06. Statistical Analysis tutorial.doc

Introduction

In this tutorial you will learn to use the basics of statistical analysis to analyze a sample data set. When you complete this tutorial you should be able to take a data set and determine the following:

- Mean
- Median
- Standard deviation
- Standard deviation of the mean
- A confidence limit for the mean

The final bullet point, setting a confidence limit for the mean, is usually the goal for chemists. Experimental results are always subject to uncertainty, assigning a confidence limit allows the chemist to draw comparisons and make meaningful conclusions about their result.

Definitions and formulas

Basic definitions and formulas are given here. Most of these formulas are all accessible on a TI-8x calculator or in MS Excel.

1. **Parent population.** A *hypothetical infinite set of observations* from which the sample data set (experimental data) is taken. We will assume the parent population follows a normal (Gaussian) distribution. (You need not be concerned with the specifics of a normal distribution to make use of this tutorial.)

2. **Sample set (experimental data).** This is the set of data you will be analyzing. This set of data represents a *random sample* of all possible observations from the *parent population*. We will analyze the sample set to make estimates about the parent population. We will also assume the sample data set follows a normal (Gaussian) distribution. The sample data set will contain n observations.

3. **Sample Mean (or average), μ**. Take the sum of the individual observations, x_i, in the sample set and then divide by the number of observations, n.

   $$mean = \mu = \frac{1}{n}\sum_{1}^{n} x_i$$

4. **Median,** $u_{1/2}$. This is the center observation of a sorted sample set. One half of the observations fall below the median while the other half are above the median.

5. **Standard deviation, *s***. This is a measure of how "spread out" the sample observations are about the mean. The smaller the value of ***s***, the more precise is the sample set. The standard deviation is the most widely accepted method in the sciences for determining the precision of a sample set. For a large sample set roughly 68% of all observations will fall within ±***s*** of the mean, and 95% of all observations are within ±2***s*** of the mean. The formula is given below but it is best to use your calculator or computer to find the standard deviation!

$$s = \sqrt{s^s} = \sqrt{\frac{1}{n-1}\sum_{1}^{n}\left(x_i - \mu\right)^2}$$

6. **Standard deviation of the mean, s_μ**. This is the "deviation" or expected error in the mean calculated from the sample set. It is an indicator of the precision of the *sample mean*. Like the standard deviation, we can expect the "true" mean to fall within one standard deviation of the mean ($\mu \pm s_\mu$) 68% of the time and to be within two standard deviations of the mean ($\mu \pm 2s_\mu$) 95% of the time for a large sample set. The standard deviation of the mean is found by dividing the sample set standard deviation by the $\sqrt{n}$.

$$s_u = \frac{s}{\sqrt{n}}$$

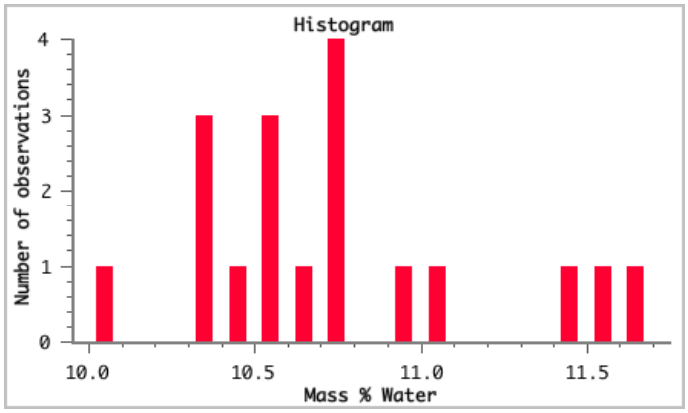
7. **Confidence limit for the mean (Students t-test)**. Once the standard deviation of the mean has been determined we will apply a two-sided Students t-test to set a confidence limit for the mean. The Students t-test will set the upper and lower limits of the mean based on two criteria, 1) the sample set size, *n*, and 2) the confidence limit that the sample mean represents the "true" mean from the parent population. The confidence limit (*CL*) is found by multiplying the standard deviation of the mean the Students t-test value, *t*:
$CL = t^*s_u$.

A worked example

In chemistry 1A we perform an experiment to determine the mass % water in a hydrated salt. Each student measures and reports the mass percent water in their salt sample. All the observations are collected for the class to analyze. The sample data set is given in Table 1. I have sorted the values so the median could readily be determined. I entered the data set in MS Excel for easy analysis.

Table 1. Sample Data Set
% water in a hydrated salt

1	**10.00**
2	**10.31**
3	**10.32**
4	**10.36**
5	**10.44**
6	**10.52**
7	**10.53**
8	**10.57**
9	**10.65**
10	**10.70**
11	**10.70**
12	**10.71**
13	**10.75**
14	**10.91**
15	**11.00**
16	**11.44**
17	**11.56**
18	**11.62**



3. **The sample mean**. There are 18 observations in this sample set. Using Excel μ = 10.72% water by mass in the hydrated salt.

4. **The sample median**. Since there is an even number of observations, the sample median will be the average of the middle two observations, observations 9 and 10:
$$\frac{10.65\% + 10.70\%}{2} = 10.68\% .$$

5. **The standard deviation, *s***. From Excel the sample standard deviation is 0.44%. On average, each observation differs from the mean by 0.44%. Of course some observations differ less than 0.44% and some differ by more than 0.44% As stated earlier, about 68% of the observations are expected to be within $\pm s$ of the mean, that is, 10.72%±0.44%. This gives a range about the mean of 10.28% to 11.16%. Looking at the sample data, those highlighted in blue fall within this range. A total of 14 out of 18 observations or 78% of the sample set.

6. **The standard deviation of the mean, s_μ**. Using the formula given:
$s_u = \frac{s}{\sqrt{n}} = \frac{0.44\%}{\sqrt{18}} = 0.10\%$. This is the standard deviation of the mean. It is interpreted in the same manner as the standard deviation except we are talking about the sample mean, not the observations from the sample data set. An interpretation of s_u is as follows. If we were to return to the lab and repeat the experiment, lets say, 9 more times we would have a total of ten 18-sample data sets. Each of these ten data sets would have a mean. These ten means (from the ten individual experiments) now become our sample data set. The standard deviation of this "set of means" would be expected to be ±0.10%, what was estimated from the single sample data set.

7. **The confidence limit (Students t-test)**. The standard deviation of the mean is a good indication of the precision of your mean. However, we would really like to know the following about the

sample mean:

"How confident am I that the sample mean represents the true mean of the parent population?"

This is where the Students t-test is needed. Think of the t-test as assigning a range to the sample mean that that should include the "true" mean with the stated confidence limit. The stated confidence limit can be as low as, say 50%, or high as, say 99.9%. What is this confidence limit you ask? It is a limit of probability that the "true" mean, the mean of the parent population, will fall within a given range around your sample mean.

To apply the t-test we need 1) a confidence limit and 2) the degrees of freedom ν, for our sample data set. The degrees of freedom equals $n-1$. For our sample set that would be 17.

Lets start with a confidence limit of 90% as a first example. From Table II we find the t value of 1.740 (in blue) that corresponds to 17 degrees of freedom at 90% confidence. This t value is then multiplied by the standard deviation of the mean to determine the range about the sample mean for 90% confidence

$$CL = t^*s_u = 1.740*0.10\% = 0.17\%.$$

We can now state the following: I am 90% confident that the true mass % water in the hydrated salt is between 10.72%±0.17%. If we apply a 99% confidence limit to our mean the t value increases to 2.898 (red) and the range to ±0.29%. With 99% confidence, the true mass % water in the hydrated salt is between 10.43% and 11.01%.

A note on the Students t-test (Table II)

If you look closely at Table II the t values decrease down a column (as sample size increases) and increase across a row (as confidence level increases). This is as expected. As sample size increases the value of the sample mean should better reflect the parent population, hence the uncertainty in the sample mean should decrease as the number of observations increases. Moving across a row the t values increase as the confidence limit increases. To state with more certainty the true value of the mean, you will need to quote a broader range about the sample mean.

Another worked example

The hydrated salt also contains an ion called oxalate. As part of the lab exercise the students also determine the mass % oxalate. This determination is more time consuming than for the mass% water, so the class has fewer observations. The mass % oxalate data is given in Table III and the statistical analysis is summarized below. Make sure you can verify this analysis on your own.

Table III.
% oxalate in a hydrated salt

1	51.8
2	52.0
3	52.4
4	52.7
5	53.1
6	53.4
7	53.4
8	53.7
9	54.4
10	54.6
11	54.7
12	56.1
13	57.3

Sample mean: 53.82%
Sample median: 53.4% (7^{th} observation)
Sample standard deviation: 1.6%
Standard deviation of the mean: 0.44%
90% confidence limit: t = 1.782, CL = 0.79%
90% confidence "true" mean: 53.82%±0.79% = 53.03% to 54.61%

Summary

This simple tutorial provides the minimum skills necessary to find a sample mean and standard deviation. You should be able to report the value of the "true" mean to within a stated confidence limit.

Table II. Two-sided Students t-test values. Calculated in MS Excel using the TINV function.
ν = n-1.

ν	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291