Lecture 2
Convergence theorems

1 Types of Convergence

To deal with convergence arguments in the next section, we have to introduce the concept of a *random experiment*. By definition, a random experiment is a triple $(\Omega, \mathcal{A}, P)$ where Ω denotes the sample space, $\mathcal{A}$ is some class of subsets of Ω and P is a function from $\mathcal{A}$ into $[0,1]$, i.e. $P : \mathcal{A} \to [0,1]$. Elements ω of sample space Ω are referred to as outcomes of a random experiment. In each random experiment, there is one realized outcome ω. Elements A of $\mathcal{A}$ are called events. We say that event A happens if realized outcome ω belongs to A, i.e. $\omega \in A$. Event A does not occur if the realized outcome $\omega \notin A$. Each event $A \in \mathcal{A}$ has some probability of happening. This probability is denoted by $P(A)$. Thus, function P defines a probability of events A in $\mathcal{A}$. It would be nice if we could define probability for all subsets of Ω. However, due to some technicalities, it is not always possible. Therefore we have to use only some class $\mathcal{A}$ of subsets of Ω, which may or may not contain all subsets of Ω. In this course, we will not talk about class $\mathcal{A}$ at all. It is mentioned here only for completeness.

Once we have incorporated the concept of the random experiment, we can represent random variables as functions on sample space Ω. Thus, if ξ is a random variable, then $\xi : \Omega \to \mathbb{R}$, i.e. for each realized outcome $\omega \in \Omega$, we have realization $\xi(\omega)$ of random variable ξ.

1.1 Definitions

Let $\xi_1, ..., \xi_n, ...$ be a sequence of random variables. Then, for any realized outcome $\omega \in \Omega$, we have a sequence of real numbers, $\xi_1(\omega), ..., \xi_n(\omega), ...$

Definition 1. We say that $\{\xi_n\}_{n=1}^{\infty}$ *converges* to some random variable ξ *almost surely* if $P\{\omega : \xi_n(\omega) \to \xi(\omega)\} = 1$. In this case we write $\xi_n \to \xi$ a.s.

Definition 2. We say that $\{\xi_n\}_{n=1}^{\infty}$ *converges* to ξ *in probability* if $P\{\omega : |\xi_n(\omega) - \xi(\omega)| > \varepsilon\} \to 0$ as $n \to \infty$ for any $\varepsilon > 0$. In this case we write $\xi_n \to_p \xi$.

Definition 3. We say that $\{\xi_n\}_{n=1}^{\infty}$ *converges* to ξ *in quadratic mean* if $E[|\xi_n - \xi|^2] \to 0$ as $n \to \infty$. In this case we write $\xi_n \to \xi$ in L_2.

Definition 4. We say that $\{\xi_n\}_{n=1}^{\infty}$ *converges* to ξ *in distribution* if $\lim_{n\to\infty} F_{\xi_n}(x) = F_\xi(x)$ for all $x \in \mathbb{R}$ where $F_\xi(x)$ is continuous. In this case we write $\xi_n \Rightarrow \xi$.

We will discuss the relationships between different types of convergence after we prove some inequalities.

1.2 Useful Inequalities

First, we have Markov inequality.

Theorem 5. *Let X be any nonnegative random variable such that $E[X]$ exists. Then for any $t > 0$, we have $P\{X \geq t\} \leq E[X]/t$.*

Proof. Since X is nonnegative,

$$\begin{aligned} E[X] &= \int_0^\infty xf(x)dx \\ &= \int_0^t xf(x)dx + \int_t^\infty xf(x)dx \\ &\geq \int_t^\infty xf(x)dx \\ &\geq t\int_t^\infty f(x)dx \\ &= tP\{X \geq t\} \end{aligned}$$

where f denotes the pdf of X. A similar argument works for discrete random variables as well. □

From the Markov inequality we can derive the Chebyshev inequality.

Theorem 6. *For any random variable X with mean μ and any $t > 0$, we have $P\{|X - \mu| \geq t\} \leq V(X)/t^2$.*

Proof. Note that $|X - \mu| \geq t$ if and only if $|X - \mu|^2 \geq t^2$. Thus, $P\{|X - \mu| \geq t\} = P\{|X - \mu|^2 \geq t^2\}$. Since $|X - \mu|^2$ is a nonnegative random variable, $P\{|X - \mu|^2 \geq t^2\} \leq E[|X - \mu|^2]/t^2 = V(X)/t^2$ by Markov inequality. □

These two inequalities are of huge importance in statistics and probability.

1.3 Relations between different types of convergence

We can use the Markov inequality to prove that convergence in quadratic mean implies convergence in probability:

Theorem 7. *If $\xi_n \to \xi$ in L_2, then $\xi_n \to_p \xi$.*

Proof. By Markov inequality,

$$P\{|\xi_n - \xi| > \varepsilon\} = P\{|\xi_n - \xi|^2 > \varepsilon^2\} \leq E[|\xi_n - \xi|^2]/\varepsilon^2 \to 0$$

for any $\varepsilon > 0$. □

Convergence in probability implies convergence in distribution:

Theorem 8. *If $\xi_n \to_p \xi$, then $\xi_n \Rightarrow \xi$.*

Proof. Note that $\xi_n \leq x$ and $\xi > x + \varepsilon$ implies $|\xi_n - \xi| > \varepsilon$. Thus,

$$\begin{aligned} F_{\xi_n}(x) &= P\{\xi_n \leq x\} \\ &= P\{\xi_n \leq x, \xi \leq x + \varepsilon\} + P\{\xi_n \leq x, \xi > x + \varepsilon\} \\ &\leq P\{\xi \leq x + \varepsilon\} + P\{|\xi_n - \xi| > \varepsilon\} \\ &= F_\xi(x + \varepsilon) + P\{|\xi_n - \xi| > \varepsilon\}. \end{aligned}$$

for any $x \in \mathbb{R}$ and $\varepsilon > 0$. Similarly,

$$F_\xi(x - \varepsilon) \leq F_{\xi_n}(x) + P\{|\xi_n - \xi| > \varepsilon\}.$$

Thus,

$$F_\xi(x - \varepsilon) - P\{|\xi_n - \xi| > \varepsilon\} \leq F_{\xi_n}(x) \leq F_\xi(x + \varepsilon) + P\{|\xi_n - \xi| > \varepsilon\}.$$

Next, if x is a point of continuity of F_ξ, for any $\delta > 0$, there exists $\varepsilon(\delta) > 0$ such that

$$F_\xi(x + \varepsilon(\delta)) - \delta \leq F_\xi(x) \leq F_\xi(x - \varepsilon(\delta)) + \delta.$$

Therefore

$$F_\xi(x) - \delta - P\{|\xi_n - \xi| > \varepsilon(\delta)\} \leq F_{\xi_n}(x) \leq F_\xi(x) + \delta + P\{|\xi_n - \xi| > \varepsilon(\delta)\}.$$

Next, since $\xi_n \to_p \xi$,by definition:

$$\lim_n |F_{\xi_n}(x) - F_\xi(x)| \leq \delta$$

So, $F_{\xi_n}(x) \to F_\xi(x)$ as $n \to \infty$ for any $x \in \mathbb{R}$ where $F_\xi(x)$ is continuous. □

As an exercise, prove that almost sure convergence implies convergence in probability. Also, one can show that if c is some constant and $\xi_n \Rightarrow c$, then $\xi_n \to_p c$.

We have mentioned here all the correct implications. Any other implication is, in general, incorrect. Let us show, for example, that convergence in quadratic mean does not follow from convergence in probability. Let $\Omega = [0, 1]$ be a sample space. Let $\xi_n(\omega) = n$ if $\omega \in [0, 1/n]$ and 0 otherwise. Then it is obvious that $\xi_n \to_p 0$. Indeed, for any $\varepsilon \in (0, 1)$, $P\{|\xi_n - 0| > \varepsilon\} = 1/n \to 0$ as $n \to \infty$ and $P\{|\xi_n - 0| > \varepsilon\} = 0$ for any $\varepsilon \geq 1$. On the other hand, $E[|\xi_n - 0|^2] = E[|\xi_n|^2] = n^2 \cdot (1/n) = n \to \infty$.

2 Slutsky theorem and Continuous mapping theorem

Let $X, X_1, ..., X_n, ...$ and $Y, Y_1, ..., Y_n, ...$ be some random variables. Let g be some continuous function. Let c be some constant. Then

1. If $X_n \to_p X$ and $Y_n \to_p Y$, then $X_n + Y_n \to_p X + Y$ and $X_n Y_n \to_p XY$.

2. If $X_n \Rightarrow X$ and $Y \to_p c$, then $X_n + Y_n \Rightarrow X + c$ and $X_n Y_n \Rightarrow cX$.

3. If $X_n \to_p X$, then $g(X_n) \to_p g(X)$.

4. If $X_n \Rightarrow X$, then $g(X_n) \Rightarrow g(X)$

The first and second statements are known as the Slutsky theorem. The third and forth statements are known as the Continuous mapping theorem. These theorems are widely used in statistics.

Note that, in general, $X_n \Rightarrow X$ and $Y_n \Rightarrow Y$ does not imply $X_n + Y_n \Rightarrow X + Y$ or $X_n Y_n \Rightarrow XY$.

3 Law of Large Numbers

Theorem 9. *If $\{X_n\}_{n=1}^{\infty}$ is a sequence of independent and identically distributed (iid) random variables with $E[X_n] = \mu$ and $V(X) = \sigma^2 < \infty$, then $\overline{X}_n := \sum_{i=1}^n X_i/n \to \mu$ in L_2 and, thus, in probability.*

Proof. By linearity of expectation, $E[\overline{X}_n] = E[\sum_{i=1}^n X_i/n] = \sum_{i=1}^n E[X_i]/n = \mu$. Thus,

$$\begin{aligned} E[|\overline{X} - \mu|^2] &= V(\overline{X}) \\ &= V(\sum_{i=1}^n X_i)/n^2 \\ &= \sum_{i=1}^n V(X_i)/n^2 \\ &= \sigma^2/n. \end{aligned}$$

Thus, $E[|\overline{X} - \mu|^2] \to 0$ as $n \to \infty$. □

Another version of the law of large numbers is

Theorem 10. *If $\{X_n\}_{n=1}^{\infty}$ is a sequence of iid random variables with $E[X_n] = \mu$ and $E[|X_n|] < \infty$, then $\overline{X}_n \to \mu$ a.s.*

Proof. Omitted. □

4 Central Limit Theorem

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of iid random variables with mean μ and variance σ^2. We have seen already that $E[\sum_{i=1}^n (X_i - \mu)/\sqrt{n}] = 0$ and $V(\sum_{i=1}^n (X_i - \mu)/\sqrt{n}) = \sigma^2$. A much more remarkable result is the Central limit theorem:

Theorem 11. $\sum_{i=1}^n (X_i - \mu)/\sqrt{n} \Rightarrow N(0, \sigma^2)$.

In the multivariate case, if $V(X_i) = E[(X_i - E[X_i])(X_i - E[X_i])^T] = \Sigma$, then $\sum_{i=1}^n (X_i - \mu)/\sqrt{n} \Rightarrow N(0, \Sigma)$.

4.1 Example

Let the probability of a newborn being a boy be, say, 0.51. What is the probability that at least half out of 100 newborns will be boys? To answer this question, let $X_i = 1$ if i-th newborn is a boy and $X_i = 0$ otherwise. Then $X_i = 1$ with probability $p = 0.51$ and $X_i = 0$ with probability $1 - p = 0.49$. Therefore $\mu = E[X_i] = 0.51$ and $\sigma^2 = p(1-p) = 0.51 \cdot 0.49$. Moreover, $X_1, ..., X_{100}$ are independent random variables. The total number of boys equals $\sum_{i=1}^{100} X_i$. Thus

$$
\begin{aligned}
P\{\sum_{i=1}^{100} X_i/100 \geq 0.5\} &= P\{\sum_{i=1}^{100}(X_i - \mu)/100 \geq -0.01\} \\
&= P\{\sum_{i=1}^{100}(X_i - \mu)/(\sigma\sqrt{100}) \geq -0.01 \cdot \sqrt{100}/\sqrt{0.51 \cdot 0.49}\} \\
&\approx 1 - \Phi(-0.01 \cdot \sqrt{100}/\sqrt{0.51 \cdot 0.49})
\end{aligned}
$$

since $\sum_{i=1}^{n}(X_i - \mu)/(\sigma\sqrt{n}) \Rightarrow N(0,1)$ as $n \to \infty$ and, hence, $P\{\sum_{i=1}^{n}(X_i - \mu)/(\sigma\sqrt{n}) \leq x\} \to \Phi(x)$ as $n \to \infty$ for any $x \in \mathbb{R}$. Note that here we used the fact that $\Phi(x)$ is continuous at all $x \in \mathbb{R}$.

5 Delta method

Let $X_1, ..., X_n, ...$ be a sequence of iid random variables with mean μ and variance σ^2. As before, $\overline{X}_n = \sum_{i=1}^{n} X_i/n$. By the Central limit theorem $\sqrt{n}(\overline{X}_n - \mu)/\sigma \Rightarrow N(0,1)$. Suppose that $g : \mathbb{R} \to \mathbb{R}$ is some twice continuously differentiable function (i.e. it has at least two derivatives, and the second derivative is continuous). The delta method allows us to derive the limit (or, as we usually say, asymptotic) distribution of $g(\overline{X}_n)$:

Theorem 12. *If $g'(\mu) \neq 0$, then $\sqrt{n}(g(\overline{X}_n) - g(\mu))/\sigma \Rightarrow N(0, (g'(\mu))^2)$.*

Proof. By the Taylor theorem with remainder, for any realization $\overline{X}_n(\omega)$, there is some $\mu_n^\star(\omega)$ between μ and $\overline{X}_n(\omega)$ such that

$$
g(\overline{X}_n(\omega)) - g(\mu) = g'(\mu)(\overline{X}_n(\omega) - \mu) + g''(\mu_n^\star(\omega))(\overline{X}_n(\omega) - \mu)^2/2. \tag{1}
$$

Thus, we have defined a new sequence of random variables, $\{\mu_n^\star\}_{n=1}^\infty$. By the Law of large numbers, $\overline{X}_n \to_p \mu$. Since $\mu_n^\star$ is between μ and $\overline{X}_n$, $\mu_n^\star \to_p \mu$ as well. By the Continuous mapping theorem, $g''(\mu_n^\star) \to_p g''(\mu)$ since $g''(x)$ is continuous. Moreover, by the Slutsky theorem, $n^{1/4}(\overline{X}_n - \mu) = \sqrt{n}(\overline{X}_n - \mu)/n^{1/4} \Rightarrow 0$ since $\sqrt{n}(\overline{X}_n - \mu) \Rightarrow N(0,1)$ and $1/n^{1/4} \to_p 0$. Thus, $n^{1/4}(\overline{X}_n - \mu) \to_p 0$. Moreover, $\sqrt{n}(\overline{X}_n - \mu)^2 = (n^{1/4}(\overline{X}_n - \mu))^2 \to_p 0$ since $f(x) = x^2$ is a continuous function. So, by the Slutsky theorem again, $\sqrt{n}g''(\mu_n^\star(\omega))(\overline{X}_n(\omega) - \mu)^2/2 \to_p 0$. In addition, by the Central limit theorem, $\sqrt{n}g'(\mu)(\overline{X}_n(\omega) - \mu) \Rightarrow N(0, \sigma^2(g'(\mu))^2)$. Multiplying equation (1) by $\sqrt{n}$ and applying the Slutsky theorem once more yields the result. □

Note that this theorem also holds when $g'(\mu) = 0$ but in this case the asymptotic distribution will be 0 (constant), i.e. degenerate. I recommend that you remember the argument used in this theorem as it is very typical in statistics and econometrics.

The Delta method has a multidimensional extension. Let $X_1, ..., X_n, ...$ be a sequence of iid $k \times 1$ random vectors with mean μ and covariance matrix Σ. Then, by the multidimensional Central limit theorem, $\sqrt{n}(\overline{X}_n - \mu) \Rightarrow N(0, \Sigma)$. Let $g : \mathbb{R}^k \to \mathbb{R}$ be a twice continuously differentiable function. Let $\tau^2 = (\partial g(\mu)/\partial \mu)^T \Sigma (\partial g(\mu)/\partial \mu)$. Here $\partial g(\mu)/\partial \mu$ is a $k \times 1$ vector with i-th component equals $\partial g(\mu)/\partial \mu_i$. Then $\sqrt{n}(g(\overline{X}_n) - g(\mu)) \Rightarrow N(0, \tau^2)$.

5.1 Example

Let $X_1, ..., X_n, ...$ be a sequence of iid random variables with mean μ and variance σ^2. What is the limiting distribution of $(\overline{X}_n)^2$? Let $g(x) = x^2$. Then $g'(\mu) = 2\mu$. Thus, by the Delta method, $\sqrt{n}((\overline{X}_n)^2 - \mu^2) \Rightarrow N(0, 4\mu^2\sigma^2)$. Note that if $\mu = 0$, then the limit distribution is degenerate.