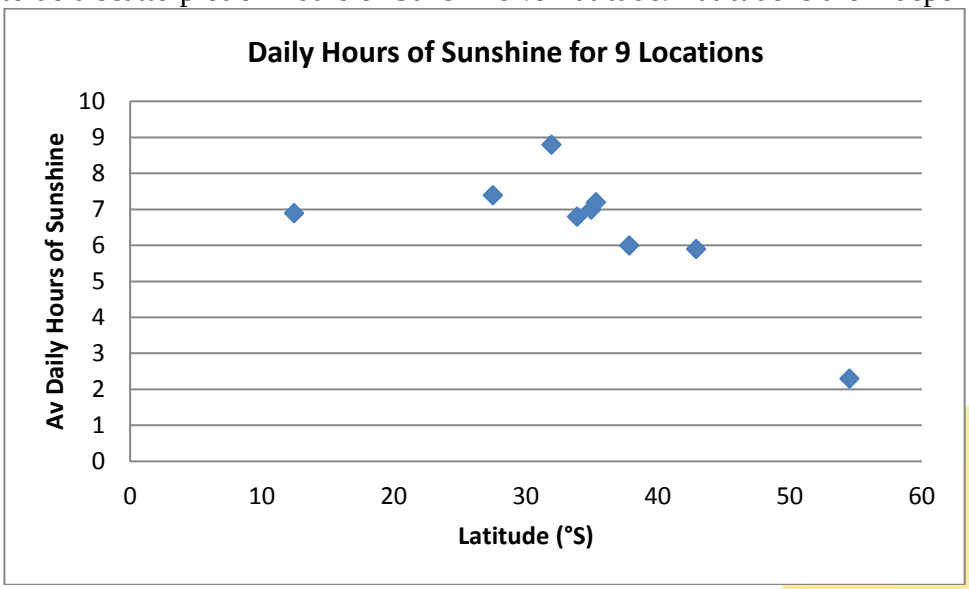
2. For the same data, calculate the range, interquartile range and standard deviation.

Mean	Mode	Median
$\bar{x} = \frac{\Sigma x}{n} = \frac{58.3}{9} = 6.48$	There is no mode	Is the 5th value. Median = 6.9
Range	IQR	SD
8.8 - 2.3 = 6.5	Q_1=5.95 Q_3=7.3 IQR=1.35	1.78 sample assumed

3. For the Latitude Data, use a 5 number summary to draw a box and whisker plot.



4. Use Excel to do a scatterplot of Hours of Sunshine vs Latitude. Latitude is the independent variable.



5. Use Excel to determine the correlation and the regression equation.

Correlation = - 0.69 Moderate negative correlation
Regression line:

$$H = -0.11L + 10.2$$

where L is the latitude

H is the av. daily hours of sunshine

Organising Data – Tables

1. (a) The speed (in km/hr) of 40 cars passing a school at 9am on a school day.

Speed	Tally	Frequency	Cumulative Frequency
10 to but less than 20 km/hr	III	3	3
20 to but less than 30 km/hr	𝍸 III	8	11
30 to but less than 40 km/hr	𝍸 III	8	19
40 to but less than 50 km/hr	𝍸 𝍸	10	29
50 to but less than 60 km/hr	𝍸 II	7	36
60 to but less than 70 km/hr	IIII	4	40
		40	

(b) Enter the data into Stem and Leaf Plot.

Stem	Leaf
1	2 4 9
2	0 2 5 5 7 7 7 8
3	1 1 2 2 3 5 8 9
4	0 0 1 2 4 5 6 6 8 9
5	0 0 2 2 2 7 8
6	1 2 4 9
	4|5 means 45 km/hr

(c) What percentage of cars were doing 40km/hr or more?

Using the Cumulative Frequency Column, we know that 19 motorists were doing less than 40 km/hr. Therefore the number doing 40km/hr or more is 40 (Total) – 19 = 21motorists.

The number doing 40km/hr or more = 21

The fraction doing 40km/hr or more = $^{21}/_{40}$

Make this a percentage: $^{21}/_{40}$ x 100 = 52.3 or approximately 53%

2. The number of students attending a class (maximum 25) for 30 lessons is given in the table below:

(a) Is the data discrete or continuous? The data is discrete because the number of student can only be a whole number.

(b) Enter the data into a frequency distribution table (groups not required). Include a relative frequency and % relative frequency column.

Number attending	Tally	Frequency	Relative Frequency	% Relative Frequency
20	I	1	$\frac{1}{30} = 0.033\overline{3}$	3.3%
21	I	1	$\frac{1}{30} = 0.033\overline{3}$	3.3%
22	II	2	$\frac{2}{30} = 0.066\overline{6}$	6.7%
23	$\cancel{\text{IIII}}$	5	$\frac{5}{30} = 0.16\overline{6}$	16.7%
24	$\cancel{\text{IIII}}$ $\cancel{\text{IIII}}$ I	11	$\frac{11}{30} = 0.36\overline{6}$	36.7%
25	$\cancel{\text{IIII}}$ $\cancel{\text{IIII}}$	10	$\frac{10}{30} = 0.333\overline{3}$	33.3%
	Total	30	1.00	100%

(c) What proportion of lessons contained 22 students? The relative frequency for 22 students is 0.067 (to 3 decimal places)

(d) What percentage of lessons were fully attended? From the table - 33.3%

3. Systolic Blood Pressures of 35 patients at a Cardiac Clinic

(a) Enter the data into a Stem and Leaf Plot. Use the key: 11|4 means 114.

Stem	Leaf
8	8
9	2 5
10	2 7 9
11	0 2 4 5 8
12	0 1 2 2 4 6 7 8
13	1 3 4 4 7 8
14	1 3 4 5 6 9
15	5
16	1
17	5
18	8
	11|4 means 114

(b) If hypertension (high blood pressure) is defined by a systolic blood pressure 140 or above, what percentage of this group are suffering hypertension? With a Stem and Leaf, the original values are retained, so counting is required. There are 10 out of 35 which is 28.5% (to 1 d.p.)

4. The Shot Put distances thrown by 27 world champion shot putters are given in the table below. The unit is metres (m).

22.25	20.19	21.39	21.25	21.19	22.07
21.72	20.37	20.45	23.09	21.19	21.22
21.07	21.55	23.12	20.91	22.58	21.97
20.22	20.38	22.37	22.19	21.70	20.54
22.67	21.58	21.72			

(a) Enter the data into a Frequency Distribution Table. Your FDT must have at least 5 groups.

Distance	Tally	Frequency	Cumulative Frequency	% Relative Frequency
20 to but less than 20.5	$\cancel{\text{IIII}}$	5	5	$\frac{5}{27}\times 100 = 18.5\%$
20.5 to but less than 21	II	2	7	$\frac{2}{27}\times 100 = 7.4\%$
21 to but less than 21.5	$\cancel{\text{IIII}}$ I	6	13	$\frac{6}{27}\times 100 = 22.2\%$
21.5 to but less than 22	$\cancel{\text{IIII}}$ I	6	19	$\frac{6}{27}\times 100 = 22.2\%$
22 to but less than 22.5	IIII	4	23	$\frac{4}{27}\times 100 = 14.8\%$
22.5 to but less than 23	II	2	25	$\frac{2}{27}\times 100 = 7.4\%$

23 to but less than 23.5	II	2	27	$\frac{2}{27}\times 100 = 7.4\%$
	Total	27		100%

(b) How many have thrown less than 22m? (Use cumulative frequency to answer this) The cumulative frequency of the group before 19, this is the sum of the frequencies less than 22.

(c) What percentage threw 21 to but less than 22m? (Use a % Relative Frequency Column) This is two groups of the table. Each group has 22.2%, so 44.4% threw 21 to but less than 22m.

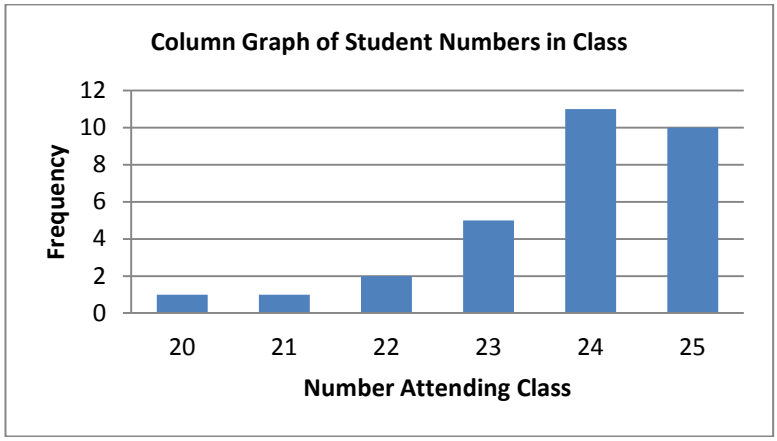
Graphs

1. The height of a plant is measured every Friday morning for twelve weeks. Which type of graph would be best to show the growth of the plant? Line

2. A 'sound and vision' shop sells CDs, DVDs, console games and computer games. Which type of graph would best display the relative sales of the different items sold? Pie

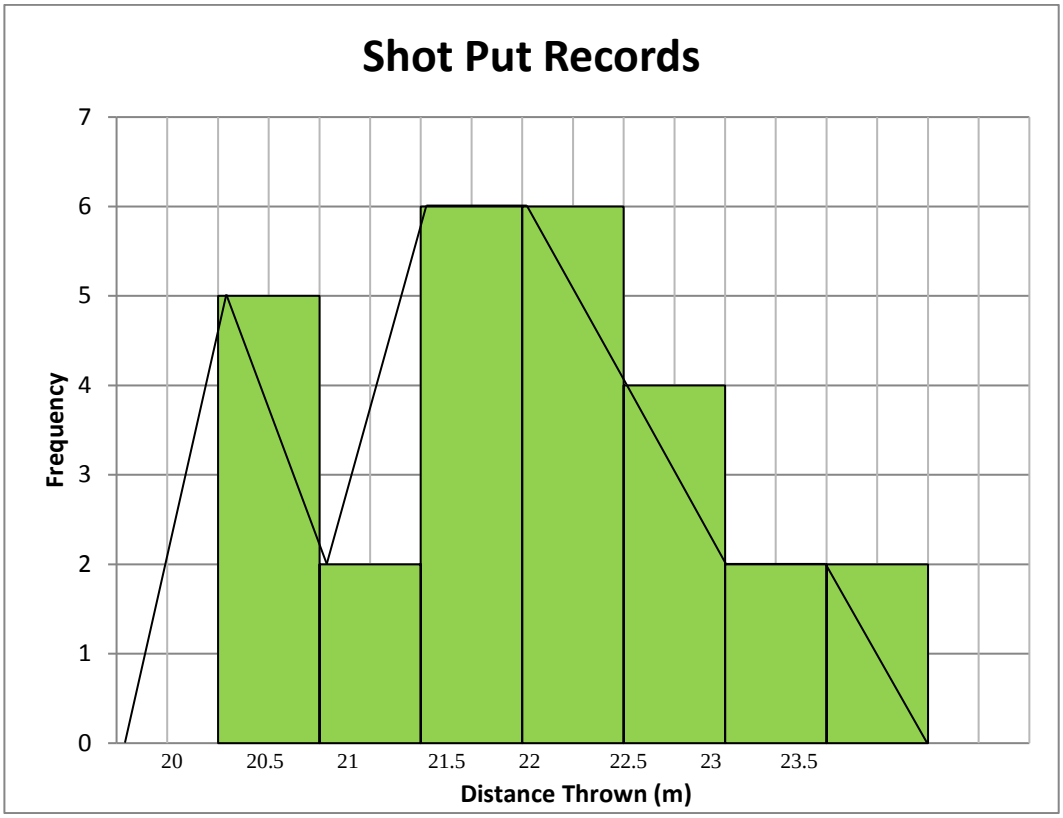
3. A student wishes to compare the number of motorcycle fatalities between the states. To get a more accurate picture, the data is collected for the years 2007, 2008 and 2009. Which type of graph allows all this data to be presented in one graph? Composite Column

4. The number of students attending a class (maximum 25) for 30 lessons is given in the table below:

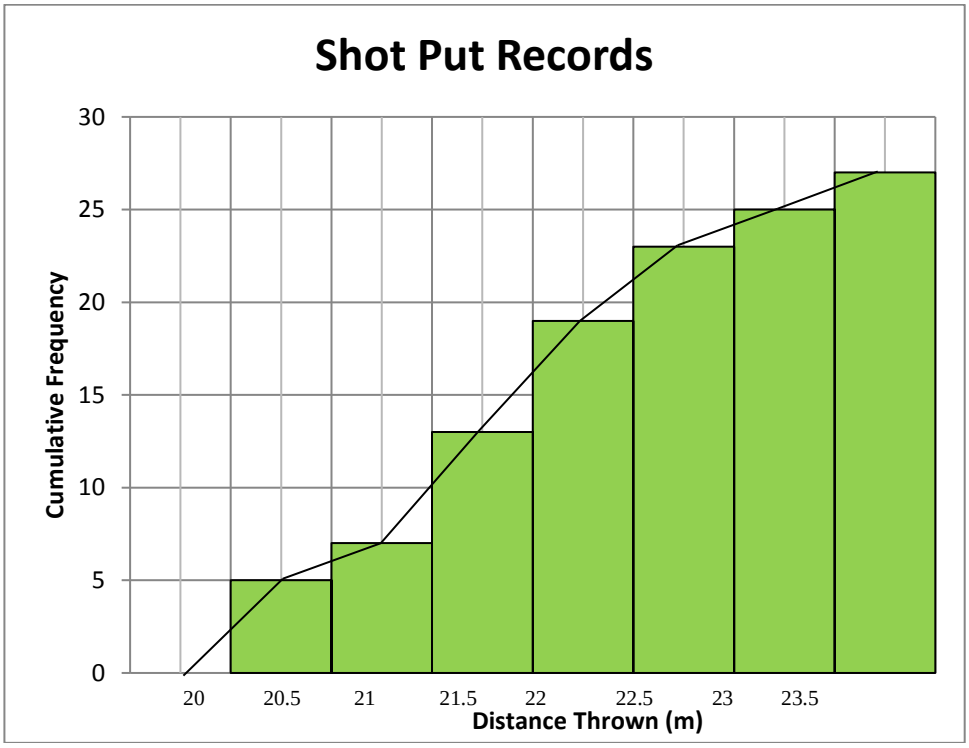


5. The Shot Put distances thrown by 27 world champion shot putters are given in the Frequency Distribution Table below. The unit is metres (m).

(a) Construct a frequency histogram for this information. As a second step, put a frequency polygon on the histogram.



(b) Construct a cumulative frequency histogram and then add an Ogive to the histogram.



Measures of Central Tendency

1. The lifetime, in hours, of a sample of 15 light bulbs is: 351, 429, 885, 509, 317, 753, 827, 737, 487, 726, 395, 773, 926, 688, 485. Calculate the mean, mode and median of the values. Do not organise into a table.

In order to calculate the median, the values must be ordered.
317, 351, 395, 429, 485, 487, 509, 688, 726, 737, 753, 773, 827, 885, 926

Mean	Mode	Median
$\overline{X} = \frac{\sum X}{n}$ $\overline{X} = \frac{317 + 351 + ... + 885 + 926}{15}$ $\overline{X} = 619.2$ The mean is approx. 619 hours. The mean is a good indicator of a central or typical value in this question.	No value is repeated – there is no mode. The mode would not significant in this example.	As there are 15 values, the position of the median is: $\frac{n+1}{2} = \frac{15+1}{2} = \frac{16}{2} = 8$ The 8th value is the median. Counting through the ordered list, the 8th value is 688. The median is also a good indicator or typical value in this question.

2. The number of children in a 10 families is: 1, 5, 2, 2, 2, 3, 1, 4, 3, 2. Calculate the mean, mode and median of the values. Do not organise into a table.

In order to calculate the median, the values must be ordered.
1, 1, 2, 2, 2, 2, 3, 3, 4, 5

Mean	Mode	Median
$\overline{X} = \frac{\sum X}{n}$ $\overline{X} = \frac{1 + 1 + ... + 4 + 5}{10}$ $\overline{X} = 2.5$ The mean is 2.5 children. The mean is a good indicator of a central or typical value in this question. However it is not possible to have 2.5 children	The mode is 2, it occurs 4 times, or has a frequency of 4. The mode is a good indicator of a typical value in this example.	As there are 15 values, the position of the median is: $\frac{n+1}{2} = \frac{10+1}{2} = \frac{11}{2} = 5.5$ The 5.5th value is the median. The 5th value is 2 and the 6th value is 2 also, so the median is 2. The median is also a good indicator or typical value in this question.

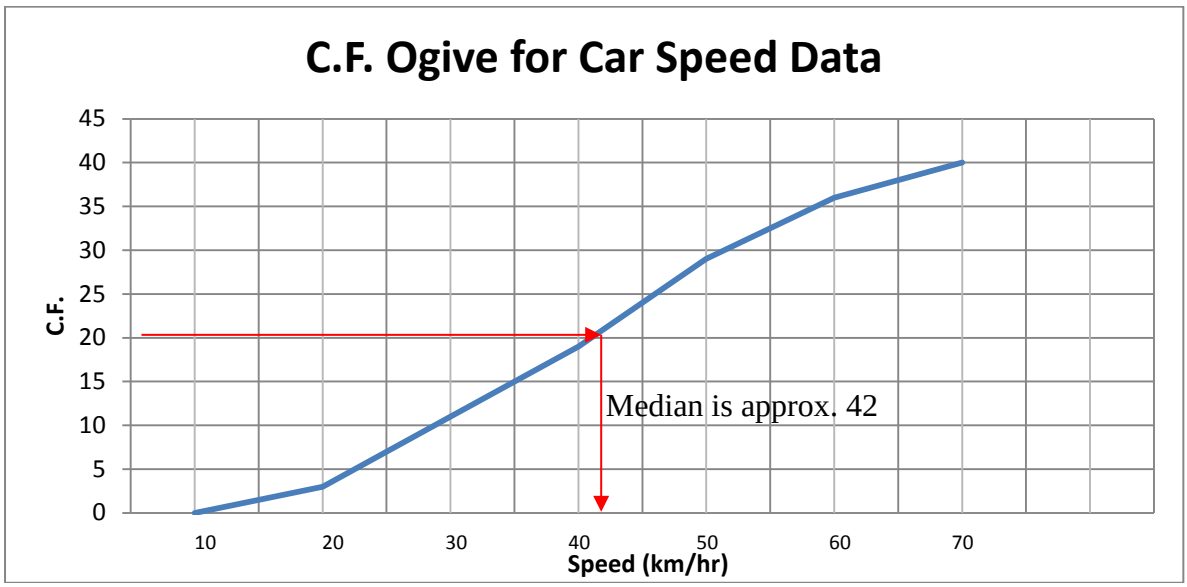
3. For the Car Speed Data, calculate the mean, mode and median of the values after organising the data in Frequency Distribution Table (This can be found in the answers to the Topic 'Graphs').

Car Speed Data (in km/hr)

Speed(x)	Group Midpoint	Frequency(f)	Cumulative Frequency	$f \times x$
10 to but less than 20 km/hr	15	3	3	45
20 to but less than 30 km/hr	25	8	11	200
30 to but less than 40 km/hr	35	8	19	280
40 to but less than 50 km/hr	45	10	29	450
50 to but less than 60 km/hr	55	7	36	385
60 to but less than 70 km/hr	65	4	40	260
		$n = \Sigma f = 40$		$\Sigma fx = 1620$

Mean	Mode	Median
$\overline{X} = \frac{\Sigma fX}{\Sigma f}$ $\overline{X} = \frac{1620}{40}$ $\overline{X} = 40.5$ The mean is approx. 40.5 km/hr. The mean is a good indicator of a central or typical value in this question.	The modal group is 40 to but less than 50 km/hr.	As there are 40 values, the position of the median is: $\frac{n+1}{2} = \frac{40+1}{2} = \frac{41}{2} = 20.5$ The 20.5th value is the median. The median is the average of the 20th and 21st values. The group '40 to but less than 50 km/hr' contains the 20th to the 29th value. $Median = L_m + \frac{\left(\frac{n+1}{2}\right) - cf_{m-1}}{f_m} \times Group\ width$ $Median = 40 + \frac{20.5-19}{10} \times 10$ $Median = 41.5$

There is also a graphical method for working out the median.



4. Students Attending Class

Number attending (x)	Frequency(f)	$f \times x$	Cumulative Frequency
20	1	20	1
21	1	21	2
22	2	44	4
23	5	115	9
24	11	264	20
25	10	250	30
Total	$\Sigma f = 30$	$\Sigma fx = 714$	

(a) Calculate the mean, mode and median.

Mean	Mode	Median
$\overline{X} = \frac{\Sigma fX}{\Sigma f}$ $\overline{X} = \frac{714}{30}$ $\overline{X} = 23.8$ The mean is approx. 23.8 students. The mean is a good indicator of a central or typical value in this question.	The modal group is 24 students.	As there are 30 values, the position of the median is: $\frac{n+1}{2} = \frac{30+1}{2} = \frac{31}{2} = 15.5$ The 15.5th value is the median. The median is the average of the 15th and 16th values. The group 24 contains the 10th to the 20th values, including the 15th and 16th values. The median is 24.

(b) Discuss the appropriateness of each measure of central tendency as a typical value.

All measures of centre are relevant. On average 24 students attend lectures.

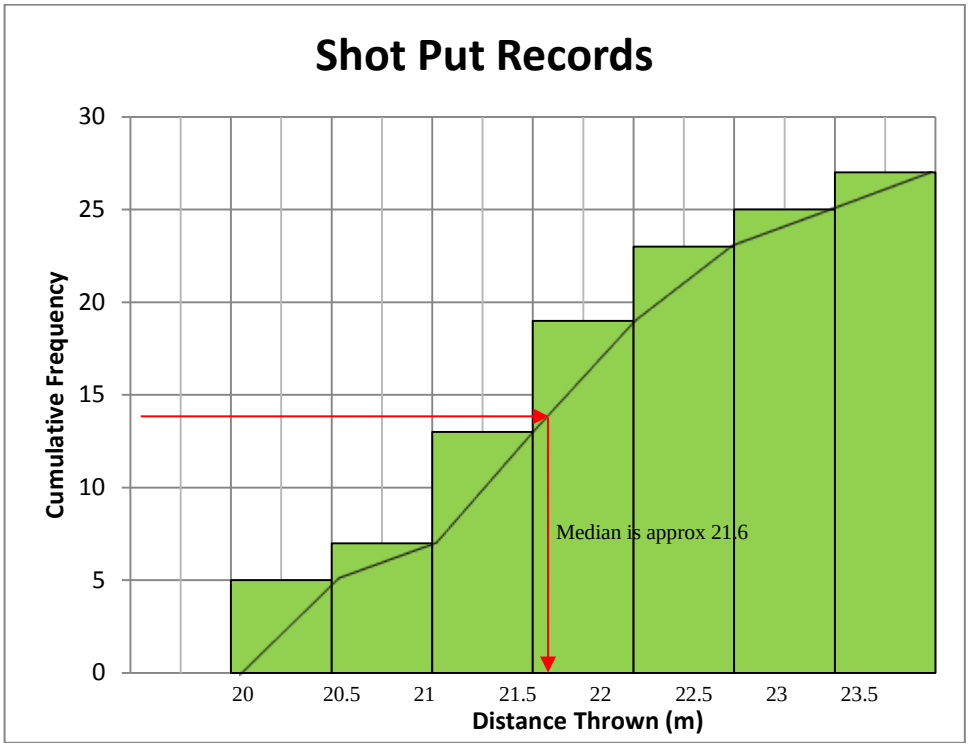
5. The Shot Put distances thrown by 27 world champion shot putters are given in the table below. The unit is metres (m).

Distance (x)	Group Centre	Frequency (f)	Cumulative Frequency	$f \times x$
20 to but less than 20.5	20.25	5	5	101.25
20.5 to but less than 21	20.75	2	7	41.5
21 to but less than 21.5	21.25	6	13	127.5
21.5 to but less than 22	21.75	6	19	130.5
22 to but less than 22.5	22.25	4	23	89
22.5 to but less than 23	22.75	2	25	45.5
23 to but less than 23.5	23.25	2	27	46.5
	Total	$\Sigma f = 27$		$\Sigma fx = 581.75$

Calculate the mean, mode and median.

Mean	Mode	Median
$\overline{X} = \frac{\Sigma fX}{\Sigma f}$ $\overline{X} = \frac{581.75}{27}$ $\overline{X} = 21.5$ The mean is approx. 21.5 m. The mean is a good indicator of a central or typical value in this question.	This data is bimodal. The modal groups are 21 to but less than 21.5 and 21.5 to but less than 22. The mode is of limited use in this example.	As there are 27 values, the position of the median is: $\frac{n+1}{2} = \frac{27+1}{2} = \frac{28}{2} = 14$ The 14th value is the median. The median is located in the group '21.5 to but less than 22 m'. $Median = L_m + \frac{\left(\frac{n+1}{2}\right) - cf_{m-1}}{f_m} \times Group\ width$ $Median = 21.5 + \frac{14-13}{6} \times 0.5$ $Median = 21.6$ *(rounded to 1dp)*

There is also a graphical method for working out the median.



Measures of Spread

1. The lifetime, in hours, of a sample of 15 light bulbs is: 351, 429, 885, 509, 317, 753, 827, 737, 487, 726, 395, 773, 926, 688, 485.

   In order to calculate the median, the values must be ordered.
   317, 351, 395, 429, 485, 487, 509, 688, 726, 737, 753, 773, 827, 885, 926

Range	SD
Range = Highest Value - Lowest Value *Range = 926 - 317* *Range = 609*	Using a Scientific Calculator $s = 203$ *(rounded to nearest whole number)*
IQR	
The first quartile (Q_1) is the $\frac{n+1}{4}$ $= \frac{15+1}{4}$ $= \frac{16}{4}$ $= 4^{th}$ value Q_1 is 429	The third quartile (Q_3) is the $\frac{3(n+1)}{4}$ $= \frac{3\times16}{4}$ $= \frac{48}{4}$ $= 12^{th}$ value Q_3 is 773
The IQR is 773 – 429 = 344	

2. The number of children in a 10 families is: 1, 5, 2, 2, 2, 3, 1, 4, 3, 2. Calculate the range, inter-quartile range and standard deviation of the values. Do not organise into a table.

   In order to calculate the median, the values must be ordered.

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

Range	SD
Range = Highest Value - Lowest Value *Range = 5 - 1* *Range = 4*	Using a Scientific Calculator $s = 1.27$ *(rounded to 2dp)* (assuming sample)
IQR	
The first quartile (Q_1) is the $\frac{n+1}{4}$ $= \frac{10+1}{4}$ $= \frac{11}{4}$ $= 2.75^{th}$ value Q_1 is 1.75	The third quartile (Q_3) is the $\frac{3(n+1)}{4}$ $= \frac{3\times 11}{4}$ $= \frac{33}{4}$ $= 8.25^{th}$ value Q_3 is 3.25
The IQR is 3.25 – 1.75 = 1.5	

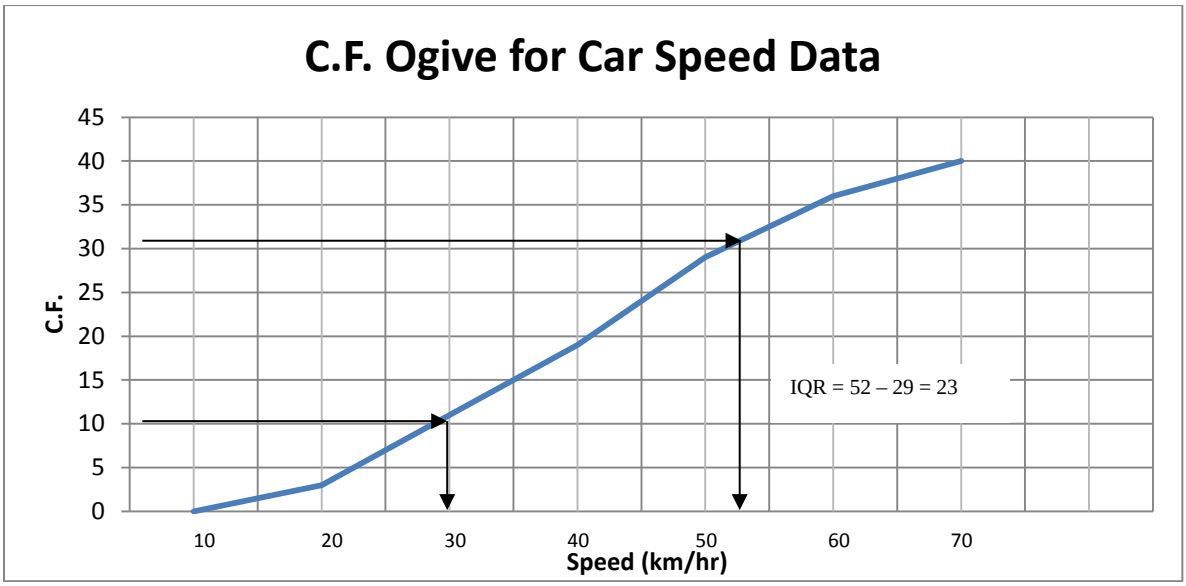
3. For the Car Speed Data, calculate the range, inter-quartile range and standard deviation of the values after organising the data in Frequency Distribution Table. The assumption is that the original values are lost and it is a sample.

Car Speed Data (in km/hr)

Speed(x)	Group Midpoint	Frequency(f)	Cumulative Frequency	$f \times x$
10 to but less than 20 km/hr	15	3	3	45
20 to but less than 30 km/hr	25	8	11	200
30 to but less than 40 km/hr	35	8	19	280
40 to but less than 50 km/hr	45	10	29	450
50 to but less than 60 km/hr	55	7	36	385
60 to but less than 70 km/hr	65	4	40	260
		$n = \Sigma f = 40$		$\Sigma fx = 1620$

Range	SD
Range = Highest Value - Lowest Value *Range = 70 - 10* *Range = 60*	Using a Scientific Calculator $s = 14.5$ *(rounded to 1dp)* (assuming sample)
IQR	
The first quartile (Q_1) is the $\frac{n+1}{4}$ $= \frac{40+1}{4}$ $= \frac{41}{4}$ $= 10.25^{th}$ value using interpolation *The first quartile* $(Q_1) = 20 + \frac{10.25-3}{8} \times 10$ $= 20 + \frac{7.25}{8} \times 10$ $= 29.1$	The third quartile (Q_3) is the $\frac{3(n+1)}{4}$ $= \frac{3 \times 41}{4}$ $= \frac{123}{4}$ $= 30.75^{th}$ value using interpolation *The third quartile* $(Q_3) = 50 + \frac{30.75-29}{7} \times 10$ $= 50 + \frac{1.75}{7} \times 10$ $= 52.5$
The IQR is 52.5 – 29.1 = 23.4	

The Ogive can also be used to calculate the quartiles and IQR



4. The number of students attending a class (maximum 25) for 30 lessons is given in the table below. The assumption is that the original values are lost and it is a sample.

Number attending (x)	Frequency(f)	$f \times x$	Cumulative Frequency
20	1	20	1
21	1	21	2
22	2	44	4
23	5	115	9
24	11	264	20
25	10	250	30
Total	$\Sigma f = 30$	$\Sigma fx = 714$	

Range	SD
Range = Highest Value - Lowest Value *Range = 25 - 20* *Range = 5*	Using a Scientific Calculator $s = 1.27$ *(rounded to 2dp)* (assuming sample)
IQR	
The first quartile (Q_1) is the $\frac{n+1}{4}$ $= \frac{30+1}{4}$ $= \frac{31}{4}$ $= 7.75^{th}$ value The 7th and 8th values are both 23. $Q_1 = 23$	The third quartile (Q_3) is the $\frac{3(n+1)}{4}$ $= \frac{3 \times 31}{4}$ $= \frac{93}{4}$ $= 23.25^{th}$ value The 23rd and 24th values are both 25. $Q_3 = 25$
The IQR is 25 – 23 = 2	

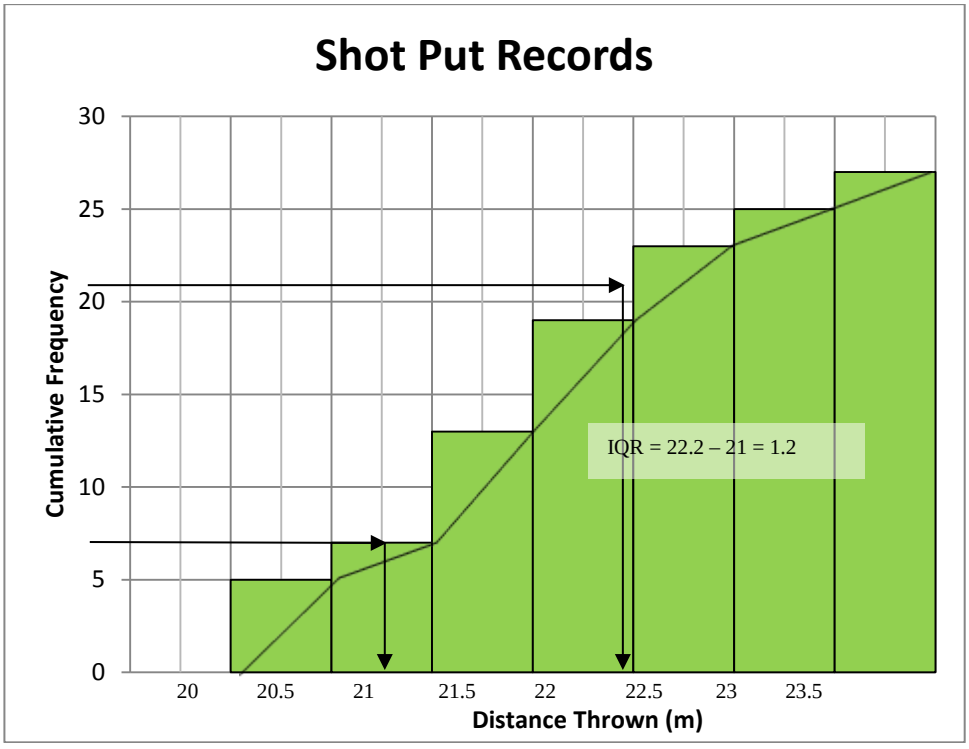
5. The Shot Put distances thrown by 27 world champion shot putters are given in the table below. The unit is metres (m). The assumption is that the original values are lost and it is a sample.

Distance (x)	Group Centre	Frequency (f)	Cumulative Frequency	$f \times x$
20 to but less than 20.5	20.25	5	5	101.25
20.5 to but less than 21	20.75	2	7	41.5
21 to but less than 21.5	21.25	6	13	127.5
21.5 to but less than 22	21.75	6	19	130.5
22 to but less than 22.5	22.25	4	23	89
22.5 to but less than 23	22.75	2	25	45.5
23 to but less than 23.5	23.25	2	27	46.5
	Total	$\Sigma f = 27$		$\Sigma fx = 581.75$

Range	SD

Range = Highest Value - Lowest Value *Range = 23.5 - 20* *Range = 3.5*	Using a Scientific Calculator $s = 0.901$ *(rounded to 3dp)* (assuming sample)
IQR	
The first quartile (Q_1) is the $\frac{n+1}{4}$ $= \frac{27+1}{4}$ $= \frac{28}{4}$ $= 7^{th}$ value Interpolation is not required here as the 7^{th} value is 21.	The third quartile (Q_3) is the $\frac{3(n+1)}{4}$ $= \frac{3\times 28}{4}$ $= \frac{84}{4}$ $= 21^{st}$ value By interpolation $\textit{The third quartile } (Q_3) = 22 + \frac{21-19}{4}\times 10$ $= 22 + \frac{2}{4}\times 0.5$ $= 22.25$
The IQR is 22.25 – 21 = 1.25	

The Ogive can also be used to calculate the quartiles and IQR



Comparing Data

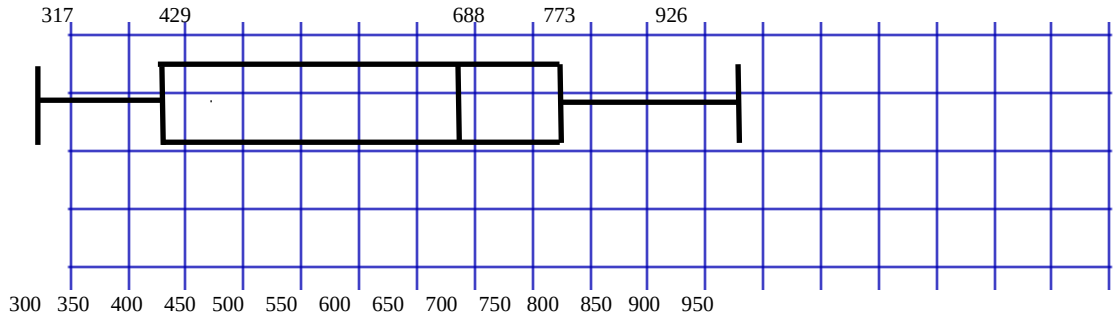
1. The lifetime, in hours, of a sample of 15 light bulbs is: 351, 429, 885, 509, 317, 753, 827, 737, 487, 726, 395, 773, 926, 688, 485.

   In order to calculate the median, the values must be ordered.
   317, 351, 395, 429, 485, 487, 509, 688, 726, 737, 753, 773, 827, 885, 926

   From the Measures of Central Tendency section, the median is 688.

   From the Measures of Spread section, the quartiles are 429 and 773.

   The lowest and highest values are 317 and 926.



2. The Car Speed Data was collected from a school zone just prior to an advertising campaign. The data collected was:

Car Speed Data (in km/hr)

12	41	44	45	28	40	32	62	46	25
31	35	31	20	59	27	49	19	58	38
22	50	46	14	33	48	25	32	52	69
40	52	57	27	61	42	39	64	52	27

After an advertising campaign to make motorists more aware of speeding in school zones, the following data was collected.

22	45	42	27	27	40	45	26	75	25
31	48	30	16	28	42	19	55	46	38
34	50	14	59	33	42	30	32	36	18
38	41	50	14	31	39	25	46	39	27

To organise the data, a back to back stem and leaf plot will be used.

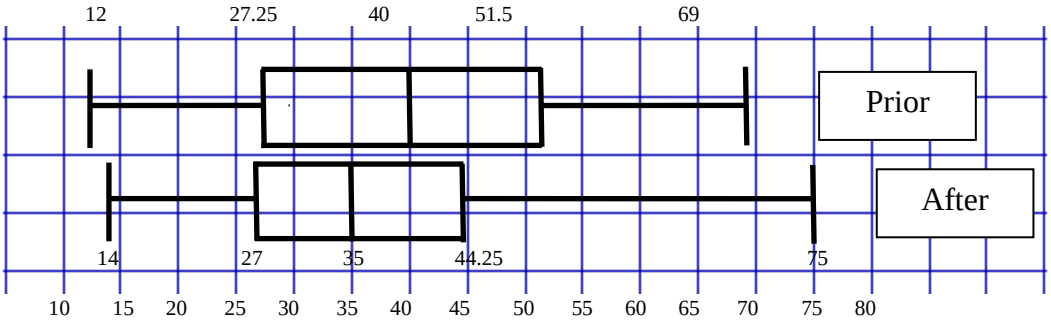
After advertising	Stem	Prior to Advertising
9 8 6 4 4	1	2 4 9
8 7 7 7 6 5 5 2	2	0 2 5 5 7 7 7 8
9 9 8 8 6 4 3 2 1 1 0 0	3	1 1 2 2 3 5 8 9
8 6 6 5 5 2 2 2 1 0	4	0 0 1 2 4 5 6 6 8 9
9 5 0 0	5	0 2 2 2 7 8 9
	6	1 2 4 9
5	7	

For the Prior to Advertising Data

Lowest Value = 12	First Quartile is the $\frac{n+1}{4}$ $=\frac{40+1}{4}$ $=\frac{41}{4}$ $=10.25^{th}$ value This means that the first quartile is a quarter of the way between the 10th and 11th values. First Quartile = 27 + 0.25 x (28-27) = 27.25	Median is the $\frac{n+1}{2}$ $=\frac{40+1}{2}$ $=\frac{41}{2}$ $=20.5^{th}$ value Median = 40	Third Quartile is the $\frac{3(n+1)}{4}$ $=\frac{3\times 41}{4}$ $=30.75^{th}$ value This means that the third quartile is three quarters of the way between the 30th and 31th values. Third Quartile = 50 + 0.75 x (52-50) =51.5	Highest Value = 69
Lowest Value = 12	First Quartile *(Q_1)* = 27.25	Median = 40	Third Quartile *(Q_3)* = 51.5	Highest Value = 69

For the After the Advertising Data

Lowest Value = 14	First Quartile is the $\frac{n+1}{4}$ $=\frac{40+1}{4}$ $=\frac{41}{4}$ $=10.25^{th}$ value This means that the first quartile is a quarter of the way between the 10th and 11th values. First Quartile = 27	Median is the $\frac{n+1}{2}$ $=\frac{40+1}{2}$ $=\frac{41}{2}$ $=20.5^{th}$ value Median = (34+36)/2=35	Third Quartile is the $\frac{3(n+1)}{4}$ $=\frac{3\times 41}{4}$ $=30.75^{th}$ value This means that the third quartile is three quarters of the way between the 30th and 31th values. Third Quartile = 42 + 0.75 x (45-42) =44.25	Highest Value = 75
Lowest Value = 14	First Quartile *(Q_1)* = 27	Median = 35	Third Quartile *(Q_3)* = 44.25	Highest Value = 75



The quartiles and the median are lower after the campaign, based on this there is some evidence of a drop in speed. The lowest and highest value have increased, these can be outliers and so unreliable. The Stem and Leaf Plot suggests that the highest is an 'odd' value.

3. The heart rate of a group of athletes was compared to the heart rate of a group of office workers after climbing a set of stairs. The data is presented below:

Athletes

96	111	88	79	101	91
104	106	121	93	103	96
85	117	126	97	83	93
112	106	110	116	91	88

Office Workers

114	107	94	113	113	97
118	103	121	127	117	131
145	132	118	108	145	126
100	138	120			

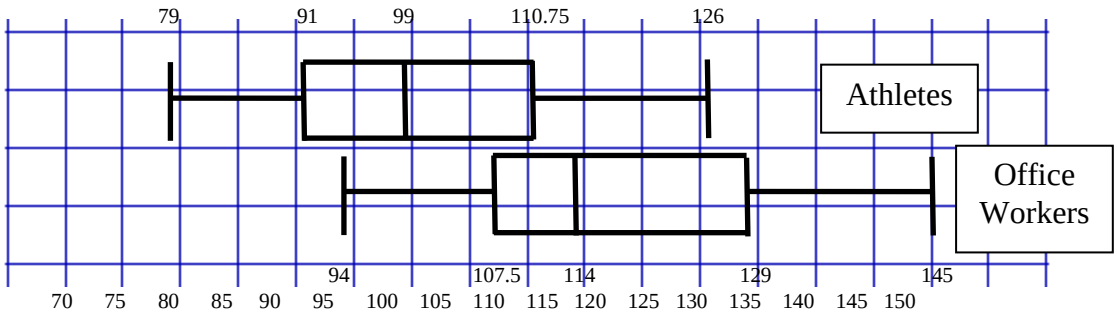
Office Workers	Stem	Athletes
	7	9
	8	3 5 8 8
7 4	9	1 1 3 3 6 6 7
8 7 3 0	10	1 3 4 6 6
8 8 7 4 3 3	11	0 1 2 6 7
7 6 1 0	12	1 6
8 2 1	13	
5 5	14	

For the Athletes

Lowest Value = 79	First Quartile is the $\frac{n+1}{4}$ $=\frac{24+1}{4}$ $=\frac{25}{4}$ $=6.25^{th}$ value This means that the first quartile is a quarter of the way between the 6th and 7th values. First Quartile = 91	Median is the $\frac{n+1}{2}$ $=\frac{24+1}{2}$ $=\frac{25}{2}$ $=12.5^{th}$ value Median = (97+101)/2=99	Third Quartile is the $\frac{3(n+1)}{4}$ $=\frac{3\times 25}{4}$ $=18.75^{th}$ value This means that the third quartile is three quarters of the way between the 18th and 19th values. Third Quartile = 110 + 0.75 x (111-110) =110.75	Highest Value = 126
Lowest Value = 79	First Quartile (Q_1) = 91	Median = 99	Third Quartile (Q_3) = 110.75	Highest Value = 126

For the Office Workers

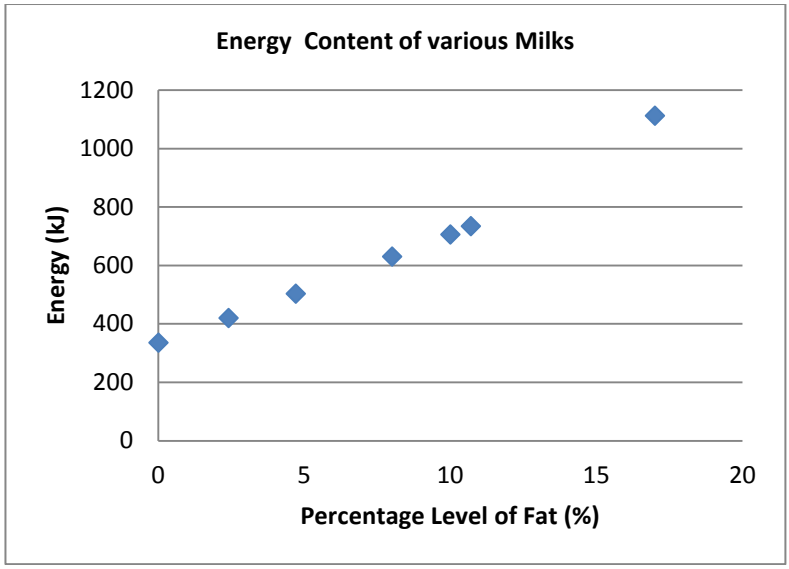
Lowest Value = 94	First Quartile is the $\frac{n+1}{4}$ $=\frac{21+1}{4}$ $=\frac{22}{4}$ $=5.5^{th}$ value First Quartile = 107 + 0.5 x (108-107) = 107.5	Median is the $\frac{n+1}{2}$ $=\frac{21+1}{2}$ $=\frac{22}{2}$ $=11^{th}$ value Median = 114	Third Quartile is the $\frac{3(n+1)}{4}$ $=\frac{3\times 22}{4}$ $=16.5^{th}$ value Third Quartile = 127 + 0.5 x (131-127) =129	Highest Value = 145
Lowest Value = 94	First Quartile (Q_1) = 107.5	Median = 114	Third Quartile (Q_3) = 129	Highest Value = 145



It is quite clear that there is a difference between the two groups.

Correlation and Regression

1. (a) Excel scatterplot. The scatterplot suggests Very Strong Correlation.



(b) The correlation coefficient = 0.989; This also suggests very strong correlation.

(c) The regression line;

$$y = 44.4x + 300 \quad or$$

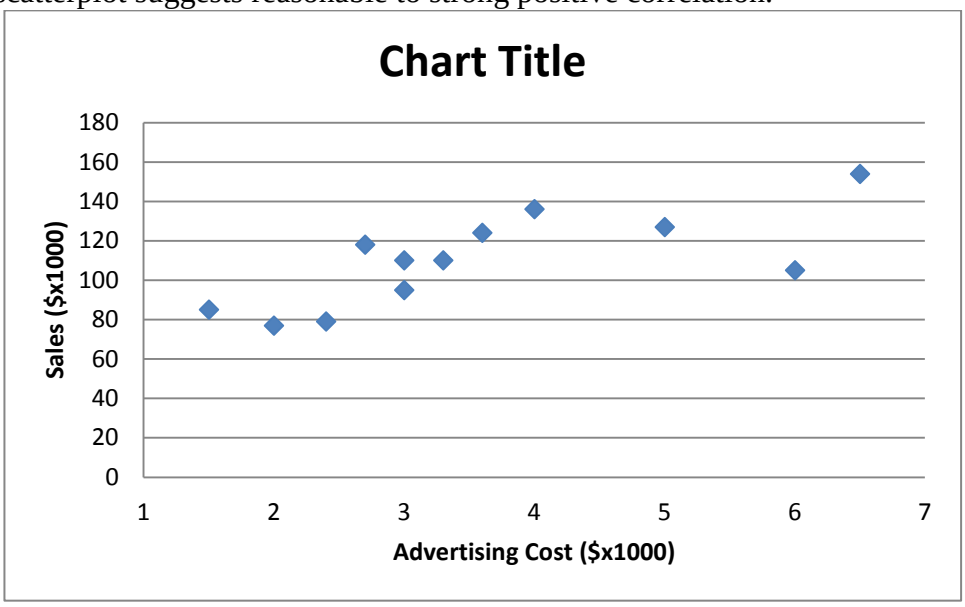
$$E = 44.4F + 300$$

where E - energy content of the milk

F - percentage Fat level of the milk

2. A large company is making note of the amount spent on advertising each month and then comparing this with the amount of monthly sales.

(a) The scatterplot suggests reasonable to strong positive correlation.



(b) The correlation coefficient = 0.735 suggesting Quite Strong Positive Correlation

(c) The regression line.

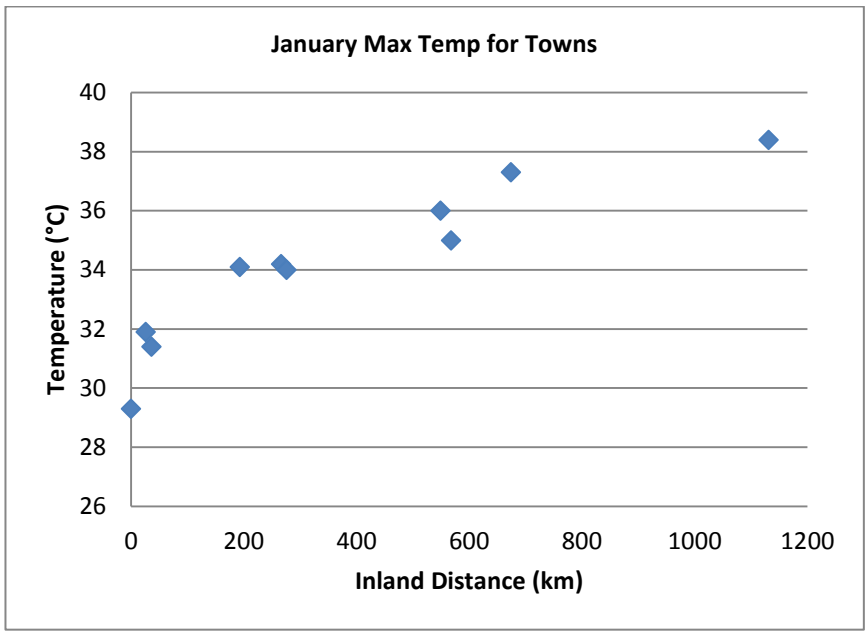
$$y = 11.2x + 70 \quad or$$

$$S = 11.2A + 70$$

where S - Sales per month ($000)

A - Advertising Costs per month ($000)

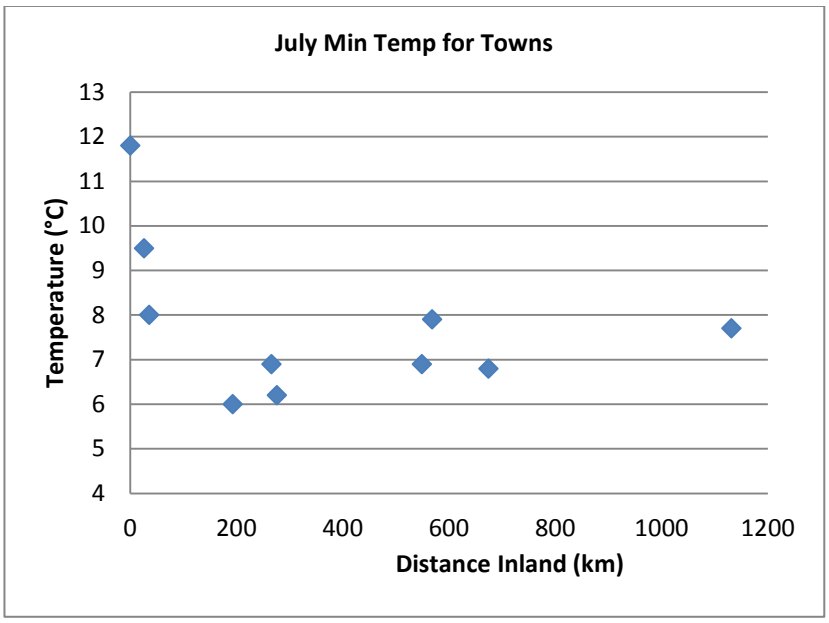
3. (a) Distance Inland vs January Average Max Temp. The scatterplot suggests strong positive correlation.



Correlation coefficient = 0.92; this suggests Strong Positive Correlation.
Regression line:

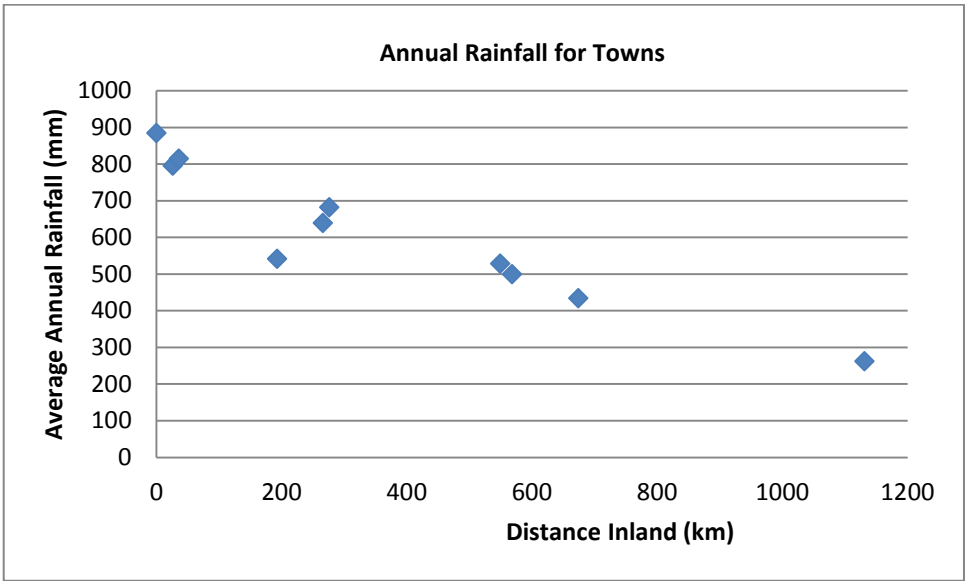
$$y = 0.0071x + 31.5 \quad or$$
$$T_{Jan\max} = 0.0071d + 31.5$$
where $T_{Jan\max}$ *- Jan Average Max Temp (°C)*
d - Inland Distance (km)

(b) Distance Inland vs July Average Minimum Temp. There is some (negative?) correlation here. It appears that once you get 200km inland that the minimum temperature stops decreasing and then increases at a lower rate.



Correlation Coefficient = 0.55; this also suggests slight positive correlation. It is not meaningful to proceed to regression.

(c) Distance Inland vs Annual Rainfall. Scatterplot suggests strong negative correlation. This means the greater the distance inland, the lower the annual rainfall.



Correlation coefficient = -0.94; this suggests Strong Positive Correlation.

Regression line:

$$y = -0.505x + 796 \quad or$$

$$R_{Av} = -0.505d + 796$$

where R_{Av} *- Average Annual Rainfall (mm)*

d - Inland Distance (km)

Comments:

Graph (a) is not surprising. Maximum temperature can be influenced by other factors such as altitude but distance from the sea is a significant one.

Graph (b) is a very interesting graph. Perhaps altitude has an effect over 200 km inland. This could lead to further investigations.

Graph (c) is expected. However rainfall can be affected by topography such as mountain ranges. This usually creates a rainy side or a rain shadow depending on the direction of the prevailing wind. There is one location (Blackwater) that is lower than the regression line suggests, perhaps this location is in a rain shadow.