Principles of Ecology BIOL412001

Laboratory 3: Spreadsheet Graphs
Due Date: 1:00 PM on the 22nd (Tues) or 24th (Thurs) of February 2005

THIS IS THE LAST OF ONLY TWO LAB ASSIGNMENTS THAT YOU MUST DO BY YOURSELF. NO GROUP ASSIGNMENTS WILL BE ACCEPTED FOR THIS LAB. THIS IS TO ENSURE THAT YOU ALL CAN USE THE SOFTWARE AT AN ACCEPTABLE LEVEL.

Introduction

This lab continues from the previous one to further develop your skills in using Microsoft Excel (XL for short) as a tool for scientific calculations and the presentation of research results. The previous lab focused on the use of XL for performing a range of calculations and other operations on data that make our lives as scientists easier and more productive. One of the most important aspects of being a scientist is the responsibility to communicate our findings to our peers and the general public in a manner that is clear, concise, and unambiguous. The most effective way of communicating complex numerical results is by using simple and clear graphs. XL has moderately sophisticated graphing capabilities that can be used (and abused) to create scientific graphs. This lab is designed to introduce you to the basic characteristics of good scientific graphs and how to produce them using XL.

This lab is written assuming you have access to Microsoft XL on one of the MS Windows operating systems. This is the standard at TSU, and all campus computers should have this software. Fortunately, there are more similarities than differences between most spreadsheet programs, but it is advisable to use MS XL as opposed to another program because you may spend a lot of time trying to find equivalent commands and functions (NOTE: I am not promoting Microsoft or any of its products). To gain the most from this lab you should work through the handout interactively with the software. This will help reinforce the various techniques and concepts we will cover.

Characteristics of good scientific graphs

As we all know, a picture is worth a thousand words, so we will use pictures wherever possible throughout this handout (that's why it is so long!). Consider Figures 1 and 2 below. Both figures are based on the same data, but which is the easiest and clearest to understand? If you said Figure 1 then you are already aware of what constitutes a good scientific graph. If you said Figure 2 then, well, I guess you really do need to read this handout **VERY** carefully!

Let's look at the specific characteristics of Figure 1 that make it so easy to understand:

1) The figure has a concise title explaining what it is showing. If the figure has a caption (a sentence or paragraph below the figure that provides additional information), then a title is not strictly necessary. You can see the use of captions in just about any scientific paper in the primary literature.
2) The axes are clearly labeled with units of measurement (scale), and the values cover an appropriate range to clearly show the trends in the data.
3) The different data series (lines) are clearly defined with different symbols (triangles vs. circles) and different line styles (solid vs. dashed).
4) There is a clear legend that identifies what the different data series represent.
5) Because there are only two axes for the data (species vs. month), the graph is presented as a two dimensional figure. This makes it very easy to see what the values are for each data point, and to compare the two data series to each other.
6) The overall format and style of the figure is clean and simple, and does not distract from the important message conveyed by the data.

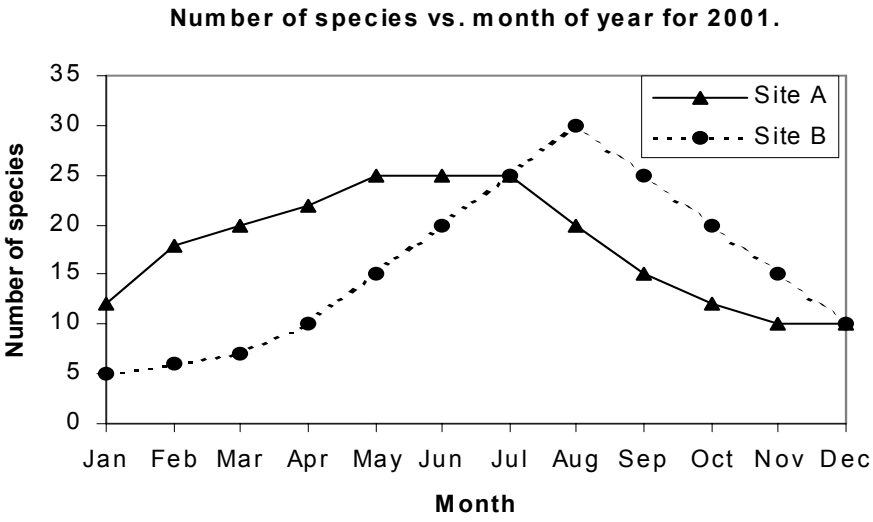


Figure 1.

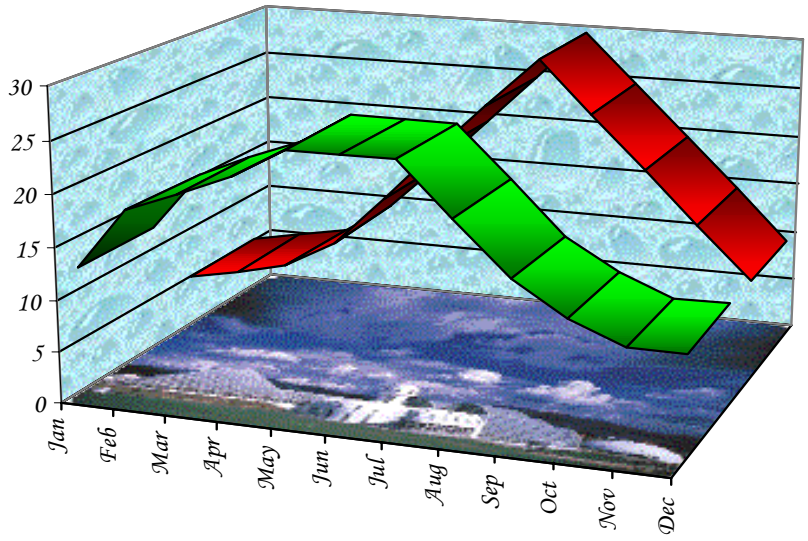


Figure 2.

Now, let's consider the deficiencies of Figure 2:

1) There is no caption or title to identify what is being graphed.
2) Even though there are values indicated on the axes, there are no units or other information indicating what the axes represent or the scale of measurement.
3) The different data series are obvious as long as you have a color display of the figure. Scientific journals are normally printed in black and white, so there would be no way to identify which data series was which if this figure were to be published (trust me – it would NEVER get published). It is good to use color in a graph for a poster or seminar presentation, but you also need some other distinguishing feature such as line or symbol style. Remember that there are lots of color-blind people out there. Relying solely on color is unfair to them.
4) There is no legend to identify what the different data series represent.
5) This is an example of a pseudo three-dimensional graph. There are only two axes (as in Figure 1), but a fake (pseudo) three-dimensional perspective has been imposed on the figure. As you can see, this makes it difficult to identify the actual value for any point on the lines (ribbons) and to

compare the two data series to each other. NEVER USE A FAKE THREE-DIMENSIONAL PERSPECTIVE ON A TWO DIMENSIONAL GRAPH. The only time you should use a 3-D perspective is when you are plotting three data axes to create a response surface.

6) The overall format and style of the figure is distractingly complex. The unnecessary use of visual textures, pictures, fancy fonts, and fake 3-D distracts from the message being conveyed by the data.

In short, if you were to submit a graph like Figure 1 for grading, you would receive full points for it. On the other hand, if you submitted a graph with any of the deficiencies listed above for Figure 2, then you would loose a significant amount of points for each deficiency.

Some useful terms

Frequency: The number of times a value occurs in a dataset. The frequency of occurrence of values in a dataset can be extremely insightful and is often the basis for many different types of statistical comparisons. Frequency is usually represented by a histogram (see below), but the first step in creating a histogram is to construct a frequency table. This type of table usually has two columns: one for each unique value that a dataset contains, or subranges of data if the values are continuous (more on this later); and one column for the number of times each unique value or subrange of values occurs. Below is an example of a frequency table based on the size of fruit sampled from an orchard. You can see that there were a total of 44 fruits that were 4 cm in circumference.

Size (cm in circumference.)	Frequency
1	22
2	34
3	56
4	44
5	16

Numerical and Categorical data: These are the two main types of data you are likely to encounter as a biologist. The distinction is important because each type requires different methods for analysis and presentation.

- **Numerical data** are numbers that have intrinsic meaning. Examples include the height and mass of an individual, or the number of leaves on a plant. These data identify quantifiable differences among objects. For example, we know that an individual who is 180 cm tall is 30 cm taller than an individual who is 150 cm tall. Similarly, a plant with 150 leaves has 50 more leaves than one with 100 leaves. Numerical data can be subdivided into continuous and discrete data. **Continuous data** can take on any value within an observed range, limited only by the accuracy and precision of the instrument used to make the measurement. For example, if we measured the height of all people in the world between the ages of 20 and 60 years of age with an accuracy of 1.0 μm, we would find a bewildering array of different heights between the range of approximately 0.9 m to 2.13 m. In contrast, **discrete data** can only taken on certain values. The most common example of discrete numerical data involves counts of objects. We may count the number of live individuals in a population, but the values can only be integer numbers. You cannot have 2.635 live individuals, but you can have 2, or 3, or 35,651 live individuals.

- **Categorical data** represent discrete categories into which objects can be placed, but those categories have no intrinsic quantitative meaning. Common categories in biology include gender, color, or location where a sample was collected. Names are usually used to identify categories, but numbers can also be used (e.g., males = 0 and females = 1). However, with categorical data, those numbers are completely arbitrary, i.e., they don't have any intrinsic meaning other than identifying the different categories. In the gender example above, the numbers (0 and 1) could

easily be reversed without any loss of meaning. Categorical data are sometimes referred to as discrete data because objects cannot span two or more categories, i.e., categories must be based on discrete characteristics of the objects. If objects are categorized by color, you can't have a value for half a color (nonsensical), or a value for two colors combined (that would be a new color or a new category).

Grouping, or binning, of data: Grouping or binning of data is done to rescale data or to change it from continuous to discrete or categorical data. This may be necessary for many reasons, but the most common that we will encounter is when we want to summarize a large data set by constructing a frequency table or histogram. For example, lets say we need to accurately determine the mass of 100 insects using a high quality balance (both accurate AND precise). It is highly unlikely that any two insects in our sample will have exactly the same mass (our balance is accurate to several milligrams), so we will probably end up with 100 different values for mass. In order to construct a frequency table for our sample, we need to take the range of masses we measured and divide it into several smaller subranges or groups (i.e., bins or categories). These may be 0.00-0.10 g, 0.11-0.20 g, 0.21-0.30 g, etc. Thus, we have taken a *continuous* numerical variable (mass), and converted it into a *discrete* numerical variable. It is now a discrete variable because the mass of each insect must be within **one** of these discrete groups. We cannot have an insect with a mass equal to 2.5 groups because this does not make sense. But we do know that the mass of an insect in the 0.11-0.20 group is smaller than the mass of an insect in the 0.21-0.30 group.

Types of graphs

Excel can construct many different types of graphs: Pies, Stars, Polar charts, 3-D, whatever. Many of them are useless for scientific graphs, or are only appropriate for very specialized types of data. Because of this, we will focus on only four types of graphs that will cover all of our needs in this course. These graphs are histograms, bar charts, line graphs, and scatter plots. Examples of each of these graphs are shown below in Figure 3.

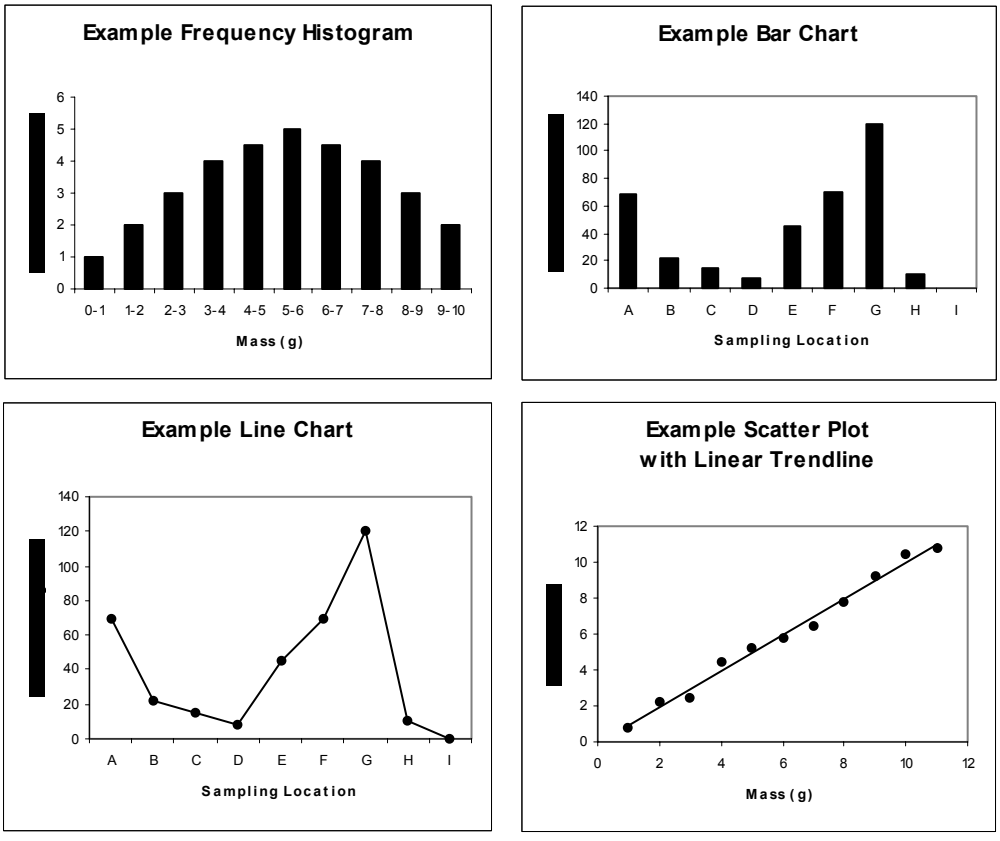


Figure 3: examples of the four types of graphs discussed in this lab.

Histograms and frequency tables

Histograms are graphs that have the frequency of occurrence of an event as the Y-axis and a discrete numerical or categorical variable as the X-axis. Discrete variables may need to be grouped if there are too many groups or classes, or if the frequencies for each group are too small. Just how many groups should be used depends on your data, but sometimes it is necessary to try several group sizes to find one that is appropriate. A simple rule is that too few groups can hide important patterns in the frequency distribution, but too many can introduce irregularities in the shape of the distribution that may obscure important generalizations.

Histograms are used to display how often a particular response or event occurs. For example, in many bird species, both the male and female adults feed the chicks. One may ask if the effort expended in feeding the chicks is comparable for both sexes? One way to demonstrate the effort would be to construct a histogram with sex as the X-axis and frequency of feeding visits to the chicks as the Y-axis.

We often hear that the heights of humans approximates a bell curve (or more accurately, a normal distribution). To demonstrate this, you might measure the height of 100 randomly chosen people of one sex, and then convert this continuous numerical variable into a discrete numerical variable by grouping the data. You would then count the frequency of heights in each group and make a graph with the discrete variable for height on the X-axis and frequency on the Y-axis. The result would look like a bell curve if height really does follow a normal distribution in the population.

To make a frequency histogram in XL requires two separate steps. The first is to make a frequency table from the data (this is the most difficult part, but is fairly simple). The second step is making a bar chart from the frequency table, which is quite easy to do.

Frequency tables: Enter the column headings and data shown below in Figure 4 into a new worksheet. The column labeled RAW DATA is a discrete numerical variable consisting of data we have collected. This could be something like the number of bird species identified at several different locations, or the number of individuals in different age classes. The first task in constructing a frequency table is to decide on the categories you want to group the data into. The example given here is simple: there are only 5 different integer values in the data (3, 4, 5, 6, 7), so these numbers make logical groupings for the frequencies and are the values shown in the column labeled CATEGORIES. We will now tell XL to count all the occurrences of numbers in the RAW DATA column that fall within each of the 5 categories listed in the CATEGORIES column. The result will be reported in the corresponding rows in the column labeled FREQUENCIES.

To do this, we will use a function in XL called FREQUENCY. First, select the cells in the FREQUENCIES column that correspond to the different categories as shown below in Figure 4. Next, select FUNCTION from under the INSERT menu. Then select STATISTICS from the list of function types in the dialog box that appears, and then scroll down to the FREQUENCY function in the list of statistical functions. We now have to enter two sets of data as indicated in Figure 5 below. The first set of data is our raw data, which is referenced in the input box labeled DATA_ARRAY in Figure 5 below. Click your mouse in this input box and then drag your mouse through the cells containing the raw data. Now click your mouse in the second input box labeled BINS_ARRAY. The cells referenced here must contain the different categories you want your raw data grouped into. In our example, these are the numbers that appear in the CATEGORIES column of Figure 4. Simply drag your mouse through that range of cells to enter their reference. **THE FOLLOWING IS VERY IMPORTANT!! HOLD DOWN THE CONTROL (CTRL) AND SHIFT KEYS TOGETHER WHILE CLICKING THE OK BUTTON IN THE DIALOG BOX SHOWN IN FIGURE 5**. The reason we need to do this is that the frequency function is a special function that depends on arrays of data. To make sure XL enters the function correctly, we need to tell XL that the function is an array function by holding down the CTRL

and SHIFT keys while clicking OK. When you have done this, you should see the same numbers on your spreadsheet as you see in the FREQUENCIES column in Figure 6 below.

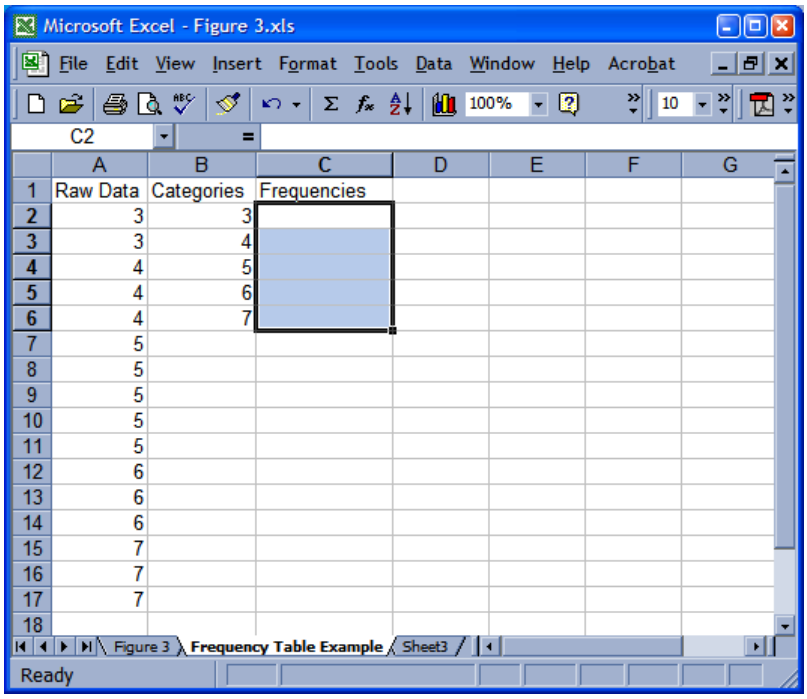


Figure 4.

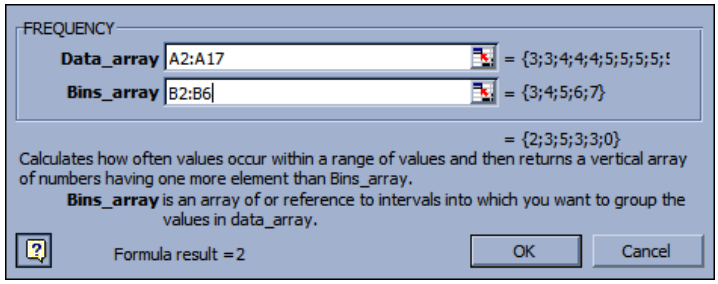


Figure 5.

Remember that the frequencies are the number of occurrences of the raw data in each of the categories you defined. This means that the sum of the frequencies should equal the total number of datum in the RAW DATA column. So an easy way to check your frequency table is to sum together all the numbers in the FREQUENCIES column, and count all the numbers in the RAW DATA column. If these numbers are the same, then your frequency table is correct. If they are different, then something went wrong (you may not have pressed the CTRL and SHIFT keys while clicking OK in the dialog box in Figure 5). You can use the SUM function and the COUNT function to do this, both of which are in the STATISTICS group of functions. I have shown the results of using these functions for our example in Figure 6 below. You should try using them yourself now so that you know how to check any frequency table you may construct in the future.

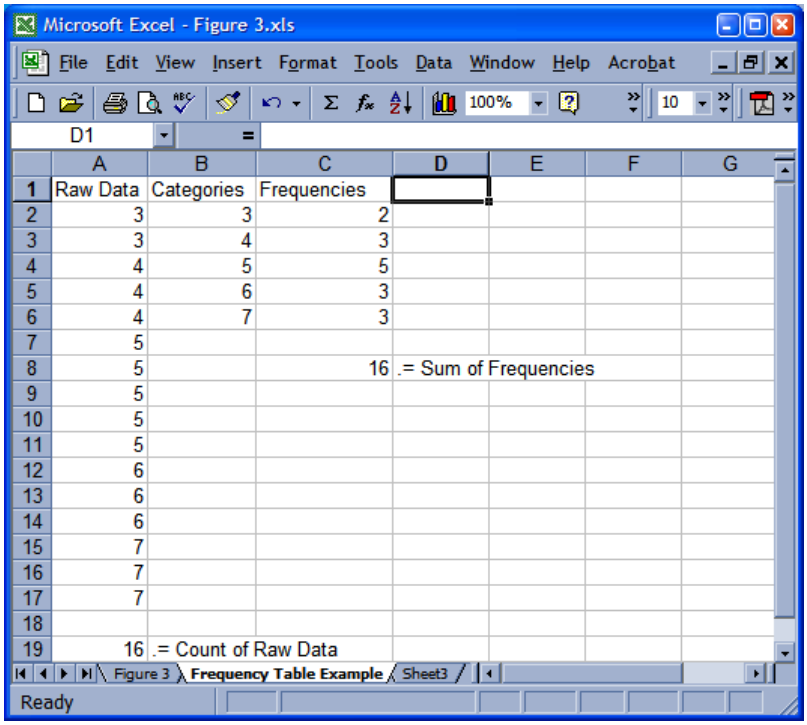


Figure 6.

How would you decide what numbers to use for your categories if your data were a continuous numerical variable? In this case, the categories should be based on the range of your data (largest value minus smallest value), and on the number of bars (groups) you want in the graph. In general, it is best to use integer values to define the boundaries for successive groups, and your groups should all cover the same range of values. The most common exceptions are usually the groups for the smallest and largest values, which may not have many data points in them. In this situation, the range of values covered may need to be larger than for the other groups, either to ensure that there are enough data points in the groups at the extremes of the histogram, or to limit the number of groups needed to cover the full range of the data. How to deal with this type of situation is discussed in the next paragraph.

When you enter values in XL to define the category boundaries (as in Figure 6 above), the numbers you enter set the upper limit for each category. In our example above, the first two categories are defined as 3 and 4 respectively. This means that the first category would in fact contain all values less than or equal to 3. Similarly, the next category would contain all values greater than 3 and less than or equal to 4. The last category defined in Figure 6 has an upper limit of 7. What would we do if we had one datum (single data point) with the value of 11? Would it make sense to include four additional categories with boundaries 8 through 11 to handle this one data point? In some situations that may be the best solution, but an alternative would be to have one category that includes ALL values greater than 7. We can do this very easily by simply including a blank cell at the end of our categories list when we enter that range of cells in the BINS_ARRAY of the FREQUENCY function dialog box (Figure 5). I have shown this range of cells in Figure 7 below, along with the additional data point (11) in Column A, and the new frequency table in column C. Notice that the new frequency table has an extra row at the bottom with the number 1 in it, representing the count of all values greater than 7. In the next section we will discuss how to make a bar chart (histogram) from our frequency table. Since the basic procedure for making a bar chart is the same as for making a line graph, both types of graphs will be covered in the next section.

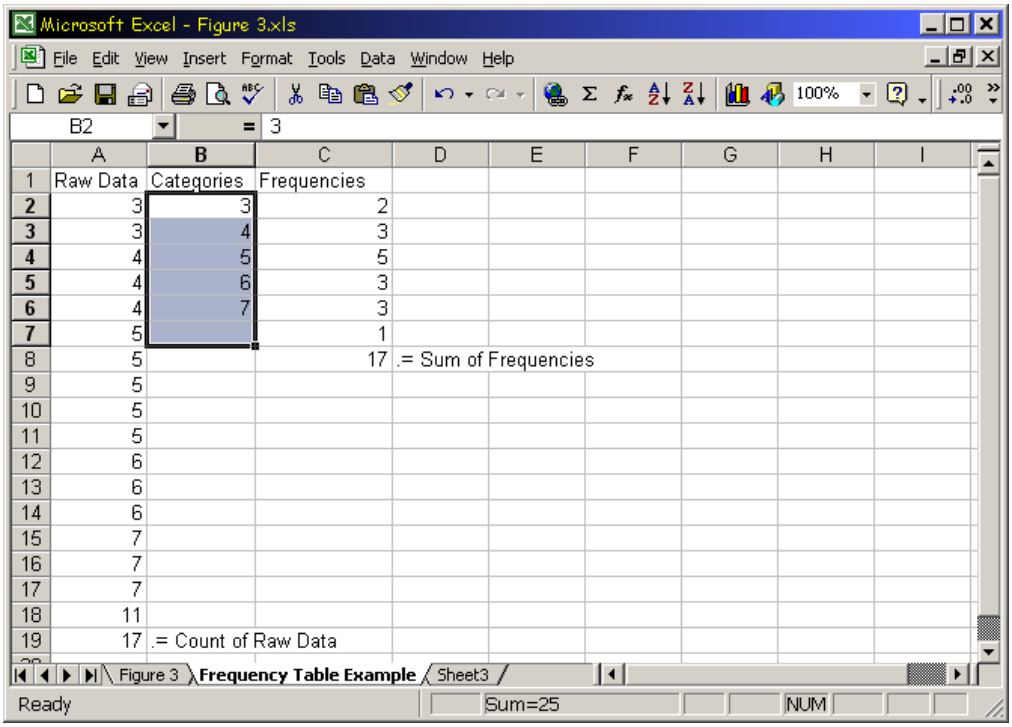


Figure 7.

Bar charts, histograms, and line graphs: Bar charts are graphs with categories (either discrete numerical data or true categorical data) along the X-axis and numerical values on the Y-axis. The bars indicate the value of each category. The only difference between a bar chart and a histogram is that the data for the bars in a histogram are frequencies. If the data used for the bars are not frequencies, then we have a bar chart. In a line chart, we use points instead of bars to indicate the values of the numerical variable, and each point is connected to the next by a line. One very important piece of information we often want to convey about some types of data is the slope of the relationship between two numerical variables. **It is absolutely critical that you always remember that the slope of a line in a line graph is essentially meaningless because the data on the X-axis are categorical**. We use scatter plots to look at the slope of the relationship between two variables, and this type of graph is covered later in the handout. In this course we are unlikely to encounter any situations where a line graph would be preferable to a bar graph. The only situation they are likely to be of any use is when you wish to imply that a data point somewhere between two categories has a numerical value somewhere between the numerical values of the two categories. As you can imagine, this would represent an ambiguous relationship based on essentially nonexistent data, and so you need to be very careful about using a line graph instead of a bar chart. One example of the reasonable use of a line chart would be a graph in which the mean weight of different age classes is plotted. The categorical data are the age classes (perhaps 5-year classes for humans). In this case, the categories are really arbitrary divisions of the continuous variable age. A bar chart would give the mean for all individuals in a five year period, but growth would look like a staircase in which individuals stay the same weight for five years and then suddenly shoot up to the next mean weight. A line graph, in which a straight line connects successive means, would imply that the mean weight changes in a more gradual fashion over time although it cannot be inferred that the change follows the straight line exactly.

Constructing the bar graph: To make a bar or line graph, you need to have the numerical data for the Y-axis in one column and another column with the labels you want to use for each category. These labels can be text (e.g., Female, Male, etc.), or they can be numbers indicating the range of discrete numerical categories (e.g., <3, 3-4, 4-5, 5-6, 6-7, >7). These data are easier to work with if they are in adjacent columns, with the numerical data first, followed by the category names.

To enter category labels for the frequency table you just made, we first have to specify the format of column D in your spreadsheet as "Text". This will ensure that XL's dumb auto-format function won't convert anything you enter to a date format. To do this, simply RIGHT CLICK on the column letter D, select the FORMAT CELLS option, then the NUMBER tab, then TEXT from the list on the LHS of the dialog box, and then OK. Now you can enter in column D the labels you want to use to identify each category for your frequency table. I suggest using the number sequence given in the paragraph above.

Next, select ONLY the numbers for the frequencies in column C by dragging the mouse through them. Then click on the INSERT menu and select the CHART option. The dialog box in Figure 8a (below) will appear. On the LHS of the dialog box you will see a list titled CHART TYPE that shows the different types of graphs you can choose from. Since we are making a histogram from our frequency table we want the vertical bar chart, which should be the first option. In XL, this graph is called a COLUMN CHART. You should see another option called a BAR CHART that has horizontal bars. In some situations a horizontal orientation may make sense, but we will always use a vertical orientation in our labs. On the RHS of the dialog box you will see a variety of different types of bar charts, including stacked bar charts and several examples of nasty psuedo-3D graphs. We want the simple bar chart option, which should be the first one in the list. Make sure this option is selected by clicking on it. There is a button underneath the examples of different graphs titled PRESS AND HOLD TO VIEW SAMPLE. This is a very useful feature because clicking on it will show what your graph will look like based on the range of cells you selected previously (Figure 8b). Click on it now to make sure your choice of graph is appropriate. If everything looks right, click on the NEXT button to get the dialog box in Figure 9a below.

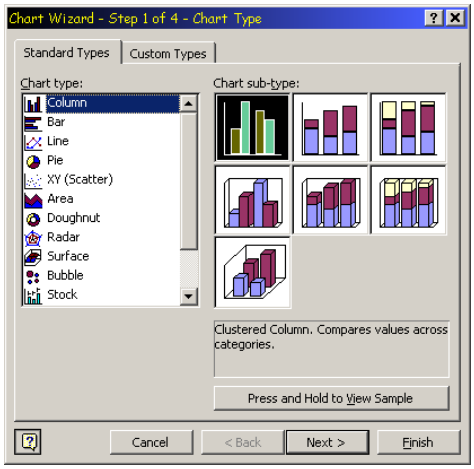


Figure 8a.

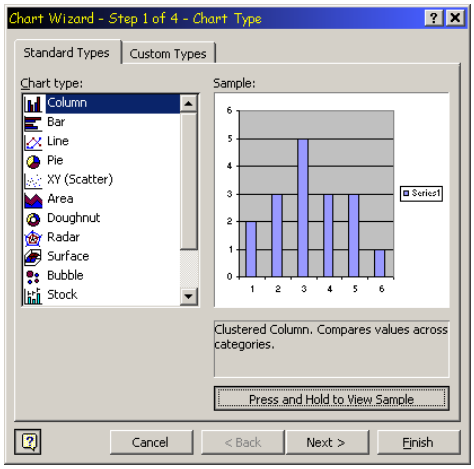


Figure 8b.

You shouldn't have to change anything in Figure 9a as long as the example graph displayed looks correct. We will add details like category labels and axis labels next. To add the category labels, click on the SERIES tab in Figure 9a to open the dialog box shown in Figure 9b. As indicated in this figure, there is a small graphic on the RHS of the input box labeled CATEGORY (X) AXIS LABELS. Click on this graphic and a small input box will open on your screen. Enter the range of cells containing your category labels by simply dragging your mouse through them, and then pressing the ENTER key. You should now see the category labels in the example graph in the dialog box as shown in Figure 9b. If everything appears correct, click NEXT to open the dialog box shown in Figure 10.

Figure 10 shows several options we can manipulate as we see fit. The first tab in the figure is labeled TITLES, and this is where we enter the CHART TITLE and axis labels. You can enter the examples shown in the figure for this exercise. The important thing to notice is that both axes have a label and units. For the X-axis the units are YEARS, and the Y-axis units are COUNTS. Even though the data in this example are imaginary, I have included units so that they make sense in our example. **You MUST always include units with your axis labels or suffer a severe deduction in grades**.

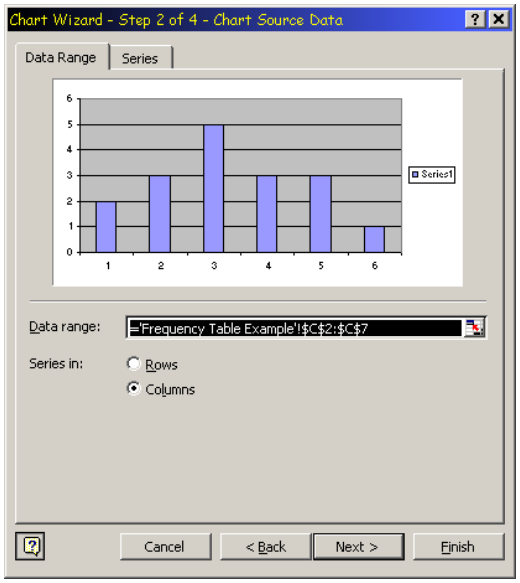


Figure 9a.

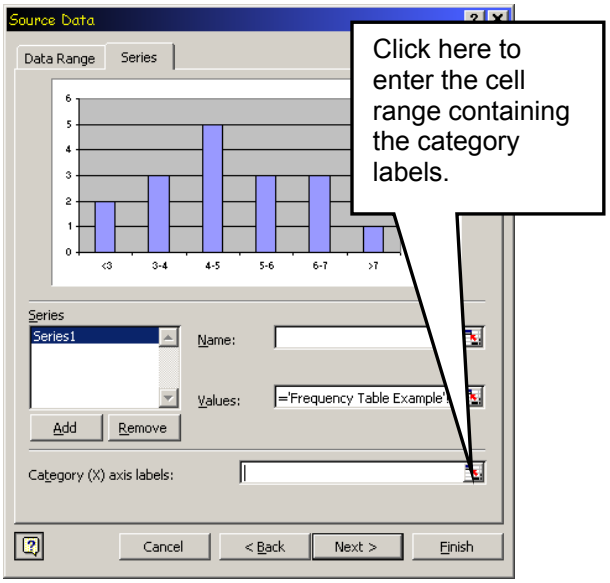


Figure 9b.

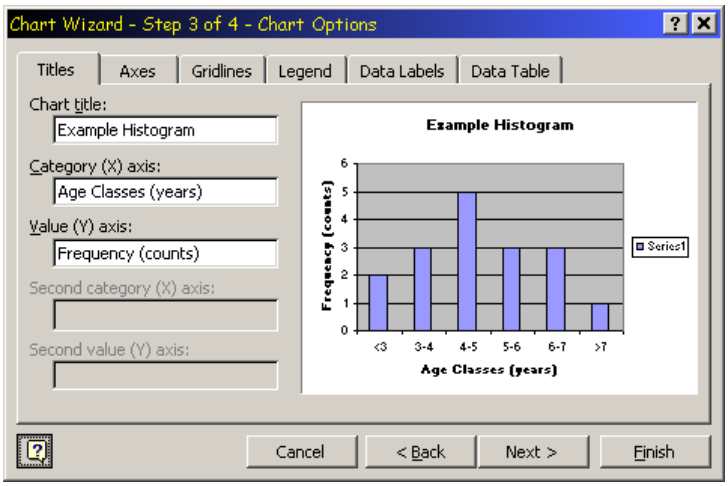


Figure 10.

I will now step through each of the other tabs shown in Figure 10 to explain some of the more important options you may change: AXES tab – usually there is no need to change anything here. GRIDLINES tab – you can select which grid lines are shown on the graph using the options on this tab. In scientific graphs, gridlines are rarely used unless absolutely necessary to enhance interpretation of the data. The reason is that they usually complicate the graph without any substantial improvement in interpretation. Your best choice here is to either accept the default (major grid lines for the Y-axis), or make sure no grid lines are selected. LEGEND tab – here you can decide whether or not a legend is shown, and if so, where on the graph the legend is displayed. This last option is not so important because the legend object can be dragged anywhere on the final graph. The first option though is important. In our example we only have one data series plotted so a legend is not required. Therefore, you should deselect the SHOW LEGEND option on this tab. If we had plotted two or more data series, the legend would be crucial for correct interpretation. The remaining two tabs (DATA LABELS and DATA TABLE) provide options that we will not use for our graphs.

When you have set all the options you think appropriate click NEXT and the last dialog box will appear asking you if you want the new graph to appear as an object on the current worksheet or as a new worksheet in the current workbook. The default is an object on the current worksheet, and this is usually the easiest to work with because it keeps the graph with the data used to create it. Select this option and click FINISH. You should see a graph similar to the one in Figure 11 below if you used all the options I suggested in this exercise.

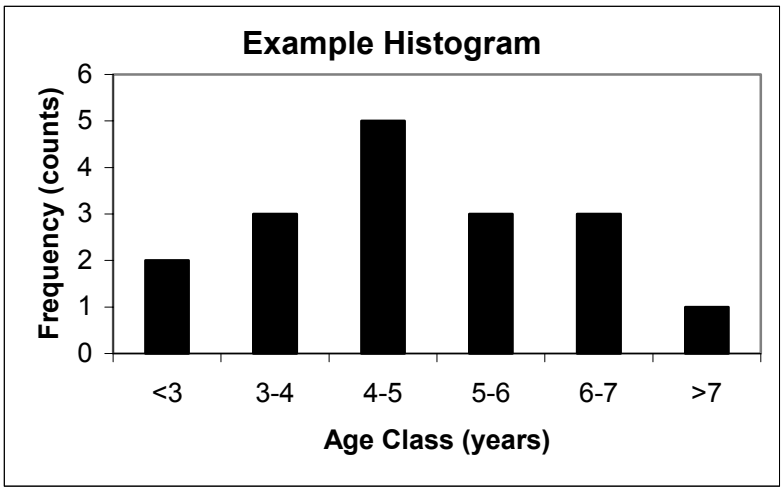


Figure 11.

Once the graph is complete there are still many options you can change by double clicking on any item in the graph itself. You can change font styles and sizes, or edit the text entered for titles and axis labels. In Figure 11, I changed the color of the bars from the default color to black by double clicking on the bar and changing the appropriate color options. I also did the same for the background color of the graph, which you should always do (using a white background enhances clarity). You can also change the size of the graph by clicking on its border and dragging the sizing handles, and you can copy the graph and paste it onto another worksheet or into another program like a word processor (which is what I did to make this document). The editing options that XL provides are extensive, but always remember that the most important feature of a scientific graph is that it conveys the message in the data simply, clearly, and concisely. Scientists (including myself) are not interested in how fancy you can make a graph.

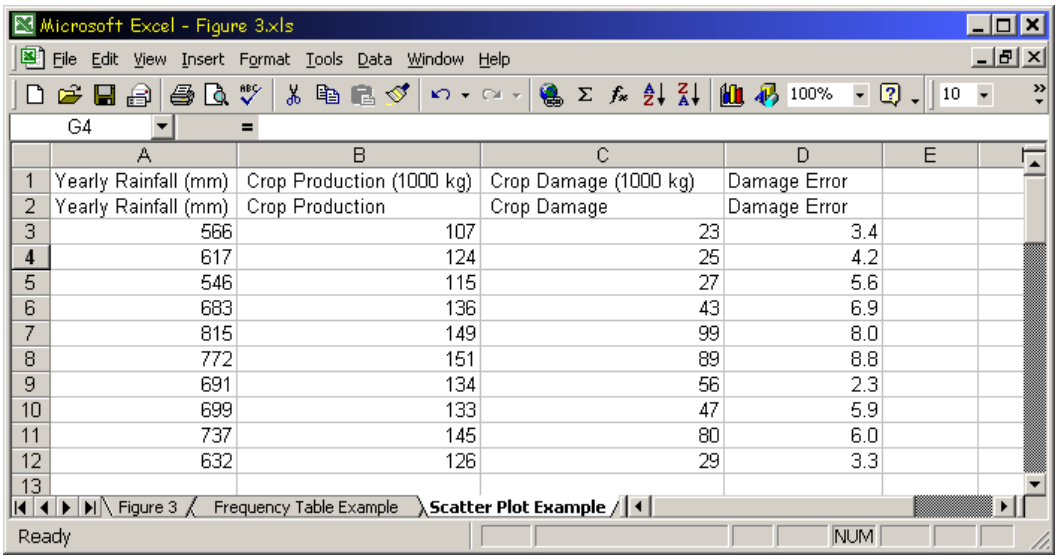
The procedure we used to make the histogram in Figure 11 is exactly the same for any bar graph or line graph. If we wanted a line graph instead of a bar graph, we would simply choose that option in Figure 8a.

Scatter plots

Scatter plots are used to show the relationship (if any) between two numerical variables by plotting data points on a Cartesian plane using pairs of coordinates (x, y) for each point. Sometimes it is assumed that one variable is measured without error and changes in that variable drive the response of the second variable. In this situation the driving variable is called the INDEPENDENT variable and is assigned to the X-axis, and the second variable is the RESPONSE or the DEPENDENT variable and is assigned to the Y-axis. If the independent variable does in fact drive the dependent variable, then a causative response exist. An example is the boiling temperature of pure water (dependent variable) and atmospheric pressure (independent variable). As atmospheric pressure decreases, the temperature at which pure water boils also decreases. Therefore, the boiling point of pure water DEPENDS on atmospheric pressure, but boiling water does not CAUSE atmospheric pressure to change (the volume of the atmosphere is so much larger than the volume of gas produced by boiling a beaker of water that the gas produced has no measurable effect on atmospheric pressure. This could be verified experimentally by using a moderately large pressure chamber and boiling a small beaker of water inside it. Setting the pressure to one value and then

boiling the beaker of water for a short time should show no change in pressure, but changing the pressure of the chamber will show a large effect on the boiling temperature of the water.

We will now create a scatter plot based on the data shown in Figure 12 below. Enter these data in a new workbook or new worksheet now. You will notice that there are two rows with column names in Figure 12 instead of just one. This is not a mistake and will be explained shortly.



	A	B	C	D	E
1	Yearly Rainfall (mm)	Crop Production (1000 kg)	Crop Damage (1000 kg)	Damage Error	
2	Yearly Rainfall (mm)	Crop Production	Crop Damage	Damage Error	
3	566	107	23	3.4	
4	617	124	25	4.2	
5	546	115	27	5.6	
6	683	136	43	6.9	
7	815	149	99	8.0	
8	772	151	89	8.8	
9	691	134	56	2.3	
10	699	133	47	5.9	
11	737	145	80	6.0	
12	632	126	29	3.3	
13					

Figure 12.

To create a scatter plot in XL, you need a minimum of two columns of data – one for each axis. **The easiest way to make a scatter plot in XL is to put the data in two adjacent columns, with the first (LHS) column containing the data for the X-axis, and the second (RHS) column containing the data for the Y-axis.** You can also plot more than one data series on the same scatter plot. This is frequently done to compare two data series together. In our example, we will compare the responses of Crop Production (1st Y-axis data) and Crop Damage (2nd Y-axis data) to Rainfall (X-axis). To achieve this, the data for the second data series must appear in a third column to the right of the data for the X-axis, as shown in Figure 12. **The rule to remember for making scatter plots is that the data for the X-axis must always be in a column to the left of the data for the Y-axis.** XL will use the column label that you enter as the text for the legend if you include one. If you want a shorter label for the legend, you need to include a second row for the column labels containing an abbreviated label, as shown in Figure 12.

If your data are set up according to the description above, then making the scatter plot itself is the essence of simplicity. Simply drag your mouse through the X- and Y-axis data, including the row for the column labels (only the abbreviated labels if you have them), select the CHART option from under the INSERT menu, select the XY (SCATTER) option from the dialog box (Figure 8a), select the specific type of scatter plot on the same dialog (normally we do not want XL to join the data points with a line, so the first option is usually the best), then click next and adjust the options of your choice in the same way as you did for the histogram example above. Let's do this now for the rainfall and crop production data we just entered. Select the first three columns of data (**NOT the column labeled DAMAGE ERROR**), including the row for the abbreviated column labels but not the extended column labels. Then follow the instructions above. When you do this, you should see a graph similar to the one shown in Figure 13 below.

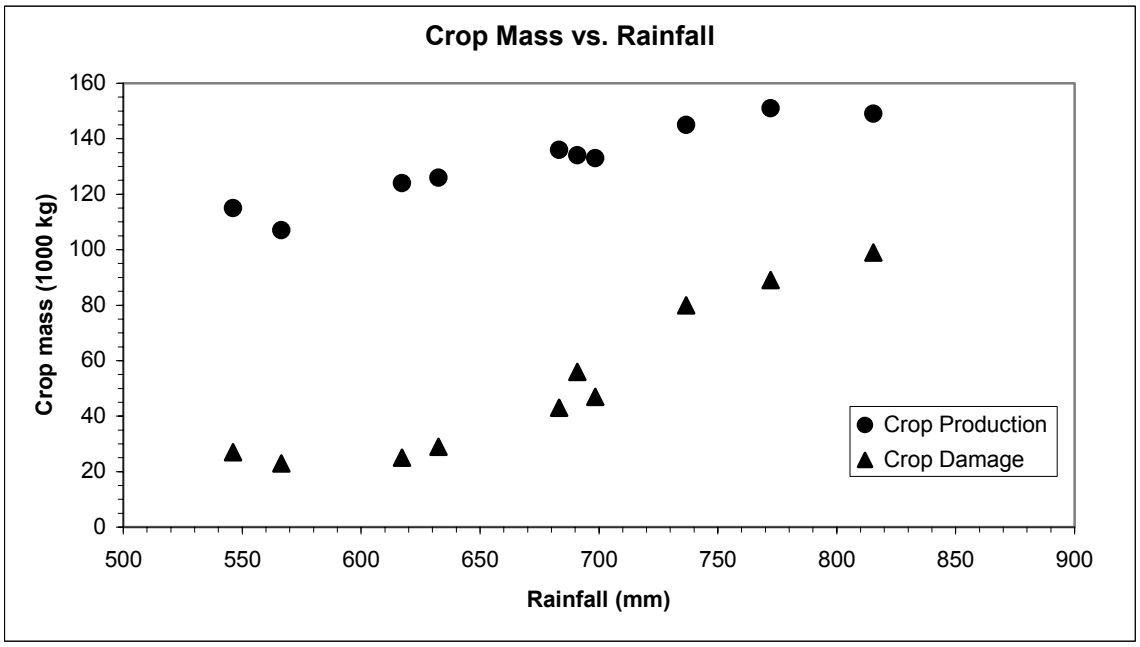


Figure 13.

After looking at Figure 13, you may be thinking that your graph only looks vaguely similar. All I have done is adjusted some of the more basic formatting options that are available by double clicking different elements of the graph in XL as described earlier. The legend can be resized and dragged anywhere on the graph. One of the most important set of options I have not mentioned yet control how the numbers and tick marks appear on the axes. To access these options, simply double click on the line that defines the axis you want to change and you should see the dialog box in Figure 14a appear.

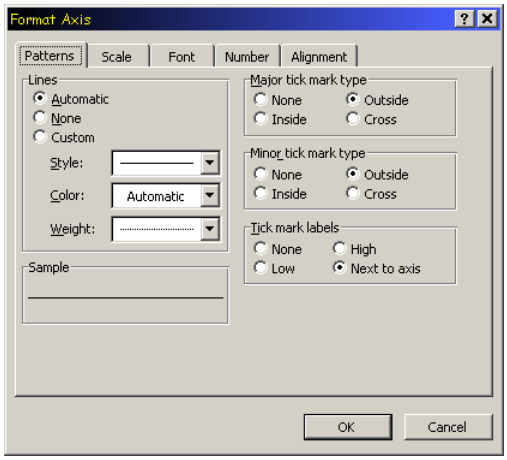


Figure 14a.

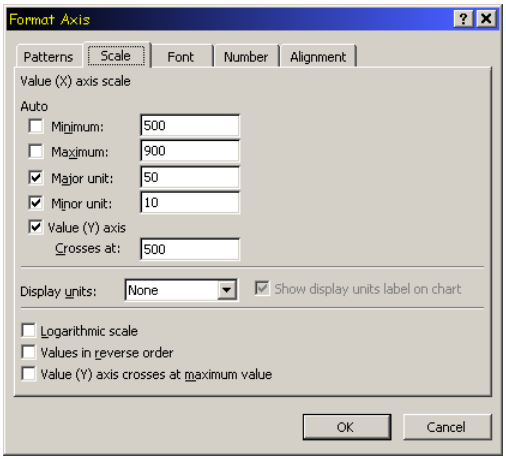


Figure 14b.

The first tab is titled PATTERNS, and this is where you can set visual characteristics of the lines used for the axis, as well as if and how axis tick marks are displayed. Usually we have no need to change the characteristics of the axis line, but we may want to change the tick marks. Tick marks are the little lines you see on the axis that identify where the different numerical values lie on the line. By default you will see what are called the MAJOR TICK MARKS, which are the ones that identify the points on the line corresponding to the numerical values that are displayed. However, you can also specify MINOR TICK MARKS, which are finer subdivisions of the numerical axis, but which do not have numbers printed near them. You can choose whether or not to display minor tick marks on this tab, but you should always have the major tick marks displayed.

The next tab labeled SCALE is shown in Figure 14b and is the most important tab for our purposes. Here you can specify the minimum and maximum numerical values for the axis, the units for the major tick marks (e.g., in the example shown here and in Figure 13, the major tick marks occur at 50-unit intervals), the units for the minor tick marks (10-unit intervals in this example), and the value at which the opposite axis crosses the one you are currently interested in. These settings are very important because they allow you to maximize the space available in the graph for showing the relationships in the data. If you look at Figure 13 again, you will notice that the X-axis does not start at zero, but instead at 500 mm and ends at 900 mm. This is perfectly OK. There is no rule that states that the numerical values for any axis must start and finish at specific values. The only thing that is required is that the range of values covered by an axis be clearly displayed. If we set the axis in Figure 13 to cover the range 0 mm to 900 mm, all our data points would be scrunched up in the RHS of the graph, which would obscure the relationships within and between each data series. By setting the X-axis to start at 500 mm, we have maximized the space available in the graph to show the important relationships in our data much more clearly. The best way to understand how these options work is to experiment with them. You can't damage anything doing this, and if you happen to mess something up, you can always use the UNDO function you learnt about last week to correct what you did.

Trend lines

Trend lines are used to summarize the relationship between two variables. There are many different types of trend lines, most of which are based on purely empirical relationships. These are the type we will use in this exercise. Later in this semester, we will use a mathematical model to fit a growth response curve to population growth data we will collect in an experiment on an aquatic plant. This type of model specifies an explicit functional relationship between the dependent and independent variable that has a clear biological interpretation. The empirical relationships we will use in the current lab do not specify any functional relationship and therefore have no biological interpretation. Their use is limited to providing an empirical summary of the relationship between variables and as a visual aid in illustrating the trend in our data. Both of these are legitimate uses of trend lines.

If you look carefully at Figure 13, or at the same figure you produced in XL, you will see that the data for Crop Production nearly forms a straight line, whereas the data for Crop Damage appears to be more curved. We are going to fit a straight line to the Crop Production data and a 3rd order polynomial curve to the Crop Damage data. To do so, click on any one of the data points for Crop Production. This should automatically select all the related data points. Then click on the CHART menu and select the ADD TRENDLINE option to open the dialog box shown in Figure 15a below. This dialog provides several different types of trend line we can choose from. For this example, select LINEAR and then click on the OPTIONS tab to open the dialog box shown in Figure 15b. In this dialog you can specify a name for the line that will appear in the legend of the figure. You can also specify if the line is to extend beyond the range of the data. For our purposes, that is never a good idea, so leave the FORECAST parameters set to zero. You can also tell XL to force the origin of the line through zero, but again, this is never a good idea for our purposes in this course, although there are legitimate situations where this may need to be done. The last two options in this dialog are quite useful. These allow you to include the equation for the line on the graph, along with a statistic called R-SQUARED (r^2). This statistic is a measure of how well the line fits the data. More formally, it is the proportion of variation in the Y-axis data that is accounted for by the line. The closer this statistic is to 1, the more closely the line fits the data. An r^2 of 1 means the data points lie perfectly on the line, irrespective of the shape of the line (i.e., straight, simple curve, complex curve). Select these two options then click OK and you should see a straight line and the equation and r^2 value as shown in Figure 16. Now we will fit a 3rd order polynomial to the Crop Damage data. Follow the exact same procedure as you did for the straight line except select the POLYNOMIAL option in the TRENDLINE dialog, and change the ORDER parameter from 2 to 3. You should now see the curve shown in Figure 16. Notice the high value for r^2 associated with this line.

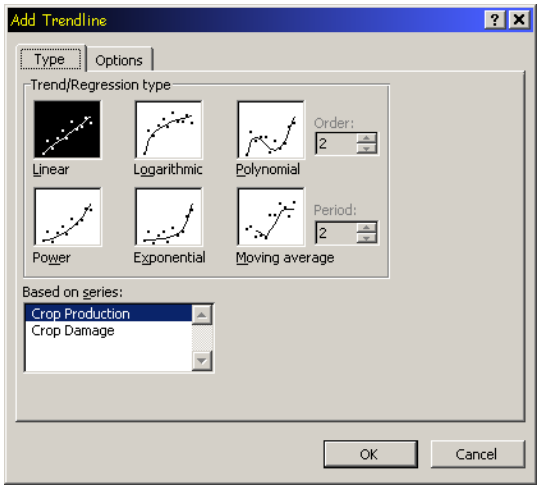


Figure 15a.

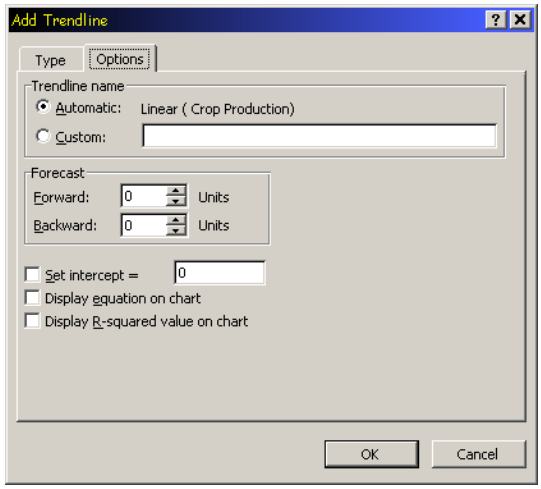


Figure 15b.

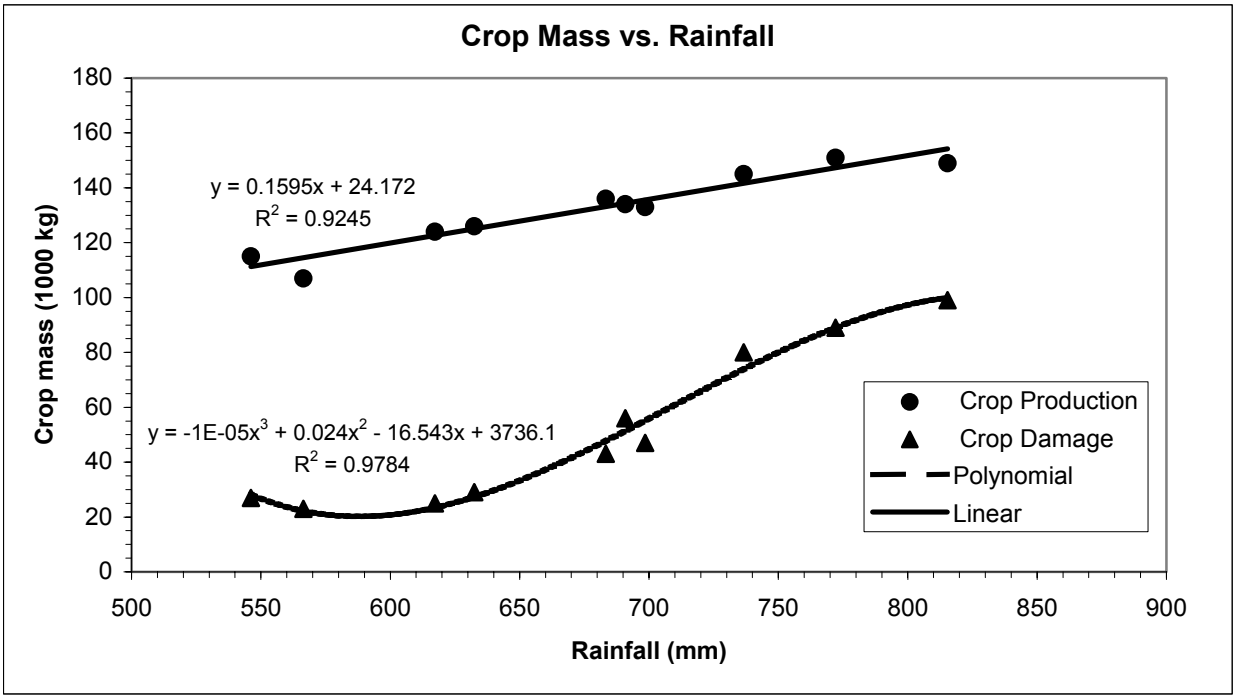


Figure 16.

Error bars

When each data point we plot are the means of several measurements, we need to include error bars to provide some indication of the variation associated with each sample the means are based on. For example, let's assume that the CROP DAMAGE data we plotted in Figure 16 are the means calculated from crop mass measured from five separate fields. We have calculated the standard error for each of these means and those values appear in the column labeled DAMAGE ERROR in our spreadsheet. We will now use these values for standard error to put error bars on each of the data points for CROP DAMAGE in Figure 16. This is quite straight forward, but you must remember that XL needs a column of numbers that specify the size of the error bars for each data point. Double click on any data point for CROP DAMAGE on the graph in XL and the dialog box shown in Figure 17 will open. Select the tab labeled Y ERROR BARS (these are the vertical error bars – you can include horizontal error bars too if appropriate) and click on the option under DISPLAY labeled BOTH. This simply means both plus and minus (upper and lower) error bars will be included. If this were a bar chart, we would only use PLUS error bars. Under the options labeled ERROR AMOUNT, select CUSTOM and click on the icon at the

RHS of the input box labeled "+". Now drag your mouse through the column of numbers labeled DAMAGE ERROR in the spreadsheet, but **DO NOT** include the column label – we only need the numbers. Then press ENTER. Next, do exactly the same for the input box labeled "–" in Figure 17. We use the same numbers for both the + and the – parts of the error bars because the standard error of the mean is symmetrical around the mean. When done, click OK and you should see the error bars in your graph as they appear in Figure 18 below.

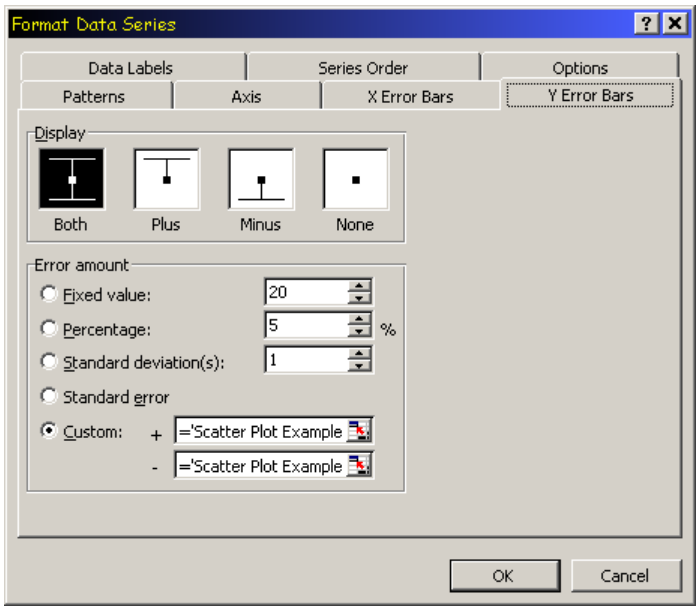


Figure 17.

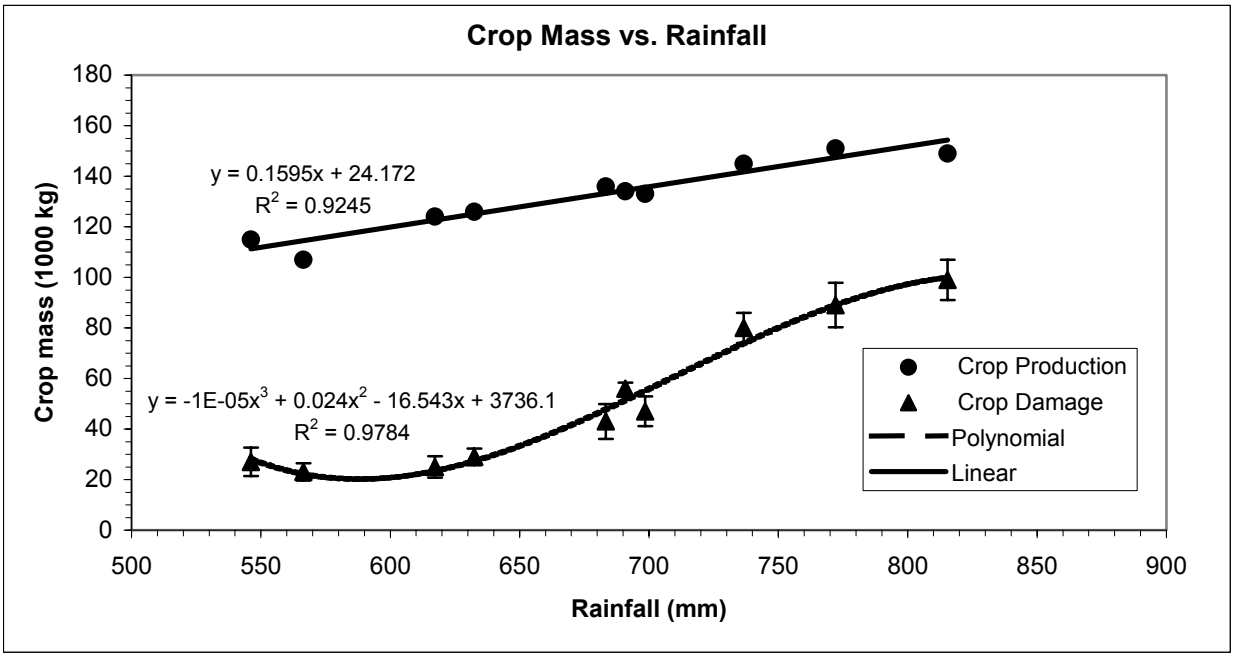


Figure 18.

Exercises:

You will find a data file with the lab handout on the course web page that you can use for the questions in this assignment. This will save you considerable time because you won't have to type all the data in by hand. The data file is write protected, meaning you won't be able to change it in any way. To use the data, open the data file in addition to a new workbook on your computer and copy and paste the data from the data file into your own workbook.

1) Make a frequency table using the data for Question 1 in the data file. These data are integer data, like you get when counting numbers of live individuals (you can't get a fraction of a live individual). Use the methodology described above to make the table (we will check to make sure you have used the correct functions). Label the two columns of the frequency table "Category Values" and "Frequencies". The values range from 1 to 17, suggesting logical category boundaries at one-unit intervals.

   Check your results by summing all of the frequencies using the SUM function, and counting the total number of data points using the COUNT function. Are these two numbers the same and why?

2) The data for Question 1 have some gaps and some low-frequency values. Regrouping the data will produce a smoother graph. Regroup the data by redoing the frequency table, but use the following category values: 0, 3, 6, 9, 15, and 18.

   Check your results again by using the SUM and COUNT functions.

3) Make a frequency table using the data for Question 3 in the data file. These data are the individual body mass from a sample of fish, which is a continuous variable. There is a wide range in body mass, from hatchlings to large adults, so the data ranges from 0.010 grams to 12.045 grams. Make two frequency tables, one with one-gram category values and another with three-gram category values.

   Check your results again by using the SUM and COUNT functions.

4) Make a histogram from the frequency table in Question 2. Don't forget to include labels for each category on the X-axis, and make sure your graph is formatted for clarity. If you are not sure how to do this, then review the discussion on formatting graphs in this handout and use the example graphs as templates for your own formatting. Remember, penalties will be applied for any of the errors/deficiencies discussed in this handout.

5) Make a bar chart from the data for Question 5 in the data file. Make sure you include the error bars (upper error bars only for bar charts) and appropriate formatting. Next, make a line graph using the same data, but this time include both upper and lower error bars.

6) Look at the data for Question 6 in the data file. These data are from an experimental field over a period of 10 years. The variables include annual rainfall (in mm) measured at the field, the total crop production (in metric tons per hectare), the error associated with the total crop production measurement, and the portion of the total crop that was damaged by insects to the point that it could not be sold (in metric tons per hectare). Total crop tonnage = useable tonnage + damaged tonnage. **Note: these data are different to the data used earlier in this handout!**

   Graph the relationship between Total Crop and Rainfall with error bars for Total Crop. Fit a linear trend line and present the equation for the line and the r^2.

   Graph the relationship between Rainfall and both Total Crop and Damaged Crop on the same

graph. Fit trend lines and present equations for each line and the r^2's. Hint: a linear trend line is not the best choice for the Damaged Crop data. Choose a trend line with a better fit.

7) Planarians will lose weight when they are starving and, of course, will gain weight when they receive sufficient food. You have a series of aquarium tanks to which you will add some planaria and a constant supply of food. You also have a model of planarian growth in these tanks. In this model, y is the predicted average total mass (g) of planarians per tank after two months in the tanks, and x is the total average mass of planarians (g) initially added to the experimental tanks. Graph the following equation (model) that relates these two variables:

$$y = \frac{x^2 - x}{x(x/12)}$$

Graph y over the range of x values from 0 to 10 grams added to the tank. From your graph (or from the values you calculated), find the point where $y = x$, i.e., the point where both y and x have the same numerical value. This is called the break-even point, where the total weight of planaria at the end of two months is equal to the total weight of planaria initially added to the tank.

Hint: you have to choose the x values to use. Begin by creating a series of integer values from 1 through 10 in a column labeled x using the INSERT SERIES option. Then use these numbers as the basis for calculating your corresponding y values. If the graph you construct indicates a large change in y values from one x value to the next, you may need to insert some intermediate x values to better represent the change in y over that range.

8) Using the data from the table below, make a chart that compares the effect of diet for each sex (you will have to decide which type of graph is best for these data). Be sure to include the error term and show all of the data on one chart. This will require entering the data in a different order than shown below.

Draw some conclusions about the effect of both sex and diet (and the interaction between these factors) based on the chart. **NOTE:** an *interaction* means that the response of mass due to one factor (e.g. diet) is modified by the different levels of the other factor (e.g. sex).

Sex	Diet	Adult Mass (grams)	Standard Error (grams)
Male	High Fat	492.4	17.6
	High Protein	532.3	21.4
	High Carbohydrate	481.2	12.3
Female	High Fat	427.2	13.8
	High Protein	477.8	21.1
	High Carbohydrate	239.1	10.1

Before you submit your assignment, make sure you check the following items:

1) YOU MUST NOT SUBMIT A GROUP ASSIGNMENT FOR THIS LAB. SEE THE BEGINNING OF THIS HANDOUT FOR DETAILS.
2) Make sure the layout of your spreadsheets is clear and logical, the answers to the questions are in the correct order, the answers are clearly identified on the spreadsheet, you have formatted your graphs for simplicity and clarity, and that you have used the appropriate number of decimal places in your answers (no more than 2 or 3 decimal places for this lab – use the CELL FORMAT option to set the number of decimal places).

HOW TO SUBMIT YOUR ASSIGNMENT

You will submit this assignment by emailing the XL file as an attachment to Saudat Adamson (our TA). Saudat's email address is:

sadamson@mytsu.tnstate.edu

YOU MUST USE YOUR MYTSU EMAIL ACCOUNT TO SEND THE FILE TO SAUDAT. THE REASONS FOR THIS ARE EXPLAINED IN THE COURSE SYLLABUS AND WILL NOT BE REPEATED HERE.

NAMING YOUR COMPUTER FILE

To expedite the significant task of grading these assignments, we request that you adhere to the following convention when naming the computer files you send us:

Lab X YyyyZzzz.xls

Where X = the lab number, the "y's" are the first 4 letters of your last name, and the "z's" are the first 4 letters of your first name. **Please pay close attention to the spaces and the capitalization**. This may seem pedantic, but it really does help us enormously. Using this format, I would name my file for this particular lab as "Lab 3 PeteAndr.xls".

Written by Dr. P. Ganter and Dr. A. Peterson, TSU.