Notes on Statistics

By S.E. Van Bramer

March 8, 1995

1. Terminology

a. **Indeterminate (random) error:** evaluate with statistics

b. **Determinate (systematic) error:** evaluate with reference standards.

c. **Gross error:** big mistake, like spilling everything on the floor.

d. **One sided probability:** Use one sided probability if comparing the size or magnitude from two different data sets (ie. a is larger than b).

e. **Two sided probability:** Use two sided probability if comparing two different data sets for a difference (ie. a is different than b).

f. **Population:** This refers to a set of all possible measurements. This is an ideal that can only be approached. Greek letters are used to symbolize population statistics

g. **Sample:** This refers to a set of actual measurements. The distinction between sample and population statistics is most important for a small number of measurements (less than 20).

h. **t-Test:** This is one of the most powerful and widely used statistical tests. The t-test (Student's t) is used to calculate the confidence intervals of a measurement when the population standard deviation (σ) is not know. Which is usually the case. The t-test is also used to compare two averages. The t-test corrects for the uncertainty of the sample standard deviation (s) caused by taking a small number of samples.

i. **Detection Limit**

i. *Action Limit;* L_c 2σ, 97.7% certain that signal observed is not random noise.

ii. *Detection Limit;* L_D 3σ, 93.3% certain to detect signal above the 2σ action limit when the analyte is at this concentration.

iii. *Quantitation Limit;* L_Q 10σ, Signal required for 10% RSD.

iv. T*ype I Error*; Type I error is identification of random noise as signal.

v. *Type II Error*; Type II error is not identifying signal that is present.

2. **Descriptive Statistics**. These statistics are used to describe a population or a sample.

a. **Population Mean (μ)** and **Sample Average (x, or $\bar{x}$)**

$$\mu = \sum_{i=1}^{N}\left(\frac{x_i}{N}\right) \qquad \text{OR} \qquad \bar{x} = \sum_{i=1}^{N}\left(\frac{x_i}{N}\right)$$

b. **Standard Deviation:** measurement of the spread in individual data points to reflect the uncertainty of a single measurement.

i. *Population standard deviation (σ).* For large sample sets (usually more than 20 measurements) or when the population mean (μ) is known.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$

ii. *Sample standard deviation (s).* For small sample sets (usually less than 20 measurements) when the sample average ($\bar{x}$) is used.

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

iii. *Pooled standard deviation (s_{pooled}).* When several small sets have the same sources of indeterminate error (ie: the same type of measurement but different samples) the standard deviations of the individual data sets may be pooled to more accurately determine the standard deviation of the analysis method.

$$s_{pooled} = \sqrt{\frac{\sum_{i=1}^{N_1}(x_i-\bar{x}_1)^2 + \sum_{j=1}^{N_2}(x_j-\bar{x}_2)^2}{N_1+N_2-2}}$$

c. **Standard Error of the Mean (σ_m).** The standard error of the mean is the uncertainty in the average. This is different from the standard deviation (σ), which is the variation for each individual measurement. Notice that when N is 1 (a single measurement) $\sigma_m = \sigma$.

i. *If σ is known*, the uncertainty in the mean is:

$$uncertainty\ (\sigma_m) = \frac{\sigma}{\sqrt{N}}$$

ii. *If σ is unknown*, use the t-score to compensate for the uncertainty in s. The value for t is obtained from a table for appropriate % confidence level and for N-1 degrees of freedom. (N-1 because one degree of freedom is used to calculate the mean.) Since the uncertainty is a range that could be greater or less than the mean, a two-sided value should be used for t.

$$uncertainty\ (s_m) = t * \frac{s}{\sqrt{N}}$$

d. **z-Score.** Normalizes data points so that the average is 0 and the standard deviation is 1. The cumulative normal distribution (z-score) shows what percentage of a normal distribution is bounded by a given value for z. One sided distributions, the distribution is the area from -∞ and z. Two sided values give the area between ± z. To illustrate this:

i. For a 1 sided distribution 97.72% of all data points will be less than 2 standard deviations above the average.

ii. For a 2 sided distribution 68.28% of all data points will be between ± 1 standard deviation from the average.

$$z = \frac{(x_i - \mu)}{\sigma}$$

Table 1. Cumulative Normal Distribution. The area under a gaussian distribution where z is the population standard deviation (σ).

z	0	1	2	3
$p_{1\ sided}$	.500	.8414	.9772	.9986
$p_{2\ sided}$	0.00	.6828	.9544	.9876

e. **Confidence Interval.** The confidence interval is the preferred method for describing the range of uncertainty in a value. The confidence interval is expressed as a range of uncertainties at a stated percent confidence. This percent confidence reflects the percent certainty that the value is within the stated range.

i. *If the population standard deviation (σ) is known*. The standard error of the mean (σ_m) combined with the z-score (from a table for the desired Confidence Level) is used to express the uncertainty in the mean as a range. This is the confidence interval at the stated certainty. The percentage used should always be stated. This method is widely used to report results with a percent certainty and is expressed as follows:

$$\bar{x} \pm z * \sigma_m \quad \text{OR} \quad \bar{x} \pm \frac{z * \sigma}{\sqrt{N}}$$

Table 2. Values of z for given Confidence Level.

Confidence Level (%)		50	68	90	95	99	99.9
z (2 sided)	0.67	1.000	1.645	1.960	2.576	3.29	
z (1 sided)	0.0	0.407	1.282	1.645	2.326	3.08	

ii. *If the population standard deviation (σ) is unknown*, the sample standard deviation (s) may be used to estimate the confidence interval. This is the preferred method for reporting the uncertainty in experimental results. It takes into account the number of measurements made, the variance in the measurements, and expresses the range at the stated percent confidence level.

$$\bar{x} \pm \frac{(t * s)}{\sqrt{N}}$$

iii. Based upon the confidence interval calculated above, an experimental result should be expressed as:

5.3 ± 1.2 at the 95% confidence level

NOTE: Values for Student's t are given in Table 3.

3. **Comparison Tests.** These tests are used to compare averages to determine if there is a significant difference between two values.

a. **Comparing the sample to the true value.** Method #1. The t-test is used to determine if there is a significant difference between an experimental average and the population mean (μ) or "true value". This method is used to compare experimental results to quality control standards and standard reference materials. This comparison is based upon the confidence interval for the sample mean calculated above. If the difference between the measured value and the true value is greater than the uncertainty in the measurement, there is a significant difference between the two values at that confidence level. This may be expressed mathematically that IF:

$$|\bar{x} - \mu| \leq t * \frac{s}{\sqrt{N}}$$

Then there is no significant difference at the stated confidence level. This could be stated as "there is no significant difference between the experimental results and the accepted value for the Standard Reference Materials at the 95% confidence interval."

b. **Comparing the sample to the true value, Method #2.** This is same test as above, but it is often easier to understand the meaning of the test by calculating an experimental value for t ($t_{experimental}$). Then the experimental t-score ($t_{experimental}$) is compared to t-critical (t_c), the value of t found in a table. $t_{experimental}$ is calculated as follows:

$$t_{experimental} = \frac{(\bar{x} - \mu)}{s} * \sqrt{N}$$

There is a significant difference between the sample average and the true value if $t_{experimental}$ is greater than t_c. t_c is chosen for N-1 degrees of freedom at the desired percent confidence interval. If the experimental value may be greater or less than the true value, use a two sided t-score. If specifically testing for a significant increase or decrease (but not both) use a single sided value for t_c.

c. **Comparing two experimental averages.** The t-test may also be used to compare two experimental averages. This is most accurately done by using the pooled standard deviation and calculating $t_{experimental}$ as:

$$t_{experimental} = \frac{|x_1 - x_2|}{s_{pooled}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

If $t_{experimental}$ is greater than $t_{critical}$ then there is a significant difference between the two means. $t_{critical}$ is determined at the appropriate confidence level from a table of the t-statistic for $N_1 + N_2 - 2$ degrees of freedom.

Table 3. t_c for Normally Distributed Data.

$P_{1\ sided}$	$t_{.60}$	$t_{.70}$	$t_{.80}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
$P_{2\ sided}$	$t_{.20}$	$t_{.40}$	$t_{.60}$	$t_{.80}$	$t_{.90}$	$t_{.95}$	$t_{.98}$	$t_{.99}$
df								
1	.325	.727	1.376	3.078	6.314	12.71	31.82	63.66
2	.289	.617	1.061	1.886	2.920	4.303	6.965	9.925
3	.277	.584	.978	1.638	2.353	3.182	4.541	5.841
4	.271	.569	.941	1.533	2.132	2.776	3.747	4.607
5	.267	.559	.920	1.476	2.015	2.571	3.365	4.032
6	.265	.553	.920	1.440	1.943	2.447	3.143	3.707
7	.263	.549	.896	1.415	1.895	2.365	2.998	3.499
8	.262	.546	.889	1.397	1.860	2.306	2.896	3.355
9	.261	.543	.883	1.383	1.833	2.262	2.821	3.250
10	.260	.542	.879	1.372	1.812	2.228	2.764	3.169
20	.257	.533	.860	1.325	1.725	2.086	2.528	2.845
∞	.283	.524	.842	1.282	1.645	1.960	2.326	2.576

4. **Q-test.** Use to identify statistical outliers in data. This test should be applied sparingly and never more than once to a single data set.

$$Q_n = \frac{|x_a - x_b|}{R}$$

R is the range of all data points, x_a is the suspected outlier, and x_b is the data point closest to x_a. At the 90% confidence interval Q for N replicate measurements is:

Table 4 Q test decision level at 90% confidence interval

N	3	4	5	6	7	8	9	10
Q	.94	.76	.64	.56	.51	.47	.44	.41

5. **Linear Regression.** Fit the line y = mx + b to linear data where x is the dependent variable, y is the independent variable, and x_i is the i'th data point, and N different standards are used. y_{ave} is the average of the y values for the standards, and x_{ave} is the average of the x values for the standards. This method assumes that there is no variance in the value for x.

$$s_{xy} = ss_{xy} = \Sigma(x - x_{ave})(y - y_{ave}) \quad \text{OR} \quad \Sigma x_i y_i - \left(\frac{\Sigma x_i \Sigma y_i}{N}\right)$$

$$s_{yy} = ss_y = \Sigma(y - y_{ave})^2 \quad \text{or} \quad \Sigma y_i^2 - \left(\frac{(\Sigma y_i)^2}{N}\right)$$

$$s_{xx} = ss_x = \Sigma(x - x_{ave})^2 \quad \text{or} \quad \Sigma x_i^2 - \left(\frac{(\Sigma x_i *)^2}{N}\right)$$

$$\text{Slope} = m = \frac{s_{xy}}{s_{xx}}$$

$$\text{Intercept} = b = y_{ave} - (m * x_{ave})$$

Assuming linear function and no replicates, the standard deviation about the regression is:

$$s_r = \sqrt{\frac{s_{yy} - (m^2 * s_{xx})}{(N - 2)}}$$

Uncertainty in $y_{predicted}$:

$$s_y = s_r * \sqrt{1 + \left(\frac{1}{N}\right) + \frac{(x - x_{ave})^2}{s_{xx}}}$$

Uncertainty in $x_{predicted}$ for an unknown with an average signal y_{unk} from M replicates:

$$s_x = \left(\frac{s_r}{m}\right) * \sqrt{\left(\frac{1}{M}\right) + \left(\frac{1}{N}\right) + \left(\frac{(y_{unk} - y_{ave})^2}{s_{xx}}\right)}$$

6. **Error Analysis**. The following techniques are used to determine how error propagates through an experimental procedure. This method is based upon combining the uncertainty for each step.

Table 5. Error Propagation.

Calculation	Example	Standard Deviation
Addition and Subtraction	$x = a + b - c$	$s_x = (s_a^2 + s_b^2 + s_c^2 ...)^{½}$.
Multiplication and Division	$x = a*b/c$	$s_x = x * [(s_a/a)^2 + (s_b/b)^2 + (s_c/c)^2]^{½}$
Exponentiation	$x = a^b$	$s_x = x * b * (s_a/a)$ (no uncertainty in b)
Logarithm	$x = \log_{10} a$	$s_x = 0.434 * (s_a/a)$
	$x = \ln a$	$s_x = s_a/a$
Antilog	$x = \text{antilog}_{10} a$	$s_x = 2.303 * x * s_a$
	$x = e^a$	$s_x = x * s_a$

EXAMPLE:

a. From a calibration curve the concentration of an unknown is 16±2 ppm

b. The solution was prepared by:
   i. dissolving 0.0452±0.0001 g of compound
   ii. in 250.0±0.1 ml of water

c. From this the weight of unknown in the compound is:
   $x = 16 \text{ ppm} * 250 \text{ ml} * (\text{mg/l}) * (1 \text{ g}/1000 \text{ mg}) * (1 \text{ l}/1000 \text{ ml})$
   $= 0.0040$ g of unknown in the compound

d. The uncertainty in this weight is:
   $s_x = x * [(s_a/a)^2 + (s_b/b)^2 + (s_c/c)^2]^{½}$
   $= 0.0040 * [(2/16)^2 + (0.1/250)^2]^{½}$
   $= 0.0040 * (0.0156 + 0.0004)^{½}$ (*NOTE: concentration uncertainty is limiting*)
   $= 0.0005$

e. Therefor the weight of the unknown is:
   0.0040±0.0005 g (or 4.0±0.5 mg)

7. **References**

a. Howard, M.; Workman, J. *Statistics in Spectroscopy*; Academic: Boston, 1991.
b. Box, G.; Hunter, W.; Hunter, J. *Statistics for Experimenters*; Wiley: New York: 1978.
c. Akhnazarova, S.; Kafarov, V. *Experiment Optinization in Chemistry and Chemical Engineering*; MIR: Moscow, 1982.
d. Skoog, D.; West, D.; Holler, F. *Analtyical Chemistry, An Introduction*; Saunders: Philadelphia, 1994.

Statistics Problem Set #1

ENVR 303, 1995. Dr. Van Bramer

1. Use the following data sets for the calculations in this problem set. These data have a random normal distribution. These are all concentrations (ppb) for replicate samples:
   a. Set #1, a large sample set: 25.160, 25.227, 24.402, 23.924, 20.730, 23.615, 23.648, 23.747, 23.613, 22.910, 25.075, 24.301, 24.611, 25.133, 24.152, 24.196, 24.775, 23.841, 24.883, 25.561

   b. Set #2, a small sample set: 22.143, 22.640, 24.084, 23.135, 24.967

   c. A pair of small data sets
      i. Set #3; 11.892, 10.491, 10.172, 12.480, 11.095
      ii. Set #4; 17.656, 16.874, 17.999, 17.825, 19.525, 17.712

2. Descriptive Statistics.
   a. Determine the average and standard deviation for each data set.

   b. Determine the standard error of the mean for each data set.

   c. Determine the following confidence intervals (2 sided) for each data set.
      i. 50%
      ii. 68%
      iii. 90%
      iv. 95%
      v. 99%
      vi. 99.9%

3. Comparison Tests.
   a. For the Set #1 and Set#2; compare the means, and determine if there is a significant difference between the two sets at the 90, 95, and 99% Confidence Interval.

   b. For Set #3 and Set #4; Compare the means, and determine if there is a significant difference between the two sets at the 90, 95, and 99% Confidence Interval.

   c. For the experiments with Set #3, and Set #4, I am really interested in finding if Set #4 is significantly larger than Set #3. Determine if this is true at the 90, 95, and 99% Confidence Interval. Think about what this means in terms of the t-score.

   d. The true values for the data sets are:
      i. Set #1 24.0
      ii. Set #2 24.0
      iii. Set #3 12.0
      iv. Set #4 17.5

      Determine if there is a significant difference between the true value and the measured value for each set at the 80, 90, 95, and 99% Confidence Interval.

4. z-Score. For set #1, normalize the data by z-scoring.

5. Rejection of outlayers. Use the Q-test to determine if 19.525 is an outlier in the second small data set.

Statistics Problem Set #2

ENVR 303, 1995. Dr. Van Bramer

1. Given the following data for mercury concentration in a soil sample, construct a calibration curve, using linear regression determine the unknown concentration. The sample was prepared from a 1.3452 g of a sample (weighed on an analytical balance). Extracting the mercury, and diluting the sample to 100.0 mL in a class A volumetric flask.
   a. Based upon propogation of error, what is the concentration of mercury in the soil sample?
   b. How would you verify the accuracy of this determination.
   c. Based on the replicat samples, what are the different LOD's. How many replicate samples would be required to measure a 0.001 ppm sample with 10% RSD?

Concentration (ppm)	Signal LOG(Po/P)
0	0.02599
0.001	0.03544
0.002	0.03447
0.005	0.05885
0.01	0.04349
0.05	0.23143, 0.22492, 0.22656, 0.27000
0.092	0.41289
0.124	0.55200
unknown	0.21535

2. A unknown sample was prepared for analysis by inductively coupled plasma emission. Solution A; 0.0153±0.0001 g of unknown was weighed out on an analytical balance, dissolved and diluted to 250.0±0.1 ml in a volumetric flask. Solution B; 100.0±0.2 ml of solution A was transferred to a second flask and diluted to 250.0±0.1 ml. Solution C; 0.0021±0.0001 g of iron was dissolved, 100.0±0.2 ml of solution A was added, and the solution diluted to 250 ml. Giving the following data, what is the weight percent of Fe in the unknown sample? Based upon the replicate samples and propagation of error what is the uncertainty in this value? Report your results as ± 90% confidence interval using the t-test. (10 points)

	Trial 1	Trial 2	Trial 3
Blank	0.788847	0.835551	0.996183
Solution B	8.799929	8.880910	8.850854
Solution C	21.71420	21.92304	21.92695

S.E. Van Bramer
March 16, 1995

ENVR 303

Problem Set #1, discussion

I have already posted some of the answers to the statistics problem set, but I also want you to spend some time thinking about what those answers mean. In an effor to address that, I am writing this document.

1. **Data sets:** What you see here is just a list of numbers. What do these numbers correspond to. What are they a measure of. When you apply statistical tests to your own data, you need to think about this. This is important, because it determins what you measure the uncertainty of. For example, lets say that set #1 is measurement of lead in water:

    a. If the set represents samples from 20 different locations in a lake, then you are measuring the variability in the lake at the time you took the samples.

    b. If the set represents samples of tap water taken on 20 consecutive days, then you are measuring the day to day variablity of the tap water.

    c. If the set represents samples of tap water taken one right after the next, then you are measuring the variability of the tap water over this time frame.

    d. If you took a single water sample and are measuring that sample 20 times, then you are measuring the variablity of the analysis.

    All of these measurements represent very different types of information. I want to make two points. First, think carefully about what it is that you want to find out and design your experiment accordingly. Each of the above experiments provides distinctly different information. What are you interested in? Second, when you describe your experiment to someone else be careful about what you say. Say what you did and what you measured. Do not make a claim for a type of information that you did not obtain. ie: from experiment d above, you can say that the concentration of lead in your sample was x +/- s. but you do not know anything about how much that sample will vary from one taken a month later at the same place.

2. **Descriptive Statistics:** These are just statistical tools that you use to represent your data set. It is a shorthand to make life easier. Instead of having to list off all twenty values from set #1, you can describe the set with statistics. When you do this, just be certain to

tell exactly which statistics you use so someone else can tell what you did.

a. POOR, Set #1 is 24.175 ± 1.066.

b. Better, Set #1 is 24.175 ± 1.066 (sample standard deviation from 20 replicat measurements)

c. Best, The average of Set #1 is 24.175 ± 0.467 at the 95% confidence interval.

3. **Comparison Tests:** These statistical tools are used to compare two things to see if there is a significant difference. You can not tell if two sets are the same, only if they are "significantly different". Some examples of what is found

a. Comparing set #1 to the "true value". $t_{experimental} = 0.735$. Then compare this to t_c from the table. Since you are comparing for a difference, this is a two sided test. You have 20 samples, so you have 19 degrees of freedom. Looking in the table that I gave you, the closest that you can find is 20 degrees of freedom (pretty close, if you want to do better interpolate between the values given, or go to a bigger table in a different source). From the table, $t_c = 1.64$. From this you can say "Based upon this experiment there is no significant difference between Set #1 and the accepted value at the 90% confidence interval".

b. Comparing Set #3 and Set #4. Now you are comparing two experimental values. First, you can improve your estimate of the population standard deviation by pooling the standard deviation from each data set. Then you may used this pooled standard deviation to determine $t_{experimental} = 12.146$. From the table, find t_c (for $N_1 + N_2 - 2$ degrees of freedom) = 2.262 (95% CI, 2 sided). Based upon this you can say "There is a significant difference between Set #3 and Set #4 at the 95% confidence interval"

c. Alternatively, if you expect Set # 4 to be larger, you may use a 1 sided $t_c = 1.833$ (95% CI, 1 sided). Based upon this you can say "Set #4 is significantly larger than Set #3 at the 95% confidence interval."

4. **z-Score**: Just to review from class, using Set #3 as an example, what I was looking for here could be shown as:

	Sample	z-score
	11.892	0.6939
	10.491	-0.7658
	10.172	-1.0982
	12.480	1.3066
	11.095	-0.1365
Avg	11.226	0.0
STDS	0.95974	0.894

The z-Score data has been normalized so that the average is 0.0 and the standard deviation is 1.0 (or close, it would work better with a larger data set). This "normalization" makes it easier to visualize the statistical distribution of the data. That some points are above and some below the average. That most of the points are within two standard deviations of the average, etc.

5. **Rejection of Data:** This test is to determine if a data point is a "statistical anomolly." When collecting data, occasionally a data point will be far removed from the average. Keeping this point can skew your results significantly, and it may be a good idea to remove the data point. THIS DOES NOT MEAN THAT THE POINT IS WRONG. If you take enough data points, eventually you will probably get one point that is "way out". Remember DATA = INFORMATION. If you carelessly throw away data points, you are loosing potentially important information. Apply this test with care, reason, and common sense. For this example.

   a. First calculate Q_n for the suspected outlier.
      $Q = (19.525 - 17.999)/(19.525 - 16.874)$
      $Q = 0.5756$

   b. From the table for Q test decision level, with 5 samples at the 90% confidnece interval $Q_{decision} = 0.64$.

   c. Since the tabulated value for Q is larger than our experimental value. There is no statistical basis for removing this data point at the 90% confidence interval. ie: we can not be 90% certain that this data point is an outlier.