

Maximum Likelihood vs. Bayesian Estimation

A comparison of parameter estimation methods

Lulu Ricketts
Apr 20, 2021 11 min read



Hands-on Tutorials



Photo by [kazuend](#) on [Unsplash](#)

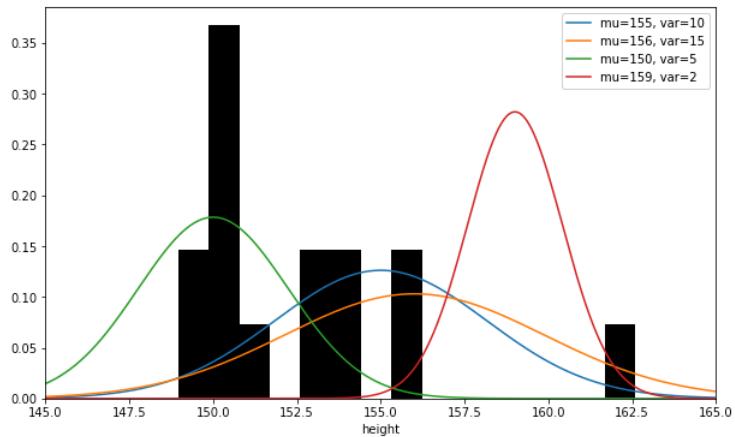
At its core, machine learning is about models. How can we represent data? In what ways can we group data to make comparisons? What distribution or model does our data come from? These questions (and many many more) drive data processes, but the latter is the basis of parameter estimation.

Maximum likelihood estimation (MLE), the [frequentist view](#), and Bayesian estimation, the [Bayesian view](#), are perhaps the two most widely used methods for parameter estimation, the process by which, given some data, we are able to estimate the model that produced that data. Why's this important? Data collected in the real world is almost never representative of the entire population (imagine how much we would need to collect!), and so by estimating distribution parameters from an observed sample population, we can gain insight to unseen data.

As a prerequisite to this article, it is important that you first understand concepts in calculus and probability theory, including joint and conditional probability, random variables, and probability density functions.

Parameter estimation deals with approximating parameters of a distribution, meaning the type of distribution is typically assumed beforehand, which determines what the unknown parameters you will be estimating are (λ for Poisson, μ and σ^2 for Gaussian). The example I use in this article will be Gaussian.

Sample problem: Suppose you want to know the distribution of tree's heights in a forest as a part of an longitudinal ecological study of tree health, but the only data available to you for the current year is a sample of 15 trees a hiker recorded. The question you wish to answer is: "With what distribution can we model the entire forest's trees' heights?"



histogram of observed data (15 samples), and 4 examples of Gaussian curves that could have produced it.
Image by author.

Quick note on notation

- θ is the unknown variables, in our Gaussian case, $\theta = (\mu, \sigma^2)$
- D is all observed data, where $_D = (x_1, x_2, \dots, x_n)$

Likelihood Function

The (pretty much only) commonality shared by MLE and Bayesian estimation is their dependence on the **likelihood** of seen data (in our case, the 15 samples). The likelihood describes the chance that each possible parameter value produced the data we observed, and is given by:

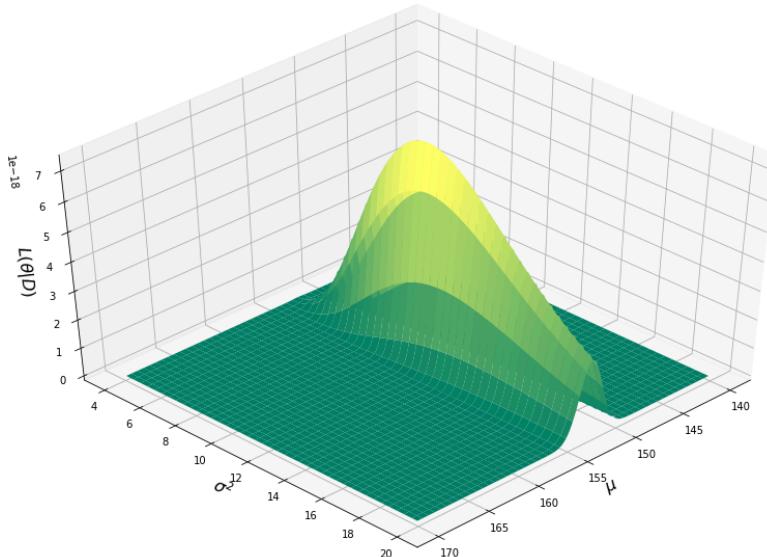
$$\mathcal{L}(\theta | D) = f(D | \theta) = \prod_{i=1}^N f(x_i | \theta)$$

likelihood function. Image by author.

Thanks to the wonderful i.i.d. assumption, all data samples are considered independent and thus we are able to forgo messy conditional probabilities.

Let's return to our problem. All this entails is knowing the values of our 15 samples, what are the probabilities that each combination of our unknown parameters (μ, σ^2) produced this set

of data? By using the Gaussian distribution function, our likelihood function is:



likelihood function over μ, σ^2 . Image by author.

Maximum Likelihood Estimation (MLE)

Awesome. Now that you know the likelihood function, calculating the maximum likelihood solution is *really* easy. It's in the name. To get our estimated parameters (θ), all we have to do is find the parameters that yield the maximum of the likelihood function. In other words, what combination of (μ, σ^2) give us that brightest yellow point at the top of the likelihood function pictured above?

To find this value, we need to apply a bit of calculus and derivatives:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^N f(x_i | \theta) = \frac{\partial}{\partial \theta} \prod_{i=1}^N f(x_i | \theta) \quad X$$

(problematic) calculation of θ . Image by author.

As you may have noticed, we run into a problem. Taking derivatives of products can get really complex and we want to avoid this. Luckily, we have a way around this issue: to instead use the log likelihood function. Recall that (1) the log of products is the sum of logs, and (2) taking the log of any function may change the values, but does not change *where* the maximum of that function occurs, and therefore will give us the same solution.

log likelihood:

$$l(\theta) = \ln f(D | \theta) = \ln \prod_{i=1}^N f(x_i | \theta) = \sum_{i=1}^N \ln f(x_i | \theta) \quad \checkmark$$

$$\text{therefore, } \hat{\theta} = \underset{\theta}{\operatorname{argmax}} l(\theta)$$

correct calculation of θ using log likelihood. Image by author.

It turns out that for a Gaussian random variable, the MLE solution is simply the mean and variance of the observed data. Therefore, for our problem, the MLE solution modeling the distribution of tree heights is a Gaussian distribution with $\mu=152.62$ and $\sigma^2=11.27$.

Bayesian Estimation

Great news! To help you on your search for the distribution of tree heights in this forest, your coworker has managed to go into the data archives and dig up the mean of tree heights in the forest for the past 10 years. With this information, you can now additionally use Bayesian estimation to solve this problem.

The central idea behind Bayesian estimation is that before we've seen any data, we already have some prior knowledge about the distribution it came from. Such prior knowledge usually comes from experience or past experiments. Before jumping into the nitty gritty of this method, however, it is vitally important to grasp the concept of Bayes' Theorem.

Bayes' Theorem

Hopefully you know, or at least heard of, Bayes' Theorem in a probabilistic context, where we wish to find the probability of one event conditioned on another event. Here, I hope to frame it in a way that'll give insight into Bayesian parameter estimation and the significance of priors.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

posterior likelihood prior
 ↓ ↓ ↓
 evidence evidence evidence

Bayes' Theorem. Image by author.

To illustrate this equation, consider the example that event A = "it rained earlier today", and event B = "the grass is wet", and we wish to calculate $P(A|B)$, the probability that it rained earlier given that the grass is wet. To do this, we must calculate $P(B|A)$, $P(B)$, and $P(A)$. The conditional probability $P(B|A)$ represents the probability that the grass is wet given that it rained. In other words, it is the **likelihood** that the grass would be wet, given it is the case that it rained.

The value of $P(A)$ is known as the **prior**: the probability that it rained regardless of whether the grass is wet or not (prior to knowing the state of the grass). This prior knowledge is key because it determines how strongly we weight the likelihood. If we are somewhere that doesn't rain often, we would be more inclined to attribute wet grass to something other than rain such

as dew or sprinklers, which is captured by a low $P(A)$ value. However, if we were somewhere that constantly rains, it is more probable that wet grass is a byproduct of the rain, and a high $P(A)$ will reflect that.

All that's left is $P(B)$, also known as the **evidence**: the probability that the grass is wet, this event acting as the evidence for the fact that it rained. An important property of this value is that its a normalizing constant for the final probability, and as you will soon see in Bayesian estimation, we substitute a normalizing factor in place of the traditional "evidence."

The equation used for Bayesian estimation takes on the same form as Bayes' theorem, the key difference being that we are now using models and probability density functions (pdfs) in place of numerical probabilities.

$$p(\theta|D) = \frac{p(D|\theta) p(\theta)}{\int p(D|\theta) p(\theta) d\theta}$$

posterior distribution likelihood function prior distribution

Bayesian Estimation. Image by author.

Notice that first, the likelihood is equivalent to the likelihood used in MLE, and second, the evidence typically used in Bayes' Theorem (which in this case would translate to $P(D)$), is replaced with an integral of the numerator. This is because (1) $P(D)$ is extremely difficult to actually calculate, (2) $P(D)$ doesn't rely on θ , which is what we really care about, and (3) its usability as a normalizing factor can be substituted for the integral value, which ensures that the integral of the posterior distribution is 1.

Recall that to solve for parameters in MLE, we took the argmax of the log likelihood function to get numerical solutions for (μ, σ^2) . In Bayesian estimation, we instead compute a distribution over the parameter space, called the **posterior pdf**, denoted as $p(\theta|D)$. This distribution represents how strongly we believe each parameter value is the one that generated our data, after taking into account both the observed data and prior knowledge.

The prior, $p(\theta)$, is also a distribution, usually of the same type as the posterior distribution. I won't get into the details of this, but when the distribution of the prior matches that of the posterior, it is known as a conjugate prior, and comes with many computational benefits. Our example will use conjugate priors.

Let's return to our problem concerning tree heights one more time. In addition to the 15 trees recorded by the hiker, we now have means for tree heights over the past 10 years.

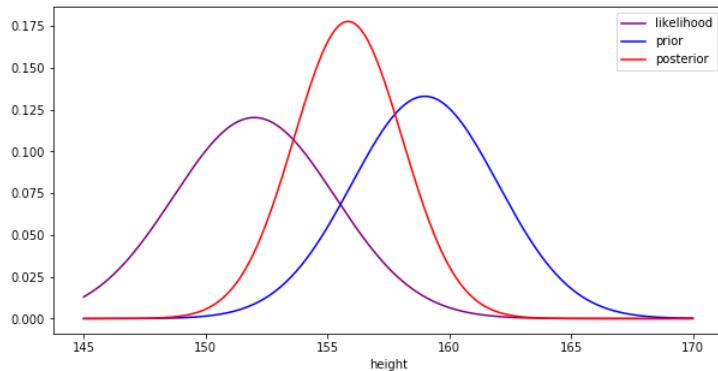
Prev. Year	1	2	3	4	5	6	7	8	9	10
-	160.5	153.8	157.3	160.5	164.0	156.4	157.1	158.1	163.5	160.6

tree heights for the previous 10 years. Image by author.

Following the assumption that this year, tree heights should fall into the distribution of all the previous year's, our prior is the Gaussian distribution with $\mu=159.2$ and $\sigma^2=9.3$.

All that's left is to calculate our posterior pdf. For this calculation, I assume a fixed $\sigma^2 = \sigma^2_{\text{MLE}} = 11.27$. In reality we would not solve Bayesian estimation this way, but multiplying Gaussians of different dimensions, like our likelihood and prior, is extremely complex and I believe simplifying calculations in this case is sufficient for understanding the process and is easier to visualize. If you'd like more resources on how to execute the full calculation, check out [these two links](#).

Multiplying the univariate likelihood and prior and then normalizing the result, we end up with a posterior Gaussian distribution with $\mu=155.85$ and $\sigma^2=7.05$.



likelihood, prior, and posterior distributions. Image by author

And there you have the result of Bayesian estimation. As you can see, the posterior distribution takes into account both the prior and likelihood to find a middle ground between them. In light of new observed data, the current posterior becomes the new prior, and a new posterior is calculated with the likelihood given by the novel data.

Predictions

We have models to describe our data, so what can we do with them? The foremost usage of these models is to make predictions on unseen future data, which essentially tell us how likely an observation is to have come from this distribution. I won't explicitly go through the calculations for our example, but the formulas are below if you'd like to on your own.

MLE Prediction

$$p(x|D) = p(x|\hat{\theta})$$

MLE prediction. Image by author.

Maximum likelihood predictions utilize the predictions of the latent variables in the density function to compute a probability. For instance, in the Gaussian case, we use the maximum likelihood solution of (μ, σ^2) to calculate the predictions.

Bayesian Prediction

$$p(x|D) = \int p(\theta|D) p(x|\theta) d\theta$$

Bayesian prediction. Image by author.

As you probably guessed, Bayesian predictions are a little more complex, using both the posterior distribution and the distribution over the random variable θ to yield the prediction of a new sample.

Concluding Remarks

When to use MLE? Bayesian estimation?

We've seen the computational differences between the two parameter estimation methods, and a natural question now is: When should I use one over the other? While there's no hard and fast rule when selecting a method, I hope you can use the following questions as rough guidelines to steer you in the right direction:

- **How much data are you working with?**

MLE, which depends solely on the outcomes of *observed* data, is notorious for becoming easily biased when the data is minimal. Consider an experiment where you flip a fair coin 3 times, and each flip comes up heads. While you know a fair coin will come up heads 50% of the time, the maximum likelihood estimate tells you that $P(\text{heads}) = 1$, and $P(\text{tails}) = 0$. In situations where observed data is sparse, Bayesian estimation's incorporation of prior knowledge, for instance knowing a fair coin is 50/50, can help in attaining a more accurate model.

- **Do you have reliable prior knowledge about your problem?**

As I just mentioned, prior beliefs can benefit your model in certain situations. However, unreliable priors can lead to a slippery slope of highly biased models that require large amounts of seen data to remedy. Make sure that if you are using priors, they are well defined and **contain relevant insight to the problem you're trying to solve**. If you are unsure about the reliability of your priors, MLE may be a better option, especially if you have a sufficient amount of data.

- **Are you limited in computational resources?**

A recurring theme in this article is that Bayesian computations are more complex than those for MLE. With modern

computational power, this difference may be inconsequential, however if you do find yourself constrained by resources, MLE may be your best bet.

Some interesting properties

Finally, having already discussed many of the differences between MLE and Bayesian estimation, I just want to provide some interesting connections between these two methods.

1. If the Bayesian prior is uniform over all values (an "non-informative prior"), Bayesian predictions will be very similar, if not equal to, MLE predictions.
 2. If the Bayesian prior is well-defined and non-zero at all points, then, as the amount of observed data approaches infinity, MLE and Bayesian predictions will converge to the same value.
-

That's all! If you got this far, thank you for reading. Any feedback is appreciated!

- [Link to code](#)
- • •

WRITTEN BY

Lulu Ricketts

[See all from Lulu Ricketts](#)

Bayesian Statistics

Deep Dives

Hands On Tutorials

Parameter Estimation

Probability Distributions

Share This Article



Towards Data Science is a community publication. Submit your insights to reach our global audience and earn through the TDS Author Payment Program.

[Write for TDS](#)

Related Articles



ARTIFICIAL INTELLIGENCE

How to Forecast Hierarchical Time Series

A beginner's guide to forecast reconciliation

Dr. Robert Kübler

August 20, 2024 13 min read



DATA SCIENCE

Hands-on Time Series Anomaly Detection using Autoencoders, with Python

Here's how to use Autoencoders to detect signals with anomalies in a few lines of...

Piero Pailungo

August 21, 2024 12 min read

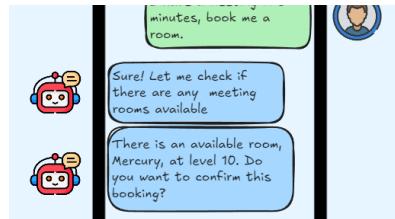
DATA SCIENCE

Solving a Constrained Project Scheduling Problem with Quantum Annealing

Solving the resource constrained project scheduling problem (RCPSP) with D-Wave's hybrid constrained quadratic model (CQM)

Luis Fernando PÉREZ ARMAS, Ph.D.

August 20, 2024 29 min read



Integrating LLM Agents with LangChain into VICA

Learn how we use LLM Agents to improve and customise transactions in a chatbot!

Ng Wei Cheng

August 20, 2024 17 min read



DATA SCIENCE

Towards Generalization on Graphs: From Invariance to Causality

This blog post shares recent papers on out-of-distribution generalization on graph-structured data

Qitian Wu

July 18, 2024 19 min read



Monte Carlo Methods Decoded

Solving complex problems with simulations

Hennie de Harder

February 15, 2024 19 min read



CINEMA

Evaluating Cinematic Dialogue - Which syntactic and semantic features are predictive of genre?

This article explores the relationship between a movie's dialogue and its genre, leveraging domain-driven data...

Christabelle Pabalan



[Subscribe to Our Newsletter](#)

Your home for data science and AI. The world's leading publication for data science, data analytics, data engineering, machine learning, and artificial intelligence professionals.

© Insight Media Group, LLC 2025

[WRITE FOR TDS](#) · [ABOUT](#) · [ADVERTISE](#) · [PRIVACY POLICY](#) ·

[TERMS OF USE](#)

[DO NOT SELL OR SHARE MY PERSONAL INFORMATION](#)