Separation Science - Statistics Unit
Thomas Wenzel
Department of Chemistry
Bates College, Lewiston ME  04240
twenzel@bates.edu
INTRODUCTION

The field of statistics is extensive, as evidenced by several departments or programs at Bates offering semester-long courses on statistics. This several page description of aspects of statistics is therefore only a small component of the field, but the concepts described are those most common and fundamental to an examination of data that is gathered when performing a chemical analysis. The purpose of chemical analysis is to identify and measure the true amount of a substance in a mixture. Unfortunately, any measurement we make is subject to a variety of errors that can result in the measured value being different from the true value. The two general types of errors are known as systematic and random errors. Two other terms that are important to define are accuracy and precision.

Accuracy refers to how close the measured value comes to the true value.

Precision refers to how close replicate measurements of the same sample come to each other.

The three pictures below representing arrows shot at a target serve to illustrate the difference between accuracy and precision.

Figure A: This person is accurate and precise. The arrows are in the center spot as desired and are all close to each other. The shots are not exactly in the center, though, so do differ from the true value. The small differences between the actual values and the true values are the result of random errors.

Figure B: This person is inaccurate, but precise. The arrows do not land near the center of the target, but the arrows are all clustered in one area of the target. This person has what we would describe as a systematic error. There is something wrong in their aim or their release of the arrow that is causing the arrows to miss the mark. Whatever the error, the person consistently makes it from shot to shot. Presumably, an archery instructor could examine this person’s technique and correct the systematic error resulting in both accuracy and precision. Note as well that the shots do not all land exactly in the same spot, so there is a small amount of random error as well.

Figure C: This person is neither accurate nor precise. The shots do not land in the center spot and, rather than landing in one clustered area, land all over the target. However, suppose as we often do with chemical measurements, we were to put a coordinate system on the target with the origin at the center of the target and then compute an average of the four shots shown. Note that this person would have an average value that exactly equaled the true value, but also note that it would be due mostly to luck. If they had been stopped at three shots or had taken a fifth, the average would not have equaled the true value.
These targets can also be used to introduce a statistical concept known as confidence. Confidence is an expression of how close a set of measured values comes to the true value. If you were told to put an apple on your head and select one of the shooters above to put an arrow through it, I expect that you would feel quite confident that shooter A would be up to the task. You might even feel confident about shooter B, once you had corrected for the systematic error. I expect that you would have absolutely no confidence that shooter C would be capable of hitting the apple.

**Systematic (determinate) errors** are caused by a defect in the analytical method or the technique by the person making the measurement. For example, consistently reading an instrument wrong or preparing a solution wrong will impart a systematic error to the result. Using a pipet improperly to transfer a solution, or using an automatic pipetter that is miscalibrated are other examples of systematic errors. It is possible to identify the source of systematic errors and correct them, although it might not be easy to do so. Often we can examine our methods for the presence of systematic errors by utilizing what are called “certified” samples that have a known amount of the substance that we are analyzing.

**Random (indeterminate) errors** are errors that occur through the normal course of performing an analysis and are those over which the analyst has no control. This includes the normal uncertainty in reading measurements such as weights on a balance and volumes using calibrated glassware. Instrumental noise is another source of random errors. Random errors are as likely to be positive as they are to be negative so performing replicate measurements leads to some cancelation of random errors.

It is also important to talk about the **absolute deviation** and the **relative deviation**. The absolute deviation of a balance that measures to four decimal points is 0.0001 g. The absolute deviation of a buret is usually 0.02 ml. The relative deviation is the absolute deviation divided by the measurement. It is often useful to convert the relative deviation to a percent. For example, the relative deviation of an object that weighed 0.0010 gram would be as follows:

\[
\frac{0.0001 \text{ g}}{0.0010 \text{ g}} \times 100 = 10 \%
\]

By contrast, the relative deviation of an object that weighed 0.1000 gram would be as follows:

\[
\frac{0.0001 \text{ g}}{0.1000 \text{ g}} \times 100 = 0.1 \%
\]

Comparing the relative deviation for these two examples should serve to illustrate why it is desirable to weigh larger samples or measure larger volumes. The relative uncertainty in a small weight such as 0.0010 g is so large that the error in any measurement performed using the solid sample will be similarly large.
SIGNIFICANT FIGURES

It is essential for anyone performing analytical chemistry to have an appreciation for significant figures. Significant figures, by definition, are those digits in a number that are known with certainty, plus the first uncertain digit. The last digit is generally considered uncertain by +/-1 in the absence of qualifying information.

Treatment of Zeros

-Zeroes that appear between other digits, such as those in 20.09 are significant. The number 20.09 has four significant figures.

-Zeroes used only to locate the decimal point, such as 0.002009 are not significant. The number 0.002009 has four significant figures.

-Zeroes that terminate a number, such as those in 0.200900, are usually significant. The number 0.200900 has six significant figures. The number 200,900 is an interesting one to consider. The number 200,900 would be considered to have six significant figures. If instead, the number was only to have four significant figures, it would have to be written as 2.009 x 10^5 to show that there are only four significant figures. So note that an exponential number has no effect on the number of significant figures.

When performing calculations, one must be careful to report the proper number of significant figures in the final number. Such answers must be rounded off, and the rounding can be done either before or after the calculation (my personal preference is to do the rounding after).

Addition and Subtraction

In addition and subtraction, the absolute uncertainties of the numbers must be compared. The value with the largest uncertainty will determine how many digits are retained in the other values. For example, consider adding the following weights:

\[
\begin{align*}
10.01 \quad & \text{g} \\
9.0012 \quad & \text{g} \\
14.032 \quad & \text{g}
\end{align*}
\]

If the uncertainty in the first weight is +/-0.01 gram, then the 0.0012 g from the second weight and 0.002 g from the third are not worth considering since they are already less than the uncertainty in the first weight. So the correct numbers to add with rounding would be:

\[
\begin{align*}
10.01 \quad & \text{g} \\
9.00 \quad & \text{g} \\
14.03 \quad & \text{g}
\end{align*}
\]
**Multiplication and Division**

In multiplication and division, the relative uncertainties in the measurements are what need to be assessed. This sounds complicated, but there is a straightforward way to determine the number of significant figures that belong in the answer: it’s the same number of significant figures as those in the value with the fewest. So if we were to multiply the three weights listed above (10.01, 9.0012, and 14.032), these numbers have 4, 5, and 5 significant figures respectively, the answer would be reported to four significant figures.

Finally, numbers that are known with absolute certainty, such as the number of measurements made, do not need to be considered when assessing significant figures.

**CENTRAL TENDENCY OF A SET OF RESULTS**

Once we have a set of data, we need some way of expressing one value that in some way represents the entire set of data. The central tendency is how this is done. The most common way of expressing the central tendency is to use what is known as the mean, \( \bar{x} \). The mean is the average of the set of measurements. An alternative way of expressing the central tendency is to use the median. If the values are arranged in rank order from lowest to highest, the median value is the one for which exactly half the values are lower and half the values are higher. If there is an even number of measurements, the median is found by averaging the two middle results. In most instances the mean is the best measure of central tendency. The median is better when there are a few measurements that are so different than most of the others that they skew the mean. As an example, suppose we were to select 30 people and determine a central tendency for their yearly income. Now suppose that one of the 30 people happened to be Bill Gates. Because Bill Gates income is so large, it would probably cause the mean income of the 30 to be higher than the income of any of the other 29 people. The mean would therefore not be representative of the population, and the median would serve as a better measure of the central tendency.

**PRECISION**

Precision is a measure of how well values cluster about the mean. It can be used to compare various analytical methods. Assuming two methods do not exhibit any systematic errors, the method with the better precision would be preferred. Precision can also be used as a check on the analyst. Two different people performing exactly the same analytical method on the same sample will often get different levels of precision. The one with the better precision is a better analyst. Precision is also used to decide whether a data point can be rejected, and, as we will see, to establish what is called a confidence limit.

There are several ways in which precision can be measured. One measure of precision is the average deviation, \( \bar{d} \). This is the average of the deviation of individual results from the mean without regard to the sign of the deviation.

\[
\bar{d} = \frac{1}{n} \sum |x_i - \bar{x}|
\]
The most common measure of precision is the standard deviation, s. The standard deviation is defined as the square root of the average of the square of the individual deviations from the mean.

\[
\sigma = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}
\]

The average and standard deviations are expressed in the same units as that used to express the mean. The relative average or standard deviations can be expressed as follows:

\[
\frac{d}{x} \cdot 100 = \% \quad \frac{s}{x} \cdot 100 = \%
\]

**Normal Distribution Curve** (otherwise known as the Gaussian curve or bell curve):

Since we are concerned with statistical treatment of a set of measurements obtained by a chemical analysis, if one were to perform the analysis an infinite number of times (this is known as the population), and then plot the frequency of occurrence of each particular measured value versus that value, a normal, or Gaussian, distribution curve would be obtained. The distinguishing feature of the normal distribution curve is the perfectly symmetrical nature of the curve about the mean. This trend is so commonly known that it is often assumed that the normal distribution curve explains all measurements of systems. This is hardly the case and many natural phenomena exhibit distribution curves that take other forms than the Gaussian curve. A plot of a Gaussian curve is shown below.
Several things are worth noting:

1. The term \( u \) is used to denote the mean of the population (an infinite number of measurements), whereas \( \bar{x} \) is used to denote the mean of a sample.

2. The term \( \sigma \) is used to denote the standard deviation of the population, whereas \( s \) is used to denote the standard deviation of a sample.

3. For every positive error, there is a corresponding negative error of equal magnitude (in other words, the curve is perfectly symmetrical).

4. The frequency of those values with a small error (plus or minus one standard deviation) is great (68%).

5. The frequency of those values with a large error (plus or minus three standard deviations) is small (0.26%).

The standard practice is to perform an analysis in triplicate. If one wanted to obtain the true value of a measurement, it is worth realizing that in the absence of systematic errors, the mean of the population (an infinite number of measurements) would equal the true value of the substance being measured. Clearly, it is impossible to carry out an infinite number of measurements, and some smaller number must be measured. But is three enough? If one considers the normal distribution curve, and realizes that only five out of 100 measurements are greater than plus or minus two deviations, the chance that the first three would have large errors is remote. Especially when one realizes that an equal number of these five of 100 measurements with large errors ought to be positive and negative. The chance of having three measurements in a row with either large negative or positive errors is extremely remote. So three measurements is a useful compromise that allows the analyst to perform the analysis in a reasonable amount of time and effort.

ASSESSING ACCURACY

Keeping in mind the assumption that an analytical method has no systematic errors, and keeping in mind that the mean of an infinite number of measurements will result in the true value, it would be highly desirable to be able to assess how close a mean arrived at by three measurements comes to the true value. **Confidence limits** are a way of performing such an assessment. A confidence limit is actually a range of values in which, if you were to then go on and perform a measurement an infinite number of times, the true value would be likely to fall. When assessing confidence limits, there must always be a level or percent (usually 90, 95, or 99%) of confidence. The higher the degree of confidence, the larger the range. It never is possible to say with 100% certainty that, if the measurement were then performed an infinite number of times, it would fall within the specified range unless the range is zero to infinity. Of course a range of zero to infinity makes no sense and defeats the purpose of performing the measurement in the first place.
Usually, in chemical analysis, confidence limits are assessed at the 95% level. In other words, after arriving at a mean and confidence limit from three measurements, if one were to then go on and perform the measurement an infinite number of times, in 19 out of 20 times the true value obtained by an infinite number of measurements would fall within the range you calculated from the three you actually measured. A gambler would jump at the chance of being able to place wagers if there was a certainty that 95% of the time he or she would be correct. If a person were on trial for a crime (say they were accused of killing their spouse by spiking their sugar bowl with arsenic), he or she would probably argue that this situation was the 1 out of 20 where the true value would fall outside the confidence limits, and that it would fall outside the range on the low side. In such a case, it might be necessary either to use a higher degree of confidence, or to show that the range within which the true value is expected is well above the normal level of arsenic in sugar.

When assessing confidence limits, the precision of the analysis is crucial. As an example, consider the following two sets of measurements for the % of substance in a sample. Note that each set of measurements has the same mean.

\[
\begin{array}{cc}
12.46 & 10.26 \\
12.38 & 11.69 \\
12.48 & 15.37 \\
\end{array}
\]

\[
\bar{x} = 12.44 \quad \bar{x} = 12.44
\]

\[
s = 0.05 \quad s = 2.64
\]

If these were a set of analytical measurements either performed by two different people using the same method, or the same person using two different methods, the first set of measurements are a sample of a very narrow Gaussian distribution, whereas the second set are a sample from a very broad Gaussian distribution. Another way we could say this is that the random errors in the first set of measurements are small, whereas the random errors in the second set of measurements are large. The confidence limits (the range in which the true value would fall if one were to conduct an infinite number of measurements) determined for the second set of measurements must be considerably larger than with the first set if it is to express the likelihood of encompassing the true value. If you just thought intuitively about guessing in each case a range in which you might reasonably expect a fourth measurement to fall, I think you would agree that the range for the first set of measurements would be much smaller than that for the second.

Confidence limits are calculated as follows, in which \( n \) is the number of measurements, \( s \) is the standard deviation, and the value \( t \) is looked up in a table.

\[
\text{Confidence limits} = \pm \frac{(t)(s)}{(n)^{1/2}}
\]
Values of $t$ for Calculating Confidence Limits

<table>
<thead>
<tr>
<th>Number of measurements</th>
<th>Degrees of Freedom</th>
<th>Probability Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n-1</td>
<td>90%</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>6.314</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2.920</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2.353</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2.132</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>2.015</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>1.943</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>1.895</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>1.860</td>
</tr>
</tbody>
</table>

The confidence limits at the 95% level for the two sets of data listed above are as follows:

$\bar{x} = 12.44$

$s = 0.05$

C.L. = +/-0.12

Range: 12.32-12.56

$\bar{x} = 12.44$

$s = 2.64$

C.L. = +/-6.55

Range: 5.89-18.99

Notice the large range for the second set of data. This is the range in which the true value would occur (95% of the time) if an infinite set of measurements were performed. The large range for the second set of measurements is caused by the large standard deviation. In this case, the range is so large that the measurements do not help much in determining the actual percentage of material in the sample.

TESTS OF SIGNIFICANCE

It is quite common to encounter situations in which sets of measurements must be compared. This can involve two people who analyzed the same sample using the same method, or one person using two different methods to analyze the same sample. Measurements of the same sample (perhaps a series of standards) on different days using the same instrument or method is one common example in which sets of measurements are compared. Since only three measurements are performed in each case, it’s quite certain that the two standard deviations and two means would not be identical to each other. If the means and standard deviations are different, it follows that the confidence limits are different as well. Even though the values are different, a useful question to ask is whether they are significantly different. We would probably agree that having $1.00 is significantly different than having $10.00, whereas we would also probably agree that having $1,000,001 is not significantly different than having $1,000,010. Of course, some miserly people might quibble over the difference between $1,000,001 and $1,000,010, showing that what constitutes a significant difference is somewhat arbitrary. In statistics, when we assess whether things are significantly different, we must then express this to some degree of confidence (like 90%, 95%, etc.). The three common ways in which we will test
significance is determining whether two standard deviations are different, whether two means are different, and whether a data point can be discarded in the calculation of a mean.

**Standard deviations (F test)**

An F value is calculated using the following expression, in which $s_1$ and $s_2$ are the two different standard deviations, so that $F > 1$. The value is then compared to an F value selected from a table. Note that the larger the difference between the two standard deviations, the larger the calculated F value. If the calculated F value is larger than that from the Table, then the two standard deviations are significantly different.

\[ F_{\text{CALC}} = \frac{s_2^2}{s_1^2} \]

**Table of F Values (95% Confidence Level)**

<table>
<thead>
<tr>
<th>$n_1$ = 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_2$</td>
<td>161</td>
<td>200</td>
<td>216</td>
<td>225</td>
<td>230</td>
</tr>
<tr>
<td>2</td>
<td>18.5</td>
<td>19.0</td>
<td>19.2</td>
<td>19.2</td>
<td>19.3</td>
</tr>
<tr>
<td>3</td>
<td>10.1</td>
<td>9.55</td>
<td>9.28</td>
<td>9.12</td>
<td>9.01</td>
</tr>
<tr>
<td>4</td>
<td>7.71</td>
<td>6.94</td>
<td>6.59</td>
<td>6.39</td>
<td>6.26</td>
</tr>
<tr>
<td>5</td>
<td>6.61</td>
<td>5.79</td>
<td>5.41</td>
<td>5.19</td>
<td>5.05</td>
</tr>
<tr>
<td>6</td>
<td>5.99</td>
<td>5.14</td>
<td>4.76</td>
<td>4.53</td>
<td>4.39</td>
</tr>
</tbody>
</table>

**Means (Null hypothesis)**

A t value is calculated using the following expression, in which $s$ is the standard deviation, $\bar{x}_1$ and $\bar{x}_2$ are the two means being compared, and $n_1$ and $n_2$ are the number of measurements used in determining each of the respective means.

\[ t_{\text{CALC}} = \frac{\bar{x}_1 - \bar{x}_2}{s} \times \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} \]

Note that the expression only has one term for the standard deviation, rather than two. In order to use the null hypothesis, the two means being compared must have been obtained from measurements with similar levels of precision (in other words, it was already shown that the standard deviations are not significantly different). If the standard deviations are not significantly different, then, statistically, it does not matter which one is used in calculating the t value. The calculated t value is then compared to a t value obtained from a table. When obtaining the value from the table, the degrees of freedom are determined as follows:

\[ \text{Degrees of freedom} = n_1 + n_2 - 2 \]
The calculated t value is then compared to the value from the table. Note that the larger the difference between the means, the larger the calculated t value. If the calculated t value is greater than that from the table, the means are significantly different.

**Rejection of a Data Point**

One of the most common assessments is determining whether a data point can be ignored in calculating a mean. It is very tempting to use one’s judgement in such a decision, and all too often people ignore or discard data based on their own seat-of-the-pants assessment that it’s “different”. In fact, discarding data on the basis of such judgement calls has gotten many scientists into trouble when it turns out the data that was discarded could not be done so on statistical grounds.

If the analyst knows that a mistake was made in performing the analysis (some of the solution was spilled, too much of something was added, a volumetric flask was overfilled, the color of a solution that is supposedly more concentrated is less intense than one that is supposedly less concentrated, etc.), the mistake is recorded in the notebook and the result is discarded (of course, usually such mistakes are realized beforehand and the mistake is corrected before undertaking the analysis). When the analyst has no evidence that a mistake was made, yet one measurement appears different than the others, the only way to discard the data is on statistical grounds. The procedure we will use to assess whether a data point can be rejected is the Q test. As with other such tests of significance, we must perform the test to some level of confidence. For our situation, we will assess at the 90% confidence level.

The procedure involves calculating a Q value using the following expression in which the questionable value is subtracted from its nearest neighbor. The range is the difference between the highest and lowest value in the set of measurements. Note that as the value becomes more suspect, meaning that there is a greater and greater difference between the suspect value and its closest neighbor, the calculated Q value approaches 1.

\[
Q_{\text{CALC}} = \frac{x_2 - x_1}{\text{Range}}
\]

It is important to note that the Q test can only be used to reject one data point. If a set of analytical measurements are obtained that appear to be bimodal (measured values of 12.47, 12.49, 12.54 and 13.21, 13.22, 13.25), something else is happening with the analysis that deserves further attention.
Table of Q Values

<table>
<thead>
<tr>
<th>N</th>
<th>Q(90%)</th>
<th>Q(96%)</th>
<th>Q(99%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.94</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>4</td>
<td>0.76</td>
<td>0.85</td>
<td>0.93</td>
</tr>
<tr>
<td>5</td>
<td>0.64</td>
<td>0.73</td>
<td>0.82</td>
</tr>
<tr>
<td>6</td>
<td>0.56</td>
<td>0.64</td>
<td>0.74</td>
</tr>
<tr>
<td>7</td>
<td>0.51</td>
<td>0.59</td>
<td>0.68</td>
</tr>
<tr>
<td>8</td>
<td>0.47</td>
<td>0.54</td>
<td>0.63</td>
</tr>
<tr>
<td>9</td>
<td>0.44</td>
<td>0.51</td>
<td>0.60</td>
</tr>
</tbody>
</table>

PROBLEMS

1. Calculate the sum and product of the following set of numbers:

   32.367    12.2681    19.17    198.245

2. The following set of measurements is obtained for the percentage of manganese in a sample of steel.

   1.01, 0.95, 0.99, 1.05, 1.06, 0.94, 0.85, 1.05, 1.05

   Can any of the measurements be rejected at the 96% level?

   Calculate the mean, average deviation, standard deviation and confidence limits (95%) for the set of measurements.

3. The following two sets of measurements are obtained for the amount of chloroform in a sample of drinking water (chloroform is produced during the water chlorination process). The two sets of measurements are obtained by two different analytical methods. The amounts are measured in parts per billion (ppb).

   57.3, 55.4, 56.2, 55.9

   57.8, 58.2, 56.5, 57.6

   Can any measurements be rejected at the 96% level?

   Are the two means significantly different (95% level)?

   Suppose the EPA requires that municipalities with chloroform levels above 50 ppb at the 95% confidence level treat the water to lower the chloroform. Would the water need to be further treated based on the measurements?