Asymptotics I: Convergence and the Law of Large Numbers

Gov 2001: Quantitative Social Science Methods I

Scott Cunningham

Harvard University

Spring 2026

Today's reading

Required

- **Aronow & Miller**, §3.2: Asymptotics (pp. 92–110)
- **Blackwell**, Ch. 3: Large-sample properties

Recommended

- **Casella & Berger**, Ch. 5: Properties of a random sample

Let's derive some MLEs together

Monday gave you the machinery — now we practice

Two examples:

1. **Poisson** — count data (easy, one parameter)
2. **Normal** — continuous data (harder, two parameters)

Same four steps every time: write likelihood → take log → differentiate → solve.

Count data are everywhere in political science

Example 1: Poisson model

Model: $X_i \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$ for $i = 1, \ldots, n$

PMF:

$$f(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots$$

Examples:

- Protests per country-month
- Bills introduced per legislator per session
- Civilian casualties per district-year
- FOIA requests per agency per quarter

One unknown parameter: $\lambda > 0$ (the rate).

Poisson log-likelihood

Joint PMF (i.i.d.):

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Log-likelihood:

$$\ell(\lambda) = \sum_{i=1}^{n} \left[x_i \log \lambda - \lambda - \log(x_i!) \right]$$

Simplify:

$$\ell(\lambda) = \log \lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log(x_i!)$$

The last term is a constant — it won't affect the maximizer.

Poisson MLE: differentiate and solve

First-order condition:

$$\frac{d\ell}{d\lambda} = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0$$

Solve:

$$\frac{1}{\lambda}\sum_{i=1}^{n} x_i = n \quad \Longrightarrow \quad \hat{\lambda}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\lambda}_{\text{MLE}} = \bar{X}$$

Plug-in and MLE coincide again — Poisson is an exponential family.

Verify it's a maximum

Second derivative:

$$\frac{d^2\ell}{d\lambda^2} = -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i < 0 \quad \text{for all } \lambda > 0$$

The log-likelihood is globally concave in λ.

$\hat{\lambda}_{\text{MLE}} = \bar{X}$ is the unique global maximum. ✓

Now both parameters are unknown

Example 2: Normal model

Model: $X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ for $i = 1, \ldots, n$

PDF:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Political science framing:

- District-level vote shares
- Measurement error in survey responses
- Ideology scores (NOMINATE, ideal points)

Two unknown parameters: μ (location) and σ^2 (spread).

The likelihood is a product over all observations

Same step as Poisson — write the joint density

Joint PDF (i.i.d.):

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i \mid \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Next: take the log to turn this product into a sum.

Take the log to get the normal log-likelihood

Log-likelihood:

$$\ell(\mu, \sigma^2) = \sum_{i=1}^{n} \log f(x_i \mid \mu, \sigma^2)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\boxed{\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Two parameters $\longrightarrow$ two partial derivatives.

Solving for $\hat{\mu}$: differentiate with respect to μ

Partial derivative:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

Solve:

$$\sum_{i=1}^{n} x_i - n\mu = 0 \quad \Longrightarrow \quad \hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{\mu}_{\text{MLE}} = \bar{X}$$

The sample mean — no surprise here.

Solving for $\hat{\sigma}^2$: differentiate with respect to σ^2

Partial derivative (treat σ^2 as a single variable):

$$\frac{\partial \ell}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

Solve:

$$\frac{n}{2\sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 \implies \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

Divides by n, not $n-1$.

What happens when we take the expectation?

Is $\hat{\sigma}^2_{\text{MLE}}$ unbiased?

Start from our MLE:

$$\mathrm{E}\left[\hat{\sigma}^2_{\text{MLE}}\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[(X_i - \bar{X})^2\right]$$

Key identity: $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$

Next: take expectations of both sides.

The MLE for variance is biased

Dividing by n undershoots

Take expectations:

$$\mathrm{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = n\sigma^2 - n \cdot \frac{\sigma^2}{n} = (n-1)\sigma^2$$

Divide by n:

$$\boxed{\mathrm{E}\left[\hat{\sigma}^2_{\mathrm{MLE}}\right] = \frac{(n-1)\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \neq \sigma^2}$$

The MLE systematically underestimates the variance.

The MLE for variance is biased — but the bias vanishes

We just showed:

$$\mathrm{E}\left[\hat{\sigma}^2_{\mathrm{MLE}}\right] = \frac{n-1}{n}\,\sigma^2 \;\neq\; \sigma^2$$

The bias: $-\sigma^2/n$

As n grows: $\frac{n-1}{n} \to 1$, so the bias shrinks to zero

Key question: What does "the bias vanishes as n grows" actually mean, formally?

That's what today is about.

The question “what happens with more data?” has many children

A brief history of asymptotic theory

1713 **Jacob Bernoulli**, *Ars Conjectandi* — first LLN, proved for Bernoulli trials

1733 **De Moivre** — normal approximation to the binomial

1867 **Chebyshev** — general inequality that makes the WLLN proof elegant

1929 **Khintchine** — WLLN without requiring finite variance

1933 **Kolmogorov** — Strong LLN, full axiomatization of probability

It started with one man in Basel, twenty years before anyone else thought to ask.

Jacob Bernoulli (1655–1705): the father of asymptotics

Basel, Switzerland — University of Basel, 1687–1705

- Trained as a **theologian** (his father wanted him in the ministry) — turned to mathematics against his family's wishes
- Part of the extraordinary **Bernoulli dynasty** — eight mathematicians across three generations
- Bitter rivalry with his younger brother **Johann**, who was arguably more talented — they publicly attacked each other's work
- One of the first to master **Leibniz's new calculus** (Newton's *Principia*: 1687)

His tombstone in Basel's cathedral bears a logarithmic spiral and the inscription *Eadem mutata resurgo* — "Though changed, I rise again the same."

Bernoulli wanted to prove that observation yields "moral certainty"

The direct problem (easy):

An urn has 3,000 white and 2,000 black pebbles. Probability of drawing white = 3/5.

The inverse problem (Bernoulli's question):

You *don't know* the ratio. Can repeated draws tell you what it is?

The judicial framing:

How many cases must a judge observe before acting with *moral certainty* — certainty sufficient for practical action, not mathematical proof?

Bernoulli's answer: For any desired precision and any desired confidence, a finite number of trials suffices.

This is the Law of Large Numbers.

It took Bernoulli twenty years to prove it

What he had: combinatorial methods, binomial coefficients, early calculus

What he lacked: no variance, no normal approximation, no Chebyshev inequality

His approach: bound the binomial tail term by term — show the central terms dominate by any desired factor as n grows

Painstaking, elementary, and very conservative: his bounds said $n = 25{,}550$ trials suffice where modern tools need far fewer

He died in 1705 before publishing. His nephew edited *Ars Conjectandi*, published 1713.

Bayes' *Essay on the Doctrine of Chances* was also published posthumously (1763, by Richard Price). Two theologians-turned-mathematicians, both solving the inverse problem, both published after death.

What happens to our estimators as we collect more data?

Monday we built two estimators for voter turnout θ:

- **Plug-in**: $\hat{\theta} = \bar{X} = 68/200 = 0.34$
- **MLE**: $\hat{\theta}_{\text{MLE}} = \bar{X} = 0.34$

We said both are "consistent" — they converge to the true θ.

Three questions:

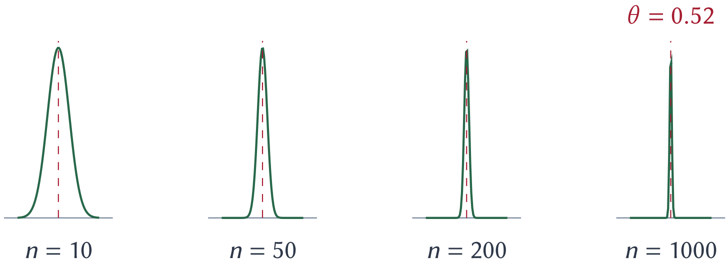
1. What does "converge" mean for a random variable?
2. Why should we believe sample averages converge?
3. What can we say about *how fast*?

Roadmap

1. **Convergence in probability**
   The formal concept: what it means for randomness to become negligible
2. **Markov → Chebyshev → Law of Large Numbers**
   A proof chain: each inequality builds on the last
3. **Consistency**
   The property we actually care about for estimators

What does it mean for a random variable to "settle down"?

Consider: $\bar{X}_n$ computed from n draws from Bernoulli(0.52)



The randomness doesn't disappear — it just becomes negligible.

Convergence in probability

Definition

X_n **converges in probability** to c, written $X_n \xrightarrow{p} c$, if for every $\varepsilon > 0$:

$$\lim_{n \to \infty} \Pr\big(|X_n - c| > \varepsilon\big) = 0$$

Notation: $\operatorname{plim} X_n = c$

In words: the probability that X_n is "far" from c goes to zero, no matter how small you set "far."

This is weaker than "X_n equals c for large n" — there's still randomness, but it becomes negligible.

Example: opinion polls with increasing sample sizes

True support $\theta = 0.52$, threshold $\varepsilon = 0.05$

$\bar{X}_n$ from Bernoulli(0.52). Exact probability via binomial:

n	$\Pr(\lvert\bar{X}_n - 0.52\rvert > 0.05)$	Interpretation
50	0.565	More likely wrong than right
200	0.162	Usually within 5 points
500	0.025	Rarely off by 5 points
2000	< 0.001	Essentially on target

For any fixed ε, the probability goes to zero. That's convergence in probability.

How we computed those probabilities

The binomial distribution does the work

Each $X_i \sim \text{Bernoulli}(\theta)$ iid, so the sum $S_n = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, \theta)$.

Since $\bar{X}_n = S_n/n$, the event $|\bar{X}_n - \theta| > \varepsilon$ is equivalent to:

$$S_n < (\theta - \varepsilon)\, n \quad \text{or} \quad S_n > (\theta + \varepsilon)\, n$$

This works for *any* θ — that's the LLN. We used $\theta = 0.52$ only to produce actual numbers.

Step 1: Compute the cutoffs

$\theta = 0.52$, $\varepsilon = 0.05$, $S_{50} \sim \text{Binomial}(50,\ 0.52)$

$\bar{X}_{50}$ is "off by more than 0.05" when $\bar{X}_{50} < 0.47$ or $\bar{X}_{50} > 0.57$.

Multiply through by $n = 50$: $\quad S_{50} < 0.47 \times 50 = 23.5 \quad \text{or} \quad S_{50} > 0.57 \times 50 = 28.5$

But S_{50} counts heads — it's an **integer**.

So: $\quad S_{50} \leq 23 \quad \text{or} \quad S_{50} \geq 29$

Step 2: Look up the binomial CDF

$S_{50} \sim \text{Binomial}(50,\ 0.52)$

We need the probability of the two tails:

$$\Pr(S_{50} \leq 23) \ + \ \Pr(S_{50} \geq 29)$$

Rewrite the upper tail using the CDF:

$$\Pr(S_{50} \leq 23) \ + \ \left[1 - \Pr(S_{50} \leq 28)\right] \ = \ 0.565$$

More than half the time, $\bar{X}_{50}$ lands farther than 0.05 from the truth.

And this is what I would do

Computing $\Pr(|\bar{X}_{50} - 0.52| > 0.05)$ in code

In R:

```r
pbinom(23, 50, 0.52) + 1 - pbinom(28, 50, 0.52)
```

In Python:

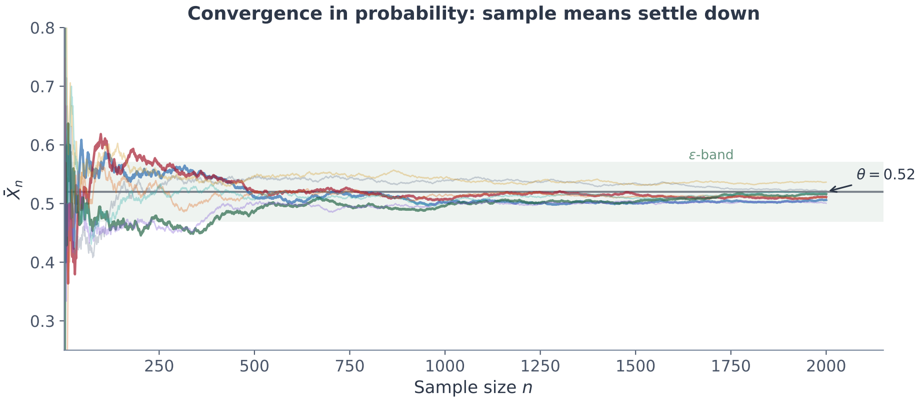
```python
from scipy.stats import binom
binom.cdf(23, 50, 0.52) + 1 - binom.cdf(28, 50, 0.52)
```

Both return 0.565. Repeat for $n = 200, 500, 2000$ to fill the table.

What convergence in probability is not

- **Not "$X_n = c$ for large n"**
  $\bar{X}_n$ is still random for any finite n; it just stays close to c with high probability
- **Not almost sure convergence**
  Almost sure: $\Pr(\lim X_n = c) = 1$ (stronger — every sample path converges)
  Convergence in probability: tails shrink, but occasional "excursions" are possible
- **Not a statement about any fixed n**
  It's a statement about the sequence $X_1, X_2, X_3, \ldots$ as $n \to \infty$

Simulation: sample paths converging to $\theta = 0.52$



Eight independent sequences of $\bar{X}_n$ from Bernoulli(0.52), $n = 1, \ldots, 2000$. All paths eventually enter the ε-band.

Useful properties of convergence in probability

If $X_n \xrightarrow{p} a$ and $Y_n \xrightarrow{p} b$, then:

- $X_n + Y_n \xrightarrow{p} a + b$
- $X_n Y_n \xrightarrow{p} ab$
- $X_n / Y_n \xrightarrow{p} a/b \quad (\text{if } b \neq 0)$

Continuous Mapping Theorem

Once you prove one convergence, you get many more for free

If g is continuous at a and $X_n \xrightarrow{p} a$, then:

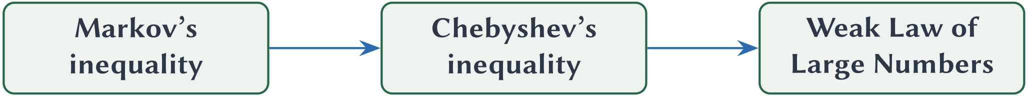
$$g(X_n) \xrightarrow{p} g(a)$$

Why this matters. Suppose you've shown $\bar{X}_n \xrightarrow{p} \mu$ by the LLN. Now you need $\bar{X}_n^2$. Squaring is continuous, so $\bar{X}_n^2 \xrightarrow{p} \mu^2$ — no new proof required.

The sample variance needs $\overline{X^2} - \bar{X}_n^2$. The LLN gives $\overline{X^2} \xrightarrow{p} E[X^2]$, and the CMT gives $\bar{X}_n^2 \xrightarrow{p} \mu^2$. Consistency of the sample variance follows immediately.

Any smooth transformation of a consistent estimator is itself consistent.

The proof chain: Markov → Chebyshev → LLN



Bounds tail using mean

Bounds tail using variance

$\bar{Y}_n \xrightarrow{p} \mu$

Each result builds on the last. The entire chain is about four lines of math.

A question only the mean can answer

Setting: You pull data from the Current Population Survey.

Mean household income: $\mathrm{E}[X] = \$100{,}000$

Question: What fraction of households earn more than \$500,000?

You don't know the shape of the distribution. You don't know the variance. All you have is the mean.

Can you say *anything* useful?

Andrey Markov (1856–1922) answered with just the mean

St. Petersburg, Russia — student of Chebyshev, succeeded him at the Russian Academy of Sciences

His insight: If a quantity is nonnegative and its average is small, it can't often be large.

- Mean income is \$100,000
- At most \$100,000/\$500,000 = 20% of households can earn ≥\$500K

Crude, but it works for **any** nonnegative quantity — no assumptions about the distribution's shape.

Markov's inequality: the formal result

Markov's Inequality

If $X \geq 0$ and $\mathrm{E}[X]$ exists, then for any $a > 0$:

$$\Pr(X \geq a) \leq \frac{\mathrm{E}[X]}{a}$$

Our CPS example: X = household income, $\mathrm{E}[X] = \$100{,}000$, $a = \$500{,}000$

$$\Pr(X \geq 500{,}000) \leq \frac{100{,}000}{500{,}000} = 0.20$$

At most 20% of households earn \$500K or more — guaranteed, regardless of the distribution.

Markov's inequality: proof

Start from the definition of expectation ($X \geq 0$):

$$\mathrm{E}[X] = \int_0^\infty x f(x)\, dx$$

Drop the part below a (it's nonnegative):

$$\mathrm{E}[X] \geq \int_a^\infty x f(x)\, dx$$

Bound $x \geq a$ in the remaining integral:

$$\int_a^\infty x f(x)\, dx \geq a \int_a^\infty f(x)\, dx = a \cdot \Pr(X \geq a)$$

$$\Pr(X \geq a) \leq \frac{\mathrm{E}[X]}{a} \quad \text{QED}$$

Markov alone is too loose — we need variance

Markov only uses the mean. It ignores how spread out X is.

Example: $X \sim \text{Bernoulli}(0.01)$

- $\Pr(X \geq 1) = 0.01$ (exact)
- Markov bound: $\Pr(X \geq 1) \leq 0.01/1 = 0.01$ (tight here)

But for $\bar{X}_n$ from Bernoulli(0.52), $n = 400$:

- We want $\Pr(|\bar{X} - 0.52| > 0.05)$
- Markov on $|\bar{X} - 0.52|$ gives a bound > 1 — useless!

Fix: Apply Markov to $(\bar{X} - 0.52)^2$ instead — that uses the variance.

Now suppose we also know the variance

Same CPS data: $\mathrm{E}[X] = \$100{,}000$

New information: $\mathrm{SD}(X) = \$40{,}000$

Question: What fraction of households are more than $80,000 away from the mean?

Markov can't help — it doesn't use the variance.

Can knowing the spread give us a tighter bound?

Pafnuty Chebyshev (1821–1894) used the variance

St. Petersburg, Russia — founder of the St. Petersburg school of mathematics, Markov's teacher

His insight (1867): The variance tells you how concentrated a distribution is around its mean.

- SD = \$40,000, and \$80,000 = 2 standard deviations
- At most $1/2^2 = 25\%$ of households can be more than 2 standard deviations from the mean

Tighter than Markov, and still works for **any** distribution with finite variance.

Chebyshev's inequality: the formal result

Chebyshev's Inequality

For any random variable X with $\mathrm{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2 < \infty$:

$$\Pr\big(|X - \mu| \geq t\big) \leq \frac{\sigma^2}{t^2}$$

Proof: Apply Markov's inequality to $(X - \mu)^2$ with threshold t^2:

$$\Pr\big((X - \mu)^2 \geq t^2\big) \leq \frac{\mathrm{E}[(X - \mu)^2]}{t^2} = \frac{\sigma^2}{t^2}$$

One line. That's the whole proof. Chebyshev is just Markov applied to the squared deviation.

Chebyshev in "k-sigma" form

Set $t = k\sigma$:

$$\Pr\left(|X - \mu| \geq k\sigma\right) \leq \frac{1}{k^2}$$

k	Chebyshev bound	Normal (for comparison)
1	$\leq 100\%$	31.7%
2	$\leq 25\%$	4.6%
3	$\leq 11.1\%$	0.3%
5	$\leq 4\%$	$< 0.001\%$

Chebyshev is conservative but holds for **any** distribution with finite variance.

Chebyshev example: how far can a poll be from the truth?

$\bar{X}_n$ from Bernoulli(0.52), $n = 400$

- $\mathrm{E}[\bar{X}] = 0.52, \quad \mathrm{Var}(\bar{X}) = \dfrac{0.52 \times 0.48}{400} = 0.000624$

Chebyshev:

$$\Pr(|\bar{X} - 0.52| \geq 0.05) \leq \frac{0.000624}{0.05^2} = \frac{0.000624}{0.0025} = 0.2496$$

Exact (binomial): ≈ 0.046

Chebyshev overestimates by 5× — but it works without knowing the distribution is binomial.

The Weak Law of Large Numbers

WLLN

If $Y_1, Y_2, \ldots$ are i.i.d. with $\mathrm{E}[Y_i] = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2 < \infty$, then:

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} \mu$$

In words: the sample mean converges in probability to the population mean.

This is why statistics works. With enough data, sample averages tell us about populations.

WLLN proof via Chebyshev (three lines)

We know:

- $\mathrm{E}[\bar{Y}_n] = \mu$
- $\mathrm{Var}(\bar{Y}_n) = \sigma^2/n$

Apply Chebyshev to $\bar{Y}_n$: for any $\varepsilon > 0$,

$$\Pr\big(|\bar{Y}_n - \mu| \geq \varepsilon\big) \leq \frac{\mathrm{Var}(\bar{Y}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n\to\infty} 0$$

$$\Pr\big(|\bar{Y}_n - \mu| \geq \varepsilon\big) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0 \qquad \text{QED}$$

One of the most elegant proofs in statistics. Three centuries of work, distilled into three lines.

What the LLN tells us — and what it doesn't

What it tells us:

- Sample averages converge to population averages
- The estimator "settles down" to the right answer
- More data $\rightarrow$ better estimates (formally)

What it does not tell us:

- How close $\bar{Y}_n$ is to μ for any *finite* n
- The *shape* of the sampling distribution
- How to construct confidence intervals

For that, we need the Central Limit Theorem — Monday.

The LLN applies to any sample average

If $g(Y)$ is any function with $\mathrm{E}[|g(Y)|] < \infty$:

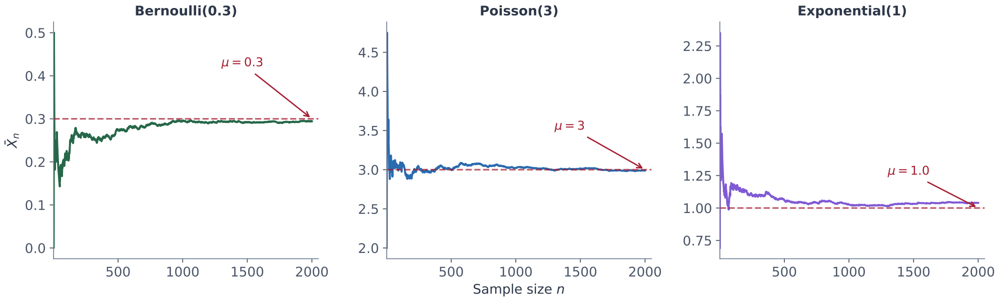
$$\frac{1}{n}\sum_{i=1}^{n} g(Y_i) \xrightarrow{p} \mathrm{E}[g(Y)]$$

Examples:

- $\frac{1}{n}\sum Y_i^2 \xrightarrow{p} \mathrm{E}[Y^2]$
- $\frac{1}{n}\sum (Y_i - \bar{Y})^2 \xrightarrow{p} \mathrm{Var}(Y)$
- $\frac{1}{n}\sum X_i Y_i \xrightarrow{p} \mathrm{E}[XY]$

Sample variances, sample covariances, sample correlations — all consistent.

Simulation: the LLN in action across three distributions



Different distributions, same story: $\bar{X}_n \to \mu$ as n grows.

Consistency: the minimal requirement for an estimator

Definition

An estimator $\hat{\theta}_n$ is **consistent** for θ if:

$$\hat{\theta}_n \xrightarrow{p} \theta \quad \text{as } n \longrightarrow \infty$$

In words: with enough data, the estimator gets arbitrarily close to the truth with high probability.

The LLN gives us consistency for free: $\bar{X}$ is consistent for μ because $\bar{X} \xrightarrow{p} \mu$.

Consistency vs. unbiasedness: two different virtues

	Unbiasedness	**Consistency**
Statement	$\mathrm{E}[\hat{\theta}] = \theta$	$\hat{\theta}_n \xrightarrow{p} \theta$
Applies to	Any fixed n	The sequence as $n \to \infty$
Meaning	Correct on average	Correct in the limit

Neither implies the other!

Unbiased but inconsistent — and vice versa

Unbiased but not consistent: $\hat{\mu} = X_1$

- $E[X_1] = \mu \quad \checkmark \quad$ (unbiased for any n)
- $\text{Var}(X_1) = \sigma^2 \quad$ (doesn't shrink!)
- Ignores $X_2, \ldots, X_n$ entirely — more data doesn't help

Biased but consistent: $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum (X_i - \bar{X})^2$

- $E[\hat{\sigma}^2_{\text{MLE}}] = \frac{n-1}{n}\sigma^2 \neq \sigma^2 \quad$ (biased)
- $\hat{\sigma}^2_{\text{MLE}} \xrightarrow{p} \sigma^2 \quad \checkmark \quad$ (consistent — bias $\rightarrow 0$, variance $\rightarrow 0$)

Consistency is about the destination; unbiasedness is about the journey.

A sufficient condition for consistency

If $\text{Bias}(\hat{\theta}_n) \to 0$ and $\text{Var}(\hat{\theta}_n) \to 0$ as $n \to \infty$, **then** $\hat{\theta}_n$ is consistent for θ.

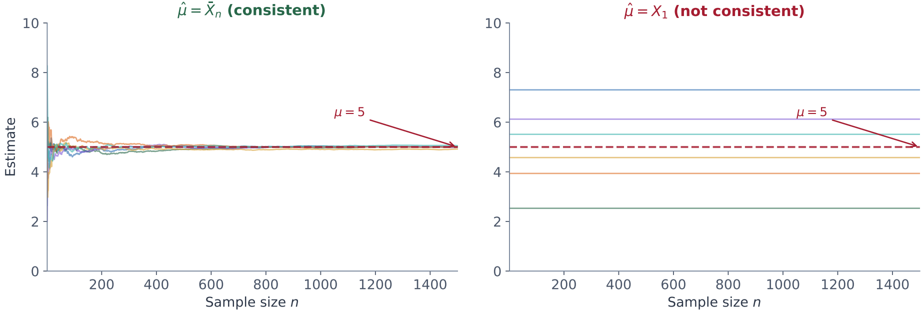
Proof idea: By Chebyshev applied to $\hat{\theta}_n$:

$$\Pr\left(|\hat{\theta}_n - \theta| \geq \varepsilon\right) \leq \frac{\text{MSE}(\hat{\theta}_n)}{\varepsilon^2} = \frac{\text{Bias}^2 + \text{Var}}{\varepsilon^2} \to 0$$

This is why we check bias and variance separately. Both must vanish for guaranteed consistency.

Sufficient but not necessary — there are consistent estimators where the bias doesn't vanish (but the variance dominates).

Simulation: consistent vs. inconsistent estimators



Left: $\bar{X}_n$ concentrates around $\mu = 5$ as n grows. **Right**: X_1 just sits wherever it landed.

MLE is consistent (under regularity conditions)

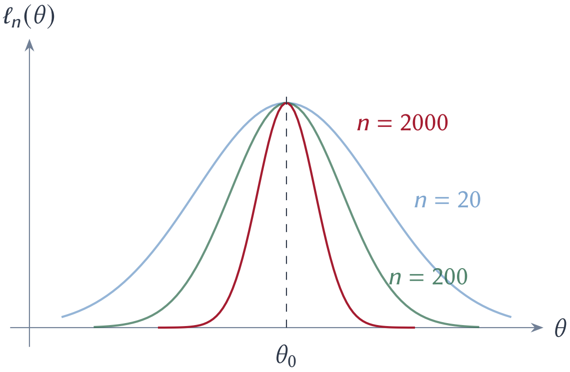
Why? Informal argument:

1. MLE maximizes $\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i \mid \theta)$
2. By the LLN: $\ell_n(\theta) \xrightarrow{p} \mathrm{E}[\log f(X \mid \theta)]$ for each θ
3. $\mathrm{E}[\log f(X \mid \theta)]$ is maximized at θ_0 (the true parameter)
4. The maximizer of ℓ_n converges to the maximizer of the limit

Step 3 uses KL divergence: $\mathrm{E}_{\theta_0}[\log f(X \mid \theta)]$ is maximized when $\theta = \theta_0$.

From 06a: even if the model is wrong, MLE finds the θ^* closest to the truth in KL divergence.

The likelihood surface sharpens around the truth



As n grows, the log-likelihood concentrates around θ_0. The MLE (the peak) has nowhere to go but the truth.

Continuous Mapping Theorem: consistency under transformations

CMT

If $X_n \xrightarrow{p} c$ and g is continuous at c, then:

$$g(X_n) \xrightarrow{p} g(c)$$

Connection to MLE invariance (from Monday):

If $\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta$, then $g(\hat{\theta}_{\text{MLE}}) \xrightarrow{p} g(\theta)$.

Example: $\hat{p} = \bar{X} \xrightarrow{p} p$ from Bernoulli data.

By CMT: $\text{odds}(\hat{p}) = \frac{\hat{p}}{1-\hat{p}} \xrightarrow{p} \frac{p}{1-p}$

Consistent estimators of transformations come free.

Among consistent estimators, some converge faster

Many estimators are consistent for μ:

- $\bar{X}_n$ (sample mean)
- Sample median
- 10%-trimmed mean

All converge to μ (for symmetric distributions). But they converge at **different rates**.

From Monday: Fisher information $I(\theta)$ and the Cramér-Rao bound set a floor:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{n\, I(\theta)}$$

MLE achieves this floor asymptotically — it's the **most efficient** consistent estimator.

Efficiency = speed of convergence. We'll formalize this with the CLT on Monday.

Key takeaways

1. **Convergence in probability**: random variables can "settle down" to constants
2. **Markov → Chebyshev → LLN**: a proof chain, each building on the last
3. **Law of Large Numbers**: sample averages converge to population averages
4. **Consistency**: the minimal requirement for a useful estimator
5. **MLE is consistent** (under regularity conditions) — the likelihood sharpens around the truth

What's still missing

The LLN says $\bar{Y}_n \to \mu$. But:

- How fast does $\bar{Y}_n$ approach μ?
- What is the *distribution* of $\bar{Y}_n - \mu$?
- How do we build *confidence intervals*?

All three questions have the same answer:

The Central Limit Theorem

Monday: the most important theorem in statistics.

Monday: The Central Limit Theorem and confidence intervals

Reading:

- A&M §3.2.3–3.2.4: CLT, convergence in distribution
- Blackwell Ch. 3: Large-sample properties (continue)

What we'll cover:

- CLT: $\sqrt{n}(\bar{Y}_n - \mu)/\sigma \xrightarrow{d} N(0, 1)$
- Asymptotic normality of MLE
- Standard errors and confidence intervals
- Connection to Fisher information from Monday's lecture

Second midterm: Wednesday, March 12.