Skill Builder Exercise 1

Error Analysis, Statistics and Graphing

This semester, most of labs we require us to calculate a numerical answer based on the data we obtain. A hard question to answer in most cases is how "good" is your answer. Today's activity will introduce you to this topic- usually called error analysis. For this one activity only, you can answer the questions on the handout. For the rest of the semester, you will need to do your work in your notebook.

Error Analysis

Percent error:

One way to determine how "good" your answer is to calculate the error. The error of a measurement is defined as the difference between the experimental and the true value. This is often expressed as **percent (%) error**, which is calculated as:

$$\textit{Percent error} = \left| \frac{Experimental - True}{True} \right| \times 100 \qquad (1)$$

where the bars on either side of the fraction denote the absolute value.

In chemical measurements, we try to eliminate errors, which can be divided into two broad types: systematic and random. *Systematic error* occurs regularly and predictably because of faulty methods or sampling techniques, defective instrumentation or calibration, and/or incorrect assumptions. *Random error* is governed by chance. Examples include a weighing error due to air currents or changes in temperature near a balance or current fluctuations for electronic instrumentation. Systematic errors always affect the measured quantity in the same direction, while random errors can make the measured quantity either too large or too small.

Accuracy is the closeness of agreement between a measured value and the true (or accepted) value. True values can never be obtained by measurement. However, we accept values obtained by skilled workers using the best instrumentation as true values for purposes of calculation or for judging our own results.

Precision describes the reproducibility of our results. A series of measurements with values that are very close to one another shows good precision. It is important to understand, though, that good precision does NOT guarantee accuracy!

Standard Deviation:

The **standard deviation** of a series of measurements that includes at least 6 independent trials may be defined as follows: let x_m represent an individual measured value, n be the number of measurements obtained, and $\bar{x}$ be the average or mean of the various independent trials or measurements.

Subtracting the mean from the individual measurements gives the deviation (d) of the measurement:

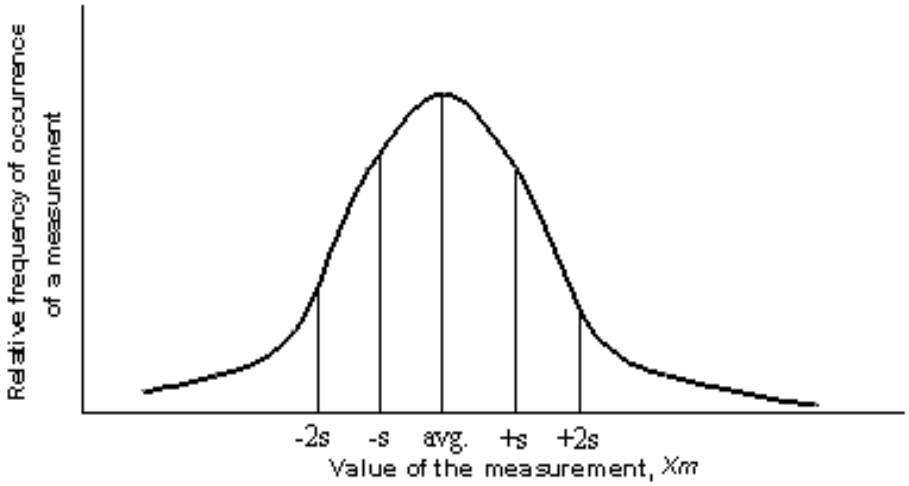
$$d = x_m - \bar{x} \qquad (2)$$

The standard deviation, *s*, is defined by:

$$s = \sqrt{\frac{\sum d^2}{n-1}} \qquad (3)$$

where $\sum d^2$ refers to the sum of all the values of d^2.

The standard deviation is used to indicate precision when a large number of measurements of the same quantity are subject to random errors only. We can understand the meaning of *s* if we plot the number of times a given value is obtained (on the y-axis) versus the values x_m (on the x-axis). Such a *normal distribution curve* is bell-shaped with the most frequent value being the average value $\bar{x}$.



If all that are present are random errors, most measurements result in values near the mean ($\bar{x}$) 68% of the measurements fall within one standard deviation while 95% of the measured values are found within two standard deviations (*2s)* of $\bar{x}$. The value of *2s* is called the uncertainty of the measurement. If we report our value of the measurement as the range $\bar{x} \pm 2s$, we are saying that $\bar{x}$ is the most probable value, and 95% of the measured values fall within this range.

Q Test:

For most of the experiments in this course, the standard deviation is impossible to reliably calculate because we perform too few measurements of a particular quantity. When there are less than 6 measured values, the ***Q Test*** is used to decide whether to reject suspected "bad" values as outliers. The Q test is a test to see if we can assume with greater than 90% confidence that one point is in error and can be thrown out. If so, the average and standard deviation can be recalculated without that point.

The equation for the Q test is:

$$Q = \frac{\left|\text{suspect - nearest (gap)}\right|}{\left|\text{largest - smallest (range)}\right|} \qquad (4)$$

n **(# of measurements)**	**3**	**4**	**5**
Q Test **value** **(90% probability)**	0.94	0.76	0.64

The use of the Q test can best be seen in an example. Suppose we measure the absorbance of a sample five times and are results are as follows.

Trial	Absorbance
Trial 1	0.223
Trial 2	0.354
Trial 3	0.246
Trial 4	0.219
Trial 5	0.231

The average and standard deviation of this data set are 0.255 +/- 0.051.

The suspect value is the one that is farthest away from the mean or average. It will always be the largest or smallest value. In our example, the value of 0.354 is the suspect value since it appears much higher than the other values.

The Q test will allow us to determine if the 0.354 value is statistically outside the normal level of variation within the data set and therefore can be thrown out or if we need to include it when determining the average and standard deviation.

To do the Q test, we need to determine the "gap" and the "range" of our data set.

The gap is the difference between the suspect value and the closest other data point while the range is the largest value in the data set minus the smallest value.

In our example,

$$Q = \frac{0.354-0.246}{0.354-0.219} = 0.800$$

The value of Q obtained is compared to the table at the top of this page. If the calculated value of Q (0.80 in our example) is more than the value of Q in the table (for five measurements, 0.64), the data point can be discarded and the average recalculated. This is the case in our example.

Graphical Representation of Data and the Use of Excel®:

Scientists answer questions by performing experiments which provide information about a given problem. After collecting sufficient data, scientists attempt to correlate their findings and derive fundamental relationships that may exist between the acquired data. Whether a set of measurements or variables are correlated can be examined by constructing a graph and calculating the coefficient of determination (also known as R^2). Microsoft Excel® is a program commonly used to construct a graph and calculate R^2. Instruction on how to use Excel® for graphing is given later.

Graphing:

A graph is a diagram that represents the variation of one factor in relation to one or more other factors. These variables can be represented on a coordinate axes.

The vertical axis is the y-axis (or ordinate), and the horizontal axis is the x-axis (or abscissa). When plotting a certain variable on a particular axis, experiments are normally designed so that you vary one property (represented by the *independent variable*) and then measure the corresponding effect on the other property (represented by the *dependent variable*).

All graphs should conform to the following guidelines:

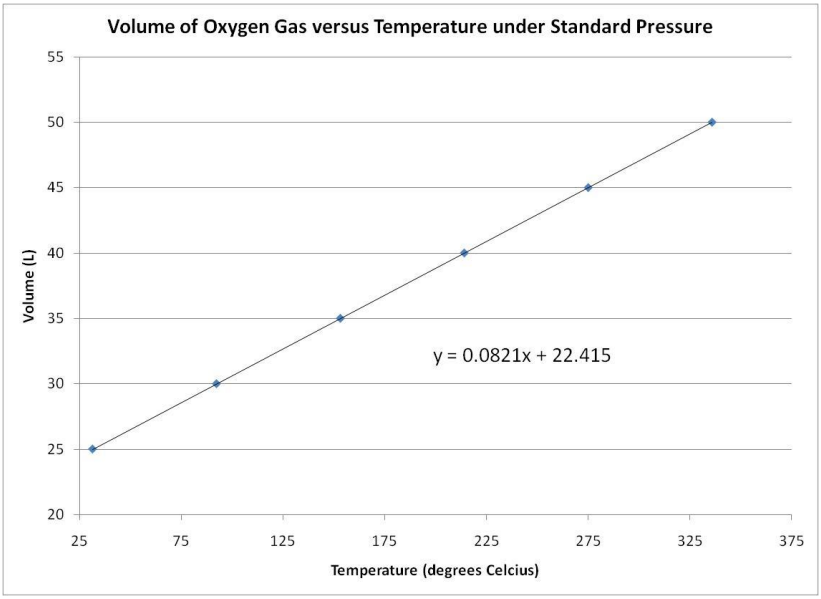
1. They should have a descriptive title.
2. The independent variable is placed on the horizontal axis; the dependent variable is plotted on the vertical axis.
3. Label both the vertical and horizontal axes with units clearly marked.
4. The scale chosen for the data should reflect the precision of the measurements. For example, if temperature is known to be +0.1 °C, you should be able to plot the value this closely. Moreover, the data points should be distributed so that the points extend throughout the entire page (as opposed to a small portion of the paper).
5. There should be a visible point on the graph for each experimental value.

Linear Graph:

Let us first examine a linear graph. Consider the following measurements made of an oxygen sample under standard pressure:

Volume (L)	**Temperature (^{o}C)**
25.00	31.49
30.00	92.38
35.00	153.28
40.00	214.18
45.00	275.08
50.00	335.97

Using graph paper or any graphing program such as Microsoft Office Excel®, one can first construct a plot of the data, where volume is determined to lie on the y-axis, and temperature is plotted on the x-axis. Once the data is plotted, a best-fitting line is constructed, and an equation of the line in slope-intercept form $\boldsymbol{y = mx + b}$ is formulated, where m = slope and b = y-intercept. That is,

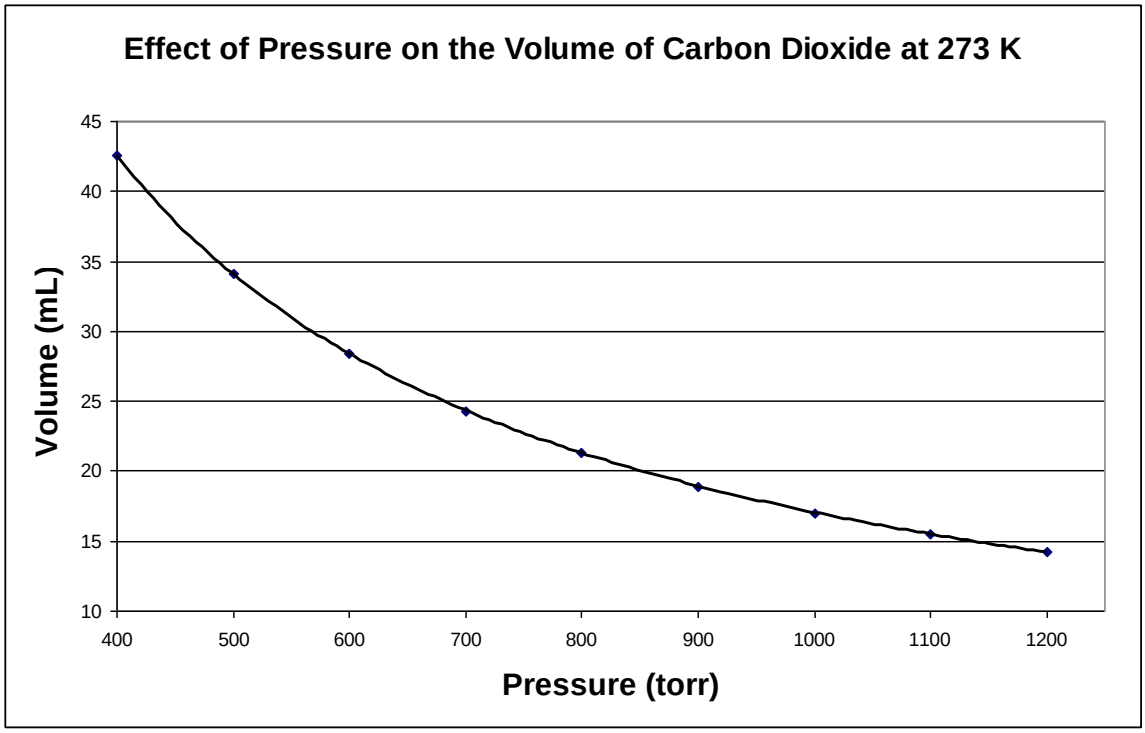


Non-Linear Graph:

Now examine an indirect function involving a hyperbola. Consider the following measurements made of a carbon dioxide gas sample at 273 K:

Volume (mL)	Pressure (torr)
42.6	400
34.1	500
28.4	600
24.3	700
21.3	800
18.9	900
17.0	1000
15.5	1100
14.2	1200

Once again, using graph paper or any graphing program such as Microsoft Office Excel®, one can construct a plot of the data, where volume is determined to lie on the y-axis, and pressure is plotted on the x-axis.

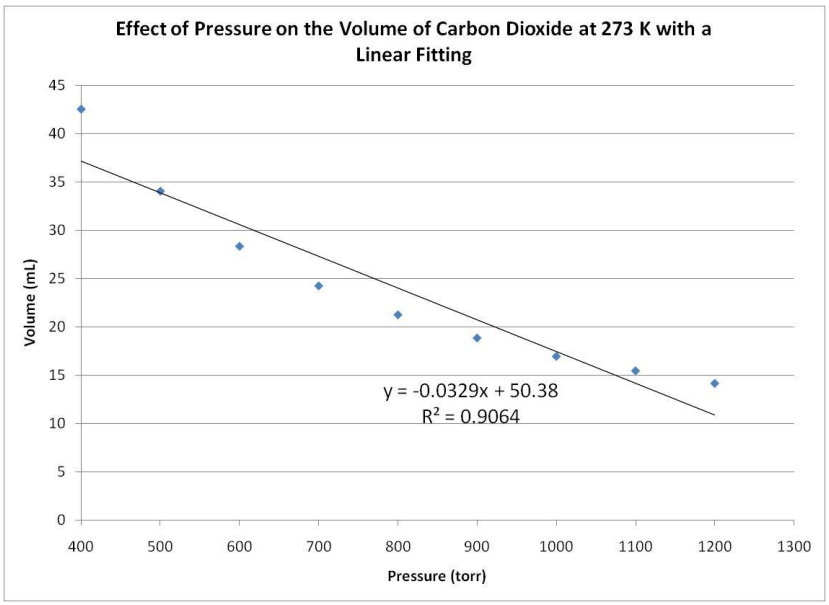


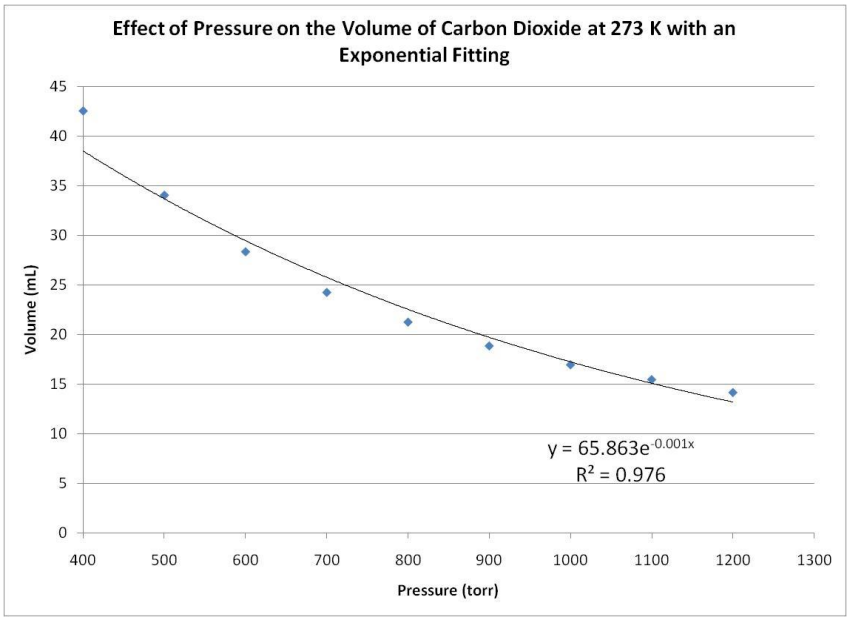
As depicted in the graph above, some chemical relationships are not linear; that is, there are no simple linear equations to represent such relationships. Instead, a plot of data for this kind of relationship gives a curved (non-linear) fit. Such a graph is useful in showing an overall chemical relationship, although the slope and the y-intercept are NOT useful for its interpretation.

Coefficient of Determination, R^2: Is x correlated with y?

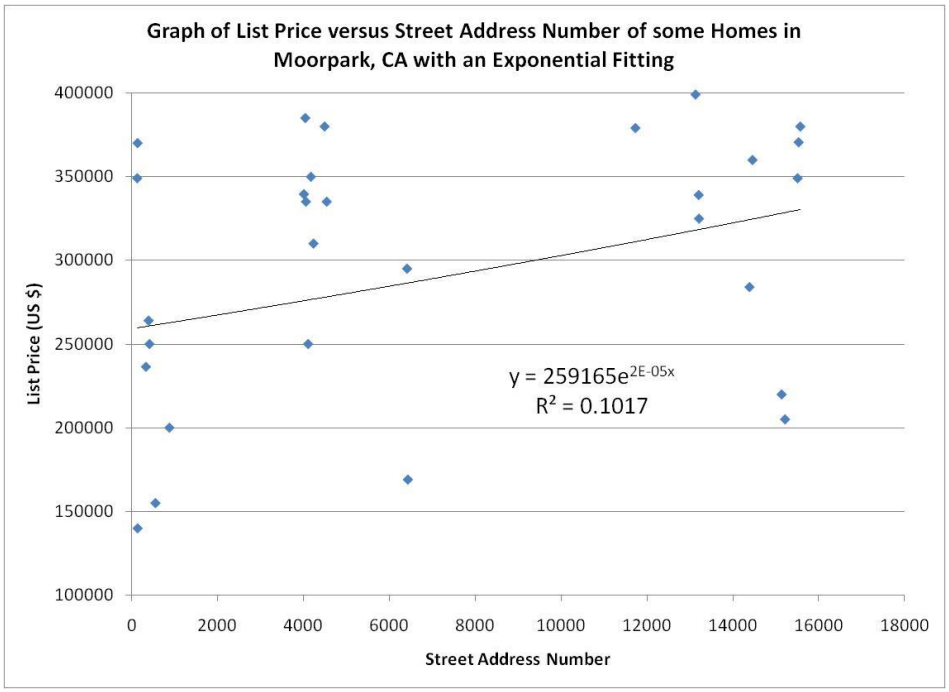
A set of (x,y) values are not always correlated in a linear or any other models/fittings. The coefficient of determination or the R^2 (or the Excel® function **RSQ)** is a measure of the correlation between the (x,y) variables. This coefficient of determination indicates how strongly a set of x values correlate with the corresponding set of y values. The R^2 value ranges from 0 to 1. A value of 1 means that data set perfectly fits a linear model or equation and value of 0 means that there is no correlation between x and y. A value of 0.8 means that 80% of the data fit the model/fitting.

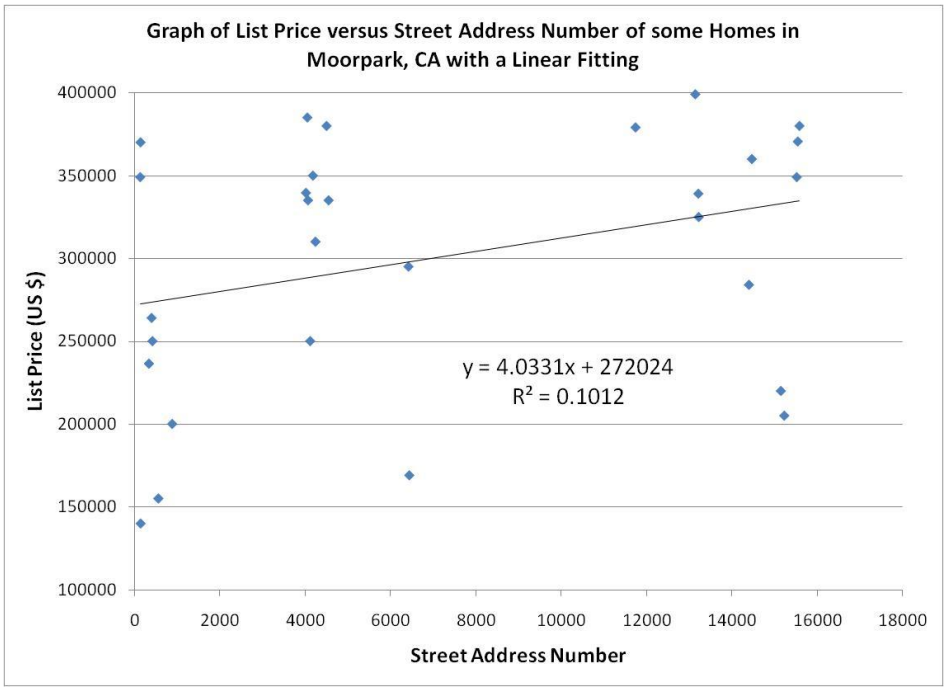
Let's examine the two graphs above (Volume vs Temperature and Volume vs Pressure). The R^2 value for the linear fitting of Volume vs Temperature is 1 (a perfect fit!). If a linear fitting is to be done on the Volume vs Pressure graph, an R^2 value of 0.9064 is obtained. Volume and pressure, in this case, are correlated but a linear model might not be the best fit. If an exponential fitting is used, an R^2 value of 0.976 is obtained (see graphs on the next page). This means that x and y are correlated and an exponential fitting better explains the correlation than a linear fitting.





One can also have a data set that is not correlated to each other. Note the two graphs below. The data gives the list prices of some of the homes for sale in Moorpark, CA and their corresponding street address number. Since street address numbers are not unique to a neighborhood, we can guess that there should not be any correlation between the two variables. The R^2 for the exponential and linear fittings are 0.1017 and 0.1012, respectively. These values are significantly lower than the ones discussed above. These low R^2 values demonstrate that there is no correlation (linear nor exponential) between list price and street address number.





Excel® calculates the R^2 value by taking the square of R (also known as Pearson Product Moment Correlation Coefficient) as defined by equation 5 below. An R^2 value equal to or greater than 0.99 "generally" means that the data has a "good" fitting to a linear model or equation.

$$r = \frac{\sum(x - x_{ave})(y - y_{ave})}{\sqrt{\sum(x - x_{ave})^2 (y - y_{ave})^2}} \qquad (5)$$

Excel® Graphing Procedure

Note: Various versions of Excel® may function a bit differently from the directions outlined

below. Please adjust accordingly. Two sets of instructions are provided for users with Microsoft Office Excel® 2007 and Excel® 97-2003, respectively. If you encounter difficulties, consult your instructor for assistance.

Microsoft Office Excel® 2007: Begin by typing the data onto your Excel® spreadsheet in (x, y) form. To plot the data, highlight the x/y coordinates, select ***Insert*** from the display menu followed by ***Scatter***, choose ***X Y (Scatter)***, and click ***OK***. To label the graph and coordinate axes, use the ***Layout*** function from the toolbar and select the appropriate ***Labels***.

Microsoft Office Excel® 97-2003: Begin by typing the data onto your Excel® spreadsheet in (x, y) form. To plot the data, highlight the x/y coordinates and select the ***Chart Wizard*** icon from the display menu. Work through the four steps of the Chart Wizard as shown below:

1. Chart Wizard – Step 1 of 4: Select ***XY (Scatter)*** and hit ***Next>***.
2. Chart Wizard – Step 2 of 4: Hit ***Next>*** a second time.
3. Chart Wizard – Step 3 of 4: Select ***Titles*** and give your graph an appropriate title, making sure to label both the x-axis and y-axis WITH UNITS IN PARENTHESES! Other options are available via various menu options, including the ability to delete the legend. When completed, hit ***Next>***.
4. Chart Wizard – Step 4 of 4: All graphs should be incorporated into your laboratory reports as full sheet inserts. Therefore, select the ***As new sheet:*** option.

Once your data is plotted, you can draw a best-fitting line for the data utilizing the trendline function. Place the mouse cursor on one of the data points and proceed to right-click. All the data points corresponding to a particular data set should now be highlighted. Right-click again, and various menu options will pop up; select ***Add Trendline…*** utilizing the left-click. If your graph involves a linear function, select ***Linear***. If your graph involves a non-linear function, select ***Power*** for this particular lab. Under the options menu, you can display an equation and r^2 value on your graph by selecting ***Display Equation on chart*** and ***Display R-squared value on chart***. Finally, click on ***OK*** (or ***Close***) and print your graph.

Problem Set

1: A student performs an experiment to calculate the specific heat capacity of copper. The student experimentally finds the answer 0.340 J/g°C. Looking up the accurate published value it is found to be 0.385 J/g°C. Solve for the student's percent error.

Recall that:

$$\textit{Percent error} = \left| \frac{Experimental - True}{True} \right| \times 100$$

2: Since 1965 dimes are composed of copper with 25% nickel on the outside. Previous to 1965, dimes were composed of 90.0% silver and 10.0% copper. The composition changed when the dime cost more in silver than it was worth.

A 1963 dime is weighed on ten different balances, and the mass was recorded below. Calculate the average and standard deviation as outlined below. (parts a-f)

Balance Number	**Mass (g) = x_m**	**$d = x_m - \bar{x}$**	**d^2**
1	2.495 g		
2	2.509 g		
3	2.507 g		
4	2.511 g		
5	2.508 g		
6	2.538 g		
7	2.512 g		
8	2.501 g		
9	2.510 g		
10	2.490 g		

a) Solve for the average value, ($\bar{x}$ **):**_____

b) Fill in the chart for all the d (deviations) and d^2 (deviations squared) values.

c) Solve for the standard deviation, s. $s = \sqrt{\frac{\sum d^2}{n-1}}$

d) Solve for the range, $\bar{x} \pm 2s$. (95% confidence interval) which the way answers are typically reported.

e) Check all data points against the range. Identify values outside the range that may be unreliable and discarded.

3: a) A student determines the concentration of a sodium hydroxide solution by titration with standardized KHP. S/he obtains the values: 0.190 M, 0.202 M, and 0.205 M. Should the value 0.190 M be rejected? Apply the *Q Test*. For three values Q must be greater than 0.94 to reject the number. Remember $Q = \frac{|\text{suspect} - \text{nearest}|}{|\text{largest} - \text{smallest}|}$

b) The student decides to repeat the experiment two more times. The five values now include: 0.190 M, 0.202 M, 0.205 M, 0.201M and 0.203M. Use the *Q Test* to see if the test if the first value may be rejected. For five values Q must be greater than 0.64 to reject the number.

c) Solve for the average molarity of the 5 measurements with and without the rejected number. Is there value in repeating an experiment several times?

4. (Take home assignment). A set of solution densities as a function of weight/volume % sugar is given below. Note that weight/volume % sugar refers to how many grams of sugar per 100 mL of solution. As an example, 9.000 % means that there are 9.000 g of sugar per 100 mL of solution. Use Excel® (or similar program) to construct a density (y-axis) versus weight/volume % sugar (x-axis) plot. Add a linear fit through the Add Trendline function and display the equation and the R^2 value on your chart. **Print out your graph.**

Examine your R^2 value and your plot. You will notice, upon visual inspection, that there are four data points that can be considered outliers. Remove these data points, one set at a time by highlighting and then deleting the x,y values on the columns. As you delete the outliers, one data set at a time, you will see that the graph, the equation and the R^2 change accordingly. Note how the R^2 value changes. By the last deletion, you will now have an R^2 value that is generally acceptable.

Print out the graph again without all four outliers, and submit to your instructor in lab next lab period. Graphs should conform to the five guidelines on page 3.

Be sure to display the equation and r^2 value.

weight/volume % sugar	density of solution (g/mL)
0.00	0.998
2.007	1.017
3.070	1.002
4.000	1.009
5.010	1.008
6.094	1.036
6.991	1.017
8.008	1.020
9.000	1.028
10.00	1.030
11.12	1.033
12.11	1.053
13.01	1.041
15.00	1.050
16.00	1.055
17.02	1.055
18.00	1.056
19.00	1.060
21.03	1.071
23.05	1.066
24.02	1.080

This lab has been adapted from the Moorpark College Chemistry Lab Manual.