Graphical Analysis Lab 1  Physics 227L

Introduction to Scientific and Engineering Graphing and Analysis of Data using Spreadsheets

Computer spreadsheets are very powerful tools that are widely used in Business, Science, and Engineering to perform calculations and record, present, and analyze data, as well as perform a variety of more sophisticated calculations that can use a number of special functions. One of the virtues of spreadsheets is their relative ease of use compared to programs like Matlab and Mathematica or programming languages like C, or Fortran. In fact, spreadsheets have many of the functions of programming languages as built in functions and can do many of the same things. You can also make your own custom functions that can be called in spreadsheets if you need to, and Microsoft Visual Basic for Excel is a full-fledged programming language (albeit with a bit of a learning curve).

We will be using Microsoft Excel for this lab. However, other spreadsheet programs are very similar and have much the same functionality. For example, OpenOffice is a free suite of software programs with Word Processor, Spreadsheet, Presentation (Power Point), Drawing, Math typesetting, and Data Base software.  http://www.openoffice.us.com/openoffice/free-open-office-download-yahoo.php?pk=845397  The Open Office spreadsheet is called Calc.

Don’t worry if you have never used Excel before, we will learn how to use the basic features of it in this lab. You should become fairly proficient in using Excel by the end of the semester, since we will use it to analyze and plot the data for almost all of the experiments. Learn how to use Excel for yourself. Don’t let your partner be the one who always makes the spreadsheet, you should take turns working on it.

Linear graphs

The basic linear equation \( y = mx + b \) is the simplest and most widely used relation to fit data to. Remember that \( m \) is the slope and \( b \) is the \( y \) intercept of the function. \( x \) is called the independent variable, and \( y \) the dependent variable, since \( y \) depends on \( x \). (of course we could have taken it the other way around, but there is a reason for how the independent and dependent variables are chosen). The independent variable is the quantity that you directly vary at will in your experiment, like change the wavelength of the light source. The dependent variable is the new value of the quantity that you are measuring that responds to or is dependent on the independent variable, like the stopping voltage in the photoelectric effect. The independent variable is usually put on the horizontal axis and the dependent variable on the vertical axis.

Example: Let’s take some data from a lab experiment on the photoelectric effect that we will perform later in the semester. The student measured the stopping voltage (which is the voltage required to make the photocurrent go to zero) as a function of the wavelength of the LED used and obtained the following data table.
Open an Excel spreadsheet and enter (type) the data (including labels and units) into two columns. It is good practice to leave a few blank rows at the top of the spreadsheet to insert information about the spreadsheet, column labels, and column units. Note the information at the top of the table about who took the data and when, also the instrument box number. You should get in the habit of doing the same kind of thing for all your experiments.

After you have typed the data in the columns, make a simple scatter plot of the data. To do this, select both columns of data, which will then have a box around it with light blue background highlight, as shown below. Click on the insert tab at the top of the spreadsheet and select the scatter plot with no lines connecting the data points, as shown below.
The graph should look similar to the one shown above. What do you notice about this graph? There are several problems with it. Excel’s biggest user base is business, so default graph formats are mostly setup for that purpose. As a result, you have to do a lot of formatting work to get a graph into a proper scientific or engineering format. You can make a graph template to avoid tedious and repetitive formatting steps, but that is a topic for later.

Let’s just briefly state what needs to be changed on the graph, and your instructor can show you how to do the steps in class if you do not know how. Much of the formatting can be done from the chart tools (green highlighted) menu at the top of the spreadsheet which appears when you click on the plot area of the graph. The data should occupy most of the graph window area, this is a failure of Excel to properly autoscale the maximum and minimum values to use on both the vertical, and especially the horizontal axis. For this you can rescale by clicking on the numbers on each scale and using the format axis box that comes up to fix the various problems with the axis and manually choose the maximum and minimum values for each axis. (Try to pick “nice” rounded or even numbers for maximum and minimum axis values) You can also do a lot of formatting by right clicking the mouse in the plot area and selecting format plot area. Note that there are no labels on the axis stating what is being plotted, the legend simply says series one, and there is no need for a legend when you only have a single set of data or a single curve. There is no box around the graph plotting area, the gridlines are not necessary, and the tick marks should be inside. Select both major and minor tick marks, the default graph only shows major tick marks. Its nice to have a title on your graph. In scientific papers this is usually done with a figure caption instead. In this lab, put a title on all your graphs.

Copy and paste each of your graphs into a new Word document for printing later. (this is to save on printing time and expense. We will be printing our graphs and comments from a word document NEAR THE END OF LAB to avoid accidental printing of numerous pages of large tables of data we will generate later). When you are done formatting, your graph should be formatted like the one shown below (notice the marker style was changed to round, you can
choose whatever you like). Format the font sizes to be 12 point on the axis labels and graph title. The default 10 point font is too small for most purposes.

Note that the data does not appear to be entirely linear when plotted this way. There seems to be an increasing slope at the two shortest wavelengths. Without knowing any theory, one typically plots the dependent variable vs. the independent variable. If the data is not a linear relationship, then one looks for other functional forms. However, the photoelectric effect is described by the following equation, (which earned Einstein the Noble Prize in physics!)

\[ hf = eV_s + W \]  \hspace{1cm} (1.)

or dividing by \( e \) and solving for \( V_s \)

\[ V_s = \left( \frac{h}{e} \right) f - \frac{W}{e} \]  \hspace{1cm} (2.)

Where \( h \) is Planck’s constant, \( e \) the electron charge, \( W \) the work function of the surface, \( V_s \) the stopping voltage, and \( f \) the frequency of the light. Equation 2 shows that if we plot the frequency of the light, not the wavelength as the independent variable and \( V_s \) as the dependent variable, we should get a straight line whose slope is the value \( \frac{h}{e} \), and intercept \( -\frac{W}{e} \). This is easily done by using the (nonlinear) relation between frequency and wavelength, \( f = \frac{c}{\lambda} \), where \( c = 2.9979 \text{ m/s} \) is the speed of light.

Make a column for the frequency for each of the wavelengths in the data table and plot the stopping voltage vs. the frequency (be careful to convert wavelengths to meters from nanometers in your formula!). Use an absolute reference to link the cell formula to a separate cell having the value of \( c \) in it, and use exponential notation (E format) for the frequency values. Add a linear trendline to your data and display the equation and \( R^2 \) values on the graph. VERY IMPORTANT: Be sure to format the trendline label with at least 4 significant figures. Use scientific notation format, since if you only pick four decimals, and the number has a large power of 10, it is not going to be correctly shown on the trendline equation and \( R^2 \) value if not expressed in scientific notation. Finally, format the aspect ratio of your graph (length to height ratio) to be approximately the golden ratio \( \left(\frac{1+\sqrt{5}}{2}\right) = 1.61803 \).
http://en.wikipedia.org/wiki/Golden_ratio) for the most aesthetically pleasing graph. Here is what your graph and spreadsheet should look like. The cell formula for frequency showing the absolute reference is highlighted so you can see how to link to a fixed cell, while the other cells are relative references. This allows you to copy the formula for the top cell into the ones directly below it, and the relative references change with each line, but the absolute ones stay the same.

Also notice that some information cells have been added to the spreadsheet on the accepted value for $h/e$, the experimental value (taken from the slope of the trendline equation), and the percent error from the accepted value, 26.8% in this case. We could/should have added a cell for the experimental value of the work function $W$, its accepted value, and percent error (1.19 and 1.5 eV and 18.4% respectively for the Cs-Sb surface).

There is a problem with this data, as it gives values for $h/e$ that are off by a considerable amount, typically 20-30%, as shown by the comparison with the accepted value on the spreadsheet. When we perform the experiment, we will see the reason for this, having to do with reverse leakage currents, and apply a correction to get better values for $h/e$ and $W$. Suffice to say the good value for $R^2 = 0.99699$ indicates the data is of high quality and linear to high degree. It indicates that the experimenter likely took good data.

**Coefficient of Determination:** What does the $R^2$ value indicate? In statistics, the coefficient of determination, denoted $R^2$ and pronounced R squared, indicates how well data points fit a line or
$R^2$ is a statistic that will give some information about the goodness of fit of a model. It provides a measure of how well observed outcomes are replicated by the model, as the proportion of total variation of outcomes explained by the model. The coefficient of determination ranges from 0 to 1 in most situations you will encounter. An $R^2$ close to one represent a trendline that is almost identical to the data points, meaning that you can use the trendline to accurately predict additional values. The closer the $R^2$ is to one, the better fit to the data, an $R^2$ of exactly 1 indicates that the regression curve perfectly fits the data.

However, you need several decimals accuracy of $R^2$ (usually 5 is good) for data that is very accurate, since there is a visual difference in the quality of fit between $R^2$ values of 0.991 and 0.9993 data fits. Also, a fit of 0.9 might sound pretty good, but actually the fit to the data can be rather poor. Often in life sciences and in data sets with several different confounding variables, $R^2$ values might be very low, like 0.5 - 0.7, yet they are considered to be indicative of a trend. For more information on this subject consult this article and the references therein (http://en.wikipedia.org/wiki/Coefficient_of_determination)

Add the section shown below to your spreadsheet with random errors added to the stopping voltage values. (the one shown is right underneath the other data and graphs we have done so far). In the spreadsheet below, the formula cell to add a variable random fractional error between plus and minus eps is shown. Adjust the amplitude of the random noise parameter eps for a few different values between 0 → 0.5 and note how the slope and $R^2$ value changes. Each time you enter a value of eps, the random numbers recalculate (even if you retype the same value and re-enter it, which is instructive, try it a couple of times and watch what happens). Note the graph here has a considerable difference in the slope from the no noise case above to a nominal 15% noise shown in the graph (2.8419 vs. 3.0284 in units of $10^{14}$ Hz), but yet the $R^2$ value is a relatively high 0.980 vs. 0.99699.
Paste 2 or 3 of your graphs with randomized data into the Word document you have open and type a couple of sentences describing what you did and observed.

Aside from data that have an intrinsic linear dependence, which you have seen in other labs, there is a large variety of data that is clearly nonlinear in the independent variable $x$ but can be transformed to new variables so that one can still obtain a linear relationship between the transformed variables, and then perform a linear least square fit. We just did a very simple case by changing from wavelength to frequency to make the photoelectric data linear. Consider now a more complicated case.

**Make a Template of your graphs:** To make a template of your graphs, just click on the plot area of a scatter plot graph so that Green Chart Tools tab comes up. Select the design tab and then go to the far left and click on the Save as Template menu item. This will then bring up a save box and you need to tell it where to save the template. This is the file location where it will have to be retrieved from when you want to use it so that you do not have to go thru all the formatting steps each time you do a graph.

**Example: Thin Lens Equation**

The geometric optics thin lens equation that is valid for lenses or mirrors if given by

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \tag{3.}$$

where $s$ is the object distance (where the light source is), $s'$ the image distance (where the source gets focused), and $f$ is the focal length of the lens. The distances and $s$ and $s'$ are measured from the front and back surfaces of the lens respectively, the so called vertices of the surfaces. When $f$ is positive ($f > 0$) the lens is a converging lens and is said to be a positive lens, when $f$ is negative ($f < 0$), the lens is a diverging lens and is said to be a negative lens.

Solving for the image distance $s'$ in terms of the object distance we have

$$s' = s \frac{f}{s-f}. \tag{4.}$$

This is clearly a nonlinear input output relationship between $s$ and $s'$. Let’s plot it for the 2 cases of $f = 1$ and $f = -1$. Make a table of values for $s$ from -10 to plus 10, say at least 100 equally spaced values and the corresponding values of $s'$ for both the $f$ =1 and $f$ = -1 cases. Then plot these two cases. Make sure you do not accidentally get a value where $s = f$, otherwise you will get numerical infinity (overflow). Comment on the features of these two plots. Paste the graphs in word and type your comments on the graphs there. What does it mean when $s'$ changes sign in the case and where does it do so? What does it mean when $s'$ does not change sign?

Often we can find a simple transformation that simplifies the behavior of a more complex function into something that we are familiar with, such as a linear function. Note the simple transformation of variables to $u = 1/s$, $v = 1/s'$ in the thin lens eqn. 3 yields the following relation

$$v = -u + 1 /f \tag{5.}$$
If we measure all distances in units of the magnitude of the focal length $|f|$, then eqn. 5 simplifies to 2 equations for a straight line.

$$v = -u + 1 \text{ positive/converging lenses,}$$  \hspace{1cm} (6.)

$$v = -u - 1 \text{ negative/diverging lens}$$  \hspace{1cm} (7.)

Note the nonlinear relation eqn. 4 between object and image distance has been transformed to be a linear relation between object and image. Both converging and diverging lenses have the same slope of -1, the only difference being a positive $y$ intercept of +1 for positive lenses vs. a negative intercept of -1 for negative lenses. Plot this relationship for both positive and negative lenses on the same graph, choose $f = +1$ for converging lenses and $f = -1$ for diverging lenses. Generate columns of equally spaced values of $u$ and the corresponding values of $v$ for converging and diverging lenses for at least 100 values of $u$ between -10 to plus 10. Plot both the converging and diverging lenses on the same $(u,v)$ graph. Paste your comments and graphs in the word document.

Note that for real object distances, $s$ is positive, and thus $u$ will also be positive. If we take $u$ to be positive, explain the difference of what happens to $v$ in the case of a converging lens vs. a diverging lens. Put this in your comments. (Hint look at your graphs and what happens to $v$ for $u$ positive in the two cases. Note that $v$ positive is a real image and $v$ negative is a virtual image).