Lecture 8: Convergence of transformations and law of large numbers

Transformation and convergence

- Transformation is an important tool in statistics.
- If X_n converges to X in some sense, we often need to check whether $g(X_n)$ converges to $g(X)$ in the same sense.
- The continuous mapping theorem provides an answer to the question in many problems.

Theorem 1.10. Continuous mapping theorem

Let $X, X_1, X_2, \ldots$ be random k-vectors defined on a probability space and g be a measurable function from $(\mathscr{R}^k, \mathscr{B}^k)$ to $(\mathscr{R}^l, \mathscr{B}^l)$. Suppose that g is continuous a.s. P_X. Then

(i) $X_n \to_{a.s.} X$ implies $g(X_n) \to_{a.s.} g(X)$;

(ii) $X_n \to_p X$ implies $g(X_n) \to_p g(X)$;

(iii) $X_n \to_d X$ implies $g(X_n) \to_d g(X)$.

Proof

(i) can be established using a result in calculus.

(iii) follows from Theorem 1.9(i): for any bounded and continuous h, $E[h(g(X_n))] \to E[h(g(X))]$, since $h \circ g$ is bounded and continuous.

We prove (ii) for the special case of $X = c$ (a constant).
From the continuity of g, for any $\varepsilon > 0$, there is a $\delta_\varepsilon > 0$ such that

$$\|g(x) - g(c)\| < \varepsilon \quad \text{whenever } \|x - c\| < \delta_\varepsilon.$$

Hence,

$$\{\omega : \|g(X_n(\omega)) - g(c)\| < \varepsilon\} \supset \{\omega : \|X_n(\omega) - c\| < \delta_\varepsilon\}$$

and

$$P(\|g(X_n) - g(c)\| \geq \varepsilon) \leq P(\|X_n - c\| \geq \delta_\varepsilon).$$

Hence $g(X_n) \to_p g(c)$ follows from $X_n \to_p c$.

Is the previous arguement still valid when c is replaced by the random vector X in the general case?
If not, how do we fix the proof?

Example 1.30.

(i) Let $X_1, X_2, \ldots$ be random variables.
If $X_n \to_d X$, where X has the $N(0,1)$ distribution, then $X_n^2 \to_d Y$, where Y has the chi-square distribution χ_1^2.

(ii) Let (X_n, Y_n) be random 2-vectors satisfying $(X_n, Y_n) \to_d (X, Y)$, where X and Y are independent random variables having the $N(0,1)$ distribution.
Then $X_n/Y_n \to_d X/Y$, which has the Cauchy distribution $C(0,1)$.

(iii) Under the conditions in part (ii), $\max\{X_n, Y_n\} \to_d \max\{X, Y\}$, which has the c.d.f. $[\Phi(x)]^2$ ($\Phi(x)$ is the c.d.f. of $N(0,1)$).

In Example 1.30(ii) and (iii), the condition that $(X_n, Y_n) \to_d (X, Y)$ cannot be relaxed to $X_n \to_d X$ and $Y_n \to_d Y$ (exercise); i.e., we need the convergence of the joint c.d.f. of (X_n, Y_n).
This is different when $\to_d$ is replaced by $\to_p$ or $\to_{a.s.}$.

The next result, which plays an important role in statistics, establishes the convergence in distribution of $X_n + Y_n$ or $X_n Y_n$ when no information regarding the joint c.d.f. of (X_n, Y_n) is provided.

Theorem 1.11 (Slutsky's theorem)

Let $X, X_1, X_2, ..., Y_1, Y_2, ...$ be random variables on a probability space. Suppose that $X_n \to_d X$ and $Y_n \to_p c$, where c is a constant. Then

(i) $X_n + Y_n \to_d X + c$;

(ii) $Y_n X_n \to_d cX$;

(iii) $X_n / Y_n \to_d X/c$ if $c \neq 0$.

Proof

We prove (i) only.
The proofs of (ii) and (iii) are left as exercises.
Let $t \in \mathscr{R}$ and $\varepsilon > 0$ be fixed constants.
Then

$$\begin{aligned} F_{X_n+Y_n}(t) &= P(X_n + Y_n \leq t) \\ &\leq P(\{X_n + Y_n \leq t\} \cap \{|Y_n - c| < \varepsilon\}) + P(|Y_n - c| \geq \varepsilon) \\ &\leq P(X_n \leq t - c + \varepsilon) + P(|Y_n - c| \geq \varepsilon) \end{aligned}$$

Proof (continued)

Similarly,

$$F_{X_n+Y_n}(t) \geq P(X_n \leq t-c-\varepsilon) - P(|Y_n - c| \geq \varepsilon).$$

If $t-c$, $t-c+\varepsilon$, and $t-c-\varepsilon$ are continuity points of F_X, then it follows from the previous two inequalities and the hypotheses of the theorem that

$$F_X(t-c-\varepsilon) \leq \liminf_n F_{X_n+Y_n}(t) \leq \limsup_n F_{X_n+Y_n}(t) \leq F_X(t-c+\varepsilon).$$

Since ε can be arbitrary (why?),

$$\lim_{n\to\infty} F_{X_n+Y_n}(t) = F_X(t-c).$$

The result follows from $F_{X+c}(t) = F_X(t-c)$.

An application of Theorem 1.11 is given in the proof of the following important result.

Theorem 1.12

Let $X_1, X_2, \ldots$ and $Y = (Y_1, \ldots, Y_k)$ be random k-vectors satisfying

$$a_n(X_n - c) \to_d Y,$$

where $c \in \mathscr{R}^k$ and $\{a_n\}$ is a sequence of positive numbers with $\lim_{n\to\infty} a_n = \infty$.

Let g be a function from $\mathscr{R}^k$ to $\mathscr{R}$.

(i) If g is differentiable at c, then

$$a_n[g(X_n) - g(c)] \to_d [\nabla g(c)]^\tau Y,$$

where $\nabla g(x)$ denotes the k-vector of partial derivatives of g at x.

(ii) Suppose that g has continuous partial derivatives of order $m > 1$ in a neighborhood of c, with all the partial derivatives of order j, $1 \le j \le m-1$, vanishing at c, but with the mth-order partial derivatives not all vanishing at c.

Then

$$a_n^m[g(X_n) - g(c)] \to_d \frac{1}{m!} \sum_{i_1=1}^{k} \cdots \sum_{i_m=1}^{k} \frac{\partial^m g}{\partial x_{i_1} \cdots \partial x_{i_m}}\bigg|_{x=c} Y_{i_1} \cdots Y_{i_m}$$

Proof

We prove (i) only.
Let

$$Z_n = a_n[g(X_n) - g(c)] - a_n[\nabla g(c)]^\tau (X_n - c).$$

If we can show that $Z_n = o_p(1)$, then by $a_n(X_n - c) \to_d Y$, Theorem 1.9(iii), and Theorem 1.11(i), result (i) holds.
The differentiability of g at c implies that for any $\varepsilon > 0$, there is a $\delta_\varepsilon > 0$ such that

$$|g(x) - g(c) - [\nabla g(c)]^\tau (x - c)| \le \varepsilon \|x - c\|$$

whenever $\|x - c\| < \delta_\varepsilon$.
Then for a fixed $\eta > 0$,

$$P(|Z_n| \ge \eta) \le P(\|X_n - c\| \ge \delta_\varepsilon) + P(a_n \|X_n - c\| \ge \eta/\varepsilon).$$

Since $a_n \to \infty$, $a_n(X_n - c) \to_d Y$ and Theorem 1.11(ii) imply $X_n \to_p c$.
By Theorem 1.10(iii), $a_n(X_n - c) \to_d Y$ implies $a_n \|X_n - c\| \to_d \|Y\|$.
Without loss of generality, we can assume that η/ε is a continuity point of $F_{\|Y\|}$.

Proof (continued)

Then

$$\limsup_n P(|Z_n| \geq \eta) \leq \lim_{n\to\infty} P(\|X_n - c\| \geq \delta_\varepsilon) + \lim_{n\to\infty} P(a_n\|X_n - c\| \geq \eta/\varepsilon) = P(\|Y\| \geq \eta/\varepsilon).$$

$Z_n \to_p 0$ follows since ε can be arbitrary.

Remarks

- In statistics, we often need a nondegenerated limiting distribution of $a_n[g(X_n) - g(c)]$ so that probabilities involving $a_n[g(X_n) - g(c)]$ can be approximated by the c.d.f. of $[\nabla g(c)]^\tau Y$, under Theorem 1.12(i).
- When $\nabla g(c) = 0$, Theorem 1.12(i) indicates that the limiting distribution of $a_n[g(X_n) - g(c)]$ is degenerated.
  In such cases the result in Theorem 1.12(ii) may be useful.

Corollary 1.1 (the δ-method)

Assume the conditions of Theorem 1.12.
If Y has the $N_k(0,\Sigma)$ distribution, then

$$a_n[g(X_n)-g(c)] \to_d N(0,[\nabla g(c)]^\tau \Sigma \nabla g(c)).$$

Example 1.31

Let $\{X_n\}$ be a sequence of random variables satisfying

$$\sqrt{n}(X_n - c) \to_d N(0,1).$$

Consider the function $g(x) = x^2$.
If $c \neq 0$, then an application of Corollary 1.1 gives that

$$\sqrt{n}(X_n^2 - c^2) \to_d N(0,4c^2).$$

If $c = 0$, $g'(c) = 0$ but $g''(c) = 2$.
Hence, an application of Theorem 1.12(ii) gives that

$$nX_n^2 \to_d [N(0,1)]^2,$$

which has the chi-square distribution χ_1^2 (Example 1.14).
The last result can also be obtained by applying Theorem 1.10(iii).

Example: Ratio estimator

$(X_1, Y_1), \ldots, (X_n, Y_n)$ are iid bivariate random vectors with finite 2nd order momemnts

$\bar{X}_n = n^{-1}\sum_i X_i$, $\bar{Y}_n = n^{-1}\sum_i Y_i$
$\mu_x = E(X_1)$, $\mu_y = E(Y_1) \neq 0$, $\sigma_x^2 =$Var(X_1), $\sigma_y^2 =$Var(Y_1),
$\sigma_{xy} =$Cov(X_1, Y_1)

By the CLT,

$$\sqrt{n}\left(\begin{pmatrix} \bar{X}_n \\ \bar{Y}_n \end{pmatrix} - \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}\right) \to_d N_2\left(0, \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}\right)$$

By the δ-method, $g(x,y) = x/y$, $\partial g/\partial x = y^{-1}$, $\partial g/\partial y = -xy^{-2}$

$$\sqrt{n}\left(\frac{\bar{X}_n}{\bar{Y}_n} - \frac{\mu_x}{\mu_y}\right) \to_d N(0, \sigma^2)$$

$$\sigma^2 = (\mu_y^{-1}, -\mu_x\mu_y^{-2})\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}\begin{pmatrix} \mu_y^{-1} \\ -\mu_x\mu_y^{-2} \end{pmatrix} = \frac{\sigma_x^2}{\mu_y^2} - \frac{\mu_x\sigma_{xy}}{\mu_y^3} + \frac{\mu_x^2\sigma_y^2}{\mu_y^4}$$

The law of large numbers

- The law of large numbers concerns the limiting behavior of a sum of random variables.
- The weak law of large numbers (WLLN) refers to convergence in probability.
- The strong law of large numbers (SLLN) refers to a.s. convergence.
- The WLLN and SLLN play important roles in establishing consistency of estimators in large sample statistical inference.

Lemma 1.6. (Kronecker's lemma)

Let $x_n \in \mathscr{R}$, $a_n \in \mathscr{R}$, $0 < a_n \le a_{n+1}$, $n = 1, 2, ...$, and $a_n \to \infty$.
If the series $\sum_{n=1}^{\infty} x_n / a_n$ converges, then $a_n^{-1} \sum_{i=1}^{n} x_i \to 0$.

Our first result gives the WLLN and SLLN for a sequence of independent and identically distributed (i.i.d.) random variables.

Theorem 1.13

Let $X_1, X_2, \ldots$ be i.i.d. random variables.

(i) (The WLLN). A necessary and sufficient condition for the existence of a sequence of real numbers $\{a_n\}$ for which

$$\frac{1}{n}\sum_{i=1}^{n} X_i - a_n \to_p 0$$

is that $nP(|X_1| > n) \to 0$, in which case we may take

$$a_n = E(X_1 I_{\{|X_1| \leq n\}}).$$

(ii) (The SLLN). A necessary and sufficient condition for the existence of a constant c for which

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_{a.s.} c$$

is that $E|X_1| < \infty$, in which case $c = EX_1$ and

$$\frac{1}{n}\sum_{i=1}^{n} c_i(X_i - EX_1) \to_{a.s.} 0$$

for any bounded sequence of real numbers $\{c_i\}$.

Proof of Theorem 1.13(i)

We prove the sufficiency.
The proof of necessity can be found in Petrov (1975).
Consider a sequence of random variables obtained by truncating X_j's at n:

$$Y_{nj} = X_j I_{\{|X_j| \leq n\}}.$$

Let

$$T_n = X_1 + \cdots + X_n, \quad Z_n = Y_{n1} + \cdots + Y_{nn}.$$

Then

$$P(T_n \neq Z_n) \leq \sum_{j=1}^{n} P(Y_{nj} \neq X_j) = nP(|X_1| > n) \to 0.$$

For any $\varepsilon > 0$, it follows from Chebyshev's inequality that

$$P\left(\left|\frac{Z_n - EZ_n}{n}\right| > \varepsilon\right) \leq \frac{\mathrm{var}(Z_n)}{\varepsilon^2 n^2} = \frac{\mathrm{var}(Y_{n1})}{\varepsilon^2 n} \leq \frac{EY_{n1}^2}{\varepsilon^2 n},$$

where the last equality follows from the fact that Y_{nj}, $j = 1, ..., n$, are i.i.d.

Proof of Theorem 1.13(i) (continued)

From integration by parts, we obtain that

$$\frac{EY_{n1}^2}{n} = \frac{1}{n}\int_{[0,n]} x^2 dF_{|X_1|}(x) = \frac{2}{n}\int_0^n xP(|X_1| > x)dx - nP(|X_1| > n),$$

which converges to 0 since $nP(|X_1| > n) \to 0$ (why?).
This proves that $(Z_n - EZ_n)/n \to_p 0$, which together with $P(T_n \neq Z_n) \to 0$ and the fact that $EY_{nj} = E(X_1 I_{\{|X_1| \leq n\}})$ imply the result.

Proof of Theorem 1.13(ii)

The proof for sufficiency is given in the textbook.
We prove the necessity.
Suppose that $T_n/n \to_{a.s.}$ holds for some $c \in \mathscr{R}$, $T_n = X_1 + \cdots + X_n$.
Then

$$\frac{X_n}{n} = \frac{T_n}{n} - c - \frac{n-1}{n}\left(\frac{T_{n-1}}{n-1} - c\right) + \frac{c}{n} \to_{a.s.} 0.$$

Proof of Theorem 1.13 (ii)

From Exercise 114, $X_n/n \to_{a.s.} 0$ and the i.i.d. assumption on X_n's imply

$$\sum_{n=1}^{\infty} P(|X_n| \geq n) = \sum_{n=1}^{\infty} P(|X_1| \geq n) < \infty,$$

which implies $E|X_1| < \infty$ (Exercise 54).
From the proved sufficiency, $c = EX_1$.

Remarks

- If $E|X_1| < \infty$, then $a_n = E(X_1 I_{\{|X_1| \leq n\}}) \to EX_1$ and result the WLLN is actually established in Example 1.28 in a much simpler way.
- On the other hand, if $E|X_1| < \infty$, then a stronger result, the SLLN, can be obtained.
- Some results for the case of $E|X_1| = \infty$ can be found in Exercise 148 and Theorem 5.4.3 in Chung (1974).

The next result is for sequences of independent but not necessarily identically distributed random variables.