Introduction to Measurement
Chem 313

Accuracy and Precision

This course is all about measurement. We will be performing some classic assays to measure the concentrations of all kinds of analytes by various techniques. Every assay can be characterized by its accuracy and its precision. To explain fully the difference between accuracy and precision, let's use the following example.

A colorimetric assay is used to measure the concentration of Co^{2+} in treated wastewater effluent being released into a local stream. Twenty samples are obtained and measured using the assay. The reported $[Co^{2+}]$ is the mean of the 20 measurements. Accuracy is a measure of how close this mean is to the true value. In practice, one would confirm the accuracy of an assay by either measuring a sample with a known concentration or analyzing the sample with a previously established assay. Accuracy can most often be improved by using careful calibration procedures. For example, we often construct standard curves for colorimetric assays. When we do this, we are in essence calibrating our method so that we can obtain accurate results. Once the assay is designed and a calibration mechanism is in place, our attention turns to precision.

Standard Deviation

The precision of our colorimetric assay is a measure of variability among the 20 measurements. It is standard practice to use standard deviation (s) as a measure of this variability.

$$s = (\{\Sigma[(x_i-m)^2]\}/n)^{1/2}$$

n = number of measurements, 20

x_i = each of the measurements, the 20 $[Co^{2+}]$

m = mean of the measurements

Essentially, this is a measure of the average absolute deviation. The difference is squared so that negative and positive differences don not cancel each other out in the calculation. We will spend most of our efforts in this class on using the concept of standard deviation to report the precision of our measurements.

As it turns out, the standard deviation can be manipulated to provide us with very specific probability measurements of the value of the true mean. The true mean should not be confused with the actual value. The actual value, as defined above in our example, is the real concentration of $[Co^{2+}]$ in the wastewater. The true mean, μ, is the average of an infinite number of measurements. The difference between the true mean and the actual value depends on the accuracy of the assay. In designing the assay the goal is to minimize this difference. In performing the assay the goal is to *estimate* the true mean by

taking a *few measurements* (because taking an infinite number of measurements would take too long).

Normal (Gaussian) Distributions

If we could take an infinite number of measurements, we could prepare a frequency graph, which plots the number of measurements that fall within various intervals. This would result in a Normal distribution. This distribution is very unique and can be mathematically formulated. There are special qualities inherent to any Normal distribution. A Normal distribution can be completely characterized by its mean, μ, and standard deviation, σ. The Greek letter, σ, is used to indicate it is the true standard deviation of an infinite distribution. The mean, μ, is given by the maximum of the Normal distribution, and $\mu \pm \sigma$ defines the range in which 2/3 or 66.7 % of the measurements fall within.

By making a finite number of measurements we can estimate both μ and σ. The mean of the finite number of measurements is an estimate of μ. Below we will discuss how we can quantify the accuracy of this estimate. σ can be estimated using the following equation for s_e, the estimate of the standard deviation.

$$s_e = (\{\Sigma[(x_i-m)^2]\}/(n-1))^{1/2}$$

Student's t test

It is useful to be able to state with a meaningful degree of confidence how sure we are that the true mean, μ, falls within a given range. To do this we use the t-table in conjunction with the following equation.

$$\mu = \pm t \cdot s_e/n^{1/2} \text{ or } \pm t \cdot s_m$$

The variable t is obtained from the Student's t-table (Table 4-2 in your textbook), $s_e/n^{1/2}$ is the standard error of the mean, s_m. You must first calculate the degrees of freedom appropriate for your given statistical analysis, and then decide the appropriate level of confidence you want to report. For simple analyses the degrees of freedom (df) is n-1. You loose one degree of freedom when you calculate the mean. For some situations, such as least squares regressions, n-2, is appropriate.

For example, if ten measurements were performed giving a mean, m, of 2.32 and an estimate of the standard deviation, s_e, of 0.64, we determine that the degrees of freedom is 9 (n-1) and $t_{95\%CL}$ is 2.262 (read from Table 4-2). We calculate.

$$2.32 = \pm (2.262)\cdot(0.64)/(10^{1/2}) = 0.46$$

Thus, we are roughly 95 % certain that the true mean is somewhere in the range of 2.32±0.46 (or somewhere between 1.86-2.78).

Let's add a word on the basis of the student t-test. In fact, as you will soon find out, this is part of the focus of the first experiment. If we were to take an infinite number of measurements and divide them into groups of ten. We could calculate the mean and estimate of the standard deviation, and the estimate of the standard deviation of the mean, s_m, for each group of ten. We could also calculate the true mean. We could then calculate a t value for each group of ten using the following equation,

$$t_{calc} = (\mu - m)/s_m$$

where μ is the true mean, m is the mean for a given set of ten, and $s_{m(ave)}$ is the average estimate of the standard deviation of the mean among the sets of ten (in other words, for each set of ten, an s_m can be calculated and then the average s_m can be calculated). So now we have a long list (in fact an infinite list) of t_{calc} values. We could construct a frequency plot of these t_{calc} values, similar to the Normal distribution described above.

This plot technically is not a Normal distribution, but it does have properties that are very specific to any group of t values that were derived from sets of ten, df =9. There is a maximum at t = 0, and the curve gradually approaches a frequency of zero at each of the sides under the curve. It can be shown mathematically that 95 % of the area under this curve falls within -2.262 to 2.262. This means that for any set of ten measurements there is a 95 % chance that the corresponding t_{calc} falls within this range. *The corollary to this is that for any set of ten measurements, there is a 95 % chance that the true mean falls within* ± $t_{calc}s_m$. Now that is truly a powerful statement. The t-distribution described above is unique for df = 9. There are different distributions that correspond to different degrees of freedom. The values for t listed in the t-table (Table 4-2 in your textbook) derive from these t-distribution plots at various degrees of freedom.

In our first experiment we will have a collection of 10000 numbers acquired from summing the result of tossing five dice 10000 times (a feat performed by Excel). You will divide the entries into 1000 sets of ten and calculate the m, s_e, s_m and t_{calc} for each of the 1000 sets. You will determine the percentage of these sets that have t_{calc} values that lie outside the range of 2.262 and -2.262 and compare it to the expected 5%.

It is hoped that your experience with this experiment will help you to appreciate why you can definitively say something about your confidence in estimating the true mean from making a finite number of measurements.

Linear Least Squares Regression Analysis

One of the foundations of science is the ability to describe mathematical relationship between variables. To accomplish this we usually seek to control one of the variables (the independent variable) and measure the other variable (the dependent variable) and determine whether there is a statistically meaningful correlation between the two variables. (I will come back to the issue of correlation shortly). If it becomes evident that there is a correlation, we often seek to define the mathematical relationship between the independent and dependent variables. Often the relationship is either linear or can be linearized, such that a plot involving the some function of the dependent variable on the y-axis and the independent variable on the x-axis shows a linear relationship. We use a procedure called Least Squares Regression Analysis (LSRA) to define the slope and intercept of this linear relationship and, as we shall see, the uncertainties associated with interpolating such plots. LSRA calculates the slope and intercept of a line that minimizes the sum of the squares of the deviations between the measured y values, y_i, and the **y** value from the regression line.

For instance, for the following set of data.

x	Y
1	0.168
2	0.254
3	0.524
4	0.556
5	0.821
Regression Statistics	
R	0.976381
R^2	0.95332
Standard Error (s_y)	0.064964
Intercept (b)	-0.0178
Slope (m)	0.1608

Here x_i and y_i are the pairs of x and y data, where x_i is the independent variable, y_i is the dependent variable, and **y**'s are the y values calculated from the best fit line drawn through the points. LSRA determines the line that can be drawn though the points that minimizes the sum of the square deviates. Using some calculus, one can obtain equations for the slope (m), intercept (b), and the errors in the slope (s_m) and intercept (s_b), the standard error of regression (s_y), the correlation coefficient (R), and others. These equations can be found in Chapter 5 of the Harris. You will not need to manually solve any of these equations because we can use Excel to calculate them for us. There will be more on this later.

The correlation coefficient (R) is a value between 0 and 1 for a plot with a positive slope or 0 and -1 for a plot with a negative slope. For a positive (negative) slope the closer R is to 1 (-1) the stronger the correlation. The correlation coefficient of zero suggests that there is no correlation at all. Correlation coefficients are often provided along with a

slope and intercept as a way of characterizing the goodness of fit. However, this is often a misuse of R. R is most useful when you want to show to what extent the variance of one variable is coupled to the variance in the other variable. The coefficient of determination, R^2, is very useful in this regard. For instance in Experiment 2 we will prepare a plot of the masses of the pennies as a function of the year in which they were minted. Let's say we find that this plot has an R of 0.82 ($R^2 = 0.67$). This suggests the 67 % of the variation in the masses of the pennies is coupled to the differences in the years they were minted, and 33 % of the variance comes from some other factors, such as, for instance, maybe the number of times a given coin exchanged owners. However, using a correlation coefficient to characterize a Beer's law plot (see below) is a little pointless because the Beer's law relationship has already been established over and over again and we already know that essentially all of the variability in the absorbance values is coupled to the variations in concentration. A better indicator of the quality of a calibration plot (see below) is the standard error of regression, s_y. s_y is essentially a measure of the average deviation of the measured y values and the **y** values given by the LSRA best fit line.

$$s_y = \{\text{sum of square deviates}/(n-2)\}^{1/2}$$

Calibration Plots

Often we will use LSRA to define a calibration curve that is used for quantitative analysis. For instance, the typical spectroscopic experiment works like this. A series of standards is prepared of known concentrations. Absorbances for each of the standard solutions are measured using a spectrometer. These absorbances are plotted as a function of concentration. A plot of absorbance vs. concentration is known as a Beer's law plot. Beer's law states that for a monochromatic beam of light (for practical purposes, light of a single wavelength) passing through a sample of thickness b, the absorbance of the radiation is given by

$$A = \varepsilon bc$$

Where c is the concentration of the absorbing analyte and ε is the molar absorptivity. The molar absorptivity is dependent on the analyte, the solvent, and the wavelength of light. Thus, there is a linear relationship between A and c. If we plot A on the y-axis as a function of c, on the x-axis, we can perform a LSRA to find the slope and intercept. The intercept should be close to zero and the slope is εb. Generally we would also prepare and measure the absorbance of some sample which we are interested of learning the concentration of (the unknown), and we use our calibration (Beer's law) plot to find the concentration of the analyte in the unknown. We then use the slope and intercept from the LSRA and the measured absorbance of the unknown to calculate c_{unk}. Well, this is fairly straightforward and simple, but how do we express the statistical error in this c_{unk}. The equation for the uncertainty in c_{unk} is the equation for s_x given on page 87 of your Harris textbook.

One of the experiments we will be performing is a standard addition analysis, which is a modified standard curve analysis that accounts for any matrix effects in the unknown that may skew an analysis based on a plain vanilla-type standard curve analysis. We will discuss the standard addition plot in greater detail when we perform the experiment. For now, it is sufficient to state that c_{unk} is given by the negative of they x-intercept in a standard additions plot. The uncertainty in the x-intercept can be calculated using the equation at the bottom of page 89.

Error Extrapolation

Often the value we want to report is not the same as the value we obtain for an unknown from LSRA. For instance, we may dilute an unknown substantially prior to measuring it with a spectrometer. There are always some errors that arise in the preparation of the unknown. These errors are directly related to the tolerances in the pipets, volumetric cylinders, burets, graduated cylinders, ect. The idea is to propagate these errors based on the equation used to calculate the final value that is being reported. The rules for propagating errors are given in Harris on pg 56, Table 3.1 and below.

Addition and Subtraction

$$a + b + c = d \qquad a - b - c = d$$

$$s_d = (s_a^2 + s_b^2 + s_c^2)^{1/2}$$

Multiplication and Division

$$a \cdot b/c = d$$

$$s_d/d = [(s_a/a)^2 + (s_b/b)^2 + (s_c/c)^2)^{1/2}\ ]$$

Exponents

$$y = x^a$$

$$s_y/y = a \cdot s_x/x$$

logs

$$y = \ln x$$

$$s_y = s_x/x$$

$$y = e^x$$

$$s_y/y = s_x$$

An example: How many mg of quinine are in a 2-L bottle of tonic water?

A 10.00 mL aliquot of tonic water is added to a 500.00 mL volumetric cylinder and diluted to the mark. A of this diluted solution was analyzed by fluorescence. Using a calibration curve of Fluorescenve vs. concentration produced from standard solutions prepared from pure quinine sulfate, the concentration of quinine in the 50 μL aliquot was found to be 1.54±0.12 mM.

First let's calculate the mg quinine in the 2-L bottle of tonic water. The concentration of the diluted solution in the 500.00 mL volumetric is 15.4±1.2 μM. The sample was diluted by a factor of 50.00 (500.00/10.00). The molecular weight of quinine is 324.417 g/mol. So, (15.4 μmol/L)·(500.00/10.00)·(2 L)·(324.417 μg/μmol)·(1 mg/1000μg) = 0.4996 mg The only quantities that have uncertainties associated with them besides the concentration of the diluted sample are the 10.00 mL from the pipet and the 500.00 mL from the volumetric flask. The typical tolerance of a 10 mL pipet is ±0.02 mL (Table 2-4) and the typical tolerance of a 500.00 mL volumetric is ±0.2 mL (Table 2-10). For our burets it is ±0.05 mL (Table 2-2).

We can reduce the above equation to

$$X = \text{Constant}\cdot[(a\pm sa)\cdot(b\pm sb)/(c\pm sc)]$$ and the rule of multiplication and division is appropriate.

$$\text{mg quinine} = (2\cdot 324.417/1000)\cdot[(15.4\pm 1.2)\cdot(500.00\pm 0.2)/(50.00\pm 0.05)]$$

$$\begin{aligned}(s_X/X)^2 &= (s_a/a)^2 + (s_b/b)^2 + (s_c/c)^2 \\ &= (1.2/15.4)^2 + (0.2/500)^2 + (0.05/50.00)^2 \\ &= 0.006073 \\ s_X &= X\,(0.006073)^{1/2} = 0.03893 \text{ mg}\end{aligned}$$

And the mg quinine in a 2L bottle is reported as 0.500±0.039 mg. Some text will tell you that you should only keep one significant figure in the error. So, you would report it as 0.50±0.04 mg, others will tell you to follow different rules. If you always keep two sig. figs. in the error, you will stay out of trouble.

Sometimes you will be faced with a problem, such as

$$X = \text{Constant}\cdot[\{(a\pm sa) - (b\pm sb)\} / (c\pm sc)]$$

In this case you must break the propagation of errors into steps.

$d \pm s_d = \{(a \pm s_a) - (b \pm s_b)\}$, use the add/sub rule

$$s_d = (s_a^2 + s_b^2)^{1/2}$$

then,

$X = \text{Constant} \cdot [(d \pm s_d) / (c \pm s_c)]$, now apply the mult/div rule

$$s_x = X\,[(s_d/d)^2 + (s_c/c)^2]^{1/2}$$

Conclusion

Well, that is most of what you will need to get you through the semester. The rest will pick up as we go. Below list the mathematical equations for the LSRA calculation, and a tutorial on using excel to help you calculate these values.

LSRA Parameters

The "best fit" line $y = mx + b$

Minimize the square of the vertical deviation - d_i^2

$$d_i^2 = (y_i - y)^2 = (y_i - (mx_i + b))^2$$

$$s_y = [\Sigma\, d_i^2 / (n-2)]^{1/2}$$

n = the number of data points

$$D = n\Sigma(x_i^2) - (\Sigma x_i)^2$$

$$m \text{ (the slope)} = [n\Sigma(x_i y_i) - \Sigma x_i \Sigma y_i]/D$$

$$b \text{ (the intercept)} = [\Sigma(x_i^2)\Sigma y_i - \Sigma x_i \Sigma(x_i y_i)]/D = y_{ave} - m x_{ave}$$

Error analysis using LSRA

error in slope $= s_m = [n s_y^2 / D]^{1/2}$

error in intercept $= s_b = [s_y^2 \Sigma(x_i^2)/D]^{1/2}$

For typical calibration plot:

error in unknown x from interpolation s_x

$$s_x = \{(s_y/m)^2\,[1/k_{unk} + 1/D\,(\,x^2 n - 2x\Sigma x_i + \Sigma(x_i^2)\,)]\}^{1/2}$$

k = number of replicate measurements of the unknown

x = the calculate x value for the unknown

For standard addition plot:

error in x intercept

$s_{x\text{-}int} = \{ (s_y/m)^2 [1/D (n x_{int}^2 - 2x_{int}\Sigma x_i + \Sigma(x_i^2)] \}^{1/2}$

x_{int} = the value for the x-intercept

Tutorial on using excel for LSRA

Let's use some made up atomic absorption data. Atomic absorption was used to measure Ca^{2+} in a Tum's tablet. A standard curve is to be prepared by measuring the absorbance of several standard solutions of known [Ca^{2+}]. The data is typed into the spreadsheet as shown below.

[

	[Ca2+]	abs
	x	y
std A	2.607	0.042
std B	5.213	0.081
std C	10.426	0.165
std D	15.639	0.238
std E	26.065	0.399

Now, let's first let Excel perform a regression analysis for us. Go to "Tools" in the tool bar at the top. If there is a "Data Analysis" under the "Tools" tab, click it. If it is not there, you will have to add it. In this case click on "Add-ins" under the "Tools" tab, check both "Analysis Tool" packs, and click "OK". When you go back into the "Tool" tab, "Analysis Tools" should now be there. Once you click on it, scroll down to "regression", click on it, and hit "OK". Input the y and x data (for example, by clicking, holding, and dragging until all of the x data is highlighted), select the "output range" box and choose a cell to place the data that is removed from your data (such as A12). Also check the "residuals" box, and hit OK! On the data sheet scroll down to A12. The results give several pieces of data. An example of the output from the regression of the above data is shown on the next page. You can see that it gives a correlation coefficient (multiple R) and R^2. It also gives the standard error of regression, s_y, just below the adjusted R. Further down the intercept and the "X variable", which is really the slope is given along with there standard errors, s_b and s_m. The upper and lower range of 95 % CL of the intercept and slope is also given. These are based on the product of $t_{(n-2)}$ and the standard errors. However, if you were reporting the error in the slope and intercept based on the 95 % confidence limit it would be proper to divide these values by $n^{1/2}$, because we would want to use the estimate of the standard deviation of the mean.

	[Ca2+]	abs			
	x	y			
std A	2.607	0.042			
std B	5.213	0.081			
std C	10.426	0.165			
std D	15.639	0.238			
std E	26.065	0.399			
	SUMxi				
sum	59.95				

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.999867
R Square	0.999733
Adjusted R Square	0.999645
Standard Error	0.00267
Observations	5

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.080249	0.080249	11252.97	1.85E-06
Residual	3	2.14E-05	7.13E-06		
Total	4	0.08027			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.002883	0.002091	1.378433	0.261878	-0.00377	0.009538
X Variable 1	0.015189	0.000143	106.08	1.85E-06	0.014733	0.015645

OK. Let's say the absorbance of the diluted extract of a Tum's tablet was 0.253. Based on the slope and intercept, the [Ca^{2+}] of this diluted extract is 16.4669 ppm Ca^{2+}. How do we determine the uncertainty associated with this number? We need to use

$$s_x = \{(s_y/m)^2 [1/k_{unk} + 1/D (x^2 n - 2x\Sigma x_i + \Sigma(x_i^2))]\}^{1/2}$$

$k_{unk} = 1$, because we only made one measurement of the unknown

$x = 16.4669$

$s_y = 0.00267$

$m = 0.015189$

$\Sigma x_i = ?$

$\Sigma(x_{i2}) = ?$

$D = n\Sigma(x_{i2}) - (\Sigma x_i)_2 = ?$

We could perform this calculation, but we first need to get Σx_i , $\Sigma(x_i^2)$, and D. These are easily calculated using the spreadsheet. Two spaces below the column of x's I will calculate the sum of the x's (Σx_i) [type "=SUM" and click, hold, and drag the x's]. Now I label the column adjacent to the y's as"x^2", and calculate the square of each of the x values. And then two spaces below the column of x^2's I will calculate the sum of the x^2's [$\Sigma(x_i^2)$]. From here D is easily calculated.

	[Ca2+]	Abs			
	x	Y	x^2		
std A	2.607	0.042	6.796449		
std B	5.213	0.081	27.17537		
std C	10.426	0.165	108.7015		
std D	15.639	0.238	244.5783		
std E	26.065	0.399	679.3842		
	SUMxi		SUM(xi)^2		D
sum	59.95		1066.636		1739.177

Now, we can plug everything in and calculate s_x.

$s_x = 0.197$

Thus, the concentration of the diluted extract is
16.6±0.2 ppm (both would be acceptable in this case).
Try to use the spreadsheet with this data to hone and test your Excel abilities.