T-test in Chemistry

By Daniel Lee & Michael Hooper

 The statistical analysis methods used most often in chemistry is the student’s t-test based on the normal t-distrbution. Data collected in most chemistry labs follows a normal distribution making this analysis methods very useful. The normal distributions are bell-shaped continuous probability functions centered on the mean or average. Although in application the values are usually just looked up in a table, the calculated probability actually comes from the properties of the integration of the function.

The “Z-test” is the simplest and easiest to use cumulative distribution function (CDF). However the Z-test requires knowledge about the population average: µ and the population standard deviation: σ which is impossible to know in real life applications.

To account for this lack of knowledge the T-test was developed. It has a similar bell curve shape but it’s distribution function takes into account uncertainty in estimating the population statistics. To the left is an example of a 95% confidence Interval graphed on to a t-distribution.

Applications of the T Distribution:

The T-test is very useful as an analytical method, various forms of it can be used to determine the significance of data, compare datasets, and compare averages with known values. Determining your hypothesis and using the T-test to determine whether or not it can be rejected with the given sets of data is vital to chemistry. Various forms of the t-test allows us to say whether a given set of data is significant given measurements of it’s expected value and uncertainties in the form of the sample mean and standard deviation. The simplest form of the T-test is the confidence interval:

$$\overbar{x}\pm \frac{T\_{(∝/2,dof)}s\_{x}}{\sqrt{n}}$$

 This gives a range of values which contains the population average. The T value used in computation depends on both the desired accuracy level and the degrees of freedom. T is defined as the value on the x axis of the t-test for which $∝/2$ area lies beyond. This means when finding a confidence interval you are finding the range of values on a t-distribution which contains 95% of the probability distribution. But how does this relate to the distribution of your data? T-distributions (and normal distributions in general) all have the same shape changed only by the mean and standard deviation. Using this property you can extrapolate information about the population by normalizing with the average and standard deviation of your data. This requires no assumptions or knowledge about values of the population standard deviation or mean. The only assumption required is that the population is normally distributed, which nearly everything measured in analytical chemistry will be. An alpha value of 0.05 means that 95% of the area of the graph is contained between T(-a) and T(a). An equivalent statement is that for $∝$ = 0.05, 5% the area of the graph will be outside the range with 2.5% on each side of the graph. The other parameter of the T-distribution is the degrees of freedom which corrects for the uncertainty in estimation of the standard deviation at various sample sizes. Degrees of freedom is equal to sample size minus one.

 How is this used? If you are asked to conduct a study and write a report on it just reporting the sample mean and standard deviation says nothing about the actual value of the mean in the underlying population. However using a t-test allows you to say that according to the data collected we can say with 95% confidence that the value of the population mean lies between two values. In scientific publications the 95% confidence interval is the standard, though sometimes more than 95% confidence will be used, almost never less. The size of the confidence interval actually makes the range of the interval larger. The range from -∞ to ∞ would be a 100% confidence interval for example. So the goal is to collect data with enough precision and accuracy to have a very narrow, high confidence (95% or above) confidence interval.

(example t distributions with different confidence levels with the area outside the interval shaded)

The t-test is very robust, rearranging and manipulating the formula gives you formulas which apply to a variety of situations. Algebraically rearranging the formula of the confidence interval and accounting for the relationships between statistics you can find formulations which are applicable to cases when you are trying to compare averages from multiple sources or the differences between statistics calculated from your data. These all have the general form:

Tcalc= $\frac{\overbar{|x}\_{1}-\overbar{x}\_{2}|}{s\_{d}}\sqrt{n}$

 Where the values of the means and the form of the standard deviation will change according to what you are testing and what your hypothesis is. If you are trying to determine if there is a difference between your mean values and an accepted value then plugging them in as $\overbar{x}\_{1} and \overbar{x}\_{2}$ into the equation along with your standard deviation and the root of your sample size and comparing it to table value is equivalent to checking if the accepted value lies inside or outside the 95% Confidence interval for your data. This simplification works because the standard deviation of an academically accepted value is by definition zero or very close to zero. So the complicated formulations of the pooled two sample standard deviations will simplify back down to this general form. The process is more complicated for non-accepted values such as if you are comparing two different experiments conducted by different analysts or institutions, but the underlying principle is the same: manipulating the numbers in a way that their differences will form their own normal distribution which can be analyzed and compared to tabulated values. This is the power of the t-test.