A Gentle Introduction to Maximum Likelihood for Undergraduate Engineering Audiences

Alp Kucukelbir

version 0.3

1 Introduction

You step on a scale to measure your weight, yet you are unsure of your scale's accuracy. So you take a second measurement of your weight. What do you do? You take the average of those two numbers. My question to you is, *why?* Why not take the median? Or the mode? Or simply *discard* the second number altogether?

Estimation theory is a beautiful and deep topic. Many engineering and computer science courses provide an elementary appreciation for it, generally by working through some examples. Yet many students struggle with the concept (prior exposure to statistics/probability) and mechanics of the examples (jargon, set-up, and computation). My experience has been that these challenges mask the true insights of estimation theory.

My goal in this gentle introduction is to work through, in eye-watering detail, every step of a simple maximum likelihood estimation problem. As a nod to my electrical engineering background, I will ground the entire discussion in a seemingly trivial example of measuring voltage with a voltmeter. Despite its simplicity, this example will motivate many aspects of estimation theory and (hopefully!) answer that first question of *why the mean?*

2 Probabilistic Modeling

The world is not deterministic. There is uncertainty (noise) in our measurements. Probability theory is a powerful framework in which to study and model uncertainty. This is why we use probability to setup estimation problems.

So let's start with our basic example. We have a circuit board and a voltmeter. Our goal is to use our voltmeter to measure the voltage at a certain point on the circuit board. In the absence of any uncertainty (noise), this is what the measurement would look like

$$x = v, \tag{2.1}$$

where x is our measurement (the thing we read off the voltmeter) and v is the *true* voltage at that point in the terminal. This may seem like a pointless equation, but it is critical to understand that *philosophically*, what you measure is **not** the true thing. The measurement x is the result of an action: in this case, plugging a voltmeter into the board and reading off the number. The true thing is the voltage v that actually exists on the board.

The goal of probabilistic modeling is twofold: 1) to *explain* the relationship between your measurement and the true thing, and 2) to *describe* the uncertainty in your measurement. In our case, point 1 is trivial: we measure the voltage directly as explained by Equation (2.1). Since we've covered point 1) above, let's move on to point 2).

I now tell you that the voltmeter is not perfect. I have determined that in reality it tends to introduce small errors in my measurements. I have realized that these errors seem to be more or less in the $[-5, 5]$ volt range while an overwhelming amount of the errors are actually between $[-3, 3]$. Therefore, I **decide** to **model** this "real-world" phenomenon as follows,

$$x = v + \eta, \tag{2.2}$$

$$\eta \sim \mathcal{N}(0, 1). \tag{2.3}$$

Now, what does this say? Remember, I have always encouraged you to read equations out aloud in plain English. So let's do that.

Equation (2.2) says that I measure some number x from my voltmeter. This number is equal to the true thing (voltage v) plus some other number η.

Equation (2.3) says that η is random. This means that if I take one measurement, I will get $x_1 = v + \eta_1$. If I then take another measurement, I will get $x_2 = v + \eta_2$. Note that the true voltage v is the same, while the thing that makes $x_1 \neq x_2$ is that $\eta_1 \neq \eta_2$.

Now, eq. (2.3) says something more about η. Not only is η random, but we also know how it is distributed. In this case, it is distributed as $\mathcal{N}(0, 1)$, a Gaussian centered around 0 with standard deviation 1. This is my model for how my voltmeter makes mistakes in its measurements.

Note that I have not said anything about estimation yet. I am simply *describing* the uncertainty in my measurements. I am **probabilistically modeling** the physics of my voltmeter. That is all. This may be a bad model of reality, in which case I will likely do a poor job of estimating the voltage. Many real-world problems were difficult to tackle until someone came along and proposed a good probabilistic model. Let's assume, for now, that this is a decent model.

In most undergraduate engineering classes, this information will *almost* always be given to you. Your job is to simply understand what is going on and make sure you grasp the subtleties between the true thing and measurements of that true thing. Probability comes in as a tool to describe the uncertainty in your measurements.

Now that we've established the model, we can talk about probabilities of measurements.

3 Probabilities of Measurements

First, let's make our lives a bit easier by combining Equations (2.2) and (2.3) into

$$x \sim \mathcal{N}(v, 1), \tag{3.4}$$

which says: my measurement x is a random number that follows a Gaussian distribution with mean equal to the true voltage v and standard deviation of 1 volt.

Another way of saying this, is by explicitly writing out the *probability* of measuring some value of x. Naturally, you should expect this to be some function of the true voltage v, and for it to take large values around v and smaller values away from v. To state this rigorously, I need to *fix* the value of the voltage v to some value, say $v = v_f$. With this, the probability of measuring some value of x given v is

$$p(x|v = v_f) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - v_f)^2}{2}\right). \tag{3.5}$$

Again, read this out in plain English. This says that, if the voltage on the circuit board happened to be some voltage v_f, then the probability of me measuring some number on my voltmeter is given by the right hand side of Equation (3.5). So let's plug some numbers in, just for fun. Let's say $v_f = 5$ volts. What is the probability of observing $x = 4.7$?

$$p(x = 4.7|v = 5) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(4.7 - 5)^2}{2}\right) \tag{3.6}$$

$$= 0.3814 \tag{3.7}$$

In comparison, what is the probability of observing $x = 2$?

$$p(x = 2|v = 5) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(2 - 5)^2}{2}\right) \tag{3.8}$$

$$= 0.0044 \tag{3.9}$$

So, we are more likely to measure 4.7 volts rather than 2 volts, given that the true voltage is 5 volts. This seems to adhere with our understanding of reality and how well our voltmeter works. Great.

In practice, we tend to drop the explicit notation of $v = v_f$. The following equation is completely equivalent to eq. (3.5)

$$p(x|v) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - v)^2}{2}\right). \tag{3.10}$$

If you get confused, simply re-introduce the explicit notation during your calculations. Do not worry: you will quickly become accustomed to the implicit notation and save the planet with your reduced ink consumption.

Up to this point, we've been discussing *single* measurements up to this point. Now we are going to switch to talking about *multiple* measurements. The reason is, just like the weighing example in the introduction, if you don't trust your voltmeter, you will want to take multiple measurements. Before we address *what to do* with those measurements, we need to introduce the concept of *likelihood functions* and their subtleties.

4 Likelihood

Let's say that we measure the voltage n times, $x_1, x_2, \ldots, x_n$, and stack all those measurements into a vector $X = \{x_1, x_2, \ldots, x_n\}$. If we know that these measurements were done

independently, we can compute the joint probability as a product of the single measurement probabilities. Therefore, the joint probability of observing all n measurements given the true voltage v (which doesn't change between measurements) is

$$p(X|v) = \prod_{i=1}^{n} p(x_i|v) \tag{4.11}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - v)^2}{2}\right). \tag{4.12}$$

This is the *likelihood function*

$$l(X|\theta) = p(X|v), \tag{4.13}$$

$$\theta = v. \tag{4.14}$$

In plain English, it says: the probability of measuring n independent numbers from my voltmeter is equal to the product of measuring each of those numbers separately. Nothing outrageous.

Now in *likelihood* parlance, we like to talk about the likelihood being parametrized by θ. This is simply another way of saying that the likelihood distribution depends on the value of θ. In our case, the measurements X will clearly depend on what the true value of the voltage v is. Therefore, our parameter θ is v, and X is the data (our measurements).

We've spent three full pages talking about all sorts of interesting things, but now we finally get to the goal of this endeavor: estimating θ from X. Or in plain English, what do you do with n measurements from your voltmeter if what you want is the true voltage v in your circuit.

5 Maximum Likelihood Estimation

Maximum Likelihood (ML) estimation is a *framework* of estimating the parameter (what you want) from your data (what you measure). Stated simply, it says this: to estimate θ, first probabilistically model your problem, then calculate the likelihood of your measurements, and finally find the θ that maximizes your likelihood. Mathematically, this looks like this

$$\widehat{\theta}_{ML} = \arg\max_{\theta} l(X|\theta), \tag{5.15}$$

where $\widehat{\theta}_{ML}$ is the maximum likelihood **estimate** of θ, and $\arg\max_\theta$ means 'find the θ that maximizes the following function.' The right hand side of this equation will be a function of the data X. This function is called the maximum likelihood **estimator**. The **estimate** is the output of the **estimator**.

Let's first understand the insight behind this. Remember what the likelihood is. The likelihood is the probability of observing n measurements as a function of the parameter θ. In our case, it is the probability of measuring n values from our voltmeter, given that the true voltage is v. What maximum likelihood tells you is, to estimate the true voltage, just find the value of v that maximizes the probability of observing the measurements. This is the "most likely" value of v.

So, I claim that, for our small example here, the maximum likelihood estimator is

$$\widehat{v}_{ML} = \overline{X} \tag{5.16}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{5.17}$$

You should verify this (do it, please!) by checking if this is indeed the maximizer of Equation (4.12). A useful trick is to take the logarithm of the likelihood, because of this property

$$\arg\max_{\theta} l(X|\theta) = \arg\max_{\theta} \log\left(l(X|\theta)\right). \tag{5.18}$$

This always holds as the logarithm is a monotonically increasing function. (It does not modify the maxima/minima of the function.)

So now we have answered the question of why we should take the mean of our voltmeter measurements! Because taking the mean is the maximum likelihood estimator of the true voltage. And so it enjoys all the wonderful properties of maximum likelihood estimation, which I presume you will cover in class.

(I will probably include a brief discussion on this in the future.)
(I will also probably expand this simple example to MAP estimation.)

6 Version History

`version 0.3:` Typos, thanks to comments from an anonymous EENG 202 alum.

`version 0.2:` Expanded for general audiences.

`version 0.1:` Initial draft, written as a supplement for Yale University's EENG 202 class.