ECMA31000: Introduction to Empirical Analysis

Asymptotics I

Joe Hardwick

University of Chicago

Autumn 2021

Outline

- This week:
  - Uses of Asymptotic Theory
  - Convergence almost surely
  - Convergence in probability
  - Convergence in r-th moment
  - Convergence in distribution
  - Relations between modes of convergence

Use of Asymptotic Theory: Estimation

- Example: Want to learn population mean wage μ.
  - Observe an iid sample of wages $\{X_i\}_{i=1}^{n}$.
  - Form an estimator $S_n\big(\{X_i\}_{i=1}^{n}\big) = \frac{1}{n}\sum_{i=1}^{n} X_i := \bar{X}_n$ which estimates μ using sample $\{X_i\}_{i=1}^{n}$.
  - In what sense is S_n is close to μ? How does this depend on n?
- Two types of results:
  - Finite sample bounds (explicit dependence on n)
  - Asymptotic results (what would happen if we could keep increasing sample size: $n \to \infty$).
- Asymptotic results often easier to derive, but our sample is only ever finite.
- Takeaway: Asymptotic properties are an approximation to finite sample properties (but not always a good one).

Use of Asymptotic Theory: Testing

- Suppose a randomly drawn height, X, is distributed according to P_X, where

$$\mathrm{E}(X) = \mu \text{ (unknown)}$$
$$Var(X) = \sigma^2 \text{ (unknown)}$$

- Suppose we wish to test $H_0 : \mu = 180cm$ vs. $H_1 : \mu > 180cm$.
- Let

$$T_n = \frac{\sqrt{n}\left(\bar{X}_n - 180\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}}.$$

- Test function

$$\phi_n(T_n) = \mathbf{1}\left(T_n > t_{n-1,1-\alpha}\right).$$

  where $t_{n-1,1-\alpha}$ is $1-\alpha$ quantile of t_{n-1} distribution.
- Probability of rejecting H_0 is known for each μ if we also know $X \sim Normal$.

Use of Asymptotic Theory: Testing

- Difficult (or impossible) to compute if P_X non-normal.
- But: If $T_n \overset{d}{\approx} \mathcal{N}(0,1)$, and $\mu = 180cm$ (H_0 true), then

$$\begin{aligned} \mathrm{P}\left(T_n > t_{n-1,1-\alpha}\right) &\approx Prob\left(\mathcal{N}(0,1) > z_{1-\alpha}\right) \\ &= 1 - \Phi\left(z_{1-\alpha}\right) \\ &= \alpha, \end{aligned}$$

  where $z_{1-\alpha}$ is the $1-\alpha$ quantile of $\mathcal{N}(0,1)$, because

$$t_{n-1,1-\alpha} \to z_{1-\alpha}$$

  for each α.
- Can approximate the rejection probability of ϕ_n by considering simpler experiment using $\mathcal{N}(0,1)$ random variable in place of T_n and $z_{1-\alpha}$ in place of $t_{n-1,1-\alpha}$.

Convergence almost surely

- We say a sequence of real numbers $\{x_n\}_{n\geq 1}$ converges to $x \in \mathbb{R}$ ($x_n \to x$) if, for each $\epsilon > 0$, there exists n_ϵ such that

$$|x_n - x| \leq \epsilon$$

  for all $n > n_\epsilon$.
- Let $\{X_n\}_{n\geq 1}$ and X be random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathrm{P})$.
- If $\omega \in \Omega$ is fixed, $\{X_n(\omega)\}_{n\geq 1}$ forms a sequence of reals.
- $X_n \to X$ pointwise if $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$.
- $X_n \overset{a.s.}{\to} X$ ("X_n converges almost surely to X") if $\mathrm{P}(\{\omega : X_n(\omega) \to X(\omega)\}) = 1$.

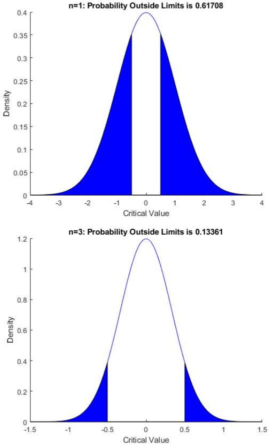
Example

- Probability Space: $([0,1], \mathcal{F}, \mathrm{P})$ where $\mathrm{P}(\{0\}) = 0$.
- Sequence

$$X_n(\omega) = 2^n \mathbf{1}\left(\omega \le \frac{1}{n}\right).$$

Example: Normal distributions

- Suppose $\{X_n\}_{n\geq 1}$ is a sequence such that $X_n \sim \mathcal{N}\left(0, \frac{1}{n^2}\right)$:



- As $n \to \infty$, a larger proportion of the density falls in the interval $\left(-\frac{1}{2}, \frac{1}{2}\right)$, so

$$\mathrm{P}\left(|X_n - 0| > \frac{1}{2}\right) \to 0.$$

Convergence in Probability

- X_n converges in probability to X ($X_n \xrightarrow{p} X$) if $\forall \epsilon > 0$:

$$\mathrm{P}\left(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}\right) \to 0.$$

- Example: Let

$$X_n = \begin{cases} 2^n & \text{with probability } \frac{1}{n} \\ 0 & \text{with probability } 1 - \frac{1}{n} \end{cases}.$$

Example: Weak LLN

- Suppose $\{X_i\}_{i\geq 1}$ are iid with $\mathrm{E}(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Then $\bar{X}_n \xrightarrow{p} \mu$.

Proof.

Using Chebyshev's inequality, for any $\epsilon > 0$

$$\begin{aligned}
\mathrm{P}\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) &\leq \frac{\mathrm{E}\left(\left|\bar{X}_n - \mu\right|^2\right)}{\epsilon^2} \\
&= \frac{\mathrm{E}\left(\sum_i (X_i - \mu)^2 + \sum_i \sum_{j\neq i} (X_i - \mu)(X_j - \mu)\right)}{n^2\epsilon^2} \\
&= \frac{n Var(X_i)}{n^2\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0,
\end{aligned}$$

where 2nd equality follows because X_i are iid and expectation is linear. Hence $\bar{X}_n \xrightarrow{p} \mu$. $\square$

Example: Estimating the CDF

- Rather than estimating a feature of F_X, we can estimate the entire distrbution:
- Let $\{X_i\}_{i\geq 1}$ be an iid sample drawn from F_X.
- Define the empirical distribution of F_X by

$$\hat{F}_X(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x).$$

- $\hat{F}_X(x)$ is just the proportion of observations less than x, and is an estimate of

$$F_X(x) = \mathrm{P}(X \leq x).$$

Example: Estimating the CDF

- $\{\mathbf{1}(X_i \leq x)\}_{i \geq 1}$ is an independent sequence because the X_i are independent.
- Also identically distributed:

$$\mathrm{P}(\mathbf{1}(X_i \leq x) = 1) = \mathrm{P}(X_i \leq x) = F_X(x).$$

- Since

$$\begin{aligned} Var(\mathbf{1}(X_i \leq x)) &= \mathrm{E}\left(\mathbf{1}(X_i \leq x)^2\right) - \mathrm{E}(\mathbf{1}(X_i \leq x))^2 \\ &= F_X(x)(1 - F_X(x)) < \infty, \end{aligned}$$

  we conclude that $\hat{F}_n(x) \xrightarrow{p} F(x)$.
- (Glivenko-Cantelli) Can in fact show that:

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \xrightarrow{p} 0.$$

Questions?

Convergence a.s.

- Recall: $X_n \stackrel{a.s.}{\to} X$ if $\mathrm{P}\left(\left\{\omega : X_n\left(\omega\right) \to X\left(\omega\right)\right\}\right) = 1$.
- Suppose ω is such that $X_n\left(\omega\right) \to X\left(\omega\right)$. Then for any $\epsilon > 0$,

$$\left|X_n\left(\omega\right) - X\left(\omega\right)\right| \leq \epsilon$$

  for $n > N\left(\epsilon, \omega\right)$. For such ω, $\left|X_n\left(\omega\right) - X\left(\omega\right)\right| > \epsilon$ only finitely many times.
- Thus, the set of ω for which $X_n\left(\omega\right) \to X\left(\omega\right)$ is the set of ω such that for any $\epsilon > 0$,

$$\left|X_n\left(\omega\right) - X\left(\omega\right)\right| > \epsilon$$

  finitely many times.
- Now note:

$$A_k := \cup_{n \geq k}\left\{\omega : \left|X_n\left(\omega\right) - X\left(\omega\right)\right| > \epsilon\right\}$$

  is the set of ω such that $\left|X_n\left(\omega\right) - X\left(\omega\right)\right| > \epsilon$ for some $n \geq k$.

Convergence a.s.

- Therefore,

$$\cap_{k \geq 1} \cup_{n \geq k} \{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}$$

is the set of ω such that $|X_n(\omega) - X(\omega)| > \epsilon$ for some n, no matter how large k gets.

- This is equivalent to saying $|X_n(\omega) - X(\omega)| > \epsilon$ <u>infinitely often</u>.
- Since $X_n \stackrel{a.s.}{\to} X$,

$$\mathrm{P}(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon \text{ infinitely often}\}) = 0,$$

or

$$\mathrm{P}(\cap_{k \geq 1} A_k) = 0.$$

Convergence a.s.

- By a similar argument to Exercise 2(b) of problem set 1:

$$\mathrm{P}\left(\cap_{k\geq 1}A_k\right) = \lim_{K\to\infty}\mathrm{P}\left(\cap_{k=1}^{K}A_k\right).$$

  Finally, notice that

$$A_k = \cup_{n\geq k}\left\{\omega : \left|X_n\left(\omega\right) - X\left(\omega\right)\right| > \epsilon\right\}$$

  is a decreasing sequence of sets: $A_k \supset A_{k+1} \supset A_{k+2}\ldots$, so $\cap_{k=1}^{K}A_k = A_K$.
- In conclusion: $X_n \overset{a.s.}{\to} X$ iff

$$\lim_{K\to\infty}\mathrm{P}\left(\cup_{n\geq K}\left\{\omega : \left|X_n\left(\omega\right) - X\left(\omega\right)\right| > \epsilon\right\}\right) = 0.$$

Convergence a.s.

- Two equivalent definitions: $X_n \stackrel{a.s.}{\to} X$ if

$$\mathrm{P}\left(\{\omega : X_n(\omega) \to X(\omega)\}\right) = 1$$

or

$$\forall \epsilon > 0,\ \lim_{n \to \infty} \mathrm{P}\left(\cup_{k \geq n} \{\omega : |X_k(\omega) - X(\omega)| > \epsilon\}\right) = 0$$

- First definition not always helpful: What if we don't know the underlying probability space?
- 2nd definition doesn't require us to know which ω converge, just need to show overall fraction of ω converges to 0.
- For example, consider

$$X_n = \begin{cases} 2^n & \text{with prob. } \frac{1}{n^2} \\ 0 & \text{with prob. } 1 - \frac{1}{n^2}. \end{cases}$$

Example

- Showing that $X_n \xrightarrow{a.s.} 0$ much easier with 2nd definition:
- Let $\epsilon > 0$ be given. Then

$$
\begin{aligned}
\mathrm{P}\left(\cup_{k\geq n}\{|X_k - 0| > \epsilon\}\right) &\leq \sum_{k=n}^{\infty} \mathrm{P}\left(|X_k - 0| > \epsilon\right) \\
&\leq \sum_{k=n}^{\infty} \mathrm{P}\left(X_k = 2^k\right) \\
&= \sum_{k=n}^{\infty} \frac{1}{k^2} \\
&= \sum_{k=1}^{\infty} \frac{1}{k^2} - \sum_{k=1}^{n-1} \frac{1}{k^2} \to 0
\end{aligned}
$$

as $n \to \infty$.

Convergence a.s. $\implies$ Convergence in Prob.

Theorem

Suppose $X_n \stackrel{a.s.}{\to} X$. Then $X_n \stackrel{p}{\to} X$.

Proof.

Note that $|X_K(\omega) - X(\omega)| > \epsilon$ implies that $|X_n(\omega) - X(\omega)| > \epsilon$ for some $n \geq K$:

$$\{\omega : |X_K(\omega) - X(\omega)| > \epsilon\} \subset \cup_{n \geq K} \{\omega : |X_n(\omega) - X(\omega)| > \epsilon\},$$

and so

$$\mathrm{P}(\{\omega : |X_K(\omega) - X(\omega)| > \epsilon\}) \leq \mathrm{P}(\cup_{n \geq K} \{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}).$$

Therefore, if $X_n \stackrel{a.s.}{\to} X$, the RHS converges to 0 as $K \to \infty$, so

$$\lim_{K \to \infty} \mathrm{P}(\{\omega : |X_K(\omega) - X(\omega)| > \epsilon\}) \leq 0.$$

Since probabilities are non-negative, conclude $X_n \stackrel{p}{\to} X$. $\square$

Difference between $\xrightarrow{a.s.}$ and $\xrightarrow{p}$

- Convergence almost surely requires that for almost all ω, and any $\epsilon > 0$,
$$|X_n(\omega) - X(\omega)| > \epsilon$$
for finitely many n.
- Convergence in probability merely requires that the fraction of ω such that $|X_n(\omega) - X(\omega)| > \epsilon$ converges to 0 as $n \to \infty$. The actual ω for which the condition holds can change with n as long as the overall fraction decreases to 0.
- Because of this, $X_n(\omega) \to X(\omega)$ is not required for any ω, as next example demonstrates.

Counterexample: $\xrightarrow{p} \not\Rightarrow \xrightarrow{a.s.}$

- Consider the space $\left([0,1], \mathcal{B}([0,1]), \mathrm{P}\right)$ and define:

$$X_1(\omega) = \begin{cases} 1 & \text{if } \omega \in [0,1] \\ 0 & \text{otherwise} \end{cases};$$

$$X_2(\omega) = \begin{cases} 1 & \text{if } \omega \in \left[0, \frac{1}{2}\right] \\ 0 & \text{otherwise} \end{cases}; \quad X_3(\omega) = \begin{cases} 1 & \text{if } \omega \in \left[\frac{1}{2}, 1\right] \\ 0 & \text{otherwise} \end{cases};$$

$$X_4(\omega) = \begin{cases} 1 & \text{if } \omega \in \left[0, \frac{1}{3}\right] \\ 0 & \text{otherwise} \end{cases}; \quad X_5(\omega) = \begin{cases} 1 & \text{if } \omega \in \left[\frac{1}{3}, \frac{2}{3}\right] \\ 0 & \text{otherwise} \end{cases}; \quad \dots$$

- If $[a,b]$ is an interval contained in $[0,1]$, we set $\mathrm{P}([a,b]) = b - a$.
- Then, for example, $\mathrm{P}_{X_1}(X_1 = 1) = 1$ and $\mathrm{P}_{X_5}(X_5 = 1) = \frac{1}{3}$.

Counterexample: $\xrightarrow{p} \not\Rightarrow \xrightarrow{a.s.}$

- As n grows large, $\mathrm{P}\left(\{\omega : X_n(\omega) = 0\}\right) \to 1$. Thus for any $\epsilon > 0$
$$\lim_{n\to\infty} \mathrm{P}\left(\{\omega : |X_n(\omega) - 0| > \epsilon\}\right) = 0,$$
and so $X_n \xrightarrow{p} 0$.

- Notice that the elements ω which satisfy the condition $|X_n(\omega) - 0| > \epsilon$ change with n.

- For every element $\omega \in [0,1]$, $X_n(\omega) = 1$ infinitely many times. So:
$$\mathrm{P}\left(\{\omega : X_n(\omega) \to 0\}\right) = 0.$$

Questions?

Difference between $\stackrel{a.s.}{\rightarrow}$ and $\stackrel{p}{\rightarrow}$

- Another way to describe difference: $X_n \stackrel{a.s.}{\rightarrow} X$ if

$$\forall \epsilon > 0, \ \lim_{n \to \infty} \mathrm{P}\left(\cup_{k \geq n} \{\omega : |X_k(\omega) - X(\omega)| > \epsilon\}\right) = 0.$$

- P refers to the probability measure on the underlying probability space, but the condition inside says "at least one X_k $(k \geq n)$ is far from X".
- So we can rewrite this using the distribution $P_{\{X_k\}_{k \geq n}, X}$ induced by $\{X_k\}_{k \geq n}, X$:

$$\forall \epsilon > 0, \ \lim_{n \to \infty} \mathrm{P}_{\{X_k\}_{k \geq n}, X}\left(\cup_{k \geq n} \{|X_k - X| > \epsilon\}\right) = 0.$$

  This no longer requires us to verify statements pointwise in $\omega \in \Omega$. It is a requirement on the joint distribution of $\left(\{X_k\}_{k \geq n}, X\right)$, as $n \to \infty$.

Difference between $\stackrel{a.s.}{\to}$ and $\stackrel{p}{\to}$

- Convergence in probability is simpler: $X_n \stackrel{p}{\to} X$ if

$$\forall \epsilon > 0, \lim_{n\to\infty} \mathrm{P}\left(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}\right) = 0.$$

- This can be rewritten as

$$\forall \epsilon > 0, \lim_{n\to\infty} \mathrm{P}_{X_n,X}\left(\{|X_n - X| > \epsilon\}\right) = 0.$$

- It is a requirement on the (bivariate) joint distribution of (X_n, X) as $n \to \infty$.

Example

- Consider again

$$X_n = \begin{cases} 2^n & \text{with probability } \frac{1}{n} \\ 0 & \text{with probability } 1 - \frac{1}{n} \end{cases}.$$

- Previously established $X_n \xrightarrow{p} 0$ always, but $X_n \xrightarrow{a.s.} 0$ only for some underlying probability spaces.
  - Note: Since $X_n \xrightarrow{p} 0$, if $X_n \xrightarrow{a.s.} X$ then $X = 0$ (almost surely).
- Can also break $X_n \xrightarrow{a.s.} 0$ by putting conditions on $\mathrm{P}_{\{X_k\}_{k \geq n}, X}$:
- Suppose $\{X_k\}_{k \geq n}$ is an independent sequence of random variables.

Example

- Adding this condition does nothing to $P_{X_n,X}$ for any n ($X_n \xrightarrow{p} 0$), but $X_n \overset{a.s.}{\nrightarrow} 0$:

$$
\begin{aligned}
&\mathrm{P}_{\{X_k\}_{k\geq n},X}\left(\cup_{k\geq n}\{|X_k - X| > \epsilon\}\right)\\
=&1 - \mathrm{P}_{\{X_k\}_{k\geq n}}\left(\cap_{k\geq n}\{|X_k - 0| \leq \epsilon\}\right)\\
=&1 - \Pi_{k=n}^{\infty}\mathrm{P}_{X_k}\left(X_k = 0\right)\\
=&1 - \Pi_{k=n}^{\infty}\left(\frac{k-1}{k}\right)\\
=&1 - \lim_{N\to\infty}\Pi_{k=n}^{N}\left(\frac{k-1}{k}\right)\\
=&1 - \lim_{N\to\infty}\left(\frac{n-1}{N}\right) = 1
\end{aligned}
$$

for all n. So we do not get convergence *a.s.*!

Questions?

Convergence in r-th moment

- Let $\{X_n\}_{n\geq 1}, X$ be random variables on some probability space $(\Omega, \mathcal{F}, \mathrm{P})$.
- X_n <u>converges in $r-$th mean to X</u> $\left(X_n \stackrel{r-\text{th}}{\to} X\right)$ for some $r > 0$ if
$$\mathrm{E}\left(|X_n - X|^r\right) \to 0.$$
- Follows from Chebyshev's inequality that $X_n \stackrel{r-\text{th}}{\to} X$ implies $X_n \stackrel{p}{\to} X$:
$$\mathrm{P}\left(|X_n - X| > \epsilon\right) \leq \frac{\mathrm{E}\left(|X_n - X|^r\right)}{\epsilon^r} \to 0.$$

Convergence in r-th moment

- Converse is false. Consider again:

$$X_n = \begin{cases} 2^n & \text{with prob. } \frac{1}{n^2}; \\ 0 & \text{with prob. } 1 - \frac{1}{n^2}. \end{cases}$$

Then $X_n \xrightarrow{a.s.} 0$ ($\implies X_n \xrightarrow{p} 0$), but

$$\mathrm{E}\left(|X_n - 0|^r\right) = \frac{2^{nr}}{n^2} \to \infty.$$

- So $\xrightarrow{a.s.}$ does not imply $\xrightarrow{r-\text{th}}$ for any $r > 0$.
- Note: In this example X_n is unbounded. $\xrightarrow{a.s.}$ <u>does</u> imply $\xrightarrow{r-\text{th}}$ under additional conditions, e.g. dominated/bounded convergence theorem.

Convergence in r-th moment

- To show $\overset{r-\text{th}}{\rightarrow}$ does not imply $\overset{a.s.}{\rightarrow}$, consider a function $f(n)$ satisfying $f(n) \in [0,1]$ for all n and $\lim_{n\to\infty} f(n) = 0$. Let

$$X_n = \begin{cases} 1 & \text{with prob. } f(n); \\ 0 & \text{with prob. } 1 - f(n). \end{cases}$$

- We have:

$$\mathrm{E}\left(|X_n - 0|^r\right) = f(n) \to 0,$$

  but (see e.g. slides 21+22) X_n does not necessarily converge a.s. to 0.

Convergence of Random Vectors

- A sequence of $(K \times 1)$ random vectors $X_n \xrightarrow{p} X$ if $\forall \epsilon > 0$:

$$\mathrm{P}\left(\|X_n - X\| > \epsilon\right) \to 0$$

as $n \to \infty$, where $\|\cdot\|$ is the euclidean norm on $\mathbb{R}^K$:

$$\|X\| = \sqrt{\sum_{i=1}^{K} X_i^2},$$

where X_i is the i-th component of X.

- $X_n \overset{a.s}{\to} X$ if:

$$\mathrm{P}\left(\{\omega : \|X_n(\omega) - X(\omega)\| \to 0\}\right) = 1.$$

- $X_n \overset{\text{r-th}}{\to} X$ if

$$\lim_{n\to\infty} \mathrm{E}\left(\|X_n - X\|^r\right) = 0.$$

Convergence of Random Vectors

1. $X_n \xrightarrow{p} X \iff X_{n,i} \xrightarrow{p} X_i$ for $i = 1, ..., k$;
2. $X_n \xrightarrow{a.s.} X \iff X_{n,i} \xrightarrow{a.s.} X_i$ for $i = 1, ..., k$;
3. $X_n \xrightarrow{r-th} X \iff X_{n,i} \xrightarrow{r-th} X_i$ for $i = 1, ..., k$.

Proof.

Part 1: Note $|X_{n,i} - X_i| \leq \|X_n - X\| \leq \sqrt{K} \max_{i \leq K} |X_{n,i} - X_i|$.
If $X_{n,i} \xrightarrow{p} X_i$ for all i, then

$$\begin{aligned}
\mathrm{P}\left(\|X_n - X\| > \epsilon\right) &\leq \mathrm{P}\left(\sqrt{K} \max_i |X_{n,i} - X_i| > \epsilon\right) \\
&= \mathrm{P}\left(\cup_{i=1}^K \left\{\sqrt{K}\,|X_{n,i} - X_i| > \epsilon\right\}\right) \\
&\leq \sum_{i=1}^K \mathrm{P}\left(|X_{n,i} - X_i| > \epsilon/\sqrt{K}\right) \to 0.
\end{aligned}$$

while if $X_n \xrightarrow{p} X$, $\mathrm{P}\left(|X_{n,i} - X_i| > \epsilon\right) \leq \mathrm{P}\left(\|X_n - X\| > \epsilon\right) \to 0$. $\square$

Convergence in Distribution

- Weakest notion of convergence (implied by $\xrightarrow{p}$).
- Sequence of random variables X_n converges in distribution to X $\left(X_n \xrightarrow{d} X\right)$ if
$$F_{X_n}(x) \to F_X(x)$$
for all x such that F_X is continuous at x.
- Why only at continuity points of F_X?
  - For X continuous, F_X continuous everywhere, so no problem.
  - For X discrete, F_X is only right-continuous.
  - Take $X_n = \frac{1}{n}$. Clearly $X_n \to 0$, but $\forall n$, $F_{X_n}(0) = 0 \neq 1 = F_X(0)$.
  - Since $X_n \to X$ in ordinary sense, should expect $X_n \xrightarrow{d} X$.
  - 0 is the only point at which $F_{X_n} \to F_X$ fails, and is the only discontinuity point of F_X.

Convergence in Distribution

- In contrast to $\xrightarrow{p}, \xrightarrow{a.s.}, \xrightarrow{\text{r-th}}$, $\{X_n\}_{n\geq 1}, X$ don't need to be defined on same probability space.
- Intuitively: $\xrightarrow{d}$ merely requires distribution of X_n converges to distribution of X.
    - For this reason, if F_X is known (e.g. standard normal), it is common to see $X_n \xrightarrow{d} \mathcal{N}(0, 1)$.
- Equivalently, $X_n \xrightarrow{d} X$ iff either

1. $\mathrm{E}(g(X_n)) \to \mathrm{E}(g(X))$ $\forall$ bounded continuous functions g.
2. $\mathrm{E}(g(X_n)) \to \mathrm{E}(g(X))$ $\forall$ bounded lipschitz functions g.

- (g is Lipschitz if $\exists K$ such that $|g(x) - g(y)| \leq K|x - y|$).

Questions?

$X_n \xrightarrow{p} X$ implies $X_n \xrightarrow{d} X$

Proof.

Suppose $X_n \xrightarrow{p} X$. We will use equivalent condition 2. to show $\xrightarrow{d}$. Let g be any bounded Lipschitz function:

$$|g(x) - g(y)| \leq K|x - y|; \qquad |g(x)| \leq B.$$

We need to show

$$\mathrm{E}(g(X_n)) \to \mathrm{E}(g(X)).$$

First note that for any $\epsilon > 0$:

$$\begin{aligned}
&|\mathrm{E}(g(X_n) - g(X))| \\
\leq &\mathrm{E}(|g(X_n) - g(X)|) \\
= &\mathrm{E}\left(|g(X_n) - g(X)| \mathbf{1}_{|X_n - X| > \epsilon}\right) + \mathrm{E}\left(|g(X_n) - g(X)| \mathbf{1}_{|X_n - X| \leq \epsilon}\right) \\
\leq &\underbrace{2B\mathrm{P}(|X_n - X| > \epsilon)}_{(1)} + \underbrace{K\mathrm{E}\left(|X_n - X| \mathbf{1}_{|X_n - X| \leq \epsilon}\right)}_{(2)}.
\end{aligned}$$

$X_n \xrightarrow{p} X$ implies $X_n \xrightarrow{d} X$

Proof.

Notice that

$$(1) = 2B\mathrm{P}\left(|X_n - X| > \epsilon\right) \to 0$$

because $X_n \xrightarrow{p} X$. Also note that

$$(2) = K\mathrm{E}\left(|X_n - X| \mathbf{1}_{|X_n - X| \leq \epsilon}\right) \leq K\epsilon.$$

In summary:

$$\left|\mathrm{E}\left(g\left(X_n\right) - g\left(X\right)\right)\right| \leq o(1) + K\epsilon,$$

where notation "$o(1)$" means a sequence converging to 0. Since ϵ can be chosen arbitrarily small, result follows. $\square$

$X_n \xrightarrow{d} X$ does not imply $X_n \xrightarrow{p} X$

- First note that $X_n \xrightarrow{d} X$ can occur even if random variables are not defined on the same probability space.
- For a more concrete example, let

$$X_n = Z = \begin{cases} 1 & \text{with probability } \frac{1}{2}; \\ -1 & \text{with probability } \frac{1}{2}; \end{cases}$$
$$X = -Z.$$

Then $X_n \xrightarrow{d} X$ since Z has the same distribution as $-Z$, but, for ϵ small,

$$\mathrm{P}\left(|X_n - X| > \epsilon\right) = \mathrm{P}\left(2\,|Z| > \epsilon\right) = 1$$

for all n, so $X_n \overset{p}{\nrightarrow} X$.

Exception! $X_n \xrightarrow{d} c$ implies $X_n \xrightarrow{p} c$

- If $X_n \xrightarrow{d} c$ for some constant c, then $X_n \xrightarrow{p} c$:

Proof.

Suppose $X_n \xrightarrow{d} c$. This means

$$F_{X_n}(x) \to \begin{cases} 0 & x < c; \\ 1 & x > c. \end{cases}$$

Need to show $\forall \epsilon > 0$:

$$\mathrm{P}(|X_n - c| > \epsilon) \to 0.$$

Rewrite

$$\begin{aligned} \mathrm{P}(|X_n - c| > \epsilon) &= \mathrm{P}(\{X_n - c > \epsilon\} \cup \{X_n - c < -\epsilon\}) \\ &= \mathrm{P}(X_n > c + \epsilon) + \mathrm{P}(X_n < c - \epsilon) \\ &= 1 - F_{X_n}(c + \epsilon) + F_{X_n}(c - \epsilon) - \mathrm{P}(X_n = c - \epsilon) \\ &\leq 1 - F_{X_n}(c + \epsilon) + F_{X_n}(c - \epsilon) \to 1 - 1 + 0 = 0. \end{aligned}$$

$\xrightarrow{d}$ for random vectors

- Let $\{X_n\}_{n\geq 1}, X$ be $(K \times 1)$ random vectors with distribution functions $\{F_{X_n}\}_{n\geq 1}, F_X$.
- X_n converges in distribution to X if

$$F_{X_n}(x) \rightarrow F_X(x)$$

  for all x such that F_X is continuous at x.
- Show in Problem Set that $X_n \xrightarrow{d} X$ implies $X_{n,i} \xrightarrow{d} X_i$ for $i = 1, \ldots, K$.

$\xrightarrow{d}$ for random vectors

- Earlier we saw that for random vectors $X_n \xrightarrow{a.s./\,p\,/\,\text{r-th}} X$, iff the components $X_{n,i} \xrightarrow{a.s./\,p\,/\,\text{r-th}} X_i$.
- NOT true for $\xrightarrow{d}$: e.g. Suppose for all n:

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} = \begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Now let

$$\begin{pmatrix} Z \\ W \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

Clearly $X_n \xrightarrow{d} Z$ and $Y_n \xrightarrow{d} W$ (since X, Y, Z, W all have <u>marginal</u> distribution $\mathcal{N}(0,1)$) but

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ Y \end{pmatrix} \overset{d}{\neq} \begin{pmatrix} Z \\ W \end{pmatrix}.$$

$\xrightarrow{d}$ for random vectors

- Takeaway: We can't simply define $\xrightarrow{d}$ for vectors by applying the definition for random variables to each component, because the marginal distributions do not contain all the information of the joint distribution.
- Exception: Suppose random vectors $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ for some constant vector c. Then:

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}.$$

- We will use this fact later to find the asymptotic distribution of OLS/IV estimators.

Cramér-Wold Device

- Although convergence of marginals doesn't imply joint convergence, we can express $\xrightarrow{d}$ for vectors in terms of $\xrightarrow{d}$ for linear combinations of elements:

Theorem

(Cramér–Wold) Let $\{X_n\}_{n\geq 1}, X$ be $(K \times 1)$ random vectors. Then:

$$X_n \xrightarrow{d} X \iff t'X_n \xrightarrow{d} t'X \text{ for all } t \in \mathbb{R}^k.$$

Continuous Mapping Theorem

- Let $g : \mathbb{R}^k \to \mathbb{R}^n$ be a function that is continuous on a set $S \subset \mathbb{R}^k$ with $\mathrm{P}(X \in S) = 1$. Then the following hold:

$$X_n \xrightarrow{a.s.} X \implies g(X_n) \xrightarrow{a.s.} g(X);$$
$$X_n \xrightarrow{p} X \implies g(X_n) \xrightarrow{p} g(X);$$
$$X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X).$$

- The theorem doesn't hold for $\xrightarrow{\text{r-th}}$: Take

$$X_n = \begin{cases} n & \text{with probability } \frac{1}{n^2}; \\ 0 & \text{with probability } 1 - \frac{1}{n^2}. \end{cases}$$

Letting $g(x) = x^2$, we see $\mathrm{E}(|X_n - 0|) = \frac{1}{n} \to 0$ but, for all n:

$$\mathrm{E}(|g(X_n) - g(0)|) = \mathrm{E}\left(\left|X_n^2 - 0\right|\right) = 1.$$

Continuous Mapping theorem

- It is important that $\mathrm{P}(X \in S) = 1$. To see this, suppose

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}.$$

- Consider the continuous function $g : \mathbb{R}^2 \to \mathbb{R}$ given by $g(x, y) = \frac{x}{y}$. g is continuous except at points $(x, 0) \in \mathbb{R}^2$. Set

$$S = \mathbb{R}^2 \setminus \{(x, 0) : x \in \mathbb{R}\}.$$

  If $c \neq 0$, then $\mathrm{P}((X, c) \in S) = 1$, and the CMT gives

$$g(X_n, Y_n) = \frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c} = g(X, c).$$

  If $c = 0$, then $\mathrm{P}((X, c) \in S) = 0$. This would lead to the nonsensical result

$$\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{0}.$$

Proofs

- $\stackrel{a.s.}{\to}$: $X_n \stackrel{a.s.}{\to} X$ means $\mathrm{P}\left(\omega : X_n(\omega) \to X(\omega)\right) = 1$. Since $\mathrm{P}\left(\omega : g \text{ is continuous at } X(\omega)\right) = 1$, it follows that

$$\mathrm{P}\left(\omega : X_n(\omega) \to X(\omega) \text{ and } g \text{ is continuous at } X(\omega)\right) = 1.$$

It follows from ordinary analysis that

$$\mathrm{P}\left(\omega : g\left(X_n(\omega)\right) \to g\left(X(\omega)\right)\right) = 1.$$

Proofs

- $\xrightarrow{p}$: Want to show that $\mathrm{P}\left(\left|g\left(X_n\right) - g\left(X\right)\right| > \epsilon\right) \to 0$ for any $\epsilon > 0$, and we know that $\mathrm{P}\left(\left|X_n - X\right| > \delta\right) \to 0$ for any $\delta > 0$.
- Given $\epsilon > 0$, define for each $\delta > 0$:

$$B_\delta = \left\{x \in \mathbb{R}^k | \exists y \in \mathbb{R}^k : |x - y| \leq \delta \text{ and } \left|g\left(x\right) - g\left(y\right)\right| > \epsilon\right\}.$$

- These are points where $|X - X_n|$ small *doesn't* imply $\left|g\left(X_n\right) - g\left(X\right)\right|$ is small.

Proofs

- Note if $X \notin B^{\delta}$ then $|g(X) - g(y)| \leq \epsilon$ for all y such that $|X - y| \leq \delta$.
- If $X \notin B^{\delta}$, $|g(X) - g(X_n)| \leq \epsilon$ when $|X - X_n| \leq \delta$. So:

$$\begin{aligned}
\mathrm{P}(|g(X) - g(X_n)| > \epsilon) &\leq \mathrm{P}(|X - X_n| > \delta) \\
&\quad + \mathrm{P}\left(\{|X - X_n| \leq \delta\} \cap \left\{X \in B^{\delta}\right\}\right) \\
&\leq \mathrm{P}(|X - X_n| > \delta) + \mathrm{P}\left(X \in B^{\delta}\right) \\
&= \mathrm{P}(|X - X_n| > \delta) + \mathrm{P}\left(X \in B^{\delta} \cap S\right),
\end{aligned}$$

since

$$\begin{aligned}
\mathrm{P}\left(X \in B^{\delta} \cap S\right) &= \mathrm{P}\left(X \in B^{\delta}\right) + \mathrm{P}(X \in S) - \mathrm{P}\left(X \in B^{\delta} \cup S\right) \\
&= \mathrm{P}\left(X \in B^{\delta}\right) + 1 - 1 = \mathrm{P}\left(X \in B^{\delta}\right).
\end{aligned}$$

Proofs

- $\mathrm{P}\left(|X - X_n| > \delta\right)$ can be made arbitrarily small for any $\delta > 0$.
- Since $B^\delta \cap S$ only contains continuity points of g, $B^\delta \cap S \downarrow \emptyset$ as $\delta \downarrow 0$ because for any $x \in S$, $x \notin B^\delta$ for δ small enough.
- Let $\delta_n \downarrow 0$. We have

$$\lim_{\delta \to 0} \mathrm{P}\left(X \in B^\delta \cap S\right) = \lim_{n \to \infty} \mathrm{P}\left(X \in B^{\delta_n} \cap S\right) = \mathrm{P}\left(X \in \emptyset\right) = 0.$$

- The proof for $\xrightarrow{d}$ is omitted.

Questions?

Example: Slutsky's Theorem

Theorem

Suppose $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{p} c$ *for some constant* c*. Then:*

$$X_n + Y_n \xrightarrow{d} X + c;$$

$$X_n Y_n \xrightarrow{d} Xc;$$

$$X_n / Y_n \xrightarrow{d} X/c \textit{ provided } c \neq 0.$$

Proof.

We stated that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ implies

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}.$$

The result follows by noting that $x + y$, xy and x/y are all continuous functions of (x, y), provided $y \neq 0$ in the last case. $\square$

Example: Slutsky's Theorem

- It is important that Y_n converges in probability to a constant! If instead, for all n

$$Y_n = (-1)^n X_n \sim \mathcal{N}(0, 1),$$

then $X_n + Y_n = 0$ for odd n, and $X_n + Y_n \sim \mathcal{N}(0, 4)$ for even n. Thus,

$$F_{X_n + Y_n}(x)$$

does not converge for any $x \in \mathbb{R}$. This means $X_n + Y_n$ cannot converge in distribution.

Example: Sample Correlation

- Suppose $\{(X_i, Y_i)\}_{i\geq 1}$ is a sequence of (2×1) iid random vectors with $\mathrm{E}\left(X_i^2\right) < \infty$, $\mathrm{E}\left(Y_i^2\right) < \infty$.
- The sample correlation between X, Y is given by

$$\hat{\rho}_{XY} = \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2}}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i - \bar{X}\bar{Y}}{\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2}\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2 - \bar{Y}^2}},$$

where $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ are the sample means.

Example: Sample Correlation

- We have already seen that $\bar{X} \xrightarrow{p} \mathrm{E}(X)$ and $\bar{Y} \xrightarrow{p} \mathrm{E}(Y)$.
- Next, note that since the vectors (X_i, Y_i) are iid, the product $X_i Y_i$ is an iid sequence of random variables.
- By the weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i \xrightarrow{p} \mathrm{E}(XY);$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathrm{E}\left(X^2\right);$$

$$\frac{1}{n}\sum_{i=1}^{n} Y_i^2 \xrightarrow{p} \mathrm{E}\left(Y^2\right).$$

Example: Sample Correlation

- It follows that

$$\left(\bar{X}, \bar{Y}, \frac{1}{n}\sum_{i=1}^{n} X_i Y_i, \frac{1}{n}\sum_{i=1}^{n} X_i^2, \frac{1}{n}\sum_{i=1}^{n} Y_i^2\right)$$
$$\xrightarrow{p} \left(\mathrm{E}\left(X\right), \mathrm{E}\left(Y\right), \mathrm{E}\left(XY\right), \mathrm{E}\left(X^2\right), \mathrm{E}\left(Y^2\right)\right).$$

- Now let

$$g\left(x, y, s, t, w\right) = \frac{s - xy}{\sqrt{t - x^2}\sqrt{w - y^2}}.$$

- g is continuous at all points except where $t = x^2$ and $w = y^2$. Provided neither X nor Y are constant random variables (if they were the sample variances would also be 0!) we get

$$\mathrm{E}\left(X^2\right) > \mathrm{E}\left(X\right)^2; \qquad \mathrm{E}\left(Y^2\right) > \mathrm{E}\left(Y\right)^2,$$

so g is continuous at
$\left(\mathrm{E}\left(X\right), \mathrm{E}\left(Y\right), \mathrm{E}\left(XY\right), \mathrm{E}\left(X^2\right), \mathrm{E}\left(Y^2\right)\right)$.

Example: Sample Correlation

- It follows by the continuous mapping theorem that:

$$
\begin{aligned}
&\frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2}} \\
\xrightarrow{p}&\frac{\mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)}{\sqrt{\mathrm{E}(X^2) - \mathrm{E}(X)^2}\sqrt{\mathrm{E}(Y^2) - \mathrm{E}(Y)^2}} \\
=&\frac{Cov(X, Y)}{\sqrt{Var(X)\,Var(Y)}},
\end{aligned}
$$

which is the population correlation coefficient.

Example

- A form of the following result will appear several times when we analyse OLS/IV estimators.
- Let $A_n \in \mathbb{R}^{P \times K}$ be a sequence of matrices converging in probability to a constant matrix A.
  - This is just the same as vector convergence: Stack the columns on top of each other!
- Let B_n be a sequence of $(K \times 1)$ random vectors such that

$$B_n \xrightarrow{d} \mathcal{N}(\mu, \Sigma).$$

Then:

$$A_n B_n \xrightarrow{d} A\mathcal{N}(\mu, \Sigma) \sim \mathcal{N}\left(A\mu, A\Sigma A'\right).$$

Example

Proof.

Since the vector consisting of the columns of A_n, denoted $vec\left(A_n\right)$ converges in probability to $vec\left(A\right)$, a constant vector, we obtain

$$\begin{pmatrix} B_n \\ vec\left(A_n\right) \end{pmatrix} \stackrel{d}{\rightarrow} \begin{pmatrix} \mathcal{N}\left(\mu, \Sigma\right) \\ vec\left(A\right) \end{pmatrix}.$$

By the continuous mapping theorem,

$$A_n B_n \stackrel{d}{\rightarrow} A\mathcal{N}\left(\mu, \Sigma\right).$$

Since $\mathrm{E}\left(AX\right) = A\mathrm{E}\left(X\right)$, and we showed in problem set 1 that $Var\left(AX\right) = AVar\left(X\right)A'$, we conclude that

$$A\mathcal{N}\left(\mu, \Sigma\right) \sim \mathcal{N}\left(A\mu, A\Sigma A'\right)$$

because linear transformations of multivariate normals are also (multivariate) normal. $\square$

Summary of implications

Questions?