

Lecture 11: Asymptotic normality and the delta method

Lecturer: Yiqiao Zhong[Report an issue](#)*Last update: October 19, 2026*

These notes support STAT 709 at the University of Wisconsin–Madison. Use them alongside the cited textbooks; corrections and questions are welcome.

Priority guide

High Core concepts and calculations to master; directly examined.

Mid Supporting concepts useful in other lectures; in scope for examinations.

Low Enrichment supporting understanding; will not be examined.

Further reading Optional self-study, references, extensions, and historical notes.

A star (*) marks optional material; priority labels give the scope of each block. Lecture-specific exemptions and the optional status of Part VI override the general exam guide.

Mid priority *Overview*

A large-sample distribution turns an estimation error into a usable uncertainty scale. We first learn when transformations preserve convergence, then linearize a smooth estimator, and finally examine where the approximation can fail. The main route is continuous mapping, Slutsky and the **delta method**, followed by examples and plug-in estimation. GMM, small-sample refinements and growing dimensions supply optional context.

Recap from earlier lectures

From Lectures 8–10: weak convergence and estimators

Reminder only: this block introduces no new exam requirements.

- **Weak convergence.** Expectations of every bounded continuous test function characterize convergence in distribution. Used in Section 1.
- **CLT and scalar projections.** The CLT describes centered fluctuations; Cramér–Wold checks vector limits through projections. Used in Section 2.
- **Stochastic orders and consistency.** Consistency gives a probability limit; o_P and O_P keep track of remainder sizes. Used in Section 2.
- **Estimator recipes.** Method of moments and maximum likelihood construct the estimators whose limiting errors we now study. Used in Section 3.

Learning goals

- Apply continuous mapping and Slutsky with their assumptions checked.
- State and prove the delta method and choose its centering and scale.
- Derive asymptotic distributions of moment and plug-in estimators.
- Explain why asymptotic normality does not by itself control tail risk.

High priority Introduction

In classical asymptotic theory, one of the most useful results is asymptotic normality. Generally speaking, if θ is an unknown parameter, and $\hat{\theta}_n$ is a reasonable estimator, then $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma_\theta^2)$ for some variance σ_θ^2 . It is the basis for deriving p -values and confidence intervals for large-sample settings.

1 Convergence of transformations

High priority Convergence of transformations

First, we introduce two useful theorems.

Theorem 1 (Continuous mapping theorem, Thm 1.10 in [5]). *Let $X, X_1, X_2, \dots$ be random vectors in $\mathbb{R}^d$ defined on a common probability space and g be a measurable function from $(\mathbb{R}^d, \mathcal{B}^d)$ to $(\mathbb{R}^l, \mathcal{B}^l)$. Suppose that g is continuous a.s. with respect to μ_X (namely the measure induced by X). Then*

$$(i) \quad X_n \xrightarrow{a.s.} X \text{ implies } g(X_n) \xrightarrow{a.s.} g(X).$$

$$(ii) \quad X_n \xrightarrow{p} X \text{ implies } g(X_n) \xrightarrow{p} g(X).$$

$$(iii) \quad X_n \xrightarrow{d} X \text{ implies } g(X_n) \xrightarrow{d} g(X).$$

Proof requirements. All three conclusions are required. The proofs of parts (i) and (ii) are required; the proof of part (iii) is optional.

Parts (i) and (ii): required. For (i), it is essentially pointwise convergence except for a zero-measure set. Here is a detailed proof. Let

$$A_0 = \{\omega \in \Omega : \lim_n X_n(\omega) = X(\omega)\}, \quad D = \{x \in \mathbb{R}^d : g \text{ is continuous at } x\}$$

By assumptions, $P(A_0^c) = 0$ and $\mu_X(D^c) = P(X^{-1}(D^c)) = 0$. Thus, the probability of the set/event $A = A_0 \cap X^{-1}(D)$ is 1. For every $\omega \in A$, we have pointwise convergence: $\lim_n X_n(\omega) = X(\omega)$ and g is continuous at $X(\omega)$, so $\lim_n g(X_n(\omega)) = g(X(\omega))$.

For (ii), our goal is to show, for any $\varepsilon > 0$, $\lim_n P(\|g(X_n) - g(X)\| > \varepsilon) = 0$. A proof is straightforward if we impose stronger assumptions such as g is uniformly continuous or Lipschitz continuous; in the Lipschitz case $\|g(X_n) - g(X)\| \leq L\|X_n - X\|$ so

$$X_n \xrightarrow{p} X \implies \lim_n P(\|X_n - X\| > \varepsilon/L) = 0 \implies \lim_n P(\|g(X_n) - g(X)\| > \varepsilon) = 0.$$

For the general case, every subsequence of a sequence converging in probability has a further subsequence converging almost surely: choose indices with $\mathbb{P}(\|X_{n_j} - X\| > 2^{-j}) < 2^{-j}$ and apply the first Borel–Cantelli lemma. Part (i) transfers that further subsequence through g . If $g(X_n)$ failed to converge in probability, there would be a subsequence with some fixed error probability bounded away from zero, contradicting this almost-sure further subsequence. This proves (ii). □

Low priority Continuous mapping: proof of part (iii)

Part (iii): optional. For (iii), first we apply Skorokhod’s theorem (or see [2, Theorem 3.2.8]): If $X_n \xrightarrow{d} X$, then there are random vectors $Y, Y_1, Y_2, \dots$ defined on a common probability space such that $\mu_Y = \mu_X$, $\mu_{Y_n} = \mu_{X_n}$, $n = 1, 2, \dots$, and $Y_n \xrightarrow{a.s.} Y$. Then we use Theorem 1 (i) and complete the proof. □

Mid priority Continuous mapping: interpretation

Instructor's Note Continuous transformations carry limits with them

Intuitively, a continuous transformation is a rule that keeps nearby inputs nearby. Once we understand the limit of an estimator, we can reuse it after changing how we report the estimator. For example, $T_n \rightarrow_p \theta$ gives $T_n^2 \rightarrow_p \theta^2$, and $Z_n \Rightarrow N(0, 1)$ gives $Z_n^2 \Rightarrow \chi_1^2$. The transformation need not be linear or invertible. A discontinuity matters only if the limit can land there: $X_n = 1/n \rightarrow 0$, but $g(x) = \mathbf{1}_{\{x > 0\}}$ gives $g(X_n) = 1 \neq g(0)$. The continuous mapping theorem carries the limit through a transformation; the delta method will zoom in to describe the size of the remaining fluctuations. See Bartlett's Stat 210B notes, slides 15–16, for the theorem and transformed-limit examples.

Low priority Practice: Another proof of continuous mapping

Exercise prompt. Use the bounded-continuous-test criterion from Lecture 10 to prove Theorem 1(iii) when g is continuous everywhere. How can the argument be extended when g is continuous only μ_X -almost surely?

See Exercise 9 in Section 7 for the full statement and complete solution.

High priority Slutsky's theorem

Theorem 2 (Slutsky's theorem; Thm. 1.11 of [5]). Let $X, X_1, X_2, \dots, Y_1, Y_2, \dots$ be random variables on a probability space. Suppose that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a fixed real number. Then

(i) $X_n + Y_n \xrightarrow{d} X + c$.

(ii) $Y_n X_n \xrightarrow{d} cX$.

(iii) $X_n/Y_n \xrightarrow{d} X/c$ if $c \neq 0$.

Study guidance. The proofs of Theorem 2 are not required. You are encouraged to derive one or two of its results on your own, then check the remaining results against Shao's textbook proof (Theorem 1.11 in [5]).

In statistics, Y_n is usually a sequence of consistent estimators, so this theorem gives a limiting characterization.

Recall from the previous lecture that $Y_n \xrightarrow{p} c$ is equivalent to $Y_n \xrightarrow{d} c$. So Theorem 2 establishes arithmetic properties about convergence in distribution. It is important that we assume Y_n converges to a constant; otherwise these conclusions do not follow from marginal convergence alone.

If we think of $Z_n = (X_n, Y_n) \in \mathbb{R}^2$ as a two-dimensional random vector, then convergence of $(X_n)_{n \geq 1}$ and $(Y_n)_{n \geq 1}$ only gives us information about the *marginal distribution* of Z_n , so we don't know how X_n and Y_n are *coupled*. Consider an example of perfect negative correlation: $X_n = -Y_n$ and both converge in distribution to $G \sim N(0, 1)$. Then, $X_n + Y_n \equiv 0$ instead of $2G$.

Transforming convergence. Continuous mapping transfers convergence through a map continuous at the limiting random vector almost surely. Slutsky allows a random denominator to be replaced by a nonzero probability limit without assuming independence.

Mid priority Practice: normalization

Exercise prompt. If $\sqrt{n}(T_n - \theta) \Rightarrow N(0, 4)$ and $V_n \rightarrow_p 2$, find the limiting law of $\sqrt{n}(T_n - \theta)/V_n$. Must the two statistics be independent?

See Exercise 1 in Section 7 for the full statement and complete solution.

2 Delta method

High priority *Delta method*

Study requirement. You are required to master both the delta-method theorem and its proof. This is a core theoretical result in mathematical statistics: it explains how a local linear approximation determines an estimator's limiting distribution.

A common scenario in statistics is to build new estimators from existing ones, so it is useful to know how to transfer the asymptotic properties. For example, if $\{X_n\}$ is a sequence of estimators, what can we say about the transformed estimators $g(X_n)$?

Theorem 3 (Delta method). *Let $X_1, X_2, \dots$ and Y be random vectors in $\mathbb{R}^k$ satisfying*

$$a_n(X_n - c) \xrightarrow{d} Y, \quad (1)$$

where $c \in \mathbb{R}^k$ and $\{a_n\}$ is a sequence of positive numbers with $\lim_{n \rightarrow \infty} a_n = \infty$. Let g be a function on $\mathbb{R}^k$ to $\mathbb{R}$. If g is differentiable at c , then

$$a_n[g(X_n) - g(c)] \xrightarrow{d} [\nabla g(c)]^\top Y,$$

where $\nabla g(x) \in \mathbb{R}^k$ denotes the gradient of g at x .

In the special case where g is a linear function, then the result follows from Cramér-Wold theorem from Lecture 10. In general, we want to argue that g can be linearized in a local neighborhood around c .

Proof. Let

$$Z_n = a_n[g(X_n) - g(c)] - a_n[\nabla g(c)]^\top (X_n - c).$$

It suffices to show $Z_n = o_p(1)$, since the conclusion will follow by combining $Z_n = o_p(1)$ and Equation 1 due to Slutsky's theorem.

The differentiability of g at c implies that for any $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that

$$|g(x) - g(c) - [\nabla g(c)]^\top (x - c)| \leq \epsilon \|x - c\| \quad (2)$$

whenever $\|x - c\| < \delta_\epsilon$. Let $\eta > 0$ be any fixed scalar. When $|Z_n| \geq \eta$, then at least one of the two events happens: $\|X_n - c\| \geq \delta_\epsilon$ and $a_n \epsilon \|X_n - c\| \geq \eta$; this is because $\|X_n - c\| < \delta_\epsilon$ implies $|Z_n| \leq a_n \epsilon \|X_n - c\|$ due to Equation 2.

By the union bound, namely $A_0 \subset A_1 \cup A_2 \Rightarrow \mathbb{P}(A_0) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2)$, we obtain

$$\mathbb{P}(|Z_n| \geq \eta) \leq \underbrace{\mathbb{P}(\|X_n - c\| \geq \delta_\epsilon)}_{I_1} + \underbrace{\mathbb{P}(a_n \epsilon \|X_n - c\| \geq \eta/\epsilon)}_{I_2}.$$

By the assumption $\lim_{n \rightarrow \infty} a_n = \infty$ and Slutsky's theorem, we have $X_n \xrightarrow{p} c$. So as $n \rightarrow \infty$,

$$I_1 = \mathbb{P}(\|X_n - c\| \geq \delta_\epsilon) \rightarrow 0$$

For I_2 , the limit is not generally zero for fixed ϵ . Instead $a_n(X_n - c) \Rightarrow Y$ implies tightness. Given $\alpha > 0$, choose M with $\limsup_n \mathbb{P}(a_n \|X_n - c\| > M) < \alpha$, then choose ϵ so that $\eta/\epsilon > M$. With this ϵ fixed, $I_1 \rightarrow 0$, and hence $\limsup_n \mathbb{P}(|Z_n| \geq \eta) \leq \alpha$. Let $\alpha \downarrow 0$. Thus $Z_n = o_p(1)$, completing the proof. $\square$

As a simple consequence, when the asymptotic distribution Y is normal, then a differentiable transformation has a first-order Gaussian limit (possibly degenerate).

Corollary 1 (delta method for asymptotic normality). Assume the weak convergence of Equation 1. If $Y \sim N(0, \Sigma)$, then

$$a_n[g(X_n) - g(c)] \xrightarrow{d} \mathcal{N}(0, [\nabla g(c)]^\top \Sigma \nabla g(c)). \quad (3)$$

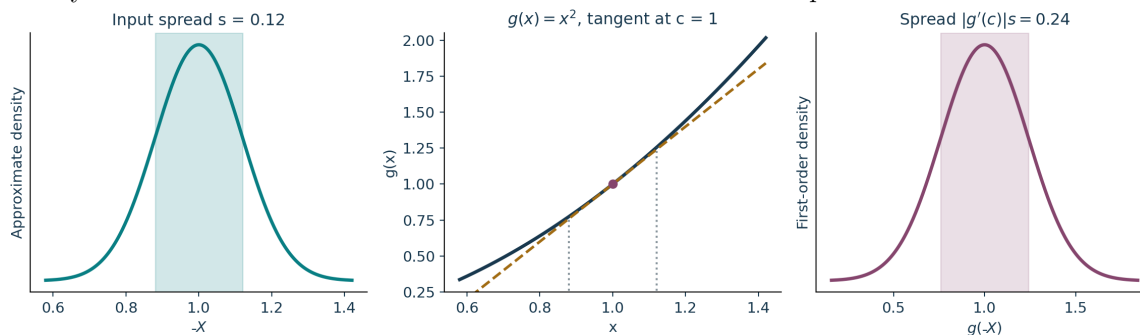
Mid priority The delta method in a picture

Instructor's Note A tangent transports uncertainty

Imagine a small horizontal cloud around c . A nonzero slope tilts and stretches this cloud; the delta method tracks that first-order stretch. At a flat point, the tangent loses the information, so the curvature and a different magnification become relevant.

Picture: linear and quadratic scales. For $g(x) = x^2$ and a displacement h , the exact change is $2ch + h^2$ and its linear approximation is $2ch$. At $c = 1$ and $h = 0.1$, these are 0.21 and 0.20; at $c = 0$ they are 0.01 and 0. The separation between the curves is h^2 . This illustrates the approximation error, not a simulated proof.

First-order uncertainty. The delta method multiplies a limiting fluctuation by the derivative at the probability limit. A zero derivative can erase this first-order limit and require a different scale.



For $g(x) = x^2$ near $c = 1$, an approximately Normal input of standard deviation $s = 0.12$ has first-order output standard deviation $|g'(1)|s = 0.24$. The output density is the delta approximation, not the exact squared-Normal density.

Mid priority Higher-order delta method

Sometimes the first-order Taylor expansion is not sufficient for our purposes. The following example is taken from [5][Example 1.31].

Example 1 (Higher-order delta method). Let $(X_n)_{n \geq 1}$ be a sequence of random variables satisfying $\sqrt{n}(X_n - c) \xrightarrow{d} N(0, 1)$. Consider the function $g(x) = x^2$.

1. If $c \neq 0$, then the (first-order) delta method gives $\sqrt{n}(X_n^2 - c^2) \xrightarrow{d} N(0, 4c^2)$.
2. If $c = 0$, then the (first-order) delta method does not give a useful result. Instead, if we consider second-order Taylor expansion then $nX_n^2 \xrightarrow{d} \chi_1^2$ (see Exercise 5), where $\chi_1^2 \stackrel{d}{=} [N(0, 1)]^2$ is the chi-square distribution.

In case 2, if we insist on using $\sqrt{n}X_n^2$ for the delta method, we get convergence to a degenerate distribution where the variance is zero. This means that we don't get useful distribution information, which motivates us to zoom in and consider a larger factor n instead of $\sqrt{n}$.

Mid priority *Plug-in estimator*

Suppose we find $T_n(X)$ such that $\sqrt{n}(T_n(X) - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. Assume $0 < \sigma^2 < \infty$. If σ^2 is unknown, replace it by a consistent estimator $\hat{\sigma}_n^2$ and take $\hat{\sigma}_n$ to be its nonnegative square root. Then by Slutsky's theorem,

$$\frac{\sqrt{n}(T_n(X) - \theta)}{\hat{\sigma}_n} \xrightarrow{d} \mathcal{N}(0, 1).$$

Mid priority *Practice: Logarithm of a sample mean*

Exercise prompt. Let X_i be i.i.d. positive variables with mean $\mu > 0$ and variance $\sigma^2 < \infty$. Find the limiting distribution of $\sqrt{n}(\log \bar{X} - \log \mu)$.

See Exercise 2 in Section 7 for the full statement and complete solution.

3 Prevalence of asymptotic normality

Mid priority *Prevalence of asymptotic normality*

In classical statistics, we expect that many common estimators have the asymptotic normality property, which forms the core of large-sample theory. For an estimation problem, we want to construct a consistent estimator with asymptotic variance (e.g., the variance in Equation 3) as small as possible, so there is less uncertainty and more accuracy in estimation.

Asymptotic variance. Let $(a_n)_{n \geq 1}$ be a positive sequence, and either $a_n \rightarrow \infty$ or $a_n \rightarrow a > 0$. Assume

$$a_n(T_n(X) - \theta) \xrightarrow{d} Y \quad \text{with } 0 < \mathbb{E}Y^2 < \infty.$$

(i) The asymptotic variance of the estimator $T_n(X)$ is defined to be

$$\frac{\text{Var}(Y)}{a_n^2}.$$

(ii) Let $T'_n(X)$ be another estimator. Asymptotic relative efficiency is defined to be the ratio between two asymptotic variances.

(iii) One important fact is

$$\mathbb{E}Y^2 \leq \liminf_{n \rightarrow \infty} \mathbb{E}[a_n^2(T_n - \theta)^2] \quad (4)$$

If $\mathbb{E}Y = 0$, then $\text{Var}(Y) = \mathbb{E}Y^2$ gives a lower bound on the limiting scaled MSE in Equation 4; this is not a pointwise finite- n inequality. Symmetry about zero with integrability is one way to obtain $\mathbb{E}Y = 0$. It is possible that we have a strict inequality in Equation 4 and a very large gap even exists; see Example 2 below.

Equation 4 is due to the following fact: if $X_n \xrightarrow{d} X$, then for any continuous (but possibly unbounded) function $f : \mathbb{R} \rightarrow \mathbb{R}$ satisfying $\mathbb{E}[|f(X)|] < \infty$ we have $\mathbb{E}[|f(X)|] \leq \liminf_{n \rightarrow \infty} \mathbb{E}[|f(X_n)|]$. (See Exercise 6 for the complete truncation proof.)

High priority *Asymptotic normality does not control tails*

Example 2. Consider i.i.d. Bernoulli random variables $X_1, \dots, X_n$ with $P(X_i = 1) = p$ where $p \in (0, 1)$ is unknown. Both MoM and MLE lead to the sample mean estimator for p : $\hat{p}(X) = \bar{X}$. Now suppose we want to estimate $\theta = p^{-1}$. Consider the plug-in estimator

$$\hat{\theta}_n = \begin{cases} 1/\bar{X}, & \bar{X} > 0, \\ n^n, & \bar{X} = 0. \end{cases}$$

We apply the delta method with transformation $g(x) = 1/x$ to obtain

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \frac{1-p}{p^3})$$

The exceptional event has probability $(1-p)^n \rightarrow 0$, so this finite definition has the displayed delta-method limit. Yet $\mathbb{E}\hat{\theta}_n \geq n^n(1-p)^n \rightarrow \infty$. The original unmodified reciprocal, interpreted as $+\infty$ at $\bar{X} = 0$, has infinite expectation for every n . Both versions preserve the lesson: the limiting distribution does not control moments.

Remark 1. Intuitively, convergence in distribution does not control the tail behavior. That is why the second moment of the limiting distribution is bounded by the liminf of the scaled second moments. To pass from second moments to variances, one must also control the means, since $\text{Var}(Z) = \mathbb{E}Z^2 - (\mathbb{E}Z)^2$. Convergence in distribution alone supplies neither moment convergence.

3.1 Method of moments

High priority *Method of moments*

Recall that we want to estimate a quantity $\theta = \theta(P) \in \mathbb{R}^k$ of an unknown distribution (either parametric or nonparametric). Suppose its k -th moment exists, and denote $\mu_j = \mathbb{E}_{X \sim P} X^j$ for $j \leq k$. Also suppose that we can find a one-one mapping $h : \mathbb{R}^k \rightarrow \mathbb{R}^k : \mu := (\mu_1, \dots, \mu_k) = h(\theta_1, \dots, \theta_k)$ and denote $g = h^{-1}$.

We use i.i.d. observations $X_1, \dots, X_n \in \mathbb{R}$ to construct empirical moments

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n X_i^j$$

If h^{-1} is continuous, then we know that by SLLN and **continuous mapping theorem**¹

$$\hat{\theta} = h^{-1}(\hat{\mu}) \xrightarrow{a.s.} \theta, \quad \text{as } n \rightarrow \infty.$$

To find the asymptotic distribution of MoM estimator $\hat{\theta}$, we make further assumptions: g is differentiable at μ and $\mathbb{E}[|X_1|^{2k}] < \infty$. Then, by an application of the delta method, $\hat{\theta}$ is asymptotically normal, satisfying

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, Dg(\mu)V_\mu Dg(\mu)^\top),$$

where Dg is the Jacobian with entry $\partial g_i / \partial \mu_j$, and $V_\mu \in \mathbb{R}^{k \times k}$ is given by

$$[V_\mu]_{ij} = \mu_{i+j} - \mu_i \mu_j.$$

¹As before, we suppress the dependence on n ; we could write $\hat{\theta}_n$, but normally it is clear from context that we view $n \rightarrow \infty$.

Mid priority *Practice: Moment estimation of Normal variance*

Exercise prompt. For i.i.d. $N(\mu, \sigma^2)$ data, apply the vector delta method to $\hat{v} = n^{-1} \sum X_i^2 - \bar{X}^2$. Find its limiting variance on the $\sqrt{n}$ scale.

See Exercise 3 in Section 7 for the full statement and complete solution.

3.2 Maximum likelihood estimation

Mid priority *Maximum likelihood estimation*

Scope of this discussion. Rigorous results on maximum likelihood estimation will be stated in STAT 710. Here we adopt a non-rigorous, physicist's approach: expand near the true parameter and see how the score fluctuation and likelihood curvature determine the estimator's uncertainty.

Let $P_\theta \in \mathcal{P}$ be an unknown distribution from a parametric family where $\theta \in \mathbb{R}^k$. Suppose we have n i.i.d. observations $X = (X_1, \dots, X_n)$ from P_θ with a common pdf $f_\theta(x)$. The log likelihood function is $\ell_n(\theta; X) = \sum_{i=1}^n \log f_\theta(X_i)$. Suppose that ℓ_n is differentiable with respect to θ . Let the MLE be $\hat{\theta}$, which, if it is an interior differentiable maximizer, satisfies

$$\nabla_\theta \ell_n(\hat{\theta}; X) = 0.$$

Taking consistency of the MLE as a starting point, let us work out the asymptotic distribution of $\hat{\theta}$.

Instructor's Note *Expanding the score like a physicist*

Keep the original mental picture: a consistent maximizer is close to the true parameter, so expand the score nearby and ask which random fluctuation moves its zero. The score is the slope; the information is the stiffness resisting that movement.

Write $s_\theta(x) = \nabla_\theta \log f_\theta(x)$ and $V = \mathbb{E}_\theta s_\theta(X_i) s_\theta(X_i)^\top$. Under a fixed support and justified differentiation under the integral, $\mathbb{E}_\theta s_\theta(X_i) = 0$ and $\mathbb{E}_\theta \nabla_\theta^2 \log f_\theta(X_i) = -V$ (Exercise 7). The log density itself generally does not have mean zero. The multivariate CLT gives

$$n^{-1/2} \nabla \ell_n(\theta) \Rightarrow N(0, V).$$

Assume V is positive definite, $\hat{\theta} \rightarrow_p \theta$ is an interior score root, and the Hessian averaged along the segment from θ to $\hat{\theta}$, divided by n , converges in probability to $-V$. The integral Taylor identity then yields

$$0 = \nabla \ell_n(\theta) + \left\{ \int_0^1 \nabla^2 \ell_n(\theta + t(\hat{\theta} - \theta)) dt \right\} (\hat{\theta} - \theta),$$

so

$$\sqrt{n}(\hat{\theta} - \theta) = V^{-1} n^{-1/2} \nabla \ell_n(\theta) + o_p(1) \Rightarrow N(0, V^{-1}).$$

The sign is positive after moving the score across the negative Hessian. The matrix V is the one-observation **Fisher information**. Under the regularity assumptions of efficiency theory, this variance is optimal among regular estimators²; Lecture 13 studies its unbiased-estimation counterpart. Consistency and a pointwise LLN alone do not justify the Hessian approximation at a random parameter.

Asymptotic likelihood. The score has mean zero under justified differentiation, while its negative expected derivative is information. A consistent interior root and control of the Hessian near the true parameter turn the score CLT into asymptotic normality.

²In local asymptotic theory, regularity requires the estimator to have a stable limiting error law under parameter perturbations of order $n^{-1/2}$. This restriction excludes isolated-point improvements such as superefficiency; a full definition belongs to advanced asymptotic theory.

3.3 Generalized method of moments

Low priority *Generalized method of moments*

Recall that we introduced a way of generalizing moments. Suppose $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P_\theta$, $X_i \in \mathbb{R}^p$, $\theta \in \Theta$, where Θ is compact for simplicity. Let $g : \mathbb{R}^p \times \Theta \rightarrow \mathbb{R}^m$ be a continuous function, such that

$$\mathbb{E}[g(X_i, \theta)] = 0.$$

Write $\Theta \subset \mathbb{R}^k$. The number m of moment equations may exceed the parameter dimension k ; the observation dimension p is separate. Let $W \in \mathbb{R}^{m \times m}$ be a positive definite matrix. Define generalized method of moments (GMM) estimator as

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^n g(X_i, \theta) \right)^T W \left(\frac{1}{n} \sum_{i=1}^n g(X_i, \theta) \right).$$

We also write $\hat{\theta}(W, X)$ or simply $\hat{\theta}(W)$ to emphasize the dependence on W .

Consistency of GMM estimators. Suppose that $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P_{\theta_0}$, $\theta_0 \in \Theta$,

- (1) (Identifiability). $W\mathbb{E}[g(X_i, \theta)] = 0$, if and only if $\theta = \theta_0$.
- (2) (Regularity). $\mathbb{E}[\sup_{\theta \in \Theta} \|g(X_i, \theta)\|] < \infty$.

Then for any fixed W ,

$$\hat{\theta}(W) \xrightarrow{P} \theta_0, \text{ as } n \rightarrow \infty.$$

Remark 2. Condition (1) is necessary, because we need to make sure the population minimizer is unique. Without identifiability, different parameter values can satisfy the same population moment equations. Repeating a single equation cannot identify several parameters unless that equation already distinguishes them.

Proof idea. Denote

$$Q_n(\theta) = \left(\frac{1}{n} \sum_{i=1}^n g(X_i, \theta) \right)^T W \left(\frac{1}{n} \sum_{i=1}^n g(X_i, \theta) \right),$$

$$Q(\theta) = (\mathbb{E}[g(X_i, \theta)])^T W (\mathbb{E}[g(X_i, \theta)]).$$

Use ULLN to show

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{P} 0.$$

The minimizer of $Q(\theta)$ is unique $\theta = \theta_0$. So,

$$\hat{\theta} \xrightarrow{P} \theta_0.$$

(Recall Homework 4 Problem 6). □

Remark 3. Consistency holds if W is replaced by $(W_n)_{n \geq 1}$ with

$$W_n \xrightarrow{P} W.$$

Optimal weighting and asymptotic efficiency. Suppose $g \in C^1(\mathbb{R}^p \times \Theta)$. Denote

$$G = \mathbb{E}[\nabla_{\theta} g(X_i, \theta_0)] \in \mathbb{R}^{m \times k}, \quad \Omega = \mathbb{E}[g(X_i, \theta_0)g(X_i, \theta_0)^T] \in \mathbb{R}^{m \times m}.$$

Asymptotic normality. In addition to conditions (1) and (2), assume that

$$(3) \quad \mathbb{E}[\sup_{\theta \in \Theta} \|\nabla g(X_i, \theta)\|] < \infty.$$

$$(4) \quad G^T W G \text{ is nonsingular, which is satisfied if } G \text{ has full column rank and } W \succ 0.$$

Then,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \underbrace{(G^T W G)^{-1} G^T W \Omega W^T G (G^T W^T G)^{-1}}_{V(W)}).$$

Optimal weighting. Suppose $\Omega \succ 0$. Then

$$V(\Omega^{-1}) \preceq V(W), \quad \forall W \succ 0.$$

Here, $A_1 \succ A_2$ means $A_1 - A_2$ is positive definite, and $A_1 \succeq A_2$ means $A_1 - A_2$ is positive semidefinite. Asymptotic efficiency. We can first derive a consistent estimator W_n of Ω^{-1} , i.e.

$$W_n \xrightarrow{P} \Omega^{-1},$$

and estimate θ_0 by $\hat{\theta}(W_n)$.

Conclusion: $\hat{\theta}(W_n)$ has the smallest asymptotic variance among all possible choices of W . GMM enlarges the scope of MoM, improves the estimation efficiency of MoM and enjoys nice theoretical guarantees.

Conditions behind optimal weighting. For the displayed asymptotic result, also require a consistent measurable minimizer, an interior θ_0 , a neighborhood on which differentiation and expectation interchange, finite covariance Ω , and a uniform LLN for the derivative. Take $G = \mathbb{E}[D_{\theta} g(X_i, \theta_0)] \in \mathbb{R}^{m \times k}$. Compactness and continuity alone do not establish this local asymptotic expansion. For symmetric $W \succ 0$, the sandwich simplifies to $(G^T W G)^{-1} G^T W \Omega W G (G^T W G)^{-1}$.

4 Limitations of the asymptotic approach

Low priority *Limitations of the asymptotic approach*

The asymptotic approach is widely used in statistical inference. Its validity depends on having a large sample size n so that the distribution of $T_n(X)$ is approximated by its asymptotic one.

4.1 Small sample size

Low priority *Small sample size*

If n is not sufficiently large, there is an advantage to doing higher-order Taylor expansion. Let $X_1, \dots, X_n$ be i.i.d. random variables and $\bar{X}$ be the sample mean. Recall that in the proof of CLT, we expand the characteristic function of $\bar{X}$ to the second order. It is possible to consider a higher-order expansion, known as an **Edgeworth expansion**. We state a result from [5][Thm. 1.16].

Let $h : \mathbb{R} \rightarrow \mathbb{R}$ be a smooth function satisfying $h(\mu) = 0$ where $\mu = \mathbb{E}X_1$. Suppose that $\sigma_h^2 = \text{Var}(X_1)[h'(\mu)]^2 > 0$. We know that applying the delta method yields $W_n := \sqrt{n}h(\bar{X})/\sigma_h \xrightarrow{d} N(0, 1)$. Then under some conditions,

$$\sup_{x \in \mathbb{R}} \left| F_{W_n}(x) - \Phi(x) - \sum_{j=1}^m \frac{p_j(x)\Phi'(x)}{n^{j/2}} \right| = o\left(\frac{1}{n^{m/2}}\right)$$

where Φ is the normal CDF and p_j is a certain polynomial of degree at most $3j - 1$.

4.2 Increasing dimensions

Low priority Increasing dimensions

So far, when we establish consistency and asymptotic normality, we assume that the dimensions d of observations are fixed. In modern applications, the dimensions are quite large, from hundreds to millions. It seems inappropriate to imagine that the sample size n goes to infinity while the dimension d is fixed.

There are mainly two approaches to address this issue:

- (i) nonasymptotic approach where we spell out the dependence on dimension;
- (ii) asymptotic approach where we assume an appropriate scaling, e.g., a common regime is proportional growth: $d = d_n$ and $\lim_{n \rightarrow \infty} d_n/n = c$ where c is a constant.

For the asymptotic approach, we need to be clear about what consistency means when the dimension is growing. We need to specify what measurement (e.g., MSE) we make to assess the error of two random vectors of growing dimensions.

Example 3 (Dimension troubles in mean estimation). Suppose that $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, I_d)$. Our goal is to estimate the unknown mean vector μ . Imagine a sequence of estimation problems of growing sample size and dimension under the proportional regime: $d = d_n$ and $\lim_{n \rightarrow \infty} d_n/n = c$, also $\mu_n \in \mathbb{R}^{d_n}$ is of growing dimension (we emphasize the dependence on n in the subscript). Assume that the norm $\rho = \|\mu_n\| > 0$ is fixed.

Let $\hat{\mu}_n = \bar{X}$ be the sample mean estimator for μ_n . It is clear that $\hat{\mu}_n \sim N(\mu_n, \frac{1}{n}I_d)$. To assess the quality of the estimator, we consider **cosine similarity**:

$$\text{sim}(\hat{\mu}_n, \mu_n) = \frac{\langle \hat{\mu}_n, \mu_n \rangle}{\|\hat{\mu}_n\|_2 \|\mu_n\|_2}.$$

A geometric interpretation of cosine similarity is the cosine of the angle between the two vectors. Let $G \sim N(0, I_d)$ be a standard normal random vector. Then

$$\text{sim}(\hat{\mu}_n, \mu_n) = \frac{\|\mu_n\|}{\|\mu_n + \frac{1}{\sqrt{n}}G\|} + \frac{1}{\sqrt{n}} \frac{\langle \mu_n / \|\mu_n\|, G \rangle}{\|\mu_n + \frac{1}{\sqrt{n}}G\|}.$$

To simplify the right-hand side, observe

$$\frac{\|\mu_n + \frac{1}{\sqrt{n}}G\|^2}{\|\mu_n\|^2} = 1 + \frac{2}{\rho^2 \sqrt{n}} \langle \mu_n, G \rangle + \frac{1}{n\rho^2} \|G\|^2$$

By LLN, $\|G\|^2 = (1 + o_p(1))d$. Also, $\langle \mu_n, G \rangle \sim N(0, \rho^2)$. So we have

$$\frac{2}{\rho^2 \sqrt{n}} \langle \mu_n, G \rangle = O_P\left(\frac{1}{\sqrt{n}}\right), \quad \frac{1}{n\rho^2} \|G\|^2 \xrightarrow{p} \frac{c}{\rho^2}.$$

Our derivation leads to

$$\frac{\|\mu_n + \frac{1}{\sqrt{n}}G\|}{\|\mu_n\|} \xrightarrow{p} \left(1 + \frac{c}{\rho^2}\right)^{1/2} \quad \frac{\langle \mu_n / \|\mu_n\|, G \rangle}{\|\mu_n + \frac{1}{\sqrt{n}}G\|} = O_P(1),$$

and thus

$$\text{sim}(\hat{\mu}_n, \mu_n) \xrightarrow{p} \left(1 + \frac{c}{\rho^2}\right)^{-1/2}.$$

If we measure estimation consistency under the cosine similarity, then the sample mean estimator is consistent if and only if $c = 0$.

Used again in Part V: High dimensions: tools and phenomena.

Low priority *Practice: Total and per-coordinate error*

Exercise prompt. In the Gaussian mean example with $d_n/n \rightarrow c$, compute $\mathbb{E} \|\bar{X} - \mu_n\|^2$ and the mean squared error per coordinate. How can their limits differ?

See Exercise 4 in Section 7 for the full statement and complete solution.

5 Further reading

Further reading *Further reading*

The book [1][Section 5.3] has an extensive introduction to the delta method and various statistical applications. Jun Shao's book [5][Section 1.5.3] also provides the theory for delta method.

For GMM, good references include [4], [3] and Wikipedia.

Key ideas

- Continuous mapping transfers convergence through a map continuous at the limit almost surely.
- Slutsky combines distributional limits with probability limits that are constants.
- The delta method linearizes smooth transformations; both its theorem and proof are core tools.
- Asymptotic normality describes central fluctuations and need not control tails.

6 Detailed reference

High priority *Transforming convergence*

Continuous mapping transfers convergence through a map continuous at the limiting random vector almost surely. Slutsky allows a random denominator to be replaced by a nonzero probability limit without assuming independence. Return to the earlier explanation.

Mid priority *First-order uncertainty*

The delta method multiplies a limiting fluctuation by the derivative at the probability limit. A zero derivative can erase this first-order limit and require a different scale. Return to the earlier explanation.

Mid priority *Asymptotic likelihood*

The score has mean zero under justified differentiation, while its negative expected derivative is information. A consistent interior root and control of the Hessian near the true parameter turn the score CLT into asymptotic normality. Return to the earlier explanation.

Looking ahead

Lecture 12 returns to a fixed sample size and asks how good an unbiased estimator can be. It uses conditional expectation and sufficiency from Lectures 5–7, rather than requiring a new asymptotic approximation. Continue to Lecture 12.

You have finished Part III—check yourself. Preview Part IV.

7 Exercises and complete solutions

Mid priority *Normalization*

Exercise 1 (Normalization).

If $\sqrt{n}(T_n - \theta) \Rightarrow N(0, 4)$ and $V_n \rightarrow_p 2$, find the limiting law of $\sqrt{n}(T_n - \theta)/V_n$. Must the two statistics be independent?

Solution.

Slutsky gives $\sqrt{n}(T_n - \theta)/V_n \Rightarrow N(0, 4)/2 = N(0, 1)$. Independence is unnecessary because the denominator converges to a nonzero constant. Define the quotient arbitrarily when $V_n = 0$; that event has probability tending to zero. □

Mid priority *Logarithm of a sample mean*

Exercise 2 (Logarithm of a sample mean).

Let X_i be i.i.d. positive variables with mean $\mu > 0$ and variance $\sigma^2 < \infty$. Find the limiting distribution of $\sqrt{n}(\log \bar{X} - \log \mu)$.

Solution.

The CLT gives $\sqrt{n}(\bar{X} - \mu) \Rightarrow N(0, \sigma^2)$. Since $g(x) = \log x$ is differentiable near μ with $g'(\mu) = 1/\mu$, the delta method gives $N(0, \sigma^2/\mu^2)$. Positivity makes the logarithm defined, and consistency keeps the sample mean near the differentiability point. □

Mid priority *Moment estimation of Normal variance*

Exercise 3 (Moment estimation of Normal variance).

For i.i.d. $N(\mu, \sigma^2)$ data, apply the vector delta method to $\hat{v} = n^{-1} \sum X_i^2 - \bar{X}^2$. Find its limiting variance on the $\sqrt{n}$ scale.

Solution.

Set $g(a, b) = b - a^2$, $Dg = (-2a, 1)$. The relevant moments are $\text{Var}(X) = \sigma^2$, $\text{Cov}(X, X^2) = 2\mu\sigma^2$ and $\text{Var}(X^2) = 4\mu^2\sigma^2 + 2\sigma^4$. The quadratic form is $4\mu^2\sigma^2 - 8\mu^2\sigma^2 + 4\mu^2\sigma^2 + 2\sigma^4 = 2\sigma^4$. Therefore $\sqrt{n}(\hat{v} - \sigma^2) \Rightarrow N(0, 2\sigma^4)$. □

Low priority *Total and per-coordinate error*

Exercise 4 (Total and per-coordinate error).

In the Gaussian mean example with $d_n/n \rightarrow c$, compute $\mathbb{E}\|\bar{X} - \mu_n\|^2$ and the mean squared error per coordinate. How can their limits differ?

Solution.

The error covariance is I_{d_n}/n , so total MSE is its trace $d_n/n \rightarrow c$. Dividing by d_n gives $1/n \rightarrow 0$. Thus each coordinate can be accurately estimated while the total vector error remains nonvanishing. The chosen loss determines what consistency means. $\square$

Mid priority *Second-order delta method*

Exercise 5 (Second-order delta method).

If $\sqrt{n}T_n \Rightarrow N(0, 1)$, prove $nT_n^2 \Rightarrow \chi_1^2$ and $\sqrt{n}T_n^2 \rightarrow_p 0$.

Solution.

Continuous mapping with x^2 gives $(\sqrt{n}T_n)^2 \Rightarrow G^2 \sim \chi_1^2$. This sequence is tight. Multiplication by $n^{-1/2} \rightarrow 0$ and Slutsky yield $\sqrt{n}T_n^2 \rightarrow_p 0$. The second scale captures curvature where the first derivative vanishes. $\square$

Mid priority *A lower bound on limiting moments*

Exercise 6 (A lower bound on limiting moments).

Prove that $Z_n \Rightarrow Z$ implies $\mathbb{E}|f(Z)| \leq \liminf_n \mathbb{E}|f(Z_n)|$ for continuous f .

Solution.

For $M > 0$, $h_M = \min\{|f|, M\}$ is bounded and continuous. Thus $\mathbb{E}h_M(Z) = \lim_n \mathbb{E}h_M(Z_n) \leq \liminf_n \mathbb{E}|f(Z_n)|$. Monotone convergence as $M \uparrow \infty$ gives the assertion, also when the left side is infinite. $\square$

Mid priority *Score and information identities*

Exercise 7 (Score and information identities).

For a positive twice differentiable density on parameter-independent support, justify $\mathbb{E}s_\theta = 0$ and $-\mathbb{E}D^2 \log f_\theta = \mathbb{E}s_\theta s_\theta^\top$, assuming both derivatives may pass under the integral.

Solution.

Since $\int f_\theta = 1$, its first and second parameter derivatives integrate to zero. The identity $s_\theta f_\theta = \nabla f_\theta$ gives the mean-zero score. Also $D^2 \log f_\theta = (D^2 f_\theta)/f_\theta - s_\theta s_\theta^\top$. Integrating against f_θ eliminates the first term and proves the second identity. These are differentiation-under-the-integral arguments, not integration by parts in the data variable. $\square$

Low priority *Why optimal GMM weighting helps*

Exercise 8 (Why optimal GMM weighting helps).

Let $\Omega \succ 0$ and G have full column rank. Put $A_W = (G^\top W G)^{-1} G^\top W$ and $A_* = (G^\top \Omega^{-1} G)^{-1} G^\top \Omega^{-1}$. Prove $A_W \Omega A_W^\top \succeq A_* \Omega A_*^\top$.

Solution.

Both matrices satisfy $A_W G = A_* G = I$. Set $D = A_W - A_*$, so $DG = 0$. Since $\Omega A_*^\top = G(G^\top \Omega^{-1} G)^{-1}$, the cross term $D \Omega A_*^\top$ vanishes. Expanding gives $A_W \Omega A_W^\top - A_* \Omega A_*^\top = D \Omega D^\top \succeq 0$. $\square$

Low priority *Another proof of continuous mapping*

Exercise 9 (Another proof of continuous mapping). Suppose $X_n \Rightarrow X$ in $\mathbb{R}^d$. Recall Lecture 10's bounded-continuous-test criterion: $\mathbb{E}h(X_n) \rightarrow \mathbb{E}h(X)$ for every bounded continuous h .

1. Prove Theorem 1(iii) when $g : \mathbb{R}^d \rightarrow \mathbb{R}^l$ is continuous everywhere by applying this criterion to a composition.
2. Extend the argument to a measurable g whose discontinuity set D_g has $\mathbb{P}(X \in D_g) = 0$. First obtain the closed-set bound $\limsup_n \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$ by approximating $\mathbf{1}_F$ from above with continuous functions. Then compare the closure of $g^{-1}(C)$ with $g^{-1}(C) \cup D_g$ for closed $C \subset \mathbb{R}^l$.

Solution. Everywhere-continuous case. For any bounded continuous $h : \mathbb{R}^l \rightarrow \mathbb{R}$, the composition $h \circ g$ is bounded and continuous. Therefore

$$\mathbb{E} h(g(X_n)) \rightarrow \mathbb{E} h(g(X)),$$

which proves $g(X_n) \Rightarrow g(X)$ by the same test-function criterion. When g is only almost surely continuous, the composition need not be continuous everywhere, so we supply the extension below.

Closed-set bound from test functions. For a nonempty closed set $F \subset \mathbb{R}^d$, let $d(x, F) = \inf_{y \in F} \|x - y\|$ and $h_m(x) = \max\{1 - m d(x, F), 0\}$. Each h_m is bounded and continuous, and $h_m \downarrow \mathbf{1}_F$. Thus, for each fixed m ,

$$\limsup_n \mathbb{P}(X_n \in F) \leq \lim_n \mathbb{E} h_m(X_n) = \mathbb{E} h_m(X).$$

Letting $m \rightarrow \infty$ and using bounded convergence gives the closed-set bound. It is immediate when F is empty.

Continuity outside a null set. Fix a closed $C \subset \mathbb{R}^l$ and put $A = g^{-1}(C)$. If $x \in \overline{A}$ and g is continuous at x , choose $x_j \in A$ with $x_j \rightarrow x$. Since $g(x_j) \in C$ and C is closed, $g(x) \in C$. Consequently

$$\overline{A} \subset A \cup D_g.$$

It follows that

$$\begin{aligned} \limsup_n \mathbb{P}(g(X_n) \in C) &\leq \limsup_n \mathbb{P}(X_n \in \overline{A}) \\ &\leq \mathbb{P}(X \in \overline{A}) \leq \mathbb{P}(g(X) \in C). \end{aligned}$$

The final inequality uses $\mathbb{P}(X \in D_g) = 0$.

Recovering convergence in distribution. Write $Z_n = g(X_n)$ and $Z = g(X)$. Taking complements of the closed-set inequality gives $\liminf_n \mathbb{P}(Z_n \in O) \geq \mathbb{P}(Z \in O)$ for every open O . At a continuity point a of the joint CDF F_Z , use the open and closed lower orthants

$$O_a = \{z : z_j < a_j \text{ for every } j\}, \quad C_a = \{z : z_j \leq a_j \text{ for every } j\}.$$

Continuity gives $\mathbb{P}(Z \in O_a) = \lim_{\epsilon \downarrow 0} F_Z(a - \epsilon \mathbf{1}) = F_Z(a) = \mathbb{P}(Z \in C_a)$. Hence

$$F_Z(a) \leq \liminf_n F_{Z_n}(a) \leq \limsup_n F_{Z_n}(a) \leq F_Z(a).$$

Thus $F_{Z_n}(a) \rightarrow F_Z(a)$ at every continuity point, proving part (iii). This is the closed-set (Portmanteau) route; see also Bartlett, slide 11, for the equivalent weak-convergence criteria. $\square$

References

- [1] Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics, Volumes I–II*. Chapman & Hall/CRC, 2015.
- [2] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 5th edition, 2019.

- [3] A.R. Hall. *Generalized Method of Moments*. Advanced texts in econometrics. Oxford University Press, 2005.
- [4] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.
- [5] Jun Shao. *Mathematical Statistics*. Springer, 2nd edition, 2003.

What changed

October 19, 2026: Corrected notation, wording and cross-references after the course-wide errata review; refreshed the linked HTML and PDF editions.

October 14, 2026. Added issue-reporting links to the webpage and PDF. The mathematics and priority assignments are unchanged in this update.