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 $(\mathbb{R}^k, \mathcal{B}^k)$ to $(\mathbb{R}^l, \mathcal{B}^l)$. Suppose that $g$ is continuous a.s. $P_X$. Then

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

(ii) $X_n \overset{p}{\to} X$ implies $g(X_n) \overset{p}{\to} g(X)$;

(iii) $X_n \overset{d}{\to} X$ implies $g(X_n) \overset{d}{\to} g(X)$. 
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 \((\mathbb{R}^k, \mathcal{B}^k)\) to \((\mathbb{R}^l, \mathcal{B}^l)\).

Suppose that \( g \) is continuous a.s. \( P_X \). Then

(i) \( X_n \rightarrow_{a.s.} X \) implies \( g(X_n) \rightarrow_{a.s.} g(X) \);
(ii) \( X_n \rightarrow_p X \) implies \( g(X_n) \rightarrow_p g(X) \);
(iii) \( X_n \rightarrow_d X \) implies \( g(X_n) \rightarrow_d g(X) \).
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
\[ \|g(x) - g(c)\| < \varepsilon \quad \text{whenever} \quad \|x - c\| < \delta_\varepsilon. \]

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 argument still valid when \( c \) is replaced by the random vector \( X \) in the general case?
If not, how do we fix the 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

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

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 argument still valid when $c$ is replaced by the random vector $X$ in the general case? If not, how do we fix the proof?
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 
\[ \| g(x) - g(c) \| < \varepsilon \quad \text{whenever} \quad \| x - c \| < \delta_\varepsilon. \]

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 argument still valid when \( c \) is replaced by the random vector \( X \) in the general case?
If not, how do we fix the proof?
Proof

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

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

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

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 $g(X_n) \rightarrow_p g(c)$ follows from $X_n \rightarrow_p c$.

Is the previous argument 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 $X_n \xrightarrow{d} X$, where $X$ has the $N(0, 1)$ distribution, then $X_n^2 \xrightarrow{d} Y$, where $Y$ has the chi-square distribution $\chi_1^2$.

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

(iii) Under the conditions in part (ii), $\max\{X_n, Y_n\} \xrightarrow{d} \max\{X, Y\}$, 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 $(X_n, Y_n) \xrightarrow{d} (X, Y)$ cannot be relaxed to $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ (exercise); i.e., we need the convergence of the joint c.d.f. of $(X_n, Y_n)$. This is different when $\xrightarrow{d}$ is replaced by $\xrightarrow{p}$ or $\xrightarrow{a.s.}$.

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.
Example 1.30.

(i) Let $X_1, X_2, \ldots$ be random variables. If $X_n \xrightarrow{d} X$, where $X$ has the $N(0, 1)$ distribution, then $X_n^2 \xrightarrow{d} Y$, where $Y$ has the chi-square distribution $\chi^2_1$.

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

(iii) Under the conditions in part (ii), $\max\{X_n, Y_n\} \xrightarrow{d} \max\{X, Y\}$, 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 $(X_n, Y_n) \xrightarrow{d} (X, Y)$ cannot be relaxed to $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ (exercise); i.e., we need the convergence of the joint c.d.f. of $(X_n, Y_n)$. This is different when $\xrightarrow{d}$ is replaced by $\xrightarrow{p}$ or $\xrightarrow{a.s.}$.

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.
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 $\chi_1^2$.

(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 $X, X_1, X_2, \ldots, Y_1, Y_2, \ldots$ be random variables on a probability space. Suppose that $X_n \overset{d}{\to} X$ and $Y_n \overset{p}{\to} c$, where $c$ is a constant. Then

(i) $X_n + Y_n \overset{d}{\to} X + c$;
(ii) $Y_nX_n \overset{d}{\to} cX$;
(iii) $X_n/Y_n \overset{d}{\to} X/c$ if $c \neq 0$.

Proof

We prove (i) only. The proofs of (ii) and (iii) are left as exercises.

Let $t \in \mathbb{R}$ and $\varepsilon > 0$ be fixed constants.

Then

$$F_{X_n+Y_n}(t) = P(X_n + Y_n \leq t)$$

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

$$\leq P(X_n \leq t - c + \varepsilon) + P(|Y_n - c| \geq \varepsilon)$$
Theorem 1.11 (Slutsky’s theorem)

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

(i) $X_n + Y_n \rightarrow_d X + c$;
(ii) $Y_nX_n \rightarrow_d cX$;
(iii) $X_n/Y_n \rightarrow_d X/c$ if $c \neq 0$.

Proof

We prove (i) only. The proofs of (ii) and (iii) are left as exercises.

Let $t \in \mathbb{R}$ and $\varepsilon > 0$ be fixed constants. Then

$$F_{X_n + Y_n}(t) = P(X_n + Y_n \leq t)$$

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

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

$$a_n(X_n - c) \xrightarrow{d} Y,$$

where $c \in \mathbb{R}^k$ and $\{a_n\}$ is a sequence of positive numbers with

$$\lim_{n \to \infty} a_n = \infty.$$

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

(i) If $g$ is differentiable at $c$, then

$$a_n[g(X_n) - g(c)] \xrightarrow{d} [\nabla g(c)]^\tau Y,$$

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 $j$, $1 \leq j \leq m - 1$, vanishing at $c$, but with the $m$th-order partial derivatives not all vanishing at $c$.

Then

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

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

whenever \( \|x - c\| < \delta_\varepsilon \).

Then for a fixed \( \eta > 0 \),

\[ P(|Z_n| \geq \eta) \leq P(\|X_n - c\| \geq \delta_\varepsilon) + P(a_n\|X_n - c\| \geq \eta/\varepsilon). \]

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 \( \eta/\varepsilon \) is a continuity point of \( F_{\|Y\|} \).
Proof (continued)

Then

\[
\limsup_{n} P(|Z_n| \geq \eta) \leq \lim_{n \to \infty} P(\|X_n - c\| \geq \delta \epsilon) + \lim_{n \to \infty} P(a_n\|X_n - c\| \geq \eta / \epsilon)
= P(\|Y\| \geq \eta / \epsilon).
\]

\(Z_n \to_p 0\) follows since \(\epsilon\) 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.
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 \( \varepsilon \) 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 \( \delta \)-method)
Assume the conditions of Theorem 1.12. If \( Y \) has the \( N_k(0, \Sigma) \) distribution, then
\[
a_n[g(X_n) - g(c)] \to_d N(0, [\nabla g(c)]^\top \Sigma \nabla g(c)).
\]

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 \( \chi_1^2 \) (Example 1.14).
The last result can also be obtained by applying Theorem 1.10(iii).
Corollary 1.1 (the δ-method)
Assume the conditions of Theorem 1.12.
If $Y$ has the $N_k(0, \Sigma)$ distribution, then
\[ a_n[g(X_n) - g(c)] \stackrel{d}{\rightarrow} N(0, [\nabla g(c)]^\tau \Sigma \nabla g(c)). \]

Example 1.31
Let \{X_n\} be a sequence of random variables satisfying
\[ \sqrt{n}(X_n - c) \stackrel{d}{\rightarrow} N(0, 1). \]
Consider the function $g(x) = x^2$.
If $c \neq 0$, then an application of Corollary 1.1 gives that
\[ \sqrt{n}(X_n^2 - c^2) \stackrel{d}{\rightarrow} N(0, 4c^2) . \]
If $c = 0$, $g'(c) = 0$ but $g''(c) = 2$.
Hence, an application of Theorem 1.12(ii) gives that
\[ nX_n^2 \stackrel{d}{\rightarrow} [N(0, 1)]^2 , \]
which has the chi-square distribution $\chi^2_1$ (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 moments

\[
\bar{X}_n = n^{-1} \sum_i X_i, \quad \bar{Y}_n = n^{-1} \sum_i Y_i
\]

\(\mu_x = E(X_1), \quad \mu_y = E(Y_1) \neq 0, \quad \sigma^2_x = \text{Var}(X_1), \quad \sigma^2_y = \text{Var}(Y_1), \quad \sigma_{xy} = \text{Cov}(X_1, Y_1)\)

By the CLT,

\[
\sqrt{n} \left( \left( \begin{array}{c} \bar{X}_n \\ \bar{Y}_n \end{array} \right) - \left( \begin{array}{c} \mu_x \\ \mu_y \end{array} \right) \right) \rightarrow \mathcal{N}_2 \left( 0, \begin{pmatrix} \sigma^2_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2_y \end{pmatrix} \right)
\]

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

\[
\sqrt{n} \left( \frac{\bar{X}_n}{\bar{Y}_n} - \frac{\mu_x}{\mu_y} \right) \rightarrow \mathcal{N}(0, \sigma^2)
\]

\[
\sigma^2 = (\mu_y^{-1}, -\mu_x \mu_y^{-2}) \begin{pmatrix} \sigma^2_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2_y \end{pmatrix} \begin{pmatrix} \mu_y^{-1} \\ -\mu_x \mu_y^{-2} \end{pmatrix} = \frac{\sigma^2_x}{\mu_y^2} - \frac{\mu_x \sigma_{xy}}{\mu_y^3} + \frac{\mu_x^2 \sigma^2_y}{\mu_y^4}
\]