lec04

January 29, 2026

1 Lecture 4: Asymptotic Normality of the MLE

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

As in Lecture 3, **our focus is on single-parameter models** where the distribution is known up to one real number θ.

The goal is to estimate θ by the method of maximum likelihood, and to examine the properties of the estimator.

1.1 Road Map for the Lecture

1. Recall what we know about MLEs in particular models, and notice a common theme.
2. Describe the common theme as a property of maximum likelihood estimation instead of something that has to be derived separately for each model.
3. Define some useful quantities and check that they behave the way we expect them to in known cases.
4. Derive the main properties – consistency, asymptotic normality – of the MLE.

1.1.1 1. A Recurring Theme

Let the sample $X_1, X_2, \dots, X_n$ be i.i.d. with the distribution in the first column below. Let $\bar{X}_n$ be the sample mean. In Lecture 1 and your probabiilty class, you established the asymptotic normality of the MLE of a parameter or parameters in each distribution, but your reasoning varied depending on the model.

Distribution	Parameter	MLE	Reason for Asymptotic Normality of the MLE
Exponential	Rate λ	$1/\bar{X}_n$	Delta method, starting with the CLT applied to $\bar{X}_n$
Bernoulli	Success probability p	$\bar{X}_n$	CLT
Poisson	Mean μ	$\bar{X}_n$	CLT

Distribution	Parameter	MLE	Reason for Asymptotic Normality of the MLE
Normal (unknown μ, σ)	Mean μ	$\bar{X}_n$	Normal for all n
Normal (unknown μ, σ)	Variance σ^2	$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$	$\frac{n}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-1}$, roughly normal for large n

When there's a recurring result like the asymptotic normality we discovered in all these cases, it's worth trying to see if there's a more general reason for it than the argument we've given in each particular case.

1.1.2 2. Describing the General Result

Let $X_1, X_2, ..., $ be i.i.d. with a "nice" distribution that has a parameter θ whose unknown true value is θ_0. Later we'll state a couple of ways in which distributions are "nice". These are regularity conditions under which we can prove our results. All the distributions in the table above are nice.

The main result is that the distribution of an MLE is asymptotically normal. We don't have to come up with a different argument for asymptotic normality for each underlying distribution.

We will prove (or almost-prove) the result in this lecture and Lecture 5. For now we'll just state it and see how it can be used.

Asymptotic Normality of the MLE Let $\hat{\theta}_n$ be the MLE of θ_0 based on $X_1, X_2, \ldots, X_n$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2)$$

We'll say more about σ^2 shortly. For now, remember that using the result above means understanding it as follows:

$\hat{\theta}_n$ is approximately normal $(\theta_0, \sigma^2/n)$ for large n.

This says that the MLE based on a large sample is approximately normal, centered at the true value of the parameter, with an SD decreasing like $1/\sqrt{n}$.

Confidence Interval for θ_0 By the normal approximation, for large n we have

$$P_{\theta_0}(|\hat{\theta}_n - \theta_0| < 2\frac{\sigma}{\sqrt{n}}) \approx 0.95$$

This allows us to build confidence intervals for θ_0 in the same way you made confidence intervals for a population mean: Start at the estimate and go 2 SDs on either side.

Thus the interval $\hat{\theta}_n \pm 2\frac{\sigma}{\sqrt{n}}$ is an approximate 95% confidence interval for θ_0.

Also as before, if you want a confidence level different from 95%, you can replace the factor of 2 by the appropriate value of z.

But What is σ^2? That's the burning question. Here is a preview of what we will discover about σ^2 in this lecture and the next. Then we'll roll up our sleeves and start doing the details.

- σ^2 is a positive number that depends on θ_0.
- $\sigma^2 = \dfrac{1}{I(\theta_0)}$ where I is a function that we will define. $I(\theta)$ will be called the Fisher information of a single observation, evaluated at θ. We'll discuss the reason for the name.

Since $I(\theta_0)$ depends on θ_0, we won't be able to calculate it exactly. But, in a move that should feel familiar, we will replace it by $I(\hat{\theta}_n)$ which we can compute based on the sample. The asymptotic normality will still hold, just as it did in when you replaced the population SD by the sample SD in confidence intervals for the population mean (see Problem 5 of Worksheet 1).

We will now define some quantities we'll need to derive asymptotic normality. The starting point is the familiar pair of likelihood and log-likelihood.

1.1.3 3.1. Deriving Asymptotic Normality: Terminology and Notation Recap

Likelihood 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).

Log-Likelihood 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.

Maximum Likelihood Estimator The **maximum likelihood estimator** (MLE) is:

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

1.1.4 3.2. The Score Function

Motivation The MLE $\hat{\theta}$ is (typically) found by setting the derivative of the log-likelihood to zero:

$$\frac{d}{d\theta} \ell_n(\theta; X) \bigg|_{\theta = \hat{\theta}_{MLE}} = 0$$

To understand the MLE more deeply — especially its sampling distribution — we need to study the derivative of the log-likelihood.

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

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

For an i.i.d. model, let $\ell_1(\theta; X_i) = \log f_\theta(X_i)$ denote the log-likelihood contribution from the single observation X_i. Then:

$$\ell_n(\theta; X) = \sum_{i=1}^{n} \ell_1(\theta; X_i)$$

and the score decomposes as:

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

The score is a sum of i.i.d. terms! The CLT will enter the picture at some point. So we'll need the mean and variance:

$$E_\theta(S_n(\theta; X)) = nE_\theta(\ell_1'(\theta; X_1))$$

$$Var_\theta(S_n(\theta; X)) = nVar_\theta(\ell_1'(\theta; X_1))$$

Important: We're differentiating with respect to θ, not with respect to X_i. Even if X_i is discrete, the score is well-defined as long as $f_\theta(x)$ is differentiable in θ.

Example 1: Exponential For **one observation** $X_i \sim$ Exponential(λ): $f_\lambda(x) = \lambda e^{-\lambda x}$

$$\ell_1(\lambda; X_i) = \log\lambda - \lambda X_i$$

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

Confirm the MLE:

$$S_n(\lambda; X) = \ell_n'(\lambda; X) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = \frac{n}{\lambda} - n\bar{X}_n$$

Setting $S_n(\lambda; X) = 0$ gives $\hat{\lambda} = 1/\bar{X}_n$, confirming our MLE.

Mean of the Score

$$E_\lambda(\ell_1'(\lambda; X_i)) = E_\lambda(\frac{1}{\lambda} - X_i) = \frac{1}{\lambda} - \frac{1}{\lambda} = 0$$

So $E(S_n(\lambda; X)) = 0$.

Example 2: Gaussian (known variance) For $X_i \sim N(\mu, \sigma^2)$ with σ^2 known: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

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

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

Once again, the mean of the score is 0: $E_\mu(S_n(\mu; X)) = 0$.

Example 3: Bernoulli For $X_i \sim \text{Bernoulli}(p)$: $f_p(x) = p^x(1-p)^{1-x}$ for $x \in \{0, 1\}$

$$\ell_1(p; X_i) = X_i \log p + (1 - X_i)\log(1-p)$$

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

Yet again, the mean of the score is 0: $E_p(S_n(p; X)) = 0$.

Yes, there's a pattern here, and we can generalize.

1.1.5 3.3. Mean of the Score

A fundamental fact: **the score has mean zero at the value of the parameter.** That is,

Theorem: $E_\theta[\ell_1'(\theta; X_i)] = 0$

For this result to hold, the θ with respect to which the expectation is taken must be the same as the θ at which ℓ_1 is evaluated.

Proof: We can write:

$$\ell_1'(\theta; X_i) = \frac{\partial}{\partial\theta}\log f_\theta(X_i) = \frac{\frac{\partial}{\partial\theta} f_\theta(X_i)}{f_\theta(X_i)}$$

Taking the expectation (integrating over the θ-distribution of X_i):

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

Now we interchange the derivative (with respect to θ) and the integral (with respect to x):

$$\int \frac{\partial}{\partial\theta} f_\theta(x)\, dx = \frac{\partial}{\partial\theta}\int f_\theta(x)\, dx = \frac{\partial}{\partial\theta}(1) = 0 \quad \square$$

(The interchange is valid under regularity conditions that hold for "nice" models.)

1.1.6 3.4. The Fisher Information

Since the score has mean zero, its variance is particularly important.

Definition 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:

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

Why "Information"? Intuitively, the Fisher information measures **how much the data tells us about** θ: - High $I(\theta)$ means ℓ_1' varies a lot $\to$ small changes in θ produce large changes in the likelihood $\to$ data is informative about θ - Low $I(\theta)$ means the likelihood is flat $\to$ data doesn't distinguish well between different θ values

We'll see this more precisely next lecture: the variance of the MLE is approximately $1/(nI(\theta))$.

Fisher Information Examples **Exponential:** $\ell_1'(\lambda; X_i) = 1/\lambda - X_i$, where $E_\lambda(X_i) = 1/\lambda$, $\mathrm{Var}_\lambda(X_i) = 1/\lambda^2$

$$I(\lambda) = Var_\lambda\left(\frac{1}{\lambda} - X_i\right) = Var_\lambda(X_i) = \frac{1}{\lambda^2}$$

Gaussian: $\ell_1'(\mu; X_i) = (X_i - \mu)/\sigma^2$

$$I(\mu) = Var_\mu\left(\frac{X_i - \mu}{\sigma^2}\right) = \frac{Var_\mu(X_i)}{\sigma^4} = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}$$

Bernoulli: $\ell_1'(p; X_i) = (X_i - p)/(p(1-p))$

$$I(p) = Var_p\left(\frac{X_i - p}{p(1-p)}\right) = \frac{Var_p(X_i)}{[p(1-p)]^2} = \frac{p(1-p)}{[p(1-p)]^2} = \frac{1}{p(1-p)}$$

Agrees with Our Claim about Asymptotic Normality In the last two cases, $I(\theta) = 1/Var_\theta(X_i)$. This is not a coincidence — it holds for nice one-parameter families called "exponential families"!

But for our purposes, the crucial pattern is that these values make sense in the main result we stated earlier (and have yet to prove):

The MLE $\hat{\theta}_n$ is approximately normal $(\theta_0, \sigma^2/n)$ for large n. Here $\sigma^2 = 1/I(\theta_0)$.

Let's see if this checks out for the Gaussian and the Bernoulli. The exponential is for you to check in Worksheet 2.

Asymptotic normality, Gaussian Case: This one is true without the word "asymptotic". The true value of the parameter is μ_0 and the MLE is $\bar{X}_n$. You know from your probability class that this MLE is normal for all n. For each n it has mean μ_0 and variance $\sigma^2/n = \dfrac{1}{n(1/\sigma^2)} = \dfrac{1}{nI(\mu_0)}$.

Asymptotic normality, Bernoulli Case: The true value of the parameter is p_0 and the MLE is the sample proportion $\bar{X}_n$. By the CLT, this is asymptotically normal. It is also unbiased, so its mean is p_0. Its variance is $\dfrac{p_0(1-p_0)}{n} = \dfrac{1}{n(1/p_0(1-p_0))} = \dfrac{1}{nI(p_0)}$.

1.2 In class, Lecture 4 stopped here. Lecture 5 will start with 3.5 below.

But at this point, the content below should be a relatively easy read. Try it!

1.2.1 3.5. 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:

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

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

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

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

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

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

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.

1.2.2 4.1. Towards Asymptotic Normality: Consistency

This was proved 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 $X_1, X_2, \ldots, X_n$.

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

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

1.2.3 4.2. Towards Asymptotic Normality: Taylor Expansion

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.

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

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

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

1.2.4 4.3. 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

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

1.2.5 4.4. 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, $|\tilde{\theta}_n - \theta_0| \leq |\hat{\theta}_n - \theta_0|$ and so $\tilde{\theta}_n \xrightarrow{P} \theta_0$.

We want to conclude that $\frac{1}{n}S_n'(\tilde{\theta}) \xrightarrow{P} -I(\theta_0)$ 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 evaluated at θ_0.

1.2.6 4.5. 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

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

as we had claimed.

1.2.7 Next Lecture

- Complete the argument that the MLE is asymptotically normal with variance $1/(nI(\theta))$
- What does it mean to be "efficient"?
- Show that the MLE is pretty much the best thing you can do, in many situations.

```
[ ]:
```