

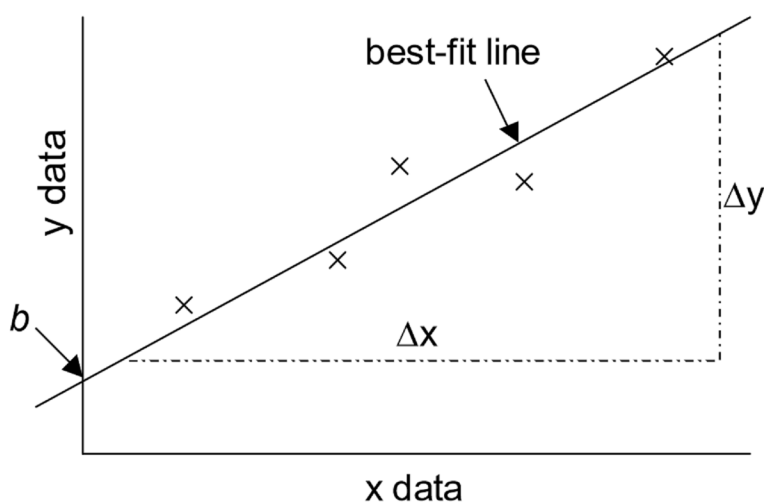
1: Using Excel for Graphical Analysis of Data (Experiment)

Objectives

- To learn to use Excel to explore a number of linear graphical relationships.

In several upcoming labs, a primary goal will be to determine the mathematical relationship between two variable physical parameters. Graphs are useful tools that can elucidate such relationships. First, plotting a graph provides a visual image of data and any trends therein. Second, via appropriate analysis, they provide us with the ability to predict the results of any changes to the system.

An important technique in graphical analysis is the transformation of experimental data to produce a straight line. If there is a direct, linear relationship between two variable parameters, the data may be fitted to the equation of line with the familiar form $y = mx + b$ through a technique known as [linear regression](#). Here m represents the slope of the line, and b represents the y -intercept, as shown in the figure below. This equation expresses the mathematical relationship between the two variables plotted, and allows for the prediction of unknown values within the parameters.



The equation for the best-fit line is

$$y = mx + b \quad (1.1)$$

where

$$b = \text{y-intercept} \quad (1.2)$$

$$m = \text{slope} \quad (1.3)$$

$$= \frac{\Delta y}{\Delta x} \quad (1.4)$$

$$= \frac{y_2 - y_1}{x_2 - x_1} \quad (1.5)$$

Computer spreadsheets are powerful tools for manipulating and graphing quantitative data. In this exercise, the spreadsheet program Microsoft Excel[®] will be used for this purpose. In particular, students will learn to use Excel in order to explore a

number of linear graphical relationships. Please note that although Excel can fit curves to nonlinear data sets, this form of analysis is usually not as accurate as linear regression.

Procedure

Part 1: Simple Linear Plot

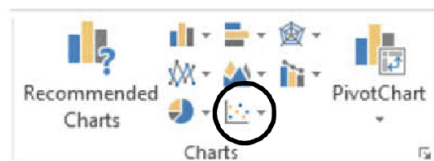
Scenario: A certain experiment is designed to measure the volume of 1 mole of helium gas at a variety of different temperatures, while keeping the gas pressure constant at 758 torr:

Temperature (K)	Volume of Helium (L)
203	14.3
243	17.2
283	23.1
323	25.9
363	31.5

1. Launch the program Microsoft Excel[®] (2016 version, found on all computers in all the computer centers on campus). Go to the Start button (at the bottom left on the screen), then click Programs, followed by Microsoft Excel[®].
2. Enter the above data into the first two columns in the spreadsheet.
 - Reserve the first row for column labels.
 - The x values must be entered to the left of the y values in the spreadsheet. Remember that the independent variable (the one that you, as the experimenter, have control of) goes on the x-axis while the dependent variable (the measured data) goes on the y-axis.
3. Highlight the set of data (not the column labels) that you wish to plot (Figure 1).

	A	B
1	Temp	Vol
2	203	14.3
3	243	17.2
4	283	23.1
5	323	25.9
6	363	31.5
7		
8		

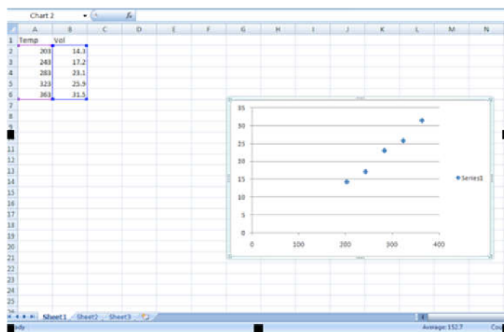
- Click on *Insert > Recommended Charts* followed by *Scatter* (Figure 2).



- Choose the scatter graph that shows data points only, with no connecting lines – the option labeled *Scatter with Only Markers* (Figure 3).

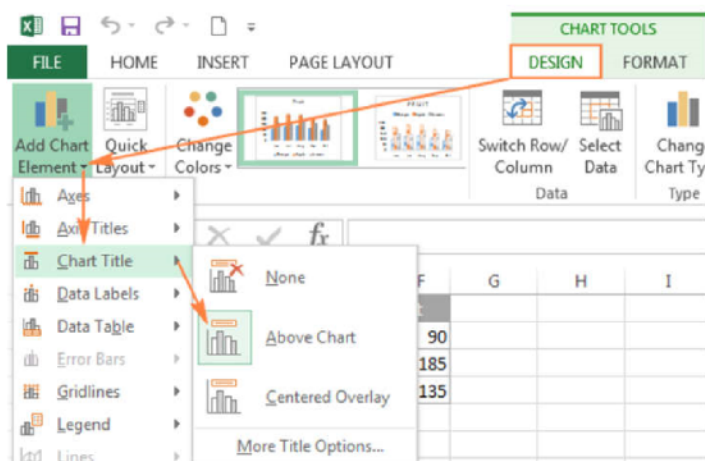


- You should now see a scatter plot on your Excel screen, which provides a preview of your graph (Figure 4).

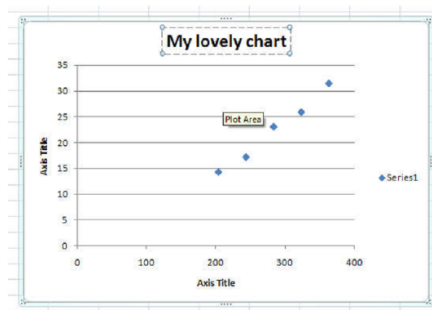


4. If all looks well, it is time to add titles and label the axes of your graph (Figure 5).

- First, click inside the chart.
- Switch to the *Design* tab, and click *Add Chart Element > Chart Title > Above Chart*
- The graph should be given a meaningful, explanatory title that starts out "Y versus X followed by a description of your system.
- Click on *Axis Titles* (select *Primary Horizontal Axis Title* and *Primary Vertical Axis Title*) to add labels to the x- and y-axes. Note that it is important to label axes with both the measurement and the units used.



- To change the titles, click the text box for each title, highlight the text and type in your new title (Figure 6).



5. Your next step is to add a trendline to the plotted data points. A trendline represents the best possible linear fit to your data. To do this you first need to "activate" the graph. Do this by clicking on any one of the data points. When you do this, all the data points will appear highlighted.

- Click the *Chart Elements* button next to the upper-right corner of the chart.
- Check the *Trendline* box.
- Click *More Options*. This will display the option shown in Figure 7.
- Notice that the *Linear* button is already selected. Now select the *Display Equation on Chart* box and the *Display R-squared*

value on *Chart* box. Then click the *Close* button.



6. The equation that now appears on your graph is the equation of the fitted trendline. The R^2 value gives a measure of how well the data is fit by the equation. The closer the R^2 value is to 1, the better the fit. Generally, R^2 values of 0.95 or higher are considered good fits. Note that the program will always fit a trendline to the data no matter how good or awful the data is. You must judge the quality of the fit and the suitability of this type of fit to your data set.
7. Print out a full-sized copy of your prepared graph and attach it to your report. Then record the following information on your report:
 - the equation of the best-fit trendline to your data
 - the slope of the trendline
 - the y-intercept of the trendline
 - whether the fit of the line to the data is good or bad, and why.
8. By graphing the five measured values, a relationship is established between gas volume and temperature. The graph contains a visual representation of the relationship (the plot) and a mathematical expression of the relationship (the equation). It can now be used to make certain predictions.

For example, suppose the 1 mole sample of helium gas is cooled until its volume is measured to be 10.5 L. You are asked to determine the gas temperature. Note that the value 10.5 L falls outside the range of the plotted data. How can you find the temperature if it doesn't fall between the known points? There are two ways to do this.

Method (1): Extrapolate the trendline and estimate where the point on the line is.

- Click on the *Layout* tab along the top menu, then *Trendline > More Trendline Options*.
- In the section labeled *Forecast* enter a number in the box labeled *Backward*, since we want to extend the trendline the backward x direction. To decide what number to enter, look at your graph to see how far back along the x-axis you need to go in order to cover the area where volume = 10.5 L. After entering a number, click *Close*, and the line on your graph should now be extended in the backward direction.
- Now use your graph to estimate the x value by envisioning a straight line down from $y = 10.5$ L to the x-axis. Record this value on your report.

Method (2): Plug this value for volume into the equation of the trendline and solve for the unknown temperature. Do this and record your answer on your report. Note that this method is generally more precise than extrapolating and "eyeballing" from the

graph.

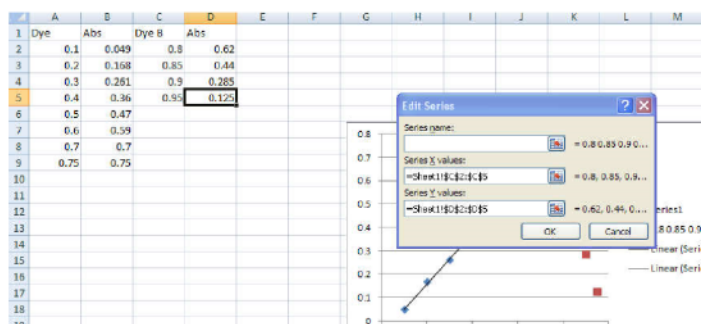
Part 2: Two Data Sets with Overlay

Scenario: In a certain experiment, a spectrophotometer is used to measure the light absorbance of several solutions containing different quantities of a red dye. The two sets of data collected are presented in the table below:

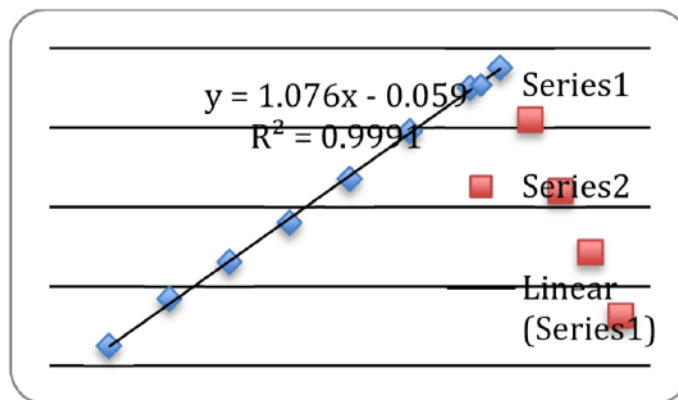
Data A		Data B	
Amount of Dye (mol)	Absorbance (unitless)	Amount of Dye (mol)	Absorbance (unitless)
0.100	0.049	0.800	0.620
0.200	0.168	0.850	0.440
0.300	0.261	0.900	0.285
0.400	0.360	0.950	0.125
0.500	0.470		
0.600	0.590		
0.700	0.700		
0.750	0.750		

You would like to see how these two sets of data relate to each other. To do this you will have to place both sets of data, as independent relationships, on the same graph. Note that this process only works when you have the same axis values and magnitudes.

1. Enter this new data on a fresh page (Sheet 2) in Excel. Be sure to label your data columns A and B. Again, remember to enter the x values to the left of the y values.
2. First, plot Data A only as an XY Scatter plot (the same way you did with the data in Part 1). Fit a trendline to this data using linear regression, and obtain the equation of this line.
3. Now you need to add Data B to this graph.
 - Activate the graph by clicking on one of the plotted data points.
 - Right-click the chart, and then choose *Select Data*. The *Select Data Source* box appears on the worksheet with the source data of the chart.
 - Click the *Add* tab and type "Data B" for the Series Name.
 - Click the little icon under *Series X values*, then highlight the x-axis values of Data B.
 - Press enter, then repeat this procedure for the *Series Y values*, highlighting the y-axis values of Data B. For each of these steps, you should see a display similar to what is shown in Figure 8. Note that slight differences may appear due to the version of Microsoft Excel[®] installed on your computer.



- Click OK twice to return to the main Excel window.
- At this point you should see the new data points (labeled as Series 2) as shown in Figure 9. You can now independently analyze this dataset by inserting a trendline as before.



4. Print out a full-sized copy of your prepared graph and attach it to your report. Then record the following information on your report:
- the equation of the best-fit trendline for Data A,
 - the equation of the best-fit trendline for Data B,
 - If these trendlines were extrapolated, they would intersect. Determine the values of x and y for the point of intersection using simultaneous equations.

Part 3: Statistical Analysis and Simple Scatter Plots

When many independent measurements are made for one variable, there is inevitably some scatter (noise) in the data. This is usually the result of random errors over which the experimenter has little control.

Scenario: Ten different students at two different colleges each measure the sulfate ion concentration in a sample of tap water:

College #1 dataset	35.9 ppm	43.2 ppm	33.5 ppm	35.1 ppm	32.8 ppm	37.6 ppm	31.9 ppm	36.6 ppm	35.0 ppm	32.0 ppm
College #2 dataset	45.1 ppm	34.2 ppm	36.8 ppm	31.0 ppm	40.7 ppm	29.6 ppm	35.4 ppm	32.5 ppm	43.5 ppm	38.8 ppm

Simple statistical analyses of these datasets might include calculations of the mean and median concentration, and the standard deviation. The mean (\bar{x}) is simply the average value, defined as the sum (Σ) of each of the measurements (x_i) in a data set divided by the number of measurements (N):

$$\bar{x} = \frac{\sum x_i}{N} \quad (1.6)$$

The median (M) is the midpoint value of a numerically ordered dataset, where half of the measurements are above the median and half are below. The median location of N measurements can be found using:

$$M = \frac{N + 1}{2} \quad (1.7)$$

When N is an odd number, the formula yields a integer that represents the value corresponding to the median location in an ordered distribution of measurements. For example, in the set of numbers (3 1 5 4 9 9 8) the median location is $(7 + 1) / 2$, or the 4th value. When applied to the numerically ordered set (1 3 4 5 8 9), the number 5 is the 4th value and is thus the median – three scores are above 5 and three are below 5. Note that if there were only 6 numbers in the set (1 3 4 5 8 9), the median location is $(6 + 1) / 2$, or the 3.5th value. In this case the median is half-way between the 3rd and 4th values in the ordered distribution, or 4.5.

Standard deviation (s) is a measure of the variation in a dataset, and is defined as the square root of the sum of squares divided by the number of measurements minus one:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N - 1}} \quad (1.8)$$

So to find s , subtract each measurement from the mean, square that result, add it to the results of each other difference squared, divide that sum by the number of measurements minus one, then take the square root of this result. The larger this value is, the greater the variation in the data, and the lower the precision in the measurements.

While the mean, median and standard deviation can be calculated by hand, it is often more convenient to use a calculator or computer to determine these values. Microsoft Excel[®] is particularly well suited for such statistical analyses, especially on large datasets.

1. Enter the data acquired by the students from **College #1** (only) into a single column of cells on a fresh page (Sheet 4) in Excel. Then in any empty cell (usually one close to the data cells), instruct the program to perform the required functions on the data. To compute the mean or average of the data entered in cells a1 through a10, for example, you must:

- click the mouse in an empty cell
- type "=average(a1:a10)"
- and press return

To obtain the median you would instead type "=median(a1:a10)". To obtain the standard deviation you would instead type "=stdev(a1:a10)".

2. Record on your report:

- The Excel calculated mean, median and standard deviation for the College #1 dataset.
- As an additional exercise, calculate the standard deviation of this dataset by hand, and compare it to the value obtained from the program.

Rejecting Outliers

Do all the measurements in the College #1 data set look equally good to you, or are there any values that do not seem to fit with the others? If so, are you allowed to reject these measurements?

Outliers are data points which lie far outside the range defined by the rest of the measurements and may skew your results to a great extent. If you determine that an outlier resulted from an obvious experimental error (e.g., you incorrectly read an instrument or prepared a solution), you may reject the point without hesitation. If, however, none of these errors is evident, you must use caution in making your decision to keep or reject a point. One rough criterion for rejecting a data point is if it lies **beyond two standard deviations from the mean or average**.

3. Using the above criteria, determine if there are any outliers in the College #1 dataset.

- Record these outlier measurements (if any) on your report.
- Then, excluding the outliers, re-calculate the mean, median and standard deviation of this data set (use Excel).

Rejecting data points cannot be done just because you want your results to look better. If you choose to reject an outlier for any reason, you must always include documentation in your lab report which clearly states:

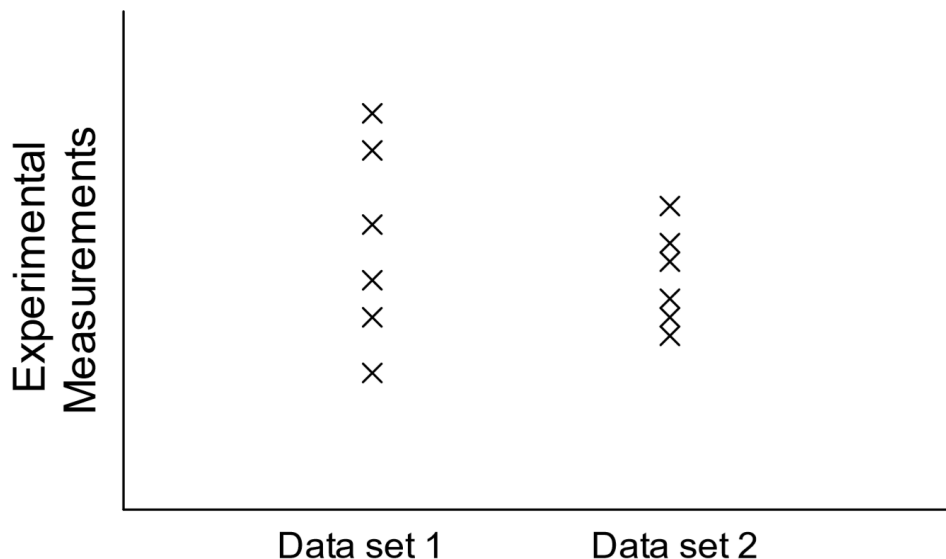
- that you did reject a point
- which point you rejected
- why you rejected it

Failure to disclose this could constitute scientific fraud.

Graphing a Scatter Plot

Unlike the linear plots created so far, a scatter plot simply shows the variation in measurements of a single variable in a given dataset, i.e., it supplies a visual representation of the "noise" in the data. The data is plotted in a column, and there is no x-y

dependence here (Figure 10). Note that datasets with a greater degree of scatter will have a higher standard deviation and consist of less precise measurements than datasets with a small degree of scatter.



To obtain such a plot using Excel, all the x values for each dataset must be identical. Thus, let the College #1 data be assigned $x = 1$, and let $x = 2$ for all the College #2 data:

Measurements by Students from College #1		Measurements by Students from College #2	
College 1	$[\text{SO}_4^{-2}]$ (ppm)	College 2	$[\text{SO}_4^{-2}]$ (ppm)
1	35.9	2	45.1
1	43.2	2	34.2
1	33.5	2	36.8
1	35.1	2	31.0
1	32.8	2	40.7
1	37.6	2	29.6
1	31.9	2	35.4
1	36.6	2	32.5
1	35.0	2	43.5
1	32.0	2	38.8

- Enter the data as shown above into the first four columns of your spreadsheet.
 - Plot the College #1 dataset as an XY Scatter Plot.
 - Now add the College #2 dataset to this graph applying the same steps you used to create your earlier graph in the section "Two Data Sets with Overlay" (Part 2).
 - Add appropriate axis labels and a title. You may also want to adjust the x-axis and y-axis scales to improve the final look of your graph.
- Print out a full-sized copy of your prepared graph and attach it to your report. Then record the following information on your report:
 - Which dataset (College #1 or College #2) show the least scatter? The greater standard deviation? The more precise measurements?

Lab Report: Using Excel for Graphical Analysis of DataName: _____ Lab Partner: None for this assignment

Date: _____ Lab Section: _____

Turn in the graphs you made for ALL three parts in this assignment

For each graph make sure the following components are in the printout:

1. Title for the graph
2. Labels for x and y axes (along with appropriate units when applicable)
3. Line equation and R^2 when appropriate.

Part 1: Simple Linear Plot

- Which set of data is plotted on the y-axis?
- Which set of data is plotted on the x-axis?
- Record the following information:
 - The equation of the fitted trendline
 - The value of the slope of this line
 - The value of the y-intercept of this line
- Is the fit of the trendline to your data good (circle one)? **Yes / No**. Explain why you think the line is a good fit to the data.
- Determine the temperature (in K) of the gas in the cold room when it has a measured volume of 10.5 L using
 - a) Extrapolation and “eyeballing”
 - b) The equation of the trendlineShow your calculations for b) below.

Part 2: Two Data Sets and Overlay

- Record the equations of the trendlines fitted to
 - Data set A:
 - Data set B:
- Perform a simultaneous equations calculation to determine the x and y values for the point of intersection between these lines. Show your work below.

Part 3: Statistical Analysis and Simple Scatter Plots

- For the College #1 data set, record the following values (determined using Excel):
 - the mean SO_4^{-2} concentration
 - the median SO_4^{-2} concentration
 - the standard deviation in the data set
- Calculate the standard deviation in the College #1 data set *by hand*. Show *all* your work below. Continue your work on an attached page if you require more space.
- Are there any outliers in the College #1 data set (circle one)? Yes / No
- If yes, which measurements are the outliers?
- Show the calculations you used to identify the outliers (or, if none, how you determined that there were none).
- Re-calculate the following values (using Excel) *excluding the outliers*:
 - the mean SO_4^{-2} concentration
 - the median SO_4^{-2} concentration

- the standard deviation in the data set
- Create a scatter plot showing both the College #1 and College #2 data. Attach a printout of your graph to this report. Be sure that your axes are properly labeled, and that your graph has an appropriate title.
- Examine your plotted data. Which data set:
 - has the larger standard deviation?
 - contains the more precise measurements?