Module 7 — Asymptotic Theory

Statistics

hbar.university

Modes of Convergence

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 7 — Asymptotic Theory. Node 7.1 — Modes of Convergence.

Modes of convergence form the foundational language for all large-sample results in statistics. Every asymptotic statement — consistency, asymptotic normality, convergence of test statistics — is expressed through one of the convergence modes defined here. This node depends on the measure-theoretic probability framework (Node 1.8) and moment theory (Node 1.9), and enables the consistency (Node 7.2) and asymptotic normality (Node 7.3) results that follow.

Formal Definition

Let $(T_n)_{n\geq 1}$ be a sequence of random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, taking values in $\mathbb{R}^k$.

Convergence in probability. $T_n \xrightarrow{p} T$ if for every $\varepsilon > 0$,

$$\mathbb{P}(|T_n - T| > \varepsilon) \to 0 \quad \text{as } n \to \infty.$$

Almost sure convergence. $T_n \xrightarrow{a.s.} T$ if

$$\mathbb{P}\Big(\lim_{n\to\infty} T_n = T\Big) = 1.$$

Equivalently, for every $\varepsilon > 0$, $\mathbb{P}(\sup_{m\geq n} |T_m - T| > \varepsilon) \to 0$ as $n \to \infty$.

Convergence in distribution. $T_n \xrightarrow{d} T$ if for every continuity point t of the CDF F_T,

$$F_{T_n}(t) \to F_T(t) \quad \text{as } n \to \infty.$$

Equivalently, $\mathbb{E}[f(T_n)] \to \mathbb{E}[f(T)]$ for every bounded continuous function f (the Portmanteau characterization).

L^p convergence. For $p \geq 1$, $T_n \xrightarrow{L^p} T$ if $\mathbb{E}[|T_n - T|^p] \to 0$. The case $p = 2$ is mean-square convergence.

Intuition

Almost sure convergence is path-level: on almost every realization ω, the sequence $T_n(\omega)$ literally converges. Convergence in probability is weaker: the probability of a large deviation vanishes, but exceptional paths may misbehave infinitely often. Convergence in distribution is weakest: it says only that the laws of T_n approach the law of T, without requiring the random variables to be defined on the same space. L^p convergence controls the average magnitude of the deviation, and its strength depends on p.

The distinction matters because different statistical guarantees require different modes. Consistency of an estimator is convergence in probability; the central limit theorem gives convergence in distribution; the strong law of large numbers gives almost sure convergence.

Structural Role

Modes of convergence are the formal language for all large-sample statements in statistics. Consistency (Node 7.2) is convergence in probability of an estimator to the true parameter. Asymptotic normality (Node 7.3) is convergence in distribution of a standardized estimator to a Gaussian. Slutsky's theorem and the continuous mapping theorem, stated below, are the primary tools for combining and transforming convergent sequences to establish distributional limits of composite statistics. The hierarchy of modes determines which conclusions transfer between settings.

Mathematical Development

Theorem (Hierarchy of modes).

$$T_n \xrightarrow{a.s.} T \implies T_n \xrightarrow{p} T \implies T_n \xrightarrow{d} T.$$

$$T_n \xrightarrow{L^p} T \implies T_n \xrightarrow{p} T.$$

Proof that a.s. implies in probability. Fix $\varepsilon > 0$. Define $A_n = \{|T_n - T| > \varepsilon\}$. Almost sure convergence means $\mathbb{P}(\limsup_n A_n) = 0$. Since $\mathbb{P}(A_n) \le \mathbb{P}(\bigcup_{m \ge n} A_m)$ and $\mathbb{P}(\bigcup_{m \ge n} A_m) \downarrow \mathbb{P}(\limsup_n A_n) = 0$, we have $\mathbb{P}(A_n) \to 0$. $\square$

Proof that L^p implies in probability. By Markov's inequality, $\mathbb{P}(|T_n - T| > \varepsilon) \le \varepsilon^{-p}\mathbb{E}[|T_n - T|^p] \to 0$. $\square$

Proof that in probability implies in distribution. Fix a continuity point t of F_T. For any $\delta > 0$,

$$F_{T_n}(t) = \mathbb{P}(T_n \le t) \le \mathbb{P}(T \le t + \delta) + \mathbb{P}(|T_n - T| > \delta).$$

Similarly, $F_{T_n}(t) \ge \mathbb{P}(T \le t - \delta) - \mathbb{P}(|T_n - T| > \delta)$. Taking $n \to \infty$ and then $\delta \to 0$ using continuity of F_T at t yields $F_{T_n}(t) \to F_T(t)$. $\square$

Converse: convergence in distribution to a constant. If $T_n \xrightarrow{d} c$ for a constant c, then $T_n \xrightarrow{p} c$. This follows because $F_c(t) = \mathbf{1}(t \ge c)$ has a single discontinuity, and pointwise convergence of CDFs at all continuity points forces the mass to concentrate at c.

Theorem (Continuous Mapping Theorem). If $T_n \xrightarrow{d} T$ and $g : \mathbb{R}^k \to \mathbb{R}^m$ is continuous $\mathbb{P}_T$-almost everywhere, then $g(T_n) \xrightarrow{d} g(T)$. The same holds with $\xrightarrow{p}$ and $\xrightarrow{a.s.}$ in place of $\xrightarrow{d}$.

Proof sketch for the distributional case. Let D_g be the set of discontinuity points of g. By assumption, $\mathbb{P}(T \in D_g) = 0$. For any bounded continuous f, the composition $f \circ g$ is continuous on $\mathbb{R}^k \setminus D_g$. By the Portmanteau theorem extended to functions continuous a.e., $\mathbb{E}[f(g(T_n))] \to \mathbb{E}[f(g(T))]$. $\square$

Theorem (Slutsky). If $T_n \xrightarrow{d} T$ and $S_n \xrightarrow{p} c$ for a constant c, then:

$$T_n + S_n \xrightarrow{d} T + c, \qquad T_n S_n \xrightarrow{d} cT, \qquad T_n / S_n \xrightarrow{d} T/c \ \ (c \neq 0).$$

Proof. Consider the map $g(t, s) = t + s$. The pair $(T_n, S_n) \xrightarrow{d} (T, c)$ since $S_n \xrightarrow{p} c$ (the joint convergence follows because convergence in probability to a constant makes the marginals asymptotically independent). Applying the continuous mapping theorem to g yields $T_n + S_n \xrightarrow{d} T + c$. The product and ratio cases follow analogously. □

Skorokhod representation theorem. If $T_n \xrightarrow{d} T$, there exist random variables T_n', T' on a common probability space with the same marginal distributions such that $T_n' \xrightarrow{a.s.} T'$. This converts distributional convergence to almost sure convergence on a possibly different probability space, often simplifying proofs.

Borel-Cantelli and almost sure convergence. The first Borel-Cantelli lemma states: if $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\limsup_n A_n) = 0$. Applied with $A_n = \{|T_n - T| > \varepsilon\}$: if $\sum_n \mathbb{P}(|T_n - T| > \varepsilon) < \infty$ for every $\varepsilon > 0$, then $T_n \xrightarrow{a.s.} T$. This provides a checkable sufficient condition for almost sure convergence from tail probability bounds.

Subsequence characterization. $T_n \xrightarrow{p} T$ if and only if every subsequence of (T_n) has a further subsequence that converges almost surely to T. This characterization is frequently used in proofs: to show convergence in probability, extract a subsequence via the diagonal argument and upgrade to almost sure convergence.

Dominated convergence and L^p convergence. If $T_n \xrightarrow{p} T$ and $|T_n|^p \leq Y$ a.s. for some integrable Y, then $T_n \xrightarrow{L^p} T$. This is the probabilistic version of the dominated convergence theorem, and it provides the bridge from convergence in probability back up to L^p convergence when a dominating bound exists.

Multivariate extension. All four modes extend to $\mathbb{R}^k$-valued random vectors by replacing $|T_n - T|$ with $\|T_n - T\|$ (any norm; all norms on $\mathbb{R}^k$ are equivalent). The Cramer-Wold device characterizes convergence in distribution of random vectors: $T_n \xrightarrow{d} T$ in $\mathbb{R}^k$ if and only if $a^\top T_n \xrightarrow{d} a^\top T$ for every $a \in \mathbb{R}^k$. This reduces multivariate distributional convergence to checking one-dimensional convergence along all projections.

Worked Examples

Example 1. Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{E}[X_i] = \mu$, $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Set $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. By the weak law of large numbers, $\bar{X}_n \xrightarrow{p} \mu$. By the strong law, $\bar{X}_n \xrightarrow{a.s.} \mu$. Since $\mathbb{E}[(\bar{X}_n - \mu)^2] = \sigma^2/n \to 0$, we also have $\bar{X}_n \xrightarrow{L^2} \mu$. By the CLT, $\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} \mathcal{N}(0, 1)$. All four modes appear in this single example.

Example 2 (Slutsky application). Suppose $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ and we estimate σ^2 by $S_n^2 = (n-1)^{-1} \sum (X_i - \bar{X}_n)^2$. Since $S_n^2 \xrightarrow{p} \sigma^2$, Slutsky gives $\sqrt{n}(\bar{X}_n - \mu)/S_n \xrightarrow{d} \mathcal{N}(0, 1)$. This justifies the standard z-type confidence interval with estimated variance.

Example 3 (Continuous mapping). Let $T_n \xrightarrow{d} \mathcal{N}(0, 1)$. Then $T_n^2 \xrightarrow{d} \chi_1^2$ by the continuous mapping theorem with $g(t) = t^2$. More generally, if $T_n \xrightarrow{d} \mathcal{N}(0, I_k)$, then $\|T_n\|^2 \xrightarrow{d} \chi_k^2$.

Example 4 (Convergence in distribution but not in probability). Let $X \sim \mathcal{N}(0, 1)$ and define $T_n = (-1)^n X$. Then $T_n \sim \mathcal{N}(0, 1)$ for every n, so $T_n \xrightarrow{d} \mathcal{N}(0, 1)$. But $T_n - T_{n+1} = (-1)^n X - (-1)^{n+1} X = 2(-1)^n X$, which does not converge to 0, so T_n does not converge in probability (or almost surely). Convergence in distribution says nothing about the joint behavior of the sequence.

Example 5 (L^p convergence vs. almost sure). Let $\Omega = [0, 1]$ with Lebesgue measure. Define $T_n = n \cdot \mathbf{1}_{[0, 1/n]}$. Then $T_n \xrightarrow{a.s.} 0$ (for any $\omega > 0$, eventually $T_n(\omega) = 0$). However, $\mathbb{E}[|T_n|^p] = n^p/n = n^{p-1} \to \infty$ for $p > 1$, so $T_n \not\xrightarrow{L^p} 0$ for $p > 1$. Almost sure convergence does not imply L^p convergence without uniform integrability.

Theorem (Uniform integrability bridge). $T_n \xrightarrow{L^1} T$ if and only if $T_n \xrightarrow{p} T$ and $\{T_n\}$ is uniformly integrable: $\lim_{M\to\infty} \sup_n \mathbb{E}[|T_n|\mathbf{1}_{|T_n|>M}] = 0$. This provides the missing link between convergence in probability and L^p convergence.

Cross-Links

AI. Convergence in distribution underpins the theoretical analysis of stochastic gradient descent: under appropriate step-size schedules, the iterate distribution converges to a stationary distribution, and Slutsky-type arguments justify plug-in variance estimation in online learning.

Economics. Asymptotic results in econometrics (consistency and normality of GMM estimators) are stated entirely in terms of these modes; the distinction between convergence in probability and almost sure convergence determines whether path-dependent economic dynamics can be analyzed.

Linear Algebra. Convergence in L^2 is norm convergence in the Hilbert space $L^2(\Omega, \mathbb{P})$; the hierarchy of modes mirrors the hierarchy of topologies (norm, weak, weak-$*$) on Banach spaces.

Quantum Mechanics. Convergence of density matrices in trace norm is the quantum analogue of L^1 convergence; weak convergence of quantum states corresponds to convergence in distribution of measurement outcomes.

Structural Compression

Invariant. The hierarchy a.s. $\Rightarrow$ in prob. $\Rightarrow$ in dist. is strict; each mode imposes a progressively weaker condition on the probability measure. $L^p \Rightarrow$ in prob. provides an independent chain.

Mechanism. Each mode controls a different aspect of the probability mass: almost sure convergence controls path behavior; convergence in probability controls tail probabilities; convergence in distribution controls only the law. L^p convergence controls moment-weighted deviations. The continuous mapping theorem and Slutsky's theorem transfer convergence through smooth operations.

Cross-System Echo. Convergence in distribution is weak convergence of measures; in functional analysis, this is weak-$*$ convergence of probability measures as linear functionals on $C_b(\mathbb{R})$. The Skorokhod representation theorem bridges distributional and pathwise convergence by constructing a coupling.

Exercises

1. Construct a sequence T_n that converges to 0 in probability but not almost surely. Verify both claims.
2. Let $T_n \xrightarrow{d} \mathcal{N}(0,1)$. Prove that $|T_n| \xrightarrow{d} |Z|$ where $Z \sim \mathcal{N}(0,1)$, and identify the distribution of $|Z|$.
3. Suppose $T_n \xrightarrow{L^2} T$. Prove that $\mathbb{E}[T_n] \to \mathbb{E}[T]$ and $\mathrm{Var}(T_n) \to \mathrm{Var}(T)$.
4. Let $X_1, \ldots, X_n$ be i.i.d. Uniform$(0, \theta)$. Show that $X_{(n)} = \max_i X_i$ satisfies $X_{(n)} \xrightarrow{a.s.} \theta$. Find the rate: determine the limit distribution of $n(\theta - X_{(n)})$.
5. Prove that if $T_n \xrightarrow{p} T$ and $T_n \xrightarrow{p} T'$, then $T = T'$ a.s. (uniqueness of limits in probability).
6. Let $T_n \xrightarrow{d} T$ and $S_n \xrightarrow{d} S$. Give a counterexample showing that $T_n + S_n$ need not converge in distribution to $T + S$ without additional conditions.
7. Argue, structurally, why convergence in distribution is the natural mode for the central limit theorem while convergence in probability is the natural mode for the law of large numbers, relating each to the type of statistical conclusion it supports.

Consistency

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 7 — Asymptotic Theory. Node 7.2 — Consistency.

Consistency is the minimal asymptotic requirement for an estimator: as the sample size grows, the estimator must converge to the true parameter. This node instantiates the convergence modes of Node 7.1 in the estimation setting, depends on the likelihood framework (Nodes 3.1, 3.2), and enables the rate-of-convergence results in asymptotic normality (Node 7.3) and the general M-estimation framework (Node 7.5).

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta$, $\theta \in \Theta \subseteq \mathbb{R}^k$. Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be a sequence of estimators.

Weak consistency. $\hat{\theta}_n$ is weakly consistent for θ if $\hat{\theta}_n \xrightarrow{p} \theta$ under P_θ for all $\theta \in \Theta$.

Strong consistency. $\hat{\theta}_n$ is strongly consistent for θ if $\hat{\theta}_n \xrightarrow{a.s.} \theta$ under P_θ for all $\theta \in \Theta$.

Fisher consistency. A functional $T(F)$ is Fisher consistent for θ if $T(F_\theta) = \theta$ for all $\theta \in \Theta$. Fisher consistency is a population-level property; combined with convergence of the plug-in estimator $T(\hat{F}_n)$ to $T(F)$, it yields statistical consistency.

Intuition

Consistency says that with enough data, the estimator gets the answer right. It does not say how quickly or how precisely — that is the role of asymptotic normality (Node 7.3). An inconsistent estimator is fundamentally broken: no amount of data can rescue it. Consistency is therefore a necessary filter, not a measure of quality.

The conceptual mechanism is always the same: the sample criterion function converges to a population criterion function, and the population criterion has a unique optimizer at the truth. Identifiability ensures the truth is distinguishable; the law of large numbers drives the convergence.

Structural Role

Consistency is the first-order asymptotic property of an estimator. It guarantees that the sampling distribution of $\hat{\theta}_n$ concentrates at the true parameter as $n \to \infty$. It is a necessary but not sufficient condition for good statistical performance: it does not control the rate of convergence, the shape of the sampling distribution, or finite-sample behavior. Rate and distributional shape are characterized in Node 7.3. Consistency of M-estimators is studied in the general framework of Node 7.5.

Mathematical Development

Sufficient condition via moments. If $\mathbb{E}_\theta[\hat{\theta}_n] \to \theta$ and $\mathrm{Var}_\theta(\hat{\theta}_n) \to 0$, then $\hat{\theta}_n \xrightarrow{p} \theta$ by Chebyshev's inequality: $\mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) \leq \mathrm{Var}(\hat{\theta}_n)/\varepsilon^2 \to 0$. This is the simplest route to consistency but requires both vanishing bias and vanishing variance.

Consistency of the MLE: Wald's proof.

Let $M_n(\theta') = n^{-1}\ell(\theta'; X_1, \ldots, X_n) = n^{-1}\sum_{i=1}^n \log p_{\theta'}(X_i)$ be the normalized log-likelihood. Define the population criterion $M(\theta') = \mathbb{E}_\theta[\log p_{\theta'}(X)]$.

Step 1: Unique maximum. By the information inequality (Gibbs' inequality / KL divergence nonnegativity),

$$M(\theta') = \mathbb{E}_\theta[\log p_{\theta'}(X)] \leq \mathbb{E}_\theta[\log p_\theta(X)] = M(\theta),$$

with equality if and only if $p_{\theta'} = p_\theta$ P_θ-a.e. Under identifiability ($p_{\theta'} = p_\theta$ a.e. implies $\theta' = \theta$), the maximum of $M(\theta')$ over Θ is uniquely attained at $\theta' = \theta$.

Step 2: Uniform convergence. Assume $\sup_{\theta' \in \Theta} |M_n(\theta') - M(\theta')| \xrightarrow{p} 0$. This holds under a uniform law of large numbers (ULLN), which requires either compactness of Θ and continuity of $\theta' \mapsto \log p_{\theta'}(x)$, or a Glivenko-Cantelli argument for the class $\{\log p_{\theta'} : \theta' \in \Theta\}$.

Step 3: Argmax continuous mapping theorem. If $M_n \to M$ uniformly and M has a unique maximizer θ, then $\hat{\theta}_n = \arg\max M_n \xrightarrow{p} \theta$.

Proof of step 3. Fix $\varepsilon > 0$. By uniqueness and continuity of M, there exists $\delta > 0$ such that $M(\theta') \leq M(\theta) - \delta$ for all $\|\theta' - \theta\| > \varepsilon$. Under uniform convergence, for large n,

$$M_n(\hat{\theta}_n) \geq M_n(\theta) \geq M(\theta) - \delta/3,$$

and $M_n(\hat{\theta}_n) \leq M(\hat{\theta}_n) + \delta/3$. If $\|\hat{\theta}_n - \theta\| > \varepsilon$, then $M(\hat{\theta}_n) \leq M(\theta) - \delta$, giving $M_n(\hat{\theta}_n) \leq M(\theta) - 2\delta/3$, a contradiction. □

Consistency of the method of moments. Let $\mu_k(\theta) = \mathbb{E}_\theta[X^k]$ and $\hat{\mu}_k = n^{-1} \sum X_i^k$. By the WLLN, $\hat{\mu}_k \xrightarrow{p} \mu_k(\theta)$. If the moment map $g : \theta \mapsto (\mu_1(\theta), \ldots, \mu_r(\theta))$ is continuous and injective, the method-of-moments estimator $\hat{\theta}_n = g^{-1}(\hat{\mu}_1, \ldots, \hat{\mu}_r)$ satisfies $\hat{\theta}_n \xrightarrow{p} \theta$ by the continuous mapping theorem.

Uniform law of large numbers (ULLN). A class of functions $\mathcal{F}$ is Glivenko-Cantelli for P if $\sup_{f \in \mathcal{F}} |n^{-1} \sum f(X_i) - \mathbb{E}[f(X)]| \xrightarrow{a.s.} 0$. Sufficient conditions include: (i) $\mathcal{F}$ has finite VC dimension; (ii) $\mathcal{F}$ is uniformly bounded and parametrized by a compact set with equicontinuity. The ULLN is the key technical ingredient linking pointwise LLN convergence to the uniform convergence needed for consistency of extremum estimators.

Rate of convergence and consistency. Consistency alone does not specify a rate. The sample mean converges at rate $n^{-1/2}$ (by the CLT). The MLE for the uniform upper bound θ converges at rate n^{-1} (superefficient). Nonparametric density estimators converge at rate $n^{-2/(4+p)}$ (slower than parametric). These distinctions motivate the refined study of asymptotic normality in Node 7.3.

Consistency under model misspecification. When the true distribution P_0 does not belong to the model $\{P_\theta : \theta \in \Theta\}$, the MLE converges to the pseudo-true parameter

$$\theta^* = \arg\min_{\theta \in \Theta} \mathrm{KL}(P_0 \| P_\theta) = \arg\max_{\theta \in \Theta} \mathbb{E}_0[\log p_\theta(X)].$$

This is the KL projection of P_0 onto the model. The proof of consistency carries through verbatim — the ULLN and argmax CMT arguments do not require correct specification — but the limit θ^* is no longer the "true" parameter. This observation is the foundation for the sandwich variance under misspecification (Node 7.3).

Consistency of Bayesian estimators. The posterior mode (MAP estimator) is consistent under the same conditions as the MLE, since the prior contributes $O(1)$ to the log-posterior while the likelihood contributes $O(n)$. More generally, the posterior distribution itself concentrates at the true parameter (posterior consistency), provided the prior assigns positive mass to every KL neighborhood of the truth (Schwartz's theorem). This connects to the Bernstein-von Mises theorem (Node 6.7).

Well-separated maximum condition. An alternative to compactness for the argmax CMT: if $M(\theta)$ has a well-separated maximum, meaning

$$\sup_{\|\theta'-\theta\|>\varepsilon} M(\theta') < M(\theta) \quad \forall \varepsilon > 0,$$

then uniform convergence $\sup_{\theta'} |M_n(\theta') - M(\theta')| \xrightarrow{p} 0$ implies $\hat{\theta}_n \xrightarrow{p} \theta$. This condition holds when M is strictly concave or when the KL divergence $\mathrm{KL}(P_\theta \| P_{\theta'}) > 0$ for $\theta' \neq \theta$ (identifiability) and M is continuous.

Consistency and efficiency are independent. The sample mean and the sample median are both consistent for the center of a symmetric distribution. However, their rates differ: under normality, the mean converges at rate σ^2/n while the median converges at rate $\pi\sigma^2/(2n)$. Consistency is a qualitative property (yes/no); efficiency is quantitative (how fast).

Consistency of the bootstrap. The bootstrap is consistent if the bootstrap distribution of $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ converges to the same limit as $\sqrt{n}(\hat{\theta}_n - \theta)$. This holds under the conditions for asymptotic normality of $\hat{\theta}_n$, providing nonparametric confidence intervals that are asymptotically correct.

Worked Examples

Example 1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(\lambda)$. The MLE is $\hat{\lambda}_n = 1/\bar{X}_n$. By the SLLN, $\bar{X}_n \xrightarrow{a.s.} 1/\lambda$. By the continuous mapping theorem with $g(x) = 1/x$, $\hat{\lambda}_n \xrightarrow{a.s.} \lambda$. The MLE is strongly consistent.

Example 2. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(0, \theta)$. The MLE is $\hat{\theta}_n = X_{(n)} = \max_i X_i$. We have $\mathbb{P}(X_{(n)} \leq t) = (t/\theta)^n$ for $0 \leq t \leq \theta$. Thus $\mathbb{P}(\theta - X_{(n)} > \varepsilon) = (1 - \varepsilon/\theta)^n \to 0$ for any $\varepsilon > 0$, so $\hat{\theta}_n \xrightarrow{p} \theta$. In fact $\sum_n \mathbb{P}(\theta - X_{(n)} > \varepsilon) < \infty$, so by Borel-Cantelli, $X_{(n)} \xrightarrow{a.s.} \theta$. Note that $\hat{\theta}_n$ is biased ($\mathbb{E}[X_{(n)}] = n\theta/(n+1)$) but consistent — consistency does not require unbiasedness.

Example 3 (Inconsistency). Let $\hat{\theta}_n = X_1$ regardless of n. Then $\hat{\theta}_n$ is unbiased but $\mathrm{Var}(\hat{\theta}_n) = \mathrm{Var}(X_1) \not\to 0$. This estimator is not consistent: unbiasedness alone is insufficient.

Example 4 (Method of moments). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Gamma}(\alpha, \beta)$ with $\mathbb{E}[X] = \alpha/\beta$ and $\mathrm{Var}(X) = \alpha/\beta^2$. The moment equations are $\hat{\mu}_1 = \hat{\alpha}/\hat{\beta}$ and $\hat{\mu}_2 - \hat{\mu}_1^2 = \hat{\alpha}/\hat{\beta}^2$. Solving: $\hat{\beta} = \hat{\mu}_1/(\hat{\mu}_2 - \hat{\mu}_1^2)$ and $\hat{\alpha} = \hat{\mu}_1^2/(\hat{\mu}_2 - \hat{\mu}_1^2)$. By the WLLN, $\hat{\mu}_k \xrightarrow{p} \mu_k$, and by the continuous mapping theorem, $(\hat{\alpha}, \hat{\beta}) \xrightarrow{p} (\alpha, \beta)$ since the moment map is continuously invertible on $(0, \infty)^2$.

Example 5 (Non-compact parameter space). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ with $\Theta = \mathbb{R}$. The MLE $\hat{\mu} = \bar{X}_n$ is consistent by the WLLN. Here the ULLN holds despite non-compact Θ because the Gaussian log-likelihood has quadratic curvature that prevents the maximizer from escaping. In contrast, for a mixture model with the number of components as a parameter, unbounded Θ can lead to inconsistency of the MLE when the likelihood is unbounded.

Cross-Links

AI. Consistency of empirical risk minimizers (ERM) is the learning-theoretic analogue: under finite VC dimension, the empirical risk minimizer converges to the population risk minimizer. The ULLN over function classes is precisely the Glivenko-Cantelli condition used in both settings.

Economics. Consistency of GMM estimators in econometrics follows the same argmax/argmin continuous mapping theorem structure, with the GMM objective replacing the log-likelihood.

Linear Algebra. The uniform convergence condition can be expressed as operator norm convergence of the empirical measure operator to the population operator in the space of bounded functions on Θ.

Quantum Mechanics. Consistency of quantum state tomography estimators — the reconstructed density matrix converges to the true state — requires analogous identifiability (informationally complete measurements) and ULLN conditions on the measurement statistics.

Structural Compression

Invariant. $\hat{\theta}_n \xrightarrow{p} \theta$: the sampling distribution of $\hat{\theta}_n$ concentrates at θ as $n \to \infty$.

Mechanism. The WLLN drives sample averages to population means; under identifiability, the population criterion is uniquely maximized at the truth; the argmax continuous mapping theorem transfers uniform convergence of the criterion to convergence of the optimizer.

Cross-System Echo. Consistency of the MLE corresponds to the convergence of empirical risk minimizers to the population risk minimizer in statistical learning theory; the structural parallel is: identifiability $\leftrightarrow$ realizability, ULLN $\leftrightarrow$ Glivenko-Cantelli, argmax CMT $\leftrightarrow$ fundamental theorem of statistical learning.

Exercises

1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Prove that $(\bar{X}_n, S_n^2)$ is strongly consistent for (μ, σ^2).
2. Show that $\hat{\theta}_n = \bar{X}_n + 1/n$ is consistent for μ even though it is biased for every n.
3. Let $\hat{\theta}_n$ be consistent for θ and let g be continuous at θ. Prove that $g(\hat{\theta}_n)$ is consistent for $g(\theta)$.
4. Give an example of a strongly consistent estimator that is not unbiased for any finite n.
5. Verify the ULLN for the family $\{\log p_\lambda : \lambda > 0\}$ where $p_\lambda(x) = \lambda e^{-\lambda x}$ on $x > 0$, by checking equicontinuity on compact subsets of $(0, \infty)$.
6. Suppose Θ is not compact. Construct an example where the MLE is inconsistent because the argmax escapes to infinity, and explain which step of Wald's proof fails.
7. Argue, structurally, why identifiability is necessary for consistency of the MLE, connecting the information inequality to the KL divergence and explaining what fails in the proof when two distinct parameters yield the same distribution.

Asymptotic Normality

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 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 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. 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.

Delta Method

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 7 — Asymptotic Theory. Node 7.4 — Delta Method.

The delta method transfers asymptotic normality from a parameter to any smooth function of that parameter. It depends on asymptotic normality (Node 7.3) and enables the general M-estimation framework (Node 7.5) and likelihood asymptotics (Node 7.6). It is the primary tool for constructing confidence intervals for derived quantities.

Formal Definition

Let $\hat{\theta}_n$ be asymptotically normal:

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V(\theta)), \quad \theta \in \Theta \subseteq \mathbb{R}^k.$$

Let $g : \Theta \to \mathbb{R}^m$ be continuously differentiable at θ with Jacobian $\dot{g}(\theta) = \partial g / \partial \theta^\top \in \mathbb{R}^{m \times k}$.

First-order delta method. If $\dot{g}(\theta) \neq 0$ (has full row rank), then

$$\sqrt{n}\,\big(g(\hat{\theta}_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}\big(0,\ \dot{g}(\theta)\, V(\theta)\, \dot{g}(\theta)^\top\big)\,.$$

Second-order delta method. When $g : \mathbb{R} \to \mathbb{R}$ with $g'(\theta) = 0$ and $g''(\theta) \neq 0$:

$$n\big(g(\hat{\theta}_n) - g(\theta)\big) \xrightarrow{d} \frac{g''(\theta)}{2}\, V(\theta)\, \chi^2_1.$$

Intuition

A smooth function of an estimator inherits the estimator's asymptotic Gaussianity because, locally, a smooth function is approximately linear. The Jacobian $\dot{g}(\theta)$ acts as a magnification factor: it stretches or compresses the Gaussian fluctuations of $\hat{\theta}_n$ into Gaussian fluctuations of $g(\hat{\theta}_n)$. When the first derivative vanishes, the linear approximation is trivial and the curvature (second derivative) determines the leading-order behavior, producing a chi-squared rather than Gaussian limit.

Structural Role

The delta method extends asymptotic normality from the parameter θ to any smooth function $g(\theta)$. It is the primary tool for constructing asymptotic confidence intervals and tests for derived quantities — odds ratios, relative risks, variance components, functions of regression coefficients — and is used directly in M-estimation (Node 7.5), variance-stabilizing transformations, and the construction of asymptotic pivots (Node 4.5).

Mathematical Development

Proof of the first-order delta method.

By a first-order Taylor expansion of g around θ:

$$g(\hat{\theta}_n) = g(\theta) + \dot{g}(\theta)(\hat{\theta}_n - \theta) + R_n,$$

where the remainder $R_n = O(\|\hat{\theta}_n - \theta\|^2)$. Since $\hat{\theta}_n - \theta = O_p(n^{-1/2})$, we have $R_n = O_p(n^{-1})$, so $\sqrt{n}R_n = O_p(n^{-1/2}) = o_p(1)$.

Therefore:

$$\sqrt{n}\big(g(\hat{\theta}_n) - g(\theta)\big) = \dot{g}(\theta) \cdot \sqrt{n}(\hat{\theta}_n - \theta) + o_p(1).$$

Since $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} Z \sim \mathcal{N}(0, V(\theta))$ and $\dot{g}(\theta)$ is a fixed matrix, Slutsky's theorem gives

$$\sqrt{n}\big(g(\hat{\theta}_n) - g(\theta)\big) \xrightarrow{d} \dot{g}(\theta)Z \sim \mathcal{N}(0,\, \dot{g}(\theta)V(\theta)\dot{g}(\theta)^\top). \quad \square$$

Scalar case. For $g : \mathbb{R} \to \mathbb{R}$ and scalar θ:

$$\sqrt{n}\big(g(\hat{\theta}_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}\big(0,\ [g'(\theta)]^2 V(\theta)\big)\,.$$

The asymptotic variance is $[g'(\theta)]^2$ times the asymptotic variance of $\hat{\theta}_n$.

Proof of the second-order delta method.

When $g'(\theta) = 0$, expand to second order:

$$g(\hat{\theta}_n) - g(\theta) = \frac{1}{2}g''(\theta)(\hat{\theta}_n - \theta)^2 + O_p(n^{-3/2}).$$

Multiplying by n:

$$n\big(g(\hat{\theta}_n) - g(\theta)\big) = \frac{g''(\theta)}{2}\left[\sqrt{n}(\hat{\theta}_n - \theta)\right]^2 + o_p(1).$$

Since $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V(\theta))$, the continuous mapping theorem gives

$$\left[\sqrt{n}(\hat{\theta}_n - \theta)\right]^2 \xrightarrow{d} V(\theta)\chi_1^2,$$

so $n(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} \frac{g''(\theta)}{2}V(\theta)\chi_1^2$. $\square$

Multivariate second-order delta method. For $g : \mathbb{R}^k \to \mathbb{R}$ with $\nabla g(\theta) = 0$, let $H = \nabla^2 g(\theta)$ be the Hessian. Then

$$n\big(g(\hat{\theta}_n) - g(\theta)\big) \xrightarrow{d} \frac{1}{2}Z^\top H Z, \quad Z \sim \mathcal{N}(0, V(\theta)).$$

This is a quadratic form in a Gaussian, expressible as a weighted sum of independent χ_1^2 variables with weights equal to the eigenvalues of $HV(\theta)$.

Variance-stabilizing transformations. Choose g so that the asymptotic variance of $g(\hat{\theta}_n)$ does not depend on θ. For scalar θ with asymptotic variance $V(\theta)$, solve

$$[g'(\theta)]^2 V(\theta) = c \quad \Longrightarrow \quad g(\theta) = \int \frac{d\theta}{\sqrt{V(\theta)}}.$$

This produces an asymptotically pivotal quantity: $\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, c)$ with c independent of θ.

Application to the MLE. For the MLE with $V(\theta) = I(\theta)^{-1}$:

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

Worked Examples

Example 1 (Log-odds / logit). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ with MLE $\hat{p} = \bar{X}_n$ and $V(p) = p(1-p)$. Consider the log-odds $g(p) = \log(p/(1-p))$. Then $g'(p) = 1/(p(1-p))$, so

$$[g'(p)]^2 V(p) = \frac{1}{p^2(1-p)^2} \cdot p(1-p) = \frac{1}{p(1-p)}.$$

Thus $\sqrt{n}(\text{logit}(\hat{p}) - \text{logit}(p)) \xrightarrow{d} \mathcal{N}(0, 1/(p(1-p)))$. A confidence interval for $\text{logit}(p)$ is $\text{logit}(\hat{p}) \pm z_{\alpha/2}/\sqrt{n\hat{p}(1-\hat{p})}$.

Example 2 (Fisher z-transform). Let r be the sample correlation coefficient estimating ρ. For bivariate normal data, $\sqrt{n}(r-\rho)$ is asymptotically $\mathcal{N}(0, (1-\rho^2)^2)$. The Fisher z-transform $g(r) = \frac{1}{2}\log\frac{1+r}{1-r} = \text{arctanh}(r)$ satisfies $g'(\rho) = 1/(1-\rho^2)$, giving

$$[g'(\rho)]^2 (1-\rho^2)^2 = 1.$$

This is a variance-stabilizing transformation: $\sqrt{n}(\text{arctanh}(r) - \text{arctanh}(\rho)) \xrightarrow{d} \mathcal{N}(0, 1)$, with asymptotic variance independent of ρ. The standard approximation uses $n-3$ in place of n for finite-sample correction.

Example 3 (Ratio of means). Let (X_i, Y_i) be i.i.d. with $\mathbb{E}[X] = \mu_X$, $\mathbb{E}[Y] = \mu_Y \neq 0$, and covariance matrix Σ. The ratio $g(\mu_X, \mu_Y) = \mu_X/\mu_Y$ has gradient $\nabla g = (1/\mu_Y, -\mu_X/\mu_Y^2)^\top$. By the delta method,

$$\sqrt{n}\left(\frac{\bar{X}}{\bar{Y}} - \frac{\mu_X}{\mu_Y}\right) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{\sigma_X^2}{\mu_Y^2} - \frac{2\mu_X\sigma_{XY}}{\mu_Y^3} + \frac{\mu_X^2\sigma_Y^2}{\mu_Y^4}\right).$$

Cross-Links

AI. The delta method underlies the Laplace approximation for posterior predictive uncertainty: the variance of a prediction $f(\theta)$ is approximated by $\nabla f(\hat{\theta})^\top \hat{V} \nabla f(\hat{\theta})$, a direct application of the first-order delta method to neural network outputs.

Economics. In econometrics, the delta method is the standard tool for inference on nonlinear functions of regression coefficients, such as elasticities, marginal effects, and impulse response functions.

Linear Algebra. The delta method is the pushforward of a Gaussian measure under a smooth map: the Jacobian acts as a linear map between tangent spaces, transforming the Gaussian by the standard formula for affine transformations of normals.

Quantum Mechanics. Error propagation in quantum parameter estimation uses the same linearization: the variance of a derived quantity $g(\theta)$ under the quantum Cramer-Rao bound is $[g'(\theta)]^2/(n \cdot F_Q(\theta))$, where F_Q is the quantum Fisher information.

Structural Compression

Invariant. Smooth functions of asymptotically normal estimators are asymptotically normal, with variance determined by the Jacobian of the transformation.

Mechanism. First-order Taylor linearization reduces $g(\hat{\theta}_n) - g(\theta)$ to a linear function of $\hat{\theta}_n - \theta$; the linear map transforms the Gaussian limit by the Jacobian. When the Jacobian vanishes, the second-order term dominates, producing a chi-squared limit.

Cross-System Echo. The delta method is the statistical instance of error propagation in measurement theory. In differential geometry, it is the pushforward of a distribution under a smooth map. The variance-stabilizing transformation is the unique reparametrization that makes the Fisher information metric flat.

Exercises

1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$. Use the delta method to find the asymptotic distribution of $g(\bar{X}_n) = 1/\bar{X}_n$ (the MLE of λ).

2. For the Poisson MLE $\hat{\lambda} = \bar{X}_n$, apply the delta method to find the asymptotic distribution of $\sqrt{\hat{\lambda}}$. Verify that $g(\lambda) = \sqrt{\lambda}$ is a variance-stabilizing transformation.

3. Prove the multivariate delta method: if $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V)$ in $\mathbb{R}^k$ and $g : \mathbb{R}^k \to \mathbb{R}^m$ is differentiable, derive the asymptotic distribution of $g(\hat{\theta}_n)$.

4. Let $\hat{p}_1$ and $\hat{p}_2$ be independent sample proportions from n_1 and n_2 observations. Use the delta method to derive the asymptotic distribution of the log-relative-risk $\log(\hat{p}_1/\hat{p}_2)$.

5. Apply the second-order delta method to $g(\theta) = (\theta - \mu)^2$ with $\hat{\theta}_n = \bar{X}_n$ and $\theta = \mu$, and identify the limiting distribution.

6. Find the variance-stabilizing transformation for the binomial proportion $\hat{p}$ (with $V(p) = p(1-p)$), and show it is $g(p) = \arcsin(\sqrt{p})$.

7. Argue, structurally, why the delta method breaks down for non-differentiable functions g, giving a specific example where $g(\hat{\theta}_n)$ is asymptotically normal but with a different variance formula, and explaining what replaces the Jacobian.

M-Estimation

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 7 — Asymptotic Theory. Node 7.5 — M-Estimation.

M-estimation provides a unified asymptotic framework that encompasses MLE, method of moments, robust estimators, and regression estimators as special cases. This node depends on consistency (Node 7.2), asymptotic normality (Node 7.3), the delta method (Node 7.4), and the likelihood framework (Node 3.2), and enables the likelihood asymptotics of Node 7.6.

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_0$.

M-estimator. An M-estimator maximizes a sample objective:

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} m(X_i, \theta).$$

Z-estimator. A Z-estimator (estimating equations estimator) $\hat{\theta}_n$ solves

$$\frac{1}{n} \sum_{i=1}^{n} \psi(X_i, \hat{\theta}_n) = 0,$$

where $\psi(x, \theta) : \mathcal{X} \times \Theta \to \mathbb{R}^k$ is the estimating function.

If $m(x, \theta)$ is differentiable in θ, the M-estimator solves the Z-estimating equations with $\psi(x, \theta) = \nabla_\theta m(x, \theta)$. The MLE is the M-estimator with $m(x, \theta) = \log p_\theta(x)$.

Intuition

M-estimation abstracts away the specific form of the objective function and asks: under what conditions does a sample optimizer converge to a population optimizer, and what is its limiting distribution? The answer separates into two independent components — consistency (the optimizer of the sample criterion converges to the optimizer of the population criterion) and asymptotic normality (the fluctuations around the limit are Gaussian with a specific covariance). The sandwich covariance formula captures the mismatch between the curvature of the objective and the variability of the estimating function, which coincide under correct model specification but diverge under misspecification.

Structural Role

M-estimation unifies MLE, method of moments, robust estimators, and regression estimators under a single asymptotic framework. It decouples the consistency argument (uniform law of large numbers applied to the criterion) from the normality argument (CLT applied to the estimating function at the solution), making both components verifiable independently. The sandwich covariance provides valid inference under misspecification and is the theoretical basis for robust standard errors in econometrics and biostatistics.

Mathematical Development

Consistency of M-estimators.

Let $M(\theta) = \mathbb{E}_{P_0}[m(X, \theta)]$ and $M_n(\theta) = n^{-1} \sum_{i=1}^{n} m(X_i, \theta)$.

Conditions: 1. $M(\theta)$ has a unique maximizer θ_0. 2. $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$ (uniform convergence). 3. Θ is compact (or the maximizer is well-separated).

Conclusion: $\hat{\theta}_n \xrightarrow{p} \theta_0$.

The proof is the same argmax continuous mapping theorem argument as in Node 7.2: uniform convergence of the criterion plus unique maximization of the limit implies convergence of the maximizers.

Consistency of Z-estimators.

Let $\Psi(\theta) = \mathbb{E}_{P_0}[\psi(X, \theta)]$. If $\Psi(\theta_0) = 0$ has a unique solution, $n^{-1} \sum_i \psi(X_i, \theta) \xrightarrow{p} \Psi(\theta)$ uniformly, and Ψ is continuous, then $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Asymptotic normality of Z-estimators.

Define the Jacobian and variance matrices:

$$A(\theta_0) = \mathbb{E}[\nabla_\theta \psi(X, \theta_0)], \qquad B(\theta_0) = \mathbb{E}[\psi(X, \theta_0)\psi(X, \theta_0)^\top].$$

Regularity conditions: (i) $A(\theta_0)$ is nonsingular; (ii) ψ is differentiable in θ with the derivative satisfying a ULLN; (iii) $B(\theta_0)$ is finite and positive definite.

Theorem. Under these conditions,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\big(0,\ A(\theta_0)^{-1} B(\theta_0) [A(\theta_0)^{-1}]^\top\big)\,.$$

Proof. Taylor-expand the sample estimating equation around θ_0:

$$0 = \frac{1}{n}\sum_{i=1}^{n} \psi(X_i, \hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} \psi(X_i, \theta_0) + \left[\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \psi(X_i, \theta_n^*)\right](\hat{\theta}_n - \theta_0),$$

where θ_n^* lies between $\hat{\theta}_n$ and θ_0. Multiply through by $\sqrt{n}$:

$$0 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi(X_i, \theta_0) + \left[\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \psi(X_i, \theta_n^*)\right]\sqrt{n}(\hat{\theta}_n - \theta_0).$$

By the CLT, $n^{-1/2} \sum_i \psi(X_i, \theta_0) \xrightarrow{d} \mathcal{N}(0, B(\theta_0))$.

By the WLLN and consistency, $n^{-1} \sum_i \nabla_\theta \psi(X_i, \theta_n^*) \xrightarrow{p} A(\theta_0)$.

Solving:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -A(\theta_0)^{-1} \cdot \frac{1}{\sqrt{n}} \sum_i \psi(X_i, \theta_0) + o_p(1).$$

By Slutsky's theorem, the limit is $-A(\theta_0)^{-1} \cdot \mathcal{N}(0, B) = \mathcal{N}(0, A^{-1} B A^{-\top})$. $\square$

Sandwich variance estimator. A consistent estimator of the asymptotic covariance is

$$\widehat{V}_n = \hat{A}_n^{-1} \hat{B}_n [\hat{A}_n^{-1}]^\top,$$

where $\hat{A}_n = n^{-1}\sum_i \nabla_\theta \psi(X_i, \hat{\theta}_n)$ and $\hat{B}_n = n^{-1}\sum_i \psi(X_i, \hat{\theta}_n)\psi(X_i, \hat{\theta}_n)^\top$.

Information matrix equality. Under correct model specification with $\psi = \nabla_\theta \log p_\theta$, the Bartlett identities give $A(\theta_0) = -I(\theta_0)$ and $B(\theta_0) = I(\theta_0)$, so $A^{-1}BA^{-\top} = I(\theta_0)^{-1}$. The sandwich reduces to the inverse Fisher information.

Connection to robust statistics. Huber's M-estimator uses $\psi(x, \theta) = \psi_c(x - \theta)$ where

$$\psi_c(u) = \begin{cases} u & |u| \le c, \\ c \cdot \operatorname{sign}(u) & |u| > c, \end{cases}$$

which corresponds to $m(x, \theta) = \rho_c(x - \theta)$ with $\rho_c(u) = u^2/2$ for $|u| \le c$ and $c|u| - c^2/2$ for $|u| > c$. This estimator is consistent for the location parameter under symmetric distributions and has bounded influence function, providing robustness against outliers at the cost of efficiency: the ARE relative to the sample mean under normality is $[2\Phi(c) - 1 - 2c\phi(c) + 2c^2(1 - \Phi(c))]^{-1} \cdot (2\Phi(c) - 1)^2$.

Connection to empirical risk minimization. In machine learning, the empirical risk $n^{-1}\sum_i L(Y_i, f_\theta(X_i))$ is an M-estimation objective with $m(x, \theta) = -L(y, f_\theta(x))$. Consistency requires the function class to be Glivenko-Cantelli; asymptotic normality requires finite VC dimension or analogous complexity control. The sandwich variance appears in the asymptotic distribution of the ERM solution.

Worked Examples

Example 1 (Huber estimator). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$ with F symmetric about μ. The Huber M-estimator with $c = 1.345$ solves $\sum_i \psi_{1.345}(X_i - \hat{\mu}) = 0$. Under $F = \mathcal{N}(\mu, 1)$: $A = -\mathbb{E}[\psi'_{1.345}(Z)] = -(2\Phi(1.345) - 1) \approx -0.8227$ and $B = \mathbb{E}[\psi_{1.345}(Z)^2] \approx 0.7101$. The asymptotic variance is $B/A^2 \approx 1.049$, giving ARE ≈ 0.953 relative to the sample mean — 95.3% efficiency at the normal with much greater robustness to outliers.

Example 2 (Heteroskedasticity-robust inference). In the linear model $Y_i = x_i^\top \beta + \varepsilon_i$ with $\mathbb{E}[\varepsilon_i | x_i] = 0$ but $\operatorname{Var}(\varepsilon_i | x_i) = \sigma^2(x_i)$ (heteroskedastic), OLS is an M-estimator with $\psi(X_i, \beta) = x_i(Y_i - x_i^\top \beta)$. Here $A = -\mathbb{E}[x_i x_i^\top]$ and $B = \mathbb{E}[\varepsilon_i^2 x_i x_i^\top]$. The sandwich variance $A^{-1}BA^{-\top}$ is the Huber-White heteroskedasticity-consistent covariance estimator, estimated by $(X^\top X)^{-1}(\sum_i \hat{e}_i^2 x_i x_i^\top)(X^\top X)^{-1}$ where $\hat{e}_i$ are OLS residuals.

Cross-Links

AI. Empirical risk minimization is M-estimation; the training loss is the M-objective, and the sandwich variance appears in the asymptotic distribution of learned parameters. Stochastic gradient descent can be analyzed as an approximate Z-estimator solving the sample score equation.

Economics. Generalized Method of Moments (GMM) is a Z-estimation framework with $\psi = g(\theta)^\top W \cdot h(X_i, \theta)$; the sandwich variance is the Eicker-Huber-White robust standard error, standard in applied econometrics.

Linear Algebra. The sandwich form $A^{-1}BA^{-\top}$ is a congruence transformation of B by A^{-1}, the same algebraic operation that governs covariance transformation under linear maps (Node 8.1).

Quantum Mechanics. Quantum state estimation via maximum likelihood on measurement outcomes is an M-estimation problem; the sandwich variance accounts for measurement-model mismatch when the assumed measurement model differs from the true POVM.

Structural Compression

Invariant. Z-estimators solving $\mathbb{E}[\psi(X, \theta_0)] = 0$ are $\sqrt{n}$-consistent and asymptotically normal with sandwich covariance $A^{-1}BA^{-\top}$.

Mechanism. Taylor linearization of the estimating equation at the solution maps the CLT for ψ into a Gaussian limit for $\hat{\theta}_n$ via the inverse Jacobian A^{-1}. The sandwich structure emerges because the variability (B) and the curvature (A) of the estimating equation are distinct objects that coincide only under correct specification.

Cross-System Echo. M-estimation is the frequentist analogue of variational inference (both minimize an objective over parameters). The sandwich variance is the standard covariance correction in generalized estimating equations (GEE) in biostatistics, in heteroskedasticity-robust standard errors in econometrics, and in quasi-likelihood theory.

Exercises

1. Show that the MLE is a special case of M-estimation with $m(x, \theta) = \log p_\theta(x)$, and verify that the sandwich reduces to $I(\theta)^{-1}$ under correct specification.
2. Derive the influence function $\mathrm{IF}(x; T, F) = -A^{-1}\psi(x, \theta_0)$ for a Z-estimator and verify that $\int \mathrm{IF}(x; T, F)\, dF(x) = 0$.
3. For the Huber M-estimator with tuning constant c, compute the breakdown point and the asymptotic efficiency at the normal as functions of c.
4. In the linear regression model with heteroskedastic errors, derive the Huber-White sandwich estimator from the general Z-estimator formula with $\psi_i = x_i(Y_i - x_i^\top \beta)$.
5. Show that the sample median is a Z-estimator with $\psi(x, \theta) = \mathrm{sign}(x - \theta)$. Compute A and B when $X \sim \mathcal{N}(\mu, \sigma^2)$ and derive the asymptotic variance $\pi\sigma^2/(2n)$.
6. Consider empirical risk minimization with the hinge loss $m(x, \theta) = -\max(0, 1 - y \cdot x^\top \theta)$. Explain why the standard M-estimation asymptotics do not apply directly (non-differentiability) and describe modifications needed.
7. Argue, structurally, why the sandwich variance is the correct asymptotic covariance even under model misspecification, while the model-based variance A^{-1} alone is invalid, connecting this to the failure of the information matrix equality.

Likelihood Asymptotics

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_k^2 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 $\Delta_n^\top I(\theta_0)^{-1}\Delta_n \xrightarrow{d} \chi_k^2$. □

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 $\Lambda_n \xrightarrow{d} \chi_q^2$ 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 $\mathbb{P}(\Lambda_n^* \le x) = F_{\chi_k^2}(x) + O(n^{-2})$, 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.