# Statistical inference

Jason Mak

July 2024

## Contents

# Chapter 6: Convergence

Given a probability space $(\Omega, \mathcal{F}, P)$, we have a sequence of finite random variables $\{X_n\}_{n=1}^{\infty}$ such that $X_n : (\Omega, \mathcal{F}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. This chapter discusses the convergence behaviours of this sequence in various situations.

Normally, it is desirable that $\{X_n\}_{n=1}^{\infty}$ is *i.i.d.* (independent and identically distributed), which means $\{X_n\}_{n=1}^{\infty}$ is a group of independent variables and $X_i \overset{d.}{=} X_j \ \forall i, j \in \mathbb{N}^*$. In this case, every random variable shares an identical *pdf* $F$.

We may use the above notation without further giving its detail meaning throughout the chapter.

## 6.1 Convergence in probability

### 6.1.1 Basics

**Definition**: convergence in probability

$\{X_n\}_{n=1}^{\infty}$ converges in probability to a random variable $X$, denoted as $X_n \xrightarrow{\text{p.}} X$ if

$$\forall \epsilon > 0, \ \lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1, \text{ or equivalently, } \lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$$

The statement holds only when $X$ is finite almost everywhere.

Note that by the completeness of real number, $P(|X_n - X| < \epsilon)$ and $P(|X_n - X| \leq \epsilon)$ are equivalent since $\epsilon$ ranges in $\mathbb{R}^+$.

If $\{\mathbf{X}_n\}_{n=1}^{\infty}$ and $\mathbf{X}$ are finite random vectors or sequences, with their corresponding distance function $d$ in metric space, define $\mathbf{X}_n \xrightarrow{\text{p.}} \mathbf{X}$ to be $d(\mathbf{X}_n, \mathbf{X}) \xrightarrow{\text{p.}} 0$.

**Proposition**: Additivity of convergence in probability

It is easy to check that convergence in probability satisfies the additivity property. If $X_n \xrightarrow{\text{p.}} X$ and $Y_n \xrightarrow{\text{p.}} Y$, then $\{X_n + Y_n\}_{n=1}^{\infty} \xrightarrow{\text{p.}} XY$.

More properties (e.g. multiplicability and continuous mapping) are discussed in 6.2.2 with the help of almost surely convergence.

**Example**: Weak law of large numbers (the simple version)

If $\{X_n\}_{n=1}^{\infty}$ is a sequence of *i.i.d.* random variables of $\mathcal{L}^2$, which guarantees $\text{var}(X_1) = \text{var}(X_2) = \cdots$ and $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \cdots$ to be finite, we have

$$\bar{X}_n \xrightarrow{\text{p.}} \mathbb{E}(X_1)$$

where $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean for $n \in \mathbb{N}^*$.
Proof:
For all $n \in \mathbb{N}^*$, we first have $\mathbb{E}(\bar{X}_n) = \mathbb{E}(X_1)$ and $\text{var}(\bar{X}_n) = \frac{\text{var}(X_1)}{n}$ . Then, by Chebyshev's inequality,

$$\forall \epsilon > 0,\ P(|\bar{X}_n - \mathbb{E}(X_1)| \geq \epsilon) = P(|\bar{X}_n - \mathbb{E}(\bar{X}_n)|^2 \geq \epsilon^2) < \frac{\text{var}(\bar{X}_n)}{\epsilon^2} = \frac{\text{var}(X_1)}{\epsilon^2} \times \frac{1}{n}$$

Since $P(|\bar{X}_n - \mathbb{E}(X_1)| \geq \epsilon) \geq 0$ and $\lim_{n\to\infty} \frac{\text{var}(X_1)}{\epsilon^2} \times \frac{1}{n} = 0$, we have $\lim_{n\to\infty} P(|\bar{X}_n - \mathbb{E}(X_1)| \geq \epsilon) = 0\ \forall \epsilon > 0$.
More versions of laws of large numbers will be shown in the following.

### 6.1.2 Weak laws of large numbers

**Theorem**: General weak law of large numbers Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of independent random variables. If

$$1)\quad \lim_{n\to\infty} \sum_{i=1}^{n} P(|X_i| > n) = 0$$

and

$$2)\quad \lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}(X_i^2 \mathbf{1}_{\{|X_i| \leq n\}}) = 0$$

then

$$\frac{S_n - a_n}{n} \xrightarrow{\text{p.}} 0$$

where

$$S_n := \sum_{i=1}^{\infty} X_i \text{ and } a_n := \sum_{i=1}^{n} \mathbb{E}(X_i \mathbf{1}_{\{|X_i| \leq n\}})$$

Proof:
Let $X'_{n,i} = X_i \mathbf{1}_{|X_i| \leq n}$ and $S'_n = \sum_{i=1}^{n} X'_{n,i}$ for $n \in \mathbb{N}^*, i \in \{1, 2, \cdots, n\}$. Then for any $n \in \mathbb{N}^*$,

$$\sum_{i=1}^{n} P(X'_{n,i} \neq X_i) = \sum_{i=1}^{n} P(|X_i| > n)$$

and

$$\mathbb{E}(S'_{n,i}) = \sum_{i=1}^{n} \mathbb{E}(X_i \mathbf{1}_{\{|X_i| \leq n\}}) = a_i$$

We will first show that the condition 1 gives $\{S_n - S'_n\}_{n=1}^{\infty} \xrightarrow{\text{p.}} 0$ and the condition 2 gives $\{\frac{S'_n - a_n}{n}\}_{n=1}^{\infty} \xrightarrow{\text{p.}} 0$. Therefore, $\frac{S_n - a_n}{n} \xrightarrow{\text{p.}} 0$ by the addibility.

Condition 1 gives $\lim_{n\to\infty}\sum_{i=1}^{n}P(X'_{n,i}\neq X_i)=0$. Since for any $\epsilon>0$,

$$P(|S_n-S'_n|>\epsilon)\leq P(S_n\neq S'_n)\leq P\left(\bigcup_{i=1}^{n}\{X_n\neq X_{n,i}\}\right)\leq\sum_{i=1}^{n}P(X'_{n,i}\neq X_i)$$

thus, $\lim_{n\to\infty}P(|S_n-S'_n|>\epsilon)=0$ and $\{S_n-S'_n\}_{n=1}^{\infty}\xrightarrow{\text{p.}}0$.
Next, we prove the remaining part. Since for any $n\in\mathbb{N}^*$, $|S'_n|\leq\sum_{i=1}^{n}|X'_{n,i}|\leq n^2\in\mathcal{L}^2$, we apply the Chebyshev's inequality, i.e. for any $\epsilon>0$,

$$P\left(\left|\frac{S_n-a_n}{n}\right|>\epsilon\right)=P\left(\left|\frac{S_n-\mathbb{E}(S_n)}{n}\right|>\epsilon\right)\leq\frac{\mathrm{var}(S_n)}{\epsilon^2}=\frac{1}{n^2\epsilon^2}\sum_{i=1}^{n}\mathrm{var}(X_{n,i})\ \forall n\in\mathbb{N}^*$$

The final equation is given by independence. Condition 2 further gives that

$$\lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(X'_{n,i})^2\right]=\lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}(X_i^2\mathbf{1}_{\{|X_i|\leq n\}})=0$$

Then, since $\mathrm{var}(X_{n,i})\leq\mathbb{E}\left[(X'_{n,i})^2\right]\ \forall 1\leq i\leq n$, we have

$$\forall\epsilon>0,\ \lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_{n,i})=0\ \Rightarrow\ \lim_{n\to\infty}P\left(\left|\frac{S_n-a_n}{n}\right|>\epsilon\right)=0$$

Thus, $\{\frac{S'_n-a_n}{n}\}_{n=1}^{\infty}\xrightarrow{\text{p.}}0$ is proved.

**Example**: applying the general weak law of convergence
The following are different situations that the 2 conditions of the general weak law hold.

1. The simple version in 6.2.1, when $\{X_n\in\mathcal{L}^2\}_{n=1}^{\infty}$ are *i.i.d.*
2. $\{X_n\in\mathcal{L}^1\}_{n=1}^{\infty}$ are *i.i.d.*, then $\lim_{n\to\infty}\frac{a_n}{n}=\mathbb{E}(X_1)$ and $\frac{S_n}{n}\xrightarrow{\text{p.}}\mathbb{E}(X_1)$ still holds as 1.
3. $\{X_n\}_{n=1}^{\infty}$ are *i.i.d.* and $\lim_{x\to\infty}xP(|X_1|>x)=0$

We verify that they satisfy the two conditions. Situation 1 is easy. For situation 2, we have

$$0\leq\lim_{n\to\infty}\sum_{i=1}^{n}P(|X_i|>n)=\lim_{n\to\infty}n\mathbb{E}(\mathbf{1}_{|X_1|>n})=\lim_{n\to\infty}\mathbb{E}(n\mathbf{1}_{|X_1|>n})\leq\lim_{n\to\infty}\mathbb{E}(|X_1|\mathbf{1}_{|X_1|>n})=0$$

because $X_1 \in \mathcal{L}^1$ gives $P(|X_1| = \infty) = 0$, and

$$\begin{aligned}
\forall \epsilon > 0,\ 0 \le \lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}(X_i^2 \mathbf{1}_{\{|X_i| \le n\}}) &= \lim_{n\to\infty} \frac{1}{n} \mathbb{E}(X_1^2 \mathbf{1}_{\{|X_1| \le n\}}) \\
&= \lim_{n\to\infty} \frac{1}{n} \left[ \mathbb{E}(X_1^2 \mathbf{1}_{\{|X_1| \le \epsilon\sqrt{n}\}}) + \mathbb{E}(X_1^2 \mathbf{1}_{\{\epsilon\sqrt{n} < |X_1| \le n\}}) \right] \\
&\le \lim_{n\to\infty} \frac{1}{n} \left[ n\epsilon^2 + \mathbb{E}(n|X_1| \mathbf{1}_{\{\epsilon\sqrt{n} < |X_1| \le n\}}) \right] \\
&= \lim_{n\to\infty} \left[ \epsilon^2 + \mathbb{E}(|X_1| \mathbf{1}_{\{\epsilon\sqrt{n} < |X_1| \le n\}}) \right] \\
&= \epsilon^2 + 0 = \epsilon^2
\end{aligned}$$

For situation 3, define $t(x) := xP(|X_1| > x)$. We have

$$\lim_{n\to\infty} \sum_{i=1}^n P(|X_i| > n) = \lim_{n\to\infty} nP(|X_1| > n) = \lim_{x\to\infty} t(x) = 0$$

and

$$\begin{aligned}
\forall n \in \mathbb{N}^*,\ \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}(X_i^2 \mathbf{1}_{\{|X_i| \le n\}}) &= \frac{1}{n} \int_{[-n,n]} x^2 \mathrm{d}P \circ X_1^{-1}(x) \int_{[-n,n]} x^2 \mathrm{d}P \circ X_1^{-1}(x) \\
&= \frac{1}{n} \int_{[-n,n]} \int_0^{|x|} (2y) \mathrm{d}y \mathrm{d}P \circ X_1^{-1}(x) \\
&= \frac{1}{n} \int_0^n 2y \int_{[-n,-y)\cup(y,n]} 1 \mathrm{d}P \circ X_1^{-1}(x) \text{ by Fubini's theorem} \\
&= \frac{1}{n} \int_0^n 2y[P(|X_1| > y) - P(|X_1| > n)] \mathrm{d}y \\
&= \frac{1}{n} \int_0^n 2t(y) \mathrm{d}y - \frac{1}{n} P(|X_1| > n) \int_0^n 2y \mathrm{d}y \\
&= \frac{2}{n} \int_0^n t(y) \mathrm{d}y - t(n)
\end{aligned}$$

Since $0 \le t(x) \le x\ \forall x \in [0, \infty)$ and $\lim_{x\to\infty} t(x) = 0$, for any $\epsilon > 0$, $\exists M > 0$ such that $|t(x)| < \epsilon$, then

$$\forall n \ge M,\ \frac{1}{n} \int_0^n t(y) \mathrm{d}y = \frac{1}{n} \left( \int_0^M t(y) \mathrm{d}y + \int_M^n t(y) \mathrm{d}y \right) \le \frac{1}{n} \left( \int_0^M y \mathrm{d}y + \int_M^n \epsilon \mathrm{d}y \right) = \epsilon + \frac{M^2 - 2M}{2n}$$

and thus,

$$0 \le \lim_{n\to\infty} \frac{1}{n} \int_0^n t(y) \mathrm{d}y \le \epsilon\ \forall \epsilon > 0 \ \Rightarrow\ \lim_{n\to\infty} \frac{1}{n} \int_0^n t(y) \mathrm{d}y = 0$$

which gives

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}(X_i^2 \mathbf{1}_{\{|X_i| \le n\}}) = \lim_{n\to\infty} \frac{2}{n} \int_0^n t(y) \mathrm{d}y - \lim_{n\to\infty} t(n) = 0$$

## 6.2 Almost surely convergence

### 6.2.1 Basics

We have shown the definition of almost surely convergence in 3.1.2. We state it again in the context of random variable, random vector and random sequence.

**Definition**: almost surely convergence

$X_n \xrightarrow{\text{a.s.}} X$ if there exist $S \in \Omega$ such that $P(S^c) = 0$ and $\lim_{n\to\infty} X_n(\omega) = X(\omega)\ \forall \omega \in S$. Or we can simply write

$$P\left(\left\{\lim_{n\to\infty} X_n = X\right\}\right) = 1$$

If $\{\mathbf{X}_n\}_{n=1}^{\infty}$ and $\mathbf{X}$ are finite random vectors (map to $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$) or sequences (map to $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$). We can also define $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$ to be $X_n := d(\mathbf{X}_n, \mathbf{X}) \xrightarrow{\text{a.s.}} 0$, where $d$ is the distance function in the corresponding metric space. $\{X_n\}_{n=1}^{\infty}$ are random variables because $d$ is continuous.

Just like the previous case, it is also easy to check that almost surely convergence satisfies the additivity property. Other properties are discussed in 6.2.2 together with convergence in probability.

**Proposition**: equivalent condition for almost surely convergence

In the case of $X$ being finite almost everywhere, $X_n \xrightarrow{\text{a.s.}} X$ is equivalent to the following statements:

1. $\forall \epsilon > 0$,
$$P\left(\liminf_{n\to\infty} \{|X_n - X| < \epsilon\}\right) = 1 \text{ or equivalently, } P\left(\limsup_{n\to\infty} \{|X_n - X| \geq \epsilon\}\right) = 0$$
2. $\forall \epsilon > 0$,
$$\lim_{n\to\infty} P\left(\bigcap_{k\geq n} \{|X_k - X| < \epsilon\}\right) = 1 \text{ or equivalently, } \lim_{n\to\infty} P\left(\bigcup_{k\geq n} \{|X_k - X| \geq \epsilon\}\right) = 0$$
3. $\forall \epsilon > 0$, $\exists A \in \mathcal{F}$ such that $P(A) < \epsilon$ and $X_n$ converge to $X$ uniformly on $A^c$, i.e.
$$\lim_{n\to\infty} \sup_{\omega\in A^c} |X_n(\omega) - X(\omega)| = 0$$
which is the definition of almost uniformly convergence.
4. (Almost surely Cauchy condition) $\{X_n\}_{n=1}^{\infty}$ is a Cauchy sequence almost surely, which means there exists $S \in \mathcal{F}$ such that $P(S^c) = 0$ and
$$\lim_{n\to\infty} \left(\sup_{i,j\geq n} |X_i(\omega) - X_j(\omega)|\right) = 0\ \forall \omega \in S$$

Proof:
The following proof only consider the subset $\Omega'$ such that $X$ is finite every where on $S$ and $P(\Omega'^c) = 0$. It does not affect the probability when taking the union of $\Omega'^c$.
To prove 1, we first have

$$\liminf_{n\to\infty} \{|X_n - X| < \epsilon\} = \{\omega \in \Omega : \exists n_0 \text{ such that } |X_n(\omega) - X(\omega)| < \epsilon \; \forall n > n_0\}$$

Thus, for any $\omega \in \{\lim_{n\to\infty} X_n = X\}$, we have $\forall \epsilon > 0$, $\omega \in \liminf_{n\to\infty} \{|X_n - X| < \epsilon\}$ and thus $\{\lim_{n\to\infty} X_n = X\} \subset \liminf_{n\to\infty} \{|X_n - X| < \epsilon\}$. Therefore, $X_n \xrightarrow{\text{a.s.}} X$ gives our new condition.
We then prove the other direction. $\forall \epsilon > 0$, let $S_\epsilon := \liminf_{n\to\infty} \{|X_n - X| < \epsilon\}$ and $S := \bigcup_{\epsilon \in \mathbb{R}+} S_\epsilon^c$. Similar to the above example, we have $S_{\epsilon_2}^c \subset S_{\epsilon_1}^c$ whenever $\epsilon_1, \epsilon_2 \in \mathbb{R}$ and $\epsilon_1 < \epsilon_2$. Thus, $S = \bigcup_{\epsilon \in \mathbb{Q}+} S_\epsilon^c \in \mathcal{F}$ and then $P(S) = 0$. We can prove that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for any $\omega \in S^c$.
2 then follows because $\forall \epsilon > 0$,

$$P\left(\liminf_{n\to\infty} \{|X_n - X| < \epsilon\}\right) = P\left(\bigcup_{n=1}^{\infty} \bigcap_{k\geq n} \{|X_k - X| < \epsilon\}\right) = \lim_{n\to\infty} P\left(\bigcap_{k\geq n} \{|X_k - X| < \epsilon\}\right)$$

by continuity from below.
We further show that 2 and 3 are equivalent. From 2, for any $\delta > 0$, and $\epsilon > 0$, there exists $N \in \mathbb{N}^*$ such that $\forall n \geq N$,

$$P\left(\bigcup_{k\geq n} \{|X_k - X| \geq \epsilon\}\right) < \delta$$

For any $\delta > 0$ and $\epsilon > 0$, take the above $N_\epsilon$ and let $A_\epsilon := \bigcup_{k\geq N_\epsilon} \{|X_k - X| \geq \epsilon\}$ such that $P(A_\epsilon) < \delta$. We claim that

$$\lim_{n\to\infty} \sup_{\omega\in A^c} |X_n(\omega) - X(\omega)| = 0$$

where $A = \bigcup_{\epsilon\in\mathbb{R}+} A_\epsilon$. Because for any $\epsilon > 0$, there exist $N_\epsilon$ and $A_\epsilon$ such that $\forall \omega \in A^c$, we have $\omega \in {A_\epsilon}^c = \bigcap_{k\geq N_\epsilon} \{|X_k - X| < \epsilon\}$, which means $|X_k(\omega) - X(\omega)| < \epsilon \; \forall k \geq N$.
To prove $P(A) < \delta$, let $\{a_i \in \mathbb{Q}+\}_{i=1}^{\infty}$ to be a countable sequence that converges to 0. Then similar to above, $\{A_{a_i}\}_{i=1}^{\infty}$ is monotonic increasing. Thus,

$$A = \bigcup_{i=1}^{\infty} A_{a_i,n} \in \mathcal{F}, \quad \text{and } P(A) = \lim_{i\to\infty} P(A_{a_i,n}) < \delta$$

The other directly, from 3 to 2, is more simple and omitted.
4 is immediately given from the equivalence between convergent sequence and Cauchy sequence, details shown in 1.1.3.

We then use the following zero-one law to discuss more equivalent statements of almost surely convergence. Zero-one law is a type of law that shows the probability of an event going to either 1 or 0.

**Theorem**: Borel zero-one law

Let $(\Omega, \mathcal{F}, P)$ be a probability measure and $\{A_n \in \mathcal{F}\}_{n=1}^{\infty}$ be a sequence of mutually independent event events, then

$$P\left(\limsup_{n\to\infty} A_n\right) = \begin{cases} 0 & , \sum_{n=1}^{\infty} P(A_n) < \infty \\ 1 & , \sum_{n=1}^{\infty} P(A_n) = \infty \end{cases}$$

Proof:

First, suppose $\sum_{n=1}^{\infty} P(A_n) < \infty$. We have

$$\begin{aligned} P\left(\limsup_{n\to\infty} A_n\right) &= P\left(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k\right) \\ &= \lim_{n\to\infty} P\left(\bigcup_{k=n}^{\infty} A_k\right) \quad \text{(since } \left\{\bigcup_{k=n}^{\infty} A_k\right\} \text{ is monotonic decreasing)} \\ &\le \lim_{n\to\infty} \sum_{k=n}^{\infty} P(A_k) \\ &= \sum_{n=1}^{\infty} P(A_n) - \lim_{n\to\infty} \sum_{k=1}^{n-1} P(A_k) \text{ (since } \sum_{n=1}^{\infty} P(A_n) \text{ converges)} \\ &= 0 \end{aligned}$$

Then, if $\sum_{n=1}^{\infty} P(A_n) = \infty$, we need to prove $P\left(\limsup_{n\to\infty} A_n\right) = 1$. Since

$$P\left(\limsup_{n\to\infty} A_n\right) = 1 - P\left(\liminf_{n\to\infty} A_n^c\right) = 1 - \lim_{n\to\infty} P\left(\bigcap_{k=n}^{\infty} A_k^c\right) \text{ (similar to above)}$$

it suffices to prove that for any $n \in \mathbb{N}^*$, $P\left(\bigcap_{k=n}^{\infty} A_k^c\right) = 0$. We first have

$$P\left(\bigcap_{k=n}^{\infty} A_k^c\right) = P\left(\bigcap_{m=n}^{\infty}\bigcap_{k=n}^{m} A_k^c\right) = \lim_{m\to\infty} P\left(\bigcap_{k=n}^{m} A_k^c\right) = \lim_{m\to\infty} \prod_{k=n}^{m} [1 - P(A_k)]$$

because of monotonicity and independence. Applying the following inequality

$$1 - x \le \mathrm{e}^{-x} \ \forall 0 \le x \le 1$$

we obtain

$$\forall m \ge n,\ \prod_{k=n}^{m} [1 - P(A_k)] \le \prod_{k=n}^{m} \mathrm{e}^{P(A_k)} = \exp\left(-\sum_{k=n}^{m} P(A_k)\right)$$

Then $\lim_{m\to\infty}\prod_{k=n}^{m}[1-P(A_k)]=0$ because

$$\prod_{k=n}^{m}[1-P(A_k)]\geq 0$$

and

$$\lim_{m\to\infty}\exp\left(-\sum_{k=n}^{m}P(A_k)\right)=\lim_{x\to-\infty}e^x=0$$

**Corollary**: Conditions for almost surely convergence using the Borel zero-one law
In the case of $X$ being finite almost everywhere, $X_n\xrightarrow{\text{a.s.}}X$ if and only if for any $\epsilon>0$, $\sum_{n=1}^{\infty}P(|X_n-X|\geq\epsilon)<\infty$.
This follows from the proposition 1 and the Borel zero-one law.

When $X$ being infinite, the Borel zero-one law is also helpful in this example.
**Example**:
If $\{X_n\}_{n=1}^{\infty}$ is *i.i.d.* and their shared *pdf* satisfied $F(x)<1\ \forall x\in\mathbb{R}$, then

$$M_n:=\sup\{X_i\}_{i=1}^{n}\xrightarrow{\text{a.s.}}\infty$$

Proof:
Because of independence,

$$\forall x\in\mathbb{R},\ P(M_n\leq x)=P(X_1\leq x,X_2\leq x,\cdots,X_n\leq x)=\prod_{i=1}^{n}P(X_i\leq x)=F(x)^n$$
$$\Rightarrow\sum_{i=1}^{\infty}P(M_i\leq x)=\sum_{i=1}^{\infty}F(x)^i=\frac{F(x)}{1-F(x)}<\infty$$

By the Borel zero-one law, $P\left(\limsup_{i\to\infty}\{M_i\leq x\}\right)=0$, and thus

$$P\left(\liminf_{i\to\infty}\{M_i>x\}\right)=P\left[\left(\limsup_{i\to\infty}\{M_i\leq x\}\right)^c\right]=1-P\left(\limsup_{i\to\infty}\{M_i\leq x\}\right)=1\ \forall x\in\mathbb{R}$$

For any $x\in\mathbb{R}$, let $S_x:=\liminf_{i\to\infty}\{M_i>x\}$ such that $P(S_x^c)=0$. By the property of limit infimum, $\forall\omega\in S_x$, $\exists n_{x,\omega}\in\mathbb{N}^*$ such that $M_i(\omega)>x\ \forall i>n_{x,\omega}$.
Let $S:=\bigcup_{x\in\mathbb{R}}S_x^c$. Then, for any $\omega\in S^c$ and any $x\in\mathbb{R}$, we have $\omega\in S_x$ and therefore $\exists n_{x,\omega}\in\mathbb{N}^*$ such that $M_i(\omega)>x\ \forall i>n_{x,\omega}$, which gives $\lim_{i\to\infty}M_i(\omega)=\infty\ \forall\omega\in S^c$.
We still need to show $S\in\mathcal{F}$. We notice that whenever $x,y\in\mathbb{R}$ and $x<y$, $\{M_i>x\}\supset\{M_i>y\}$ and thus $S_x\supset S_y$, $S_x^c\subset S_y^c$. Therefore,

$$S=\bigcup_{x\in\mathbb{R}}S_x^c=\bigcup_{i\in\mathbb{Z}}\bigcup_{i-1\leq x<i}S_x^c=\bigcup_{i\in\mathbb{Z}}S_i^c\text{ is the countable union of sets }\Rightarrow\ S\in\mathcal{F}$$

We also have $P(S)\leq\sum_{i\in\mathbf{Z}}P(S_i^c)=0$, thus finishing the proof.

### 6.2.2 Connection with convergence in probability

**Theorem**: Almost surely convergence implies convergence in probability
If $X_n \xrightarrow{\text{a.s.}} X$ and $X$ are finite almost everywhere, then $X_n \xrightarrow{\text{p.}} X$.
Proof:

$$\begin{aligned}\forall \epsilon > 0,\ 0 &= \lim_{n\to\infty} P\left(\bigcup_{k\geq n}\{|X_k - X| \geq \epsilon\}\right) \text{ (from the proposition)}\\ &\geq \lim_{n\to\infty} P\left(\{|X_n - X| \geq \epsilon\}\right) \geq 0\end{aligned}$$

Thus, $\lim_{n\to\infty} P\left(\{|X_n - X| \geq \epsilon\}\right) = 0$ and $X_n \xrightarrow{\text{p.}} X$.

**Theorem**: equivalence of two types of convergences for monotone sequence
If $\{X_n\}_{n=1}^{\infty}$ is an either monotonic increasing or decreasing sequence of random variables, then $X_n \xrightarrow{\text{p.}} X$ implies $X_n \xrightarrow{\text{a.s.}} X$, which means they are equivalent.
Proof:
The monotonicity gives either $X \overset{a.s.}{\geq} X_n$ (monotonic increasing case) or $X \overset{a.s.}{\leq} X_n$ (monotonic decreasing case) for any $n \in \mathbb{N}^*$, which can be proved by contradiction. Then, For any $\epsilon > 0$,

$$P(|X_n - X| < \epsilon) = P\left(\bigcap_{k\geq n}\{|X_n - X| < \epsilon\}\right)$$

Thus, $\lim_{n\to\infty} P\left(\bigcap_{k\geq n}\{|X_n - X| < \epsilon\}\right) = 1$ and $X_n \xrightarrow{\text{a.s.}} X$.

**Proposition**: connections between the two types of convergence
Given a sequence of finite variables $\{X_n\}_{n=1}^{\infty}$, we say it satisfies the Cauchy in probability condition if and only if for any $\epsilon > 0$ and $\delta > 0$, there exists $N \in \mathbb{N}^*$ such that

$$\forall n, m \in \mathbb{N}^* \text{ and } n, m \geq N,\ P(|X_n - X_m| > \epsilon) < \delta$$

Then we have the following:

1. If $\{X_n\}_{n=1}^{\infty}$ satisfy the Cauchy in probability condition, then there exists a sub-sequence $\{X_{n_i}\}_{i=1}^{\infty}$ such that $\{X_{n_i}\}$ converges almost surely to a random variable $X$.
2. $\{X_n\}_{n=1}^{\infty}$ satisfies the Cauchy in probability condition if and only if it converges in probability to a random variable: $X_n \xrightarrow{\text{p.}} X$

3. 1 and 2 together give that if $X_n \xrightarrow{\text{p.}} X$, their exists a sub-sequence of $\{X_n\}_{n=1}^{\infty}$ that converges almost surely to $X$. Further, $X_n \xrightarrow{\text{p.}} X$ if and only if for every sub-sequence $\{X_{n_i}\}_{i=1}^{\infty}$, there exists a further sub-sequence that converges almost surely to $X$:

$$\{X_{n_{i_j}}\}_{j=1}^{\infty} \xrightarrow{\text{a.s.}} X, \text{ where } X \text{ is finite almost everywhere}$$

Proof:
We show the construct of the subset that satisfies statement 1. Let $n_0 = 0$ and for index $i \geq 1$,

$$n_i := \inf\{N \in \mathbb{N}^* : N > n_{i-1} \text{ and } \forall n \geq N,\ P(|X_n - X_N| > 2^{-i}) < 2^{-i}\}$$

From the definition, the infimum must exist and be finite, and this makes $P(|X_{n_{i+1}} - X_{n_i}| > 2^{-i}) < 2^{-i}\ \forall i \in \mathbb{N}^*$. Thus,

$$\sum_{i=1}^{\infty} P(|X_{n_{i+1}} - X_{n_i}| > 2^{-i}) < \sum_{i=1}^{\infty} 2^{-i} < \infty$$

By the Borel zero-one law, $P(\limsup_{i\to\infty}\{|X_{n_{i+1}} - X_{n_i}| > 2^{-i}\}) = 0$. Let $S = \limsup_{i\to\infty}\{|X_{n_{i+1}} - X_{n_i}| > 2^{-i}\}$. For any $\omega \in S^c$, there exists $N$ such that $|X_{n_{i+1}}(\omega) - X_{n_i}(\omega)| \leq 2^{-i}\ \forall i \geq N$, which gives

$$\begin{aligned}
&\forall i > N,\ |X_{n_i}(\omega) - X_{n_j}(\omega)| \leq \sum_{k=i}^{\infty} |X_{n_{i+1}}(\omega) - X_{n_i}(\omega)| \leq \sum_{k=i}^{\infty} 2^{-k} = 2^{1-i}\ \forall j > i \\
&\Rightarrow \begin{cases} \sup_{j>i} X_{n_j}(\omega) \leq X_{n_i}(\omega) + 2^{1-i} \\ \inf_{j>i} X_{n_j}(\omega) \geq X_{n_i}(\omega) - 2^{1-i} \end{cases} \\
&\Rightarrow \sup_{j>i} X_{n_j}(\omega) - 2^{1-i} \leq X_{n_i}(\omega) \leq \inf_{j>i} X_{n_j}(\omega) + 2^{1-i}
\end{aligned}$$

Since

$$\lim_{i\to\infty}\left[\sup_{j>i} X_{n_j} - 2^{1-i}(\omega)\right] = \limsup_{i\to\infty} X_{n_i}(\omega)$$

$$\lim_{i\to\infty}\left[\inf_{j>i} X_{n_j}(\omega) + 2^{1-i}\right] = \liminf_{i\to\infty} X_{n_i}(\omega)$$

both exist and $\liminf_{i\to\infty} X_{n_i}(\omega) \leq \limsup_{i\to\infty} X_{n_i}(\omega)$, we have

$$\liminf_{i\to\infty} X_{n_i}(\omega) = \lim_{i\to\infty} X_{n_i}(\omega) = \limsup_{i\to\infty} X_{n_i}(\omega)$$

which means $\{X_{n_i}(\omega)\}_{i=1}^{\infty}$ converges for all $\omega \in S^c$. Thus, $\lim_{i\to\infty} X_{n_i}$ exist almost surely.
Next, we prove the $\Rightarrow$ direction of 2. Continuing from 1, we claim that $\{X_{n_i}\}_{i=1}^{\infty} \xrightarrow{\text{a.s.}} X$ gives $X_n \xrightarrow{\text{p.}} X$, where $\{n_i\}_{i=1}^{\infty}$ is what we have constructed. For $\epsilon > 0$, we first have

$$\begin{aligned}
&\forall n, j \in \mathbb{N}^*,\ \{|X_n - X| > \epsilon\} \subset \left\{|X_n - X_{n_i}| > \frac{\epsilon}{2}\right\} \cup \left\{|X - X_{n_i}| > \frac{\epsilon}{2}\right\} \\
&\qquad \Rightarrow P(|X_n - X| > \epsilon) \leq P(|X_n - X_{n_i}| > \frac{\epsilon}{2}) + P(|X - X_{n_i}| > \frac{\epsilon}{2})
\end{aligned}$$

Therefore, for any $\epsilon, \delta > 0$, Cauchy in probability gives that there exists a big enough $N_1$ such that $P(|X_n - X_{n_i}| > \frac{\epsilon}{2}) < \frac{\delta}{2}$ whenever $n, n_i \geq N_1$. And since $\{X_{n_i}\}_{i=1}^{\infty} \xrightarrow{\text{a.s.}} X$ gives $\{X_{n_i}\}_{i=1}^{\infty} \xrightarrow{\text{p.}} X$, there exists a big enough $i_0$ such that $P(|X - X_{n_i}| > \frac{\epsilon}{2}) < \frac{\delta}{2}$ whenever $i \geq i_0$. Pick $N = \max(N_1, n_{i_0})$, we then have $P(|X_n - X| > \epsilon) < \delta \ \forall n \geq N$, which gives $\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0 \ \forall \epsilon > 0$.
We continue to prove its inverse. Since for $\epsilon > 0$,

$$\forall n, m \in \mathbb{N}^*,\ P(|X_n - X_m| > \epsilon) \leq P(|X_n - X| > \frac{\epsilon}{2}) + P(|X_m - X| > \frac{\epsilon}{2})$$

Given $X_n \xrightarrow{\text{p.}} X$, we can pick a large enough $N$ for any $\delta > 0$ such that $\forall n, m \geq N$, $P(|X_n - X| > \frac{\epsilon}{2}) < \frac{\delta}{2}$ and $P(|X_m - X| > \frac{\epsilon}{2}) < \frac{\delta}{2}$. Thus it gives $P(|X_n - X_m| > \epsilon) < \delta$ and the sequence satisfies the Cauchy in probability condition.
Finally, we prove statement 3. Because any sub-sequence of $\{X_n\}_{n=1}^{\infty}$ converges to $X$ in probability, which at the same time satisfies the Cauchy in probability condition, we have shown that it has a further sub-sequence converging almost surely to $X$.
We lastly prove the inverse by contradiction. Suppose $X_n \xrightarrow{\text{p.}} X$ does not hold, there exist $\epsilon_0 > 0$ and a sub-sequence $\{X_{n_i}\}_{i=1}^{\infty}$ such that $\lim_{i\to\infty} P(|X_{n_i} - X| > \epsilon_0)$ does not exists or not equal to 0, which means $\exists \delta_0 > 0$ such that $P(|X_{n_i} - X| > \epsilon_0) > \delta_0$ for infinitely many $i \in \mathbb{N}^*$. Then, for any sub-sequence of $\{X_{n_i}\}_{i=1}^{\infty}$,

$$P\left(\bigcup_{k\geq j}\left\{|X_{n_{i_k}} - X| > \epsilon_0\right\}\right) > P\left(\left\{|X_{n_{i_j}} - X| > \epsilon_0\right\}\right)$$
$$\text{does not converges to } 0\ \forall j \in \mathbb{N}^*, \text{ where } \{i_j\}_{j=1}^{\infty} \text{ is a index set}$$

By the equivalent condition for almose surely convergence,

$$\{X_{n_{i_j}}\}_{j=1}^{\infty} \xrightarrow{\text{a.s.}} X$$

does not hold.

**Continuous mapping theorem** (for almost surely convergence and convergence in probability)
Let $f : \bar{\mathbb{R}} \mapsto \bar{\mathbb{R}}$ be a continuous function.

1. If $X_n \xrightarrow{\text{a.s.}} X$, then $f(X_n) \xrightarrow{\text{a.s.}} f(X)$
2. If $X_n \xrightarrow{\text{p.}} X$, then $f(X_n) \xrightarrow{\text{p.}} f(X)$

Proof:
We have proved the surely case in 1.3.2. The almost surely case directly follows by picking a subset

with zero measure.

To prove 2, we make use of the equivalent condition of convergence in probability that for any sub-sequence of $\{X_n\}_{n=1}^{\infty}$, there exists a further sub-sequence converging almost surely to $X$. Since such further sub-sequence satisfies the continuous mapping, so do $\{X_n\}_{n=1}^{\infty}$.

**Proposition**: Multiplicability of almost surely convergence and convergence in probability

Let $X$, $Y$, $\{X_n\}_{n=1}^{\infty}$ and $\{Y_n\}_{n=1}^{\infty}$ to be finite random variables.

1. If $X_n \xrightarrow{\text{a.s.}} X$ and $Y_n \xrightarrow{\text{a.s.}} Y$, then $\{X_nY_n\}_{n=1}^{\infty} \xrightarrow{\text{a.s.}} XY$.
2. If $X_n \xrightarrow{\text{p.}} X$ and $Y_n \xrightarrow{\text{p.}} Y$, and $\{X_nY_n\}_{n=1}^{\infty} \xrightarrow{\text{p.}} XY$.

We first proof the almost surely case.

Take $S \in \mathcal{F}$ such that $P(S^c) = 0$ and $\{X_n\}_{n=1}^{\infty}$, $\{Y_n\}_{n=1}^{\infty}$ converge to $X, Y$ on $S$. For any $\epsilon > 0$, we need to take a large enough $n_0$ such that

$$|X_n(\omega)Y_n(\omega) - X(\omega)Y(\omega)| < \epsilon \ \forall \omega \in S, n \geq n_0$$

Since

$$0 \leq |X_n(\omega)Y_n(\omega) - X(\omega)Y(\omega)| \leq |X_n(\omega)||Y_n(\omega) - Y(\omega)| + |Y(\omega)||X_n(\omega) - X(\omega)|$$

This is possible because we take $n_1, n_2$ and $n_0 = \max(n_1, n_2)$ such that

$$|Y_n(\omega) - Y(\omega)| < \frac{\epsilon}{2\max\{|X_n(\omega)| : n \in \mathbb{N}^*\}} \ \forall \omega \in S, n \geq n_1$$

and

$$|X_n(\omega) - X(\omega)| < \frac{\epsilon}{2|Y(\omega)|} \ \forall \omega \in S, n \geq n_1$$

We omit the discussion of the trivial case when $\max\{|X_n(\omega)| : n \in \mathbb{N}^*\} = 0$.

For the convergence in probability case, it suffices to show that any sub-sequence, $\{X_{n_i}Y_{n_i}\}_{i=1}^{\infty}$, has a further sub-sequence that converges to $XY$ almost surely. We pick the index $\{i_j\}_{j=1}^{\infty}$ satisfying

$$\{X_{n_{i_j}}\}_{j=1}^{\infty} \xrightarrow{\text{a.s.}} X \text{ and } \{Y_{n_{i_j}}\}_{j=1}^{\infty} \xrightarrow{\text{a.s.}} Y$$

This is possible because we can first find a sub-sequence that satisfies the first convergence. Then we find a further sub-sequence that satisfies the other. It gives

$$\{X_{n_{i_j}}Y_{n_{i_j}}\}_{j=1}^{\infty} \xrightarrow{\text{a.s.}} XY$$

and thus $\{X_nY_n\}_{n=1}^{\infty} \xrightarrow{\text{p.}} XY$.

### 6.2.3 Strong laws of large numbers

We state the theorem with examples in the beginning. The goal of this section is to prove it.

**Strong laws of large numbers**

If $\{X_n\}_{n=1}^{\infty}$ are independent and $S_n := \sum_{i=1}^{n} X_n$ for $n \in \mathbb{N}^*$, the following are two types of strong laws of large numbers with different conditions.

1. (A more general strong law) When $X_n \in \mathcal{L}^1$ $\forall n \in \mathbb{N}^*$ and there exists a monotonic increasing sequence $\{c_n \in \mathbb{R}^+\}_{n=1}^{\infty}$ such that

$$\lim_{n\to\infty} c_n = \infty \text{ and } \sum_{n=1}^{\infty} \operatorname{var}\left(\frac{X_n}{c_n}\right) < \infty$$

   then

$$\left\{\frac{S_n - \mathbb{E}(S_n)}{c_n}\right\}_{n=1}^{\infty} \xrightarrow{\text{a.s.}} 0$$

2. (Kolmogorov's strong law) When $\{X_n\}_{n=1}^{\infty}$ are *i.i.d.*, then $X_1 \in \mathcal{L}^1$ gives

$$\left\{\frac{S_n}{n}\right\}_{n=1}^{\infty} \xrightarrow{\text{a.s.}} \mathbb{E}(X_1)$$

   Conversely, given the *i.i.d.* case, if $\left\{\frac{S_n}{n}\right\}_{i=1}^{\infty}$ converges almost surely to a real constant $c$, then $X_1 \in \mathcal{L}^1$ and $c = \mathbb{E}(X_1)$.

Their proofs are based on the following lemmas.

**Lemma**: Skorohod's inequality

Suppose $\{X_n\}_{n=1}^{\infty}$ are independent and $a \in \mathbb{R}^+$ is fixed. Let $S_n := \sum_{i=1}^{n} X_n$ and $c(n) := \sup_{1\le i\le n} P(|S_n - S_i| > \alpha)$ for $n \in \mathbb{N}^*$. If $c(n) < 1$ for some $n \in \mathbb{N}^*$, then

$$P\left(\left\{\sup_{1\le i\le n} |S_i| > 2a\right\}\right) \le \frac{1}{1-c(n)} P\left(|S_n| > a\right)$$

Specially, when $\{X_n\}_{n=1}^{\infty}$ are *i.i.d*, $c(n) = \sup_{1\le i\le n} P(|S_i| > \alpha)$. Proof:

Let $Y := \inf\{i : |S_i| > 2a\}$ be a (discrete) random variable. For all $n \in \mathbb{N}^*$, we have

$$\forall 1 \le i \le n,\ \{|S_n - S_i| \le a, Y = i\} \subset \{|S_n| > a, Y = i\}$$

and

$$\forall n \in \mathbb{N}^*,\ \left\{\sup_{1\le i\le n} |S_i| > 2a\right\} = \{Y \le n\} = \bigcup_{i=1}^{n} \{Y = i\}$$

Then,

$$
\begin{aligned}
P(|S_n > a|) &\geq \sum_{i=1}^{n} P(\{|S_n| > a, Y = i\}) \\
&\geq \sum_{i=1}^{n} P(\{|S_n - S_i| \leq a, Y = i\}) \\
&\geq \sum_{i=1}^{n} P(|S_n - S_i| \leq a) P(Y = i) \text{ because of independence} \\
&= \sum_{i=1}^{n} [1 - P(|S_n - S_i| \geq a)] \times P(Y = i) \\
&\geq \sum_{i=1}^{n} [1 - c(n)] \times P(Y = i) \\
&= [1 - c(n)] P(Y \leq n) \\
&= [1 - c(n)] P\left(\left\{\sup_{1 \leq i \leq n} |S_i| > 2a\right\}\right)
\end{aligned}
$$

Note that the independence comes from that $\{|S_n - S_i| \leq a\}$ is an event on $\sigma\left(\{X_j\}_{j=i+1}^{n}\right)$, while $\{Y = i\} = \{S_i \geq 2a \text{ and } S_j < 2a \ \forall 1 \leq j < i\}$ is an event on $\sigma\left(\{X_j\}_{j=1}^{i}\right)$.

**Lemma**: equivalence of convergence in probability and almost surely convergence for the sum of independent random variables (Lévy's theorem)

Suppose $\{X_n\}_{n=1}^{\infty}$ are independent. Let $S_n := \sum_{i=1}^{n} X_n$, then $S_n \xrightarrow{\text{p.}} S$ gives $S_n \xrightarrow{\text{a.s.}} S$, which means they are equivalent.

Proof:

It suffices to prove that if $\{X_n\}_{n=1}^{\infty}$ satisfies the Cauthy in probability condition, then it also satisfies the almost surely Cauthy condition, which is

$$
Y_n := \sup_{i,j \geq n} |S_i - S_j| \xrightarrow{\text{a.s.}} 0
$$

Since $\{Y_n\}_{n=1}^{\infty}$ is monotonic decreasing, we only need to show $Y_n \xrightarrow{\text{p.}} 0$. We also have

$$
\forall n \in \mathbb{N}^*,\ Y_n = \sup_{i,j \geq n} |S_i - S_n + S_n - S_j| \leq \sup_{i \geq n} |S_i - S_n| + \sup_{j \geq n} |S_j - S_n| = 2 \sup_{i \geq 0} |S_{n+i} - S_n|
$$

Thus, it suffices to prove $Z_n := \sup_{i \geq 0} |S_{n+i} - S_n| \xrightarrow{\text{p.}} 0$. For any $\epsilon > 0$, we further have $\forall n \in \mathbb{N}^*$,

$$
P(|Z_n - 0| > \epsilon) = P\left(\bigcup_{i=1}^{\infty} \left\{\sup_{0 \leq j \leq i} |S_{n+j} - S_n| > \epsilon\right\}\right) = \lim_{i \to \infty} P\left(\left\{\sup_{0 \leq j \leq i} |S_{n+j} - S_n| > \epsilon\right\}\right)
$$

The following proves that for any small $\delta > 0$, there exists $N \in \mathbb{N}^*$ such that

$$
P\left(\left\{\sup_{0 \leq j \leq i} |S_{n+j} - S_n| > \epsilon\right\}\right) < \delta \ \forall i \in \mathbb{N}^*, n \geq N
$$

and thus $P(|Z_n - 0| > \epsilon) < \delta\ \forall n \geq N$.
First, for any $\epsilon > 0$ and $\delta > 0$, Cauthy in probability gives that there exists $N_0 \in \mathbb{N}^*$ such that

$$P\left(|S_{n+i} - S_i| > \frac{\epsilon}{2}\right) \leq \delta\ \forall n \geq N_0, i \in \mathbb{N}^*$$

For any $n \geq N_0$, let $X'_i = X_{n+i}$ and $S'_i = \sum_{j=1}^{i} X'_i = S_{n+i} - S_n$, then we have

$$\forall j \in \mathbb{N}^*,\ P\left(|S'_j| > \frac{\epsilon}{2}\right) \leq \delta$$

and

$$\forall i \in \mathbb{N}^*,\ c(i) := \sup_{1\leq j\leq i} P\left(|S'_i - S'_j| > \frac{\epsilon}{2}\right) = \sup_{1\leq j\leq i} P\left(|S_{n+i} - S'_{n+j}| > \frac{\epsilon}{2}\right) \leq \delta$$

By Skorohod's inequality, we obtain that $\forall i \in \mathbb{N}^*$,

$$P\left(\left\{\sup_{0\leq j\leq i} |S_{n+j} - S_n| > \epsilon\right\}\right) = P\left(\left\{\sup_{0\leq j\leq i} |S'_j| > \epsilon\right\}\right) \leq \frac{1}{1 - c(i)} \times P\left(|S'_j| > \frac{\epsilon}{2}\right) \leq \frac{\delta}{1-\delta}$$

Since we can take $\delta < \frac{1}{2}$ arbitrarily small, we have

$$\forall i \in \mathbb{N}^*, n \geq N_0,\ P\left(\left\{\sup_{0\leq j\leq i} |S_{n+j} - S_n| > \epsilon\right\}\right) < \frac{1}{1-\delta} \times \delta < 2\delta$$

Reset $\delta' = \frac{\delta}{2}$ and get its corresponding $N_0$, we proved the lemma.

The above lemma is a bridge from weak laws of large numbers to strong laws of large numbers. But since $\{S_n\}_{n=1}^{\infty}$ goes to infinity in most cases, we cannot use the result directly, but need some transformation.
We further have the following lemma.

**Lemma**: Kolmogorov's Convergence Criterion
Suppose $\{X_n \in \mathcal{L}^2\}_{n=1}^{\infty}$ are independent. If $\sum_{i=1}^{\infty} \mathrm{var}(X_i) < \infty$, then

$$\left\{\sum_{i=1}^{n} [X_i - \mathbb{E}(X)]\right\}_{n=1}^{\infty} \text{ converges almost surely}$$

Proof:
For any $n \in \mathbb{N}^*$, we first have $\mathbb{E}(X_n)$ being finite. Let $Y_n := X_n - \mathbb{E}(X_n)$, then $\mathbb{E}(Y_n) = 0$ and $\mathrm{var}(Y_n) = \mathbb{E}(Y_n^2)$. Then, let $S_n := \sum_{i=1}^{n} Y_n\ \forall n \in \mathbb{N}^*$. By the above theorem, it suffices to prove that $\{S_n\}_{n=1}^{\infty}$ satisfies the Cauchy in probability condition.
We use Chebyshev's theorem. For any $\epsilon > 0$,

$$P(|S_n - S_m| > \epsilon) \leq \frac{\mathrm{var}(S_n - S_m)}{\epsilon^2} = \frac{\sum_{i=n}^{m} \mathrm{var}(X_i)}{\epsilon^2}\ \forall n, m \in \mathbb{N}^*, n \leq m$$

Then the Cauchy in probability condition is satisfied because $\sum_{i=1}^{\infty} \text{var}(X_i) < \infty$.

With the above lemma, we are able to prove the first case of law of large numbers.

**Proof of strong laws of large numbers (1)**

First, the Kolmogorov's Convergence Criterion shows that

$$\left\{ \sum_{i=1}^{n} \left[ \frac{X_n}{c_n} - \frac{\mathbb{E}(X_n)}{c_n} \right] \right\} \text{ converges almost surely}$$

because $\sum_{n=1}^{\infty} \text{var}\left(\frac{X_n}{c_n}\right) < \infty$. Take $A \in \mathcal{F}$ be the set on which the above converges and $P(A^c) = 0$. For any $\omega \in A$, let $a_n := X_n(\omega) - \mathbb{E}(X_n)\ \forall n \in \mathbb{N}^*$. We next show that

$$\lim_{n\to\infty} \frac{1}{c_n} \sum_{i=1}^{n} a_i = 0$$

given $\sum_{i=1}^{\infty} \frac{a_i}{c_i}$ converges, $\{c_n \in \mathbb{R}^+\}_{n=1}^{\infty}$ is monotonic increasing and $\lim_{n\to\infty} c_n = \infty$. It thus proves the theorem.

For any $n \in \mathbb{N}^*$, let $r_n := \sum_{i=n}^{\infty} \frac{a_i}{c_i}$ , then

$$\frac{a_n}{c_n} = r_n - r_{n+1} \ \Rightarrow\ a_n = c_n(r_n - r_{n+1})$$

and

$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} c_i(r_i - r_{i+1}) = \sum_{i=1}^{n} r_i(c_i - c_{i-1}) - c_n r_{n+1}$$

(We let $c_0 = 0$ for convenience, also satisfying the monotonicity.) Next, we have $\lim_{n\to\infty} r_n = 0$, and for any $\epsilon > 0$, there exist $N_0$ such that $\forall n \geq N_0$, $|r_n| < \epsilon$, and therefore,

$$\begin{aligned}
\left| \frac{1}{c_n} \sum_{i=1}^{n} a_i \right| &\leq \frac{1}{c_n} \sum_{i=1}^{N_0-1} |r_i|(c_i - c_{i-1}) + \frac{1}{c_n} \sum_{i=N_0}^{n} |r_i|(c_i - c_{i-1}) + \left| \frac{c_n r_{n+1}}{c_n} \right| \\
&< \frac{f(N_0)}{c_n} + \frac{1}{c_n} \sum_{i=N_0}^{n} \epsilon(c_i - c_{i-1}) + |r_{n+1}| \text{ (where } f(N_0) := \sum_{i=1}^{N_0-1} |r_i|(c_i - c_{i-1}) < \infty) \\
&< \frac{f(N_0)}{c_n} + \frac{\epsilon}{c_n} \times (c_n - c_{N_0-1}) + \epsilon \\
&= 2\epsilon + \frac{g(N_0)}{c_n} \text{ (where } g(N_0) := f(N_0) - c_{N_0-1} < \infty \text{ )}
\end{aligned}$$

Since we can retake $\epsilon' = \frac{1}{3}\epsilon$, get the corresponding $N_0'$ and take a large enough $N_1$ such that $\frac{g(N_0')}{c_n} < \epsilon\ \forall n \geq N_1$, it guarantees that $\left| \frac{1}{c_n} \sum_{i=1}^{n} a_i \right| < \epsilon\ \forall n \geq N_1$.

**Proof of strong laws of large numbers (2)**

Now we prove the Kolmogorov's strong law. Define $X_n' := X_n \mathbf{1}_{|X_n| \leq n}$ and $S_n' := \sum_{i=1}^{n} X_n'$ for

$n \in \mathbb{N}^*$. We will prove that *i.i.d.* and $X_1 \in \mathcal{L}^1$ together give $\left\{\frac{S'_n}{n}\right\}_{i=1}^{\infty} \xrightarrow{\text{a.s.}} \mathbb{E}(X_1)$ and then $\left\{\frac{S_n}{n}\right\}_{i=1}^{\infty} \xrightarrow{\text{a.s.}} \mathbb{E}(X_1)$.

First, the strong law of large number (1) gives that

$$\left\{\frac{S'_n - \mathbb{E}(S'_n)}{n}\right\}_{n=1}^{\infty} \xrightarrow{\text{a.s.}} 0 \quad (\text{taking } c_n = n)$$

when

$$\sum_{n=1}^{\infty} \text{var}\left(\frac{X'_n}{n}\right) < \infty$$

which holds because

$$\begin{aligned}
\lim_{N\to\infty} \sum_{n=1}^{N} \text{var}\left(\frac{X'_n}{n}\right) &\leq \lim_{N\to\infty} \sum_{n=1}^{N} \frac{1}{n^2} \mathbb{E}\left[(X'_1)^2\right] \\
&= \lim_{N\to\infty} \sum_{n=1}^{N} \frac{1}{n^2} \mathbb{E}\left[X_1^2 \mathbf{1}_{|X_1|\leq n}\right] \\
&= \lim_{N\to\infty} \sum_{n=1}^{N} \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}\left[X_1^2 \mathbf{1}_{i-1<|X_1|\leq i}\right] \\
&= \lim_{N\to\infty} \sum_{i=1}^{N} \mathbb{E}\left[X_1^2 \mathbf{1}_{i-1<|X_1|\leq i}\right] \sum_{n=i}^{N} \frac{1}{n^2} \\
&= \lim_{N\to\infty} \mathbb{E}\left[X_1^2 \mathbf{1}_{0<|X_1|\leq 1}\right] \left(1 + \sum_{n=2}^{N} \frac{1}{n^2}\right) + \sum_{i=2}^{N} \mathbb{E}\left[X_1^2 \mathbf{1}_{i-1<|X_1|\leq i}\right] \sum_{n=i}^{N} \frac{1}{n^2} \\
&< \lim_{N\to\infty} \mathbb{E}\left[X_1^2 \mathbf{1}_{0<|X_1|\leq 1}\right] \left(1 + \frac{1}{2-1}\right) + \sum_{i=2}^{N} \mathbb{E}\left[X_1^2 \mathbf{1}_{i-1<|X_1|\leq i}\right] \times \frac{1}{i-1} \\
&\qquad (\text{since } \forall i \geq 2,\ \sum_{n=i}^{N} \frac{1}{n^2} < \sum_{n=i}^{\infty} \frac{1}{n^2} < \sum_{n=i}^{\infty} \int_{n-1}^{n} \frac{1}{x^2} \mathrm{d}x = \int_{i-1}^{\infty} \frac{1}{x^2} \mathrm{d}x = \frac{1}{i-1}) \\
&= 2\mathbb{E}\left[X_1^2 \mathbf{1}_{0<|X_1|\leq 1}\right] + \lim_{N\to\infty} \sum_{i=2}^{N} \mathbb{E}\left[X_1^2 \mathbf{1}_{i-1<|X_1|\leq i}\right] \times \frac{2}{i} \ (\text{since } \frac{1}{i-1} \leq \frac{2}{i}\ \forall i \geq 2) \\
&< 2\mathbb{E}\left[|X_1| \mathbf{1}_{0<|X_1|\leq 1}\right] + \lim_{N\to\infty} \sum_{i=2}^{N} \mathbb{E}\left[i|X_1| \mathbf{1}_{i-1<|X_1|\leq i}\right] \times \frac{2}{i} \\
&= 2\mathbb{E}\left[|X_1| \mathbf{1}_{0<|X_1|\leq 1}\right] + \lim_{N\to\infty} \sum_{i=2}^{N} 2\mathbb{E}\left[|X_1| \mathbf{1}_{i-1<|X_1|\leq i}\right] \\
&= 2\mathbb{E}(|X_1|) < \infty
\end{aligned}$$

Then, $\left\{\frac{S'_n}{n} - \mathbb{E}(X_1)\right\}_{i=1}^{\infty} \xrightarrow{\text{a.s.}} 0$ because

$$X'_n \xrightarrow{\text{a.s.}} X_1 \Rightarrow \lim_{n\to\infty} \mathbb{E}(X'_n) = \mathbb{E}(X_1) \ \Rightarrow \ \lim_{n\to\infty} \frac{\mathbb{E}(S'_n)}{n} = \mathbb{E}(X_1) \text{ (since it is a limit of the average)}$$

Further, to obtain $\left\{\frac{S_n}{n}\right\}_{i=1}^{\infty} \xrightarrow{\text{a.s.}} \mathbb{E}(X_1)$, it suffices to prove $\left\{\frac{S_n}{n} - \frac{S'_n}{n}\right\}_{i=1}^{\infty} \xrightarrow{\text{a.s.}} 0$. From the definition of $X'_n$, we have

$$\sum_{n=1}^{\infty} P(X'_n \neq X_n) = \sum_{n=1}^{\infty} P(|X_1| > n) \leq \sum_{n=1}^{\infty} \int_{n-1}^{n} P(|X_1| > x)\mathrm{d}x = \mathbb{E}(|X_1|) < \infty$$

By the Borel zero-one law, $P\left(\liminf_{n\to\infty}\{X'_n = X_n\}\right) = 1$. Let $S := \liminf_{n\to\infty}\{X'_n = X_n\}$. We have $P(S^c) = 1$ and $\forall \omega \in S$, $\exists N_\omega$ such that $X(\omega) = X'(\omega)$. Therefore,

$$\lim_{n\to\infty} \frac{S_n(\omega) - S'_n(\omega)}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} [X(\omega) - X'(\omega)] = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{N_\omega} [X(\omega) - X'(\omega)] = 0 \ \forall \omega \in S$$

which gives $\left\{\frac{S_n}{n} - \frac{S'_n}{n}\right\}_{i=1}^{\infty} \xrightarrow{\text{a.s.}} 0$.
Then we prove the other direction. Suppose $\left\{\frac{S_n}{n}\right\}_{i=1}^{\infty} \xrightarrow{\text{a.s.}} c$ for $c \in \mathbb{R}$. We first have

$$\forall n \geq 2,\ \frac{X_n}{n} = \frac{S_n}{n} - \frac{S_{n-1}}{n} = \left(\frac{S_n}{n} - \frac{S_{n-1}}{n-1}\right) + \left(\frac{S_{n-1}}{n-1} - \frac{S_{n-1}}{n}\right) = \left(\frac{S_n}{n} - \frac{S_{n-1}}{n-1}\right) + \frac{S_{n-1}}{n-1} \times \frac{1}{n}$$

which converges to 0 almost surely when $n$ goes to infinity. The Borel zero-one law then gives

$$\sum_{n=1}^{\infty} P\left(\frac{|X_n|}{n} > \epsilon\right) < \infty$$

Further,

$$\begin{aligned}
\mathbb{E}(|X_1|) &= \int_0^{\infty} P(|X_1| > x)\mathrm{d}x \\
&= \int_0^{\infty} \sum_{n=1}^{\infty} \mathbf{1}_{(n-1,n]}(x) \times P(|X_1| > x)\mathrm{d}x \\
&= \sum_{n=1}^{\infty} \int_{n-1}^{n} P(|X_1| > x)\mathrm{d}x \text{ (by monotone convergence theorem)} \\
&\leq \sum_{n=1}^{\infty} \int_{n-1}^{n} P(|X_1| > n-1)\mathrm{d}x \\
&= \sum_{n=1}^{\infty} P(|X_1| > n-1) \\
&= P(|X_1| > 0) + \sum_{n=1}^{\infty} P\left(\frac{|X_n|}{n} > 1\right) \\
&< 1 + \infty \text{ (by taking } \epsilon = 1)
\end{aligned}$$

To complete the proof, we lastly need to show $c = \mathbb{E}(X_1)$. This is given by the fact that $c \neq \mathbb{E}(X_1)$ violates the weak law of large numbers.

## 6.3 Convergence in mean

**Definition**: Convergence in mean

For a $p \in \mathbb{N}^*$, when the $p$-th moment of $X$ and $X_n$ exist for all $n \in \mathbb{N}^*$, we say $\{X_n\}_{n=1}^{\infty}$ converges in the $p$-th mean (a.k.a. converges in $\mathcal{L}^p$ space) if

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|^p) = 0$$

denoted as

$$X_n \xrightarrow{L^p} X$$

It is necessary to ensure $\mathbb{E}(|X_n - X|^p)$ exists for $n \in \mathbb{N}^*$.

Notice that by $\mathbb{E}(|X_n - X|) \geq |\mathbb{E}(X_n - X)| = |\mathbb{E}(X_n) - \mathbb{E}(X)| \geq 0$, $X_n \xrightarrow{L^1} X$ implies $\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)$.

**Theorem**: convergence in mean gives convergence in probability

If $X_n \xrightarrow{L^p} X$ for some $p \in \mathbb{N}^*$, then $X_n \xrightarrow{\text{p.}} X$.

The proof makes use of Markov's inequality. For any $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) = P(|X_n - X|^p > \epsilon^p) \leq \frac{\mathbb{E}(|X_n - X|^p)}{\epsilon^p}$$

Then $\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$ holds by the Sandwich theorem.

**Theorem**: Dominated convergence theorem gives convergence in $\mathcal{L}^1$ space

If there exists a random variable $Y \in \mathcal{L}^1$ such that $X_n \leq Y \ \forall n \in \mathbb{N}^*$, then both

1. $X_n \xrightarrow{\text{a.s.}} X$
2. $X_n \xrightarrow{\text{p.}} X$

gives $X_n \xrightarrow{L^1} X$ and hence $\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)$.

If $X_n \xrightarrow{\text{a.s.}} X$, then $|X_n - X| \xrightarrow{\text{a.s.}} 0$, $X \leq Y$ and $|X_n - X| \leq 2Y$. The dominated convergence theorem gives

$$\lim_{n\to\infty} \mathbb{E}(X_n - X) = 0$$

We loosen the condition into $X_n \xrightarrow{\text{p.}} X$. It suffices to show that every convergent sub-sequence of $\{\mathbb{E}(|X_n - X|)\}_{n=1}^{\infty}$ converges to 0. Let $\{n_i\}_{i=1}^{\infty}$ to be the index set of a sub-sequence, then there exists a further sub-sequence, say, with index set $\{i_j\}_{j=1}^{\infty}$, that converges almost surely to $X$. 1 gives

$$\lim_{j\to\infty} \mathbb{E}(|X_{n_{i_j}} - X|) = 0$$

Since $\{X_{n_i}\}_{i=1}^{\infty}$ converges,

$$\lim_{i\to\infty} \mathbb{E}(|X_{n_i} - X|) = 0$$

also holds. It thus gives the result.

# Chapter 7: Distribution

## 7.1 Idea of studying distribution

**Definition**: Equation in distribution
Two random variables $X$ (on probability space $(\Omega_1, \mathcal{F}_1, P_1)$) and $Y$ (on $(\Omega_2, \mathcal{F}_2, P_2)$) are equal in distribution if for any $A \in \bar{\mathbb{R}}$, $P_1(X \in A) = P_2(Y \in A)$, written

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

We know that this condition is equivalent to $F_X(x) = F_Y(x)\ \forall x \in \bar{\mathbb{R}}$, where $F_X$, $F_Y$ are their *pdf.*

We write $X \sim D$ meaning $X$ follows a certain distribution $D$. Then if $X \sim D$ and $Y \sim D$, we have $X \overset{d.}{=} Y$. The distribution $D$ can be regarded as a hypothetical random variable $X_D$ on $(\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))$ whose measure $P$ is generated by its *pdf.* $F_D = F_X$.

A distribution $D$ is determined by its *pdf.* or probability density if exists. Some famous distribution can be characterised by some special notions and coefficients. Representations of famous distributions are $U(l, r)$ (uniform distribution), $N(\mu, \sigma^2)$ (normal distribution), $B(n, x)$ (binomial distribution) etc.

If $X \overset{d.}{=} Y$, the following are equal:

1. Their *pdf.*
2. Their expectations, moments and variances (if exist)
3. Their probability density functions (if exist)
4. Their moment generating functions $M_X(t) := \mathbb{E}[e^{tX}]$ if it is finite and defined on $\mathbb{R}$ (to be discussed in 7.3)

5. Their characteristic function (to be discussed in 7.3)

If we are not concerned with the actual probability space, it suffices to study its distribution $D$, whose results apply to any random variable with the same distribution. We can write the notations $\mathbb{E}(D)$, $\mathrm{var}(D)$, $F_D$, $f_D$, $D \overset{a.s.}{\geq} 0$ the same way as random variables. We also write $\mathrm{d}D$ to simply denote $\mathrm{d}P \circ X^{-1}$.

**Example**: Normal distribution
Distribution with *pdf.*

$$F(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma}} \exp \frac{-(t-\mu)^2}{2\sigma^2} \mathrm{d}t$$

or density on the Lebesgue measure space

$$f(x) := \frac{1}{\sqrt{2\pi\sigma}} \exp \frac{-(t-\mu)^2}{2\sigma^2}$$

where $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$ is the normal distribution, written $N(\mu, \sigma^2)$. It possesses the following properties

1. $\mathbb{E}[N(\mu, \sigma^2)] = \mu$
2. $\mathrm{var}[N(\mu, \sigma^2)] = \sigma^2$
3. Moment generating function of $N(\mu, \sigma^2)$ is $M(t) = \exp[t\mu + \frac{1}{2}\sigma^2 t^2]$
4. Since the moment generating function is defined on $\mathbb{R}$, if $X \sim N(\mu, \sigma^2)$, then $X \in \mathcal{L}^\sigma$, and we can compute its moment by differentiating its moment generating function

## 7.2 Convergence in distribution

### 7.2.1 Basics

**Definition**: Convergence in distribution
Let $\{F_n\}_{n=1}^{\infty}$ and $F$ be the *pdf.* of random variables $\{X_n\}_{n=1}^{\infty}$ and $X$ respectively. If

$$\forall x \in \mathbb{R} \text{ such that } F \text{ is continuous at } x, \lim_{n\to\infty} F_n(x) = F(X)$$

then $\{X_n\}_{n=1}^{\infty}$ converges to $X$ in distribution, written

$$X_n \xrightarrow{\mathrm{d.}} X$$

It is obvious that if $X_n \xrightarrow{\text{d.}} X_0$ and $X_n \overset{d.}{=} Y_n$ $\forall n \in \mathbb{N}$, then $X_n \xrightarrow{\text{d.}} Y$. So, in the context of convergence in distribution, it is the distributions instead of the random variables that matter. We may also write $D_n \xrightarrow{\text{d.}} D$, where $\{D_n\}_{n=1}^{\infty}$ and $D$ are distributions.
$\lim_{n\to\infty} X_n \overset{d.}{=} X$ is a stronger condition than $X_n \xrightarrow{\text{d.}} X$, since the latter one only considers continuous points, which allows $X_n \xrightarrow{\text{d.}} 0$ (degenerate into a constant) when $X_n \sim U(0, \frac{1}{n})$.

**Proposition**: Density condition of convergence in distribution. (Scheffé's Lemma)
If $\{f_n\}_{n=1}^{\infty}$ and $f$ are densities of $\{X_n\}_{n=1}^{\infty}$ and $X$ with reference to $(R, \mathcal{B}, \mu)$, and $\lim_{n\to\infty} f_n \overset{a.s.}{=} f$, then

$$\forall A \in \mathcal{B}(\bar{\mathbb{R}}),\ \lim_{n\to\infty} P(X_n \in A) = P(X \in A)$$

which is a stronger condition that implies convergence in distribution.
Proof:
For any $A \in \mathcal{B}(\bar{\mathbb{R}})$, $P(X \in A \setminus R) = 0$ for $X$ and $\{X_n\}_{n=1}^{\infty}$. Thus, we only consider the set $B := A \cap R \in \mathcal{B}$. We have

$$\begin{aligned}
\lim_{n\to\infty} |f_n - f| \overset{a.s.}{=} 0 \ \Rightarrow\ 0 \le \lim_{n\to\infty} \left| \int_R f_n - f \mathrm{d}\mu \right| &\le \lim_{n\to\infty} \int_R |f_n - f| \mathrm{d}\mu = \int_R \lim_{n\to\infty} |f_n - f| \mathrm{d}\mu = 0 \\
\Rightarrow \lim_{n\to\infty} \left( \int_B f_n - f \mathrm{d}\mu + \int_{R\setminus B} |f_n - f| \mathrm{d}\mu \right) &= \lim_{n\to\infty} \int_R f_n - f \mathrm{d}\mu = 0 \\
\Rightarrow \lim_{n\to\infty} \left| \int_B f_n - f \mathrm{d}\mu \right| &= \lim_{n\to\infty} \left| \int_{R\setminus B} f_n - f \mathrm{d}\mu \right| \text{ (since both sides are bounded)}
\end{aligned}$$

Then,

$$\begin{aligned}
\forall n \in \mathbb{N}^*,\ |P(X_n \in A) - P(X \in A)| &= |P(X_n \in B) - P(X \in B)| \\
&= \left| \int_B f_n - f \mathrm{d}\mu \right| \\
&= \frac{1}{2} \left( \left| \int_B f_n - f \mathrm{d}\mu \right| + \left| \int_{R\setminus B} f_n - f \mathrm{d}\mu \right| \right) \\
&\le \frac{1}{2} \left| \int_R f_n - f \mathrm{d}\mu \right|
\end{aligned}$$

which gives $\lim_{n\to\infty} |P(X_n \in A) - P(X \in A)| = 0$ $\forall A \in \mathcal{B}(\bar{\mathbb{R}})$. Take $A := [-\infty, x]$ where $F(x)$ is continuous, it implies $X_n \xrightarrow{\text{d.}} X$. Note that the converse does not hold.

**Continuous mapping theorem** (for convergence in distribution)
If $f : \bar{\mathbb{R}} \mapsto \mathbb{R}$ is continuous, then $X_n \xrightarrow{\text{d.}} X$ implies $\{f(X_n)\}_{n=1}^{\infty} \xrightarrow{\text{d.}} f(X)$. In fact, it also holds for

function $f$ (not necessarily measurable) that

$$P \circ X^{-1}(\{x \in \bar{\mathbb{R}} : f \text{ is not continuous at } x\}) = 0$$

It is necessary to first show that the set of discontinuous points is in $\mathcal{B}(\bar{\mathbb{R}})$ as follows

$$\begin{aligned}
&\{x \in \bar{\mathbb{R}} : f \text{ is not continuous at } x\} \setminus \{-\infty, +\infty\} \\
=&\{x \in \mathbb{R} : \exists \epsilon > 0, \forall \delta > 0, \exists x_1, x_2 \text{ such that } |x_1 - x| < \delta, |x_2 - x| < \delta, \text{ and } |f(x_1) - f(x_2)| > \epsilon\} \\
=& \bigcup_{\epsilon \in \mathbb{Q}^+} \bigcap_{\delta \in \mathbb{Q}^+} \{x \in \mathbb{R} : \exists x_1, x_2 \in B(x, \epsilon), \text{ and } |f(x_1) - f(x_2)| > \epsilon\} \\
=& \bigcup_{\epsilon \in \mathbb{Q}^+} \bigcap_{\delta \in \mathbb{Q}^+} \{\text{an open set}\} \quad \text{(proved by definition)}
\end{aligned}$$

The proof of the theorem need the interaction between convergence almost surely, which is discussed in the following section.

### 7.2.2 Connection with other convergence

**Theorem**: Almost surely convergence implies convergence in probability
If $X_n \xrightarrow{\text{a.s.}} X$, then $X_n \xrightarrow{\text{p.}} X$.
Proof:
Let $S$ be the set such that $\lim_{n\to\infty} X_n(\omega) = X(\omega)\ \forall \omega \in S$ and $P(S^c) = 0$. Then, for any $\epsilon > 0$, $\liminf_{n\to\infty} \{|X_n - X| < \epsilon\} \supset S$ by the proposition 1 of almost surely convergence. Further, $\forall x \in \mathbb{R}$,

$$S \cap \{X \le x - \epsilon\} \subset S \cap \liminf_{n\to\infty} \{X_n \le x\} \subset S \cap \limsup_{n\to\infty} \{X_n \le x\} \subset S \cap \{X \le x\}$$

(The first subseting is obvious, while the last subsetting can be proved by contradiction.) Then, by using Fatou's lemma twice, we have

$$F(x - \epsilon) \le P\left(S \cap \liminf_{n\to\infty} \{X_n \le x\}\right) = P\left(\liminf_{n\to\infty} \{X_n \le x\}\right) \le \liminf_{n\to\infty} P(X_n \le x)$$

$$\limsup_{n\to\infty} P(X_n \le x) \le P\left(\limsup_{n\to\infty} \{X_n \le x\}\right) = P\left(S \cap \limsup_{n\to\infty} \{X_n \le x\}\right) \le F(x)$$

When $F$ is continuous at $x$, all the inequality becomes equality by taking $\epsilon \to 0$. Then

$$F(x) = \liminf_{n\to\infty} F_n(x) = \limsup_{n\to\infty} F_n(x) = \lim_{n\to\infty} F_n(x)$$

Although $X_n \xrightarrow{\text{d.}} X$ does not imply $X_n \xrightarrow{\text{a.s.}} X$, the following theorem still states a reverse connection.
**Baby Skorohod Theorem**

Given random variables $\{X_n\}_{n=1}^{\infty}$ and $X$ on the probability space $(\Omega, \mathcal{F}, P)$, if $X_n \xrightarrow{\text{d.}} X$, we can construct random variables $\{X'_n\}_{n=1}^{\infty}$ and $X'$ on $([0,1], \mathcal{B}([0,1]), P')$ such that $X' \overset{d.}{=} X$, $X'_n \overset{d.}{=} X\ \forall n \in \mathbb{N}^*$ and $X'_n \xrightarrow{\text{a.s.}} X'$ in the following way.

Let $\{F_n\}_{n=1}^{\infty}$ and $F$ to be the *pdf.* of $X$. Define $\{Q_n\}_{n=1}^{\infty}$, $Q$ to be quantile function by

$$\forall p \in [0,1],\ Q(p) := \inf\{x \in \bar{\mathbb{R}} : F(x) \geq p\} \text{ (same for } Q_n)$$

Thus, $Q(p) \leq q \iff F(q) \geq p$ by the properties of *pdf.* Further let $Y \sim U(0,1)$ on $([0,1], \mathcal{B}([0,1]), P')$ whose *pdf.* is $F_Y(x) = x\ \forall x \in [0,1]$. Let

$$X' := Q(Y), X'_n := Q_n(Y)\ \forall n \in \mathbb{N}^*$$

Then, $X \overset{d.}{=} X'$ because

$$\forall y \in \bar{\mathbb{R}},\ F_{X'}(y) = P'[Q(Y) \leq y] = P'(Y \leq F(y)) = F_Y[F(y)] = F(y)$$

(Likewise for $X_n \overset{d.}{=} X'_n$) In fact, this is a practical way to generate another random variable with a uniform distributed variable and a *pdf.*

It remains to prove $X'_n \xrightarrow{\text{a.s.}} X$. We will show a lemma that when $p \in S := (0,1) \cap \{p : Q \text{ is continuous at } p\}$,

$$\lim_{n\to\infty} Q_n[Y(p)] = \lim_{n\to\infty} Q_n(p) = Q(p) = Q(Y[p])$$

which will give the result because the discontinuous points of $Q$ are at most countable (same as $F$), thus $P'([0,1] \setminus A) = 0$.

**Proof** of Lemma:

When $p \in (0,1)$ and $Q$ is continuous at $p$,

$$\lim_{n\to\infty} Q_n(p) = Q(p)$$

The following will use the following properties of $Q$, they are directly from the definition.

$$\forall p \in [0,1], q \in \bar{\mathbb{R}},\ Q(p) \leq q \iff F(q) \geq p, \text{ and } Q(p) > q \iff F(q) < p$$

First, since $F$ is continuous except on at most countable numbers of points, for any $\epsilon > 0$, there exists $x$ on which $F$ is continuous such that

$$Q(p) - \epsilon < x < Q(p) \text{ when } |Q(p)| < \infty, \text{ or } \epsilon < x \text{ when } Q(p) = +\infty$$

Then, the properties of $Q$ and $\lim_{n\to\infty} F_n(x) = F(x)$ gives that for any large enough $n$,

$$Q(p) - \epsilon < x < Q_n(p) \text{ when } |Q(p)| < \infty, \text{ or } \epsilon < x < Q_n(p) \text{ when } Q(p) = +\infty$$

Taking $\epsilon \to 0$ (finite case) or $\epsilon \to +\infty$ (infinite case), we conclude

$$\liminf_{n\to\infty} Q_n(p) \geq Q(p) \text{ when } |Q(p)| < \infty, \text{ or } \liminf_{n\to\infty} Q_n(p) = +\infty \text{ when } Q(p) = +\infty$$

The other direction is similar, but in finite case, we need an arbitrary $p' > p$. For any $\epsilon > 0$, there exists $x$ on which $F$ is continuous such that

$$Q(p') < x < Q(p') + \epsilon \;\Rightarrow\; Q(x) \geq p' > p$$

The convergence gives $Q_n(x) \geq p$ for any large enough $n$, thus

$$\limsup_{n\to\infty} Q_n(p) \leq Q(p')$$

Since $Q$ is continuous at point $t$, the result follows by taking $p' \to p^+$.

**Proof** of continuous mapping theorem
Since $X_n \xrightarrow{\text{d.}} X$, construct the above $X'_n \overset{d.}{=} X_n$ and $X' \overset{d.}{=} X$ such that $X'_n \xrightarrow{\text{a.s.}} X'$. The continuous mapping theorem of almost surely convergence gives $\{f(X'_n)\}_{n=1}^{\infty} \xrightarrow{\text{a.s.}} f(X')$. Even if $f$ is discontinuous on a set $A$ such that $P'(X' \in A) = P(X \in A) = 0$, almost surely convergence still holds by discarding $A$.
It further gives $\{f(X'_n)\}_{n=1}^{\infty} \xrightarrow{\text{d.}} f(X')$ and thus, $\{f(X_n)\}_{n=1}^{\infty} \xrightarrow{\text{a.s.}} f(X)$.

**Corollary**: Convergence in distribution implies convergence of mean
Continued from above, the dominated convergence theorem gives $\lim_{n\to\infty} \mathbb{E}[f(X'_n)] = \mathbb{E}[f(X')]$ whenever $f$ is bounded (then it is bounded by a $\mathcal{L}^1$ function). Then,

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$$

Or if $|X|$ is bounded by some $\mathcal{L}^1$ function,

$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}(X)$$

### 7.2.3 Extension of definition into random vector

To facilitate the extension, we show some equivalent definitions of convergence in distribution.

**Proposition**: Equivalent definitions of convergence in distribution (Portmanteau theorem)
$X_n \xrightarrow{\text{d.}} X$ if and only if

1. For any bounded and continuous $f : \bar{\mathbb{R}} \mapsto \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}\left[f(X_n)\right] = \mathbb{E}\left[f(X)\right]$$

2. For any $A \in \mathcal{B}(\bar{\mathbb{R}})$ such that $P[X \in \partial(A)]$,

$$\lim_{n\to\infty} P(X_n \in A) = P(X \in A)$$

where

$$\partial(A) := \left\{x \in \bar{\mathbb{R}} : \exists\{a_n \in A\}_{n=1}^{\infty},\ \lim_{n\to\infty} a_n = x \text{ and } \exists\{b_n \in A^c\}_{n=1}^{\infty},\ \lim_{n\to\infty} b_n = x\right\}$$

is the boundary of $A$. (We can prove that $\partial(A)^c$ is open. Thus, $\partial(A)$ is close and in $\mathcal{B}(\bar{\mathbb{R}})$.)

$X_n \xrightarrow{\text{d.}} X$ implies 1 by the continuous mapping theorem. We then prove the other direction. Given $x \in \mathbb{R}$ to be an arbitrary point on which $F$ is continuous. We construct these two bounded and continuous functions. For any $k \in \mathbb{N}^*$, define

$$g_k(t) := \begin{cases} 1 & , t \le x \\ 1 - k(t-x) & , x < t \le x + \frac{1}{k} \\ 0 & , t > x + \frac{1}{k} \end{cases}$$

$$\text{and } h_k(t) := \begin{cases} 1 & , t \le x - \frac{1}{k} \\ -k(t-x) & , x - \frac{1}{k} < t \le x \\ 0 & , t > x \end{cases}$$

Then, for any $n \in \mathbb{N}^*$,

$$\mathbb{E}[h_k(X_n)] = \int_{\bar{\mathbb{R}}} h_k \mathrm{d}P \circ X_n^{-1} \le \int_{\bar{\mathbb{R}}} \mathbf{1}_{(-\infty,x]} \mathrm{d}P \circ X_n^{-1} = F_n(x) \le \int_{\bar{\mathbb{R}}} g_k \mathrm{d}P \circ X_n^{-1} = \mathbb{E}[g_k(X_n)]$$

Since $\lim_{n\to\infty} \mathbb{E}[h_k(X_n)] = \mathbb{E}[h_k(X)]$, $\lim_{n\to\infty} \mathbb{E}[g_k(X_n)] = \mathbb{E}[g_k(X)]$ and by the monotone convergence theorem,

$$\lim_{k\to\infty} \mathbb{E}[g_k(X)] = \mathbb{E}[\mathbf{1}_{(-\infty,x]}] = F(x) \text{ and } \lim_{k\to\infty} \mathbb{E}[h_k(X)] = \mathbb{E}[\mathbf{1}_{(-\infty,x)}] = F(x)$$

We can conclude $\lim_{n\to\infty} F_n(x) = f(x)$.
Condition 2 implies $X_n \xrightarrow{\text{d.}} X$ because $\partial[-\infty, x] = \{x\}$ has its measure 0. For the other direction, we will show that $\forall A \in \bar{\mathbb{R}}$ such that $P[X \in \partial(A)] = 0$,

$$\lim_{n\to\infty} \mathbb{E}[\mathbf{1}_A(X_n)] = \mathbb{E}[\mathbf{1}_A(X)]$$

By the (almost surely) continuous mapping theorem, it suffers to show that the set $B$ on which $\mathbf{1}_A$ is discontinuous has measure $P \circ X^{-1}(B) = 0$. We will claim that $B = \partial(A)$, thus the result follows. To show $\partial(A) \subset B$, we have for any $x \in \partial(A)$, there exist $\{a_n \in A\}_{n=1}^{\infty}$ and $\{b_n \in A^c\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = x$. Since

$$\lim_{n\to\infty} \mathbf{1}_A(a_n) = 1, \ \lim_{n\to\infty} \mathbf{1}_A(b_n) = 0$$

$\mathbf{1}_A$ is discontinuous at $x$.
Next, we show $B \subset \partial(A)$. For any $x \in B$, there exist $\{x_n \in \bar{\mathbb{R}}\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} x_n = x$, but $\lim_{n\to\infty} \mathbf{1}_A(x_n) \neq \mathbf{1}_A(x)$ (or limit does not exist). Assume $\mathbf{1}_A(x) = 1$ (the other case is similar). Then there exists a substring $\{x_{n_i}\}_{i=1}^{\infty}$ such that

$$\lim_{i\to\infty} x_{n_i} = x, \text{ but } \forall i \in \mathbb{N}^*, \ \mathbf{1}_A(x_{n_i}) = 0 \Rightarrow x_{n_i} \in A^c$$

There also exists a sequence $a_n := x \in A \ \forall n \in \mathbb{N}^*$ such that $\lim_{n\to\infty} a_n = x$. These two sequence gives the condition of $x \in \partial(A)$.

**Definition**: Convergence in distribution of random vector
Let $\{\mathbf{X}_n\}_{n=1}^{\infty}$ and $\mathbf{X}$ be some random vector of $\bar{\mathbb{R}}^k$, where $k \in \mathbb{N}^*$, we define

$$\mathbf{X}_n \xrightarrow{\text{d.}} \mathbf{X} \ \text{ (i.e. convergence in distribution to } \mathbf{X})$$

if for any bounded and continuous function, $f : \bar{\mathbb{R}}^k \mapsto \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}\left[f(\mathbf{X}_n)\right] = \mathbb{E}\left[f(\mathbf{X})\right]$$

**Proposition**: Convergence in probability implies convergence in distribution
If $X_n \xrightarrow{\text{p.}} X$, then $X_n \xrightarrow{\text{d.}} X$.
Proof:
For any bounded and continuous function $f : \mathbb{R} \mapsto \mathbb{R}$, $\{f(X_n)\}_{n=1}^{\infty} \xrightarrow{\text{p.}} f(X)$ and thus $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$. It satisfies the equivalent definition of $X_n \xrightarrow{\text{d.}} X$.

**Proposition**: Convergence in distribution to a constant implies convergence in probability
If $X_n \xrightarrow{\text{d.}} c$, where $c \in \mathbb{R}$ is a constant, then $X_n \xrightarrow{\text{p.}} c$. Proof:
For any $\epsilon > 0$,

$$0 \leq \lim_{n\to\infty} P(X_n < c - \epsilon) \leq \lim_{n\to\infty} F_n(c - \epsilon) = 0 \text{ and } \lim_{n\to\infty} P(X_n \leq c + \epsilon) = \lim_{n\to\infty} F_n(c + \epsilon) = 1$$

Thus $\lim_{n\to\infty} P(|X_n - c| \leq \epsilon) = 1 - 0 = 1$.

**Slutsky's theorem**

Given $X_n \xrightarrow{\text{d.}} X$ with their probability space $(\Omega, \mathcal{F}, P)$, if $\{Y_n\}_{n=1}^{\infty}$ is another sequence of random variables on $(\Omega, \mathcal{F}, P)$ such that $Y_n \xrightarrow{\text{p.}} 0$ (or equivalently, $Y_n \xrightarrow{\text{d.}} 0$), then

$$\{X_n + Y_n\}_{n=1}^{\infty} \xrightarrow{\text{d.}} X + c$$

As a corollary,

1. If $\{X_n - Y_n\}_{n=1}^{\infty} \xrightarrow{\text{p.}} 0$ then $Y_n \xrightarrow{\text{d.}} X$.
2. If $Y_n \xrightarrow{\text{d.}} c \in \mathbb{R}$, (or $Y_n \xrightarrow{\text{p.}} c$), then $\{X_n + Y_n\}_{n=1}^{\infty} \xrightarrow{\text{d.}} X + c$

Proof:

For any bounded and continuous $f : \bar{\mathbb{R}} \mapsto \mathbb{R}$, $\lim_{n\to\infty} \mathbb{E}\left[f(X_n)\right] = \mathbb{E}\left[f(X)\right]$. It suffices to show that $\lim_{n\to\infty} \mathbb{E}\left[f(X_n + Y_n)\right] = \mathbb{E}\left[f(X)\right]$. We can further restrict such $f|_{\mathbb{R}}$ to be uniform continuous on $\mathbb{R}$ because the functions we constructed in the proof (i.e. $g_k$, $h_k$) are uniform continuous, which gives

$$\lim_{\epsilon\to 0^+} \sup\{|f(a) - f(b)| : a, b \in \mathbb{R} \text{ and } |a - b| \leq \epsilon\} = 0$$

Let $C_\epsilon := \sup\{|f(a) - f(b)| : a, b \in \mathbb{R} \text{ and } |a - b| \leq \epsilon\}$ and $M := \sup_{x\in\bar{\mathbb{R}}}\{|f(x)|\}$ be the bound. We have for any $\epsilon > 0$, $n \in \mathbb{N}^*$

$$\begin{aligned}
&\Big|\mathbb{E}\left[f(X_n + Y_n)\right] - \mathbb{E}\left[f(X)\right]\Big| \\
\leq&\Big|\mathbb{E}\left[f(X_n + Y_n)\right] - \mathbb{E}\left[f(X_n)\right]\Big| + \Big|\mathbb{E}\left[f(X_n)\right] - \mathbb{E}\left[f(X)\right]\Big| \\
\leq&\mathbb{E}\left(\big|f(X_n + Y_n) - f(Y_n)\big|\right) + \Big|\mathbb{E}\left[f(X_n)\right] - \mathbb{E}\left[f(X)\right]\Big| \\
=&\mathbb{E}\left(\big|f(X_n + Y_n) - f(Y_n)\big|\mathbf{1}_{|Y_n \leq \epsilon|}\right) + \mathbb{E}\left(\big|f(X_n + Y_n) - f(Y_n)\big|\mathbf{1}_{|Y_n > \epsilon|}\right) + \Big|\mathbb{E}\left[f(X_n)\right] - \mathbb{E}\left[f(X)\right]\Big| \\
\leq&\mathbb{E}\left(C_\epsilon\right) + P(|Y_n| > \epsilon)\mathbb{E}\left(2M\right) + \Big|\mathbb{E}\left[f(X_n)\right] - \mathbb{E}\left[f(X)\right]\Big| \\
=&C_\epsilon + 2MP(|Y_n| > \epsilon) + \Big|\mathbb{E}\left[f(X_n)\right] - \mathbb{E}\left[f(X)\right]\Big|
\end{aligned}$$

Since $\lim_{n\to\infty} P(|Y_n| > \epsilon) = 0$, $\lim_{n\to\infty} \big|\mathbb{E}\left[f(X_n)\right] - \mathbb{E}\left[f(X)\right]\big| = 0$ and $\lim_{\epsilon\to 0^+} C_\epsilon = 0$, we have

$$\lim_{n\to\infty} \Big|\mathbb{E}\left[f(X_n + Y_n)\right] - \mathbb{E}\left[f(X)\right]\Big| = 0$$

by taking $\epsilon \to 0^+$, which gives the result.

## 7.3 Characteristic function of distribution

### 7.3.1 Remarks on complex function

The character function is a $\mathbb{R} \to \mathbb{C}$ function. We begin with some remarks on the calculus of complex function.

**Definition**: Complex function
A function $f : S \mapsto \mathbb{C}$ is a complex function if it can be decomposed into $f = g + \mathrm{i}h$, where $g, h : S \to \mathbb{R}$.
We write $\mathrm{Re}f := g$ and $\mathrm{Im}f := h$ to represent its real part and imaginary part respectively.
Define $\|f\| = \sqrt{g^2 + h^2}$ to be its norm.

Since $\mathbb{C}$ is a metric space with its distance function $d(z_1, z_2) = \|z_1 - z_2\|$ $\forall z_1, z_2 \in \mathcal{C}$. The limit, continuity, open set and Borel $\sigma$-algebra $\mathcal{B}(\mathbb{C})$ are defined accordingly. It is easy to verify that a sequence of complex numbers is convergent if and only if both their real part and imaginary part are convergent.

**Proposition**: Ratio criteria for convergent complex series
For any sequence $\{z_n \in \mathbb{C}\}_{n=1}^{\infty}$, $\lim_{n\to\infty} \sum_{i=1}^{n} z_i$ exists if

$$\lim_{n\to\infty} \sum_{i=1}^{n} \|z_i\| \text{ exists and be finite}$$

which can be given by

$$R := \lim_{n\to\infty} \frac{\|z_{n+1}\|}{\|z_n\|} < 1$$

We first prove the second half, if $R < 1$, taking an arbitrary $r \in (R, 1)$, there exists $N \in \mathbb{N}^*$ such that $\|z_{n+1}\| < r\|z_n\|$ $\forall n \geq N$. Then $\{\|z_n\|\}_{n=N}^{\infty}$ is upper-bounded by a convergent geometric series. Since $\sum_{i=1}^{n} \|z_n\|$ is increasing as $n$ increases, we have $\sum_{i=1}^{\infty} \|z_i\|$ exists and be finite.
Then, $\sum_{i=1}^{\infty} |\mathrm{Re}z_i|$ and $\sum_{i=1}^{\infty} |\mathrm{Im}z_i|$ both exist, since they are bounded by $\sum_{i=1}^{\infty} \|z_i\|$. It remains to prove that both $\sum_{i=1}^{\infty} \mathrm{Re}z_i$ and $\sum_{i=1}^{\infty} \mathrm{Im}z_i$ exist, which will gives the existence of $\sum_{i=1}^{\infty} z_i = \sum_{i=1}^{\infty} \mathrm{Re}z_i + \mathrm{i} \sum_{i=1}^{\infty} \mathrm{Im}z_i$.
We only show the part for $\sum_{i=1}^{\infty} \mathrm{Re}z_i$ since the other is the same. We first have

$$\forall n \in \mathbb{N}^*, |\mathrm{Re}z_n| + \mathrm{Re}z_n \geq 0 \text{ and } \sum_{i=1}^{n} |\mathrm{Re}z_i| + \mathrm{Re}z_i \leq 2 \sum_{i=1}^{n} |\mathrm{Re}z_i|$$

Thus, $\sum_{i=1}^{\infty} |\mathrm{Re}z_i| + \mathrm{Re}z_i$ exists and is finite since it is monotonic increasing and bounded. Then $\sum_{i=1}^{\infty} \mathrm{Re}z_i$ exists and is finite because it is the difference of two convergent series.

We define its integration and differentiation as follows.

**Definition**: Lebesgue integral of complex function

Given a measure space $(S, \mathcal{B}, \mu)$ and a measurable complex value function $f : (S, \mathcal{B}) \mapsto (\mathbb{C}, \mathcal{B}(\mathbb{C}))$. If $\mathrm{Re}f, \mathrm{Im}f \in \mathcal{L}^1$, then $f$ is Lebesgue measurable (write $f \in \mathcal{L}^1$) and define

$$\int_S f \mathrm{d}\mu = \int_S \mathrm{Re}f \mathrm{d}\mu + \mathrm{i} \int_S \mathrm{Im}f \mathrm{d}\mu$$

It suffices that $\|f\| \in \mathcal{L}^1$ gives $f \in \mathcal{L}^1$, because $\mathrm{Re}f, \mathrm{Im}f \leq \|f\|$. But the converse does not hold, though we can still compute $\int_S \|f\| \mathrm{d}\mu$ since $\|f\| \in \mathcal{L}^+$.

For complex value random variable $X : (\Omega, \mathcal{F}) \mapsto (\mathbb{C}, \mathcal{B}(\mathbb{C}))$, we can also write $\mathbb{E}(X)$ as the expectation form to represent the integral.

**Proposition**: about Lebesgue integral of complex function

1. Linearity
2. Rule of integral of transformation
3. $\| \int_S f \mathrm{d}\mu \| \leq \int_S \|f\| \mathrm{d}\mu$
4. Dominated convergence theorem, where the dominant is given by $\|f_n\| \overset{a.s.}{\leq} g$ and $g \in \mathcal{L}^1 : \mathbb{R} \to \mathbb{R}$

Proof of 1 and 2 is simply by considering $\mathrm{Re}f, \mathrm{Im}f$ separately. 3 holds when $\int_S \|f\| \mathrm{d}\mu = +\infty$. For finite cases, we use the Cauchy-Schwarz inequality, $ab + cd \leq \sqrt{a^2 + c^2} \times \sqrt{b^2 + d^2}$ $\forall a, b, c, d \in \mathbb{R}$. We have $\forall x \in S$ such that $\|f(x)\| < \infty$,

$$\mathrm{Re}f(x) \left( \int_S \mathrm{Re}f \mathrm{d}\mu \right) + \mathrm{Im}f \left( \int_S \mathrm{Im}f \mathrm{d}\mu \right) \leq \|f(x)\| \times \sqrt{\left( \int_S \mathrm{Re}f \mathrm{d}\mu \right)^2 + \left( \int_S \mathrm{Im}f \mathrm{d}\mu \right)^2}$$

Integrating both side gives (since $\mu(\{x : \|f(x)\| = \infty\}) = 0$)

$$\left( \int_S \mathrm{Re}f \mathrm{d}\mu \right)^2 + \left( \int_S \mathrm{Im}f \mathrm{d}\mu \right)^2 \leq \sqrt{\left( \int_S \mathrm{Re}f \mathrm{d}\mu \right)^2 + \left( \int_S \mathrm{Im}f \mathrm{d}\mu \right)^2} \times \int_S \|f\| \mathrm{d}\mu$$

i.e.

$$\left\| \int_S f \mathrm{d}\mu \right\| \leq \sqrt{\left\| \int_S f \mathrm{d}\mu \right\|} \times \int_S \|f\| \mathrm{d}\mu$$

When $\| \int_S f \mathrm{d}\mu \| = 0$, the result $\int_S \|f\| \mathrm{d}\mu \geq 0$ always hold. When $\| \int_S f \mathrm{d}\mu \| > 0$, divide both sides by $\sqrt{\| \int_S f \mathrm{d}\mu \|}$ and the result follows.

To prove 4, note that $\|f_n\| \overset{a.s.}{\leq} g$ gives $\mathrm{Re}f, \mathrm{Im}f \overset{a.s.}{\leq} g \ \forall n \in \mathbb{N}^*$. The dominated convergence of these two real functions gives the result.

**Definition**: Differentiation of complex function
A function $f : \mathbb{C} \mapsto \mathbb{C}$ is differentiable at $z \in \mathbb{C}$ if the limit

$$\lim_{h \to 0, h \in \mathbb{C}} \frac{f(z+h) - f(h)}{h} = 0$$

exists. Define the derivative $f'(z)$ to be the above limit on which it exists.

The following propositions and theorems of real function can be extended to complex function. We just list them without proof.

1. Linearity of derivative
2. Rule of transforming integral and the chain rules
3. Integrate by parts and rule of differentiation of product

**Example**: Exponential function, sine and cosine It begins with the fact that $z' = 1$ and $c' = 0$ ($c \in \mathbb{C}$ is a constant). By the chain rule, $(z^n)' = nz^{n-1}$ for any $n \in \mathbb{N}^*$. Then we define

1. Exponential function $\exp : \mathbb{C} \mapsto \mathbb{C}$
$$\exp(z) := \sum_{n=0}^{\infty} \frac{z^n}{n!} \text{ (the summation always convergent, given by the ratio text)}$$
2. Sine function $\sin : \mathbb{C} \mapsto \mathbb{C}$
$$\sin(z) := \frac{\exp(\mathrm{i}z) - \exp(-\mathrm{i}z)}{2\mathrm{i}}$$
3. Cosine function $\cos : \mathbb{C} \mapsto \mathbb{C}$
$$\cos(z) := \frac{\exp(\mathrm{i}z) + \exp(-\mathrm{i}z)}{2}$$

with the following properties

1. $\exp(0) = 1$ and $\exp(z_1 + z_2) = \exp(z_1)\exp(z_2) \ \forall z_1, z_2 \in \mathbb{C}$
2. $\forall z \in \mathbb{C}$, $\sin(z)^2 + \cos(z)^2 = 1$

3. The Euler equality: $\exp(ix) = \cos(x) + i\sin(x)\ \forall x \in \mathbb{R}$

4. Pi is defined by $\pi := \inf\{x \in \mathbb{R}^+, \sin(x) = 0\}$. Any other proposition about pi holds with this definition

5. Euler's number is defined by $e := \exp(1)$. Any other proposition about this number holds with this definition

**Example**: Differentiating exp

We will show that exp is continuous and differentiable $(\exp z)' = \exp z$. We proceed by the following steps

1. Whenever $\|z\| < 1$, $\|\exp z - 1\| \leq 2\|z\|$. Because

$$\|\exp z - 1\| \leq \sum_{n=1}^{\infty} \frac{\|z^n\|}{n!} = \|z\| + \sum_{n=2}^{\infty} \frac{\|z\|^n}{n!}$$

and

$$\begin{aligned}\sum_{n=2}^{\infty} \frac{\|z\|^n}{n!} &\leq \sum_{n=2}^{\infty} \|z\| \times \left(\frac{\|z\|}{2}\right)^{n-1} \quad \text{(bounded by a convergent geometric series)} \\ &= \frac{\frac{\|z\|^2}{2}}{1 - \frac{\|z\|}{2}} \leq \|z\| \quad \text{(since } \|z\| < 1\text{, by simple calculus)}\end{aligned}$$

2. $\lim_{z\to 0} \exp z = 1$, since $0 \leq \lim_{z\to 0} \|\exp z - 1\| \leq \lim_{z\to 0} 2\|z\| = 0$

3. $\exp z$ is continuous, since $\lim_{h\to 0} \exp(z+h) = \lim_{h\to 0} (\exp z)(\exp h) = \exp z$

4. exp is differentiable at $z = 0$. Because we have

$$\lim_{h\to 0} \frac{\exp(h) - \exp(0)}{h} = \lim_{h\to 0} \sum_{n=0}^{\infty} \frac{h^n}{(n+1)!}$$

whose limit exists and equals to 1 since when $\|z\| \leq 1$,

$$\left\| \sum_{n=0}^{\infty} \frac{z^n}{(n+1)!} - 1 \right\| \leq \frac{3}{2}\|z\| \text{ (similar to 1)}$$

5. $(\exp z)' = \exp z$, since

$$\forall z \in \mathbb{C},\ \lim_{h\to 0} \frac{\exp(h+z) - \exp(z)}{h} = \exp z \times \lim_{h\to 0} \frac{\exp(h) - \exp(0)}{h} = \exp z$$

6. By the chain rule, $(\sin z)' = \cos z$, and $(\cos z)' = -\sin z$

The following propositions of exp will be useful in next section.

**Proposition**: another approach to exp

$\forall z \in \mathbb{C},\ \lim_{n\to\infty}\left(1+\frac{z}{n}\right)^n = \exp z$

Proof:

$$\forall n \in \mathbb{N},\ \left(1+\frac{z}{n}\right)^n = \sum_{k=0}^{n} \frac{z^k}{n^k}\binom{n}{k} = \sum_{k=0}^{n} \frac{z^k}{k!}\prod_{j=0}^{k-1}\frac{n-j}{n} \leq \sum_{k=0}^{n}\frac{z^k}{k!}\times 1$$

$$\Rightarrow\ \forall M \in \{0,1,\cdots,n\},\ \left\|\left(1+\frac{z}{n}\right)^n - \exp z\right\|$$

$$\leq \left\|\sum_{k=0}^{n}\frac{z^k}{k!}\left(1-\prod_{j=0}^{k-1}\frac{n-j}{n}\right)\right\| + \left\|\sum_{k=n+1}^{\infty}\frac{z^k}{k!}\right\|$$

$$\leq \sum_{k=0}^{M}\frac{\|z\|^k}{k!}\left(1-\prod_{j=0}^{k-1}\frac{n-j}{n}\right) + \sum_{k=M+1}^{n}\frac{\|z\|^k}{k!} + \sum_{k=n+1}^{\infty}\frac{\|z\|^k}{k!}$$

$$\leq \sum_{k=0}^{M}\frac{\|z\|^k}{k!}\left(1-\prod_{j=0}^{k-1}\frac{n-j}{n}\right) + 2\sum_{k=M+1}^{\infty}\frac{\|z\|^k}{k!}$$

(For convenience, assume $\prod_{j=0}^{-1}\frac{n-j}{n}=1$.) We also have that for any $k \in \mathbb{N}$, $\lim_{n\to\infty}\prod_{j=0}^{k-1}\frac{n-j}{n}=1$, and the convergence of $\sum_{n=0}^{\infty}\frac{\|z\|^n}{n!}$ gives $\lim_{M\to\infty}\sum_{n=M}^{\infty}\frac{\|z\|^n}{n!}=0$. For any $\epsilon>0$, we can pick $M$ such that $2\sum_{k=M+1}^{\infty}\frac{\|z\|^k}{k!}\leq\frac{\epsilon}{2}$. Then we can pick $N \geq M$ such that for any $n \geq N$,

$$\frac{\|z\|^k}{k!}\left(1-\prod_{j=0}^{k-1}\frac{n-j}{n}\right) \leq \frac{\epsilon}{2(M+1)}\quad \forall k \in \{0,1,\cdots,M\}$$

Then it gives $\lim_{n\to\infty}\|\left(1+\frac{z}{n}\right)^n - \exp z\| = 0$.

**Proposition**: Expansion of $\exp(ix)$

For any $x \in \mathbb{R}$,

$$\left\|\exp(\mathrm{i}x) - \sum_{n=0}^{N}\frac{(\mathrm{i}x)^n}{n!}\right\| \leq \min\left\{\frac{|x|^{N+1}}{(N+1)!}, \frac{2|x|^N}{N!}\right\}\ \forall N \in \mathbb{N}^*$$

We call $\sum_{n=0}^{N}\frac{(\mathrm{i}x)^n}{n!}$ the expansion of $\exp(\mathrm{i}x)$ at $N$. Then RHS is the error term, often denoted as $\xi_N$.

The proof begins with integration by part. For any $x \in \mathbb{R}$,

$$\forall n \in \mathbb{N},\ \int_0^x (x-t)^n \exp(\mathrm{i}t)\mathrm{d}t = \left[-\frac{(x-t)^{n+1}}{n+1}\exp(\mathrm{i}t)\right]\Big|_0^x + \int_0^x \frac{\mathrm{i}(x-t)^{n+1}}{n+1}\exp(\mathrm{i}t)\mathrm{d}t$$

$$= \frac{x^{n+1}}{n+1} + \frac{\mathrm{i}}{n+1}\int_0^x (x-t)^{n+1}\exp(\mathrm{i}t)\mathrm{d}t$$

When $n = 0$, we have

$$x + \mathrm{i}\int_0^x (x-t)\exp(\mathrm{i}t)\mathrm{d}t = \int_0^x \exp(\mathrm{i}t)\mathrm{d}t = \frac{\exp(\mathrm{i}x) - 1}{\mathrm{i}}$$

Thus, $\exp(\mathrm{i}x) = 1 + \mathrm{i}x + \mathrm{i}^2 \int_0^x (x-t)\exp(\mathrm{i}t)\mathrm{d}t$. Further expand the last term,

$$\begin{aligned}
\exp(\mathrm{i}x) &= 1 + \mathrm{i}x + \frac{\mathrm{i}^2 x^2}{2!} + \frac{\mathrm{i}^3}{2!}\int_0^x (x-t)^2 \exp(\mathrm{i}t)\mathrm{d}t \\
&= 1 + \mathrm{i}x + \frac{\mathrm{i}^2 x^2}{2!} + \frac{\mathrm{i}^3 x^3}{3!} + \frac{\mathrm{i}^4}{3!}\int_0^x (x-t)^3 \exp(\mathrm{i}t)\mathrm{d}t \\
&= \left[\sum_{n=0}^{N} \frac{(\mathrm{i}x)^n}{n!}\right] + \frac{\mathrm{i}^{N+1}}{N!}\int_0^x (x-t)^N \exp(\mathrm{i}t)\mathrm{d}t \ \forall N \in \mathbb{N}^*
\end{aligned}$$

Therefore,

$$\begin{aligned}
\left\|\exp(\mathrm{i}x) - \sum_{n=0}^{N}\frac{(\mathrm{i}x)^n}{n!}\right\| = \left\|\frac{\mathrm{i}^{N+1}}{N!}\int_0^x (x-t)^N \exp(\mathrm{i}t)\mathrm{d}t\right\| &\leq \frac{1}{N!}\left|\int_0^x |x-t|^N \|\exp(\mathrm{i}t)\|\mathrm{d}t\right| \\
&\leq \frac{1}{N!} \times \left|\int_0^x (x-t)^N \mathrm{d}t\right| = \frac{|x|^{N+1}}{(N+1)!}
\end{aligned}$$

To prove another error term, rewrite our first formula by replacing $n$ by $N-1$,

$$\int_0^x (x-t)^{N-1}\exp(\mathrm{i}t)\mathrm{d}t = \frac{x^N}{N} + \frac{\mathrm{i}}{N}\int_0^x (x-t)^N \exp(\mathrm{i}t)\mathrm{d}t$$

Then, by substitution,

$$\begin{aligned}
\exp(\mathrm{i}x) - \sum_{n=0}^{N}\frac{(\mathrm{i}x)^n}{n!} &= \frac{\mathrm{i}^{N+1}}{N!}\int_0^x (x-t)^N \exp(\mathrm{i}t)\mathrm{d}t \\
&= \frac{\mathrm{i}^N}{(N-1)!}\int_0^x (x-t)^{N-1}\exp(\mathrm{i}t)\mathrm{d}t - \frac{(\mathrm{i}x)^N}{N!}
\end{aligned}$$

which gives

$$\left\|\exp(\mathrm{i}x) - \sum_{n=0}^{N}\frac{(\mathrm{i}x)^n}{n!}\right\| \leq \frac{1}{(N-1)!}\left|\int_0^x (x-t)^{N-1}\mathrm{d}t\right| + \left|\frac{(\mathrm{i}x)^N}{N!}\right| = 2\frac{|x|^N}{N!}$$

Thus, we get the second upper bound of LHS.

The reason why we need two bounds is that when we integrate the error term, the first bound often gives a inner limit, while the second bound often gives a dominant $\mathcal{L}^1$ term, thus dominated convergence theorem can apply.

#### 7.3.2 The characteristic function

**Definition**: Characteristic function of distribution

Given a finite random variable $X \sim D$, define $\varphi : \mathbb{R} \mapsto \mathbb{C}$

$$\varphi_X(t) := \mathbb{E}[\exp(\mathrm{i}tX)] = \int_{\mathbb{R}} \exp(\mathrm{i}tX)\mathrm{d}D = \int_{\mathbb{R}} \cos(tX)\mathrm{d}D + \mathrm{i}\int_{\mathbb{R}} \sin(tX)\mathrm{d}D$$

Since $\|\exp(\mathrm{i}tX)\| = \|\cos(tX) + \mathrm{i}\sin(tX)\| = 1$, $\varphi_X$ always exists and bounded.

Another way to express the character function is

$$\varphi_X(t) = \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{(\mathrm{i}tX)^n}{n!}\right]$$

Thus, it is closely related to moment. If $X \in \mathcal{L}^N$ (i.e. $\mathbb{E}(X^n) < \infty \ \forall n \in \{1, 2, \cdots, N\}$),

$$\varphi_X(t) = \sum_{n=0}^{N} \frac{(\mathrm{i}t)^n}{n!}\mathbb{E}(X^n) + \mathbb{E}\left[\sum_{n=N+1}^{\infty} \frac{(\mathrm{i}tX)^n}{n!}\right]$$

We can also apply the dominated convergence theorem. When $X \in \mathcal{L}^\infty$ and $\lim_{n\to\infty} \frac{|tX|^n}{n!} = 0 \ \forall t \in \mathbb{R}$, then

$$\varphi_X(t) = \sum_{n=0}^{\infty} \frac{(\mathrm{i}t)^n}{n!}\mathbb{E}(X^n) \ \forall t \in \mathbb{R}$$

because

$$\sup_{N\in\mathbb{N}^*} \left\| \sum_{n=0}^{N} \frac{(\mathrm{i}tX)^n}{n!} \right\| = \sup_{N\in\mathbb{N}^*} \left\| \exp(\mathrm{i}tX) - \sum_{n=0}^{N} \frac{(\mathrm{i}x)^n}{n!} \right\| \leq \sup_{N\in\mathbb{N}^*} \frac{2|tX|^N}{N!} \in \mathcal{L}^1$$

Note that if the moment generating function is defined on $\mathbb{R}$ and finite, then

$$\sum_{n=1}^{\infty} \frac{|tX|^n}{n!} = \mathbb{E}[\exp(|tX|)] < \mathbb{E}[\exp(tX)] + \mathbb{E}[\exp(-tX)] < \infty \ \forall t \in \mathbb{R}$$

which is a sufficient condition that gives both $\lim_{n\to\infty} \frac{|tX|^n}{n!} = 0$ and $X \in \mathcal{L}^\infty$.

We will later show that the characteristic function, as well as the moment generating function (if defined on $\mathbb{R}$ and finite) uniquely determines a distribution, thus we may also write $\varphi_D$ and $M_D$ for both function.

**Proposition**: properties of characteristic function

1. $\|\varphi_X(t)\| \leq 1$

2. $\varphi_X(0) = 1$

3. If the distribution $D$ is symmetric, then $\varphi(t) \in \mathbb{R}\ \forall t \in \mathbb{R}$

4. $\varphi_X$ is uniform continuous

5. If $X \in \mathcal{L}^N$, $\varphi_X$ is $N$-times differentiable and $\varphi_X^{(N)}(t) = \mathbb{E}[(\mathrm{i}X)^N \exp(\mathrm{i}tX)]$

6. If $\varphi_X$ is $N$-times differentiable at 0, then $\mathbb{E}(\mathrm{i}^N X^N) = \varphi_X^{(N)}(0)$ exists.

7. Followed from 6, if $N$ is even, then $\mathbb{E}(X^n)$ exists for $n \in \{1, 2, \cdots, N\}$, which are $\mathbb{E}(X^N) = \frac{\varphi_X^{(N)}(0)}{\mathrm{i}^N}$.

8. If $Y := aX + b$, then $\varphi_Y(t) = \exp(itb)\varphi_X(at)$

9. If $X$ and $Y$ are independent, then $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$

Proofs of some properties:

3 is given by

$$\begin{aligned}
\forall h, t \in \mathbb{R},\ \|\varphi_X(t+h) - \varphi_Y(t)\| &\leq \mathbb{E}\|\exp[\mathrm{i}(t+h)X] - \exp(\mathrm{i}tX)\| \\
&= \mathbb{E}[\|\exp(\mathrm{i}tX)\| \times \|\exp(\mathrm{i}hX) - 1\|] \\
&\leq \mathbb{E}\|\exp(\mathrm{i}hX) - 1\|
\end{aligned}$$

Since $\lim_{h\to 0}[\exp(\mathrm{i}hX) - 1] = 0$, we have $\lim_{h\to 0}\sup_{t\in\mathbb{R}}\|\varphi_X(t+h) - \varphi_Y(t)\| = 0$ and the result follows.

To show 5, we first show $\varphi_X'(t) = \mathbb{E}[\mathrm{i}X\exp(\mathrm{i}tX)]$. For any $t \in \mathbb{R}$,

$$\begin{aligned}
&\forall h \in \mathbb{R}, \left\|\frac{\varphi_X(t+h) - \varphi_X(t)}{h} - \mathbb{E}[\mathrm{i}X\exp(\mathrm{i}tX)]\right\| \\
&\leq \mathbb{E}\left[\|\exp(\mathrm{i}tX)\| \times \frac{1}{|h|} \times \|\exp(\mathrm{i}hX) - 1 - \mathrm{i}hX\|\right] \\
&\leq \mathbb{E}\left[1 \times \frac{1}{|h|} \times \min\left\{\frac{|hX|^2}{2!}, \frac{2|hX|}{1!}\right\}\right] \quad \text{(by expanding at 2)} \\
&= \mathbb{E}\left(\min\left\{\frac{1}{2}|h||X|^2, 2|X|\right\}\right)
\end{aligned}$$

Let $\xi_h := \min\left\{\frac{1}{2}|h||X|^2, 2|X|\right\}$. When $|X| \in \mathcal{L}^1$, $\xi_h \leq 2|X| \in \mathcal{L}^1$ and $\lim_{h\to 0}\xi_h = 0$, then dominated convergence theorem applies,

$$0 \leq \lim_{h\to 0}\left\|\frac{\varphi_X(t+h) - \varphi_X(t)}{h} - \mathbb{E}[\mathrm{i}X\exp(\mathrm{i}tX)]\right\| \leq \lim_{h\to 0}\mathbb{E}(\xi_h) = \mathbb{E}\left(\lim_{h\to 0}\xi_h\right) = 0$$

which gives

$$\varphi_X'(t) = \frac{\varphi_X(t+h) - \varphi_X(t)}{h} = \mathbb{E}[\mathrm{i}X\exp(\mathrm{i}tX)]\ \forall t \in \mathbb{R}$$

By induction, we can differentiate more times and the result follows.
To prove 6, we first differentiate $\varphi_X(0)$ once,

$$\varphi_X'(0) = \lim_{h\to 0}\mathbb{E}\left[\frac{\exp(\mathrm{i}hX)-1}{h}\right] = \lim_{h\to 0}\mathbb{E}\left[\frac{1}{h}\sum_{n=1}^{\infty}\frac{(\mathrm{i}Xh)^n}{n!}\right] = \lim_{h\to 0}\mathbb{E}\left[\mathrm{i}X\sum_{n=0}^{\infty}\frac{(\mathrm{i}Xh)^n}{(n+1)!}\right]$$

Since $\varphi_X(0)$ exists, the limit exists, which is $\mathbb{E}(\mathrm{i}X)$. Then, if $\varphi_X^{(2)}(0)$ exists, $\varphi_X'(t)$ exists and is finite for $t$ in a small enough open ball $B(0,r)$, which is

$$\varphi_X'(t) = \lim_{h\to 0}\frac{\varphi_X(t+h)-\varphi_X(t)}{h} = \lim_{h\to 0}\mathbb{E}\left[\exp(\mathrm{i}tX)\times\frac{\exp(\mathrm{i}hX)-1}{h}\right] = \mathbb{E}[\mathrm{i}X\exp(\mathrm{i}tX)]$$

We can compute $\varphi_X^{(2)}(0)$ using $\varphi_X'(t)$. By induction, we obtain $\varphi_X^{(N)}(0) = \mathbb{E}(\mathrm{i}^N X^N)$. The existence of LHS gives the existence of RHS.
7 follows immediately since it further gives the existence of $\mathbb{E}(X^N)$ when $N$ is even. However, when $N$ is odd, $\mathbb{E}(X^N)$ may not exist even if $\mathbb{E}(\mathrm{i}^N X^N)$ exists.

**Example**: Characteristic function of normal distribution
When $X \sim N(0,1)$, we first have its moment generating function $M_N(t) = \exp\left(-\frac{t^2}{2}\right)$. Since $X \in \mathcal{L}^\infty$, we can expand both sides

$$\forall t \in \mathbb{R},\ \sum_{n=0}^{\infty}\frac{t^n}{n!}\mathbb{E}(X^n) = \sum_{n=0}^{\infty}\frac{(-1)^n t^{2n}}{2^n n!}$$

Comparing the coefficient, we have $\mathbb{E}(X^n) = 0$ when $n$ is odd, and

$$E(X^n) = \frac{n!}{(\frac{n}{2})!2^{\frac{n}{2}}}$$

when $n$ is even. Then we obtain the characteristic function of $N(0,1)$

$$\varphi_N(t) := \varphi_X(t) = \sum_{n=0}^{\infty}\frac{(\mathrm{i}t)^n}{n!}\mathbb{E}(X^n) = \sum_{n=0}^{\infty}\frac{(\mathrm{i}t)^{2n}}{(2n)!}\times\frac{(2n)!}{(n)!2^n} = \sum_{n=0}^{\infty}\frac{(-1)^n t^{2n}}{2^n n!} = \exp\left(-\frac{t^2}{2}\right)$$

**Theorem**: uniqueness of characteristic function The characteristic function uniquely determines the probability distribution. It suffices to find a way to express the *pdf*. We will show that

$$F_X(x) = \lim_{\sigma\to\infty,\sigma\in\mathbb{N}^*}\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{x}\int_{-\infty}^{\infty}\exp(-\mathrm{i}\theta\sigma y)\varphi(\sigma y)f_N(y)\mathrm{d}y\mathrm{d}\theta$$

for all $x$ on which $F_X$ is continuous, where $f_N(y) = \frac{1}{\sqrt{2\pi}}\exp\frac{-y^2}{2}$ is the density of the standard normal distribution. For the discontinuous points, which are at most countable, they are naturally given by $F_X(x) = \lim_{h \to x^+} F_X(h) = 0$.

Proof:

Suppose $X \sim D$. We begin with the following lemma. For any $\theta \in \mathbb{R}$ and other distribution $D'$ on $\mathbb{R}$, we have

$$\begin{aligned}\int_{\mathbb{R}} \exp(-\mathrm{i}\theta y)\varphi(y)\mathrm{d}D'(y) &= \int_{\mathbb{R}} \exp(-\mathrm{i}\theta y) \int_{\mathbb{R}} \exp(\mathrm{i}yx)\mathrm{d}D\mathrm{d}D' \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} \exp[-\mathrm{i}y(x-\theta)]\mathrm{d}D'\mathrm{d}D \text{ (by Fubini theorem)} \\ &= \int_{\mathbb{R}} \varphi_{D'}(x-\theta)\mathrm{d}D \ \text{ where } \varphi_{D'} \text{ is its characteristic function}\end{aligned}$$

Then, let $D'$ be the normal distribution $N(0,\sigma^2)$ with density $f_\sigma(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-x^2}{2\sigma^2} = \sigma^{-1} f_N(\sigma^{-1}x)$ and characteristic function $\varphi_\sigma(t) := \varphi_N(\sigma t) = \exp \frac{-\sigma^2 t^2}{2}$. We have $\forall \theta \in \mathbb{R}$,

$$\int_{-\infty}^{\infty} \exp(-\mathrm{i}\theta\sigma y)\varphi(\sigma y)f_N(y)\mathrm{d}y = \int_{-\infty}^{\infty} \exp(-\mathrm{i}\theta y)\varphi(y)f_\sigma(y)\mathrm{d}y = \int_{\mathbb{R}} \exp\left[-\frac{\sigma^2(z-\theta)^2}{2}\right]\mathrm{d}D(z)$$

Since RHS $\in \mathcal{L}^+(\mu)$, integrate $\theta$ both sides from $-\infty$ to $x \in \mathbb{R}$ gives

$$\begin{aligned}&\int_{-\infty}^{x}\int_{-\infty}^{\infty} \exp(-\mathrm{i}\theta\sigma y)\varphi(\sigma y)f_N(y)\mathrm{d}y\mathrm{d}\theta \\ =&\int_{-\infty}^{x}\int_{\mathbb{R}} \exp\left[-\frac{\sigma^2(z-\theta)^2}{2}\right]\mathrm{d}D(z)\mathrm{d}\theta \\ =&\int_{\mathbb{R}}\int_{-\infty}^{x} \exp\left[-\frac{\sigma^2(z-\theta)^2}{2}\right]\mathrm{d}\theta\mathrm{d}D(z) \text{ (Fubini's theorem)} \\ =&\int_{\mathbb{R}}\int_{-\infty}^{x-z} \exp\left[-\frac{\sigma^2 s^2}{2}\right]\mathrm{d}s\mathrm{d}D(z) \text{ (change of variable } s := \theta - z) \\ =&\int_{\mathbb{R}}\int_{-\infty}^{x-z} \frac{\sqrt{2\pi}}{\sigma} f_{N(0,\sigma^{-2})}(s)\mathrm{d}s\mathrm{d}D(z) \\ =&\frac{\sqrt{2\pi}}{\sigma}\int_{\mathbb{R}} P\left[N\left(0,\sigma^{-1}\right) \le x-z\right]\mathrm{d}D(z) \\ =&\frac{\sqrt{2\pi}}{\sigma}P(\sigma^{-1}N + X \le x)\end{aligned}$$

(Assume $N$ with standard normal distribution is also a random variable on the same probability space as $X$.) Since $\lim_{\sigma\to\infty}\sigma^{-1}N = 0$, we have $\{\sigma^{-1}N + X\}_{\sigma=1}^{\infty} \xrightarrow{\text{d.}} X$ by Slutsky's theorem. Therefore, for any $x$ on which $F_X$ is continuous,

$$F(x) = \lim_{\sigma\to\infty,\sigma\in\mathbb{N}^*} P(\sigma^{-1}N + X \le x) = \lim_{\sigma\to\infty,\sigma\in\mathbb{N}^*} \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{x}\int_{-\infty}^{\infty}\exp(-\mathrm{i}\theta\sigma y)\varphi(\sigma y)f_N(y)\mathrm{d}y\mathrm{d}\theta$$

**Corollary**: Fourier inversion

Suppose $\varphi$ is the characteristic function for distribution $D$ on $\mathbb{R}$. If $\|\varphi\| \in \mathcal{L}^1(\lambda)$, then $D$ has a probability density $f$ with regard to $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ (Lebesgue measure space) given by

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \exp(-\mathrm{i}tx)\varphi(t)\mathrm{d}t$$

From the above proof, we can observe that random variable $Y_\sigma := \sigma^{-1}N + X$ has a density for any $\sigma \in \mathbb{N}^*$, which is

$$\begin{aligned} f_{Y_\sigma}(\theta) :&= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-\mathrm{i}\theta\sigma y)\varphi(\sigma y) f_N(y)\mathrm{d}y \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-\mathrm{i}\theta t)\varphi(t) f_N(\sigma^{-1}t)\mathrm{d}t \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\mathrm{i}\theta t - \frac{t^2}{2\sigma^2}\right)\varphi(t)\mathrm{d}t \end{aligned}$$

Since for any $t \in \mathbb{R}$

$$\left\| \exp\left(-\mathrm{i}\theta t - \frac{t^2}{2\sigma^2}\right)\varphi(t) \right\| \leq \|\exp\left(-\mathrm{i}\theta t\right)\| \times \|\exp 0\| \times \|\varphi(t)\| \leq \|\varphi(t)\| \in \mathcal{L}^1$$

and

$$\lim_{\sigma\to\infty} \exp\left(-\mathrm{i}\theta t - \frac{t^2}{2\sigma^2}\right)\varphi(t) = \exp(-\mathrm{i}\theta t)\varphi(t)$$

Thus, by the dominated convergence theorem,

$$\forall x \in \mathbb{R}, \lim_{\sigma\in\mathbb{N}^*, \sigma\to\infty} f_{Y_\sigma}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\mathrm{i}xt\right)\varphi(t)\mathrm{d}t$$

Suppose the limit to be $f(x)$. It remains to prove that it is a valid probability density of $X$. First, $f \geq 0$ since $f_{Y_\sigma}$ are densities and thus nonnegative for $\sigma \in \mathbb{N}^*$.
Then, we show $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$ from the fact that $\{Y_\sigma\}_{\sigma=1}^{\infty} \xrightarrow{\mathrm{d.}} X$. Let

$$A := \{x \in \mathbb{R} : F_X \text{ is continuous at } x\} \quad \text{where } F_X \text{ is the } \textit{pdf.} \text{ of X}$$

Since $\mathbb{R} \setminus A$ is countable, we partition $\mathbb{R}$ into $\mathbb{R} = \bigcup_{i=1}^{\infty}[l_i, r_i)$ such that for any $i \in \mathbb{N}^*$, $l_i < r_i$, $l_i, r_i \in A$, and $[l_i, r_i)$ are mutually disjoint. By dominated convergence theorem, for any $i \in \mathbb{N}^*$,

$$\begin{aligned} & f_{Y_\sigma}\mathbf{1}_{[l_i,r_i)} \leq \frac{\int_{-\infty}^{\infty} ||\varphi(t)||\mathrm{d}t}{2\pi}\mathbf{1}_{[l_i,r_i)} \in \mathcal{L}^1(\lambda)\ \forall \sigma \in \mathbb{N}^* \\ \Rightarrow & \int_{l_i}^{r_i} f(x)\mathrm{d}x = \lim_{\sigma\in\mathbb{N}^*, \sigma\to\infty} \int_{l_i}^{r_i} f_{Y_\sigma}(x)\mathrm{d}x = \lim_{\sigma\in\mathbb{N}^*, \sigma\to\infty} P(Y_\sigma \in [l_i, r_i)) \end{aligned}$$

Using the equivalent condition for convergence in distribution,

$$\forall i \in \mathbb{N}^*, \int_{l_i}^{r_i} f(x)\mathrm{d}x = \lim_{\sigma\in\mathbb{N}^*, \sigma\to\infty} P(Y_n \in [l_i, r_i)) = P(X \in [l_i, r_i)) \in [0, 1]$$

because $P(X \in \partial[l_i, r_i)) = 0 \ \forall i \in \mathbb{N}^*$. Then

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \sum_{i=1}^{\infty} \int_{l_i}^{r_i} f(x)\mathrm{d}x = \sum_{i=1}^{\infty} P(X \in [l_i, r_i)) = P(X \in \mathbb{R})$$

Therefore, $f_{Y_\sigma}$ converges to a valid density $f$, which gives $\{Y_\sigma\}_{\sigma=1}^{\infty}$ converges in distribution to, say $X'$, with density $f$. We have $X' \overset{d.}{=} X$ and $f$ is the probability density of $f$.

**Corollary**: Uniqueness of moment generating function

Let $\mathcal{D}$ be the set of distributions whose moment generating function is defined on $\mathbb{R}$ and is finite. For any two distributions $D_1, D_2 \in \mathcal{D}$ whose moment generating function $M_{D_1}$, $M_{D_2}$ is the same, we have $D_1 \overset{d.}{=} D_2$.

Proof

Under the restriction of moment generation function, the characteristic function of $D_1$ and $D_2$ can be represented as

$$\forall i \in \{1, 2\},\ \varphi_{D_i}(t) = \sum_{n=0}^{\infty} \frac{(\mathrm{i}t)^n}{n!}\mathbb{E}(D_i^n) \ \forall t \in \mathbb{R}$$

while their moment generating functions are

$$M_{D_1}(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\mathbb{E}(D_1^n) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\mathbb{E}(D_2^n) = M_{D_1}(t) \ \forall t \in \mathbb{R}$$

By comparing the coefficient of $t^n$, $D_1$ and $D_2$ have the same moments up to order infinity. Therefore, they have the same characteristic function and thus equal in distribution.

### 7.3.3 New criteria for convergence in distribution

This section proves the following theorem that introduces another approach to prove convergence in distribution via characteristic function.

**Theorem**: Continuity criteria of convergence in distribution

Let $\{X_n\}_{n=1}^{\infty}$ be some finite random variables with characteristic functions $\{\varphi_n := \varphi_{X_n}\}_{n=1}^{\infty}$. First, if $\{X_n\}_{n=1}^{\infty}$ converges to a finite random variable, $X_n \xrightarrow{\text{d.}} X$, then $\varphi_X(t) = \lim_{n\to\infty} \varphi_n(t) \ \forall t \in \mathbb{R}$. Conversely, if

1. $\lim_{n\to\infty} \varphi_n(t)$ exists for all $t \in \mathbb{R}$

2. and $\varphi := \lim_{n\to\infty} \varphi_n$ is continuous at $t = 0$

then, $\{X_n\}_{n=1}^{\infty}$ converges to a finite random variable whose characteristic function is $\varphi$.
The first is easy to verify by continuous mapping theorem (then $\{\mathrm{Re}\exp(\mathrm{i}tX_n)\}_{n=1}^{\infty} \xrightarrow{\text{d.}} \mathrm{Re}\exp(\mathrm{i}tX)$ and $\{\mathrm{Im}\exp(\mathrm{i}tX_n)\}_{n=1}^{\infty} \xrightarrow{\text{d.}} \mathrm{Im}\exp(\mathrm{i}tX)$) and dominated convergence theorem (then $\lim_{n\to\infty} \mathrm{Re}\varphi_n(t) = \mathrm{Re}\varphi_X(t)$ and $\lim_{n\to\infty} \mathrm{Im}\varphi_n(t) = \mathrm{Im}\varphi_X(t)$).
The proof of the second part requires the concepts of tightness and relative compactness on a set of distribution.

**Definition**: Tightness
Let $\Pi$ be a set of distributions. $\Pi$ is defined to be tight if it satisfies one of the following three equivalent conditions

1. For any $\epsilon > 0$, there exists a compact set $K \subset \mathbb{R}$ such that $P(D \in K) > 1 - \epsilon \ \forall D \in \Pi$
2. For any $\epsilon > 0$, there exists a finite interval $I \subset \mathbb{R}$ such that $P(D \in I^c) > 1 - \epsilon \ \forall D \in \Pi$
3. For any $\epsilon > 0$, there exists $M \in \mathbb{R}$ such that $F_D(M) - F_D(-M) > 1 - \epsilon \ \forall D \in \Pi$

**Proposition**: about tight distribution set
If $\Pi$ is tight, then for any sequence of distribution $\{D_n \in \Pi\}_{n=1}^{\infty}$, there exists a subsequence that converges in distribution to a distribution (or we say it is convergent in distribution).

$$\{D_{n_i}\}_{i=1}^{\infty} \xrightarrow{\text{d.}} D$$

We construct the subsequence using the fact that any sequence has a convergent subsequence. We also use the countability and denseness of $\mathbb{Q}$. Apply an index set to $\mathbb{Q}$, i.e. $\mathbb{Q} = \{q_n\}_{n=1}^{\infty}$.
Let $\{F_n\}_{n=1}^{\infty}$ be the *pdf.* of an arbitrary sequence of distribution $\{D_n \in \Pi\}_{n=1}^{\infty}$. When $n = 1$, take the subsequence $\{D\}_{N_1}$, where $N_1 \subset \mathbb{N}^*$ is the index set, such that $\{F_n(q_1)\}_{n\in N_1}$ converges.
Using induction, suppose we have the subsequence $\{D\}_{N_k}$ such that

$$\lim_{n\to\infty, n\in N_k} F_n(q_i) \text{ exists for any } i \in \{1, 2, \cdots, k\}$$

We take a further subsequence with index set $N_{k+1} \subset N_k$ such that $\{F_n(q_{k+1})\}_{n\in N_{k+1}}$ converges. Then, $\{F_n(q_i)\}_{n\in N_i}$ converges for any $i \in \{1, 2, \cdots, k+1\}$. It results in a subsequence $\{D_{n_i}\}_{i=1}^{\infty}$ such that

$$F(q) := \lim_{i\to\infty} F_{n_i}(q) \text{ exists } \forall q \in \mathbb{Q}$$

We first have $F(q_1) \leq F(q_2)$ when $q_1 \leq q_2, q_1, q_2 \in \mathbb{Q}$. Extend $F$ to $\mathbb{R}$ by

$$F(x) = \lim_{q \to x^+, q \in \mathbb{Q}} F(q)$$

The limit exists because the sequence $\{F(q)\}$, where $q \in \mathbb{Q}$ and $q \to x^+$, is monotonic decreasing and bounded from below. We will claim that $F$ is a *pdf.*, say, of distribution $D$, and $\{D_{n_i}\}_{i=1}^{\infty} \xrightarrow{\text{d.}} D$
We first have $F$ to be monotonic increasing and right continuous.
For any $x$ on which $F$ is continuous, we can take two sequence $\{L_k \in \mathbb{Q}\}_{i=k}^{\infty}$ and $\{R_k \in \mathbb{Q}\}_{k=1}^{\infty}$ such that $\{L_k\}_{k=1}^{\infty}$ is monotonic increasing, $\{R_k \in \mathbb{Q}\}_{k=1}^{\infty}$ is monotonic decreasing and $\lim_{k \to \infty} L_k = \lim_{k \to \infty} R_k = x$. Then, for any $k \in \mathbb{N}^*$,

$$F_{n_i}(L_k) \leq F_{n_i}(x) \leq F_{n_i}(R_k) \ \forall i \in \mathbb{N}^*$$

Taking $i \to \infty$,

$$F(L_k) \leq \liminf_{i \to \infty} F_{n_i}(x) \leq \limsup_{i \to \infty} F_{n_i}(x) \leq F(R_k)$$

Taking $k \to \infty$, since $F$ is continuous at $x$, all the inequality become equality and $\lim_{i \to \infty} F_{n_i}(x) = F(x)$.
It remains to show that $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$, which needs the tightness of $\Pi$.
For any $\epsilon > 0$, we can take a large enough $M$ such that

$$\sup_{n \in \mathbb{N}^*} \{1 - F_n(M) + F_n(-M)\} \leq \sup_{n \in \mathbb{N}^*} P(D_n \in [-M, M]^c) < \epsilon$$

We further enlarge $M$ (if necessary) so that $F$ is continuous on $M$ and $-M$. Taking $n \to \infty$,

$$\epsilon > \limsup_{n \to \infty} \{1 - F_n(M) + F_n(-M)\} = 1 - F(M) + F(-M)$$

which gives $F(M) - F(-M) > 1 - \epsilon$. Since $F(M) \leq 1$ and $F(-M) \geq 0$, we have $F(M) > 1 - \epsilon$ and $F(-M) < \epsilon$. Therefore, $\lim_{M \to -\infty} F(x) = 0$ and $\lim_{M \to +\infty} F(x) = 1$.

**Corollary**: Tight sequence of distribution
A sequence of distribution $\{D_n\}_{n=1}^{\infty}$ is defined to be tight if for any $\epsilon > 0$,

$$\limsup_{n \to \infty} P(D_n \in [-M, M]^c) \leq \epsilon$$

The proposition of tightness also applies. There exists a subsequence of $\{D_n\}_{n=1}^{\infty}$ that is convergent in distribution.
This is because there always exists a large enough $N$ such that $\sup_{n \geq N} P(D_n \in [-M, M]^c) \leq 2\epsilon$. We can use this fact in proof of $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$ above and get the result.

Here a lemma before we can prove the continuity criteria

**Lemma**:

Let $D$ be a distribution with characteristic function $\varphi$, then for any $x \in \mathbb{R}^+$,

$$P(D \in [-x,x]^c) \leq \alpha x \int_0^{x^{-1}} [1 - \mathrm{Re}\varphi(t)]\mathrm{d}t,$$

where

$$\alpha := \inf\left\{1 - \frac{\sin a}{a} : a \in \mathbb{R} \text{ and } |a| \geq 1\right\}^{-1} \text{ is a constant in } \mathbb{R}^+$$

Proof:

Since $\mathrm{Re}\varphi(t) = \int_{\mathbb{R}} \cos(ty)\mathrm{d}D(y)$, for any $x \in \mathbb{R}^+$,

$$\begin{aligned}
x\int_0^{x^{-1}} [1 - \mathrm{Re}\varphi(t)]\mathrm{d}t &= x\int_0^{x^{-1}} \int_{\mathbb{R}} [1 - \cos(ty)]\mathrm{d}D(y)\mathrm{d}t \\
&= \int_{\mathbb{R}} x\int_0^{x^{-1}} [1 - \cos(ty)]\mathrm{d}t\mathrm{d}D(y) \text{ (Fubini)} \\
&= \int_{\mathbb{R}} 1 - \frac{\sin(x^{-1}y)}{x^{-1}y}\mathrm{d}D(y) \\
&\geq \int_{\{y\in\mathbb{R}: x^{-1}|y|>1\}} 1 - \frac{\sin(x^{-1}y)}{x^{-1}y}\mathrm{d}D(y) \\
&\geq \int_{\{y\in\mathbb{R}:|y|>x\}} \inf\left\{1 - \frac{\sin a}{a} : a \in \mathbb{R} \text{ and } |a| \geq 1\right\}\mathrm{d}D(y) \\
&= \alpha^{-1}P(D \in [-x,x]^c)
\end{aligned}$$

**Proof** of the second part of the continuity criteria of convergence in distribution

We first claim that those criteria give $\{D_n\}_{n=1}^{\infty}$ (the distributions of $\{X_n\}_{n=1}^{\infty}$) to be a tight sequence. For any $M \in \mathbb{N}^*$,

$$\limsup_{n\to\infty} P(D_n \in [-M,M]^c) \leq \limsup_{n\to\infty} \alpha M \int_0^{M^{-1}} [1 - \mathrm{Re}\varphi_n(t)]\mathrm{d}t$$

For any $t \in \mathbb{N}^*$, $\lim_{n\to\infty}\varphi_n(t) = \varphi(t)$ (condition 1) implies $\lim_{n\to\infty} 1 - \mathrm{Re}\varphi_n(t) = 1 - \mathrm{Re}\varphi(t)$. By dominated convergence theroem,

$$\limsup_{n\to\infty} P(D_n \in [-M,M]^c) \leq \limsup_{n\to\infty} \alpha M \int_0^{M^{-1}} [1 - \mathrm{Re}\varphi_n(t)]\mathrm{d}t = \int_0^{M^{-1}} [1 - \mathrm{Re}\varphi(t)]\mathrm{d}t$$

Since $\varphi(0) = \lim_{n\to\infty}\varphi_n(0) = 1$ and $\varphi$ is continuous at 0 (condition 2), we have $\lim_{t\to 0}[1-\mathrm{Re}\varphi(t)] = 0$. Then, for any $\epsilon > 0$, there exists sufficiently large $M \in \mathbb{R}$ such that

$$1 - \mathrm{Re}\varphi(t) < \frac{\epsilon}{\alpha}\ \forall |t| \leq M^{-1} \ \Rightarrow\ \alpha M \int_0^{M^{-1}} [1 - \mathrm{Re}\varphi(t)]\mathrm{d}t \leq \alpha M \int_0^{M^{-1}} \frac{\epsilon}{\alpha}\mathrm{d}t \leq \epsilon$$

By the proposition of tight distribution set, there exists some subsequences $\{D_{n_i}\}_{i=1}^{\infty}$ that is convergent in distribution. We further claim that all of these subsequences converge in distribution to the same distribution, which gives the convergence of $\{D_n\}_{n=1}^{\infty}$ in distribution.
Suppose two subsequences of $\{D_n\}_{n=1}^{\infty}$ converge to $D_1$ and $D_2$ in distribution respectively. Their characteristic functions must be the same, $\varphi$ (by the first part of this theorem). Since characteristic function uniquely determines a distribution, $D_1 \overset{d}{=} D_2$.

**Example**: Binomial distribution converges to Poison distribution
Let $\lambda \in \mathbb{R}^+$ be a constant. The sequence of Binomial variable $\left\{X_n \sim B(n, \frac{\lambda}{n})\right\}_{n=1}^{\infty}$ converges in distribution to a Poisson distribution

$$X_n \xrightarrow{\text{d.}} Po(\lambda)$$

The probability mass functions (i.e. density with regard to counting measure $C$) and characteristic of both distributions are given as

$$\forall x \in \{0, 1, \cdots, n\},\ f_{B(n,p)}(x) = p^x(1-p)^{n-x}\binom{n}{x}$$

$$\varphi_{B(n,p)}(t) = \int_{\{0,1,\cdots,n\}} \exp(\mathrm{i}tx) f_{B(n,p)}(x)\mathrm{d}C = \sum_{x=0}^{n} \left[p\exp(\mathrm{i}t)\right]^x (1-p)^{n-x}\binom{n}{x} = \left[1 - p + p\exp(\mathrm{i}t)\right]^n$$

$$\forall x \in \mathbb{N},\ f_{Po(\lambda)}(x) = \frac{\lambda^x}{x!}\exp(-\lambda)$$

$$\varphi_{Po(\lambda)}(t) = \int_{\mathbb{N}} \exp(\mathrm{i}tx) f_{Po(\lambda)}(x)\mathrm{d}C = \exp(-\lambda)\sum_{x=0}^{\infty} \frac{\left[\lambda\exp(\mathrm{i}t)\right]^x}{x!} = \exp\left[\lambda\exp(\mathrm{i}t) - \lambda\right]$$

By the continuity criteria, since $\varphi_{Po(\lambda)}$ is continuous at 0, it suffers to prove the point-wise convergence of their characteristic functions.

$$\varphi_n(t) := \varphi_{B\left(n,\frac{\lambda}{n}\right)}(t) = \left[1 - \frac{\lambda}{n} + \frac{\lambda}{n}\exp(\mathrm{i}t)\right]^n = \left[1 + \frac{\lambda\exp(\mathrm{i}t) - \lambda]}{n}\right]^n$$

Then,

$$\forall t \in \mathbb{R},\ \lim_{n\to\infty} \varphi_n(t) = \lim_{n\to\infty}\left[1 + \frac{\lambda\exp(\mathrm{i}t) - \lambda]}{n}\right]^n = \exp\left[\lambda\exp(\mathrm{i}t) - \lambda\right]] = \varphi_{Po(\lambda)}(t)$$

### 7.3.4 The central limit theorem

This is one of the fundamental theorem in Statistics.
**Central limit theorem**

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of *i.i.d.* finite random variables. If $\mu := \mathbb{E}(X_1)$ and $\sigma^2 := \mathrm{var}(X_1)$ exists and are both finite, then

$$\left\{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right\}_{n=1}^{\infty} \xrightarrow{\text{d.}} N(0,1) \quad \text{where } \bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \ \forall n \in \mathbb{N}^*$$

Let $Y_n := \frac{X-\mu}{\sigma}$ so that $\mathbb{E}(Y_n) = 0$ and $\mathrm{var}(Y_n) = 1$ for any $n \in \mathbb{N}^*$. Also, define $\bar{Y}_n$ accordingly. To prove $\{\sqrt{n}\bar{Y}_n\}_{n=1}^{\infty} \xrightarrow{\text{d.}} N(0,1)$, we define the characteristic function for any $n \in \mathbb{N}^*$,

$$\varphi_n(t) := \varphi_{\bar{Y}_n}(\sqrt{n}t) = \left[\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n = \mathbb{E}\left[\exp\left(\frac{\mathrm{i}tY_1}{\sqrt{n}}\right)\right] \quad \forall t \in \mathbb{R}$$

It suffices to prove that $\lim_{n\to\infty} \varphi_n = \varphi_N = \exp\frac{-t^2}{2}$. For all $n \in \mathbb{N}^*$ and $t \in \mathbb{R}$, we expand $\varphi_n(t)$ at 2,

$$\varphi_n(t) = \left[1 + \frac{\mathrm{i}t}{\sqrt{n}}\mathbb{E}(Y_1) + \frac{\mathrm{i}^2t^2}{2n}\mathbb{E}(Y_1^2) + \xi_n(t)\right]^n = \left[1 - \frac{t^2}{2n} + \xi_n(t)\right]^n$$

where

$$\xi_n(t) = \mathbb{E}\left(\min\left\{\frac{|tY_1|^3}{6\sqrt{n}^3}, \frac{|tY_1|^2}{n}\right\}\right)$$

Using an inequality (which can be proved by induction)

$$\forall n \in \mathbb{N}^* \text{ and } a, b \in \mathbb{C} \text{ such that } \|a\|, \|b\| \le 1,\ \|a^n - b^n\| \le n\|a-b\|$$

we can show that the error term does not affect the limit,

$$0 \le \lim_{n\to\infty}\left\|\varphi_n(t) - \left(1 - \frac{t^2}{2n}\right)^n\right\| \le \lim_{n\to\infty} n\|\xi_n(t)\| = \lim_{n\to\infty}\mathbb{E}\left(\min\left\{\frac{|tY_1|^3}{6\sqrt{n}}, |tY_1|^2\right\}\right) = 0$$

The last equality is given by the dominated convergence theorem. Finally,

$$\lim_{n\to\infty}\varphi_n(t) = \lim_{n\to\infty}\left(1 - \frac{t^2}{2n}\right)^n = \exp\left(-\frac{t^2}{2}\right)$$

gives the result.

# Chapter 8: Point estimation

## 8.1 General background of point estimation

### 8.1.1 Population

A population is a set $U$ of all elements of the experimental interest, which can be finite, countably infinite or uncountably infinite. Each element possesses one or more attributes for the experimental purpose, which can be represented as a map

$$X_U : U \to S$$

where $S$ is the set of values of the attribute. (Not to confuse with random variable $X$.) For example, all the citizens in a country constitute a population ($U$), and we are concerned with the attributes of gender ($X_{U,1}$), height ($X_{U,2}$), and weight ($X_{U,3}$) of each person.
Some numerical parameters, denoted as,

$$\{\theta_i\}_{i\in I}, \text{ where } I \text{ is a index set}$$

describe or summarise the behaviour of an attribute over the population
For example, when $U$ is finite with size $N$ and $S$ is a numerical set, we can define the population mean

$$\mu := \frac{1}{N}\sum_{x\in U} X_U(x)$$

and population variance

$$\sigma^2 := \frac{1}{N}\sum_{x\in U} [X_U(x) - \mu]^2$$

to be population parameters.
However, the attributions and parameters of the whole population may be difficult or costly to obtain. Some even only exist hypothetically. They are often estimated from a sample.

### 8.1.2 Sample, statistic and point estimation

A sample is a portion of the population obtained by a known sampling method, which is often finite with a known sample size $n$. It can be represented as a sequence $\{\omega_i \in U\}_{i=1}^n$ or vector $[\omega_1, \omega_2, \cdots, \omega_n]^T \in U^n$, whose entries $\omega_i$ are elements of the population.
When the randomness of the sampling method is presented, we can assign a probability measure, with a proper interpretation of probability, to evaluate how likely a specific sample will be chosen. We can model it by a probability space

$$(\Omega, \mathcal{S}, P)$$

Where $\Omega \subset U^n$ is the sample space, i.e. the set of all probable sequences or vectors obtained from the population, $\mathcal{S}$ is a properly chosen $\sigma$-algebra (simply the power set of $\Omega$ if it is finite), and $P$ is the probability measure.
Each $\omega_i$ element in the sample $[\omega_1, \omega_2, \cdots, \omega_n]^T \in \Omega$ maps to an attribute $\{x_i\}_{i=1}^n$ by the function

$X_U$, the same as population. The attribute, if numeric, can be represented as a measurable vector or sequence

$$\mathbf{X} : (\Omega, \mathcal{S}) \mapsto (\bar{\mathbb{R}}^n, \mathcal{B}(\bar{\mathbb{R}}^n))$$

such that

$$\mathbf{X}([\omega_1, \omega_2, \cdots, \omega_n]^T) = [X_U(\omega_1), X_U(\omega_2), \cdots, X_U(\omega_n)]$$

Since a random vector of $n$ entries is equivalent to the collection of $n$ random variable, another representation is $\mathbf{X} = [X_1, X_2, \cdots, X_n]^T$ or $\{X_i\}_{i=1}^n$, where

$$X_i : (\Omega, \mathcal{S}) \mapsto (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}}))\ \forall i \in \{1, 2, \cdots, n\}$$

such that $X_i$ is a function that only takes the entry $\omega_i$ as input

$$X_i([\omega_1, \omega_2, \cdots, \omega_n]^T) = X_i(\omega_i) = X_U(\omega_i)$$

We can therefore define the (joint) *pdf.* for $\mathbf{X}$ and any of $X_i$.

Since the attributes $\mathbf{X}$ or $\{X_i\}_{i=1}^n$ contain information of the experimental interest. We may simply say $\mathbf{X}$ or $\{X_i\}_{i=1}^n$ is the sample itself when the other information of $\{\omega_i\}_{i=1}^n$ is outside the scope of discussion.
In real practice, the attributes of an obtained sample is known and fixed, we call it the realization of sample $\{X_i\}_{i=1}^n$, denoted as $\{x_i\}_{i=1}^n$, which is one possible output of $\{X_i\}_{i=1}^n$.

Though the sample size $n$ is finite, it is necessary to study the asymptotic behaviour when $n$ is sufficiently large and heading towards infinity. For example, we are concerned about whether $X_n$ converges to a random variable almost surely, in probability or in distribution.
In this case, we should also assume the sampling method is repeatable so that the strategy for obtaining a sample with any size $n \in \mathbb{R}$ is possible and independent of $n$. Thus the asymptotic behaviour of $\{X_n\}_{n=1}^\infty$ also depends on the sampling method.

With extra assumptions or modelling of the population and proper random sampling method, we can assert more restrictions for the sample. For example, we may presume that the distribution of $\mathbf{X}$ is determined by some population parameters $\{\theta_i : i \in I\}$, i.e.,

$$\forall n \in \mathbb{R}, \mathbf{x} \in \mathbb{R}^n,\ F(\mathbf{x}) = F\left(\{x_i\}_{i=1}^n\right) = G\left(\{\theta_i : i \in I\}; \{x_i\}_{i=1}^n\right)$$

where $F$ is the joint *pdf.* of $\mathbf{X}$ and $G$ is a modelling function.
We may also want $\{X_i\}_{i=1}^n$ to be mutually independent and identically distributed. then,

- The *pdf.* of every $X_i$ is the same for any $i \in \{1, 2, \cdots, n\}$, possibly determined by the population parameter $\{\theta_i : i \in I\}$
- The asymptotic behaviour of $\{X_i\}_{n=1}^{\infty}$ is simply the behaviour of a infinite sequence of *i.i.d.* random variables.

**Definition**: Statistic
Statistics is a real function of the sample. It may summarise or conclude the information obtained from the sample, denoted as

$$s, \text{ or } s(\mathbf{X}) : \bar{\mathbb{R}}^n \mapsto \mathbb{R}$$

**Definition**: Point estimator and point estimate
Point estimator is the statistic

$$\hat{\theta}_i(X_1, X_2, \cdots, X_n) = \hat{\theta}_i(\mathbf{X}) \; \forall i \in I$$

that is designed to infer the population parameters $\{\theta_i\}_{i \in I}$. We may simply write $\hat{\theta}_i$, $\hat{\theta}$ (when there is only one parameter to estimate), $\hat{\theta}_{i,n}$ or $\hat{\theta}_n$ (when the sample size matters). The same as sample, the asymptotic behaviour of $\hat{\theta}_n$ when $n$ goes to infinity is also studied.
As a function, and in most cases, measurable function, of $\{X_i\}_{n=1}^{\infty}$, the estimators are also random variables of the same probability space $(\Omega, \mathcal{S}, P)$.
The realisations of estimators given an obtained sample are known as estimates, which is the same function that applies to a realization of sample,

$$\hat{\theta}_i(x_1, x_2, \cdots, x_n) \; \forall i \in I$$

For example, the sample mean

$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_n$$

is the estimator of population mean $\mu$. The sample variance

$$s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_n - \bar{X})^2 = \frac{1}{n-1} \left[ \left( \sum_{i=1}^{n} X_n^2 \right) - n\bar{X}^2 \right]$$

is the estimator of population variance $\sigma^2$.

### 8.1.3 Criteria for a good point estimator

There can be infinite number of estimators for a population parameter. The following shows some criteria for how we can evaluate whether an estimator is good or not.

**Definition**: Unbiasedness
An estimator is unbiased if its expectation is equal to the population parameter, i.e.

$$\mathbb{E}(\hat{\theta}) = \theta$$

or its biased, defined as follows, is 0,

$$\text{Bias}(\hat{\theta}) := \mathbb{E}(\hat{\theta} - \theta) = 0$$

**Definition**: asymptotical unbiasedness
An estimator is asymptotically unbiased if it tends to be unbiased when the sample size goes to infinity, i.e.

$$\lim_{n\to\infty} \mathbb{E}(\hat{\theta}_n) = \theta$$

**Definition**: Mean square error
The mean square error of a estimator $\hat{\theta}$ to population parameter $\theta$ is

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

For an estimator to be good, we want its mean square error to be minimized. Note that when $\text{var}(\hat{\theta})$ exists,

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left\{\left[\hat{\theta} - \mathbb{E}(\hat{\theta})\right]^2 + \mathbb{E}\left(\hat{\theta} - \theta\right)\left[2\hat{\theta} - \theta - \mathbb{E}(\theta)\right]\right\} = \text{var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$$

Thus, we prefer the estimator with 0 bias and less variance. The comparison of variances between two unbiased variables is known as the comparison of efficiency.

**Definition**: Relative efficiency
Let $\hat{\theta}, \tilde{\theta}$ be two unbiased estimations of the same parameter $\theta$, the efficiency of $\hat{\theta}$ relative to $\tilde{\theta}$ is defined to be

$$\text{Effi}(\hat{\theta}, \tilde{\theta}) = \frac{\text{var}(\tilde{\theta})}{\text{var}(\hat{\theta})}$$

$\hat{\theta}$ is more efficient than $\tilde{\theta}$ if $\text{Effi}(\hat{\theta}, \tilde{\theta}) > 1$.

**Definition**: uniformly minimum-variance unbiased estimator

Let $\Theta$ be a set of all unbiased estimators with finite variances for population parameter $\theta$, an estimator $\hat{\theta}$ is called the uniformly minimum-variance unbiased estimator (UMVUE), aka. minimum-variance unbiased estimator (MVUE), if

$$\text{Effi}(\hat{\theta}, \tilde{\theta}) = \frac{\text{var}(\tilde{\theta})}{\text{var}(\hat{\theta})} \geq 1 \ \forall \tilde{\theta} \in \Theta$$

With some desired conditions, $\inf\{\text{var}(\hat{\theta}) : \hat{\theta} \in \Theta\}$ can be obtained. Details to be discuss in the next section.

**Definition**: Consistency

An estimator is consistent if it converges in probability to the population parameter when the sample size goes to infinity, i.e.

$$\hat{\theta}_n \xrightarrow{\text{p.}} \theta$$

## 8.2 Likelihood function

### 8.2.1 Definition of Likelihood function

When the distribution of a random sample $\{X_i\}_{i=1}^n$ has a density that is determined by some population parameter $\{\theta_i : i \in I\}$. We can define the likelihood function that helps to get an estimator.

**Definition**: Likelihood function

When the following assumption is made,

1. Suppose the sample size is $n \in \mathbb{N}^*$, the joint density $f : R^n \to \mathbb{R}$ of $\{X_i\}_{i=1}^n$ exists with reference measure $\mu$ being the product measure of $\{(R_n, \mathcal{B}_n, \mu_i)\}_{i=1}^n$
2. There exists a finite number of population parameters $\{\theta_i\}_{i=1}^I$ (or denoted as $\Theta$ on the parameter space $S_p \subset \mathbb{R}^I$, where $I \in \mathbb{N}^*$), that determine the joint density of $\{X_i\}_{i=1}^n$ with a fix reference measure $(R, \mathcal{B}, \mu)$, i.e. $\exists g : \mathbb{R}^n \mapsto \mathbb{R}$ with parameter $\Theta \in S_p$ such that

$$f\left([x_1, x_2, \cdots, x_n]^T\right) = f\left([x_1, x_2, \cdots, x_n]^T; \Theta\right) \ \forall [x_1, x_2, \cdots, x_n]^T \in R$$

We can define the likelihood function to be

$$L(\Theta; \mathbf{x}) := L([\theta_1, \theta_2, \cdots, \theta_I]^T; [x_1, x_2, \cdots, x_n]^T) := g\left([x_1, x_2, \cdots, x_n]^T; \Theta\right)$$

which is a $S_p \mapsto \mathbb{R}$ function and takes $\mathbf{x} \in R$ as parameters. The parameters can be further regarded as random variables (i.e. a function from $\Omega$).

For any $\Theta \in S_p$, $f\left([X_1, X_2, \cdots, X_n]^T; \Theta\right) \overset{a.s.}{>} 0$. Further it is desirable that the set

$$\{\mathbf{X} \in R : f(\mathbf{X}; \Theta) > 0\}$$

does not depend on the choices of $\Theta \in S_p$. Then the following is defined.
**Definition**: Log-likelihood function
When $S := \{\mathbf{X} \in R : L(\Theta; \mathbf{X}) > 0\}$ is irrelevant to different choices of $\Theta$, we define the logarithm of the likelihood function,

$$l(\Theta; \mathbf{x}) := \ln L(\Theta; \mathbf{x}) \text{ whenever } \mathbf{x} \in S$$

in which $\ln : \mathbb{R}^+ \mapsto \mathbb{R}$ is defined to be the inverse function of $\exp : \mathbb{R} \mapsto \mathbb{R}^+$ possessing the following

1. $\forall x, y \in \mathbb{R}^+,\ \ln(xy) = \ln(x) + \ln(y)$

2. $\forall x \in \mathbb{R}^+,\ \ln(x)' = \frac{1}{x}$ Thus it is monotonic increasing.

**Example**: (Log-)likelihood function in *i.i.d.* case
If $\mathbf{X} = [X_1, X_2, \cdots, X_n]^T$ composes of *i.i.d.* random variables $\{X_i\}_{i=1}^n$, and $f_1(*, \Theta) : R_1 \mapsto \mathbb{R}$ (with the reference measure $\mu_1$ on $(R_1, \mathcal{B}_1)$ and parameter $\Theta \in S_p$) is the probability density for every $X_i$, then

1. $$L(\Theta, \mathbf{x}) = f\left([x_1, x_2, \cdots, x_n]^T; \Theta\right) = \prod_{i=1}^n f_1(x_i, \Theta)$$

2. $$l(\Theta, \mathbf{x}) = \sum_{i=1}^n \ln f_1(x_i, \Theta)$$

### 8.2.2 Bound of estimator variance and Fisher information

**Theorem**: Cramér–Rao lower bound of estimator variance (single unbiased estimator case)
Under the assumptions above that the likelihood function is defined, plus the condition stated below, we obtain the lower bound of the variance of an unbiased estimator $\hat{\theta}$ of $\theta$. (When there are multiple

parameters, this bound applies to any of them, e.g. $\theta_i, i \in \{1, 2, \cdots, I\}$, while assuming that the true values of parameters other than $\theta_i$ are attained)

$$\mathrm{var}(\hat{\theta}) \geq \mathbb{E}\left\{\left[\frac{\partial}{\partial\theta} l(\theta, \mathbf{X})\right]^2\right\}^{-1}$$

It holds only when the reference measure $(R, \mathcal{B}, \mu)$ $(R \in \mathcal{B}(\mathbb{R}^n), P(\mathbf{X} \notin R) = 0)$ and the parameter space $S_p \subset \mathbb{R}^I$ is chosen such that

1. (Existence of log-likelihood function) $\forall \mathbf{x} \in R, \theta \in S_p,\ 0 < L(\theta, \mathbf{x}) < +\infty$. The choice of $R$ does not depend on $\theta$
2. (Partial differentiability) $S_p \in \mathbb{R}^I$ is a open, and $\frac{\partial}{\partial\theta} L(\theta, \mathbf{x})$ exists and is finite for all $\theta \in S_p$ and almost every $\mathbf{x} \in R$
3. (Condition to interchange integration and differentiation) For any $\theta \in S_p$, there exists an open set $B \subset \mathbb{R}$ containing $\theta_i$ and a function $g : R \mapsto [0, +\infty], g \in \mathcal{L}^1(\mu)$ such that

   $$\forall t \in B \setminus \{\theta\},\ \left|\frac{\left[\hat{\theta}(\mathbf{x}) - t\right] L(\theta, \mathbf{x}) - \left[\hat{\theta}(\mathbf{x}) - \theta\right] L(\theta, \mathbf{x})}{t - \theta}\right| \leq g(\mathbf{x}) \text{ for } \mathbf{x} \in R \text{ almost everywhere}$$

   which gives $\left[\hat{\theta}(\mathbf{x}) - \theta\right] \frac{\partial}{\partial\theta} L(\theta, \mathbf{x}) \in \mathcal{L}^1(\mu)$ and

   $$\int_R \left[\hat{\theta}(\mathbf{x}) - \theta\right] \frac{\partial}{\partial\theta} L(\theta, \mathbf{x}) \mathrm{d}\mu(\mathbf{x}) = \frac{\partial}{\partial\theta} \int_R \left[\hat{\theta}(\mathbf{x}) - \theta\right] L(\Theta, \mathbf{x}) \mathrm{d}\mu(\mathbf{x})\ \forall\Theta \in S_p$$

   by dominated convergence theorem.

Remark:
The following will use the term "differentiable with the dominated condition" to express conditions 2 and 3 as a whole.
Here are some stronger substitutions for condition 3

1. $\mu(R) < \infty$
2. Both $L(\theta, \mathbf{x})$ and $\hat{\theta}(\mathbf{x}) L(\theta, \mathbf{x})$ is differentiable with the dominated condition such that it is possible to interchange the integration and partial differentiation.

Proof:
From the unbiasedness, we have

$$\forall\theta \in S_p,\ 0 = \mathbb{E}(\hat{\theta} - \theta) = \int_R \left[\hat{\theta}(\mathbf{x}) - \theta\right] L(\theta, \mathbf{x}) \mathrm{d}\mu(\mathbf{x})$$

Thus we can take the partial derivative of both sides with regard to $\theta$ and further interchange integration and differentiation

$$\begin{aligned}
\forall\theta \in S_p,\ 0 &= \frac{\partial}{\partial\theta}\int_R \left[\hat{\theta}(\mathbf{x}) - \theta\right] L(\theta, \mathbf{x})\mathrm{d}\mu(\mathbf{x}) \\
&= \int_R \frac{\partial}{\partial\theta}\left[\hat{\theta}(\mathbf{x}) - \theta\right] L(\theta, \mathbf{x})\mathrm{d}\mu(\mathbf{x}) \\
&= \int_R \left[\hat{\theta}(\mathbf{x}) - \theta\right] \frac{\partial}{\partial\theta} L(\theta, \mathbf{x}) - L(\theta, \mathbf{x})\mathrm{d}\mu(\mathbf{x}) \text{ (by the chain rule)} \\
&= \int_R \left[\hat{\theta}(\mathbf{x}) - \theta\right] \frac{\partial}{\partial\theta} L(\theta, \mathbf{x})\mathrm{d}\mu(\mathbf{x}) - \int_R L(\theta, \mathbf{x})\mathrm{d}\mu(\mathbf{x}) \\
&= \int_R \left[\hat{\theta}(\mathbf{x}) - \theta\right] \frac{\partial}{\partial\theta} L(\theta, \mathbf{x})\mathrm{d}\mu(\mathbf{x}) - 1
\end{aligned}$$

Then, since

$$\frac{\partial}{\partial\theta} l(\theta, \mathbf{x}) = \frac{\partial}{\partial\theta}\ln L(\theta, \mathbf{x}) = \frac{1}{L(\theta, \mathbf{x})} \times \frac{\partial}{\partial\theta} L(\theta, \mathbf{x})\ \forall\theta \in S_p, \mathbf{x} \in R$$

We plug in $\frac{\partial}{\partial\theta} L(\theta, \mathbf{x}) = L(\theta, \mathbf{x}) \times \frac{\partial}{\partial\theta} l(\theta, \mathbf{x})$,

$$1 = \int_R \left[\hat{\theta}(\mathbf{x}) - \theta\right] \frac{\partial}{\partial\theta} L(\theta, \mathbf{x})\mathrm{d}\mu(\mathbf{x}) = \int_R \left[\hat{\theta}(\mathbf{x}) - \theta\right] L(\theta, \mathbf{x}) \times \frac{\partial}{\partial\theta} l(\theta, \mathbf{x})\mathrm{d}\mu(\mathbf{x})$$

We use the Cauchy-Schwarz inequality $\left(\int_S fg\mathrm{d}\mu\right)^2 \leq \int_S f^2\mathrm{d}\mu \times \int_S g^2\mathrm{d}\mu\ \forall f, g \in \mathcal{L}^+(\mu)$,

$$\begin{aligned}
1^2 &= \left\{\int_R \left[\hat{\theta}(\mathbf{x}) - \theta\right] L(\theta, \mathbf{x}) \times \frac{\partial}{\partial\theta} l(\theta, \mathbf{x})\mathrm{d}\mu(\mathbf{x})\right\}^2 \\
&\leq \left\{\int_R \left|\hat{\theta}(\mathbf{x}) - \theta\right| \sqrt{L(\theta, \mathbf{x})} \times \sqrt{L(\theta, \mathbf{x})}\left|\frac{\partial}{\partial\theta} l(\theta, \mathbf{x})\right|\mathrm{d}\mu(\mathbf{x})\right\}^2 \\
&\leq \int_R \left[\hat{\theta}(\mathbf{x}) - \theta\right]^2 L(\theta, \mathbf{x})\mathrm{d}\mu(\mathbf{x}) \times \int_R L(\theta, \mathbf{x}) \left[\frac{\partial}{\partial\theta} l(\theta, \mathbf{x})\right]^2 \mathrm{d}\mu(\mathbf{x}) \\
&= \mathbb{E}\left\{\left[\hat{\theta}(\mathbf{X}) - \theta\right]^2\right\} \times \mathbb{E}\left\{\left[\frac{\partial}{\partial\theta} l(\theta, \mathbf{x})\right]^2\right\}
\end{aligned}$$

Since $\mathbb{E}\left[\hat{\theta}(\mathbf{X}) - \theta\right] = 0$, we have finally

$$\operatorname{var}\left(\hat{\theta}\right) = \operatorname{var}\left[\hat{\theta}(\mathbf{X}) - \theta\right] = \mathbb{E}\left\{\left[\hat{\theta}(\mathbf{X}) - \theta\right]^2\right\} \geq \mathbb{E}\left\{\left[\frac{\partial}{\partial\theta} l(\theta, \mathbf{x})\right]^2\right\}^{-1}$$

Note that if $\mathbb{E}\left\{\left[\frac{\partial}{\partial\theta} l(\theta, \mathbf{x})\right]^2\right\} = +\infty$, $\operatorname{var}(\theta) \geq 0$, which means we have no extra information about it.

We make a definition of the denominator of the Cramér–Rao lower bound. It is a way of measuring the amount of information that a sample carries about unknown parameters with many properties of our interests.

**Definition**: Fisher information

With the same assumptions and conditions of Cramér–Rao lower bound applies, we define

$$\mathcal{I}(\theta) := \mathbb{E}\left\{\left[\frac{\partial}{\partial\theta} l(\theta, \mathbf{X})\right]^2\right\}$$

to be the Fisher information for parameter $\theta$ obtained from sample $\mathbf{X}$.

Then the Cramér–Rao inequality can be express as

$$\operatorname{var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}(\theta)}$$

**Proposition**: properties about Fisher information

The following further requires $L(\theta, \mathbf{X})$ to be differentiable with the dominated condition.

1. $\mathbb{E}\left[\frac{\partial}{\partial\theta} l(\theta, \mathbf{X})\right] = 0$, then obviously, $\mathcal{I}(\theta) = \operatorname{var}\left[\frac{\partial}{\partial\theta} l(\theta, \mathbf{X})\right]$

2. If $L(\theta, \mathbf{X})$ is twice partial differentiable with respect to $\theta$ with the dominated condition, (i.e. $\frac{\partial}{\partial\theta} L(\theta, \mathbf{X})$ is differentiable with the dominated condition), then

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} l(\theta, \mathbf{X})\right]$$

3. If $\mathbf{X} = [X_1, X_2, \cdots, X_n]^T$ composes of *i.i.d.* random variables $\{X_i\}_{i=1}^n$, and $f_1(*, \theta) : R_1 \mapsto \mathbb{R}$ (with the reference measure $\mu_1$ on $(R_1, \mathcal{B}_1)$ and parameter $\theta$) is the probability density for every $X_i$, then

$$\mathcal{I}(\theta) = n\mathbb{E}\left\{\left[\frac{\partial}{\partial\theta} f_1(\theta, X_1)\right]^2\right\}$$

Proof:

We first show 1.

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial}{\partial\theta} l(\theta, \mathbf{x})\right] =& \int_R \left[\frac{\partial}{\partial\theta} \ln L(\theta, \mathbf{x})\right] L(\theta, \mathbf{x}) \mathrm{d}\mu(\mathbf{x}) \\
=& \int_R \left[\frac{\partial}{\partial\theta} L(\theta, \mathbf{x})\right] \frac{1}{L(\theta, \mathbf{x})} L(\theta, \mathbf{x}) \mathrm{d}\mu(\mathbf{x}) \\
=& \frac{\partial}{\partial\theta} \int_R L(\theta, \mathbf{x}) \mathrm{d}\mu(\mathbf{x}) = \frac{\partial}{\partial\theta} 1 = 0
\end{aligned}$$

Then 2 is similar.

$$
\begin{aligned}
\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}l(\theta,\mathbf{x})\right] &= \int_R \left[\frac{\partial^2}{\partial^2\theta}\ln L(\theta,\mathbf{x})\right] L(\theta,\mathbf{x})\mathrm{d}\mu(\mathbf{x}) \\
&= \int_R \left\{\left[\frac{\partial^2}{\partial\theta^2}L(\theta,\mathbf{x})\right]\times\frac{1}{L(\theta,\mathbf{x})} - \left[\frac{\partial}{\partial\theta}L(\theta,\mathbf{x})\right]^2\times\frac{1}{L(\theta,\mathbf{x})^2}\right\} L(\theta,\mathbf{x})\mathrm{d}\mu(\mathbf{x}) \\
&= \int_R \left[\frac{\partial^2}{\partial\theta^2}L(\theta,\mathbf{x})\right]\mathrm{d}\mu(\mathbf{x}) - \int_R\left[\frac{\partial}{\partial\theta}l(\theta,\mathbf{x})\right]^2 L(\theta,\mathbf{x})\mathrm{d}\mu(\mathbf{x}) \\
&= \frac{\partial^2}{\partial\theta^2}\int_R L(\theta,\mathbf{x})\mathrm{d}\mu(\mathbf{x}) - \mathbb{E}\left\{\left[\frac{\partial}{\partial\theta}l(\theta,\mathbf{x})\right]^2\right\} \\
&= 0 - \mathcal{I}(\theta)
\end{aligned}
$$

To prove 3, we note that

$$
\begin{aligned}
\mathcal{I}(\theta) &= \mathbb{E}\left\{\left[\frac{\partial}{\partial\theta}l(\theta,\mathbf{X})\right]^2\right\} \\
&= \mathbb{E}\left\{\left[\sum_{i=1}^n\frac{\partial}{\partial\theta}\ln f_1(X_i,\theta)\right]^2\right\} \\
&= \mathbb{E}\left\{\sum_{i=1}^n\left[\frac{\partial}{\partial\theta}\ln f_1(X_i,\theta)\right]^2\right\} + \mathbb{E}\left\{2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[\frac{\partial}{\partial\theta}\ln f_1(X_i,\theta)\right]\left[\frac{\partial}{\partial\theta}\ln f_1(X_j,\theta)\right]\right\} \\
&= \sum_{i=1}^n\mathbb{E}\left\{\left[\frac{\partial}{\partial\theta}\ln f_1(X_i,\theta)\right]^2\right\} + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}\left[\frac{\partial}{\partial\theta}\ln f_1(X_i,\theta)\right]\mathbb{E}\left[\frac{\partial}{\partial\theta}\ln f_1(X_j,\theta)\right] \\
&= n\mathbb{E}\left\{\left[\frac{\partial}{\partial\theta}\ln f_1(\theta,X_1)\right]^2\right\} \quad \text{(by 1)}
\end{aligned}
$$

**Proposition**: More cases of lower bound of estimator variance
When some similar regularity conditions (i.e. existence of log-likelihood, differentiability, dominated condition to interchange differentiation and integration) are satisfied, the following hold

1. For a more general statistics $s(\mathbf{X})$ who has finite expectation dependent on a parameter $\theta$,

$$\mathbb{E}[s(\mathbf{X})] = \phi(\theta) < \infty$$

   we have

$$\mathrm{var}[s(\mathbf{X})] \geq \frac{1}{I_n(\theta)}\left[\phi'(\theta)\right]^2$$

2. Multivariable?

Definition: Sufficient estimator

### 8.2.3 Maximum likelihood function

**Definition**: Maximum likelihood estimator
With the assumption stated above, the maximum likelihood estimators of $\{\theta_i\}_{i=1}^I$ is $\{\hat{\theta}_i\}_{i=1}^I$ on the parameter space $S_p \subset \mathbb{R}^I$ that maximise the likelihood function, denoted as

$$[\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_I]^T = \hat{\Theta}(\mathbf{X}) = \underset{\Theta \in S_p}{\operatorname{argmax}}\ L(\Theta, \mathbf{X}) := \{\Theta_0 \in \mathbb{R}^I : L(\Theta_0, \mathbf{X}) \geq L(\Theta, \mathbf{X})\ \forall \Theta \in S_p\}$$

while its realisation is the corresponding $\hat{\Theta}(\mathbf{x})$.
$\hat{\Theta}$ is not unique from the definition. However, we usually choose the model with nice enough $g$ and proper $S_p$ to ensure the uniqueness.
Since ln is monotonic increasing, it is equivalent for $\hat{\Theta}$ to be the one that maximises the log-likelihood function $l(\Theta; \mathbf{X})$.

# Chapter 9: Hypothesis testing