Appendix E: Graphing Data

You will often make scatter diagrams and line graphs to illustrate the data that you collect. Scatter diagrams are often used to show the relationship between two variables. For example, in an absorbance spectrum, the variables would be the wavelength of light and the amount of light absorbed. Although this data is recorded in a table, a scatter diagram can illustrate in a more visual way the relationship between the two data sets: absorbance and wavelength.

The first consideration for a graph, is whether the graph is needed, and if so, the type of graph to be used. For accuracy, a well constructed table of data usually gives more information than a graph. The values obtained and their variability are readily apparent in a table, and interpolation (reading the graph) is unnecessary. For visual impact, however, nothing is better than a graphic display.

There are a variety of graph types to be chosen from; e.g. line graphs, bar graphs and pie graphs. Each of these has its own characteristics and subdivisions. One also has to decide upon singular or multiple graphs, two-dimensional or three dimensional displays, presence or absence of error bars, and the aesthetics of the display. The latter include such details as legend bars, axes labels, titles and selection of the symbols to represent data, and patterns for bar graphs.

THE BASICS

Perhaps the number one rule for graphic display has to do with the axes. Given a two-dimensional graph, with two values (x and y), which value is x and which is y? The answer is always the same - the KNOWN value (the "independent variable") is always the ordinate (x) value. The value that is MEASURED (the "dependent variable") is the abscissa (y) value. For example, in a standard curve of absorption in spectrophotometry, the known concentrations of the standards are placed on the x axis, while the measured absorbance would be on the y.

The scales should always be arranged with the lowest value on the left of the x axis, and the lowest value at the bottom of the y value. The range of each scale should be determined by the lowest and higest value of your data, with the scale rounded to the nearest tenth, hundredth, thousandth, etc. That is, if the data ranges from 12 to 93, the scale should be from 10 to 100. It is not necessary to always range from 0, unless you wish to demonstrate the relationship of the data to this value (as for example in a Lineweaver-Burke plot of enzyme kinetics, or a Spectrophotometric standard curve).

The number of integrals placed on the graph will be determined by the point you wish to make, but in general, one should use about ten divisions of the scale. For our range of 12 to 93, an appropriate scale would be from 0 to 100, with an integral of 10. Placing smaller integrals on the scale does not convey more information, but merely adds a lot of confusing marks to the graph. The user can estimate the values of 12 and 93 from such a scale without having every possible value ticked off on a scale.

An important rule of scale deals with multiple graphs drawn separately. If the graphs are to be compared, the scales MUST REMAIN CONSTANT. Nothing is more disconcerting than to be shown two graphs with varying axis values and being asked to compare the two. It would be better to merely tabulate the data than to graphically present it.

LINE GRAPH VS. BAR GRAPH OR PIE GRAPH

If the presentation is to highlight various data as a percent of the total data, then a pie graph is ideal. Pie graphs might be used for example to demonstrate the composition of the white cell differential count. They are the most often used graph type for business use, particularly in displaying budget details.

Pie Graphs are circular presentations which are drawn by summing your data and computing the percent of the total for each data entry. These percent values are then converted to portions of a circle (by multiplying the percent by 360 °) and drawing the appropriate arc of a circle to represent the percent. By connecting the arc to the center point of the circle, the pie is divided into wedges, the sizes of which demonstrate the relative size of the data to the total. If one or more wedges are to be highlighted, that wedge can be drawn slightly out of the perimeter of the circle for what is referred to as an exploded view.

More typical of data presented in cell biology, however, are the line graph and the bar graph. There is no hard and fast rule for choosing between these graph types, except where the data is non-continuous. Then, a bar graph must be used. In general, line graphs are used to demonstrate data which is related on a continuous scale, whereas bar graphs are used to demonstrate discontinuous or interval data.

Distance is an example of a continuous variable. We may choose to collect the data in 1 mm intervals, or 1 cm. The range is continuous from 0 to the limit of our measurements. That is we may wish to measure the value at 1 mm, or 1.2 mm or 1.23 mm or 1.23445 mm. The important point is that the 2 mm position is 2x the point at 1 mm. There is a linear relationship between the values to be placed on the x axis. Therefore a linear graph would be appropriate, with the dots connected by a single line. If we choose to ignore the 1.2 and 1.23 and round these down to a value of 1, then a bar graph would be more appropriate. This latter technique (dividing the data in appropriate intervals and plotting as a bar graph) is known as a Histogram.

Having decided that the data has been collected as a continuous series, and that the data will be plotted on a linear graph, there are still decisions to be made. Should the data be placed on the graph as individual points with no lines connecting them (a scatter diagram)? Should a line be drawn between the points (known as a Dot-to-Dot)? Should the points be plotted, but curve smoothing be applied? If the latter, what type of smoothing?

There are many algorithms for curve fitting, with the two most commonly used being linear regression and polynomial regression. It is important to decide BEFORE graphing the data, which of these is appropriate.

Linear regression is used when there is good reason to suspect a linear relationship within the data (as for example in a spectrophotometric standard following the Beer-Lambert law). In general, the y value can be calculated from the equation for a straight line, $y = mx + b$, where m is the slope and b is the y- intercept.

Computer programs for this can be very misleading. Any set of data can be entered into a program to calculate and plot linear regression. It is important that there be a valid reason for supposing linearity before using this function, however. It is important that use of linear regression must be warranted by the relationship within the data, not by the individual drawing the graph.

If the data collected involves two or more sets of data having a common x axis, but varying y axes (or values), then a multiple graph may be used. The rules for graphing apply to each set of data, with the following provision: make sure that you have a figure legend that identifies which data points belong to each line.

A. SETTING UP A GRAPH

To make a scatter diagram, simply plot ordered pairs of number on graph paper. The horizontal line on the graph paper is identified as the x-axis (or abscissa) and the vertical line is the y-axis (or ordinate**). Each axis is labeled with an appropriate unit of measurement. Each increment of these lines represents the same amount of the measurement.** For example, if you are drawing a scatter diagram of an absorbance spectrum and if one square of the x-axis represents 10 nm of wavelength of light, each other square also represents 10 nm of the wavelength of light along the x-axis. (This is called a “linear scale”.) Similarly, if each For square of the y-axis represents 0.0 10 absorbance unit of light absorbed, each other square also represents 0.010 absorbance unit along the y-axis. In other words, each axis has a consistent scale, even though the two axes do not use the same linear scale.

How do you know which variable is to be on the x-axis, and which is to be on the y-axis? The x-axis should be the **independent variable,** or the parameter that you selected to study. In an absorbance spectrum, it would be the wavelengths that you chose to measure the absorbance of light at. The y-axis should be the **dependent variable,** or the data that you got from your measurement. In an absorbance spectrum, the dependent variable would be the absorbance that you measured at the wavelength of light that you selected. (You can think of it as the values for dependent variables measured in your data sets depend on which independent variables you chose.)

B. APPROXIMATING A "BEST FIT" LINE FOR A SCATTER DIAGRAM

If you look at your scatter plot and the middle points on the graph are close to forming a straight line, it is reasonable to conclude that the relationship between the independent and dependent variables is linear. The straight line defines this linear relationship. If, for example, you have studied the effects of a particular fertilizer on fruit production in apple trees, your independent variable is the amount of fertilizer applied, while your dependent variable is the weight of apples harvested at each level of fertilizer application. If you were to plot these results, you would have a line that tells you exactly how much fruit you could expect from apple trees at a given level of fertilizer application.

It is unlikely that your data points in the apple experiment described above would all lie on a perfect line, due to random variation in conditions and trees. To be able to get the best approximation of fruit yield per fertilizer application, you would want to know the average result. The best way to do this is to draw the "**best-fit straight line**" though your scatter plot. The most valid best fit straight lines that illustrate a linear relationship are determined by using a type of statistical analysis call linear regression analysis, which you will learn to use in Lab Exercise 6. For this exercise, practice approximating a best fit line in the following way:

- If your data points seem to form a basically straight line after you have made a scatter diagram illustrating your data, place a straight edge on your graph along the data points. Move the straight edge until it is as close as possible to as many points as possible, and draw a line along the straight edge. There should be approximately the same number of data points on each side of your line, and the line should minimize the distance of the data points from the line as possible. Notice that your line may or may not pass through any particular plotted data point.

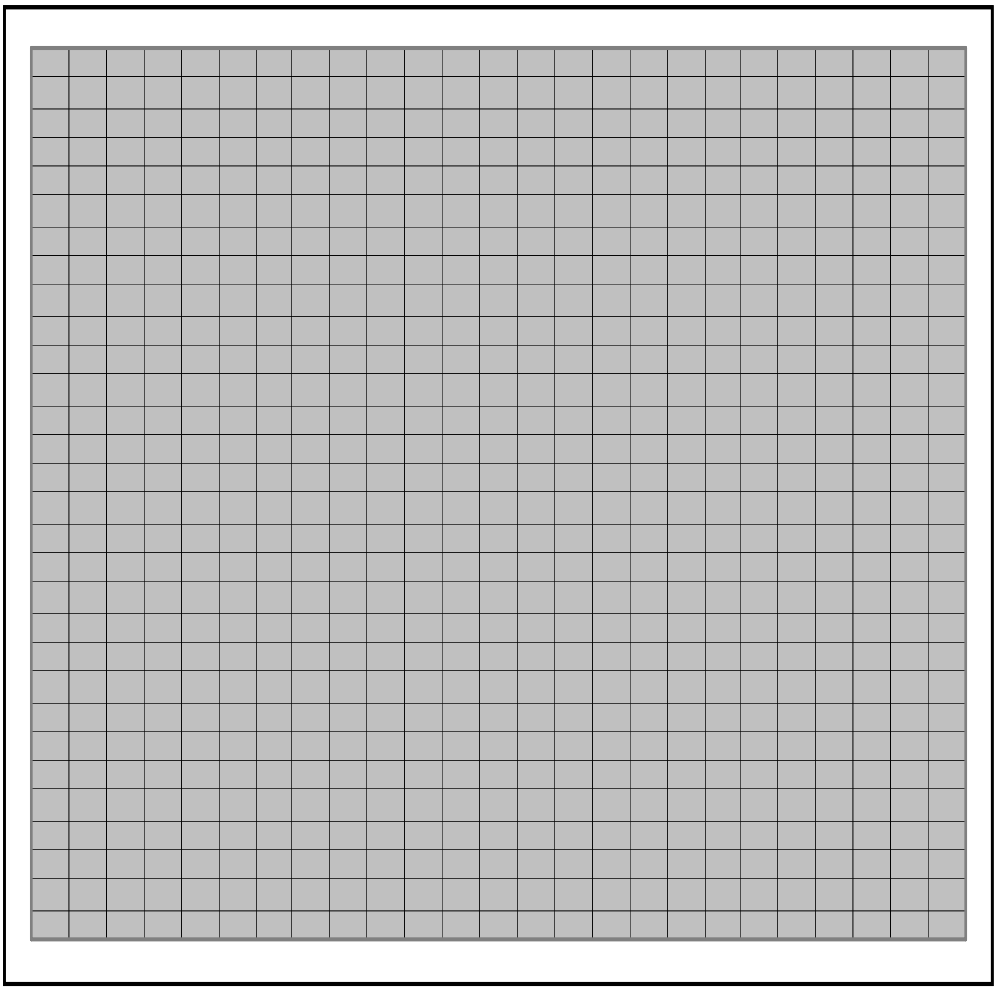
- If your data points seem to form a curve rather than a straight line, there is a nonlinear relationship between your dependent and independent variables. You will have to approximate your line by drawing the curve freehand rather than by using a straight-edge. Try for the same effect, however– do not connect-the-dots. The line should be a smooth curve, which may or may not pass through any particular data point.

- Often, you will be graphing data that illustrates a relationship that is generally linear but then the linearity breaks down at the extremes of changing conditions. In these cases, some of the data points on your graph will form a nearly straight line. This is an indication that at a certain point, the data that you collected no longer have a linear relationship. However, the information in the linear part of the graph may still be valuable. In that case, place your straight edge so that it is as close as possible to as many points that lie along the *linear* part of your graph as possible.

Table 1. Checklist for preparing a scatter plot and linear plot

1. **Your graph should always be given a brief title to explain what relationship you are studying.**
2. **Plan how to mark off the units of measurement on each axis so that you completed graph will nearly fill the page.**
3. **Both axes should be clearly labeled and marked with appropriate units of measurement.**
4. **Both axes should have a linear scale, meaning that the same increments are consistently the same distance apart. The size of the increments on one axis does not have to be the same as that of the other, but they must both be a linear scale.**
5. **The x-axis should be the independent variable. This is the variable that the experimenter chooses and can change. The y-axis should have the dependant variable, or the one that the experimenter observes as he chooses the independent variable to measure.**
6. **You may draw a line on your scatter plot to better illustrate any pattern that is revealed. If you find a linear relationship between your independent and dependent variable, draw a best-fit straight line through the points that are consistent with the linear relationship. If there is no linear relationship, you can leave your scatter plot as is, or draw curved lines between your data if you wish.**

7. **Plan how to mark off the units of measurement on each axis so that you completed graph will nearly fill the page.**
8. **Both axes should be clearly labeled and marked with appropriate units of measurement.**
9. **Both axes should have a linear scale, meaning that the same increments are consistently the same distance apart. The size of the increments on one axis does not have to be the same as that of the other, but they must both be a linear scale.**
10. **The x-axis should be the independent variable. This is the variable that the experimenter chooses and can change. The y-axis should have the dependant variable, or the one that the experimenter observes as he chooses the independent variable to measure.**
11. **You may draw a line on your scatter plot to better illustrate any pattern that is revealed. If you find a linear relationship between your independent and dependent variable, draw a best-fit straight line through the points that are consistent with the linear relationship. If there is no linear relationship, you can leave your scatter plot as is, or draw curved lines between your data if you wish.**

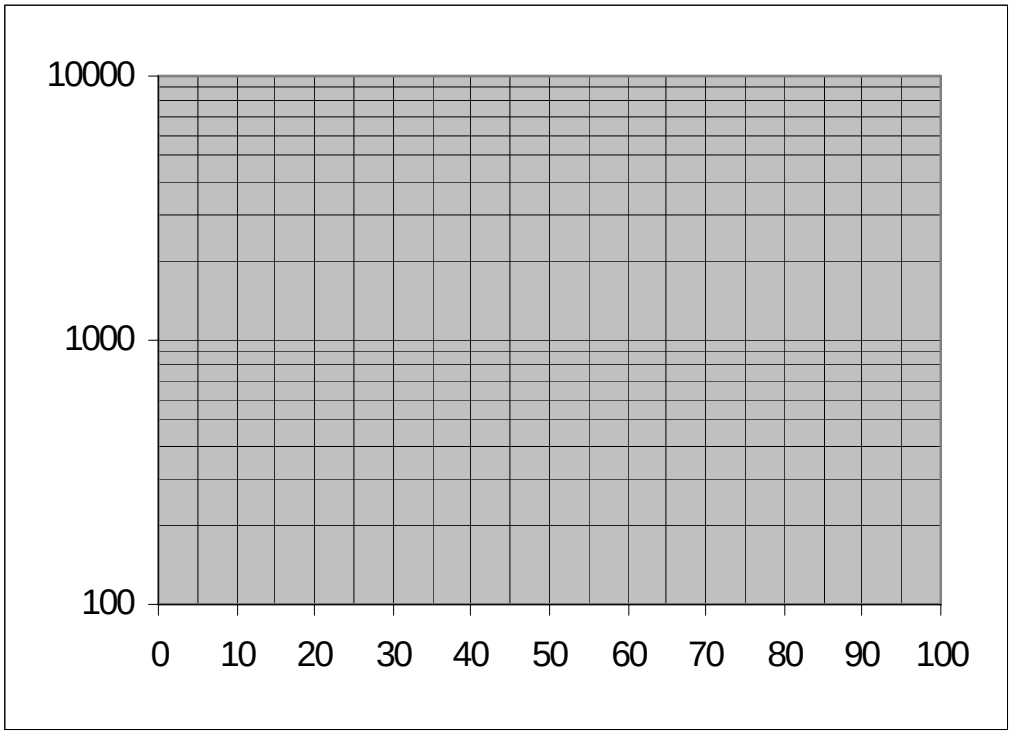


C. GRAPHING A SEMILOG PLOT

Your data may not have a linear relationship, in which case a straight line cannot be draw on for your data. Often in the biological sciences, the relationship is exponential rather than linear. That is, one value doubles for each increase in the other value. For example, each time a cell divides, the number of cells is divided. This means that if you repeatedly measure the cell count of a culture over a given interval of time, the cell count will not rise linearly with time, but rather exponentially with time. If you graph this relationship on semilog paper, the line will be linear.

On semilog paper, the X-axis is linear (each increment is spaced equally and represents an equal unit of measurement), but the Y-axis is exponential (each increment is NOT spaced equally and does NOT equal the same unit of measurement). The hard part of using semilog paper is deciding what units lie along the Y-axis.

You will notice that there are heavier tracing and lighter tracings of the Y-axis grids. The heavy tracing represent a "decade": the light tracings within the decade are assigned numbers that are equally spaced. For example, the units within a decade would be 1, 2, 3, 4, 5, 6, 7, 8, 9, to 10. Or, depending on what your data is, the units within a decade would be 100, 200, 300, 400, 500, 600, 700, 800, 900, to 1,000. The important thing to notice is that the decade ABOVE the first decade rises 10 times as fast. For example, if your first decade is 1, 2, 3, 4, 5, 6, 7, 8, 9, to 10, the the second decade is 10, 20, 30, 40, 50, 60, 70, 80, 90, to 100, and the third decade would be 100, 200, 300, 400, 500, 600, 700, 800, 900, to 1,000.



D. GRAPHING WITH MICROSOFT EXCEL

You may find an excellent animated tutorial for the use of Microsoft Excel spreadsheets and graphing at the website www.geospiza.com. Click on "Education" at the top left and follow the menu. The steps used to graph with Excel are listed here for your quick reference:

1. Enter data in a table. Select the data and column headings for your graph.
2. Select the CHART ICON (upper right, tricolor bar graph).
3. Select the XY(SCATTER) "Chart Type". Select the "Chart Subtype". Click on NEXT.
4. Make sure that the correct data appear in the data range box in the "Data Range" box. Under "Series", make sure that the X and Y value(s) correspond to the correct columns. Click on NEXT.
5. Under "Chart Options", add a title for the chart and titles for the X and Y axes. Be sure to add the correct units of measurement. Click on NEXT.
6. If you want to, you can edit the title as follows: double click on the graph title. Choose the font and click on OK. You can also edit the titles for the X or Y axes by double clicking on them from the graph. For example, you can change the alignment or font of the title. Click on OK.
7. You can also change the range of your X and Y axes by double clicking on the axis line. Under "Scale", you can set the minimum value so that it's closer to the earliest data point and set the maximum so that it's closer to the last data point. Be sure to Change the "Value the axis crosses at" to be consistent with the "minimum" value. Click on OK.
8. You can add grid lines to your graph by clicking on the graph and selecting "Chart Options". Make your selections under "gridlines" and click OK.
9. Under "Chart Location", you can select AS NEW SHEET if you want the graph to appear separately, and AS OBJECT IN if you want the graph to appear with you data table ("Sheet 1"). Click on FINISH.

E. MAKING A SEMILOG PLOT WITH MICROSOFT EXCEL.

Follow the instructions for "Graphing with Microsoft Excel" (above) with the following modification. After Step #8, you will need to edit the scale for the Y axis to change it to log scale.

1. Double click on the Y-axis of your graph. Under "Scale", set you minimum and maximum units. Set the "Major unit" and "Minor unit" to 10, and set the "Value (X) axis crosses at" to 1. Select the logarithmic scale and click on OK.

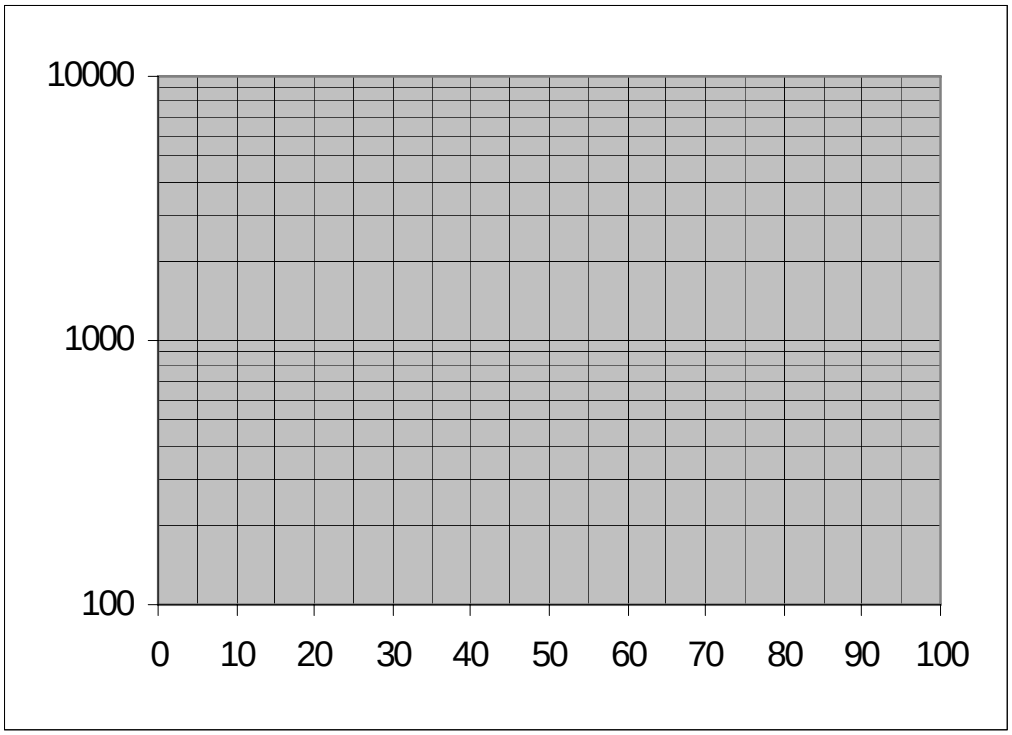
F. LINEAR REGRESSION WITH MICROSOFT EXCEL

You will often be drawing graphs of standard curves which will be used later for extrapolation of raw data of measurements. You can do your extrapolation manually from your graph, or you may derive an equation for your best-fit straight line and use the equation for your determinations. The equation

$$\mathbf{Y = mX + b}, \quad \text{where } \mathbf{m} \text{ stands for the slope and } \mathbf{b} \text{ stands for the Y-intercept}$$

can be derived manually by inspection of a graph, using the methods that you learned in algebra classes. Alternatively, you can use a scientific calculator or a spreadsheet program to derive the equation automatically using a program called "linear regression". The following instructions will show you how to use Microsoft Excel to derive an equation from your data.

1. Make a graph, following instructions above.
2. Make Linear Regression Line by selecting the graph
3. Select add trendline from chart menu
4. Select Linear from option and do for Series 1
5. Compare y-intercept, and slope from line. Select insert function and chose SLOPE select x and y coordinates for Line. *remember that you cannot do slope with a 0 so enter in .01 for that value or chose another one*
6. Select insert function and chose INTERCEPT select x and y coordinates for the line.
7. The Correlation Coefficient or "R" value is a measure of how well your data agrees with a linear relationship. A closer the number is to one, the better your data is.


10000
1000
100
0
10
20
30
40
50
60
70
80
90
100