Lecture 22 – MLE and Wald test

Last Modified: November 12, 2026

Outline:

22.1 Asymptotic Properties of MLE 1
22.2 Application to hypothesis testing 2

22.1 Asymptotic Properties of MLE

How to find an asymptotically efficient estimator?

Recall that if $X_i \overset{\text{iid}}{\sim} f_\theta(\theta)$ for a regular statistical model, then the MLE $\hat{\theta}_n$ for θ can be expressed as the solution to

$$\frac{\partial}{\partial\theta}\ell_{\underline{x}}(\theta) = 0,$$

where $\ell_{\underline{x}}(\theta) = \sum_{i=1}^n \log f_\theta(x_i)$ is the log-likelihood function.

Theorem 1 (Consistency)**.** Under certain regularity conditions (the MLE must be unique, f_θ is differentiable, etc.), the MLE $\hat{\theta}_n$ satisfies

$$\hat{\theta}_n \xrightarrow{p} \theta$$

This means the MLE is consistent.

Theorem 2 (Asymptotic Normality)**.** Under further regularity conditions, the MLE is asymptotically normal with variance given by the Cramér–Rao Lower Bound, i.e.,

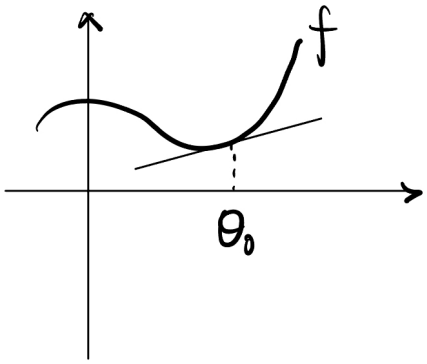
$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{I_X(\theta)}\Big)$$

This means that the MLE is *asymptotically efficient*.

Proof sketch. Using the Taylor series expansion, we can approximate a smooth function $f(\theta)$ about a point θ_0 using with polynomials:

$$f(\theta) = f(\theta_0) + (\theta - \theta_0)f'(\theta_0) + \frac{1}{2}(\theta - \theta_0)^2 f''(\theta_0) + \cdots$$

which is a local approximation close to θ_0.



Applying the expansion to the score function $\ell'_{\underline{x}}(\theta) = \frac{\partial}{\partial\theta}\ell_{\underline{x}}(\theta)$ and then evaluating at $\hat{\theta}_n$ yields

$$\underbrace{\ell'_{\underline{x}}(\hat{\theta}_n)}_{0} = \ell'_{\underline{x}}(\theta) + (\hat{\theta}_n - \theta)\ell''_{\underline{x}}(\theta) + \cdots$$

where the left-hand side is 0 by the definition of the MLE.
Dropping the higher order terms in the expansion and solving $\hat{\theta} - \theta$ leads to

$$\hat{\theta}_n - \theta \approx -\frac{\ell'_{\underline{x}}(\theta)}{\ell''_{\underline{x}}(\theta)}$$

Scaling both side by $\sqrt{n}$ then gives

$$\sqrt{n}(\hat{\theta}_n - \theta) \approx \frac{\frac{1}{\sqrt{n}}\ell'_{\underline{x}}(\theta)}{-\frac{1}{n}\ell''_{\underline{x}}(\theta)}$$

The numerator on the right-hand side satisfies

$$\begin{aligned}\frac{1}{\sqrt{n}}\ell'_{\underline{X}}(\theta) &= \sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f_\theta(X_i)\\ &\xrightarrow{d} N(0, I_X(\theta))\end{aligned}$$

where the last step follows from the CLT with the fact that $\frac{\partial}{\partial\theta}\log f_\theta(X_i)$ are iid with

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f_\theta(X_i)\right] = 0, \qquad \mathsf{Var}\left(\frac{\partial}{\partial\theta}\log f_\theta(X_i)\right) = I_X(\theta)$$

Meanwhile, by the LLN,

$$-\frac{1}{n}\ell''_{\underline{X}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} -\frac{\partial^2}{\partial\theta^2}\log f_\theta(X_i) \xrightarrow{p} \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right] = I_X(\theta)$$

Combining the two limits and using Slutsky, we have

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{\frac{1}{\sqrt{n}}\ell'_{\underline{X}}(\theta)}{-\frac{1}{n}\ell''_{\underline{X}}(\theta)} \xrightarrow{d} N\left(0, \frac{1}{I_X(\theta)}\right)$$

□

Corollary 1. Let $\hat{\theta}_n$ be the MLE for θ and let g be a differentiable function. Under regularity conditions,

$$g(\hat{\theta}_n) \xrightarrow{p} g(\theta), \quad \text{and} \quad \sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}\left(0, \frac{[g'(\theta)]^2}{I_X(\theta)}\right)$$

Thus, $g(\hat{\theta}_n)$ is also asymptotically efficient for $g(\theta)$.

22.2 Application to hypothesis testing

Knowing the approximate distribution of an asymptotically normal estimator lets us design hypothesis tests at given level α.

Standard error

The **standard error** of an estimator is its standard deviation:

$$\mathrm{se}(\hat{\theta}_n) = \sqrt{\mathrm{Var}_\theta(\hat{\theta}_n)}.$$

It measures how much the estimator varies from sample to sample, in the same units as the parameter being estimated.

Suppose an estimator is asymptotically normal:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, v), \qquad v > 0.$$

For large n, we use the approximation $\hat{\theta}_n \approx \mathcal{N}(\theta, v/n)$. The standard deviation in this normal approximation, $\sqrt{v/n}$, is called the **asymptotic standard error**. If $\hat{v}_n$ is a consistent estimator of v, the **estimated standard error** is

$$\widehat{\mathrm{se}}(\hat{\theta}_n) = \sqrt{\frac{\hat{v}_n}{n}}.$$

For example, for a regular MLE, $v = 1/I(\theta)$, and we can use

$$\widehat{\mathrm{se}}(\hat{\theta}_n) = \frac{1}{\sqrt{nI(\hat{\theta}_n)}}.$$

Slutsky's theorem then gives

$$\frac{\hat{\theta}_n - \theta}{\widehat{\mathrm{se}}(\hat{\theta}_n)} \xrightarrow{d} \mathcal{N}(0, 1).$$

22.2.1 Two-sided Wald test

Consider the hypotheses

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0.$$

The **Wald statistic** measures the distance between the estimate and the null value in estimated standard errors:

$$T_n = \frac{\hat{\theta}_n - \theta_0}{\widehat{\mathrm{se}}(\hat{\theta}_n)}.$$

Under H_0, $T_n \approx \mathcal{N}(0, 1)$. The **two-sided Wald test** therefore rejects when

$$|T_n| = \frac{|\hat{\theta}_n - \theta_0|}{\widehat{\mathrm{se}}(\hat{\theta}_n)} \geq z_{\alpha/2},$$

which gives asymptotic level α. For a regular MLE, the statistic becomes

$$T_n = \sqrt{nI(\hat{\theta}_n)}(\hat{\theta}_n - \theta_0).$$

For this simple null, we may also use $1/\sqrt{nI(\theta_0)}$ as the standard error, since $\theta = \theta_0$ under H_0.

22.2.2 One-sided Wald test

Now consider

$$H_0 : \theta \le \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0.$$

The **one-sided asymptotic level-α Wald test** rejects when

$$T_n = \frac{\hat{\theta}_n - \theta_0}{\widehat{\text{se}}(\hat{\theta}_n)} > z_\alpha.$$

Let's verify the level heuristically. Write $s_n = \widehat{\text{se}}(\hat{\theta}_n) > 0$. When the true value is $\theta \le \theta_0$, we have

$$T_n = \frac{\hat{\theta}_n - \theta}{s_n} + \frac{\theta - \theta_0}{s_n} \le \frac{\hat{\theta}_n - \theta}{s_n} \approx \mathcal{N}(0, 1).$$

The extra term is nonpositive, making rejection less likely. Therefore,

$$\mathbb{P}_\theta(T_n \ge z_\alpha) \le \mathbb{P}_\theta\left(\frac{\hat{\theta}_n - \theta}{s_n} \ge z_\alpha\right) \approx \alpha.$$

For a regular MLE, the rule is

$$\sqrt{nI(\hat{\theta}_n)}(\hat{\theta}_n - \theta_0) > z_\alpha.$$

22.2.3 Worked example: testing an exponential rate

Suppose $X_1, \ldots, X_n$ are independent waiting times with density

$$f_\theta(x) = \theta e^{-\theta x}, \qquad x > 0, \quad \theta > 0.$$

Here θ is the *rate*, so $\mathbb{E}_\theta[X_i] = 1/\theta$: a larger rate means a shorter average waiting time. We want to test

$$H_0 : \theta \le \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0$$

using a one-sided Wald test with asymptotic level α.

Step 1: Find the MLE. The log-likelihood and its derivative are

$$\ell_{\underline{x}}(\theta) = n \log \theta - \theta \sum_{i=1}^n x_i, \qquad \ell'_{\underline{x}}(\theta) = \frac{n}{\theta} - \sum_{i=1}^n x_i.$$

Setting the derivative equal to zero gives $\hat{\theta}_n = 1/\overline{X}_n$. Since $\ell''_{\underline{x}}(\theta) = -n/\theta^2 < 0$, this is the maximizer.

Step 2: Find the Fisher information and estimated standard error. For one observation, $\log f_\theta(x) = \log \theta - \theta x$, so

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X)\right] = \frac{1}{\theta^2}.$$

Thus $\hat{\theta}_n \approx \mathcal{N}(\theta, \theta^2/n)$, and we estimate its standard error by

$$\widehat{\text{se}}(\hat{\theta}_n) = \frac{1}{\sqrt{nI(\hat{\theta}_n)}} = \frac{\hat{\theta}_n}{\sqrt{n}}.$$

Step 3: Write the rejection rule. The Wald statistic simplifies to

$$T_n = \frac{\hat{\theta}_n - \theta_0}{\hat{\theta}_n/\sqrt{n}} = \sqrt{n}\left(1 - \frac{\theta_0}{\hat{\theta}_n}\right) = \sqrt{n}(1 - \theta_0 \overline{X}_n).$$

We reject when $T_n > z_\alpha$, or equivalently when

$$\overline{X}_n < \frac{1 - z_\alpha/\sqrt{n}}{\theta_0}.$$

This makes sense: small observed waiting times provide evidence for a larger rate.

Step 4: Apply the test to data. Suppose $n = 100$, $\overline{X}_n = 0.80$ minutes, and we test $\theta \leq 1$ against $\theta > 1$ at asymptotic level $\alpha = 0.05$. Then

$$\hat{\theta}_n = \frac{1}{0.80} = 1.25, \qquad \widehat{\mathrm{se}}(\hat{\theta}_n) = \frac{1.25}{10} = 0.125, \qquad T_n^{\mathrm{obs}} = \frac{1.25 - 1}{0.125} = 2.$$

Since $2 > z_{0.05} \approx 1.645$, we reject H_0. There is evidence that the rate exceeds one per minute, or equivalently that the mean waiting time is less than one minute. The approximate one-sided p-value is

$$p \approx P(Z \geq 2) = 1 - \Phi(2) \approx 0.0228, \qquad Z \sim \mathcal{N}(0, 1),$$

where Φ is the standard normal CDF. These cutoffs and the p-value use a large-sample normal approximation, rather than the exact sampling distribution.