Data Analysis, Standard Error, and Confidence Limits

E80 Spring 2011 Notes

We Believe in the Truth

We frequently assume (believe) when making measurements of something (like the mass of a rocket motor) that there is a *true value*, μ, of the measurement and that each individual measurement has some random error in it. We further assume (believe) that the true measurement lies at the center of a distribution of the noisy measurements, and that the distribution is normal (Gaussian) with a *true standard deviation* of σ. The question then arises: From our set of actual (noisy) measurements can we estimate the value of μ, and how certain are we of the estimate? In other words, if we measured the mass of a rocket motor a bunch of times, how close is our estimator to the actual mass, and how certain are we?

Sample Mean of a Set of Measurements

For a set of N measurements

$$x_1, x_2, \cdots, x_N,$$

we can calculate the *sample mean*,

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

which we use as an estimate of the true value of the measured quantity, μ. If we knew the true value, we could calculate the *error* in each measurement as

$$\varepsilon_i = x_i - \mu.$$

However, since we don't know the true value, but only the sample mean, we can calculate the *residuals*

$$e_i = x_i - \overline{x}.$$

Because our mean depends on our measurements, only $N-1$ of our residuals are independent. We have lost a degree of freedom in calculating our residuals, rather than our errors.

We can characterize our residuals with the *sample variance* and *sample standard deviation*. The sample variance is

$$S^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N} e_i^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} x_i^2 - N\left(\overline{x}^2\right)\right].$$

The last form is useful for calculations. The sample standard deviation is

$$S = \sqrt{S^2}.$$

The sample standard deviation is an estimate of σ, the true spread of the distribution of the measurements, but it doesn't relate directly to how far the sample mean is from the true value. That distance is related to the *Standard Error*, which in turn relates to the true standard deviation by

$$SE = \frac{\sigma}{\sqrt{N}}.$$

Since we believe in σ, but usually don't have a way to determine σ, we'll use S and the *Estimated Standard Error,*

$$ESE = \frac{S}{\sqrt{N}}.$$

For a set of measurement with sufficient individual measurements, the probability that the true value of the measurement is within a given range of the sample mean follows the normal distribution with the estimated standard error replacing the sample standard deviation, e.g., for a sample mean of 42.000 and a sample standard deviation of 0.100 in a set of 200 measurements, the estimated standard error is

$$ESE = \frac{S}{\sqrt{N}} = \frac{0.100}{\sqrt{200}} = 0.0071.$$

The likely range of the true value relative to the sample mean is usually reported as a *confidence interval*, e.g.,

$$\bar{x} = 42.000 \pm 0.007(68\% \text{confidence interval}),$$

which means that we are 68-percent certain that μ is within ±0.007 of 42.000.

The confidence interval is usually calculated by multiplying the estimated standard error by a constant related to the area under a standard normal curve, $(ESE \times k)$. For example, if we had sufficient measurements (and the 200 above qualifies), and we wanted a 95% confidence interval, the fraction of the area under a standard normal curve from -1.96 to $+1.96$ is 95%, so one would commonly calculate

$$\bar{x} = 42.000 \pm 0.0071 \times 1.96 = 42.000 \pm 0.014(95\% \text{confidence}).$$

However, as the number of measurements decreases, the normal distribution under-reports the uncertainty. One must use the Student's (W.S. Gossett) *t*-value to accurately estimate the confidence interval. The confidence interval, $\pm\lambda$, is given by

$$\lambda = tESE = \frac{tS}{\sqrt{N}},$$

and t is the Student's *t*-value determined given the degrees of freedom and the desired confidence limit. A portion of the two-tailed table follows:

SIGNIFICANCE LEVEL FOR TWO-TAILED TEST						
df	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
10	1.372	1.812	2.228	2.764	3.169	4.587
20	1.325	1.725	2.086	2.528	2.845	3.850
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

For the example above, if we only had 11 measurements with the same sample mean of 42.000 and sample standard deviation of 0.100, the degrees of freedom would be $df = N - 1 = 11 - 1 = 10$ and the value of t for 95% confidence (5% significance) would be 2.228 as opposed to 1.960 for the normal distribution and

$$\lambda = tESE = \frac{tS}{\sqrt{N}} = 2.228\frac{0.100}{\sqrt{11}} = 2.228(0.030) = 0.067\,,$$

and we would report

$$\bar{x} = 42.000 \pm 0.067(95\%\,\text{confidence})\,.$$

A subtlety that often escapes students is that $\pm\lambda$ is the uncertainty in the average or sample mean of the measurements, but the uncertainty in any individual measurement is governed by *S*, the sample standard deviation. This difference is even more important in linear regression.

Linear Regression

For a set of N measurement pairs, $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$, we can assume that the measurements are linearly related by a function of the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,,$$

where the error, ε_i, is the difference between the true value of y_i and the measured value of y_i. x_i is assumed to either be known exactly, or to contain much less error than y_i.

$$\varepsilon_i = y_i - y_{i(true)} = y_i - \left(\beta_0 + \beta_1 x_i\right).$$

The true values of the set of y's and the true values of β_0 and β_1 are most likely unknown, so we will again work with the residuals:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

and

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i,$$

where

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right).$$

And the circumflex (^) indicates an estimated value, just as $\bar{x}$ is an estimate for μ and S is an estimate for σ. The most common form of linear regression involves minimizing the *Sum of the Squared Residuals* (SSE).

$$SSE = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right]^2.$$

The results of the minimization (the derivation of which can be found many places) are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2},$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{x}$ and $\bar{y}$ are the usual means.

The equivalent of the sample standard deviation for linear regression is the *Root Mean Squared Residual* (RMSE) or S_e.

$$S_e = \sqrt{\frac{SSE}{N-2}} = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N-2}}.$$

The $N-2$ in the denominator comes from the fact that we have lost two degrees of freedom in our residuals because we calculated both $\hat{\beta}_0$ and $\hat{\beta}_1$ from our data.

The sample standard error for $\hat{\beta}_0$ is similar to the sample standard error for a single parameter but has a term to account for the linear fit

$$S_{\beta_0} = S_e \sqrt{\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}} \, .$$

The expression for S_{β_1} just has a term for the linear fit

$$S_{\beta_1} = S_e \sqrt{\frac{1}{\sum_{i=1}^{N} (x_i - \bar{x})^2}} \, .$$

As before, one must use the Student's (W.S. Gossett) *t*-value to accurately estimate the confidence intervals for both $\hat{\beta}_0$ and $\hat{\beta}_1$. The confidence interval, $\pm\lambda_{\beta_0}$, is given by

$$\lambda_{\beta_0} = tS_{\beta_0} \, ,$$

and $\pm\lambda_{\beta_1}$ by

$$\lambda_{\beta_1} = tS_{\beta_1} \, .$$

However, the degrees of freedom used in the table, $df = N - 2$, as explained above.

Sometimes, after calculating the linear fit, one wants to know the confidence interval in y calculated for a specific x, $\pm\lambda_y$. The sample standard error for y is given by

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

And the confidence interval as

$$\lambda_y = tS_y \, .$$

This calculation gives you the confidence interval for the *calculated average* y if you set the experiment to x and repeated the measurement a number of times. Note that this question is very different than asking what the *spread* of measured values would be for y if you set the experiment to x and repeated the measurement a number of times.

For generalized least-square parameter estimations, there are equivalent expressions that either can be derived from first principles, or found in the statistics literature.

Propagation of Errors

Often one needs to calculate a quantity based on several other measured quantities. The question arises: How do errors in the other measured quantities affect the calculated quantities. In particular, assume you have a function

$$F = F(x, y, z, \cdots) \, .$$

How do you calculate the uncertainty or confidence interval in F given the uncertainties or confidence intervals in $x,y,z,\cdots$?

Assume that the residuals are a reasonable approximation for the errors and that the errors are small. Then we can do a Taylor series expansion of F about the true values of the variables, and only keep the first-order terms

$$F-F_{true}=\frac{\partial F}{\partial x}\left(x-x_{true}\right)+\frac{\partial F}{\partial y}\left(y-y_{true}\right)+\frac{\partial F}{\partial z}\left(z-z_{true}\right)+\cdots.$$

With the approximate substitution $\varepsilon_x = x - x_{true}$, etc. we have

$$\varepsilon_F=\frac{\partial F}{\partial x}\varepsilon_x+\frac{\partial F}{\partial y}\varepsilon_y+\frac{\partial F}{\partial z}\varepsilon_z+\cdots.$$

If the errors are systematic, known, and small (so that the linear approximations are accurate) the above expression complete with the algebraic signs on the derivatives will permit one to calculate the error in F fairly accurately.

However, the more common case is that the errors are random variables, ands one makes the assumptions that the errors are uncorrelated, i.e., for a set of data ε_{x_i}, ε_{y_i}, and ε_{z_i} are completely independent of each other. In such a case the uncertainties add in a Root-Sum-of-Squares sense

$$\varepsilon_F=\sqrt{\left(\frac{\partial F}{\partial x}\right)^2\varepsilon_x^2+\left(\frac{\partial F}{\partial y}\right)^2\varepsilon_y^2+\left(\frac{\partial F}{\partial z}\right)^2\varepsilon_z^2+\cdots}.$$

An example will help to clarify the use of the equations: Suppose we want to calculate the resistance of an unknown resistor, R_T, which is the bottom half of a voltage divider with known resistor R_1 on top, and measured input and output voltages V_{in} and V_{out}. The equation for the resistance of R_T is

$$R_T=\frac{R_1 V_{out}}{V_{in}-V_{out}}.$$

The desired expansion is (using a differential for the Taylor series)

$$dR_T=\frac{\partial R_T}{\partial R_1}dR_1+\frac{\partial R_T}{\partial V_{in}}dV_{in}+\frac{\partial R_T}{\partial V_{out}}dV_{out}=\frac{V_{out}}{V_{in}-V_{out}}dR_1+\frac{-R_1V_{out}}{\left(V_{in}-V_{out}\right)^2}dV_{in}+\frac{R_1V_{in}}{\left(V_{in}-V_{out}\right)^2}dV_{out}.$$

Assuming the residuals are good estimates for errors and that the errors are small

$$e_{R_T}=\frac{V_{out}}{V_{in}-V_{out}}e_{R_1}+\frac{-R_1V_{out}}{\left(V_{in}-V_{out}\right)^2}e_{V_{in}}+\frac{R_1V_{in}}{\left(V_{in}-V_{out}\right)^2}e_{V_{out}}.$$

If we knew the exact small values for the residuals, we could use the equation as is, but if we wanted to use the Standard Deviations, Standard Errors, or Confidence Intervals, and we can assume they are uncorrelated, we would add them in the RSS sense

$$e_{R_T} = \sqrt{\frac{V_{out}^2}{\left(V_{in}-V_{out}\right)^2}e_{R_1}^2 + \frac{R_1^2 V_{out}^2}{\left(V_{in}-V_{out}\right)^4}e_{V_{in}}^2 + \frac{R_1^2 V_{in}^2}{\left(V_{in}-V_{out}\right)^4}e_{V_{out}}^2} \ .$$

To be explicit, the e's are replaced by the standard deviation, the standard error, or the confidence interval as appropriate. As a numerical example, assume R_1 is a 20 kΩ ± 1% resistor and V_{in} and V_{out} are both measured by a fully-accurate 12-bit DAQ set to a ±5 V range. The smallest resolvable voltage in a DAQ is the range divided by the number of distinct values, which is calculated as

$$10\text{V}\frac{1}{2^{12}} = \frac{10\text{V}}{4096} = 0.027\text{V} \ .$$

The uncertainty in an individual voltage measurement is ±1/2 LSB (Least Significant Bit) or

$$\frac{\pm 0.027\text{V}}{2} = \pm 0.013\text{V} \ .$$

If $V_{in} = 3.000 \pm 0.013\text{V}$ and $V_{out} = 1.000 \pm 0.013\text{V}$, then the uncertainty in R_T is

$$e_{R_T} = \sqrt{\frac{1^2}{\left(3-1\right)^2}200\Omega^2 + \frac{20\text{k}\Omega^2 1^2}{\left(3-1\right)^4}0.013^2 + \frac{20\text{k}\Omega^2 3^2}{\left(3-1\right)^4}0.013^2} = 230\Omega$$

The calculated value of R_T with the uncertainty is

$$R_T = \frac{R_1 V_{out}}{V_{in}-V_{out}} = \frac{20\text{k}\Omega \bullet 1.000}{3.000-1.000} = 10.00\text{k}\Omega \pm 0.23\text{k}\Omega \ .$$

Often a simplification of the error formula will aid in the calculation. In our example, we can substitute R_T in the error formula:

$$e_{R_T} = \frac{R_T}{R_1}e_{R_1} + \frac{-R_T}{\left(V_{in}-V_{out}\right)}e_{V_{in}} + \frac{R_T\left(V_{in}/V_{out}\right)}{\left(V_{in}-V_{out}\right)}e_{V_{out}} \ ,$$

which is somewhat easier to calculate and also aids in picking component values in a design to minimize the error.

Also, the error terms (the things that get squared under the radical, not the individual residuals) that are 10% or less than the maximum error term can usually be dropped from the calculation because

$$\sqrt{100^2 + 10^2} = \sqrt{10000+100} = \sqrt{10100} = 100.5 \approx 100$$

A shortcut for the calculus-challenged who already have the formula entered in a spreadsheet or other calculation aid is to calculate actual differences in the answer due to the uncertainties in the factors and add the calculated differences in the RSS sense.

Quantization Error

We've neglected one important item in all of these calculations. We've assumed that the individual measurements are made to infinite precision (but because of noise, not to infinite accuracy). Any actual measurement will have a quantization error. In other words, it will be measured to only a finite number of digits, and the last digit will have some uncertainty in it. All of the formulas we have derived assume that the true standard deviation of the measurement is significantly larger than the quantization error. If the standard deviation of the measurement is a factor of ten larger than the quantization error you can pretty much ignore quantization and use the formulas as is. However, if the true standard deviation and the quantization error are comparable, you have to include the quantization error or noise in your calculation. For the purposes of E80 we will use the following procedure:

1. Calculate the quantization range, *q*. As explained above, for a fully-accurate 12-bit DAQ set to a ±5 V range, *q* is calculated as

$$q = 10\text{V}\frac{1}{2^{12}} = \frac{10\text{V}}{4096} = 0.027\text{V}\,.$$

For a digital instrument, like a DMM, it is typically 1 least significant digit. If we assume that quantity being measured has an equally likely chance of having a value anywhere in a quantization range, the uncertainty in a given measurement is $\pm q/2$, but for a series of measurements, the standard deviation is $q/\sqrt{12}$.

2. If $S > \frac{10q}{\sqrt{12}}$, you can ignore quantization in your calculation.

3. If $\frac{q}{\sqrt{12}} < S < \frac{10q}{\sqrt{12}}$, you need to include quantization. Add $q/\sqrt{12}$ to your sample standard deviation in the RSS sense.

$$S_{used} = \sqrt{S^2 + \frac{q^2}{12}}\,.$$

4. If $S < \frac{q}{\sqrt{12}}$, report your confidence interval as $\pm q/2$, and plan to take a stochastic signals class to learn how to do the calculations properly. The Digital Signal Processing (DSP) world has a variety of techniques to deal with these issues, such as dithering, noise shaping, and oversampling. Yes, I know, the confidence interval in 4 is larger than in 3. It’s counterintuitive but correct.