Likelihood Asymptotics

Statistics · Module 7 — Asymptotic Theory

Node 7.6

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 7 — Asymptotic Theory. Node 7.6 — Likelihood Asymptotics.

Likelihood asymptotics studies the large-sample geometry of the log-likelihood surface and the asymptotic behavior of all inference procedures derived from it. This node depends on asymptotic normality (Node 7.3), M-estimation (Node 7.5), and the exponential family and sufficiency theory (Nodes 2.3, 2.5). It is the capstone of Module 7, unifying the Wald, Score, and Likelihood Ratio tests under a single asymptotic framework.

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}$, $\theta_0 \in \Theta \subseteq \mathbb{R}^k$. Define the log-likelihood $\ell(\theta) = \sum_{i=1}^n \log p_\theta(X_i)$ and the normalized score $\Delta_n = n^{-1/2}\nabla_\theta \ell(\theta_0)$.

Local Asymptotic Normality (LAN). The model $\{P_\theta^{(n)}\}$ satisfies LAN at θ_0 if for every $h \in \mathbb{R}^k$,

$$\log \frac{L(\theta_0 + h/\sqrt{n})}{L(\theta_0)} = h^\top \Delta_n - \frac{1}{2} h^\top I(\theta_0) h + o_p(1),$$

where $\Delta_n \xrightarrow{d} \mathcal{N}(0, I(\theta_0))$.

Wilks' theorem. Under $H_0 : \theta = \theta_0$,

$$\Lambda_n = 2[\ell(\hat{\theta}_n) - \ell(\theta_0)] \xrightarrow{d} \chi^2_k.$$

Intuition

LAN says that at the local scale $1/\sqrt{n}$ around the truth, every regular parametric model looks like a Gaussian shift experiment. The log-likelihood ratio is a quadratic function of the local parameter h plus Gaussian noise — the same structure as observing $Z = h + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I^{-1})$. All first-order asymptotic inference is determined by this Gaussian approximation. The three classical test statistics (Wald, Score, LR) are simply three different ways to measure distance from the null in this locally Gaussian world.

Structural Role

LAN is the deepest structural statement about likelihood-based inference. It provides the theoretical foundation for: (i) the asymptotic equivalence of Wald, Score, and LR tests (Nodes 4.4, 5.4); (ii) the Cramer-Rao efficiency bound (Node 3.4); (iii) the Bernstein-von Mises theorem (Node 6.7); and (iv) the

asymptotic optimality of the MLE. Every regular parametric model is, locally, a Gaussian shift experiment — this is the fundamental simplification that makes asymptotic inference tractable.

Mathematical Development

Score asymptotics. At the true parameter θ_0, the score contributions $s_i = \nabla_\theta \log p_{\theta_0}(X_i)$ are i.i.d. with $\mathbb{E}[s_i] = 0$ (score identity) and $\mathrm{Cov}(s_i) = I(\theta_0)$ (Fisher information). By the multivariate CLT,

$$\Delta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i \xrightarrow{d} \mathcal{N}(0, I(\theta_0)).$$

Proof of LAN. Taylor-expand the log-likelihood ratio:

$$\ell(\theta_0 + h/\sqrt{n}) - \ell(\theta_0) = \frac{h^\top}{\sqrt{n}} \nabla \ell(\theta_0) + \frac{1}{2n} h^\top \nabla^2 \ell(\theta_n^*) h,$$

where θ_n^* lies between θ_0 and $\theta_0 + h/\sqrt{n}$. The first term is $h^\top \Delta_n$. For the second term, $n^{-1}\nabla^2 \ell(\theta_n^*) \xrightarrow{p} -I(\theta_0)$ by the WLLN and continuity. Thus the second term converges to $-h^\top I(\theta_0) h/2$, establishing LAN. □

Asymptotic equivalence of the three test statistics.

Under $H_0 : \theta = \theta_0$:

Wald statistic:

$$W_n = n(\hat{\theta}_n - \theta_0)^\top I(\hat{\theta}_n)(\hat{\theta}_n - \theta_0).$$

Score (Rao) statistic:

$$S_n = \frac{1}{n} \nabla \ell(\theta_0)^\top I(\theta_0)^{-1} \nabla \ell(\theta_0) = \Delta_n^\top I(\theta_0)^{-1} \Delta_n.$$

Likelihood ratio statistic:

$$\Lambda_n = 2[\ell(\hat{\theta}_n) - \ell(\theta_0)].$$

All three converge to χ^2_k under H_0, and $W_n - \Lambda_n = o_p(1)$, $S_n - \Lambda_n = o_p(1)$.

Proof sketch for Wilks' theorem. By LAN with $h = \sqrt{n}(\hat{\theta}_n - \theta_0)$:

$$\Lambda_n = 2[\ell(\hat{\theta}_n) - \ell(\theta_0)] = 2h^\top \Delta_n - h^\top I(\theta_0) h + o_p(1).$$

Using the asymptotic normality of the MLE, $h = \sqrt{n}(\hat{\theta}_n - \theta_0) = I(\theta_0)^{-1}\Delta_n + o_p(1)$. Substituting:

$$\Lambda_n = 2\Delta_n^\top I(\theta_0)^{-1}\Delta_n - \Delta_n^\top I(\theta_0)^{-1} I(\theta_0) I(\theta_0)^{-1} \Delta_n + o_p(1) = \Delta_n^\top I(\theta_0)^{-1}\Delta_n + o_p(1).$$

Since $\Delta_n \xrightarrow{d} \mathcal{N}(0, I(\theta_0))$, the quadratic form $\Delta_n^\top I(\theta_0)^{-1}\Delta_n \xrightarrow{d} \chi^2_k$. □

Composite null hypothesis. For $H_0 : \theta \in \Theta_0$ where Θ_0 is a smooth $(k-q)$-dimensional submanifold of Θ, the LR statistic satisfies $\Lambda_n \xrightarrow{d} \chi^2_q$ under H_0, where $q = k - \dim(\Theta_0)$ is the number of constraints.

Bartlett correction. The LR statistic admits a refinement: defining $\Lambda_n^* = \Lambda_n/(1 + a/n)$ where a depends on the third and fourth cumulants of the log-likelihood, the corrected statistic satisfies $\mathbb{P}(\Lambda_n^* \le x) = F_{\chi^2_k}(x) + O(n^{-2})$, improving the $O(n^{-1})$ error of the uncorrected statistic. This correction is exact for exponential families.

Profile likelihood. For a partition $\theta = (\psi, \lambda)$ where ψ is the parameter of interest and λ is the nuisance parameter, the profile likelihood is $L_P(\psi) = \max_\lambda L(\psi, \lambda)$. The profile log-likelihood ratio $2[\ell_P(\hat{\psi}) - \ell_P(\psi_0)] \xrightarrow{d} \chi^2_{\dim(\psi)}$ under $H_0 : \psi = \psi_0$, eliminating the nuisance parameter.

Asymptotic efficiency of the MLE. Under LAN, the MLE achieves the asymptotic minimax bound: for any loss function $L(\theta, \hat{\theta}) = w(\hat{\theta} - \theta)$ with w subconvex and symmetric, no regular estimator can have smaller asymptotic risk than the MLE. This is the Hajek-Le Cam convolution theorem: any regular estimator $\hat{\theta}_n$ satisfies $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} Z + W$ where $Z \sim \mathcal{N}(0, I^{-1})$ and W is independent noise. The MLE has $W = 0$.

Contiguity. Two sequences P_n, Q_n are contiguous if $P_n(A_n) \to 0 \implies Q_n(A_n) \to 0$. Under LAN, $P^{(n)}_{\theta_0 + h/\sqrt{n}}$ and $P^{(n)}_{\theta_0}$ are contiguous. Le Cam's third lemma: if (T_n, Λ_n) converges jointly under P_{θ_0}, the distribution of T_n under the contiguous alternative is obtained by tilting the joint limit. This enables power calculations for local alternatives.

Power under local alternatives. Under $\theta = \theta_0 + h/\sqrt{n}$, the LR statistic has a noncentral chi-squared distribution:

$$\Lambda_n \xrightarrow{d} \chi^2_k(\delta), \quad \delta = h^\top I(\theta_0) h.$$

The power of the LR test at level α is $\mathbb{P}(\chi^2_k(\delta) > \chi^2_{k,\alpha})$, which increases with the noncentrality δ. The Wald and Score tests have the same asymptotic power function under local alternatives, confirming first-order equivalence.

Higher-order asymptotics. The first-order theory treats all three test statistics as equivalent. Second-order analysis reveals differences: - The LR statistic admits Bartlett correction (improved χ^2 approximation to order $O(n^{-2})$). - The Wald statistic is not invariant under reparametrization: testing $H_0 : \theta = \theta_0$ and $H_0 : g(\theta) = g(\theta_0)$ can give different results. - The Score statistic requires only the restricted (null) estimate, making it computationally simpler.

In practice, the LR test is generally preferred for its reparametrization invariance and Bartlett correctability.

Asymptotic optimality of the Neyman-Pearson test. For testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_0 + h/\sqrt{n}$, the LR test is asymptotically most powerful among all tests of the same level. This follows from the Neyman-Pearson lemma applied to the locally Gaussian limit experiment.

Worked Examples

Example 1 (LR test for normal mean). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$. Test $H_0 : \mu = 0$. The LR statistic is $\Lambda_n = 2[\ell(\bar{X}) - \ell(0)] = 2 \cdot n[\bar{X}^2/2] = n\bar{X}^2$. Under H_0, $\sqrt{n}\bar{X} \sim \mathcal{N}(0, 1)$, so $\Lambda_n \sim \chi^2_1$ exactly. The Wald statistic is $W_n = n\bar{X}^2$ and the Score statistic is $S_n = n\bar{X}^2$ — all three coincide exactly for the normal location model.

Example 2 (LR test for Poisson rate). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$. Test $H_0 : \lambda = \lambda_0$. The MLE is $\hat{\lambda} = \bar{X}$. The LR statistic is $\Lambda_n = 2n[\bar{X} \log(\bar{X}/\lambda_0) - (\bar{X} - \lambda_0)]$. Under H_0, $\Lambda_n \xrightarrow{d} \chi^2_1$. The Wald statistic is $W_n = n(\bar{X} - \lambda_0)^2/\bar{X}$ and the Score statistic is $S_n = n(\bar{X} - \lambda_0)^2/\lambda_0$. Here the three statistics differ at finite n; the Bartlett correction improves the chi-squared approximation for Λ_n.

Cross-Links

AI. The likelihood ratio test underpins model selection criteria (AIC, BIC) that penalize the LR statistic by model complexity. In deep learning, the local quadratic approximation of the loss surface (LAN analogue) is the basis for second-order optimization methods (natural gradient, Fisher-information-based preconditioning).

Economics. The LR, Wald, and Score (Lagrange multiplier) tests are the trinity of econometric testing; the Wald test is most common in practice due to its computational simplicity (requires only the unrestricted estimate).

Linear Algebra. The LAN expansion is a second-order Taylor expansion of a function on a Riemannian manifold, where the Fisher information is the metric tensor. The chi-squared limit of the LR statistic is the squared Riemannian distance from the null.

Quantum Mechanics. Quantum Local Asymptotic Normality (qLAN) extends LAN to quantum state estimation: n copies of a quantum state ρ_θ are asymptotically equivalent to a quantum Gaussian shift experiment, with the quantum Fisher information replacing $I(\theta)$.

Structural Compression

Invariant. At local scale $1/\sqrt{n}$, the log-likelihood ratio is a quadratic Gaussian process; all first-order asymptotic inference is determined by the Fisher information $I(\theta_0)$.

Mechanism. Taylor expansion of the log-likelihood to second order; CLT applied to the score; WLLN applied to the Hessian; the resulting Gaussian quadratic form produces χ^2 limits for all three test statistics. Contiguity transfers these results to local alternatives.

Cross-System Echo. LAN is the statistical instance of a general principle in information geometry: the Fisher information metric is the unique Riemannian metric on the manifold of probability distributions invariant under sufficient statistics. The Hajek-Le Cam theory establishes that LAN models are, asymptotically, Gaussian shift experiments — the simplest possible inference problem.

Exercises

1. Verify the LAN expansion explicitly for the Bernoulli model $p_\theta(x) = \theta^x(1-\theta)^{1-x}$ by computing Δ_n, $I(\theta_0)$, and the remainder term.
2. For the exponential family $p_\eta(x) = h(x)\exp(\eta x - A(\eta))$, show that the LAN expansion is exact to second order (the remainder is identically zero).
3. Prove that the Wald, Score, and LR statistics are asymptotically equivalent under H_0 by expressing each as $\Delta_n^\top I(\theta_0)^{-1}\Delta_n + o_p(1)$.
4. Compute the Bartlett correction factor a for the Poisson model and verify that $\Lambda_n/(1 + a/n)$ has improved chi-squared approximation.
5. For the normal model $\mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$, construct the profile likelihood for μ by maximizing over σ^2, and show that the profile LR statistic for $H_0 : \mu = \mu_0$ equals $n\log(1 + T^2/(n-1))$ where T is the usual t-statistic.
6. Use Le Cam's third lemma to derive the power of the LR test for $H_0 : \mu = 0$ against the local alternative $\mu = h/\sqrt{n}$ in the $\mathcal{N}(\mu, 1)$ model.
7. Argue, structurally, why LAN implies that all regular parametric models are asymptotically equivalent to Gaussian shift experiments, and explain why this makes the Fisher information the only relevant local quantity for asymptotic inference.
