36-705: Intermediate Statistics **Fall 2019**

Lecture 19: October 11

Lecturer: Siva Balakrishnan

Today we will try to address the question of what is the asymptotic distribution of the MLE. This is analogous to the CLT which gave the asymptotic distribution of averages.

In some cases, we can do this directly. For instance, if $X_1, \ldots, X_n \sim \text{Ber}(p)$ then the MLE is just the average:

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

and so we know by the CLT:

$$\sqrt{n}\frac{\widehat{p} - p}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1),$$

which tells us the asymptotic distribution of the MLE.

More generally, however the MLE need not be a simple average of i.i.d. terms, but the main take-away is that asymptotically it often behaves like one.

19.1 Reminder

Just to remind you - in the lecture on the Cramér-Rao bound, we defined the score,

$$s(\theta) = \sum_{i=1}^{n} \nabla \log(p(X_i;\theta)),$$

which is the gradient of the log-likelihood, and the Fisher Information,

$$I(\theta) = \mathbb{E}[s(\theta)s(\theta)^T].$$

We showed that $s(\theta)$ has mean 0, so $I(\theta) = \text{Var}(s(\theta))$. The Fisher information is alternatively the expected Hessian of the log-likelihood:

$$I_n(\theta) = \mathbb{E}\left[\sum_{i=1}^{n} \nabla^2 \log p(X_i;\theta)\right].$$

It is worth remembering that the score is data-dependent, while the Fisher Information is not (it is an expectation over the data so does not depend on the values of $X_1, \ldots, X_n$).

Let $\widehat{\theta}$ denote the MLE. The rough goal for most of today is to show that (under enough regularity conditions),

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, [I_1(\theta)]^{-1}).$$

19.2 Counterexample

The usual counterexample to the above convergence in distribution is the MLE for the uniform distribution.

For the uniform distribution most regularity conditions fail. Formally, we observe $X_1, \ldots, X_n \sim U[0, \theta]$ and want to estimate θ. The log-likelihood:

$$\mathcal{LL}(\theta) = \log \left[\frac{1}{\theta^n} \mathbb{I}(\theta \geq \max_{i=1}^{n} X_i) \right].$$

The MLE is just $\widehat{\theta} = \max_{i=1}^{n} X_i$. Firstly, you should observe that the log-likelihood is not differentiable at the MLE, so the Fisher information is not defined at the MLE.

Another thing that we used frequently in defining the equivalent forms of the Fisher information was to exchange derivatives (with respect to θ) and integrals (with respect to X). This in general does not work if the domain of integration depends on the parameter with respect to which we are taking the derivative. For the uniform distribution the domain of density depends on the parameter.

On the other hand, things are usually nice for exponential families. They will automatically satisfy all the regularity conditions (provided it is identifiable, i.e. say full-rank and minimal) and the MLE is extremely well-behaved in such models.

Returning to the uniform case, we can directly analyze the distribution of the MLE. In Lecture 4, we showed the following:

$$n(\widehat{\theta} - \theta) \xrightarrow{d} -\text{Exp}(1/\theta)$$

(we did this when $\theta = 1$ but you can work out the general case). So it should be clear that, $\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \delta_0$, where δ_0 is a point mass at 0 and it does not have a Gaussian limit.

19.3 MLE asymptotics

We will only attempt a heuristic calculation here. If you are curious to see a rigorous proof with minimal regularity assumptions you should look at Van der Vaart's book on Asymptotic Statistics. Here is a list of some sufficient regularity conditions:

1. The dimension of the parameter space does not change with n, i.e. $\theta \in \mathbb{R}^d$ and d is fixed. We have seen that if d grows the MLE need not even be consistent.
2. $p(x;\theta)$ is a smooth (thrice differentiable) function of θ,
3. We can interchange differentiation with respect to θ and integration over X. This in turn requires that the range of X does not depend on θ, and some integrability conditions on $p(x;\theta)$.
4. The parameter θ is identifiable.
5. If the parameter space is restricted, i.e. $\theta \in \Theta$ for some set Θ then θ is in the interior of the set Θ (i.e. cannot be on its boundary).

We will focus on the case when the parameter is one-dimensional, although everything carries over almost exactly in the general (fixed) d case.

Theorem 19.1 *Under the regularity conditions above,*

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, 1/I(\theta)).$$

We note that under the conditions of the theorem one can verify that the MLE is consistent, i.e. that $\widehat{\theta} \xrightarrow{p} \theta$. The basic idea is to verify that under the differentiability assumptions on the density, we can effectively treat the parameter space as compact, then derive a uniform law of large numbers, and then apply the proof from the previous lecture notes. This is a complicated technical proof but you can look it up by searching for Wald's proof of the consistency of the MLE.

The proof will use all the facts about scores and the Fisher information that we derived earlier.

Proof: This is not a complete proof. I will try to point out why the various regularity conditions are needed.

To begin with let us note the following fact: if $\widehat{\theta} \xrightarrow{p} \theta$, then

$$\mathbb{E}_\theta[-\nabla^2_\theta \log p(X;\widehat{\theta})] \xrightarrow{p} \mathbb{E}_\theta[-\nabla^2_\theta \log p(X;\theta)] = I(\theta).$$

Roughly, this is saying that as $\widehat{\theta}$ gets close to θ the Hessian of the log-likelihood at $\widehat{\theta}$ gets close to the Hessian of log-likelihood at θ. This is just a smoothness assumption on the Hessian, which is why we assumed that $p(X;\theta)$ is thrice differentiable.

Since $\widehat{\theta}$ maximizes the log-likelihood we know that the derivative of the log-likelihood at $\widehat{\theta}$ must be 0, i.e.

$$\mathcal{LL}'(\widehat{\theta}) = 0.$$

Formally you need to know that $\widehat{\theta}$ is not on the boundary of the parameter space. To prove this you will need to use the fact that θ is not on the boundary and that $\widehat{\theta} \xrightarrow{p} \theta$.

By a Taylor expansion of the derivative of the log-likelihood we obtain that,

$$0 = \mathcal{LL}'(\widehat{\theta}) = \mathcal{LL}'(\theta) + (\widehat{\theta} - \theta)\mathcal{LL}''(\widetilde{\theta}),$$

where $\widetilde{\theta}$ is some point in between $\widehat{\theta}$ and θ. This in turn gives us that,

$$(\widehat{\theta} - \theta) = \frac{\mathcal{LL}'(\theta)}{-\mathcal{LL}''(\widetilde{\theta})},$$

so that,

$$\sqrt{n}(\widehat{\theta} - \theta) = \frac{\frac{\mathcal{LL}'(\theta)}{\sqrt{n}}}{-\frac{\mathcal{LL}''(\widetilde{\theta})}{n}}.$$

We will look at the numerator and denominator separately. The denominator is:

$$-\frac{\mathcal{LL}''(\widetilde{\theta})}{n} = \frac{1}{n}\sum_{i=1}^{n} -\nabla_\theta^2 \log p(X_i; \widetilde{\theta}) \xrightarrow{p} \mathbb{E}_\theta[-\nabla_\theta^2 \log p(X; \widetilde{\theta})] \xrightarrow{p} \mathbb{E}_\theta[-\nabla_\theta^2 \log p(X; \theta)]$$

where the last step uses the fact that $\widetilde{\theta} \xrightarrow{p} \theta$.

The numerator is just the score function, i.e.

$$\begin{aligned}\frac{1}{\sqrt{n}}\mathcal{LL}'(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_\theta \log p(X_i; \theta) = \sqrt{n} \times \frac{1}{n}\sum_{i=1}^{n} [\nabla_\theta \log p(X_i; \theta) - \mathbb{E}[\nabla_\theta \log p(X; \theta)]] \\ \xrightarrow{d} N(0, \mathrm{Var}(\nabla_\theta \log p(X; \theta))) \xrightarrow{d} N(0, I(\theta)),\end{aligned}$$

where we used the facts that the score has mean 0, that the variance of the score is the Fisher information and that by the CLT $\sqrt{n}$ times an average of i.i.d. terms minus its expectation converges in distribution to a normal.

Putting the pieces together via Slutsky's theorem we obtain that,

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \frac{1}{I(\theta)} N(0, I(\theta)) \xrightarrow{d} N(0, 1/I(\theta)),$$

which is what we wanted to prove. ■

Example: Suppose that $X_1, \ldots, X_n \sim \mathrm{Exp}(\theta)$, then the log-likelihood,

$$\mathcal{LL}(\theta) = n \log \theta - \theta \sum_{i=1}^{n} X_i.$$

The score function:

$$s(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} X_i,$$

and the Fisher information,

$$I(\theta) = \frac{n}{\theta^2}.$$

The MLE is $\widehat{\theta} = \frac{1}{\overline{X}}$. So we can use the above result to conclude that,

$$\widehat{\theta} - \theta \xrightarrow{d} N\left(0, \frac{\theta^2}{n}\right).$$

19.4 Influence Functions and Regular Asymptotically Linear Estimators

We could have followed a similar proof as above to conclude that the MLE can be written as:

$$\widehat{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n} \frac{\nabla_\theta \log p(X_i;\theta)}{I(\theta)} + \text{Remainder},$$

where the remainder is small (roughly proportional to the previous term multiplied by $[I(\widetilde{\theta}) - I(\theta)] \to 0$).

The term,

$$\psi(x) = \frac{\nabla_\theta \log p(x;\theta)}{I(\theta)},$$

is called the *influence function*, i.e. as you can see above it measures the influence each single observation has on the estimator $\widehat{\theta}$, i.e.

$$\widehat{\theta} \approx \theta + \frac{1}{n}\sum_{i=1}^{n} \psi(X_i).$$

An estimator is often called *robust* if the function ψ is bounded, i.e. each observation can exert a limited influence on the estimator. Almost every estimator we have seen so far is non-robust in this sense.

For instance, for $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with σ known say, it is easy to check that the MLE satisfies,

$$\widehat{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta),$$

so that the influence of any point on the MLE is $X_i - \theta$ which is certainly unbounded. This means that if I corrupted a single point X_i then the MLE could be arbitrarily bad.

Thinking of a complex predictor like a deep neural network, one can try to obtain some information about the predictor by trying to compute the influence function of training images on the final predictor. A paper that did this (and quite a bit more) won ICML's best paper a few years ago.

Returning to the expression:

$$\widehat{\theta} \approx \theta + \frac{1}{n}\sum_{i=1}^{n} \psi(X_i).$$

Estimators that satisfy this type of expansion are called asymptotically linear estimators (many non-MLE estimators also satisfy expansions of this form). There is a classical result due to Le Cam that any sufficiently well-behaved (regular) estimator is asymptotically linear. It is not easy to prove (see Van Der Vaart's book). This together with the Cramér-Rao lower bound implies that the MLE is the "best regular asymptotically linear estimator".

19.5 Asymptotic Relative Efficiency

Once you restrict attention to asymptotically linear estimators, comparing estimators in terms of their MSE boils down to comparing their variances. Specifically, if

$$\begin{aligned} \sqrt{n}(W_n - \tau(\theta)) &\rightsquigarrow N(0, \sigma_W^2) \\ \sqrt{n}(V_n - \tau(\theta)) &\rightsquigarrow N(0, \sigma_V^2) \end{aligned}$$

then the *asymptotic relative efficiency (ARE)* is

$$\text{ARE}(V_n, W_n) = \frac{\sigma_W^2}{\sigma_V^2}.$$

Example 19.2 *Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. The mle of λ is $\overline{X}$. Let*

$$\tau = \mathbb{P}(X_i = 0).$$

So $\tau = e^{-\lambda}$. Define $Y_i = I(X_i = 0)$. This suggests the estimator

$$W_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Another estimator is the mle

$$V_n = e^{-\hat{\lambda}}.$$

The delta method gives

$$\text{Var}(V_n) \approx \frac{\lambda e^{-2\lambda}}{n}.$$

We have

$$\begin{aligned} \sqrt{n}(W_n - \tau) &\rightsquigarrow N(0, e^{-\lambda}(1 - e^{-\lambda})) \\ \sqrt{n}(V_n - \tau) &\rightsquigarrow N(0, \lambda e^{-2\lambda}). \end{aligned}$$

So

$$\text{ARE}(W_n, V_n) = \frac{\lambda}{e^{\lambda} - 1} \leq 1. \quad \square$$

Since the mle is efficient, we know that, in general, $\text{ARE}(W_n, \text{mle}) \leq 1$.

19.6 Multivariate Case

Now let $\theta = (\theta_1, \ldots, \theta_k)$. In this case we have

$$\sqrt{n}(\hat{\theta} - \theta) \rightsquigarrow N(0, I^{-1}(\theta))$$

where $I^{-1}(\theta)$ is the inverse of the Fisher information matrix. The approximate standard error of $\hat{\theta}_j$ is $\sqrt{I^{-1}_{jj}/n}$. If $\tau = g(\theta)$ with $g : \mathbb{R}^k \to \mathbb{R}$ then by the delta method,

$$\sqrt{n}(\hat{\tau} - \tau) \rightsquigarrow N(0, (g')^T I^{-1} g')$$

where g' is the gradient of g evaluated at θ.