Convergence and limit theorems

Robert J. Carroll

Last revised: 5 May 2026

---

1 Motivation

The questions "does the polling estimator converge to the true population vote share as the sample size grows?" and "what does its sampling distribution look like for moderate sample sizes?" are the workhorse questions of empirical political-science inference. The first is a question about consistency, answered by the law of large numbers: the sample mean converges to the population mean in a precise probabilistic sense. The second is a question about the shape of fluctuations around the limit, answered by the central limit theorem: appropriately rescaled sample averages are approximately normally distributed, and this is what underwrites the standard $\pm 1.96\,\sigma/\sqrt{n}$ confidence intervals reported in every survey.

This handout closes the probability-and-measure cluster by working through these convergence theorems, plus their cousins. The first piece of business is to recognize that a sequence of random variables can "converge" in several different senses — four standard ones, with logical relations among them but each useful in a different context (for an estimator's consistency, for an asymptotic distribution, for a limit theorem in functional analysis). Section 2 lays out the four modes and their logical relations. Section 3 covers the Borel–Cantelli lemmas, which are the standard tool for almost-sure-convergence arguments. Sections 4 and 5 are the headline acts: the law of large numbers and the central limit theorem.

The cluster as a whole has been heavy on machinery, and this handout is where the machinery starts paying off. Once we know that sample averages converge to expectations, and that they do so with normally-distributed fluctuations of order $1/\sqrt{n}$, every empirical aggregation argument in political-science modeling acquires a rigorous backstop.

2 Modes of convergence

When a political scientist says "the polling estimator is consistent — it converges to the true population vote share as the sample grows," which kind of convergence is being claimed? It turns out there are at least four distinct technical answers, each useful for different theorems and different applications. The strong-law-of-large-numbers sense is *almost-sure* convergence: a single sequence of polls converges, sample by sample. The weak-law sense is convergence *in probability*: the chance of being far from the limit shrinks to zero. The asymptotic-distribution sense (used in confidence intervals) is convergence *in distribution*: the CDFs line up in the limit, but the random variables themselves need not. The mean-square sense (used in econometric efficiency arguments) is convergence in L^2: the mean squared error goes to zero. The four modes are listed below in roughly decreasing order of strength, with the precise logical relations and the standard counter-examples making up the rest of the section.

Definition 1 (Almost-sure convergence)**.** $X_n \to X$ *almost surely* (a.s., or *with probability one*) if

$$\mathbb{P}\left(\left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

We write $X_n \xrightarrow{\text{a.s.}} X$.

This is convergence "pointwise on Ω, except possibly on a null set." It is the strongest of the four modes and the closest to the everyday-calculus notion of convergence.

Definition 2 (Convergence in probability)**.** $X_n \to X$ *in probability* if for every $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| \geq \epsilon) \to 0 \quad \text{as } n \to \infty.$$

We write $X_n \xrightarrow{\mathbb{P}} X$.

Convergence in probability says "the probability of being far from the limit goes to zero," for any tolerance level. Notice that for any single n, it could be that $|X_n - X|$ is large with positive probability; the condition only constrains how that probability shrinks with n.

Definition 3 (Convergence in L^p)**.** For $p \geq 1$, $X_n \to X$ *in* L^p if $\mathbb{E}[|X_n - X|^p] \to 0$ as $n \to \infty$. We write $X_n \xrightarrow{L^p} X$.

The case $p = 2$ (mean-square convergence) is the most common in practice. L^p convergence is the natural mode in functional analysis: the random variables are viewed as elements of the Banach space $L^p(\Omega, \mathcal{F}, \mathbb{P})$, and convergence is in the p-norm.

Definition 4 (Convergence in distribution)**.** $X_n \to X$ *in distribution* (or *weakly*) if $F_{X_n}(x) \to F_X(x)$ at every continuity point x of F_X. We write $X_n \xrightarrow{d} X$ or $X_n \Rightarrow X$.

Convergence in distribution is convergence of the CDFs, not of the random variables themselves. Two important features. First, the random variables need not be defined on the same probability space; the only thing that matters is the distributions $\mathbb{P}_{X_n}, \mathbb{P}_X$. Second, the "at continuity points of F_X" qualifier is essential: a sequence of continuous distributions can converge in distribution to a discrete distribution, and at the discrete jump points the CDFs need not agree in the limit.

Theorem 5 (Implications among modes)**.** *For sequences of random variables* (X_n) *and a random variable* X *on the same probability space:*

- $X_n \xrightarrow{a.s.} X$ *implies* $X_n \xrightarrow{\mathbb{P}} X$.
- $X_n \xrightarrow{L^p} X$ *implies* $X_n \xrightarrow{\mathbb{P}} X$ *(for* $p \geq 1$*).*
- $X_n \xrightarrow{\mathbb{P}} X$ *implies* $X_n \xrightarrow{d} X$.

The converses fail in general; the first two are linked but neither is stronger than the other.[1]

[1]The converse implications fail with standard counter-examples. *In probability* $\not\Rightarrow$ *a.s.*: on $\Omega = [0,1]$ with Lebesgue measure, take X_n to be the indicator of a sequence of intervals of shrinking but periodically resetting length — e.g., $\mathbf{1}_{[0,1/2]}, \mathbf{1}_{[1/2,1]}, \mathbf{1}_{[0,1/3]}, \mathbf{1}_{[1/3,2/3]}, \mathbf{1}_{[2/3,1]}, \ldots$. Then $X_n \xrightarrow{\mathbb{P}} 0$ (the lengths go to 0) but for almost every ω, $X_n(\omega) = 1$ for infinitely many n, so $X_n(\omega) \not\to 0$. *A.s.* $\not\Rightarrow$ L^p: take $X_n = n\mathbf{1}_{[0,1/n]}$; then $X_n \to 0$ a.s. but $\mathbb{E}[X_n] = 1$ for every n, so no L^p convergence to 0. *In distribution* $\not\Rightarrow$ *in probability*: take X standard normal and $X_n = -X$ for every n; both have the same distribution (by symmetry of the normal), so $X_n \xrightarrow{d} X$, but $|X_n - X| = 2|X|$ does not go to 0 in probability. The diagram of implications and the ways they fail is the basic geometry of convergence in probability theory; Billingsley (1995) chapter 5 and Durrett (2019) chapter 2.3 work through the details. The dominated convergence theorem provides one important converse-with-extra-hypothesis: a.s. convergence *plus* a uniform integrable bound implies L^p convergence.

The intuition: a.s. and L^p are both "strong" modes that imply the "moderate" mode (in probability), which in turn implies the "weak" mode (in distribution). The two strong modes are not comparable: a.s. controls the random variables sample-wise but says nothing about their averages; L^p controls the averages but says nothing sample-wise. In practice, statisticians prove consistency in probability, then upgrade to a.s. via Borel–Cantelli when they need it, and use convergence in distribution for asymptotic normality.

3 Borel–Cantelli lemmas

How likely is it that a particular kind of event happens *infinitely often*? In an extended series of independent surveys, must some surveys be seriously off, just by chance? In a long sequence of elections, must wave-elections occur infinitely often? Questions of this shape — about whether infinitely many of a sequence of events occur, with positive probability or with probability one — are tail-event questions, and the right tool for them is the pair of *Borel–Cantelli lemmas*. The lemmas relate the probability of infinitely many occurrences to the convergence of $\sum \mathbb{P}(A_n)$, and they are the standard means of upgrading convergence-in-probability statements to almost-sure ones (a route we use in the proof of the strong law of large numbers in §4).

For events $A_1, A_2, \ldots$, the *lim sup* and *lim inf* are the "infinitely often" and "eventually" events:

$$\limsup_n A_n := \bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_k = \{A_k \text{ occurs for infinitely many } k\},$$

$$\liminf_n A_n := \bigcup_{n=1}^{\infty} \bigcap_{k \geq n} A_k = \{A_k \text{ occurs for all sufficiently large } k\}.$$

The Borel–Cantelli lemmas relate the probabilities of these limit events to the convergence of $\sum \mathbb{P}(A_n)$. They are the workhorse tools for arguments about whether infinitely many events occur with positive probability, and they are how one upgrades convergence-in-probability statements to almost-sure ones.

Theorem 6 (Borel–Cantelli I)**.** *If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\limsup_n A_n) = 0$ — i.e., almost surely, only finitely many of the A_n occur.*

Proof. $\mathbb{P}(\limsup A_n) = \mathbb{P}\left(\bigcap_n \bigcup_{k \geq n} A_k\right) \leq \mathbb{P}\left(\bigcup_{k \geq n} A_k\right) \leq \sum_{k \geq n} \mathbb{P}(A_k)$ by countable subadditivity, for every n. The right side is the tail of a convergent series and tends to 0 as $n \to \infty$. □

Theorem 7 (Borel–Cantelli II)**.** *If the events A_n are independent and $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then $\mathbb{P}(\limsup_n A_n) = 1$ — i.e., almost surely, infinitely many of the A_n occur.*

Proof sketch. $1 - \mathbb{P}(\limsup A_n) = \mathbb{P}(\liminf A_n^c) = \mathbb{P}\left(\bigcup_n \bigcap_{k \geq n} A_k^c\right) \leq \sum_n \mathbb{P}\left(\bigcap_{k \geq n} A_k^c\right)$. By independence, $\mathbb{P}\left(\bigcap_{k=n}^{N} A_k^c\right) = \prod_{k=n}^{N} (1 - \mathbb{P}(A_k)) \leq \prod_{k=n}^{N} e^{-\mathbb{P}(A_k)} = e^{-\sum_{k=n}^{N} \mathbb{P}(A_k)}$. Letting $N \to \infty$ and using $\sum \mathbb{P}(A_k) = \infty$, the bound goes to 0. □

The two lemmas are sharp: independence is essential for BC2 (without it, A_n could all be the same event of probability $1/2$, so $\sum \mathbb{P}(A_n) = \infty$ but $\mathbb{P}(\limsup A_n) = 1/2$, not 1). The asymmetry — BC1 needs no independence, BC2 does — is a standard small fact worth remembering.

4 The law of large numbers

The polling estimator $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ for i.i.d. samples X_i should converge to $\mathbb{E}[X_1] = \mu$. The law of large numbers is the formal statement, in two strengths.

Theorem 8 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}[|X_1|] < \infty$ and $\mathbb{E}[X_1] = \mu$. Then*

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{\mathbb{P}} \mu.$$

Proof under the stronger assumption $\mathbb{E}[X_1^2] < \infty$. $\mathbb{E}[\bar{X}_n] = \mu$ by linearity. By independence, $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(X_1)/n$. By Chebyshev's inequality (from the previous handout),

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\mathrm{Var}(X_1)}{n\epsilon^2} \to 0.$$

Removing the second-moment hypothesis to recover the theorem under only $\mathbb{E}[|X_1|] < \infty$ requires a truncation argument; we omit the details. □

Theorem 9 (Strong law of large numbers, Kolmogorov)**.** *Under the same hypothesis (X_i i.i.d., $\mathbb{E}[|X_1|] < \infty$, $\mathbb{E}[X_1] = \mu$),*

$$\bar{X}_n \xrightarrow{a.s.} \mu.$$

The strong law is genuinely deeper than the weak law and was the first major theorem to require the full machinery of measure-theoretic probability. The standard proof uses the Borel–Cantelli lemmas to upgrade convergence in probability to a.s. convergence, with a truncation-and-summation argument controlling the deviations.[2]

Example 10 (Polling consistency)**.** Continuing Example 10 of the previous handout: the polling estimator $\bar{X}_n$ for the vote share of L converges almost surely to the true population vote share p. The strong law gives the substantive content: as the sample grows, the estimator gets close to p and stays close — not just on average across hypothetical re-sampled studies (a weak-law claim), but on essentially every sample path. This is what consistency *means* for a real-world survey, and it is the formal version of the practical rule of thumb that "larger surveys are more reliable."

5 The central limit theorem

The law of large numbers tells us $\bar{X}_n \to \mu$. The next question is: how fast, and with what shape? The answer, due to de Moivre, Laplace, Lyapunov, Lindeberg, and Lévy in successive refinements, is that the appropriate rescaling of $\bar{X}_n - \mu$ is approximately normally distributed.

[2]The history is itself part of the story. The weak law in its modern generality is due to Khintchine (1929); the strong law to Kolmogorov (1930), with a crucial earlier special case (i.i.d. Bernoulli) due to Borel (1909) — the same Borel of Borel–Cantelli, and indeed Borel's strong law is the historical origin of the lemmas. The shift from weak to strong reflected the broader shift to measure-theoretic foundations in the 1920s and 1930s. For applied work the practical distinction is small: in any specific application either both laws hold or both fail (the conditions are the same), and the difference is just whether the convergence is a single-trajectory statement (strong) or a population-level statement (weak). Polling consistency is a strong-law claim — a single sequence of polls converges; a weak-law-only world would let any individual sequence wander while only the distribution of estimates concentrates. Durrett (2019) chapter 2 gives the careful treatment.

Theorem 11 (Central limit theorem, classical i.i.d. form)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}[X_1] = \mu$, $\mathrm{Var}(X_1) = \sigma^2 \in (0, \infty)$. Then*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1),$$

where $N(0,1)$ denotes the standard normal distribution.

The standard proof uses *characteristic functions*: $\phi_X(t) := \mathbb{E}[e^{itX}]$ is the Fourier transform of the distribution, and convergence in distribution is equivalent to pointwise convergence of characteristic functions (Lévy's continuity theorem). One shows that the characteristic function of the LHS converges to $e^{-t^2/2}$, the characteristic function of $N(0,1)$, by a Taylor expansion using the moment hypotheses. We omit the calculation; Billingsley (1995) or Durrett (2019) carry it out.

The substantive content. The deviation $\bar{X}_n - \mu$ is of order $\sigma/\sqrt{n}$ — shrinking, but slowly. Multiplied by $\sqrt{n}$ to compensate, the distribution becomes approximately normal as n grows. The scaling rate $1/\sqrt{n}$ is what underwrites the standard $\pm 1.96\,\sigma/\sqrt{n}$ confidence intervals: at the 95% level, $\mathbb{P}(|Z| \le 1.96) \approx 0.95$ for $Z \sim N(0,1)$, so $\mathbb{P}(|\bar{X}_n - \mu| \le 1.96\,\sigma/\sqrt{n}) \to 0.95$ as $n \to \infty$.

Example 12 (Survey margins of error)**.** A survey samples n voters and reports the proportion supporting L as $\bar{X}_n$, where X_i are i.i.d. Bernoulli(p). The CLT gives

$$\frac{\bar{X}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{d} N(0,1),$$

and a 95% confidence interval for p is approximately $\bar{X}_n \pm 1.96\sqrt{p(1-p)/n}$. The worst-case width (at $p = 1/2$) is $1.96/\sqrt{n}$, which produces the rule-of-thumb that a sample of $n = 1000$ has a margin of error of about ± 3 percentage points. This is the actual quantitative content of survey methodology, and it is one of the most consequential applications of the CLT in political science.

The classical CLT extends in many directions: independent but non-identically-distributed summands (Lindeberg–Feller), dependent summands under various mixing conditions (martingale CLTs, mixing CLTs), random vectors (multivariate CLT), and so on. For applied political science the i.i.d. case covers most of what one needs in survey settings, and the more sophisticated versions are deployed when the design departs from i.i.d. sampling (cluster sampling, time-series data, network data).

6 What's next

This closes the probability-and-measure cluster. The reader has the formal apparatus to read any rigorous probabilistic argument in political-economy theory: a probability space, random variables and their distributions, expectations and conditional expectations, the inequalities, and the asymptotic theorems.

The natural follow-on is into specific applied directions, none of which is yet a planned handout in this project but each of which builds directly on the foundations now in place:

- *Statistical inference.* The LLN and CLT are the foundation of point estimation, hypothesis testing, and confidence intervals. The bridge from probability theory to political-science methodology is what comes next when one moves from the foundations into applied empirical work.

- *Stochastic processes.* Sequences of random variables indexed by time, with dependence structure — Markov chains, martingales, Brownian motion. These are essential for dynamic models in political-economy theory and for time-series analysis in empirical work.
- *Decision theory under risk and ambiguity*, the next applied cluster of this project: the von Neumann–Morgenstern axioms (choice under risk) and Savage's subjective expected utility (choice under ambiguity) build directly on probability theory plus the order-theoretic machinery of preferences from earlier in the project.

For broader treatments, Billingsley (1995), Durrett (2019), and Williams (1991) all develop the limit theorems and convergence modes carefully. Williams (1991) is the gentlest first reading and includes martingale theory.

7 Exercises

Exercise 13. Prove that almost-sure convergence implies convergence in probability. (*Hint*: $|X_n - X| \geq \epsilon$ for some $n \geq m$ is the event $\bigcup_{n \geq m} \{|X_n - X| \geq \epsilon\}$, whose probability is bounded by $\mathbb{P}(\sup_{n \geq m} |X_n - X| \geq \epsilon)$. Use that this last quantity goes to 0 as $m \to \infty$ on the event of a.s. convergence.)

Exercise 14. Prove that L^p convergence implies convergence in probability via Markov's inequality applied to $|X_n - X|^p$.

Exercise 15. Construct a sequence of random variables X_n on $[0, 1]$ with Lebesgue measure that converges to 0 in probability but not almost surely. (See the footnote in §2 for the standard construction; spell it out in detail and verify both the in-probability convergence and the almost-sure failure.)

Exercise 16 (Borel–Cantelli applied to polling)**.** Suppose at each time n a survey is conducted and let A_n be the event "the survey's reported vote share is more than 5 percentage points off from the true population share." Suppose the surveys are independent and $\mathbb{P}(A_n) = c_n$ for some sequence c_n. Use Borel–Cantelli I and II to characterize when "a.s. only finitely many surveys are seriously off" versus "a.s. infinitely many surveys are seriously off." Comment on what $c_n \to 0$ at rate $1/n^{1/2}$ versus $1/n$ versus $1/n^2$ implies for each.

Exercise 17 (Polling consistency from Chebyshev)**.** Continuing Example 10: directly from Chebyshev's inequality, give an explicit upper bound on $\mathbb{P}(|\bar{X}_n - p| \geq 0.02)$ as a function of n. With this Chebyshev bound, determine the smallest n such that $\mathbb{P}(|\bar{X}_n - p| \geq 0.02) \leq 0.05$ at $p = 0.5$. Compare this bound with the CLT-derived n from the worked example in §5. Why is the Chebyshev bound much more conservative?

Exercise 18. Let X_n be a sequence with $\mathbb{P}(X_n = 1/n) = 1$ for each n (deterministic). Show that $X_n \to 0$ in all four modes (a.s., probability, L^p for any p, distribution). Now let Y_n be independent with $\mathbb{P}(Y_n = n) = 1/n$ and $\mathbb{P}(Y_n = 0) = 1 - 1/n$. Show that $Y_n \to 0$ in probability and in distribution, but *not* in L^1 (compute $\mathbb{E}[Y_n]$). Where does Borel–Cantelli II tell us Y_n goes a.s.?

Exercise 19 (CLT with non-identically-distributed votes)**.** Suppose voter i votes for L with probability p_i, with p_i varying across voters but bounded away from 0 and 1, and the votes are independent. Let $S_n = \sum_{i=1}^{n} X_i$ be the total vote count. Compute $\mathbb{E}[S_n]$ and $\mathrm{Var}(S_n)$. State

without proof (it is a Lindeberg–Feller CLT) that $(S_n - \mathbb{E}[S_n])/\sqrt{\mathrm{Var}(S_n)} \xrightarrow{d} N(0,1)$ as $n \to \infty$, under the boundedness assumption. What does this tell us about the asymptotic distribution of vote totals in a heterogeneous electorate?

Exercise 20. Let X_n be uniform on $[0, 1/n]$. Show $X_n \to 0$ in L^1, in probability, and a.s. (Verify each in turn.)

Exercise 21 (Aggregation and the LLN)**.** Suppose policy preferences in a population are summarized by individual ideal points $X_1, X_2, \ldots$, drawn i.i.d. from some distribution with mean μ (the population mean preference) and finite variance. The strong LLN says the sample average $\bar{X}_n \to \mu$ a.s. Argue that the median voter theorem's prediction ("the median voter's ideal point is decisive") is, in the limit, also a statement about μ when the underlying distribution is symmetric. (*Hint*: for a symmetric distribution, mean equals median, so the LLN-derived sample mean is also the asymptotic sample median.)

Exercise 22. Suppose $X_n \xrightarrow{d} X$ where X has a continuous distribution function F. Show that for any $a < b$ in $\mathbb{R}$, $\mathbb{P}(a < X_n \leq b) \to \mathbb{P}(a < X \leq b)$. Why does the continuity hypothesis on F matter? Give a counter-example with a discontinuous limit distribution.

References

Billingsley, Patrick (1995). *Probability and Measure*. 3rd ed. New York: Wiley.

Durrett, Rick (2019). *Probability: Theory and Examples*. 5th ed. Cambridge: Cambridge University Press.

Williams, David (1991). *Probability with Martingales*. Cambridge: Cambridge University Press.