Dealing With Errors in Instrumental Analysis

I. Systematic Errors

Whenever a measurement is made two questions need to be asked.

First, am I measuring what I think I am measuring, or are other effects interfering with my determination?

Second, how "good" is the measurement?

The first question deals with the concept of **accuracy**. For example, suppose that the percentage of tin in an alloy is measured. The amount determined experimentally is the best guess at the true amount. The actual difference between the true value and the experimentally determined value is called the **bias** in the measurement, where

$$\text{experimental value - true value = bias} \qquad \text{(A.1)}$$

Bias is therefore a measure of the accuracy of a determination. Unfortunately, most real samples have no known, true value, and the bias cannot be determined.

Factors which can affect accuracy are called **systematic errors**. Some examples are:

a. end point errors in titrations
b. non-linear curve in a calibration curve
c. background interference
d. less than 100% current efficiency in coulometry
e. diffusion current in potentiometry
f. contamination

Systematic errors always need to be accounted for in the analysis. Unfortunately, these errors cannot be dealt with through statistics. Instead, they must be detected and corrected by other means. Most often, a systematic error in an analysis is detected through the analysis of a known standard or of a reference material for which the concentration is known. In these cases, a bias *can* be measured. A systematic error present in the analysis will lead to analytical results that are always ("systematically") high or low. Systematic errors are also detected by repetition of an analysis under slightly different conditions. When systematic error is important, such analyses often show trends.

II. Random Errors

The second question which was asked has to do with the concept of precision. If a measurement was repeated a number of times in the exact same manner, the numerical value of the results would change slightly - but randomly - each time. This slight change is due to the failure to control all experimental variables exactly the same way in each run. The variation in results contains a random error component. The average of this error is near zero if the error is indeed random. The magnitude of the error or precision is a measure of the variation in the results. Thus, **precision** is a measure of how reproducible the results are. Statisticians have developed well-defined ways to describe random errors, some of which are discussed below.

Repeated Measurements

The standard method for examining the precision of a measurement is to perform the analysis on the same sample several times, so that replicate data are available for statistical analysis. For replicate data, the **standard deviation** s is

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}} \qquad \text{(A.2)}$$

where $\overline{x} = \sum_{i=1}^{n} x_i / n$

is the mean, and n is the number of replicate analyses x_i.

Upon closer examination of the standard deviation one notices a limitation: it does not take into account the number of replicates analyzed in considering the precision of a series of analyses.

For example, chemist A has a standard deviation of 5 after doing 20 analyses and chemist B also has a standard deviation of 5, but he performed only 3 repeated analyses. It would seem logical that one would have more confidence in chemist A's results than in chemist B, but the equal standard deviation values imply that the precision of both sets of results is the same.

Confidence Intervals

To clarify this apparent dilemma, a confidence interval (C.I.) can be defined for each series of results. Mathematically, this is done as follows:

$$\text{C.I} = \frac{(t_\alpha * s)}{\sqrt{n}} \qquad \text{(A.3)}$$

where the quantity t_α is determined from a t-table, s is the standard deviation given by equation (A.2), n is the number of replicates performed, and the "degrees of freedom" defining t_α are (n - 1) because one is used in calculating the mean value of x. A degree of freedom can be thought of as an independent value - with the mean and n-1 values, we can calculate the n^{th} value, so it is not an independent value. The confidence interval explicitly includes the number of analyses in its definition so this helps clarify the effect of the number of replicates on the precision. The confidence interval has another helpful property: a 95% confidence level means that the true value falls within the limits of the confidence interval generated from the t value and the equation A.3.

To use the t-table, pick as the **column** the confidence level a desired and as the **row** the degrees of freedom as stated. Where the column intersects with the row is the t-value which is used.

Now we return to the example. Using the t-table, the confidence interval calculated in equation A.3 is reported with an associated confidence level. In the example, chemist A reports a 95% confidence interval of +/- 2.3 and, chemist B a 95% confidence interval of +/- 12.4. Obviously, chemist A's results are much more precise than those of chemist B.

III. Linear Regression

With the advent of modern instrumentation, the chemist is often faced with the task of calibrating instruments by using various concentrations of standards. If there is a linear correlation between the output signal and the concentration of the standard, the chemist can use a standard regression technique to determine the parameters that best fit the data.

Again, the chemist is faced with the problem of noise in the data, and it is quite obvious that one would need a more powerful technique to quantify these than what was used with repetitive analysis. The statistics will be a little more difficult to calculate, but the basic concept is the same as previously developed. First, we need to define the regression method, and then develop the statistics. The equation for a straight line is

$$y = bx + a \tag{A.4}$$

where b is the slope, a is the y-intercept, x is the independent variable (usually concentration), and where y is the dependent variable (usually instrumental output). If all the error is taken to be in y, the best straight line fit to the data can be obtained by using a standard linear regression, from which

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{A.5}$$

$$a = \bar{y} - b\bar{x} \tag{A.6}$$

The slope b and the y-intercept a will contain some noise because the data used to obtain them also contain noise.

To show how linear regression works more explicitly, let's consider the following example. Chemists A and B have obtained the following data, and each has developed a linear calibration curve by least squares.

A) data

x	y
1	1.1
2	1.9
3	3.1
4	3.9
5	5.1

for which a=0.020 and b=1.00.

B) data

x	y
1	1.3
2	1.7
3	3.3
4	3.7
5	5.3

for which a=0.060 and b=1.00.

Which set of data gives the greater confidence? By examining the data, it can immediately be pointed out that chemist A's data are more precise. To quantify this observation, however, something more detailed is needed. The following equations can be used to determine the standard deviation of the slope and y-intercept in a linear regression

$$s_b = \frac{s_{y/x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad \text{(A.7)}$$

$$s_a = s_{y/x}\sqrt{\frac{\sum_{i=1}^{n} x^2_i}{\sum_{i=1}^{n} n(x_i - \bar{x})^2}} \qquad \text{(A.8)}$$

$$s_{y/x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \qquad \text{(A.9)}$$

where yi is the **fitted** y-value calculated from the least squares estimates of a and b using our assumed line $y_i = bx_i + a$, and n is the number of points in the calibration curve. As was found in the previous discussion, the standard deviation does not take into consideration the number of replicated points in a series of repeated measurements. Once again a confidence interval may be calculated by using the t-table to establish a confidence level. The equation for the confidence interval is modified slightly as follows:

$$\text{C.I.} = t_a * s_a \qquad \text{(A.10)}$$

$$\text{C.I.} = t_a * s_b \qquad \text{(A.11)}$$

Now, the degrees of freedom used in determining t_a are (n - 2) for both A.10 and A.11 because 2 means were calculated, one for x and one for y.

Determining Unknown Concentrations by Regression

Quite often it is necessary to use a calibration curve to determine an unknown concentration of a sample. By measuring the instrument output and using our formula as determined by the regression analysis, the concentration of the unknown x is easily found to be

$$\hat{x} = (y\text{-}a)/b \qquad \text{(A.12)}$$

To find the error associated with the calculated concentration the following equation is used to determine the corresponding standard deviation

$$s_{x_0} = \frac{s_{y/x}}{b}\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(y_o - \bar{y})^2}{b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{A.13}$$

where m is the number of replicated measurements of the response of the unknown sample, y_i yielding mean y_o. Notice that in this equation there are two quantities which affect the standard deviation first, the number of points n in the calibration curve; and second, the m repetitive trials taken. The number of points taken in the calibration curve affects the standard deviation in a complicated manner and is difficult to analyze; however, for most analyses 5 to 7 standards should be used in defining the calibration curve, to minimize time and to maximize precision. If improved precision is necessary, 3 or 4 repeated measurements of the unknown sample can be obtained. To determine the confidence level, the following calculation is performed:

$$\text{C.L.} = t_\alpha * s_{x_o} \tag{A.14}$$

where the degrees of freedom used in determining t_α are (n-3) when x_o is calculated from the mean y_o.

Limit of Detection and Limit of Quantitation

Another set of quantities that one frequently needs to establish when dealing with calibration curves is the limit of detection and limit of quantitation. For most work done in analytical chemistry, the following definitions are used:

$$\text{Limit of detection} = a + 3s_{y/x} \tag{A.15}$$

$$\text{Limit of quantitation} = a + 10s_{y/x} \tag{A.16}$$

where a is the y-intercept and $s_{y/x}$ is calculated using equation (A.9). It should also be noted that there is no standard deviation or confidence level associated with the limits of detection or quantitation.

Standard Additions

The method of standard addition is a technique used to eliminate matrix effects in complex samples. This technique most often utilizes a linear regression to calculate the parameters from which the concentration of the sample can be determined. Most instrumental analysis texts cover the technique in detail. To analyze the error associated with this method, the following equation can be used to determine the standard deviation of the unknown concentration:

$$s_{x_E} = \frac{s_{y/x}}{b}\sqrt{\frac{1}{n} + \frac{\overline{y}^2}{b^2 \sum_{i=1}^{n}(x_i - \overline{x})^2}} \tag{A.17}$$

where $s_{y/x}$ is calculated using equation (A.9), b is the slope of the standard additions line and where n is the number of points defining the standard additions line. (Note that equation A.17 is a modified version of equation A.13 where m =∞ and y_o=0.) To determine the confidence level for the unknown concentration, the following equation is used

$$\text{C.L.} = t_\alpha * s_{x_E} \tag{A.18}$$

where the degrees of freedom used in determining t_α are (n - 2). It should be noted that the more points taken to determine the parameters, the better the precision.

IV. Propagation of Random Errors In Computation

There is often a need to estimate the error in a result that has been computed from two or more data, each of which has an error associated with it. The way in which the individual errors accumulate depends upon the arithmetic relationship between the terms containing the errors. Consider the following summation problem:

0.50(+/-0.02) + 4.10(+/-0.03) - 1.97(+/-0.05) = 2.63(+/- ?).

The numbers in parentheses are the absolute indeterminate errors expressed as standard deviations. The uncertainty associated with the sum could be as large as +/-0.10 if the signs of the three individual standard deviations happened to be all positive or all negative. On the other hand, under fortuitous circumstances, the three uncertainties could combine to give an accumulated error of zero. Neither of these is as probable as a combination leading to an uncertainty intermediate between the extremes.

The **most probable uncertainty** in the case of sums or differences can be found by taking the square root of the sum of the squares of the individual standard deviations. This is easily demonstrated by propagation of error. In the present example:

$$s_y^2 = s_a^2 + s_b^2 + s_c^2 \tag{A.19}$$

where s_y is the standard deviation of the sum and s_a, s_b and s_c are the standard deviations of the individual terms.

If y = f(a,b,c) then the general form for propagating random errors for any function where the result y is dependent upon the experimental variables a, b, and c is

$$s_y^2 = \left(\frac{\partial y}{\partial a}\right)^2 s_a^2 + \left(\frac{\partial y}{\partial b}\right)^2 s_b^2 + \left(\frac{\partial y}{\partial c}\right)^2 s_c^{\ 2} \qquad \text{(A.20)}$$

Thus for the example above where

$$y = a + b + c \qquad \text{(A.21)}$$

and

$$\left(\frac{\partial y}{\partial a}\right) = \left(\frac{\partial y}{\partial b}\right) = \left(\frac{\partial y}{\partial c}\right) = 1 \qquad \text{(A.22)}$$

then

$$s_y^{\ 2} = (1)^2 s_a^{\ 2} + (1)^2 s_b^{\ 2} + (1)^2 s_c^{\ 2} \qquad \text{(A.23)}$$

or

$$s_y = \sqrt{s_a^2 + s_b^2 + s_c^2} \qquad \text{(A.24)}$$

and

$$s_y = \sqrt{(0.02)^2 + (0.03)^2 + (0.05)^2} = +/- 0.06$$

This is the familiar rule for propagating error in a sum or difference.

For an equation which contains only products and quotients, we can use a similar approach to propagate error. For example, if

$$y = (a * b) / c \qquad \text{(A.25)}$$

then, because equation (A.20) applies here as well

and $\left(\frac{\partial y}{\partial a}\right) = \frac{b}{c} \quad \left(\frac{\partial y}{\partial b}\right) = \frac{a}{b} \quad \left(\frac{\partial y}{\partial c}\right) = \frac{-ab}{c}$

it follows that

$$s_y^2 = \left(\frac{b}{c}\right)^2 s_a^2 + \left(\frac{a}{b}\right)^2 s_b^2 + \left(\frac{-ab}{c}\right)^2 s_c^2$$

Then divide each side by (a * b)^2 and then multiply each side by c^2 to get

$$\left(\frac{c}{ab}\right) s_y^2 = \left(\frac{1}{a}\right)^2 s_a^2 + \left(\frac{1}{b}\right)^2 s_b^2 + \left(\frac{-1}{c}\right)^2 s_c^2$$

Now we define $(s_y)_r = \frac{c}{ab}\, s_y = \frac{s_y}{y}$

and $(s_a)_r = \frac{s_a}{a}$ $(s_b)_r = \frac{s_b}{b}$ $(s_c)_r = \frac{s_c}{c}$

and substitute to get

$$(s_y)_r = \sqrt{(s_a)_r^2 + (s_b)_r^2 + (s_c)_r^2} \qquad \text{(A.26)}$$

This is the familiar rule for propagating error in a product or quotient.

References

1. J. C. Miller and J. N. Miller, *Statistics For Analytical Chemistry*, Ellis Horwood Limited, 1984.

2. D. A. Skoog and J.J. Leary, *Principles of Instrumental Analysis*, Fourth Edition, Saunders, 1991. Appendix 1.