lec05

February 3, 2026

1 Lecture 5: Asymptotic Normality and Efficiency of the MLE

Data 145, Spring 2026: Evidence and Uncertainty
Instructors: Ani Adhikari, William Fithian

As in lectures 3 and 4, **our focus is on single-parameter models** where the distribution is known up to one real number θ. We are using the method of maximum likelihood to estimate $\hat{\theta}$.

1.1 Road Map for the Lecture

1. Quickly recap some terminology and notation.
2. Derive an alternative way to calculate the Fisher information. (Provided as part of Lecture 4 notes)
3. Derive (apart from some technical issues) the asymptotic normality of the MLE and identify its variance. (Provided as Part of Lecture 4 notes)
4. Define the relative efficiency of two estimators.
5. Examine the efficiency of the MLE by applying the Cramér-Rao theorem.

1.1.1 1. Recap: Terminology and Notation

Given i.i.d. data $X_1, \dots, X_n$ from density f_θ, the **likelihood function** is:

$$\text{Lik}(\theta; X) = \prod_{i=1}^{n} f_\theta(X_i)$$

This is the joint density of the data, viewed as a function of θ (with data held fixed).

The **log-likelihood** is:

$$\ell_n(\theta; X) = \log \text{Lik}(\theta; X) = \sum_{i=1}^{n} \log f_\theta(X_i)$$

The subscript n reminds us that X is a sample of size n.

The **maximum likelihood estimator** (MLE) is:

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} \text{Lik}(\theta; X) = \arg\max_{\theta\in\Theta} \ell_n(\theta; X)$$

The **score function** (or just **score**) is the derivative of the log-likelihood with respect to θ:

$$S_n(\theta; X) = \ell_n'(\theta; X) = \frac{\partial}{\partial\theta}\ell_n(\theta; X)$$

The **Fisher information** in a single observation X_i is defined by:

$$I(\theta) = Var_\theta(\ell_1'(\theta; X_i)) = E_\theta(\ell_1'(\theta; X_i)^2)$$

The second equality uses $E_\theta(\ell_1'(\theta; X_i)) = 0$.

For n i.i.d. observations, the **total Fisher information** is $nI(\theta)$, since:

$$Var_\theta(S_n(\theta; X)) = n \cdot Var_\theta(\ell_1'(\theta; X_i)) = nI(\theta)$$

1.1.2 2. An Alternative Formula for the Fisher Information

There's another way to compute Fisher information that's often more convenient for calculations.

Assume that ℓ is twice differentiable.

Theorem: $I(\theta) = -E_\theta[\ell_1''(\theta; X_i)]$

Proof: We showed that $E_\theta[\ell_1'(\theta; X_i)] = 0$ for all θ. Differentiate both sides with respect to θ:

$$0 = \frac{d}{d\theta}E_\theta[\ell_1'(\theta; X_i)] = \frac{d}{d\theta}\int \ell_1'(\theta; x) f_\theta(x)\, dx$$

Switch the derivative and integral, and use the product rule of derivatives:

$$0 = \int \ell_1''(\theta; x) \cdot f_\theta(x)\, dx + \int \ell_1'(\theta; x) \cdot \frac{d}{d\theta}f_\theta(x)\, dx$$

The first integral is $E_\theta[\ell_1''(\theta; X_i)]$. For the second, note that $\frac{d}{d\theta}f_\theta(x) = \ell_1'(\theta; x) \cdot f_\theta(x)$, so:

$$\int \ell_1'(\theta; x) \cdot \frac{d}{d\theta}f_\theta(x)\, dx = \int \ell_1'(\theta; x)^2 f_\theta(x)\, dx = E_\theta[\ell_1'(\theta; X_i)^2] = I(\theta)$$

Therefore: $0 = E_\theta[\ell_1''(\theta; X_i)] + I(\theta)$, giving $I(\theta) = -E_\theta[\ell_1''(\theta; X_i)]$. □

Interpretation: The Fisher information equals the negative expected curvature of the log-likelihood. More curvature at the maximum means the MLE is more precisely determined.

Check the New Formula in the Normal Case If the sample is i.i.d. normal (μ, σ^2) for a known σ^2, we have seen that

$$\ell_1'(\mu; X_i) = \frac{X_i - \mu}{\sigma^2}$$

so

$$\ell_1''(\mu; X_i) = -\frac{1}{\sigma^2}$$

Note that **this is a constant** so its expectation is just itself, and it agrees with $Var_\mu(\ell_1'(\mu; X_i))$ calculated earlier.

Check the New Formula in the Exponential Case If the sample is i.i.d. exponential with rate λ, we have seen that

$$\ell_1'(\lambda; X_i) = \frac{1}{\lambda} - X_i$$

so

$$\ell_1''(\lambda; X_i) = -\frac{1}{\lambda^2}$$

Once again, it's a constant, so its expectation is just itself.

1.1.3 3.1. Towards Asymptotic Normality: Consistency

This was shown in Lecture 3 (apart from some care required to establish uniform convergence instead of pointwise convergence; but don't worry about that).

Let $\hat{\theta}_n$ be the MLE of the true parameter θ_0 based on $X_1, X_2, \dots, X_n$.

Then $\hat{\theta}_n$ is a consistent estimator of θ_0. That is, $\hat{\theta}_n \overset{P}{\to} \theta_0$.

Here the P in the $\overset{P}{\to}$ symbol is the true underlying probability distribution, that is, P_{θ_0}.

1.1.4 3.2. Towards Asymptotic Normality: Taylor Expansion

First, the statement of asymptotic normality of the MLE:

Theorem: Let θ_0 be the true value of θ. Under regularity conditions that we have assumed without stating, the MLE $\hat{\theta}_n$ is asymptotically normal:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\to} N\left(0, \frac{1}{I(\theta_0)}\right) \qquad \text{as } n \to \infty$$

Organizing a derivation: The MLE is obtained by setting the derivative of the log-likelihood to be 0. Since the derivative of the log-likelihood is the score function, we have $0 = S_n(\hat{\theta}_n; X)$.

Since $\hat{\theta}_n$ and the true θ_0 are likely to be close for large n, use a Taylor expansion of $S_n\hat{\theta}_n; X)$ about θ_0. For ease of notation, we will suppress the sample X from now on. But it's there, and it's the reason the equalities below are equalities of random variables.

$$0 \;=\; S_n(\hat{\theta}_n) \;=\; S_n(\theta_0) + (\hat{\theta}_n - \theta_0)S_n'(\tilde{\theta}_n)$$

for some point $\tilde{\theta}_n$ between $\hat{\theta}_n$ and θ_0.

Note that we're assuming S_n is differentiable.

Rewrite the above to see that

$$\hat{\theta}_n - \theta_0 \;=\; \frac{S_n(\theta_0)}{-S_n'(\tilde{\theta}_n)}$$

and hence

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \;=\; \frac{\frac{1}{\sqrt{n}}S_n(\theta_0)}{-\frac{1}{n}S_n'(\tilde{\theta}_n)}$$

1.1.5 3.3. Towards Asymptotic Normality: The Numerator

$$S_n(\theta_0) = \sum_{i=1}^{n} \ell_1'(\theta_0; X_i)$$

This is a sum of i.i.d. random variables with common mean 0 and common variance $I(\theta_0)$. Here the mean and variance are calculated using the true θ_0 as the parameter.

By the CLT,

$$\frac{S_n(\theta_0)}{\sqrt{nI(\theta_0)}} \xrightarrow{d} N(0,1)$$

and hence

$$\frac{S_n(\theta_0)}{\sqrt{n}} \xrightarrow{d} N(0, I(\theta_0))$$

1.1.6 3.4. Towards Asymptotic Normality: The Denominator

First note that for any θ,

$$\frac{1}{n}S_n'(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell^{"}_1(\theta)$$

This is the mean of an i.i.d. sample. By the Weak Law of Large Numbers,

$$\frac{1}{n}S_n'(\theta) \xrightarrow{P} E_\theta(\ell^{"}_1(\theta)) = -I(\theta)$$

By consistency of the MLE, $\hat{\theta}_n \xrightarrow{P} \theta_0$, where the probabilities are calculated using the true θ_0.

By the same "squeezing" argument as in the derivation of the delta method, $|\tilde{\theta}_n - \theta_0| \le |\hat{\theta}_n - \theta_0|$ and so $\tilde{\theta}_n \xrightarrow{P} \theta_0$.

We want to conclude that $\frac{1}{n}S_n'(\tilde{\theta}_n) \xrightarrow{P} -I(\theta_0)$ when the probabilities are calculated using the true θ_0. But we don't quite have that, and it takes some work and regularity conditions to prove.

It's fine to simply assume that we have enough regularity to make it work, and therefore the denominator converges in probability to the constant $-I(\theta_0)$.

In fact, in all our examples we've seen that $\ell_1''(\theta; X_i)$ is a constant (that is, a non-random quantity) involving θ. Thich implies $\frac{1}{n}S_n'(\tilde{\theta}_n)$ is that same quantity for every n. So the "convergence", if we still want to call it that, is automatically to that quantity.

1.1.7 3.5. Asymptotic Normality

Now use Slutsky's theorem to see that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to a normal $(0, I(\theta_0)$ random variable times the constant $-1/I(\theta_0)$. That constant is squared in the calculation of the variance, so

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right)$$

as we had claimed.

From a practical perspective, the result says that for large n, the distribution of $\hat{\theta}_n$ is approximately normal with mean θ_0 and variance $1/nI(\theta_0)$.

So if the sample is large, the MLE is almost unbiased, it has a variance we can estimate, and its distribution is approximately normal This allows us to construct confidence intervals for θ_0, as you have seen in exercises.

That's a great property of the MLE. But could we do better? Let's look at a way to compare to estimators of the same parameter.

1.1.8 4. Relative Efficiency, and Efficiency

One measure of how accurate an estimate is relative to another is *efficiency*.

Suppose we are estimating a parameter θ. The **relative efficiency** of two unbiased estimators T_1 and T_2 is defined as $\dfrac{Var_\theta(T_1)}{Var_\theta(T_2)}$.

T_1 is considered a better estimator than T_2 if this ratio is less than 1.

It's important that we have assumed both estimators to be unbiased. If they had different biases, then we'd have to deal with that before comparing variances.

Let T be an unbiased estimator of θ based on an i.i.d. sample $X_1, X_2, \dots, X_n$. We will define the **efficiency** of T as $\dfrac{1/nI(\theta)}{Var_\theta(T)}$, and we will call T **efficient** if its efficiency is 1.

The Cramér-Rao theorem will help us see what this has to do with the MLE.

1.1.9 5. Efficiency and the Cramér-Rao Bound

Cramér-Rao Theorem For any unbiased estimator T of θ based on an i.i.d. sample of size n,

$$Var_\theta(T) \ \geq \ \frac{1}{nI(\theta)}$$

This allows us to say that MLE $\hat{\theta}_n$ is asymptotically efficient. We know that for large n it is almost unbiased and its variance is close to this lower bound. In other words, if we can calculate an MLE, it will be pretty much the best estimator we could get.

But keep in mind that the Cramér-Rao theorem is not an asymptotic result. It is true for all n.

Proof: For the score function $S_n(\theta; X)$, we know that $E_\theta(S_n(\theta; X)) = 0$. We also know that $Var(S_n(\theta; X)) = nI(\theta)$.

Let $\rho(T, S_n)$ be the correlation between T and $S_n(\theta; X)$. Then $\rho(T, S_n) \leq 1$. Thus (skipping the subscript θ in all expectations and variances), we have

$$\frac{Cov(T, S_n)}{SD(T)SD(S_n)} \leq 1$$

which is equivalent to

$$Var(T) \geq \frac{Cov^2(T, S_n)}{Var(S_n)}$$

So

$$Var(T) \geq \frac{Cov^2(T, S_n)}{nI(\theta)}$$

The Cramér-Rao theorem would be true if $Cov(T, S_n) = 1$. Let's see if we can show this.

First note that since $E(S_n) = 0$,

$$Cov(T, S_n) = 1 \iff E(TS_n) = 1$$

We'll have to find that expected product. What we have going for us is that T is unbiased. So let's use that.

To simplify notation, let $\underset{\sim}{x} = x_1, x_2, \ldots, x_n$ and $d\underset{\sim}{x}$ denote $dx_1 dx_2 \ldots dx_n$.

Since T is unbiased,

$$\theta = E(T) = \int_{\underset{\sim}{x}} T(\underset{\sim}{x}) \left[\Pi_{i=1}^n f_\theta(x_i)\right] d\underset{\sim}{x}$$

Differentiate both sides with respect to θ:

$$1 = \frac{d}{d\theta} \int_{\underset{\sim}{x}} T(\underset{\sim}{x}) \left[\Pi_{i=1}^n f_\theta(x_i)\right] d\underset{\sim}{x}$$

Under regularity conditions that we assume are met, we can switch the derivative and integal on the right side to get

$$\begin{aligned} 1 &= \int_{\underset{\sim}{x}} \frac{d}{d\theta} T(\underset{\sim}{x}) \left[\Pi_{i=1}^n f_\theta(x_i)\right] d\underset{\sim}{x} \\ &= \int_{\underset{\sim}{x}} T(\underset{\sim}{x}) \frac{d}{d\theta} \left[\Pi_{i=1}^n f_\theta(x_i)\right] d\underset{\sim}{x} \end{aligned}$$

Recall that $\ell_1(\theta; x_k) = \log(f_\theta(x_k))$ and therefore

$$\ell_1'(\theta; x_k) = \frac{f_\theta'(x_k)}{f_\theta(x_k)}$$

By the product rule for derivatives, $\frac{d}{d\theta}\left[\Pi_{i=1}^n f_\theta(x_i)\right]$ is the sum of n terms, of which term k is equal to

$$\begin{aligned} f_\theta'(x_k)\left[\Pi_{i \neq k} f_\theta(x_i)\right] &= \ell_1'(\theta; x_k) f_\theta(x_k)\left[\Pi_{i \neq k} f_\theta(x_i)\right] \\ &= \ell_1'(\theta; x_k)\left[\Pi_{i=1}^n f_\theta(x_i)\right] \end{aligned}$$

Plug this into our earlier equation to get

$$1 \;=\; \int_{\underset{\sim}{x}} T(\underset{\sim}{x}) \sum_{i=1}^n \ell_1'(\theta; x_i)\left[\Pi_{i=1}^n f_\theta(x_i)\right] d\underset{\sim}{x} = \int_{\underset{\sim}{x}} T(\underset{\sim}{x}) S_n(\theta; \underset{\sim}{x})\left[\Pi_{i=1}^n f_\theta(x_i)\right] d\underset{\sim}{x} = E(T \cdot S_n)$$

just as we had hoped. This establishes the Cramér-Rao bound.

Example: Efficiency and the Bernoulli MLE Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli (p). In a probability class, you found the sample proportion $\hat{p}_n$ to be the MLE of p.

From Lecture 4, the Fisher information is $I(p) = \dfrac{1}{p(1-p)}$.

Fix n. Since $\hat{p}_n$ is an i.i.d. sample proportion, we know that $E_p(\hat{p}_n) = p$ and $Var_p(\hat{p}_n) = \dfrac{p(1-p)}{n} = \dfrac{1}{nI(p)}$.

So $\hat{p}_n$ is unbiased and efficient for all n.

```
[ ]:
```