Statistics, Data Analysis, and Machine Learning for Physicists
Timothy Brandt
Spring 2020

Lecture 2

In this lecture we will go through the likelihood in the example of N noisy measurements $\{x_i\}$ of a single quantity y. We will now assume the noise to be Gaussian. Please note that this is a standard assumption (and often a pretty good one), but is almost never right in detail. Assuming an incorrect model for your errors can lead to serious mistakes. The first homework will present realistic examples of how this can happen, even when your errors are actually Gaussian.

In our case, we return to our N noisy measurements $\{x_i\}$. Each of these x_i is then equal to y plus some random noise δ_i; the random noise is different for each measurement. I cannot know what the δ_i are–if I knew them, I would just subtract them off and get y with no uncertainties! I will not assume that I know the values of the δ_i, but rather that I know *the distribution from which they were drawn.* Here, I assume that to be the Gaussian distribution with zero mean:

$$\frac{dp}{d\delta_i} = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{\delta_i^2}{2\sigma_i^2}\right]. \tag{1}$$

It is standard to assume that your errors are drawn from distributions with zero mean. If they are drawn from distributions with a nonzero mean, you would probably call that (possibly unknown) mean a bias or a nuisance parameter and try to fit it out. In this case, I will also assume that I know the variance of the distribution from which the δ_i are drawn. Often this is a reasonable assumption: I have an idea of how my measurement apparatus works, so I can estimate the scatter in measurement values that it will produce.

With this assumption that our errors are Gaussian with known variance, we can write down the *likelihood function*, the probability of our measured $\{x_i\}$ given an assumed value for y. Let's first write down the probability for one of them, x_1. If the actual, physical value that I am trying to measure is y, then in order for me to have actually measured x_1, that means that

$$x_1 = y + \delta_1 \tag{2}$$

and hence

$$\delta_1 = x_1 - y. \tag{3}$$

I can now use my assumption that I know the distribution from which δ_i was drawn to write down

$$\mathcal{L}_1 = \frac{dp(x_1|y)}{dx_1} = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{\delta_1^2}{2\sigma_1^2}\right] = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{(x_1-y)^2}{2\sigma_1^2}\right]. \tag{4}$$

This is the likelihood for y given a single measurement x_1. Note that, since δ_i is a continuous random variable, the likelihood is a probability density. Equation (4) is then the probability of δ_1 lying in the range $[x_1 - y, x_1 - y + \delta x)$. The factor δx will drop out of our analyses later.

For this example, with a single measurement x_1, can anybody guess the value of y for which the likelihood is maximized? You could differentiate with respect to y, but it isn't really necessary in this case. The maximum likelihood is given by $y = x_1$, i.e., my measurement is exactly correct. This is because 0 is the mode of the Gaussian distribution. As a quick reminder, we discussed the mean and variance of a probability distribution last time. The mode is the value with the maximum probability or probability density. The mode might, but might not, be the same as the mean. In this case the mean and the mode are the same.

Now we'll account for the other measurements, $\{x_i\}$ for $i > 1$. Each additional measurement has a likelihood that looks just like Equation (4). If I assume two things:

1. The value of y (i.e. the physical parameter I am trying to measure) is the same for all measurements; and

2. Each measurement is independent;

then the likelihood of getting all of my x_i is simply the product of the individual likelihoods. This is just the rule for combining probabilities of independent events. I then have

$$\mathcal{L} = \prod_{i=1}^{N} \mathcal{L}_i \tag{5}$$

$$= \prod_{i=1}^{N} \frac{\delta x}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{\delta_i^2}{2\sigma_i^2}\right] \tag{6}$$

$$= \prod_{i=1}^{N} \frac{\delta x}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x_i - y)^2}{2\sigma_i^2}\right]. \tag{7}$$

Products are not fun to work with, so I take a logarithm to turn the product into a sum. The log likelihood is then given by

$$\ln \mathcal{L} = \sum_{i=1}^{N} \ln \frac{\delta x}{\sqrt{2\pi}\sigma_i} - \sum_{i=1}^{N} \frac{(x_i - y)^2}{2\sigma_i^2}. \tag{8}$$

The first term is just a constant: N times the log of our small increment in x (to account for the fact that the p is a probability density). The second term might look familiar: it is almost χ^2. One more bit of manipulation gives:

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - y)^2}{\sigma_i^2} = -2 \ln \mathcal{L} + \text{constant}. \tag{9}$$

One way of asking for the best model is to ask for the model that maximizes the likelihood (or, equivalently, the log likelihood). Equation (9) shows that this is also equivalent to finding the model y that minimizes χ^2. I can find this model by differentiating Equation (9) with respect to y. This is a standard technique in statistical analyses. Remember in the back of your mind, though, that Equation (9) is built on the assumption of Gaussian errors with zero mean and known variance. If I differentiate, I obtain

$$\frac{d\chi^2}{dy} = -2 \sum_{i=1}^{N} \frac{(x_i - y)}{\sigma_i^2} \tag{10}$$

Since the likelihood is evidently smooth, an extremum will have $\frac{d\chi^2}{dy} = 0$, or

$$\frac{d\chi^2}{dy} = 0 = -2 \sum_{i=1}^{N} \frac{(x_i - y)}{\sigma_i^2} \tag{11}$$

$$\sum_{i=1}^{N} \frac{y}{\sigma_i^2} = \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \tag{12}$$

$$y \sum_{i=1}^{N} \frac{1}{\sigma_i^2} = \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \tag{13}$$

$$y = \left(\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}\right) \left(\sum_{i=1}^{N} \frac{1}{\sigma_i^2}\right)^{-1}. \tag{14}$$

So the maximum likelihood value is the *weighted average* of my measurements $\{x_i\}$, where the weights are the inverse variances. Note that there is only one extremum and that it must be a minimum because $\chi^2 \to \infty$

as $y \to \pm\infty$. I have therefore found the *global* best-fit model. Finding the global best-fit isn't always this easy.

The next step is to estimate the uncertainty in my inferred value of y. To do this, I can test how sensitive my value of y is to a change in a *single* measurement x_i,

$$\frac{dy}{dx_i} = \frac{1}{\sigma_i^2}\left(\sum_{j=1}^{N}\frac{1}{\sigma_j^2}\right)^{-1} \tag{15}$$

(I have changed the summation over i to summation over j to avoid confusion). The variance of a sum of independent measurements is just the sum of the variances. You can see that for two measurements as follows:

$$\begin{aligned}
\sigma_{a+b}^2 = \langle((a-a_0)+(b-b_0))^2\rangle &= \langle(a-a_0)^2+(b-b_0)^2+2(a-a_0)(b-b_0)\rangle && (16)\\
&= \langle(a-a_0)^2+(b-b_0)^2+2(a-a_0)(b-b_0)\rangle && (17)\\
&= \langle(a-a_0)^2\rangle+\langle(b-b_0)^2\rangle+\langle 2(a-a_0)(b-b_0)\rangle && (18)\\
&= \sigma_a^2+\sigma_b^2+\langle a-a_0\rangle\langle b-b_0\rangle && (19)\\
&= \sigma_a^2+\sigma_b^2 && (20)
\end{aligned}$$

The crucial step here is the assumption that a and b are independent, so that $\langle(a-a_0)(b-b_0)\rangle = 0$. If this is true, then I can write the variance in y due to all of my measurement errors as the sum of the variances due to the individual measurement errors:

$$\begin{aligned}
\sigma_y^2 &= \sum_{i=1}^{N}\left(\frac{dy}{dx_i}\right)^2\sigma_i^2 && (21)\\
&= \sum_{i=1}^{N}\left(\frac{1}{\sigma_i^2}\left(\sum_{j=1}^{N}\frac{1}{\sigma_j^2}\right)^{-1}\right)^2\sigma_i^2 && (22)\\
&= \sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(\sum_{j=1}^{N}\frac{1}{\sigma_j^2}\right)^{-2} && (23)\\
&= \left(\sum_{j=1}^{N}\frac{1}{\sigma_j^2}\right)^{-1}. && (24)
\end{aligned}$$

So that's it, the square of my uncertainty on the measured value of y.

Once I have my maximum likelihood model (and its uncertainty), I can ask whether this exercise was meaningful, in the sense that I can at least partially check my assumption about independent Gaussian errors. *Is my maximum likelihood model a good fit to the data?* Note that this is only one check–errors can pass this check even if they are significantly correlated and/or non-Gaussian. Still, it's almost always worth doing.

Looking back at Equation (9), this is equal to

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{x_i-y}{\sigma_i}\right)^2 = \sum_{i=1}^{N}\left(\frac{\delta_i}{\sigma_i}\right)^2 \tag{25}$$

(assuming, of course, that my model is now correct and that $x_i - y$ are actually my realizations of the errors). Under these assumptions, δ_i/σ_i is a Gaussian random variable of unit variance. The sum of the squares of many Gaussian random variables forms a χ^2-distribution. What is the mean/expectation value of the χ^2 distribution? Well, each term has a mean of 1, since it's just the variance of a Gaussian random variable.

So, if these are all independent, the mean of the χ^2 distribution should be N, my number of data points. In fact, I took some information away when I fit for the maximum likelihood value. Remember that when I had just a single measurement, the maximum likelihood was when $\chi^2 = 0$. It turns out that the mean of the χ^2 distribution is equal to the number of data points N minus the number n of independent free parameters that I fit, in this case 1. The number $N - n$ is referred to as the number of *degrees of freedom*, and χ^2 is often normalized to χ^2_{dof}, which has a mean value of 1.

What is the variance of χ^2 about this mean? To calculate this, we'll use the fact that the fourth moment of the Gaussian distribution with zero mean is

$$\langle x^4 \rangle = 3\sigma^4. \tag{26}$$

You can get this in the same way as the variance of a Gaussian, through integration by parts (with one additional step). Armed with this, let's compute

$$\left\langle \left(\chi^2 - N\right)^2 \right\rangle = \left\langle \left(\sum_{i=1}^{N} \left(\frac{\delta_i}{\sigma_i}\right)^2 - N \right)^2 \right\rangle \tag{27}$$

$$= \left\langle \left(\sum_{i=1}^{N} \left(\frac{\delta_i}{\sigma_i}\right)^2 \right)^2 + N^2 - 2N \left(\sum_{i=1}^{N} \frac{\delta_i}{\sigma_i} \right)^2 \right\rangle \tag{28}$$

$$= \left\langle \left(\sum_{i=1}^{N} \left(\frac{\delta_i}{\sigma_i}\right)^4 \right) + \left(\sum_{i=1}^{N} \sum_{j \neq i} \left(\frac{\delta_i}{\sigma_i}\right)^2 \frac{\delta_i}{\sigma_i} \right) + N^2 - 2N\,(N) \right\rangle \tag{29}$$

$$= \left\langle 3N + N(N-1) + N^2 - 2N^2 \right\rangle \tag{30}$$

$$= 2N. \tag{31}$$

So the variance of the χ^2 distribution is $2N$, where again N is the number of degrees of freedom. Typically, if your model is a good fit to the data and your χ^2 is really drawn from the appropriate χ^2 distribution you should find a best-fit χ^2 roughly equal to your number of data points minus free parameters, where "roughly" means within $\pm\sqrt{2N}$. Values of χ^2 that are much smaller than the number of degrees of freedom could indicate overestimated errors, while large values of χ^2 could indicate that the model is a poor fit, that the data have underestimated error bars, that the errors are non-Gaussian and/or correlated, or something else.

Overestimating the errors on individual measurements will generally increase your estimated error on a model parameter. In other words, you will be doing better than you think you are. It is much more common for people to underestimate their errors on individual measurements. In this case, you will typically overstate your case: you will claim an incorrectly precise constraint on your model.