Environmental Systems and Societies

*How to organize and write the data collection and processing section of a lab report/Internal Assessement*

**Learning Goal:** You will be able to create a data table(s), process the data appropriately and present the findings of the processing graphically.

**Success Criteria:** The data table should organize the raw data into an easy to read format. The processing of the data should be meaningful and appropriate for the goal of the investigation and the graphical representation should be clear and easy to understand.

**Part one:** Data tables.

The purpose of a data table is to organize the data in a methodical way. Generally we try to follow a standardized method that is understood by the scientific community. Occasionally one may need to develop a different way to organize the information, but in general it should not be necessary.

Step One: Creating a data table. The first step in organizing your data is determining your variables. You need to know which one you are changing (Independent Variable) and which one you are measuring (Dependent Variable).

Step Two: Make a T-chart:

Label the left side with the name of your independent variable. label the right side with the name of your dependent variable:

Temperature

Height

Step three: Determine how the variables will be measured. Remember to use Scientific units when you measure (Metric), the meter for length, the gram for mass, the liter for volume and Celsius for temperature. Write the abbreviation for your unit of measure in parenthesis next to your label:

Temperature (˚C)

Height (cm)

Step four: Setting up for multiple trials. You will need to divide the dependent variable side of the T-chart into smaller columns. You will probably need to add a row above the new columns explaining what they are.

|  |  |
| --- | --- |
| Temperature (˚C) | Height (cm) |
| Trial 1 | Trial 2 | Trial 3 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Step five: Listing the values for the independent variable. You will write the values that you used when manipulating your independent variable in the left column under the label. They should be sequential and have the same number of significant figures. *You should not include the unit of measure next to these values.*

|  |  |
| --- | --- |
| Temperature (˚C) | Height (cm) |
| Trial 1 | Trial 2 | Trial 3 |
| 25 |  |  |  |
| 30 |  |  |  |
| 35 |  |  |  |

Step six: Writing a title. Just listing the names of your variables is not enough. You will need to write a descriptive, yet specific title for your data. The simplest way to do this is write the dependent variable first then “versus” then the independent variable second. You do not include your units of measure in the title:

|  |
| --- |
| Height of Plant versus Temperature |
| Temperature (˚C) | Height (cm) |
| Trial 1 | Trial 2 | Trial 3 |
| 25 |  |  |  |
| 30 |  |  |  |
| 35 |  |  |  |

Step seven: Recording raw data. You are now ready to record the results from your experiment. It is fine to do this by hand. If you write poorly, then take the time to enter it into the computer so there is no mistaking the value you intended. The raw data should be recorded with the level of precision for every value. Do not try to be overly precise. If the most of your data is precise to the hundredth: 13.45 cm do not put some in at the thousandth: 14.678 round it to: 14.68. This is a warning sign to the people grading the internal assessment that you do not understand the appropriate level of precision for your data. Error on the less precise.

|  |  |
| --- | --- |
| Temperature (˚C) | Height (cm) |
| Trial 1 | Trial 2 | Trial 3 |
| 25 | 13.4 | 13.5 | 13.2 |
| 30 | 13.9 | 14.1 | 14.0 |
| 35 | 12.8 | 12.5 | 12.4 |

You have now correctly constructed a data table and recorded the data appropriately.

**Part two:** Processing the data.

Now that you have organized your raw data, it is time to decide how to process the data. Processing data is essentially plugging the raw data into formulas, which give answers that are representational of *something*. The word something is the tricky part. The *something* might be a population estimate using a formula like the Lincoln-Peterson index. The *something* might also be the central tendency (the average) of the data. Or the *something* might be an equation to gauge the probability that two sets of data are related to each other. Your job is to determine what *something* is relevant to your investigation. It would be a waste of time to use your data to determine population size, if the data has nothing to do with populations. The following is a limited list of possible ways to process the data. There are hundreds of formulas to find the *something* specific to each unique investigation, at this point you do not need to worry about most of those. Here is the list:

**The Central Tendency:**

The central tendency is essentially a number that represents the whole data set. For example if we are testing to see how water affects plant growth we get the following data set:

|  |
| --- |
| Growth of plant versus Amount of water |
| Amount of water (ml) | Growth of plant (cm) |
|  | Trial 1 | Trial 2 | Trial 3 | Trial 4 |
| 10 | 3 | 5 | 5 | 3 |
| 15 | 2 | 2 | 2 | 10 |
| 20 | 1 | 1 | 1.5  | 1 |

**The Mean**

If we add the amount of growth for all the trials that received 10 ml of water together we get 12. If we divide that value by the number of trials we get 4. This is the *mean* of the data set. If we want to generalize we would expect our plant to grow 4 cm if we gave it 10 ml of water. If we look at this data set this prediction would make sense because we have two values above the mean and two values below the mean and they are all only 1 away from the mean.

**The Median**

But if we look at the data for 15 ml of water our mean is 4 cm of growth also. However this result is a little more suspect because if we based our actions only on the mean we would say more water would give us the same amount of growth as 10 ml of water. But a quick glance at the data tells us that 3 out of the four trials produced less growth than the 10 ml mean. And the one trial that produced more growth was significantly higher, which should register as weird. The 10 in this data set is considered to be an outlier. When ever you have an outlier another way to look for the central tendency is to calculate the median. To do this list the data in numerical order:

2,2,2,10

Now find the middle (you can count from each end until you get to the middle:

2, 2 center 2, 10 so the median is between the second and third number, just average them if it is an even data set: 2 + 2 = 4/2 = 2 so the median score is 2 which makes more sense given 3 out of 4 of the measurements are the same as the median. So your conclusion for this set of data might be if we gave the plants 15 ml of water we would expect the plant to grow 2 cm.

**The Mode**

When we look at the data that corresponds to the 15 ml of water we can see that the number that appears the most is 2 cm. That is called the *mode*. Sometimes it is easy to see, but when you have large amounts of data, it is likely that you will need to put the data sets in order, then it will just be a matter of tallying up the presence of the data and finding the one that appears the most. In the third data set you can see that the one that appears the most is 1. It also is the median. The mean is slightly higher at 1.13 cm. In all three cases you can conclude that more water seems to be leading to less plant growth.

*Sources for central tendencies:*

<http://onlinestatbook.com/2/summarizing_distributions/what_is_ct.html>

<http://www.purplemath.com/modules/meanmode.htm>

**The Standard Deviation**

I find it beneficial to calculate all three and then try to explain how they are pertinent to my results. Seldom are the data sets so obvious as the ones above and in reality instead of a number we want to see trends, more to the point we want to know what numbers in our data set fall into what we would call “normal”. What I mean by this is if we give 10 ml of water to a group of plants, not all of the plants will grow 4 cm some will grow 2.97 cm some will grow 5.36 cm. Occasionally we get one that grows 7.08 cm it would be nice to know if the 7 cm grow falls in the normal range or should we essentially ignore it. For that we use *standard deviation.*

 There are really valid reasons for finding the deviations, *squaring them* and finding that sum and then finding the square root, but I will try not to bog you down with those explanations. If you would like them go here for a really easy to follow explanation: <http://www.mathsisfun.com/data/standard-deviation.html> But here is the gist of the method.

1. Find the mean of your data set (sum of the data divided by how many pieces of data you have).
2. Find out how far each data point is from the mean (the measurement – the mean). Now square that number for each value. (This is called the squared difference)
3. Next add the squared values. Now divide that sum by the total number of values (the average of the squared values)

That is it to determine the standard deviation. Here is an example using the 10 ml data set from above.

1. The mean is 4.
2. The squared difference: -12 = 1, 12 = 1, 12 = 1, -12 = 1
3. The average of the squared differences: 1+1+1+1 = 4 / 4 = 1

So the standard deviation from the mean is one. So what?

Here is how this information is helpful, It tells us that any value between 3 and 5 cm is normal growth for the plant when given this much water. The other way to say it is 3-5 cm of growth is within one standard deviation of the mean. If all of your data is within one standard deviation then there is not a lot of variance but if your data ranges between 2 and 6 cm that is two standard deviations and it implies more variability and therefore less certainty.

**The T-test**

You can use a t-test to calculate the probability that the mean values of your control group and your experimental group are different. You might say can’t I just look at the mean values to see that? The answer is no, because you have to take into account the amount of variance in your data set. *Should you do this?* It depends if you are a lawyer trying a case your case is usually stronger if you have more evidence and if the evidence is difficult to refute. Statistics function as the evidence for the case you will make in your conclusion. The more evidence you have and the more it agrees with each other, the stronger conclusion you will write. It may sound daunting, but it is really not that tough and I will try to help you if you need it. If you would like to try and include a t-test in your analysis and discussion, feel free (where appropriate). For a discussion and instructions on when and how to use those methods try: <http://www.socialresearchmethods.net/kb/stat_t.php>

**The Pearson’s Correlation Coefficient**

Remember as you build your argument the quantity and quality of your evidence will determine if the scientific community will agree or disagree with your findings. The Pearson’s correlation coefficient is one more piece of evidence you can use to strengthen your case. It is used when you are trying to see if one variable is related to another. In essence there are three scenarios for correlation. The first is a positive correlation and this can mean as one variable increases, so does the other one. It also can mean as one variable decreases, so does the other one. So a positive correlation means the variables do the *same* thing. The second is a negative correlation. This means that as one variable goes up, the other variable goes down, in other words: a negative correlation means the variables do the opposite thing. The last scenario is no correlation, meaning that as one variable goes up, the other might go up or down or nothing. In other words there is no discernable relationship between the variables. The most common way to calculate correlation between variables is the Pearson correlational coefficient.

The Pearson correlational coefficient assigns a value between -1 and 1 to the relationship between variables. If the value is close to -1 then it is considered to have a likelihood of a negative correlation. If the value is close to 1 then it is considered to have a likelihood of a positive correlation. If the value is close to 0 then it is considered to have a likelihood of no correlation. The letter *r* is used to represent the correlation coefficient.

The following guidelines have been proposed to use as a gauge:

|  |  |
| --- | --- |
|   | Coefficient, *r* |
| Strength of Association | Positive | Negative |
| Small | .1 to .3 | -0.1 to -0.3 |
| Medium | .3 to .5 | -0.3 to -0.5 |
| Large | .5 to 1.0 | -0.5 to -1.0 |

Remember that these values are guidelines and whether an association is strong or not will also depend on what you are measuring. There are many other things to consider when assessing the strength of association between two variables for a more detailed explanation go here: <https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php>

To calculate the Pearson’s correlation coefficient you must meet certain requirements:

*Requirements*

* Scale of measurement should be interval or ratio
* Variables should be approximately normally distributed
* The association should be linear
* There should be no outliers in the data

\*Follow the link on the preceding page for more information on these assumptions.

*Equation:*



Follow this link to find an online calculator to help you determine Pearson’s correlation coefficient: <http://www.socscistatistics.com/tests/pearson/Default2.aspx>

\*The coefficient of determination is *r*2 and that is an expression of how much variance there is in your data sets.

**Part three:** Presentation of processed data:

There are two parts to this section: the first is creating a new data table(s), which display your processed data in organized ways. The second is the creation of the appropriate graph(s), which illustrate what you have calculated and help to show the trends in the data.

**Data tables:** essentially you will construct the data tables as described in section one. You will put the variables on the left side and the processed data on the right. You may record several different calculations in one table, but each one should be labeled clearly. For example:

|  |
| --- |
| Standard Deviation and Pearson’s Correlation Coefficient for amount of water and growth of plant |
| Variable | Standard Deviation | Pearson’s Correlation Coefficient (r) |
| Amount of water (ml) | 1.4 | -0.23 |
| Growth of plant (cm) | 0.4 | -0.23 |

You may also create a data table for *each* method of statistical analysis.

**Graphs**: Once you have processed your data it is time to present what you have calculated. This serves two purposes: 1) It allows for easy inspection of the relationships between variables. 2) It allows for easier identification of trends in the data, which may imply a relationship between variables. The first step in creating a graph is to determine which type of graph is appropriate for your data. There are several types of graphs, but for this class we focus on four main categories. They are:

* Line graphs—these are used when you are graphing continuous data. The most common type of continuous data is time, but it is appropriate if the data is collected at *sequential* intervals like temperature or pH.
* Bar graphs or Histograms—these are used when you are comparing categories. The categories might be locations (quadrats or lakes), seasons, or populations.
* Pie charts—These are used when you are comparing parts of a whole. The parts might be percent of time spent in each stage of insect growth, percentages of a population, or percent of world food production by country. Notice many of these include percentages, which imply parts of a whole.
* Scatter plot—These are used to look for relationships between two variables. The **raw** data is used here and is plotted on the graph then a line of best fit is created. If done in excel the line will be assigned an *r*-value which corresponds to the Pearson’s correlation coefficient. If not you will need to determine if the line points upward (suggesting a positive correlation) or downward (suggesting a negative correlation) or flat (suggesting no correlation).

No matter the type of graph you make there are rules for the construction of the graph:

**A title:** all graphs need a title; it is made exactly as you did when you made the data table. It is the dependent variable vs. the independent variable.

**Axis or pie slices:** For a pie chart you need to make a circle, then divide the circle by the percentages you are using. If doing by hand you will need to approximate the size of the pie to match the percentage of the whole that your measurement is taking. If it is your data is 25% of the whole, that is easy because it is a ¼ of the circle, but if it is 18% that becomes trickier. The way I do it is imagine the circle is a clock. Each hour on the clock counts as about 8%, so if I have a measurement that is 18% I would make my slice from 12:00 to 2:00 then I would add about ¼ of the space between 2:00 and 3:00 (1 hour = 8% 15 minutes or 1/4 = 2%). Like so:

12

1

3

2

For the next slice I would calculate the size and put it in place and continue until the whole circle is filled and the percentages equal 100. You should not have slices that are unaccounted for. If you do readjust your pie sizes—but remember they should be proportional so if you have a slice that is 18% and one that is 38% the second one should be roughly double the size of the first.

For all of the other types of graphs you just need to begin with two lines one represents the y-axis and is vertical and the other represents the x-axis and is horizontal. The question is where to put them. If you remember from math class all numbers will fall on a number line. Zero is in the middle of the number line and the rest of the numbers go on forever in each direction (For the x-axis the negative numbers go left and the positive numbers go right. For the y-axis the negative numbers go down and the positive numbers go up.) If we put these two number lines together it makes this picture:

Quadrant one:

X and Y both positive

Quadrant two:

X negative and Y positive

Quadrant three:

X and Y both negative

Quadrant four:

X positive and Y negative

Usually when we make a graph in early high school we focus on are numbers being in the first quadrant (how do you really have negative growth for a plant?) But as we get into more complicated investigations we realize that we do have negative values that we need to graph. When this happens we simply include all or part of the other quadrants (usually just quadrants 1 and 4 because you don’t typically have negative values for your independent variable). The zero is where the lines meet. So to draw the axis’s for this type of graph it would look like this:

For most purposes we only need the first quadrant, so we would draw the x and y axis’s like this:

By this point you should have your axis (or circle) and your title like this:

Plant growth versus Amount of water

**Labels:** You now have to name each axis. Your y-axis should be named after your dependent variable. Your x-axis should be named after your independent variable like so:

Plant growth versus Amount of water

Amount of water

Plant Growth

**Units of Measure**: For some reason this is an often-missed part of a graph. You need to tell the viewer of the graph how you are measuring each variable. You don’t do this in the title, you do it next to the labels. The units of measure are always S.I. units (International System of Units) the easy way to think of this is to always use the metric system. Here is an example:

Plant growth versus Amount of water

Amount of water (ml)

Plant Growth (cm)

**Intervals:** You must now divide each axis into segments. Each segment is represented by an interval. The interval is evenly spaced and sequential. You also only need to put the intervals that represent your data range. If your data doesn’t start until 45 then don’t start your intervals until 40. If they end at 83 then have the intervals end at 85. The cool consequence of this is it will automatically enlarge your graph. Remember if your axis starts at zero you will need to add a break in the graph to represent the jump from zero to 45. The other problem people have making graphs is how far apart to make the intervals. The simple answer is count how many lines you have on your graph paper (if no lines I go with number of data points) and divide your range with the number of lines. If your data range is 85 – 45 that is a range of 40 (you can round if you need to) and you have 10 lines (or data points) then divide 40 by 10 and get an interval of 4. What if it is 3.81—depends, if the data points are whole numbers I would round to a whole number (4) if the data points are about .5 apart I would round to .5 (4, 4.5, 5…). If the data points are even closer I would consider rounding to hundredths or even more precise, it just depends on your data. For our example the range of plant growth was from 3 cm to 5 cm and the range of watering was from 10 ml to 20 ml. I would take the range of 2 for plant growth and divide it by 4 to get an interval of 0.5 and I would take the ml range of 10 and divide by 4 and get an interval of 2.5 ml. Here is how it would look:

Plant growth versus Amount of water

Amount of water (ml)

Plant Growth (cm)

2.5 5.0 7.5 10 12.5

0.5 1.0 1.5 2.0 2.5

**A Key:** Before you plot your data you need to determine if you need a key. If you are plotting one line or two well-labeled bars of data you will not need a key. Move on. If your data is composed of two lines or more or several pieces of pie, or groups of categories that are repeated on a single graph you need a key. All you do is make a box to the right of the graph with a list of variables and a symbol or color or dashed line next to them that will be on your graph. The point is to make the graph easily understandable. Example of a key:

Key

Tap water

Lake water

**The Data:** The final part of your graph is the data. To plot a point you simply look at your values for your dependent and independent variables. These correspond to positions on the y and x axis respectively. So the dependent variable is the amount of water given and it is 10 ml and the mean growth is 1.5 cm then you would find 10 on the y-axis and 4 on the x-axis and put a point. That is it. It would look like this:

Plant growth versus Amount of water

Amount of water (ml)

Plant Growth (cm)

2.5 5.0 7.5 10 12.5

0.5 1.0 1.5 2.0 2.5

If it is a line graph you connect all of the dots at the end (sequentially). If it is a bar graph you would create a bar from the axis up to the data point. If it is a scatterplot you estimate a line of best fit. If it is a pie chart you estimate the size of your pie slice.