**DATA ANALYSIS – a how to guide**

Data analysis is a key skill in most science assignments – including experiment reports and research assignments. All data is unique and analysis of data is therefore sometimes very difficult. However, there are some key similarities in the analysis of all data, and these similarities are presented here.

There are four key and common stages in data analysis. These are in order:

**Identifying the trends in the data**

**Identifying the uncertainty and the limitations in the data**

**Making conclusions.**

**Evaluating the reliability and the validity of the conclusions**

Each of these stages is in all science research reports, and they therefore play a major role of science reports in our senior school. Each of these sounds complicated, but can be explained as a simple step by step process you need to follow when examining data. Once you understand the process of data analysis, you can practice it and easily become a Jedi data master.

On the following pages is a step by step guide to each of the stages of data analysis.

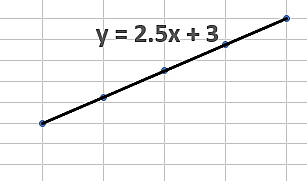
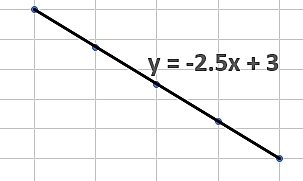
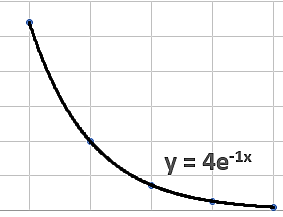
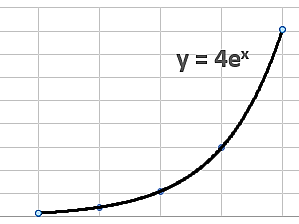
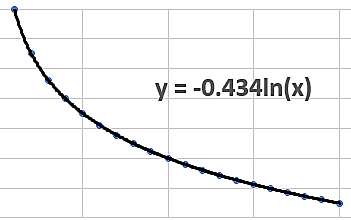
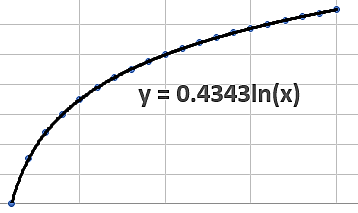
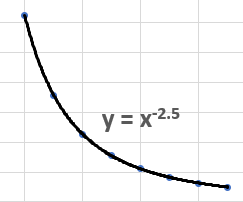
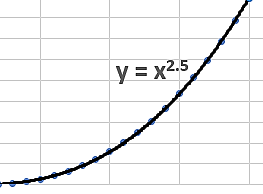
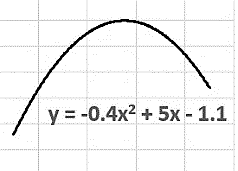
**Identifying and describing Trends and Relationships.**

Generally speaking trends and relationships are the same idea. “Trend” or “trend line” is used a lot when identifying or describing the line of best fit on a graph. Relationship is a broader term, and can be used instead of trend, but also has a wider range of use. Both words can be used to identify or describe how the dependent variable (on the y axis) and the independent variable (on the x axis) are connected to each other. However the word relationship is more useful when talking about the maths that connects two variables.

There is a very simple way to ensure you do a good job identifying trends and relationships – simply follow the steps below.

* Start with the “most important” trend, then work through to the least important. All trends need to be relevant.
* Write a para for each trend you can see. Each graph probably has a trend, and if you have more than one graph, there may be trends across several graphs.
* For each paragraph, follow the same sequence of argument. See next bullet for each para.
* Starting each paragraph with a very simple sentence identifying the trend in the data. The sentence should go along the lines of…

“As the (*independent variable*) increases, the (*dependent variable*) increases/decreases (*use which ever applies*)”.

* The second sentence is a description of the trend. This can be simple, or more complex, depending on the data (shape of the trend line), and how smart you are. Most trends can be described as linear, or exponential, or polynomial, or power, or logarithmic. How to describe each of these trends is shown below.
* Linear. These are straight lines going up (positive slope) or going down (negative slope). A typical description would be “The trend appears linear, with the rate of change in the *dependent variable* being constant at “quote the slope of the line”, as *independent variable* increases. If the line goes through the origin (0,0), the replace the word linear with the word proportional. Proportional trends/relationships are linear ones which go through the 0,0 point.
* Exponential. A line which curves up or down – these can be hard to tell apart from power lines. A typical description (for one which curves up) would be “The trend appears exponential, with the rate of change in the *dependent variable* increasing as *independent variable* increases. Switch out the word increasing, for decreasing if the exponential line curves down.
* Logarithmic. A line which curves up towards a maximum value (it levels out), or a line which curves down towards a minimum value (levels out). A typical description (for one which curves up) would be “The trend appears best described by a logarithmic relationship, with the rate of change in the *dependent variable* decreasing to zero as *independent variable* increases. Ths works for both types of lines. You should also add an estimate of the maximum or minimum value the dependent variable reaches.
* Power. A power trend line (or index line) is very similar to the exponential curve, only it has a more symmetrical arc, and often you have to look at the math equation to identify the difference. Scientists often use the term exponential to describe a trend which is actually best described by a power trend line. Mathematicians would be horrified by this mistake. Use the same description as exponential, but replace the word “exponential” with the word “power”.
* Polynomial. A polynomial trendline is a curved line where the data goes up and down. This is useful for complex data, say in biology where there is seasonal variation in the data. These descriptions are difficult, but a typical description would be “The trend appears best described by a polynomial relationship, with the *dependent variable* increasing to a maximum at quote data and then decreasing as *independent variable* increases.
* The rest of your paragraph will be used to quote data from the table or graph which supports the identification and description of the graph. You may have done enough of this in your description, but if not, you must use data.
* At the end of the paragraph you must interpret what this trend means. This interpretation should point the way towards your conclusion, but not actually be your conclusion. If you get stuck for ideas, look at the graph and see if any of the slope, y axis intercept, max or min values, have some significant meaning. These points are all potentially very meaningful and can be used to make an interpretation of the data.

**Identifying Uncertainty and Limitations.**

Do this is two parts – first discuss the uncertainty, then the limitations.

Uncertainty is a word used to describe how certain you are that your data is precise. This does not mean how accurate or correct your data is, just how certain you are that the data was measured with precision. Data that you definitely know is precise would have “low uncertainty”. Data which you believe is not very precise would have “high uncertainty”.

* Start this section with a simple statement stating the amount of uncertainty in the data. Something along the lines of “There is a (*use an appropriate word*) amount of uncertainty in the data”. For the rest of the paragraph, you need to justify this statement by explaining how you determine the uncertainty – see next bullet point.
* There are three ways to identify the amount of uncertainty in the data. You need to apply all three ways. There is uncertainty in the data even if only one of these ways identifies it.
* The first way to identify uncertainty is from the data values. Data values are often written an additional ± value beside them. This ± value is the uncertainty in the data – but you have to convert it as a percentage in order to use it for uncertainty. Your teacher will expect to see percentage uncertainty in your arguments, not absolute (the actual ± value) uncertainty. Sometimes the uncertainty is shown on a graph as “error bars” (note – they are called error bars, but are actually uncertainty bars) which extend out from the data points. If all you have is the graph, then you need to calculate the percentage uncertainty from the error bars (your teacher can show you how). Generally speaking, uncertainty which is less than 5% is considered low uncertainty. Anything higher than ten percent is considered to be significantly uncertain. In some situations where it is very difficult to measure values – such as health effects in humans - uncertainty can be higher and the data still consider reasonably certain.
* The second way to identify uncertainty is to see if there was data which was not used in the final results. Scientists sometimes label or discuss data in the trials which they consider outliers, and often label them anomalies. These anomalies are not used to calculate averages. This means the anomalies do not show in the final set of results. Because the anomalies have been removed, the final data set can seem to have low uncertainty. However, the fact that there are anomalies in the raw data means there is uncertainty in the data.
* The third way to identify uncertainty in the trend, not so much uncertainty in the data. Look at the data points and the trendline on a graph. Data points which are very close to a trend line suggests that the trend is very consistent, and that there is low uncertainty in the trend. If the data points are not close to the trend line, the trend is not consistent, and that there is a significant amount of uncertainty in the data. The degree to which the data predicts the trend line is measured by the R2 value in excel, and high R2 values means your trend is consistent, and has low uncertainty. R2 is a very precise term however, so do not use it to say your data is accurate.

Limitations are things which mean your experiment process did not necessarily measure what it was intend to. This is a difficult idea so you have to think carefully. However almost all data contains limitations, you just have to identify them and explain what they are.

* Start this section with a simple statement along the lines of “There appear to be (*use an appropriate word*) limitations in the data”. For the rest of the paragraph, you need to identify and briefly explain each limitation – see next bullet point.
* There are three ways to identify the limitations. You need to apply all three ways. Only discuss the limitations you find, although if you cannot identify any, you do need to explain this.
* The first way to identify limitation is to look at how much data was measured. Generally speaking there needs to be at least 5 variations of the independent variable, and at least three trials for each variation. Five variations of the independent variable means there will be five data points on a graph – which is the minimum to use to accurately identify a trend. If there is less than five data points on the graph the data is limited in its ability to identify a trend. If there are no trials and no averages used, the data is also limited in its precision.
* The second way to identify limitations is to look at the range of the independent variable. Sometimes this is not a real-life range, but one the scientist used so the experiment/investigation could be completed in a shorter amount of time. Say a scientist was measuring the rate of decay in human skeletal remains – which may take many years to do in real life conditions. Scientists may choose to use tiny bone fragments rather than whole bones to speed things up. This is not “wrong”, scientists do this all the time, and it is perfectly okay. However, the trend/relationship may be limited because the range of masses in the bone fragments is not representative of real-life ranges.
* The third way to identify limitations is by identifying any controlled variables which were not actually controlled. All controlled variables should be kept constant so that it is only the independent variable causing the change in the dependent variable. Sometimes the controlled variables cannot be kept constant, or were not kept constant. This is quite common in health studies in humans. Studies which investigated the effect of smoking on lung cancer had to try and control all the other factors which cause lung cancer control. Controlling variables is practically impossible when studying long term health risks in humans. If controlled variables were not well controlled, this limits the accuracy of any trend which has been identified.

**Conclusions.**

The conclusion section is usually relatively short. Each conclusion has its own paragraph, and each paragraph follows the same sequence. The first paragraph has to be the main conclusion made from the data, followed by one paragraph each for the lesser important conclusions.

* Start each paragraph with a simple statement of the actual conclusion you have made.
* Follow this statement with a description of the trend supporting the conclusion. If you have a mathematical equation, include it here to support your opening statement. If you do not have a mathematical equation, elaborate and further explain your conclusion or the trend which supports it.
* Explain the real-life implications of the conclusion. That is, explain how this conclusion is relevant and important. For a good grade you must connect you conclusion and real life implications to your research question. This is sometimes difficult, but is a key criterion. By doing this you show that you understand the significance of the data, the trend, and the conclusion.

**Evaluation of reliability and validity**

This sounds complicated, but is actually quite easy. It is easier if you discuss reliability first and then discuss the validity; so, do this in two parts.

Reliability generally refers to whether your data is repeatable. That is, if anyone else was to do the investigation using the same method, would they get the same result? They should, if you had a reliable method. So this part is really not about the data, but about whether your method was reliable. You do not have to guess this, as the uncertainty you identified in the data and in the trend (back in the uncertainty section) can be used here to justify the reliability of the method. If you have identified significant uncertainty in the data or in the trend (only need one), the method is not reliable. If the uncertainty in both the data and the trend/relationship is low, then the method is reliable. Technically reliability should only take one paragraph, and there is a sequence you can follow for this paragraph.

* Start the paragraph by stating if the method reliable, not reliable, or somewhere in between. Then justify this by explaining the uncertainty you identified earlier in the data and/or the trend. Then try to identify why that uncertainty occurred. You will have to try and identify erros you made in the methodology. This last bit is the hard part.

Validity generally refers to whether the trends you identified and the conclusions you made are valid. That is; are the trends real, and are the conclusions meaningful? Again, you don’t have to guess this as the limitations you identified earlier can be used to justify your decision about validity. There is also one additional way to evaluate validity. This is by comparing your conclusions to existing theory and determining if there is any actual error in your conclusion. You can only do this if you know the “correct” answer you were supposed to get, so this option is not always possible. Evaluating validity should also only take one paragraph, and the sequence you should follow is outlined below.

* Start the paragraph by stating if the conclusion(s) made in the investigation is/are valid, not valid, or somewhere in between.
* To justify this, start with determining the amount of error in the conclusion. This is only possible if your conclusion can be compared to a “known” value. Determine the percent difference between the two values

( ). Obviously, if there is a low percentage error, your conclusion is valid; and vice versa. If the conclusion cannot be compared to a “known” value, you can still argue whether there is error by comparing it with existing theory. If your conclusion contradicts an existing well accepted theory, you may have significant error and your conclusion may not be valid, or it is valid, and everyone else is wrong!

* Once you have dealt with error, use the limitations you identified earlier to justify if the conclusion is valid. Limitations which are significant means the conclusion is not likely to be valid, at least not valid for situations outside the conditions used in the investigation. You should then explain how that limitation could have been avoided.