Handout 09: Maximum Likelihood Estimation (MLE)

Consider once again a random sample of size from n from some distribution parameterized by some unknown θ, $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{G}_\theta$, where the distribution has a pdf or pmf equal to $f(\theta; x)$.[1] The **likelihood (function)** $\mathcal{L}$ is the joint probability of observing any particular set of data $x_1, \ldots, x_n$ as a function of θ. Basically, how likely it is to observe this specific configuration of the data. We can write the likelihood as:

$$\mathcal{L}(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(\theta; x_i).$$

This form comes the fact that the data are assumed to be sampled i.i.d. from the distribution $\mathcal{G}_\theta$.[2] Frequently, it is useful to work with the log-likelihood function l, which has the following form:[3]

$$l(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \log\left[f(\theta; x_i)\right]$$

Now, consider wanting to find a point estimator θ given a sample of data. One very common technique is to take the **maximum likelihood estimator (MLE)**, which finds the value of θ which maximizes the chance of observering our data. Mathematically, we have:

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\arg\max} \left[\mathcal{L}(\theta; x_1, \ldots, x_n)\right],$$

Where Θ are the set of allowed values that θ can take on. Maximizing the likelihood is the same as maximizing the log-likelihood; this is often the form that is easier to compute.

The MLE has a number of important properties:

- **asymptotically unbiased**: while the MLE is often biased, in the limit of large n, the bias will limit to zero.
- **consistent**: the MLE is consistent under very weak regularity conditions.
- **invariant**: the MLE for estimating a transformation of θ will be equal to the transformation of the MLE of θ. For example, the MLE for the standard deviation will always be the square root of the MLE for the variance.
- **efficent**: this is a stronger form of consistency. In short, the rate of convergence to θ has n grows is optimal.

Due to these properties, the fact that the MLE is a very intuitive type of estimator, and (as we will soon seen), we can compute approximate hypothesis tests using them, the MLE is a very popular technique for generating a point estimator in any case where it can be computed.

[1] Throughout this handout, any many others to come, I will use the convention of using a semicolon to seperate the inputs of functions that have both random and non-random inputs. Often, we will drop the random inputs (here, x) when it is clear from the context.

[2] Please review the Joint Distributions section of the probability review handout for you want more details.

[3] Like many statistical texts, I use the convention that $log()$ is the natural logarithm. We never need anything other than the natural log.

Example (Poisson rate) Let's see an example of how we compute the MLE estimator for the estimation of the parameter $\lambda > 0$ from a sample of data that follows the Poisson distribution. To start, we write down the log-likelihood:

$$\begin{aligned} l(\lambda; x_1, \ldots, x_n) &= \sum_{i=1}^{n} \log \left[f(\lambda; x_i) \right] \\ &= \sum_{i=1}^{n} \log \left[\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right] \\ &= \sum_{i=1}^{n} \left[x_i \cdot \log(\lambda) - \lambda - \log(x_i!) \right] \end{aligned}$$

We want to find the maximum, so let's take the derivative with respect to λ (note that the pesky $x_i!$ term disappears):

$$\frac{\partial}{\partial \lambda} l(\lambda; x_1, \ldots, x_n) = \sum_{i=1}^{n} \left[\frac{x_i}{\lambda} - 1 \right]$$

Now, we want to find the maximum of this function. So, let's set the derivative equal to zero. This is usually the step where we replace the generic parameter (λ) with the hat version ($\hat{\lambda}$):

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \left[\frac{x_i}{\hat{\lambda}} - 1 \right] \\ 0 &= \sum_{i=1}^{n} \left[\frac{x_i}{\hat{\lambda}} \right] - n \\ n &= \sum_{i=1}^{n} \left[\frac{x_i}{\hat{\lambda}} \right] \\ \hat{\lambda} &= \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \bar{x}. \end{aligned}$$

So, the MLE for λ is just the sample mean. The average of a Poisson distribution is λ, so this should seem very reasonable, if not particularly exciting.

When working with the MLE for a family of distributions with multiple parameters, we repeat this process by taking the derivative with respect to each parameter and setting them all equal to zero and solving the system of equations.