(Last Accessed 9-5-2019)

Appendix II
Review of Math Skills and Data Analysis

Scientific Notation

Since much of what a chemist is concerned about involves submicroscopic particles (e.g., atoms, electrons, and protons), we often need to write very large or small numbers. After all, there are 602,000,000,000,000,000,000,000 atoms in a mole. Writing numbers like Avogadro's number can be avoided by using scientific notation. Scientific equation uses powers of ten to more simply represent numbers. For example, Avogadro's number can be written 6.02×10^{23}. There are many other ways to write the same number using powers of ten (e.g., 602.0×10^{21}). This is called *exponential notation*. *Scientific notation*, like all exponential notation, is written as a product of two separate parts. The first is a number between 1 and 9.999 and the second is a multiple of 10 express by an exponent. Numbers less than one are written with negative exponents:

Examples:

$940{,}000 = 9.4 \times 10^5$ $\quad$ $0.0000000357 = 3.57 \times 10^{-8}$

Logarithms

Logarithms are often used in science to represent data that covers large ranges (e.g., x changes from 1×10^{-3} to $1\text{x}10^{-12}$). This data would be difficult to graph on one page and still read the numbers. In order to graph data like this conveniently, we use the logarithm function.

The logarithm of a positive number is the exponent or power of a given base that is required to produce that number. For example, in base 10 (which is used *for common logarithms*), $10^1 = 10$ so the logarithm of 10 is 1; $10^2 = 100$ so the logarithm of 100 is 2. In other words, ask yourself, "100 is ten to the *what* power?" Logarithms of multiples of ten are easy to figure out in your head but for other numbers you will need a table (in the olden days) or a calculator. Numbers between 10 and 100 will have logarithms between 1 and 2. Numbers less than 1 will have negative logarithms ($0.1 = 10^{-1}$, so $\log 0.1 = -1$). *Natural logarithms* use the base e ($e = 2.71828\ldots$). Converting common logarithms (log) to natural logarithms (ln) involves a conversion of 2.303:

$$\ln X = 2.303 \log X$$

The reverse operation can be performed by taking the "antilogarithm." This amounts to determining what number corresponds to the given logarithm. In other words, ask yourself, "ten to this power is *what*?" (E.g., the "antilogarithm" of –1 is 0.1). Your calculator will most likely use an inverse function to calculate either the common or the natural logarithms.

Because logarithms are exponents, all properties of exponents are also properties of logarithms. For example:

1. The log of a product of two numbers is equal to the sum of the logs of those numbers ($\log A \times B = \log A + \log B$);

2. The log of the quotient of two numbers is equal to the difference between the log of the numerator and the log of the denominator (log A/B = log A – log B);
3. The log of the n^{th} power of a number is equal to n times the log of the number (log $A^n = n$ log A); and
4. The log of the n^{th} root of a number is the log of the number divided by n (log A^{-n} = (log A)/n).

Graphs

Graphs are used by scientists to get a visual representation of data. Graphs show how two variables are related by plotting one against the other. The plotted points are then connected by a smooth curve or a straight line. Usually data are plotted in a Cartesian coordinate system in which two axes are drawn perpendicular to each other, the horizontal one being the *x*-axis and the vertical one being the *y*-axis. Each axis is divided into scaled units usually of equal length. The scale on the *x*-axis does not need to be the same as the scale on the *y*-axis. The variable over which we have little or no control, the independent variable, is plotted on the *x*-axis while the dependent variable is plotted on the *y*-axis. Once the data is plotted and a smooth curve is drawn through the points, a mathematical equation can be determined that defines the line and, therefore, the relationship between the variables. On a straight-line graph values beyond those plotted can be predicted by extrapolation and values within the plotted data can be predicted by interpolation.

General Graphing Procedure

Use the entire page when making a graph, leaving space on the left side and the bottom for labeling the axes. It is not necessary to start at 0. Use a range that represents your data. For example, if your data ranges from 255 kJ/mole to 890 kJ/mole, your axis should range from 250 to 900. Label set intervals on each axis (every 10 kJ/mole may be appropriate for the above example). Be sure you plot the independent variable along the *x*-axis and the dependent variable along the *y*-axis. Label the axes clearly with the property plotted and the units used. Plot the data and draw the best straight line or smooth curve.

Interpreting Graphs

Straight-line graphs correspond to linear equations. The general formula for a line is $y = mx + b$ where m is the slope of the line and b is the y intercept. The slope of the line is defined as the increase in y (Δy) for a given increase in x (Δx) or

$$\text{Slope} = m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x}$$

Curved lines do not fit the general equation for a line because on a curve the slope is constantly changing. Sometimes a set of data that yield a curve can be manipulated to give a straight line. For example, if the volume of a gas is plotted against its pressure, the resulting graph will be curved. However, if the volume is plotted against 1/pressure the result is a straight line. Often it is desirable to find a straight-line relationship for data that is curved. The function that allows the data to fall on a straight line can often give important information about the system.

Data Analysis

Numbers by themselves have little or no meaning. Scientists are constantly analyzing patterns, relationships, and accuracy in numbers. To understand anything about a number it must have units. The number 6 has no meaning by itself, but 6 feet or 6 gallons convey something about the measurement taken. How good is the measurement? All measurements are subject to error. Experimental measurements may only be as precise as the measurements used to obtain them. Results should include an estimate of the experimental error involved.

Significant Figures

One way to indicate uncertainty in a measurement is to use significant figures. The last digit reported in the number is the digit that is estimated, the one in doubt. For example, when reading a ruler that has millimeter markings on it, you can estimate between the lines giving a measurement of 2.543 with uncertainty in the last digit. Zeros are significant unless they are used to position the decimal point. The number 5,700 does not contain enough information to determine the number of significant figures. Use scientific notation to show significant figures in cases like this (5.70×10^3).

Measurement	Significant Figures	Notes
3.794	4	
2.940	4	trailing 0 significant
0.0078	2	leading 0's position decimal point
1.36×10^{12}	3	first factor show significant figures

The uncertainty of a measurement affects the uncertainty of any quantity calculated from that measurement. The final result should indicate the overall uncertainty. The uncertainty depends on the operations carried out.

1. *Addition/Subtraction* – Look at the numbers involved in the calculation. The answer is certain only to the decimal place of the least certain number involved. Calculate all trailing digits and then round to show the correct number of significant figures. For example:

   15.783
   12.08 ← hundredth place is least certain
   0.0796
   27.9426 or 27.94

2. *Multiplication/Division* – Look at the numbers involved in the calculation. The answer can contain only the number of significant figures of the least certain number involved. For example,

$$\frac{3.13 \times 12}{5.792} = 6.4848 = 6.4$$

   two sig. figs.

3. *Logarithms/Antilogarithms* – The numbers to the right of the decimal point in the log should have the same number of significant figures as the original number. Likewise, for the antilog, the number of digits to the right of the decimal equals the number of digits in the answer. For example,

$$\log 4.378 = 0.64127 = 0.6413$$
$$\log 4.37 \ = 0.64048 = 0.640$$
$$\log x = 5.328 = 212813.90 = 2.13 \times 10^5$$

Rounding is necessary when the mathematical result produces more digits than are significant. If the portion of the number to be discarded begins with 4 or less, digits are dropped. If the portion of the number to be discarded begins with 5 or greater, the preceding digit should be incremented by one unless the discarded part is exactly 5 (e.g., 500). In that case, the preceding digit is left unchanged if it is even and incremented if it is odd. For example, rounding to the nearest 0.01 would result in

1.79485 becoming 1.79,
0.782 becoming 0.08,
5.5653 becoming 5.57,
5.5650 becoming 5.56, and
5.5550 becoming 5.55

Rounding should be done as a final step in the calculation. Do not round at each step or you will loose accuracy.

Types of Errors

We all know that errors exist in scientific data acquisition. How important the error or uncertainty is depends on the situation. The normal range, or acceptable blood sugar levels, in humans is quite wide before low or high blood sugar is diagnosed. This gives the clinician a fairly wide error margin for any determination. However, some drugs given for cancer treatment are very toxic to humans and the blood levels have to be monitored quite accurately. There may be a small window in which the drug is at high enough levels to kill cancer cells without killing the patient. For obvious reasons the clinician in this case may not be allowed much error in an analysis.

Errors fall into three main types:
1) indeterminate errors
2) determinate errors
3) gross errors

Indeterminate errors are random errors that are statistically inherent in repeated measurements. Indeterminate errors are often reflected in the precision of an expert analysis. These errors are usually symmetrically disposed about the mean.

Determinate errors are errors in measurement that can be ascribed to some cause. These kinds of errors usually cause all the measurements to be high or low of the true value. Sources of determinate errors are instrument errors, method errors and personal errors. Determinate errors are usually much larger than indeterminate.

Gross errors are just that - large errors that occur only occasionally. These are usually due to carelessness, laziness or inexperience on the part of the experimenter. These include reading scales incorrectly, transposing numbers in lab books, spilling solutions, etc.

Calculating and Reporting Error

When collecting data in lab, you must be able to ascertain if it is “good” data. All data has uncertainty in it. Learning how to describe the uncertainty in your data and to estimate the reliability of your data when you report it to others is a crucial skill. Usually a scientist does more than one run on an experiment, usually 2-6 repetitions. This allows her/him to calculate a central value and look at the variation in the data set. You can use the median or mean as the central value of a data set.

The median is the middle value (odd # data points) in an ordered data set. The mean or average of a data set is calculated by dividing the sum of the replicate measurements (data points) by the number of replicates done (# data points). It is defined as

$$\bar{x} = \frac{\Sigma_i x_i}{n}$$

$\bar{x} = mean$

$x_i = individual\ datum$

$n = \#\ in\ population$

$\Sigma = sum$

The standard deviation indicates how closely the data are clustered about the mean. For a finite set of data, the standard deviation is defined as

$$s = \sqrt{\frac{\sum_i \left(x_i - \bar{x}\right)^2}{n-1}}$$

Variation in lab data is expected and there are several ways to describe it. These different ways to describe uncertainty in data say something about the precision of the data. Precision is the agreement between two or more measurements that have been obtained in exactly the same fashion.

Range is the difference between the largest and smallest value in the data set.

Relative Standard Deviation (RSD): This is the standard deviation divided by the mean. It is usually expressed in parts per hundred (%) or parts per thousand (ppt).

$\%RSD = \frac{s}{\bar{x}}(100)$: also know as the coefficient of variation (CV)

RSD is used frequently by chemists because it gives a clearer picture of data quality than just the standard deviation. Why?

A scientist must know how to describe data to others. Normally you report x ± s as a minimum piece of information and RSD, if you like. Are you still wondering how well you really did? If so, you want to know how accurate you were. Accuracy is how close a set of data comes to the accepted value or "true value". We can never know the real true value unless you have reason to know that the unknown was made correctly or there is some absolute standard for comparison. There are two ways to describe your accuracy.

Absolute error: $\bar{x} - \mu$, where μ is the "true value". Note that you don't take the absolute value of this difference. You report + or - to indicate whether your value was high or low.

% Relative error: This is the absolute error divided by the "true" value, μ, and reported as a %.

$$\text{\% Relative error} = \frac{\bar{x} - \mu}{\mu}(100)$$

This can also be reported in parts per thousand relative error. Relative error is the descriptor used most often to describe your accuracy.

Confidence Interval and Applying Statistics

The exact true mean, μ, can never be known without an infinite number of measurements. Instead we use statistical theory to set limits around the experimental mean, x, wherein μ is expected to fall with a given degree of probability. These limits are called confidence limits or confidence intervals.

For example, the use of significant figures plays an important role when reporting the results of a quantitative analysis. In general, the significant figures can never imply that a result is known more accurately than the uncertainty of the measurement indicates. Example: 1237.37 ± 0.4 (CV) is incorrect; 1237.4 ± 0.4 (CV) is correct.

$$\text{Confidence limit for finding } \mu = \bar{x} \pm \frac{t \cdot s}{\sqrt{n}}$$

$\bar{x}$ = *mean*
s = *standard deviation*
n = *# data points*
t = *student's t (see table below)*
(student was the pseudonym for W.S. Gossett, a statistician)

(Last Accessed 9-5-2019)

Values of Student's *t*

Degrees of Freedom (n-1)	Confidence Level (%) 50	90	*95*	98	99	99.5	99.9
1	1.000	6.314	*12.706*	31.821	63.657	127.32	636.619
2	0.816	2.920	*4.303*	6.965	9.925	14.089	31.598
3	0.765	2.353	*3.182*	4.541	5.841	7.453	12.924
4	0.741	2.132	*2.776*	3.747	4.604	5.598	8.610
5	0.727	2.015	*2.571*	3.365	4.032	4.773	6.869
6	0.718	1.943	*2.447*	3.143	3.707	4.317	5.959
7	0.711	1.895	*2.365*	2.998	3.500	4.029	5.408
8	0.706	1.860	*2.306*	2.896	3.355	3.832	5.041
9	0.703	1.833	*2.262*	2.821	3.250	3.690	4.781
10	0.700	1.812	*2.228*	2.764	3.169	3.581	4.587
15	0.691	1.753	*2.131*	2.602	2.947	3.252	4.073
20	0.687	1.725	*2.086*	2.528	2.845	3.153	3.850
25	0.684	1.708	*2.068*	2.485	2.787	3.078	3.725
30	0.683	1.697	*2.042*	2.457	2.750	3.030	3.646
40	0.681	1.684	*2.021*	2.423	2.704	2.971	3.551
60	0.679	1.671	*2.000*	2.390	2.660	2.915	3.460
120	0.677	1.658	*1.980*	2.358	2.617	2.860	3.373