Review on Probability

김지수 (Jisu KIM)

딥러닝의 통계적 이해 (Deep Learning: Statistical Perspective), 2025 2nd Semester

Probability Spaces

A probability space is a triple $(\Omega, \mathcal{F}, P)$ where Ω is a set of "outcomes," $\mathcal{F}$ is a set of "events," and $P : \mathcal{F} \to [0, 1]$ is a function that assigns probabilities to events.

Definition. Let Ω be a set. A nonempty collection $\mathcal{F}$ of subsets of Ω is called σ-algebra (or field) if

(i) if $A \in \mathcal{F}$ then $\Omega \backslash A \in \mathcal{F}$, and

(ii) if $A_1, A_2, \cdots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Example. $\mathcal{F} = \{\phi, \Omega\}$ trivial $\sigma-$field

$\mathcal{F} = 2^{\Omega} = \{A | \ A \subset \Omega\}$: power set $\Longrightarrow \sigma-$field

Without P, $(\Omega, \mathcal{F})$ is called a measurable space, i.e., it is a space on which we can put a measure.

Definition. A measure is a nonnegative countably additive set function; that is, for an σ-algebra $\mathcal{F}$, a function $\mu : \mathcal{F} \to [0, \infty]$ is a measure if

(i) $\mu(A) \geq \mu(\phi) = 0$ for all $A \in \mathcal{F}$, and

(iii) For $A_1, A_2, \cdots \in \mathcal{F}$ with $A_i \cap A_j = \phi$ for any $i \neq j$,

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

Definition. (1) $\mu(\Omega) < \infty \Longrightarrow$finite measure

(2) $\mu(\Omega) = 1 \Longrightarrow$probability measure

(3) $\exists$a partition $A_1, A_2, \cdots$ with $\bigcup_{i=1}^{\infty} A_i = \Omega$ and $\mu(A_i) < \infty \Longrightarrow \sigma-$finite measure

Theorem ([1, Theorem 1.1.1])**.** *Let μ be a measure on $(\Omega, \mathcal{F})$.*

(i) Monotonicity. If $A \subset B$ then $\mu(A) \leq \mu(B)$.

(ii) Subadditivity. If $A \subset \bigcup_{i=1}^{\infty} A_i$ then $\mu(A) \leq \sum_{i=1}^{\infty} \mu(A_i)$.

(iii) Continuity from below. $A_n \uparrow A$ (i.e. $A_1 \subset A_2 \subset \cdots$ and $A = \bigcup_{i=1}^{\infty} A_i$) then $\mu(A_i) \uparrow \mu(A)$.

(iv) Continuity from above. $A_n \downarrow A$ (i.e. $A_1 \supset A_2 \supset \cdots$ and $A = \bigcap_{i=1}^{\infty} A_i$) with $\mu(A_1) < \infty$ then $\mu(A_i) \downarrow \mu(A)$.

Definition. Let $\mathcal{A}$ be a class of subsets of Ω. Then $\sigma(\mathcal{A})$ denotes the smallest $\sigma-$algebra that contains $\mathcal{A}$.

For any any $\mathcal{A}$, such $\sigma(\mathcal{A})$ exists and is unique.

Definition. Borel $\sigma-$field on $\mathbb{R}^d$, denoted by $\mathcal{R}^d$, is the smallest $\sigma-$field containing all open sets.

Theorem ([1, Theorem 1.1.4]). *There is a unique measure μ on $(\mathbb{R}, \mathcal{R})$ with*

$$\mu((a,b]) = b - a.$$

Such measure is called Lebesgue measure.

Example ([1, Example 1.1.3]). Product space
$(\Omega_i, \mathcal{F}_i, \mathcal{P}_i)$: sequence of probability spaces
Let $\Omega = \Omega_1 \times \cdots \times \Omega_n = \{(\omega_1, \cdots, \omega_n) | \; \omega_i \in \Omega_i\}$
$\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_n =$the $\sigma-$field generated by $A_1 \times \cdots \times A_n$, where $A_i \in \mathcal{F}_i$
$P = P_1 \times \cdots \times P_n$ (i.e. $P(A_1 \times \cdots \times A_n) = P_1(A_1) \cdots P_n(A_n)$)

Distribution and Random Variables

Definition. Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ are measurable spaces. A mapping $X : \Omega \to S$ is a measurable map from $(\Omega, \mathcal{F})$ to $(S, \mathcal{S})$ if

$$\text{for all } B \in \mathcal{S}, \;\; X^{-1}(B) \coloneqq \{\omega \in \Omega : \;\; X(\omega) \in B\} \in \mathcal{F}.$$

If $(S, \mathcal{S}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $d > 1$ then X is called a random vector. If $d = 1$, X is called a random variable.

Example. A trivial but useful example of a random variable is indicator function 1_A of a set $A \in \mathcal{F}$:

$$1_A(\omega) = \begin{cases} 1 & \omega \in A, \\ 0 & \omega \notin A. \end{cases}$$

If X is a random variable, then X induces a probability measure on $\mathbb{R}$.

Definition. The probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined as $\mu(A) = P(X \in A)$ for all $A \in \mathcal{B}(\mathbb{R})$ is called the distribution of X.

Remark. The distribution can be defined similarly for random vectors.

The distribution of a random variable X is usually described by giving its distribution function.

Definition. The distribution function $F(x)$ of a random variable X is defined as $F(x) = P(X \leq x)$.

Theorem ([1, Theorem 1.2.1]). *Any distribution function F has the following properties:*
(i) F is nondecreasing.
(ii) $\lim_{n\to\infty} F(x) = 1, \;\; \lim_{n\to-\infty} F(x) = 0$.
(iii) F is right continuous. i.e. $\lim_{y\downarrow x} F(y) = F(x)$.
(iv) $P(X < x) = F(x-) = \lim_{y\uparrow x} F(x)$.
(v) $P(X = x) = F(x) - F(x-)$.

Theorem ([1, Theorem 1.2.2]). *If F satisfies (i) (ii) (iii) in [1, Theorem 1.2.1], then it is the distribution function of some random variable. That is, there exists a triple $(\Omega, \mathcal{F}, P)$ and a random variable X such that $F(x) = P(X \leq x)$.*

Theorem. *If F satisfies (i) (ii) (iii), then there uniquely exists a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that for all $a < b$,*

$$\mu((a,b]) = F(b) - F(a).$$

Definition. If X and Y induce the same distribution μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we say X and Y are equal in distribution. We write

$$X \stackrel{d}{=} Y.$$

Definition. When the distribution function $F(x) = P(X \leq x)$ has the form $F(x) = \int_{-\infty}^{x} f(y)dy$, then we say X has the density function f.

Remark. f is not unique, but unique up to Lebesque measure 0.

Theorem ([1, Theorem 1.3.4])**.** *If $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ and $f : (S, \mathcal{S}) \to (T, \mathcal{T})$ are measurable maps, then $f(X)$ is measurable.*

Theorem. *$f : (S, \mathcal{S}) \to (T, \mathcal{T})$ and suppose $\mathcal{S} = \sigma(\text{open sets})$, $\mathcal{T} = \sigma(\text{open sets})$. Then, if f is continuous then f is measurable.*

Theorem ([1, Theorem 1.3.5])**.** *If $X_1, \cdots, X_n$ are random variables and $f : (\mathbb{R}^n, \mathcal{R}^n) \to (\mathbb{R}, \mathcal{R})$ is measurable, then $f(X_1, \cdots, X_n)$ is a random variable.*

Theorem ([1, Theorem 1.3.6])**.** *If $X_1, \cdots, X_n$ are random variables then $X_1 + \cdots + X_n$ is a random variable.*

Remark. If X, Y are random variables, then

$$cX \ (c \text{ is scalar}),\ X \pm Y,\ XY,\ \sin(X),\ X^2,\ \cdots,$$

are all random variables.

Theorem ([1, Theorem 1.3.7])**.** $\inf_n X_n,\ \sup_n X_n,\ \limsup_n X_n,\ \liminf_n X_n$ *are random variables.*

Integration

Let μ be a σ-finite measure on $(\Omega, \mathcal{F})$.

Definition. For any predicate $Q(\omega)$ defined on Ω, we say Q is true ($\mu-$)almost everywhere (or a.e.) if $\mu(\{\omega : Q(\omega) \ is \ false\}) = 0$

Step 1.

Definition. φ is a simple function if $\varphi(\omega) = \sum_{i=1}^{n} a_i 1_{A_i}$ with $A_i \in \mathcal{F}$

If φ is a simple function and $\varphi \geq 0$, we let

$$\int \varphi d\mu = \sum_{i=1}^{n} a_i \mu(A_i)$$

Step 2.

Definition. If f is measurable and $f \geq 0$ then we let

$$\int f d\mu = \sup\{\int \varphi d\mu :\ 0 \leq \varphi \leq f \text{ and } \varphi \text{ simple}\}$$

We define the integral of f over the set E:

$$\int_E f d\mu := \int f \cdot 1_E d\mu$$

Step 3.

Definition. We say measurable f is integrable if $\int |f| d\mu < \infty$. Let

$$f^+(x) := f(x) \vee 0, \qquad \text{and} \qquad f^-(x) := (-f(x)) \vee 0,$$

where $a \vee b = \max(a, b)$. We define the integral of f by

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu.$$

we can also define $\int f d\mu$ if $\int f^+ d\mu = \infty$ and $\int f^- d\mu < \infty$, or $\int f^+ d\mu < \infty$ and $\int f^- d\mu = \infty$

Theorem ([1, Theorem 1.4.7]). *Suppose f and g are integrable.*

(i) If $f \geq 0$ a.e. then $\int f d\mu \geq 0$
(ii) $\forall a \in \mathbb{R}$, $\int a f d\mu = a \int f d\mu$
(iii) $\int f + g d\mu = \int f d\mu + \int g d\mu$
(iv) If $g \leq f$ a.e. then $\int g d\mu \leq \int f d\mu$
(v) If $g = f$ a.e. then $\int g d\mu = \int f d\mu$
(vi) $|\int f d\mu| \leq \int |f| d\mu$

Definition. If X is a random variable on $(\Omega, \mathcal{F}, P)$, we define its expected value to be $\mathbb{E}(X) = \int_\Omega X dP$

We also write $\mathbb{E}(X; A) = \int_A X dP$.

Several techniques of integration

- The pushforward measure of a transformation T is $T_*\mu := \mu(T^{-1}(A))$. The change of variables formula for pushforward measures is
$$\int_\Omega f \circ T d\mu = \int_{T(\Omega)} f dT_*\mu.$$
Now, consider a probability space $(\Omega, \mathcal{F}, P)$, and consider a measurable map $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ as a transformation. Then the distribution measure μ_X of X is in fact the pushforward measure $\mu_X(A) = P(X \in A) = P(X^{-1}(A))$, and hence the change of variable formula becomes
$$\mathbb{E}_P[f(X)] = \int_\Omega f(X(\omega)) dP(\omega) = \int_{X(\Omega)} f(x) d\mu_X(x).$$

- For Lebesgue measure λ and Riemann integrable function f, $\int_{[a,b]} f d\lambda$ is the same as the Riemann integral $\int_a^b f(x)dx$.

- $\int f d\delta_x = f(x)$, where δ_x is the Dirac-delta measure, i.e., $\delta_x(A) = I(x \in A)$.

- For a random variable $X \geq 0$,
$$\begin{aligned}
\mathbb{E}_P[X] = \int_\Omega X(\omega) dP(\omega) &= \int_\Omega \int_{[0, X(\omega)]} dt dP(\omega) \\
&= \int_{\{(\omega,t) \in \Omega \times [0,\infty) : 0 \leq t \leq X(\omega)\}} dt \times dP(\omega) \\
&= \int_0^\infty \int_{\{\omega \in \Omega : X(\omega) \geq t\}} dP(\omega) dt \\
&= \int_0^\infty P(X \geq t) dt.
\end{aligned}$$

Independence

Definition. Let $(\Omega, \mathcal{F}, P)$ be probability space. Two events $A, B \in \mathcal{F}$ are independent (독립) if

$$P(A \cap B) = P(A)P(B).$$

Two random variables X and Y are independent if for all $C, D \in \mathcal{R}$,

$$P(X \in C,\ Y \in D) = P(X \in C)P(Y \in D).$$

Two σ-fields $\mathcal{F}_1$ and $\mathcal{F}_2$ ($\subset \mathcal{F}$) are independent if for all $A \in \mathcal{F}_1$ and $B \in \mathcal{F}_2$, A and B are independent.

Definition. $\sigma-$fields $\mathcal{F}_1, \cdots, \mathcal{F}_n$ are independent if for all $A_i \in \mathcal{F}_i$,

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i),$$

Random variables $X_1, \cdots, X_n$ are independent if for all $B_i \in \mathcal{R}$,

$$P\left(\bigcap_{i=1}^{n} \{X_i \in B_i\}\right) = \prod_{i=1}^{n} P(X_i \in B_i).$$

Sets $A_1, \cdots, A_n$ are independent if for all $I \subset \{1, \cdots, n\}$,

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i)$$

Remark. the definition of independent events is not enough to assume pairwise independent, which is $P(A_i \cap A_j) = P(A_i)P(A_j),\ i \neq j$. It is clear that indenendent events are pairwise independent, but converse is not true.

Example. Let $X_1,\ X_2,\ X_3$be independent random variables with $P(X_i = 0) = P(X_i = 1) = \frac{1}{2}$

Let $A_1 = \{X_2 = X_3\},\ A_2 = \{X_3 = X_1\}$ and $A_3 = \{X_1 = X_2\}$. These events are pairwise independent but not independent.

Theorem ([1, Theorem 2.1.12])**.** *Suppose X and Y are independent, and $f, g : \mathbb{R} \to \mathbb{R}$ are measurable functions with $f, g \geq 0$ or $\mathbb{E}\left|f(X)\right|, \mathbb{E}\left|g(X)\right| < \infty$, then*

$$\mathbb{E}_P[f(X)g(Y)] = \mathbb{E}_P[f(X)]\mathbb{E}_P[g(Y)].$$

Conditional Expectation

Definition. Let $(\Omega, \mathcal{F}_0, P)$ be a given probability space. Suppose a σ-field $\mathcal{F} \subset \mathcal{F}_0$ and a random variable $X \in \mathcal{F}_0$ are given. $E(X|\mathcal{F})$ (conditional expectation of X given $\mathcal{F}$) is a random variable Y such that

(i) $Y \in \mathcal{F}$, i.e., is $\mathcal{F}$ measurable

(ii) for all $A \in \mathcal{F}$, $\int_A X dP = \int_A Y dP$.

Any Y satisfying (i) and (ii) is said to be a version of $E(X|\mathcal{F})$.

Lemma ([1, Lemma 4.1.1])**.** *If Y satisfies (i) and (ii), then it is integrable.*

Remark. Uniqueness.

Suppose there are two random variables Y and Y' satisfying (i) and (ii) of the definition of the conditional expectation. Then $Y = Y'$ a.s.

Remark. Existence

$E(X|\mathcal{F})$ exists.

Example ([1, Example 4.1.3])**.** If $X \in \mathcal{F}$, $\mathbb{E}(X|\mathcal{F}) = X$

Example ([1, Example 4.1.4])**.** If X is independent of $\mathcal{F}$, then $\mathbb{E}(X|\mathcal{F}) = \mathbb{E}(X)$

Example ([1, Example 4.1.5])**.** Let $\Omega_1, \Omega_2, \cdots$ be a countable partition of Ω into disjoint sets and let $\mathcal{F} = \sigma(\Omega_1, \Omega_2, \cdots)$

Then $E(X|\mathcal{F})(\omega) = \sum_{k=1}^{\infty} c_k I(\omega \in \Omega_k)$

where $c_k = \begin{cases} \frac{\int_{\Omega_k} X dP}{P(\Omega_k)} & if\ P(\Omega_k) > 0 \\ arbitrary & if\ P(\Omega_k) = 0 \end{cases}$

Definition. $P(A|\mathcal{F}) = E(1_A|\mathcal{F})$

$P(A|B) = P(A \cap B)/P(B)$

Remark. $P(A|\sigma(B)) = \begin{cases} P(A|B) & if\ \omega \in B \\ P(A|B^c) & if\ \omega \in B^c \end{cases}$

Definition. Conditional expectation given random variable

$E(X|Y) = E(X|\sigma(Y))$

Definition. Conditional expectation given $Y = y$, i.e. $E(X|Y = y)$.

Consider $E(X|Y)$, which is $\sigma(Y)$-measurable. Then there exists a measurable function $h : \mathbb{R} \to \mathbb{R}$ s.t. $\mathbb{E}(X|Y) = h(Y)$ (Exercise 1.3.8) Now, we can define

$$\mathbb{E}(X|Y = y) = h(y)$$

Definition. $P(A|Y = y) = E(I_A|Y = y)$

Example ([1, Example 4.1.6])**.** $(X, Y) \sim pdf\ f(x, y)$ (w.r.t. Lebesque measure). Then

$$\mathbb{E}(g(X)|Y) = \frac{\int g(x)f(x, Y)dx}{\int f(x, Y)dx},$$

provided $f(x, y) > 0\ \forall(x, y)$.

Example ([1, Example 4.1.7])**.** Suppose X and Y are independent. Let φ be a function wit $\mathbb{E}|\varphi(X, Y)| < \infty$ and let $g(x) = \mathbb{E}(\varphi(x, Y))$. Then

$$\mathbb{E}(\varphi(X, Y)|X) = g(X).$$

Example. Convolution formula

$X \perp Y$

Let $\varphi_z(x, y) = I(x + y \le z)$

Then $g(x) = E(\varphi_z(x, Y)) = P(Y \le z - x)$

Hence $P(X + Y \le z|X) = F_Y(z - X)$

which implies

$P(X + Y \le z) = E(P(X + Y \le z|X))$

$= E(F_Y(z - X))$

$= \int_{-\infty}^{\infty} F_Y(z - x)dF_X(x) = F_X * F_Y$

Properties

Theorem ([1, Theorem 4.1.9]). *(a) Linearlity.*

$$\mathbb{E}(aX + Y|\mathcal{F}) = a\mathbb{E}(X|\mathcal{F}) + \mathbb{E}(Y|\mathcal{F}).$$

(b) Monotonicity.
If $X \leq Y$, then

$$\mathbb{E}(X|\mathcal{F}) \leq \mathbb{E}(Y|\mathcal{F}).$$

(c) Monotone convergence theorem.
If $X_n \geq 0$ and $X_n \uparrow X$ with $E|X| < \infty$, then

$$\mathbb{E}(X_n|\mathcal{F}) \uparrow \mathbb{E}(X|\mathcal{F}).$$

Theorem ([1, Theorem 4.1.10]). *Jensen Inequality*
If φ is convex and $E|X| < \infty$ and $E|\varphi(X)| < \infty$, then

$$\varphi(E(X|\mathcal{F})) \leq E(\varphi(X)|\mathcal{F}).$$

Theorem ([1, Theorem 4.1.11]). *Conditional expectation is a contraction in L^p, $p \geq 1$, i.e.,*
i.e.,

$$\mathbb{E}(|\mathbb{E}(X|\mathcal{F})|^p) \leq \mathbb{E}|X|^p.$$

Theorem ([1, Theorem 4.1.12]). *If $\mathcal{F} \subset \mathcal{G}$ and $E(X|\mathcal{G}) \in \mathcal{F}$, then*

$$\mathbb{E}(X|\mathcal{F}) = \mathbb{E}(X|\mathcal{G}).$$

Theorem ([1, Theorem 4.1.13]). *If $\mathcal{F}_1 \subset \mathcal{F}_2$, then*
(i) $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_1)|\mathcal{F}_2) = \mathbb{E}(X|\mathcal{F}_1)$
(ii) $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_2)|\mathcal{F}_1) = \mathbb{E}(X|\mathcal{F}_1)$

Theorem ([1, Theorem 4.1.14]). *If $X \in \mathcal{F}$ and $E|X| < \infty$, then*

$$\mathbb{E}(XY|\mathcal{F}) = X\mathbb{E}(Y|\mathcal{F}).$$

Theorem ([1, Theorem 4.1.15]). *Suppose $EX^2 < \infty$, then $\mathbb{E}(X|\mathcal{F})$ is a random variable $Y \in \mathcal{F}$ that minimizes $\mathbb{E}(X-Y)^2$ among all random variables$\in \mathcal{F}$*

Weak laws of large numbers (큰 수의 약법칙)

Various modes of convergence

$\{X_n\}$ and X are random variables defined on $(\Omega, \mathcal{F}, P)$

Definition. $X_n \to X$ almost surely (a.s.) (with probability 1(w.p. 1), almost everywhere(a.e.)) if $P\{\omega : X_n(\omega) \to X(\omega)\} = 1$

Equivalent definition : $\forall \epsilon,\ \lim_{m\to\infty} P\{\omega : |X_n(\omega) - X(\omega)| \leq \epsilon\ \forall n \geq m\} = 1$
or $\forall \epsilon,\ \lim_{m\to\infty} P\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\ \forall n \geq m\} = 0$

Definition. $X_n \to X$ in probability (확률수렴) (in pr, $\xrightarrow{p}$) if $\lim_{n\to\infty} P\{|X_n - X| > \epsilon\} = 0$

Theorem. $X_n \to X$ *a.s.* $\Longrightarrow X_n \xrightarrow{p} X$

Remark. $X_n \xrightarrow{p} X \not\Rightarrow X_n \to X$ a.s.

Definition. $X_n \to X$ in L_p, $0 < p < \infty$
if $\lim_{n\to\infty} E(|X_n - X|^p) = 0$ provided $E|X_n|^p < \infty$, $E|X|^p < \infty$.

Theorem. $X_n \to X$ *in* $L_p \implies X_n \xrightarrow{p} X$

Theorem. *(Chebyshev inequality, 체비셰프 부등식)*

$$P(|X| \geq \epsilon) \leq \frac{E|X|^p}{\epsilon^p}$$

Remark. $X_n \xrightarrow{p} X \not\Rightarrow X_n \to X$ in L_p

Example. $\Omega = [0,1],\ \mathcal{F} = \mathcal{B}[0,1],\ P = Unif[0,1]$
$X(\omega) = 0,\ X_n(\omega) = nI(0 \leq \omega \leq \frac{1}{n})$
Then $P\{|X_n(\omega) - X(\omega)| > \epsilon\} = P\{0 \leq \omega \leq \frac{1}{n}\} = \frac{1}{n} \to 0$
But $E|X_n - X| = E|X_n| = 1$

Theorem. $X_n \xrightarrow{p} X$ *and there exists a random variables* Z *s.t.*
$|X_n| \leq Z$ *and* $E|Z|^p < \infty$
Then $X_n \to X$ *in* L_p.

Remark. If $E|X| < \infty$, then
$\lim_{n\to\infty} \int_{A_n} |X| dP \to 0$ whenever $P(A_n) \to 0$

2..2.1. L_2 weak law

Theorem ([1, Theorem 2.2.3])**.** *Let* $X_1, X_2, \cdots$ *be uncorrelated random variables with* $EX_i = \mu$ *and* $Var(X_i) \leq C < \infty$ *Let* $S_n = \sum_{i=1}^{n} X_i$. *Then* $\frac{S_n}{n} \to \mu$ *in* L_2 *and also in probability.*

Theorem ([1, Theorem 2.2.14])**.** *Weak law of large numbers (큰 수의 약법칙, 대수의 약법칙)*
Let $X_1, X_2, \cdots$ *be i.i.d. random variables with* $E|X_i| < \infty$. *Let* $S_n = X_1 + \cdots + X_n$ *and let* $\mu = EX_1$. *Then* $\frac{S_n}{n} \to \mu$ *in probability.*

Weak Convergence

We define weak convergence for random variables, but most of the results can be generalized to measurable maps $X_n, X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$, where S is equipped with a metric ρ.

Definition. A sequence of random vectors $\{X_n\}$ converges weakly or converges in distribution (분포수렴) to a limit X ($X_n \Rightarrow X,\ X_n \xrightarrow{w} X,\ X_n \xrightarrow{d} X$) if

$$\mathbb{E}_P\left[g(X_n)\right] \to \mathbb{E}_P\left[g(X)\right], \qquad \text{for all } g \in C_b(\mathbb{R}),$$

where $C_b(\mathbb{R})$ is a set of continuous and bounded functions. We analogously define $P_n \xrightarrow{d} P$ for probability measures $\{P_n\}$ and P, i.e., $\int g(x)dP_n(x) \to \int g(x)dP(x)$ for all $g \in C_b(\mathbb{R})$. We also analogously define $F_n \xrightarrow{d} F$ ($F_n \Rightarrow F,\ F_n \xrightarrow{w} F$) for distribution functions $\{F_n\}$ and F, i.e., $\int g(x)dF_n(x) \to \int g(x)dF(x)$ for all $g \in C_b(\mathbb{R})$.

Theorem ([1, Theorem 3.2.9]). *A sequence of distribution function F_n converges weakly to a limit F if and only if $F_n(y) \to F(y)$ for all continuity points of F.*

Example ([1, Example 3.2.1]). Let $X_1, X_2, \cdots$ be i.i.d. with $P(X_1 = 1) = P(X_1 = -1) = \frac{1}{2}$, and let $S_n = X_1 + \cdots + X_n$. Then

$$F_n(y) = P(S_n/\sqrt{n} \le y) \to \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \ \forall y \in \mathbb{R}.$$

That is, $F_n \xrightarrow{w} \mathcal{N}(0,1)$.

Example ([1, Example 3.2.3]). Let $X \sim F$ and $X_n = X + \frac{1}{n}$. Then

$$F_n(x) = P(X_n \le x) = F(x - \frac{1}{n}) \to F(x-).$$

Hence $F_n(x) \to F(x)$ only when $F(x) = F(x-)$, i.e. only if x is a continuity point of F. Still, $X_n \xrightarrow{d} X$.

Example ([1, Example 3.2.4]). Let $X_p \sim Geo(p)$, i.e. $P(X_p \ge m) = (1-p)^{m-1}$. Then

$$P(X_p > \frac{x}{p}) = (1-p)^{\frac{x}{p}} \to e^{-x}, \qquad \text{as } p \to 0.$$

In words, pX_p converges weakly to an exponential distribution.

Central Limit Theorem (중심극한정리)

Theorem ([1, Theorem 3.4.1]). *Let $X_1, X_2, \cdots$ be i.i.d. with $\mathbb{E}X_i = \mu$ and $Var(X_i) = \sigma^2 > 0$. If $S_n = X_1 + \cdots + X_n$, then*

$$(S_n - n\mu)/(\sqrt{n}\sigma) \xrightarrow{d} \mathcal{N}(0,1).$$

Theorem ([1, Theorem 3.4.17]). *Berry-Essen theorem*

Let $X_1, X_2, \cdots$ be i.i.d. with $EX_i = 0$, $EX_i^2 = \sigma^2$ and $E|X_1|^3 = \rho < \infty$. Let $F_n(x)$ be the distribution function of $(X_1 + \cdots + X_n)/(\sigma\sqrt{n})$ and $\Phi(x)$ be the standard normal distribution. Then

$$\sup_x |F_n(x) - \Phi(x)| \le 3\rho/(\sigma^3\sqrt{n}).$$

Stochastic Order Notation

The classical order notation should be familiar to you already.

1. We say that a sequence $a_n = o(1)$ if $a_n \to 0$ as $n \to \infty$. Similarly, $a_n = o(b_n)$ if $a_n/b_n = o(1)$.
2. We say that a sequence $a_n = O(1)$ if the sequence is eventually bounded, i.e. for all n large, $|a_n| \le C$ for some constant $C \ge 0$. Similarly, $a_n = O(b_n)$ if $a_n/b_n = O(1)$.
3. If $a_n = O(b_n)$ and $b_n = O(a_n)$ then we use either $a_n = \Theta(b_n)$ or $a_n \asymp b_n$.

When we are dealing with random variables we use stochastic order notation.

1. We say that $X_n = o_P(1)$ if for every $\epsilon > 0$, as $n \to \infty$

   $$\mathbb{P}(|X_n| \ge \epsilon) \to 0,$$

   i.e. X_n converges to zero in probability.

2. We say that $X_n = O_P(1)$ if for every $\epsilon > 0$ there is a finite $C(\epsilon) > 0$ such that, for all n large enough:

$$\mathbb{P}\left(|X_n| \geq C(\epsilon)\right) \leq \epsilon.$$

The typical use case: suppose we have $X_1, \ldots, X_n$ which are i.i.d. and have finite variance, and we define:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

1. $\hat{\mu} - \mu = o_P(1)$ (Weak Law of Large Number)
2. $\hat{\mu} - \mu = O_P(1/\sqrt{n})$ (Central Limit Theorem)

Proposition. *1. $X_n \xrightarrow{P} X$ implies $X_n \xrightarrow{d} X$, and this implies $X_n = O_p(1)$. Also, $X_n = o_p(1)$ implies $X_n = O_p(1)$.*

2. (a) $O_p(1) + O_p(1) = O_p(1)$

(b) $O_p(1) + o_p(1) = O_p(1)$

(c) $o_p(1) + o_p(1) = o_p(1)$

(d) $O_p(1) \cdot O_p(1) = O_p(1)$

(e) $O_p(1) \cdot o_p(1) = o_p(1)$

(f) $o_p(1) \cdot o_p(1) = o_p(1)$

References

[1] Rick Durrett. *Probability—theory and examples*, volume 49 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. Fifth edition.