
# Asymptotic Distributions of MLEs

Sepehr Akbari

*4/3/2026*

Department of Math & CS, Lake Forest College

# Motivation

Let $X_1, \ldots, X_n \sim \text{Pois}(\lambda)$ represent number of hourly arrivals at a store, where $\lambda$ is the unknown average arrival rate. The Maximum Likelihood Estimator (MLE) for $\lambda$ is

$$\hat{\lambda} = \bar{X}$$

Let $\lambda = 8$. Consider we observe $n = 5$ hours of data. We can simulate this in R (`seed=450`) to compute $\hat{\lambda}$:

$$X = (11, 9, 7, 7, 8) \quad \Rightarrow \quad \hat{\lambda} = 8.4$$

Notice that if we have observed a different 5 hours, we would get a different sample and thus a different $\hat{\lambda}$. Therefore, $\hat{\lambda}$ is a *random variable (RV)*.

# Motivation

How "good" is our point estimate $\hat{\lambda} = 8.4$? Is $\lambda$ likely between $8.3$ and $8.5$? Or between $5.0$ and $11.0$?

To answer this, we need to understand the distribution of $\hat{\lambda}$.

For a limited $n$, it's difficult to derive the exact pdf/pmf of $\hat{\lambda}$. Instead, we can study the *asymptotic distribution* of $\hat{\lambda}$, which describes how $\hat{\lambda}$ behaves as we collect more data ($n \to \infty$).

This allows us to perform inference and construct reliable confidence intervals (CIs) for $\lambda$, without needing the exact finite-sample distribution of $\hat{\lambda}$.

So... what is the asymptotic distribution of $\hat{\lambda}$?

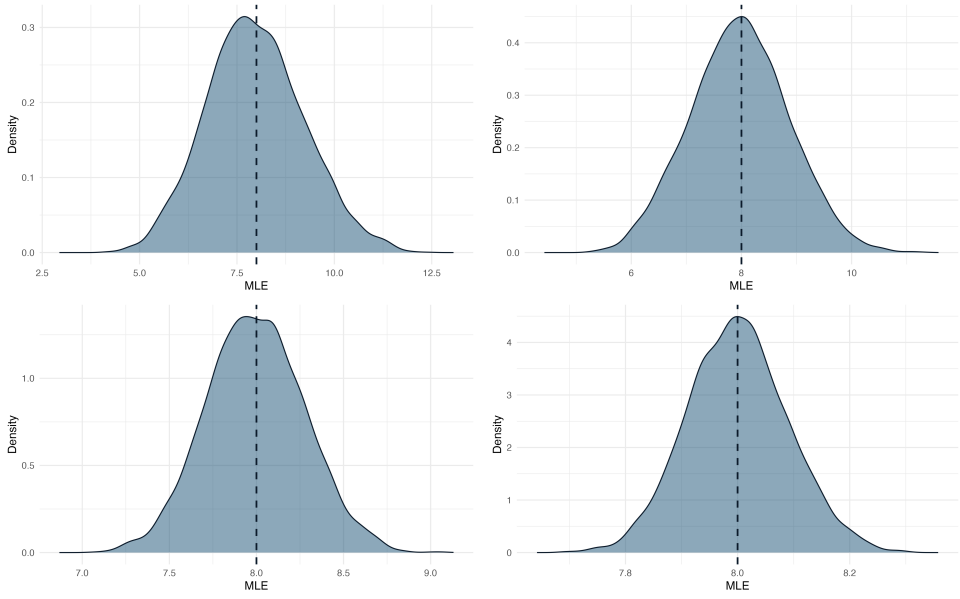


**Figure 1:** PDF of $\hat{\lambda}$, for $n = (5, 10, 100, 1000)$ simulated hours.

# Asymptotic Normality

Notice that

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \text{where } X_i \sim \text{Pois}(\lambda)$$

By the Central Limit Theorem (CLT), the mean of a sufficiently large number of i.i.d. RVs always converges to a Normal distribution. As $n \to \infty$:

$$\mathbb{E}(\hat{\lambda}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i) \qquad = \lambda$$

$$\text{Var}(\hat{\lambda}) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(X_i) = \frac{\lambda}{n}$$

Therefore, asymptotically; $\hat{\lambda} \sim \mathcal{N}\left(\lambda, \frac{\lambda}{n}\right)$.

# Asymptotic Normality

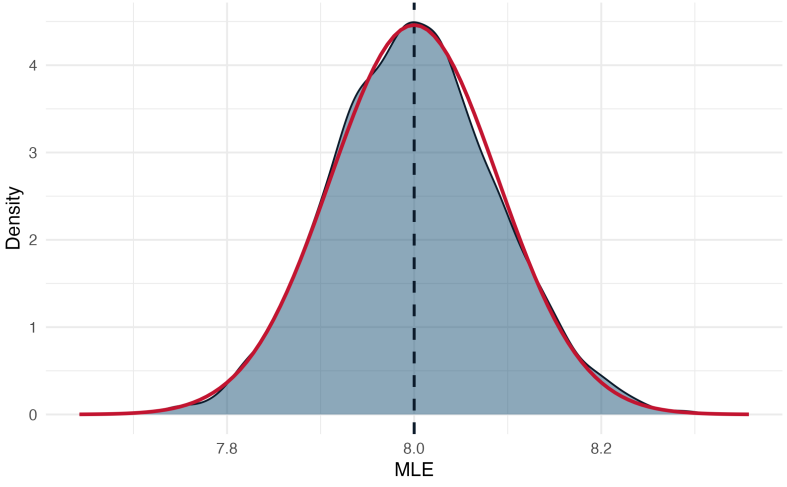


**Figure 2:** PDF of $\hat{\lambda}$ for $n = 1000$ simulated hours. In red: $\mathcal{N}\left(\lambda = 8, \frac{\lambda}{n} = 0.008\right)$ density.

# Score Function

Every MLE is asymptotically Normal. What if the MLE is not as straightforward as $\bar{X}$?

Recall the log-likelihood function of some unknown parameter $\theta$ is

$$\log L(\theta) = \sum_{i=1}^{n} \log f(X_i; \theta)$$

which we take the first derivative of with respect to $\theta$ (the *Score function*):

$$S(\theta) = \frac{\partial[\log L(\theta)]}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial[\log f(X_i; \theta)]}{\partial \theta}$$

Since $S(\theta)$ is a sum of i.i.d. RVs, by the CLT, $S(\theta)$ is asymptotically Normal.

# Curvature

Observe that

$$S(\hat{\theta}) = \left.\frac{\partial[\log L(\theta)]}{\partial\theta}\right|_{\theta=\hat{\theta}} = 0$$

We can approximate $S(\hat{\theta})$ using a first-order Taylor expansion around the true parameter $\theta$:

$$0 = S(\hat{\theta}) \approx \frac{\partial[\log L(\theta)]}{\partial\theta} + (\hat{\theta} - \theta)\frac{\partial^2[\log L(\theta)]}{\partial\theta^2}$$

We call the second derivative the *curvature* of the $\log L(\theta)$ at $\theta$.

We can solve for the estimation error $(\hat{\theta} - \theta)$, and determine the distribution of $\hat{\theta}$.

## Estimation Error

Solving for the estimation error, we get:

$$(\hat{\theta} - \theta) \approx -\frac{\partial[\log L(\theta)]/\partial\theta}{\partial^2[\log L(\theta)]/\partial\theta^2} = -\frac{S(\theta)}{\text{curvature}}$$

As $n \to \infty$:

- *Numerator.* $S(\theta)$ is asymptotically Normal (by the CLT).
- *Denominator.* The curvature is a sum of random variables. By the Law of Large Numbers, it converges to its expected value (a constant).

Therefore, $\hat{\theta}$ is asymptotically Normal.

We know the distribution of $\hat{\theta}$. What are its mean and variance?

# Asymptotic Bias

Since we have a valid pdf, we have:

$$\int_{-\infty}^{\infty} f(x;\theta)\ dx = 1 \quad \Rightarrow \quad \int_{-\infty}^{\infty} \frac{\partial f(x;\theta)}{\partial \theta}\ dx = 0$$

Using the identity $\frac{\partial[\log f(x;\theta)]}{\partial \theta} = \frac{1}{f(x;\theta)} \cdot \frac{\partial f(x;\theta)}{\partial \theta}$, we get:

$$\frac{\partial f(x;\theta)}{\partial \theta} = f(x;\theta) \cdot \frac{\partial[\log f(x;\theta)]}{\partial \theta} \quad \Rightarrow \quad \int_{-\infty}^{\infty} \frac{\partial[\log f(x;\theta)]}{\partial \theta} \cdot f(x;\theta)\ dx = 0$$

So $\mathbb{E}[S_i(\theta)] = 0$. Observe $S(\theta) = \sum S_i(\theta) \implies \mathbb{E}[S(\theta)] = 0$.

$$\hat{\theta} \approx \theta + \frac{S(\theta)}{-\text{curvature}} \quad \Rightarrow \quad \mathbb{E}[\hat{\theta}] \approx \theta + \frac{\mathbb{E}[S(\theta)]}{-\text{curvature}} = \theta$$

Therefore, $\hat{\theta}$ is asymptotically *unbiased*.

# Asymptotic Variance

The variance of a single observation's score, $S_i(\theta) = \partial[\log f(X;\theta)]/\partial\theta$ , can be found using our expected value identity,

$$\begin{aligned} 0 &= \int_{-\infty}^{\infty} \left[ \frac{\partial^2[\log f(x;\theta)]}{\partial\theta^2} f(x;\theta) + \left( \frac{\partial[\log f(x;\theta)]}{\partial\theta} \right)^2 f(x;\theta) \right] dx \\ &= \mathbb{E}\left[ \frac{\partial^2[\log f(X;\theta)]}{\partial\theta^2} \right] + \mathbb{E}\left[ \left( \frac{\partial[\log f(X;\theta)]}{\partial\theta} \right)^2 \right] = \mathbb{E}[S_i(\theta)]^2 + \mathbb{E}[S_i(\theta)^2] \end{aligned}$$

Recall that $\text{Var}(S_i(\theta)) = \mathbb{E}[S_i(\theta)^2] - (\mathbb{E}[S_i(\theta)])^2 = \mathbb{E}[S_i(\theta)^2]$. Therefore,

$$\text{Var}\left( \frac{\partial[\log f(X;\theta)]}{\partial\theta} \right) = -\mathbb{E}\left[ \frac{\partial^2[\log f(X;\theta)]}{\partial\theta^2} \right] \coloneqq I(\theta)$$

which is defined as the *Fisher Information*.

# Asymptotic Distribution of MLEs

Since $\text{Var}(S_i(\theta)) = I(\theta)$, we get:

$$\text{Var}(S(\theta)) = \sum_{i=1}^{n} \text{Var}(S_i(\theta)) = n \cdot I(\theta)$$

Recall our estimation error approximation. As $n \to \infty$, by law of large numbers, the total curvature converges to $\mathbb{E}(\text{curvature}) = n \cdot (-I(\theta))$:

$$\text{Var}(\hat{\theta}) \approx \text{Var}\left(\frac{S(\theta)}{-\text{curvature}}\right) = \text{Var}\left(\frac{S(\theta)}{n \cdot I(\theta)}\right) = \frac{n \cdot I(\theta)}{(n \cdot I(\theta))^2} = \frac{1}{n \cdot I(\theta)}$$

Therefore, as $n \to \infty$:

$$\hat{\theta} \sim \mathcal{N}\left(\theta, \frac{1}{n \cdot I(\theta)}\right)$$

# Example

We can apply our general result to the Poisson arrival rate ($\lambda$):

$$\log f(x;\lambda) = x\log\lambda - \lambda - \log x! \quad \Rightarrow \quad \frac{\partial^2[\log f(x;\lambda)]}{\partial\lambda^2} = -\frac{x}{\lambda^2}$$

The Fisher Information for a single hour is given by

$$I(\lambda) = -\mathbb{E}\left[-\frac{X}{\lambda^2}\right] = \frac{\mathbb{E}[X]}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

can be used to find the asymptotic variance of $\hat{\lambda}$,

$$\text{Var}(\hat{\lambda}) \approx \frac{1}{n \cdot I(\lambda)} = \frac{\lambda}{n} \qquad \Rightarrow \quad \hat{\lambda} \sim \mathcal{N}\left(\lambda, \frac{\lambda}{n}\right)$$

which matches our earlier result.

MLEs are great...
but are they the best?

# Efficiency of MLEs

What does “best” mean?

We love MLEs because they are asymptotically Normal, unbiased, and shrink in variance as we collect more data.

A better estimator would be unbiased, but with a smaller variance. In other words, it would require a smaller $n$ to achieve the same level of precision as the MLE.

This is the definition of an *efficient* estimator. $\hat{\theta}$ is efficient if it has the smallest variance among all unbiased estimators of $\theta$.

# Estimator Comparison

Let $\phi = u(X_1, \ldots, X_n)$ be any unbiased estimator for $\theta$. So, $\mathbb{E}(\phi) = \theta$.

We desire to find the minimum $\text{Var}(\phi)$ among all $\phi$.

Recall our Score function which represents the "information" in our sample

$$S(\theta) = \frac{\partial[\log L(\theta)]}{\partial \theta}$$

where $\mathbb{E}(S(\theta)) = 0$ and $\text{Var}(S(\theta)) = n \cdot I(\theta)$.

To see how well our arbitrary estimator $\phi$ performs, we look at its correlation with the information in our sample, by calculating their *covariance*, $\text{Cov}(\phi, S(\theta))$.

# Covariance

The covariance of $\phi$ and $S(\theta)$ is defined as

$$\begin{aligned}\text{Cov}(\phi, S(\theta)) &= \mathbb{E}[\phi \cdot S(\theta)] - \mathbb{E}[\phi] \cdot \mathbb{E}[S(\theta)] \\ &= \mathbb{E}[\phi \cdot S(\theta)] - (\theta)(0) \\ &= \mathbb{E}[\phi \cdot S(\theta)]\end{aligned}$$

We can evaluate this using the joint pdf and $L(\theta) = \prod f(X_i; \theta)$ :

$$\begin{aligned}\mathbb{E}[\phi \cdot S(\theta)] &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \phi \cdot \left( \frac{\partial [\log L(\theta)]}{\partial \theta} \right) L(\theta) \, dx_1 \dots dx_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \phi \cdot \frac{\partial L(\theta)}{\partial \theta} \, dx_1 \dots dx_n \\ &= \frac{\partial}{\partial \theta} \left[ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \phi \cdot L(\theta) \, dx_1 \dots dx_n \right] = \frac{\partial}{\partial \theta} \mathbb{E}(\phi) = \frac{\partial \theta}{\partial \theta} = 1\end{aligned}$$

$\therefore \text{Cov}(\phi, S(\theta)) = 1.$

# Rao-Cramér Inequality

To relate covariance and variance, consider the correlation coefficient $\rho$, which states for any two RVs, their correlation is bounded between $-1$ and $1$:

$$\rho^2 = \frac{[\text{Cov}(\phi, S(\theta))]^2}{\text{Var}(\phi) \cdot \text{Var}(S(\theta))} \leq 1$$

Plugging in our values, we get

$$\frac{1^2}{\text{Var}(\phi) \cdot (n \cdot I(\theta))} \leq 1 \quad \Rightarrow \quad \text{Var}(\phi) \geq \frac{1}{n \cdot I(\theta)}$$

known as the *Rao-Cramér Lower Bound (RCLB)*.

RCLB dictates that no unbiased estimator can have a lower variance than this.

# Asymptotic Efficiency

Recall that as $n \to \infty$, the distribution of the MLE is given by

$$\hat{\theta} \sim \mathcal{N}\left(\theta, \frac{1}{n \cdot I(\theta)}\right) \quad \Rightarrow \quad \text{Var}(\hat{\theta}) = \frac{1}{n \cdot I(\theta)}$$

which is exactly equal to the Rao-Cramér Lower Bound.

Therefore, MLEs are *asymptotically efficient*. They achieve the lowest possible variance among all unbiased estimators as $n \to \infty$.

*That's why we—asymptotically—love MLEs!*

# What if we have multiple parameters?

# Gradients

We've assumed $\theta$ was a single *scalar*. Many models need estimating multiple parameters simultaneously (e.g., both $\mu$ and $\sigma^2$ in a Normal distribution).

Let our parameters be a $k$-dimensional vector:

$$\theta = \left[\theta_1, \theta_2, \ldots, \theta_k\right]^T$$

Our log-likelihood is now a multi-dimensional surface.

Therefore, $S(\theta)$ is no longer a single number, but a vector of partial derivatives,

$$S(\theta) = \nabla \log L(\theta) = \left[\frac{\partial[\log L(\theta)]}{\partial \theta_1}, \ldots, \frac{\partial[\log L(\theta)]}{\partial \theta_k}\right]^T$$

which is called the *Gradient*.

## Curvature in Multi-Dimensions

In single dimensions, curvature was the second derivative. In $k$-dimensions, curvature is described by the matrix of all second-order partial derivatives

$$H_{i,j} = \frac{\partial^2[\log L(\theta)]}{\partial\theta_i \partial\theta_j}$$

known as the *Hessian Matrix* $(H)$, which describes the "shape" of the peak we are standing on.

Fisher Information also turns into the *Fisher Information Matrix*:

$$I_{i,j}(\theta) = -\mathbb{E}(H_{i,j})$$

*Remark.* Understanding the geometry of this Hessian is how methods like Natural Gradient Descent find the most efficient path to the minimum loss.

# Multivariate Asymptotic Distribution

Recall, asymptotic variance was the inverse of the Fisher Information,

$$\text{Var}(\hat{\theta}) = \frac{1}{n \cdot I(\theta)} = (n \cdot I(\theta))^{-1}$$

which in linear algebra, becomes matrix inversion.

Therefore, the MLE vector $\hat{\theta}$, converges to a *Multivariate Normal Distribution*:

$$\hat{\theta} \sim \mathcal{N}_k \left( \theta, \frac{1}{n} I^{-1}(\theta) \right)$$

Amazingly, $I^{-1}(\theta)$ gives us the *Covariance Matrix* of our estimators, revealing how the uncertainty in one parameter correlates with uncertainty in another.

## Thank You!

Questions?

**Sepehr Akbari**
akbaris79@lakeforest.edu
*Department of Math & CS, Lake Forest College*

# What if we have multiple parameters?

# Gradients

We've assumed $\theta$ was a single *scalar*. Many models need estimating multiple parameters simultaneously (e.g., both $\mu$ and $\sigma^2$ in a Normal distribution).

Let our parameters be a $k$-dimensional vector:

$$\theta = \left[\theta_1, \theta_2, \ldots, \theta_k\right]^T$$

Our log-likelihood is now a multi-dimensional surface.

Therefore, $S(\theta)$ is no longer a single number, but a vector of partial derivatives,

$$S(\theta) = \nabla \log L(\theta) = \left[\frac{\partial[\log L(\theta)]}{\partial \theta_1}, \ldots, \frac{\partial[\log L(\theta)]}{\partial \theta_k}\right]^T$$

which is called the *Gradient*.

## Curvature in Multi-Dimensions

In single dimensions, curvature was the second derivative. In $k$-dimensions, curvature is described by the matrix of all second-order partial derivatives

$$H_{i,j} = \frac{\partial^2[\log L(\theta)]}{\partial\theta_i \partial\theta_j}$$

known as the *Hessian Matrix* $(H)$, which describes the "shape" of the peak we are standing on.

Fisher Information also turns into the *Fisher Information Matrix*:

$$I_{i,j}(\theta) = -\mathbb{E}(H_{i,j})$$

*Remark.* Understanding the geometry of this Hessian is how methods like Natural Gradient Descent find the most efficient path to the minimum loss.

# Multivariate Asymptotic Distribution

Recall, asymptotic variance was the inverse of the Fisher Information,

$$\text{Var}(\hat{\theta}) = \frac{1}{n \cdot I(\theta)} = (n \cdot I(\theta))^{-1}$$

which in linear algebra, becomes matrix inversion.

Therefore, the MLE vector $\hat{\theta}$, converges to a *Multivariate Normal Distribution*:

$$\hat{\theta} \sim \mathcal{N}_k \left( \theta, \frac{1}{n} I^{-1}(\theta) \right)$$

Amazingly, $I^{-1}(\theta)$ gives us the *Covariance Matrix* of our estimators, revealing how the uncertainty in one parameter correlates with uncertainty in another.

## Thank You!

Questions?

**Sepehr Akbari**
akbaris79@lakeforest.edu
*Department of Math & CS, Lake Forest College*