Name: _____ Section: _____

Error Analysis, Statistics and Graphing Workshop

Percent 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:

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

Note: At times a true value may not be known or possible. If two experimental values are determined and the true value is unknown, then the percent difference may be calculated. The percent difference is the absolute value of the difference between the two experimental values divided by the average value and multiplied by 100.

$$\%\ \text{difference} = \frac{|\text{Value 1} - \text{Value 2}|}{(\text{Value 1} + \text{Value 2}) / 2} \times 100\ \%$$

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 more difficult to define and is governed by chance. Examples include a weighing error due to air currents or changes in temperature near a balance and line 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 is a sign of good precision. It is important to understand, though, that good precision does NOT guarantee accuracy!

Name: _____ Section: _____

Standard Deviation:

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

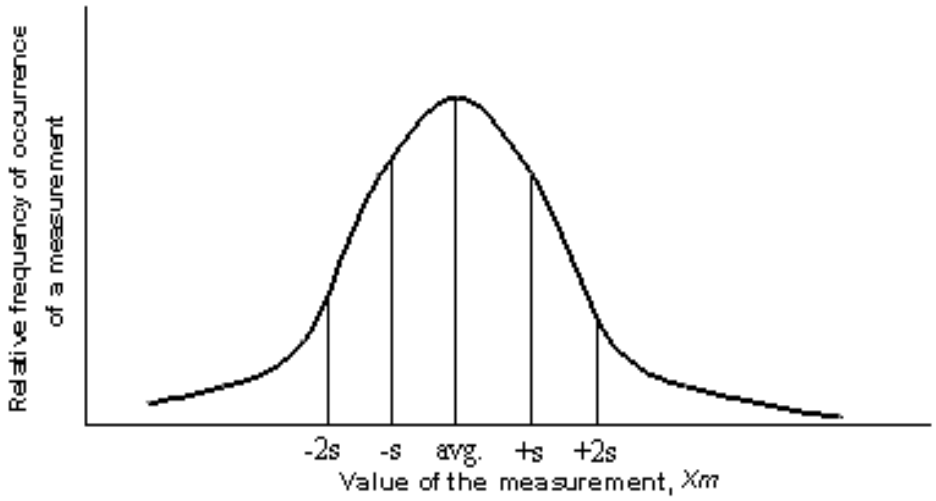
$$d = x_m - \bar{x} \qquad (2)$$

and 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 $\bar{x}$.



Most measurements result in values near $\bar{x}$; 68% of the measurements fall within one standard deviation while 95% of the measured values are found within 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.

$$\text{range} = \bar{x} \pm 2s \qquad (4)$$

Name: _____ Section: _____

Q Test:

For most of the experiments in this course, the standard deviation is impossible to calculate because we perform too few measurements of a particular quantity. When there are such few measured values (< 6), the ***Q Test*** is used to decide whether to reject suspected "bad" values as outliers.

$$Q = \left| \frac{\text{suspect} - \text{nearest}}{\text{largest} - \text{smallest}} \right| \qquad (5)$$

***n* (# of measurements)**	**3**	**4**	**5**	**6 or more**
***Q Test* value (90% probability)**	0.94	0.76	0.64	Do not use

If *Q* for a set of data is larger than those listed in the table, the suspect value may be rejected, and the average of the other values is reported. Please note that the sign of *Q* is NOT important since we are using the absolute value.

Graphical Representation of Data and the Use of Excel®:

Scientists answer posed 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:

Graphical representations of data illustrate relationships among data visually. 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 conventionally 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.

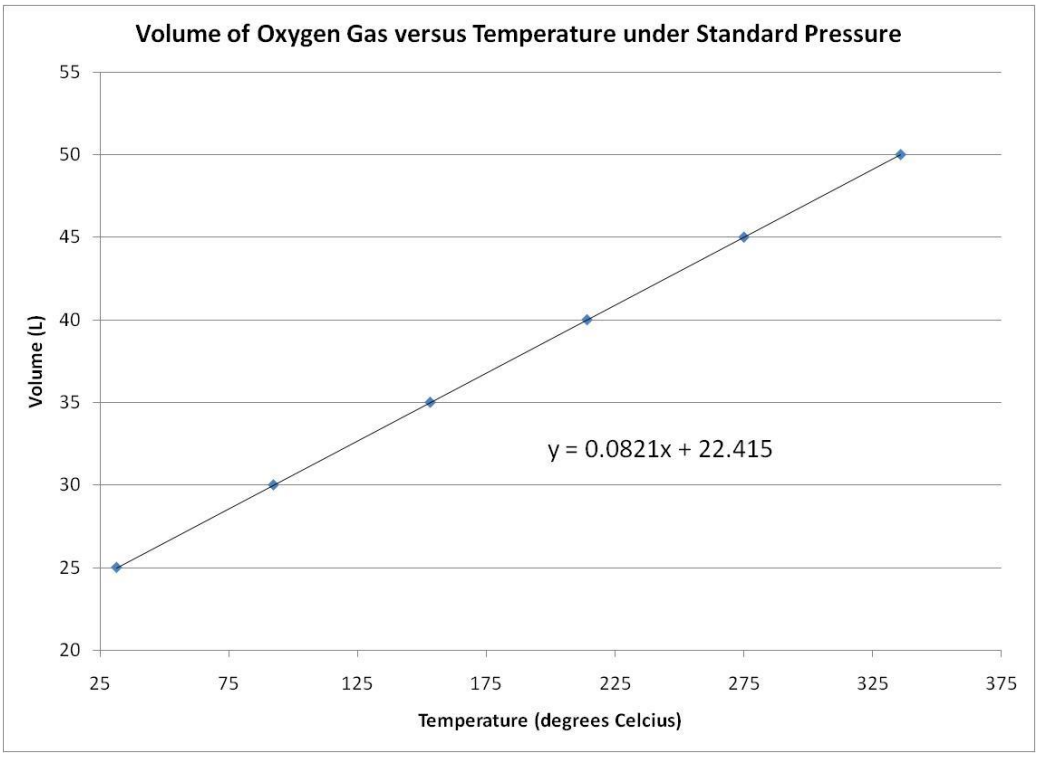
Name: _____ Section: _____

Linear Graph:

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

Temperature (°C)	**Volume (L)**
31.49	25.00
92.38	30.00
153.28	35.00
214.18	40.00
275.08	45.00
335.97	50.00

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,



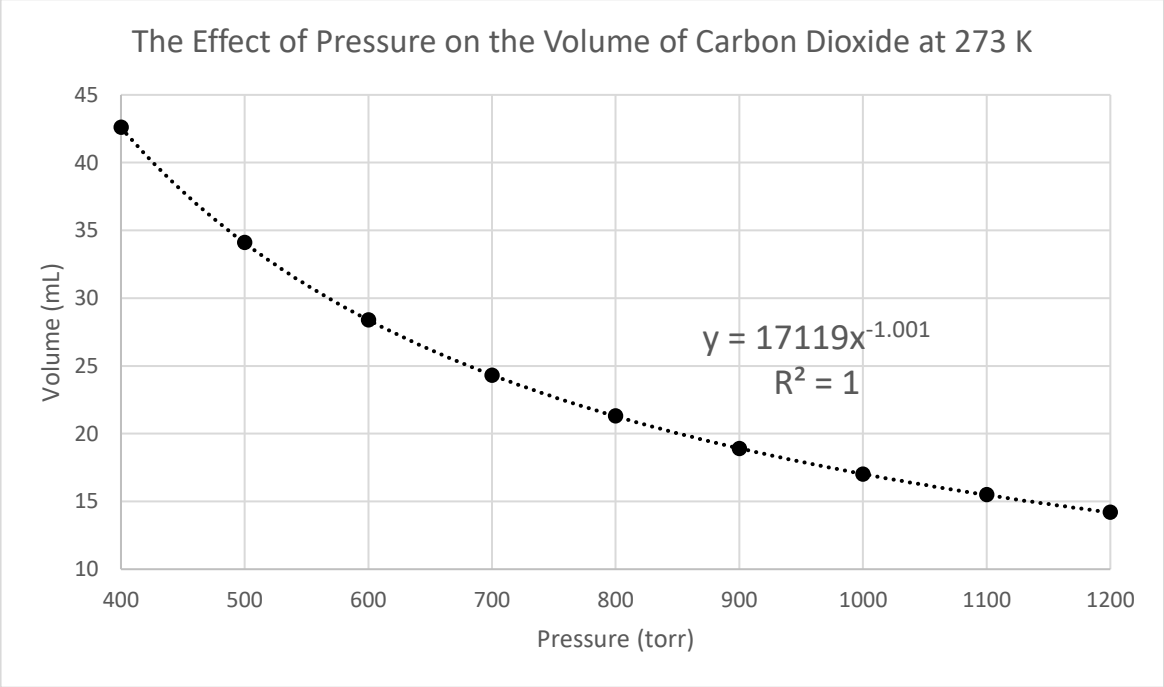
Name: _____ Section: _____

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:

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

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.



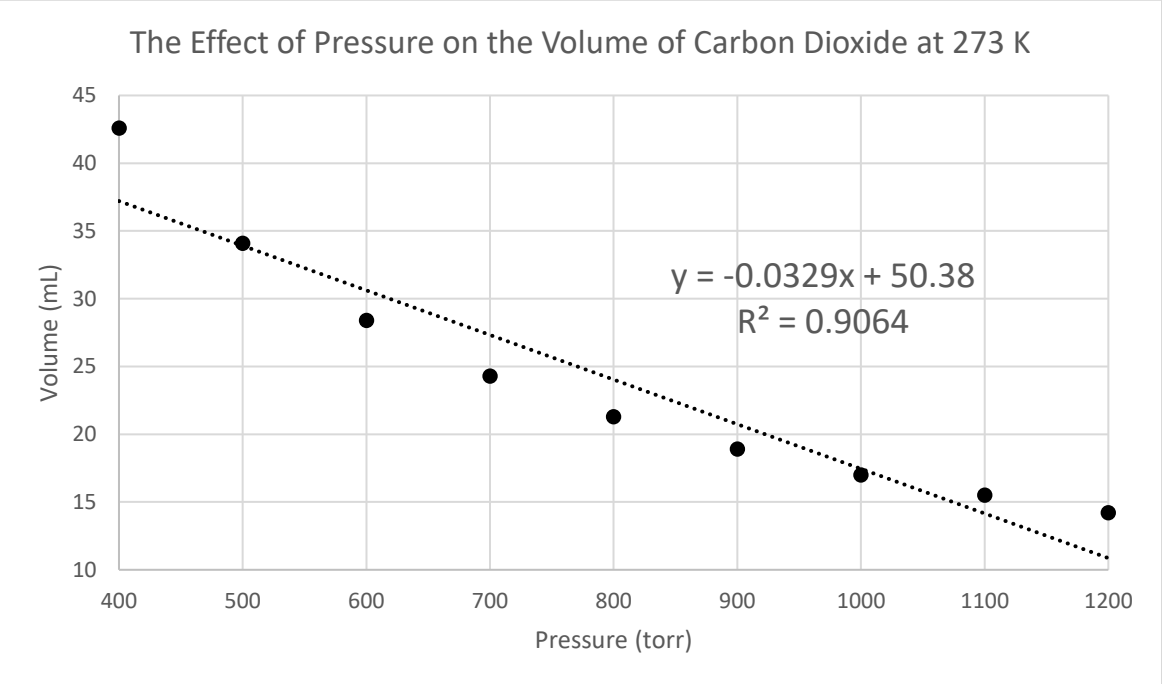
As depicted in the graph above, some chemical relationships are not linear; that is, there are no simple linear equations to represent such relationships. Instead, a plot of data for this kind of relationship gives a curved (non-linear) fit. Such a graph is useful in showing an overall chemical relationship,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

Name: _____ Section: _____

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

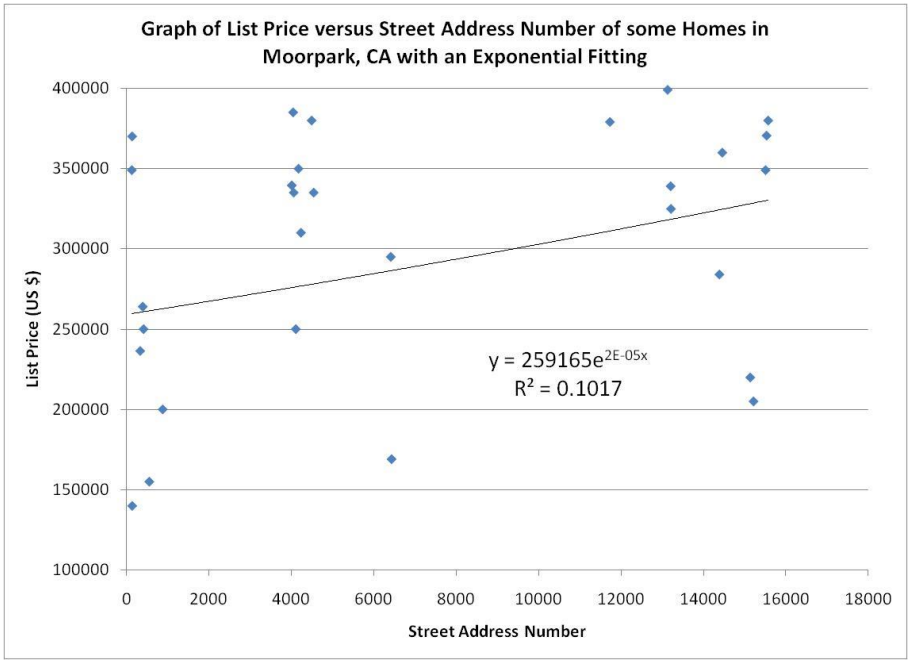
A set of (x,y) values are not always correlated in a linear or any other models/fittings. The coefficient of determination or the r^2 (or the Excel® function **RSQ)** is a measure of the correlation or linear dependence (in the case of a linear fitting) between the (x,y) variables. This coefficient of 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.

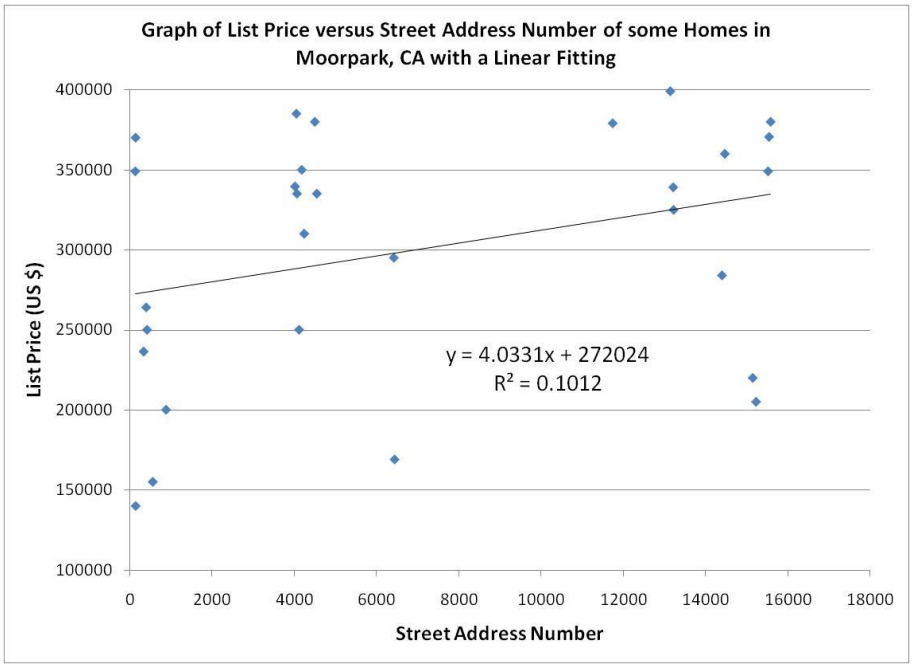
Let's examine the two graphs above (Temperature vs Volume and Pressure vs Volume). The r^2 value for the linear fitting of Temperature vs Volume is 1 (a perfect fit!). If a linear trendline is used on the Pressure vs Volume graph, an r^2 value of 0.9064 (below) is obtained. Pressure and volume, in this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



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 easily conclude 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), 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 the trendline equation.

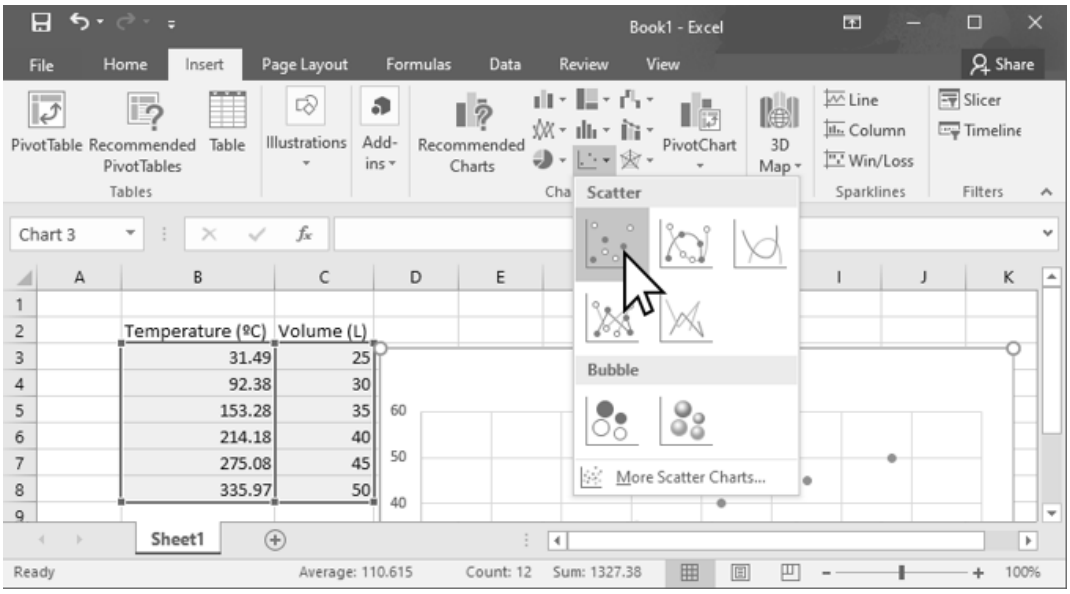
$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 (y-\bar{y})^2}} \qquad (6)$$

Name: _____ Section: _____

Excel® Procedure

Note that various versions of Excel® may function a bit differently from the directions outlined below (which work on department-owned laptop computers):

Put the title for your x-axis (include units) in one Excel® cell (box). In the cell to the right, put the title for your y-axis. Using these boxes as headings, input the numeric data (like a table) in the cells under these titles (each box should contain one number; each row represents one data point in x,y format). Click and drag your mouse to highlight just the numeric boxes. From the "Insert" tab, choose a "Scatter" plot. (See example, below.)



Your graph must include a meaningful Chart Title and Axis Titles (with units). These Chart Elements can be added to your graph by clicking on the "+" icon in the upper right corner of your graph. Your instructor may request additional Chart Elements.

To add a Trendline, right click on any data point on your graph and choose "Display Trendline" from the menu that appears. The format trendline pane will appear on the right side of your screen. Linear should be selected by default, but other types of trendlines may be tested. From this pane, you should check the box next to "Display Equation on chart." Your instructor may also ask you to check the box for "Display R-squared value on chart."

Name: _____ Section: _____

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.

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

2. Since 1965, dimes are composed of copper with 25% nickel on the outside. A Roosevelt Type dime (1946 to 1964), designed by John R Sinnockis, is 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 is recorded.

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 and d^2 values.

Name: _____ Section: _____

c) Solve for the standard deviation, s.

$$s = \sqrt{\frac{\sum d^2}{n-1}}$$

d) Solve for the range, $\bar{x} \pm 2s$.

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

f) Do you suspect the differing values are due to random errors or systematic errors or possibly both? Explain. How might you test your hypothesis?

g) Solve for the new average value, $\bar{x}$, removing values outside the range in part (e).

Note: Once unreliable data points are discarded the process should be repeated to recalculate $\bar{x}$, d, s, and range values. Only the recalculation of $\bar{x}$ is required for the problem set today, but you are encouraged to recalculate d, s and range on your own.

Name: _____ Section: _____

3a) 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.

$$Q = \left| \frac{\text{suspect} - \text{nearest}}{\text{largest} - \text{smallest}} \right|$$

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 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 measurements from part b with and without the rejected number. Is there value in repeating an experiment several times?

Name: _____ Section: _____

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. 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 two graphs, *(1) use all data points, (2) without all four outliers,* and submit to your instructor in lab next week. Graphs should conform to the five guidelines given in the introduction.

   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