

## STATISTICS PRIMER FOR CHEMISTRY STUDENTS

Statistics is the study of looking at a small sample and attempting to say something meaningful about the total population of interest. In addition to calculating values such as means and medians you will often try to determine the error possible in your calculation. You usually assume that the variable of interest fluctuates randomly and has a normal distribution, or Gauss distribution. A normal distribution means that the probability of obtaining a certain value is symmetric about the mean. Most random distributions that you will encounter will be normal distributions. The actual function for the normal distribution, called the density, is given by:

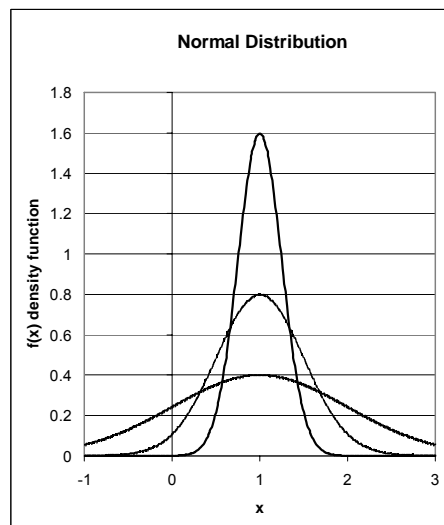
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Where  $\mu$  and  $\sigma$  are the mean and standard deviation of the population - this will be explained below. Plots of the density are shown below in Figure A-1. Note that as the standard deviation gets smaller the normal distribution gets narrower and taller (the area under all the curves will be equal to 1.00 because the area represents the probability of all measurements - consider a finite example, the roll of a die. You know the outcome is a value between 1 and 6 and the probability of any given outcome is 1/6 - and the sum of all possible outcomes is 1.0.) This means that any measurement you make will be more likely to be closer to the mean as the standard deviation gets smaller. When the distribution isn't normal (examples are: skewed - when a test is too easy and everyone gets a 100%, bimodal - when there is an additional factor that separates the population into two distinct groups, for example, physical strength in humans would have a bimodal distribution because the males would have a different average strength than the females, etc.) then our concepts of mean, median, etc. don't really have the same meanings.

Figure 1.

Plots of normal distributions where the means are at 1.0 but the standard deviations are 0.25, 0.50, and 1.00 (narrowest to widest peaks). It is useful as a benchmark to note the percent of the data points in a normal distribution that fall within 1, 2, or 3  $\sigma$  values of the mean,  $\mu$ .

Deviation from the mean	% of Data points
$\mu \pm 1\sigma$	68.26
$\mu \pm 2\sigma$	95.46
$\mu \pm 3\sigma$	99.73



What are these terms mean, average, standard deviation, etc? Note that the symbols  $\mu$  and  $\sigma$  are used to represent the mean and standard deviation of the total population of interest while the sample you are measuring has a mean and standard deviation represented as  $\bar{x}$  and  $s$ , respectively.

Suppose you have made N separate experimental measurements of the same quantity or property. The experimental values that you measured are:  $x_1, x_2, \dots, x_N$ .

The **Average** ( $\bar{x}$ ,  $\bar{x}$  bar): You can average all of the experimental measurements to get a more precise value for the quantity. If the experimental measurements include only random errors, the average will also be a more accurate value for that quantity.

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i = \frac{(x_1 + x_2 + \dots + x_N)}{N}$$

In MS Excel the average of a data set is most efficiently calculated by using the function **AVERAGE(data range)**, where the data range is something like a column of data from A1 to A10 given by (A1:A10).

**Sample Standard Deviation (s):** The standard deviation is a measure of the spread of the experimental values in your data set. If the experimental measurements include large random errors, the sample standard deviation will be large. Similarly, if the property that you are measuring naturally takes on a large range of values (even without experimental errors), the sample standard deviation will be large. (Note that for small data sets s can deviate from the true standard deviation,  $\sigma$ , for the data set, but as the data set gets larger s will approach  $\sigma$ .)

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

$$= \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_N - \bar{x})^2}{N-1}}$$

In MS Excel the standard deviation of a data set is most efficiently calculated by using the function **STDEV(data range)**, where the data range is something like a column of data from A1 to A10 given by (A1:A10).

**Confidence Interval for the mean:** The distribution of the means has a standard deviation,  $\sigma_M$ , obtained from measuring sets of samples, behaves somewhat differently than a normal distribution. For measurements with replaced samples,

$$\mu_M = \mu \quad \text{and} \quad \sigma_M = \frac{\sigma}{\sqrt{N}}$$

Figure 2. shows the relative relations between the total population and the distribution of means obtained by sampling a subset of the population. These values are correct for large samples, but

for small samples, say less than 50, then the distribution is slightly perturbed from this and the Student t-distribution should be used.

The confidence interval for the mean gives you an estimate of how close you are to the true mean based on your experimental sampling of the total population. At best, the average of any set of experimental measurements of a mean is an approximation of the true value of that mean. Suppose that the true value is  $\mu$ . You can use the *average and standard deviation of a set of measurements* to establish a confidence interval for  $\mu$ . This involves calculating upper and lower limiting values such that there is a specified probability (or confidence level) that  $\mu$  lies between the limiting values.

$$\bar{x} - \frac{t \times s}{\sqrt{N}} \leq \mu \leq \bar{x} + \frac{t \times s}{\sqrt{N}}$$

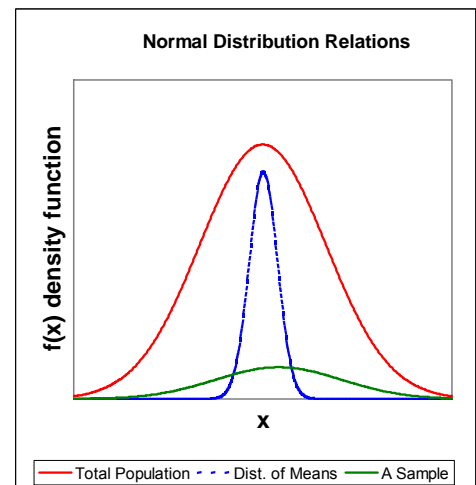
Another way to write the confidence interval is:

$$\mu = \bar{x} \pm \frac{t \times s}{\sqrt{N}}$$

In these equations the t, obtained from a table, is a correction to take into account the Student t-distribution that describes the distribution of the mean more accurately for a small sampling set. If you leave t out you would have a confidence interval corresponding to  $2\sigma_M$  centered at  $\bar{x}$ .

Figure 2.

This set of plots shows the distribution of a total population, the distribution of means from sampling a subset of the total population, and the distribution of a sample population. Note that these plots are not normalized to one.



You can choose different confidence levels by choosing different values for t from Table 1 below. To use the table, you need to know how many “degrees of freedom” exist in your data set. The “degrees of freedom” is defined as:

$$\text{Degrees of freedom} = N - 1$$

(One degree of freedom is already used in the average.)

**One-Tail versus Two-Tails** The other thing you need to know when selecting the t value is whether you are testing a hypothesis that is “one-tailed” or “two-tailed” in the probability distribution

(normal distribution or bell shaped curve, where the “tails” are the wings, or extremes of the curve). What this means is that a one-tailed test is a proposal that the parameter of interest is greater than a certain value that corresponds to the point on the probability distribution where 5% probability is less than that point and 95% probability is above that point (or the opposite, the parameter of interest is less than a certain value that corresponds to the point on the probability distribution where 95% probability is less than that point and 5% probability is above that point). This might be something like a government agency testing the weight of bread made by a bakery and requiring 95% certainty that a loaf of bread weighs more than 1.5 pounds - the concern is only that the weight is above the minimum, not how much above. The attached table is calculated for the one-tailed test. To use the table for a situation where you want a 95% certainty that the result is between the two tails - like the location of the mean in the confidence interval, or to test to see if two means are the same you have to use the column corresponding to ½ of the probability of rejecting the hypothesis. This means that for 95% certainty, for two-tail conditions, the sum of the tails must add up to 5% probability of failure - this means each tail can only contribute 2.5%, or that the column under 97.5% certainty (one-tail) is the correct set of t values for 95% certainty for a two tail test.

You can obtain a narrower interval for a given confidence level by making more experimental measurements (increasing N).

**Example Calculation:**

Suppose you are manufacturing heart valves. You take ten samples from the manufacturing line and weigh them on a balance. The ten masses are; 11.49, 11.53, 11.63, 11.61, 11.62, 11.51, 11.52, 11.57, 11.60, and 11.56 g.

The average mass of the heart valves is:

$$[11.49+11.53+11.63+11.61+11.62+11.51+11.52+11.57+11.60+11.56] \div 10$$

$$= 11.56 \text{ g}$$

The sample standard deviation of the heart valve masses is:

$$s = \sqrt{\frac{(11.49-11.56)^2 + (11.53-11.56)^2 + (11.63-11.56)^2 + (11.61-11.56)^2 + (11.62-11.56)^2 + (11.51-11.56)^2 + (11.52-11.56)^2 + (11.57-11.56)^2 + (11.60-11.56)^2 + (11.56-11.56)^2}{9}}$$

$$s = \sqrt{\frac{(0.0049) + (0.0009) + (0.0049) + (0.0025) + (0.0036) + (0.0025) + (0.0016) + (0.0001) + (0.0016) + (0.0000)}{9}}$$

$$s = \sqrt{0.0025} = 0.05 \text{ g}$$

The 95% confidence interval is:

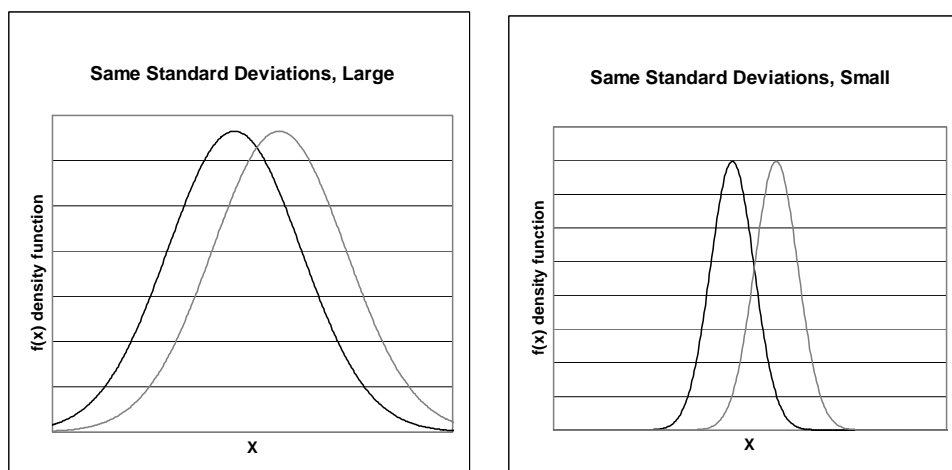
$$\mu = 11.56 \pm \frac{2.262 \times 0.05}{\sqrt{10}} = 11.56 \pm 0.036 \text{ g}$$

This means that there is a 95% probability that the true average mass of all heart valves is between 11.52 and 11.60 g.

### TESTING TWO AVERAGE VALUES - ARE THEY EQUAL?

First, just because the means you calculate for two sets of samples are different numbers, doesn't mean that they are *statistically* different means. This is illustrated in Figure 3. Whether the means are really different depends on how much the distributions overlap, or how big the standard deviations of the two means are relative to the separation of the means

Figure 3. Same pair of means, different standard deviations.



If you need to determine whether two average values are significantly different, you can use a paired t-test. First, you make the hypothesis that the means are the same. Then calculate t according to the formula:

$$t_{calc} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{[(N_1 - 1)(s_1)^2 + (N_2 - 1)(s_2)^2] \cdot [N_1 + N_2]}{(N_1 + N_2 - 2) \cdot N_1 N_2}}}$$

Using the t-table (Table 1) look in the row that corresponds to:

$$\text{Degrees of Freedom} = N_1 + N_2 - 2.$$

and in the column that corresponds to the required degree of confidence (DOC) for a two-tail test:

$$\text{DOC}(\text{two-tail}) \rightarrow \text{DOC}(\text{one-tail}) + (100 - \text{DOC}(\text{one-tail}))/2$$

(For example,  $\text{DOC}(\text{two-tail}) = 95\%$  corresponds to  $[95 + (100 - 95)/2] = 97.5\%$  in the table,)

The two averages are significantly different if the calculated value of  $t_{\text{calc}}$  is outside the range determined from tabulated value of  $t_{\text{table}}$  for the confidence level you desire.

Averages significantly different if:  $|t_{\text{table}}| > t_{\text{calc}}$

If  $|t_{\text{table}}| < t_{\text{calc}}$  then your hypothesis was true and the means are statistically the same.

### Using MS Excel Function TTEST :

The t test in MS Excel returns the probability that the two means are the same - which is a bit different from the above method which give a value that you compare to a range determined from the table.

In MS Excel the t test of two data set is most efficiently calculated by using the function **TTEST(array1,array2,tails,type)**, where array is a data range is something like a column of data from A1 to A10 given by (A1:A10), tails refers to whether you are testing for a greater or lesser value (outside of one tail of the distribution), or is you are looking to see if the result is outside of or between both tails - as in this case. Type is referring to the sample variances that you are assuming - you probably want to assume type = 2 for equal variance - as in sets of data taken on the same system, but at different times.

### Using MS Excel TOOLS t-test:

Going to the TOOLS menu in Excel and selecting Data Analysis followed by t-test

The following data uses the weights from the above example and adds to them an increasing amount of weight to shift the mean. The data sets are tested to see at what point the means are no longer the same within a 95% confidence level. Three methods are used, the MS Excel TTEST function, the MS Excel t-test from the tools menu, and the above formula and the Student t-table at the end of this document. All three give comparable results.

Data Set #	1	2	3	4	5	6	7
Mean Shift	0 g	0.01 g	0.02 g	0.03 g	0.04 g	0.05 g	0.06 g
1	11.49	11.5	11.51	11.52	11.53	11.54	11.55
2	11.53	11.54	11.55	11.56	11.57	11.58	11.59
3	11.63	11.64	11.65	11.66	11.67	11.68	11.69
4	11.61	11.62	11.63	11.64	11.65	11.66	11.67
5	11.62	11.63	11.64	11.65	11.66	11.67	11.68
6	11.51	11.52	11.53	11.54	11.55	11.56	11.57
7	11.52	11.53	11.54	11.55	11.56	11.57	11.58
8	11.57	11.58	11.59	11.6	11.61	11.62	11.63

<b>9</b>	11.6	11.61	11.62	11.63	11.64	11.65	11.66
<b>10</b>	11.56	11.57	11.58	11.59	11.6	11.61	11.62
<b>#</b>	1	2	3	4	5	6	7
<b>N</b>	10	10	10	10	10	10	10
<b>xbar</b>	11.564	11.574	11.584	11.594	11.604	11.614	11.624
<b>s</b>	0.0499	0.0499	0.0499	0.0499	0.0499	0.0499	0.0499

+/- Confidence interval = 0.036 for t(95%) = 2.262

**Excel paired T-test (TTEST Function) goes to a value of 1 as the samples become identical**

paired T test (1,1) =	1.0000	
paired T test (1,2) =	0.6596	Probabilities of being the same
paired T test (1,3) =	0.3822	
paired T test (1,4) =	0.1958	
paired T test (1,5) =	0.0900	
paired T test (1,6) =	0.0380	<== Less than 95% probable at this point
paired T test (1,7) =	0.0150	

**Calculation of tcalc based on t-test formula given in this document and student t-table**

For 18 degrees of freedom, if tcalc is in the range (-2.101 to 2.101), there is a 95% confidence level that the two means are identical.

tcalc (sets 1 & 1) =	0.0000	
tcalc (sets 1 & 2) =	0.4478	
tcalc (sets 1 & 3) =	0.8956	
tcalc (sets 1 & 4) =	1.3434	
tcalc (sets 1 & 5) =	1.7912	
tcalc (sets 1 & 6) =	2.2391	<== Less than 95% probable at this point
tcalc (sets 1 & 7) =	2.6869	

**Calculations from using the tools menu t-Test**

t-Test: Two-Sample Assuming Equal Variances

	t Stat	same as tcalc above
Data Sets = (1,2)	-0.4478	
Data Sets = (1,3)	-0.8956	
Data Sets = (1,4)	-1.3434	
Data Sets = (1,5)	-1.7912	
Data Sets = (1,6)	-2.2390	<== Less than 95% probable at this point
Data Sets = (1,7)	-2.6868	
t Critical one-tail	1.7340	= F(t)*100 for 95% CL with 18 DOF
t Critical two-tail	2.1009	= F(t)*100 for 97.5% CL with 18 DOF

## Calculating a Best Fit Line and Estimation of Error

If one has a data set made up of a variable and a dependent variable with a linear relationship that relates them then we are talking about a set of  $N$  paired data points,  $(x_1, y_1) \dots (x_N, y_N)$ , and a line,  $y = mx + b$ . It is worth taking a brief look at how the “best fit”, “linear regression”, or “least squares” line is determined. The basic notion is that we need to find a straight line from which the data points have the smallest deviation. This is done by minimizing the sum of the squares of the deviations of the data point  $y$  values from the fitted line  $\tilde{y}$  values,  $\tilde{y} = mx + b$ , where the slope is  $m$  and the  $y$ -intercept is  $b$ .

$$\text{minimize } \sum_{i=1}^N (y_i - \tilde{y}_i)^2$$

To begin, one has to collect the “sum of the squares” of the  $X$  values,  $Y$  values, and  $XY$  pair (note:  $S_{XX}$  refers to the sum of the squares of the  $X$  values etc.). These take the following form:

$$S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad S_{YY} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad S_{XY} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Once in hand, these three sums allow us to determine the slope and intercept of the least squares lines as well as the correlation coefficient. The slope and  $y$ -intercept of the least squares line is obtained from:

$$\text{Slope} = m = \frac{S_{XY}}{S_{XX}} \quad \text{and} \quad y\text{-intercept} = b = \bar{y} - m\bar{x}$$

**The  $r^2$  value, termed the coefficient of determination**, is derived from the correlation coefficient,  $r$ , which comes from the expression:

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

The closer either  $r$  or  $r^2$  approaches a value of 1, the better our straight line is considered to “fit” the data. A good fit is typically associated with an  $r^2$  value of at least 0.90.

We are going to do a linear regression analysis on the data to generate the best straight line through the set of data points. This is a very easy task in Excel. Go the **Chart** on the main tool bar, then select **Add Trendline**. In the new window select '**Linear**' (**regression**), then move to the 'Options' window and highlight 'Display equation on chart' and 'Display r-squared value on chart'. Hit OK, and see the results. You now have the equation of the line, and an estimator of how well the line represents the data ( $r^2$ ).

Armed with the equation of the line, we can estimate  $\tilde{y}^*$  for a given x, say  $x^*$ . This would be:

$$\tilde{y}^* = mx^* + b$$

However, to predict  $\tilde{y}^*$  properly from these data, you must provide not simply a single number for the result, but the range allowed by the uncertainty (confidence interval) contained within the data. If the  $x^*$  of interest is in the middle of the experimental data at  $\bar{x}$ , then the spread takes a somewhat familiar form:

$$\tilde{y}^* = (mx^* + b) \pm \left( t_\alpha \frac{S_{x|Y}}{\sqrt{N}} \right)$$

where  $S_{x|Y}$  is similar to a standard deviation, but in this case it is the deviation of the  $y_i$  values from the best fit line (see below). However, as our x value of interest,  $x^*$ , gets further from the range of the sample data one expects the best fit line to diverge more and more from the true line due to the limited data sampled. This introduces an additional term to account from the linear deviation at more distant ranges from the sampled data.

$$\tilde{y}^* = (mx^* + b) \pm \left( t_\alpha S_{x|Y} \sqrt{\frac{1}{N} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \right)$$

The spread in the Y value is defined by the following expression.

Eq.4 
$$\text{Spread}(Y) = S_{x|Y} \sqrt{\left( \frac{1}{N} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right)}$$

where:

$$S_{x|Y} = \sqrt{\frac{\sum_{i=1}^N (Y_i - \tilde{Y}_i)^2}{N-2}} \quad \tilde{Y}_i = \text{Values from fitted line}$$

Note from the expression that as our test value for  $x^*$  approaches the value of the mean,  $\bar{x}$ , the spread approaches a minimum.

To evaluate this expression we need to obtain a value for the sum of the squares of the X values,  $S_{xx}$ , and the value for  $S_{x|Y}$ .

The spread(Y) value is added to and subtracted from the Y value you obtained from your equation of the line as shown below in equation 5. The 't<sub>α</sub>', or 't-distribution' value, comes from a table (commonly found in reference books such as a CRC Handbook of Chemistry and Physics) and depends on the value we set for our confidence level, and the number of (x,y) data pairs. Recall we have chosen 95 % for our confidence level. Obtain t<sub>α</sub> from the table in Appendix A).

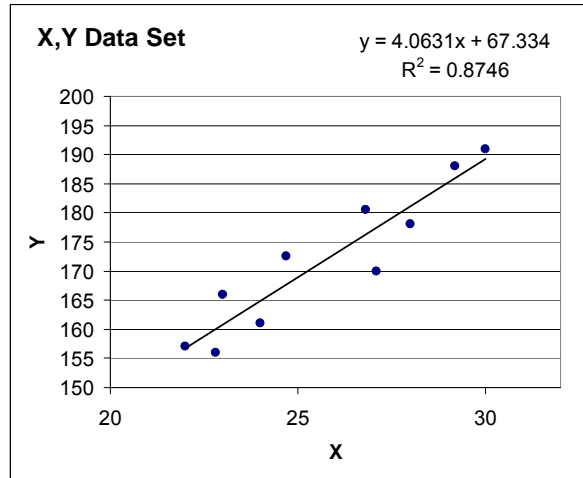
Eq. 5      Upper limit:  $Y + t_\alpha[\text{Spread}(Y)]$       Lower limit:  $Y - t_\alpha[\text{Spread}(Y)]$

**Example Calculation.**

Figure 4. A sample data set of (x,y) values and a best fit line.

Below are the numerical values of the data points and various calculated intermediate values from an Excel spreadsheet.

Notice that the Y test data point is contained in the calculated spread. (These are from actual lab data.)



	<b>Xi</b>	<b>Yi</b>		
1	24.70	172.50	<b>X test =</b>	27.00
2	22.00	157.00	<b>Y test =</b>	177.00
3	27.10	170.00		
4	24.00	161.00	<b>N =</b>	10
5	28.00	178.00		
6	22.80	156.00	<b>Xbar =</b>	25.76
7	30.00	191.00		
8	23.00	166.00		
9	26.80	180.50		
10	29.20	188.00		

	<b>Xi</b>	<b>(Xi-Xbar)^2</b>	<b>yi* = m xi + b</b>	<b>(yi-yi*)^2</b>
1	24.70	1.12	167.69	23.11
2	22.00	14.14	156.72	0.08
3	27.10	1.80	177.44	55.42
4	24.00	3.10	164.85	14.81
5	28.00	5.02	181.10	9.62
6	22.80	8.76	159.97	15.79
7	30.00	17.98	189.23	3.14
8	23.00	7.62	160.79	27.19
9	26.80	1.08	176.23	18.27
10	29.20	11.83	185.98	4.09

<b>SXX =</b>	72.44	<b>talpha =</b>	2.306
<b>SX Y =</b>	4.63	<b>(n-2=8, 95%, 2-tails)</b>	
<b>m=</b>	4.06	<b>Y*+(t*spreadY)=</b>	180.76
<b>b=</b>	67.33	<b>Y*-(t*spreadY) =</b>	173.32
<b>root term =</b>	0.35		
<b>spreadY=</b>	1.61		

**Table 1. Values of *t* at Various Confidence Levels of Probability**

Degrees of Freedom	Confidence Level				
	75%	90%	95%	97.5%	99.5%
1	1.000	3.078	6.314	12.706	63.657
2	.816	1.886	2.920	4.303	9.925
3	.765	1.638	2.353	3.182	5.841
4	.741	1.533	2.132	2.776	4.604
5	.727	1.476	2.015	2.571	4.032
6	.718	1.440	1.943	2.447	3.707
7	.711	1.415	1.895	2.365	3.499
8	.706	1.397	1.860	2.306	3.355
9	.703	1.383	1.833	2.262	3.250
10	.700	1.372	1.812	2.228	3.169
11	.697	1.363	1.796	2.201	3.106
12	.695	1.356	1.782	2.179	3.055
13	.694	1.350	1.771	2.160	3.012
14	.692	1.345	1.761	2.145	2.977
15	.691	1.341	1.753	2.131	2.947
16	.690	1.337	1.746	2.120	2.921
17	.689	1.333	1.740	2.110	2.898
18	.688	1.330	1.734	2.101	2.878
19	.688	1.328	1.729	2.093	2.861
20	.687	1.325	1.725	2.086	2.845
30	.683	1.310	1.697	2.042	2.750
40	.681	1.303	1.684	2.021	2.704
60	.679	1.296	1.671	2.000	2.660
120	.677	1.289	1.658	1.980	2.617
∞	.674	1.282	1.645	1.960	2.576

from the *CRC Handbook of Chemistry & Physics*, 69<sup>th</sup> ed., p. A-105.