Maximum likelihood estimation (MLE) illustrated with simple examples

Likelihood function Let X be a random variable following an absolutely continuous probability distribution with density function f depending on a parameter θ. Then the function $\mathcal{L}(\theta|x) = f_\theta(x)$ considered as a function of θ, is the likelihood function (of θ, given the outcome x of X).

For example, if the density is the univariate Normal distribution, then :

$$\Theta = [\mu, \sigma]^T \text{ and } f_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

1 Example 1 : A coin tossing experiment

Description Consider a coin tossing experiment where a coin is tossed n times. Let the result of the n successive coin tosses be recorded as $H, T, \ldots, H$ for example. Based on the observed data (the result of n successive coin tosses) we would like to infer the probability that the coin yields a Head on a random coin toss.

If we assume a discrete distribution as the underlying probability model, there is effectively only one parameter of this distribution, namely the probability of a head. Formally, if X is the random variable,

- The sample space (i.e. set of all possible outcomes) is {H,T}
- The probability distribution is given by :

$$P(X = H) = p_H \quad P(X = T) = 1 - p_H$$

- The set of parameters (generally referred to as Θ) consists of only one parameter in this case. i.e. $\Theta = [p_H]$

In Maximum Likelihood Estimation, we define a likelihood function which usually is defined in terms of probability of observing the given data or something along similar lines (for continuous random variables, density function is used).

Let $\vec{o}$ denote the observed data (i.e. the sequence of coin tosses), let the likelihood function be denoted as : $\mathcal{L}(\Theta|\vec{o}$ (the arrow on top is to emphasize that the there are multiple observations i.e. o is a vector).

In the above case, a good choice for the likelihood function is just the probability of observing the data i.e. the probability of getting the sequence of coin tosses.

Suppose we tossed the coin 7 times and got the sequence H, H, T, H, T, T, H. Assuming that the coin tosses are independent, i.e. the result of a coin toss does not depend on outcomes of previous (or later) coin tosses in any way, we get

$$P(\vec{o}) = P(H, H, T, H, T, T, H) = p_H^4(1 - p_H)^3$$

Note : Since we assumed independence of each coin toss, the probability of observing a particular sequence just depends on the number of heads (and hence tails) in that sequence. i.e.
$\{H, H, T, H, T, T, H\}$,$\{H, H, T, T, T, H, H\}$,$\{T, H, H, H, T, T, H\}$ etc.
are all equivalent in this sense.

In general, if there were n coin tosses and n_H Heads were observed we have

$$P(\vec{o}) = p_H^{n_H}(1 - p_H)^{n-n_H} \tag{1}$$

Defining the likelihood function as the probability of getting the observed sequence, from 1 we have :

$$\mathcal{L}(\Theta|\vec{o}) = P(\vec{o}) = p_H^{n_H}(1 - p_H)^{n-n_H} \tag{2}$$

In a number of cases, it is more convenient mathematically to deal with the log of the likelihood than the likelihood itself (the log-likelihood is almost always just the natural logarithm of the likelihood). In our example above, the log likelihood can be defined as :

$$LL(\Theta|\vec{o}) = log(\mathcal{L}(\Theta|\vec{o}) = n_H(log(p_H)) + (n - n_H)log((1 - p_H)) \tag{3}$$

Since log() is monotonic, maximizing the likelihood is equivalent to maximizing the log likelihood. Applying first order conditions of maximization, since there is only one parameter p_H we differentiate with respect to p_H and equate to 0. This yields,

$$\begin{aligned} \frac{d}{dp_H}LL(\Theta|\vec{o}) = & \quad \tfrac{d}{dp_H}n_H(log(p_H)) + (n - n_H)log((1 - p_H)) \\ = & \quad \tfrac{n_H}{p_H} - \tfrac{n-n_H}{1-p_H} \\ = & \quad 0 \qquad \text{first order maximization condition} \end{aligned}$$

This yields the Maximum Likelihood Estimate (MLE) of p_H (which we denote as p_H^*) as :

$$p_H^* = \frac{n_H}{n} \tag{4}$$

Note : To be thorough, one would need to verify the second order condition for maximization also i.e. that the second derivative is negative. Here we have skipped this step for the sake of brevity.

In other words, the Maximum Likelihood estimate of p_H (denoted as p_H^*) is the proportion of heads observed in our experiment.