Asymptotic Normality

Statistics · Module 7 — Asymptotic Theory

Node 7.3

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 7 — Asymptotic Theory. Node 7.3 — Asymptotic Normality.

Asymptotic normality is the second-order asymptotic characterization of an estimator: beyond consistency (Node 7.2), it specifies the rate of convergence and the shape of the limiting sampling distribution. This result depends on the modes of convergence (Node 7.1), consistency (Node 7.2), and moment theory (Node 1.9), and enables the delta method (Node 7.4), M-estimation theory (Node 7.5), and likelihood asymptotics (Node 7.6).

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta$, $\theta \in \Theta \subseteq \mathbb{R}^k$. A consistent estimator $\hat{\theta}_n$ is **asymptotically normal** with asymptotic variance $V(\theta)$ if

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V(\theta)) \quad \text{under } P_\theta,$$

where $V(\theta) \in \mathbb{R}^{k \times k}$ is positive definite. The rate $\sqrt{n}$ is the canonical parametric rate, meaning $\hat{\theta}_n - \theta = O_p(n^{-1/2})$.

Intuition

Asymptotic normality says that for large n, the estimator $\hat{\theta}_n$ behaves approximately as $\theta + n^{-1/2}Z$ where $Z \sim \mathcal{N}(0, V(\theta))$. The Gaussian shape arises because the estimator is determined by an average of many small independent contributions, and the CLT applies. The matrix $V(\theta)$ controls the size and shape of the confidence ellipsoid. When $V(\theta) = I(\theta)^{-1}$, the estimator achieves the Cramer-Rao bound and extracts all available information from the data.

Structural Role

Asymptotic normality is the canonical second-order characterization of estimator performance. It enables construction of confidence intervals and hypothesis tests (Modules 4-5), establishes efficiency comparisons via asymptotic relative efficiency, provides the foundation for the delta method (Node 7.4), and underlies the Bernstein-von Mises theorem (Node 6.7) connecting frequentist and Bayesian asymptotics.

Mathematical Development

Asymptotic normality of the MLE.

Under regularity conditions — (R1) the parameter space Θ is open, (R2) the support of p_θ does not depend on θ, (R3) $\log p_\theta(x)$ is three times differentiable in θ with the third derivative bounded in expectation, (R4) the Fisher information $I(\theta) = \mathbb{E}_\theta[\nabla \log p_\theta(X) \nabla \log p_\theta(X)^\top]$ is positive definite — the MLE satisfies

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0,\, I(\theta)^{-1}).$$

Proof. The MLE satisfies the score equation $\nabla_\theta \ell(\hat{\theta}_n) = 0$, where $\ell(\theta) = \sum_{i=1}^n \log p_\theta(X_i)$. Taylor-expand around the true θ:

$$0 = \nabla_\theta \ell(\theta) + \nabla_\theta^2 \ell(\theta^*) \cdot (\hat{\theta}_n - \theta),$$

where θ^* lies between $\hat{\theta}_n$ and θ (componentwise, via the mean value theorem). Rearranging and multiplying by $n^{-1/2}$:

$$\sqrt{n}(\hat{\theta}_n - \theta) = -\left[\frac{1}{n}\nabla_\theta^2 \ell(\theta^*)\right]^{-1} \frac{1}{\sqrt{n}} \nabla_\theta \ell(\theta).$$

CLT for the score. Since $\mathbb{E}_\theta[\nabla \log p_\theta(X_i)] = 0$ (the score identity) and $\mathrm{Cov}_\theta(\nabla \log p_\theta(X_i)) = I(\theta)$, the multivariate CLT gives

$$\frac{1}{\sqrt{n}} \nabla_\theta \ell(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla \log p_\theta(X_i) \xrightarrow{d} \mathcal{N}(0, I(\theta)).$$

WLLN for the Hessian. By the law of large numbers and consistency of $\hat{\theta}_n$ (hence $\theta^* \xrightarrow{p} \theta$),

$$\frac{1}{n}\nabla_\theta^2 \ell(\theta^*) \xrightarrow{p} \mathbb{E}_\theta[\nabla^2 \log p_\theta(X)] = -I(\theta).$$

The last equality uses the second Bartlett identity: $\mathbb{E}[\nabla^2 \log p_\theta] = -I(\theta)$.

Combining via Slutsky. By Slutsky's theorem,

$$\sqrt{n}(\hat{\theta}_n - \theta) = I(\theta)^{-1} \cdot \frac{1}{\sqrt{n}} \nabla_\theta \ell(\theta) + o_p(1) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1}). \quad \square$$

Proof for exponential families. Let $p_\theta(x) = h(x) \exp(\eta(\theta)^\top T(x) - A(\eta(\theta)))$ with natural parameter $\eta = \eta(\theta)$. The score is $\nabla_\theta \ell(\theta) = \sum_i [\dot{\eta}(\theta)]^\top (T(X_i) - \nabla_\eta A(\eta))$. The Fisher information is $I(\theta) = \dot{\eta}(\theta)^\top \nabla_\eta^2 A(\eta)\, \dot{\eta}(\theta)$. Since $\nabla^2 \log p_\theta$ does not depend on the data (in the natural parametrization, it equals $-\nabla_\eta^2 A(\eta)$), the WLLN step is exact: the sample Hessian equals the expected Hessian. The regularity conditions (R3) are automatically satisfied, making the exponential family the cleanest setting for asymptotic normality.

Asymptotic relative efficiency. For two asymptotically normal estimators $\hat{\theta}_n^{(1)}$ and $\hat{\theta}_n^{(2)}$ with scalar θ,

$$\mathrm{ARE}(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}) = \frac{V_2(\theta)}{V_1(\theta)}.$$

The Cramer-Rao lower bound (Node 3.4) states $V(\theta) \geq I(\theta)^{-1}$ for all regular estimators. An estimator achieving $V(\theta) = I(\theta)^{-1}$ is **asymptotically efficient**. The MLE is asymptotically efficient under the regularity conditions above.

Sandwich variance under misspecification. When the model is misspecified ($P_0 \notin \{P_\theta : \theta \in \Theta\}$), the MLE converges to the pseudo-true parameter $\theta^* = \arg\min_\theta \mathrm{KL}(P_0 \| P_\theta)$. The asymptotic variance takes the sandwich form:

$$V = A(\theta^*)^{-1} B(\theta^*) A(\theta^*)^{-\top},$$

where $A(\theta^*) = \mathbb{E}_0[-\nabla^2 \log p_{\theta^*}(X)]$ and $B(\theta^*) = \mathbb{E}_0[\nabla \log p_{\theta^*}(X) \nabla \log p_{\theta^*}(X)^\top]$. Under correct specification, the second Bartlett identity gives $A = B = I(\theta)$ and the sandwich reduces to $I(\theta)^{-1}$.

Connection to Fisher information. The Fisher information matrix $I(\theta)$ plays three roles simultaneously: (i) it is the variance of the score, (ii) it is the negative expected Hessian of the log-likelihood, (iii) it is the inverse of the asymptotic variance of the efficient estimator. These three characterizations coincide under correct specification and regularity.

Asymptotic normality of the sample mean. As a baseline: for i.i.d. X_i with mean μ and finite variance σ^2, the CLT gives $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. This is the prototype for all asymptotic normality results. The MLE generalizes this: the score $\nabla \log p_\theta(X_i)$ plays the role of $X_i - \mu$, and the Fisher information $I(\theta)$ plays the role of σ^2.

Non-regular cases. Asymptotic normality at rate $\sqrt{n}$ requires regularity. For the uniform model $\mathrm{Uniform}(0, \theta)$, the MLE $X_{(n)}$ converges at rate n (not $\sqrt{n}$) and the limit distribution is exponential, not Gaussian. The regularity condition that fails is differentiability of the support with respect to θ. For the Laplace location model, the MLE converges at rate $\sqrt{n}$ but the proof requires different techniques (the score has a discontinuity).

Multivariate CLT statement. The multivariate CLT, which underpins the score asymptotics, states: if $X_1, \ldots, X_n$ are i.i.d. in $\mathbb{R}^k$ with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Cov}(X_i) = \Sigma$, then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

The proof uses the Cramer-Wold device: show $a^\top \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, a^\top \Sigma a)$ for all $a \in \mathbb{R}^k$ using the univariate CLT, then invoke the equivalence between convergence of all one-dimensional projections and convergence of the joint distribution.

Worked Examples

Example 1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Bernoulli}(p)$. The MLE is $\hat{p} = \bar{X}_n$. The Fisher information is $I(p) = 1/(p(1-p))$. By asymptotic normality, $\sqrt{n}(\hat{p} - p) \xrightarrow{d} \mathcal{N}(0, p(1-p))$. This recovers the CLT for the sample proportion. The asymptotic 95% confidence interval is $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$.

Example 2. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with both parameters unknown, $\theta = (\mu, \sigma^2)$. The Fisher information matrix is

$$I(\theta) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}.$$

The MLE is $(\bar{X}_n, \hat{\sigma}_n^2)$ where $\hat{\sigma}_n^2 = n^{-1} \sum (X_i - \bar{X}_n)^2$. By asymptotic normality,

$$\sqrt{n} \begin{pmatrix} \bar{X}_n - \mu \\ \hat{\sigma}_n^2 - \sigma^2 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right).$$

The off-diagonal zeros reflect asymptotic independence of $\bar{X}_n$ and $\hat{\sigma}_n^2$, which is exact under normality.

Example 3 (Comparing estimators via ARE). For $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, the sample mean has asymptotic variance σ^2 and the sample median has asymptotic variance $\pi\sigma^2/2$. The ARE of the mean relative to the median is $(\pi\sigma^2/2)/\sigma^2 = \pi/2 \approx 1.57$: the mean is 57% more efficient at the normal. However, for heavy-tailed distributions (e.g., Cauchy), the median is consistent while the mean is not even defined, illustrating that ARE is model-dependent.

Cross-Links

AI. In stochastic gradient descent, the iterate θ_n satisfies an analogous asymptotic normality result under decreasing step sizes: $\sqrt{n}(\theta_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, V)$ where V depends on the noise covariance and the Hessian of the loss, mirroring the sandwich structure.

Economics. Asymptotic normality of GMM estimators with the efficient weighting matrix yields variance $(G^\top S^{-1} G)^{-1}$, the econometric analogue of the inverse Fisher information.

Linear Algebra. The proof uses the spectral properties of the Fisher information matrix: positive definiteness ensures invertibility, and the quadratic form $h^\top I(\theta) h$ defines a Riemannian metric on the parameter space.

Quantum Mechanics. The quantum Cramer-Rao bound gives the minimum variance for estimating a quantum state parameter, with the quantum Fisher information replacing $I(\theta)$; the classical asymptotic normality result holds for quantum measurements in the limit of many identical preparations.

Structural Compression

Invariant. $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})$: the standardized MLE converges in distribution to a Gaussian with variance equal to the inverse Fisher information.

Mechanism. Score equation linearized via Taylor expansion; CLT applied to the score; WLLN applied to the Hessian; Slutsky yields the Gaussian limit. The Fisher information emerges as both the score variance and the negative expected Hessian.

Cross-System Echo. Asymptotic normality is the statistical manifestation of the central limit theorem for estimating functions. In information geometry, the Fisher information defines the unique invariant Riemannian metric on the statistical manifold, and asymptotic normality states that the MLE is a normal coordinate chart centered at the truth.

Exercises

1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$. Compute the Fisher information and verify that the MLE $\hat{\lambda} = \bar{X}_n$ achieves the asymptotic variance $I(\lambda)^{-1} = \lambda$.

2. Verify the second Bartlett identity $\mathbb{E}[\nabla^2 \log p_\theta(X)] = -I(\theta)$ for the exponential distribution $p_\lambda(x) = \lambda e^{-\lambda x}$.

3. For the Cauchy location model $p_\theta(x) = \pi^{-1}(1 + (x - \theta)^2)^{-1}$, the MLE exists and is consistent but the Fisher information involves a nontrivial integral. Compute $I(\theta)$ and state the asymptotic distribution of the MLE.

4. Show that the sample median of $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ is asymptotically normal with variance $\pi\sigma^2/(2n)$, and compute its ARE relative to the sample mean.

5. Derive the sandwich variance formula when $X_i \overset{\text{i.i.d.}}{\sim} \text{Exp}(1)$ but the working model is $\mathcal{N}(\mu, \sigma^2)$. Identify the pseudo-true parameters.

6. Let $\hat{\theta}_n$ be asymptotically normal with variance $V(\theta)$. Prove that $a^\top \hat{\theta}_n$ is asymptotically normal with variance $a^\top V(\theta) a$ for any fixed $a \in \mathbb{R}^k$.

7. Argue, structurally, why the $\sqrt{n}$ rate is universal for parametric models while nonparametric rates are slower, connecting the rate to the dimension of the parameter space and the CLT.

Statistics · Module 7 — Asymptotic Theory · Node 7.3
Concepts: central-limit-theorem, convergence, fisher-information
Prerequisites: 7.1, 7.2, 1.9
Enables: 7.4, 7.5, 7.6

hbar.university — own.your.cognition