ACE2046 Quantitative Techniques
Statistical Computing

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 1, 2012-13

1 Basics (Revision)

1.1 Introduction and Definitions

Statistics is the study of random variation, i.e. where the outcome of an experiment cannot be predicted exactly. It involves the collection and interpretation of data to enable us to make inferences from the sample data to the population from which the sample has been taken. It allows us to distinguish real differences from random variation.

- **Random variables:** The quantities measured in a study. A particular outcome is an **observation**.
- **Data:** Several different observations collected together.
- **Population:** The collection of all possible outcomes.
- **Sample:** In practice, we almost always cannot observe the whole population. Instead we observe a **sample**, or sub–set, of the population. It is essential that this sample is **representative** of the population, and so we usually take a **random sample**, in which all members of the population are equally likely to be selected.

Variables can be

- **qualitative** (e.g. flower colour, seed shape) or
- **quantitative** (e.g. seedling height, weight of elephant).

Quantitative variables can be

- **discrete** (only a countable number of values – e.g. number of germinating seeds) or
- **continuous** (e.g. survival times, weight, height).

1.1.1 Example data sets

Milk yields from cows : Milk yields are compared from cows grazed on two different types of pasture over a period of time. The results were:

A:	17.4	19.8	16.3	21.3	24.3	20.2	19.3
B:	18.8	23.2	17.8	24.8	25.2	84.1	

Eye colour of students : Fifty students were randomly sampled and a note was made of their eye colour. The results obtained were:

Blue	Brown	Green	Grey
22	18	6	4

Students and television : Twenty Newcastle students were asked how many TV programmes they watched in the the preceding week. The results are shown below.

$$32, 21, 5, 15, 39, 28, 28, 11, 8, 26, 15, 25, 15, 24, 24, 44, 12, 13, 12, 21.$$

Students eating fruit : The number of portions of fruit consumed in a day by 100 students was recorded.

The results were:

No. of portions of fruit per day	0	1	2	3	4	5
Frequency	22	23	35	14	5	1

Cholesterol and diet : The fasting cholesterol concentration in plasma of people eating 3 different amounts of whole grain (in servings per day) was measured and the concentration recorded (in mmol/L).

The results were

Figure 1: Cross section of a colon crypt

6 servings per day									
4.83	4.80	5.17	5.02	5.11	4.86	5.20	4.89	5.11	4.59
4.92	5.14	4.83	4.95	4.95	5.14				
3 servings per day									
5.13	5.10	4.77	5.22	5.28	5.28	4.89	4.98	4.89	5.16
4.86	5.10	5.04	5.13						
0 servings per day									
5.26	5.32	5.28	5.02	5.14	5.14	5.02	5.14	4.96	5.14

Colon cancer cells : The number of damaged cells in colon crypts caused by two different preparations of a carcinogen was tested by splitting eight colon crypts into two halves and applying one preparation to one half and the other to the other half.

Crypt number	1	2	3	4	5	6	7	8
Prep. A:	31	20	18	17	9	8	10	8
Prep. B:	18	17	14	11	11	7	5	6

1.2 Summary measures

1.2.1 Introduction

For quantitative variables, it is useful to summarise the data in two ways:

- Measures of **location** – a quantity which is 'typical' of the data, or 'central' in the data;
- Measures of **spread** – which quantifies the variability in the data.

Before we begin, we need to establish our notation. Suppose we have a random sample of size n, we denote each value by $x_1, x_2, \ldots, x_n$. So if we had a sample $\{0, 3, 2, 0\}$ then $n = 4$, and $x_1 = 0$, $x_2 = 3$, $x_3 = 2$, and $x_4 = 0$.

1.2.2 Measures of location: The sample mean

This is the most widely used measure. It is defined by:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

which can be more conveniently written as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

So in practice, we add up all the observations, and then divide this by how many observations we have. Sometimes data are tabulated. Thus if we have values $x_1, x_2, \ldots, x_k$ occurring with frequencies $f_1, f_2, \ldots, f_k$, then the sample mean is given by:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i f_i \ .$$

Example: Milk yields : The mean milk yield from cows grazed on pasture A is given by

$$\begin{aligned}\overline{x}_A &= \frac{17.4 + 19.8 + 16.3 + 21.3 + 24.3 + 20.2 + 19.3}{7} \\ &= \frac{138.9}{7} \\ &= 19.8,\end{aligned}$$

and from pasture B we have

$$\begin{aligned}\overline{x}_B &= \frac{18.8 + 23.2 + 17.8 + 24.8 + 25.2 + 84.1}{6} \\ &= \frac{163.9}{6} \\ &= 32.3.\end{aligned}$$

Example: Students eating fruit :

Here we have an example of a frequency table. If data are given in this form, we need to use the alternative formula for the mean:

$$\begin{aligned}\overline{x} &= \frac{1}{100}\left\{(0 \times 22) + (1 \times 23) + (2 \times 35) + (3 \times 14) + (4 \times 5) + (5 \times 1)\right\} \\ &= \frac{1}{100}\left\{0 + 23 + 70 + 42 + 20 + 5\right\} \\ &= \frac{1}{100} \times 160 \\ &= 1.6.\end{aligned}$$

The **mode** of this dataset is 2, since this occurs 35 times – more than any other count.

1.2.3 Measures of location: The sample median

This is the middle observation when the data are ranked from smallest to largest. So we put the data in ascending order and simply 'pick out' the one in the middle. This is fine if we have an odd number of observations. However, if we have an even number of observations, there is no one value "in the middle", and so we have to take an average of the middle *two* values.

$$\text{median} = \begin{cases} \frac{n+1}{2}th \text{ smallest observation} & n \text{ odd} \\ \text{average of } \frac{n}{2}th \text{ and } \left(\frac{n}{2}+1\right)th \text{ smallest observations} & n \text{ even.} \end{cases}$$

Example: Milk yields : Now do you notice the large observation of 84.1 from a cow who grazed on pasture B? Such an outlying observation could have greatly influenced the calculation of the mean yield (and might be due to recording error). Here, it might be more appropriate to use the median as a measure of location. Placing the observations in ascending numerical order, we get:

$$17.8,\ 18.8,\ 23.2,\ 24.8,\ 25.2,\ 84.1.$$

We have an even number of observations, so for the median we calculate the average of the middle *two* observations:

$$M_B = \frac{23.2 + 24.8}{2} = 24,$$

which is quite a bit smaller than $\overline{x}_B = 32.3$!

For pasture A, we have the following data

$$16.3, 17.4, 19.3, 19.8, 20.2, 21.3, 24.3$$

so we have a median of $M_A = 19.8$.

1.2.4 Measures of location: The mode

This is the value which occurs most often. It is usually only used if the data are discrete.

1.2.5 Measures of location: Comparison

So when do we use the mean and not the median? Or when should we use the median and not the mode? Here are a few pointers:

- If the distribution of the data is reasonably **symmetric**, mean ≈ median. If it is also **unimodal** then mean ≈ median ≈ mode. So we could use any. However, since the mathematical properties of $\overline{x}$ are much easier to determine than that of the median, for example, we almost always use $\overline{x}$;
- If the distribution is **asymmetric**, $\overline{x}$ can be distorted by occasional extreme observations, and we prefer to use the median which is more 'robust'. Because the median just picks out the observation in the middle, it doesn't matter how extreme the largest or smallest observations are;
- Sometimes with asymmetric data, it is useful to transform the data to try to obtain a symmetric distribution which is easier to analyse.

1.2.6 Measures of spread

Measures of spread attempt to quantify the variability of data, or how "spread out" observations are. If you are asked to summarise some data, you should always give at least one measure of location *and* one measure of spread.

1.2.7 Measures of spread: The range

This is the difference between the largest and smallest observations:

$$\text{range} = x_{\max} - x_{\min}.$$

Example: Students and Television The range for this dataset is:

$$39 - 5 = 34$$

1.2.8 Measures of spread: The sample variance

The sample variance is the average squared distance of observations from the mean:

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n - 1},$$

which, like the sample mean, can be written in a more convenient form:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Here, we divide by $n-1$ to correct for bias caused by the fact that $\overline{x}$ is an estimate of the true mean, and subject to error. For tabulated data, we have the alternative formula

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(x_i - \overline{x})^2.$$

The **sample standard deviation** is the square root of the sample variance, and is preferred as a summary measure as it is in the units of the original data. We denote this by s.

Example: Milk yields from cows : We have already calculated measures of location for these data. For cows which grazed on pasture A, there were no obvious outliers in the data, and so the sample standard deviation is probably the measure of spread to calculate. The variance is given by

$$\begin{aligned} s^2 &= \frac{(17.4-19.8)^2 + (19.8-19.8)^2 + \cdots + (19.3-19.8)^2}{6} \\ &= \frac{5.76 + 0 + 12.25 + 2.25 + 20.25 + 0.16 + 0.25}{6} \\ &= 6.82. \end{aligned}$$

Thus the standard deviation is simply

$$s = \sqrt{6.82} = 2.61.$$

1.2.9 Measures of Spread: The interquartile range

The **lower quartile** has one quarter of the data less than it, and the **upper quartile** has three–quarters of the data less than it. So

$$\text{LQ} = \frac{(n+1)}{4}th \text{ smallest observation, and}$$
$$\text{UQ} = \frac{3(n+1)}{4}th \text{ smallest observation.}$$

The interquartile range is simply the difference between these two quartiles:

$$\text{IQR} = \text{UQ - LQ} \ .$$

Example: Milk yields from cows : For data set A, we reorder the data to get:

$$16.3, 17.4, 19.3, 19.8, 20.2, 21.3, 24.3$$

So the quartiles are:

$$\text{LQ} = 17.4 \quad \text{and} \quad \text{UQ} = 21.3$$

and the IQR = 21.3-17.4 = 3.9

1.2.10 Measures of spread: Comparison

Just like we had a choice of summary measures for location, we have a choice of measures for spread too. When do we use the standard deviation and not the interquartile range? When can we use the range? Here are a few pointers:

- As with location summaries, if the distribution of the data is reasonably symmetric, it doesn't really matter whether we use the standard deviation or the interquartile range, though the mathematical properties of s are much easier to determine and so s is often preferred; the simpler *range* is rarely used at all;
- If the distribution of our data is asymmetric, the interquartile range should be used. This is because when calculating the IQR, only the position of observations is taken into account, not the magnitude of observations. Thus, the IQR is much less affected by outliers, and so is more 'representative', than the standard deviation;
- The mean and standard deviation go hand–in–hand, as do the median and IQR, as summaries of location and spread.

1.3 Using Minitab

`Minitab` can be used to calculate all the statistics described in this section, via the commands:

```
Stat -> Basic statistics -> Display Desc. Statistics
```

For the Cholesterol and diet data set, this would give:

```
Variable  C1   N  N*    Mean  SE Mean   StDev  Minimum      Q1  Median      Q  Maximum
C2        0   10   0  5.1420   0.0376  0.1191   4.9600  5.0200  5.1400  5.2650  5.3200
          3   14   0  5.0593   0.0428  0.1603   4.7700  4.8900  5.1000  5.1750  5.2800
          6   16   0  4.9694   0.0422  0.1687   4.5900  4.8375  4.9500  5.1325  5.2000
```

1.4 Graphical presentation of data

The summary measures discussed above provide convenient ways of quantifying the location of a "typical" data point, and how "spread out" the data are around this typical value. However, before deciding which measures of location and spread we should use, we need an idea of what the dataset looks like. Reasonably symmetric data, for example, would lead us to use the mean and standard deviation; the median and interquartile range should be used if the dataset is asymmetric. The following graphical methods provide ways of looking at the shape of the distribution of our sample data.

1.4.1 Bar charts

A bar chart can be used to represent **qualitative** data. There is no scale along the x–axis, as this axis represents the qualitative attribute (such as eye colour or car type) that we are investigating. The y–axis represents the number of people with, say, brown eyes or Toyota cars. There should be clear spaces between the bars. Figure 2 shows a bar chart from `Minitab` .

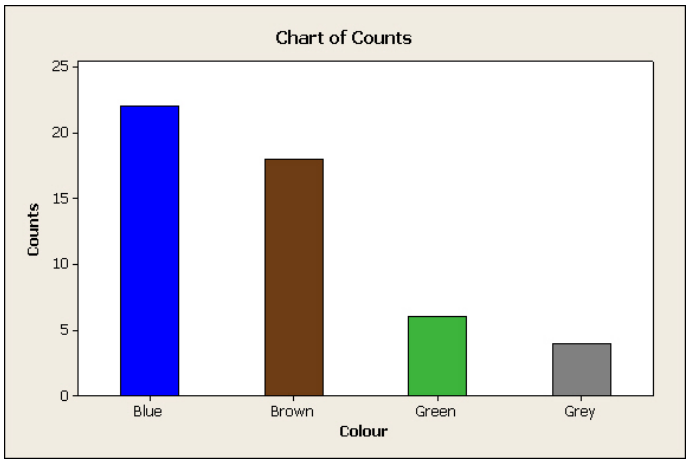

Figure 2: Example of a bar chart in `Minitab`

Figure 2 was generated using

```
Graphs -> Barcharts
```

We can also use a bar chart for discrete quantitative data. For example figure 3 shows the "Students eating fruit" data.

1.4.2 Histograms

A histogram is used to display data which are **quantitative** and **continuous**. Thus, the x–axis has a numeric scale, and there are no spaces between consecutive bars. Figure 4 shows a

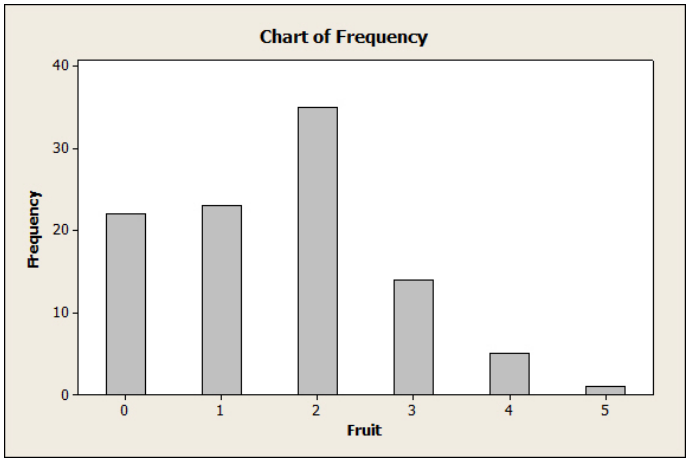

Figure 3: Bar chart for discrete data

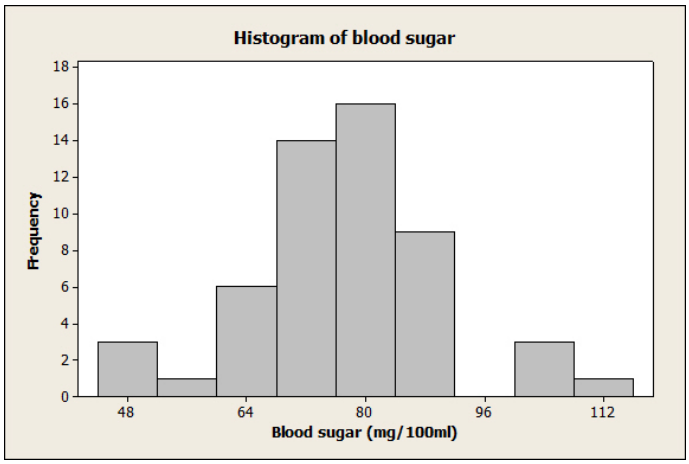

Figure 4: Example of a histogram in `Minitab`

histogram produced in `Minitab` , representing fasting blood sugar concentrations (mg/100ml) for 53 non-pregnant women. Notice there are no gaps between bars as there are with bar charts.

Figure 4 was generated using

```
Graph -> Histogram
```

1.4.3 Box and whisker plots

This is a very useful graphical summary of the spread of the data, and can be used for both discrete and continuous datasets. These plots are most useful for comparing *two or more* datasets, drawn alongside each other on the same scale. "Outliers" are identified and shown by asterisks (*). Using the remaining data, a rectangular box is drawn from the lower quartile to the upper quartile with a dividing line at the median. From the ends of the box, lines ("whiskers") are drawn to the maximum and minimum observations (apart from any outliers). Figure 5 shows three such plots, which compare the cholesterol level and diet.

Figure 5 was generated using

```
Graph -> Boxplot
```

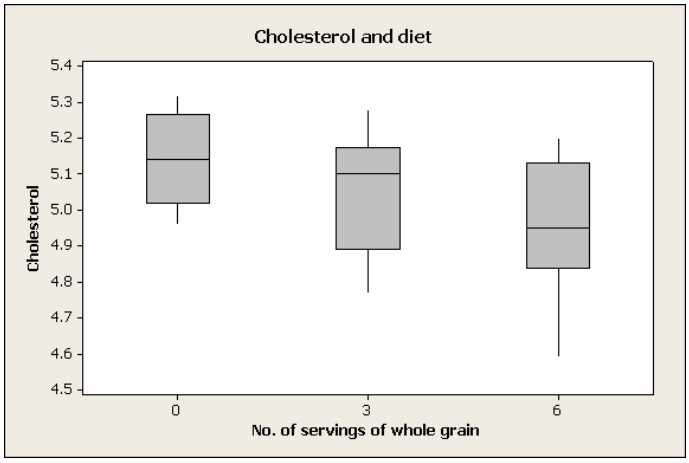

Figure 5: Box and whisker plots in `Minitab`

1.4.4 Stem and leaf plots

Stem and leaf diagrams provide a convenient way of representing **quantitative** data. Table 1 shows a stem and leaf diagram from `Minitab` . The numbers on the left–hand–side of the vertical line represent "tens" (this is the "stem"), and then the readings from the experiment are placed in the appropriate row (second digit only). The main attraction to such a diagram is that it gives us an idea of the spread of the data (it's almost like a bar chart or histogram sideways) while still displaying the raw observations.

0	5	8							
1	1	2	2	3	5	5			
2	1	1	4	4	5	6	8	8	8
3	2	9							
4	4								

Table 1: Stem and leaf plot to show the number of TV programmes watched by students

Table 1 was generated using

```
Graph -> Stem and leaf
```

1.5 Scatterplots

If you have two variables, e.g. weight and height, then we can display them using a scatter plot. To do scatterplots in `Minitab` we do:

```
Graph -> Scatterplots
```

For example figure 6 shows the heights (cm) and weights (kg) of thirty eleven-year-old girls from a school in Bradford.

We can also use a scatter plot when we want to plot how a variable changes over time. We simply use time as the x variable. There is also a special time series plot available in `Minitab` but this would not be appropriate if we had unequally spaced measurements, as is often the case in experiments.

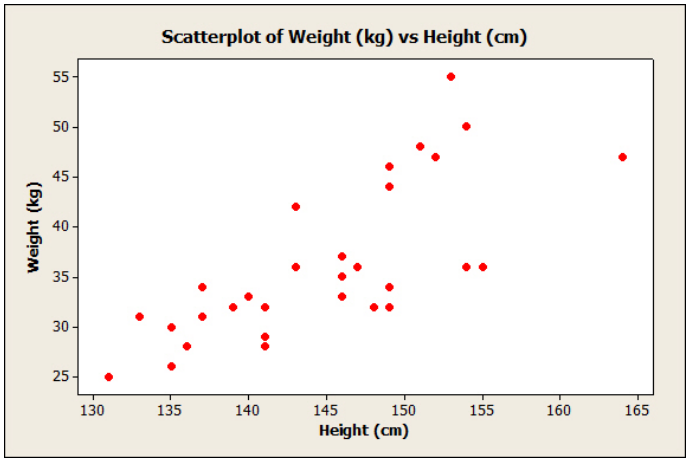

Figure 6: Scatter plot of heights and weights.

1.6 Practical 1

Questions

1. The following data come from an experiment to investigate the effect of a drug on the growth of chickens. Two groups of chickens were reared in identical conditions. However, one group had standard feed and the other had standard feed plus a low dose of the drug. Later, the weight (in pounds) of all the chicks was measured.

Control	3.93	3.78	3.88	3.93	3.84	3.75	3.98
	3.84	3.62	3.84	3.99	2.89	3.60	3.05
Drug	3.96	3.94	4.02	4.06	3.94	4.09	4.17
	4.12	4.23	3.87	4.31	4.02	3.98	3.90

   (a) Using `Minitab` , type the "Control" data into column `C1` of the worksheet, and the "Drug" data into column C2.

   (b) Using the drop-down menus

   ```
   Stat -> Basic  Statistics -> Display Descriptive Statistics
   ```

   obtain the mean and standard deviation for each group. Write your answers in the spaces below:

	Mean	Std dev.	Median	N
Control				
Drug				

   (c) Using the Graph menu, produce histograms of the two groups. Use

   ```
   Graph -> Histogram -> Simple
   ```

   entering `C1` and `C2` as your graph variables. Try putting the two histograms side by side on the same graph. Use

   ```
   Multiple Graphs - In separate panels of the same graph
   ```

   Experiment with the different options available.

   (d) Using the Graph menu, produce boxplots of the two groups. Try to overlay the boxplots on the same page (use Graph - Boxplot - Multiple Y's, Simple). Experiment with the options on Multiple Graphs

(e) Using (b), (c) and (d), compare and contrast the two groups (when you describe a dataset, always comment on the location and spread of the data). Do you think the drug has been effective?

2. The data in the file `ground.txt` refer to some measurements on samples of ground wheat.

   (a) Open a new worksheet in Minitab using

   ```
   File -> New -> Minitab Worksheet
   ```

   (b) Read the data into Minitab. You can do this directly as follows.

   ```
   File -> Other files -> Import special text
   ```

   Enter `c1-c8` for the columns where the data are to be stored. Click `Ok`.
   Enter the following as the File Name.

   `http://www.mas.ncl.ac.uk/~nmf16/teaching/ace2046/ground.txt`

   Click `Open`.
   Alternatively save the data to a file from the module Web page (accessible via Blackboard) and then read the data from this file.

   (c) The percentage protein in the samples is in column 2. Name this column `Protein` .

   (d) Produce a histogram of the data in column 2. Use

   `Graph -> Histogram`

   (e) Do you think that the data look normally distributed?

3. The data in the file `bloodfat1.txt` give concentrations of plasma cholesterol and plasma triglycerides (mg/dl) for 51 male patients where there was no evidence of heart disease.

   (a) Open a new worksheet in Minitab.

   (b) Read the data into Minitab. You can do this directly as follows.

   ```
   File -> Other files -> Import special text
   ```

   Enter `c1` for the column where the data are to be stored. Click `Ok`.
   Enter the following as the File Name.

   `http://www.mas.ncl.ac.uk/~nmf16/teaching/ace2046/bloodfat1.txt`

   Click `Open`.
   Alternatively save the data to a file from the module Web page (accessible via Blackboard) and then read the data from this file.

(c) Name this column `Blood Fat`.

(d) Use

```
Graph -> Histogram -> With Fit
```

to produce a histogram with a superimposed normal distribution, with C1 as your Graph variable.

(e) Produce a normal probability plot of the data in C1. Use

```
Graph -> Probability Plot -> Single
```

with C1 as your graph variable.

(f) Do you think the sample looks reasonably normally distributed? YES/NO
Are there any reasons why you might have doubts about this?

(g) Carry out a one-sample t-test to test the assertion that the population mean length is equal to 150 mg/dl. Your hypotheses are:

$$H_0 : \mu = 150\text{mg/dl} \quad \text{versus} \quad H_A : \mu \neq 150\text{mg/dl} .$$

Use

```
Stat -> Basic Statistics -> 1-Sample t
```

to perform the test in `Minitab` . Enter `C1` for Samples in columns, and enter the Hypothesed mean (150). Write down your mean, test statistic, p-value and 95% confidence interval below:
Mean:= _____mg/dl
Test statistic t =_____, p-value = _____,
95% confidence interval: (_____, _____)

(h) Complete the following paragraph, deleting where appropriate:

Since our p-value is GREATER THAN/LESS THAN 0.05, we HAVE/DO NOT HAVE a significant result at the 5% level of significance. Thus, we should RETAIN/REJECT the null hypothesis H_0. The evidence DOES/DOES NOT suggest that the population mean blood fat concentration differs from 150 mg/dl. The 95% confidence interval CONTAINS/DOES NOT CONTAIN the hypothesised value, further supporting this conclusion.

4. When you have finished please see me to confirm that you have completed the practical.