Lecture 11: Standard Error, Propagation of Error, Central Limit Theorem in the Real World

5 October 2005

1 Standard Error

Quick point of terminology: last time, when we talked about getting at the sampling distribution of summary statistics, we mostly looked at their means — the law of large numbers, in particular, is about the mean of the sample distribution. There's also going to be a variance or standard deviation. It's a bit unfortunate, terminologically, but the standard deviation of a sample statistic is called its **standard error**. The main tool for getting at standard errors is the central limit theorem. Recall that \overline{X} has mean μ and variance σ^2/n , so it has standard deviation σ/\sqrt{n} .

2 Propagation of Error

In many experimental lab courses, you learn a rather mysterious-looking formula for the error bars of derived or calculated quantities. It says that if you have a quantity z which is a function of measured quantities x and y, i.e., z = h(x, y), then

$$\sigma_z = \sqrt{\left(\frac{\partial h}{\partial x}\right)^2} \sigma_x^2 + \left(\frac{\partial h}{\partial y}\right)^2 \sigma_y^2$$

where σ_z is the standard deviation of z, and similarly for the other variables. (This formula, and everything which follows, extends in the natural way to functions of more than two variables.)

We are now in a position to see exactly where this formula comes from, and when it's actually valid.

We assume that each of the input quantities x and y is really a random variable, X, Y, which has some average value (μ_x, μ_y) , plus fluctuations around it which represent noise in our apparatus, errors of procedure, gremlins, etc. The value of z we calculate is therefore also a random quantity, Z, because if the fluctuations had come out differently, we'd be plugging different numbers into the function h, and getting a different answer. The question we want to answer is *how* different that result would, probably, be.

Let's start by Taylor-expanding h, making the expansion around the mean values of the input variables:

$$h(X,Y) \approx h(\mu_x,\mu_y) + \left(\frac{\partial h}{\partial x}\right)(X-\mu_X) + \left(\frac{\partial h}{\partial y}\right)(Y-\mu_Y) + \text{higher order terms}$$

A Taylor expansion like this is only valid if the neglected higher-order terms, like, $\frac{1}{2} \left(\frac{\partial^2 h}{\partial x^2}\right) (X - \mu_X)^2$, are small compared to the included terms, like $\left(\frac{\partial h}{\partial x}\right) (X - \mu_X)$. So we want

$$\begin{pmatrix} \frac{\partial h}{\partial x} \end{pmatrix} (X - \mu_X) \gg \frac{1}{2} \left(\frac{\partial^2 h}{\partial x^2} \right) (X - \mu_X)^2 \\ \left(\frac{\partial h}{\partial x} \right) \gg \frac{1}{2} \left(\frac{\partial^2 h}{\partial x^2} \right) (X - \mu_X) \\ \frac{2 \left(\frac{\partial h}{\partial x^2} \right)}{\left(\frac{\partial^2 h}{\partial x^2} \right)} \gg (X - \mu_X)$$

And similarly for Y. We can have this happen either if $X - \mu_X$ is always very small, or if the ratio of the first to the second derivative is always very large — that is, the function h is smooth.

Assumption 1: Measurement errors are small, where the scale for smallness is set by the ratio of first to second derivatives.

If Assumption 1 holds, and we can use our Taylor expansion, we've reexpressed h as a linear combination of random variables, and we know how to handle linear combinations. First, the mean:

$$\mathbf{E}\left[Z\right] = \mathbf{E}\left[h(X,Y)\right] \approx h(\mu_X,\mu_Y) + \mathbf{E}\left[\left(\frac{\partial h}{\partial x}\right)(X-\mu_X)\right] + \mathbf{E}\left[\left(\frac{\partial h}{\partial y}\right)(Y-\mu_Y)\right] \\ = h(\mu_X,\mu_Y) + \left(\frac{\partial h}{\partial x}\right)\mathbf{E}\left[X-\mu_X\right] + \left(\frac{\partial h}{\partial y}\right)\mathbf{E}\left[Y-\mu_Y\right] \\ = h(\mu_X,\mu_Y) + \left(\frac{\partial h}{\partial x}\right)(\mathbf{E}\left[X\right]-\mu_X) + \left(\frac{\partial h}{\partial y}\right)(\mathbf{E}\left[Y\right]-\mu_Y) \\ = h(\mu_X,\mu_Y) + \left(\frac{\partial h}{\partial x}\right)(\mu_X-\mu_X) + \left(\frac{\partial h}{\partial y}\right)(\mu_Y-\mu_Y) \\ = h(\mu_X,\mu_Y)$$

Now we compute the variance:

$$\operatorname{Var}\left(Z\right) = \operatorname{Var}\left(h(X,Y)\right) \approx \operatorname{Var}\left(h(\mu_x,\mu_y) + \left(\frac{\partial h}{\partial x}\right)(X-\mu_X) + \left(\frac{\partial h}{\partial y}\right)(Y-\mu_Y)\right)$$
$$= \operatorname{Var}\left(\left(\frac{\partial h}{\partial x}\right)(X-\mu_X) + \left(\frac{\partial h}{\partial y}\right)(Y-\mu_Y)\right)$$

We can drop $h(\mu_X, \mu_Y)$ because it's just a constant, but now we need to make an additional assumption. **Assumption 2**: The measurement errors in the input variables are independent.

$$\operatorname{Var}(Z) \approx \operatorname{Var}\left(\left(\frac{\partial h}{\partial x}\right)(X-\mu_X)\right) + \operatorname{Var}\left(\left(\frac{\partial h}{\partial y}\right)(Y-\mu_Y)\right)$$
$$= \left(\frac{\partial h}{\partial x}\right)^2 \operatorname{Var}(X-\mu_X) + \left(\frac{\partial h}{\partial y}\right)^2 \operatorname{Var}(Y-\mu_Y)$$
$$= \left(\frac{\partial h}{\partial x}\right)^2 \operatorname{Var}(X) + \left(\frac{\partial h}{\partial y}\right)^2 \operatorname{Var}(Y)$$
$$= \left(\frac{\partial h}{\partial x}\right)^2 \sigma_X^2 + \left(\frac{\partial h}{\partial y}\right)^2 \sigma_Y^2$$

Taking the square root of Var(Z) to get the standard deviation gives us the usual formula for propagation of error.

The most important special case for this is when the values of x and y we plug in to the formula are themselves obtained by averaging many measurements — that X, above, is really \overline{X} , and Y is really \overline{Y} . Let's make the following assumptions.

Assumption 3: Measurement errors are independent from one measurement to the next.

Assumption 4: There are many measurements of each variable.

In this case, we can use the central limit theorem to say more about \overline{X} and \overline{Y} . The mean values of \overline{X} and \overline{Y} are still the population means, μ_X and μ_Y . But now the standard deviations we plug in are standard errors, $s_x = \sigma_X/\sqrt{n}$ and $s_y = \sigma_Y/\sqrt{n}$. Also, \overline{X} and \overline{Y} are Gaussian. Since a linear combination of independent Gaussians is Gaussian, Z is also Gaussian.

So we have the following result.

Suppose $Z = h(\overline{X}, \overline{Y})$, where \overline{X} is the sample mean of measured values of X, and likewise for Y. Then, if Assumptions 1–4 hold, Z is approximately Gaussian, with mean $h(\mu_X, \mu_Y)$, and variance

$$\left(\frac{\partial h}{\partial x}\right)^2 \frac{\sigma_X^2}{n} + \left(\frac{\partial h}{\partial y}\right)^2 \frac{\sigma_Y^2}{n}$$

where n is the number of measurements of each input variable, and σ_X^2 is the true (population) variance of X.

3 The Law of Large Numbers and Central Limit Theorem in Real Data

The law of large numbers and the central limit theorem I've presented assume *independent* data-points. While we can *create* independent data, and a lot of experimental technique, survey design methods, etc., is about *ensuring* our data

are independent, phenomena in the natural world are rarely so cooperative as to be *completely* independent. Fortunately, the asymptotic laws still generally hold when the data values are *not too dependent*. Making this precise involves some mathematics way beyond the scope of this course — though I strongly encourage you to take a course in stochastic processes, where you'll learn all about it — but we can convince ourselves of its validity experimentally in many cases.

Here's one particular case: the wind-tunnel data which we first saw back in the second lecture. We'll look at the acceleration measurements. These are weakly correlated (the correlation between successive values is about -0.017, which is small, but definitely not zero). The equivalent of looking at a sample of n independent draws from the distribution is to look at at a **time average** of Tsuccessive values from the series: $\frac{1}{T} \sum_{i=1}^{T} a_{k+i-1}$. This time average is going to depend on the starting position, k, as well as on the length of the interval over which we average, T. If the system we're looking at is well-behaved, though, our initial starting point (k) makes less and less difference as we look at longer and longer intervals $(T \to \infty)$, just as, with independent samples, the sample mean always converges to the population mean. With a dependent system, what we hope is that the time average converges to the **space average**, which is the mean over the sample space:

$$\lim_{T \to \infty} \frac{1}{T} \sum_{i=1}^{T} a_{k+i-1} = \int af(a) da$$

where f(a) is system's density in the sample space — the fraction of the time it spends near the point a. If this happens, we say that the system is **ergodic**. Ergodicity is extremely important for statistics, because it means that any sufficiently long sequence of data is **representative** of the whole process, and we can use it to make reliable inferences about the system as a whole. It's also extremely important to making statistical mechanics and thermodynamics work. Unfortunately, the math needed to really handle ergodicity is fairly complicated¹, but we can see it demonstrated in our data. After all, if the equation above holds, then, starting from any position k, the time averages should get closer and closer as T gets larger and larger. If we histogram the time averages (Fig. 1), we see that this is indeed the case — they become more and more tightly peaked around a common central value.

If the values a_t are ergodic, then so is any function of a_t . In particular, if we look at the indicator function which says whether or not $a_t \in B$ for some set B, this will also converge on a limiting value, which is the probability of B. We saw something like this in lecture 2, but let's look at it again for the acceleration. Here I've chosen the set $B = [0.05, 0.06] \cup [-0.03, -0.02]$ — i.e., two intervals on either side of zero. (There's no particular interest to this region, I just chose it to show that this works on pretty much any event you like.) As you can see in Fig. 2, the time-average of the number of measurements falling in B converges to

¹Though you might try reading Michael C. Mackey, *Time's Arrow: The Origins of Ther*modynamic Behavior (Dover Books, 2003).



Figure 1: Distribution of the values of time averages. Filled circles: individual measurements. Open circles: averages of successive measurements (T = 2). Squares: averages over thirty time-steps (T = 30); diamonds: T = 100; triangles: T = 1000. Note that as we average over longer and longer times, the distribution gets narrower and narrower, while the center does not move. This indicates that time-averages are converging to a common value, independent of when we start observing the acceleration — that the system is ergodic.

a stable value, no matter when we start making our measurements. (Remember that the sampling rate here is 30kHZ, so 3000 time-steps is one second.)

At this point, we should be pretty much convinced that the law of large numbers holds in this data — that reasonably long samples all look alike, and are all representative of the process as a whole. What about the central limit theorem? More specifically, do the time averages approach a Gaussian distribution?

One way to check this would be to compare the histograms of the timeaverages, as in Fig. 1, to Gaussian density functions with the same mean and variance. But then we'd have to assess whether two more-or-less bell-shaped wiggly curves are good matches, and we'd rather do something easier. The something easier is provided by probability plots, which you read about last week.

Remember how probability plots work: along the horizontal axis, I've plotted all the different values seen in the data, in order. Each value falls at a a certain quantile of the data: the ith largest value is bigger than or equal to a fraction i/nof the sample values. Now, for any distribution, the quantile function Q(p) is the inverse of the cumulative distribution F(x). Just as F(x) answers "what is the probability that a random value is $\leq x$?", Q(p) answers "What value is at least as large as a fraction p of the samples?" The vertical axis gives Q(F(x)), where F(x) is the CDF of the data, and Q is the quantile function of a theoretical distribution, here the standard Gaussian. If the data really does come from the theoretical distribution, then $Q = F^{-1}$, and we should get a straight line, up to sampling error. If not, we'll get something curved. One wrinkle is that a sample from any Gaussian distribution, plotted against the standard Gaussian, should give a straight line, because all Gaussian distributions can be standardized by a linear transformation. So plotting the data against a standard Gaussian lets us check normality.

The next few figures give Gaussian probability plots for the individual acceleration measurements (T = 1), averages over successive pairs of measurements (T = 2), and then over times of length 30, 100 and 1000. What you can see is that the probability plots come closer and closer to straight lines, over more and more of their range, until at T = 1000 we've got something which is really very Gaussian indeed. So it looks like the central limit theorem holds in this real-world, correlated data too.

However, there's an important caveat here. If the CLT worked just like it did in the case of independent data, then the variance of the time-averages should be approximately Var (A)/T. We know that the variance is getting smaller we can see that in Fig. 1 — but is it getting smaller like 1/T?

T	Variance of time averages	$\operatorname{Var}\left(A\right)/T$
1	$7.50 \cdot 10^{-4}$	$7.50 \cdot 10^{-4}$
2	$3.69\cdot10^{-4}$	$3.75\cdot10^{-4}$
30	$2.83 \cdot 10^{-5}$	$2.50 \cdot 10^{-5}$
100	$6.21 \cdot 10^{-6}$	$7.50 \cdot 10^{-6}$
1000	$2.55 \cdot 10^{-7}$	$7.50 \cdot 10^{-7}$

The variance in the time-averages is actually getting smaller *faster* than the

Convergence of frequencies to probabilities



Figure 2: Convergence of relative frequencies to long-run probabilities for real data. The horizontal axis shows time, in steps of 1/30,000 second. The vertical axis shows the fraction of measures to date which fall into the region $B = [-0.03, -0.02] \cup [0.05, 0.06]$. Gray horizontal line: long-run average of this fraction (probability). Solid line: time-averages starting from the first measurement. Dashed line: time-averages starting from the 100,00th measurement. Dotted line: time-averages starting from the 900,000th measurement.



Figure 3: Gaussian probability plot of the acceleration values. Here and in the other probability plots, the diagonal line connects values at the first quartile to those at the third quartile; it serves as a rough guide to the eye.



Figure 4: Gaussian probability plot of the means of pairs of successive accelerations



Figure 5: Gaussian probability plot of the means of thirty successive accelerations. The horizontal line through zero is just a graphics bug.



Figure 6: Gaussian probability plot of the means of 100 successive accelerations



Figure 7: Gaussian probability plot of the means of 1,000 successive accelerations

CLT would predict in independent data. This is basically because the correlation between a_t and a_{t+1} is negative — if one of them fluctuates above the mean value, the other one is apt to move below it, so they're even more likely to cancel out fluctuations around the mean than independent measurements are. The moral of this story is that while time averages converge, and they tend to have a Gaussian distribution when you look at enough of them, you can't, necessarily, assume that they'll have the same Gaussian distribution as if measurements were all independent of one another.