Calibration and Linear Regression Analysis: A Self-Guided Tutorial

Part 1 – Instrumental Analysis with Excel: The Basics

CHM314 Instrumental Analysis

Department of Chemistry, University of Toronto

Dr. D. Stone (*prepared by J. Ellis*)

1 Introduction - Instrumental Analysis and Calibration

Instrumental analysis is very important in all areas of analytical chemistry. Modern analytical chemistry is a quantitative science, meaning that the desired result is almost always numeric. We need to know there is 55 μg of mercury in a sample of water, or 20 mM glucose in a blood sample. Quantitative results for analytical chemistry are obtained using devices or instruments that allow us to determine the concentration of a chemical in a sample from an observable signal.

Before an instrument can be successfully used to determine a concentration, it must be *calibrated* for values in the range that it is to be used. This normally involves testing samples of known concentration, known as standards, and measuring the corresponding signal from the device. This is performed over the entire operating range of the instrument, or at least within the linear range, to statistically generate a *calibration curve* for the device. In this brief tutorial, the basic fundamentals of calibration curve determination will be covered, including linear regression and correlation. As well, the basics of statistical analysis using Microsoft™ Excel will be presented.

A calibration curve is an equation relating the output signal of an instrument, such as an electrical voltage or current, to the quantity that the instrument measures. In the simplest form, the calibration curve will take the form of the equation of a straight line, with a slope and a *y*-intercept, determined by statistical analysis of the calibration data. Other equations can be used, such as logarithmic or polynomial fits. These will be explained in later sections.

Throughout this tutorial, all steps and examples are performed using MS Excel. However, the information and techniques can be applied to any spreadsheet program. Spreadsheet software can be easily applied to statistical problems, including calibration curve generation, but more sophisticated software exists to facilitate the computations, such as mathematical packages like Matlab or Mathematica, and statistical software like SAS or SPSS. In fact, Excel comes with a built-in Analysis Toolpak, which will be discussed at the end of the tutorial. While these packages automate many of the computations needed in statistical processing, the results are generally more difficult to interpret. The steps presented in this section, and throughout the tutorial, will provide you with a working knowledge of spreadsheets and their use in basic statistical instrumental analysis.

2   Microsoft™ Excel Basics

In this section, data manipulation using MS Excel is introduced, including importing, copying and pasting data and entering equations. A basic understanding of the Windows operating system is assumed, including the ability to navigate within Windows, and find and open files.

Throughout this tutorial, different text formats will be used to indicate different actions or operations. Keystrokes (things you need to type) will be denoted with the `Courier` font. Menus will be denoted with **bold Arial**, followed by an arrow, then the menu item in Arial. So if you see **File**→New, this means, click on the **File** menu at the top of the screen and select the New option. Options, buttons in dialog boxes and the Enter key will also use this font. Dialog box titles will be in **bold Times**. Functions will be denoted with `ALL-CAPS COURIER`.

2.1 Using Excel

1. Open Excel. One way to do this is to click the **Start** button, select Run and type `Excel`, then press `Enter`.

2. Open a new file by selecting **File**→New or pressing `Ctrl-N`. A new spreadsheet file will appear. You can also open an existing file by selecting **File**→Open, or pressing `Ctrl-O`. In this case, a dialog box will open, showing a directory tree. You can navigate to the file you would like to open. However, we will work with a new file, into which you will import numerical data.

2.1.1 Cells in Excel

Data is entered and manipulated in Excel within cells. Each block on the screen when you look at Excel is a cell. A cell can contain letters, numbers or equations. Equations operate on other cells in the spreadsheet to calculate values. The cells are divided into columns, designated by letters, and rows, designated by numbers, and are denoted by the letter, followed by the number. For example, the first cell on a worksheet is A1. You can enter data into cells in a number of ways. The easiest is simply by typing the desired value into the cell and pressing the Enter key, but this can become tedious if there is a list or series of data that you wish to enter. This can be done by *pasting* a series of data or entering an equation. You can also plot a graph of your data.

2.1.2 Pasting a Series

Pasting a series of data is useful if there is a regular pattern to the data, or if it is useful to view the form of an equation. It is of much use in analysis, but is a good introductory exercise as it illustrates some useful techniques. In this, we will see how to make three columns of regularly spaced data.

1. Open new worksheet, and place the cursor in the top-left most cell (A1).

2. Type `1` and press Enter. The number 1 should appear on the right side of the cell, and the cursor should move to cell A2.

3. Type `2` and press Enter. The first two cells in column A should now be filled with 1 and 2.

4. Select cells A1 and A2 by clicking and holding the mouse button on cell A1, dragging to cell A2, and releasing the mouse button.

5. Using the mouse, place the arrow at the bottom right corner of the selection, the *handle*, which is marked by a small square. The cursor should change to indicate that you are on the handle.

6. Click and drag the handle down to cell A10. This will fill cells A1-10 with the numbers 1 through 10. You can drag the handle to different cells to create different length series.

7. To create series with different intervals, simply change the first two numbers. In cells B1 and B2, enter 0 and 0.1, respectively.
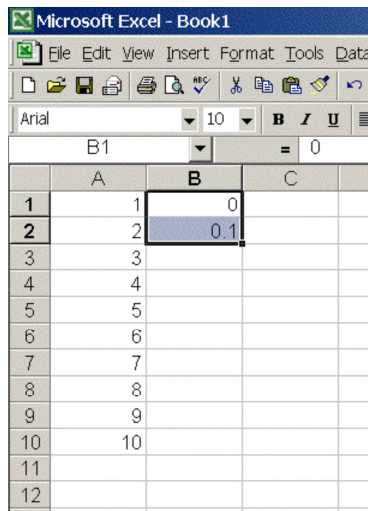
8. Drag the handle to B16 to create the series.

9. You can have decreasing and negative series as well.

2.1.3 Equations

We will now see how to manipulate data using equations. This is useful when you want to test out a calibration curve, or use the calibration equation to analyze experimental data. You can use Excel to generate complex equations, however we will only treat very simple ones here. Before we begin, note the table of operators below used in numerical computing. These are not exactly the same as you would see written elsewhere, but they mean the same thing.

Task	Operator	Example	Result
Multiplication	*	2*3	6
Division	/	4/2	2
Exponent	^	2^3	8
Order of operations	(..)	2*3+5	11
		2*(3+5)	30
Power of ten	e or E	3.2e+4	32000
		3.2e-2	0.032

Continuing from the previous example:

1. In cell C1, beside the 0 from the second series, type =, then select cell B1. Cell C1 should now contain the phrase =B1. The =-sign tells Excel that the text following the = is part of an equation.

2. Continue the equation by typing *2+5 (do not press Enter yet) so cell C1 should read B1*2+5. Now press enter and the cell should read 5. This is the result of 0×2+5.

3. Now copy this equation to the other cells in column C by dragging the handle of cell C1 (the square in the bottom right corner of the cell) to C16. This fills in the values for all the cells, and should contain the values 5 to 8. Note that the equation for each cell, which you can view along the top of the worksheet in

the area marked f_x, has changed to reflect the referencing cell. Excel does this automatically to allow easy copying of equations.

	C16		=	=B16*2+5	
	A	B	C	D	
1	1	0	5		
2	2	0.1	5.2		
3	3	0.2	5.4		
4	4	0.3	5.6		
5	5	0.4	5.8		
6	6	0.5	6		
7	7	0.6	6.2		
8	8	0.7	6.4		
9	9	0.8	6.6		
10	10	0.9	6.8		
11		1	7		
12		1.1	7.2		
13		1.2	7.4		
14		1.3	7.6		
15		1.4	7.8		
16		1.5	8		
17					
18					

The series in column C1 is the equation for the straight line $y = 2x + 5$ for $x = 0\ldots1.5$. This will become more apparent in the next section. However, we will first try one more equation.

4. In column E, create a series from -1 to 1 with an increment of 0.2.

5. In cell F1, type the equation `=E1^2+$A$2` and drag it over the whole series in column E.

The `$`-sign tells Excel that when copying the equation into other cells, it should always use the value in A2, and not change it based on the cell referencing. This is useful for defining constants. You can also use the $-modifier once only, such as $A2. This means that the column remains constant, and only the rows change as the equation is copied.

This series represents the parabola $y = x^2 + 2$ for $x = -1\ldots1$.

2.1.4 Plotting Data in Charts

A useful way to view and present your data is with charts. There are many types of charts available with Excel, but the most useful for calibration curves is the X-Y scatter plot. We will use this tool to plot the straight line $y = 2x + 5$ and the parabola $y = x^2 + 2$.

1. With the mouse, highlight columns B and C, rows 1 to 16.

2. Select **Insert**→Chart. The **Chart Wizard** dialog box will appear.

3. In Step 1, there are many types of charts to choose from. Select the XY (Scatter) and click Next.

4. In Step 2, we can choose to plot multiple series on the same chart. We won't do that here. In fact, Excel has already specified the proper axes, so we do not need to change anything.[1] Click Next.

---

[1] For more information on plotting multiple series on one chart, check the Excel Help under Plotting.
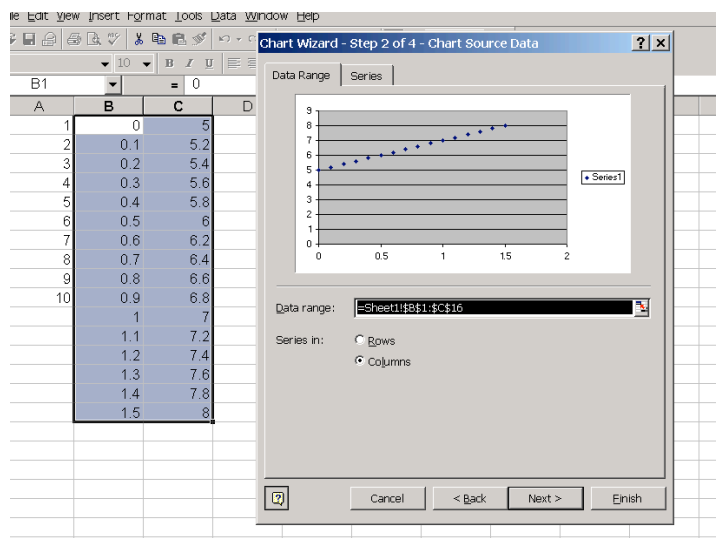
5. In Step 3, you can enter chart titles, axis titles, and other display characteristics on the chart. Change what you want, then click Next.

6. In Step 4, you can specify whether the chart should be shown in the current worksheet, or whether it should stand alone in a separate worksheet. Click Finish and the chart should appear.

7. Repeat these steps to plot the parabola, highlighting instead columns E and F.

Your plots should look like this:
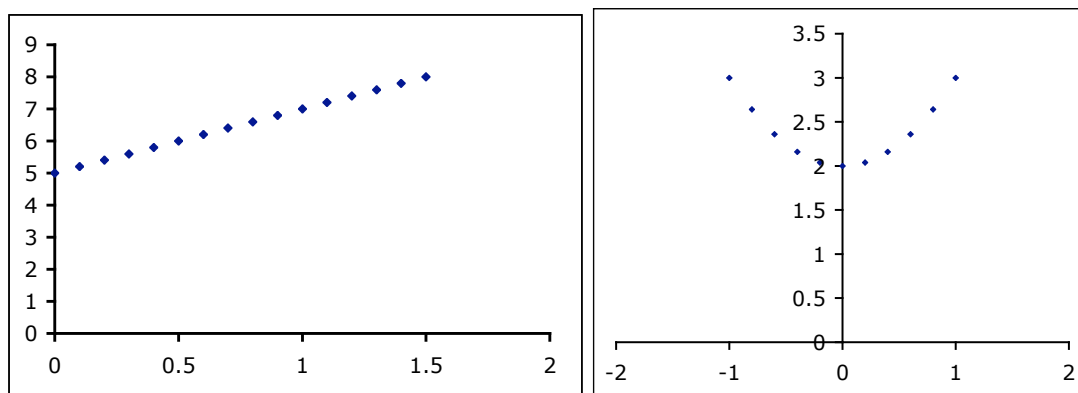
2.2 Excel Functions

In this section, we consider two very useful Excel functions that you will use to develop calibration curves. Excel functions are built-in formulas that perform frequent operations. The two functions we will review here are SUM and AVERAGE. Functions can be entered as part of an equation, using the = sign, or entered in a cell on its own.

You will enter fluorescence intensities and use this data to generate a calibration curve related concentrations in $pg \cdot ml^{-1}$ to the intensities, from Miller and Miller[2]. This data will then be used in all the subsequent examples. Enter the following data in the second two columns of an Excel spreadsheet. Column B should contain the fluorescence intensities and column C the concentrations.

---

[2] Miller JN and Miller JC. **Statistics and Chemometrics for Analytical Chemistry**, Pearson Education Ltd: United Kingdom (2000), 4th Edition.

Fluorescence Intensities	Concentration (pg/ml)
2.1	0
5.0	2
9.0	4
12.6	6
17.3	8
21.0	10
24.7	12

2.2.1 The SUM Function

The SUM function in Excel is used to add all the elements in a series of data. This is fundamental for statistical analysis, since all applications involve the sum of a series of numbers or samples. Its use is straightforward when operating on a series of data, such as the fluorescence intensities in the previous exercise. Using this data as a starting point, we will see how the SUM function works.

1. If it is not already open, find and open the fluorescence data from the previous exercise.

2. In the cell directly under the concentration values (should be cell B9, from the previous exercise), type =sum) then highlight the cells you want to sum (in this case, B2-B8). The formula in the cell should now read =sum(B2:B8. Close the bracket and press Enter. The value in the cell should now be 42. This should come as no surprise.

3. Do the same for the fluorescence intensity data. The sum should be 91.7.

Throughout this tutorial, whenever you see the symbol $\sum$ in a formula, it indicates summation over a series of samples or data.

2.2.2 The AVERAGE Function

The AVERAGE function, which calculates the mean of a set of samples, is also very useful in statistical analysis. A full definition of the arithmetic mean of a set of data is quite complicated and involved. A simple definition is that the mean is the expected result of any process. It is important to not confuse the population mean with the sample mean. The sample mean is the mean for a set of discrete samples n, given by the formula $\bar{x} = \frac{1}{n}\sum x_i$, where the x_i are all the discrete samples. The average of a set is denoted by the symbol $\bar{x}$, so any formula containing this indicates the use of the AVERAGE function.

The population mean is the expected value for a process as n approaches infinity. For instance, if we take a very large number of samples (normally on the order of 10^6 or even higher), we would approach the population mean. It is the true expected value, and the sample mean is an approximation of. This is normally denoted μ, however it is not encountered in this tutorial and is only presented here as to make you aware that there is a difference.

The AVERAGE function is used in the same way in Excel as the SUM function, except the word sum is replaced with average. To calculate the average fluorescence intensity, the cell where you wish to do this should look like =average(A2:A9). The calculated value should be 13.1.

You are now ready to start calculating calibration curves using Excel, which will be covered in the next section.

## 3   The Calibration Curve and Correlation Coefficient

A calibration curve is an equation that permits us to calculate a desired experimental result in terms of another. In the simplest form, this is given as the equation for a straight line, where the x-value is the input and the y-value is the output. This is a best-fit curve through a series of experimental sample data, and is the series of all points that are average $\{x,y\}$ pairs of data, for the range of x and y, $\{\bar{x},\bar{y}\}$. Once we know the equation for the average line, we can determine how well it fits the actual experimental data, using the product-moment correlation coefficient, or, for simplicity, the correlation coefficient, R. This is a measure of how close the data points are to the line. If the correlation coefficient is ±1, it is a perfect fit and the line accurately describes the data. An R of 0 indicates no linear correlation, and the straight line does not describe the data at all. An $|R|$ value close to 1 is desirable. The sign of R indicates the slope of the regression line. The square of the correlation coefficient, R^2, is also a common measure.

The simplest way to determine a calibration curve from plotted data in Excel is to use the Trendline option on a chart. From this you can view the best-fit curve line and display the equation and correlation coefficient. While this does not provide accurate information, it is a good first test of the correlation. To display a trendline on a plot on the fluorescence:

1. Plot the fluorescence intensity data on a chart, with the concentration on the x-axis and the fluorescence intensities on the y-axis.

2. On the plot, highlight the data series of interest and press the right mouse button. This will open a context-sensitive menu.

3. Select Add Trendline… A dialog box will appear.
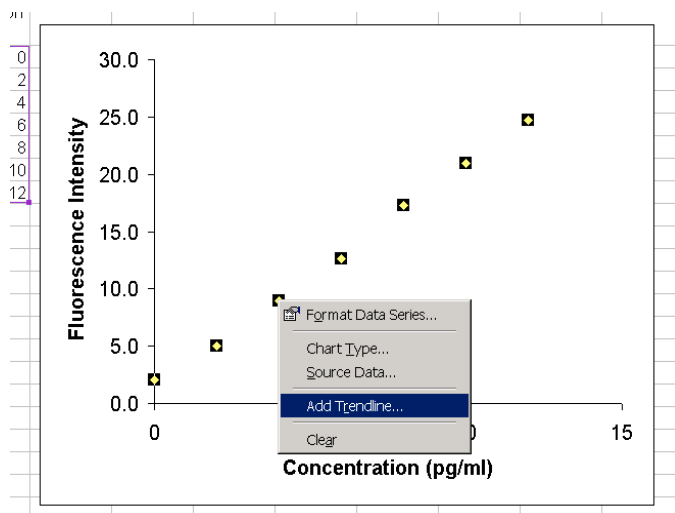
4. In the **Type** tab, select Linear in the **Trend/Regression Type** box and press OK. A trendline will appear on your plot, which is the best-fit line through the data points.
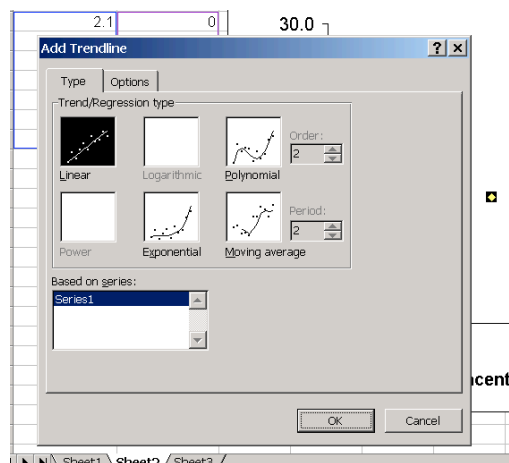
You can display the equation of the trendline and the correlation coefficient R^2 on your plot. Once you have added the trendline,

1. Right-click on the trendline (not on the data points). A context menu will appear.

2. Select Format Trendline... A dialog box will appear.

3. In the **Options** tab, check the Display equation on chart and Display R-squared value on chart. Excel displays R^2 as the correlation coefficient.

4. Click OK, and the equation and R^2 values. You can see that these are the same as those calculated previously.
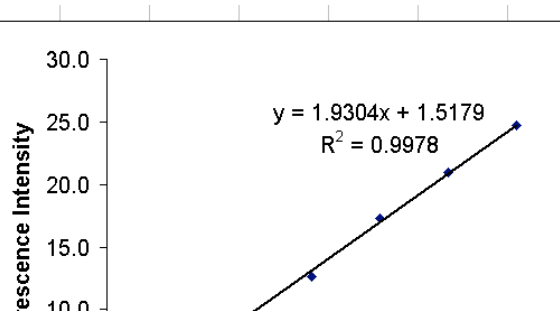
In future sections, we will discuss the meaning of the calibration equation and the correlation coefficient, as well as how to calculate them and determine the statistical error associated with them.