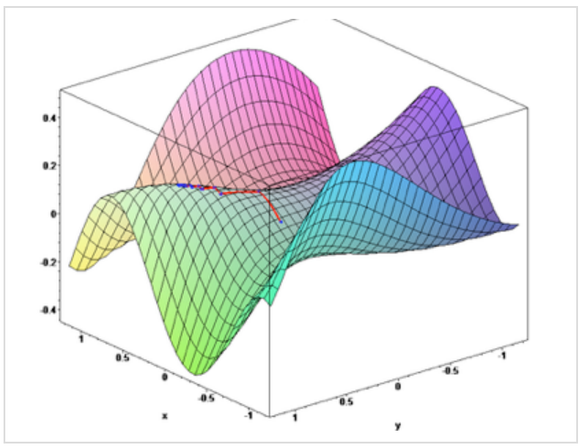
Maximum Likelihood Estimation [Brief Concept Intro]



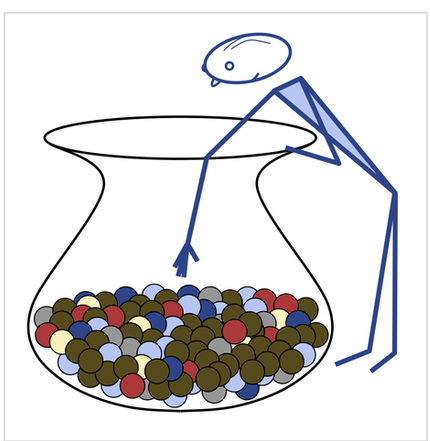
Let's say you are trying to estimate the height of a group of people somewhere. If the group is small enough, you can just measure all of them and be done with it. But in real life, the groups are pretty large and you cannot measure each and every person. So we end up having a model which will estimate the height of a person. For example, if you are surveying a group of professional basketball players, you may have a model which will be centered around 6'7" with a variance of a couple of inches. But how do we get this model in the first place? How do we know if this model is accurate enough to fit the entire group?

Say hello to maximum likelihood estimation

Maximum likelihood estimation (MLE) is a way to estimate the underlying model parameters using a subset of the given set. As in, let's say the group has 50,000 people. We obviously cannot go through all of them to estimate our model. So we pick a small subset of, say, 200 people to build our model. MLE will take this subset and estimate the underlying parameters that will fit the entire dataset (in our case, 50,000).

To get to that model, MLE relies on a mathematical expression known as the 'likelihood function'. This is a function of the sample data. The likelihood of a set of data is the probability of obtaining that particular set of data using the chosen model. This model is basically a probability distribution model, and this expression contains the unknown model parameters. The values of these parameters that maximize the sample likelihood are known as the 'Maximum Likelihood Estimates'. Simple enough, right? Let's move forward.

What is the problem we are trying to solve?



Although we don't want to go into too much mathematical detail here, it's nice to understand the basic framework. Suppose we have a random sample X_1, X_2, ... X_n. When we talk about random samples, it means that values are not fixed and they are distributed in a particular way. We assume that the probability distribution of the sample depends on some unknown parameter. Let's call that parameter 'θ'. So to connect this to our sample, it means that our data points are basically the observed values of these random samples and they are governed by the parameter θ. Let's say you trying to estimate an unknown variable whose value can vary between 5.5 and 6.8. If you measure it to be 6.1, then that is your data point. The fact that is varies between 5.5 and 6.8 is quantified by a "random sample".

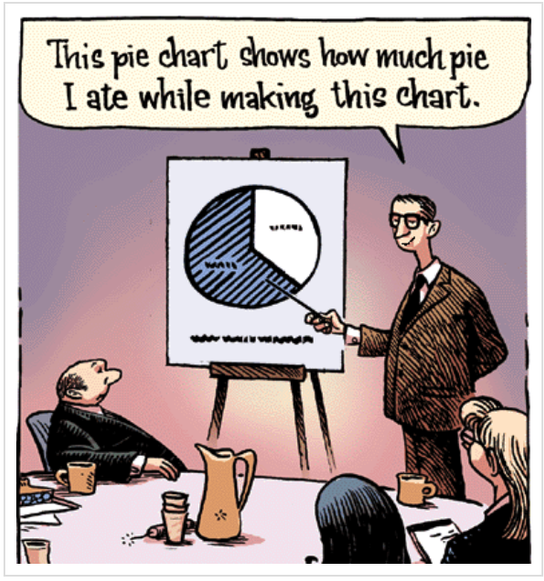
Our primary goal here will be to find a point estimator $u(X_1, X_2, \ldots X_n)$, such that $u(x_1, x_2, \ldots x_n)$ is a good point estimate of θ. Ok now what is this "point estimator"? A point estimator basically calculates a value which is the best estimate of the unknown parameter. Here, $x_1, x_2, \ldots x_n$ are the observed values of the random sample. For example, let's say we are dealing with a random sample $X_1, X_2, \ldots X_n$. If we plan to take this random sample $X_1, X_2, \ldots X_n$ for which the Xi are assumed to be normally distributed with mean μ and variance σ^2, then our goal will be to find a good estimate of μ using the data $x_1, x_2, \ldots x_n$ that we obtained from our specific random sample. We are basically trying to estimate the mean for the whole system using a particular instance.

How do we solve it?

~~You can take a moment here and think about it! What would be a good way to approach this problem?~~ We can see that a good estimate of the unknown parameter, θ, would be the value of θ that maximizes the likelihood of getting the data we observed. That's what we have been talking all along, right? We want that value which will describe the entire system. Since it's probabilistic, we want the parameter to be the most likely parameter that will work well. Hence we have the name "maximum likelihood". Mystery solved!

As we can see here, the idea behind the method of maximum likelihood estimation is fairly straightforward. But how would we implement this method in practice? Well, suppose we have a random sample $X_1, X_2, \ldots X_n$ for which the probability density function of each Xi is $f(x_i; \theta)$. Then, the joint probability density function of $X_1, X_2, \ldots X_n$ is given by $L(\theta)$. Now we can look at a big scary equation that describes this function, but we don't want to do that. We just need to know that X_i are independent and we just take the product of the indexed terms to get the joint probability density function. Alright, so where do we go from here? One reasonable way to proceed is to treat the likelihood function, $L(\theta)$, as a function of θ, and find the value of θ that maximizes it.

Why MLE?



Maximum likelihood estimation is a totally analytic maximization procedure. Now wait a minute, what does that even mean? It means that when we are dealing with MLE, we can readily calculate something using well defined mathematical expressions. It is an important property because a lot of procedures don't have this kind of flexibility! MLE applies to every form of censored or multicensored data. You might ask, why would the data be censored? Well, it's not like we deliberately censor it. It may so happens that we might not have access to all the data we want in real life. So sometimes, part of the data is censored and you have to do your calculations without that data. This makes it harder for other models to deal with because they don't behave nicely when they don't have full access. But MLE doesn't have that problem. Moreover, MLE's and Likelihood Functions generally have very desirable large sample properties:

- They become unbiased minimum variance estimators as the sample size increases. To cut the jibber jabber out, this means that the performance increases as we feed more data. MLE's become increasingly confident about the output as the size increases. This is a very good thing!
- They have approximate normal distributions and approximate sample variances that can be calculated and used to generate confidence bounds. This means that the behavior of MLE is well defined. The behavior of a mathematical framework is very important because we need to understand what it's going to do under extreme conditions. As discussed in the previous point, getting a confidence bound is critical.
- Likelihood functions can be used to test hypotheses about models and parameters. This point is pretty self explanatory. These functions lend themselves to this kind of analysis.

What are the drawbacks?

There are only a couple of drawbacks to MLE's, but they are important. With small samples, MLE's may not be very precise and may even generate a line that lies above or below the data points. If we have small numbers of failures (say 5 to 10), MLE's can be heavily biased. This is expected behavior! Sometimes, the computation power required can be high. If we want to calculate MLE's, we would often need specialized software for solving complex non-linear equations. Given that we have really powerful machines these days, this is not that much of a problem.