

Ivan A. Canay

Large Sample Theory

| *Econ 480-1 :: Ver. October 9, 2026*



Contents

Foreword	xi
I Foundations	1
1 Conditional Expectation	3
1.1 What Does Knowing X Tell Us About Y ?	3
1.2 Some Vocabulary	4
1.3 The Definition	5
1.4 Working with the Definition	6
1.5 Conditional Expectation is the Best Predictor	10
1.6 Key Concepts	13
1.7 Concluding Remarks	13
1.8 Problems	14
II Convergence and Limit Theorems	17
2 Inequalities and the LLN	19
2.1 Three Prototypical Problems	19
2.2 Bounding Probabilities by Moments	20
2.3 Concentration: Exponential Bounds at Every n	21
2.4 Modes of Convergence	23
2.5 The Law of Large Numbers	28
2.6 Three Intervals for the Coin	30
2.7 Key Concepts	31
2.8 Concluding Remarks	31
2.9 Problems	32
3 The Analog Principle and Uniform Laws	35
3.1 Continuous Mapping	35
3.2 The Analog Principle	37
3.3 Glivenko–Cantelli	39
3.4 Uniform Laws of Large Numbers	42
3.5 Key Concepts	46
3.6 Concluding Remarks	47
3.7 Problems	47

4	Convergence in Distribution	51
4.1	Convergence in Distribution	51
4.2	The Portmanteau Theorem	54
4.3	Continuous Mapping, Asymptotic Equivalence and Slutsky	55
4.4	Orders in Probability	59
4.5	Key Concepts	61
4.6	Concluding Remarks	62
4.7	Problems	62
5	Characteristic Functions and the CLT	65
5.1	Characteristic Functions	65
5.2	The Central Limit Theorem	68
5.3	Studentization	71
5.4	Key Concepts	74
5.5	Concluding Remarks	75
5.6	Problems	75
6	Heterogeneous Data: Triangular Arrays	79
6.1	Averages of Non-Identically Distributed Data	79
6.2	Triangular Arrays	80
6.3	The Lindeberg–Feller Theorem	81
6.4	An Application: Heteroskedastic Sums	84
6.5	Key Concepts	88
6.6	Concluding Remarks	88
6.7	Problems	89
7	The Delta Method and Influence Functions	91
7.1	The Delta Method	91
7.2	Asymptotic Linearity and Influence Functions	95
7.3	Key Concepts	99
7.4	Concluding Remarks	99
7.5	Problems	100
III	Inference	103
8	Asymptotic Inference and Its Limits	105
8.1	What an Asymptotic Statement Promises	105
8.2	How Good Is the Approximation?	106
8.3	Is the Approximation Uniform?	109
8.4	Local Asymptotics	111
8.5	Key Concepts	115
8.6	Concluding Remarks	115
8.7	Problems	116

9 The Case of the Difference in Means	119
9.1 Parameters Built from Conditional Expectations	119
9.2 Three Representations, One Estimator	120
9.3 Inference	122
9.4 Key Concepts	125
9.5 Concluding Remarks	126
9.6 Problems	126
 IV Appendix	 129
A Background Results	131
A.1 Facts from Integration Theory	131
A.2 Radon–Nikodym and the Existence of Conditional Expectation	131
A.3 Jensen’s Inequality	133
A.4 The Cauchy–Schwarz Inequality	134
A.5 Hoeffding’s Lemma	135
A.6 The Continuous Mapping Theorem	135
A.7 The Portmanteau Theorem	137
A.8 Characteristic Function Estimates	138
A.9 The Lindeberg–Feller Theorem	140



Foreword

These notes were prepared for the first half of Econ 480-1 at Northwestern University, the opening course of the PhD sequence in econometrics. The sequence was taught for many years by Joel Horowitz, who has now retired, and it is being rebuilt. During this transition year Federico Bugni and I are splitting the course: I teach the first half, on large sample theory, and Federico teaches the second, on M-estimation.

The material here is technical, and we have made an effort to present it with a modern spin — triangular arrays rather than only i.i.d. sums, uniformity rather than only pointwise asymptotics, influence functions rather than only case-by-case calculations — without giving up the classical foundations those ideas rest on. Each chapter opens by saying why the topic is covered and closes by saying where it will be used, in this course and later in the sequence.

Most of you will not do research in econometric theory. You will, however, spend the next several years reading papers whose central claims rest on arguments of the kind developed here, in whatever field you end up working in. The aim of this half of the course is that you can follow those arguments: see what is being assumed, where the approximation enters, and when it is likely to fail. That seems to us worth doing carefully at least once.



Part I

Foundations



1

Conditional Expectation

Many of the quantities econometricians care about are *conditional expectations*, or functions of them. Almost every procedure for learning such a quantity from data is a *normalized sum*, or functions of it. This chapter is about the first object, and only about that. Chapters 2 through 8 build the machinery for the second, and Chapter 9 returns here with that machinery in hand.

Why we cover this

Conditional expectation is the object almost every parameter in this sequence is built from, and it is worth defining once, carefully, rather than relying on the version you met in a first course — which was almost certainly stated only for discrete X , or only for jointly continuous (Y, X) , and which gave you a formula rather than a definition. The definition below covers both cases and presumes neither a density nor a functional form. It also fixes a habit of thought we will use all quarter: an object is defined by the property that characterizes it, and formulas are consequences.

1.1 What Does Knowing X Tell Us About Y ?

Start with the simplest instance of the question. Let (Y, X) be a random vector with Y real-valued and $X \in \{0, 1\}$, and write $\pi = P\{X = 1\} \in (0, 1)$. Knowing X means knowing which of two groups an observation belongs to, and the natural summary of Y within a group is its population mean in that group. Write

$$\mu_0 := E[Y|X = 0] , \quad \mu_1 := E[Y|X = 1] ,$$

so that the conditional expectation of Y given X is the *random variable* taking the value μ_0 when $X = 0$ and μ_1 when $X = 1$:

$$E[Y|X] = \mu_0(1 - X) + \mu_1 X . \tag{1.1}$$

Three features of this trivial case survive into full generality. First, $E[Y|X]$ is a *random variable*, and specifically a function of X — not a number. This is the most common source of confusion about the object. Second, it is determined by the joint distribution of (Y, X) alone: no model, no assumption about functional form. Third, the difference $\mu_1 - \mu_0$ compares two population means and carries no causal interpretation on its own. Reading it as the effect of X requires an assumption about how X came to take the value it did, and nothing in the joint distribution supplies one.

1.2 Some Vocabulary

The definition in the next section quantifies over Borel sets, and the proofs that follow use three standard facts about integrals.

- A σ -*algebra* on a set Ω is a collection of subsets containing Ω and closed under complements and countable unions. Its elements are the sets to which the probability P assigns a number; sets outside it have no probability, which is why the restriction is needed at all.
- The *Borel σ -algebra* on $\mathbf{R}^k$ is the smallest σ -algebra containing every open set, and a *Borel set* is one of its elements. Virtually every set you will ever write down — intervals, rectangles, half-spaces, and their countable unions and intersections — is Borel. The quantifier “for every Borel set B ” should therefore be read as “for virtually every familiar subset of $\mathbf{R}^k$ ”.
- A function $g : \mathbf{R}^k \rightarrow \mathbf{R}$ is *Borel measurable* if $g^{-1}(B)$ is Borel for every Borel $B \subseteq \mathbf{R}$. Every continuous function, every indicator of a Borel set, and every pointwise limit of such functions is Borel measurable. From here on, “function of X ” means “Borel measurable function of X ”.
- An event N is *null* if $P\{N\} = 0$. Two random variables V and U are *equal almost surely*, written $V = U$ a.s., if $P\{V \neq U\} = 0$. Statements about random variables are almost always statements up to null events, and (1.2) below determines $E[Y|X]$ only in that sense. Any particular random variable satisfying the defining property is called a *version* of the conditional expectation; different versions differ only on a null event, so it does no harm to speak of “the” conditional expectation.

Example 1.1 (Two versions) Let $X \sim \text{Uniform}[0, 1]$ and $Y = 0$. Both $m_1(X) = 0$ and $m_2(X) = \mathbf{1}\{X = 0\}$ are versions of $E[Y | X]$. They differ as functions, since $m_1(0) = 0$ and $m_2(0) = 1$, but $P\{m_1(X) \neq m_2(X)\} = P\{X = 0\} = 0$. Changing a conditional expectation on a null event produces another version of it. ■

Hint 1.1 (Three facts from integration theory) (F1) Simple approximation. Every bounded Borel measurable g is the pointwise limit of simple functions $g_n = \sum_{j \leq J_n} c_{n,j} \mathbf{1}\{\cdot \in B_{n,j}\}$ with $\sup_n \sup_x |g_n(x)| \leq \sup_x |g(x)|$. (F2) Dominated convergence. If $V_n \rightarrow V$ a.s. and $|V_n| \leq U$ a.s. for an integrable U , then $E[V_n] \rightarrow E[V]$. (F3) Monotone convergence. If $0 \leq V_n \uparrow V$ a.s., then $E[V_n] \uparrow E[V]$.

These are Facts A.1–A.3 of Appendix A, where the standing background results of the course are collected.

Cf. Billingsley [1995, §13, p. 182] for measurability and simple approximation; his Theorem 16.2, p. 208, for monotone convergence and Theorem 16.4, p. 209, for dominated convergence. The Cauchy–Schwarz inequality, used twice below, is Fact A.7 of the appendix.

1.3 The Definition

The definition says that $E[Y|X]$ is the function of X that reproduces the expectation of Y on every event determined by X .

Definition 1.1 Let (Y, X) be a random vector with Y taking values in $\mathbf{R}$ and X in $\mathbf{R}^k$, and suppose $E[|Y|] < \infty$. The conditional expectation of Y given X , written $E[Y|X]$, is any random variable $m(X)$ with $E[|m(X)|] < \infty$ satisfying

$$E[(Y - m(X)) \mathbf{1}\{X \in B\}] = 0 \quad \text{for every Borel set } B \subseteq \mathbf{R}^k. \quad (1.2)$$

Equivalently, $E[Y \mathbf{1}\{X \in B\}] = E[m(X) \mathbf{1}\{X \in B\}]$ for every Borel B : on any event that X determines, $m(X)$ carries the same expectation as Y itself.

Such an $m(X)$ exists whenever $E[|Y|] < \infty$, by the Radon–Nikodym theorem; Section A.2 of the appendix gives the construction. It is determined up to null events, which the next lemma establishes.

Lemma 1.1 (Uniqueness) If $m(X)$ and $\tilde{m}(X)$ both satisfy (1.2), then $P\{m(X) = \tilde{m}(X)\} = 1$.

PROOF. Subtracting the two instances of (1.2) gives

$$E[(\tilde{m}(X) - m(X)) \mathbf{1}\{X \in B\}] = 0 \quad \text{for every Borel } B.$$

Take $B = \{x \in \mathbf{R}^k : m(x) > \tilde{m}(x)\}$, which is Borel because m and $\tilde{m}$ are Borel measurable. The random variable $(\tilde{m}(X) - m(X)) \mathbf{1}\{X \in B\}$ is then non-positive everywhere and strictly negative exactly on the event $\{X \in B\}$. A non-positive random variable with expectation zero is zero almost surely, so $P\{X \in B\} = 0$. Exchanging the roles of m and $\tilde{m}$ gives

$P\{\tilde{m}(X) > m(X)\} = 0$, and the two together give the claim. ■

Cf. [Lehmann and Romano \[2005, §2.4, p. 37\]](#) define the same object by their (2.19)–(2.20), and [Billingsley \[1995, §34, p. 445\]](#) by his (34.1). Both obtain uniqueness from the Radon–Nikodym construction rather than directly.

Condition (1.2) looks forbidding because of the quantifier over all Borel sets. It is not, and the binary case of Section 1.1 shows why. Take $B = \{0\}$: since $\mathbf{1}\{X \in \{0\}\} = 1 - X$, the condition reads $E[(Y - m(X))(1 - X)] = 0$, and dividing by $P\{X = 0\} = 1 - \pi$ gives $m(0) = E[Y|X = 0] = \mu_0$. Taking $B = \{1\}$ gives $m(1) = \mu_1$ the same way. On a two-point set the quantified condition therefore contains exactly two pieces of information, and they are the two group means of (1.1). Every other choice of B — there are only four — is implied by these two. The next example is the same computation with more categories.¹

Example 1.2 (Categorical conditioning) Let X take values in a finite set $\{x_1, \dots, x_K\}$ with $\pi_k = P\{X = x_k\} > 0$ for each k , and let $E[|Y|] < \infty$. Take $B = \{x_k\}$ in (1.2). Since $m(X)\mathbf{1}\{X = x_k\} = m(x_k)\mathbf{1}\{X = x_k\}$,

$$E[Y\mathbf{1}\{X = x_k\}] = m(x_k)P\{X = x_k\},$$

so that $m(x_k) = E[Y\mathbf{1}\{X = x_k\}]/\pi_k$. Equivalently, $E[Y|X] = \sum_{k \leq K} E[Y|X = x_k]\mathbf{1}\{X = x_k\}$, the elementary formula. Two things are worth noticing. The formula was *derived* from (1.2) rather than assumed, and it required only the K sets $\{x_k\}$: because every Borel B meets the range of X in a union of these, the remaining instances of (1.2) follow by adding up. When X is not discrete no such finite reduction is available, and the quantifier is doing real work. ■

Cf. [Billingsley \[1995, Example 34.1, p. 446\]](#) carries out this computation for an arbitrary countable partition.

1.4 Working with the Definition

Everything in this chapter follows from (1.2). The next lemma says that the defining identity, stated for indicators, holds against every bounded function of X .

¹Although there are infinitely many Borel sets $B \subseteq \mathbf{R}$, they generate only four distinct events when $X \in \{0, 1\}$: $\emptyset$, $\{X = 0\}$, $\{X = 1\}$, and Ω . The conditions for the first and last are trivial or implied by those for $\{X = 0\}$ and $\{X = 1\}$.

Lemma 1.2 (Extension) *Let $E[|Y|] < \infty$ and $m(X) = E[Y|X]$. Then*

$$E[(Y - m(X))g(X)] = 0 \quad (1.3)$$

for every bounded Borel measurable $g : \mathbf{R}^k \rightarrow \mathbf{R}$.

PROOF. Write $D = Y - m(X)$, so $E[|D|] \leq E[|Y|] + E[|m(X)|] < \infty$. If $g = \sum_{j \leq J} c_j \mathbf{1}\{\cdot \in B_j\}$ is simple, then $E[Dg(X)] = \sum_{j \leq J} c_j E[D \mathbf{1}\{X \in B_j\}] = 0$ by (1.2) and linearity, each term being finite because $|c_j D|$ is integrable.

For general bounded g , say $|g| \leq C$, take simple $g_n \rightarrow g$ pointwise with $|g_n| \leq C$, as provided by (F1) of Hint 1.1. Then $Dg_n(X) \rightarrow Dg(X)$ pointwise and $|Dg_n(X)| \leq C|D|$, which is integrable. Dominated convergence (F2) gives $E[Dg_n(X)] \rightarrow E[Dg(X)]$, and the left-hand side is zero for every n . ■

Lemma 1.2 is used in nearly every argument that follows. The properties of conditional expectation now follow in a few lines each.

Lemma 1.3 (Properties) *Let $E[|Y|] < \infty$ and $E[|U|] < \infty$. Then, provided the expressions involved are defined,*

- P1.** *If $Y = f(X)$ for a Borel measurable f , then $E[Y|X] = f(X)$ a.s.*
- P2.** *$E[Y + U|X] = E[Y|X] + E[U|X]$ a.s.*
- P3.** *$E[f(X)Y|X] = f(X)E[Y|X]$ a.s., for every bounded Borel measurable f .*
- P4.** *$E[E[Y|X]] = E[Y]$.*
- P5.** *If $Y \perp\!\!\!\perp X$ then $E[Y|X] = E[Y]$ a.s.*

PROOF. **P1.** The candidate $m(X) = f(X)$ is integrable by assumption and makes the integrand in (1.2) identically zero. Lemma 1.1 says there is no other.

P2. Let m_Y and m_U denote versions of the two conditional expectations. Then $m_Y(X) + m_U(X)$ is integrable, and for every Borel B ,

$$\begin{aligned} E[(Y + U - m_Y(X) - m_U(X))\mathbf{1}\{X \in B\}] \\ = E[(Y - m_Y(X))\mathbf{1}\{X \in B\}] \\ + E[(U - m_U(X))\mathbf{1}\{X \in B\}] = 0, \end{aligned}$$

so $m_Y(X) + m_U(X)$ satisfies (1.2) for $Y + U$.

P3. We must show $f(X)m(X)$ satisfies (1.2) for the variable $f(X)Y$, where $m = E[Y|X]$. It is integrable because f is bounded. For Borel B , the function $x \mapsto f(x)\mathbf{1}\{x \in B\}$ is bounded and Borel measurable, so Lemma 1.2 applies to it and gives

$$E[(f(X)Y - f(X)m(X))\mathbf{1}\{X \in B\}] = E[(Y - m(X))f(X)\mathbf{1}\{X \in B\}] = 0.$$

P4. Take $B = \mathbf{R}^k$ in (1.2).

P5. The constant $E[Y]$ is integrable, and if $Y \perp\!\!\!\perp X$ then Y and $\mathbf{1}\{X \in B\}$ are independent, so $E[Y\mathbf{1}\{X \in B\}] = E[Y]P\{X \in B\} = E[E[Y]\mathbf{1}\{X \in B\}]$. ■

Cf. [Lehmann and Romano \[2005, Lemma 2.4.1, p. 39\]](#) for P1–P3 and their [Lemma 2.4.2, eq. \(2.21\), p. 39](#), for P4, both stated for a statistic T in place of X . The route through [Lemma 1.2](#) is what makes P3 two lines here.

Remark 1.1 P4 is used more often than the rest combined, and P3 is the one students misapply: it requires f to be a function of the conditioning variable, and nothing else. Both are instances of the same idea — once X is known, any function of X behaves like a constant. Note that boundedness is sufficient but not necessary for P3; the identity remains valid under suitable integrability conditions. ■

The converse of P5 fails, and three conditions get confused in applied work. Say that Y is *mean independent* of X if $E[Y|X] = E[Y]$ a.s. Independence implies mean independence by P5, and mean independence implies that X and Y are uncorrelated whenever both have finite second moments: take $g(X) = X$ in [Lemma 1.4](#) below, so that $E[XY] = E[Xm(X)] = E[X]E[Y]$. The three are therefore strictly nested, each requiring more moments than the last to be well defined, and neither implication reverses; [Problem 1.3](#) asks for the counterexamples.

Example 1.3 (Signal plus noise) Let $Y = g(X) + U$ where g is Borel measurable with $E[|g(X)|] < \infty$ and $E[U|X] = 0$. Then by P2, P1 and the hypothesis,

$$E[Y|X] = E[g(X)|X] + E[U|X] = g(X) .$$

No linearity or smoothness on g was used. [Figure 1.1\(a\)](#) shows a sample from such a model with $g(x) = \sin(1.5x) + 0.35x$: the conditional expectation is the curve, and it is not a line, not monotone, and in no way special. Note also what the argument needs. If we assume only $E[UX] = 0$ rather than $E[U|X] = 0$ — a strictly weaker condition, as [Section 1.5](#) shows —, the conclusion fails, and [Problem 1.2](#) asks you to produce a counterexample. ■

Example 1.4 (Coarse and fine information) Let $h : \mathbf{R}^k \rightarrow \mathbf{R}^d$ be Borel measurable. Knowing $h(X)$ is knowing less than knowing X , and the *tower property* says the coarser conditional expectation can be computed in two stages:

$$E[E[Y|X] \mid h(X)] = E[Y|h(X)] \quad \text{a.s.} \quad (1.4)$$

To see it, write $m(X) = E[Y|X]$ and let ℓ denote a version of $E[m(X)|h(X)]$. For any Borel $B \subseteq \mathbf{R}^d$ the set $h^{-1}(B)$ is Borel, so [\(1.2\)](#) applied to Y with the set $h^{-1}(B)$ gives $E[Y\mathbf{1}\{h(X) \in B\}] = E[m(X)\mathbf{1}\{h(X) \in B\}]$, and [\(1.2\)](#) applied to $m(X)$ with B gives $E[m(X)\mathbf{1}\{h(X) \in B\}] = E[\ell\mathbf{1}\{h(X) \in B\}]$.

Chaining the two shows ℓ satisfies (1.2) for Y and the conditioning variable $h(X)$, which is (1.4). Taking h constant recovers P4.

A die makes the two stages concrete. Let X be the face of a fair die, $Y = \mathbf{1}\{X \geq 4\}$, and $h(X) = \mathbf{1}\{X \text{ is even}\}$. Knowing X determines Y , so $E[Y|X] = \mathbf{1}\{X \geq 4\}$, taking the values 0, 0, 0, 1, 1, 1 across the six faces. Knowing only the parity does not: the odd faces $\{1, 3, 5\}$ carry $(0+0+1)/3 = 1/3$ and the even faces $\{2, 4, 6\}$ carry $(0+1+1)/3 = 2/3$, which is (1.4) evaluated on each of the two sets. Figure 1.1(b) draws it. Note that the two sets h merges are not intervals: coarsening is not the same as grouping neighbouring values, and nothing in (1.4) suggests otherwise. ■

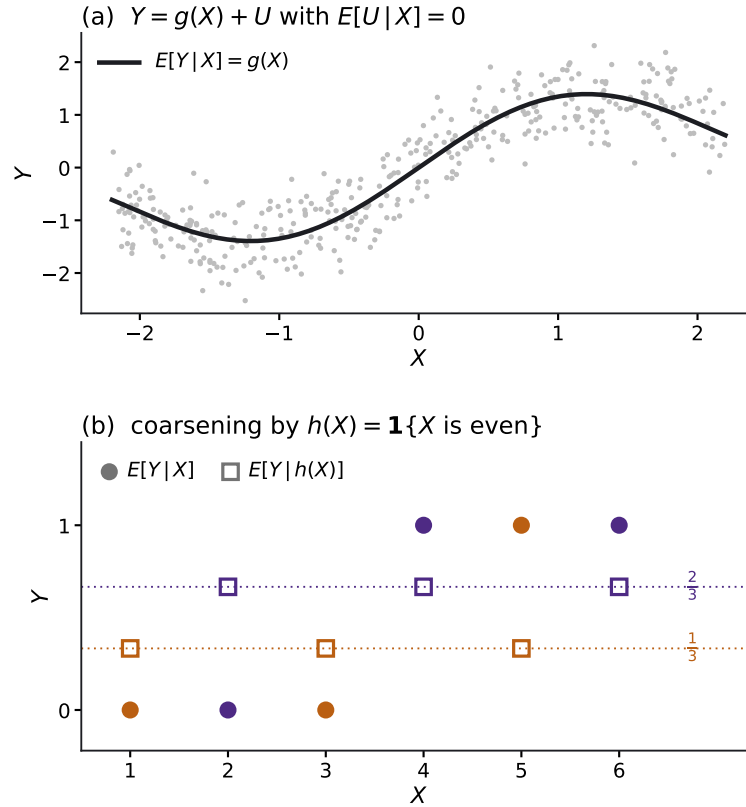


FIGURE 1.1: Two ways of seeing a conditional expectation. In panel (a) the conditional expectation is the curve g , recovered from $Y = g(X) + U$ with $E[U|X] = 0$; nothing about it is linear or monotone. Panel (b) is the die of Example 1.4: filled circles are $E[Y|X] = \mathbf{1}\{X \geq 4\}$ and open squares are $E[Y|h(X)]$ for $h(X) = \mathbf{1}\{X \text{ is even}\}$, colour marking the two sets h merges. Averaging the circles of one colour gives the square of that colour, which is the tower property (1.4); the colours alternate because parity does not group neighbouring faces.

1.5 Conditional Expectation is the Best Predictor

Under a second-moment condition, $E[Y|X]$ is exactly the function of X that predicts Y with the smallest expected squared error.

Throughout the section, assume $E[Y^2] < \infty$ and write $m(X) = E[Y|X]$. Let

$$\mathcal{M} = \{g(X) : g : \mathbf{R}^k \rightarrow \mathbf{R} \text{ Borel measurable and } E[g^2(X)] < \infty\}$$

denote the square-integrable functions of X . Note that $\mathcal{M}$ is a linear space and contains every bounded function of X .

The first thing to establish is that $m(X)$ belongs to $\mathcal{M}$ at all. Definition 1.1 asks only for integrability, so this needs an argument.

Theorem 1.1 (L^2 contraction) *If $E[Y^2] < \infty$ then $E[m^2(X)] \leq E[Y^2]$. In particular $m(X) \in \mathcal{M}$.*

PROOF. Fix $c > 0$ and set $g_c(x) = m(x)\mathbf{1}\{|m(x)| \leq c\}$, a bounded Borel measurable function. Lemma 1.2 gives $E[(Y - m(X))g_c(X)] = 0$, that is,

$$E[Y g_c(X)] = E[m(X)g_c(X)] = E[m^2(X)\mathbf{1}\{|m(X)| \leq c\}] =: a_c,$$

where $a_c \leq c^2 < \infty$. Applying the Cauchy–Schwarz inequality to the left-hand side,

$$a_c = E[Y g_c(X)] \leq (E[Y^2])^{1/2} (E[g_c^2(X)])^{1/2} = (E[Y^2])^{1/2} a_c^{1/2}.$$

If $a_c = 0$ the bound $a_c \leq E[Y^2]$ is immediate; otherwise divide by $a_c^{1/2} > 0$ and square, giving $a_c \leq E[Y^2]$. This holds for every $c > 0$.

Finally, $m(X)$ is finite almost surely because it is integrable, so $m^2(X)\mathbf{1}\{|m(X)| \leq c\} \uparrow m^2(X)$ a.s. as $c \uparrow \infty$. Monotone convergence (F3) gives $a_c \uparrow E[m^2(X)]$, and hence $E[m^2(X)] \leq E[Y^2]$. ■

Cf. The inequality is the quadratic case of conditional Jensen, Billingsley [1995, eq. (34.7)]. The truncation argument above establishes square-integrability rather than presuming it.

Conditioning therefore contracts: predicting Y from X can only shrink second moments. Square-integrability in turn carries Lemma 1.2 beyond bounded functions.

Lemma 1.4 (Extension in L^2) *If $E[Y^2] < \infty$ then $E[(Y - m(X))g(X)] = 0$ for every $g(X) \in \mathcal{M}$.*

PROOF. Write $D = Y - m(X)$ and note $E[D^2] \leq 2E[Y^2] + 2E[m^2(X)] \leq$

$4E[Y^2] < \infty$ by Theorem 1.1. For $g(X) \in \mathcal{M}$ and $c > 0$ put $g_c = g\mathbf{1}\{|g| \leq c\}$, which is bounded, so $E[Dg_c(X)] = 0$ by Lemma 1.2. Then

$$\begin{aligned} |E[Dg(X)]| &= |E[Dg(X)] - E[Dg_c(X)]| = |E[Dg(X)\mathbf{1}\{|g(X)| > c\}]| \\ &\leq (E[D^2])^{1/2} (E[g^2(X)\mathbf{1}\{|g(X)| > c\}])^{1/2}, \end{aligned}$$

by Cauchy–Schwarz. Since $E[g^2(X)] < \infty$ and $\mathbf{1}\{|g(X)| > c\} \rightarrow 0$ a.s. as $c \rightarrow \infty$, dominated convergence (F2) sends the second factor to zero. The left-hand side does not depend on c , so it is zero. ■

Theorem 1.2 (Best predictor) Suppose $E[Y^2] < \infty$. Then

- (i) $E[(Y - m(X))g(X)] = 0$ for every $g(X) \in \mathcal{M}$; and
- (ii) $m(X)$ solves $\min_{g(X) \in \mathcal{M}} E[(Y - g(X))^2]$, and any other solution equals $m(X)$ almost surely.

PROOF. Part (i) is Lemma 1.4. For (ii), let $g(X) \in \mathcal{M}$ and write $Y - g(X) = (Y - m(X)) + (m(X) - g(X))$. Squaring and taking expectations,

$$\begin{aligned} E[(Y - g(X))^2] &= E[(Y - m(X))^2] + 2E[(Y - m(X))(m(X) - g(X))] \\ &\quad + E[(m(X) - g(X))^2], \end{aligned} \tag{1.5}$$

every term being finite because Y , $m(X)$ and $g(X)$ are all square-integrable. Since $\mathcal{M}$ is a linear space, $m(X) - g(X) \in \mathcal{M}$, so the middle term vanishes by (i). Hence

$$E[(Y - g(X))^2] = E[(Y - m(X))^2] + E[(m(X) - g(X))^2] \geq E[(Y - m(X))^2],$$

so $m(X)$ is a solution. If $g(X)$ is also a solution then the two outer terms are equal, forcing $E[(m(X) - g(X))^2] = 0$; a non-negative random variable with zero expectation is zero almost surely. ■

Cf. van der Vaart [1998, Theorem 11.1, p. 153] states the same characterization for projections onto a linear space of square-integrable variables, and observes that it does not by itself assert existence. Our route avoids that issue: Theorems 1.1 and 1.2 establish existence and optimality of the minimizer.

Part (i) of Theorem 1.2 says the prediction error $Y - E[Y|X]$ is uncorrelated with *every* square-integrable function of X , not merely with X itself. That is much stronger than zero correlation with X , and it is used again in Chapter 7 to compute influence functions. It is also what the following decomposition exploits.

Theorem 1.3 (Law of total variance) If $E[Y^2] < \infty$ then

$$\text{Var}[Y] = E[\text{Var}[Y|X]] + \text{Var}[E[Y|X]],$$

where $\text{Var}[Y|X] = E[(Y - E[Y|X])^2|X]$.

PROOF. Write $Y - E[Y] = (Y - m(X)) + (m(X) - E[Y])$ and square. The cross term has expectation zero: the second factor lies in $\mathcal{M}$ by Theorem 1.1, so Theorem 1.2(i) applies. The first squared term has expectation $E[(Y - m(X))^2] = E[E[(Y - m(X))^2|X]] = E[\text{Var}[Y|X]]$ by P4, and the second is $\text{Var}[m(X)]$ because $E[m(X)] = E[Y]$, again by P4. ■

Cf. Billingsley [1995, Problem 34.10(b), p. 456] states this decomposition; his 34.10(a) is the related fact that conditioning on more information reduces dispersion.

In the binary case this reads $\text{Var}[Y] = (1 - \pi)\sigma_0^2 + \pi\sigma_1^2 + \pi(1 - \pi)(\mu_1 - \mu_0)^2$ with $\sigma_x^2 = \text{Var}[Y|X = x]$: variation within groups plus variation between them. It also sharpens Theorem 1.1, the gap there being the expected conditional variance.

Where we will use this

- **In this class.** Chapters 2 through 8 build the tools for learning objects like this one from a finite sample; Chapter 9 returns to conditional expectation and carries one parameter built from it, a difference of two group means, from definition to confidence interval. In the second half of the quarter, estimators are defined by minimizing an expected loss over a restricted class — the same logic as Theorem 1.2, with the negative log-likelihood or a moment criterion in place of squared error.
- **Later in the sequence.** The winter quarter is largely about estimating a conditional expectation when X is high-dimensional and no functional form is credible. The spring quarter is about when a difference of conditional expectations may be read causally.
- **In recent research.** Much of the current machine-learning-flavored work in econometrics estimates conditional expectations as intermediate objects with flexible learners, then corrects for the bias that flexibility introduces. That programme only makes sense because $E[Y|X]$ is defined, as above, without reference to any model.

1.6 Key Concepts

- **A conditional expectation is a prediction, not a cause.** $E[Y|X]$ tells you what to expect when you *observe* a given value of X . It says nothing about what would happen if you *set* X to that value; reading it causally requires assumptions this chapter does not make.
- **Definition.** $E[Y|X]$ is any integrable $m(X)$ with $E[(Y - m(X))\mathbf{1}\{X \in B\}] = 0$ for every Borel B . It exists whenever $E[|Y|] < \infty$, and is unique up to null events.
- **Version.** Any random variable with that property. Two versions agree almost surely.
- **Law of iterated expectations.** $E[E[Y|X]] = E[Y]$, and more generally $E[E[Y|X] | h(X)] = E[Y|h(X)]$.
- **Best predictor.** If $E[Y^2] < \infty$, then $E[Y|X]$ minimizes $E[(Y - g(X))^2]$ over square-integrable functions g of X , and does so uniquely up to null events.

1.7 Concluding Remarks

The only result taken on faith in this chapter is the existence of $E[Y|X]$ under $E[|Y|] < \infty$. It follows from the Radon–Nikodym theorem, and the construction is worth reading once: Billingsley [1995, §34, p. 445], with the Radon–Nikodym theorem itself at his §32, p. 419. A fully general treatment conditions on a σ -algebra rather than on a random vector, which costs nothing extra once the definition is written as (1.2), though we will not need it. Lehmann and Romano [2005, §2.4] give the same definition in the form we use, stated for a statistic.

The order of presentation is deliberate. Definition 1.1 is the definition, and the best-predictor property of Theorem 1.2 is a theorem about it. Several references, van der Vaart [1998] among them, do the reverse and *define* the conditional expectation as a projection. That is elegant but requires $E[Y^2] < \infty$ from the start.

The two references used most heavily in the rest of these notes are Jiang [2010], which is organized around techniques rather than results and is written at exactly the level of this course, and Vershynin [2026] for the concentration inequalities of Chapter 2.

1.8 Problems

Problem 1.1 Let X take the values $\{1, 2, 3\}$ with probabilities $(0.5, 0.3, 0.2)$, and suppose $E[|Y|] < \infty$ with

$$E[Y\mathbf{1}\{X = 1\}] = -0.5, \quad E[Y\mathbf{1}\{X = 2\}] = 0, \quad E[Y\mathbf{1}\{X = 3\}] = 0.4.$$

1. Find a function $m : \{1, 2, 3\} \rightarrow \mathbf{R}$ such that (1.2) holds for each of the three singletons $\{1\}$, $\{2\}$ and $\{3\}$. Use this function to write $E[Y|X]$ as a random variable, and deduce that (1.2) holds for every Borel set.
2. Let $h(x) = \mathbf{1}\{x \geq 2\}$. Compute $E[Y|h(X)]$ from Definition 1.1 and check the tower property (1.4) numerically.

Problem 1.2 Let $Y = g(X) + U$ with $E[|g(X)|] < \infty$.

1. Show that $E[U|X] = 0$ implies $E[Y|X] = g(X)$, and give an example with g nonlinear in which $E[Y|X]$ is not a linear function of X .
2. Now suppose only that $E[UX] = 0$. Construct a joint distribution for (X, U) with $E[|U|] < \infty$ — taking (X, U) bounded is easiest — such that $E[UX] = 0$ but $E[U|X] \neq 0$, and conclude that $E[Y|X] \neq g(X)$ for that distribution. Explain in one sentence which step of part (a) fails.

Problem 1.3 Independence, mean independence, and zero correlation are three strictly nested conditions.

1. Give (Y, X) with $E[Y|X] = E[Y]$ but Y not independent of X . (Try making the conditional variance depend on X .)
2. Give (Y, X) with $\text{Cov}[X, Y] = 0$ but $E[Y|X] \neq E[Y]$. (Try a symmetric X and $Y = X^2$.)
3. State the moment conditions each of the three notions requires in order to be well defined.

Problem 1.4 Let $h : \mathbf{R}^k \rightarrow \mathbf{R}$ be Borel measurable and suppose $E[Y^2] < \infty$.

1. Show that $\mathcal{M}_h = \{g(h(X)) : g : \mathbf{R} \rightarrow \mathbf{R} \text{ Borel measurable and } E[g^2(h(X))] < \infty\} \subseteq \mathcal{M}$.
2. Using Theorem 1.2, conclude that $E[(Y - E[Y|X])^2] \leq E[(Y - E[Y|h(X)])^2]$: coarser information predicts no better.
3. Deduce that $\text{Var}[E[Y|h(X)]] \leq \text{Var}[E[Y|X]]$, and interpret this in terms of Theorem 1.3.

Problem 1.5 The proof of Lemma 1.1 used, without proof, that a non-positive random variable with expectation zero vanishes almost surely. This problem supplies it, and then puts it to work twice. Throughout, V is a random variable with $E[|V|] < \infty$.

1. Suppose $P\{V \geq 0\} = 1$ and $E[V] = 0$. Show that $P\{V = 0\} = 1$. Begin from the almost-sure bound $j^{-1}\mathbf{1}\{V \geq 1/j\} \leq V\mathbf{1}\{V \geq 1/j\} \leq V$, valid because $V \geq 0$ almost surely, and finish with a countable union.
2. Deduce the form used in the text: if $P\{V \leq 0\} = 1$ and $E[V] = 0$ then $P\{V = 0\} = 1$. Then say precisely which random variable this is applied to in the proof of Lemma 1.1, and why that variable is non-positive.
3. Let $E[|Y|] < \infty$, so that $E[Y|X]$ is defined. Prove that if $P\{Y \geq 0\} = 1$ then $P\{E[Y|X] \geq 0\} = 1$. Take $B = \{x : m(x) < -1/j\}$ in (1.2) and let $j \rightarrow \infty$. Where in the argument is it used that B is Borel?

Bibliography

- P. Billingsley. *Probability and Measure*. Wiley-Interscience, 1995.
- J. Jiang. *Large Sample Techniques for Statistics*. Springer Texts in Statistics. Springer, New York, 2010.
- E. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, New York, 3rd edition, 2005.
- A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.
- R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, second edition, 2026.



Part II

**Convergence and Limit
Theorems**



2

Inequalities, Modes of Convergence, and the Law of Large Numbers

In the previous chapter we defined the object: $E[Y|X]$, the best prediction of Y given X , determined by the joint distribution and by nothing else. We never asked how one would learn it from a finite sample. Every method in this course begins by answering that question in the same way, replacing a population expectation by its sample analogue, and the whole of large-sample theory is an attempt to say when the replacement is legitimate. This chapter asks the question in its simplest form. *When, and in what sense, may we replace a population expectation by its sample analogue?*

Why we cover this

Every estimator in this course is built by replacing an expectation with an average. The law of large numbers is the theorem that licenses the replacement, the modes of convergence say in what sense it holds, and the inequalities of Sections 2.2 and 2.3 say how good the replacement is at a given sample size. None of the four ideas is difficult on its own. What matters is knowing which one a given claim rests on, because “the sample mean is close to the population mean” can mean at least three different things, and the difference between them decides whether a confidence interval is honest.

2.1 Three Prototypical Problems

Before building machinery to learn a feature of a distribution from data, it is worth naming what “learn” will mean, and fixing notation for the rest of the course. Let $W_1, \dots, W_n$ be independent and identically distributed with distribution P , where P is known only to lie in some class $\mathbf{P}$, and let $\theta(P)$ denote a feature of P . Three problems recur throughout the sequence.

1. *Estimation.* Construct $\hat{\theta}_n = \hat{\theta}_n(W_1, \dots, W_n)$ that is, in a sense to be made precise, a good guess for $\theta(P)$.

2. *Testing.* Construct $\phi_n = \phi_n(W_1, \dots, W_n)$ taking values in $[0, 1]$, giving the probability with which one rejects the hypothesis that $\theta(P) = \theta_0$.
3. *Confidence regions.* Construct a set $C_n = C_n(W_1, \dots, W_n)$ containing $\theta(P)$ with probability approximately $1 - \alpha$.

Writing W for the generic observation keeps the two running examples of this course notationally separate from the conditioning variable: in the first, $W_i = X_i$ is a scalar; in the second, $W_i = (Y_i, X_i)$ is a pair.

Example 2.1 (The coin) Let $W_1, \dots, W_n$ be i.i.d. Bernoulli(q) with $0 < q < 1$ and let $\theta(P) = q$. The natural estimator is $\bar{W}_n = n^{-1} \sum_{i \leq n} W_i$. Everything here is available in closed form: $E[\bar{W}_n] = q$, $\text{Var}[\bar{W}_n] = q(1 - q)/n$, and $n\bar{W}_n$ is exactly Binomial(n, q), so every approximation we develop can be checked against the truth. ■

The class $\mathbf{P}$ matters as much as the parameter. Almost everything in this course is stated for classes restricted only by a moment condition, rather than for parametric families, because that is what an econometrician is usually willing to assume.

That choice has a price, and the price is why this course exists. If $\mathbf{P}$ is the normal family, the exact distribution of $\bar{W}_n$ is available at every n ; if it is every distribution with a finite second moment, it is not, and no formula recovers it. Exact finite-sample statements therefore require assumptions few would defend, and the alternative is to approximate. This chapter delivers the two kinds of approximation available: an *inequality*, which bounds the error at every n at the cost of being conservative, and a *limit theorem*, which is sharper but exactly correct only at a sample size no one has. We develop large-sample theory not because we believe $n = \infty$, but because for these model classes the limit is the only sharp description on offer.

2.2 Bounding Probabilities by Moments

Every bound in this course descends from one idea: an event of interest is contained in an event on which some non-negative random variable is large, and a non-negative random variable cannot be large too often without having a large expectation.

Lemma 2.1 (Markov and Chebyshev) Let V be a random variable and $a > 0$.

- (i) If $V \geq 0$ then $P\{V \geq a\} \leq a^{-1} E[V]$.
- (ii) For any V and any $p > 0$, $P\{|V| \geq a\} \leq a^{-p} E[|V|^p]$.

(iii) If $E[V^2] < \infty$ then $P\{|V - E[V]| \geq a\} \leq a^{-2} \text{Var}[V]$.

PROOF. For (i), note the pointwise inequality $a\mathbf{1}\{V \geq a\} \leq V\mathbf{1}\{V \geq a\} \leq V$, valid because $V \geq 0$. Taking expectations gives $aP\{V \geq a\} \leq E[V]$. Part (ii) is (i) applied to the non-negative variable $|V|^p$ and the level a^p , since $\{|V| \geq a\} = \{|V|^p \geq a^p\}$. Part (iii) is (ii) with $p = 2$ applied to $V - E[V]$. ■

Cf. [Jiang \[2010, eq. \(5.77\), p. 152\]](#), who states the sharper form $P\{V > a\} \leq a^{-1}E[V\mathbf{1}\{V > a\}]$ and derives Markov's inequality from it; [Vershynin \[2026, Proposition 1.6.2 and Corollary 1.6.3, p. 16\]](#).

Lemma 2.1 is crude — it uses nothing about V beyond one moment — and that is exactly why it is useful. It is also the engine of almost every consistency proof in the course: to show an estimator is close to its target, bound a moment of the difference and divide. Sharper bounds, in Section 2.3, come from applying part (i) not to $|V|$ but to $e^{\lambda V}$, which uses every moment at once.

Applied to a sample mean, part (iii) already gives the bound that carries the law of large numbers. If $W_1, W_2, \dots$ are i.i.d. with $E[W_i] = \mu$ and $\text{Var}[W_i] = \sigma^2 < \infty$, then $\text{Var}[\bar{W}_n] = \sigma^2/n$ by independence, and so

$$P\{|\bar{W}_n - \mu| > \varepsilon\} \leq \frac{\sigma^2}{n\varepsilon^2} \quad \text{for every } \varepsilon > 0 \text{ and every } n. \quad (2.1)$$

We use (2.1) twice: in Section 2.3 as the benchmark that concentration improves upon, and in Section 2.5 to prove the weak law.

Two further inequalities, Jensen's and Cauchy–Schwarz's, are used repeatedly in this course but never proved in class. They are stated with their proofs in Sections A.3 and A.4 of the appendix, and cited from here on by number.

2.3 Concentration: Exponential Bounds at Every n

Chebyshev's inequality implies that a sample mean is within ε of its expectation except with probability of order $1/n$; see (2.1). When the summands are bounded, far more is true: the exceptional probability decays exponentially in n . The device is to apply Markov's inequality to $e^{\lambda S}$ and then choose λ well, an argument known as the *Chernoff bound*.

The argument needs one input, which we take from the appendix. *Hoeffding's lemma*, Fact A.8, states that a centered random variable supported on an interval of length $M - m$ satisfies $E[e^{\lambda V}] \leq \exp\{\lambda^2(M - m)^2/8\}$ for every λ . In other words, boundedness alone forces the moment generating function to be no larger than that of a normal variable.

Theorem 2.1 (Hoeffding) Let $V_1, \dots, V_n$ be independent with $m_i \leq V_i \leq M_i$ almost surely, and write $S_n = \sum_{i \leq n} (V_i - E[V_i])$. Then for every $t > 0$,

$$P\{S_n \geq t\} \leq \exp\left\{\frac{-2t^2}{\sum_{i \leq n} (M_i - m_i)^2}\right\}, \quad (2.2)$$

and the same bound holds for $P\{S_n \leq -t\}$; consequently $P\{|S_n| \geq t\}$ is at most twice the right-hand side.

PROOF. Fix $\lambda > 0$. Since $v \mapsto e^{\lambda v}$ is increasing, the event $\{S_n \geq t\}$ coincides with $\{e^{\lambda S_n} \geq e^{\lambda t}\}$, and Markov's inequality applied to the non-negative variable $e^{\lambda S_n}$ gives

$$P\{S_n \geq t\} \leq e^{-\lambda t} E[e^{\lambda S_n}].$$

The summands are independent, so the expectation factors, and Fact A.8 applies to each centered term $V_i - E[V_i]$, which lies in an interval of the same length $M_i - m_i$:

$$E[e^{\lambda S_n}] = \prod_{i \leq n} E[e^{\lambda(V_i - E[V_i])}] \leq \exp\left\{\frac{1}{8}\lambda^2 \sum_{i \leq n} (M_i - m_i)^2\right\}.$$

Writing $D = \sum_{i \leq n} (M_i - m_i)^2$, we therefore have $P\{S_n \geq t\} \leq \exp\{-\lambda t + \lambda^2 D/8\}$ for every $\lambda > 0$. The exponent is a quadratic in λ minimized at $\lambda = 4t/D$, and substituting that value gives $-4t^2/D + 2t^2/D = -2t^2/D$, which is (2.2). The bound for $P\{S_n \leq -t\}$ follows by applying (2.2) to $-V_1, \dots, -V_n$, and the two-sided statement by adding the two. ■

Cf. *Vershynin [2026, Theorem 2.2.6, p. 30]* states exactly (2.2), with the Rademacher case at his *Theorem 2.2.1, p. 28*. *Jiang [2010, Example 5.6, p. 135]* derives a bound of the same shape by the same Chernoff device, building it from the elementary inequality $e^x \leq 1 + x + ax^2$ of his *Example 5.5*.

Applied to the coin of Example 2.1, where $V_i = W_i \in [0, 1]$ so that $D = n$, Theorem 2.1 gives

$$P\{|\bar{W}_n - q| \geq \varepsilon\} \leq 2e^{-2n\varepsilon^2} \quad \text{for every } n \text{ and every } q. \quad (2.3)$$

Compare this with what Chebyshev's inequality delivers. Since $\text{Var}[\bar{W}_n] = q(1 - q)/n \leq 1/(4n)$,

$$P\{|\bar{W}_n - q| \geq \varepsilon\} \leq \frac{1}{4n\varepsilon^2}. \quad (2.4)$$

Both hold at every n and neither involves an unknown quantity, but the first decays exponentially in n and the second only polynomially. In other words, both answer the question this chapter opened with — how good is the replacement of q by $\bar{W}_n$ at a given sample size — and the answers differ enormously. Section 2.6 turns them into intervals and compares them with the one used in practice.

2.4 Modes of Convergence

An estimator is a sequence of random variables, and there is more than one way for such a sequence to settle down. Three notions matter for us; a fourth, convergence in distribution, waits until Chapter 4 because it is a statement about distributions rather than about random variables.

Definition 2.1 Let V_n , $n \geq 1$, and V be random variables defined on a common probability space.

- (i) $V_n \rightarrow V$ in probability, written $V_n \xrightarrow{P} V$, if $P\{|V_n - V| > \varepsilon\} \rightarrow 0$ for every $\varepsilon > 0$.
- (ii) $V_n \rightarrow V$ in L^q for $q > 0$, written $V_n \xrightarrow{L^q} V$, if $E[|V_n - V|^q] \rightarrow 0$.
- (iii) $V_n \rightarrow V$ almost surely, written $V_n \xrightarrow{a.s.} V$, if $P\{V_n \rightarrow V\} = 1$.

An estimator $\hat{\theta}_n$ with $\hat{\theta}_n \xrightarrow{P} \theta(P)$ for every $P \in \mathbf{P}$ is called consistent.

The three definitions differ in what they hold fixed, and the difference is easiest to see on Figure 2.1(a), which draws six independent sample paths $n \mapsto \bar{W}_n$ for the coin. Both definitions fix ε first and then let n grow, which is why the band drawn there has a fixed half-width. Convergence in probability is a statement about a *vertical slice*: fix n , look across the paths, and ask what fraction of them lie outside the band. It requires that fraction to vanish as the slice moves right. Almost sure convergence is a statement about the *paths themselves*: it requires that all but a null set of them eventually stay inside, and that this hold for every $\varepsilon > 0$, the point of entry moving right as the band narrows and depending on the realization as well as on ε . Convergence in L^q is a statement about neither. An event of vanishing probability can destroy any of the three; what distinguishes L^q is that the *magnitude* of the error on that event still counts, so a rare deviation ruins the convergence if it is large enough.

Remark 2.1 (Reading almost sure convergence) Two misreadings of Definition 2.1(iii) are worth heading off, because both are common and both matter.

The first is the order of the limit and the probability. Almost sure convergence puts the limit inside: the event $\{V_n \rightarrow V\}$ is a statement about the entire infinite sequence, and we require it to have probability one. Convergence in probability puts the limit outside, applying it to a sequence of numbers $P\{|V_n - V| > \varepsilon\}$, one for each n . The second says nothing about any single realization; the first says something about almost every realization.

The second misreading is to conclude that almost sure convergence, being stronger, is the notion we should always want. It is the natural one when the

sequence itself is the object of study — in the bootstrap, or when a numerical algorithm is iterated. For first-order asymptotic inference, convergence in probability is normally all that is needed, and the tools that manipulate it — the continuous mapping theorem, Slutsky’s theorem, the o_p and O_p calculus — are built for that mode. It is the default notion of consistency here. ■

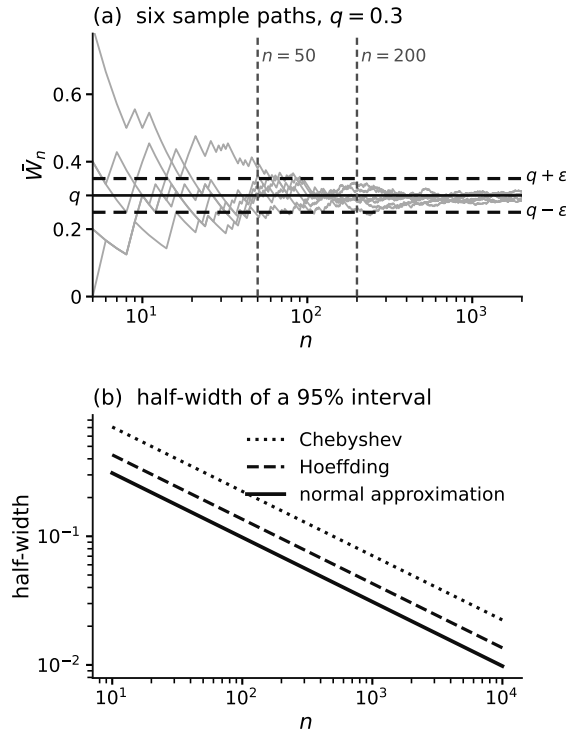


FIGURE 2.1: The two questions of this chapter, on the coin. Panel (a) shows six independent sample paths of $\bar{W}_n$ for a Bernoulli(0.3) coin, with a band of fixed half-width $\varepsilon = 0.05$ about q : three of the six lie outside it at $n = 50$ and none at $n = 200$. Panel (b) plots the half-width of three 95% intervals against n , in the worst case over q : Chebyshev, Hoeffding, and the normal approximation of Chapter 5.

One extension of Definition 2.1(i) is worth recording now, because it is how every statement about the power of a test is eventually phrased.

Definition 2.2 A sequence V_n diverges to $+\infty$ in probability, written $V_n \xrightarrow{P} +\infty$, if $P\{V_n \leq M\} \rightarrow 0$ for every $M \in \mathbf{R}$.

This is not a special case of Definition 2.1(i), since $+\infty$ is not a random variable, but the logic is the same one about a vertical slice: eventually almost all of the mass sits above any level one names. For the coin with $q > 0$ the

sequence $n\bar{W}_n$ diverges in this sense, since $\bar{W}_n \xrightarrow{P} q$ puts $\bar{W}_n$ above $q/2$ with probability tending to one, and then $n\bar{W}_n > nq/2 > M$ for all large n . The reason to have the notion is that a test statistic behaves this way under a fixed alternative: if it diverges in probability while the critical value stays bounded, the probability of rejection tends to one, which is what it means for a test to be *consistent*.

Lemma 2.2 $V_n \xrightarrow{L^q} V$ for some $q > 0$ implies $V_n \xrightarrow{P} V$, and $V_n \xrightarrow{\text{a.s.}} V$ implies $V_n \xrightarrow{P} V$.

PROOF. For the first, Markov's inequality in the form of Lemma 2.1(ii) gives, for every $\varepsilon > 0$,

$$P\{|V_n - V| > \varepsilon\} \leq \varepsilon^{-q} E[|V_n - V|^q] \rightarrow 0.$$

For the second, fix $\varepsilon > 0$ and set $A_n = \{|V_n - V| > \varepsilon\}$. On the event $\{V_n \rightarrow V\}$, only finitely many A_n occur, so $\mathbf{1}\{A_n\} \rightarrow 0$ almost surely. Since $\mathbf{1}\{A_n\} \leq 1$ and constants are integrable, dominated convergence gives $P\{A_n\} = E[\mathbf{1}\{A_n\}] \rightarrow 0$. ■

Cf. Jiang [2010, Theorem 2.1, p. 22, and Theorem 2.7, p. 23]. Dominated convergence is Fact A.2.

Neither implication reverses, and neither of the two stronger notions implies the other. Two examples settle all of it. Both live on the unit interval with P the uniform distribution, so that U denotes a Uniform[0, 1] variable.

Example 2.2 (The moving window) Write each $n \geq 1$ as $n = 2^k + r$ with $0 \leq r < 2^k$, and set $V_n = \mathbf{1}\{U \in (r2^{-k}, (r+1)2^{-k}]\}$. As n increases the window sweeps across $[0, 1]$, halving in width each time k increases. Then $E[|V_n|^q] = 2^{-k} \rightarrow 0$ for every $q > 0$, so $V_n \xrightarrow{L^q} 0$ and hence $V_n \xrightarrow{P} 0$. But for every $u \in (0, 1]$ the window covers u for infinitely many n , so $V_n(u) = 1$ infinitely often and $V_n(u)$ does not converge. The one point it does converge at is $u = 0$, which the half-open intervals never cover; the set on which convergence fails therefore has probability one. Convergence in L^q , and in probability, therefore does not imply almost sure convergence. ■

Example 2.3 (The tall spike) Let $V_n = e^n \mathbf{1}\{U \leq 1/n\}$. For any fixed $u > 0$ we have $u > 1/n$ once n is large, so $V_n(u) = 0$ eventually and $V_n \xrightarrow{\text{a.s.}} 0$. But for every $q > 0$,

$$E[|V_n|^q] = e^{qn} \cdot \frac{1}{n} \rightarrow \infty,$$

so V_n fails to converge to 0 in L^q for every q , and $E[V_n] \rightarrow \infty$ while $V_n \rightarrow 0$ almost surely. Almost sure convergence, and convergence in probability,

therefore do not imply convergence in L^q , nor do they imply convergence of expectations.

The height of the spike is doing all the work. With $V_n = n \mathbf{1}\{U \leq 1/n\}$ one gets $E[|V_n|^q] = n^{q-1}$, which fails to vanish only for $q \geq 1$: that sequence separates almost sure convergence from L^1 but not from L^q for $q < 1$. Problem 2.3 works out where the threshold lies. ■

Remark 2.2 (Consistency vs mean squared error) Read Example 2.3 as an estimator. Fix θ and let $\hat{\theta}_n$ be $\theta + V_n$, with V_n the spike above. Then $\hat{\theta}_n \rightarrow \theta$ almost surely, and hence in probability, so $\hat{\theta}_n$ is consistent, while

$$E[(\hat{\theta}_n - \theta)^2] = \frac{e^{2n}}{n} \rightarrow \infty.$$

Consistency constrains probabilities; mean squared error constrains moments, and it is the strictly stronger requirement, being convergence in L^2 . The two are often conflated in both directions: a simulation reporting a shrinking mean squared error has shown more than consistency, and an estimator with an unbounded mean squared error may still be perfectly consistent. The gap is exactly a vanishing probability attached to an exploding value, which is what Theorem 2.2 below rules out. ■

The next criterion turns a summable sequence of probabilities into almost sure convergence, and is the standard route to the strong law.

Lemma 2.3 (Borel–Cantelli) Let $A_1, A_2, \dots$ be events, and suppose that $\sum_n P\{A_n\}$ converges. Then the probability that infinitely many of them occur is zero.

PROOF. The event that infinitely many A_n occur is $\bigcap_{m \geq 1} \bigcup_{n \geq m} A_n$, so for every m ,

$$P\{A_n \text{ i.o.}\} \leq P\left\{\bigcup_{n \geq m} A_n\right\} \leq \sum_{n \geq m} P\{A_n\}.$$

The right-hand side is the tail of a convergent series, so it tends to zero as $m \rightarrow \infty$, and the left-hand side does not depend on m . ■

Cf. Jiang [2010, Lemma 2.5, p. 44]; Billingsley [1995, Theorem 4.3, p. 59]. The converse statement, that independence and a divergent sum force the events to occur infinitely often, is Billingsley's Theorem 6.3, p. 88; Problem 2.4 uses it.

Remark 2.3 (What summability buys) Suppose the series $\sum_n P\{|V_n - V| > \varepsilon\}$ converges for every $\varepsilon > 0$. Fix one such ε and apply Lemma 2.3 to the events A_n in that series: outside a null set, only finitely many of them occur.

To conclude that $V_n \rightarrow V$ almost surely one further step is required, since the exceptional event may depend on ε : note that

$$\{V_n \not\rightarrow V\} = \bigcup_{k \geq 1} \{|V_n - V| > 1/k \text{ i.o.}\} ,$$

a countable union of null events, hence null. In other words, convergence in probability asks these probabilities to tend to zero, and almost sure convergence follows if they do so fast enough to be summable.

Summability is sufficient and not necessary, and Example 2.3 already shows why: there $P\{|V_n| > \varepsilon\} = 1/n$ for every small ε , which is not summable, and yet $V_n \rightarrow 0$ almost surely. What almost sure convergence is equivalent to is

$$P\left\{\sup_{m \geq n} |V_m - V| > \varepsilon\right\} \longrightarrow 0 \quad \text{for every } \varepsilon > 0 ,$$

a statement about the whole tail of the sequence rather than about one term at a time. Summability is the version of it we can usually check, and it explains why Section 2.3 is a prerequisite for the strong law: an exponential bound is summable, a polynomial one need not be. ■

One gap remains, the one Example 2.3 exposed: convergence in probability does not deliver convergence of expectations. The condition that closes it is that the sequence not put non-vanishing mass on large values.

Definition 2.3 A sequence $V_1, V_2, \dots$ is uniformly integrable if

$$\lim_{\lambda \rightarrow \infty} \sup_{n \geq 1} E[|V_n| \mathbf{1}\{|V_n| > \lambda\}] = 0 .$$

Uniform integrability fails in Example 2.3: once $e^n > \lambda$ the indicator is redundant, and $E[|V_n| \mathbf{1}\{|V_n| > \lambda\}] = e^n/n$, which diverges rather than vanishing. The milder spike $V_n = n \mathbf{1}\{U \leq 1/n\}$ fails too, with the same expectation equal to one for every $n > \lambda$. It holds, for instance, whenever $\sup_n E[|V_n|^{1+\delta}] < \infty$ for some $\delta > 0$, which is why a bounded moment slightly above the one of interest is such a common assumption.

Theorem 2.2 Suppose $V_n \xrightarrow{P} V$ and $E[|V_n|] < \infty$ for every n . Then $V_1, V_2, \dots$ is uniformly integrable if and only if $V_n \xrightarrow{L^1} V$ with $E[|V|] < \infty$, in which case $E[V_n] \rightarrow E[V]$.

Cf. Jiang [2010, Theorem 2.17, p. 33], who states the equivalence for L^p and adds the further equivalent condition $E[|V_n|^p] \rightarrow E[|V|^p] < \infty$; Billingsley [1995, §16, “Uniform Integrability”, p. 216].

2.5 The Law of Large Numbers

We can now answer the question the chapter opened with. Almost every estimator in this sequence is built by writing the parameter as a feature of the distribution and then replacing the distribution by the empirical one. If $\theta(P)$ solves $E[g(W, \theta)] = 0$, the estimator solves $n^{-1} \sum_{i \leq n} g(W_i, \hat{\theta}_n) = 0$; if $\theta(P)$ minimizes $E[m(W, \theta)]$, the estimator minimizes $n^{-1} \sum_{i \leq n} m(W_i, \theta)$. Method of moments, maximum likelihood, least squares and GMM are all instances. The construction is called the *analog principle*, and it is only as good as the approximation of $E[g(W, \theta)]$ by its sample analogue. The law of large numbers is the theorem that makes it legitimate.

Theorem 2.3 (Weak law of large numbers) *Let $W_1, \dots, W_n$ be i.i.d. with $E[W_i] = \mu$ and $\text{Var}[W_i] = \sigma^2 < \infty$. Then $\bar{W}_n \xrightarrow{P} \mu$.*

PROOF. This is (2.1), whose right-hand side tends to zero as $n \rightarrow \infty$ for every fixed $\varepsilon > 0$. ■

Cf. The Chebyshev proof is standard. Finite variance is not needed: Jiang [2010, Theorem 6.1, p. 174] shows that for i.i.d. summands, $\bar{W}_n \xrightarrow{P} \mu$ holds if and only if $tP\{|W_1| > t\} \rightarrow 0$ and the truncated mean converges, a condition implied by $E[|W_1|] < \infty$. The proof of that version proceeds by truncation, a technique we use repeatedly from here on.

Remark 2.4 (What the variance assumption is doing) A finite variance buys the one-line proof and nothing else: the weak law holds under a finite mean alone, as the source note records. What an infinite variance does put at risk is what comes after the law of large numbers: the usual $\sqrt{n}$ rate, the Gaussian limit of Chapter 5, and the conventional variance-based standard errors built on it need not remain valid. ■

The weak law is a statement about the vertical slice. With Section 2.3 in hand we can say something about the paths as well, at least when the data are bounded.

Theorem 2.4 (Strong law, bounded case) *Let $W_1, W_2, \dots$ be i.i.d. with $a \leq W_i \leq b$ almost surely and $E[W_i] = \mu$. Then $\bar{W}_n \xrightarrow{a.s.} \mu$.*

PROOF. Fix $\varepsilon > 0$ and apply Theorem 2.1 to $W_1, \dots, W_n$, for which $D = n(b-a)^2$ and $S_n = n(\bar{W}_n - \mu)$. Taking $t = n\varepsilon$,

$$P\{|\bar{W}_n - \mu| \geq \varepsilon\} \leq 2 \exp\left\{\frac{-2n\varepsilon^2}{(b-a)^2}\right\}. \quad (2.5)$$

The right-hand side is a geometric sequence in n , hence summable, so Lemma

2.3 applied to $A_n = \{|\bar{W}_n - \mu| \geq \varepsilon\}$ gives $P\{|\bar{W}_n - \mu| \geq \varepsilon \text{ i.o.}\} = 0$. Since $\varepsilon > 0$ was arbitrary, Remark 2.3 converts this into $P\{\bar{W}_n \rightarrow \mu\} = 1$. ■

Cf. The route through an exponential bound and Borel–Cantelli is the standard one for bounded summands, and it uses nothing beyond Theorem 2.1 and Lemma 2.3 as proved above. The general strong law, Theorem 2.5, requires a different argument; see Billingsley [1995, Theorem 6.1, p. 85].

Chebyshev’s bound (2.1) is of order $1/n$ and $\sum_n 1/n$ diverges, so it delivers convergence in probability and stops there; Hoeffding’s decays exponentially, the series converges, and Borel–Cantelli upgrades the same deviations to a statement about the paths — under the boundedness assumption.

Boundedness can be dropped, and the general statement is not merely more general: it is sharp.

Theorem 2.5 (Strong law of large numbers) *Let $W_1, W_2, \dots$ be i.i.d. Then $\bar{W}_n \xrightarrow{\text{a.s.}} \mu$ for some $\mu \in \mathbf{R}$ if and only if $E[|W_1|] < \infty$, in which case $\mu = E[W_1]$.*

Cf. Jiang [2010, Theorem 6.6, §6.3] states exactly this and obtains the classical sufficiency half as a corollary; he also gives versions for independent but not identically distributed summands, and versions in which the normalizing constant is not n . Billingsley [1995, Theorem 6.1, p. 85] and Vershynin [2026, Theorem 1.7.1, p. 17] give sufficiency only.

Integrability is therefore not an artifact of the proof: it is exactly what almost sure convergence of a sample mean requires, and Theorem 2.4 is the case in which the argument is cheap. The weak law asks for less. There are distributions with an infinite mean for which $\bar{W}_n$ still converges in probability, and for those, by Theorem 2.5, it does not converge almost surely.

Two qualifications are worth naming. The law of large numbers says nothing about *how fast*: the error vanishes, but the theorem supplies no rate, and providing one is the work of Section 2.3 at finite n and of the central limit theorem in Chapter 5 in the limit. And the versions stated here assume the data are i.i.d., which is enough for most of our purposes but is not what econometricians usually have. Laws of large numbers are available for summands that are independent but not identically distributed, for triangular arrays, and under various forms of dependence. Two of these come later: Chapter 6 replaces the identical-distribution assumption with triangular arrays, and the spring quarter takes up dependence within clusters.

Remark 2.5 (Where this is going) Take the estimator defined by minimizing a sample criterion, $\hat{\theta}_n \in \arg \min_{\theta \in \Theta} M_n(\theta)$ with $M_n(\theta) = n^{-1} \sum_{i \leq n} m(W_i, \theta)$, and let θ_0 minimize $M(\theta) = E[m(W, \theta)]$. A consistency proof for $\hat{\theta}_n$ has three steps:

- (i) $M_n(\theta) \xrightarrow{P} M(\theta)$ for each fixed θ — this chapter;

- (ii) the convergence in (i) holds *uniformly* over $\theta \in \Theta$ — Chapter 3;
- (iii) θ_0 is well separated from the rest of Θ , and therefore $\hat{\theta}_n \xrightarrow{P} \theta_0$ — the second half of the quarter.

This chapter justifies replacing one fixed population expectation by its sample analogue. M-estimation requires doing so simultaneously at every candidate parameter value, and that is the subject of the next lecture. ■

2.6 Three Intervals for the Coin

Return to the coin of Example 2.1 and ask for an interval that contains q with probability at least 0.95. Three constructions are now available, and they differ in what they assume and in what they deliver.

Chebyshev's bound (2.4) gives the interval $\bar{W}_n \pm (4n\alpha)^{-1/2}$; Hoeffding's bound (2.3) gives $\bar{W}_n \pm \{\log(2/\alpha)/2n\}^{1/2}$. Both are valid at every n and for every q , and neither requires knowing anything about P beyond the range of the data. A third interval, $\bar{W}_n \pm z_{1-\alpha/2}\{\bar{W}_n(1 - \bar{W}_n)/n\}^{1/2}$, comes from the normal approximation and is the one reported in practice; we cannot yet justify it, and Chapter 5 is largely about doing so. At $n = 1000$ and $\alpha = 0.05$, the three half-widths are 0.071, 0.043 and, at worst over q , 0.031. Figure 2.1(b) shows how the comparison behaves as n grows.

The ordering is not an accident. The finite-sample intervals pay for their honesty in width; the asymptotic interval is narrower because it estimates the variance from the data rather than bounding it, and is exactly correct only in a limit no sample attains. Which one is right depends on what you are willing to assume. Chapter 8 shows that the asymptotic interval can be worse than this comparison suggests.

Where we will use this

- **In this class.** The analog principle of Section 2.5 is how every estimator in the second half of the quarter is defined, and Remark 2.5 is the skeleton of its consistency proof. Chebyshev's inequality supplies step (i) there and reappears in every consistency argument through Chapter 7; the modes of convergence are the vocabulary in which every asymptotic statement of the course is phrased.
- **Later in the sequence.** The Chernoff device of Section 2.3 is where the finite-sample bounds of the winter quarter begin, and they are what allows anything to be said at all when the number of covariates is comparable to, or larger than, the sample size.
- **In recent research.** Non-asymptotic bounds of the kind in Section

2.3 are the standard currency of the statistical learning literature, where results are stated as inequalities holding at every n rather than as limits. The two styles answer different questions, and it is worth being able to read both.

2.7 Key Concepts

- **The analog principle.** Write the parameter as a feature of P , then replace P by the empirical distribution. The law of large numbers is what licenses the substitution; everything else in the course refines it.
- **Markov and Chebyshev.** $P\{V \geq a\} \leq a^{-1}E[V]$ for $V \geq 0$, and $P\{|V - E[V]| \geq a\} \leq a^{-2} \text{Var}[V]$. Trade a probability for a moment; for a sample mean this is (2.1).
- **The Chernoff bound.** Apply Markov to $e^{\lambda V}$ and optimize over λ . This turns a polynomial bound into an exponential one, and gives Hoeffding's inequality for bounded summands, valid at every n .
- **Modes of convergence.** In probability is about the vertical slice, almost surely about the paths, L^q about the magnitude of the error. Summability of the deviation probabilities is what separates the first two. Consistency is the first, and it is not the same as a vanishing mean squared error.
- **The law of large numbers.** $\bar{W}_n \xrightarrow{P} \mu$ under a finite variance, which is Chebyshev, and $\bar{W}_n \xrightarrow{\text{a.s.}} \mu$ exactly when $E[|W_1|] < \infty$.

2.8 Concluding Remarks

The organization of this chapter follows Jiang [2010], whose Chapter 5 collects the inequalities and whose Chapter 2 develops the modes of convergence with a generous supply of counterexamples; Examples 2.2 and 2.3 are standard and appear there in various forms. The treatment of concentration follows Vershynin [2026], Chapter 2, which is the natural place to read further — in particular on sub-Gaussian tails and maxima, on sub-exponential variables and Bernstein's inequality, which we have not covered and which are the right tools when the summands are unbounded but have light tails. Billingsley [1995]

is the reference for the measure-theoretic statements, including the general strong law and uniform integrability.

2.9 Problems

Problem 2.1 Let $V \geq 0$ with $E[V] = 1$.

1. Show that $P\{V \geq 10\} \leq 1/10$, and construct a distribution for V attaining the bound.
2. Suppose in addition that $\text{Var}[V] = 1$. Using Chebyshev's inequality, bound $P\{V \geq 10\}$ and compare with part (a).
3. Explain in one sentence why the bound in part (a) cannot be improved over the class of all $V \geq 0$ with $E[V] = 1$. (The bound in part (b) is a different matter; it is not sharp over the class it quantifies over.)

Problem 2.2 Let $W_1, \dots, W_n$ be i.i.d. Bernoulli(q).

1. Using (2.3), find the smallest n guaranteeing $P\{|\bar{W}_n - q| \geq 0.01\} \leq 0.05$ for every q .
2. Repeat using (2.4), and compare.
3. Now suppose you knew $q \leq 0.01$. Which of the two bounds can exploit that information, and which cannot? Explain what this suggests about when Hoeffding's inequality is and is not the right tool.

Problem 2.3 Let $U \sim \text{Uniform}[0, 1]$ and let $V_n = n^\gamma \mathbf{1}\{U \leq 1/n\}$ for a constant $\gamma > 0$.

1. For which γ does $V_n \xrightarrow{P} 0$? For which does $V_n \xrightarrow{\text{a.s.}} 0$?
2. Fix $q > 0$. For which γ does $V_n \xrightarrow{L^q} 0$? For which γ is the sequence uniformly integrable?
3. Deduce that for every fixed q there is a sequence separating almost sure convergence from convergence in L^q , and explain why the sequence of Example 2.3 does the job for all q at once.

Problem 2.4 Let $V_1, V_2, \dots$ be independent with $P\{V_n = 1\} = p_n$ and $P\{V_n = 0\} = 1 - p_n$.

1. Show that $V_n \xrightarrow{P} 0$ if and only if $p_n \rightarrow 0$.
2. Show, using Lemma 2.3, that $\sum_n p_n < \infty$ implies $V_n \xrightarrow{\text{a.s.}} 0$.

3. Take $p_n = 1/n$. Show that $V_n \xrightarrow{P} 0$ but that V_n does not converge to 0 almost surely. (You may use, without proof, that for independent events a divergent sum of probabilities forces the events to occur infinitely often.)

Problem 2.5 Let $W_1, W_2, \dots$ be i.i.d. with $E[W_i] = \mu$, and consider the two routes to a law of large numbers used in Section 2.5.

1. Suppose $\text{Var}[W_i] = \sigma^2 < \infty$. Chebyshev's inequality bounds $P\{|\bar{W}_n - \mu| \geq \varepsilon\}$ by $\sigma^2/(n\varepsilon^2)$. Show that this bound is not summable in n , so that Lemma 2.3 cannot be applied to it directly.
2. Show that the bound is summable along the subsequence $n = k^2$, and conclude that $\bar{W}_{k^2} \rightarrow \mu$ almost surely as $k \rightarrow \infty$. (This is the first step of one standard proof of the general strong law; the work lies in controlling what happens between consecutive squares.)
3. Explain, in two or three sentences, why Theorem 2.4 needs no such device.

Bibliography

- P. Billingsley. *Probability and Measure*. Wiley-Interscience, 1995.
- J. Jiang. *Large Sample Techniques for Statistics*. Springer Texts in Statistics. Springer, New York, 2010.
- R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, second edition, 2026.



3

The Analog Principle and Uniform Laws of Large Numbers

The previous chapter showed in what sense a sample average converges to a population expectation. That settles the problem when the parameter is an expectation, but quite often that is not the case. This chapter takes two steps away from the sample average. The first is short: a continuous function of sample averages converges to the same function of the expectations, which covers variances, ratios and group-mean differences. The second is not. Some estimators are evaluated at a point that the data choose, and for these the law of large numbers has to hold at every candidate point at once.

Why we cover this

A sample quantile inverts the empirical distribution function at a point the data determine. A maximum likelihood estimator sits wherever the sample log-likelihood happens to peak. Each reads a random function at a location that is itself random, and accuracy at a location named in advance says nothing about the location actually used. Everything in this chapter exists to close that gap.

3.1 Continuous Mapping

Chapter 2 defined convergence in probability for a real-valued sequence. The estimators below are functions of several averages at once, so the definition has to cover random vectors.

Definition 3.1 *Let V_n and V be random vectors in $\mathbf{R}^k$ and let $\|\cdot\|$ denote the Euclidean norm. Then $V_n \xrightarrow{P} V$ if*

$$P\{\|V_n - V\| > \varepsilon\} \longrightarrow 0 \quad \text{for every } \varepsilon > 0 .$$

Nothing is lost by checking coordinates one at a time. Write $V_n = (V_{n,1}, \dots, V_{n,k})$ and $V = (V_1, \dots, V_k)$, and suppose $V_{n,j} \xrightarrow{P} V_j$ for each $j \leq k$.

If $\|V_n - V\| > \varepsilon$ then some coordinate deviates by more than $\varepsilon/\sqrt{k}$, since otherwise the norm would be at most ε . The event $\{\|V_n - V\| > \varepsilon\}$ is therefore contained in the union of the k events $\{|V_{n,j} - V_j| > \varepsilon/\sqrt{k}\}$, and

$$P\{\|V_n - V\| > \varepsilon\} \leq \sum_{j \leq k} P\{|V_{n,j} - V_j| > \varepsilon/\sqrt{k}\}, \quad (3.1)$$

where each of the finitely many terms on the right tends to zero. The converse direction is immediate from $|V_{n,j} - V_j| \leq \|V_n - V\|$. In other words, convergence in probability of a vector is the same thing as convergence of each of its coordinates, and the weak law of Section 2.5 applies to a vector of sample averages one coordinate at a time.

The step in (3.1) that bounds the probability of a union by the sum of the probabilities is the *union bound*. It returns twice in this chapter, both times to replace an uncountable family of deviations by a finite one.

The corresponding statement for convergence in distribution is false. Convergence of each coordinate in distribution does not give convergence of the vector, and Chapter 4 gives the counterexample; keeping the two apart matters, because the manipulations of this section are available for one mode and not the other.

Theorem 3.1 (Continuous mapping) *Let $g : \mathbf{R}^k \rightarrow \mathbf{R}^m$ be Borel measurable and write*

$$C_g = \{v \in \mathbf{R}^k : g \text{ is continuous at } v\}$$

for its set of continuity points, which is a Borel set. If $P\{V \in C_g\} = 1$ and $V_n \xrightarrow{P} V$, then $g(V_n) \xrightarrow{P} g(V)$.

The proof is in Section A.6. For almost sure convergence the statement is immediate: on the event where $V_n \rightarrow V$ and $V \in C_g$, which has probability one, continuity of g at V gives $g(V_n) \rightarrow g(V)$. It is the version in probability that requires an argument, because there is no single event on which the sequence converges.

One thing only is being asked beyond measurability: that the limit be a point of continuity with probability one. Nothing further is required of g , and in every application below the limit is a constant, so the requirement reduces to checking that one point.

For most functions the requirement is vacuous. Sums, differences, products, maxima and minima are continuous on all of $\mathbf{R}^k$, so $C = \mathbf{R}^k$ and the theorem applies with no side condition: if a vector of sample averages converges, so does any polynomial in its coordinates.

Ratios are the first real case. The function $g(a, b) = a/b$ is continuous at every point with $b \neq 0$; where $b = 0$ it has to be assigned some value, and any Borel choice serves, since beyond measurability the theorem asks nothing of g away from its continuity set. Consistency of a ratio of averages therefore requires the population denominator to be nonzero, and the requirement cannot

be dropped. Take the deterministic sequence $V_n = 1/n$, which converges to 0, and $g(v) = 1/v$; then $g(V_n) = n$ diverges. Matrix inversion behaves the same way. The map $A \mapsto A^{-1}$ is continuous on the set of matrices with nonzero determinant, which is open, so the inverse of a sample second-moment matrix converges to the inverse of the population one exactly when the latter is invertible.

Indicators are the case the rest of this chapter turns on. For fixed t the function $g(v) = \mathbf{1}\{v \leq t\}$ is continuous everywhere except at t , so $C_g = \mathbf{R} \setminus \{t\}$ and the theorem applies whenever $P\{V = t\} = 0$. When V has an atom at t the conclusion fails: with $V_n = t + 1/n$ we have $\mathbf{1}\{V_n \leq t\} = 0$ for every n , while $\mathbf{1}\{V \leq t\} = 1$. What the theorem allows, then, is any set of discontinuities the limit misses, however large that set is.

Example 3.1 (The coin) Let $W_1, \dots, W_n$ be i.i.d. Bernoulli(q) as in Example 2.1. The variance of a single draw is $q(1 - q)$, and the natural estimator replaces q by $\bar{W}_n$. Since $\bar{W}_n \xrightarrow{P} q$ by the weak law and $g(v) = v(1 - v)$ is continuous on all of $\mathbf{R}$, Theorem 3.1 gives

$$\bar{W}_n(1 - \bar{W}_n) \xrightarrow{P} q(1 - q) .$$

No separate argument is needed, and none of the closed-form structure of the binomial is used. ■

3.2 The Analog Principle

Example 3.1 followed a recipe worth stating in general. A parameter is a feature of the distribution, written $\theta(P)$; the data determine a distribution of their own; and the estimator is the same feature evaluated at that distribution.

Definition 3.2 Let $W_1, \dots, W_n$ be i.i.d. with distribution P on $\mathbf{R}^k$. The empirical distribution $\hat{P}_n$ places mass $1/n$ at each observation, so that

$$\hat{P}_n\{A\} = \frac{1}{n} \sum_{i \leq n} \mathbf{1}\{W_i \in A\} \quad (3.2)$$

for every Borel $A \subseteq \mathbf{R}^k$. When $k = 1$, its distribution function is the empirical c.d.f.

$$\hat{F}_n(t) = \hat{P}_n\{(-\infty, t]\} = \frac{1}{n} \sum_{i \leq n} \mathbf{1}\{W_i \leq t\} , \quad t \in \mathbf{R} . \quad (3.3)$$

Two features of $\hat{P}_n$ are used constantly. It is random: each realization of the sample produces a probability distribution, and the map from samples to

distributions is the object under study. And integration against it is averaging, since (3.2) gives

$$\int g(w) d\hat{P}_n(w) = \frac{1}{n} \sum_{i \leq n} g(W_i) \quad (3.4)$$

for every Borel g for which the right-hand side is defined.

The *analog principle*, or *plug-in principle*, is to estimate $\theta(P)$ by $\theta(\hat{P}_n)$. Its content is entirely in the choice of θ : once the parameter is written as a feature of the distribution, (3.4) determines the estimator. The mean of P gives the sample mean. The distribution function gives $\hat{F}_n$. The variance,

$$\sigma^2(P) = \int w^2 dP(w) - \left(\int w dP(w) \right)^2, \quad (3.5)$$

gives

$$\begin{aligned} \hat{\sigma}_n^2 &= \sigma^2(\hat{P}_n) = \int w^2 d\hat{P}_n(w) - \left(\int w d\hat{P}_n(w) \right)^2 \\ &= \frac{1}{n} \sum_{i \leq n} W_i^2 - \bar{W}_n^2, \end{aligned} \quad (3.6)$$

which is the usual estimator.

Consistency of (3.6) is Section 3.1 and nothing else. Suppose $E[W^2] < \infty$ and set

$$A_n = \frac{1}{n} \sum_{i \leq n} W_i^2, \quad B_n = \bar{W}_n.$$

Both are sample averages of integrable functions of a single observation, so the strong law, Theorem 2.5, gives $A_n \xrightarrow{\text{a.s.}} E[W^2]$ and $B_n \xrightarrow{\text{a.s.}} E[W]$, hence

$$(A_n, B_n) \xrightarrow{P} (E[W^2], E[W])$$

by (3.1) and Lemma 2.2. Since $\hat{\sigma}_n^2 = g(A_n, B_n)$ for $g(a, b) = a - b^2$, which is continuous on $\mathbf{R}^2$, Theorem 3.1 gives

$$\hat{\sigma}_n^2 \xrightarrow{P} E[W^2] - E[W]^2 = \text{Var}[W].$$

In other words, a plug-in estimator that is a continuous function of finitely many sample averages is consistent, and the argument is always this one.

Example 3.2 (The two groups) Let $W = (Y, X)$ with $X \in \{0, 1\}$ and $\pi = P\{X = 1\} \in (0, 1)$, and let $\mu_x = E[Y|X = x]$, the two numbers that Section 1.1 showed the conditional expectation reduces to. Writing

$$\mu_x = \frac{E[Y \mathbf{1}\{X = x\}]}{P\{X = x\}}$$

makes μ_x a ratio of two expectations, and the analog principle returns the group average

$$\hat{\mu}_{n,x} = \frac{n^{-1} \sum_{i \leq n} Y_i \mathbf{1}\{X_i = x\}}{n^{-1} \sum_{i \leq n} \mathbf{1}\{X_i = x\}} .$$

The estimator is undefined when no observation falls in group x , so set $\hat{\mu}_{n,x} = 0$ on that event; the event has probability $(1 - P\{X = x\})^n \rightarrow 0$, so the convention has no effect on the limit. With it, $\hat{\mu}_{n,x} = g(A_{n,x}, B_{n,x})$ for the averages $A_{n,x} = n^{-1} \sum_{i \leq n} Y_i \mathbf{1}\{X_i = x\}$ and $B_{n,x} = n^{-1} \sum_{i \leq n} \mathbf{1}\{X_i = x\}$ and the Borel function $g(a, b) = a/b$ for $b \neq 0$, $g(a, 0) = 0$, which is continuous at every point with $b \neq 0$. If $E[|Y|] < \infty$ the two averages converge to $E[Y \mathbf{1}\{X = x\}]$ and $P\{X = x\}$, and that limit is a point of continuity because $P\{X = x\} > 0$. Theorem 3.1 therefore gives

$$\hat{\mu}_{n,1} - \hat{\mu}_{n,0} \xrightarrow{P} \mu_1 - \mu_0 .$$

The limit is a comparison of two population means, and Chapter 9 carries this estimator from here to a confidence interval. ■

Everything so far has been pointwise: one functional at a time, one value of t at a time. That is the right level of generality for the sample variance, whose definition names no free argument. It is the wrong level for an object evaluated at a point the data select, and the rest of the chapter is about those.

3.3 Glivenko–Cantelli

The empirical c.d.f. is the natural first example of what uniform convergence buys. Start with t fixed. Then $\hat{F}_n(t)$ is the sample average of the i.i.d. Bernoulli variables $\mathbf{1}\{W_i \leq t\}$, whose common mean is $F(t)$, so the strong law gives

$$\hat{F}_n(t) \xrightarrow{\text{a.s.}} F(t) , \quad (3.7)$$

and Hoeffding's inequality supplies the rate $P\{|\hat{F}_n(t) - F(t)| \geq \varepsilon\} \leq 2e^{-2n\varepsilon^2}$, exactly as for the coin at (2.3).

Finitely many values of t cost nothing more. For $t_1 < \dots < t_m$ the vector $(\hat{F}_n(t_1), \dots, \hat{F}_n(t_m))$ is a vector of sample averages, so it converges to $(F(t_1), \dots, F(t_m))$ by (3.1), and any continuous function of those m values is consistent for the same function of F by Theorem 3.1. The estimated probability of an interval, $\hat{F}_n(b) - \hat{F}_n(a)$, is the simplest instance.

The step Sections 3.1 and 3.2 cannot take is to a value of t that is not fixed in advance, and two familiar objects require it. The Kolmogorov–Smirnov statistic $\sup_t |\hat{F}_n(t) - F(t)|$ is the obvious one, since the supremum is the

statistic. The sample quantile is the instructive one: for $\tau \in (0, 1)$ the plug-in estimator of $\theta_\tau = F^{-1}(\tau)$ is

$$\hat{\theta}_{n,\tau} = \inf\{t \in \mathbf{R} : \hat{F}_n(t) \geq \tau\} , \quad (3.8)$$

which reads $\hat{F}_n$ near θ_τ — but at which t it reads it is decided by the data, and (3.7) at any nominated t says nothing about the value at a random one. Controlling $\hat{F}_n$ at every t at once removes the difficulty, and for the empirical c.d.f. that control is available.

Theorem 3.2 (Glivenko–Cantelli) *Let $W_1, W_2, \dots$ be i.i.d. real-valued with distribution function F , and write $\|\hat{F}_n - F\|_\infty = \sup_{t \in \mathbf{R}} |\hat{F}_n(t) - F(t)|$. Then $\|\hat{F}_n - F\|_\infty \xrightarrow{a.s.} 0$.*

PROOF. The strategy is to replace the supremum over $\mathbf{R}$ by a maximum over finitely many points, at a cost that can be made arbitrarily small. Monotonicity of $\hat{F}_n$ and F is what makes the replacement possible: between two grid points neither function can move by more than F moves across the cell, up to the deviations at the endpoints.

The grid. Fix $\varepsilon > 0$ and let k be the smallest integer with $k\varepsilon > 1$. Set $t_0 = -\infty$, $t_k = +\infty$, and for $1 \leq i \leq k-1$,

$$t_i = \inf\{t \in \mathbf{R} : F(t) \geq i\varepsilon\} ,$$

reading the infimum of the empty set as $+\infty$. The points are nondecreasing, and consecutive ones may coincide; a cell with $t_{i-1} = t_i$ is empty and contributes nothing. They satisfy

$$F(t_i-) - F(t_{i-1}) \leq \varepsilon \quad \text{for every } i \leq k , \quad (3.9)$$

where $F(t-)$ denotes the left limit and we read $F(t_0) = 0$, $F(t_k-) = 1$. Indeed $F(t) < i\varepsilon$ for every $t < t_i$, so $F(t_i-) \leq i\varepsilon$, while right-continuity of F gives $F(t_{i-1}) \geq (i-1)\varepsilon$; subtracting yields (3.9). Any point at which F jumps by more than ε is one of the t_i , which is why the construction survives atoms.

The cells. Fix i and let $t_{i-1} \leq t < t_i$. Since $\hat{F}_n$ and F are nondecreasing,

$$\begin{aligned} \hat{F}_n(t) - F(t) &\leq \hat{F}_n(t_i-) - F(t_{i-1}) \\ &= \{\hat{F}_n(t_i-) - F(t_i-)\} + \{F(t_i-) - F(t_{i-1})\} , \\ \hat{F}_n(t) - F(t) &\geq \hat{F}_n(t_{i-1}) - F(t_i-) \\ &= \{\hat{F}_n(t_{i-1}) - F(t_{i-1})\} - \{F(t_i-) - F(t_{i-1})\} , \end{aligned}$$

and the second brace is at most ε in both lines by (3.9). Taking the supremum over $t \in \mathbf{R}$ and then over the finitely many cells,

$$\|\hat{F}_n - F\|_\infty \leq \max_{0 \leq i \leq k} \left\{ |\hat{F}_n(t_i) - F(t_i)| \vee |\hat{F}_n(t_i-) - F(t_i-)| \right\} + \varepsilon . \quad (3.10)$$

The finitely many points. For each fixed t , both $\hat{F}_n(t) \rightarrow F(t)$ and $\hat{F}_n(t-) \rightarrow F(t-)$ almost surely, by the strong law applied to the i.i.d. variables $\mathbf{1}\{W_i \leq t\}$ and $\mathbf{1}\{W_i < t\}$, whose means are $F(t)$ and $F(t-)$. A finite union of null sets is null, so outside a single null set the maximum in (3.10) tends to zero, and therefore $\limsup_n \|\hat{F}_n - F\|_\infty \leq \varepsilon$ almost surely.

Removing ε . Apply the above to $\varepsilon = 1/m$ for $m = 1, 2, \dots$ and intersect the countably many almost sure events, which is again almost sure. On that event $\limsup_n \|\hat{F}_n - F\|_\infty \leq 1/m$ for every m , hence the limit is zero. ■

Cf. *van der Vaart [1998, Theorem 19.1, p. 266]* gives this argument; *Billingsley [1995, Theorem 20.6, p. 269]* builds the grid from quantiles instead.

Figure 3.1(a) is the picture. The staircase converges to F everywhere at once, and the quantity driven to zero is the largest vertical gap, not the gap at any nominated point.

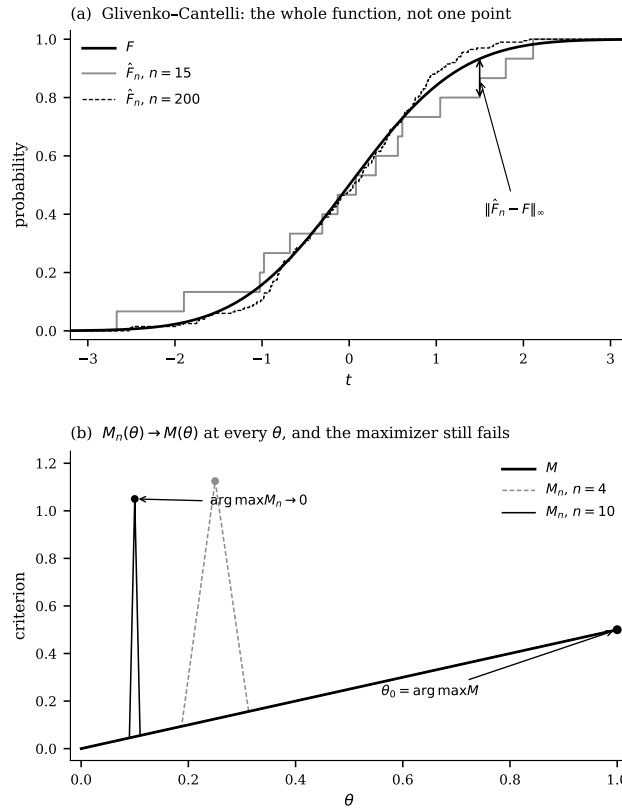


FIGURE 3.1: Panel (a) shows the empirical c.d.f. of a standard normal sample at $n = 15$ and $n = 200$, with the largest vertical gap marked for the smaller sample; Theorem 3.2 drives that gap to zero, not merely the gap at a fixed t . Panel (b) draws the criterion functions of Example 3.3.

The device has a name, since Section 3.4 uses it again. *Discretize, then bound the finitely many pieces at once*: a supremum over an uncountable index set is sandwiched between values at finitely many indices, chosen so that the sandwich is tight to within ε , and those are controlled by a law of large numbers and the union bound. What varies is the source of the tightness — monotonicity here, continuity and an integrable bound in Section 3.4.

Remark 3.1 The inequality of Dvoretzky, Kiefer and Wolfowitz states that the pointwise Hoeffding bound holds for the supremum as well:

$$P\{\|\hat{F}_n - F\|_\infty \geq \varepsilon\} \leq 2e^{-2n\varepsilon^2} \quad \text{for every } n \text{ and every } F. \quad (3.11)$$

Uniformity over t is therefore free, and (3.11) turns Theorem 3.2 into a confidence band for F valid at every n . van der Vaart [1998, p. 268] records that the constant 2 cannot be improved. ■

3.4 Uniform Laws of Large Numbers

Now replace the index t by a parameter. Throughout, $\Theta \subseteq \mathbf{R}^d$ is a parameter space, $m(w, \theta)$ is a fixed function, and

$$M_n(\theta) = \frac{1}{n} \sum_{i \leq n} m(W_i, \theta), \quad M(\theta) = E[m(W, \theta)], \quad (3.12)$$

with the estimator $\hat{\theta}_n$ a maximizer of M_n over Θ and the estimand θ_0 a maximizer of M . Remark 2.5 wrote the same skeleton for a minimizer, and nothing changes, since minimizing M_n is maximizing $-M_n$.

For each fixed θ the weak law gives $M_n(\theta) \xrightarrow{P} M(\theta)$ whenever $E[|m(W, \theta)|] < \infty$. The next example shows that this is not enough. It is deterministic, so no probabilistic subtlety is hiding in it.

Example 3.3 (A runaway maximizer) Let $\Theta = [0, 1]$, let $M(\theta) = \theta/2$, and for $n \geq 3$ let

$$M_n(\theta) = \frac{\theta}{2} + \left(1 - n^2|\theta - n^{-1}|\right)^+,$$

a triangular spike of height one centred at $1/n$ with half-width $1/n^2$, added to M . Fix $\theta \in (0, 1]$: once n is large enough that $1/n + 1/n^2 < \theta$, the spike vanishes at θ and $M_n(\theta) = M(\theta)$. At $\theta = 0$ it also vanishes, since $1/n - 1/n^2 > 0$ for $n \geq 2$. So $M_n(\theta) \rightarrow M(\theta)$ at every point of Θ . Yet $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| = 1$ for every n , the maximizer of M_n is $1/n$, and the maximizer of M is 1. The estimator converges to the wrong place, and to a place that maximizes nothing. Figure 3.1(b) draws it. ■

A spike of fixed height and shrinking width escapes pointwise detection while dominating the maximum. The following display says which quantity rules it out.

Remark 3.2 (Why the supremum is the right object) Write the gap between the population criterion at θ_0 and at the estimator as a sum of three pieces:

$$\begin{aligned} M(\theta_0) - M(\hat{\theta}_n) &= \{M(\theta_0) - M_n(\theta_0)\} + \{M_n(\theta_0) - M_n(\hat{\theta}_n)\} \\ &\quad + \{M_n(\hat{\theta}_n) - M(\hat{\theta}_n)\} . \end{aligned} \quad (3.13)$$

The middle term is at most zero, because $\hat{\theta}_n$ maximizes M_n . The outer two are each bounded by $\sup_{\theta} |M_n(\theta) - M(\theta)|$; the third bound matters only because the supremum runs over all of Θ , not at a fixed θ . There the argument is evaluated at $\hat{\theta}_n$, chosen after the data are seen, where pointwise convergence at every θ is no help. Sending the supremum to zero therefore forces $M(\hat{\theta}_n) \xrightarrow{P} M(\theta_0)$, and Example 3.3 shows what fails when it does not. ■

Definition 3.3 M_n converges to M uniformly in probability over Θ if

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0 ,$$

and uniformly almost surely if the same supremum tends to zero almost surely.

The uniform law of large numbers delivers Definition 3.3 from conditions that can be checked.

Theorem 3.3 (Uniform law of large numbers) Let $W_1, W_2, \dots$ be i.i.d. with distribution P , let $\Theta \subseteq \mathbf{R}^d$ be compact, let $m : \mathbf{R}^k \times \Theta \rightarrow \mathbf{R}$ be such that $w \mapsto m(w, \theta)$ is Borel measurable for every $\theta \in \Theta$, and suppose that

- (i) $\theta \mapsto m(w, \theta)$ is continuous on Θ for P -almost every w ;
- (ii) there is a function d with $|m(w, \theta)| \leq d(w)$ for every $\theta \in \Theta$ and P -almost every w , and $E[d(W)] < \infty$.

Then M is finite and continuous on Θ , and with M_n and M as in (3.12),

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{a.s.} 0 .$$

The function d in (ii) is an *envelope* for the family $\{m(\cdot, \theta) : \theta \in \Theta\}$: a single integrable function dominating all of them at once. The proof is the discretization of Section 3.3 run again, with the grid a finite collection of balls in Θ and the tightness coming from continuity rather than monotonicity.

PROOF. Finiteness is immediate from (ii), since $|M(\theta)| \leq E[d(W)] < \infty$ for every θ . Fix $\varepsilon > 0$.

Step 1: brackets from small balls. For $\theta \in \Theta$ and $\rho > 0$ write $B(\theta, \rho) = \{\vartheta \in \Theta : \|\vartheta - \theta\| < \rho\}$ and define

$$u(w, \theta, \rho) = \sup_{\vartheta \in B(\theta, \rho)} m(w, \vartheta) , \quad \ell(w, \theta, \rho) = \inf_{\vartheta \in B(\theta, \rho)} m(w, \vartheta) .$$

Both are measurable in w : by (i) the supremum over the ball agrees, for almost every w , with the supremum over a countable dense subset of it, and a countable supremum of measurable functions is measurable. Both are bounded in absolute value by $d(w)$, hence integrable.

As ρ decreases the ball shrinks, so $u(w, \theta, \rho)$ is nonincreasing and $\ell(w, \theta, \rho)$ nondecreasing in ρ , and by the continuity in (i),

$$u(w, \theta, \rho) \downarrow m(w, \theta) \quad \text{and} \quad \ell(w, \theta, \rho) \uparrow m(w, \theta) \quad \text{as } \rho \downarrow 0$$

for almost every w . Both are dominated by d , so applying dominated convergence, Fact A.2, along any sequence $\rho_1 > \rho_2 > \dots \downarrow 0$ — which suffices, because both quantities are monotone in ρ — gives $E[u(W, \theta, \rho)] \downarrow M(\theta)$ and $E[\ell(W, \theta, \rho)] \uparrow M(\theta)$. Hence for each $\theta \in \Theta$ there is a radius $\rho(\theta) > 0$ with

$$E[u(W, \theta, \rho(\theta))] \leq M(\theta) + \varepsilon \quad \text{and} \quad E[\ell(W, \theta, \rho(\theta))] \geq M(\theta) - \varepsilon . \quad (3.14)$$

Continuity of M follows from the same display: for $\vartheta \in B(\theta, \rho)$ we have $\ell(\cdot, \theta, \rho) \leq m(\cdot, \vartheta) \leq u(\cdot, \theta, \rho)$ pointwise, so $|M(\vartheta) - M(\theta)| \leq E[u(W, \theta, \rho)] - E[\ell(W, \theta, \rho)]$, and the right-hand side can be made as small as desired by choosing ρ small.

Step 2: finitely many balls suffice. The balls $\{B(\theta, \rho(\theta)) : \theta \in \Theta\}$ form an open cover of Θ , which is compact, so finitely many of them cover it: there are $\theta_1, \dots, \theta_K$ with $\Theta = \bigcup_{j \leq K} B_j$, where $B_j = B(\theta_j, \rho(\theta_j))$. Write $u_j(w) = u(w, \theta_j, \rho(\theta_j))$ and $\ell_j(w) = \ell(w, \theta_j, \rho(\theta_j))$.

Step 3: the law of large numbers on each piece. Each u_j and each ℓ_j is an integrable function of a single observation, so Theorem 2.5 applies to the i.i.d. variables $u_j(W_1), u_j(W_2), \dots$ and gives

$$\frac{1}{n} \sum_{i \leq n} u_j(W_i) \xrightarrow{\text{a.s.}} E[u_j(W)] , \quad \frac{1}{n} \sum_{i \leq n} \ell_j(W_i) \xrightarrow{\text{a.s.}} E[\ell_j(W)] , \quad (3.15)$$

for each $j \leq K$; and K is finite, so all $2K$ statements hold outside a single null set.

Step 4: assembling. Fix $\theta \in \Theta$ and pick j with $\theta \in B_j$. Then $\ell_j(W_i) \leq m(W_i, \theta) \leq u_j(W_i)$ for every i , so $M_n(\theta) \leq n^{-1} \sum_{i \leq n} u_j(W_i)$; and $M(\theta) \geq$

$E[\ell_j(W)] \geq M(\theta_j) - \varepsilon$ by (3.14), while $E[u_j(W)] \leq M(\theta_j) + \varepsilon$. Therefore

$$\begin{aligned} M_n(\theta) - M(\theta) &\leq \frac{1}{n} \sum_{i \leq n} u_j(W_i) - E[\ell_j(W)] \\ &= \left\{ \frac{1}{n} \sum_{i \leq n} u_j(W_i) - E[u_j(W)] \right\} \\ &\quad + \left\{ E[u_j(W)] - E[\ell_j(W)] \right\}, \end{aligned}$$

whose second brace is at most 2ε . The symmetric argument with ℓ_j bounds $M_n(\theta) - M(\theta)$ from below by the corresponding deviation minus 2ε . Since the bound holds for every θ and the index j ranges over a finite set,

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \leq D_n + 2\varepsilon, \quad D_n = \max_{j \leq K} \{A_{n,j} \vee B_{n,j}\}, \quad (3.16)$$

where

$$A_{n,j} = \left| \frac{1}{n} \sum_{i \leq n} u_j(W_i) - E[u_j(W)] \right|, \quad B_{n,j} = \left| \frac{1}{n} \sum_{i \leq n} \ell_j(W_i) - E[\ell_j(W)] \right|.$$

By (3.15) the maximum D_n tends to zero almost surely, so the supremum on the left of (3.16) has limit superior at most 2ε , almost surely. Applying this along $\varepsilon = 1/m$, $m = 1, 2, \dots$, and intersecting countably many almost sure events completes the proof. ■

Cf. van der Vaart [1998, Example 19.8, p. 272] states the result in this form and obtains it from bracketing, his Theorem 19.4, p. 270. The covering device is the one Jiang [2010, §1.4, pp. 9–11] runs inside Wald's proof.

Remark 3.3 What the proof needs is that the family $\{m(\cdot, \theta) : \theta \in \Theta\}$ be approximable by finitely many brackets. The three hypotheses are one way to buy that. Continuity makes the brackets close as the balls shrink, and it is continuity in the *parameter* only: nothing is assumed about $w \mapsto m(w, \theta)$ beyond measurability, so the least-absolute-deviation criterion $m(w, \theta) = -|y - x'\theta|$ is covered even though $|\cdot|$ is not differentiable. Compactness turns the cover by small balls into a finite one, which makes it a condition on the index set and not on the estimator. The envelope does two jobs, making dominated convergence available in Step 1 and the strong law available in Step 3.

Other routes buy the same thing. Theorem 3.2 is one: its index set is all of $\mathbf{R}$, which is not compact, and monotonicity supplies the brackets that continuity supplies here. Families too large for either device belong to empirical process theory, where the hypothesis is a size condition on the family itself and the index set plays no role. The conclusion above is almost sure convergence because Step 3 invokes the strong law; the same argument with the weak law gives the version in probability. ■

Example 3.4 (The coin's log-likelihood) Take the coin of Example 2.1 and let

$$m(w, \theta) = w \log \theta + (1 - w) \log(1 - \theta) ,$$

the log-likelihood contribution of a Bernoulli observation, on $\Theta = [\delta, 1 - \delta]$ for a fixed $\delta \in (0, 1/2)$. Condition (i) holds because $\theta \mapsto m(w, \theta)$ is continuous on Θ for each of the two values of w , and (ii) holds with the constant envelope $d(w) = \log(1/\delta)$, since $|\log \theta|$ and $|\log(1 - \theta)|$ are at most $\log(1/\delta)$ on Θ . Theorem 3.3 therefore gives

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\text{a.s.}} 0 , \quad M(\theta) = q \log \theta + (1 - q) \log(1 - \theta) .$$

The restriction to $[\delta, 1 - \delta]$ is not cosmetic. Continuity holds on all of $(0, 1)$, but there the envelope condition fails, since $|m(w, \theta)|$ is unbounded as θ approaches either endpoint, and the parameter space is not compact. ■

Where we will use this

- **In this class.** Theorem 3.3 verifies the uniform convergence condition assumed throughout the second half of the quarter, where extremum estimators are shown to be consistent by the argument of Remark 3.2. The word *uniform* returns in Chapter 8 with a different index, over distributions rather than parameters.
- **Later in the sequence.** The winter quarter estimates conditional expectations with function classes too rich for a single integrable envelope, and the tools that replace Theorem 3.3 measure the size of the class by entropy. In the spring, the empirical distribution of Definition 3.2 is what the bootstrap resamples from.
- **In recent research.** Estimators built on flexible first-stage learners are the current version of the same problem: the nuisance function is drawn from a class too large for any envelope argument, and rate conditions and sample splitting replace the uniform law.

3.5 Key Concepts

- **The analog principle.** Write the parameter as a feature of P , replace P by the empirical distribution.
- **The continuous mapping theorem.** $V_n \xrightarrow{P} V$ and g continuous at V with probability one give $g(V_n) \xrightarrow{P} g(V)$.

- **Uniformity is about where the object is evaluated.** A statement holding at each fixed index is silent at an index chosen after seeing the data. Every use of uniformity in this course traces back to that sentence.
- **Glivenko–Cantelli.** $\|\hat{F}_n - F\|_\infty \rightarrow 0$ almost surely: the empirical c.d.f. converges as a function, not merely at each point.
- **The uniform law of large numbers.** On a compact Θ , with $m(w, \cdot)$ continuous and an integrable envelope, $\sup_\theta |M_n(\theta) - M(\theta)| \rightarrow 0$ almost surely.

3.6 Concluding Remarks

The uniform laws of this chapter follow [van der Vaart \[1998\]](#), whose Chapter 19 is the place to read further: Theorem 19.1 is the Glivenko–Cantelli argument used here, and the bracketing and entropy conditions of his Sections 19.2 and 19.4 replace our envelope when the family of functions is too large for one. [Jiang \[2010, §1.4\]](#) runs the covering argument of Theorem 3.3 directly on a likelihood ratio, presented as an illustration of the ε – δ technique. [Billingsley \[1995, §20\]](#) gives Glivenko–Cantelli in a measure-theoretic setting.

Figure 3.1 is generated by `fig-uniform.py`, in this chapter’s folder under `code/`.

3.7 Problems

Problem 3.1 Let $W_1, \dots, W_n$ be i.i.d. with $E[W^2] < \infty$ and $\text{Var}[W] > 0$, and let $\hat{\sigma}_n^2$ be as in (3.6).

1. Show that the usual estimator $\{n/(n-1)\}\hat{\sigma}_n^2$ is also consistent for $\text{Var}[W]$, and explain in one sentence why the divisor cannot matter asymptotically.
2. Let $\hat{\gamma}_n = n^{-1} \sum_{i \leq n} (W_i - \bar{W}_n)^3 / \hat{\sigma}_n^3$ be the plug-in estimator of the skewness, with the convention $\hat{\gamma}_n = 0$ on the event $\{\hat{\sigma}_n^2 = 0\}$, where the ratio is undefined. State the additional moment condition your argument needs — part (a) did not require it — and prove consistency using (3.1) and Theorem 3.1. Where exactly does the assumption $\text{Var}[W] > 0$ enter, and what does it have to do with the continuity set in Theorem 3.1?
3. Give an example of a feature $\theta(P)$ for which the plug-in $\theta(\hat{P}_n)$ is not consistent, and say which hypothesis of Theorem 3.1 fails.

Problem 3.2 Let $W_1, W_2, \dots$ be i.i.d. with distribution function F , fix $\tau \in (0, 1)$, and suppose F is continuous and strictly increasing on a neighbourhood of $\theta_\tau = F^{-1}(\tau)$. Let $\hat{\theta}_{n,\tau}$ be as in (3.8).

1. Fix $\varepsilon > 0$ and let $\eta = \min\{F(\theta_\tau + \varepsilon) - \tau, \tau - F(\theta_\tau - \varepsilon)\}$. Show that $\eta > 0$, and that $\|\hat{F}_n - F\|_\infty < \eta$ implies $|\hat{\theta}_{n,\tau} - \theta_\tau| \leq \varepsilon$.
2. Conclude from Theorem 3.2 that $\hat{\theta}_{n,\tau} \xrightarrow{\text{a.s.}} \theta_\tau$.
3. Explain, in two or three sentences, why the pointwise statement (3.7) at the single point $t = \theta_\tau$ does not suffice for part (b).

Problem 3.3 Let $W = (Y, X)$ with $X \in \mathbf{R}^d$, let $\Theta \subseteq \mathbf{R}^d$ be compact, and consider the two criteria

$$m_{\text{LS}}(w, \theta) = -(y - x'\theta)^2, \quad m_{\text{LAD}}(w, \theta) = -|y - x'\theta|.$$

1. For each, exhibit an envelope d as in Theorem 3.3(ii) and state the moment condition on (Y, X) making it integrable. Which of the two requires the weaker condition, and why?
2. Verify the continuity condition for m_{LAD} , and say why the fact that $|\cdot|$ is not differentiable is irrelevant to Theorem 3.3.
3. Now take $\Theta = \mathbf{R}^d$. Show that no integrable envelope exists for m_{LS} unless $X = 0$ almost surely, and explain which steps of the proof of Theorem 3.3 this breaks.

Problem 3.4 Return to Example 3.3, and modify it so that the failure is stochastic rather than deterministic. Let $U_1, U_2, \dots$ be i.i.d. Uniform $[0, 1]$, let $\Theta = [0, 1]$, and set

$$m(u, \theta) = \frac{\theta}{2} + \mathbf{1}\{u \leq \theta\}, \quad \text{so that} \quad M_n(\theta) = \frac{\theta}{2} + \hat{F}_n(\theta).$$

1. Compute $M(\theta)$ and its maximizer over Θ .
2. Show that Theorem 3.3 does not apply, by identifying which of its two hypotheses fails. Show nonetheless that $\sup_\theta |M_n(\theta) - M(\theta)| \xrightarrow{\text{a.s.}} 0$, and explain why this is not a contradiction.
3. Deduce that the conditions of Theorem 3.3 are sufficient and not necessary, and say what the example suggests about which of the two hypotheses does the real work.

Bibliography

- P. Billingsley. *Probability and Measure*. Wiley-Interscience, 1995.
- J. Jiang. *Large Sample Techniques for Statistics*. Springer Texts in Statistics. Springer, New York, 2010.
- A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.



4

Convergence in Distribution

The previous two chapters were about whether an estimator settles down at a number. This one is about the shape of what is left over. Every standard error, test and confidence interval is a statement about how far $\hat{\theta}_n$ typically sits from $\theta(P)$, and answering that requires a fourth mode of convergence — one that describes the distribution of the rescaled error rather than the error itself.

Why we cover this

Sampling distributions almost never have closed forms. The distribution of a regression coefficient, a two-stage least squares estimator or a test statistic is unavailable at any finite n , and convergence in distribution is a clear sense in which any of them can be described. The machinery below is what converts that description into arithmetic: rules for adding, multiplying and transforming quantities whose limits are known, and a bookkeeping notation for discarding the terms that do not matter.

4.1 Convergence in Distribution

Definition 4.1 Let V_n , $n \geq 1$, and V be random vectors in $\mathbf{R}^k$ with distribution functions F_n and F . Then V_n converges in distribution to V , written $V_n \xrightarrow{d} V$, if

$$F_n(v) \longrightarrow F(v) \quad \text{at every } v \text{ at which } F \text{ is continuous.}$$

Two features separate this from the modes of Definition 2.1. It is a statement about distributions and not about random variables, so V_n and V need not live on a common probability space; writing $V_n \xrightarrow{d} V$ is an abuse of notation for a statement about F_n and F . And the convergence is not required everywhere.

Example 4.1 (Why continuity points) Let V_n be the constant $1/n$ and V the constant 0. Then $F_n(v) = \mathbf{1}\{v \geq 1/n\}$ and $F(v) = \mathbf{1}\{v \geq 0\}$, and

$F_n(v) \rightarrow F(v)$ at every $v \neq 0$. At $v = 0$ the convergence fails outright: $F_n(0) = 0$ for every n while $F(0) = 1$. Since $1/n \rightarrow 0$ in every other sense available, a definition demanding convergence at every v would be useless, and the point at which it fails is the single discontinuity of F .

The atom is the same obstruction that appeared at Theorem 3.1, where $\mathbf{1}\{v \leq t\}$ could not be passed through a limit when $P\{V = t\} > 0$. Definition 4.1 handles it by excluding those points rather than by assuming they are absent. ■

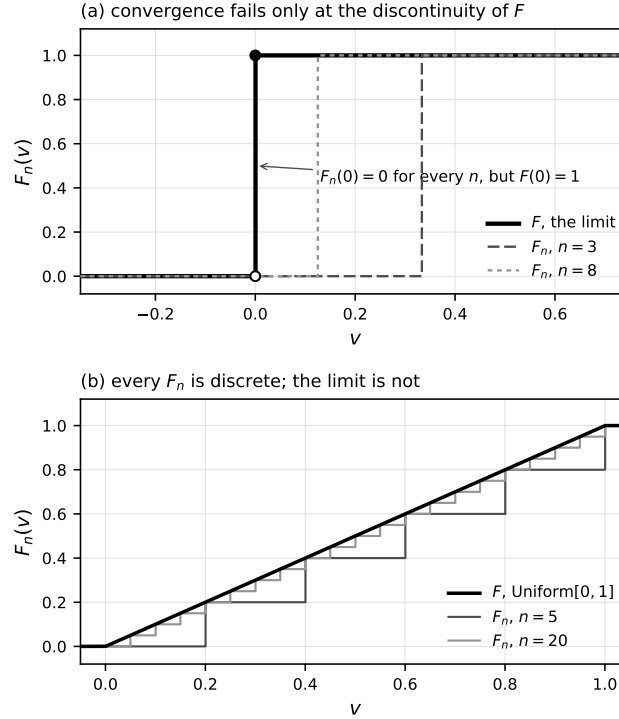


FIGURE 4.1: What the definition does and does not require. Panel (a) is Example 4.1: the distribution functions agree with the limit at every v except $v = 0$, the single point at which F is discontinuous, and there the disagreement is total at every n . Panel (b) takes V_n uniform on $\{1/n, \dots, n/n\}$, the sequence of Problem 4.1. Each V_n is discrete and puts mass $1/n$ at each of its n points, so the mass at every fixed point tends to zero; the distribution functions nonetheless converge to that of a Uniform[0, 1] variable.

Example 4.2 (The coin) Let $W_1, \dots, W_n$ be i.i.d. Bernoulli(q) as in Example 2.1, and write

$$R_n = \frac{\sqrt{n}(\bar{W}_n - \theta(P))}{\sigma(P)} = \frac{\sqrt{n}(\bar{W}_n - q)}{\sqrt{q(1-q)}}. \quad (4.1)$$

This is a *root*: a function of the data and of the distribution P , rather than of

the data alone. Since $n\bar{W}_n$ is exactly $\text{Binomial}(n, q)$, the distribution function of R_n is available in closed form at every n and is a step function with jumps of order $n^{-1/2}$. Chapter 5 shows that $R_n \xrightarrow{d} Z$ with Z standard normal, so while the distribution function of R_n is discontinuous everywhere its jumps sit, the limit is continuous everywhere: no pointwise agreement between the two is being claimed at any finite n . ■

The relation to the modes of Chapter 2 is one implication and one partial converse. Recall that $\xrightarrow{P}$ is defined for variables sharing a probability space.

Lemma 4.1 *Let V_n and V be random vectors.*

- (i) *If $V_n \xrightarrow{P} V$ then $V_n \xrightarrow{d} V$.*
- (ii) *If $V_n \xrightarrow{d} c$ for a constant c , then $V_n \xrightarrow{P} c$.*

PROOF. Part (i) is Theorem 4.3 below by setting (V_n, U_n) in that theorem to (V, V_n) here. For (ii), begin with $k = 1$. The limit has distribution function $F(v) = \mathbf{1}\{v \geq c\}$, continuous everywhere except at c . Fix $\varepsilon > 0$. Then $c - \varepsilon$ and $c + \varepsilon/2$ are continuity points of F , so

$$\begin{aligned} P\{|V_n - c| > \varepsilon\} &\leq F_n(c - \varepsilon) + 1 - F_n(c + \varepsilon/2) \\ &\longrightarrow F(c - \varepsilon) + 1 - F(c + \varepsilon/2) = 0 . \end{aligned}$$

The bound holds for every $\varepsilon > 0$, which is the definition of $V_n \xrightarrow{P} c$.

For $k > 1$, each coordinate projection $v \mapsto v_j$ is continuous on $\mathbf{R}^k$, so Theorem 4.2 gives $V_{n,j} \xrightarrow{d} c_j$ for every $j \leq k$. The scalar case just proved turns each of these into $V_{n,j} \xrightarrow{P} c_j$, and (3.1) assembles the coordinates into $V_n \xrightarrow{P} c$. ■

Cf. van der Vaart [1998, Theorem 2.7, p. 10] collects both directions; Jiang [2010, Theorem 2.9, p. 26].

The converse in (ii) can fail when the limit is nonconstant, and the reason is that convergence in distribution says nothing about where V_n sits relative to V . If Z is standard normal then $-Z$ has the same distribution as Z , so the constant sequence $V_n = Z$ converges in distribution to $-Z$ while $|V_n - (-Z)| = 2|Z|$ never becomes small.

Example 4.3 (Coordinates are not enough) Convergence in probability of a random vector is convergence of each of its coordinates, by the union bound of Chapter 3. The corresponding statement here is false. Let Z be standard normal and set

$$V_n = Z , \quad U_n = (-1)^n Z .$$

Each coordinate converges in distribution: V_n is standard normal for every n , and so is U_n , since $-Z$ and Z have the same distribution. The pair does not converge. For even n the vector (V_n, U_n) is distributed as (Z, Z) , which is supported on the diagonal; for odd n it is distributed as $(Z, -Z)$, supported on the antidiagonal. The two distributions are different and the sequence alternates between them. In particular $V_n + U_n$ equals $2Z$ along the even subsequence and 0 along the odd one. ■

Example 4.3 is the reason the rest of this chapter is not a formality. Adding, multiplying and dividing sequences requires knowing the joint distribution of the limit, and marginal statements do not supply it. Section 4.3 identifies the one case in which they do.

Remark 4.1 (Polya) The definition asks only for pointwise convergence of the distribution functions. When the limit is continuous, that convergence is automatically uniform:

$$\sup_{v \in \mathbf{R}} |F_n(v) - F(v)| \longrightarrow 0.$$

The proof is the grid argument of Theorem 3.2 with the monotonicity of F_n and F doing the same work, and continuity of F replacing the appeal to the strong law. Uniformity in the index, obtained from monotonicity, is therefore a phenomenon that has now appeared twice on two different objects. ■

Cf. van der Vaart [1998, Lemma 2.11, p. 12].

4.2 The Portmanteau Theorem

Distribution functions may be awkward to manipulate. Nothing in Definition 4.1 tells us what happens to $g(V_n)$, or to $V_n + U_n$. The following theorem gives six other forms, each equivalent to the definition and each useful for a different purpose.

Theorem 4.1 (Portmanteau) *Let V_n and V be random vectors in $\mathbf{R}^k$. The following are equivalent.*

- (i) $F_n(v) \rightarrow F(v)$ at every continuity point of F .
- (ii) $E[f(V_n)] \rightarrow E[f(V)]$ for every bounded continuous $f : \mathbf{R}^k \rightarrow \mathbf{R}$.
- (iii) $E[f(V_n)] \rightarrow E[f(V)]$ for every bounded Lipschitz $f : \mathbf{R}^k \rightarrow \mathbf{R}$.
- (iv) $\liminf_n E[f(V_n)] \geq E[f(V)]$ for every nonnegative continuous $f : \mathbf{R}^k \rightarrow \mathbf{R}$.

- (v) $\liminf_n P\{V_n \in G\} \geq P\{V \in G\}$ for every open G .
- (vi) $\limsup_n P\{V_n \in K\} \leq P\{V \in K\}$ for every closed K .
- (vii) $P\{V_n \in B\} \rightarrow P\{V \in B\}$ for every Borel B with $P\{V \in \partial B\} = 0$.

The proof is in Section A.7. A set B as in (vii), whose topological boundary carries no mass under the limit, is a *P-continuity set*; part (i) is the special case $B = (-\infty, v_1] \times \cdots \times (-\infty, v_k]$, whose boundary lies in a union of k hyperplanes through v .

Each form answers a different question. Parts (ii) and (iii) are about expectations of *bounded* functions, and the restriction cannot be dropped: $E[V_n] \rightarrow E[V]$ does not follow from $V_n \xrightarrow{d} V$. Let V_n equal n with probability $1/n$ and 0 otherwise; then $V_n \xrightarrow{d} 0$, while $E[V_n] = 1$ for every n . Convergence of a mean, a variance or any other moment therefore needs a separate hypothesis ruling out mass that escapes to infinity, and uniform integrability, Definition 2.3, is the usual one. Parts (v) and (vi) are about probabilities of regions, as with the rejection region of a test, and the one-sided inequalities are what survive there. Part (vii) delivers both at once when the region is well behaved.

Example 4.4 (Two-sided regions) Suppose $R_n \xrightarrow{d} Z$ with Z standard normal, as in Example 4.2. Take $B = (-z, z)$ for $z > 0$. Its boundary is $\{-z, z\}$, and the standard normal has no atoms, so $P\{Z \in \partial B\} = 0$ and (vii) gives

$$P\{|R_n| < z\} \rightarrow P\{|Z| < z\}.$$

Whether the endpoints are included makes no difference to the limit for the same reason. What would change the answer is an atom of the limit *on the boundary*, not discreteness of the limit as such: a discrete limit placing no mass at $\pm z$ still satisfies $P\{V \in \partial B\} = 0$, and (vii) applies to it unchanged. When the limit does charge an endpoint — a possibility that is not exotic, since R_n itself is discrete at every n — the choice of endpoint convention matters, and only the inequalities (v) and (vi) remain available. ■

4.3 Continuous Mapping, Asymptotic Equivalence and Slutsky

The three results of this section are what make convergence in distribution usable, and all three come out of Theorem 4.1.

Theorem 4.2 (Continuous mapping) Let $g : \mathbf{R}^k \rightarrow \mathbf{R}^m$ be Borel measurable and let C_g be its set of continuity points, as in Theorem 3.1. If $P\{V \in C_g\} = 1$ and $V_n \xrightarrow{d} V$, then $g(V_n) \xrightarrow{d} g(V)$.

PROOF. By Theorem 4.1(vi), it is enough to prove that

$$\limsup_n P\{g(V_n) \in K\} \leq P\{g(V) \in K\}$$

for every closed $K \subseteq \mathbf{R}^m$. If g were continuous everywhere, $g^{-1}(K)$ would be closed and Portmanteau could be applied to it directly. Here $g^{-1}(K)$ need not be closed, so we replace it by its closure. Writing $\bar{A}$ for the closure of A ,

$$\begin{aligned} \limsup_n P\{g(V_n) \in K\} &\stackrel{(1)}{=} \limsup_n P\{V_n \in g^{-1}(K)\} \\ &\stackrel{(2)}{\leq} \limsup_n P\{V_n \in \overline{g^{-1}(K)}\} \\ &\stackrel{(3)}{\leq} P\{V \in \overline{g^{-1}(K)}\} \\ &\stackrel{(4)}{=} P\{V \in \overline{g^{-1}(K)} \cap C_g\} \\ &\stackrel{(5)}{\leq} P\{V \in g^{-1}(K)\} \\ &\stackrel{(6)}{=} P\{g(V) \in K\} . \end{aligned}$$

Here, (1) and (6) follow from the definition of the inverse image, and (2) follows from $g^{-1}(K) \subseteq \overline{g^{-1}(K)}$. The set $\overline{g^{-1}(K)}$ is closed, so (3) follows from the Portmanteau theorem. Equality (4) uses $P\{V \in C_g\} = 1$.

For (5), we claim that

$$\overline{g^{-1}(K)} \cap C_g \subseteq g^{-1}(K) .$$

Indeed, take $v \in \overline{g^{-1}(K)} \cap C_g$. By the definition of closure, there is a sequence $v_j \in g^{-1}(K)$ such that $v_j \rightarrow v$. Thus $g(v_j) \in K$ for every j . Since $v \in C_g$, g is continuous at v , and hence

$$g(v_j) \rightarrow g(v) .$$

Because K is closed, this implies $g(v) \in K$, proving the claimed inclusion. ■

Cf. van der Vaart [1998, Theorem 2.3, p. 7]. Continuity is meant here as continuity of g as a function on $\mathbf{R}^k$ at the point in question, not continuity of its restriction to C_g ; that is what licenses the closure step, where the approximating v_j need not lie in C_g .

Theorem 4.3 (Asymptotic equivalence) *Let U_n and V_n be random vectors on the same probability space. If $V_n \xrightarrow{d} V$ and $U_n - V_n \xrightarrow{P} 0$, then $U_n \xrightarrow{d} V$.*

PROOF. By part (iii) of Theorem 4.1, it is enough to show that

$$E[f(U_n)] \rightarrow E[f(V)]$$

for every bounded Lipschitz function f . Fix such an f , and choose $B, L < \infty$ such that

$$|f(v)| \leq B \quad \text{and} \quad |f(u) - f(v)| \leq L\|u - v\| \quad \text{for all } u, v .$$

Add and subtract $E[f(V_n)]$. The triangle inequality gives

$$\begin{aligned} |E[f(U_n)] - E[f(V)]| &\leq |E[f(U_n)] - E[f(V_n)]| \\ &\quad + |E[f(V_n)] - E[f(V)]| . \end{aligned} \tag{4.2}$$

The second term tends to zero because $V_n \xrightarrow{d} V$ and f is bounded and Lipschitz. It remains to control the first term.

The convergence $U_n - V_n \xrightarrow{P} 0$ does not directly imply convergence of $E[\|U_n - V_n\|]$, so we cannot simply take expectations in the Lipschitz inequality. Instead, fix $\varepsilon > 0$ and split according to whether U_n and V_n are within ε of each other. Let $A_{n,\varepsilon} = \{\|U_n - V_n\| \leq \varepsilon\}$. Then

$$\begin{aligned} |E[f(U_n)] - E[f(V_n)]| &\leq E[|f(U_n) - f(V_n)|] \\ &= E[|f(U_n) - f(V_n)| \mathbf{1}_{\{A_{n,\varepsilon}\}}] \\ &\quad + E[|f(U_n) - f(V_n)| \mathbf{1}_{\{A_{n,\varepsilon}^c\}}] \\ &\leq L\varepsilon + 2B P\{\|U_n - V_n\| > \varepsilon\} . \end{aligned}$$

The probability in the last line tends to zero because $U_n - V_n \xrightarrow{P} 0$. Therefore

$$\limsup_{n \rightarrow \infty} |E[f(U_n)] - E[f(V_n)]| \leq L\varepsilon .$$

This holds for every $\varepsilon > 0$, so

$$|E[f(U_n)] - E[f(V_n)]| \longrightarrow 0 .$$

Both terms on the right-hand side of (4.2) now tend to zero, and so part (iii) of Theorem 4.1 gives $U_n \xrightarrow{d} V$. ■

Cf. van der Vaart [1998, Theorem 2.7(iv), p. 10]; Billingsley [1995, Theorem 25.4, p. 332]. This is the only place in the chapter where the Lipschitz form is used, and it is why Theorem 4.1 carries it.

Asymptotic equivalence is the licence to replace a complicated sequence by a simpler one with the same limit, and it is used in that direction throughout the second half of the sequence: an estimator is expanded, the leading term is a sample average, and the remainder is shown to vanish in probability. It also delivers the one case in which marginal convergence determines a joint limit.

Theorem 4.4 (Slutsky) *Let V_n and V take values in $\mathbf{R}^k$, let U_n take values in $\mathbf{R}^m$, and suppose $V_n \xrightarrow{d} V$ and $U_n \xrightarrow{P} c$ for a constant $c \in \mathbf{R}^m$. Then*

$(V_n, U_n) \xrightarrow{d} (V, c)$. Consequently $V_n + U_n \xrightarrow{d} V + c$ when $m = k$, and when $m = 1$,

$$U_n V_n \xrightarrow{d} cV, \quad \text{and} \quad \frac{V_n}{U_n} \xrightarrow{d} \frac{V}{c} \quad \text{if } c \neq 0.$$

PROOF. The pair (V_n, c) converges in distribution to (V, c) , since for bounded continuous f on $\mathbf{R}^{k+m}$ the map $v \mapsto f(v, c)$ is bounded and continuous and part (ii) applies. The two pairs are asymptotically equivalent, because

$$\|(V_n, U_n) - (V_n, c)\| = \|U_n - c\| \xrightarrow{P} 0,$$

so Theorem 4.3 gives $(V_n, U_n) \xrightarrow{d} (V, c)$. The remaining conclusions follow from Theorem 4.2 applied to $g(v, u) = v + u$, to $g(v, u) = uv$, and to the Borel function equal to v/u for $u \neq 0$ and to 0 for $u = 0$. Only the last needs a side condition: it is continuous at every point with $u \neq 0$, and the limit (V, c) is such a point exactly when $c \neq 0$. ■

Cf. van der Vaart [1998, Lemma 2.8, p. 11]; Jiang [2010, Theorem 2.13, p. 30]. Deriving it from Theorems 4.3 and 4.2 rather than directly is what makes the role of the constant visible.

Remark 4.2 (What the constant is doing) The proof uses c constant exactly once, in the claim that $(V_n, c) \xrightarrow{d} (V, c)$, and that step is where a random limit would break it. Slutsky's theorem is not false when $U_n \xrightarrow{d} U$ with U random; it is unavailable, because the hypothesis no longer determines the joint distribution of the limit. Example 4.3 is the counterexample in its simplest form: there $V_n \xrightarrow{d} Z$ and $U_n \xrightarrow{d} Z$, while $V_n + U_n$ has no limit at all. If joint convergence $(V_n, U_n) \xrightarrow{d} (V, U)$ is available by some other route, Theorem 4.2 applies directly and nothing more is needed. ■

Example 4.5 (Studentizing the coin) For the coin, $\theta(P) = q$ and $\sigma(P) = \sqrt{q(1-q)}$. We show in Chapter 5 that the standardized root of Example 4.2 satisfies $R_n \xrightarrow{d} Z$. Its denominator depends on the unknown q , and replacing $\sigma(P)$ by the plug-in estimator of Example 3.1 produces the *studentized root*

$$\tilde{R}_n = \frac{\sqrt{n}(\bar{W}_n - q)}{\sqrt{\bar{W}_n(1 - \bar{W}_n)}}, \quad (4.3)$$

set equal to zero on the event $\{\bar{W}_n \in \{0, 1\}\}$, where the denominator vanishes. That event has probability $q^n + (1-q)^n$, which tends to zero, so the convention cannot affect any limit below.

Both R_n and $\tilde{R}_n$ depend on q as well as on the data, so neither is computable from a sample. What changes is where the dependence sits: $\tilde{R}_n$ involves the parameter only through the centring, while R_n involves it in the denominator as well.

Writing the studentized root as the standardized one times a correction isolates the only new ingredient:

$$\tilde{R}_n = R_n \cdot \frac{\sqrt{q(1-q)}}{\sqrt{\bar{W}_n(1-\bar{W}_n)}}.$$

Example 3.1 gives $\bar{W}_n(1-\bar{W}_n) \xrightarrow{P} q(1-q)$, and the map $v \mapsto \{q(1-q)/v\}^{1/2}$, extended by zero for $v \leq 0$, is Borel on $\mathbf{R}$ and continuous at every $v > 0$; the limit $q(1-q)$ is such a point because $q \in (0, 1)$. Theorem 3.1 therefore sends the correction to 1 in probability, and Theorem 4.4 gives $\tilde{R}_n \xrightarrow{d} Z$. ■

4.4 Orders in Probability

The manipulations above are always performed on expansions with several terms, most of which are discarded. The following notation makes the discarding explicit and, more to the point, makes it checkable.

Definition 4.2 Let V_n be random vectors and a_n positive constants.

(i) V_n is bounded in probability, written $V_n = O_P(1)$, if for every $\varepsilon > 0$ there is $M < \infty$ with $\sup_n P\{\|V_n\| > M\} < \varepsilon$.

(ii) $V_n = o_P(1)$ means $V_n \xrightarrow{P} 0$.

More generally $V_n = O_P(a_n)$ means $V_n/a_n = O_P(1)$, and $V_n = o_P(a_n)$ means $V_n/a_n = o_P(1)$.

The notation is an abuse: $O_P(1)$ is not an object but a property, and $V_n = O_P(1)$ should be read as " V_n has the property". The equality sign is likewise not symmetric. Each occurrence of $o_P(1)$ stands for *some* sequence tending to zero in probability, and two occurrences in the same argument need not stand for the same one, so occurrences may not be cancelled against each other: $A_n = B_n + o_P(1)$ and $C_n = D_n + o_P(1)$ say nothing about $A_n - C_n$ beyond what the rules of Lemma 4.3 deliver. Rearrangements that follow from those rules are legitimate — $A_n = B_n + o_P(1)$ does give $B_n = A_n + o_P(1)$, since the negative of a null sequence is null — and rearrangements that treat $o_P(1)$ as a fixed quantity are not.

Lemma 4.2 If $V_n \xrightarrow{d} V$ then $V_n = O_P(1)$.

PROOF. Fix $\varepsilon > 0$. The set of M with $P\{\|V\| = M\} > 0$ is at most countable, so there are arbitrarily large M for which $\{v : \|v\| \leq M\}$ is a P -continuity set, and for such an M part (vii) of Theorem 4.1 gives $P\{\|V_n\| > M\} \rightarrow$

$P\{\|V\| > M\}$. Choose M large enough that $P\{\|V\| > M\} < \varepsilon/2$; then $P\{\|V_n\| > M\} < \varepsilon$ for all n beyond some n_0 . Enlarging M finitely often to cover $n \leq n_0$ completes the proof. ■

Cf. [Jiang \[2010, Lemma 3.16, p. 58\]](#), with sufficient conditions at his [Theorem 3.1, p. 59](#).

Together with [Lemma 4.1](#), this places the two orders relative to everything else: convergence in probability to a constant is $o_P(1)$ after centring, and convergence in distribution is $O_P(1)$. The rules below are then the whole calculus.

Lemma 4.3 (The calculus) *Let $r_n = o_P(1)$, $s_n = o_P(1)$ and $A_n = O_P(1)$, $B_n = O_P(1)$ be real-valued. Then*

- (i) $r_n + s_n = o_P(1)$;
- (ii) $A_n + B_n = O_P(1)$ and $A_n B_n = O_P(1)$;
- (iii) $r_n A_n = o_P(1)$;
- (iv) $r_n + A_n = O_P(1)$;
- (v) $(1 + r_n)^{-1} = O_P(1)$.

PROOF. We prove (iii), which is the only one whose statement is not immediate from the definitions, and which is the rule used most often. Fix $\varepsilon > 0$ and $\delta > 0$. Since $A_n = O_P(1)$, choose $M < \infty$ with $P\{|A_n| > M\} < \varepsilon/2$ for every n . On the complementary event $|r_n A_n| \leq M|r_n|$, so

$$\begin{aligned} P\{|r_n A_n| > \delta\} &\leq P\{|A_n| > M\} + P\{|r_n| > \delta/M\} \\ &< \frac{\varepsilon}{2} + P\{|r_n| > \delta/M\}. \end{aligned} \quad (4.4)$$

The remaining probability tends to zero because $r_n \xrightarrow{P} 0$ and δ/M is a fixed positive number, so it is below $\varepsilon/2$ for all large n . Hence $P\{|r_n A_n| > \delta\} < \varepsilon$ eventually, and as ε and δ were arbitrary, $r_n A_n \xrightarrow{P} 0$.

The order of quantifiers is what carries the argument: M is chosen before δ/M is used as a tolerance for R_n , and it works uniformly in n because $O_P(1)$ is a statement holding for all n at once rather than in the limit. The remaining parts are proved the same way and are left as [Problem 4.4](#). ■

Cf. [Jiang \[2010, Theorem 3.2, p. 61\]](#) states the rules; his [§3.5, p. 62](#), treats the unspecified constant.

Example 4.6 (Reading an expansion) Almost every asymptotic argument in econometrics reduces to a display of the form

$$\sqrt{n}(\hat{\theta}_n - \theta(P)) = \frac{1}{\sqrt{n}} \sum_{i \leq n} \psi_P(W_i) + o_P(1), \quad (4.5)$$

with $E[\psi_P(W)] = 0$. Granting a central limit theorem for the leading term, Theorem 4.3 says the remainder may be discarded and the left-hand side inherits its limit. That is the entire content of the manipulation, and it is why the remainder has to be shown to be $o_P(1)$ rather than merely small: a term that is $O_P(1)$ and not $o_P(1)$ contributes to the limit, and one that is $O_P(n^{-1/2})$ before multiplication by $\sqrt{n}$ need not vanish afterwards. ■

Where we will use this

- **In this class.** Theorem 4.4 and the calculus of Lemma 4.3 are used in every argument from Chapter 5 onward and in every consistency and normality proof of the second half of the quarter, where they appear as steps rather than as results. The representation (4.5) is constructed in Chapter 7, where it is computed for least squares and for a difference of group means. The open- and closed-set forms of Theorem 4.1 are the language in which the uniformity failures of Chapter 8 are stated.
- **Later in the sequence.** Bootstrap validity in the spring is a statement about convergence in distribution of a conditional law, and Theorem 4.1 is what gives that statement content.
- **In recent research.** Inference under weak identification and on boundaries turns on the fact that convergence in distribution can hold at every fixed P while failing uniformly over $\mathbf{P}$; the distinction is invisible without the vocabulary of this chapter.

4.5 Key Concepts

- **Convergence in distribution.** $F_n \rightarrow F$ at continuity points of F . A statement about distributions, not about random variables: V_n and V need not share a probability space.
- **Portmanteau.** Six useful equivalent reformulations of the definition, in terms of expectations of bounded functions and probabilities of sets.
- **Joint vs Marginal.** Marginal convergence in distribution does not give joint convergence; so sums and products need a separate theorem.
- **Slutsky.** If one of the two sequences has a *constant* limit, marginal convergence determines the joint limit, and sums, products and ratios behave as expected.
- **Orders in probability.** $o_P(1)$ is convergence to zero in probability;

$O_P(1)$ is boundedness in probability, uniform in n . Convergence in distribution implies $O_P(1)$. The rules let a remainder be discarded without hand-waving.

4.6 Concluding Remarks

van der Vaart [1998, Chapter 2] is the reference for this material and states Theorem 4.1 in the seven-part form used here, at his Lemma 2.2, p. 6. Billingsley [1995, §25] proves the same equivalences through the Skorokhod representation, which constructs random variables converging almost surely with the prescribed distributions and is worth seeing as an alternative organizing idea. Jiang [2010, Chapter 3] is the best treatment of the O_P and o_P calculus in the reference set.

Figure 4.1 is generated by `fig-convdist.py`, in this chapter's folder under `code/`.

4.7 Problems

Problem 4.1 Let V_n be uniformly distributed on $\{1/n, 2/n, \dots, n/n\}$.

1. Show that $V_n \xrightarrow{d} U$ with U uniform on $[0, 1]$, and identify the set of v at which $F_n(v) \rightarrow F(v)$ fails, if any.
2. The distribution of V_n is discrete for every n and the limit is continuous. Explain why this is consistent with Definition 4.1, and what would go wrong with a definition requiring convergence of probability mass functions.
3. Show that $E[V_n] \rightarrow E[U]$, and say which part of Theorem 4.1 does and does not apply to the function $f(v) = v$.

Problem 4.2 Let $U_1, \dots, U_n$ be independent and uniform on the unit interval, and let M_n be their maximum.

1. Show that $M_n \xrightarrow{P} 1$ and that $n(1 - M_n) = O_P(1)$ but not $o_P(1)$.
2. Show that $\sqrt{n}(1 - M_n) \xrightarrow{P} 0$, and explain why this does not contradict part (a).

3. Let $\hat{\theta}_n = M_n$ estimate $\theta = 1$. Show that $M_n - \theta = O_P(n^{-1})$, and that $a_n(M_n - \theta)$ has a nondegenerate limit in distribution only when a_n is proportional to n . A sample mean instead has rate $n^{-1/2}$. Explain why no result in this chapter could have told you either rate.

Problem 4.3 Let Z be standard normal, $V_n = Z$ and $U_n = (-1)^n Z$ as in Example 4.3.

1. Verify that $U_n \xrightarrow{d} Z$ and that $V_n U_n$ does not converge in distribution.
2. Now let $U'_n = Z + n^{-1}$. Show that (V_n, U'_n) does converge in distribution, identify the limit, and say which theorem of this section delivers it.
3. Explain in two or three sentences why Theorem 4.4 is unavailable in part (a) but not needed in part (b).

Problem 4.4 Prove parts (i), (ii), (iv) and (v) of Lemma 4.3. For (v), state carefully what happens on the event $\{r_n = -1\}$ and why it does not affect the conclusion. Then give an example of sequences $r_n = o_P(1)$ and $A_n = O_P(1)$ for which $r_n A_n$ is not $O_P(n^{-1})$, so that (iii) cannot be sharpened without further assumptions.

Problem 4.5 Let $V_n \xrightarrow{d} V$.

1. Give an example with $P\{V_n \in B\} \not\rightarrow P\{V \in B\}$ for some Borel B , and identify ∂B .
2. Show that the inequality in part (v) of Theorem 4.1 can be strict, using an open interval and a limit with an atom at an endpoint.
3. Deduce from (v) and (vi) that $P\{V_n \in B\} \rightarrow P\{V \in B\}$ whenever $P\{V \in \partial B\} = 0$, which is part (vii).

Bibliography

- P. Billingsley. *Probability and Measure*. Wiley-Interscience, 1995.
- J. Jiang. *Large Sample Techniques for Statistics*. Springer Texts in Statistics. Springer, New York, 2010.
- A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.



5

Characteristic Functions and the Central Limit Theorem

Chapter 4 built the vocabulary for describing a limiting distribution and the rules for manipulating one. It did not produce a single limit. This chapter produces the one that matters: a rescaled sample average, of almost any distribution whatever, is approximately normal. The route runs through a transform that turns sums of independent variables into products.

Why we cover this

Most reported standard errors, t statistics and confidence intervals rest on a normal approximation, and the central limit theorem is what licenses it. Two things about the theorem are worth having in advance. Beyond independence and a common distribution it asks almost nothing of the data — a finite variance and nothing more — which is why the same approximation is used across settings that have little else in common. And it describes a *root* — the average centred at the unknown mean and divided by the unknown scale — so getting from there to something a sample determines takes an estimated variance and the machinery of Chapter 4.

5.1 Characteristic Functions

Definition 5.1 *The characteristic function of a random vector V in $\mathbf{R}^k$ is*

$$\varphi_V(t) = E[e^{it'V}] = E[\cos(t'V)] + i E[\sin(t'V)] , \quad t \in \mathbf{R}^k .$$

Unlike a moment generating function, φ_V exists for every distribution and every t : the integrand has modulus one, so both expectations above are finite whatever the tails of V . That is the whole reason for working with $e^{it'V}$ rather than $e^{t'V}$.

Lemma 5.1 *Let V and U be random vectors in $\mathbf{R}^k$.*

(i) $\varphi_V(0) = 1$ and $|\varphi_V(t)| \leq 1$ for every t .

(ii) φ_V is uniformly continuous on $\mathbf{R}^k$.

(iii) For $a \in \mathbf{R}^m$ and B an $m \times k$ matrix, $\varphi_{a+BV}(t) = e^{it'a} \varphi_V(B't)$.

(iv) If V and U are independent, then $\varphi_{V+U} = \varphi_V \varphi_U$.

PROOF. Part (i) is immediate: $e^{i0} = 1$, and $|\mathbb{E}[e^{it'V}]| \leq \mathbb{E}|e^{it'V}| = 1$ by Jensen's inequality and $|e^{ix}| = 1$ for all $x \in \mathbf{R}$.

For (ii), fix t and h in $\mathbf{R}^k$ and factor the difference:

$$|\varphi_V(t+h) - \varphi_V(t)| = |\mathbb{E}[e^{it'V}(e^{ih'V} - 1)]| \leq \mathbb{E}|e^{ih'V} - 1| ,$$

using $|e^{it'V}| = 1$. The bound does not involve t , which is what makes the continuity uniform. As $h \rightarrow 0$ the integrand tends to zero pointwise and is bounded by 2, so dominated convergence, Fact A.2, sends the right-hand side to zero.

Part (iii) substitutes: $t'(a + BV) = t'a + (B't)'V$, and the factor $e^{it'a}$ is a constant that comes out of the expectation.

Part (iv) is where independence enters, and it is the property the rest of the chapter is built on. Independence of V and U makes $e^{it'V}$ and $e^{it'U}$ independent, so the expectation of the product is the product of the expectations:

$$\varphi_{V+U}(t) = \mathbb{E}[e^{it'V} e^{it'U}] = \mathbb{E}[e^{it'V}] \mathbb{E}[e^{it'U}] = \varphi_V(t) \varphi_U(t) ,$$

which is the claim. ■

Cf. *van der Vaart [1998, §2.3, p. 13]*; *Billingsley [1995, §26, p. 342]*.

Part (iv) is the point of the whole construction. A sum of independent variables has a distribution given by a convolution, which is awkward, and a characteristic function given by a product, which is not.

Example 5.1 (The coin) For W Bernoulli(q), the expectation is a two-term sum:

$$\varphi_W(t) = (1-q)e^{it \cdot 0} + q e^{it \cdot 1} = 1 - q + q e^{it} .$$

By Lemma 5.1(iv) the characteristic function of $n\bar{W}_n$, a sum of n independent copies, is $(1 - q + q e^{it})^n$, which is the binomial characteristic function. ■

Example 5.2 (The standard normal) Let Z be standard normal with density ϕ . Differentiating under the integral, which dominated convergence permits because $\mathbb{E}|Z| < \infty$,

$$\varphi'_Z(t) = \mathbb{E}[iZ e^{itZ}] = i \int z e^{itz} \phi(z) dz .$$

The normal density satisfies $z\phi(z) = -\phi'(z)$, which turns the integral into one that can be integrated by parts:

$$\begin{aligned} \int z e^{itz} \phi(z) dz &= - \int e^{itz} \phi'(z) dz \\ &= - \left[e^{itz} \phi(z) \right]_{-\infty}^{\infty} + \int it e^{itz} \phi(z) dz = it \varphi_Z(t), \end{aligned}$$

the boundary term vanishing because $\phi(z) \rightarrow 0$ at both ends while $|e^{itz}| = 1$. Multiplying by the leading i gives

$$\varphi_Z'(t) = i \cdot it \varphi_Z(t) = -t \varphi_Z(t), \quad \varphi_Z(0) = 1.$$

The unique solution is

$$\varphi_Z(t) = e^{-t^2/2}. \quad (5.1)$$

In other words, the characteristic function of a standard normal has the same Gaussian form as its density, and (5.1) is the target every calculation in Section 5.2 is aiming at. ■

Two results make characteristic functions usable. Both are stated and neither is proved; the arguments are analysis rather than probability, and nothing later in the course reuses their technique.

Theorem 5.1 (Uniqueness) *Two random vectors have the same distribution if and only if they have the same characteristic function.*

Cf. Billingsley [1995, Theorem 26.2, p. 346], where the inversion formula recovering the distribution function from φ_V is given first and uniqueness follows from it.

Theorem 5.2 (Lévy continuity theorem) *Let V_n be random vectors in $\mathbf{R}^k$ with characteristic functions φ_{V_n} .*

- (i) *If $V_n \xrightarrow{d} V$ then $\varphi_{V_n}(t) \rightarrow \varphi_V(t)$ for every $t \in \mathbf{R}^k$.*
- (ii) *If $\varphi_{V_n}(t) \rightarrow \psi(t)$ for every t and ψ is continuous at 0, then ψ is the characteristic function of some random vector V , and $V_n \xrightarrow{d} V$.*

Cf. van der Vaart [1998, Theorem 2.13, p. 14]; Billingsley [1995, Theorem 26.3, p. 349]; Jiang [2010, Theorem 2.11, p. 28], who calls it the Lévy–Cramér theorem.

Direction (i) needs nothing new — it is Theorem 4.1(ii) applied to the bounded continuous functions $v \mapsto \cos(t'v)$ and $v \mapsto \sin(t'v)$. Direction (ii) is the substantial half, and it is the one Section 5.2 uses: it converts a statement about a limit of transforms, which can be computed, into a statement about a limit of distributions, which cannot. The continuity requirement at 0 is not decoration. Without it the limit ψ can fail to be a characteristic function at all, because mass is free to escape to infinity. Take V_n normal with variance n , so that $\varphi_{V_n}(t) = e^{-nt^2/2} \rightarrow \mathbf{1}\{t = 0\}$. The limit is continuous at every $t \neq 0$ and discontinuous at 0, and it is the characteristic function of no distribution.

Corollary 5.1 (Cramér–Wold) *Let V_n and V be random vectors in $\mathbf{R}^k$. Then $V_n \xrightarrow{d} V$ if and only if*

$$t'V_n \xrightarrow{d} t'V \quad \text{for every } t \in \mathbf{R}^k .$$

PROOF. If $V_n \xrightarrow{d} V$, then $t'V_n \xrightarrow{d} t'V$ by Theorem 4.2, since $v \mapsto t'v$ is continuous everywhere.

Conversely, suppose $t'V_n \xrightarrow{d} t'V$ for every t . Fix t and apply Theorem 5.2(i) to the *scalar* sequence $t'V_n$, evaluating the resulting characteristic functions at the single point 1:

$$\varphi_{V_n}(t) = \mathbb{E}[e^{it'V_n}] = \varphi_{t'V_n}(1) \longrightarrow \varphi_{t'V}(1) = \varphi_V(t) .$$

This holds for every $t \in \mathbf{R}^k$, and the limit φ_V is continuous at 0 by Lemma 5.1(ii), so Theorem 5.2(ii) gives $V_n \xrightarrow{d} V$. ■

Cf. van der Vaart [1998, Lemma 2.15, p. 15]; Jiang [2010, Theorem 2.14, p. 31]; Billingsley [1995, Theorem 29.4, p. 383].

Corollary 5.1 reduces every multivariate limit in this course to a family of scalar ones. It is why Section 5.2 can prove a one-dimensional theorem and collect the k -dimensional version for free.

5.2 The Central Limit Theorem

One technical ingredient is needed before the theorem, and it is the only part of the argument that is not a short calculation.

Lemma 5.2 (Second-order expansion) *Let X be real-valued with $E[X] = 0$ and $E[X^2] = 1$. Then*

$$\varphi_X(s) = 1 - \frac{s^2}{2} + r(s) , \quad \text{where} \quad \frac{r(s)}{s^2} \longrightarrow 0 \quad \text{as } s \rightarrow 0 . \quad (5.2)$$

The proof is in Section A.8. What it says is that a characteristic function is flat to first order and curves downward at rate $E[X^2]$, and that a finite second moment is exactly enough to control the remainder — no third moment is required, which is why the central limit theorem below assumes only a variance.

Theorem 5.3 (Lindeberg–Lévy) *Let $W_1, W_2, \dots$ be i.i.d. with $E[W] = \mu$ and $\text{Var}[W] = \sigma^2 \in (0, \infty)$. Then*

$$R_n = \frac{\sqrt{n}(\bar{W}_n - \mu)}{\sigma} \xrightarrow{d} Z , \quad (5.3)$$

with Z standard normal.

The proof has four steps. Step 1 removes the centring and scaling, Step 2 converts a sum into a product, Step 3 expands a single factor, and Step 4 takes the limit and recognizes it. Only Step 3 requires anything beyond a line of algebra.

PROOF. *Step 1: reduce to standardized summands.* Set $X_i = (W_i - \mu)/\sigma$. These are i.i.d. with $E[X_i] = 0$ and $E[X_i^2] = 1$, and

$$R_n = \frac{1}{\sqrt{n}} \sum_{i \leq n} X_i. \quad (5.4)$$

So it is enough to prove that the right-hand side of (5.4) converges to Z , and from here μ and σ play no role.

Step 2: turn the sum into a product. Fix $t \in \mathbf{R}$. By Lemma 5.1(iii) with $B = n^{-1/2}$, and then by independence and Lemma 5.1(iv) applied $n-1$ times,

$$\varphi_{R_n}(t) = E\left[\exp\left\{\frac{it}{\sqrt{n}} \sum_{i \leq n} X_i\right\}\right] = \prod_{i \leq n} E[e^{itX_i/\sqrt{n}}] = \left[\varphi_X\left(\frac{t}{\sqrt{n}}\right)\right]^n, \quad (5.5)$$

the last equality because the X_i are identically distributed. This is the step the transform exists for: the left-hand side concerns a sum of n dependent-looking pieces, the right-hand side is one function raised to a power.

Step 3: expand the single factor. At $t = 0$ both sides of (5.5) equal one for every n , so assume $t \neq 0$ from here. The argument $t/\sqrt{n}$ then tends to zero as n grows, so Lemma 5.2 applies to it. Writing r for the remainder there,

$$\varphi_X\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + r\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{z_n}{n}, \quad z_n = -\frac{t^2}{2} + nr\left(\frac{t}{\sqrt{n}}\right). \quad (5.6)$$

The remainder is negligible at exactly the right rate. Since $n(t/\sqrt{n})^2 = t^2$, multiplying and dividing by $(t/\sqrt{n})^2$ gives

$$nr\left(\frac{t}{\sqrt{n}}\right) = t^2 \cdot \frac{r(t/\sqrt{n})}{(t/\sqrt{n})^2} \longrightarrow 0,$$

by (5.2) with $s = t/\sqrt{n}$, so $z_n \rightarrow -t^2/2$. The identity $n(t/\sqrt{n})^2 = t^2$ is what the scaling in (5.4) is for: dividing by $\sqrt{n}$ makes the n copies of the second-order term combine into a quantity that neither vanishes nor diverges, and it is the only power of n that does so.

Step 4: take the limit and identify it. Combining (5.5) and (5.6) gives $\varphi_{R_n}(t) = (1 + z_n/n)^n$ with $z_n \rightarrow -t^2/2$, so Fact A.10 yields

$$\varphi_{R_n}(t) \longrightarrow e^{-t^2/2} \quad \text{for every } t \in \mathbf{R}.$$

At $t = 0$ the convergence holds trivially, as noted at the start of Step 3, so it holds for every t . By (5.1) the limit is the characteristic function of Z , and it

is continuous at 0. Theorem 5.2(ii) therefore gives $R_n \xrightarrow{d} Z$. ■

Cf. *van der Vaart* [1998, Proposition 2.17, p. 16]; *Billingsley* [1995, Theorem 27.1, p. 357]; *Vershynin* [2026, Theorem 1.7.3, p. 18]. All three run the same four steps.

Remark 5.1 (What is not claimed) Beyond the i.i.d. assumption the hypothesis is a finite, nonzero variance and nothing else, so the conclusion covers distributions with no density, no moments beyond the second, and no resemblance to a normal at any sample size — the coin among them. What it does not supply is a rate. Theorem 5.3 says the distribution functions converge; it says nothing about how large n must be for the approximation to be “good”, and that question is taken up in Chapter 8. Nor does it say anything about $\bar{W}_n$ at a fixed n , which is the object practitioners actually hold. ■

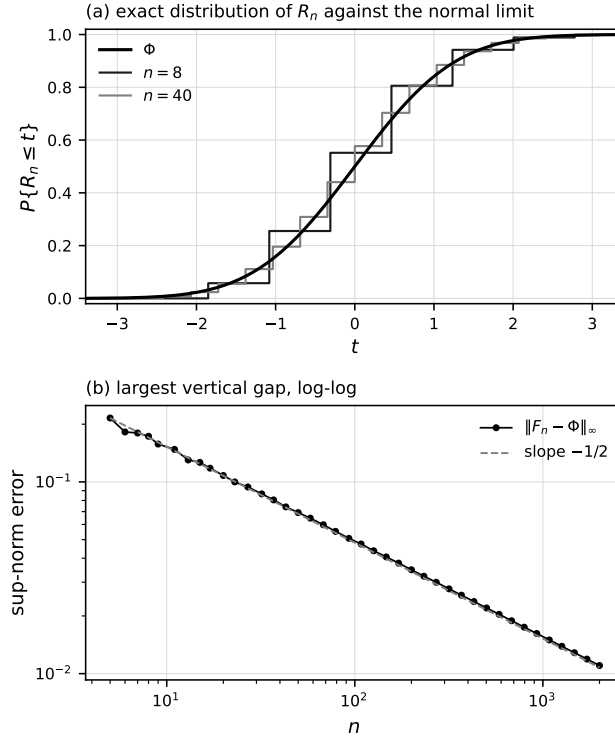


FIGURE 5.1: The central limit theorem on the coin, $q = 0.3$. Panel (a) shows the exact distribution function of R_n at $n = 8$ and $n = 40$ against Φ . Panel (b) plots $\|F_n - \Phi\|_\infty$ against n on log scales, with a reference line of slope $-1/2$.

Example 5.3 (The coin) Take $W_1, \dots, W_n$ i.i.d. $\text{Bernoulli}(q)$ with $q \in (0, 1)$, as in Example 2.1. Then $\mu = q$ and $\sigma^2 = q(1 - q) \in (0, \infty)$, so Theorem

5.3 applies and

$$R_n = \frac{\sqrt{n}(\bar{W}_n - q)}{\sqrt{q(1-q)}} \xrightarrow{d} Z ,$$

which is the limit assumed in Examples 4.2 and 4.5 and now established. Figure 5.1(a) draws the exact distribution function of R_n , available in closed form because $n\bar{W}_n$ is Binomial(n, q), against Φ at two sample sizes. The convergence is a statement about a step function approaching a continuous one, and Remark 4.1 says the largest vertical gap must vanish, since Φ is continuous. Panel (b) plots that gap, and the plot is consistent with a rate of $n^{-1/2}$ — a claim Theorem 5.3 does not make and Chapter 8 justifies. ■

The multivariate version needs no new argument.

Theorem 5.4 (Multivariate central limit theorem) *Let $W_1, W_2, \dots$ be i.i.d. random vectors in $\mathbf{R}^k$ with $E[W] = \mu$ and finite covariance matrix $\Sigma = \text{Var}[W]$. Then*

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

PROOF. By Corollary 5.1 it is enough to show that $t'\sqrt{n}(\bar{W}_n - \mu)$ converges in distribution to $t'N(0, \Sigma)$, which is $N(0, t'\Sigma t)$, for each fixed $t \in \mathbf{R}^k$. Fix t and set $Y_i = t'(W_i - \mu)$, a scalar. The Y_i are i.i.d. with $E[Y_i] = 0$ and $\text{Var}[Y_i] = t'\Sigma t$, finite because Σ is, and

$$t'\sqrt{n}(\bar{W}_n - \mu) = \frac{1}{\sqrt{n}} \sum_{i \leq n} Y_i .$$

If $t'\Sigma t > 0$, Theorem 5.3 applies to the Y_i and gives the claim. If $t'\Sigma t = 0$ then $Y_i = 0$ almost surely, the left-hand side is identically zero, and the limit $N(0, 0)$ is the point mass at zero, so the claim holds there too. ■

Cf. van der Vaart [1998, Example 2.18, p. 16]; Billingsley [1995, Theorem 29.5, p. 385].

5.3 Studentization

Theorem 5.3 is not yet usable. Its conclusion involves σ , which is unknown, and dividing by an estimate changes the statement. The result below is the one that actually gets applied, and it is worth stating as a theorem rather than reassembling each time.

Theorem 5.5 (Studentization) *Suppose $\sqrt{n}(\hat{\theta}_n - \theta(P)) \xrightarrow{d} N(0, \sigma^2(P))$ with $\sigma^2(P) > 0$, and let $\hat{\sigma}_n^2$ be a nonnegative estimator with $\hat{\sigma}_n^2 \xrightarrow{P} \sigma^2(P)$.*

Write $\hat{\sigma}_n = (\hat{\sigma}_n^2)^{1/2}$ and set the ratio below equal to zero on the event $\{\hat{\sigma}_n = 0\}$. Then

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta(P))}{\hat{\sigma}_n} \xrightarrow{d} Z . \quad (5.7)$$

PROOF. Write the ratio as a product,

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta(P))}{\hat{\sigma}_n} = \frac{\sqrt{n}(\hat{\theta}_n - \theta(P))}{\sigma(P)} \cdot h(\hat{\sigma}_n^2) ,$$

where $h(v) = \sigma(P)/\sqrt{v}$ for $v > 0$ and $h(v) = 0$ for $v \leq 0$. This separates a factor with a known limit from one converging to a constant. The convention on $\{\hat{\sigma}_n = 0\}$ is what makes the left-hand side defined everywhere, and h is the Borel function on $\mathbf{R}$ that Theorem 3.1 requires.

The first factor converges in distribution to Z , by Theorem 4.2 applied to division by the constant $\sigma(P) > 0$. For the second, h is continuous at every $v > 0$, and the limit $\sigma^2(P)$ is such a point, so Theorem 3.1 gives $h(\hat{\sigma}_n^2) \xrightarrow{P} h(\sigma^2(P)) = 1$. Theorem 4.4, with the second factor supplying the constant limit, delivers (5.7). ■

Three hypotheses are doing separable work, and it is worth knowing which fails in a given application. The limit distribution comes from Theorem 5.3 or one of its successors. Consistency of $\hat{\sigma}_n^2$ is a separate problem, usually solved by the analog principle of Section 3.2. And $\sigma^2(P) > 0$ is what keeps the ratio from being a division by something arbitrarily small; when it fails, (5.7) fails with it, and that failure is not a technicality.

Example 5.4 (The two groups) Let $W = (Y, X)$ with $X \in \{0, 1\}$ and $\pi = P\{X = 1\} \in (0, 1)$, let $\hat{\mu}_{n,x}$ be the group averages of Example 3.2, and suppose $\sigma_x^2 = \text{Var}[Y|X = x] \in (0, \infty)$ for $x = 0, 1$. Write $\hat{\pi}_n = n^{-1} \sum_{i \leq n} X_i$.

The group averages are ratios, not averages, so the theorems above do not apply to them directly. Rearranging the definition of $\hat{\mu}_{n,1}$ turns it into a ratio of two objects that *are* averages:

$$\sqrt{n}(\hat{\mu}_{n,1} - \mu_1) = \frac{1}{\hat{\pi}_n} \cdot \frac{1}{\sqrt{n}} \sum_{i \leq n} X_i(Y_i - \mu_1) , \quad (5.8)$$

and symmetrically for group zero, with $1 - X_i$, μ_0 and $1 - \hat{\pi}_n$ in place of X_i , μ_1 and $\hat{\pi}_n$. Both are set to zero when the corresponding group is empty, as in Example 3.2, an event whose probability tends to zero.

Now apply the tools in order. The numerators in (5.8) and its counterpart are the two coordinates of a sample average of

$$G_i = (X_i(Y_i - \mu_1), (1 - X_i)(Y_i - \mu_0)) ,$$

which has mean zero, since $E[X(Y - \mu_1)] = E[X(E[Y|X] - \mu_1)] = 0$ by the law

of iterated expectations and the same for the other coordinate. Its covariance matrix is diagonal, because $X(1-X) = 0$ forces the cross term to vanish, with diagonal entries $E[X(Y - \mu_1)^2] = \pi\sigma_1^2$ and $(1-\pi)\sigma_0^2$. Theorem 5.4 therefore gives

$$\frac{1}{\sqrt{n}} \sum_{i \leq n} G_i \xrightarrow{d} N(0, \text{diag}\{\pi\sigma_1^2, (1-\pi)\sigma_0^2\}) .$$

The weak law gives $\hat{\pi}_n \xrightarrow{P} \pi$, so Theorem 4.4 makes the pair converge jointly, and Theorem 4.2 applied to $g(a, b, p) = a/p - b/(1-p)$, continuous at every point with $0 < p < 1$, delivers

$$\sqrt{n} \{(\hat{\mu}_{n,1} - \hat{\mu}_{n,0}) - (\mu_1 - \mu_0)\} \xrightarrow{d} N\left(0, \frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1-\pi}\right) ,$$

the variance being $\pi\sigma_1^2/\pi^2 + (1-\pi)\sigma_0^2/(1-\pi)^2$ because the two limits are uncorrelated and jointly normal, hence independent.

Estimating that variance by replacing σ_x^2 with the within-group sample variance and π with $\hat{\pi}_n$, each consistent by Section 3.2, gives $\hat{\tau}_n^2 = \hat{\sigma}_1^2/\hat{\pi}_n + \hat{\sigma}_0^2/(1-\hat{\pi}_n)$, and Theorem 5.5 turns the display into a studentized root with a standard normal limit. It is still a root: $\mu_1 - \mu_0$ remains in the centring. In practice, this becomes a *statistic* — a function of the data only — under a hypothesis about $\mu_1 - \mu_0$. For instance, if the hypothesis is that the two means are the same, $H_0 : \mu_1 - \mu_0 = 0$, then the statistic is

$$T_n = \frac{\sqrt{n}(\hat{\mu}_{n,1} - \hat{\mu}_{n,0})}{\hat{\tau}_n} , \quad (5.9)$$

which converges to Z when our hypothesis is true — and explains why the qualifier “under the null hypothesis” is often present. That is the two-sample t statistic, and the calculation above is the whole of its justification. Note what the sequence of steps was: rewrite the estimator as a smooth function of sample averages, apply the central limit theorem to the averages, the law of large numbers to the remaining pieces, and then Slutsky and the continuous mapping theorem to assemble them. Every asymptotic normality argument in the rest of the course has this shape. ■

Remark 5.2 (Another route to the theorem) The proof of Theorem 5.3 is elementary but indirect: it works entirely with transforms and never compares R_n to a normal variable. Lindeberg’s argument does the comparison directly. Replace the summands $X_1, \dots, X_n$ by independent standard normal variables one at a time and bound the change in $E[f(R_n)]$ at each swap, for smooth bounded f . Because the variable removed and the one put in its place agree in mean and variance, the first two terms of a Taylor expansion of f cancel at every swap, and what survives is a third-order term. That term is not automatically small under a finite second moment, and controlling it requires truncation — the same two-sided device as Fact A.11, and in the triangular

arrays of Chapter 6 it is Lindeberg's condition that supplies the control. We do not give the argument.

Both routes extend beyond this chapter: Lindeberg–Feller can be proved by characteristic functions or by replacement, and Billingsley [1995] takes the first road. Replacement is the one that carries into the high-dimensional central limit theorems of the winter quarter, where no useful transform is available. ■

Where we will use this

- **In this class.** Theorem 5.3 supplies the limit that Chapters 7 and 9 attach to every estimator through the expansion (4.5), and Theorem 5.5 is the form in which it is used. Chapter 6 replaces the i.i.d. hypothesis by a triangular array, and Chapter 8 asks how good the approximation is and whether it holds uniformly. In the second half of the quarter, a central limit theorem for the score is a stated condition in every asymptotic normality result.
- **Later in the sequence.** The winter quarter needs central limit theorems in growing dimension, where the replacement argument of Remark 5.2 rather than the transform argument is what scales. In the spring, the bootstrap is justified by a central limit theorem applied conditionally on the sample.
- **In recent research.** Inference with many covariates or many instruments turns on central limit theorems in which the dimension grows with n , and the Berry–Esseen bounds of Chapter 8 are how their accuracy is measured.

5.4 Key Concepts

- **Characteristic functions.** $\varphi_V(t) = E[e^{it'V}]$ exists always, determines the distribution, and turns sums of independent variables into products. That last property is why the central limit theorem is provable.
- **Lévy's continuity theorem.** Pointwise convergence of characteristic functions to a limit continuous at 0 is equivalent to convergence in distribution. It converts a computation into a distributional statement.
- **Cramér–Wold.** A vector converges in distribution exactly when every scalar projection does, which reduces multivariate limits to univariate ones.
- **The central limit theorem.** Beyond the i.i.d. assumption, a finite

nonzero variance is the only hypothesis. Standardized sample averages converge to the standard normal.

- **Studentization.** The usable result is the central limit theorem, plus a consistent variance estimator, plus Slutsky.

5.5 Concluding Remarks

van der Vaart [1998, §2.3] covers this material in the order used here, and Billingsley [1995, §§26–27] gives the analysis in full, including the inversion formula behind Theorem 5.1 and the tightness argument behind Theorem 5.2. Vershynin [2026, Theorem 1.7.3] states the central limit theorem alongside the non-asymptotic bounds that measure how far from it a given n leaves you.

5.6 Problems

Problem 5.1 Let V be a random variable with characteristic function φ_V .

1. Show that $\varphi_{-V}(t) = \overline{\varphi_V(t)}$, and deduce that V and $-V$ have the same distribution if and only if φ_V is real-valued.
2. Compute the characteristic function of a $\text{Uniform}[-a, a]$ variable and verify directly that it is real-valued.
3. Suppose $E[V^2] < \infty$ and, without loss of generality, that $E[V] = 0$ and $E[V^2] = 1$. Show that φ_V is twice differentiable with $\varphi'_V(0) = 0$ and $\varphi''_V(0) = -1$, and deduce Lemma 5.2 from Taylor's theorem. Which step of this route needs justifying most carefully, and what does the proof in Section A.8 do instead?

Problem 5.2 Continue Example 5.1, with $W_1, \dots, W_n$ i.i.d. $\text{Bernoulli}(q)$.

1. Write down the characteristic function of R_n in (5.3) explicitly, using Lemma 5.1(iii) and (iv).
2. Verify the four steps of the proof of Theorem 5.3 directly on this expression, computing the expansion of $\varphi_X(s)$ by hand rather than citing Lemma 5.2.
3. Explain what goes wrong when $q = 0$, and identify which hypothesis of Theorem 5.3 fails.

Problem 5.3 Each part gives a sequence of characteristic functions. Decide whether the corresponding sequence of distributions converges, and if so to what.

1. $\varphi_n(t) = e^{-nt^2/2}$.
2. $\varphi_n(t) = e^{-t^2/(2n)}$.
3. $\varphi_n(t) = \cos(t/n)$.
4. $\varphi_n(t) = e^{itn}$.

In each case say which hypothesis of Theorem 5.2(ii) holds and which fails, and where the mass has gone when convergence fails.

Problem 5.4 Let Z be standard normal and let S be independent of Z with $P\{S = 1\} = P\{S = -1\} = 1/2$. Put $V = (Z, SZ)$.

1. Show that both coordinates of V are standard normal, and that V is not bivariate normal.
2. Find a $t \in \mathbf{R}^2$ for which $t'V$ is not normally distributed.
3. Explain how this is consistent with Corollary 5.1, and what it shows about deducing joint normality from marginal normality.

Problem 5.5 Let $W_1, \dots, W_n$ be i.i.d. with $E[W^2] < \infty$ and $\text{Var}[W] = \sigma^2 > 0$, and let $\hat{\sigma}_n^2$ be the plug-in estimator (3.6).

1. Verify the hypotheses of Theorem 5.5 for $\hat{\theta}_n = \bar{W}_n$, saying where the second moment is used in each and why no higher moment is needed.
2. Show that the conclusion is unchanged if $\hat{\sigma}_n^2$ is replaced by $\{n/(n-1)\}\hat{\sigma}_n^2$, and explain in one sentence why no argument about the divisor is needed.
3. Suppose instead that $\hat{\sigma}_n^2 \xrightarrow{P} \tau^2$ with $0 < \tau^2 \neq \sigma^2$, as happens under misspecified standard errors. Show that the studentized root still converges in distribution, identify the limit, and say what a test that ignores this would get wrong.

Bibliography

- P. Billingsley. *Probability and Measure*. Wiley-Interscience, 1995.
- J. Jiang. *Large Sample Techniques for Statistics*. Springer Texts in Statistics. Springer, New York, 2010.

- A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.
- R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, second edition, 2026.



6

Heterogeneous Data and Triangular Arrays

Every limit theorem so far has assumed that the observations are independent and identically distributed. In many applications the observations are independent but not identically distributed: variances differ across observations, groups have different sizes, and estimators built with a bandwidth or evaluated at a moving parameter value change the summands themselves as the sample grows. This chapter drops the identical distribution and keeps the independence.

Why we cover this

The summands of an estimator need not be identically distributed. With a fixed regression design and heteroskedastic errors their variances differ across observations; with groups of unequal size the group averages differ in precision; and with weights that change with n the summands themselves change with the sample size. All of these fall outside Chapter 5. The device that covers them is to let the whole collection of summands change with the sample size, and the condition that replaces the identical distribution asks that the share of the total variance contributed by large outcomes vanish.

6.1 Averages of Non-Identically Distributed Data

Before asking about limiting distributions it is worth asking whether the average still settles down. It does, under a condition on the variances alone, and the argument is the Chebyshev bound of Chapter 2 with the summands no longer sharing a distribution.

Theorem 6.1 *Let $W_1, W_2, \dots$ be independent with $E[W_i] = \mu_i$ and $\text{Var}[W_i] = \sigma_i^2 < \infty$, and write $\bar{\mu}_n = n^{-1} \sum_{i \leq n} \mu_i$. If*

$$\frac{1}{n^2} \sum_{i \leq n} \sigma_i^2 \longrightarrow 0, \quad (6.1)$$

then $\bar{W}_n - \bar{\mu}_n \xrightarrow{P} 0$.

PROOF. Independence makes the variance of a sum the sum of the variances, so

$$\text{Var}[\bar{W}_n] = \frac{1}{n^2} \sum_{i \leq n} \sigma_i^2,$$

and $E[\bar{W}_n] = \bar{\mu}_n$. Chebyshev's inequality, Lemma 2.1(ii), then gives

$$P\{|\bar{W}_n - \bar{\mu}_n| > \varepsilon\} \leq \frac{1}{\varepsilon^2} \cdot \frac{1}{n^2} \sum_{i \leq n} \sigma_i^2 \quad \text{for every } \varepsilon > 0,$$

whose right-hand side tends to zero by (6.1). ■

Cf. Jiang [2010, Theorem 6.2, p. 175]; the array version is his Theorems 6.3 and 6.4, pp. 176–177.

Two things about Theorem 6.1 are worth keeping. The centring is $\bar{\mu}_n$, an average of means that need not converge to anything, so the conclusion is that the average tracks a moving target rather than a fixed one. And (6.1) holds whenever the variances are bounded, since then $n^{-2} \sum_{i \leq n} \sigma_i^2 \leq \bar{\sigma}^2/n$; what it rules out is variances that grow fast enough for the total to be of order n^2 . That happens in two ways. A small number of observations may have variances large enough to dominate the sum, and this is the case usually pictured; but the growth may equally be spread across the row, as with $\sigma_i^2 = i$, where the total is $n(n+1)/2$ and no fixed number of observations carries a nonvanishing share. Problem 6.1 asks for both.

Example 6.1 (Groups of unequal size) Suppose the first m_n observations have variance σ_a^2 and the rest variance σ_b^2 , with $m_n/n \rightarrow \lambda \in [0, 1]$. Then $n^{-2} \sum_{i \leq n} \sigma_i^2 = n^{-1} \{(m_n/n)\sigma_a^2 + (1 - m_n/n)\sigma_b^2\} \rightarrow 0$, so Theorem 6.1 applies whatever λ is — including $\lambda = 0$, where one group is asymptotically negligible. Heterogeneity of this kind costs nothing at the level of the law of large numbers. Section 6.3 shows that it is not free at the level of the central limit theorem. ■

6.2 Triangular Arrays

A sequence of observations is not general enough for what follows. When summands are rescaled, reweighted or re-centred as the sample grows, the object being summed at sample size n is not the object that was being summed at $n-1$, and the notation has to allow the entire collection to change.

Definition 6.1 A triangular array is a collection of random variables $\{X_{n,i} : 1 \leq i \leq k_n\}$, indexed by $n \geq 1$, with $k_n \rightarrow \infty$, in which $X_{n,1}, \dots, X_{n,k_n}$ are independent for each fixed n . Its row sums are

$$S_n = \sum_{i \leq k_n} X_{n,i} .$$

Nothing is assumed across rows: the variables in row n need not be related to those in row $n - 1$, and there is no single probability space on which the whole array must live. That freedom is the point. Independence *within* a row, by contrast, is a genuine restriction and is what everything below uses.

The device is less exotic than it looks, because the central limit theorem of Chapter 5 is already a statement about an array. Its conclusion concerned the root

$$R_n = \frac{\sqrt{n}(\bar{W}_n - \mu)}{\sigma} ,$$

from (5.3). Writing the average out and moving the constants inside the sum,

$$R_n = \frac{1}{\sqrt{n}} \sum_{i \leq n} \frac{W_i - \mu}{\sigma} = \sum_{i \leq n} X_{n,i} , \quad X_{n,i} = \frac{W_i - \mu}{\sigma \sqrt{n}} , \quad (6.2)$$

so R_n is the row sum S_n of the array with $k_n = n$ and summands $X_{n,i}$. The summands change with n because of the $\sqrt{n}$, even though the underlying observations do not, so Chapter 5 proved an array result and simply had no need to say so.

Three situations force the generality to be needed: the observations are heterogeneous, as in Example 6.1; the estimator carries a bandwidth or a weight that shrinks with n , so that every summand depends on the sample size; or the distribution generating the data is itself indexed by n .

Remark 6.1 (Identical rows) Definition 6.1 does not ask the summands within a row to share a distribution, and Sections 6.3 and 6.4 need that freedom. The case in which they do share one — row n drawn i.i.d. from a distribution P_n that itself depends on the sample size — is nonetheless the special case worth naming, since it is what (6.2) is and what Chapter 8 runs on throughout. ■

6.3 The Lindeberg–Feller Theorem

Throughout this section $\{X_{n,i}\}$ is a triangular array with

$$\mathbb{E}[X_{n,i}] = 0 , \quad \sigma_{n,i}^2 = \text{Var}[X_{n,i}] < \infty , \quad \sum_{i \leq k_n} \sigma_{n,i}^2 = 1 , \quad (6.3)$$

the last being a normalization rather than an assumption: if the row variances sum to $s_n^2 \in (0, \infty)$, replace $X_{n,i}$ by $X_{n,i}/s_n$.

Definition 6.2 (Lindeberg's condition) *The array satisfies Lindeberg's condition if*

$$L_n(\varepsilon) = \sum_{i \leq k_n} E[X_{n,i}^2 \mathbf{1}\{|X_{n,i}| > \varepsilon\}] \longrightarrow 0 \quad \text{for every } \varepsilon > 0. \quad (6.4)$$

The quantity $L_n(\varepsilon)$ is the share of the total variance contributed by outcomes larger than ε . Since the total is one by (6.3), (6.4) says that in the limit all of the variance comes from summands that are individually small.

Summands do not arrive normalized, so it is worth recording the form in which the condition is actually checked. Given independent, mean-zero $Y_{n,1}, \dots, Y_{n,k_n}$ with $s_n^2 = \sum_{i \leq k_n} \text{Var}[Y_{n,i}] \in (0, \infty)$, put $X_{n,i} = Y_{n,i}/s_n$. Then (6.3) holds and (6.4) reads

$$\frac{1}{s_n^2} \sum_{i \leq k_n} E[Y_{n,i}^2 \mathbf{1}\{|Y_{n,i}| > \varepsilon s_n\}] \longrightarrow 0 \quad \text{for every } \varepsilon > 0. \quad (6.5)$$

The recipe is therefore: centre the summands, compute s_n^2 , divide by s_n , and verify (6.5). Note that the truncation level moves with s_n , which is what makes the condition scale-free.

Theorem 6.2 (Lindeberg–Feller) *Let $\{X_{n,i}\}$ be a triangular array satisfying (6.3) and Lindeberg's condition (6.4). Then*

$$S_n = \sum_{i \leq k_n} X_{n,i} \xrightarrow{d} Z,$$

with Z standard normal.

The proof is in Section A.9. It is the four-step argument of Theorem 5.3 with one change: where that proof expanded a single characteristic function and raised it to the power n , this one expands each of the k_n factors separately and compares the product to $e^{-t^2/2}$ term by term. Lindeberg's condition is what makes the comparison work, and it enters at exactly the place where the identical distribution was used before.

Two consequences of (6.4) are short enough to prove here, and between them they are what one actually uses.

Lemma 6.1 *Lindeberg's condition implies $\max_{i \leq k_n} \sigma_{n,i}^2 \rightarrow 0$.*

PROOF. Fix $\varepsilon > 0$ and split a single variance at the level ε , according to whether the variable exceeds it:

$$\begin{aligned} \sigma_{n,i}^2 &= E[X_{n,i}^2 \mathbf{1}\{|X_{n,i}| \leq \varepsilon\}] + E[X_{n,i}^2 \mathbf{1}\{|X_{n,i}| > \varepsilon\}] \\ &\leq \varepsilon^2 + E[X_{n,i}^2 \mathbf{1}\{|X_{n,i}| > \varepsilon\}]. \end{aligned}$$

The second term is one summand of $L_n(\varepsilon)$, and all the summands are non-negative, so it is at most $L_n(\varepsilon)$. The bound holds for every i , hence for the maximum:

$$\max_{i \leq k_n} \sigma_{n,i}^2 \leq \varepsilon^2 + L_n(\varepsilon) .$$

Letting $n \rightarrow \infty$ with ε fixed gives $\limsup_n \max_i \sigma_{n,i}^2 \leq \varepsilon^2$, and $\varepsilon > 0$ was arbitrary. ■

Remark 6.2 (What dominance misses) Lemma 6.1 is the formal version of the intuition that no single summand may matter on its own: with the total variance normalized to one, the largest individual share vanishes.

It is not the whole condition though. The vanishing maximum constrains variances only, while (6.4) constrains tails: it asks that the contribution of *large outcomes*, summed across the whole row, be negligible, and an array can have vanishing individual variances while a rare large outcome still carries a fixed share of the total. Lindeberg's condition is therefore strictly the stronger requirement, and it is the one the theorem needs. ■

Cf. Jiang [2010, Example 6.7, p. 184] exhibits an array for which Lindeberg's condition fails.

Lindeberg's condition may be awkward to work with directly, since it involves a truncated second moment at every ε . The following condition involves one moment and no truncation.

Lemma 6.2 (Lyapunov's condition) Suppose that for some $\delta > 0$,

$$\sum_{i \leq k_n} E \left[|X_{n,i}|^{2+\delta} \right] \rightarrow 0 . \quad (6.6)$$

Then Lindeberg's condition holds.

PROOF. Fix $\varepsilon > 0$. On the event $\{|X_{n,i}| > \varepsilon\}$ we have $1 < |X_{n,i}|^\delta / \varepsilon^\delta$, so multiplying the integrand of $L_n(\varepsilon)$ by that ratio only increases it:

$$E \left[X_{n,i}^2 \mathbf{1}_{\{|X_{n,i}| > \varepsilon\}} \right] \leq \frac{1}{\varepsilon^\delta} E \left[|X_{n,i}|^{2+\delta} \mathbf{1}_{\{|X_{n,i}| > \varepsilon\}} \right] \leq \frac{1}{\varepsilon^\delta} E \left[|X_{n,i}|^{2+\delta} \right] ,$$

the last step dropping the indicator from a nonnegative integrand. Summing over i gives $L_n(\varepsilon) \leq \varepsilon^{-\delta} \sum_{i \leq k_n} E |X_{n,i}|^{2+\delta}$, which tends to zero by (6.6) since ε is fixed. ■

Cf. Billingsley [1995, Theorem 27.3, p. 362]; Jiang [2010, Theorem 6.11, p. 182] states Lindeberg–Feller with Lyapunov's condition as the corollary.

Lindeberg's condition is a uniform integrability requirement on the squared summands, and uniform integrability is harder to state than a moment bound.

That is why econometric assumptions so often sit just above the second moment: Lyapunov's condition replaces the truncated second moments with a sum of $2 + \delta$ moments, and uniform bounds on those moments make it easy to establish, as Example 6.3 shows.

Example 6.2 (One summand that dominates) Let ξ take the values $\pm 1/\sqrt{2}$ with equal probabilities, so that $E[\xi] = 0$ and $\text{Var}[\xi] = 1/2$. Let $Y_2, Y_3, \dots$ be i.i.d. with mean zero and variance one, independent of ξ , and set

$$k_n = n, \quad X_{n,1} = \xi, \quad X_{n,i} = \frac{Y_i}{\sqrt{2(n-1)}} \quad \text{for } 2 \leq i \leq n.$$

The row variances sum to $1/2 + (n-1) \cdot \{2(n-1)\}^{-1} = 1$, as (6.3) requires.

Write $A_n = \sum_{2 \leq i \leq n} X_{n,i}$, so that $S_n = \xi + A_n$ with the two pieces independent. Their characteristic functions therefore multiply, by Lemma 5.1(iv). The second factor is handled by Chapter 5: A_n is $1/\sqrt{2}$ times a standardized average of i.i.d. variables, so Theorem 5.3 gives $A_n \xrightarrow{d} N(0, 1/2)$ and Theorem 5.2(i) turns that into $\varphi_{A_n}(t) \rightarrow e^{-t^2/4}$. The first factor does not depend on n at all, and is $\varphi_\xi(t) = \cos(t/\sqrt{2})$. Hence

$$\varphi_{S_n}(t) \rightarrow \cos\left(\frac{t}{\sqrt{2}}\right) e^{-t^2/4}, \quad (6.7)$$

which is continuous at 0, so Theorem 5.2(ii) applies and S_n does converge in distribution — but not to Z . The limit (6.7) is not $e^{-t^2/2}$, and the discrepancy is visible: the cosine factor contributed by ξ is still sitting there.

Lindeberg's condition fails, and Lemma 6.1 says where. Here $\sigma_{n,1}^2 = 1/2$ for every n , so the largest individual variance does not vanish. The first observation never stops mattering, and the limit remembers its distribution. ■

Remark 6.3 (Clusters) Example 6.2 already explains an assumption that will look arbitrary when it appears in the spring. With clustered data the summands are sums over groups rather than over observations, so the row of the array has one entry per cluster and its length is the number of clusters. The standard requirement that no cluster account for a non-vanishing share of the total variation is the conclusion of Lemma 6.1 for that row; Lindeberg's condition asks for more, since it also controls the tails of the cluster sums. Example 6.2 shows what a dominant cluster does: it survives into the limit and prevents normality. ■

6.4 An Application: Heteroskedastic Sums

The sums that appear in econometrics are rarely averages of the data themselves. They are averages of terms like $X_i U_i$, in which a covariate multiplies a

disturbance, and the pair (X_i, U_i) need not have the same distribution across i . This is the case Theorem 6.2 was built for, and it is worth working through once in full.

Example 6.3 (A heteroskedastic score) Let $x_1, x_2, \dots$ be a fixed sequence of real numbers — a non-random design, which keeps the calculation clean — and let $U_1, U_2, \dots$ be independent with

$$\mathbb{E}[U_i] = 0, \quad \mathbb{E}[U_i^2] = \omega_i^2, \quad \mathbb{E}[|U_i|^{2+\delta}] \leq C < \infty$$

for some $\delta > 0$ and all i . The object of interest is

$$G_n = \frac{1}{\sqrt{n}} \sum_{i \leq n} x_i U_i, \quad (6.8)$$

whose variance is $v_n^2 = n^{-1} \sum_{i \leq n} x_i^2 \omega_i^2$, a number rather than a random variable because the design is fixed. Assume $v_n^2 \rightarrow v^2 \in (0, \infty)$ and $n^{-1} \sum_{i \leq n} |x_i|^{2+\delta} \leq D < \infty$.

Write G_n/v_n as a row sum by setting $\zeta_{n,i} = x_i U_i / (v_n \sqrt{n})$, which has mean zero and whose row variances sum to one by construction. Lyapunov's condition is now a computation, and the constants x_i pass through the expectation because they are not random:

$$\sum_{i \leq n} \mathbb{E}[|\zeta_{n,i}|^{2+\delta}] = \frac{1}{v_n^{2+\delta} n^{1+\delta/2}} \sum_{i \leq n} |x_i|^{2+\delta} \mathbb{E}[|U_i|^{2+\delta}] \leq \frac{CD}{v_n^{2+\delta} n^{\delta/2}},$$

where the inequality uses the two uniform moment bounds. Since $v_n \rightarrow v > 0$, the right-hand side is of order $n^{-\delta/2}$ and tends to zero. Lemma 6.2 therefore gives Lindeberg's condition, Theorem 6.2 gives $G_n/v_n \xrightarrow{d} Z$, and Theorem 4.4 with $v_n \rightarrow v$ converts this into

$$G_n \xrightarrow{d} N(0, v^2).$$

Nothing here required the U_i to share a distribution. What it required was that the normalized row's $2+\delta$ moments vanish, which is what the two uniform bounds deliver: the factor $n^{-\delta/2}$ is the whole of the argument, and it comes from $2+\delta$ moments rather than two.

With a random design the conclusion can survive, but the argument needs changing: the natural normalization is then the variance given the design, which is random, and dividing every summand by a common random quantity can destroy independence within the row. The usual route is to argue conditionally on the design, which is left to the second half of the quarter. ■

Remark 6.4 (Where else arrays are unavoidable) Two further settings force Definition 6.1 rather than merely permit it, and both return later in the sequence. A kernel or local polynomial estimator uses a bandwidth $h_n \rightarrow 0$,

so every summand is reweighted at every sample size and no fixed sequence of summands exists at all; that is why the asymptotics of regression discontinuity designs in the spring are array asymptotics and cannot be otherwise. And Chapter 8 studies procedures along sequences of distributions P_n that drift toward the point where the procedure fails, which makes the distribution of each summand depend on n by construction. ■

Remark 6.5 (Why the distribution is allowed to move) The last of these is the case of Remark 6.1, row n drawn i.i.d. from a distribution P_n , and it deserves a word now because the reason for it is not obvious.

Naive asymptotics has trouble expressing the idea that a parameter is *small*. Letting $n \rightarrow \infty$ increases the information the sample carries, while the parameter value determines how hard the problem is — for example, how close q sits to zero, how weak an instrument is, how near a coefficient lies to a boundary. If the parameter is held fixed while n grows, the information eventually swamps the difficulty: for the coin, the standardized limit is Z at $q = 0.5$ and at $q = 0.001$ alike, although at any given n the normal approximation is far less accurate at the second.

Letting the distribution move with n is how that is repaired: the problem is made harder at the same rate as the sample grows, so the limit records the difficulty instead of erasing it. Example 6.4 does this on the coin, and Chapter 8 makes it the organizing device. ■

Example 6.4 (The drifting coin) Let $W_{n,1}, \dots, W_{n,n}$ be independent Bernoulli variables with common success probability q_n , which moves with the sample size and may approach 0 or 1. Write

$$\lambda_n = n q_n(1 - q_n) , \quad (6.9)$$

the variance of the total number of successes $n\bar{W}_n$.

When $\lambda_n \rightarrow \infty$. Setting $X_{n,i} = (W_{n,i} - q_n)/\sqrt{\lambda_n}$ gives a row of independent, mean-zero variables with variance $1/n$ each, so (6.3) holds. Lindeberg's condition then needs no computation: $|W_{n,i} - q_n| \leq 1$, so

$$\max_{i \leq n} |X_{n,i}| \leq \frac{1}{\sqrt{\lambda_n}} \rightarrow 0 ,$$

and for any fixed $\varepsilon > 0$ the indicator in (6.4) is identically zero once $\lambda_n > \varepsilon^{-2}$. Theorem 6.2 gives

$$\frac{\sqrt{n}(\bar{W}_n - q_n)}{\sqrt{q_n(1 - q_n)}} \xrightarrow{d} Z , \quad (6.10)$$

which Theorem 5.3 cannot deliver, since the observations are drawn from a different distribution at each n .

Nothing in that argument asked q_n to stay away from the boundary. It asked only that $\lambda_n \rightarrow \infty$, and that is the honest condition: the success probability may drift to 0 as fast as one likes provided the expected number of

successes still grows. With $q_n = n^{-1/2}$, for instance, $\lambda_n \approx \sqrt{n}$ and (6.10) holds.

When it does not. Take $q_n = \lambda/n$ for a fixed $\lambda > 0$, so that $\lambda_n \rightarrow \lambda$ and the expected number of successes stays bounded. Example 5.1 and Lemma 5.1(iv) give the characteristic function of the total directly. Writing $z_n = nq_n(e^{it} - 1) = \lambda(e^{it} - 1)$,

$$\varphi_{n\bar{W}_n}(t) = (1 - q_n + q_n e^{it})^n = \left(1 + \frac{z_n}{n}\right)^n,$$

so Fact A.10 yields $\varphi_{n\bar{W}_n}(t) \rightarrow \exp\{\lambda(e^{it} - 1)\}$, the characteristic function of a $\text{Poisson}(\lambda)$ variable, so Theorem 5.2(ii) gives $n\bar{W}_n \xrightarrow{d} \text{Poisson}(\lambda)$. No centring or rescaling turns this limit into a normal one.

So one family of experiments produces two different limits, and which appears depends on how fast q_n reaches the boundary relative to n . The boundary itself is not what breaks the approximation; the quantity that matters is λ_n , the information the sample carries. This is the phenomenon Chapter 8 is built around: a statement can hold at every fixed $q \in (0, 1)$ and still fail along sequences approaching the boundary, because the sequence outruns the approximation. ■

Where we will use this

- **In this class.** Theorem 6.2 is what covers the estimators of Chapters 7 and 9 when the data are not identically distributed, and Example 6.4 is the calculation Chapter 8 runs along drifting sequences to expose the difference between pointwise and uniform validity. In the second half of the quarter, a central limit theorem for the score is a stated condition; under i.i.d. sampling Chapter 5 supplies it, and Lyapunov's condition is what supplies it once the data are heterogeneous.
- **Later in the sequence.** Kernel and local polynomial estimators in the winter, and regression discontinuity designs in the spring, produce summands that are reweighted at every sample size, so their asymptotics are array asymptotics by construction. Clustered standard errors are the same device applied to sums over groups rather than observations.
- **In recent research.** Uniformity is the standard by which procedures are now judged in models where the difficulty of the problem is itself a parameter — moment inequalities, entry games with multiple equilibria, and weak instruments — and the drifting sequences of Remark 6.5 are how the claims in that literature are stated and tested.

6.5 Key Concepts

- **The law of large numbers survives heterogeneity cheaply.** Independence and $n^{-2} \sum_{i \leq n} \sigma_i^2 \rightarrow 0$ are enough for the average to track the average of the means, and bounded variances suffice for the second condition.
- **Triangular arrays.** Letting the whole row change with n is what allows summands to be rescaled, reweighted, or drawn from a distribution that itself depends on n . The i.i.d. central limit theorem is already an array result.
- **Lindeberg's condition.** With the row variances summing to one, the share contributed by outcomes above any fixed threshold vanishes. It implies that the largest individual variance vanishes, but it asks for more than that, because it constrains tails and not only variances.
- **Lyapunov's condition.** A uniformly bounded $2 + \delta$ moment, summed and rescaled, is the checkable sufficient condition, and it is why moment assumptions in econometrics sit just above the second.
- **What can go wrong.** Without Lindeberg's condition the limit need not be normal. A summand that keeps a fixed share of the total variance can survive into the limit, and the variance can also sit in rare large outcomes with no dominant summand, as for the coin with $q_n = \lambda/n$.

6.6 Concluding Remarks

Billingsley [1995, §27] is the reference for Theorem 6.2 and proves it by the characteristic function argument used in Section A.9, with Lyapunov's condition at his Theorem 27.3. Jiang [2010, §6.4] gives the same material with more worked arrays, including a Bernoulli array at his Example 6.5 and, at Example 6.7, an array where Lindeberg's condition fails. van der Vaart [1998, Example 2.28, p. 21] runs the linear regression case in the form of Example 6.3.

6.7 Problems

Problem 6.1 Let $W_1, W_2, \dots$ be independent with $E[W_i] = \mu_i$ and $\text{Var}[W_i] = \sigma_i^2$.

1. Show that $\sigma_i^2 = \sqrt{i}$ satisfies (6.1) and that $\sigma_i^2 = i$ does not.
2. Take $\mu_i = 0$ and $\sigma_i^2 = i$. Find a normalization a_n with $a_n^{-1} \sum_{i \leq n} W_i \xrightarrow{P} 0$, and say precisely what the variances alone do and do not tell you about the natural choice $a_n = n$.
3. Theorem 6.1 centres at $\bar{\mu}_n$ rather than at a limit. Give a sequence μ_i for which $\bar{\mu}_n$ does not converge, and explain what the theorem still asserts.

Problem 6.2 Verify that the array (6.2) built from i.i.d. observations with mean μ and variance $\sigma^2 \in (0, \infty)$ satisfies (6.3), and show directly that it satisfies Lindeberg's condition. Where does your argument use that the observations are identically distributed, and what does it need in place of that if they are not?

Problem 6.3 Let $X_{n,i}$, $1 \leq i \leq n$, be independent with $P\{X_{n,i} = 1 - p_i\} = p_i$ and $P\{X_{n,i} = -p_i\} = 1 - p_i$, so that $E[X_{n,i}] = 0$ and $\text{Var}[X_{n,i}] = p_i(1 - p_i)$.

1. Normalize the row so that (6.3) holds, and write down $L_n(\varepsilon)$.
2. Show that if $\sum_{i \leq n} p_i(1 - p_i) \rightarrow \infty$ then Lindeberg's condition holds, and conclude that the normalized row sum is asymptotically normal.
3. Show that the conclusion can fail when $\sum_i p_i(1 - p_i)$ converges, and identify which hypothesis breaks.

Problem 6.4 1. Show that Lyapunov's condition (6.6) for some $\delta > 0$ implies it for every smaller positive δ' , assuming the relevant moments exist.

2. Give an array satisfying Lindeberg's condition for which (6.6) fails for every $\delta > 0$, so that Lemma 6.2 is strictly one-directional.
3. In Example 6.3, suppose the bound on $E|U_i|^{2+\delta}$ is dropped and only $E[U_i^2] = \omega_i^2 \leq C$ is assumed. Show that the argument given no longer works, and say what would have to be verified instead.

Problem 6.5 Return to Example 6.2.

1. Compute $L_n(\varepsilon)$ for small ε and confirm that it does not tend to zero.

2. Modify the example by replacing $X_{n,1}$ with $\xi/\sqrt{c_n}$ for a sequence $c_n \rightarrow \infty$, and rescale the rest of the row as $X_{n,i} = \gamma_n Y_i/\sqrt{n-1}$ for $2 \leq i \leq n$, choosing γ_n so that the row variances again sum to one. Find γ_n , show that Lindeberg's condition now holds, and identify the limit of S_n .
3. Explain in two or three sentences why the limit in the original example remembers the distribution of ξ , while every summand in part (b) is forgotten.

Bibliography

- P. Billingsley. *Probability and Measure*. Wiley-Interscience, 1995.
- J. Jiang. *Large Sample Techniques for Statistics*. Springer Texts in Statistics. Springer, New York, 2010.
- A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

7

The Delta Method, Asymptotic Linearity and Influence Functions

Chapters 5 and 6 produce limiting distributions for sample averages. Many estimators are not themselves sample averages: estimators of variances, ratios, correlations and regression coefficients are smooth functions of averages. This chapter shows that a smooth function of an asymptotically normal estimator is asymptotically normal, and then extracts from that argument the object which turns out to organize the rest of the sequence.

Why we cover this

One theorem carries the normal approximation from an average to smooth functions of averages, and it does so by a single Taylor expansion. The expansion also writes the estimator, to first order, as an average of a function of one observation at a time. That function determines the limiting distribution and the asymptotic variance, gives the first-order effect of a single observation, and is the starting point for comparing estimators of the same parameter. Quite an influential function indeed.

7.1 The Delta Method

Theorem 7.1 (Delta method) *Let $g : \mathbf{R}^k \rightarrow \mathbf{R}^m$ be Borel measurable and differentiable at $\theta \in \mathbf{R}^k$, with derivative matrix $g'(\theta)$ of dimension $m \times k$. Let $\hat{\theta}_n$ be random vectors and $a_n \rightarrow \infty$ constants with*

$$a_n(\hat{\theta}_n - \theta) \xrightarrow{d} V .$$

Then

$$a_n\{g(\hat{\theta}_n) - g(\theta)\} \xrightarrow{d} g'(\theta) V . \quad (7.1)$$

The proof is four steps. Step 1 shows the estimator is close to θ , which

is what lets a statement about g near θ be used at all. Step 2 is the Taylor expansion. Step 3 shows the remainder vanishes at the rate a_n . Step 4 assembles.

PROOF. Differentiability at θ means that the function

$$\rho(\vartheta) = g(\vartheta) - g(\theta) - g'(\theta)(\vartheta - \theta) \quad (7.2)$$

satisfies $\|\rho(\vartheta)\|/\|\vartheta - \theta\| \rightarrow 0$ as $\vartheta \rightarrow \theta$.

Step 1: $\hat{\theta}_n$ is consistent. By Lemma 4.2, convergence in distribution makes $a_n(\hat{\theta}_n - \theta) = O_P(1)$, and $a_n^{-1} \rightarrow 0$ is $o_P(1)$ as a deterministic sequence. Their product is $o_P(1)$ by Lemma 4.3(iii), so

$$\hat{\theta}_n - \theta = a_n^{-1} \cdot a_n(\hat{\theta}_n - \theta) \xrightarrow{P} 0.$$

This is what makes a statement about ϑ near θ applicable at $\hat{\theta}_n$.

Step 2: expand. By the definition (7.2) of ρ , evaluated at $\hat{\theta}_n$ and multiplied by a_n ,

$$a_n\{g(\hat{\theta}_n) - g(\theta)\} = g'(\theta) a_n(\hat{\theta}_n - \theta) + a_n\rho(\hat{\theta}_n). \quad (7.3)$$

This is an identity, not yet an approximation. The first term is a fixed linear map applied to something with a known limit; everything turns on the second.

Step 3: the remainder is $o_P(1)$. Define $h : \mathbf{R}^k \rightarrow \mathbf{R}$ by

$$h(\vartheta) = \frac{\|\rho(\vartheta)\|}{\|\vartheta - \theta\|} \text{ for } \vartheta \neq \theta, \quad h(\theta) = 0,$$

which is Borel because g is. Differentiability of g at θ says exactly that $h(\vartheta) \rightarrow 0 = h(\theta)$ as $\vartheta \rightarrow \theta$, so h is continuous at θ . Step 1 makes the limit of $\hat{\theta}_n$ the constant θ , which therefore lies in the continuity set of h with probability one, and Theorem 3.1 gives

$$h(\hat{\theta}_n) \xrightarrow{P} 0.$$

Now factor the remainder as $a_n\|\rho(\hat{\theta}_n)\| = h(\hat{\theta}_n) \cdot a_n\|\hat{\theta}_n - \theta\|$, valid on $\{\hat{\theta}_n \neq \theta\}$ and trivially true elsewhere. The first factor is $o_P(1)$ and the second is $O_P(1)$, so Lemma 4.3(iii) gives $a_n\rho(\hat{\theta}_n) \xrightarrow{P} 0$.

Nothing in this step used more than continuity of h at the single point θ , which is all differentiability at a point supplies.

Step 4: assemble. The map $v \mapsto g'(\theta)v$ is linear, hence continuous everywhere, so Theorem 4.2 gives $g'(\theta)a_n(\hat{\theta}_n - \theta) \xrightarrow{d} g'(\theta)V$. By Step 3 the two sides of (7.3) differ by $o_P(1)$, and Theorem 4.3 says that a difference tending to zero in probability does not change a limit in distribution. This gives (7.1). ■

Cf. van der Vaart [1998, Theorem 3.1, p. 26].

The most common application takes $a_n = \sqrt{n}$ and V normal, in which case (7.1) says a smooth function of an asymptotically normal estimator is asymptotically normal with a sandwiched variance.

Corollary 7.1 If $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \Sigma)$ and g is differentiable at θ , then

$$\sqrt{n}\{g(\hat{\theta}_n) - g(\theta)\} \xrightarrow{d} N(0, g'(\theta) \Sigma g'(\theta)') . \quad (7.4)$$

PROOF. Apply Theorem 7.1 with $V \sim N(0, \Sigma)$. A linear image of a normal vector is normal with mean $g'(\theta) \cdot 0 = 0$ and covariance $\text{Var}[g'(\theta)V] = g'(\theta)\Sigma g'(\theta)'$. ■

Example 7.1 (The sample variance) Suppose $W_1, \dots, W_n$ are i.i.d. with a finite fourth moment, and let $\hat{\sigma}_n^2$ be the plug-in estimator (3.6). Section 3.2 wrote it as a function of two averages, $\hat{\sigma}_n^2 = g(A_n, B_n)$ with

$$A_n = \frac{1}{n} \sum_{i \leq n} W_i^2 , \quad B_n = \bar{W}_n , \quad g(a, b) = a - b^2 .$$

The pair (A_n, B_n) is a vector of sample averages of (W^2, W) , which has a finite covariance matrix Σ because $E[W^4] < \infty$, so Theorem 5.4 gives $\sqrt{n}\{(A_n, B_n) - (E[W^2], \mu)\} \xrightarrow{d} N(0, \Sigma)$. The derivative is $g'(a, b) = (1, -2b)$, so at the limit point it is $(1, -2\mu)$, and Corollary 7.1 gives

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \xrightarrow{d} N(0, E[(W - \mu)^4] - \sigma^4) ,$$

where the variance is $(1, -2\mu)\Sigma(1, -2\mu)'$ written out. The fourth moment is used only here, to apply Theorem 5.4 to (W^2, W) : consistency of $\hat{\sigma}_n^2$ needed two moments, and this route to a limiting distribution needs four. ■

Cf. *van der Vaart [1998, Example 3.2, p. 26]*, with the joint limit of the mean and the variance at his *Example 3.4, p. 28*.

When the derivative vanishes

When $g'(\theta) = 0$, (7.1) gives a degenerate limit: the error is smaller than a_n^{-1} , but its distribution at that scale is not described. Carrying the expansion one term further identifies the next scale.

Theorem 7.2 (Second-order delta method) Let $g : \mathbf{R} \rightarrow \mathbf{R}$ be Borel measurable and twice differentiable at θ , with $g'(\theta) = 0$, and suppose $a_n \rightarrow \infty$ and $a_n(\hat{\theta}_n - \theta) \xrightarrow{d} V$. Then

$$a_n^2\{g(\hat{\theta}_n) - g(\theta)\} \xrightarrow{d} \frac{1}{2}g''(\theta)V^2 . \quad (7.5)$$

PROOF. The four steps are unchanged; only the expansion and the bookkeeping on rates differ. With $g'(\theta) = 0$, Taylor's theorem at θ gives

$$g(\vartheta) - g(\theta) = \frac{1}{2}g''(\theta)(\vartheta - \theta)^2 + \rho_2(\vartheta) , \quad \frac{|\rho_2(\vartheta)|}{(\vartheta - \theta)^2} \longrightarrow 0 ,$$

so multiplying by a_n^2 and evaluating at $\hat{\theta}_n$,

$$a_n^2 \{g(\hat{\theta}_n) - g(\theta)\} = \frac{1}{2} g''(\theta) \{a_n(\hat{\theta}_n - \theta)\}^2 + a_n^2 \rho_2(\hat{\theta}_n) .$$

The remainder is $o_P(1)$ by the argument of Step 3, run with $h_2(\vartheta) = |\rho_2(\vartheta)|/(\vartheta - \theta)^2$ in place of h and with $\{a_n(\hat{\theta}_n - \theta)\}^2 = O_P(1)$ in place of $a_n \|\hat{\theta}_n - \theta\|$. The leading term converges in distribution to $\frac{1}{2} g''(\theta) V^2$ by Theorem 4.2, since $v \mapsto \frac{1}{2} g''(\theta) v^2$ is continuous, and Theorem 4.3 discards the remainder. ■

Cf. *van der Vaart* [1998, §3.3, p. 31].

Example 7.2 (The coin at $q = 1/2$) Take the coin of Example 2.1 and the plug-in estimator $\hat{\sigma}_n^2 = \bar{W}_n(1 - \bar{W}_n)$ of the variance $q(1 - q)$ of a single draw, which Example 3.1 showed is consistent. Here $g(v) = v(1 - v)$, so $g'(v) = 1 - 2v$ and $g''(v) = -2$.

For $q \neq 1/2$ the derivative $g'(q) = 1 - 2q$ is nonzero, Theorem 5.3 gives $\sqrt{n}(\bar{W}_n - q) \xrightarrow{d} N(0, q(1 - q))$, and Corollary 7.1 gives

$$\sqrt{n} \{ \bar{W}_n(1 - \bar{W}_n) - q(1 - q) \} \xrightarrow{d} N(0, (1 - 2q)^2 q(1 - q)) .$$

At $q = 1/2$ that variance is zero, so the limit at rate $\sqrt{n}$ is degenerate; and the derivative vanishes exactly where $q(1 - q)$ is largest. Theorem 7.2 applies instead, with $a_n = \sqrt{n}$ and $V \sim N(0, 1/4)$, and yields

$$n \left\{ \bar{W}_n(1 - \bar{W}_n) - \frac{1}{4} \right\} \xrightarrow{d} \frac{1}{2} (-2) V^2 = -\frac{1}{4} \chi_1^2 ,$$

using $V = Z/2$. Three features of this limit are worth noticing. The rate is n rather than $\sqrt{n}$, so the estimator is more accurate at $q = 1/2$ than anywhere else. The limit is not normal. And it is supported on the negative half-line, which says the estimator sits below the truth with probability tending to one — as it must, since $\bar{W}_n(1 - \bar{W}_n)$ cannot exceed $1/4$ and $q(1 - q)$ equals it. ■

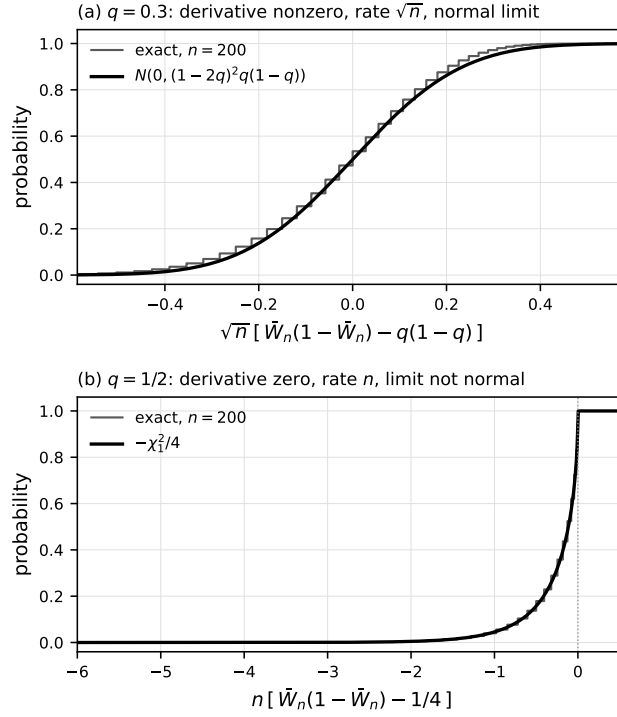


FIGURE 7.1: The two regimes of Example 7.2, with the exact distribution function of the rescaled estimator drawn against its limit at $n = 200$. Panel (a) takes $q = 0.3$, where $g'(q) \neq 0$: the rate is $\sqrt{n}$ and the limit is normal. Panel (b) takes $q = 1/2$, where $g'(q) = 0$: the rate is n , the limit is $-\chi_1^2/4$, and it puts no mass to the right of zero.

7.2 Asymptotic Linearity and Influence Functions

The proof of Theorem 7.1 produced more than (7.1). Equation (7.3) together with Step 3 says that the estimator equals a linear function of $\hat{\theta}_n - \theta$ plus a negligible term, and for a smooth function of an average this writes the estimation error as an average plus a negligible remainder. That structure is worth naming, because many estimators in the sequence have it.

Definition 7.1 Let $W_1, \dots, W_n$ be i.i.d. with distribution P . An estimator $\hat{\theta}_n$ of $\theta(P) \in \mathbf{R}^d$ is asymptotically linear at P with influence function $\psi_P : \mathbf{R}^k \rightarrow \mathbf{R}^d$ if $E[\psi_P(W)] = 0$, $E\|\psi_P(W)\|^2 < \infty$, and

$$\sqrt{n}\{\hat{\theta}_n - \theta(P)\} = \frac{1}{\sqrt{n}} \sum_{i \leq n} \psi_P(W_i) + r_n, \quad r_n = o_P(1). \quad (7.6)$$

Taking $d = 1$ gives the scalar case, in which $E\|\psi_P(W)\|^2$ is the variance of $\psi_P(W)$.

The definition is exactly the display of Example 4.6, where the point was that the remainder has to be $o_P(1)$ and not merely small. Once the representation is established, the limiting distribution and asymptotic variance follow from a single function of one observation.

Theorem 7.3 *If $\hat{\theta}_n$ is asymptotically linear at P with influence function ψ_P , then*

$$\sqrt{n}\{\hat{\theta}_n - \theta(P)\} \xrightarrow{d} N(0, E[\psi_P(W)\psi_P(W)']) .$$

PROOF. The vectors $\psi_P(W_1), \psi_P(W_2), \dots$ are i.i.d. with mean zero and covariance matrix $E[\psi_P(W)\psi_P(W)']$, finite because $E\|\psi_P(W)\|^2 < \infty$. Theorem 5.4 applies to their standardized average and gives $n^{-1/2} \sum_{i \leq n} \psi_P(W_i) \xrightarrow{d} N(0, E[\psi_P(W)\psi_P(W)'])$. By (7.6) the left-hand side differs from that sum by $o_P(1)$, so Theorem 4.3 transfers the limit. ■

Thus, once asymptotic linearity is established, ψ_P gives the normal limit at rate $\sqrt{n}$ and its asymptotic variance $E[\psi_P\psi_P']$.

Theorem 7.4 *Let $W_1, \dots, W_n$ be i.i.d. with $\mu = E[W] \in \mathbf{R}^k$ and a finite covariance matrix, let $g : \mathbf{R}^k \rightarrow \mathbf{R}^d$ be Borel measurable and differentiable at μ , and let $\theta(P) = g(\mu)$ with $\hat{\theta}_n = g(\bar{W}_n)$. Then $\hat{\theta}_n$ is asymptotically linear with*

$$\psi_P(w) = g'(\mu)(w - \mu) . \quad (7.7)$$

PROOF. Apply (7.3) with $\hat{\theta}_n = \bar{W}_n$, $\theta = \mu$ and $a_n = \sqrt{n}$, which is legitimate because Theorem 5.4 supplies the required convergence in distribution and Step 3 supplies the remainder bound:

$$\sqrt{n}\{g(\bar{W}_n) - g(\mu)\} = g'(\mu)\sqrt{n}(\bar{W}_n - \mu) + o_P(1) .$$

Since $\sqrt{n}(\bar{W}_n - \mu) = n^{-1/2} \sum_{i \leq n} (W_i - \mu)$ and $g'(\mu)$ is a fixed $d \times k$ matrix, the leading term is $n^{-1/2} \sum_{i \leq n} g'(\mu)(W_i - \mu)$, which is (7.6) with ψ_P as claimed. It has mean zero, and its variance is finite because that of W is. ■

Example 7.3 (Three influence functions) Theorem 7.4 is a recipe: differentiate, evaluate at the population mean, apply to the centred observation.

The mean. With g the identity, $\psi_P(w) = w - \mu$ and $E[\psi_P^2] = \text{Var}[W]$. Nothing is gained, but it fixes the scale: the influence function of a sample mean is the observation itself, centred.

The variance. From Example 7.1, with W replaced by the pair (W^2, W) and $g' = (1, -2\mu)$,

$$\psi_P(w) = (w^2 - E[W^2]) - 2\mu(w - \mu) = (w - \mu)^2 - \sigma^2 ,$$

after collecting terms. The influence function of the sample variance is the centred squared deviation, and $E[\psi_P^2] = E[(W - \mu)^4] - \sigma^4$ reproduces the variance computed in Example 7.1.

The two groups. Let $W = (Y, X)$ with $X \in \{0, 1\}$, $\pi = P\{X = 1\}$, and $\theta(P) = \mu_1 - \mu_0$. Rearranging (5.8) and its counterpart as in Example 5.4,

$$\psi_P(y, x) = \frac{x(y - \mu_1)}{\pi} - \frac{(1 - x)(y - \mu_0)}{1 - \pi}. \quad (7.8)$$

Its two terms cannot both be nonzero, since $x(1 - x) = 0$, so $E[\psi_P^2] = \sigma_1^2/\pi + \sigma_0^2/(1 - \pi)$, which is the variance Example 5.4 obtained by a longer route. The influence function delivers directly what took a multivariate central limit theorem, a law of large numbers and two applications of Slutsky. ■

Example 7.4 (Least squares) Let $(Y_1, X_1), \dots, (Y_n, X_n)$ be i.i.d. with $X_i \in \mathbf{R}^k$, let $Y_i = X_i' \beta + U_i$ with $E[X_i U_i] = 0$, and suppose $Q = E[XX']$ is nonsingular and $E\|XU\|^2 < \infty$. Write $\hat{Q}_n = n^{-1} \sum_{i \leq n} X_i X_i'$, and set $\hat{\beta}_n = 0$ on the event that $\hat{Q}_n$ is singular. Since $\hat{Q}_n \xrightarrow{P} Q$ and the determinant is continuous, $\det \hat{Q}_n \xrightarrow{P} \det Q \neq 0$, so that event has probability tending to zero and cannot affect a limit. Off it, the least squares estimator satisfies

$$\sqrt{n}(\hat{\beta}_n - \beta) = \hat{Q}_n^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i \leq n} X_i U_i,$$

which is an identity, obtained by substituting $Y_i = X_i' \beta + U_i$ into the normal equations. Now separate the two factors. The weak law and Theorem 3.1 applied to matrix inversion, continuous at every nonsingular matrix, give $\hat{Q}_n^{-1} \xrightarrow{P} Q^{-1}$; and the average of the $X_i U_i$ is $O_P(1)$ by Theorem 5.4 and Lemma 4.2. Writing $\hat{Q}_n^{-1} = Q^{-1} + (\hat{Q}_n^{-1} - Q^{-1})$ therefore splits the display into

$$\sqrt{n}(\hat{\beta}_n - \beta) = Q^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i \leq n} X_i U_i + o_P(1) \cdot O_P(1),$$

and the second term is $o_P(1)$ by Lemma 4.3(iii). So least squares is asymptotically linear with

$$\psi_P(y, x) = Q^{-1} x (y - x' \beta), \quad (7.9)$$

and Theorem 7.3 gives $\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, Q^{-1} \Omega Q^{-1})$ with $\Omega = E[XX'U^2]$. That is the sandwich formula, and (7.9) is where it comes from. ■

Remark 7.1 (Reading ψ_P) Beyond supplying the variance, ψ_P has a reading that explains the name. Equation (7.6) says $\hat{\theta}_n - \theta(P) \approx n^{-1} \sum_{i \leq n} \psi_P(W_i)$, so in the leading term a single observation W_i contributes $\psi_P(W_i)/n$ to the estimation error, and $\psi_P(w)$ is the first-order effect of an

observation at w . The sample mean's $\psi_P(w) = w - \mu$ is unbounded; the sample median's, under a positive density at the median, is bounded. This is a first-order reading under sampling from P , not a statement about what one arbitrary observation does to a given sample.

It also makes estimators comparable. Two estimators of the same $\theta(P)$ have two influence functions, and by Theorem 7.3 the one with the smaller $E[\psi_P^2]$ has the smaller asymptotic variance. Asking which influence functions are attainable for a given $\theta(P)$ is how efficiency is often defined. ■

Remark 7.2 (M-estimation in one display) Let $\hat{\theta}_n$ maximize $M_n(\theta) = n^{-1} \sum_{i \leq n} m(W_i, \theta)$ over Θ , as in Chapter 3, and suppose m is smooth in θ . Rather more than that is needed, and the second half of the quarter supplies it: $\hat{\theta}_n$ consistent for an interior θ_0 , so that the first-order condition $\nabla M_n(\hat{\theta}_n) = 0$ holds with probability approaching one; a central limit theorem for the score $n^{-1/2} \sum_i \nabla m(W_i, \theta_0)$; convergence of the Hessian uniformly in a neighbourhood of θ_0 ; and a nonsingular limiting Hessian H . Granting all of that, expanding the first-order condition about θ_0 and controlling the remainder as in Step 3 gives

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -H^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i \leq n} \nabla m(W_i, \theta_0) + o_P(1), \quad (7.10)$$

where $H = E[\nabla^2 m(W, \theta_0)]$, so the influence function is $\psi_P(w) = -H^{-1} \nabla m(w, \theta_0)$. Compare (7.9): least squares is the case $m(w, \theta) = -\frac{1}{2}(y - x'\theta)^2$, for which $\nabla m(w, \theta) = x(y - x'\theta)$ and $H = -Q$. The second half of the quarter establishes conditions under which (7.10) holds, and extends it to estimators defined by moment conditions rather than maximization. ■

Where we will use this

- **In this class.** Theorem 7.1 converts the limits of Chapters 5 and 6 into limits for anything smooth built from them, and the representation (7.6) is what Chapter 9 assembles into a confidence interval for a difference in means. Chapter 8 asks when the normal approximation these arguments produce is accurate and whether it holds uniformly; Example 7.2 is already a warning, since the limit changes shape at a single parameter value.
- **Later in the sequence.** Influence functions are the language of semiparametric efficiency and of double robustness in the spring, and, under additional conditions, the bootstrap reproduces the distribution of the leading sum in (7.6).
- **In recent research.** Some estimators with machine-learning first stages are constructed so that the error in estimating the nuisance function has no first-order effect on (7.6); under conditions on the

accuracy of the first stage, they remain asymptotically linear in the sense of Definition 7.1.

7.3 Key Concepts

- **The delta method.** A function differentiable at the limit point carries convergence in distribution through, with the derivative acting linearly on the limit. The proof is a Taylor expansion plus the o_P/O_P calculus.
- **Degenerate derivatives.** When $g'(\theta) = 0$ the first-order limit is degenerate, and if $g''(\theta) \neq 0$ the rate improves to a_n^2 , with limit $\frac{1}{2}g''(\theta)V^2$; when V is normal this is a scaled χ_1^2 , not a normal.
- **Asymptotic linearity.** Many estimators equal an average of $\psi_P(W_i)$ plus $o_P(1)$. The remainder must be shown to be $o_P(1)$, not merely small.
- **Influence functions.** ψ_P gives the limiting distribution, the asymptotic variance $E[\psi_P^2]$, and the first-order effect of one observation. Comparing estimators of the same parameter is comparing influence functions.
- **The recipe.** Write the estimator as a smooth function of averages, differentiate at the population value, and read off ψ_P . The limiting distribution and asymptotic variance follow.

7.4 Concluding Remarks

van der Vaart [1998, Chapter 3] is the reference for the delta method, with the sample variance at his Example 3.2, the degenerate case at his §3.3, and variance-stabilizing transformations at his §3.2 — an application we do not cover, in which g is chosen so that the limiting variance does not depend on the parameter. His Chapter 20 develops influence functions properly, as derivatives of the map $P \mapsto \theta(P)$; §20.1 has the definition and Examples 20.2 and 20.5 compute the mean and the quantile. The functional delta method of his Theorem 20.8, p. 297, is what extends Theorem 7.1 to functionals of the empirical distribution.

Figure 7.1 is generated by `fig-degenerate.py`, in this chapter's folder under `code/`.

7.5 Problems

Problem 7.1 Let $W = (Y, X)$ with $E[Y] = \mu_Y$, $E[X] = \mu_X \neq 0$, and a finite covariance matrix, and let $\hat{\theta}_n = \bar{Y}_n / \bar{X}_n$ estimate $\theta(P) = \mu_Y / \mu_X$.

1. Apply Corollary 7.1 to obtain the limiting distribution, and write the asymptotic variance in terms of $\text{Var}[Y]$, $\text{Var}[X]$ and $\text{Cov}[Y, X]$.
2. Compute the influence function and verify that $E[\psi_P^2]$ agrees with part (a).
3. The estimator is undefined when $\bar{X}_n = 0$. Explain why this does not affect either answer, and say which hypothesis of Theorem 7.1 would fail if $\mu_X = 0$.

Problem 7.2 Let $W_1, \dots, W_n$ be i.i.d. with $W > 0$ almost surely, $\mu = E[W]$ and $\text{Var}[W] = \sigma^2 \in (0, \infty)$, so that $\bar{W}_n > 0$ and $\log \bar{W}_n$ is always defined.

1. Find the limiting distribution of $\sqrt{n}(\log \bar{W}_n - \log \mu)$.
2. A transformation g is called variance stabilizing if the asymptotic variance in (7.4) does not depend on the parameter. For the coin, with $\hat{\theta}_n = \bar{W}_n$ and $\sigma^2(q) = q(1 - q)$, find g with $g'(q)^2 q(1 - q)$ constant, and identify the resulting transformation.
3. Explain in two or three sentences why such a g is useful for constructing confidence intervals.

Problem 7.3 Return to Example 7.2.

1. Verify directly that $\bar{W}_n(1 - \bar{W}_n) \leq 1/4$ for every realization, and explain how this is consistent with the sign of the limit.
2. Suppose $q_n = 1/2 + c/\sqrt{n}$ for a constant $c \neq 0$ and all n large enough that this lies in $(0, 1)$, the finitely many earlier values being defined arbitrarily in $(0, 1)$; this is a drifting sequence of the kind in Example 6.4. Find the limiting distribution of $n\{\bar{W}_n(1 - \bar{W}_n) - q_n(1 - q_n)\}$ and compare it with the limit at $q = 1/2$.
3. What does part (b) suggest about the accuracy of the normal approximation from Corollary 7.1 when q is close to, but not equal to, $1/2$?

Problem 7.4 1. Show that if $\hat{\theta}_n$ is asymptotically linear with influence function ψ_P , then ψ_P is determined up to changes on a set of probability zero.

2. Let $\hat{\theta}_n$ and $\tilde{\theta}_n$ both be asymptotically linear for the same $\theta(P)$, with influence functions ψ_P and $\tilde{\psi}_P$. Show that $\sqrt{n}(\hat{\theta}_n - \tilde{\theta}_n) \xrightarrow{P} 0$ if and only if $\psi_P = \tilde{\psi}_P$ almost surely.

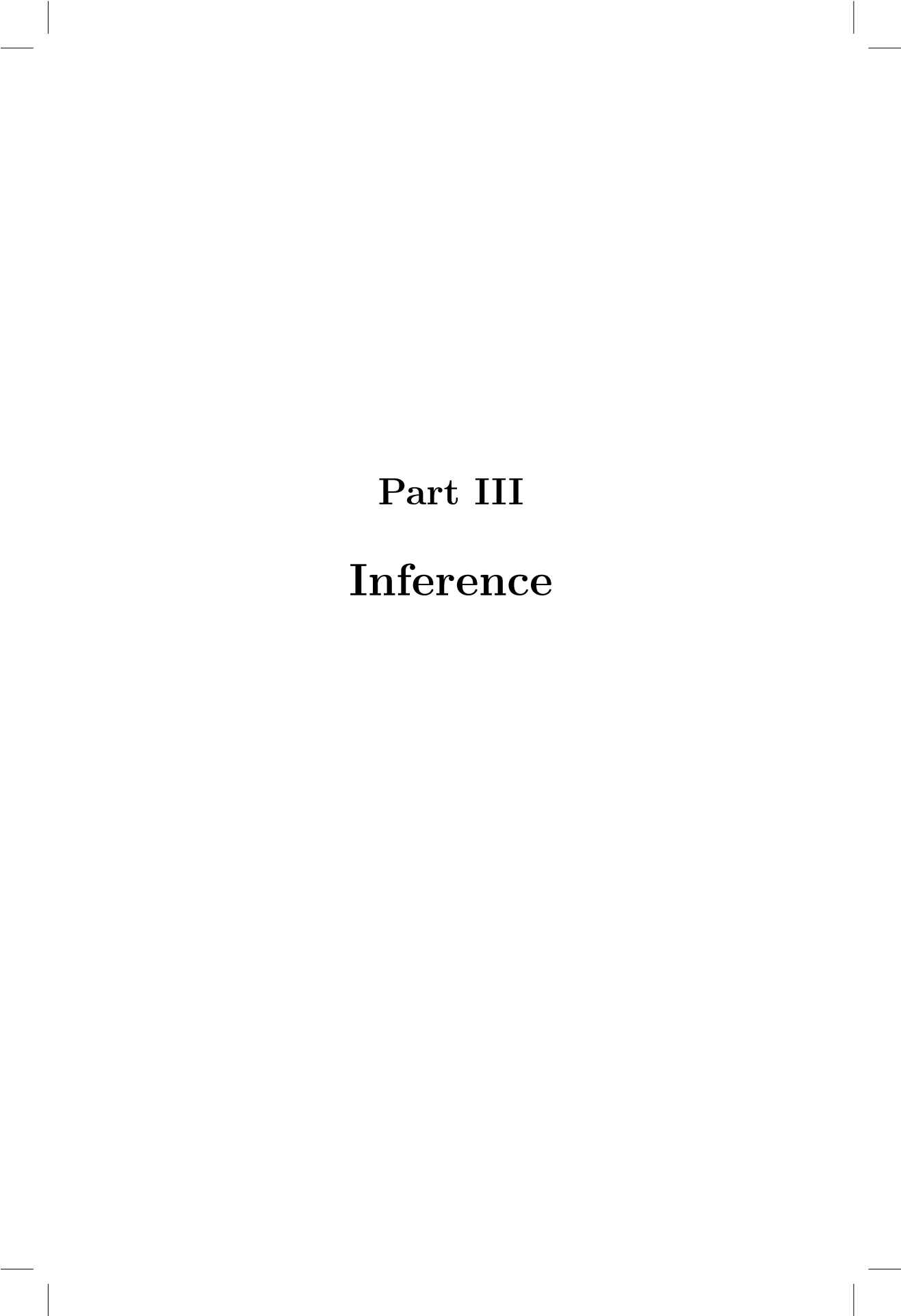
3. Give two estimators of the same $\theta(P)$ with different influence functions, and say which has the smaller asymptotic variance.

Problem 7.5 Verify (7.8) by writing $\hat{\mu}_{n,1} - \hat{\mu}_{n,0}$ as a function of three sample averages and applying Theorem 7.4. State the moment conditions you need, and say where the assumption $\pi \in (0, 1)$ enters.

Bibliography

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.





Part III

Inference



8

Asymptotic Inference and Its Limits

Most results so far have been limits: statements about what happens as n grows, with nothing said about the sample in hand. This chapter asks two questions such statements leave open. How accurate is the approximation at a given n ? And does the approximation hold across the whole model, or only at one distribution at a time? For the coin, both turn out to be governed by the same quantity, and the second question has an answer that changes how asymptotic claims should be read.

Why we cover this

“Asymptotically valid” is a weaker claim than it sounds. A procedure can be correct in the limit at every distribution in a model and still fail badly, at every sample size, at some distribution in that model — because the distribution where it fails moves as n grows. Knowing which claim one has is the difference between a correct paper and an incorrect one. The same machinery also illustrates a question the earlier chapters could not even pose: given two tests that are both asymptotically valid, which is better?

8.1 What an Asymptotic Statement Promises

We first define the validity claims used in this chapter; the second half of the quarter develops them fully.

A model is a set $\mathbf{P}$ of distributions, and a null hypothesis is a subset $\mathbf{P}_0 \subseteq \mathbf{P}$. A *test* is a sequence of statistics ϕ_n taking values in $[0, 1]$, read as the probability of rejecting; a test taking only the values 0 and 1 is *nonrandomized*, and every test in this chapter is of that kind. Its *power function* is

$$\beta_n(P, \phi_n) = \mathbb{E}_P[\phi_n] ,$$

the probability of rejection when the data are generated by P . A test has level α in finite samples if $\sup_{P \in \mathbf{P}_0} \beta_n(P, \phi_n) \leq \alpha$ for every n . The asymptotic versions come in two forms, and the difference between them is the subject of

Section 8.3:

$$\limsup_{n \rightarrow \infty} \beta_n(P, \phi_n) \leq \alpha \quad \text{for each } P \in \mathbf{P}_0, \quad (8.1)$$

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathbf{P}_0} \beta_n(P, \phi_n) \leq \alpha. \quad (8.2)$$

Condition (8.1) is *pointwise* asymptotic validity and (8.2) is *uniform* asymptotic validity. The supremum in (8.2) sits inside the limit, and that is the whole difference.

Unwinding the definitions shows what is at stake. Uniform validity says that for every $\varepsilon > 0$ there is a single N_ε with $\beta_n(P, \phi_n) \leq \alpha + \varepsilon$ for all $n \geq N_\varepsilon$ and *every* $P \in \mathbf{P}_0$: one sample size that works throughout the model. Pointwise validity gives an $N_{\varepsilon, P}$ that may depend on P and may grow without bound as P ranges over $\mathbf{P}_0$, in which case no sample size works for all of them at once. Since the null hypothesis does not say which P generated the data, the second guarantee may be worth very little.

The same distinction applies to confidence sets. A sequence C_n covers uniformly at level $1 - \alpha$ if

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathbf{P}} P\{\theta(P) \in C_n\} \geq 1 - \alpha, \quad (8.3)$$

and pointwise if the infimum is dropped and the statement made for each fixed P .

8.2 How Good Is the Approximation?

The central limit theorem says a distribution function converges. It says nothing about how far apart the two functions are at any particular n , and the following theorem does.

Theorem 8.1 (Berry–Esseen) *Let $W_1, \dots, W_n$ be i.i.d. with $E[W] = \mu$, $\text{Var}[W] = \sigma^2 \in (0, \infty)$ and $\rho = E|W - \mu|^3 < \infty$, and let $R_n = \sqrt{n}(\bar{W}_n - \mu)/\sigma$. Then there is an absolute constant C , not depending on n or on the distribution of W , with*

$$\sup_{x \in \mathbf{R}} |P\{R_n \leq x\} - \Phi(x)| \leq \frac{C \rho}{\sigma^3 \sqrt{n}}. \quad (8.4)$$

Cf. Vershynin [2026, Theorem 2.1.4, p. 27]. We do not prove it.

Four features of (8.4) are worth separating. First, it bounds a supremum over x , so it is a statement about the whole distribution function at once — the object Remark 4.1 said must vanish, now supplied with a rate.

Second, it holds at every n , not in a limit, which puts it with the inequalities

of Chapter 2 — Markov, Chebyshev, Hoeffding — rather than with the limit theorems that came after them. A bound holding at every sample size is the kind of statement that answers a question about the sample one has, and Theorem 8.1 is the first time such a bound has been available for the normal approximation itself.

Third, the distribution enters through the single quantity ρ/σ^3 , which is unaffected by shifting or rescaling W . It is a measure of how heavy the tails are relative to the spread, and it is therefore an answer to the question of *what* makes the bound large, not merely how large it is.

Lastly, it costs a third moment where the central limit theorem needed only a second. That is the price of a rate, and it is a general pattern: convergence is cheap, speed is not.

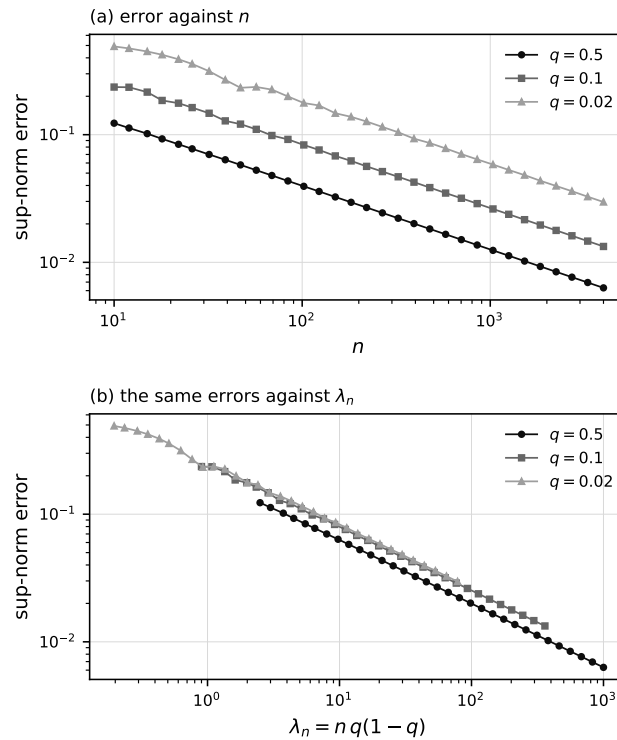


FIGURE 8.1: The largest vertical gap between the exact distribution function of R_n and Φ , for the coin at three values of q . Panel (a) plots it against n ; the three curves are far apart, and the approximation at $q = 0.02$ with $n = 1000$ is worse than at $q = 0.5$ with $n = 100$. Panel (b) plots the same numbers against $\lambda_n = nq(1-q)$, and they line up almost exactly.

Example 8.1 (The coin) Take $W \sim \text{Bernoulli}(q)$. Then $\sigma^2 = q(1-q)$ and

$$\rho = q(1-q)^3 + (1-q)q^3 = q(1-q)\{(1-q)^2 + q^2\},$$

so the quantity governing (8.4) is

$$\frac{\rho}{\sigma^3 \sqrt{n}} = \frac{(1-q)^2 + q^2}{\sqrt{n} q(1-q)} = \frac{(1-q)^2 + q^2}{\sqrt{\lambda_n}}, \quad (8.5)$$

with $\lambda_n = nq(1-q)$ as in (6.9), the variance of the number of successes. The numerator lies between $1/2$ and 1 whatever q is, so the bound is of order $\lambda_n^{-1/2}$. In this sense λ_n rather than n is the effective sample size in the bound.

That is the same λ_n that Chapter 6 produced, by a completely different route. There it decided whether a normal limit exists at all; here it controls the bound on the approximation error at a given n . ■

Remark 8.1 (Edgeworth expansions) Theorem 8.1 bounds the error. An Edgeworth expansion describes it. Under a smoothness hypothesis on the distribution of W — *Cramér’s condition*, that $\limsup_{|t| \rightarrow \infty} |\varphi_W(t)| < 1$ — and a finite fourth moment, the distribution function of R_n equals $\Phi(x)$ plus an explicit term of order $n^{-1/2}$ proportional to the skewness of W , followed by a term of order n^{-1} and a remainder $o(n^{-1})$, uniformly in x . When W is symmetric the skewness term vanishes and the normal-approximation error is $O(n^{-1})$.

The smoothness requirement is not a technicality, and the coin is the reason. A *lattice* variable, one supported on a set of the form $\{a + bk : k \in \mathbb{Z}\}$, satisfies $|\varphi_W(2\pi m/b)| = 1$ for every integer m ; those frequencies run off to infinity, so Cramér’s condition fails. A Bernoulli variable is lattice. For lattice distributions the expansion carries an additional discontinuous term of order $n^{-1/2}$ that survives even under symmetry. Figure 8.1 shows it: at $q = 1/2$ the summands are symmetric, yet the error still falls at rate $n^{-1/2}$ rather than n^{-1} .

The same expansion is what lies behind the claim that the bootstrap refines the normal approximation. A procedure that removes the $n^{-1/2}$ term is accurate to a higher order than one that does not, but the statement holds for particular statistics under regularity conditions of this kind, not for resampling as such. ■

Cf. Lehmann and Romano [2005, §11.4.1, p. 459].

Remark 8.2 (Where bounds replace limits) Theorem 8.1 looks like a technical refinement of the central limit theorem. In one setting it is not a refinement but a replacement, and that setting is now a large part of the field.

Suppose the object of interest is a vector in $\mathbf{R}^p$ with $p = p_n$ growing along with n — one coordinate per covariate, per moment condition, or per hypothesis being tested simultaneously. The fixed-dimensional convergence results of Chapter 4 then do not directly give the approximation needed, because the object being approximated lives in a different space at each sample size. What can still be true is a bound: that the distribution of a standardized vector of

sample averages and a Gaussian vector of the same dimension assign nearly the same probability to every rectangle in $\mathbf{R}^p$, with an error that is explicit in n and p . Under suitable moment and variance conditions such bounds remain small even when p exceeds n , and they are what makes simultaneous inference over many parameters possible.

The same shift explains the shape of much recent work. Many guarantees for estimators built on machine-learning first stages are stated as finite-sample bounds, and the validity of the bootstrap in high dimensions is proved as a Gaussian approximation bound rather than as a limit theorem. It is also not a new theme in these notes: the concentration inequalities of Section 2.3 were the first appearance of the same move, from statements about limits to statements that hold at every n , and Vershynin [2026] develops those tools. ■

8.3 Is the Approximation Uniform?

The word *uniform* has now been used for three different things, and the third is the one meant here. Theorem 3.3 was uniform over a parameter θ indexing a criterion function. Remark 4.1 and Theorem 8.1 are uniform over x , the point at which a distribution function is evaluated. What follows is uniformity over P : over which distribution generated the data.

How much that quantifier asks depends on how large the model is, and it is worth naming three classes to see the range. For real-valued W , take

$$\mathbf{P}_{\text{Ber}} = \{\text{Bernoulli}(q) : q \in (0, 1)\} , \quad \mathbf{P}_{\text{N}} = \{N(\mu, \sigma^2) : \mu \in \mathbf{R}, \sigma^2 > 0\} ,$$

and

$$\mathbf{P}_2 = \left\{ P \text{ on } \mathbf{R} : \int w^2 dP(w) < \infty \right\} .$$

The first two are finite-dimensional parametric models and the third is non-parametric, restricted only by a moment condition. A procedure can behave well over a parametric family and badly over a class like $\mathbf{P}_2$, and Section 2.1 recorded that the third kind is what this course usually assumes.

That third quantifier is the one that matters for inference, because it ranges over what the hypothesis does not pin down. Testing $E[W] = c_0$ says nothing about $\text{Var}[W]$, so the supremum runs over every value the variance might take; the same is true of the within-group variances in a comparison of means. Those unrestricted features are the *nuisance parameters*, and a procedure is only as good as its worst case over them. The coin is atypical here in the way it has been atypical throughout: $\text{Bernoulli}(q)$ is a one-parameter family, so there is no nuisance parameter separate from q , and the supremum over the model is simply a supremum over q .

Equation (8.5) contains a warning, though it must be read carefully. The bound is small when λ_n is large, and λ_n depends on q as well as on n . For

any fixed $q \in (0, 1)$ it grows without bound, so the approximation eventually becomes good; but no single sample size makes the bound small for all q at once, because q can always be chosen small enough to spoil it.

What that shows is a failure of the *guarantee*, not of the approximation: an upper bound that is large is consistent with an error that is small, and (8.5) by itself proves nothing about uniform behaviour. It is a reason to look, and looking is worthwhile, because on the coin the failure is real and can be computed exactly.

Example 8.2 (The Wald interval for the coin) The interval that Chapter 4 was building toward is

$$C_n = \left[\bar{W}_n \pm z_{1-\alpha/2} \sqrt{\frac{1}{n} \bar{W}_n (1 - \bar{W}_n)} \right], \quad (8.6)$$

obtained by inverting the studentized root $\tilde{R}_n$ of Example 4.5 on the event $\{0 < \bar{W}_n < 1\}$, and defined by the display itself otherwise. That qualification is not bookkeeping: the boundary samples are excluded from the inversion and are exactly the ones that break what follows. For each fixed $q \in (0, 1)$ the coverage of (8.6) tends to $1 - \alpha$, which is what Example 4.5 established.

Now fix n and let q be small. The event $\{\bar{W}_n = 0\}$ has probability $(1 - q)^n$, and on it the half-width of (8.6) is zero, so $C_n = \{0\}$. Since $q > 0$, the interval misses. Hence

$$P_q\{q \in C_n\} \leq 1 - (1 - q)^n,$$

which tends to zero as $q \downarrow 0$ with n held fixed. So

$$\inf_{q \in (0, 1)} P_q\{q \in C_n\} = 0 \quad \text{for every } n,$$

and (8.3) fails as badly as it can. The interval is pointwise asymptotically valid and uniformly worthless: it guarantees nothing for a given sample size. ■

The calculation is elementary, and it is worth seeing what it does and does not say. It does not say the interval is bad for any particular q ; for every fixed q the coverage does converge. It says that the q at which coverage is poor moves as n grows, always staying ahead of the sample size, so that no n is ever large enough for the whole model.

Remark 8.3 (Sequences, not points) Making that precise requires a sequence rather than a point. Taking $q_n = \lambda/n$ for fixed $\lambda > 0$ gives $P\{\bar{W}_n = 0\} = (1 - \lambda/n)^n \rightarrow e^{-\lambda}$, so the limit superior of coverage along that sequence is at most $1 - e^{-\lambda}$, which is far below $1 - \alpha$ when λ is small. This is the regime of Example 6.4, where the same sequence produced a Poisson limit rather than a normal one, and it is why the failure occurs: along $q_n = \lambda/n$ the statistic is not approximately normal at all.

Drifting sequences are not a trick for manufacturing counterexamples. They are what a supremum over $\mathbf{P}$ forces. A quantity like $\sup_P \beta_n(P, \phi_n)$

is approached, for each n , at some P_n near the worst case, and those P_n generally move. Studying $\sup_P \beta_n$ therefore means studying $\beta_n(P_n, \phi_n)$ along sequences, and passing to a subsequence when necessary to make the limit exist. That is the whole technique. ■

Remark 8.4 (A convention that was not innocent) Example 4.5 set $\tilde{R}_n$ to zero on the event $\{\bar{W}_n \in \{0, 1\}\}$ and dismissed it, on the ground that its probability $q^n + (1 - q)^n$ tends to zero. That is true for every fixed q and false uniformly: $\sup_q \{q^n + (1 - q)^n\} = 1$ for every n . The event we agreed to ignore is exactly the one that destroys uniform coverage. Conventions of that kind are safe pointwise and are worth revisiting whenever a uniform claim is being made. ■

Where this bites in practice is a short list, and every item on it has the same structure: a parameter of the model controls how hard the problem is, and the difficulty can be made to keep pace with the sample size. Inference on a coefficient identified only through a weak instrument can degrade as the instrument's strength approaches zero. Inference on a parameter at the boundary of its space — a variance that may be zero, an inequality that may bind — can degrade as the boundary is approached. And inference after a model has been selected on the same data can degrade near the parameter values at which the selection decision is close. Whether a given procedure fails in a given model is something to be checked rather than assumed, but in each of these settings standard procedures often do fail, and in each case the diagnosis is a drifting sequence. This also means that *how* the asymptotics are done matters: some devices illuminate a phenomenon that is present in finite samples, and others hide it. The next section illustrates one such case.

8.4 Local Asymptotics

Drifting sequences often lead to constructive work, and the clearest instance is a question none of the machinery so far can even pose: given two tests of the same null hypothesis, both asymptotically of level α , which one should be used?

Two tests of a location hypothesis

Let the observations $W_1, \dots, W_n$ be i.i.d. from P_θ , which has density $f(w - \theta)$ on the real line, with f symmetric about zero and its median at zero unique, and suppose $\sigma_0^2 = \int x^2 f(x) dx \in (0, \infty)$. Symmetry gives $E_\theta[W] = \theta$ and $\text{med}_\theta[W] = \theta$, so the same θ is both the mean and the median. Write F for the distribution function of P_0 , so that $F(0) = 1/2$, and assume F is

differentiable at zero with $f(0) > 0$. The hypotheses are

$$H_0 : \theta = 0 \quad \text{against} \quad H_1 : \theta > 0 .$$

Nothing here pins down f : the normal, the uniform on $[-1, 1]$, the logistic and the Laplace densities all qualify.

The first test uses the mean. With $\hat{\sigma}_n^2 = n^{-1} \sum_{i \leq n} (W_i - \bar{W}_n)^2$, Theorem 5.5 gives $\sqrt{n} \bar{W}_n / \hat{\sigma}_n \xrightarrow{d} Z$ under P_0 , so

$$\phi_{1,n} = \mathbf{1} \left\{ \frac{\sqrt{n} \bar{W}_n}{\hat{\sigma}_n} > z_{1-\alpha} \right\} \quad (8.7)$$

is asymptotically of level α , since $z_{1-\alpha}$ is a continuity point of Φ .

The second uses the median. Under H_0 the median is zero, so positive and negative observations should be equally frequent. The indicators $\mathbf{1}\{W_i > 0\}$ are Bernoulli(q) with $q = P_\theta\{W_i > 0\}$, hence with variance $q(1-q)$; under H_0 symmetry gives $q = 1/2$ and the variance is $1/4$. Writing $\hat{q}_n = n^{-1} \sum_{i \leq n} \mathbf{1}\{W_i > 0\}$ for the sample proportion, Theorem 5.3 therefore gives $2\sqrt{n}(\hat{q}_n - 1/2) \xrightarrow{d} Z$ under P_0 , and

$$\phi_{2,n} = \mathbf{1}\{2\sqrt{n}(\hat{q}_n - \frac{1}{2}) > z_{1-\alpha}\} \quad (8.8)$$

is also asymptotically of level α . This is the *sign test*. Under H_0 the count $n\hat{q}_n$ is Binomial($n, 1/2$) for every admissible f , so its null distribution does not depend on f and the level statement holds uniformly over this class.

Why fixed alternatives cannot choose between them

The natural comparison is by power. Fix $\theta > 0$ and compute.

For the first test, the manipulation that makes everything work is to re-express the probability under P_θ as a probability under P_0 . Set $V_i = W_i - \theta$, which is distributed according to P_0 when W_i is distributed according to P_θ . Then $\bar{W}_n = \bar{V}_n + \theta$, and, crucially, $\hat{\sigma}_n$ is unchanged, since centring at the sample mean removes the shift:

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i \leq n} (W_i - \bar{W}_n)^2 = \frac{1}{n} \sum_{i \leq n} (V_i - \bar{V}_n)^2 .$$

Therefore

$$\beta_n(P_\theta, \phi_{1,n}) = P_0 \left\{ \frac{\sqrt{n} \bar{V}_n}{\hat{\sigma}_n} > z_{1-\alpha} - \frac{\sqrt{n} \theta}{\hat{\sigma}_n} \right\} . \quad (8.9)$$

The left-hand side of the inequality converges in distribution to Z and the right-hand side diverges to $-\infty$ in probability, in the sense of Definition 2.2, since $\hat{\sigma}_n \xrightarrow{P} \sigma_0$ and $\theta > 0$ is fixed. So $\beta_n(P_\theta, \phi_{1,n}) \rightarrow 1$.

For the sign test the same substitution gives $E_\theta[\mathbf{1}\{W_i > 0\}] = P_0\{V_i > -\theta\} = 1 - F(-\theta)$, and symmetry with $\theta > 0$ makes $F(-\theta) < 1/2$. Hence

$\hat{q}_n \xrightarrow{P} 1 - F(-\theta) > 1/2$ by the weak law, the statistic in (8.8) diverges to $+\infty$ in probability, and $\beta_n(P_\theta, \phi_{2,n}) \rightarrow 1$ as well.

Both powers tend to one at every fixed alternative. A test with that property is called *consistent*. Both tests are consistent, so consistency does not distinguish them.

Local asymptotic power

The difficulty is the one named in Remark 6.5. Holding θ fixed while n grows lets the information overwhelm the problem, and in the limit every alternative is detected. But at the sample size actually in hand, power is not one uniformly over $\theta > 0$: it is near one for θ far from zero and well below one for θ close to zero, and what counts as close depends on n . To obtain a nontrivial limiting power comparison for these tests, let the alternative approach the null as the sample grows.

Take

$$\theta_n = \frac{h}{\sqrt{n}}, \quad h > 0, \quad (8.10)$$

which is the rate at which the two effects balance: $\sqrt{n}\theta_n = h$ neither vanishes nor diverges. The data are now a triangular array whose rows are i.i.d. from P_{θ_n} , the case named in Remark 6.1.

For the t -test, (8.9) still holds with θ_n in place of θ , and now $\sqrt{n}\theta_n = h$ exactly. Since $\hat{\sigma}_n \xrightarrow{P} \sigma_0$ and $\sqrt{n}\bar{V}_n/\hat{\sigma}_n \xrightarrow{d} Z$ under P_0 , Theorem 4.4 gives

$$\beta_n(P_{\theta_n}, \phi_{1,n}) \longrightarrow 1 - \Phi\left(z_{1-\alpha} - \frac{h}{\sigma_0}\right). \quad (8.11)$$

For the sign test the indicators are still Bernoulli, but their success probability now moves with the sample size, so Chapter 5 does not apply and Chapter 6 does. Each $\mathbf{1}\{W_{n,i} > 0\}$ is Bernoulli(q_n) with variance $q_n(1 - q_n)$, where

$$q_n = E_{\theta_n}[\mathbf{1}\{W_{n,i} > 0\}] = P_{\theta_n}\{W_{n,i} > 0\} = 1 - F(-\theta_n).$$

Differentiability of F at zero gives

$$F(-\theta_n) = F(0) - f(0)\theta_n + o(\theta_n) = \frac{1}{2} - \frac{f(0)h}{\sqrt{n}} + o(n^{-1/2}), \quad (8.12)$$

so $q_n = 1/2 + f(0)h/\sqrt{n} + o(n^{-1/2})$ and $q_n(1 - q_n) \rightarrow 1/4$. Apply Theorem 6.2 to the array

$$\zeta_{n,i} = \frac{\mathbf{1}\{W_{n,i} > 0\} - q_n}{\sqrt{n q_n(1 - q_n)}},$$

whose row variances sum to one. Lindeberg's condition holds for the reason it held in Example 6.4: the summands are bounded by $\{n q_n(1 - q_n)\}^{-1/2}$,

which tends to zero because $q_n(1 - q_n) \rightarrow 1/4 > 0$, so the indicator in (6.4) is eventually zero. Hence

$$\frac{\sqrt{n}(\hat{q}_n - q_n)}{\sqrt{q_n(1 - q_n)}} \xrightarrow{d} Z ,$$

and multiplying by the constant sequence $2\sqrt{q_n(1 - q_n)}$, which converges to one, gives $2\sqrt{n}(\hat{q}_n - q_n) \xrightarrow{d} Z$. Splitting the statistic of (8.8) into a centred part and a drift,

$$2\sqrt{n}\left(\hat{q}_n - \frac{1}{2}\right) = 2\sqrt{n}(\hat{q}_n - q_n) + 2\sqrt{n}\left(q_n - \frac{1}{2}\right) \xrightarrow{d} Z + 2f(0)h ,$$

since the second term converges to $2f(0)h$ by (8.12). Therefore

$$\beta_n(P_{\theta_n}, \phi_{2,n}) \longrightarrow 1 - \Phi(z_{1-\alpha} - 2f(0)h) . \quad (8.13)$$

Which test is better

Both limits have the form $1 - \Phi(z_{1-\alpha} - ch)$, which is increasing in c . So the comparison reduces to two numbers:

$$\text{the } t\text{-test wins} \iff \frac{1}{\sigma_0} > 2f(0) \iff 2f(0)\sigma_0 < 1 , \quad (8.14)$$

and the comparison does not depend on h or on α . The ratio of the squared constants, $\{2f(0)\sigma_0\}^2$, is the *asymptotic relative efficiency* of the sign test to the t -test.

Example 8.3 (Two densities) For the normal density with variance σ_0^2 , $f(0) = (\sigma_0\sqrt{2\pi})^{-1}$, so $2f(0)\sigma_0 = 2/\sqrt{2\pi} \approx 0.798 < 1$ and the t -test wins, with a relative efficiency of $2/\pi \approx 0.64$: the sign test needs about 1.57 times as many observations for the same local power.

For the Laplace density $f(x) = (2b)^{-1}e^{-|x|/b}$ we have $f(0) = 1/(2b)$ and $\sigma_0^2 = 2b^2$, so $2f(0)\sigma_0 = \sqrt{2} \approx 1.414 > 1$ and the sign test wins, with a relative efficiency of 2 in its favour.

The two tests can have different power at a given sample size even though both are consistent. The local-power calculation describes that difference for alternatives close to the null and shows how the ranking depends on f . ■

Where we will use this

- **In this class.** The vocabulary of Section 8.1 is developed properly in the second half of the quarter, where asymptotic size and uniform asymptotic validity are defined in the form used here and applied to extremum estimators, and where local alternatives are used to compute power. Chapter 9 constructs the interval that Section 8.3 tests, for a difference in means.

- **Later in the sequence.** High-dimensional inference in the winter uses bounds of the kind in Remark 8.2, and the bootstrap in the spring is a device for improving on the normal approximation in the sense of Remark 8.1.
- **In recent research.** Uniformity is the standard by which procedures are judged in every setting where a parameter controls the difficulty of the problem: weak instruments, moment inequalities, entry games with multiple equilibria, and inference after model selection.

8.5 Key Concepts

- **Pointwise and uniform validity.** A supremum over the model inside the limit rather than outside it. Pointwise validity supplies a sample size that may depend on the unknown distribution, and so need not supply one sample size that works throughout the model.
 - **Berry–Esseen.** The error in the normal approximation is at most $C\rho/(\sigma^3\sqrt{n})$, uniformly in x and at every n . The bound depends on the distribution only through ρ/σ^3 .
 - **One quantity, two questions.** For the coin the bound is of order $\lambda_n^{-1/2}$, with $\lambda_n = nq(1-q)$ — the same quantity that decided in Chapter 6 whether a normal limit exists.
 - **Uniformity can fail completely.** The Wald interval for the coin has coverage tending to $1 - \alpha$ at every fixed q and infimum coverage zero at every n , because the difficult q moves with the sample size.
 - **Local asymptotics.** Letting the alternative approach the null at rate $n^{-1/2}$ produces limiting power strictly between α and one, which ranks tests that fixed alternatives cannot distinguish.
-

8.6 Concluding Remarks

The material of Section 8.4 follows Ivan’s Econ 481 lecture notes, which in turn are based on notes generously shared by Azeem Shaikh. [Lehmann and Romano \[2005, §11.4.1\]](#) is the reference for Edgeworth expansions, and their Chapter

13 develops asymptotic relative efficiency and local asymptotic power in the generality Section 8.4 only illustrates. Vershynin [2026, §2.1] states Theorem 8.1.

8.7 Problems

Problem 8.1 Let $W_1, \dots, W_n$ be i.i.d. with a finite third moment and $\sigma^2 > 0$.

1. Show that ρ/σ^3 is unchanged if W is replaced by $a + bW$ for constants a and $b \neq 0$, and explain why any bound of the form (8.4) must have this property.
2. Compute ρ/σ^3 for the exponential distribution with mean one, and use (8.4) to find a sample size guaranteeing that the normal approximation is accurate to within 0.01, taking $C = 1/2$.
3. Compare that sample size with the one (8.5) gives for a coin with $q = 1/2$, and say in one sentence what accounts for the difference.

Problem 8.2 This problem derives the limiting distribution of the sample median, which the delta method of Chapter 7 cannot reach. Let $W_1, W_2, \dots$ be i.i.d. with distribution function F , let θ satisfy $F(\theta) = 1/2$, and suppose F is differentiable at θ with $f(\theta) > 0$. Let $\hat{\theta}_n$ be the sample median and take n even, so that $\hat{\theta}_n$ is the $(n/2)$ th largest observation.

1. Show that $P\{\sqrt{n}(\hat{\theta}_n - \theta) \leq x\} = P\{\sum_{i \leq n} Z_i \leq n/2 - 1\}$, where $Z_i = \mathbf{1}\{W_i > \theta + x/\sqrt{n}\}$.
2. With $\mu_n = E[Z_i]$ and $\sigma_n^2 = \text{Var}[Z_i]$, rewrite the right-hand side as a probability about a standardized average, and use Theorem 8.1 together with $E|Z_i - \mu_n|^3 \leq 8$ to show that the error in replacing it by Φ vanishes.
3. Show that the resulting argument of Φ converges to $2f(\theta)x$, and conclude that $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, 1/\{4f^2(\theta)\})$.
4. Where would the argument break if $f(\theta) = 0$?

Problem 8.3 Return to the Wald interval (8.6).

1. For $n = 100$ and $\alpha = 0.05$, find a value of q for which the coverage is below $1/2$, and verify your answer.
2. Show that the same failure occurs at q near one, and identify the event responsible.

3. The interval $\{c \in (0, 1) : |\sqrt{n}(\bar{W}_n - c)| \leq z_{1-\alpha/2}\sqrt{c(1-c)}\}$ replaces the estimated variance by the value implied by c . Show that it is never empty, and explain in two or three sentences why it avoids the particular failure in part (a). Say also what your argument does not establish about its coverage.

Problem 8.4 Continue the location model of Section 8.4.

1. Compute $2f(0)\sigma_0$ for the uniform density on $[-1, 1]$ and for the logistic density, and say which test is preferred in each case.
2. Two tests with local power $1 - \Phi(z_{1-\alpha} - c_i h)$ can be compared by the sample sizes they need for equal power. Show that if the sign test is given m observations and the t -test n , the two have the same limiting power along $\theta_n = h/\sqrt{n}$ when $m/n \rightarrow \{2f(0)\sigma_0\}^{-2}$, and evaluate that ratio for the normal.
3. Let f be the symmetric mixture $(1 - \epsilon)$ times the standard normal density plus ϵ times the $N(0, \tau^2)$ density, with $\epsilon \in (0, 1)$ fixed. Compute $2f(0)\sigma_0$ and show that it exceeds one once τ is large enough, so that a small amount of contamination reverses the ranking of the two tests. Explain in two or three sentences why the t -test is the one that suffers.

Bibliography

- E. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, New York, 3rd edition, 2005.
- R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, second edition, 2026.



9

The Case of the Difference in Means

The course opened by asking what $E[Y|X]$ is and how it might be learned from data. Every chapter since has supplied a piece of the answer, and none of them has produced a number. This chapter takes the simplest nontrivial parameter built from a conditional expectation and carries it the whole distance: what it is, how to estimate it, what the estimator converges to, how fast, how to estimate that, and what interval to report.

Why we cover this

A difference between two group averages is one of the most common estimands in empirical economics. Producing it correctly uses the analog principle, the continuous mapping theorem, the central limit theorem, Slutsky's theorem, an influence function and a studentization argument — six tools, drawn from four chapters, for one number.

9.1 Parameters Built from Conditional Expectations

Chapter 1 showed that $E[Y|X]$ is the best predictor of Y given X . It is not itself a number, and the parameters that get reported are features of it. Three are worth naming.

The *best linear predictor* is the closest linear approximation to $E[Y|X]$ in the mean-square sense of Section 1.5. When Y and the coordinates of X have finite second moments and $E[XX']$ is nonsingular, it is $X'\beta$ with

$$\beta = E[XX']^{-1}E[XY] , \quad (9.1)$$

a function of two expectations, estimated by the analog principle and analysed in Example 7.4. Note that, by the law of iterated expectations, $E[XY] = E[XE[Y|X]]$, which explicitly writes β as a function of $E[Y|X]$.

The *moment conditions* of the second half of the quarter are the same idea one step further out. A conditional restriction such as $E[U|X] = 0$, with $U = Y - X'\beta$, implies $E[g(X)U] = 0$ for every bounded g , an entire family of unconditional restrictions. Selecting finitely many produces a system

$E[g(W, \theta_0)] = 0$, which determines θ_0 under an identification condition, and estimators are built by making the sample analogue of that system small. Estimation of this kind is developed in the second half of the quarter.

The *difference in means* is the case this chapter follows. Let $W = (Y, X) \sim P$, observed as an i.i.d. sample $W_i = (Y_i, X_i)$, $i \leq n$, with $X \in \{0, 1\}$ and $\pi = P\{X = 1\} \in (0, 1)$, write $\mu_x = E[Y|X = x]$, and set

$$\theta(P) = \mu_1 - \mu_0 . \quad (9.2)$$

Section 1.1 showed that a conditional expectation on a binary X contains exactly two numbers, μ_0 and μ_1 , so (9.2) is one important scalar feature of $E[Y|X]$ — the contrast between the two, with the common level discarded. It has been the running example since Chapter 3.

Remark 9.1 (The causal reading) Under *random assignment*, meaning that X is statistically independent of the outcomes a unit would record under each of the two conditions, the contrast (9.2) equals an average treatment effect. That is why it is reported as often as it is, and it is the reason the warning attached to this example since Chapter 3 was worded as it was.

Making the statement precise requires vocabulary — potential outcomes, and the assumptions under which they are identified — that the spring quarter develops properly, and none of it is needed here. Everything below concerns $\mu_1 - \mu_0$, a feature of P , and holds whether or not the causal reading is available. What the assumption buys is an interpretation of the number, not a change in how it is estimated or how its standard error is computed. ■

9.2 Three Representations, One Estimator

The analog principle of Section 3.2 gives the estimator directly: replace P by $\hat{P}_n$ in (9.2), which returns the difference of the two group averages of Example 3.2,

$$\hat{\theta}_n = \hat{\mu}_{n,1} - \hat{\mu}_{n,0} , \quad \hat{\mu}_{n,x} = \frac{\sum_{i \leq n} Y_i \mathbf{1}\{X_i = x\}}{\sum_{i \leq n} \mathbf{1}\{X_i = x\}} , \quad (9.3)$$

with the convention of Example 3.2 in force: $\hat{\mu}_{n,x} = 0$ when group x is empty. Group one is empty with probability $(1 - \pi)^n$ and group zero with probability π^n , so the probability that either is empty is $\pi^n + (1 - \pi)^n$, which tends to zero. Two other estimators are written down routinely for the same parameter, and it is worth knowing at the outset that they are not alternatives.

Proposition 9.1 Write $\hat{\pi}_n = n^{-1} \sum_{i \leq n} X_i$ and suppose $0 < \hat{\pi}_n < 1$, so that both groups are non-empty. Then the following three quantities are equal.

(i) The difference in group averages $\hat{\theta}_n$ of (9.3).

(ii) The slope coefficient $\hat{b}_n$ in the least squares fit of Y_i on a constant and X_i .

(iii) The inverse-probability-weighted average

$$\hat{\theta}_n^{\text{ipw}} = \frac{1}{n} \sum_{i \leq n} \left\{ \frac{X_i Y_i}{\hat{\pi}_n} - \frac{(1 - X_i) Y_i}{1 - \hat{\pi}_n} \right\}. \quad (9.4)$$

PROOF. Throughout, write $\overline{XY} = n^{-1} \sum_{i \leq n} X_i Y_i$ and note the two identities

$$\overline{XY} = \hat{\pi}_n \hat{\mu}_{n,1}, \quad \bar{Y}_n = \hat{\pi}_n \hat{\mu}_{n,1} + (1 - \hat{\pi}_n) \hat{\mu}_{n,0}, \quad (9.5)$$

both immediate from (9.3): the first because X_i is an indicator, the second by splitting the sum over the two groups.

(iii) equals (i). Since X_i is an indicator, $n^{-1} \sum_{i \leq n} X_i Y_i / \hat{\pi}_n = \overline{XY} / \hat{\pi}_n = \hat{\mu}_{n,1}$ by the first identity in (9.5), and the same computation with $1 - X_i$ gives $\hat{\mu}_{n,0}$ for the second term. Subtracting gives $\hat{\theta}_n$.

(ii) equals (i). The least squares slope on a single regressor is the ratio of the sample covariance to the sample variance. For a binary regressor the denominator is $\hat{\pi}_n(1 - \hat{\pi}_n)$, since $X_i^2 = X_i$ gives $\overline{X^2} - \bar{X}_n^2 = \hat{\pi}_n - \hat{\pi}_n^2$. For the numerator, substitute both identities of (9.5):

$$\begin{aligned} \overline{XY} - \hat{\pi}_n \bar{Y}_n &= \hat{\pi}_n \hat{\mu}_{n,1} - \hat{\pi}_n \{ \hat{\pi}_n \hat{\mu}_{n,1} + (1 - \hat{\pi}_n) \hat{\mu}_{n,0} \} \\ &= \hat{\pi}_n (1 - \hat{\pi}_n) (\hat{\mu}_{n,1} - \hat{\mu}_{n,0}). \end{aligned}$$

Dividing by $\hat{\pi}_n(1 - \hat{\pi}_n)$, which is nonzero by hypothesis, gives $\hat{b}_n = \hat{\theta}_n$. ■

So there is one estimator here, not three, and no statistical question about which to use. What differs is the kind of generalization each representation invites once a vector of covariates Q also matters. The regression form invites modelling the *outcome*: put Q on the right-hand side and read off the coefficient on X . The weighted form invites modelling the *assignment* instead, replacing $\hat{\pi}_n$ by a fitted $P\{X = 1 \mid Q\}$, the propensity score. The difference in averages invites neither model but a *partition*: form the contrast within cells of Q and average the results. Here the three coincide because there is no Q . With one, they are different estimators resting on different assumptions, they need not agree or target the same parameter, and choosing among them — or combining them — is a subject of the spring.

The weighted form also shows where the influence function of Chapter 7 comes from, and the derivation is worth doing because the obvious shortcut is wrong. One cannot simply replace $\hat{\pi}_n$ by π in (9.4): the error is $O_P(n^{-1/2})$, which need not vanish after multiplication by $\sqrt{n}$ and can therefore affect the limit. The estimator must be centred first.

Centring is an exact identity. Since $\{n\hat{\pi}_n\}^{-1} \sum_{i \leq n} X_i(Y_i - \mu_1) = \hat{\mu}_{n,1} - \mu_1$, and similarly for group zero,

$$\hat{\theta}_n - \theta(P) = \frac{1}{n} \sum_{i \leq n} \left\{ \frac{X_i(Y_i - \mu_1)}{\hat{\pi}_n} - \frac{(1 - X_i)(Y_i - \mu_0)}{1 - \hat{\pi}_n} \right\}, \quad (9.6)$$

with no approximation anywhere. *Now* the denominators may be replaced. The first summand differs from the one with π in place of $\hat{\pi}_n$ by

$$\left(\frac{1}{\hat{\pi}_n} - \frac{1}{\pi} \right) \cdot \frac{1}{n} \sum_{i \leq n} X_i (Y_i - \mu_1) ,$$

and only one of the two factors has to be small at a rate. Assume, as Step 2 below does, that $\text{Var}[Y|X = x] < \infty$. Then $X(Y - \mu_1)$ has mean zero, since $E[X(Y - \mu_1)] = \pi E[Y - \mu_1|X = 1] = 0$, and variance $\pi\sigma_1^2 < \infty$, so the central limit theorem makes the average $O_P(n^{-1/2})$. The first factor needs no rate at all: $\hat{\pi}_n \xrightarrow{P} \pi > 0$ and $v \mapsto 1/v$ is continuous at π , so Theorem 3.1 gives $1/\hat{\pi}_n - 1/\pi = o_P(1)$. By the calculus of Lemma 4.3, the product is $o_P(n^{-1/2})$, and the same holds for group zero. So (9.6) equals the average of ψ_P below, plus a remainder that vanishes after multiplication by $\sqrt{n}$. That is the content of (7.8), and centring before substituting is the step that makes it true.

9.3 Inference

Everything needed is now in place. The argument has five steps, each quoting a different chapter.

Step 1: the estimator is consistent. Example 3.2 wrote $\hat{\mu}_{n,x}$ as a ratio of two sample averages and applied the weak law and Theorem 3.1, the ratio being continuous wherever its denominator is nonzero and the limit $P\{X = x\}$ being positive. Hence $\hat{\theta}_n \xrightarrow{P} \theta(P)$.

Step 2: the root is asymptotically linear. Suppose $\sigma_x^2 = \text{Var}[Y|X = x] \in (0, \infty)$ for $x = 0, 1$. By (7.8), $\hat{\theta}_n$ is asymptotically linear with influence function

$$\psi_P(y, x) = \frac{x(y - \mu_1)}{\pi} - \frac{(1 - x)(y - \mu_0)}{1 - \pi} , \quad (9.7)$$

which is (9.6) with π in place of $\hat{\pi}_n$, as the calculation above establishes. Theorem 7.3 then gives

$$R_n = \frac{\sqrt{n} \{\hat{\theta}_n - \theta(P)\}}{\tau(P)} \xrightarrow{d} Z , \quad \tau^2(P) = \frac{\sigma_1^2}{\pi} + \frac{\sigma_0^2}{1 - \pi} , \quad (9.8)$$

the variance being $E[\psi_P^2]$, computed in Example 5.4: the two terms of (9.7) cannot both be nonzero, because $x(1 - x) = 0$, so the cross term vanishes and the squares contribute $\pi\sigma_1^2/\pi^2$ and $(1 - \pi)\sigma_0^2/(1 - \pi)^2$.

Step 3: the variance can be estimated. The within-group variance is the plug-in estimator of σ_x^2 built from the observations in group x ,

$$\hat{\sigma}_{n,x}^2 = \frac{\sum_{i \leq n} (Y_i - \hat{\mu}_{n,x})^2 \mathbf{1}\{X_i = x\}}{\sum_{i \leq n} \mathbf{1}\{X_i = x\}} , \quad (9.9)$$

again set to zero when group x is empty. Then put

$$\hat{\tau}_n^2 = \frac{\hat{\sigma}_{n,1}^2}{\hat{\pi}_n} + \frac{\hat{\sigma}_{n,0}^2}{1 - \hat{\pi}_n}, \quad (9.10)$$

defined to be zero when either group is empty. As written, (9.9) is not a ratio of sample averages, because $\hat{\mu}_{n,x}$ appears inside the sum. Expanding the square puts it in that form,

$$\hat{\sigma}_{n,x}^2 = \frac{\sum_{i \leq n} Y_i^2 \mathbf{1}\{X_i = x\}}{\sum_{i \leq n} \mathbf{1}\{X_i = x\}} - \hat{\mu}_{n,x}^2,$$

where the first term is a ratio of the same shape as in Step 1 and is consistent for $E[Y^2|X = x]$ by Section 3.2, finite because σ_x^2 and μ_x are, and the second is consistent for μ_x^2 by Step 1 and Theorem 3.1. So $\hat{\sigma}_{n,x}^2 \xrightarrow{P} E[Y^2|X = x] - \mu_x^2 = \sigma_x^2$. Since also $\hat{\pi}_n \xrightarrow{P} \pi$ by the weak law, Theorem 3.1 applied to a function continuous wherever $0 < \pi < 1$ gives $\hat{\tau}_n^2 \xrightarrow{P} \tau^2(P)$.

Step 4: studentize. Steps 2 and 3 are exactly the hypotheses of Theorem 5.5, so

$$\tilde{R}_n = \frac{\sqrt{n} \{\hat{\theta}_n - \theta(P)\}}{\hat{\tau}_n} \xrightarrow{d} Z, \quad (9.11)$$

with that theorem's convention carried over: (9.11) is set to zero on $\{\hat{\tau}_n = 0\}$, which by Step 3 has probability tending to zero because $\tau^2(P) > 0$. This is still a root: $\theta(P)$ sits in the numerator and no sample determines it.

Step 5: a statistic, and an interval. Fixing a value for $\theta(P)$ turns (9.11) into something computable. For $H_0 : \theta(P) = \delta_0$ the statistic is

$$T_n = \frac{\sqrt{n} (\hat{\theta}_n - \delta_0)}{\hat{\tau}_n}, \quad (9.12)$$

of which Example 5.4 treated the case $\delta_0 = 0$. When the null is true, (9.12) coincides with (9.11) and therefore converges in distribution to Z , so the test rejecting at $|T_n| > z_{1-\alpha/2}$ has pointwise asymptotic level α by Example 4.4. Inverting instead of fixing gives the interval

$$C_n = \left[\hat{\theta}_n \pm z_{1-\alpha/2} \frac{\hat{\tau}_n}{\sqrt{n}} \right], \quad (9.13)$$

whose coverage tends to $1 - \alpha$ at every P satisfying the conditions above. The bridge is that on $\{\hat{\tau}_n > 0\}$, and only there, $\theta(P) \in C_n$ exactly when $|\tilde{R}_n| \leq z_{1-\alpha/2}$: the two statements say the same thing about the same quantity, once divided through by $\hat{\tau}_n/\sqrt{n}$. On $\{\hat{\tau}_n = 0\}$ the equivalence fails — the convention makes $\tilde{R}_n$ zero, so the root event always holds, while C_n has collapsed to the single point $\hat{\theta}_n$ and may well miss — but that event has probability tending to zero, so it does not affect the limit.

That is the whole argument, and the chapters it uses are worth listing: the parameter is a feature of the conditional expectation of Chapter 1; the estimator comes from the analog principle of Chapter 3 and is consistent by the continuous mapping theorem there; its influence function is Chapter 7 and its limit is the central limit theorem of Chapter 5; the variance estimate is Chapter 3 again and the studentization is Chapter 5; the conversion to a statement about probabilities is Chapter 4; and the words *pointwise asymptotic level* are Chapter 8's, and were chosen deliberately.

Remark 9.2 (When few units are treated) They were chosen deliberately because (9.8) contains a warning. With the two conditional laws of Y held fixed, the asymptotic variance $\tau^2(P)$ increases without bound as $\pi \downarrow 0$. The realized interval (9.13) need not widen with it — a treated group of one observation has estimated variance zero — and the quality of the normal approximation depends on how many observations enter the smaller group, rather than on n alone.

To say what happens, hold the two conditional laws of Y given $X = x$ fixed — so that μ_x and σ_x^2 do not move — and let only the assignment probability drift, along a sequence $\pi_n \rightarrow 0$. Then the quantity that matters is not n but the number of units in the smaller group. The count $\sum_{i \leq n} X_i$ is Binomial(n, π_n), whose variance $n\pi_n(1 - \pi_n)$ is the λ_n of (6.9), and Example 6.4 says which regime one is in. If $n\pi_n \rightarrow \infty$, the treated group is itself a growing sample and the argument above survives the drift. If $n\pi_n \rightarrow \lambda \in (0, \infty)$, the number of treated units is Poisson in the limit, it is zero with probability tending to $e^{-\lambda}$, and there is no sample from which to compute $\hat{\mu}_{n,1}$ at all. Between those two, Example 8.2 is the model for what goes wrong: coverage is correct at every fixed π and fails uniformly over π when $\mu_1 \neq 0$ (Problem 9.4), because the hard value of π moves with the sample size.

So the honest summary of (9.13) is that, with the conditional laws fixed, the normal approximation requires the smaller group to grow, and the relevant count is of order $n\pi(1 - \pi)$ rather than n — which reduces to $n\pi$ when π is the small one, and to $n(1 - \pi)$ in the mirror case. That is the same sentence the coin has been repeating since Chapter 6, now said about the object this course set out to study. ■

Where we will use this

- **In this class.** This chapter uses Chapters 1, 3, 4, 5, 7 and 8 and introduces nothing. The second half of the quarter replaces the difference in means by an extremum estimator and repeats the same five steps in that generality, with the influence function of Chapter 7 supplied by the first-order condition rather than by direct calculation.
- **Later in the sequence.** The winter quarter takes up the regression

representation of Proposition 9.1 when the covariates are numerous, and the spring takes up the weighted representation when assignment depends on covariates, where the three estimators stop coinciding and the choice among them becomes the subject. The causal reading of Remark 9.1 is developed there.

- **In recent research.** Inference when the smaller group is small relative to the sample — rare treatments, rare outcomes, few clusters — is an active area, and Remark 9.2 illustrates the common concern: the nominal sample size need not measure the information available for inference.

9.4 Key Concepts

- **Conditional expectations generate the parameters.** The best linear predictor and the difference in means are functions of $E[Y|X]$, and the moment restrictions of the second half arise from a conditional restriction on it. All are estimated by the analog principle.
 - **Three representations, one estimator.** The difference in group averages, the least squares slope on a binary regressor, and the inverse-probability-weighted average are algebraically identical. They differ only in what they suggest doing when covariates are added.
 - **The influence function does the work.** Once ψ_P is known, the limiting distribution and the asymptotic variance $E[\psi_P^2] = \sigma_1^2/\pi + \sigma_0^2/(1-\pi)$ follow without further calculation.
 - **Root, then statistic.** Studentizing removes the unknown scale but leaves a root. Fixing the parameter by a hypothesis, or inverting over it, is what produces something a sample determines.
 - **The effective sample size.** With the conditional laws fixed, the normal approximation requires the smaller group to grow: the relevant count is of order $n\pi(1-\pi)$, not n . That quantity tracks the expected size of the *smaller* group, to within a factor of at most two, and it is Chapter 6's λ_n again. Coverage holds at every fixed π and fails uniformly when $\mu_1 \neq 0$, for the reason Chapter 8 gives.
-

9.5 Concluding Remarks

There is almost nothing to cite for this chapter, which is the point of it: every result used has been proved earlier in Canay [2026]. Readers who want the causal interpretation of Remark 9.1 developed properly should turn to Canay [2027], where potential outcomes are defined and the assumptions under which (9.2) identifies a treatment effect are stated and examined. Those who want the extremum-estimator version of Section 9.3 will meet it in the second half of this quarter, and will find that the five steps are the same ones.

The course began with a projection and a question about how to learn it from data. It ends with a number computed from a sample, an interval around it, and a reasonably precise account of when that interval means what it appears to mean. The rest of the sequence is this argument made to work in harder circumstances.

9.6 Problems

Problem 9.1 *Verify the two identities (9.5) directly from the definitions, and then complete the following.*

1. *Show that the intercept in the regression of part (ii) of Proposition 9.1 equals $\hat{\mu}_{n,0}$, and interpret the two coefficients.*
2. *Show that the residual from that regression equals $Y_i - \hat{\mu}_{n,X_i}$, and deduce that the residual sum of squares is the within-group sum of squares.*
3. *Explain what goes wrong in each of the three representations when $\hat{\pi}_n = 0$, and why the convention adopted in (9.10) is harmless.*

Problem 9.2 *Let ψ_P be as in (9.7).*

1. *Verify that $E[\psi_P(Y, X)] = 0$, saying where the law of iterated expectations is used.*
2. *Compute $E[\psi_P^2]$ and confirm (9.8).*
3. *Write down the influence functions of the two group-mean estimators $\hat{\mu}_{n,1}$ and $\hat{\mu}_{n,0}$, each viewed as an estimator based on the full sample. Show that their difference is ψ_P , and explain why both must be evaluated on the same observation.*

Problem 9.3 *Suppose $\sigma_1^2 \neq \sigma_0^2$.*

1. Show that the “pooled” variance estimator, which replaces both $\hat{\sigma}_{n,x}^2$ in (9.10) by the sample variance of all n residuals from the regression in Proposition 9.1(ii), is not consistent for $\tau^2(P)$ unless $\pi = 1/2$ or $\sigma_1^2 = \sigma_0^2$.
2. Using Problem 5.5(c), find the limiting distribution of the studentized root when the pooled estimator is used, and say when a test based on it over-rejects.
3. Which representation in Proposition 9.1 makes this failure easiest to overlook, and why?

Problem 9.4 Consider a sequence with $\pi_n \rightarrow 0$ and $n\pi_n \rightarrow \infty$.

1. Show that $\sqrt{n\pi_n}(\hat{\mu}_{n,1} - \mu_1)$ has a normal limit under conditions you should state, and identify its variance.
2. Deduce the rate at which the length of the interval (9.13) shrinks, and compare it with $n^{-1/2}$.
3. Suppose instead $n\pi_n \rightarrow \lambda \in (0, \infty)$, with the conditional laws of Y given $X = x$ held fixed. Show that $P\{\hat{\pi}_n = 0\} \rightarrow e^{-\lambda}$. Then show that on the event $\{\hat{\pi}_n = 0\}$ the interval (9.13), under the conventions of (9.3) and (9.10), collapses to the single point $-\hat{\mu}_{n,0}$. Show that if $\mu_1 \neq 0$ then the conditional coverage $P\{\theta(P_n) \in C_n \mid \hat{\pi}_n = 0\}$ tends to zero, and say where you used the fact that $\hat{\mu}_{n,0}$ is consistent for μ_0 on that event. Conclude, in two or three sentences, what this implies about uniform coverage over $\pi \in (0, 1)$.

Bibliography

- I. A. Canay. Large sample theory. Lecture notes, Econ 480-1, Northwestern University, 2026.
- I. A. Canay. Identification and inference in modern econometrics. Lecture notes, Econ 480-3, Northwestern University, 2027.



Part IV

Appendix



A

Background Results

This appendix collects results that the course uses but does not cover: standard facts from integration theory, an existence theorem taken on faith, inequalities that are stated in the chapters and used there, and proofs that belong on record rather than on the board.

A.1 Facts from Integration Theory

These three facts are used throughout, first in Chapter 1 and repeatedly thereafter. They are restated as Hint 1.1 at the point of first use.

Fact A.1 (Simple approximation) *Every bounded Borel measurable $g : \mathbf{R}^k \rightarrow \mathbf{R}$ is the pointwise limit of simple functions $g_n = \sum_{j \leq J_n} c_{n,j} \mathbf{1}_{\{\cdot \in B_{n,j}\}}$, with $B_{n,j}$ Borel and $\sup_n \sup_x |g_n(x)| \leq \sup_x |g(x)|$.*

Fact A.2 (Dominated convergence) *If $V_n \rightarrow V$ almost surely and $|V_n| \leq W$ almost surely for some integrable W , then V is integrable and $E[V_n] \rightarrow E[V]$.*

Fact A.3 (Monotone convergence) *If $0 \leq V_n \uparrow V$ almost surely, then $E[V_n] \uparrow E[V]$.*

Cf. Billingsley [1995, §13, p. 182] for measurability and simple approximation; his Theorem 16.2, p. 208, for monotone convergence, and Theorem 16.4, p. 209, for dominated convergence.

A.2 Radon–Nikodym and the Existence of Conditional Expectation

Chapter 1 defines $E[Y|X]$ by a functional equation and asserts, without proof, that a solution exists. The assertion is a corollary of one theorem from measure theory, and the shape of the argument is worth seeing even if the proof is not.

Let P and Q be measures on a measurable space $(\Omega, \mathcal{A})$. Then Q is *absolutely continuous* with respect to P , written $Q \ll P$, if $P\{A\} = 0$ implies $Q\{A\} = 0$ for every $A \in \mathcal{A}$: whatever P regards as negligible, Q regards as negligible too.

Fact A.4 (Radon–Nikodym) *Let P and Q be σ -finite measures on $(\Omega, \mathcal{A})$ with $Q \ll P$. Then there is a measurable $L : \Omega \rightarrow [0, \infty)$, unique up to P -null sets, with*

$$Q\{A\} = \int_A L dP \quad \text{for every } A \in \mathcal{A} .$$

The function $L \equiv dQ/dP$ is called the Radon–Nikodym derivative, or density, of Q with respect to P .

What the theorem buys is a change of category: a hypothesis about *sets* is converted into the existence of a *function*. Once Q is known to be small wherever P is, all of Q can be recovered from P by reweighting point by point, and the weight is a single measurable function. The familiar cases are special ones. Taking P to be Lebesgue measure makes L an ordinary probability density; taking P and Q to be two candidate distributions for the same data makes L the likelihood ratio.

Existence of the conditional expectation is this theorem applied to a well-chosen pair. Suppose $E[|Y|] < \infty$, let P_X denote the distribution of X , and define, for Borel B ,

$$\nu(B) = E[Y \mathbf{1}\{X \in B\}] .$$

Then ν is a signed measure on the Borel sets — countably additive by dominated convergence, and finite because Y is integrable — and it vanishes wherever P_X does, since $P_X\{B\} = 0$ forces $Y \mathbf{1}\{X \in B\} = 0$ almost surely. So $\nu \ll P_X$, and Fact A.4, applied to the positive and negative parts of ν separately, supplies a Borel measurable m with

$$\nu(B) = \int_B m dP_X = E[m(X) \mathbf{1}\{X \in B\}] \quad \text{for every Borel } B ,$$

which rearranges to $E[(Y - m(X)) \mathbf{1}\{X \in B\}] = 0$. In other words, the conditional expectation *is* the Radon–Nikodym derivative of the measure $B \mapsto E[Y \mathbf{1}\{X \in B\}]$ with respect to the distribution of X , and the defining equation of Chapter 1 says nothing more than that. The a.s. uniqueness that Chapter 1 needs is proved there directly, at Lemma 1.1, so only the existence half of the theorem is taken on faith.

Cf. Billingsley [1995, §32, p. 419] for the Radon–Nikodym theorem, and his §34, p. 445, for this construction of the conditional expectation; Lehmann and Romano [2005, §2.4, p. 37]. The absolute-continuity notation and the likelihood-ratio reading follow the treatment of contiguity in Ivan’s advanced notes, where the same derivative reappears as the object whose law under one measure determines expectations under the other.

A.3 Jensen's Inequality

Cited from Chapter 2 onward. Its main use in this course is the *information inequality*: for any density f other than the true f_0 , applying Fact A.5 to the logarithm gives

$$\mathbb{E}_0[\log\{f(W)/f_0(W)\}] \leq \log \mathbb{E}_0[f(W)/f_0(W)] \leq 0 ,$$

so the population log-likelihood is largest at the truth. The second inequality holds because $\mathbb{E}_0[f(W)/f_0(W)]$ integrates f over the set where f_0 is positive, and equals one exactly when $f_0 > 0$ wherever f is. This is the identification condition behind the consistency of maximum likelihood, established in the second half of the quarter.

Fact A.5 (Jensen) *Let $\varphi : \mathbf{R} \rightarrow \mathbf{R}$ be convex and let V be a random variable with $E[|V|] < \infty$ and $E[|\varphi(V)|] < \infty$. Then $\varphi(E[V]) \leq E[\varphi(V)]$.*

PROOF. Write $c = E[V]$. A convex function lies above each of its supporting lines: there exists $s \in \mathbf{R}$ with

$$\varphi(v) \geq \varphi(c) + s(v - c) \quad \text{for every } v \in \mathbf{R} ,$$

where s may be taken to be any subgradient of φ at c , for instance $s = \varphi'(c)$ when φ is differentiable. Substituting V for v and taking expectations, the right-hand side has expectation $\varphi(c) + s(E[V] - c) = \varphi(c)$, which gives the claim. ■

Cf. Jiang [2010, eq. (5.56), p. 146], with the integral form at his (5.57) and an application to the log-likelihood at his Example 5.10. The supporting-line argument is the standard one.

The likelihood argument needs slightly more than Fact A.5: it needs to know when the inequality is strict, since that is what turns “the population log-likelihood is maximized at the truth” into “it is *uniquely* maximized at the truth”. Equality is possible only in the degenerate case.

Fact A.6 (Strict Jensen) *Let $\varphi : I \rightarrow \mathbf{R}$ be strictly convex on an interval I and let V be a random variable taking values in I with $E[|V|] < \infty$ and $E[|\varphi(V)|] < \infty$. If $\varphi(E[V]) = E[\varphi(V)]$, then $V = E[V]$ almost surely.*

PROOF. Write $c = E[V]$, which lies in I , and let s be a subgradient of φ at c , so that the function

$$g(v) = \varphi(v) - \varphi(c) - s(v - c)$$

is non-negative on I , as in the proof of Fact A.5. Strict convexity makes the inequality strict away from the point of contact: $g(v) > 0$ for every $v \in I$ with

$v \neq c$. Now $E[g(V)] = E[\varphi(V)] - \varphi(c) = 0$ by hypothesis, and a non-negative random variable with zero expectation vanishes almost surely, so $g(V) = 0$ almost surely and therefore $V = c$ almost surely. ■

Cf. The argument is the supporting-line proof of Fact A.5 read one step further. The strict version is not stated in any book of our reference set: Jiang [2010, eq. (5.56), p. 146] gives the inequality without a condition for equality, and his Example 5.10, p. 146, gives the information inequality in its non-strict form. Jiang [2010, §1.4, p. 9] does use the strict form, in the Cramér argument, without isolating it. The statement and proof above are therefore ours.

A.4 The Cauchy–Schwarz Inequality

Cited from Chapter 2 onward, and used twice in Chapter 1, at Theorem 1.1 and Lemma 1.4. It is what makes a covariance finite whenever two second moments are.

Fact A.7 (Cauchy–Schwarz) *If $E[V^2] < \infty$ and $E[U^2] < \infty$ then $E[|VU|] < \infty$ and*

$$(E[VU])^2 \leq E[V^2] E[U^2] .$$

PROOF. First, VU is integrable: expanding $(|V| - |U|)^2 \geq 0$ gives the pointwise bound

$$|VU| \leq \frac{1}{2}V^2 + \frac{1}{2}U^2 ,$$

whose right-hand side has finite expectation by hypothesis. In particular $E[VU]$ is well defined and finite, and every term in the expansion below is finite.

If $E[U^2] = 0$ then $U = 0$ a.s. and both sides of the claim vanish, so assume $E[U^2] > 0$. For every $\alpha \in \mathbf{R}$,

$$0 \leq E[(V - \alpha U)^2] = E[V^2] - 2\alpha E[VU] + \alpha^2 E[U^2] .$$

The right-hand side is a quadratic in α that is never negative, so its discriminant is not positive: $4(E[VU])^2 - 4E[V^2]E[U^2] \leq 0$. ■

Cf. Jiang [2010, Example 5.3, p. 130] gives the numerical version, and recovers it at p. 132 as a special case of Hölder’s inequality.

A.5 Hoeffding's Lemma

Fact A.8 (Hoeffding's lemma) *Let V be a random variable with $E[V] = 0$ and $m \leq V \leq M$ almost surely. Then for every $\lambda \in \mathbf{R}$,*

$$E[e^{\lambda V}] \leq \exp\left\{\frac{1}{8}\lambda^2(M-m)^2\right\}. \quad (\text{A.1})$$

PROOF. Since $E[V] = 0$ and $m \leq V \leq M$, we have $m \leq 0 \leq M$, and we may assume $m < M$, the case $m = M$ being trivial. The function $v \mapsto e^{\lambda v}$ is convex, so on the interval $[m, M]$ it lies below the chord joining its endpoints:

$$e^{\lambda v} \leq \frac{M-v}{M-m} e^{\lambda m} + \frac{v-m}{M-m} e^{\lambda M} \quad \text{for } v \in [m, M].$$

Substituting V and taking expectations, the terms in V drop out because $E[V] = 0$:

$$E[e^{\lambda V}] \leq \frac{M}{M-m} e^{\lambda m} - \frac{m}{M-m} e^{\lambda M}.$$

Write $p = -m/(M-m) \in [0, 1]$ and $h = \lambda(M-m)$, so that $\lambda m = -ph$. The right-hand side is then $(1-p)e^{\lambda m} + pe^{\lambda M} = e^{L(h)}$ with

$$L(h) = -ph + \log(1 - p + pe^h).$$

Direct differentiation gives $L(0) = 0$, $L'(0) = 0$, and

$$L''(h) = \frac{(1-p)pe^h}{(1-p+pe^h)^2} = \theta(1-\theta) \leq \frac{1}{4},$$

where $\theta = pe^h/(1-p+pe^h) \in (0, 1)$. A second-order Taylor expansion of L about zero therefore gives $L(h) \leq h^2/8$, which is (A.1). ■

Cf. The chord-plus-Taylor argument is Hoeffding's own. Vershynin [2026, Theorem 2.2.1, p. 28] proves the Rademacher case by the same route, and his Exercise 2.10 sets the general statement as an exercise.

A.6 The Continuous Mapping Theorem

Theorem 3.1 is stated in Chapter 3 and used from there on. For almost sure convergence it follows directly from continuity, as noted in that chapter; the work is in the version for convergence in probability, and the route is to reduce it to the almost sure case one subsequence at a time.

Fact A.9 (Subsequence criterion) *Let V_n and V be random vectors on a common probability space. Then $V_n \xrightarrow{P} V$ if and only if every subsequence of $\{V_n\}$ has a further subsequence converging to V almost surely.*

PROOF. Suppose $V_n \xrightarrow{P} V$ and let $\{V_{n_j}\}$ be a subsequence, which also converges in probability. Choose indices $n_{j_1} < n_{j_2} < \dots$ with

$$P\{\|V_{n_{j_k}} - V\| > 2^{-k}\} \leq 2^{-k} ,$$

which is possible because for each fixed tolerance the probability tends to zero. These bounds are summable, so Borel–Cantelli, Lemma 2.3, says that with probability one only finitely many of the events $\{\|V_{n_{j_k}} - V\| > 2^{-k}\}$ occur; on that event $\|V_{n_{j_k}} - V\| \leq 2^{-k}$ eventually, hence tends to zero.

Conversely, suppose the subsequence property holds and V_n does *not* converge in probability. Then there are $\varepsilon, \eta > 0$ and a subsequence along which $P\{\|V_{n_j} - V\| > \varepsilon\} \geq \eta$ for every j . That subsequence has a further subsequence converging almost surely, hence in probability by Lemma 2.2, which contradicts the bound. ■

Cf. Jiang [2010, Theorem 2.3, p. 23]. The forward half is the standard Borel–Cantelli argument and uses only Lemma 2.3.

PROOF OF THEOREM 3.1. Measurability of g makes $g(V_n)$ and $g(V)$ random vectors, so the statement is well posed, and C_g is Borel because it is the intersection over $m \geq 1$ of the open sets on which the oscillation of g is below $1/m$.

Take any subsequence of $\{g(V_n)\}$, indexed by n_j . Since $V_{n_j} \xrightarrow{P} V$, Fact A.9 supplies a further subsequence n_{j_k} along which $V_{n_{j_k}} \rightarrow V$ almost surely. Let A be the event on which that convergence holds, and $B = \{V \in C_g\}$; both have probability one, so does $A \cap B$. Fix an outcome in $A \cap B$. Then $V_{n_{j_k}} \rightarrow V$ and g is continuous at the point V , so

$$g(V_{n_{j_k}}) \longrightarrow g(V) .$$

This holds on an event of probability one, so $g(V_{n_{j_k}}) \rightarrow g(V)$ almost surely.

Every subsequence of $\{g(V_n)\}$ therefore has a further subsequence converging to $g(V)$ almost surely, and Fact A.9, applied in the other direction, gives $g(V_n) \xrightarrow{P} g(V)$. ■

Cf. van der Vaart [1998, Theorem 2.3, p. 7]; Jiang [2010, Theorem 2.12, p. 30]. The subsequence route is what allows the theorem to be stated for the full continuity set rather than for an open subset of it.

A.7 The Portmanteau Theorem

Theorem 4.1 is stated in Chapter 4 and every other result of that chapter is derived from it. Write F_n and F for the distribution functions of V_n and V , and $\bar{A}$, A° and ∂A for the closure, interior and boundary of a set A .

One observation is used twice below. For any family of pairwise disjoint Borel sets, at most countably many can have positive probability under P . In particular, for each coordinate j the hyperplanes $\{v : v_j = s\}$, $s \in \mathbf{R}$, are disjoint, so all but countably many of them are null for V ; a rectangle whose faces avoid those countably many values has $P\{V \in \partial B\} = 0$, and such rectangles can be chosen with corners arbitrarily close to any prescribed points.

PROOF OF THEOREM 4.1. We prove (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (v) $\Leftrightarrow$ (vi) $\Rightarrow$ (vii) $\Rightarrow$ (i), and then fold in (iv).

(i) $\Rightarrow$ (ii). Let f be continuous with $|f| \leq B$, and fix $\varepsilon > 0$. Choose a rectangle $R = (a_1, b_1] \times \cdots \times (a_k, b_k]$ whose faces are null for V and with $P\{V \in R\} > 1 - \varepsilon$. By (i) and inclusion-exclusion, $P\{V_n \in Q\} \rightarrow P\{V \in Q\}$ for every rectangle Q whose corners are continuity points of F , since $P\{V_n \in Q\}$ is a fixed finite signed combination of values of F_n at those corners; in particular $P\{V_n \in R\} \rightarrow P\{V \in R\}$.

The function f is uniformly continuous on the compact set $\bar{R}$, so there is $\rho > 0$ with $|f(u) - f(v)| < \varepsilon$ whenever $u, v \in \bar{R}$ and $\|u - v\| < \rho$. Partition R into finitely many disjoint rectangles $Q_1, \dots, Q_J$ of diameter less than ρ , each with faces null for V , and pick $v_j \in Q_j$. The step function $h = \sum_{j \leq J} f(v_j) \mathbf{1}\{\cdot \in Q_j\}$ satisfies $|f - h| \leq \varepsilon$ on R and $|h| \leq B$, so

$$|E[f(V_n)] - E[h(V_n)]| \leq \varepsilon + 2B P\{V_n \notin R\},$$

and the same bound holds with V in place of V_n . Since $E[h(V_n)] = \sum_{j \leq J} f(v_j) P\{V_n \in Q_j\} \rightarrow E[h(V)]$, a finite sum of convergent terms,

$$\begin{aligned} \limsup_n |E[f(V_n)] - E[f(V)]| &\leq 2\varepsilon + 2B \left(\limsup_n P\{V_n \notin R\} + P\{V \notin R\} \right) \\ &\leq 2\varepsilon + 4B\varepsilon, \end{aligned}$$

using $P\{V_n \notin R\} \rightarrow P\{V \notin R\} < \varepsilon$. As $\varepsilon > 0$ was arbitrary, (ii) follows.

(ii) $\Rightarrow$ (iii). A bounded Lipschitz function is bounded and continuous.

(iii) $\Rightarrow$ (v). Let G be open. If $G = \mathbf{R}^k$ both sides equal one, so assume otherwise and write $d(v, G^c) = \inf\{\|v - u\| : u \notin G\}$, which is Lipschitz with constant one. For $m \geq 1$ set

$$f_m(v) = \min\{1, m d(v, G^c)\},$$

a bounded Lipschitz function with values in $[0, 1]$. Each f_m vanishes off G , so

$f_m \leq \mathbf{1}\{\cdot \in G\}$, and $f_m \uparrow \mathbf{1}\{\cdot \in G\}$ pointwise, since $d(v, G^c) > 0$ for every $v \in G$ by openness. Hence for each fixed m ,

$$\liminf_n P\{V_n \in G\} \geq \liminf_n E[f_m(V_n)] = E[f_m(V)] ,$$

and letting $m \rightarrow \infty$ on the right, monotone convergence, Fact A.3, gives $E[f_m(V)] \uparrow P\{V \in G\}$.

(v) $\Leftrightarrow$ (vi). Apply one to the complement of the set in the other, using $P\{V_n \in K\} = 1 - P\{V_n \in K^c\}$ and $\liminf_n (-x_n) = -\limsup_n x_n$.

(v) and (vi) $\Rightarrow$ (vii). Let B satisfy $P\{V \in \partial B\} = 0$. Then B° is open and $\bar{B}$ is closed, and $P\{V \in B^\circ\} = P\{V \in \bar{B}\} = P\{V \in B\}$, since the two differ from B only inside ∂B . Therefore

$$\begin{aligned} P\{V \in B\} &= P\{V \in B^\circ\} \leq \liminf_n P\{V_n \in B\} \\ &\leq \limsup_n P\{V_n \in B\} \\ &\leq P\{V \in \bar{B}\} = P\{V \in B\} , \end{aligned}$$

where the outer inequalities use $B^\circ \subseteq B \subseteq \bar{B}$ together with (v) and (vi). The two ends agree, so the limit exists and equals $P\{V \in B\}$.

(vii) $\Rightarrow$ (i). Let v be a continuity point of F and take $B = (-\infty, v_1] \times \cdots \times (-\infty, v_k]$. Its boundary is contained in the union of the k hyperplanes $\{u : u_j = v_j, u \leq v \text{ in the remaining coordinates}\}$, and continuity of F at v forces $P\{V \in \partial B\} = 0$. Then (vii) reads $F_n(v) \rightarrow F(v)$.

(iv). That (ii) implies (iv) follows by truncation: for continuous $f \geq 0$ and $M > 0$ the function $f \wedge M$ is bounded and continuous, so $\liminf_n E[f(V_n)] \geq \lim_n E[(f \wedge M)(V_n)] = E[(f \wedge M)(V)]$, and letting $M \uparrow \infty$ gives $E[f(V)]$ by monotone convergence. Conversely (iv) implies (v) by the same approximation used in (iii) $\Rightarrow$ (v), since each f_m there is nonnegative and continuous. So (iv) sits inside the cycle. ■

Cf. van der Vaart [1998, Lemma 2.2, p. 6] states all seven parts; Billingsley [1995, Theorem 25.8, p. 335] proves a three-part version through the Skorokhod representation of his Theorem 25.6.

A.8 Characteristic Function Estimates

Two ingredients of the central limit theorem, Theorem 5.3, are collected here. The first is the limit that Step 4 of that proof invokes; the second is Lemma 5.2, which is Step 3 and is the only part of the argument that is not a short calculation.

Fact A.10 (Exponential limit) *Let z_n be complex numbers with $z_n \rightarrow z$. Then $(1 + z_n/n)^n \rightarrow e^z$.*

PROOF. Choose $M < \infty$ with $|z_n| \leq M$ for every n , which is possible because z_n converges. Two elementary bounds are needed. For complex a, b ,

$$|a^n - b^n| = \left| (a - b) \sum_{j < n} a^j b^{n-1-j} \right| \leq n \max\{|a|, |b|\}^{n-1} |a - b|, \quad (\text{A.2})$$

since the sum has n terms and each has modulus at most $\max\{|a|, |b|\}^{n-1}$. And from the exponential series, for complex w ,

$$|e^w - 1 - w| = \left| \sum_{k \geq 2} \frac{w^k}{k!} \right| \leq \frac{|w|^2}{2} e^{|w|}. \quad (\text{A.3})$$

Apply (A.2) with $a = 1 + z_n/n$ and $b = e^{z_n/n}$. Both have modulus at most $e^{M/n}$, using $|1 + z_n/n| \leq 1 + M/n \leq e^{M/n}$, so the maximum raised to the power $n - 1$ is at most e^M . By (A.3) with $w = z_n/n$, the factor $|a - b|$ is at most $M^2 e^M / (2n^2)$. Hence

$$|(1 + z_n/n)^n - e^{z_n}| \leq n \cdot e^M \cdot \frac{M^2 e^M}{2n^2} = \frac{M^2 e^{2M}}{2n} \rightarrow 0,$$

and $e^{z_n} \rightarrow e^z$ by continuity of the exponential. Combining the two gives the claim. ■

The expansion itself rests on a bound for the error in truncating e^{ix} , in a form that carries two different estimates at once.

Fact A.11 (Truncation bound) *For every real x ,*

$$\left| e^{ix} - 1 - ix + \frac{x^2}{2} \right| \leq \min\left\{ \frac{|x|^3}{6}, x^2 \right\}. \quad (\text{A.4})$$

PROOF. Taylor's theorem with integral remainder, applied to $f(x) = e^{ix}$ with $f^{(k)}(y) = i^k e^{iy}$, gives for each $m \geq 0$

$$e^{ix} = \sum_{k \leq m} \frac{(ix)^k}{k!} + \frac{i^{m+1}}{m!} \int_0^x (x - y)^m e^{iy} dy.$$

Taking $m = 2$ and bounding the integrand by its modulus, which is $(|x| - u)^2$ after the substitution $y = \pm u$, gives

$$\left| e^{ix} - 1 - ix + \frac{x^2}{2} \right| \leq \frac{1}{2} \int_0^{|x|} (|x| - u)^2 du = \frac{|x|^3}{6},$$

which is the first bound. Taking $m = 1$ instead gives $|e^{ix} - 1 - ix| \leq$

$\int_0^{|x|} (|x| - u) du = x^2/2$, and adding the term $x^2/2$ by the triangle inequality gives the second. ■

Both halves of (A.4) are needed. The cubic bound is the sharp one for small $|x|$ and is what produces the rate; the quadratic bound is what keeps the estimate integrable when X has no third moment, which is the case Lemma 5.2 is stated for.

PROOF OF LEMMA 5.2. Let X satisfy $E[X] = 0$ and $E[X^2] = 1$, and define $r(s) = \varphi_X(s) - 1 + s^2/2$.

The remainder is an expectation of the truncation error. Take expectations of $e^{isX} - 1 - isX + s^2X^2/2$ term by term, which linearity permits since each term is integrable. The first contributes $\varphi_X(s)$, the term -1 contributes -1 , the term $-isX$ contributes $-isE[X] = 0$, and the last contributes $s^2E[X^2]/2 = s^2/2$. Hence

$$r(s) = E\left[e^{isX} - 1 - isX + \frac{s^2X^2}{2}\right],$$

so the two moment conditions are used here and nowhere else.

Bounding it. Apply Fact A.11 with $x = sX$, then move the modulus inside the expectation and factor out s^2 . Writing $g_s(x) = \min\{|s||x|^3/6, x^2\}$,

$$\frac{|r(s)|}{s^2} \leq \frac{1}{s^2} E\left[\min\left\{\frac{|sX|^3}{6}, (sX)^2\right\}\right] = E[g_s(X)].$$

Letting $s \rightarrow 0$. For each fixed x , the first entry of the minimum tends to zero as $s \rightarrow 0$, so $g_s(x) \rightarrow 0$. And $g_s \leq x^2$ for every s , by the second entry, with $E[X^2] = 1 < \infty$. Dominated convergence, Fact A.2, applied along any sequence $s_m \rightarrow 0$, gives $E[g_{s_m}(X)] \rightarrow 0$; as the sequence was arbitrary, $r(s)/s^2 \rightarrow 0$. ■

Cf. Billingsley [1995, Theorem 26.1, p. 342, and §27] for the truncation bound and this expansion; van der Vaart [1998, Proposition 2.17, p. 16] uses the same estimate inside the proof of the central limit theorem. The dominating function x^2 is what allows a finite second moment to suffice.

A.9 The Lindeberg–Feller Theorem

Theorem 6.2 is stated in Chapter 6. The proof below is the four-step argument of Theorem 5.3 carried out one factor at a time, and it reuses the truncation bound of Fact A.11 at the step where Lindeberg’s condition enters. One elementary inequality is needed first.

Fact A.12 (Product comparison) *Let $a_1, \dots, a_k$ and $b_1, \dots, b_k$ be complex numbers with $|a_i| \leq 1$ and $|b_i| \leq 1$ for every i . Then*

$$\left| \prod_{i \leq k} a_i - \prod_{i \leq k} b_i \right| \leq \sum_{i \leq k} |a_i - b_i|. \quad (\text{A.5})$$

PROOF. Telescope, replacing the a 's by b 's one at a time:

$$\prod_{i \leq k} a_i - \prod_{i \leq k} b_i = \sum_{i \leq k} \left(\prod_{j < i} a_j \right) (a_i - b_i) \left(\prod_{j > i} b_j \right),$$

since consecutive terms of the sum cancel in pairs. Each of the two products has modulus at most one, so the modulus of the i -th term is at most $|a_i - b_i|$, and the triangle inequality gives (A.5). ■

PROOF OF THEOREM 6.2. Write $\varphi_{n,i}$ for the characteristic function of $X_{n,i}$ and fix $t \in \mathbf{R}$. At $t = 0$ every quantity below equals one, so assume $t \neq 0$. By independence within the row and Lemma 5.1(iv),

$$\varphi_{S_n}(t) = \prod_{i \leq k_n} \varphi_{n,i}(t). \quad (\text{A.6})$$

The target is $\varphi_{S_n}(t) \rightarrow e^{-t^2/2}$, and the route is to compare the product (A.6) with $\prod_i (1 - t^2 \sigma_{n,i}^2/2)$ and then that product with $e^{-t^2/2}$.

Step 1: each factor, term by term. Apply Fact A.11 with $x = tX_{n,i}$ and take expectations, using $E[X_{n,i}] = 0$ and $E[X_{n,i}^2] = \sigma_{n,i}^2$ exactly as in the proof of Lemma 5.2:

$$\left| \varphi_{n,i}(t) - \left(1 - \frac{t^2 \sigma_{n,i}^2}{2} \right) \right| \leq E \left[\min \left\{ \frac{|tX_{n,i}|^3}{6}, (tX_{n,i})^2 \right\} \right].$$

Now split the expectation at the level ε , using the cubic bound where $|X_{n,i}| \leq \varepsilon$ and the quadratic bound where it is not:

$$E \left[\min \{ \cdot \} \right] \leq \frac{|t|^3 \varepsilon}{6} \sigma_{n,i}^2 + t^2 E \left[X_{n,i}^2 \mathbf{1} \{ |X_{n,i}| > \varepsilon \} \right].$$

Summing over the row and using $\sum_i \sigma_{n,i}^2 = 1$ together with Definition 6.2,

$$\sum_{i \leq k_n} \left| \varphi_{n,i}(t) - \left(1 - \frac{t^2 \sigma_{n,i}^2}{2} \right) \right| \leq \frac{|t|^3 \varepsilon}{6} + t^2 L_n(\varepsilon). \quad (\text{A.7})$$

Lindeberg's condition sends $L_n(\varepsilon) \rightarrow 0$ for fixed ε , so the limit superior of the left-hand side is at most $|t|^3 \varepsilon/6$; as $\varepsilon > 0$ was arbitrary, the sum tends to zero. This is the only place the hypothesis is used, and it is where the identical distribution was used in Chapter 5.

Step 2: from the factors to the products. Both families are bounded by

one in modulus: $|\varphi_{n,i}(t)| \leq 1$ by Lemma 5.1(i), and $0 \leq 1 - t^2\sigma_{n,i}^2/2 \leq 1$ for all large n , because Lemma 6.1 makes $\max_i \sigma_{n,i}^2$ smaller than $2/t^2$ eventually. Fact A.12 and (A.7) therefore give

$$\left| \varphi_{S_n}(t) - \prod_{i \leq k_n} \left(1 - \frac{t^2 \sigma_{n,i}^2}{2} \right) \right| \longrightarrow 0 .$$

Step 3: the second product. Write $w_{n,i} = t^2\sigma_{n,i}^2/2$, so that $\sum_i w_{n,i} = t^2/2$ and $\max_i w_{n,i} \rightarrow 0$. For $0 \leq w \leq 1/2$ the logarithm satisfies $|\log(1-w) + w| \leq w^2$, so summing over the row,

$$\left| \sum_{i \leq k_n} \log(1-w_{n,i}) + \frac{t^2}{2} \right| \leq \sum_{i \leq k_n} w_{n,i}^2 \leq \left(\max_{i \leq k_n} w_{n,i} \right) \sum_{i \leq k_n} w_{n,i} = \frac{t^2}{2} \max_{i \leq k_n} w_{n,i} ,$$

which tends to zero. Exponentiating, $\prod_i (1 - w_{n,i}) \rightarrow e^{-t^2/2}$.

Step 4: conclude. Combining Steps 2 and 3, $\varphi_{S_n}(t) \rightarrow e^{-t^2/2}$ for every t . The limit is the characteristic function of Z by (5.1) and is continuous at 0, so Theorem 5.2(ii) gives $S_n \xrightarrow{d} Z$. ■

Cf. Billingsley [1995, Theorem 27.2, p. 359]; Jiang [2010, Theorem 6.11, p. 182]; van der Vaart [1998, Proposition 2.27, p. 20].

Bibliography

- P. Billingsley. *Probability and Measure*. Wiley-Interscience, 1995.
- J. Jiang. *Large Sample Techniques for Statistics*. Springer Texts in Statistics. Springer, New York, 2010.
- E. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, New York, 3rd edition, 2005.
- A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.
- R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, Cambridge, second edition, 2026.