HKU STAT6009/STAT7609 Study Notes

Ka Long (Leo) Chiu*

Last Updated: 2025-12-18

Contents

*email ✉: leockl@connect.hku.hk; personal website 🔗: https://leochiukl.github.io

1 Modes of Convergence

1.0.1 In STAT7609, there are two main topics to be studied: (i) measure theoretic probability theory, and (ii) theories for statistical methods, with (i) forming the foundation for (ii). Since the content for (i) has substantial overlap with the topics covered in STAT7610 (link for the study notes), here our emphasis is on the part whose discussion is absent/relatively light in STAT7610. It is recommended for the reader to be familiar with the topics covered in STAT7610 (at least to some degree), which is rather important to have a swift learning experience here.

1.0.2 **A few useful results.** The following results in probability theory are quite useful in STAT7609, so they are reviewed/introduced below for reference.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$ be a collection of events.

(a) *(First Borel-Cantelli lemma)* If $\sum_{i=1}^{\infty} \mathbb{P}(A_i) < \infty$, then $\mathbb{P}(\{\omega \in A_n \text{ io}\}) = 0$.

(b) *(Second Borel-Cantelli lemma)* If $\{A_i\}_{i\in\mathbb{N}}$ is pairwise independent and $\sum_{i=1}^{\infty} \mathbb{P}(A_i) = \infty$, then $\mathbb{P}(\{\omega \in A_n \text{ io}\}) = 1$.

[Note: It is indeed more common to assume that $\{A_i\}_{i\in\mathbb{N}}$ is independent, which makes this result easier to be proved. However, this assumption can be weakened to pairwise independence, although the proof in that case becomes harder; see Chung (2007, Theorem 4.2.5).]

(c) *(Borel 0-1 law)* If $\{A_i\}_{i\in\mathbb{N}}$ is pairwise independent, then:

$$\begin{cases} \mathbb{P}(\{\omega \in A_n \text{ io}\}) = 0 \iff \sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty, \\ \mathbb{P}(\{\omega \in A_n \text{ io}\}) = 1 \iff \sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty. \end{cases}$$

(d) *(Tail probability bounds)* Let $h : [0, \infty) \to [0, \infty)$ be a strictly increasing function and X be a random variable. Then, $\mathbb{P}(|X| \geq x) \leq \mathbb{E}[h(|X|)]/h(x)$ for all $x > 0$. [Note: The inequality with $h(x) = x$ and $h(x) = x^2$ is known as **Markov's inequality** and **Chebyshev's inequality**, respectively.]

1.1 Definitions and Equivalent Criteria of Modes of Convergence

1.1.1 **Definitions.** In STAT7609, we will focus on convergence of random *variables* (one-dimensional case), unlike the treatment in STAT7610 (which considers random vectors generally).

Let X be a random variable, and $\{X_n\}$ be a sequence of random variables. Let $X_n \sim F_n$ for every $n \in \mathbb{N}$ and $X \sim F$. Then, $\{X_n\}$ **converges to X in distribution (or weakly)**, denoted by $X_n \xrightarrow[n\to\infty]{\text{d}} X$, if $\lim_{n\to\infty} F_n(x) = F(x)$ for all continuity points x of F (i.e., x at which F is continuous).

Now, suppose also that X and all X_n's are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then:

- $\{X_n\}$ **converges almost surely** to X, denoted by $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X$, if $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$.
- $\{X_n\}$ **converges in probability** to X, denoted by $X_n \xrightarrow[n\to\infty]{\text{p}} X$, if for all $\varepsilon > 0$, we have $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$.
- $\{X_n\}$ **converges to X in L^p (or in the pth mean)**, denoted by $X_n \xrightarrow[n\to\infty]{L^p} X$, if $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0$ with $X_n \in L^p$ for all $n \in \mathbb{N}$ and $X \in L^p$, where $p > 0$.

1.1.2 **Equivalent criteria for almost sure convergence.** To establish the almost sure convergence, sometimes the following equivalent criteria are helpful (some of them will be used in later proofs), which resemble the definition of convergence for real-valued sequences learnt in calculus/analysis.

Proposition 1.1.a (Equivalent criteria for almost sure convergence)**.** The following are equivalent.

(a) $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X$.

(b) For all $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_m - X| < \varepsilon \text{ for all } m \geq n) = \lim_{n\to\infty} \mathbb{P}(\bigcap_{m=n}^{\infty} \{|X_m - X| < \varepsilon\}) = 1$.

(c) For all $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_m - X| \geq \varepsilon \text{ for some } m \geq n) = \lim_{n\to\infty} \mathbb{P}(\bigcup_{m=n}^{\infty} \{|X_m - X| \geq \varepsilon\}) = 0$.

(d) $\sup_{m\geq n} |X_m - X| \xrightarrow[n\to\infty]{\text{p}} 0$, i.e., for all $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\big(\big[\sup_{m\geq n} |X_m - X|\big] > \varepsilon\big) = 0$

(e) For all $\varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon \text{ io}) = 0$.

Moreover, "$< \varepsilon$", "$\geq \varepsilon$", and "$> \varepsilon$" above can be replaced by "$\leq \varepsilon$", "$> \varepsilon$", and "$\geq \varepsilon$", respectively.

Proof.

- "(a) $\iff$ (b)": Note that

$$\begin{aligned}\lim_{n\to\infty} X_n(\omega) = X(\omega) &\iff \forall \varepsilon > 0, \exists n \in \mathbb{N} \text{ s.t. } \forall m \geq n,\ |X_m(\omega) - X(\omega)| < \varepsilon \\ &\iff \forall k \in \mathbb{N}, \exists n \in \mathbb{N} \text{ s.t. } \forall m \geq n,\ |X_m(\omega) - X(\omega)| < 1/k.\end{aligned}$$

Therefore, we can write

$$\begin{aligned}\Big\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big\} &= \bigcap_{k=1}^{\infty} \bigcup_{n=1}^{\infty} \underbrace{\bigcap_{m=n}^{\infty} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| < 1/k\}}_{\text{increasing in } n} \\ &= \bigcap_{k=1}^{\infty} \underbrace{\lim_{n\to\infty} \bigcap_{m=n}^{\infty} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| < 1/k\}}_{\text{decreasing in } k} \qquad (1) \\ &= \lim_{k\to\infty} \lim_{n\to\infty} \bigcap_{m=n}^{\infty} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| < 1/k\} \qquad (2)\end{aligned}$$

Now we are ready to establish the equivalence. Assume that (a) holds. Then, for all $k_0 \in \mathbb{N}$, by Equation (1) we have

$$\begin{aligned}1 &\overset{\text{(assumption)}}{=} \mathbb{P}\Big(\Big\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big\}\Big) \\ &= \mathbb{P}\left(\bigcap_{k=1}^{\infty} \lim_{n\to\infty} \bigcap_{m=n}^{\infty} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| < 1/k\}\right) \\ &\leq \mathbb{P}\left(\lim_{n\to\infty} \bigcap_{m=n}^{\infty} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| < 1/k_0\}\right) \\ &= \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{m=n}^{\infty} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| < 1/k_0\}\right) \leq 1,\end{aligned}$$

which implies that $\lim_{n\to\infty} \mathbb{P}(\bigcap_{m=n}^{\infty} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| < 1/k_0\}) = 1$. Now fix any $\varepsilon > 0$. As we have $1/M < \varepsilon$ for some sufficiently large $M \in \mathbb{N}$, it follows from above that

$$1 = \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{m=n}^{\infty} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| < 1/M\}\right) \leq \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{m=n}^{\infty} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| < \varepsilon\}\right) \leq 1,$$

implying that $\lim_{n\to\infty} \mathbb{P}(\bigcap_{m=n}^{\infty} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| < \varepsilon\}) = 1$, thus establishing (b).
Now assume that (b) holds. By Equation (2), we have

$$\begin{aligned}\mathbb{P}\Big(\Big\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big\}\Big) &= \mathbb{P}\left(\lim_{k\to\infty} \lim_{n\to\infty} \bigcap_{m=n}^{\infty} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| < 1/k\}\right) \\ &= \lim_{k\to\infty} \underbrace{\lim_{n\to\infty} \mathbb{P}\left(\bigcap_{m=n}^{\infty} \{\omega \in \Omega : |X_m(\omega) - X(\omega)| < 1/k\}\right)}_{=1 \text{ by (b)}} \\ &= 1,\end{aligned}$$

hence (a) holds.

- "(b) $\iff$ (c)": It follows from De Morgan's law.
- "(c) $\iff$ (d)": Fix any $\varepsilon > 0$. Let

$$A_{n,\varepsilon} := \bigcup_{m=n}^{\infty} \{|X_m - X| \geq \varepsilon\} \quad \text{and} \quad B_{n,\varepsilon} := \left\{ \left[\sup_{m \geq n} |X_m - X| \right] \geq \varepsilon \right\}$$

for all $n \in \mathbb{N}$.

Claim: $A_{n,\varepsilon} \subseteq B_{n,\varepsilon} \subseteq A_{n,\varepsilon/2}$ for all $n \in \mathbb{N}$.

Proof. The first subset inclusion holds since

$$\begin{aligned} \omega \in B_{n,\varepsilon}^c &\implies \sup_{m \geq n} |X_m(\omega) - X(\omega)| < \varepsilon \\ &\implies |X_m(\omega) - X(\omega)| < \varepsilon \quad \forall m \geq n \\ &\implies \omega \in \bigcap_{m=n}^{\infty} \{|X_m(\omega) - X(\omega)| < \varepsilon\} = A_{n,\varepsilon}^c. \end{aligned}$$

The second subset inclusion holds since

$$\begin{aligned} \omega \in A_{n,\varepsilon/2}^c &\implies |X_m(\omega) - X(\omega)| < \varepsilon/2 \quad \forall m \geq n \\ &\implies \sup_{m \geq n} |X_m(\omega) - X(\omega)| \leq \varepsilon/2 < \varepsilon \\ &\implies \omega \in B_{n,\varepsilon}^c. \end{aligned}$$

□

The equivalence of (c) and (d) then follows from this chain of subset inclusions.

- "(c) $\iff$ (e)": The equivalence follows from noting that

$$\begin{aligned} \mathbb{P}(|X_n - X| > \varepsilon \text{ io}) &= \mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} \{|X_m - X| > \varepsilon\} \right) \\ &= \mathbb{P}\left(\lim_{n \to \infty} \bigcup_{m=n}^{\infty} \{|X_m - X| > \varepsilon\} \right) \\ &= \lim_{n \to \infty} \mathbb{P}\left(\bigcup_{m=n}^{\infty} \{|X_m - X| > \varepsilon\} \right) \end{aligned}$$

for all $\varepsilon > 0$.

By inspecting the argument above, we can observe that "$< \varepsilon$", "$\geq \varepsilon$", and "$> \varepsilon$" can be replaced by "$\leq \varepsilon$", "$> \varepsilon$", and "$\geq \varepsilon$", respectively. □

1.1.3 **Cauchy criteria for almost sure convergence.** To establish that $\{X_n\}$ converges almost surely, i.e., $X_n \xrightarrow[n \to \infty]{\text{a.s.}} X$ for some random variable X (with X often being unknown), the following Cauchy criteria are helpful, which resemble the well-known Cauchy convergence criterion for real-valued sequences learnt in calculus/analysis.

Proposition 1.1.b (Cauchy criteria for almost sure convergence)**.** The following are equivalent.

(a) X_n converges almost surely.

(b) For all $\varepsilon > 0$, $\lim_{n \to \infty} \mathbb{P}(|X_m - X_\ell| < \varepsilon \text{ for all } m, \ell \geq n) = \lim_{n \to \infty} \mathbb{P}(\bigcap_{m=n}^{\infty} \bigcap_{\ell=n}^{\infty} |X_m - X_\ell| < \varepsilon) = 1$.

(c) For all $\varepsilon > 0$, $\lim_{n \to \infty} \mathbb{P}(|X_m - X_\ell| < \varepsilon \text{ for some } m, \ell \geq n) = \lim_{n \to \infty} \mathbb{P}(\bigcup_{m=n}^{\infty} \bigcup_{\ell=n}^{\infty} |X_m - X_\ell| \geq \varepsilon) = 0$.

(d) $\sup_{m,\ell\geq n}|X_m - X_\ell| \xrightarrow[n\to\infty]{\text{p}} 0$.

Moreover, "$<\varepsilon$" and "$\geq\varepsilon$" above can be replaced by "$\leq\varepsilon$" and "$>\varepsilon$", respectively.

Proof.

- "(a) $\iff$ (b)": Similar to the proof for "(a) $\iff$ (b)" of Proposition 1.1.a, but using the Cauchy criterion for convergence of real-valued sequences instead of the definition.
- "(b) $\iff$ (c)": It again follows from De Morgan's law.
- "(c) $\iff$ (d)": Similar to the proof for "(c) $\iff$ (d)" of Proposition 1.1.a.

□

1.1.4 **Cauchy criteria for convergence in probability.** Apart from almost sure convergence, there are also Cauchy criteria for *convergence in probability* as follows:

Proposition 1.1.c (Cauchy criteria for convergence in probability)**.** The following are equivalent.

(a) X_n converges in probability (i.e., $X_n \xrightarrow[n\to\infty]{\text{p}} X$ for some random variable X).

(b) For all $\varepsilon > 0$, $\lim_{m,n\to\infty}\mathbb{P}(|X_m - X_n| < \varepsilon) = 1$ (i.e., for all $\varepsilon' > 0$, there exists $N = N(\varepsilon') \in \mathbb{N}$ such that $\left|\mathbb{P}(|X_m - X_n| < \varepsilon) - 1\right| < \varepsilon'$ for all $m, n \geq N$).

(c) For all $\varepsilon > 0$, $\lim_{m,n\to\infty}\mathbb{P}(|X_m - X_n| \geq \varepsilon) = 0$.

Moreover, "$<\varepsilon$" and "$\geq\varepsilon$" above can be replaced by "$\leq\varepsilon$" and "$>\varepsilon$", respectively.

Proof. Omitted. □

1.2 Relationships Between Modes of Convergence

1.2.1 **Implications between modes of convergence.** We start with establishing implications between modes of convergence for the one-dimensional case. While they have already been shown as a special case in STAT7610, the arguments used in the proof here are a bit different and tend to be simpler (since we focus on the special one-dimensional case, not the general case).

Theorem 1.2.a (Implications between modes of convergence)**.**

(a) $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X \implies X_n \xrightarrow[n\to\infty]{\text{p}} X$.

(b) $X_n \xrightarrow[n\to\infty]{L^p} X \implies X_n \xrightarrow[n\to\infty]{\text{p}} X$ for all $p > 0$.

(c) $X_n \xrightarrow[n\to\infty]{\text{p}} X \implies X_n \xrightarrow[n\to\infty]{\text{d}} X$.

(d) For all $q > p > 0$, $X_n \xrightarrow[n\to\infty]{L^q} X \implies X_n \xrightarrow[n\to\infty]{L^p} X$.

Proof.

(a) Assume $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X$. By Proposition 1.1.a, we have $\lim_{n\to\infty}\mathbb{P}(\bigcup_{m=n}^{\infty}\{|X_m - X| \geq \varepsilon\}) = 0$. Hence,

$$0 \leq \lim_{n\to\infty}\mathbb{P}(|X_n - X| > \varepsilon) \leq \lim_{n\to\infty}\mathbb{P}\left(\bigcup_{m=n}^{\infty}\{|X_m - X| \geq \varepsilon\}\right) = 0,$$

implying that $X_n \xrightarrow[n\to\infty]{\text{p}} X$.

(b) Assume $X_n \xrightarrow[n\to\infty]{L^p} X$. Applying [1.0.2]d with $h(x) = x^p$ where $p > 0$, we get

$$0 \le \lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) \le \lim_{n\to\infty} \frac{\mathbb{E}[|X_n - X|^p]}{\varepsilon^p} \overset{\text{(assumption)}}{=} 0$$

for all $\varepsilon > 0$, thus $X_n \xrightarrow[n\to\infty]{\text{p}} X$.

(c) Assume $X_n \xrightarrow[n\to\infty]{\text{p}} X$. Let F_n be the CDF of X_n for all $n \in \mathbb{N}$, and F be the CDF of X. Fix any $\varepsilon > 0$ and $x \in \mathbb{R}$ at which F is continuous.

Establishing an upper bound on $F_n(x)$. We have

$$\begin{aligned} F_n(x) &= \mathbb{P}(X_n \le x \cap |X_n - X| \le \varepsilon) + \mathbb{P}(X_n \le x \cap |X_n - X| > \varepsilon) \\ &\le \mathbb{P}(X \le x - (X_n - X) \cap |X_n - X| \le \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon) \\ &\le \mathbb{P}(X \le x + \varepsilon \cap |X_n - X| \le \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon) \\ &\le \mathbb{P}(X \le x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon) \\ &= F(x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon) \end{aligned}$$

for all $n \in \mathbb{N}$.

Establishing a lower bound on $F_n(x)$. Also, we have

$$\begin{aligned} F_n(x) &= 1 - \mathbb{P}(X_n > x \cap |X_n - X| \le \varepsilon) - \mathbb{P}(X_n > x \cap |X_n - X| > \varepsilon) \\ &\ge 1 - \mathbb{P}(X > x - (X_n - X) \cap |X_n - X| \le \varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \\ &\ge 1 - \mathbb{P}(X > x - \varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \\ &= F(x - \varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \end{aligned}$$

for all $n \in \mathbb{N}$.

Combining upper and lower bounds and letting $n \to \infty$. From above, for all $n \in \mathbb{N}$ we have

$$F(x - \varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \le F_n(x) \le F(x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon).$$

Letting $n \to \infty$ (taking $\liminf_{n\to\infty}$ for the first inequality and $\limsup_{n\to\infty}$ for the second inequality since we do not know whether $\lim_{n\to\infty} F_n(x)$ exists at this stage) gives

$$F(x - \varepsilon) \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le F(x + \varepsilon),$$

by the assumption that $X_n \xrightarrow[n\to\infty]{\text{p}} X$.

Since F is continuous at x, letting $\varepsilon \to 0^+$ yields $F(x) \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le F(x)$, implying that $\liminf_{n\to\infty} F_n(x) = \limsup_{n\to\infty} F_n(x) = F(x)$, hence $\lim_{n\to\infty} F_n(x) = F(x)$.

(d) Assume $X_n \xrightarrow[n\to\infty]{L^q} X$. For all $n \in \mathbb{N}$, we have

$$0 \le \mathbb{E}[|X_n - X|^p]^{1/p} \overset{\text{(property)}}{\le} \mathbb{E}[|X_n - X|^q]^{1/q} \xrightarrow[n\to\infty]{\text{(assumption)}} 0,$$

implying that $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0$, hence $X_n \xrightarrow[n\to\infty]{L^p} X$.

□

1.2.2 **Equivalence between convergence to constant in probability and in distribution.** We have $X_n \xrightarrow[n\to\infty]{\text{d}} c \iff X_n \xrightarrow[n\to\infty]{\text{p}} c$ where $c \in \mathbb{R}$ is a constant.

Proof. "⇐" follows readily from Theorem 1.2.a, so it suffices to prove "⇒". Let F_n be the CDF of X_n for all $n \in \mathbb{N}$, and F be the CDF of c (a random variable taking the value c always). Assume $X_n \xrightarrow[n\to\infty]{\text{d}} c$ and fix any $\varepsilon > 0$. Then, for all $n \in \mathbb{N}$ we have

$$0 \le \mathbb{P}(|X_n - c| > \varepsilon) \le \mathbb{P}(X_n > c + \varepsilon) + \mathbb{P}(X_n \le c - \varepsilon) = 1 - F_n(c + \varepsilon) + F_n(c - \varepsilon).$$

Since $\lim_{n\to\infty} F_n(c+\varepsilon) = F(c+\varepsilon) = 1$ and $\lim_{n\to\infty} F_n(c-\varepsilon) = F(c-\varepsilon) = 0$ (as F is continuous at $c \pm \varepsilon$), we get $\lim_{n\to\infty} \mathbb{P}(|X_n - c| > \varepsilon) = 0$, thus $X_n \xrightarrow[n\to\infty]{\text{p}} c$. □

1.2.3 Dominated convergence in probability implies convergence in L^p.

Lemma 1.2.b. If $X_n \xrightarrow[n\to\infty]{\text{p}} X$ and $|X_n| \le Y$ a.s. for all $n \in \mathbb{N}$, then $|X| \le Y$ a.s.

Proof. Fix any $\delta > 0$. For all $n \in \mathbb{N}$, we have

$$\begin{aligned} 0 \le \mathbb{P}(|X| > Y + \delta) &= \mathbb{P}(|X| > Y + \delta \cap |X_n| \le Y) + \mathbb{P}(|X| > Y + \delta \cap |X_n| > Y) \\ &\le \mathbb{P}(|X| > |X_n| + \delta \cap |X_n| \le Y) + \mathbb{P}(|X_n| > Y) \\ &\le \mathbb{P}(|X| > |X_n| + \delta) + \mathbb{P}(|X_n| > Y) \\ &\overset{\text{(triangle inequality, assumption)}}{\le} \mathbb{P}(|X_n - X| > \delta) \overset{\text{(assumption)}}{\underset{n\to\infty}{\longrightarrow}} 0. \end{aligned}$$

Hence, we have $\mathbb{P}(|X| > Y + \delta) = 0$. Since this holds for all $\delta > 0$, we have

$$\mathbb{P}(|X| > Y) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} \{|X| > Y + 1/n\}\right) = \mathbb{P}\Big(\lim_{n\to\infty} \{|X| > Y + 1/n\}\Big) = \lim_{n\to\infty} \mathbb{P}(|X| > Y + 1/n) = 0,$$

thus $|X| \le Y$ a.s. □

Lemma 1.2.c. If $\mathbb{E}[|X|] < \infty$ and $\lim_{n\to\infty} \mathbb{P}(A_n) = 0$, then $\lim_{n\to\infty} \mathbb{E}[|X|\mathbf{1}_{A_n}] = 0$.

Proof. Since $\mathbb{E}[|X|] < \infty$, we have $|X| < \infty$ a.s., and thus $\lim_{n\to\infty} |X|\mathbf{1}_{\{|X|>a\}} = 0$ a.s. for all $a > 0$. Furthermore, we have $|X|\mathbf{1}_{\{|X|>n\}} \le |X|$ for all $a > 0$. Hence, applying DCT gives $\lim_{a\to\infty} \mathbb{E}\big[|X|\mathbf{1}_{\{|X|>a\}}\big] = 0$. So, for all $\varepsilon > 0$, there exists $M > 0$ such that $\mathbb{E}\big[|X|\mathbf{1}_{\{|X|>M\}}\big] < \varepsilon$. It follows that for all $n \in \mathbb{N}$ we have

$$0 \le \mathbb{E}[|X|\mathbf{1}_{A_n}] = \mathbb{E}\big[|X|\mathbf{1}_{\{|X|>M\}}\mathbf{1}_{A_n}\big] + \mathbb{E}\big[|X|\mathbf{1}_{\{|X|\le M\}}\mathbf{1}_{A_n}\big] \le \mathbb{E}\big[|X|\mathbf{1}_{\{|X|>M\}}\big] + \mathbb{E}[M\mathbf{1}_{A_n}] < \varepsilon + M\mathbb{P}(A_n).$$

Letting $\varepsilon \to 0^+$ then yields $0 \le \mathbb{E}[|X|\mathbf{1}_{A_n}] \le M\mathbb{P}(A_n) \overset{\text{(assumption)}}{\underset{n\to\infty}{\longrightarrow}} 0$, hence $\lim_{n\to\infty} \mathbb{E}[|X|\mathbf{1}_{A_n}] = 0$. □

Theorem 1.2.d (Dominated convergence in probability implies convergence in L^p)**.** Suppose that $|X_n| \le Y$ a.s. for all $n \in \mathbb{N}$ with $\mathbb{E}[Y^p] < \infty$ for some $p > 0$ *(domination)*. Then, $X_n \xrightarrow[n\to\infty]{\text{p}} X \implies X_n \xrightarrow[n\to\infty]{L^p} X$.

Proof. By Lemma 1.2.b, we have $|X| \le Y$ a.s., and thus $|X_n - X| \le |X_n| + |X| \le 2Y$ a.s. for all $n \in \mathbb{N}$. Fix any $\varepsilon > 0$. Then, for all $n \in \mathbb{N}$,

$$\begin{aligned} 0 \le \mathbb{E}[|X_n - X|^p] &= \mathbb{E}\big[|X_n - X|^p\mathbf{1}_{\{|X_n - X|>\varepsilon\}}\big] + \mathbb{E}\big[|X_n - X|^p\mathbf{1}_{\{|X_n - X|\le\varepsilon\}}\big] \\ &\le \mathbb{E}\big[(2Y)^p\mathbf{1}_{\{|X_n - X|>\varepsilon\}}\big] + \mathbb{E}\big[\varepsilon^p\mathbf{1}_{\{|X_n - X|\le\varepsilon\}}\big] \\ &\le 2^p\mathbb{E}\big[Y^p\mathbf{1}_{\{|X_n - X|>\varepsilon\}}\big] + \varepsilon^p. \end{aligned}$$

As $Y \ge |X_n| \ge 0$ a.s., we know that $Y^p = |Y^p|$ a.s. Also, by assumption we have $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$. Thus, letting $n \to \infty$ and applying Lemma 1.2.c yields

$$0 \le \liminf_{n\to\infty} \mathbb{E}[|X_n - X|^p] = \limsup_{n\to\infty} \mathbb{E}[|X_n - X|^p] \le \varepsilon^p.$$

Letting $\varepsilon \to 0^+$ then implies that $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0$, as desired. □

1.2.4 Bounded convergence in probability implies convergence in L^p.

Corollary 1.2.e. Suppose that $|X_n| \le M$ a.s. for all $n \in \mathbb{N}$ and some $M > 0$. Then, $X_n \xrightarrow[n\to\infty]{\text{p}} X \implies X_n \xrightarrow[n\to\infty]{L^p} X$ for all $p > 0$.

Proof. Take $Y = M$ in Theorem 1.2.d. □

1.2.5 **Convergence in probability or L^r sufficiently fast implies a.s. convergence.**

(a) If $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) < \infty$ for all $\varepsilon > 0$, then $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X$.

(b) If $\sum_{n=1}^{\infty} \mathbb{E}[|X_n - X|^p] < \infty$ for some $p > 0$, then $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X$.

Proof.

(a) For all $\varepsilon > 0$, we have $\mathbb{P}(\bigcup_{m=n}^{\infty}\{|X_m - X| > \varepsilon\}) \le \sum_{m=n}^{\infty} \mathbb{P}(|X_m - X| > \varepsilon) \overset{\text{(assumption)}}{\underset{n\to\infty}{\longrightarrow}} 0$, implying that $\lim_{n\to\infty} \mathbb{P}(\bigcup_{m=n}^{\infty}\{|X_m - X| > \varepsilon\}) = 0$. Hence, by Proposition 1.1.a, we have $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X$.

(b) By [1.0.2]d we have $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) \le \sum_{n=1}^{\infty} \mathbb{E}[|X_n - X|^p]/\varepsilon^p \overset{\text{(assumption)}}{<} \infty$ for all $\varepsilon > 0$. Then, the result follows from (a).

□

[Note: The condition $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) < \infty$ in (a) requires $\mathbb{P}(|X_n - X| > \varepsilon)$ to converge to zero sufficiently fast such that the infinite sum is finite (similar for (b)). Note that it is possible to have $\{a_n\} \underset{n\to\infty}{\longrightarrow} 0$ while $\sum_{n=1}^{\infty} a_n = \infty$, e.g., $a_n = 1/n$ for all $n \in \mathbb{N}$.]

1.2.6 **Convergence in probability implies a.s. convergence for a subsequence.** If $X_n \xrightarrow[n\to\infty]{\text{p}} X$, then $X_{n_k} \xrightarrow[k\to\infty]{\text{a.s.}} X$ for some non-random positive integers $n_1 < n_2 < \cdots$.

Proof. By assumption we have $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$ for all $\varepsilon > 0$. Therefore, for all $k \in \mathbb{N}$, there exists sufficiently large $n_k \in \mathbb{N}$ such that $\mathbb{P}(|X_{n_k} - X| > 1/k) \le 2^{-k}$. By raising the values of some n_k's if needed, we may assume that $n_1 < n_2 < \cdots$. Then, we have

$$\sum_{k=1}^{\infty} \mathbb{P}(|X_{n_k} - X| > \varepsilon) = \underbrace{\sum_{k:1/k\ge\varepsilon} \mathbb{P}(|X_{n_k} - X| > \varepsilon)}_{<\infty \text{ (as a finite sum)}} + \underbrace{\sum_{k:1/k<\varepsilon} \mathbb{P}(|X_{n_k} - X| > \varepsilon)}_{\le \sum_{k:1/k<\varepsilon} \mathbb{P}\left(|X_{n_k} - X| > 1/k\right) \le \sum_{k:1/k<\varepsilon} 2^{-k} < \infty} < \infty.$$

So, the result follows from [1.2.5]. □

1.2.7 **Convergence in distribution implies a.s. convergence in another probability space.**

Theorem 1.2.f (Skorokhod's representation theorem)**.** Suppose that $X_n \xrightarrow[n\to\infty]{\text{d}} X$. Then there exist random variables $Y, Y_1, Y_2, \ldots$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $Y \overset{\text{d}}{=} X$, $Y_n \overset{\text{d}}{=} X_n$ for all $n \in \mathbb{N}$, and $Y_n \xrightarrow[n\to\infty]{\text{a.s.}} Y$.

Proof. Omitted. □

[Note: The X_n's and X may not be (and are often not) defined on that probability space $(\Omega, \mathcal{F}, \mathbb{P})$.]

1.3 Uniform Integrability

1.3.1 **Definition and characterization of uniform integrability.** Recall from STAT7610 that a collection $\{X_i\}_{i\in I} \subseteq L^1$ is said to be **uniformly integrable** (u.i.) if $\lim_{a\to\infty} \sup_{i\in I} \mathbb{E}\left[|X_i| \mathbf{1}_{\{|X_i|>a\}}\right] = 0$.

Also, we have the following characterization of uniform integrability:

Theorem 1.3.a. A collection $\{X_i\}_{i\in I} \subseteq L^1$ is u.i. iff

(a) *(uniform bounded first absolute moments)* $\sup_{i\in I} \mathbb{E}[|X_i|] < \infty$, and

(b) *(uniform absolute continuity)* for all $\varepsilon > 0$, there exists $\delta > 0$ such that $\sup_{i\in I} \mathbb{E}[|X_i|\mathbf{1}_A] < \varepsilon$ for every $A \in \mathcal{F}$ with $\mathbb{P}(A) < \delta$.

Proof. "⇒": Assume $\{X_i\}$ is u.i. Note that for all $a > 0$ and all $A \in \mathcal{F}$, we have

$$\begin{aligned}
\sup_{i\in I} \mathbb{E}[|X_i|] &\le \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|\le a\}}\big] + \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] \\
&\le a + \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big], \\
\sup_{i\in I} \mathbb{E}[|X_i|\mathbf{1}_A] &\le \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|\le a\}}\mathbf{1}_A\big] + \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\mathbf{1}_A\big] \\
&\le a\mathbb{P}(A) + \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big].
\end{aligned}$$

Fix any $\varepsilon > 0$. By uniform integrability, there exists sufficiently large $M > 0$ such that $\sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>M\}}\big] < \varepsilon/2$. Also, we choose $\delta = \varepsilon/2M$. Therefore, by above we have

$$\sup_{i\in I} \mathbb{E}[|X_i|] \le M + \varepsilon < \infty,$$

establishing (a), and

$$\sup_{i\in I} \mathbb{E}[|X_i|\mathbf{1}_A] < M(\varepsilon/2M) + \varepsilon/2 = \varepsilon$$

whenever $\mathbb{P}(A) < \delta$, establishing (b).

"⇐": Assume (a) and (b) hold. Let $M := \sup_{i\in I} \mathbb{E}[|X_i|] < \infty$. Fix any $\varepsilon > 0$, and consider the $\delta > 0$ from (b). Then, by Markov's inequality, for all $i \in I$ and all $a > M/\delta$, we have

$$\mathbb{P}(|X_i| > a) \le \frac{\mathbb{E}[|X_i|]}{a} \le \frac{\sup_{i\in I} \mathbb{E}[|X_i|]}{a} = \frac{M}{a} < \delta,$$

which implies by (b) that $\sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] < \varepsilon$. Hence, by the definition of limit, we have $\lim_{a\to\infty} \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] = 0$, as desired. □

1.3.2 Properties of uniform integrability.

(a) The singleton $\{X\}$ with $X \in L^1$ is u.i.

(b) If $|X_i| \le Y_i$ for all $i \in I$ and $\{Y_i\}_{i\in I}$ is u.i., then $\{X_i\}_{i\in I}$ is u.i.

(c) If $\{X_i\}_{i\in I}$ and $\{Y_i\}_{i\in I}$ are both u.i., then $\{X_i + Y_i\}_{i\in I}$ is u.i.

(d) $\{X_i\}_{i\in I}$ is u.i. iff $\{X_i^+\}_{i\in I}$ and $\{X_i^-\}_{i\in I}$ are both u.i.

(e) If $\{X_n\}_{n\in\mathbb{N}}$ is u.i., then every subsequence of $\{X_n\}_{n\in\mathbb{N}}$ is u.i.

(f) If $\mathbb{E}[\sup_{i\in I} |X_i|] < \infty$, then $\{X_i\}_{i\in I}$ is u.i.

(g) Let $\psi \ge 0$ be a function satisfying that $\lim_{x\to\infty} \psi(x)/x = \infty$. If $\sup_{i\in I} \mathbb{E}[\psi(|X_i|)] < \infty$, then $\{X_i\}_{i\in I}$ is u.i.

Proof.

(a) Follow the argument in the first part of the proof of Lemma 1.2.c.

(b) Note that for all $i \in I$, $Y_i \ge |X_i| \ge 0$ and so $Y_i = |Y_i|$. It then follows from noting that for all $a > 0$,

$$0 \le \sup_{i\in I} \mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] \overset{\text{(assumption)}}{\le} \sup_{i\in I} \mathbb{E}\big[|Y_i|\mathbf{1}_{\{|Y_i|>a\}}\big] \overset{\text{(assumption)}}{\underset{a\to\infty}{\longrightarrow}} 0.$$

(c) We use the characterization of u.i. from Theorem 1.3.a. First, we have

$$\sup_{i\in I}\mathbb{E}[|X_i+Y_i|] \le \underbrace{\sup_{i\in I}\mathbb{E}[|X_i|]}_{\substack{\text{(assumption)}\\ <\ \infty}} + \underbrace{\sup_{i\in I}\mathbb{E}[|Y_i|]}_{\substack{\text{(assumption)}\\ <\ \infty}} < \infty.$$

Second, fix any $\varepsilon > 0$. By assumption, there exists $\delta > 0$ such that $\sup_{i\in I}\mathbb{E}[|X_i|\mathbf{1}_A] < \varepsilon/2$ and $\sup_{i\in I}\mathbb{E}[|Y_i|\mathbf{1}_A] < \varepsilon/2$ for all $A \in \mathcal{F}$ with $\mathbb{P}(A) < \delta$. Thus, for all $A \in \mathcal{F}$ with $\mathbb{P}(A) < \delta$, we have

$$\sup_{i\in I}\mathbb{E}[|X_i+Y_i|\mathbf{1}_A] \le \sup_{i\in I}\mathbb{E}[|X_i|\mathbf{1}_A] + \sup_{i\in I}\mathbb{E}[|Y_i|\mathbf{1}_A] < \varepsilon/2 + \varepsilon/2 = \varepsilon,$$

as desired.

(d) Note that $|X_i| = \underbrace{X_i^+}_{\ge 0} + \underbrace{X_i^-}_{\ge 0}$ for all $i \in I$.

"⇒": For all $a > 0$, we have

$$0 \le \sup_{i\in I}\mathbb{E}\Big[|X_i^+|\mathbf{1}_{\{|X_i^+|>a\}}\Big] \le \sup_{i\in I}\mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] \overset{\text{(assumption)}}{\underset{a\to\infty}{\longrightarrow}} 0,$$
$$0 \le \sup_{i\in I}\mathbb{E}\Big[|X_i^-|\mathbf{1}_{\{|X_i^-|>a\}}\Big] \le \sup_{i\in I}\mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] \overset{\text{(assumption)}}{\underset{a\to\infty}{\longrightarrow}} 0.$$

"⇐": It follows from (c).

(e) Fix any subsequence $\{X_{n_k}\}_{k\in\mathbb{N}}$ of $\{X_n\}_{n\in\mathbb{N}}$. Then, for all $a > 0$, we have

$$0 \le \sup_{k\in\mathbb{N}}\mathbb{E}\Big[|X_{n_k}|\mathbf{1}_{\{|X_{n_k}|>a\}}\Big] \le \sup_{n\in\mathbb{N}}\mathbb{E}\big[|X_n|\mathbf{1}_{\{|X_n|>a\}}\big] \overset{\text{(assumption)}}{\underset{a\to\infty}{\longrightarrow}} 0,$$

so the result follows.

(f) Take $Y_i = \sup_{k\in I}|X_k|$ for all $i \in I$ in (b) (note that $|X_i| \le \sup_{k\in I}|X_k|$ for all $i \in I$). By assumption, we know that the singleton $\{Y_i\}_{i\in I} = \{\sup_{k\in I}|X_k|\}$ is u.i. by (a). So the result follows from (b).

(g) Fix any $\varepsilon > 0$. Let $C := \sup_{i\in I}\mathbb{E}[\psi(X_i)] < \infty$ and $M := 1/\varepsilon > 0$. Then by assumption, there exists $x_0 > 0$ such that $\psi(x)/x > M$ for all $x \ge x_0$. Hence, for all $a \ge x_0$, we have

$$0 \le \sup_{i\in I}\mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] \le \sup_{i\in I}\mathbb{E}\left[\frac{\psi(|X_i|)}{M}\mathbf{1}_{\{|X_i|>a\}}\right] \le \underbrace{\frac{1}{M}}_{\varepsilon}\underbrace{\sup_{i\in I}\mathbb{E}[\psi(X_i)]}_{C} = C\varepsilon.$$

Letting $\varepsilon \to 0^+$ and then $a \to \infty$ then yields $\lim_{a\to\infty}\sup_{i\in I}\mathbb{E}\big[|X_i|\mathbf{1}_{\{|X_i|>a\}}\big] = 0$, as desired.

□

[Note: For (h), the condition is only "slightly stronger" than the *uniform bounded first absolute moment* in Theorem 1.3.a; for instance, we can take $\psi(x) = x^{1+\delta}$ on $[0,\infty)$ with $\delta > 0$. Then, this property suggests that having *uniform bounded* $1+\delta$*th absolute moment* ($\sup_{i\in I}\mathbb{E}\big[|X_i|^{1+\delta}\big] < \infty$) is enough for establishing the uniform integrability (even if δ is very small!).]

1.3.3 The notion of uniform integrability allows us to obtain more relationships between different kinds of convergence, as illustrated in the following. But before that, let us first introduce an inequality that will be helpful in some later proofs.

Lemma 1.3.b (C_r-inequality)**.** For all $x, y \in \mathbb{R}$, we have

$$|x+y|^r \le C_r(|x|^r + |y|^r)$$

where $r > 0$ and $C_r = \begin{cases} 1 & \text{if } 0 < r < 1, \\ 2^{r-1} & \text{if } r \ge 1. \end{cases}$

Proof. For the case $r \geq 1$, it follows from the inequality $|(x+y)/2|^r \leq (|x|^r + |y|^r)/2$ for all $x, y \in \mathbb{R}$, which holds as $t \mapsto |t|^r$ is convex. For the case $0 < r < 1$, it follows from the inequality $\lambda^r + (1-\lambda)^r \geq \lambda + (1-\lambda) = 1$ with $\lambda = |x|/(|x|+|y|) \in [0,1]$, for all $x, y \in \mathbb{R}$. □

1.3.4 **Convergence in L^p implies convergence of pth absolute moments.** If $X_n \xrightarrow[n\to\infty]{L^p} X$, then $\lim_{n\to\infty} \mathbb{E}[|X_n|^p] = \mathbb{E}[|X|^p]$.

Proof. Consider first the case $0 < p < 1$. By Lemma 1.3.b, for all $x, y \in \mathbb{R}$, we have $|x+y|^p \leq |x|^p + |y|^p$, which implies that

$$\begin{cases} |x|^p \leq |x-y|^p + |y|^p \implies |x|^p - |y|^p \leq |x-y|^p, \\ |y|^p \leq |y-x|^p + |x|^p = |x-y|^p + |x|^p \implies |x|^p - |y|^p \geq -|x-y|^p. \end{cases}$$

Thus, we have $||x|^p - |y|^p| \leq |x-y|^p$ for all $x, y \in \mathbb{R}$. Hence,

$$0 \leq |\mathbb{E}[|X_n|^p] - \mathbb{E}[|X|^p]| = |\mathbb{E}[|X_n|^p - |X|^p]| \leq \mathbb{E}\big[\big||X_n|^p - |X|^p\big|\big] \leq \mathbb{E}[|X_n - X|^p] \overset{\text{(assumption)}}{\underset{n\to\infty}{\longrightarrow}} 0,$$

which implies that $\lim_{n\to\infty} \mathbb{E}[|X_n|^p] = \mathbb{E}[|X|^p]$.

Now consider the case $p > 1$. By Minkowski's inequality, we have

$$\begin{cases} \|X_n\|_p \leq \|X_n - X\|_p + \|X\|_p \implies \|X_n\|_p - \|X\|_p \leq \|X_n - X\|_p, \\ \|X\|_p \leq \|X - X_n\|_p + \|X_n\|_p = \|X_n - X\|_p + \|X_n\|_p \implies \|X_n\|_p - \|X\|_p \geq -\|X_n - X\|_p. \end{cases}$$

Therefore, we have

$$\big|\|X_n\|_p - \|X\|_p\big| \leq \|X_n - X\|_p,$$

meaning that

$$0 \leq \Big|\mathbb{E}[|X_n|^p]^{1/p} - \mathbb{E}[|X|^p]^{1/p}\Big| \leq \mathbb{E}[|X_n - X|^p]^{1/p} \overset{\text{(assumption)}}{\underset{n\to\infty}{\longrightarrow}} 0.$$

This implies that $\lim_{n\to\infty} \mathbb{E}[|X_n|^p] = \mathbb{E}[|X|^p]$. □

1.3.5 **Convergence in probability with uniform integrability implies convergence in L^p.**

Theorem 1.3.c (Vitali's theorem)**.** Suppose that $X_n \xrightarrow[n\to\infty]{\text{p}} X$ and $X_n \in L^p$ for all $n \in \mathbb{N}$, with $p > 0$. Then the following are equivalent.

(a) $\{|X_n|^p\}_{n\in\mathbb{N}}$ is u.i.

(b) $X_n \xrightarrow[n\to\infty]{L^p} X$.

(c) $\lim_{n\to\infty} \mathbb{E}[|X_n|^p] = \mathbb{E}[|X|^p]$ with $\mathbb{E}[|X|^p] < \infty$.

Proof.

- "(a) $\implies$ (b)": Assume (a) holds.
  Showing that $\mathbb{E}[|X|^p] < \infty$. Since $X_n \xrightarrow[n\to\infty]{\text{p}} X$, by [1.2.6] there exists a subsequence $\{X_{n_k}\}_{k\in\mathbb{N}}$ of $\{X_n\}_{n\in\mathbb{N}}$ such that $X_{n_k} \xrightarrow[n\to\infty]{\text{a.s.}} X$, thus $|X_{n_k}|^p \xrightarrow[n\to\infty]{\text{a.s.}} |X|^p$ by continuous mapping theorem. Then, we have

$$\begin{aligned} \mathbb{E}[|X|^p] = \mathbb{E}\left[\lim_{k\to\infty} |X_{n_k}|^p\right] = \mathbb{E}\left[\liminf_{k\to\infty} |X_{n_k}|^p\right] \\ \overset{\text{(Fatou)}}{\leq} \liminf_{k\to\infty} \underbrace{\mathbb{E}[|X_{n_k}|^p]}_{\leq \sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|^p]} \leq \sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|^p] \overset{\text{((a), Theorem 1.3.a)}}{<} \infty. \end{aligned}$$

Showing that $\{|X_n - X|^p\}$ is u.i. By Lemma 1.3.b, we have $|X_n - X|^p \le C_p(|X_n|^p + |X|^p)$. By (a), $\{|X_n|^p\}$ is u.i. Also, by [1.3.2] we know that $\{|X|^p\}$ is u.i. Hence, applying [1.3.2], $\{|X_n|^p + |X|^p\}$ is u.i. As C_p does not depend on n, it is straightforward to show that $\{C_p(|X_n|^p + |X|^p)\}$ is u.i. also. Hence, by [1.3.2] again, $\{|X_n - X|^p\}$ is u.i.

Showing that $X_n \xrightarrow[n\to\infty]{L^p} X$. Fix any $\varepsilon > 0$. Since $\{|X_n - X|^p\}$ is u.i., by Theorem 1.3.a, there exists $\delta > 0$ such that $\sup_{m\in\mathbb{N}} \mathbb{E}[|X_m - X|^p \mathbf{1}_A] < \varepsilon$ for all $A \in \mathcal{F}$ with $\mathbb{P}(A) < \delta$. By assumption, we have $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$, and thus there exists sufficiently large $N \in \mathbb{N}$ such that $\mathbb{P}(|X_n - X| > \varepsilon) < \delta$ for all $n \ge N$.
Therefore, for all $n \ge N$, we have

$$\begin{aligned} 0 \le \mathbb{E}[|X_n - X|^p] &= \mathbb{E}\big[|X_n - X|^p \mathbf{1}_{\{|X_n - X| \le \varepsilon\}}\big] + \mathbb{E}\big[|X_n - X|^p \mathbf{1}_{\{|X_n - X| > \varepsilon\}}\big] \\ &\le \varepsilon^p + \underbrace{\mathbb{E}\big[|X_n - X|^p \mathbf{1}_{\{|X_n - X| > \varepsilon\}}\big]}_{\le \sup_{m\in\mathbb{N}} \mathbb{E}\left[|X_m - X|^p \mathbf{1}_{\{|X_n - X| > \varepsilon\}}\right] < \varepsilon} = \varepsilon^p + \varepsilon. \end{aligned}$$

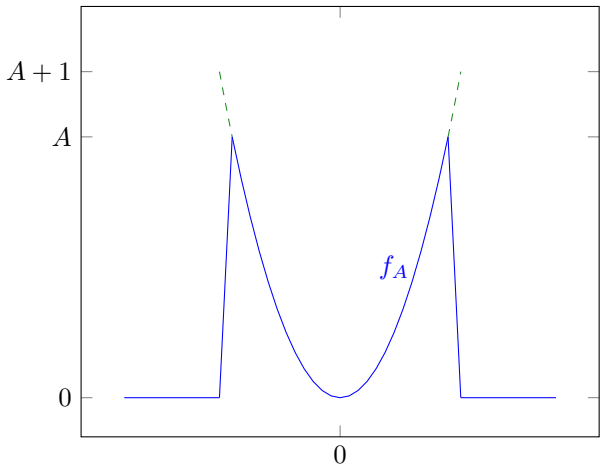
Letting $\varepsilon \to 0^+$ and then $n \to \infty$ gives $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0$, as desired.

- "(b) $\implies$ (c)": It follows from [1.3.4].
- "(c) $\implies$ (a)": Assume (c) holds.

  Constructing a specific nonnegative and continuous function. Fix any $A > 0$ and construct a nonnegative and continuous function $f_A : \mathbb{R} \to \mathbb{R}$ satisfying that

$$f_A(x) \begin{cases} = |x|^p & \text{if } |x|^p \le A, \\ \le |x|^p & \text{if } A < |x|^p \le A + 1, \\ = 0 & \text{if } |x|^p > A + 1. \end{cases}$$

  An example of such a f_A is illustrated below, which is constructed by performing linear intercorrelation for the part where $A < |x|^r \le A + 1$.



  Applying Portmanteau theorem. With $X_n \xrightarrow[n\to\infty]{\text{p}} X$, by Theorem 1.2.a we have $X_n \xrightarrow[n\to\infty]{\text{d}} X$. Since f_A is bounded and continuous by construction, by Portmanteau theorem we get $\lim_{n\to\infty} \mathbb{E}[f_A(X_n)] = \mathbb{E}[f_A(X)]$. Furthermore, note that we have $0 \le |x|^p \mathbf{1}_{\{|x|^p \le A\}} \le f_A(x) \le |x|^p \mathbf{1}_{\{|x|^p \le A+1\}}$ for all $x \in \mathbb{R}$ by construction. Thus,

$$\liminf_{n\to\infty} \mathbb{E}\big[|X_n|^p \mathbf{1}_{\{|X_n|^p \le A+1\}}\big] \ge \liminf_{n\to\infty} \mathbb{E}[f_A(X_n)] = \mathbb{E}[f_A(X)] \ge \mathbb{E}\big[|X|^p \mathbf{1}_{\{|X|^r \le A\}}\big],$$

  implying that

$$\limsup_{n\to\infty} (-\mathbb{E}\big[|X_n|^p \mathbf{1}_{\{|X_n|^p \le A+1\}}\big]) = -\liminf_{n\to\infty} \mathbb{E}\big[|X_n|^p \mathbf{1}_{\{|X_n|^p \le A+1\}}\big] \le -\mathbb{E}\big[|X|^p \mathbf{1}_{\{|X|^p \le A\}}\big].$$

By (c), we have $\lim_{n\to\infty}\mathbb{E}[|X_n|^p]=\mathbb{E}[|X|^p]$ with $\mathbb{E}[|X|^p]<\infty$. Therefore, we get

$$\lim_{n\to\infty}\mathbb{E}[|X_n|^p]-\limsup_{n\to\infty}\mathbb{E}\big[|X_n|^r\mathbf{1}_{\{|X_n|^r\le A+1\}}\big]\le\mathbb{E}[|X|^p]-\mathbb{E}\big[|X|^p\mathbf{1}_{\{|X|^p\le A\}}\big],$$
$$\implies 0\le\limsup_{n\to\infty}\mathbb{E}\big[|X_n|^r\mathbf{1}_{\{|X_n|^r>A+1\}}\big]\le\mathbb{E}\big[|X|^p\mathbf{1}_{\{|X|^p>A\}}\big]$$

Fix any $\varepsilon>0$. Then, there exists sufficiently large $A_0>0$ such that $\mathbb{E}\big[|X|^p\mathbf{1}_{\{|X|^p>A\}}\big]<\varepsilon$ for all $A>A_0$ by Lemma 1.2.c. This implies that $\limsup_{n\to\infty}\mathbb{E}\big[|X_n|^r\mathbf{1}_{\{|X_n|^p>A+1\}}\big]<\varepsilon$, and hence

$$\sup_{m\ge n_0}\mathbb{E}\big[|X_m|^p\mathbf{1}_{\{|X_m|^p>A+1\}}\big]<\varepsilon$$

for some $n_0\in\mathbb{N}$; otherwise we would have

$$\limsup_{n\to\infty}\mathbb{E}\big[|X_n|^p\mathbf{1}_{\{|X_n|^p>A+1\}}\big]=\lim_{n\to\infty}\sup_{m\ge n}\mathbb{E}\big[|X_n|^p\mathbf{1}_{\{|X_n|^p>A+1\}}\big]\ge\varepsilon,$$

contradiction.
Since $X_n\in L^p$ for all $n\in\mathbb{N}$ by assumption, by Lemma 1.2.c we know that there exists $A_1>0$ such that $\mathbb{E}\big[|X_n|^p\mathbf{1}_{\{|X_n|^p>A+1\}}\big]<\sup_{m\ge n_0}\mathbb{E}\big[|X_m|^p\mathbf{1}_{\{|X_m|^p>A+1\}}\big]$ for all $A>A_1$ and all $n=1,\dots,n_0-1$. Hence, for all $A>\max\{A_0,A_1\}$, we have

$$\sup_{m\in\mathbb{N}}\mathbb{E}\big[|X_m|^p\mathbf{1}_{\{|X_m|^p>A+1\}}\big]=\sup_{m\ge n_0}\mathbb{E}\big[|X_m|^p\mathbf{1}_{\{|X_m|^p>A+1\}}\big]<\varepsilon.$$

This implies that $\{|X_n|^p\}_{n\in\mathbb{N}}$ is u.i.

□

1.3.6 **Convergence in distribution with uniform integrability implies convergence of pth absolute moments.** If $X_n\xrightarrow[n\to\infty]{\text{d}}X$ and $\{|X_n|^p\}$ is u.i., then $\lim_{n\to\infty}\mathbb{E}[|X_n|^p]=\mathbb{E}[|X|^p]$.

Proof. With $X_n\xrightarrow[n\to\infty]{\text{d}}X$, by Skorokhod's representation theorem (Theorem 1.2.f), there exist random variables $Y,Y_1,Y_2,\dots$ such that $Y\overset{\text{d}}{=}X$, $Y_n\overset{\text{d}}{=}X_n$, and $Y_n\xrightarrow[n\to\infty]{\text{a.s.}}Y$. Since $\{|X_n|^p\}$ is u.i., $\{|Y_n|^p\}$ is also u.i. by the equality in distribution. Since $Y_n\xrightarrow[n\to\infty]{\text{a.s.}}Y\implies Y_n\xrightarrow[n\to\infty]{\text{p}}Y$ by Theorem 1.2.a, applying Theorem 1.3.c gives $\lim_{n\to\infty}\mathbb{E}[|Y_n|^p]=\mathbb{E}[|Y|^p]$. By the equality in distribution again, we have $\lim_{n\to\infty}\mathbb{E}[|X_n|^p]=\mathbb{E}[|X|^p]$, as desired. □

1.4 Closed Operations of Convergence

1.4.1 **Sums/differences.**

(a) If $X_n\xrightarrow[n\to\infty]{\text{p}}X$ and $Y_n\xrightarrow[n\to\infty]{\text{p}}Y$, then $X_n\pm Y_n\xrightarrow[n\to\infty]{\text{p}}X\pm Y$.

(b) If $X_n\xrightarrow[n\to\infty]{\text{a.s.}}X$ and $Y_n\xrightarrow[n\to\infty]{\text{a.s.}}Y$, then $X_n\pm Y_n\xrightarrow[n\to\infty]{\text{a.s.}}X\pm Y$.

(c) If $X_n\xrightarrow[n\to\infty]{L^p}X$ and $Y_n\xrightarrow[n\to\infty]{L^p}Y$, then $X_n\pm Y_n\xrightarrow[n\to\infty]{L^p}X\pm Y$.

Proof. By definition, we see that $X_n\xrightarrow[n\to\infty]{*}X\implies -X_n\xrightarrow[n\to\infty]{*}-X$, where $*$ can be "p", "a.s.", or "L^p". Hence, it suffices to prove the "+" part only.

(a) For all $\varepsilon>0$ and $n\in\mathbb{N}$, we have

$$0\le\mathbb{P}(|X_n+Y_n-(X+Y)|>\varepsilon)\overset{\text{(triangle)}}{\le}\mathbb{P}(|X_n-X|>\varepsilon/2\cup|Y_n-Y|>\varepsilon/2)$$
$$\le\mathbb{P}(|X_n-X|>\varepsilon/2)+\mathbb{P}(|Y_n-Y|>\varepsilon/2)\xrightarrow[n\to\infty]{\text{(assumption)}}0.$$

So, letting $n\to\infty$ yields the desired result.

(b) Assume $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X$ and $Y_n \xrightarrow[n\to\infty]{\text{a.s.}} Y$. Then, $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for all $\omega \in \Omega \setminus N_1$ and $\lim_{n\to\infty} Y_n(\omega) = Y(\omega)$ for all $\omega \in \Omega \setminus N_2$, with $\mathbb{P}(N_1) = \mathbb{P}(N_2) = 0$. This implies that for all $\omega \in \Omega \setminus (N_1 \cup N_2)$, we have $\lim_{n\to\infty}(X_n(\omega) + Y_n(\omega)) = X(\omega) + Y(\omega)$. Since $\mathbb{P}(N_1 \cup N_2) = 0$, the result follows.

(c) For all $n \in \mathbb{N}$, we have

$$0 \le \mathbb{E}[|X_n + Y_n - (X+Y)|^p] = \mathbb{E}[|(X_n - X) + (Y_n - Y)|^p]$$
$$\overset{\text{(Lemma 1.3.b)}}{\le} C_p(\mathbb{E}[|X_n - X|^p] + \mathbb{E}[|Y_n - Y|^p]) \overset{\text{(assumption)}}{\xrightarrow[n\to\infty]{}} 0.$$

The result then follows by letting $n \to \infty$.

□

[⚠ Warning: We do NOT have an analogous result for convergence in distribution. For example, take $X_n \sim \mathrm{N}(0,1)$ and $Y_n = -X_n \sim \mathrm{N}(0,1)$ for all $n \in \mathbb{N}$, with $X, Y \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$. Then, while we have $X_n \xrightarrow[n\to\infty]{\text{d}} X$ and $Y_n \xrightarrow[n\to\infty]{\text{d}} Y$, we do not have $X_n + Y_n \xrightarrow[n\to\infty]{\text{d}} X + Y \sim \mathrm{N}(0,2)$, as $X_n + Y_n = 0$ for all $n \in \mathbb{N}$.]

1.4.2 Products.

(a) If $X_n \xrightarrow[n\to\infty]{\text{p}} X$ and $Y_n \xrightarrow[n\to\infty]{\text{p}} Y$, then $X_nY_n \xrightarrow[n\to\infty]{\text{p}} XY$.

(b) If $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X$ and $Y_n \xrightarrow[n\to\infty]{\text{a.s.}} Y$, then $X_nY_n \xrightarrow[n\to\infty]{\text{a.s.}} XY$.

Proof. Exercise. □

1.4.3 Continuous mappings.

Theorem 1.4.a (Continuous mapping theorem)**.** Let X be a random variable and $\{X_n\}$ be a sequence of random variables. Suppose that $g : \mathbb{R} \to \mathbb{R}$ is a function satisfying that $\mathbb{P}(\{\omega \in \Omega : g \text{ is discontinuous at } X(\omega)\}) = 0$ *(continuous a.s. with respect to* $\mathbb{P}_X$*)*. Then:

(a) $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X \implies g(X_n) \xrightarrow[n\to\infty]{\text{a.s.}} g(X)$.

(b) $X_n \xrightarrow[n\to\infty]{\text{p}} X \implies g(X_n) \xrightarrow[n\to\infty]{\text{p}} g(X)$.

(c) $X_n \xrightarrow[n\to\infty]{\text{d}} X \implies g(X_n) \xrightarrow[n\to\infty]{\text{d}} g(X)$.

Proof. Consider first the special case where g is continuous.

(a) By continuity of g, we have

$$A := \left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} \subseteq \left\{\omega \in \Omega : \lim_{n\to\infty} g(X_n(\omega)) = g(X(\omega))\right\} =: B.$$

Thus, we get $1 \overset{\text{(assumption)}}{=} \mathbb{P}(A) \le \mathbb{P}(B) \le 1$, implying that $\mathbb{P}(B) = 1$ as desired.

(b) Fix any $\varepsilon, \varepsilon_0 > 0$, and choose sufficiently large $M > 0$ such that $\mathbb{P}(|X| \ge M) < \varepsilon_0$.
On the closed and bounded (hence compact) interval $[-(M+\varepsilon), M+\varepsilon]$, g is uniformly continuous by Heine-Cantor theorem. Hence, there exists $0 < \delta \le \varepsilon$ such that $|g(x) - g(y)| < \varepsilon$ whenever $|x - y| \le \delta$ and $|x| \le M$.
Claim: If $|g(X_n) - g(X)| > \varepsilon$ and $|X_n - X| \le \delta$, then $|X| \ge M$.
Proof. Assume to the contrary that $|g(X_n) - g(X)| > \varepsilon$, $|X_n - X| \le \delta$, and $|X| < M$. With $|X_n - X| \le \delta$ and $|X| < M$, we would then get $|g(X_n) - g(X)| < \varepsilon$, contradiction. □

Therefore, for all $n \in \mathbb{N}$ we have

$$\begin{aligned}
0 &\leq \mathbb{P}(|g(X_n) - g(X)| > \varepsilon) \\
&= \mathbb{P}(|g(X_n) - g(X)| > \varepsilon \cap |X_n - X| > \delta) + \mathbb{P}(|g(X_n) - g(X)| > \varepsilon \cap |X_n - X| \leq \delta) \\
&\overset{\text{(claim)}}{\leq} \mathbb{P}(|X_n - X| > \delta) + \mathbb{P}(|X| \geq M) \overset{\text{(assumption)}}{\underset{n\to\infty}{\longrightarrow}} \mathbb{P}(|X| \geq M) < \varepsilon_0.
\end{aligned}$$

So, letting $n \to \infty$ and then $\varepsilon_0 \to 0^+$ yields $\lim_{n\to\infty} \mathbb{P}(|g(X_n) - g(X)| > \varepsilon) = 0$, as desired.

(c) With $X_n \xrightarrow[n\to\infty]{\text{d}} X$, by Skorokhod's representation theorem (Theorem 1.2.f), there exist random variables $Y, Y_1, Y_2, \ldots$ such that $Y \overset{\text{d}}{=} X$, $Y_n \overset{\text{d}}{=} X_n$ for all $n \in \mathbb{N}$, and $Y_n \xrightarrow[n\to\infty]{\text{a.s.}} Y$. By (a), we have $g(Y_n) \xrightarrow[n\to\infty]{\text{a.s.}} g(Y)$, which implies that $g(Y_n) \xrightarrow[n\to\infty]{\text{d}} g(Y)$ by Theorem 1.2.a. By the equality in distribution, we then have $g(X_n) \xrightarrow[n\to\infty]{\text{d}} g(X)$, as desired.

For the general case where g is continuous a.s. with respect to $\mathbb{P}_X$, we only show (a) and (c):

(a) Let $D := \{\omega \in \Omega : g \text{ is discontinuous at } X(\omega)\}$. Then, we have

$$\begin{aligned}
A &= \left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} \\
&\subseteq \left\{\omega \in \Omega : \lim_{n\to\infty} g(X_n(\omega)) = g(X(\omega))\right\} \cup \{\omega \in \Omega : g \text{ is discontinuous at } X(\omega)\} \\
&= B \cup D.
\end{aligned}$$

Hence, $1 \overset{\text{(assumption)}}{=} \mathbb{P}(A) \leq \mathbb{P}(B \cup D) \leq \mathbb{P}(B) + \mathbb{P}(D) \overset{\text{(assumption)}}{=} \mathbb{P}(B) \leq 1$, implying that $\mathbb{P}(B) = 1$, as desired.

(c) With (a) established for this general case, we can follow the same line of argument as part (c) previously.

□

1.4.4 **Slutsky's theorem.** While [1.4.1] and [1.4.2] suggest that algebraic operations do not preserve convergence in distribution in general, *Slutsky's theorem* suggests an important special case where the convergence in distribution is preserved, namely that one of the convergences in distribution is towards a *constant*:

Theorem 1.4.b (Slutsky's theorem)**.** Let $X_n \xrightarrow[n\to\infty]{\text{d}} X$ and $Y_n \xrightarrow[n\to\infty]{\text{d}} c$ where $c \in \mathbb{R}$. Then:

(a) $X_n + Y_n \xrightarrow[n\to\infty]{\text{d}} X + c$.

(b) $X_n Y_n \xrightarrow[n\to\infty]{\text{d}} cX$.

(c) $X_n / Y_n \xrightarrow[n\to\infty]{\text{d}} X/c$ if $c \neq 0$.

Proof. We only prove (a) here and leave the rest as exercise. Let F_W denote the CDF of a random variable W. Fix any $\varepsilon > 0$ and $x \in \mathbb{R}$ at which F_{X+c} is continuous, thus F_X is continuous at $x - c$.

Establishing an upper bound on $F_n(x)$. For all $n \in \mathbb{N}$,

$$\begin{aligned}
F_{X_n+Y_n}(x) &= \mathbb{P}(X_n + Y_n \leq x) \\
&= \mathbb{P}(X_n + Y_n \leq x \cap |Y_n - c| \leq \varepsilon) + \mathbb{P}(X_n + Y_n \leq x \cap |Y_n - c| > \varepsilon) \\
&\leq \underbrace{\mathbb{P}(X_n + c \leq x - (Y_n - c) \cap |Y_n - c| \leq \varepsilon)}_{\leq \mathbb{P}(X_n + c \leq x + \varepsilon \cap |Y_n - c| \leq \varepsilon) \leq \mathbb{P}(X_n + c \leq x + \varepsilon)} + \mathbb{P}(|Y_n - c| > \varepsilon) \\
&\leq \mathbb{P}(X_n + c \leq x + \varepsilon) + \mathbb{P}(|Y_n - c| > \varepsilon) \\
&= F_{X_n}(x - c + \varepsilon) + \mathbb{P}(|Y_n - c| > \varepsilon).
\end{aligned}$$

Establishing a lower bound on $F_n(x)$. For all $n \in \mathbb{N}$,

$$\begin{aligned}
F_{X_n+Y_n}(x) &= 1 - \mathbb{P}(X_n + Y_n > x) \\
&= 1 - \mathbb{P}(X_n + Y_n > x \cap |Y_n - c| \leq \varepsilon) - \mathbb{P}(X_n + Y_n > x \cap |Y_n - c| > \varepsilon) \\
&\geq 1 - \mathbb{P}(X_n + Y_n > x - (Y_n - c) \cap |Y_n - c| \leq \varepsilon) - \mathbb{P}(|Y_n - c| > \varepsilon) \\
&\geq 1 - \mathbb{P}(X_n + c > x - \varepsilon) - \mathbb{P}(|Y_n - c| > \varepsilon) \\
&= F_{X_n}(x - c - \varepsilon) - \mathbb{P}(|Y_n - c| > \varepsilon)
\end{aligned}$$

Combining upper and lower bounds and letting $n \to \infty$. For all $n \in \mathbb{N}$, we have

$$F_{X_n}(x - c - \varepsilon) - \mathbb{P}(|Y_n - c| > \varepsilon) \leq F_{X_n+Y_n}(x) \leq F_{X_n}(x - c + \varepsilon) + \mathbb{P}(|Y_n - c| > \varepsilon).$$

Since a CDF has at most countably many (jump) discontinuities, we can choose an arbitrarily small $\varepsilon > 0$ such that $x - c - \varepsilon$ and $x - c + \varepsilon$ are also continuity points of F_X. Letting $n \to \infty$ then gives

$$F_X(x - c - \varepsilon) \leq \liminf_{n\to\infty} F_{X_n+Y_n}(x) \leq \limsup_{n\to\infty} F_{X_n+Y_n}(x) \leq F_X(x - c + \varepsilon).$$

Since F_X is continuous at $x - c$, letting $\varepsilon \to 0^+$ yields

$$F_X(x - c) \leq \liminf_{n\to\infty} F_{X_n+Y_n}(x) \leq \limsup_{n\to\infty} F_{X_n+Y_n}(x) \leq F_X(x - c),$$

implying that $\liminf_{n\to\infty} F_{X_n+Y_n}(x) = \limsup_{n\to\infty} F_{X_n+Y_n}(x) = F_X(x-c) = F_{X+c}(x)$, as desired. $\square$

2 Limit Theorems

2.0.1 Utilizing the notions of modes of convergence learnt in Section 1, in Section 2 we will study various *limit theorems* which are about convergence behaviours of random variables that are more "meaningful" in statistics, e.g., sample means $\bar{X}_n := n^{-1}\sum_{i=1}^n X_i$ and sample sums $S_n := \sum_{i=1}^n X_i$.

2.0.2 **A simple limit theorem.** As a "warm-up", let us first establish the following relatively simple limit theorem:

Proposition 2.0.a. Suppose that X_i's are uncorrelated and $\sup_{n\in\mathbb{N}} \mathbb{E}\big[X_n^2\big] < \infty$ *(uniform bounded second moments)*. Denote $\mu_i := \mathbb{E}[X_i]$ for all $i \in \mathbb{N}$, $\bar{X}_n := n^{-1}\sum_{i=1}^n X_i$, and $\bar{\mu}_n := n^{-1}\sum_{i=1}^n \mu_i$ for all $n \in \mathbb{N}$. Then:

(a) $\bar{X}_n - \bar{\mu}_n \xrightarrow[n\to\infty]{L^2} 0$.

(b) $\bar{X}_n - \bar{\mu}_n \xrightarrow[n\to\infty]{\text{a.s.}} 0$.

Proof.

(a) Note that $\sigma_i^2 := \mathrm{Var}(X_i) = \mathbb{E}\big[X_i^2\big] - \mathbb{E}[X_i]^2 \le \mathbb{E}\big[X_i^2\big] \overset{\text{(assumption)}}{\le} M$ for all $i \in \mathbb{N}$. We have

$$\begin{aligned} 0 \le \mathbb{E}\big[(\bar{X}_n - \bar{\mu}_n)^2\big] &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n \mu_i\right)^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n (X_i - \mu_i)\right)^2\right] \\ &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}(X_i, X_j) \\ &\overset{\text{(uncorrelated)}}{=} \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\big[(X_i - \mu_i)^2\big] = \frac{1}{n^2}\sum_{i=1}^n \sigma_i^2 \le \frac{M}{n} \underset{n\to\infty}{\longrightarrow} 0. \end{aligned}$$

Therefore, $\lim_{n\to\infty} \mathbb{E}\big[(\bar{X}_n - \bar{\mu}_n)^2\big] = 0$, as desired.

(b) Let $S_n := \sum_{i=1}^n X_i$ for all $n \in \mathbb{N}$. Then, we can write $\bar{X}_n - \bar{\mu}_n = (S_n - \mathbb{E}[S_n])/n$.
Establishing the result for a subsequence. By the uncorrelatedness assumption, we have $\mathrm{Var}(S_n) = \sum_{i=1}^n \sigma_i^2 \le nM$. Thus, by Chebyshev's inequality, we have

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| > n\varepsilon) \le \frac{\mathrm{Var}(S_n)}{n^2\varepsilon^2} \le \frac{M}{n\varepsilon^2}$$

for all $n \in \mathbb{N}$ and $\varepsilon > 0$. Let $n_k = k^2$ for all $k \in \mathbb{N}$, which forms indices for a subsequence. Now consider:

$$\sum_{n=1}^\infty \mathbb{P}\left(\frac{|S_{n^2} - \mathbb{E}[S_{n^2}]|}{n^2} > \varepsilon\right) = \sum_{k=1}^\infty \mathbb{P}\left(\frac{|S_{n_k} - \mathbb{E}[S_{n_k}]|}{n_k} > \varepsilon\right) = \sum_{k=1}^\infty \mathbb{P}(|S_{n_k} - \mathbb{E}[S_{n_k}]| > n_k\varepsilon) \le \sum_{k=1}^\infty \frac{M}{k^2\varepsilon^2} < \infty.$$

By [1.2.5], we have

$$\frac{S_{n^2} - \mathbb{E}[S_{n^2}]}{n^2} \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

Filling in the gaps. To establish the desired result, we need to "fill in the gaps" for the convergence. WLOG, for all $i \in \mathbb{N}$ we can assume that $\mu_i = 0$; if not, we can replace X_i by

$X_i - \mathbb{E}[X_i]$. Let $D_n := \max_{n^2 \le k < (n+1)^2} |S_k - S_{n^2}|$ for all $n \in \mathbb{N}$. Then, we have

$$\begin{aligned}
\mathbb{E}\big[D_n^2\big] &= \mathbb{E}\left[\left(\max_{n^2 \le k < (n+1)^2} |S_k - S_{n^2}|\right)^2\right] \overset{\text{(max. achieved at the same index)}}{=} \mathbb{E}\left[\max_{n^2 \le k < (n+1)^2} |S_k - S_{n^2}|^2\right] \\
&\le \sum_{k=n^2}^{(n+1)^2-1} \mathbb{E}\big[(S_k - S_{n^2})^2\big] \sum_{k=n^2}^{(n+1)^2-1} \mathbb{E}\big[(X_{n^2+1} + \cdots + X_k)^2\big] \overset{\text{(uncorrelated)}}{=} \sum_{k=n^2}^{(n+1)^2-1} \sum_{j=n^2+1}^{k} \sigma_j^2 \\
&\le \sum_{k=n^2}^{(n+1)^2-1} \sum_{j=n^2}^{(n+1)^2-1} \underbrace{\sigma_j^2}_{\le M} \le (2n+1)^2 M \le 9n^2 M.
\end{aligned}$$

Thus, by Chebyshev's inequality, we have $\mathbb{P}\big(D_n > n^2\varepsilon\big) \le 9M/(n^2\varepsilon^2)$. Using a similar argument as before, we can conclude that $D_n/n^2 \xrightarrow[n\to\infty]{\text{a.s.}} 0$. Hence, for all $n^2 \le k < (n+1)^2$ with $n \in \mathbb{N}$, we have

$$0 \le \frac{|S_k|}{k} = \frac{|S_{n^2} + (S_k - S_{n^2})|}{k} \le \frac{|S_{n^2}| + |S_k - S_{n^2}|}{n^2} \le \frac{|S_{n^2}|}{n^2} + \frac{|D_n|}{n^2}.$$

Letting $k \to \infty$ then gives $|S_k|/k \xrightarrow[k\to\infty]{\text{a.s.}} 0$ (we have $n \to \infty$ as $k \to \infty$, viewing $n = n(k)$ as a function of k), and hence $\bar{X}_n \xrightarrow[n\to\infty]{\text{a.s.}} 0$, as desired.

□

[Note: The method used in (b) is known as the **subsequence method**, which is helpful for proving other results also. The idea is to first establish the result for a certain subsequence and then fill in the gaps.]

2.1 Weak Laws of Large Numbers

2.1.1 **Equivalent sequences.** To establish weak law of large numbers (under *pairwise* independence assumption), the notion of *equivalent sequences* will be used. Two sequences of random variables $\{X_n\}$ and $\{Y_n\}$ on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are said to be **equivalent** if $\sum_{n=1}^{\infty} \mathbb{P}(X_n \ne Y_n) < \infty$.

Intuitively, this condition suggests that the probabilities for the random variables X_n and Y_n to be different are "limited" across the sequences; the two sequences are "largely the same". Perhaps this concept can be better understood by studying the following result, which suggests that the convergence behaviours of sample sums and "generalized" sample means from the two sequences are the same, a.s.:

Proposition 2.1.a. Suppose that $\{X_n\}$ and $\{Y_n\}$ are equivalent. Then:

(a) *(Sample sums)* $\sum_{n=1}^{\infty}(X_n - Y_n)$ converges a.s.

(b) *("Generalized" sample means)* If $a_n \nearrow \infty$ (i.e., the sequence $\{a_n\}$ of real numbers is increasing and $a_n \to \infty$), then

$$\frac{1}{a_n}\sum_{i=1}^{n}(X_i - Y_i) \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

Proof.

(a) By assumption, we have $\sum_{n=1}^{\infty} \mathbb{P}(X_n \ne Y_n) < \infty$. Therefore, by Borel-Cantelli lemma ([1.0.2]a), we have $\mathbb{P}(X_n \ne Y_n \text{ io}) = 0$, which implies that $\mathbb{P}(X_n = Y_n \text{ abfm}) = 1$. It follows that there exists a null set N (i.e, $\mathbb{P}(N) = 0$) such that for all $\omega \in \Omega \setminus N$, there exists $n_0 = n_0(\omega)$ such that $X_n(\omega) = Y_n(\omega)$ for all $n \ge n_0$, meaning that $X_n(\omega) - Y_n(\omega)$ is nonzero only for finitely many n's, and therefore the partial sum $\sum_{i=1}^{n}(X_i(\omega) - Y_i(\omega))$ must converge as $n \to \infty$. It follows that $\sum_{n=1}^{\infty}(X_n - Y_n)$ converges a.s.

(b) For all $\omega \in \Omega \setminus N$, we have

$$\lim_{n\to\infty} \frac{1}{a_n} \sum_{i=1}^{n} (X_i(\omega) - Y_i(\omega)) = \underbrace{\left(\lim_{n\to\infty} \frac{1}{a_n} \right)}_{0 \text{ as } a_n \nearrow \infty} \cdot \underbrace{\sum_{i=1}^{\infty} (X_i(\omega) - Y_i(\omega))}_{<\infty} = 0,$$

so the result follows.

□

Due to this result, a helpful technique for establishing limit theorems is to replace the sequence $\{X_n\}$ in consideration by another equivalent sequence $\{Y_n\}$, often by the method of *truncation* (which will be illustrated in proofs later), and then we work with $\{Y_n\}$ and show convergence results about sample sums and "generalized" sample means from $\{Y_n\}$, which would also be applicable for $\{X_n\}$ by this result.

2.1.2 Weak law of large numbers (WLLN) under pairwise independence assumption.

Theorem 2.1.b (WLLN under pairwise independence)**.** Let X_i's be *pairwise* independent and identically distributed random variables with finite mean $\mu = \mathbb{E}[X_1]$. Let $\bar{X}_n := n^{-1} \sum_{i=1}^{n} X_i$. Then, $\bar{X}_n \xrightarrow[n\to\infty]{\text{p}} \mu$.

Proof. Note that

$$\mathbb{E}[|X_1|] < \infty \overset{\text{(can show)}}{\iff} \sum_{n=1}^{\infty} \mathbb{P}(|X_1| > n) \overset{\text{(identically distributed)}}{=} \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > n) < \infty.$$

Introducing truncated random variables that are equivalent to the original ones. Now define the *truncated* random variable Y_n by $Y_n = X_n \mathbf{1}_{\{|X_n| \le n\}}$ for all $n \in \mathbb{N}$. Then, $\{X_n\}$ and $\{Y_n\}$ are equivalent since

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n \neq Y_n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > n) < \infty.$$

Let $\bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$. By Proposition 2.1.a, we know that $\bar{X}_n - \bar{Y}_n \xrightarrow[n\to\infty]{\text{a.s.}} 0$, which implies by Theorem 1.2.a that $\bar{X}_n - \bar{Y}_n \xrightarrow[n\to\infty]{\text{p}} 0$. Therefore, to establish that $\bar{X}_n \xrightarrow[n\to\infty]{\text{p}} \mu$, it suffices to show that $\bar{Y}_n \xrightarrow[n\to\infty]{\text{p}} \mu$ in view of [1.4.1].

Showing that $\mathbb{E}[\bar{Y}_n] \xrightarrow[n\to\infty]{} \mu$. We have

$$\begin{aligned}
0 \le \left| \mathbb{E}[\bar{Y}_n] - \mu \right| &= \frac{1}{n} \left| \sum_{i=1}^{n} (\mathbb{E}[Y_i] - \mathbb{E}[X_i]) \right| \\
&= \frac{1}{n} \left| \sum_{i=1}^{n} (\mathbb{E}[X_i \mathbf{1}_{\{|X_i| \le i\}}] - \mathbb{E}[X_i]) \right| \\
&= \frac{1}{n} \left| -\sum_{i=1}^{n} \mathbb{E}[X_i \mathbf{1}_{\{|X_i| > i\}}] \right| \\
&\le \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[|X_i| \mathbf{1}_{\{|X_i| > i\}}] \overset{(a_n \xrightarrow[n\to\infty]{} 0 \implies n^{-1} \sum_{i=1}^{n} a_i \xrightarrow[n\to\infty]{} 0)}{\xrightarrow[n\to\infty]{}} 0.
\end{aligned}$$

Therefore, $\mathbb{E}[\bar{Y}_n] \xrightarrow[n\to\infty]{} \mu$.

Preparation for showing that $\mathrm{Var}(\bar{Y}_n) \xrightarrow[n\to\infty]{} 0$. Since X_i's are pairwise independent, Y_i's are also pairwise independent, which implies that they are uncorrelated. Thus,

$$\mathrm{Var}\left(\sum_{i=1}^{n} Y_i \right) = \sum_{i=1}^{n} \mathrm{Var}(Y_i) \le \sum_{i=1}^{n} \mathbb{E}[Y_i^2] = \sum_{i=1}^{n} \mathbb{E}[X_i^2 \mathbf{1}_{\{|X_i| \le i\}}] \overset{\text{(identically distributed)}}{=} \sum_{i=1}^{n} \mathbb{E}[X_1^2 \mathbf{1}_{\{|X_1| \le i\}}],$$

and hence

$$\operatorname{Var}\big(\bar{Y}_n\big) = \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^n Y_i\right) \le \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\big[X_1^2\mathbf{1}_{\{|X_1|\le i\}}\big]$$

Finding a sufficiently tight upper bound on $\operatorname{Var}\big(\bar{Y}_n\big)$. While a simple method to determine an upper bound on $\operatorname{Var}\big(\bar{Y}_n\big)$ is as follows:

$$\sum_{i=1}^n \mathbb{E}\big[X_1^2\mathbf{1}_{\{|X_1|\le i\}}\big] \le \sum_{i=1}^n i\mathbb{E}\big[|X_1|\mathbf{1}_{\{|X_1|\le i\}}\big] \overset{\text{(drop indicator)}}{\le} \sum_{i=1}^n i\mathbb{E}[|X_1|] = \frac{1}{2}n(n+1)\mathbb{E}[|X_1|],$$

and thus

$$0 \le \operatorname{Var}\big(\bar{Y}_n\big) \le \frac{n(n+1)\mathbb{E}[|X_1|]}{2n^2},$$

the resulting upper bound is too crude and not sufficiently tight because it may not converge to 0 as $n \to \infty$. To tighten the upper bound, we proceed to introduce another level of truncation.

Introducing another level of truncation to show that $\operatorname{Var}\big(\bar{Y}_n\big) \underset{n\to\infty}{\longrightarrow} 0$. Let $\{a_n\}$ be a sequence of integers such that $0 \le a_n \le n$ for all $n \in \mathbb{N}$, $a_n \nearrow \infty$ and $a_n/n \underset{n\to\infty}{\longrightarrow} 0$. Then, we have

$$\begin{aligned}
\operatorname{Var}\left(\sum_{i=1}^n Y_i\right) &\le \sum_{i=1}^n \mathbb{E}\big[X_1^2\mathbf{1}_{\{|X_1|\le i\}}\big] \\
&= \sum_{i=1}^{a_n} \mathbb{E}\big[X_1^2\mathbf{1}_{\{|X_1|\le i\}}\big] + \sum_{i=a_n+1}^{n} \mathbb{E}\big[X_1^2\mathbf{1}_{\{|X_1|\le i\}}\big] \\
&= \sum_{i=1}^{a_n} \mathbb{E}\big[X_1^2\mathbf{1}_{\{|X_1|\le i\}}\big] + \sum_{i=a_n+1}^{n} \underbrace{\mathbb{E}\big[X_1^2\mathbf{1}_{\{|X_1|\le a_n\le i\}}\big]}_{\text{with another level of truncation}} + \sum_{i=a_n+1}^{n} \mathbb{E}\big[X_1^2\mathbf{1}_{\{a_n<|X_1|\le i\}}\big] \\
&\le \sum_{i=1}^{a_n} \underbrace{i}_{\le a_n}\mathbb{E}\big[|X_1|\mathbf{1}_{\{|X_1|\le i\}}\big] + \sum_{i=a_n+1}^{n} a_n\mathbb{E}\big[|X_1|\mathbf{1}_{\{|X_1|\le a_n\le i\}}\big] + \sum_{i=a_n+1}^{n} \underbrace{i}_{\le n}\mathbb{E}\big[|X_1|\mathbf{1}_{\{a_n<|X_1|\le i\}}\big] \\
&\le \underbrace{a_n\sum_{i=1}^{a_n}\mathbb{E}[|X_1|] + a_n\sum_{i=a_n+1}^{n}\mathbb{E}[|X_1|]}_{na_n\mathbb{E}[|X_1|]} + \underbrace{n\sum_{i=a_n+1}^{n}\mathbb{E}\big[|X_1|\mathbf{1}_{\{|X_1|>a_n\}}\big]}_{\le\sum_{i=1}^n \mathbb{E}\left[|X_1|\mathbf{1}_{\{|X_1|>a_n\}}\right]} \\
&\le na_n\mathbb{E}[|X_1|] + n^2\mathbb{E}\big[|X_1|\mathbf{1}_{\{|X_1|>a_n\}}\big].
\end{aligned}$$

This implies that

$$0 \le \operatorname{Var}\big(\bar{Y}_n\big) \le \underbrace{\frac{a_n}{n}}_{\underset{n\to\infty}{\longrightarrow}0}\mathbb{E}[|X_1|] + \underbrace{\mathbb{E}\big[|X_1|\mathbf{1}_{\{|X_1|>a_n\}}\big]}_{\underset{n\to\infty}{\longrightarrow}0 \text{ by Lemma 1.2.c}} \underset{n\to\infty}{\longrightarrow} 0,$$

and hence $\operatorname{Var}\big(\bar{Y}_n\big) \underset{n\to\infty}{\longrightarrow} 0$.

Completing the proof by Chebyshev's inequality. With $\mathbb{E}\big[\bar{Y}_n\big] \underset{n\to\infty}{\longrightarrow} \mu$ and $\operatorname{Var}\big(\bar{Y}_n\big) \underset{n\to\infty}{\longrightarrow} 0$, by Chebyshev's inequality, for all $\varepsilon > 0$ and $n \in \mathbb{N}$ we have

$$0 \le \mathbb{P}\big(|\bar{Y}_n - \mu| > \varepsilon\big) \le \frac{\mathbb{E}\big[(\bar{Y}_n-\mu)^2\big]}{\varepsilon^2} = \frac{\operatorname{Var}\big(\bar{Y}_n-\mu\big) + \mathbb{E}\big[\bar{Y}_n-\mu\big]^2}{\varepsilon^2} = \frac{\operatorname{Var}\big(\bar{Y}_n\big) + \big(\mathbb{E}\big[\bar{Y}_n\big]-\mu\big)^2}{\varepsilon^2} \underset{n\to\infty}{\longrightarrow} 0,$$

thus $\bar{Y}_n \underset{n\to\infty}{\xrightarrow{\text{p}}} \mu$, as desired. □

2.1.3 **Kolmogorov-Feller WLLN.** A classical and generalized form of WLLN is given by the *Kolmogorov-Feller WLLN*:

Theorem 2.1.c (Kolmogorov-Feller WLLN)**.** Let X_i's be independent random variables with cumulative distribution function (CDF) given by $F_i(x) = \mathbb{P}(X_i \le x)$ for all $i \in \mathbb{N}$. Let $\{a_n\}$ be a sequence such that $a_n > 0$ for all $n \in \mathbb{N}$ and $a_n \nearrow \infty$. Then, the following are equivalent:

(a) $(1/a_n)\sum_{i=1}^n X_i \xrightarrow[n\to\infty]{\text{p}} 0$.

(b) As $n \to \infty$,

i. $\sum_{i=1}^n \mathbb{P}(|X_i| \ge a_n) \to 0$.

ii. $\mathbb{E}\big[\sum_{i=1}^n (X_i/a_n)\mathbf{1}_{\{|X_i|<a_n\}}\big] \to 0$
(or $\mathbb{E}[\sum_{i=1}^n Y_{ni}] \to 0$ with $Y_{ni} := (X_i/a_n)\mathbf{1}_{\{(X_i/a_n)<1\}}$ for all $n, i \in \mathbb{N}$).

iii. $\mathrm{Var}\big(\sum_{i=1}^n (X_i/a_n)\mathbf{1}_{\{|X_i|<a_n\}}\big) \to 0$
(or $\mathrm{Var}(\sum_{i=1}^n Y_{ni}) \to 0$ with $Y_{ni} := (X_i/a_n)\mathbf{1}_{\{(X_i/a_n)<1\}}$ for all $n, i \in \mathbb{N}$).

Proof. Omitted. □

[Note: By taking $a_n = n$ for all $n \in \mathbb{N}$ and letting $Y_i := X_i\mathbf{1}_{\{|X_i|<n\}}$ for all $i \in \mathbb{N}$, (ii) and (iii) become $\mathbb{E}\big[\bar{Y}_n\big] \to 0$ and $\mathrm{Var}\big(\bar{Y}_n\big) \to 0$ as $n \to \infty$. By Chebyshev's inequality, we know that for all $\varepsilon > 0$, $\mathbb{P}\big(|\bar{Y}_n| > \varepsilon\big) \le \mathbb{E}\big[\bar{Y}_n^2\big]/\varepsilon^2 = (\mathrm{Var}\big(\bar{Y}_n\big) + \mathbb{E}\big[\bar{Y}_n\big]^2)/\varepsilon^2$. Letting $n \to \infty$ then yields $\lim_{n\to\infty} \mathbb{P}\big(|\bar{Y}_n| > \varepsilon\big) = 0$, i.e., $\bar{Y}_n \xrightarrow[n\to\infty]{\text{p}} 0$. Together with (i) which implies that $\{X_n\}$ and $\{Y_n\}$ are equivalent (as $\mathbb{P}(|X_i| \ge n) = \mathbb{P}(X_i \ne Y_i)$ here), we can obtain that $\bar{X}_n \xrightarrow[n\to\infty]{\text{p}} 0$. This illustrates how (b) $\implies$ (a) is established under this special case.]

2.2 Strong Laws of Large Numbers

2.2.1 For *weak* law of large numbers in Section 2.1, we focus on convergence *in probability* (relatively "weak" convergence). Next, in Section 2.2, we would like to establish *almost sure* convergence (relatively "strong" convergence) of sample sums and "generalized" sample means.

2.2.2 **Maximal inequalities.** Recall from Proposition 1.1.a that an equivalent criterion for the almost sure convergence $X_n \xrightarrow[n\to\infty]{\text{a.s.}} X$ is that for all $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\big(\big[\sup_{m\ge n} |X_m - X|\big] \ge \varepsilon\big) = 0$. Since we can write

$$\begin{aligned}
\mathbb{P}\left(\left[\sup_{m\ge n} |X_m - X|\right] \ge \varepsilon\right) &= \mathbb{P}\left(\bigcup_{\ell=n}^{\infty}\left\{\max_{n\le m\le \ell} |X_m - X| \ge \varepsilon\right\}\right) \\
&= \mathbb{P}\left(\lim_{\ell\to\infty}\left\{\max_{n\le m\le \ell} |X_m - X| \ge \varepsilon\right\}\right) \\
&= \lim_{\ell\to\infty} \underbrace{\mathbb{P}\left(\max_{n\le m\le \ell} |X_m - X| \ge \varepsilon\right)}_{\text{"maximal" probability}},
\end{aligned}$$

having *maximal inequalities*, i.e., inequalities about "maximal" probabilities, would be helpful for establishing almost sure convergence.

2.2.3 **Hajek-Renyi maximal inequality.**

Proposition 2.2.a (Hajek-Renyi maximal inequality)**.** Let X_i's be independent random variables with $\mathbb{E}[X_i] = 0$ and $\sigma_i = \mathrm{Var}(X_i) < \infty$ for all $i \in \mathbb{N}$. Let $\{c_k\}$ be a positive and decreasing sequence (i.e., $c_k > 0$ and $c_{k+1} \le c_k$ for all $k \in \mathbb{N}$). Then, for all $\varepsilon > 0$ and $m \le n$, we have

$$\mathbb{P}\left(\max_{m\le k\le n} c_k|S_k| \ge \varepsilon\right) \le \frac{1}{\varepsilon^2}\left(c_m^2\sum_{k=1}^{m}\sigma_k^2 + \sum_{k=m+1}^{n} c_k^2\sigma_k^2\right).$$

where $S_k = \sum_{i=1}^k X_i$ for all $k \in \mathbb{N}$.

Proof. **Introducing some helper notations.** Let:

- *(First occurrence for "$\geq \varepsilon$" for $c_k|S_k|$ as k goes from m to n)* $E_m := \{c_m|S_m| \geq \varepsilon\}$ and $E_j := \{\max_{m\leq k<j}(c_k|S_k|) < \varepsilon \text{ and } c_j|S_j| \geq \varepsilon\}$ for all $j = m+1, \ldots, n$. We can see that $E_m, \ldots, E_n$ are disjoint.
- $A := \{\max_{m\leq k\leq n}(c_k|S_k|) \geq \varepsilon\} = \biguplus_{j=m}^{n} E_j$.
- $Y := c_m^2 S_m^2 + \sum_{k=m+1}^{n} c_k^2(S_k^2 - S_{k-1}^2)$.

Since $\mathbb{E}\big[S_k^2\big] = \sum_{j=1}^{k} \sigma_j^2$, we can rewrite the maximal inequality as:

$$\sum_{j=m}^{n} \mathbb{P}(E_j) = \mathbb{P}(A) \leq \frac{1}{\varepsilon^2}\mathbb{E}[Y],$$

or equivalently,

$$\mathbb{E}[Y] \geq \varepsilon^2 \sum_{j=m}^{n} \mathbb{P}(E_j),$$

meaning that it suffices to establish such a lower bound on $\mathbb{E}[Y]$.

Establishing a lower bound on $\mathbb{E}[Y]$ by showing the nonnegativity of Y. Note that

$$\begin{aligned} Y &= c_m^2 S_m^2 + \sum_{k=m+1}^{n} c_k^2 S_k^2 - \sum_{k=m+1}^{n} c_k^2 S_{k-1}^2 = \sum_{k=m}^{n} c_k^2 S_k^2 - \sum_{k=m+1}^{n} c_k^2 S_{k-1}^2 \\ &\overset{\text{(adjust index)}}{=} \sum_{k=m}^{n} c_k^2 S_k^2 - \sum_{k=m}^{n-1} c_{k+1}^2 S_k^2 = \underbrace{c_n^2 S_n^2}_{\geq 0} + \sum_{k=m}^{n-1} \underbrace{(c_k^2 - c_{k+1}^2)}_{\geq 0 \text{ as } c_{k+1} \leq c_k} \underbrace{S_k^2}_{\geq 0} \geq 0. \end{aligned}$$

Therefore, we have

$$\mathbb{E}[Y] \overset{(Y\geq 0)}{\geq} \mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}\Big[Y\mathbf{1}_{\biguplus_{j=m}^{n} E_j}\Big] = \sum_{j=m}^{n} \mathbb{E}\big[Y\mathbf{1}_{E_j}\big].$$

Establishing a lower bound on each $\mathbb{E}\big[Y\mathbf{1}_{E_j}\big]$. For each $j = m, \ldots, n$, we have

$$\begin{aligned} \mathbb{E}\big[Y\mathbf{1}_{E_j}\big] &\overset{\text{(above expression)}}{=} \mathbb{E}\left[\left(c_n^2 S_n^2 + \sum_{k=m}^{n-1}(c_k^2 - c_{k+1}^2)S_k^2\right)\mathbf{1}_{E_j}\right] \\ &= c_n^2 \mathbb{E}\big[S_n^2\mathbf{1}_{E_j}\big] + \sum_{k=m}^{n-1}(c_k^2 - c_{k+1}^2)\mathbb{E}\big[S_k^2\mathbf{1}_{E_j}\big] \\ &\overset{(m\leq j)}{\geq} c_n^2 \mathbb{E}\big[S_n^2\mathbf{1}_{E_j}\big] + \sum_{k=j}^{n-1}(c_k^2 - c_{k+1}^2)\mathbb{E}\big[S_k^2\mathbf{1}_{E_j}\big] \end{aligned}$$

In view of the appearance of the expression $\mathbb{E}\big[S_k^2\mathbf{1}_{E_j}\big]$ for all $k = j, \ldots, n$, we also establish a lower bound on each of them as follows:

$$\begin{aligned} \mathbb{E}\big[S_k^2\mathbf{1}_{E_j}\big] &= \mathbb{E}\big[(S_j + (S_k - S_j))^2\mathbf{1}_{E_j}\big] \\ &= \mathbb{E}\big[S_j^2\mathbf{1}_{E_j}\big] + \mathbb{E}\big[(S_k - S_j)^2\mathbf{1}_{E_j}\big] + 2\underbrace{\mathbb{E}\big[S_j\mathbf{1}_{E_j}(S_k - S_j)\big]}_{=\mathbb{E}\left[S_j\mathbf{1}_{E_j}\right]\mathbb{E}[S_k - S_j]=0 \text{ as } S_j\mathbf{1}_{E_j} \text{ and } S_k - S_j \text{ are independent}} \\ &= \mathbb{E}\big[S_j^2\mathbf{1}_{E_j}\big] + \mathbb{E}\big[(S_k - S_j)^2\mathbf{1}_{E_j}\big] \geq \mathbb{E}\big[S_j^2\mathbf{1}_{E_j}\big] \underset{(c_j|S_j|\geq\varepsilon \text{ on } E_j)}{\geq} \mathbb{E}\left[\frac{\varepsilon^2}{c_j^2}\mathbf{1}_{E_j}\right] = \frac{\varepsilon^2\mathbb{P}(E_j)}{c_j^2}. \end{aligned}$$

Note that the resulting lower bound does not depend on k. Therefore, we have

$$\mathbb{E}\big[Y\mathbf{1}_{E_j}\big] \geq \underbrace{\left[\sum_{k=j}^{n-1}(c_k^2 - c_{k+1}^2) + c_n^2\right]}_{c_j^2} \frac{\varepsilon^2\mathbb{P}(E_j)}{c_j^2} = \varepsilon^2\mathbb{P}(E_j).$$

Completing the proof by combining all the lower bounds. It follows that

$$\mathbb{E}[Y] \geq \sum_{j=m}^{n} \varepsilon^2\mathbb{P}(E_j) = \varepsilon^2 \sum_{j=m}^{n} \mathbb{P}(E_j),$$

as desired. □

2.2.4 Kolmogorov maximal inequality.

Proposition 2.2.b (Kolmogorov maximal inequality)**.** Let X_i's be independent random variables with $\sigma_i^2 = \mathrm{Var}(X_i) < \infty$ for all $i \in \mathbb{N}$, and $S_k = \sum_{i=1}^{k} X_i$ for all $k \in \mathbb{N}$. For all $\varepsilon > 0$:

(a) *(Upper bound)* If $\mathbb{E}[X_i] = 0$ for all $i \in \mathbb{N}$, then

$$\mathbb{P}\left(\max_{1\leq k\leq n} |S_k| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(S_n)}{\varepsilon^2}$$

for all $n \in \mathbb{N}$.

(b) *(Lower bound I)* If $\mathbb{E}[X_i] = 0$ and $|X_i| \leq C < \infty$ for all $i \in \mathbb{N}$, where C is a constant, then

$$\mathbb{P}\left(\max_{1\leq k\leq n} |S_k| \geq \varepsilon\right) \geq 1 - \frac{(\varepsilon + C)^2}{\mathrm{Var}(S_n)}$$

for all $n \in \mathbb{N}$.

(c) *(Lower bound II)* If $\mathbb{E}[|X_i|] < \infty$ and $|X_i - \mathbb{E}[X_i]| \leq C < \infty$ for all $i \in \mathbb{N}$, where C is a constant, then

$$\mathbb{P}\left(\max_{1\leq k\leq n} |S_k| \geq \varepsilon\right) \geq 1 - \frac{(4\varepsilon + 2C)^2}{\mathrm{Var}(S_n)}$$

for all $n \in \mathbb{N}$.

Proof.

(a) Take $m = 1$ and $c_k = 1$ for all $k \in \mathbb{N}$ in Hajek-Renyi maximal inequality (Proposition 2.2.a).

(b) **Introducing some helper notations.** We reuse the notations used in the proof for Hajek-Renyi maximal inequality (Proposition 2.2.a) with $m = 1$ and $c_k = 1$ for all $k \in \mathbb{N}$ here:

- $E_1 := \{|S_1| \geq \varepsilon\}$ and $E_j := \{\max_{1\leq k<j} |S_k| < \varepsilon \text{ and } |S_j| \geq \varepsilon\}$ for all $j = 2, \ldots, n$.
- $A := \{\max_{1\leq k\leq n} |S_k| \geq \varepsilon\} = \biguplus_{j=1}^{n} E_j$.

Establishing an upper bound on $\mathbb{E}\big[S_n^2\mathbf{1}_A\big]$. First write

$$\begin{aligned}
\mathbb{E}\big[S_n^2\mathbf{1}_A\big] &= \sum_{j=1}^{n} \mathbb{E}\big[S_n^2\mathbf{1}_{E_j}\big] = \sum_{j=1}^{n} \left(\mathbb{E}\big[S_j^2\mathbf{1}_{E_j}\big] + \mathbb{E}\big[(S_n - S_j)^2\mathbf{1}_{E_j}\big] + 2\underbrace{\mathbb{E}\big[S_j\mathbf{1}_{E_j}(S_n - S_j)\big]}_{=0 \text{ by independence}}\right) \\
&= \sum_{j=1}^{n} \mathbb{E}\big[S_j^2\mathbf{1}_{E_j}\big] + \sum_{j=1}^{n} \mathbb{E}\big[(S_n - S_j)^2\mathbf{1}_{E_j}\big] \\
&\overset{(S_n - S_j \text{ and } \mathbf{1}_{E_j} \text{ are independent})}{=} \sum_{j=1}^{n} \mathbb{E}\big[S_j^2\mathbf{1}_{E_j}\big] + \sum_{j=1}^{n} \mathbb{E}\big[(S_n - S_j)^2\big]\mathbb{P}(E_j).
\end{aligned}$$

Since $S_j^2 = |S_{j-1} + X_j|^2 \overset{\text{(triangle)}}{\leq} (|S_{j-1}| + |X_j|)^2 \overset{\text{(assumption)}}{\leq} (|S_{j-1}| + C)^2$, we have

$$\sum_{j=1}^n \mathbb{E}\big[S_j^2 \mathbf{1}_{E_j}\big] \leq \sum_{j=1}^n \mathbb{E}\big[(|S_{j-1}| + C)^2 \mathbf{1}_{E_j}\big] \overset{(|S_{j-1}|<\varepsilon \text{ on } E_j)}{\leq} \sum_{j=1}^n \mathbb{E}\big[(\varepsilon + C)^2 \mathbf{1}_{E_j}\big] = (\varepsilon + C)^2 \mathbb{P}(A).$$

Also, by independence we have $\mathbb{E}\big[(S_n - S_j)^2\big] = \mathbb{E}\big[(X_{j+1} + \cdots + X_n)^2\big] = \sum_{k=j+1}^n \mathbb{E}\big[X_k^2\big] \leq \sum_{k=1}^n \mathbb{E}\big[X_k^2\big] = \mathbb{E}\big[S_n^2\big]$. Thus,

$$\sum_{j=1}^n \mathbb{E}\big[(S_n - S_j)^2\big]\mathbb{P}(E_j) \leq \sum_{j=1}^n \mathbb{E}\big[S_n^2\big]\mathbb{P}(E_j) = \mathbb{E}\big[S_n^2\big]\mathbb{P}(A).$$

Therefore, we get

$$\mathbb{E}\big[S_n^2 \mathbf{1}_A\big] \leq \big((\varepsilon + C)^2 + \mathbb{E}\big[S_n^2\big]\big)\mathbb{P}(A).$$

Establishing an lower bound on $\mathbb{E}\big[S_n^2 \mathbf{1}_A\big]$. On the other hand, we have

$$\mathbb{E}\big[S_n^2 \mathbf{1}_A\big] = \mathbb{E}\big[S_n^2\big] - \mathbb{E}\big[S_n^2 \mathbf{1}_{A^c}\big] \overset{(|S_n|<\varepsilon \text{ on } A^c)}{\geq} \mathbb{E}\big[S_n^2\big] - \mathbb{E}\big[\varepsilon^2 \mathbf{1}_{A^c}\big] = \mathbb{E}\big[S_n^2\big] - \varepsilon^2 \mathbb{P}(A^c) = \mathbb{E}\big[S_n^2\big] - \varepsilon^2 + \varepsilon^2 \mathbb{P}(A).$$

Completing the proof by combining the upper and lower bounds. It follows that

$$\mathbb{E}\big[S_n^2\big] - \varepsilon^2 + \varepsilon^2 \mathbb{P}(A) \leq \mathbb{E}\big[S_n^2 \mathbf{1}_A\big] \leq \big((\varepsilon + C)^2 + \mathbb{E}\big[S_n^2\big]\big)\mathbb{P}(A),$$

and hence

$$\mathbb{E}\big[S_n^2\big] - \varepsilon^2 \leq \big((\varepsilon + C)^2 + \mathbb{E}\big[S_n^2\big] - \varepsilon^2\big)\mathbb{P}(A).$$

Thus,

$$\mathbb{P}(A) \geq \frac{\mathbb{E}\big[S_n^2\big] - \varepsilon^2}{\mathbb{E}[S_n^2] + (\varepsilon + C)^2 + \varepsilon^2} = 1 - \frac{(\varepsilon + C)^2}{\mathbb{E}[S_n^2] + \underbrace{(\varepsilon + C)^2 + \varepsilon^2}_{\geq 0 \text{ as } C > 0}} \geq 1 - \frac{(\varepsilon + C)^2}{\mathbb{E}[S_n^2]} = 1 - \frac{(\varepsilon + C)^2}{\operatorname{Var}(S_n)},$$

as desired.

(c) Omitted; see Chung (2007, Theorem 5.3.2) for a proof.

□

2.2.5 **Mean criterion for a.s. convergence of series.**

Theorem 2.2.c. Let X_i's be nonnegative random variables. If $\sum_{n=1}^\infty \mathbb{E}[X_n] < \infty$, then $\sum_{n=1}^\infty X_n$ converges a.s, i.e., $S_n = \sum_{i=1}^n X_i \xrightarrow[n\to\infty]{\text{a.s.}} S$ for some random variable S.

Proof. By assumption, we have $\lim_{n\to\infty} \mathbb{E}[S_n] = \lim_{n\to\infty} \sum_{k=1}^n \mathbb{E}[X_k] = \sum_{k=1}^\infty \mathbb{E}[X_k] < \infty$. Therefore, $\{\mathbb{E}[S_n]\}$ is a convergent (hence Cauchy) sequence. By Markov's inequality, for all $m \geq n$ and all $\varepsilon > 0$, we have

$$0 \leq \mathbb{P}(|S_m - S_n| > \varepsilon) \leq \frac{1}{\varepsilon}\mathbb{E}[|S_m - S_n|] \overset{\text{(nonnegativity)}}{=} \frac{1}{\varepsilon}(\mathbb{E}[S_m] - \mathbb{E}[S_n]) \overset{\text{(Cauchy)}}{\longrightarrow}_{m,n\to\infty} 0.$$

This implies that $\lim_{m,n\to\infty} \mathbb{P}(|S_m - S_n| > \varepsilon) = 0$, and hence $S_n \xrightarrow[n\to\infty]{\text{p}} S$ for some random variable S by Proposition 1.1.c. Then, by [1.2.6], we have $S_{n_k} \xrightarrow[k\to\infty]{\text{a.s.}} S$ for some non-random positive integers $n_1 < n_2 < \cdots$. Since $\{S_n\}$ is an increasing sequence by the nonnegativity of X_i's, for all $n \in \mathbb{N}$, there exists $k \in \mathbb{N}$ such that

$$S_{n_k} \leq S_n \leq S_{n_{k+1}}.$$

As $S_{n_k} \xrightarrow[k\to\infty]{\text{a.s.}} S$ and $S_{n_{k+1}} \xrightarrow[k\to\infty]{\text{a.s.}} S$, we conclude that $S_n \xrightarrow[n\to\infty]{\text{a.s.}} S$, as desired. □

[Note: This proof illustrates another usage of the *subsequence method*.]

2.2.6 **Variance criterion for a.s. convergence of series.**

Theorem 2.2.d. Let X_i's be independent random variables with $\mathbb{E}[X_i] = 0$ and $\sigma_i^2 = \mathrm{Var}(X_i) < \infty$ for all $i \in \mathbb{N}$. If $\sum_{n=1}^{\infty} \mathrm{Var}(X_n) < \infty$, then $\sum_{n=1}^{\infty} X_n$ converges a.s.

[Note: Since the random variables are assumed to have zero mean, the variance is just equal to the second moment, so this can also be seen as a "second moment criterion" for a.s. convergence of series. In practice, to apply this result we may need to first subtract each random variable in consideration by its mean and then use this result on the random variables obtained after the subtractions.]

Proof. Here we use Cauchy criterion and the subsequence method, like the proof of Theorem 2.2.c. Let $S_n = \sum_{i=1}^{n} X_i$ for all $n \in \mathbb{N}$. By Chebyshev's inequality, for all $m \geq n$ and all $\varepsilon > 0$, we have,

$$0 \leq \mathbb{P}(|S_m - S_n| > \varepsilon) \leq \frac{1}{\varepsilon^2}\mathbb{E}\big[(S_m - S_n)^2\big] = \frac{1}{\varepsilon^2}\mathrm{Var}\left(\sum_{k=m+1}^{n} X_k\right)$$
$$\overset{\text{(independence)}}{=} \frac{1}{\varepsilon^2}\sum_{k=m+1}^{n}\mathrm{Var}(X_k) = \frac{1}{\varepsilon^2}(\mathrm{Var}(S_m) - \mathrm{Var}(S_n)) \overset{\text{(Cauchy)}}{\underset{m,n\to\infty}{\longrightarrow}} 0.$$

This implies that $\lim_{m,n\to\infty}\mathbb{P}(|S_m - S_n| > \varepsilon) = 0$, and hence $S_n \xrightarrow[n\to\infty]{\text{p}} S$ for some random variable S by Proposition 1.1.c. Then, by [1.2.6], we have $S_{n_k} \xrightarrow[k\to\infty]{\text{a.s.}} S$ for some non-random positive integers $n_1 < n_2 < \cdots$.

For all $n \in \mathbb{N}$, there exists $k \in \mathbb{N}$ such that $n_k < n \leq n_{k+1}$, and we have

$$0 \leq |S_n - S| \overset{\text{(triangle)}}{\leq} |S_{n_k} - S| + |S_n - S_{n_k}| \leq \underbrace{|S_{n_k} - S|}_{=:A_k} + \underbrace{\max_{n_k < j \leq n_{k+1}} |S_j - S_{n_k}|}_{=:B_k}. \tag{3}$$

From above, we immediately have $A_k \xrightarrow[k\to\infty]{\text{a.s.}} 0$. To show that $B_k \xrightarrow[k\to\infty]{\text{a.s.}} 0$ also, we apply Kolmogorov maximal inequality (Proposition 2.2.b) as follows. We first write

$$B_k = \max_{n_k < j \leq n_{k+1}}\left|\sum_{i=n_k+1}^{j} X_i\right| = \max_{n_k+1 \leq j \leq n_{k+1}}\left|\sum_{i=n_k+1}^{j} X_i\right| \overset{\text{(adjust indices)}}{=} \max_{1 \leq j \leq n_{k+1}-n_k}\left|\sum_{i=1}^{j} X'_i\right| = \max_{1 \leq j \leq n_{k+1}-n_k}|S'_j|$$

where $X'_i = X_{n_k+i}$ for all $i = 1, \ldots, j$ and $S'_j = \sum_{i=1}^{j} X'_i$ for all $j = 1, \ldots, n_{k+1} - n_k$. Now, the resulting expression matches with the one in Kolmogorov maximal inequality (Proposition 2.2.b), so we have

$$\sum_{k=1}^{\infty}\mathbb{P}(|B_k| \leq \varepsilon) \leq \sum_{k=1}^{\infty}\frac{\mathrm{Var}\left(S'_{n_{k+1}-n_k}\right)}{\varepsilon^2} = \sum_{k=1}^{\infty}\frac{\mathrm{Var}\left(\sum_{i=n_k+1}^{n_{k+1}} X_i\right)}{\varepsilon^2} = \frac{1}{\varepsilon^2}\sum_{k=1}^{\infty}\sum_{i=n_k+1}^{n_{k+1}}\mathrm{Var}(X_i) \leq \frac{1}{\varepsilon^2}\sum_{i=1}^{\infty}\mathrm{Var}(X_i) < \infty.$$

for all $\varepsilon > 0$. It then follows by [1.2.5] that $B_k \xrightarrow[k\to\infty]{\text{a.s.}} 0$.

Therefore, letting $n \to \infty$ in (3) (we have $k \to \infty$ as $n \to \infty$, viewing $k = k(n)$ as a function of n) yields $S_n \xrightarrow[n\to\infty]{\text{a.s.}} S$. □

2.2.7 **Kolmogorov strong law of large numbers (SLLN).** Using the variance criterion for a.s. convergence of series (Theorem 2.2.d) and Kronecker lemma (Lemma 2.2.e below), we can establish the *Kolmogorov SLLN*.

Lemma 2.2.e (Kronecker lemma)**.** If $a_n \nearrow \infty$ and $\sum_{n=1}^{\infty} y_n/a_n$ converges, then $(1/a_n)\sum_{i=1}^{n} y_i \to 0$ as $n \to \infty$.

Proof. Omitted. □

Theorem 2.2.f (Kolmogorov SLLN)**.** Let X_i's be independent random variables with $\mu_i := \mathbb{E}[X_i]$ and $\mathbb{E}\big[X_i^2\big] < \infty$ for all $i \in \mathbb{N}$. Denote $\bar{X}_n := n^{-1}\sum_{i=1}^n X_i$ and $\bar{\mu}_n := n^{-1}\sum_{i=1}^n \mu_i$ for all $n \in \mathbb{N}$. If $\sum_{i=1}^\infty \mathbb{E}\big[X_i^2\big]/i^2 < \infty$, then $\bar{X}_n - \bar{\mu}_n \xrightarrow[n\to\infty]{\text{a.s.}} 0$.

Proof. Let $Y_i = (X_i - \mathbb{E}[X_i])/i$, which has zero mean, for all $i \in \mathbb{N}$. Note that

$$\sum_{i=1}^\infty \mathrm{Var}(Y_i) = \sum_{i=1}^\infty \mathrm{Var}\left(\frac{X_i - \mathbb{E}[X_i]}{i}\right) = \sum_{i=1}^\infty \frac{\mathrm{Var}(X_i)}{i^2} = \sum_{i=1}^\infty \frac{\mathbb{E}\big[X_i^2\big] - \mathbb{E}[X_i]^2}{i^2} \le \sum_{i=1}^\infty \frac{\mathbb{E}\big[X_i^2\big]}{i^2} \overset{\text{(assumption)}}{<} \infty.$$

Therefore, by Theorem 2.2.d, $\sum_{n=1}^\infty Y_n = \sum_{n=1}^\infty (X_n - \mathbb{E}[X_n])/n$ converges a.s. By Kronecker lemma (Lemma 2.2.e), we then have $n^{-1}\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \xrightarrow[n\to\infty]{\text{a.s.}} 0$, as desired.

□

2.2.8 **Kolmogorov three series theorem.** The Kolmogorov strong law of large numbers (Theorem 2.2.f) only works when each random variable X_i has finite *second* moment, which limits its applicability. To weaken this assumption and extend the applicability, we need a more sophisticated technique, which involves the usage of *Kolmogorov three series theorem*.

Theorem 2.2.g (Kolmogorov three series theorem)**.** Let X_n's be independent random variables, and $Y_n = X_n \mathbf{1}_{\{|X_n| \le A\}}$ for all $n \in \mathbb{N}$ with $A > 0$. Then, $\sum_{n=1}^\infty X_n$ converges a.s. iff for some $A > 0$:

(a) $\sum_{n=1}^\infty \mathbb{P}(|X_n| > A) = \sum_{n=1}^\infty \mathbb{P}(X_n \neq Y_n) < \infty$.

(b) $\sum_{n=1}^\infty \mathbb{E}[Y_n]$ converges.

(c) $\sum_{n=1}^\infty \mathrm{Var}(Y_n) < \infty$.

Proof. "⇐": Assume that the three series converge. By (c), we have $\sum_{n=1}^\infty \mathrm{Var}(Y_n - \mathbb{E}[Y_n]) < \infty$. As each $Y_n - \mathbb{E}[Y_n]$ has zero mean, by Theorem 2.2.d, we know that $\sum_{n=1}^\infty (Y_n - \mathbb{E}[Y_n])$ converges a.s. Since $\sum_{n=1}^\infty \mathbb{E}[Y_n]$ converges by (b), $\sum_{n=1}^\infty Y_n$ also converges a.s. By (a), $\{X_n\}$ and $\{Y_n\}$ are equivalent, so by Proposition 2.1.a we conclude that $\sum_{n=1}^\infty X_n$ also converges a.s., as desired.

"⇒": Assume that $\sum_{n=1}^\infty X_n$ converges a.s.

(a) Note that $X_n = S_n - S_{n-1} \xrightarrow[n\to\infty]{\text{a.s.}} 0$. Therefore, by Proposition 1.1.a, we have $\mathbb{P}(|X_n| > A \text{ io}) = 0$. By Borel 0-1 law, we must then have $\sum_{n=1}^\infty \mathbb{P}(|X_n| > A) < \infty$, establishing (a).

(c) From (a), we know that $\{X_n\}$ and $\{Y_n\}$ are equivalent. Thus, $\sum_{n=1}^\infty Y_n$ converges a.s. also by assumption.
Since $|Y_n - \mathbb{E}[Y_n]| \le |Y_n| + |\mathbb{E}[Y_n]| \le 2A < \infty$ for all $n \in \mathbb{N}$, applying Kolmogorov maximal inequality (Proposition 2.2.b) gives

$$\begin{aligned} 1 \ge \mathbb{P}\left(\max_{n\le j\le r}\left|\sum_{i=n}^j Y_i\right| \ge \varepsilon\right) &\ge 1 - \frac{(4\varepsilon + 4A)^2}{\mathrm{Var}(\sum_{i=n}^r Y_i)} \\ = 1 - \frac{(4\varepsilon + 4A)^2}{\mathrm{Var}(\sum_{i=n}^r Y_i)} &\overset{\text{(independence)}}{=} 1 - \frac{(4\varepsilon + 4A)^2}{\sum_{i=n}^r \mathrm{Var}(Y_i)} \end{aligned}$$

for all $\varepsilon > 0$ and $n \in \mathbb{N}$.
Assume to the contrary that $\sum_{n=1}^\infty \mathrm{Var}(Y_n) = \infty$. Then, letting $r \to \infty$ would yield

$$\lim_{r\to\infty} \mathbb{P}\left(\max_{n\le j\le r}\left|\sum_{i=n}^j Y_i\right| \ge \varepsilon\right) = \mathbb{P}\left(\sup_{j\ge n}\left|\sum_{i=n}^j Y_i\right| \ge \varepsilon\right) = 1$$

for all $\varepsilon > 0$ and $n \in \mathbb{N}$. Since $\varepsilon > 0$ and $n \in \mathbb{N}$ can be arbitrarily large, this implies that $\sum_{n=1}^\infty Y_n$ cannot converge a.s., contradiction.

(b) Recall that $\sum_{n=1}^{\infty} Y_n$ converges a.s. by (a). By (c), we have $\sum_{n=1}^{\infty} \mathbb{E}\big[(Y_n - \mathbb{E}[Y_n])^2\big] < \infty$, which implies by Theorem 2.2.d that $\sum_{n=1}^{\infty}(Y_n - \mathbb{E}[Y_n])$ converges a.s. It follows that $\sum_{n=1}^{\infty} \mathbb{E}[Y_n]$ converges a.s., establishing (b).

□

As a corollary, we have the following criterion for the a.s. convergence of *absolute* random series, which involves only two series:

Corollary 2.2.h (Kolmogorov two series thoerem for a.s. absolute convergence)**.** Let X_n's be independent. Then $\sum_{n=1}^{\infty} |X_n|$ converges a.s. iff $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| \geq A) < \infty$ and $\sum_{n=1}^{\infty} \mathbb{E}\big[|X_n| \mathbf{1}_{\{|X_n| \leq A\}}\big] < \infty$ for some $A > 0$.

Proof. By Kolmogorov three-series theorem (Theorem 2.2.g), it suffices to show that the convergence of the two series implies that $\sum_{n=1}^{\infty} \mathrm{Var}\big(|X_n| \mathbf{1}_{\{|X_n| < A\}}\big) < \infty$.

So we now assume that $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| \geq A) < \infty$ and $\sum_{n=1}^{\infty} \mathbb{E}\big[|X_n| \mathbf{1}_{\{|X_n| \leq A\}}\big] < \infty$. Then, we hve

$$\sum_{n=1}^{\infty} \mathrm{Var}\big(|X_n| \mathbf{1}_{\{|X_n| < A\}}\big) \leq \sum_{n=1}^{\infty} \mathbb{E}\big[X_n^2 \mathbf{1}_{\{|X_n| < A\}}\big] \leq A \underbrace{\sum_{n=1}^{\infty} \mathbb{E}\big[|X_n| \mathbf{1}_{\{|X_n| < A\}}\big]}_{< \infty \text{ by assumption}} < \infty,$$

as desired. □

2.2.9 **SLLN for independent random variables with sufficient moment conditions.**

Proposition 2.2.i (SLLN for independent random variables)**.** Let X_n's be independent random variables. Assume that:

(a) For all $n \in \mathbb{N}$, $g_n : \mathbb{R} \to \mathbb{R}$ is an even and positive function which is increasing function on $(0, \infty)$, which satisfies at least one of the following conditions:
   i. $x/g_n(x)$ is increasing on $(0, \infty)$.
   ii. $x/g_n(x)$ is decreasing on $(0, \infty)$, $x^2/g_n(x)$ is increasing on $(0, \infty)$, and $\mathbb{E}[X_n] = 0$.
   iii. $x^2/g_n(x)$ is increasing on $(0, \infty)$, and X_n is symmetric about 0, i.e., $X_n \overset{\mathrm{d}}{=} -X_n$.

(b) $\sum_{n=1}^{\infty} \mathbb{E}[g_n(X_n)/g_n(a_n)] < \infty$ with $a_n > 0$ for all $n \in \mathbb{N}$.

Then,

$$\sum_{n=1}^{\infty} \frac{X_n}{a_n} \text{ converges a.s.}$$

Assuming further that $a_n \nearrow \infty$, we get

$$\frac{1}{a_n} \sum_{i=1}^{n} X_i \xrightarrow[n \to \infty]{\text{a.s.}} 0.$$

Proof. Let F_n be the CDF of X_n and $Y_n = X_n \mathbf{1}_{\{|X_n| < a_n\}}$ for all $n \in \mathbb{N}$. Then, we can write $Y_n/a_n = (X_n/a_n)\mathbf{1}_{\{|X_n|/a_n\}} < 1$ for all $n \in \mathbb{N}$. By Kolmogorov three-series theorem with $A = 1$, to establish the a.s. convergence of $\sum_{n=1}^{\infty} X_n/a_n$, it suffices to show the convergence of three series.

Showing that $\sum_{n=1}^{\infty} \mathbb{P}(|X_n|/a_n \geq 1) < \infty$. Since g_n is even and increasing on $(0, \infty)$, we have $|X_n| \geq a_n > 0 \implies g_n(X_n) \geq g_n(a_n)$. Thus, by (b) of the assumption, we have

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n| \geq a_n) \leq \sum_{n=1}^{\infty} \underbrace{\mathbb{P}(g_n(X_n) \geq g_n(a_n))}_{= \mathbb{P}(|g_n(X_n)| \geq g_n(a_n))} \overset{\text{(Markov)}}{\leq} \sum_{n=1}^{\infty} \frac{\overbrace{\mathbb{E}[g_n(X_n)]}^{\mathbb{E}[g_n(X_n)]}}{g_n(a_n)} < \infty.$$

Showing that $\sum_{n=1}^{\infty} \mathbb{E}[Y_n/a_n]$ converges. We consider the three cases for part (a) of the assumption.

- For the case where (i) holds, since $|Y_n| < a_n$ by construction, we have $|Y_n|/g_n(|Y_n|) \le a_n/g_n(a_n)$, which implies that $|Y_n|/a_n \le g_n(|Y_n|)/g_n(a_n))$ for all $n \in \mathbb{N}$. It follows that

$$\left|\sum_{n=1}^{\infty}\mathbb{E}\left[\frac{Y_n}{a_n}\right]\right| \overset{\text{(triangle)}}{\le} \sum_{n=1}^{\infty}\mathbb{E}\left[\frac{|Y_n|}{a_n}\right] \le \sum_{n=1}^{\infty}\mathbb{E}\left[\frac{g_n(|Y_n|)}{g_n(a_n)}\right] \le \sum_{n=1}^{\infty}\mathbb{E}\left[\frac{g_n(|X_n|)}{g_n(a_n)}\right] \overset{\text{(even)}}{=} \sum_{n=1}^{\infty}\mathbb{E}\left[\frac{g_n(X_n)}{g_n(a_n)}\right] \overset{\text{(b)}}{<} \infty.$$

- For the case where (ii) holds, $|X_n| \ge a_n$ implies

$$\frac{|X_n|}{g_n(|X_n|)} \le \frac{a_n}{g_n(a_n)},$$

and thus

$$\frac{|X_n|}{a_n} \le \frac{g_n(|X_n|)}{g_n(a_n)} = \frac{g_n(X_n)}{g_n(a_n)},$$

for all $n \in \mathbb{N}$. Together with $\mathbb{E}[X_n] = 0$, we then get

$$\begin{aligned}\left|\sum_{n=1}^{\infty}\mathbb{E}\left[\frac{Y_n}{a_n}\right]\right| &= \left|\sum_{n=1}^{\infty}\mathbb{E}\left[\frac{X_n}{a_n}\mathbf{1}_{\{|X_n|<a_n\}}\right]\right| \overset{(\mathbb{E}[X_n]=0)}{=} \left|-\sum_{n=1}^{\infty}\mathbb{E}\left[\frac{X_n}{a_n}\mathbf{1}_{\{|X_n|\ge a_n\}}\right]\right| \\ &\le \left|\sum_{n=1}^{\infty}\mathbb{E}\left[\frac{X_n}{a_n}\right]\right| \overset{\text{(triangle)}}{\le} \sum_{n=1}^{\infty}\mathbb{E}\left[\frac{|X_n|}{a_n}\right] \overset{\text{(above)}}{\le} \sum_{n=1}^{\infty}\mathbb{E}\left[\frac{g_n(X_n)}{g_n(a_n)}\right] \overset{\text{(b)}}{<} \infty.\end{aligned}$$

- For the case where (iii) holds, since X_n is symmetric about 0, we have $\mathbb{E}\big[X_n\mathbf{1}_{\{|X_n|<a_n\}}\big] = \mathbb{E}\big[-X_n\mathbf{1}_{\{|-X_n|<a_n\}}\big] = -\mathbb{E}\big[X_n\mathbf{1}_{\{|X_n|<a_n\}}\big]$, which implies that $\mathbb{E}[Y_n] = \mathbb{E}\big[X_n\mathbf{1}_{\{|X_n|<a_n\}}\big] = 0$, for all $n \in \mathbb{N}$. Thus, we immediately get $\sum_{n=1}^{\infty}\mathbb{E}[Y_n/a_n] = 0$.

Showing that $\sum_{n=1}^{\infty}\mathrm{Var}(Y_n/a_n) < \infty$.

Claim: $|X_n| < a_n$ implies $X_n^2/a_n^2 \le g_n(X_n)/g_n(a_n)$ for all $n \in \mathbb{N}$.

Proof. We again consider the three cases for part (a) of the assumption.

- For the case where (i) holds, $|X_n| < a_n$ implies

$$\frac{|X_n|}{g_n(X_n)} \le \frac{a_n}{g_n(a_n)},$$

and thus

$$\frac{X_n^2}{a_n^2} \le \frac{g_n(X_n)^2}{g_n(a_n)^2} \le \frac{g_n(X_n)}{g_n(a_n)},$$

since we have $|X_n| < a_n \implies g(X_n) \le g(a_n) \implies g_n(X_n)/g_n(a_n) \le 1$.

- For the case where (ii) or (iii) holds, $x^2/g_n(x)$ is increasing on $(0,\infty)$. Therefore, $|X_n| < a_n$ implies

$$\frac{X_n^2}{g_n(X_n)} \le \frac{a_n^2}{g_n(a_n)}.$$

Multiplying both sides by $g_n(X_n)/a_n^2 > 0$ then yields

$$\frac{X_n^2}{a_n^2} \le \frac{g_n(X_n)}{g_n(a_n)}.$$

□

By the claim, we then have

$$\sum_{n=1}^{\infty}\mathrm{Var}\left(\frac{Y_n}{a_n}\right) \le \sum_{n=1}^{\infty}\mathbb{E}\left[\frac{Y_n^2}{a_n^2}\right] = \sum_{n=1}^{\infty}\mathbb{E}\left[\frac{X_n^2}{a_n^2}\mathbf{1}_{\{|X_n|<a_n\}}\right] \le \sum_{n=1}^{\infty}\mathbb{E}\left[\frac{g(X_n)}{g(a_n)}\mathbf{1}_{\{|X_n|<a_n\}}\right] \le \sum_{n=1}^{\infty}\mathbb{E}\left[\frac{g(X_n)}{g(a_n)}\right] \overset{\text{(b)}}{<} \infty,$$

as desired.

So, we have established that

$$\sum_{n=1}^{\infty} \frac{X_n}{a_n} \text{ converges a.s.}$$

By Kronecker lemma, we also have

$$\frac{1}{a_n} \sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

□

Choosing $g_n(x) = |x|^r$ with $0 < r \leq 2$ for all $n \in \mathbb{N}$ in Proposition 2.2.i, we get the following:

Corollary 2.2.j (SLLN for independent random variables with sufficient moment conditions)**.** Let X_n's be independent random variables, $a_n \nearrow \infty$ with $a_n > 0$ for all $\in \mathbb{N}$. Assume that $\sum_{n=1}^{\infty} \mathbb{E}[|X_n|^r]/a_n^r < \infty$ for some $0 < r \leq 2$. Then,

(a) $(1/a_n)\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\text{a.s.}} 0$ if $0 < r \leq 1$.

(b) $(1/a_n)\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \xrightarrow[n\to\infty]{\text{a.s.}} 0$ if $1 < r < 2$.

Proof. Take $g_n(x) = |x|^r$ with $0 < r \leq 2$ for all $n \in \mathbb{N}$.

(a) With $0 < r \leq 1$, part (a)(i) of the assumption in Proposition 2.2.i is satisfied, as $x/|x|^r$ is increasing on $(0, \infty)$. Note that the assumption that $\sum_{n=1}^{\infty} \mathbb{E}[|X_n|^r]/a_n^r < \infty$ is equivalent to part (b) of the assumption in Proposition 2.2.i, so (b) holds also. Therefore, by Proposition 2.2.i, we readily have $(1/a_n)\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\text{a.s.}} 0$.

(b) With $1 < r < 2$, $x/|x|^r$ is decreasing on $(0, \infty)$. So, for the purpose of applying Proposition 2.2.i we consider $X_n - \mathbb{E}[X_n]$, which has zero mean always, and thus part (a)(ii) of the assumption in Proposition 2.2.i is fulfilled. Now, we show that part (b) of the assumption holds also:

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{\mathbb{E}[g_n(X_n - \mathbb{E}[X_n])]}{g_n(a_n)} &= \sum_{n=1}^{\infty} \frac{\mathbb{E}[|X_n - \mathbb{E}[X_n]|^r]}{\underbrace{|a_n|^r}_{a_n^r}} \overset{(C_r\text{-inequality})}{\leq} \sum_{n=1}^{\infty} \frac{C_r \mathbb{E}[|X_n|^r + |\mathbb{E}[X_n]|^r]}{a_n^r} \\
&= \sum_{n=1}^{\infty} \frac{C_r(\mathbb{E}[|X_n|^r] + |\mathbb{E}[X_n]|^r)}{a_n^r} \\
&\overset{(|\mathbb{E}[X_n]| \leq \mathbb{E}[|X_n|] = \|X_n\|_1 \leq \|X_n\|_r = (\mathbb{E}[|X_n|^r])^{1/r})}{\leq} \sum_{n=1}^{\infty} \frac{C_r(\mathbb{E}[|X_n|^r] + \mathbb{E}[|X_n|^r])}{a_n^r} \\
&= 2C_r \underbrace{\sum_{n=1}^{\infty} \frac{\mathbb{E}[|X_n|^r]}{a_n^r}}_{<\infty \text{ by assumption}} < \infty.
\end{aligned}$$

Thus, by Proposition 2.2.i, we have $(1/a_n)\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \xrightarrow[n\to\infty]{\text{a.s.}} 0$.

□

2.2.10 Kolmogorov SLLN for iid random variables.

Lemma 2.2.k. Let X be a random variable. Then,

$$\sum_{n=1}^{\infty} \mathbb{P}(|X| \geq n) \leq \mathbb{E}[|X|] \leq 1 + \sum_{n=1}^{\infty} \mathbb{P}(|X| \geq n).$$

Proof. It follows from taking expectation on the following chain of inequalities:

$$\sum_{n=1}^{\infty} \mathbf{1}_{\{|X|\geq n\}} \leq |X| \leq 1 + \sum_{n=1}^{\infty} \mathbf{1}_{\{|X|\geq n\}}.$$

□

Theorem 2.2.l (Kolmogorov SLLN for iid random variables)**.** Let X_i's be iid random variables and $S_n = \sum_{i=1}^n X_i$ for all $n \in \mathbb{N}$. Then:

(a) $\mathbb{E}[|X_1|] < \infty \implies S_n/n \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}[X_1]$.

(b) $\mathbb{E}[|X_1|] = \infty \implies \limsup_{n\to\infty} |S_n|/n = \infty$ a.s.

Proof.

(a) Assume that $\mathbb{E}[|X_1|] < \infty$. Let $Y_n = X_n \mathbf{1}_{\{|X_n|\leq n\}}$ for all $n \in \mathbb{N}$. Then, we have

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n \neq Y_n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > n) = \sum_{n=1}^{\infty} \mathbb{P}(|X_1| > n) \overset{\text{(Lemma 2.2.k)}}{\leq} \mathbb{E}[|X_1|] < \infty.$$

Thus, $\{X_n\}$ and $\{Y_n\}$ are equivalent. So, it suffices to show that $n^{-1}\sum_{i=1}^n Y_i \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}[X_1]$.

Showing that $n^{-1}\sum_{i=1}^n \mathbb{E}[Y_i] \xrightarrow[n\to\infty]{} \mathbb{E}[X_1]$**.** For all $n \in \mathbb{N}$, we have

$$\mathbb{E}[Y_n] = \mathbb{E}\big[X_n \mathbf{1}_{\{|X_n|\leq n\}}\big] = \mathbb{E}\big[X_1 \mathbf{1}_{\{|X_1|\leq n\}}\big] \overset{\text{(DCT)}}{\xrightarrow[n\to\infty]{}} \mathbb{E}[X_1].$$

This implies that $n^{-1}\sum_{i=1}^n \mathbb{E}[Y_i] \xrightarrow[n\to\infty]{} \mathbb{E}[X_1]$.

Showing that $n^{-1}\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) \xrightarrow[n\to\infty]{\text{a.s.}} 0$**.** By Corollary 2.2.j on Y_n's with $a_n = n$ for all $n \in \mathbb{N}$ and $r = 2$, it suffices to show that $\sum_{n=1}^{\infty} \mathbb{E}\big[Y_n^2\big]/n^2 < \infty$.
We proceed as follows:

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{\mathbb{E}\big[Y_n^2\big]}{n^2} &= \sum_{n=1}^{\infty} \frac{\mathbb{E}\big[Y_1^2\big]}{n^2} = \sum_{n=1}^{\infty} \frac{1}{n^2} \mathbb{E}\big[X_1^2 \mathbf{1}_{\{|X_1|\leq n\}}\big] = \sum_{n=1}^{\infty}\sum_{k=1}^{n} \frac{1}{n^2} \mathbb{E}\big[X_1^2 \mathbf{1}_{\{k-1<|X_1|\leq k\}}\big] \\
&= \sum_{k=1}^{\infty}\sum_{n=k}^{\infty} \frac{1}{n^2} \mathbb{E}\big[X_1^2 \mathbf{1}_{\{k-1<|X_1|\leq k\}}\big] = \sum_{k=1}^{\infty} \left[\mathbb{E}\big[X_1^2 \mathbf{1}_{\{k-1<|X_1|\leq k\}}\big] \left(\sum_{n=k}^{\infty} \frac{1}{n^2}\right)\right] \\
&\overset{\text{(can show: } \sum_{n=k}^{\infty} 1/n^2 \leq C/k \text{ for some } C>0)}{\leq} \sum_{k=1}^{\infty} \left[\mathbb{E}\big[X_1^2 \mathbf{1}_{\{k-1<|X_1|\leq k\}}\big] \left(\frac{C}{k}\right)\right] \\
&\leq \sum_{k=1}^{\infty} \left[\mathbb{E}\big[k|X_1| \mathbf{1}_{\{k-1<|X_1|\leq k\}}\big] \left(\frac{C}{k}\right)\right] \\
&= C \sum_{k=1}^{\infty} \mathbb{E}\big[|X_1| \mathbf{1}_{\{k-1<|X_1|\leq k\}}\big] = C\mathbb{E}[|X_1|] < \infty.
\end{aligned}$$

Completing the proof. With $n^{-1}\sum_{i=1}^n \mathbb{E}[Y_i] \xrightarrow[n\to\infty]{} \mathbb{E}[X_1]$ and $n^{-1}\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) \xrightarrow[n\to\infty]{\text{a.s.}} 0$, we get $n^{-1}\sum_{i=1}^n Y_i \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}[X_1]$, as desired.

(b) Assume that $\mathbb{E}[|X_1|] = \infty$. Then, for all $A > 0$, we have $\mathbb{E}[|X_1|/A] = \infty$. Therefore, we have

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > An) = \sum_{n=1}^{\infty} \mathbb{P}(|X_1|/A > n) \overset{\text{(Lemma 2.2.k)}}{=} \infty.$$

Hence, by Borel-Cantelli lemma ([1.0.2]b), we have

$$\begin{aligned}1 &= \mathbb{P}(|X_n| > An \text{ io}) = \mathbb{P}(|S_n - S_{n-1}| > An \text{ io}) \\ &= \mathbb{P}(|S_n| > An/2 \text{ or } |S_{n-1}| > A(n-1)/2 \text{ io}) = \mathbb{P}(|S_n| > An/2 \text{ io}) = \mathbb{P}(|S_n|/n > A/2 \text{ io}) \\ &= \mathbb{P}\left(\limsup_{n\to\infty}\left\{\frac{|S_n|}{n} > \frac{A}{2}\right\}\right) = \mathbb{P}\left(\left\{\limsup_{n\to\infty}\frac{|S_n|}{n} > \frac{A}{2}\right\}\right).\end{aligned}$$

This implies that for all $m \in \mathbb{N}$, there exists a null set N_m such that for all $\omega \in \Omega \setminus N_m$, we have

$$\limsup_{n\to\infty}\frac{|S_n(\omega)|}{n} > \frac{m}{2}.$$

Now let $N = \bigcup_{m=1}^{\infty} N_m$, which is still a null set. Then, for all $\omega \in \Omega \setminus N$, we have

$$\limsup_{n\to\infty}\frac{|S_n(\omega)|}{n} > \frac{A}{2} \quad \text{for all } A > 0 \implies \limsup_{n\to\infty}\frac{|S_n(\omega)|}{n} = \infty.$$

As $\mathbb{P}(\Omega \setminus N) = 1$, we have $\limsup_{n\to\infty} |S_n|/n = \infty$ a.s.

□

Corollary 2.2.m (SLLN for iid random variables with necessary and sufficient moment condition)**.** Let X_i's be iid random variables and $S_n = \sum_{i=1}^n X_i$ for all $n \in \mathbb{N}$. Then, $\bar{X}_n = S_n/n \xrightarrow[n\to\infty]{\text{a.s.}} c$ iff $\mathbb{E}[X_1] = c$, where $c \in \mathbb{R}$.

Proof. "⇐": It follows from Theorem 2.2.l.
"⇒": Assume $S_n/n \xrightarrow[n\to\infty]{\text{a.s.}} c$. Then, we have

$$\frac{X_n}{n} = \frac{S_n}{n} - \frac{S_{n-1}}{n} = \frac{S_n}{n} - \frac{S_{n-1}}{n-1}\cdot\frac{n-1}{n} \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

Therefore, we have $\mathbb{P}(|X_n| \geq n \text{ io}) = \mathbb{P}(|X_n|/n \geq 1 \text{ io}) = 0$ by Proposition 1.1.a. Thus,

$$\sum_{n=1}^{\infty}\mathbb{P}(|X_1| \geq n) = \sum_{n=1}^{\infty}\mathbb{P}(|X_n| \geq n) \overset{\text{(Borel 0-1 law)}}{<} \infty,$$

which implies by Lemma 2.2.k that $\mathbb{E}[|X_1|] < \infty$. Then, by Theorem 2.2.l, we have $S_n/n \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}[X_1]$. By the uniqueness of limit, we then conclude that $\mathbb{E}[X_1] = c$. □

2.2.11 **Marcinkiewicz SLLN for iid random variables.** By Theorem 2.2.l, we know that if we have $\mathbb{E}[|X_1|] < \infty$, then we get $S_n/n \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}[X_1]$, or equivalently, $(1/n)\sum_{i=1}^n (X_i - \mathbb{E}[X_1]) \xrightarrow[n\to\infty]{\text{a.s.}} 0$. This means that the sum $\sum_{i=1}^n (X_i - \mathbb{E}[X_1])$ of deviations from mean needs to be multiplied by a normalization factor $1/n$, which converges to 0 fast enough, to have a.s. convergence to zero. The following kind of SLLN suggests that if *higher* absolute moment is available, then the normalization factor is allowed to converge to 0 at a slower speed, while still ensuring that the sum can converge to zero a.s. after the multiplication. Intuitively, it indicates that the availability of higher absolute moment makes the sum of deviations from mean "more stable", and thus a "smaller degree" of normalization is needed to obtain the desired convergence.

Proposition 2.2.n (Marcinkiewicz SLLN for iid random variables)**.** Let X_i's be iid random variables, and $0 < r < 2$. Then,

$$\frac{1}{n^{1/r}}\sum_{i=1}^n (X_i - a) \xrightarrow[n\to\infty]{\text{a.s.}} 0$$

iff $\mathbb{E}[|X_1|^r] < \infty$, where $a = \mathbb{E}[X_1]$ if $1 \leq r < 2$, and a can be arbitrary real number if $0 < r < 1$ (usually we consider $a = 0$).

Proof. "⇒": Let $S_n = \sum_{i=1}^n (X_i - a)$ for all $n \in \mathbb{N}$. Assume $S_n/n^{1/r} \xrightarrow[n\to\infty]{\text{a.s.}} 0$. Then, we have

$$\frac{X_n}{n^{1/r}} = \frac{a}{n^{1/r}} + \frac{S_n}{n^{1/r}} - \frac{S_{n-1}}{n^{1/r}} = \underbrace{\frac{a}{n^{1/r}}}_{\xrightarrow[n\to\infty]{} 0} + \underbrace{\frac{S_n}{n^{1/r}}}_{\xrightarrow[n\to\infty]{\text{a.s.}} 0} - \underbrace{\frac{S_{n-1}}{(n-1)^{1/r}}}_{\xrightarrow[n\to\infty]{\text{a.s.}} 0} \cdot \underbrace{\frac{(n-1)^{1/r}}{n^{1/r}}}_{\xrightarrow[n\to\infty]{} 1} \xrightarrow[n\to\infty]{\text{a.s.}} 0,$$

which implies by Proposition 1.1.a that $\mathbb{P}\big(|X_n| \geq n^{1/r} \text{ io}\big) = \mathbb{P}\big(|X_n/n^{1/r}| > 1 \text{ io}\big) = 0$. Hence, by Borel 0-1 law, we have

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

which implies that $\mathbb{E}[|X_1|^r] < \infty$ by Lemma 2.2.k.

"⇐": Assume that $\mathbb{E}[|X_1|^r] < \infty$. Let $Y_n = X_n \mathbf{1}_{\{|X_n| \leq n^{1/r}\}}$ for all $n \in \mathbb{N}$. Then, we have

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n \neq Y_n) = \sum_{n=1}^{\infty} \mathbb{P}\Big(|X_n| > n^{1/r}\Big) = \sum_{n=1}^{\infty} \mathbb{P}\Big(|X_1| > n^{1/r}\Big) = \sum_{n=1}^{\infty} \mathbb{P}(|X_1|^r > n) \overset{\text{(Lemma 2.2.k)}}{\leq} \mathbb{E}[|X_1|^r] < \infty.$$

Thus, $\{X_n\}$ and $\{Y_n\}$ are equivalent. So, it suffices to show that $(1/n^{1/r}) \sum_{i=1}^n (Y_i - a) \xrightarrow[n\to\infty]{\text{a.s.}} 0$. We consider three cases.

Case 1: $r = 1$**.** It follows from Theorem 2.2.l.

Case 2: $0 < r < 1$**.** By Corollary 2.2.j on Y_n's with $a_n = n^{1/r}$ for all $n \in \mathbb{N}$, it suffices to show that $\sum_{n=1}^{\infty} \mathbb{E}[|Y_n|^r]/(n^{1/r})^r < \infty$, because $a/n^{1/r} \underset{n\to\infty}{\longrightarrow} 0$ for all $a \in \mathbb{R}$.

We proceed as follows:

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{\mathbb{E}[|Y_n|^r]}{(n^{1/r})^r} \leq \sum_{n=1}^{\infty} \frac{\mathbb{E}[|Y_n|]}{n^{1/r}} &= \sum_{n=1}^{\infty} \frac{1}{n^{1/r}} \mathbb{E}\Big[|X_n| \mathbf{1}_{\{|X_n| \leq n^{1/r}\}}\Big] = \sum_{n=1}^{\infty} \frac{1}{n^{1/r}} \mathbb{E}\big[|X_1| \mathbf{1}_{\{|X_1|^r \leq n\}}\big] \\
&= \sum_{n=1}^{\infty} \sum_{k=1}^{n} \frac{1}{n^{1/r}} \mathbb{E}\big[|X_1| \mathbf{1}_{\{k-1 < |X_1|^r \leq k\}}\big] = \sum_{k=1}^{\infty} \sum_{n=k}^{\infty} \frac{1}{n^{1/r}} \mathbb{E}\big[|X_1| \mathbf{1}_{\{k-1 < |X_1|^r \leq k\}}\big] \\
&= \sum_{k=1}^{\infty} \left[\mathbb{E}\big[|X_1| \mathbf{1}_{\{k-1 < |X_1|^r \leq k\}}\big] \left(\sum_{n=k}^{\infty} \frac{1}{n^{1/r}} \right) \right] \\
&\leq \sum_{k=1}^{\infty} \left[\mathbb{E}\Big[k^{1/r} \mathbf{1}_{\{k-1 < |X_1|^r \leq k\}}\Big] \left(\frac{C}{k^{1/r-1}} \right) \right] \\
&= \sum_{k=1}^{\infty} \left[k^{1/r} \mathbb{P}(k-1 < |X_1|^r \leq k) \left(\frac{C}{k^{1/r-1}} \right) \right] \\
&= C \sum_{k=1}^{\infty} k \mathbb{P}(k-1 < |X_1|^r \leq k) \leq C(1 + \mathbb{E}[|X_1|^r]) < \infty,
\end{aligned}$$

where $C > 0$ is a constant.

Case 3: $1 < r < 2$ **with** $a = \mathbb{E}[X_1]$**.** By Corollary 2.2.j on Y_n's with $a_n = n^{1/r}$ for all $n \in \mathbb{N}$, it suffices to show that $\sum_{n=1}^{\infty} \mathbb{E}[|Y_n|^r]/(n^{1/r})^r < \infty$.

We proceed as follows:

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{\mathbb{E}[|Y_n|^r]}{(n^{1/r})^r} &\le \sum_{n=1}^{\infty}\frac{\mathbb{E}\big[Y_n^2\big]}{(n^{1/r})^2} = \sum_{n=1}^{\infty}\frac{1}{n^{2/r}}\mathbb{E}\Big[X_1^2\mathbf{1}_{\{|X_1|\le n^{1/r}\}}\Big] = \sum_{n=1}^{\infty}\frac{1}{n^{2/r}}\mathbb{E}\big[X_1^2\mathbf{1}_{\{|X_1|^r\le n\}}\big]\\ &= \sum_{n=1}^{\infty}\sum_{k=1}^{n}\frac{1}{n^{2/r}}\mathbb{E}\big[X_1^2\mathbf{1}_{\{k-1<|X_1|^r\le k\}}\big] = \sum_{k=1}^{\infty}\sum_{n=k}^{\infty}\frac{1}{n^{2/r}}\mathbb{E}\big[X_1^2\mathbf{1}_{\{k-1<|X_1|^r\le k\}}\big]\\ &= \sum_{k=1}^{\infty}\left[\mathbb{E}\big[X_1^2\mathbf{1}_{\{k-1<|X_1|^r\le k\}}\big]\left(\sum_{n=k}^{\infty}\frac{1}{n^{2/r}}\right)\right] \le \sum_{k=1}^{\infty}\left[k^{2/r}\mathbb{E}\big[\mathbf{1}_{\{k-1<|X_1|^r\le k\}}\big]\left(\frac{C}{k^{2/r}-1}\right)\right]\\ &= C\sum_{k=1}^{\infty}k\mathbb{P}(k-1<|X_1|^r\le k) \le C(1+\mathbb{E}[|X_1|^r] < \infty,\end{aligned}$$

where $C>0$ is a constant. □

2.3 Central Limit Theorems

2.3.1 An important type of convergence behaviour of great interest in statistics is *asymptotic normality* (convergence to a normal random variable in distribution), since it forms the theoretical basis for many statistical methodologies. To establish asymptotic normality, various types of *central limit theorems* (CLTs) are used, and we will study them here.

2.3.2 **Classical CLT.** Let us start with reviewing the classical CLT, which is well-known and should be familiar to you already:

Theorem 2.3.a (Classical CLT)**.** Let $X_1,\dots,X_n$ be iid random variables with $\mathbb{E}[X_1]=0$ and $\sigma:=\sqrt{\mathbb{E}[X_1^2]}<\infty$. Then, $\sqrt{n}\bar{X}_n/\sigma \xrightarrow[n\to\infty]{\mathrm{d}} \mathrm{N}(0,1)$.

[Note: "$\xrightarrow[n\to\infty]{\mathrm{d}}\mathrm{N}(0,1)$" is a shorthand for "$\xrightarrow[n\to\infty]{\mathrm{d}} Z\sim\mathrm{N}(0,1)$". It is common to use this kind of shorthand to simplify the presentation of convergence in distribution.]

Proof. Omitted. □

[Note: Here we focus on the case where the random variables all have zero mean for simplicity. If the random variables have a mean different from zero, we can simply subtract each of them by its mean and apply this result to the modified random variables (with zero mean always).]

To motivate what follows, let us write

$$\frac{\sqrt{n}\bar{X}_n}{\sigma} = \frac{n\bar{X}_n}{\sqrt{n\sigma^2}} = \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n\mathbb{E}[X_i^2]}} = \sum_{i=1}^n X_{n,i}$$

where $X_{n,i}:=X_i/\sqrt{\sum_{i=1}^n\mathbb{E}[X_i^2]}$ for all $n\in\mathbb{N}$ and $i=1,\dots,n$. Geometrically, the $X_{n,i}$'s indeed form a *triangular array*, as illustrated below:

$$\begin{array}{cccccc} X_{1,1} & & & & & \\ X_{2,1} & X_{2,2} & & & & \\ X_{3,1} & X_{3,2} & X_{3,3} & & & \\ X_{4,1} & X_{4,2} & X_{4,3} & X_{4,4} & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \end{array}$$

From this expression, we see that the classical CLT can be viewed as asserting the convergence in distribution for *row sums* of random variables from the triangular array above: $\sum_{i=1}^n X_{n,i}$. The random variables in the array satisfy that (i) $\mathbb{E}[X_{n,i}] = 0$ for all $n \in \mathbb{N}$ and $i = 1, \dots, n$, (ii) $\sum_{i=1}^n \mathbb{E}\big[X_{n,i}^2\big] = 1$ for all $n \in \mathbb{N}$, and (iii) $X_{n,1}, \dots, X_{n,n}$ are independent; in words, these conditions can be expressed as: (i) each random variable in the array has zero mean, (ii) every row sum of second moments is 1, and (iii) the random variables in each row are independent.

This observation leads us to develop more general CLTs for triangular arrays of random variables satisfying the three conditions above. Two notable generalizations among them are *Lindeberg-Feller CLT* and *Lyapunov CLT*, which will be discussed in the following.

2.3.3 Lindeberg-Feller CLT.

Proposition 2.3.b (Lindeberg-Feller CLT for triangular arrays)**.** For all $n \in \mathbb{N}$, let $X_{n,1}, \dots, X_{n,n}$ be independent random variables satisfying that $\mathbb{E}[X_{n,i}] = 0$ for all $i = 1, \dots, n$ and $\sum_{i=1}^n \mathbb{E}\big[X_{n,i}^2\big] = 1$. Then the following are equivalent:

(a) *(Lindeberg condition)* For all $\varepsilon > 0$, $\sum_{i=1}^n \mathbb{E}\big[X_{n,i}^2 \mathbf{1}_{\{|X_{n,i}| \geq \varepsilon\}}\big] \underset{n\to\infty}{\longrightarrow} 0$.

(b) $\max_{1 \leq i \leq n} \mathbb{E}\big[X_{n,i}^2\big] \underset{n\to\infty}{\longrightarrow} 0$ and $\sum_{i=1}^n X_{n,i} \underset{n\to\infty}{\xrightarrow{\mathrm{d}}} \mathrm{N}(0,1)$.

Proof. Omitted. □

Theorem 2.3.c (Lindeberg-Feller CLT)**.** Let $X_1, \dots, X_n$ be independent random variables with $\mathbb{E}[X_i] = 0$ and $0 < \sigma_i^2 := \mathrm{Var}(X_i) < \infty$ for all $i = 1, \dots, n$. Let $S_n := \sum_{i=1}^n X_i$ and $B_n := \sqrt{\sum_{i=1}^n \sigma_i^2}$. Then the following are equivalent:

(a) *(Lindeberg condition)* For all $\varepsilon > 0$, $B_n^{-2} \sum_{i=1}^n \mathbb{E}\big[X_i^2 \mathbf{1}_{\{|X_i| \geq \varepsilon B_n\}}\big] \underset{n\to\infty}{\longrightarrow} 0$.

(b) $\max_{1 \leq i \leq n} \sigma_i^2 / B_n^2 \underset{n\to\infty}{\longrightarrow} 0$ and $S_n / B_n \underset{n\to\infty}{\xrightarrow{\mathrm{d}}} \mathrm{N}(0,1)$.

Proof. Take $X_{n,i} = X_i / B_n$ in Proposition 2.3.b. □

2.3.4 Lyapunov CLT.

An alternative to Lindeberg-Feller CLT is *Lyapunov CLT*. While it is indeed just a simple consequence of Lindeberg-Feller CLT, it offers a sufficient condition for asymptotic normality which is often easier to check than the Lindeberg condition, since it does not involve indicator function in the expectation, but as a tradeoff, it involves a slightly higher moment.

Proposition 2.3.d (Lyapunov CLT for triangular arrays)**.** For all $n \in \mathbb{N}$, let $X_{n,1}, \dots, X_{n,n}$ be independent random variables satisfying that $\mathbb{E}[X_{n,i}] = 0$ for all $i = 1, \dots, n$ and $\sum_{i=1}^n \mathbb{E}\big[X_{n,i}^2\big] = 1$. If $\sum_{i=1}^n \mathbb{E}\big[|X_{n,i}|^{2+\delta}\big] \underset{n\to\infty}{\longrightarrow} 0$ for some $\delta > 0$ *(Lyapunov condition)*, then $\sum_{i=1}^n X_{n,i} \underset{n\to\infty}{\xrightarrow{\mathrm{d}}} \mathrm{N}(0,1)$.

Proof. Under the Lyapunov condition, the Lindeberg condition also holds since

$$
\begin{aligned}
0 \leq \sum_{i=1}^n \mathbb{E}\big[X_{n,i}^2 \mathbf{1}_{\{|X_{n,i}| \geq \varepsilon\}}\big] &= \sum_{i=1}^n \mathbb{E}\big[|X_{n,i}|^{2+\delta} |X_{n,i}|^{-\delta} \mathbf{1}_{\{|X_{n,i}| \geq \varepsilon\}}\big] \\
&\leq \frac{1}{\varepsilon^\delta} \sum_{i=1}^n \mathbb{E}\big[|X_{n,i}|^{2+\delta} \mathbf{1}_{\{|X_{n,i}| \geq \varepsilon\}}\big] \leq \frac{1}{\varepsilon^\delta} \underbrace{\sum_{i=1}^n \mathbb{E}\big[|X_{n,i}|^{2+\delta}\big]}_{\underset{n\to\infty}{\longrightarrow} 0 \text{ (Lyapunov condition)}} \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}
$$

□

Theorem 2.3.e (Lyapunov CLT)**.** Let $X_1, \dots, X_n$ be independent random variables with $\mathbb{E}[X_i] = 0$ and $0 < \sigma_i^2 := \mathrm{Var}(X_i) < \infty$ for all $i = 1, \dots, n$. Let $S_n := \sum_{i=1}^n X_i$ and $B_n := \sqrt{\sum_{i=1}^n \sigma_i^2}$. If $B_n^{-(2+\delta)} \sum_{i=1}^n \mathbb{E}\big[|X_i|^{2+\delta}\big] \underset{n\to\infty}{\longrightarrow} 0$ for some $\delta > 0$ *(Lyapunov condition)*, then $S_n / B_n \underset{n\to\infty}{\xrightarrow{\mathrm{d}}} \mathrm{N}(0,1)$.

Proof. Take $X_{n,i} = X_i/B_n$ in Proposition 2.3.d. □

2.3.5 **An alternative form of Lindeberg condition.** Let $\Lambda_n(\varepsilon) := B_n^{-2}\sum_{i=1}^n \mathbb{E}\big[X_i^2\mathbf{1}_{\{|X_i|\geq\varepsilon B_n\}}\big]$ and $\Lambda_n^*(\varepsilon) := B_n^{-2}\sum_{i=1}^n \mathbb{E}\big[X_i^2\mathbf{1}_{\{|X_i|\geq\varepsilon B_i\}}\big]$, for all $n\in\mathbb{N}$ and $\varepsilon>0$. The Lindeberg condition in Theorem 2.3.c, namely $\Lambda_n(\varepsilon)\underset{n\to\infty}{\longrightarrow}0$ for all $\varepsilon>0$, holds iff we have $\Lambda_n^*(\varepsilon)\underset{n\to\infty}{\longrightarrow}0$ for all $\varepsilon>0$.

Proof. "⇐": Assume $\Lambda_n^*(\varepsilon)\underset{n\to\infty}{\longrightarrow}0$ for all $\varepsilon>0$. Then, since $B_i\leq B_n$ for all $i=1,\dots,n$, we have

$$0\leq\Lambda_n(\varepsilon)\leq\Lambda_n^*(\varepsilon)\underset{n\to\infty}{\longrightarrow}0,$$

for all $n\in\mathbb{N}$ and $\varepsilon>0$, which implies the Lindeberg condition.

"⇒": Assume $\Lambda_n(\varepsilon)\underset{n\to\infty}{\longrightarrow}0$. Then, for all $\varepsilon>0$ and $\delta>0$, we have

$$\begin{aligned}
0\leq\Lambda_n^*(\varepsilon) &= B_n^{-2}\sum_{i=1}^n\mathbb{E}\big[X_i^2\mathbf{1}_{\{|X_i|\geq\varepsilon B_i\}}\big]\\
&= B_n^{-2}\left(\sum_{i:B_i\leq\delta B_n}\mathbb{E}\big[X_i^2\mathbf{1}_{\{|X_i|\geq\varepsilon B_i\}}\big]+\sum_{i:B_i>\delta B_n}\mathbb{E}\big[X_i^2\mathbf{1}_{\{|X_i|\geq\varepsilon B_i\}}\big]\right)\\
&\leq B_n^{-2}\left(\sum_{i:B_i\leq\delta B_n}\mathbb{E}\big[X_i^2\big]+\sum_{i:B_i>\delta B_n}\mathbb{E}\big[X_i^2\mathbf{1}_{\{|X_i|\geq\varepsilon B_n\}}\big]\right)\\
&\leq B_n^{-2}\left(\sum_{i:B_i\leq\delta B_n}\mathbb{E}\big[X_i^2\big]+\sum_{i=1}^n\mathbb{E}\big[X_i^2\mathbf{1}_{\{|X_i|\geq\varepsilon B_n\}}\big]\right)\\
&\leq B_n^{-2}\delta^2B_n^2+\Lambda_n(\varepsilon)\quad(\text{we have } B_1^2\leq\cdots\leq B_n^2)\\
&= \delta^2+\Lambda_n(\varepsilon)\underset{n\to\infty}{\longrightarrow}\delta^2
\end{aligned}$$

for all $n\in\mathbb{N}$. Letting $n\to\infty$ and then $\delta\to0^+$ yields $\Lambda_n^*(\varepsilon)\underset{n\to\infty}{\longrightarrow}0$ for all $\varepsilon>0$, as desired. □

2.3.6 **Polya's theorem.** Polya's theorem suggests that convergence to a random variable with *continuous CDF* in distribution is indeed equivalent to uniform convergence. In particular, this applies to various types of asymptotic normality established above, and all those results actually assert uniform convergence also.

Theorem 2.3.f (Polya's theorem)**.** Let $X\sim F$ and $X_n\sim F_n$ for all $n\in\mathbb{N}$. Suppose that F is continuous. Then, $X_n\xrightarrow[n\to\infty]{\mathrm{d}}X$ iff $\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|\underset{n\to\infty}{\longrightarrow}0$ *($\{F_n\}$ converges uniformly to F).*

Proof. Omitted. □

3 Fundamentals of Statistics

3.0.1 After learning probabilistic concepts that form the foundation of the study of statistics, starting from Section 3 we will have discussions on theories for statistical methods. Let us start by covering some fundamental concepts of statistics (that may already be familiar to you), which pave the way for the study of further statistical topics.

3.1 Probability Models

3.1.1 **Three types of probability models.** In our context here, a **probability model** is viewed as a *family of cumulative distribution functions (CDFs)*. Depending on the nature of the CDFs in the family, we can classify probability models into the following three types:

(a) *Parametric:* A **parametric model** is a family $\mathcal{F} = \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ of CDFs indexed by *parameters* $\boldsymbol{\theta} \in \Theta$, where $\Theta \subseteq \mathbb{R}^p$ is the **parameter space**, with $p \in \mathbb{N}$ being fixed.

(b) *Non-parametric:* A **non-parametric model** is a probability model that is not a parametric model.

(c) *Semi-parametric:* A **semi-parametric model** is a mixture of parametric and non-parametric models.

3.1.2 **Identifiability of parametric models.** Under a parametric model $\mathcal{F} = \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, a central goal of statistics is to *infer* the unknown parameter $\boldsymbol{\theta}$ based on an iid sample $X_1, \ldots, X_n$ distributed according to a certain CDF $F_{\boldsymbol{\theta}}$ in the parametric model. An important underlying assumption behind this kind of statistical inference is that *knowing the CDF $F_{\boldsymbol{\theta}}$ amounts to knowing the parameter $\boldsymbol{\theta}$*, or more formally, the parametric model is *identifiable*.

A parametric model $\mathcal{F} = \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is said to be **identifiable** if $F_{\boldsymbol{\theta}} = F_{\widetilde{\boldsymbol{\theta}}} \implies \boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$, or equivalently, $\boldsymbol{\theta} \neq \widetilde{\boldsymbol{\theta}} \implies F_{\boldsymbol{\theta}} \neq F_{\widetilde{\boldsymbol{\theta}}}$. Otherwise (i.e., $F_{\boldsymbol{\theta}} = F_{\widetilde{\boldsymbol{\theta}}}$ for some $\boldsymbol{\theta} \neq \widetilde{\boldsymbol{\theta}}$), the parametric model $\mathcal{F}$ is said to be **unidentifiable**.

Remarks:

- Alternatively, we can replace the CDF $F_{\boldsymbol{\theta}}$ by the respective PDF/PMF $f_{\boldsymbol{\theta}}$ in the definition, where two f's are considered equal as long as they are equal almost everywhere (at all points except possibly on a set with zero measure).
- It is often (much) easier to *disprove* identifiability than proving it, since we just need to find a counterexample for the disproof.

3.1.3 **Examples about identifiability.**

(a) The family of $\mathrm{N}(\mu, \sigma^2)$ CDFs with $\boldsymbol{\theta} = (\mu, \sigma^2)$ is identifiable.

(b) The family of $\mathrm{N}(\mu_1 + \mu_2, \sigma^2)$ CDFs with $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma^2)$ is unidentifiable.

Proof.

(a) Assume $f_{\boldsymbol{\theta}}(x) = f_{\widetilde{\boldsymbol{\theta}}}(x)$ for almost all $x \in \mathbb{R}$, with $\widetilde{\boldsymbol{\theta}} = (\widetilde{\mu}, \widetilde{\sigma}^2)$. Then,

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{(x-\mu)^2/2\sigma^2} = \frac{1}{\sqrt{2\pi\widetilde{\sigma}^2}} e^{(x-\widetilde{\mu})^2/2\widetilde{\sigma}^2} \quad \text{for almost all } x \in \mathbb{R}.$$

Taking natural logarithm on both sides gives

$$-\frac{1}{2\sigma^2}(x-\mu)^2 - \frac{1}{2}\ln(2\pi\sigma^2) = -\frac{1}{2\widetilde{\sigma}^2}(x-\widetilde{\mu})^2 - \frac{1}{2}\ln(2\pi\widetilde{\sigma}^2) \quad \text{for almost all } x \in \mathbb{R}.$$

By equating coefficients, we get $(\mu, \sigma^2) = (\widetilde{\mu}, \widetilde{\sigma}^2)$.

(b) Take $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma^2) = (1, 1, 1)$ and $\widetilde{\boldsymbol{\theta}} = (\widetilde{\mu}_1, \widetilde{\mu}_2, \widetilde{\sigma}^2) = (0, 2, 1)$. Then, we can see that $f_{\boldsymbol{\theta}} = f_{\widetilde{\boldsymbol{\theta}}}$ while $\boldsymbol{\theta} \neq \widetilde{\boldsymbol{\theta}}$.

□

3.1.4 In STAT7609, we focus on two special and important *parametric* models: *exponential family* and *location-scale family*.

3.2 Exponential Family

3.2.1 **Definition.** A parametric model $\mathcal{F}$ is called a **k-parameter exponential family**, denoted by $\text{Exp}(k)$, if for each $\boldsymbol{\theta} \in \Theta$, the associated probability density/mass function (PDF/PMF) can be expressed as

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = e^{\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{T}(\boldsymbol{x}) - B(\boldsymbol{\theta})} h(\boldsymbol{x}),$$

where $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \ldots, \eta_k(\boldsymbol{\theta}))$ and $\boldsymbol{T}(\boldsymbol{x}) = (T_1(\boldsymbol{x}), \ldots, T_k(\boldsymbol{x}))$, with $B(\boldsymbol{\theta})$ depending on $\boldsymbol{\theta}$ (possibly) but not $\boldsymbol{x}$, and $h(\boldsymbol{x})$ depending on $\boldsymbol{x}$ (possibly) but not $\boldsymbol{\theta}$.

[Note: The pdf/pmf of many well-known distributions can be expressed in this form, e.g., normal, gamma, Binomial, Poisson, geometric, etc.]

For f to be a valid PDF/PMF, we must have $\int f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 1$; throughout we will just use "$\int \cdot\, \mathrm{d}\boldsymbol{x}$" to stand for integrating/summing over the whole support for convenience. Hence, we can deduce that

$$B(\boldsymbol{\theta}) = \ln\left(\int e^{\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{T}(\boldsymbol{x})} h(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \right),$$

which serves as a normalization factor and is determined by other quantities in the expression.

3.2.2 **Canonical form.** Treating $\boldsymbol{\eta} := \boldsymbol{\eta}(\boldsymbol{\theta})$ as the parameter (known as the **natural parameter**), we can obtain the *canonical form* of the exponential family. Let

$$\Xi = \left\{ \boldsymbol{\eta} \in \mathbb{R}^k : \int e^{\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{T}(\boldsymbol{x})} h(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} < \infty \right\}$$

denote the **natural parameter space**, which contains every possible natural parameter (the condition ensures that the PDF/PMF f can be normalized to become valid). Then, the **canonical form** of exponential family is given by

$$f_{\boldsymbol{\eta}}(\boldsymbol{x}) = e^{\boldsymbol{\eta}^T \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\eta})} h(\boldsymbol{x})$$

for each $\boldsymbol{\eta} \in \Xi$. [Note: A nice feature about the canonical form of exponential family is that the natural parameter space Ξ is *convex* (while the original parameter space Θ is not necessarily convex).]

3.2.3 **Full rank.** An exponential family in the canonical form is said to be of **full rank** if:

(a) *(Minimality)* Both $\eta_1, \ldots, \eta_k$ and $T_1, \ldots, T_k$ are linearly independent, i.e., none of the η's (T's) can be expressed as a linear combination of the other η's (T's).

(b) *(Containing rectangle)* The natural parameter space Ξ contains a k-dimensional rectangle.

[Note: It is possible that Ξ does not contain any k-dimensional rectangle. For example, if Ξ is a curve in $\mathbb{R}^k$, then it contains no k-dimensional rectangle.]

3.2.4 **Closed operations for exponential family.** For convenience, for a random variable/vector X, we write $X \in \text{Exp}(k)$ if its CDF F_X belongs to a k-parameter exponential family.

(a) If $X_1, \ldots, X_n \in \text{Exp}(k)$ are iid, then $(X_1, \ldots, X_n) \in \text{Exp}(k)$.

(b) If $\boldsymbol{X} \in \text{Exp}(k)$, then $\boldsymbol{T}(\boldsymbol{X}) \in \text{Exp}(k)$, where $\boldsymbol{T}$ is found in the expression for the exponential family.

Proof.

(a) By the iid property, the PDF/PMF of $(X_1, \ldots, X_n)$ is

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \prod_{i=1}^{n} e^{\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{T}(\boldsymbol{x}_i) - B(\boldsymbol{\theta})} h(\boldsymbol{x}_i), = e^{\boldsymbol{\eta}(\boldsymbol{\theta})^T \left[\sum_{i=1}^{n} \boldsymbol{T}(\boldsymbol{x}_i)\right] - nB(\boldsymbol{\theta})} \prod_{i=1}^{n} h(\boldsymbol{x}_i),$$

which is in the form as specified by the exponential family.

(b) We only prove for the case where $\boldsymbol{X}$ is a discrete. For all $\boldsymbol{t} = (t_1, \ldots, t_k)$, we have

$$\begin{aligned}\mathbb{P}(\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{t}) &= \sum_{\boldsymbol{x}: T(\boldsymbol{x}) = \boldsymbol{t}} \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = \sum_{\boldsymbol{x}: T(\boldsymbol{x}) = \boldsymbol{t}} e^{\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{T}(\boldsymbol{x}) - B(\boldsymbol{\theta})} h(\boldsymbol{x}) \\ &= \sum_{\boldsymbol{x}: T(\boldsymbol{x}) = \boldsymbol{t}} e^{\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{t} - B(\boldsymbol{\theta})} h(\boldsymbol{x}) = e^{\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{t}} h_0(\boldsymbol{t})\end{aligned}$$

where $h_0(\boldsymbol{t}) = \sum_{\boldsymbol{x}: T(\boldsymbol{x}) = \boldsymbol{t}} h(\boldsymbol{x})$.

□

3.2.5 **Calculating moments for exponential family.** Due to the special structure of exponential family, we have a specialized formula for computing moments for exponential family, in the canonical form.

Proposition 3.2.a. Let $\boldsymbol{X} \in \text{Exp}(k)$, with the PDF/PMF given by $f_\eta(\boldsymbol{x}) = e^{\eta T(\boldsymbol{x}) - A(\eta)} h(\boldsymbol{x})$. Suppose that η is an interior point of Ξ. Then, the moment generating function and cumulant generating function of $T(\boldsymbol{X})$ exist and are given by

$$M(s) = e^{A(s+\eta) - A(\eta)} \quad \text{and} \quad C(s) \overset{\text{(definition)}}{=} \ln M(s) = A(s+\eta) - A(\eta),$$

for all s in a neighbourhood of 0. Particularly, we have $\mathbb{E}[T(\boldsymbol{X})] = A'(\eta)$ and $\text{Var}(T(\boldsymbol{X})) = A''(\eta)$.

Proof. For all s in a sufficiently small neighbourhood of 0 such that $s + \eta \in \Xi$ still (such a neighbourhood exists since η is an interior point), the moment generating function is

$$\begin{aligned}M(s) &= \mathbb{E}\left[e^{sT(\boldsymbol{X})}\right] = \int e^{(s+\eta)T(\boldsymbol{x}) - A(\eta)} h(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\ &= e^{A(s+\eta) - A(\eta)} \int \underbrace{e^{(s+\eta)T(\boldsymbol{x}) - A(s+\eta)} h(\boldsymbol{x})}_{f_{s+\eta}(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} = e^{A(s+\eta) - A(\eta)},\end{aligned}$$

and thus the cumulant generating function is $C(s) = \ln M(s) = A(s+\eta) - A(\eta)$.

It follows that

$$\mathbb{E}[T(\boldsymbol{X})] = C'(0) = A'(\eta) \quad \text{and} \quad \text{Var}(T(\boldsymbol{X})) = C''(0) = A''(\eta).$$

□

3.3 Location-Scale Family

3.3.1 **Definition.** Let $F(x)$ be a CDF of a random variable X.

(a) The **location family** is the parametric model given by $\mathcal{F} = \{F(x - a) : a \in \mathbb{R}\}$ (i.e., the family of CDFs of $X + a$'s).

(b) The **scale family** is the parametric model given by $\mathcal{F} = \{F(x/b) : b > 0\}$ (i.e., the family of CDFs of bX's).

(c) The **location-scale family** is the parametric model given by $\mathcal{F} = \{F((x-a)/b) : a \in \mathbb{R}, b > 0\}$ (i.e., the family of CDFs of $a + bX$'s).

Here, a is called the **location parameter** and b is called the **scale parameter**.

For the continuous case and assuming the CDF is differentiable, the PDFs in location, scale, and location-scale families take the following expressions (f denotes the PDF of X):

(a) *Location family:* $f(x-a)$, where $a \in \mathbb{R}$.

(b) *Scale family:* $(1/b)f(x/b)$, where $b > 0$.

(c) *Location-scale family:* $(1/b)f((x-a)/b)$, where $a \in \mathbb{R}$ and $b > 0$.

3.3.2 **Examples.** While the location and scale parameters are sometimes the mean and standard deviation, it is not necessarily the case.

(a) The family of $\mathrm{N}(\mu, \sigma^2)$ CDFs is a location-scale family, with μ being the location parameter and σ being the scale parameter.

(b) The family of Cauchy(x_0, γ) CDFs is a location-scale family with x_0 being the location parameter and γ being the scale parameter; here, neither mean nor standard deviation exists.

3.4 Cramér-Rao Lower Bound

3.4.1 After having some explorations of parametric models, our next step is to analyze the statistical inference on the unknown parameter $\boldsymbol{\theta}$, or some function of $\boldsymbol{\theta}$. The procedure of statistical inference often goes as follows. We first fix a parametric model $\mathcal{F} = \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ in our consideration. Given a sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_{\boldsymbol{\theta}}$, (a function of) the unknown parameter is estimated by a (point) *estimator* $\boldsymbol{\delta} := \boldsymbol{\delta}(\boldsymbol{X})$, which is a function of $\boldsymbol{X} := \boldsymbol{X}_n := (X_1, \ldots, X_n)$ (and cannot depend on $\boldsymbol{\theta}$, which is what we do not know!).

Hence, the analysis of the statistical inference boils down to evaluations of the "quality" of the estimator $\boldsymbol{\delta}(\boldsymbol{X})$ constructed.

3.4.2 **Basic properties of estimator.**

- $\boldsymbol{\delta}$ is **unbiased** for $\boldsymbol{\theta}$ if we have $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\delta}(\boldsymbol{X})] = \boldsymbol{\theta}$ (componentwise) for all $\boldsymbol{\theta}$. Expressing differently, $\boldsymbol{\delta}(\boldsymbol{X})$ is unbiased for $\boldsymbol{\theta}$ if its **bias**, given by $\mathrm{Bias}_{\boldsymbol{\theta}}(\boldsymbol{\delta}(\boldsymbol{X})) := \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\delta}(\boldsymbol{X})] - \boldsymbol{\theta}$, equals $\mathbf{0}$.
  Also, $\boldsymbol{\delta}$ is **asymptotically unbiased** for $\boldsymbol{\theta}$ if $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\delta}(\boldsymbol{X}_n)] \underset{n\to\infty}{\longrightarrow} \boldsymbol{\theta}$ for all $\boldsymbol{\theta}$.
  [Note: The notation $\mathbb{E}_{\boldsymbol{\theta}}[\cdot]$ emphasizes that the expectation is taken under the parameter $\boldsymbol{\theta}$.]
- $\boldsymbol{\delta}$ is **(weakly) consistent** for $\boldsymbol{\theta}$ if $\boldsymbol{\delta}(\boldsymbol{X}_n) \underset{n\to\infty}{\xrightarrow{\text{p}}} \boldsymbol{\theta}$ for all $\boldsymbol{\theta}$, i.e., for all $\boldsymbol{\theta}$ and $\varepsilon > 0$, we have $\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}}(\|\boldsymbol{\delta}(\boldsymbol{X}_n) - \boldsymbol{\theta}\| > \varepsilon) = 0$.
- $\boldsymbol{\delta}$ is **strongly consistent** for $\boldsymbol{\theta}$ if $\boldsymbol{\delta}(\boldsymbol{X}_n) \underset{n\to\infty}{\xrightarrow{\text{a.s.}}} \boldsymbol{\theta}$ for all $\boldsymbol{\theta}$.
- $\boldsymbol{\delta}$ is **asymptotically normal** if $\sqrt{n}(\boldsymbol{\delta}(\boldsymbol{X}_n) - \boldsymbol{\theta}) \underset{n\to\infty}{\xrightarrow{\text{d}}} \mathrm{N}_p(\mathbf{0}, \Sigma)$ for all $\boldsymbol{\theta}$.

An estimator $\boldsymbol{\delta}$ is said to be "**CAN**" if it is consistent and asymptotically normal. In the evaluation of estimators, we are often interested in whether the "CAN" property is satisfied or not. Indeed, it can be shown that if an estimator is asymptotically normal, then it must be consistent, meaning that showing the asymptotic normality alone is enough for establishing the "CAN" property. Nevertheless, to show asymptotic normality, the consistency is often helpful, so it is still common to first show consistency *(often easier)*, and then show asymptotic normality *(often harder)*.

3.4.3 **Mean squared errors.** A notable measurement of error for estimators is *mean squared error*. In the case $p = 1$, the **mean squared error** of $\delta(\boldsymbol{X})$ is $\mathrm{MSE}_\theta(\delta(\boldsymbol{X})) := \mathbb{E}_\theta\big[(\delta(\boldsymbol{X}) - \theta)^2\big]$, and the **root mean squared error** of $\delta(\boldsymbol{X})$ is $\mathrm{RMSE}_\theta(\delta(\boldsymbol{X})) := \sqrt{\mathrm{MSE}_\theta(\delta(\boldsymbol{X}))}$, which carries the same unit as $\delta(\boldsymbol{X})$ (like the idea of standard deviation).

The mean squared error is an useful metric for comparing the qualities of estimators. If $\mathrm{MSE}_\theta(\delta_1(\boldsymbol{X})) < \mathrm{MSE}_\theta(\delta_2(\boldsymbol{X}))$, then $\delta_1(\boldsymbol{X})$ is said to be **relatively more efficient** than $\delta_2(\boldsymbol{X})$.

3.4.4 **Uniformly minimum variance unbiased estimator (UMVUE).** With the concept of *relative efficiency* arising from the comparison of mean squared error, one would then naturally be interested in studying the "most" efficient estimator. In such a discussion, the following *bias-variance decomposition* of mean squared error is helpful.

Proposition 3.4.a (Bias-variance decomposition of mean squared error)**.** We have $\mathrm{MSE}_\theta(\delta(\boldsymbol{X})) = \mathrm{Var}_\theta(\delta(\boldsymbol{X})) + \mathrm{Bias}_\theta(\delta(\boldsymbol{X}))^2$.

Proof. The decomposition follows from the following chain of equalities:

$$\begin{aligned}\mathrm{MSE}_\theta(\delta) &= \mathbb{E}_\theta\Big[(\delta-\theta)^2\Big] = \mathbb{E}_\theta\Big[(\delta - \mathbb{E}_\theta[\delta] + \mathbb{E}_\theta[\delta] - \theta)^2\Big] \\ &= \underbrace{\mathbb{E}_\theta\Big[(\delta-\mathbb{E}_\theta[\delta])^2\Big]}_{\mathrm{Var}_\theta(\delta)} + \underbrace{\mathbb{E}_\theta\Big[(\mathbb{E}_\theta[\delta]-\theta)^2\Big]}_{(\mathbb{E}_\theta[\delta]-\theta)^2=\mathrm{Bias}_\theta(\delta)^2} + \underbrace{2\mathbb{E}_\theta[(\delta-\mathbb{E}_\theta[\delta])(\mathbb{E}_\theta[\delta]-\theta)]}_{2(\mathbb{E}_\theta[\delta]-\theta)\mathbb{E}_\theta[(\delta-\mathbb{E}_\theta[\delta])]=0} \\ &= \mathrm{Var}_\theta(\delta) + \mathrm{Bias}_\theta(\delta)^2.\end{aligned}$$

□

This bias-decomposition makes it clear that there are only two sources to mean-squared error, namely *bias* and *variance*. Here, our idea is to focus on only *unbiased estimators* (with zero bias), and compare their mean squared errors or equivalently variances in this case. Expressing this in mathematical terms, if the estimators δ_1 and δ_2 are unbiased, then we have that δ_1 is relatively more efficient than δ_2 iff $\mathrm{Var}_\theta(\delta_1) < \mathrm{Var}_\theta(\delta_2)$. Therefore, the "most" efficient unbiased estimator would be the one with the smallest variance, and it is said to be *uniformly minimum variance unbiased estimator*. To be more precise, δ is a **uniformly minimum variance unbiased estimator (UMVUE)** for θ if $\mathrm{Var}_\theta(\delta) \le \mathrm{Var}_\theta\left(\widetilde{\delta}\right)$ for all θ *(uniformity)* and all unbiased estimators $\widetilde{\delta}$ for θ. [Note: It can be shown that UMVUEs are unique *almost surely*, i.e., if both δ_1 and δ_2 are UMVUEs for θ, then $\mathbb{P}(\delta_1 = \delta_2) = 1$.]

3.4.5 **Cramér-Rao lower bound.** A common approach for obtaining UMVUEs is via the *Cramér-Rao lower bound*. It is based on the following result, which provides a lower bound on the variance of *every* unbiased estimator, under some regularity conditions; such a lower bound is called the **Cramér-Rao lower bound (CRLB)**, and an unbiased estimator achieving CRLB must be a UMVUE (under the regularity conditions), but not vice versa.

Theorem 3.4.b (Cramér-Rao lower bound)**.** Let $\boldsymbol{X} = (X_1, \ldots, X_n) \sim f_{\boldsymbol{\theta}}$, $g(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\delta(\boldsymbol{X})]$, and $\ell(\boldsymbol{\theta}) := \ln f_{\boldsymbol{\theta}}(\boldsymbol{X})$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. Assume that δ is scalar-valued (so g is scalar-valued also), ℓ and g are of class C^1, the **Fisher information (matrix)** $I_{\boldsymbol{X}}(\boldsymbol{\theta}) := \mathbb{E}\big[(\partial\ell/\partial\boldsymbol{\theta})(\partial\ell/\partial\boldsymbol{\theta})^T\big]$ is invertible for all $\boldsymbol{\theta}$, and the following regularity conditions hold:

(a) $\dfrac{\partial}{\partial\boldsymbol{\theta}}\displaystyle\int f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int \frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ for all $\boldsymbol{\theta}$.

(b) $\dfrac{\partial}{\partial\boldsymbol{\theta}}\displaystyle\int \delta(\boldsymbol{x}) f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int \delta(\boldsymbol{x})\frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ for all $\boldsymbol{\theta}$.

Then, we have

$$\mathrm{Var}(\delta(\boldsymbol{X})) \ge \left(\frac{\partial g}{\partial\boldsymbol{\theta}}\right)^T (I_{\boldsymbol{X}}(\boldsymbol{\theta}))^{-1}\frac{\partial g}{\partial\boldsymbol{\theta}}$$

and the equality holds *(CRLB is attained)* iff $\delta(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{\theta}}[\delta(\boldsymbol{X})] = \delta(\boldsymbol{X}) - g(\boldsymbol{\theta}) \overset{\text{a.s.}}{=} \boldsymbol{\lambda}_0^T\,\partial\ell/\partial\boldsymbol{\theta}$ for some nonrandom $\boldsymbol{\lambda}_0 \in \mathbb{R}^p$ *(a.s. linear dependence between δ and $\partial\ell/\partial\boldsymbol{\theta}$)*.

Remarks:

- The nonrandom $\boldsymbol{\lambda}_0 \in \mathbb{R}^p$ may possibly depend on $\boldsymbol{\theta}$, but $\delta(\boldsymbol{X})$ cannot possibly depend on $\boldsymbol{\theta}$! To check whether the CRLB can be attained, we often examine whether the "form" of $\partial\ell/\partial\boldsymbol{\theta}$ is somewhat "similar to" $\delta(\boldsymbol{X}) - g(\boldsymbol{\theta})$.

- For a scalar-valued function h, we have

$$\frac{\partial h}{\partial \boldsymbol{\theta}} := \begin{bmatrix} \partial h/\partial \theta_1 \\ \vdots \\ \partial h/\partial \theta_p \end{bmatrix},$$

  with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ being a column vector.
- When the "integrand" in $\int$ is a vector (or matrix in general), "$\int$" (referring to integrating/summing over the support) is taken in the componentwise sense.
- A scalar-valued function is said to be of **class** C^k if all its partial derivatives of at most k exist and are continuous, and a vector-valued function is of **class** C^k if each of its component functions is of class C^k. In STAT7609, we often deal with C^∞ function (or *smooth* function), meaning that all its partial derivatives of any order exist and are continuous.

Proof. For convenience, let $\boldsymbol{Y} = (Y_1, \ldots, Y_p) := \partial \ell/\partial \boldsymbol{\theta}$. Note that

$$\begin{aligned}
\mathbb{E}[\boldsymbol{Y}] &= \int \left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln f_{\boldsymbol{\theta}}(\boldsymbol{x}) \right) f_{\boldsymbol{\theta}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\
&= \int \left(\frac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \right) \frac{1}{f_{\boldsymbol{\theta}}(\boldsymbol{x})} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int \frac{\partial}{\partial \boldsymbol{\theta}} \ln f_{\boldsymbol{\theta}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \overset{\text{(a)}}{=} \frac{\partial}{\partial \boldsymbol{\theta}} \int f_{\boldsymbol{\theta}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \frac{\partial}{\partial \boldsymbol{\theta}} 1 = \mathbf{0}.
\end{aligned}$$

Hence, the Fisher information matrix can also be expressed as $\operatorname{Var}(\boldsymbol{Y})$.

Claim: We have

$$\operatorname{Var}(\delta) \geq \operatorname{Cov}(\delta, \boldsymbol{Y}) \operatorname{Var}(\boldsymbol{Y})^{-1} \operatorname{Cov}(\delta, \boldsymbol{Y})^T$$

and the equality holds iff δ and $\boldsymbol{Y}$ are linearly dependent (i.e., $\delta - \boldsymbol{\lambda}_0^T \boldsymbol{Y} = c$ for some $c \in \mathbb{R}$ and $\boldsymbol{\lambda}_0 \in \mathbb{R}^p$) a.s. Here, $\operatorname{Cov}(\delta, \boldsymbol{Y}) = \begin{bmatrix} \operatorname{Cov}(\delta, Y_1) & \cdots & \operatorname{Cov}(\delta, Y_p) \end{bmatrix}$ is a row vector and $\operatorname{Var}(\boldsymbol{Y}) = [\operatorname{Cov}(Y_i, Y_j)]_{i,j=1}^{p,p}$ is a matrix.

Proof. For all $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p$, we have

$$\begin{aligned}
0 \leq \operatorname{Var}\left(\delta - \boldsymbol{\lambda}^T \boldsymbol{Y}\right) &= \operatorname{Cov}\left(\delta - \boldsymbol{\lambda}^T \boldsymbol{Y}, \delta - \boldsymbol{\lambda}^T \boldsymbol{Y}\right) \\
&= \operatorname{Var}(\delta) - 2 \operatorname{Cov}\left(\delta, \boldsymbol{\lambda}^T \boldsymbol{Y}\right) + \operatorname{Cov}\left(\boldsymbol{\lambda}^T \boldsymbol{Y}, \boldsymbol{\lambda}^T \boldsymbol{Y}\right) \\
&= \operatorname{Var}(\delta) - 2 \operatorname{Cov}(\delta, \boldsymbol{Y}) \boldsymbol{\lambda} + \boldsymbol{\lambda}^T \operatorname{Var}(\boldsymbol{Y}) \boldsymbol{\lambda} =: g(\boldsymbol{\lambda}).
\end{aligned}$$

Setting $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0 := (\operatorname{Var}(\boldsymbol{Y}))^{-1} \operatorname{Cov}(\delta, \boldsymbol{Y})^T$ gives

$$\begin{aligned}
0 \leq g(\boldsymbol{\lambda}_0) &= \operatorname{Var}(\delta) - 2 \operatorname{Cov}(\delta, \boldsymbol{Y}) \boldsymbol{\lambda}_0 + \boldsymbol{\lambda}_0^T \operatorname{Var}(\boldsymbol{Y}) \boldsymbol{\lambda}_0 \\
&= \operatorname{Var}(\delta) - 2 \operatorname{Cov}(\delta, \boldsymbol{Y}) \operatorname{Var}(\boldsymbol{Y})^{-1} \operatorname{Cov}(\delta, \boldsymbol{Y})^T + \operatorname{Cov}(\delta, \boldsymbol{Y}) \operatorname{Var}(\boldsymbol{Y})^{-1} \operatorname{Cov}(\delta, \boldsymbol{Y})^T \\
&= \operatorname{Var}(\delta) - \operatorname{Cov}(\delta, \boldsymbol{Y}) \operatorname{Var}(\boldsymbol{Y})^{-1} \operatorname{Cov}(\delta, \boldsymbol{Y})^T,
\end{aligned}$$

or

$$\operatorname{Var}(\delta) \geq \operatorname{Cov}(\delta, \boldsymbol{Y}) \operatorname{Var}(\boldsymbol{Y})^{-1} \operatorname{Cov}(\delta, \boldsymbol{Y})^T.$$

The equality holds iff $\operatorname{Var}\left(\delta - \boldsymbol{\lambda}_0^T \boldsymbol{Y}\right) = g(\boldsymbol{\lambda}_0) = 0$ iff $\delta - \boldsymbol{\lambda}_0^T \boldsymbol{Y} = c$ a.s., where $c \in \mathbb{R}$ and $\boldsymbol{\lambda}_0 \in \mathbb{R}^p$. □

Note that

$$\begin{aligned}
\frac{\partial g}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}}[\delta(\boldsymbol{X})] = \frac{\partial}{\partial \boldsymbol{\theta}} \int \delta(\boldsymbol{x}) f_{\boldsymbol{\theta}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\
&\overset{\text{(b)}}{=} \int \delta(\boldsymbol{x}) \frac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int \delta(\boldsymbol{x}) \left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln f_{\boldsymbol{\theta}}(\boldsymbol{x}) \right) f_{\boldsymbol{\theta}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \mathbb{E}[\delta \boldsymbol{Y}] \overset{(\mathbb{E}[\boldsymbol{Y}] = \mathbf{0})}{=} \operatorname{Cov}(\delta, \boldsymbol{Y}).
\end{aligned}$$

Thus, by the claim we have

$$\operatorname{Var}(\delta) \geq \left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)^T (I_{\boldsymbol{X}}(\boldsymbol{\theta}))^{-1} \frac{\partial g}{\partial \boldsymbol{\theta}}$$

and the equality holds iff $\delta - \boldsymbol{\lambda}_0^T(\partial \ell/\partial \boldsymbol{\theta}) \overset{\text{a.s.}}{=} c$ where $\boldsymbol{\lambda}_0 \in \mathbb{R}^p$ and $c \in \mathbb{R}$. Taking expectation on this equation, we have $\mathbb{E}[\delta] - 0 = c$, implying that $c = \mathbb{E}[\delta]$. Hence, this equation is equivalent to $\delta - \mathbb{E}[\delta] \overset{\text{a.s.}}{=} \boldsymbol{\lambda}_0^T(\partial \ell/\partial \boldsymbol{\theta})$ for some $\boldsymbol{\lambda}_0 \in \mathbb{R}^p$. □

[⚠ Warning: Some regularity conditions are often *violated* when the support of $f_{\boldsymbol{\theta}}$ depends on the parameter $\boldsymbol{\theta}$, e.g., for uniform distribution.]

3.4.6 **Properties of Fisher information matrix.** To compute the CRLB in Theorem 3.4.b, we need to invert the Fisher information matrix $I_{\boldsymbol{X}}(\boldsymbol{\theta})$. To find out what $I_{\boldsymbol{X}}(\boldsymbol{\theta})$ is (efficiently), the following properties can be helpful.

Assume that the regularity condition (a) in Theorem 3.4.b holds, and ℓ is of class C^1, where $\ell(\boldsymbol{\theta}) := \ln f_{\boldsymbol{\theta}}(\boldsymbol{X})$.

(a) If $\boldsymbol{X} \sim f_{\boldsymbol{\theta}}$ and $\boldsymbol{Y} \sim g_{\boldsymbol{\theta}}$ with $\boldsymbol{X}$ and $\boldsymbol{Y}$ being independent, then $I_{(\boldsymbol{X},\boldsymbol{Y})}(\boldsymbol{\theta}) = I_{\boldsymbol{X}}(\boldsymbol{\theta}) + I_{\boldsymbol{Y}}(\boldsymbol{\theta})$.

(b) If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_{\boldsymbol{\theta}}$, then $I_{\boldsymbol{X}}(\boldsymbol{\theta}) = \sum_{i=1}^n I_{X_i}(\boldsymbol{\theta}) = nI_{X_1}(\boldsymbol{\theta})$.

(c) ⭐ Suppose that $\boldsymbol{X} \sim f_{\boldsymbol{\theta}}$, ℓ is of class C^2, and the following additional regularity condition holds:

$$\frac{\partial}{\partial \boldsymbol{\theta}^T} \int \frac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \int \frac{\partial^2}{\partial \boldsymbol{\theta}^T\, \partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \quad \text{for all } \boldsymbol{\theta} \in \Theta.$$

Then, $I_{\boldsymbol{X}}(\boldsymbol{\theta}) = -\mathbb{E}\big[\partial^2 \ell/\partial \boldsymbol{\theta}^T\, \partial \boldsymbol{\theta}\big]$.

[Note: For a column d-vector-valued function $\boldsymbol{h} = (h_1, \ldots, h_n)$, we have

$$\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}^T} := \begin{bmatrix} \dfrac{\partial \boldsymbol{h}}{\partial \theta_1} & \cdots & \dfrac{\partial \boldsymbol{h}}{\partial \theta_p} \end{bmatrix} := \begin{bmatrix} \dfrac{\partial h_1}{\partial \theta_1} & \cdots & \dfrac{\partial h_1}{\partial \theta_p} \\ \dfrac{\partial h_2}{\partial \theta_1} & \cdots & \dfrac{\partial h_2}{\partial \theta_p} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial h_d}{\partial \theta_1} & \cdots & \dfrac{\partial h_d}{\partial \theta_p} \end{bmatrix},$$

and analogously, we have

$$\frac{\partial \boldsymbol{h}^T}{\partial \boldsymbol{\theta}} := \begin{bmatrix} \dfrac{\partial \boldsymbol{h}^T}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial \boldsymbol{h}^T}{\partial \theta_p} \end{bmatrix} := \begin{bmatrix} \dfrac{\partial h_1}{\partial \theta_1} & \cdots & \dfrac{\partial h_d}{\partial \theta_1} \\ \dfrac{\partial h_1}{\partial \theta_2} & \cdots & \dfrac{\partial h_d}{\partial \theta_2} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial h_1}{\partial \theta_p} & \cdots & \dfrac{\partial h_d}{\partial \theta_p} \end{bmatrix},$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is a column vector; here we can observe that $(\partial \boldsymbol{h}/\partial \boldsymbol{\theta}^T)^T = \partial \boldsymbol{h}^T/\partial \boldsymbol{\theta}$.

Moreover, the notation $\dfrac{\partial^2}{\partial \boldsymbol{\theta}^T\, \partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x})$ refers to $\dfrac{\partial}{\partial \boldsymbol{\theta}^T}\left(\dfrac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x})\right)$, where the latter expression can be understood in the above way (note that the expression in the inner parenthesis is a column vector). Particularly, both sides of the equation for the regularity condition above are $p \times p$ matrices.]

Proof.

(a) Note that

$$\begin{aligned}
I_{(\boldsymbol{X},\boldsymbol{Y})}(\boldsymbol{\theta}) &= \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ln[f_{\boldsymbol{\theta}}(\boldsymbol{X})g_{\boldsymbol{\theta}}(\boldsymbol{Y})]\left(\frac{\partial}{\partial\boldsymbol{\theta}}\ln[f_{\boldsymbol{\theta}}(\boldsymbol{X})g_{\boldsymbol{\theta}}(\boldsymbol{Y})]\right)^T\right] \\
&= \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ln f_{\boldsymbol{\theta}}(\boldsymbol{X})\left(\frac{\partial}{\partial\boldsymbol{\theta}}\ln f_{\boldsymbol{\theta}}(\boldsymbol{X})\right)^T\right] + \mathbb{E}\left[+\frac{\partial}{\partial\boldsymbol{\theta}}\ln g_{\boldsymbol{\theta}}(\boldsymbol{X})\left(\frac{\partial}{\partial\boldsymbol{\theta}}\ln g_{\boldsymbol{\theta}}(\boldsymbol{X})\right)^T\right] \\
&\quad \underbrace{+\mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ln f_{\boldsymbol{\theta}}(\boldsymbol{X})\right]\mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\ln g_{\boldsymbol{\theta}}(\boldsymbol{X})\right)^T\right] + \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ln g_{\boldsymbol{\theta}}(\boldsymbol{X})\right]\mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\ln f_{\boldsymbol{\theta}}(\boldsymbol{X})\right)^T\right]}_{O \text{ since } \mathbb{E}[\partial\ell/\partial\boldsymbol{\theta}]=\mathbf{0}} \\
&= I_{\boldsymbol{X}}(\boldsymbol{\theta}) + I_{\boldsymbol{Y}}(\boldsymbol{\theta}).
\end{aligned}$$

(b) The first equality follows by applying (a). The second equality follows by noting that $I_{X_i}(\boldsymbol{\theta}) = I_{X_1}(\boldsymbol{\theta})$ for all $i = 1, \ldots, n$ since $X_1, \ldots, X_n$ are identically distributed.

(c) Note that

$$\int \frac{\partial^2}{\partial\boldsymbol{\theta}^T\,\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \frac{\partial}{\partial\boldsymbol{\theta}^T}\int \frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \frac{\partial}{\partial\boldsymbol{\theta}^T}\frac{\partial}{\partial\boldsymbol{\theta}}\underbrace{\int f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}_{1} = O.$$

Also, we have

$$\begin{aligned}
\frac{\partial^2\ell}{\partial\boldsymbol{\theta}^T\,\partial\boldsymbol{\theta}} &= \frac{\partial}{\partial\boldsymbol{\theta}^T}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\ln f_{\boldsymbol{\theta}}(\boldsymbol{X})\right) = \frac{\partial}{\partial\boldsymbol{\theta}^T}\left(\frac{\frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{X})}{f_{\boldsymbol{\theta}}(\boldsymbol{X})}\right) \\
&= \frac{\frac{\partial^2}{\partial\boldsymbol{\theta}^T\,\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{X})}{f_{\boldsymbol{\theta}}(\boldsymbol{X})} - \frac{\left(\frac{\partial}{\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{X})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}^T} f_{\boldsymbol{\theta}}(\boldsymbol{X})\right)}{f_{\boldsymbol{\theta}}(\boldsymbol{X})^2} \\
&= \frac{\frac{\partial^2}{\partial\boldsymbol{\theta}^T\,\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{X})}{f_{\boldsymbol{\theta}}(\boldsymbol{X})} - \frac{\partial}{\partial\boldsymbol{\theta}}\ln f_{\boldsymbol{\theta}}(\boldsymbol{X})\frac{\partial}{\partial\boldsymbol{\theta}^T}\ln f_{\boldsymbol{\theta}}(\boldsymbol{X}) = \frac{\frac{\partial^2}{\partial\boldsymbol{\theta}^T\,\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{X})}{f_{\boldsymbol{\theta}}(\boldsymbol{X})} - \frac{\partial\ell}{\partial\boldsymbol{\theta}}\frac{\partial\ell}{\partial\boldsymbol{\theta}^T}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2\ell}{\partial\boldsymbol{\theta}^T\,\partial\boldsymbol{\theta}}\right] &= \int \frac{\partial^2\ell}{\partial\boldsymbol{\theta}^T\,\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int\left(\frac{\partial^2}{\partial\boldsymbol{\theta}^T\,\partial\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) - \frac{\partial\ell}{\partial\boldsymbol{\theta}}\frac{\partial\ell}{\partial\boldsymbol{\theta}^T}\right) f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\
&= O - \mathbb{E}\left[\frac{\partial\ell}{\partial\boldsymbol{\theta}}\frac{\partial\ell}{\partial\boldsymbol{\theta}^T}\right] = -I_{\boldsymbol{X}}(\boldsymbol{\theta}),
\end{aligned}$$

as desired.

□

3.4.7 **Properties of CRLB.**

(a) Let $\boldsymbol{X} \sim f_{\boldsymbol{\theta}}$ and $\boldsymbol{\eta} \in \Xi \subseteq \mathbb{R}^m$ be an alternative parameter. Suppose that we can write $\boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\eta})$ where $\boldsymbol{h} : \mathbb{R}^m \to \mathbb{R}^p$ is of class C^1.

i. We have

$$I_{\boldsymbol{X}}(\boldsymbol{\eta}) = \frac{\partial\boldsymbol{\theta}^T}{\partial\boldsymbol{\eta}} I_{\boldsymbol{X}}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\eta}^T}$$

[Note: As a special case, with $m = p = 1$, we have $I_{\boldsymbol{X}}(\eta) = h'(\eta)^2 I_{\boldsymbol{X}}(\theta)$.]

ii. *(Reparameterization invariance of CRLB)* Assume further the assumptions in Theorem 3.4.b hold, $m = p$, and $\partial\boldsymbol{\theta}^T/\partial\boldsymbol{\eta}$ is invertible (or equivalently, $\partial\boldsymbol{\theta}/\partial\boldsymbol{\eta}^T$ is invertible). Then, after replacing the parameter $\boldsymbol{\theta}$ by $\boldsymbol{\eta}$, the CRLB remains the same.

(b) If $p = 1$, then the CRLB simplifies to $g'(\theta)^2/I_{\boldsymbol{X}}(\theta)$, where $I_{\boldsymbol{X}}(\theta) = \mathbb{E}\big[\ell'(\theta)^2\big]$ (which equals $-\mathbb{E}[\ell''(\theta)]$ under the conditions specified in [3.4.6]).

Proof.

(a) i. Note that

$$I_{\boldsymbol{X}}(\boldsymbol{\eta}) = \mathbb{E}\left[\frac{\partial \ell}{\partial \boldsymbol{\eta}}\left(\frac{\partial \ell}{\partial \boldsymbol{\eta}}\right)^T\right] \overset{\text{(chain rule)}}{=} \mathbb{E}\left[\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\eta}}\frac{\partial \ell}{\partial \boldsymbol{\theta}}\left(\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\eta}}\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right)^T\right]$$
$$= \frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\eta}}\mathbb{E}\left[\frac{\partial \ell}{\partial \boldsymbol{\theta}}\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right)^T\right]\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}^T} = \frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\eta}} I_{\boldsymbol{X}}(\boldsymbol{\theta})\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}^T}.$$

ii. The CRLB for $\boldsymbol{\eta}$ is given by

$$\left(\frac{\partial g}{\partial \boldsymbol{\eta}}\right)^T (I_{\boldsymbol{X}}(\boldsymbol{\eta}))^{-1}\frac{\partial g}{\partial \boldsymbol{\eta}} \overset{\text{((i), chain rule)}}{=} \left(\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\eta}}\frac{\partial g}{\partial \boldsymbol{\theta}}\right)^T \left(\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\eta}} I_{\boldsymbol{X}}(\boldsymbol{\theta})\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}^T}\right)^{-1}\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\eta}}\frac{\partial g}{\partial \boldsymbol{\theta}}$$
$$= \left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)^T \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}^T}\left(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}^T}\right)^{-1} I_{\boldsymbol{X}}(\boldsymbol{\theta})^{-1}\left(\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\eta}}\right)^{-1}\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\eta}}\frac{\partial g}{\partial \boldsymbol{\theta}}$$
$$= \left(\frac{\partial g}{\partial \boldsymbol{\theta}}\right)^T I_{\boldsymbol{X}}(\boldsymbol{\theta})^{-1}\frac{\partial g}{\partial \boldsymbol{\theta}},$$

which is the same as the CRLB for $\boldsymbol{\theta}$.

(b) This is because with $p = 1$, $\partial g/\partial \boldsymbol{\theta}$ simplifies to the scalar $g'(\theta)$ and $I_{\boldsymbol{X}}(\boldsymbol{\theta})$ simplifies to the scalar $I_{\boldsymbol{X}}(\theta) = \mathbb{E}\big[\ell'(\theta)^2\big]$, which equals $-\mathbb{E}[\ell''(\theta)]$ under the conditions specified in [3.4.6].

□

3.4.8 **CRLB for exponential family in the canonical form.** Let $\boldsymbol{X} \in \text{Exp}(k)$, with the PDF/PMF given by $f_\eta(\boldsymbol{x}) = e^{\eta T(\boldsymbol{x}) - A(\eta)}h(\boldsymbol{x})$. Suppose that $\Xi \subseteq \mathbb{R}$ is open (so that every $\eta \in \Xi$ is an interior point) and $I_{\boldsymbol{X}}(\eta) > 0$. Then,

(a) $\text{Var}(T(\boldsymbol{X})) = I_{\boldsymbol{X}}(\eta)$.

(b) $\text{Var}(T(\boldsymbol{X})) = 1/I_{\boldsymbol{X}}(\nu)$ where $\nu = \mathbb{E}[T(\boldsymbol{X})]$ (so T attains the CRLB for ν).

Proof.

(a) We have $\ell(\eta) := \ln f_\eta(\boldsymbol{X}) = \eta T(\boldsymbol{X}) - A(\eta) + \ln(h(\boldsymbol{X}))$, and $\ell'(\eta) = T(\boldsymbol{X}) - A'(\eta)$. By Proposition 3.2.a, we have $\mathbb{E}[T(\boldsymbol{X})] = A'(\eta)$, and thus $\ell'(\eta) = T(\boldsymbol{X}) - \mathbb{E}[T(\boldsymbol{X})]$. It follows that

$$I_{\boldsymbol{X}}(\eta) = \mathbb{E}\big[\ell'(\eta)^2\big] = \mathbb{E}\Big[\big(T(\boldsymbol{X}) - \mathbb{E}[T(\boldsymbol{X})]\big)^2\Big] = \text{Var}(T(\boldsymbol{X})).$$

(b) We have $\nu = \mathbb{E}[T(\boldsymbol{X})] = A'(\eta)$. Then, we have

$$\frac{\mathrm{d}\nu}{\mathrm{d}\eta} = A''(\eta) \overset{\text{(Proposition 3.2.a)}}{=} \text{Var}(T(\boldsymbol{X})) \overset{\text{(a)}}{=} I_{\boldsymbol{X}}(\eta) > 0.$$

It then follows by [3.4.7] that

$$I_{\boldsymbol{X}}(\eta) = \left(\frac{\mathrm{d}\nu}{\mathrm{d}\eta}\right)^2 I_{\boldsymbol{X}}(\nu) = I_{\boldsymbol{X}}(\eta)^2 I_{\boldsymbol{X}}(\nu) \implies \text{Var}(T(\boldsymbol{X})) = I_{\boldsymbol{X}}(\eta) = 1/I_{\boldsymbol{X}}(\nu).$$

□

3.4.9 **Fisher information matrix for location-scale family.** Let $X_1, \ldots, X_n$ be iid with a common PDF $(1/b)f((x-a)/b)$ where $a \in \mathbb{R}$ and $b > 0$. Suppose that $f(x) > 0$ for all $x \in \mathbb{R}$ and f is of class C^1, and the regularity condition (a) in Theorem 3.4.b holds.

(a) *(Location family)* If $b = b_0$ is known, then

$$I_{\boldsymbol{X}}(a) = \frac{n}{b_0^2} \int \frac{f'(x)^2}{f(x)} \, \mathrm{d}x.$$

(b) *(Scale family)* If $a = a_0$ is known, then

$$I_{\boldsymbol{X}}(b) = \frac{n}{b^2} \int \frac{(xf'(x) + f(x))^2}{f(x)} \, \mathrm{d}x.$$

(c) *(Location-scale family)* We have

$$I_{\boldsymbol{X}}(a, b) = \frac{n}{b^2} \begin{bmatrix} \int \frac{f'(x)^2}{f(x)} \, \mathrm{d}x & \int \frac{f'(x)(xf'(x)+f(x))}{f(x)} \, \mathrm{d}x \\ \int \frac{f'(x)(xf'(x)+f(x))}{f(x)} \, \mathrm{d}x & \int \frac{(xf'(x)+f(x))^2}{f(x)} \, \mathrm{d}x \end{bmatrix}.$$

Proof.

(a) We have $f_a(x) = (1/b_0) f((x-a)/b_0)$, so $\ln f_a(x) = -\ln b_0 + \ln f((x-a)/b_0)$. Hence,

$$\frac{\partial}{\partial a} \ln f_a(x) = -\frac{1}{b_0} \frac{f'((x-a)/b_0)}{f((x-a)/b_0)}.$$

It follows that

$$\begin{aligned} I_{\boldsymbol{X}}(a) &\overset{[3.4.6]}{=} nI_{X_1}(a) = n\mathbb{E}\left[\left(\frac{\partial}{\partial a} \ln f_a(X_1)\right)^2\right] = n\mathbb{E}\left[\left(\frac{1}{b_0} \frac{f'((X_1-a)/b_0)}{f((X_1-a)/b_0)}\right)^2\right] \\ &= \frac{n}{b_0^3} \int \frac{f'((x-a)/b_0)^2}{f((x-a)/b_0)^2} \cdot f\left(\frac{x-a}{b_0}\right) \mathrm{d}x \overset{\text{(change of variables)}}{=} \frac{n}{b_0^2} \int \frac{f'(u)^2}{f(u)} \, \mathrm{d}u. \end{aligned}$$

(b) We have $f_b(x) = (1/b) f((x-a_0)/b)$, so $\ln f_b(x) = -\ln b + \ln f((x-a_0)/b)$. Hence,

$$\frac{\partial}{\partial b} \ln f_b(x) = -\frac{1}{b} - \frac{x-a_0}{b^2} \frac{f'((x-a_0)/b)}{f((x-a_0)/b)}.$$

It follows that

$$\begin{aligned} I_{\boldsymbol{X}}(b) &\overset{[3.4.6]}{=} nI_{X_1}(b) = n\mathbb{E}\left[\left(\frac{\partial}{\partial b} \ln f_b(X_1)\right)^2\right] = n\mathbb{E}\left[\left(\frac{1}{b} + \frac{X_1-a_0}{b^2} \frac{f'((X_1-a_0)/b)}{f((X_1-a_0)/b)}\right)^2\right] \\ &= \frac{n}{b^3} \int \left(1 + \frac{x-a_0}{b} \frac{f'((x-a_0)/b)}{f((x-a_0)/b)}\right)^2 \cdot f\left(\frac{x-a_0}{b}\right) \mathrm{d}x \\ &\overset{\text{(change of variables)}}{=} \frac{n}{b^2} \int \left(1 + u\frac{f'(u)}{f(u)}\right)^2 f(u) \, \mathrm{d}u = \frac{n}{b^2} \int \frac{(uf'(u)+f(u))^2}{f(u)} \, \mathrm{d}u. \end{aligned}$$

(c) We have $f_{a,b}(x) = (1/b) f((x-a)/b)$, so $\ln f_{a,b}(x) = -\ln b + \ln f((x-a)/b)$. Hence,

$$\frac{\partial}{\partial a} \ln f_{a,b}(x) = -\frac{1}{b} \frac{f'((x-a)/b)}{f((x-a)/b)} \quad \text{and} \quad \frac{\partial}{\partial b} \ln f_{a,b}(x) = -\frac{1}{b} - \frac{x-a}{b^2} \frac{f'((x-a)/b)}{f((x-a)/b)}.$$

It suffices to show that

$$\mathbb{E}\left[\left(\frac{\partial}{\partial a} \ln f_a(X_1)\right)\left(\frac{\partial}{\partial b} \ln f_b(X_1)\right)\right] = \frac{1}{b^2} \int \frac{f'(x)(xf'(x)+f(x))}{f(x)} \, \mathrm{d}x.$$

This can be established as follows:

$$\begin{aligned} \mathbb{E}\left[\left(\frac{\partial}{\partial a} \ln f_a(X_1)\right)\left(\frac{\partial}{\partial b} \ln f_b(X_1)\right)\right] &= \int \frac{1}{b} \frac{f'((x-a)/b)}{f((x-a)/b)} \left(\frac{1}{b} + \frac{x-a}{b^2} \frac{f'((x-a)/b)}{f((x-a)/b)}\right) \frac{1}{b} f\left(\frac{x-a}{b}\right) \mathrm{d}x \\ &= \frac{1}{b^2} \int \frac{f'(u)}{f(u)} \left(1 + u\frac{f'(u)}{f(u)}\right) f(u) \, \mathrm{d}u = \frac{1}{b^2} \int \frac{f'(u)(uf'(u)+f(u))}{f(u)} \, \mathrm{d}u. \end{aligned}$$

□

4 The Method of Maximum Likelihood

4.0.1 Theorem 3.4.b provides a way to establish that a given estimator $\delta(\boldsymbol{X})$ is a UMVUE. However, how we can get an estimator at the first place (which is then to be evaluated)? A famous and widely used approach is the *method of maximum likelihood*, and we will introduce it and evaluate its properties in Section 4.

4.1 Maximum Likelihood Estimator

4.1.1 **A motivating example.** The development of the method of maximum likelihood is based on the intuitive idea that the estimator $\boldsymbol{\delta} := \boldsymbol{\delta}(\boldsymbol{X})$ should give the *most likely* candidate of the parameter $\boldsymbol{\theta}$, given the observed sample $\boldsymbol{X}$.

To see this more concretely, consider the following example. Let $X \sim f_\theta$ be a discrete random variable, where $\theta \in \Theta = \{\theta_1, \theta_2\}$. The PMF $f_\theta(x)$ is given by

$f_\theta(x)$	$x=0$	$x=1$	$x=2$	$x=3$
$\theta=\theta_1$	0.5	0.1	0.1	0.3
$\theta=\theta_2$	0.2	0.3	0.2	0.3

Suppose we have a single observation X_1.

- If $X_1 = 0$, then $\theta = \theta_1$ is more likely.
- If $X_1 = 1$, then $\theta = \theta_2$ is more likely.
- If $X_1 = 2$, then $\theta = \theta_2$ is more likely.
- If $X_1 = 3$, then $\theta = \theta_1$ and $\theta = \theta_2$ are equally likely.

In view of these, a natural estimator $\delta(X_1)$ of θ is

$$\delta(X_1) = \begin{cases} \theta_1 & \text{if } X_1 = 0, \\ \theta_2 & \text{if } X_1 = 1, 2, \\ \theta_1 \text{ or } \theta_2 & \text{if } X_1 = 3. \end{cases}$$

This kind of estimator is an example of a *maximum likelihood estimator* (we are "maximizing the likelihood" in the process of constructing this estimator); this example also illustrates that maximum likelihood estimator may not be unique.

4.1.2 **Definition.** Let $\boldsymbol{X} \sim f_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$.

- *(Log-)likelihood function:* With $\boldsymbol{X} = \boldsymbol{x}$ observed, the functions of $\boldsymbol{\theta}$ defined by $L(\boldsymbol{\theta}) := L(\boldsymbol{\theta}|\boldsymbol{x}) := f_{\boldsymbol{\theta}}(\boldsymbol{x})$ and $\ell(\boldsymbol{\theta}) := \ell(\boldsymbol{\theta}|\boldsymbol{x}) := \ln L(\boldsymbol{\theta})$ are called the **likelihood function** and the **log-likelihood function**, respectively.
- *Maximum likelihood estimator:* If $\boldsymbol{\delta}(\boldsymbol{x})$ is a (global) maximizer of the likelihood function $L(\boldsymbol{\theta}|\boldsymbol{x})$ over Θ for each possible realization $\boldsymbol{x}$ of $\boldsymbol{X}$, then $\boldsymbol{\delta} := \boldsymbol{\delta}(\boldsymbol{X})$ is said to be a **maximum likelihood estimator (MLE)** of $\boldsymbol{\theta}$.

Since the natural log function $\ln(\cdot)$ is strictly increasing, the maximization problem $\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}|\boldsymbol{x})$ is equivalent to the maximization problem $\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta}|\boldsymbol{x})$. Hence, we can replace the likelihood function $L(\boldsymbol{\theta}|\boldsymbol{x})$ by the log-likelihood function $\ell(\boldsymbol{\theta}|\boldsymbol{x})$ in the definition of MLE. This can often (substantially) simplify the work needed for finding MLE, because the likelihood function $L(\boldsymbol{\theta}|\boldsymbol{x})$, as a joint PMF/PDF, often involves products; taking natural log converts them to sums, which are usually easier to deal with.

4.1.3 **Existence of MLE.** Without further assumptions, in general MLE *may not exist*, due to the lack of (optimal) solution to the maximization problem $\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta}|\boldsymbol{x})$. Nevertheless, by the *extreme value*

theorem, a maximizer is guaranteed to exist under the further assumptions that (i) the log-likelihood function ℓ is continuous and (ii) Θ is closed and bounded.

For these further assumptions, (ii) is more often violated. For instance, if $\Theta = \Theta_1 = (0,1)$ or $\Theta = \Theta_2 = (0,\infty)$, then (ii) is not fulfilled. In such cases, we may modify the parameter space Θ suitably such that (ii) is satisfied, by extending/restricting Θ as we see fit. For example, $\Theta_1 = (0,1)$ may be extended to $\Theta_1^* = [0,1] \supseteq \Theta_1$, and $\Theta_2 = (0,\infty)$ may be restricted to $\Theta_2^* = [\underline{\theta}, \overline{\theta}] \subseteq \Theta_2$, where $\underline{\theta} > 0$ and $\overline{\theta} < \infty$ are respectively the minimum and maximum possible values of the parameter, possibly determined from "external information".

4.1.4 **General steps for finding MLE.** Since finding MLE boils down to solving the maximization problem $\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta})$, we can follow standard steps from optimization theory to find MLE.

The first procedure assumes that ℓ is of class C^1 and concave, on a convex set Θ with nonempty interior. [Note: A sufficient condition for the concavity of ℓ is that the Hessian $\nabla^2 \ell := \partial^2 \ell / \partial \boldsymbol{\theta}^T \, \partial \boldsymbol{\theta}$ is negative semidefinite for every point in Θ.

A $p \times p$ matrix A is **negative semidefinite** if $\boldsymbol{x}^T A \boldsymbol{x} \leq 0$ for all $\boldsymbol{x} \in \mathbb{R}^p$. It can be shown that A is negative semidefinite iff all its eigenvalues are nonnegative iff the determinant of every symmetric submatrix of A is positive; a symmetric submatrix of A refers to a submatrix of A obtained by selecting the same set of row and column indices.]

(1) *(Finding candidates of MLE in the interior of Θ)* Solve $\partial \ell / \partial \boldsymbol{\theta} = \mathbf{0}$ (called the **likelihood equation**) to obtain all candidates of MLE in the interior of Θ.

(2) *(Obtaining MLE by appealing to concavity)* Since ℓ is concave, any candidate of MLE obtained in (1) (if exists) is a MLE.

The second procedure assumes that ℓ is of class C^1, Θ is open, and $\ell(\boldsymbol{\theta}) \to -\infty$ as $\|\boldsymbol{\theta}\| \to \infty$ or $\boldsymbol{\theta}$ tends to the boundary of Θ (i.e., tends to an arbitrary boundary point of Θ), which ensures that MLE exists (see also Bickel and Doksum (2015, Lemma 2.3.1)).

Proof. Let $\boldsymbol{\theta}_0 \in \Theta$ and $c = \ell(\boldsymbol{\theta}_0)$. With $\ell(\boldsymbol{\theta}) \to -\infty$ as $\|\boldsymbol{\theta}\| \to \infty$, there exists $M > 0$ such that $\ell(\boldsymbol{\theta}) < c$ whenever $\|\boldsymbol{\theta}\| > M$. Now consider the set $K = \{\boldsymbol{\theta} \in \bar{\Theta} : \|\boldsymbol{\theta}\| \leq M\}$, where $\bar{\Theta}$ is the *closure* of Θ, which is the union of Θ and its boundary. Note that K is closed and bounded. Also, by defining $\ell(\boldsymbol{\theta}_b) = -\infty$ for all boundary points $\boldsymbol{\theta}_b$ of Θ, we can extend the domain of ℓ to $\bar{\Theta}$ while preserving its continuity (treating it as an extended-real-valued function). Now, with $K \subseteq \bar{\Theta}$, we know that there is a maximizer $\boldsymbol{\theta}^*$ of ℓ over K, by the extreme value theorem.

With $\ell(\boldsymbol{\theta}_0) = c \in \mathbb{R}$, we must have $\|\boldsymbol{\theta}_0\| \leq M$ (otherwise from above we would have $\ell(\boldsymbol{\theta}_0) < c$) and hence $\boldsymbol{\theta}_0 \in K$. Thus, the maximizer $\boldsymbol{\theta}^*$ cannot possibly be a boundary point of Θ and we must have $\boldsymbol{\theta}^* \in \Theta$. Furthermore, for all $\boldsymbol{\theta} \in \Theta \setminus K$, we have $\|\boldsymbol{\theta}\| > M$ and thus

$$\ell(\boldsymbol{\theta}) < c = \ell(\boldsymbol{\theta}_0) \overset{(\text{max. on } K)}{\leq} \ell(\boldsymbol{\theta}^*).$$

This implies that $\boldsymbol{\theta}^*$ is a maximizer of ℓ over Θ, and hence a MLE. □

After clarifying why a MLE exists under these assumptions, we provide the steps below.

(1) *(Finding candidates of MLE in the interior of Θ)* Solve the likelihood equation $\partial \ell / \partial \boldsymbol{\theta} = \mathbf{0}$ to obtain all candidates of MLE in the interior of Θ.

(2) *(Comparing log-likelihoods among candidates)* The candidates of MLE are those from (1). By comparing the log-likelihoods of these candidates (more feasible if there are only finitely many candidates), we can identify a MLE (which has the largest loglikelihood among these candidates).

The third procedure assumes that ℓ is of class C^1, and Θ is closed and bounded (so that MLE is guaranteed to exist).

(1) *(Finding candidates of MLE in the interior of Θ)* Solve the likelihood equation $\partial \ell / \partial \boldsymbol{\theta} = \mathbf{0}$ to obtain all candidates of MLE in the interior of Θ.

(2) *(Comparing log-likelihoods among candidates)* The candidates of MLE are now (i) those from (1) and (ii) all boundary points of Θ. By comparing the log-likelihoods of these candidates (more feasible if there are only finitely many candidates), we can identify a MLE (which has the largest loglikelihood among these candidates).

In STAT7609, a MLE can often be found in the ways suggested above (as long as the assumptions are satisfied). If a MLE cannot be determined by the methods above, then more advanced techniques may be needed for finding a MLE.

4.2 Properties of MLE

4.2.1 **Invariance of MLE.** A very attractive property of MLE is its *invariance*, which means that any transformation would preserve the MLE nature: If $\boldsymbol{\delta}$ is a MLE of $\boldsymbol{\theta}$, then $\boldsymbol{g}(\boldsymbol{\delta})$ is a MLE of $\boldsymbol{\eta} := \boldsymbol{g}(\boldsymbol{\theta})$, where $\boldsymbol{g}$ is any function on Θ.

4.2.2 **Invariance when $\boldsymbol{g}$ is bijective.** Showing the invariance is quite straightforward when $\boldsymbol{g}$ is bijective, as the following proof illustrates.

Proposition 4.2.a. Let $\boldsymbol{X} \sim f_{\boldsymbol{\theta}}$ and $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta})$ where $\boldsymbol{g}$ is a bijective function on Θ. If $\boldsymbol{\delta}$ is a MLE of $\boldsymbol{\theta}$, then $\boldsymbol{g}(\boldsymbol{\delta})$ is a MLE of $\boldsymbol{g}(\boldsymbol{\theta})$.

Proof. Since $\boldsymbol{g}$ is bijective, we can write $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = f_{\boldsymbol{g}^{-1}(\boldsymbol{\eta})}(\boldsymbol{x}) =: h_{\boldsymbol{\eta}}(\boldsymbol{x})$, where $h_{\boldsymbol{\eta}}$ is the PMF/PDF of $\boldsymbol{X}$ indexed by the parameter $\boldsymbol{\eta}$. Then, the likelihood function for $\boldsymbol{\theta}$ is given by $L(\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\boldsymbol{x})$, and the likelihood function for $\boldsymbol{\eta}$ is given by $L^*(\boldsymbol{\eta}) = h_{\boldsymbol{\eta}}(\boldsymbol{x})$.

Since $\boldsymbol{\delta}$ is a MLE, $L(\boldsymbol{\theta})$ is maximized at $\boldsymbol{\theta} = \boldsymbol{\delta}$. Thus, we have

$$L^*(\boldsymbol{\eta}) = h_{\boldsymbol{\eta}}(\boldsymbol{x}) = f_{\boldsymbol{g}^{-1}(\boldsymbol{\eta})}(\boldsymbol{x}) = L(\boldsymbol{g}^{-1}(\boldsymbol{\eta})) \overset{\text{(maximizer)}}{\leq} L(\boldsymbol{\delta})$$

for all $\boldsymbol{\eta} \in \boldsymbol{g}(\Theta)$. Note that the equality is attained when $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\delta})$, so $L^*(\boldsymbol{\eta})$ is maximized at $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta})$. It follows that $\boldsymbol{g}(\boldsymbol{\delta})$ is a MLE of $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\delta})$, as desired. □

4.2.3 **Invariance when $\boldsymbol{g}$ is general function.** However, in general, $\boldsymbol{g}$ may not be injective (and hence not bijective). It is possible for $\boldsymbol{g}$ to map two *different* parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in Θ to the same $\boldsymbol{\eta}$: $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta}_1) = \boldsymbol{g}(\boldsymbol{\theta}_2)$. Consequently, even if the value of $\boldsymbol{\eta}$ is known, we are unable to identify what the parameter value in Θ should be. In particular, this causes some troubles in determining an appropriate "likelihood" given a value of $\boldsymbol{\eta}$, say $\boldsymbol{\eta}_0$. When $\boldsymbol{g}$ is bijective, the likelihood at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$ is just the value of the re-indexed PMF/PDF with $\boldsymbol{\eta} = \boldsymbol{\eta}_0$, which can be determined by evaluating the likelihood at $\boldsymbol{\theta} = \boldsymbol{g}^{-1}(\boldsymbol{\eta}_0)$. However, in the general case, it may not be possible to "re-index" the PMF/PDF, and we cannot just evaluate the likelihood at $\boldsymbol{\theta} = \boldsymbol{g}^{-1}(\boldsymbol{\eta}_0)$ as $\boldsymbol{g}^{-1}$ may not exist.

These troubles motivate us to develop the likelihood theory in such a general case as follows. Fix any $\boldsymbol{\eta}$. Let $\Theta_{\boldsymbol{\eta}} := \{\boldsymbol{\theta} \in \Theta : \boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{\eta}\}$ denote the set collecting all parameters in Θ that are mapped to $\boldsymbol{\eta}$ by g. To compute the "likelihood" at $\boldsymbol{\eta}$, there are multiple candidates resulting from the likelihoods L at different $\boldsymbol{\theta}$'s in $\Theta_{\boldsymbol{\eta}}$. As we are uncertain about which $\boldsymbol{\theta}$ to choose from, it is perhaps natural to choose a $\boldsymbol{\theta}$ that maximizes the likelihood L over $\Theta_{\boldsymbol{\eta}}$ (since ultimately we would be maximizing likelihood). This leads to the notion of *induced likelihood*.

The **induced likelihood function** for $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta})$ is defined by $\widetilde{L}(\boldsymbol{\eta}) := \widetilde{L}(\boldsymbol{\eta}|\boldsymbol{x}) = \sup_{\boldsymbol{\theta} \in \Theta_{\boldsymbol{\eta}}} L(\boldsymbol{\theta}|\boldsymbol{x})$. [Note: If there is a maximizer of L over $\Theta_{\boldsymbol{\eta}}$, then the supremum coincides with the maximum (so a $\boldsymbol{\theta}$ that maximizes the likelihood L over $\Theta_{\boldsymbol{\eta}}$ is indeed chosen).]

With this notion of induced likelihood, we can extend the MLE theory as follows. If $\boldsymbol{\delta}(\boldsymbol{x})$ is a maximizer of the *induced* likelihood function $\widetilde{L}(\boldsymbol{\eta}|\boldsymbol{x})$ over $\boldsymbol{g}(\Theta)$ for each possible realization $\boldsymbol{x}$ of $\boldsymbol{X}$, then $\boldsymbol{\delta} := \boldsymbol{\delta}(\boldsymbol{X})$ is said to be a **maximum likelihood estimator (MLE)** of $\boldsymbol{g}(\boldsymbol{\theta})$.

Now, we can state the invariance property of MLE for the general case below.

Theorem 4.2.b. Let $\boldsymbol{X} \sim f_{\boldsymbol{\theta}}$ and $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta})$, where $\boldsymbol{g}$ is any function on Θ. If $\boldsymbol{\delta}$ is a MLE of $\boldsymbol{\theta}$, then $\boldsymbol{g}(\boldsymbol{\delta})$ is a MLE of $\boldsymbol{g}(\boldsymbol{\theta})$.

Proof. Let $\widetilde{L}(\boldsymbol{\eta})$ be the induced likelihood function for $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta})$. Since $\boldsymbol{\delta}$ is a MLE of $\boldsymbol{\theta}$, for all $\boldsymbol{\eta} \in \boldsymbol{g}(\Theta)$ we have

$$\widetilde{L}(\boldsymbol{\eta}) \le \sup_{\boldsymbol{\eta}\in\boldsymbol{g}(\Theta)} \widetilde{L}(\boldsymbol{\eta}) = \sup_{\boldsymbol{\eta}\in\boldsymbol{g}(\Theta)} \sup_{\boldsymbol{\theta}\in\Theta_{\boldsymbol{\eta}}} L(\boldsymbol{\theta}) = \sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}) = L(\boldsymbol{\delta}).$$

Note that

$$\widetilde{L}(\boldsymbol{g}(\boldsymbol{\delta})) = \sup_{\boldsymbol{\theta}\in\Theta:\boldsymbol{g}(\boldsymbol{\theta})=\boldsymbol{g}(\boldsymbol{\delta})} L(\boldsymbol{\theta}) = L(\boldsymbol{\delta}),$$

so for the inequality above, the equality is attained when $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\delta})$. This implies that $\boldsymbol{g}(\boldsymbol{\delta})$ is a MLE of $\boldsymbol{g}(\boldsymbol{\theta})$. □

4.2.4 **Asymptotic efficiency.** As mentioned in [3.4.2], we are often interested in investigating the "CAN" property of an estimator. So, in what follows, we will have discussions on the "CAN" property of MLE, focusing on the case $p = 1$.

Actually, MLE not only enjoys the "CAN" property, but is also *asymptotically efficient*. Let $\boldsymbol{X}_n = (X_1, \dots, X_n) \sim f_\theta$ be a random sample. An estimator δ of $g(\theta)$ is said to be **asymptotically efficient** (AE) if $\sqrt{n}(\delta(\boldsymbol{X}_n) - g(\theta)) \xrightarrow[n\to\infty]{\mathrm{d}} \mathrm{N}\big(0, g'(\theta)^2/I_{X_1}(\theta)\big)$. This means that δ is asymptotically normal and also attains the CRLB asymptotically, in the sense that $\mathrm{Var}(\delta(\boldsymbol{X}_n)) \approx g'(\theta)^2/(nI_{X_1}(\theta))$ which equals the CRLB in Theorem 3.4.b (under the regularity conditions), when n is large.

4.2.5 **Strong consistency and asymptotic efficiency of MLE.** Under some regularity conditions, we can show that MLE is strongly consistent and asymptotically efficient a.s., as follows.

Theorem 4.2.c. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$ where $\theta \in \Theta \subseteq \mathbb{R}$, $\ell_1(\theta) := \ell_1(\theta|X) := \ln f_\theta(X)$, $\ell(\theta) := \ell(\theta|\boldsymbol{X}) := \ln \prod_{i=1}^n f_\theta(X_i) = \sum_{i=1}^n \ell_1(\theta|X_i)$, and θ_0 be the true parameter. Assume that ℓ_1 is of class C^3, θ_0 is an interior point of Θ, $0 < I_{X_1}(\theta) = \mathbb{E}\big[(\partial\ell_1(\theta)/\partial\theta)^2\big] < \infty$ for all θ, and the following regularity conditions hold:

(a) For all θ sufficiently close to θ_0, there exists a nonnegative function $H_\theta(x)$ such that $|\ell_1'''(\theta)| \le H_\theta(x)$ and $\mathbb{E}_\theta[H(X_1)] < \infty$.

(b) $\dfrac{\partial}{\partial\theta}\displaystyle\int f_\theta(x)\,\mathrm{d}x = \int \frac{\partial}{\partial\theta} f_\theta(x)\,\mathrm{d}x$ for all θ.

(c) $\dfrac{\partial}{\partial\theta}\displaystyle\int \frac{\partial}{\partial\theta} f_\theta(x)\,\mathrm{d}x = \int \frac{\partial^2}{\partial\theta^2} f_\theta(x)\,\mathrm{d}x$ for all θ.

(d) $\dfrac{\partial}{\partial\theta}\displaystyle\int \ell_1'''(\theta|x)\,\mathrm{d}x = \int \frac{\partial}{\partial\theta} \ell_1'''(\theta|x)\,\mathrm{d}x$ for all θ.

Then, with probability 1, the likelihood equation $\ell'(\theta) = 0$ yields a sequence of solutions $\delta(\boldsymbol{X}_n)$ satisfying

(a) *(strong consistency)* $\delta(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{\text{a.s.}} \theta_0$, and

(b) *(asymptotic efficiency)* $\sqrt{n}(\delta(\boldsymbol{X}_n) - \theta_0) \xrightarrow[n\to\infty]{\mathrm{d}} \mathrm{N}\big(0, I_{X_1}(\theta_0)^{-1}\big)$

where $\boldsymbol{X} := \boldsymbol{X}_n := (X_1, \dots, X_n)$.

[Note: Provided that ℓ is concave, every solution to the likelihood equation is a MLE. So in this case, this result can be seen as asserting the a.s. strong consistency and asymptotic efficiency of MLE.]

Proof.

(a) Let $s(\theta) := s(\theta|\boldsymbol{X}_n) := \ell'(\theta|\boldsymbol{X}_n)/n = \sum_{i=1}^n \ell_1'(\theta|X_i)/n$. Then, for all θ sufficiently close to θ_0, we have

$$|s''(\theta)| \le \frac{1}{n}\sum_{i=1}^n \left|\frac{\partial^3 \ln f_\theta(X_i)}{\partial\theta^3}\right| \le \frac{1}{n}\sum_{i=1}^n H(X_i) =: \bar{H}(\boldsymbol{X}) =: \bar{H}(\boldsymbol{X}_n).$$

Hence, by Taylor's theorem, for all θ sufficiently close to θ_0 we have

$$\begin{aligned} s(\theta) &= s(\theta_0) + s'(\theta_0)(\theta-\theta_0) + \frac{1}{2}s''(\xi)(\theta-\theta_0)^2 \\ &= s(\theta_0) + s'(\theta_0)(\theta-\theta_0) + \frac{1}{2}\bar{H}(\boldsymbol{X})\eta^*(\theta-\theta_0)^2 \end{aligned}$$

where ξ is between θ and θ_0, and $\eta^* := s''(\xi)/\bar{H}(\boldsymbol{X})$ with $|\eta^*| \le 1$.
By SLLN, we have:

$$\begin{aligned} s(\theta_0|\boldsymbol{X}_n) \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}[s(\theta_0)] &= \frac{1}{n}\sum_{i=1}^n \int \frac{\frac{\partial}{\partial\theta}f_{\theta_0}(x_i)}{f_{\theta_0}(x_i)} f_{\theta_0}(x_i)\,\mathrm{d}x_i = \frac{1}{n}\sum_{i=1}^n \int \frac{\partial}{\partial\theta} f_{\theta_0}(x_i)\,\mathrm{d}x_i \\ &= \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}\int f_{\theta_0}(x_i)\,\mathrm{d}x_i = 0, \\ s'(\theta_0|\boldsymbol{X}_n) \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}[s'(\theta_0)] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\ell_1''(\theta_0|X_i)] = \mathbb{E}[\ell''(\theta_0|X_1)] \overset{[3.4.6],\,(c)}{=} -I_{X_1}(\theta_0), \\ \bar{H}(\boldsymbol{X}_n) &\xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}_{\theta_0}[H(X_1)] < \infty. \end{aligned}$$

Fix any $\varepsilon > 0$. By the definition of convergence, for all sufficiently large n, we have, a.s.,

$$\begin{aligned} |s(\theta_0)| &\le \varepsilon^2, \\ |s'(\theta_0) + I_{X_1}(\theta_0)| &\le \varepsilon^2, \\ |\bar{H}(\boldsymbol{X}_n) - \mathbb{E}_{\theta_0}[H(X_1)]| &\le \varepsilon^2. \end{aligned}$$

With sufficiently small $\varepsilon > 0$ and sufficiently large n, by Taylor's theorem we then have

$$\begin{aligned} s(\theta_0+\varepsilon) &= s(\theta_0) + s'(\theta_0)\varepsilon + \frac{1}{2}\bar{H}(\bar{\boldsymbol{X}}_n)\eta^*\varepsilon^2 \\ &\overset{(|\eta^*|\le 1)}{\le} \varepsilon^2 + \big(-I_{X_1}(\theta_0)+\varepsilon^2\big)\varepsilon + \frac{1}{2}\big(\mathbb{E}_{\theta_0}[H(X_1)]+\varepsilon^2\big)\varepsilon^2 \\ &= -I_{X_1}(\theta_0)\varepsilon + \varepsilon^2 + \frac{1}{2}\big(\mathbb{E}_{\theta_0}[H(X_1)]+\varepsilon^2\big)\varepsilon^2 + \varepsilon^3 \\ &< 0, \end{aligned}$$

and similarly,

$$s(\theta_0-\varepsilon) = s(\theta_0) - s'(\theta_0)\varepsilon + \frac{1}{2}\bar{H}(\bar{\boldsymbol{X}}_n)\eta^*\varepsilon^2 > 0,$$

with probability 1 (we can ensure $\theta_0-\varepsilon, \theta_0+\varepsilon \in \Theta$ by having ε sufficiently small since θ_0 is an interior point). Now, since $s(\theta)$ is continuous, by the intermediate value theorem we know that the likelihood equation $s(\theta) = \ell'(\theta) = 0$ has a solution $\delta_{n_\varepsilon,\varepsilon}$ in $[\theta_0-\varepsilon, \theta_0+\varepsilon]$, with probability 1. Then, for each $k \in \mathbb{N}$, we let $\delta(\boldsymbol{X}_k) = \delta_{n_{1/k},1/k}$. By construction, we then have $\delta(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{\text{a.s.}} \theta_0$.

(b) With $\delta(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{\text{a.s.}} \theta_0$, for all sufficiently large n, we have that $\delta(\boldsymbol{X}_n)$ is sufficiently close to θ_0 a.s., and thus by Taylor's theorem we have, a.s.,

$$0 = s(\delta(\boldsymbol{X}_n)) = s(\theta_0) + s'(\theta_0)(\delta(\boldsymbol{X}_n)-\theta_0) + \frac{1}{2}\bar{H}(\boldsymbol{X}_n)\eta^*(\delta(\boldsymbol{X}_n)-\theta_0)^2$$

with $|\eta^*| \le 1$. Therefore,

$$\underbrace{\sqrt{n}s(\theta_0)}_{\xrightarrow[n\to\infty]{\text{d}}\,\mathrm{N}(0,I_{X_1}(\theta_0))\text{ by CLT}} = \sqrt{n}(\delta(\boldsymbol{X}_n)-\theta_0)\Big(\underbrace{-s'(\theta_0)}_{\xrightarrow[n\to\infty]{\text{a.s.}} I_{X_1}(\theta_0)} - \underbrace{\frac{1}{2}\bar{H}(\boldsymbol{X}_n)\eta^*(\delta(\boldsymbol{X}_n)-\theta_0)}_{\xrightarrow[n\to\infty]{\text{a.s.}} 0}\Big),$$

and hence by Slutsky's theorem,

$$\sqrt{n}(\delta(\boldsymbol{X}_n) - \theta_0) \xrightarrow[n\to\infty]{\mathrm{d}} \mathrm{N}\big(0, I_{X_1}(\theta_0)^{-1}\big),$$

as desired.

□

5 Asymptotic and Nonparametric Statistics

5.0.1 **Stochastic order relations.** Before starting the discussion of asymptotic and nonparametric statistics in Section 5, let us first introduce some preliminary notions that are helpful for simplifying our presentation.

(a) *(Big O in probability: Stochastic boundedness)* A sequence $\{X_n\}$ of random variables is said to be **stochastically bounded**, denoted by $X_n = O_p(1)$, if for all $\varepsilon > 0$, there exists $M_\varepsilon > 0$ and $N_\varepsilon \in \mathbb{N}$ such that $\mathbb{P}(|X_n| > M_\varepsilon) < \varepsilon$ for all $n \geq N_\varepsilon$ *(bounded with "high probability" for the "tail" part).*

(b) *(Small o in probability: Convergence to zero in probability)* For a sequence $\{X_n\}$ of random variables, we write $X_n = o_p(1)$ if $X_n \xrightarrow[n\to\infty]{\mathrm{p}} 0$.

(c) For two sequences $\{X_n\}$ and $\{Y_n\}$ of random variables, we write $X_n = O_p(Y_n)$ ($X_n = o_p(Y_n)$ resp.) if $X_n/Y_n = O_p(1)$ ($X_n/Y_n = o_p(1)$ resp.).

5.0.2 **Basic results about stochastic order relations.**

(a) If $X_n \xrightarrow[n\to\infty]{\mathrm{d}} X$ for some random variable X, then $X_n = O_p(1)$.

(b) If $X_n = o_p(1)$, then $X_n = O_p(1)$.

Proof.

(a) Assume that $X_n \xrightarrow[n\to\infty]{\mathrm{d}} X$. By continuous mapping theorem, we have $|X_n| \xrightarrow[n\to\infty]{\mathrm{d}} |X|$. Since the CDF $F_{|X|}$ has at most countably many jumps, for all $\varepsilon > 0$ there exists sufficiently large $M_\varepsilon > 0$, at which $F_{|X|}$ is continuous, such that $\mathbb{P}(|X| > M_\varepsilon) < \varepsilon/2$. Since M_ε is a continuity point of $F_{|X|}$, we have

$$\mathbb{P}(|X_n| > M_\varepsilon) = 1 - F_{|X_n|}(M_\varepsilon) \underset{n\to\infty}{\longrightarrow} 1 - F_{|X|}(M_\varepsilon) = \mathbb{P}(|X| > M_\varepsilon),$$

where $F_{|X_n|}$ denotes the CDF of X_n for each $n \in \mathbb{N}$. Therefore, there exists $N_\varepsilon \in \mathbb{N}$ such that for all $n \geq N_\varepsilon$, $|\mathbb{P}(|X_n| > M_\varepsilon) - \mathbb{P}(|X| > M_\varepsilon)| < \varepsilon/2$, and hence

$$\mathbb{P}(|X_n| > M_\varepsilon) \overset{\text{(triangle)}}{\leq} |\mathbb{P}(|X_n| > M_\varepsilon) - \mathbb{P}(|X| > M_\varepsilon)| + \mathbb{P}(|X| > M_\varepsilon) < \varepsilon.$$

It follows that $X_n = O_p(1)$.

(b) With $X_n = o_p(1)$, we have $X_n \xrightarrow[n\to\infty]{\mathrm{p}} 0$ and thus $X_n \xrightarrow[n\to\infty]{\mathrm{d}} 0$. Hence, the result follows by (a) (with $X = 0$).

□

5.1 Delta Method

5.1.1 The *central limit theorem* (Theorem 2.3.a to be specific) plays a fundamental role in many statistical methods, by establishing the asymptotic normality of a quantity often used in statistical procedures: *sample mean*. However, some statistical methods do not work with sample means directly, and instead consider some *transformations* of sample means (or more generally, transformations of asymptotically normal estimators). To find out their asymptotic distributions, a commonly used tool is the *delta method*.

5.1.2 **Delta method in the univariate case.**

Theorem 5.1.a (Delta method)**.** Let $\{\delta_n\}$ be a sequence of random variables satisfying that $\sqrt{n}(\delta_n - \theta) \xrightarrow[n\to\infty]{\mathrm{d}} \mathrm{N}(0, \sigma^2)$, and g be a function of class C^1 with $g'(\theta) \neq 0$. Then,

$$\sqrt{n}(g(\delta_n) - g(\theta)) \xrightarrow[n\to\infty]{\mathrm{d}} \mathrm{N}\big(0, g'(\theta)^2\sigma^2\big).$$

Proof. By Taylor's theorem, for all $n \in \mathbb{N}$ we have

$$g(\delta_n) - g(\theta) = g'(\xi_n)(\delta_n - \theta)$$

for some ξ_n between θ and δ_n. From the asymptotic normality of δ_n, we can get $\delta_n \xrightarrow[n\to\infty]{\text{p}} \theta$. Since ξ_n is between θ and δ_n, for all $\varepsilon > 0$ we have $0 \le \mathbb{P}(|\xi_n - \theta| > \varepsilon) \le \mathbb{P}(|\delta_n - \theta| > \varepsilon) \xrightarrow[n\to\infty]{} 0$, and hence $\xi_n \xrightarrow[n\to\infty]{\text{p}} \theta$. By the continuous mapping theorem, we then have $g'(\xi_n) \xrightarrow[n\to\infty]{\text{p}} g'(\theta)$.
Therefore, applying Slutsky's theorem gives

$$\sqrt{n}(g(\delta_n) - g(\theta)) = \underbrace{g'(\xi_n)}_{\xrightarrow[n\to\infty]{\text{p}} g'(\theta)} \cdot \underbrace{\sqrt{n}(\delta_n - \theta)}_{\xrightarrow[n\to\infty]{\text{d}} \mathrm{N}(0,\sigma^2)} \xrightarrow[n\to\infty]{\text{d}} \mathrm{N}\big(0, g'(\theta)^2\sigma^2\big).$$

□

Remarks:

- As revealed in the proof, a key idea in deriving the delta method is to consider some "changes" ("delta") and apply Taylor theorem accordingly; this can explain the name "delta method".
- Often, we consider the case $\delta_n = \bar{X}_n$ for all $n \in \mathbb{N}$; the asymptotic normality assumption is satisfied due to the CLT.

5.1.3 **Delta method with some zero derivatives.** In Theorem 5.1.a, we assume that $g'(\mu) \neq 0$ to avoid the degenerateness of the resulting normal distribution. But actually, even if $g'(\mu) = 0$, we can say something more meaningful about the asymptotic behaviour of the transformed estimator, provided that some additional assumptions are satisfied and some suitable adjustments are made. As you may expect, *Taylor's theorem* is again the key for deriving such a result.

Proposition 5.1.b. Let $\{\delta_n\}$ be a sequence of random variables satisfying that $\sqrt{n}(\delta_n - \theta) \xrightarrow[n\to\infty]{\text{d}} \mathrm{N}\big(0, \sigma^2\big)$, and g be a function of class C^m with $g^{(j)}(\theta) = 0$ for all $j = 1, \dots, m-1$ and $g^{(m)}(\theta) \neq 0$. Then,

$$\frac{g(\delta_n) - g(\theta)}{\frac{1}{m!}g^{(m)}(\theta)(\sigma/\sqrt{n})^m} \xrightarrow[n\to\infty]{\text{d}} \mathrm{N}(0,1)^m$$

where $\mathrm{N}(0,1)^m$ denotes the distribution of X^m with $X \sim \mathrm{N}(0,1)$.

Proof. By Taylor's theorem, for all $n \in \mathbb{N}$ we have

$$g(\delta_n) - g(\theta) = \sum_{j=1}^{m-1} \frac{1}{j!} g^{(j)}(\xi_n)(\delta_n - \theta)^j + \frac{1}{m!} g^{(m)}(\xi_n)(\delta_n - \theta)^m$$

for some ξ_n between θ and δ_n. Similar to the proof of Theorem 5.1.a, we have $\xi_n \xrightarrow[n\to\infty]{\text{p}} \theta$. By the continuous mapping theorem, we then have $g^{(j)}(\xi_n) \xrightarrow[n\to\infty]{\text{p}} g^{(j)}(\theta)$ for every $j = 1, \dots, m$.
By the continuous mapping theorem, for every $j = 1, \dots, m$ we have

$$\left(\frac{\delta_n - \theta}{\sigma/\sqrt{n}}\right)^j \xrightarrow[n\to\infty]{\text{d}} \mathrm{N}(0,1)^j,$$

and

$$\frac{g^{(m)}(\xi_n)}{g^{(m)}(\theta)} \xrightarrow[n\to\infty]{\text{p}} \frac{g^{(m)}(\theta)}{g^{(m)}(\theta)} = 1.$$

Therefore, applying Slutsky's theorem gives

$$\frac{g(\delta_n)-g(\theta)}{\frac{1}{m!}g^{(m)}(\theta)(\sigma/\sqrt{n})^m}=\sum_{j=1}^{m-1}\underbrace{\underbrace{C_j}_{\text{constant}}\underbrace{\frac{g^{(j)}(\xi_n)}{}}_{\xrightarrow[n\to\infty]{\text{p}}0}\underbrace{\left(\frac{\delta_n-\theta}{\sigma/\sqrt{n}}\right)^j}_{\xrightarrow[n\to\infty]{\text{d}}\mathrm{N}(0,1)^j}}_{\xrightarrow[n\to\infty]{\text{d}}0}+\underbrace{\frac{g^{(m)}(\xi_n)}{g^{(m)}(\theta)}}_{\xrightarrow[n\to\infty]{\text{p}}1}\underbrace{\left(\frac{\delta_n-\theta}{\sigma/\sqrt{n}}\right)^m}_{\xrightarrow[n\to\infty]{\text{d}}\mathrm{N}(0,1)^m}\xrightarrow[n\to\infty]{\text{d}}\mathrm{N}(0,1)^m.$$

□

As an example, we take $g(x)=\ln(1+x)^2$ and $\theta=0$. Then, $g'(\theta)=2\ln 1=0$ and $g''(\theta)=2/(1+0)=2\neq 0$. So we can apply Proposition 5.1.b with $m=2$ to get

$$\frac{\ln(1+\delta_n)^2}{(\sigma/\sqrt{n})^2}\xrightarrow[n\to\infty]{\text{d}}\chi_1^2,$$

where χ_1^2 is the chi-squared distribution with 1 degree of freedom (which is the same as $\mathrm{N}(0,1)^2$).

5.1.4 **Delta method in the multivariate case.** Using a similar technique as the one used in the proof of the univariate delta method, we can establish the delta method in the multivariate case.

Theorem 5.1.c (Delta method)**.** Let $\{\boldsymbol{\delta}_n\}$ be a sequence of random vectors satisfying that $\sqrt{n}(\boldsymbol{\delta}_n-\boldsymbol{\theta})\xrightarrow[n\to\infty]{\text{d}}\mathrm{N}_p(\mathbf{0},\Sigma)$, and g be a function of class C^1 with $\partial g/\partial\boldsymbol{\theta}\neq\mathbf{0}$. Then,

$$\sqrt{n}(g(\boldsymbol{\delta}_n)-g(\boldsymbol{\theta}))\xrightarrow[n\to\infty]{\text{d}}\mathrm{N}\left(0,\left(\frac{\partial g}{\partial\boldsymbol{\theta}}\right)^T\Sigma\left(\frac{\partial g}{\partial\boldsymbol{\theta}}\right)\right).$$

Proof. To show the multivariate version of delta method, we need to extend the notion of mode of convergence to the multivariate case (random vectors) also; it is not done here for simplicity, but is covered in STAT7610 (we shall take relevant results about multivariate modes of convergence, learnt in STAT7610, for granted here).

By Taylor's theorem, for all $n\in\mathbb{N}$ we have

$$g(\boldsymbol{\delta}_n)-g(\boldsymbol{\theta})=\left(\frac{\partial g}{\partial\boldsymbol{\theta}}(\boldsymbol{\xi}_n)\right)^T(\boldsymbol{\delta}_n-\boldsymbol{\theta})$$

for some $\boldsymbol{\xi}_n$ satisfying that $\|\boldsymbol{\xi}_n-\boldsymbol{\theta}\|<\|\boldsymbol{\delta}_n-\boldsymbol{\theta}\|$. From the asymptotic normality of $\boldsymbol{\delta}_n$, we can get $\boldsymbol{\delta}_n\xrightarrow[n\to\infty]{\text{p}}\boldsymbol{\theta}$, which then implies that $\boldsymbol{\xi}_n\xrightarrow[n\to\infty]{\text{p}}\boldsymbol{\theta}$. By the continuous mapping theorem, we then have $\partial g/\partial\boldsymbol{\theta}\,(\boldsymbol{\xi}_n)\xrightarrow[n\to\infty]{\text{p}}\partial g/\partial\boldsymbol{\theta}$.

Therefore, applying Slutsky's theorem gives

$$\sqrt{n}(g(\boldsymbol{\delta}_n)-g(\boldsymbol{\theta}))=\underbrace{\left(\frac{\partial g}{\partial\boldsymbol{\theta}}(\boldsymbol{\xi}_n)\right)^T}_{\xrightarrow[n\to\infty]{\text{p}}(\,\partial g/\partial\boldsymbol{\theta}\,)^T}\underbrace{(\boldsymbol{\delta}_n-\boldsymbol{\theta})}_{\xrightarrow[n\to\infty]{\text{d}}\mathrm{N}_p(\mathbf{0},\Sigma)}\xrightarrow[n\to\infty]{\text{d}}\mathrm{N}\left(0,\left(\frac{\partial g}{\partial\boldsymbol{\theta}}\right)^T\Sigma\left(\frac{\partial g}{\partial\boldsymbol{\theta}}\right)\right).$$

□

5.1.5 **Variance-stabilizing transformations.** Now, let us illustrate an application of the delta method in statistics, about the construction of *confidence interval*.

By Theorem 5.1.a and Slutsky's theorem, for iid random sample $X_1,\ldots,X_n$ with finite mean μ and variance $\sigma^2>0$ and a function g of class C^1, we have

$$\sqrt{n}\frac{g(\bar{X}_n)-g(\mu)}{g'(\bar{X}_n)S_n}\xrightarrow[n\to\infty]{\text{d}}\mathrm{N}(0,1)$$

with $\bar{X}_n = \sum_{i=1}^n X_i/n \xrightarrow[n\to\infty]{\text{p}} \mu$ and $S_n^2 = (n-1)\sum_{i=1}^n (X_i - \bar{X}_n)^2 \xrightarrow[n\to\infty]{\text{p}} \sigma^2$ being the sample mean and sample variance, respectively. In view of this asymptotic result, we can construct a corresponding asymptotic $(1-\alpha)$-confidence interval for $g(\mu)$ as follows:

$$\left[g(\bar{X}_n) \pm \frac{g'(\bar{X}_n) S_n}{\sqrt{n}} \Phi^{-1}(1-\alpha/2) \right]$$

where Φ^{-1} denotes the standard normal quantile function. While the confidence interval constructed here is a valid one, we see that the coefficient of $\Phi^{-1}(1-\alpha/2)$ (which may be viewed as the square root of the "asymptotic variance" of $g(\bar{X}_n)$) is rather *unstable* and can vary significantly across different samples with the same sample size. As a result, the widths of the confidence intervals obtained in this way are somewhat unstable, which is an undesirable trait.

To *stabilize* the asymptotic variance, we can utilize **variance-stabilizing transformation**, which works as follows. Consider an asymptotically normal estimator δ_n satisfying that

$$\frac{\delta_n - \theta}{\sigma(\theta)/\sqrt{n}} \xrightarrow[n\to\infty]{\text{d}} \mathrm{N}(0,1).$$

By the delta method, for a transformation (function) g of class C^1 we have

$$\frac{g(\delta_n) - g(\theta)}{g'(\theta)\sigma(\theta)/\sqrt{n}} \xrightarrow[n\to\infty]{\text{d}} \mathrm{N}(0,1).$$

To stabilize the asymptotic variance of $g(\delta_n)$ (make it independent of θ), we choose a transformation g such that

$$g'(\theta)\sigma(\theta) = k \iff g'(\theta) = \frac{k}{\sigma(\theta)} \iff g(\theta) = \int \frac{k}{\sigma(\theta)} \,\mathrm{d}\theta$$

for some constant k, where $\int k/\sigma(\theta)\,\mathrm{d}\theta$ is an antiderivative of $k/\sigma(\theta)$ with respect to θ (indefinite integral).

5.1.6 **An example of variance-stabilizing transformation.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Pois}(\lambda)$ (Poisson distribution), with mean λ and standard deviation $\sigma(\lambda) = \sqrt{\lambda}$. By CLT, we have

$$\frac{\bar{X}_n - \lambda}{\sigma(\lambda)/\sqrt{n}} \xrightarrow[n\to\infty]{\text{d}} \mathrm{N}(0,1).$$

Now we consider the variance-stabilizing transformation approach:

$$g(\lambda) = \int \frac{k}{\sigma(\lambda)} \,\mathrm{d}\lambda = \int \frac{k}{\sqrt{\lambda}} \,\mathrm{d}\lambda = 2k\sqrt{\lambda} + C$$

where C is a real number. Here, we take $k = 1/2$ and $C = 0$. Then, the transformation g is given by $g(\lambda) = \sqrt{\lambda}$. By the delta method, we then have

$$\frac{\sqrt{\bar{X}_n} - \sqrt{\lambda}}{(2\sqrt{\lambda})^{-1}\sqrt{\lambda}/\sqrt{n}} = \frac{\sqrt{\bar{X}_n} - \sqrt{\lambda}}{1/(2\sqrt{n})} \xrightarrow[n\to\infty]{\text{d}} \mathrm{N}(0,1).$$

Correspondingly, we can construct the following asymptotic $(1-\alpha)$-confidence interval for $\sqrt{\lambda}$:

$$\left[\sqrt{\bar{X}_n} \pm \frac{1}{2\sqrt{n}} \Phi^{-1}(1-\alpha/2) \right];$$

here the coefficient of $\Phi^{-1}(1-\alpha/2)$ is a constant (stabilized) when the sample size n is fixed.

To convert it to an asymptotic $(1-\alpha)$-confidence interval for λ, we can apply g^{-1} (which is the square function in this case):

$$\left[\left(\sqrt{\bar{X}_n} \pm \frac{1}{2\sqrt{n}} \Phi^{-1}(1-\alpha/2) \right)^2 \right].$$

5.2 Quantiles

5.2.1 Now we proceed to discuss some *nonparametric* statistics. Particularly, here we will focus on estimation of *quantiles*, which can be seen as "inverse" of the CDF. Given a CDF F, the **pth quantile** of F is defined as $\xi_p := F^{-1}(p) = \inf\{t \in \mathbb{R} : F(t) \geq p\}$, where $p \in (0,1)$.

The **quantile function** associated to F is given by $F^{-1}(p),\ p \in (0,1)$. Geometrically, the graph of the quantile function F^{-1} is indeed the mirror image of the original CDF F along the $45°$ line: F^{-1} jumps when F is flat and F^{-1} is flat when F jumps.

Remarks:

- While it is possible to define quantile at $p = 0$ or $p = 1$, those cases are often not of much interest and so we will exclude them here.
- *(Coinciding with the usual inverse under continuity and strict increasingness)* If the CDF F is continuous and strictly increasing (which holds true in "nice" case), then the quantile function just coincides with the usual inverse function of F for every $p \in (0,1)$. Hence, the pth quantile can be found by solving the equation $F(x) = p$ for x in that scenario.
- We can observe from the definition of quantile that, in general, the *minimum value* among the solutions to the equation $F(x) = p$, if exists, is the pth quantile.

5.2.2 **Properties of quantiles.** The following result provides some properties of quantiles that can facilitate our analysis of them.

Theorem 5.2.a. For all $p \in (0,1)$, we have:

(a) $F^{-1}(F(x)) \leq x$ for all $x \in \mathbb{R}$.

(b) $\lim_{x \to \xi_p^-} F(x) =: F(\xi_p-) \leq p \leq F(\xi_p)$.

(c) $\xi_p \leq x \iff p \leq F(x)$ for all $x \in \mathbb{R}$.

(d) $F^{-1}(p)$ is increasing and left-continuous.

(e) If F is continuous at ξ_p, then $F(\xi_p) = p$.

Proof.

(a) Note that
$$F^{-1}(F(x)) = \inf\{t \in \mathbb{R} : F(t) \geq F(x)\} \leq x,$$
since $x \in \{t \in \mathbb{R} : F(t) \geq F(x)\}$.

(b) Let $S = \{t \in \mathbb{R} : F(t) \geq p\}$. Fix any $\varepsilon > 0$. Note that we must have $F(\xi_p - \varepsilon) < p$; otherwise, $\xi_p - \varepsilon \in S$ and ξ_p would not be a lower bound of S, contradiction. Also, we must have $F(\xi_p + \varepsilon) \geq p$; otherwise, $\xi_p + \varepsilon$ could serve as an even greater lower bound of S, contradiction. Hence, we have $F(\xi_p - \varepsilon) < p \leq F(\xi_p + \varepsilon)$. Letting $\varepsilon \to 0^+$ then yields $F(\xi_p-) \leq p \leq F(\xi_p)$. [⚠ Warning: Upon taking limit, "$<$" becomes "$\leq$".]

(c) "$\Leftarrow$": Assume that $p \leq F(x)$. Then, we have $x \in \{t \in \mathbb{R} : F(t) \geq p\}$, and thus $F^{-1}(p) \leq x$.
"$\Rightarrow$": Assume that $\xi_p \leq x$. Then, we have
$$p \overset{\text{(b)}}{\leq} F(\xi_p) \overset{(F \text{ increasing})}{\leq} F(x).$$

(d) **Increasingness.** Fix any $0 < p_1 \leq p_2 < 1$. Then, we have $S_2 := \{t \in \mathbb{R} : F(t) \geq p_2\} \subseteq \{t \in \mathbb{R} : F(t) \geq p_1\} =: S_1$. Hence, $F^{-1}(p_1) = \inf S_1 \leq \inf S_2 = F^{-1}(p_2)$.
Left continuity. Fix any sequence $\{p_n\}$ that converges to p, with $p_n \leq p_{n+1}$ and $0 < p_n < 1$ for all $n \in \mathbb{N}$. Then, we have $p = \sup\{p_n : n \in \mathbb{N}\}$ by the monotone convergence theorem (MCT), and hence $F^{-1}(p_n) \leq F^{-1}(p_{n+1})$ and $F^{-1}(p_n) \leq F(p)$ for all $n \in \mathbb{N}$, due to the increasingness of F^{-1} shown above.

Thus, by MCT again, we have $F^{-1}(p_n) \underset{n\to\infty}{\longrightarrow} b$ for some $b \in \mathbb{R}$, with $b = \sup\{F^{-1}(p_n) : n \in \mathbb{N}\}$. As b is the least upper bound of $\{F^{-1}(p_n) : n \in \mathbb{N}\}$, for all $n \in \mathbb{N}$ we have $F^{-1}(p_n) \le b \le F^{-1}(p)$, which implies by (c) that $p_n \le F(b)$. Therefore, $p = \lim_{n\to\infty} p_n \le F(b)$, so $b \in \{t \in \mathbb{R} : F(t) \ge p\}$. With $F^{-1}(p)$ being a lower bound of this set, we must have $F^{-1}(p) \le b$.

It follows that $\lim_{n\to\infty} F^{-1}(p_n) = b = F^{-1}(p)$, as desired.

(e) As F is continuous at ξ_p, we have $F(\xi_p-) = F(\xi_p)$. Hence, the result directly follows from (b).

□

The property in part (c) is particularly helpful for dealing with quantiles. It allows us to convert inequalities involving *quantiles* ξ_p into inequalities involving the *CDF* F, which are often easier to handle (and we are familiar with the CDF).

5.2.3 **Quantile transformation.** A main reason why the notion of *quantiles* receives quite a lot of attention is due to the following *quantile transformation* result, which is fundamental for simulation.

Proposition 5.2.b (Quantile transformation)**.** Let F be a CDF and $U \sim \mathrm{U}(0,1)$. Then, $F^{-1}(U) \sim F$.

Proof. For every $x \in \mathbb{R}$, by (c) of Theorem 5.2.a we get

$$\mathbb{P}\big(F^{-1}(U) \le x\big) \overset{\text{(c)}}{=} \mathbb{P}(U \le F(x)) = F(x).$$

□

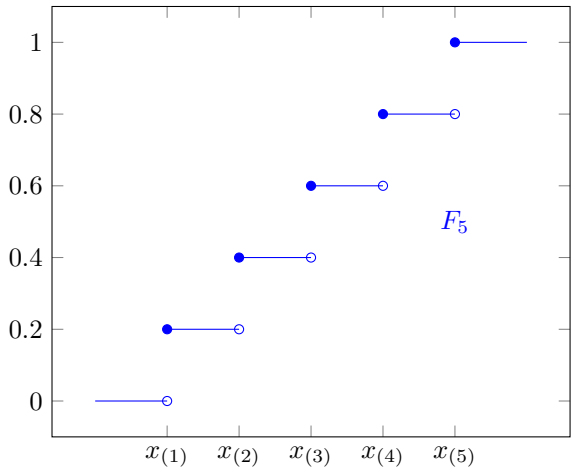
This result suggests that to generate random variates from any distribution with a known CDF, we can first generate random variates from the $\mathrm{U}(0,1)$ distribution, and then apply the quantile function F^{-1} to each of them.

5.2.4 **Sample quantiles.** After having basic understanding about what quantiles are, we then proceed to discuss how we *estimate* them. Given a random sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, it is perhaps natural to estimate the pth quantile ξ_p by the pth quantile of the *empirical CDF*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \le x\}},$$

which assigns a mass $1/n$ to each of the observation in the random sample. It can be verified that the empirical CDF is indeed a valid (discrete) CDF, and the value $F(x)$ can be interpreted as the proportion of X_i's that are less than or equal to x.

To plot an empirical CDF, we can first sort the realized sample $x_1, \ldots, x_n$ in the ascending order to yield the order statistics $x_{(1)}, \ldots, x_{(n)}$, and then we can easily plot it as a "staircase function" as follows:



The **pth sample quantile** of F is given by $\widehat{\xi}_p := F_n^{-1}(p) = \inf\{t \in \mathbb{R} : F_n(t) \geq p\}$, where $p \in (0,1)$. From the plot of the empirical CDF, it is not hard to deduce that $\widehat{\xi}_p = X_{(k)}$ iff $(k-1)/n < p \leq k/n$ iff $\lceil np \rceil = k$, where $X_{(k)}$ is the kth smallest one in the sample $X_1, \ldots, X_n$ and $\lceil \cdot \rceil$ denotes the ceiling function. Hence, we can express the pth sample quantile as $X_{(\lceil np \rceil)}$.

5.2.5 **Consistency of sample quantiles.** Let us continue our discussion by evaluating the properties of sample quantile $\widehat{\xi}_p$, viewed as an estimator of ξ_p (by considering each p individually, we can consider this as a special "parametric estimation", with ξ_p treated as a "parameter").

The following result establishes an upper bound on the probability of being "not close to" the actual quantile ξ_p (under some conditions), which allows us to show the strong consistency of the sample quantile (Theorem 5.2.d).

Lemma 5.2.c. Suppose that $F(\xi_p - \delta) < p < F(\xi_p + \delta)$ for all $\delta > 0$ *(F is not flat at ξ_p).* For all $\varepsilon > 0$ and $n \in \mathbb{N}$ (sample size), we have

$$\mathbb{P}\Big(\Big|\widehat{\xi}_p - \xi_p\Big| > \varepsilon\Big) \leq 4e^{-2n\delta_\varepsilon^2}$$

for some $\delta_\varepsilon > 0$ (possibly depending on ε but not n).

Proof. Fix any $\varepsilon > 0$. Then, we have

$$\begin{aligned}
\mathbb{P}\Big(\widehat{\xi}_p > \xi_p + \varepsilon\Big) &= \mathbb{P}\big(F_n^{-1}(p) > \xi_p + \varepsilon\big) \overset{\text{(Theorem 5.2.a)}}{=} \mathbb{P}(p > F_n(\xi_p + \varepsilon)) \\
&= \mathbb{P}(F_n(\xi_p + \varepsilon) - F(\xi_p + \varepsilon) < p - F(\xi_p + \varepsilon)) \\
&= \mathbb{P}(F_n(\xi_p + \varepsilon) - F(\xi_p + \varepsilon) < -(F(\xi_p + \varepsilon) - p)) \\
&\leq \mathbb{P}(|F_n(\xi_p + \varepsilon) - F(\xi_p + \varepsilon)| > F(\xi_p + \varepsilon) - p) \\
&\leq \mathbb{P}\left(\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > \underbrace{F(\xi_p + \varepsilon) - p}_{\delta_{\varepsilon,1} > 0}\right) \\
&\leq 2e^{-2n\delta_{\varepsilon,1}^2}.
\end{aligned}$$

where the last inequality follows from the *Dvoretzky-Kiefer-Wolfowitz (DKW) inequality.*

Also, we have

$$\begin{aligned}
\mathbb{P}\Big(\widehat{\xi}_p < \xi_p - \varepsilon\Big) &= \mathbb{P}\big(F_n^{-1}(p) \leq \xi_p - \varepsilon\big) \overset{\text{(Theorem 5.2.a)}}{=} \mathbb{P}(p \leq F_n(\xi_p - \varepsilon)) \\
&= \mathbb{P}(F_n(\xi_p - \varepsilon) - F(\xi_p - \varepsilon) \geq p - F(\xi_p - \varepsilon)) \\
&\leq \mathbb{P}(|F_n(\xi_p - \varepsilon) - F(\xi_p - \varepsilon)| > (p - F(\xi_p - \varepsilon))/2) \\
&\leq \mathbb{P}\left(\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > \underbrace{(p - F(\xi_p - \varepsilon))/2}_{\delta_{\varepsilon,2} > 0}\right) \\
&\overset{\text{(DKW)}}{\leq} 2e^{-2n\delta_{\varepsilon,2}^2}.
\end{aligned}$$

Hence,

$$\mathbb{P}\Big(\Big|\widehat{\xi}_p - \xi_p\Big| > \varepsilon\Big) = \mathbb{P}\Big(\widehat{\xi}_p > \xi_p + \varepsilon\Big) + \mathbb{P}\Big(\widehat{\xi}_p < \xi_p - \varepsilon\Big) \leq 2e^{-2n\delta_{\varepsilon,1}^2} + 2e^{-2n\delta_{\varepsilon,2}^2} \leq 4e^{-2n\delta_\varepsilon^2},$$

where $\delta_\varepsilon = \min\{\delta_{\varepsilon,1}, \delta_{\varepsilon,2}\} > 0$. □

Theorem 5.2.d (Strong consistency of sample quantiles)**.** Suppose that $F(\xi_p - \delta) < p < F(\xi_p + \delta)$ for all $\delta > 0$. Then, for all $p \in (0,1)$, $\widehat{\xi}_p \xrightarrow[n \to \infty]{\text{a.s.}} \xi_p$.

Proof. For all $\varepsilon > 0$, by Lemma 5.2.c we have

$$\sum_{n=1}^{\infty} \mathbb{P}\Big(\Big|\widehat{\xi}_p - \xi_p\Big| > \varepsilon\Big) \le \sum_{n=1}^{\infty} 4e^{-2n\delta_\varepsilon^2} \overset{\text{(integral test)}}{<} \infty.$$

The result then follows by [1.2.5]. □

5.2.6 **Asymptotic normality of sample median.** After investigating *consistency* of sample quantiles, we proceed to discuss *asymptotic normality*. There are indeed multiple methods to establish asymptotic normality, and here we will cover two of them: (i) a direct approach for sample *median* and (ii) Bahadur representation for sample quantiles; **median** refers to the 0.5th quantile and **sample median** refers to the 0.5th sample quantile. Since the pth sample quantile is given by $X_{(\lceil np \rceil)}$, we can see that the sample median is given by

$$\widehat{m} = \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ is odd,} \\ X_{(\frac{n}{2})} & \text{if } n \text{ is even.} \end{cases}$$

Proposition 5.2.e (Asymptotic normality of sample median)**.** Let m and $\widehat{m}$ denote the median and sample median, respectively. Assume that $F' = f$ (density) exists and is continuous at m, with $f(m) > 0$. Then,

$$\sqrt{n}(\widehat{m} - m) \xrightarrow[n\to\infty]{\text{d}} \mathrm{N}\big(0, 1/(4f(m)^2)\big).$$

Proof. First consider the case where n is odd. Fix any $x \in \mathbb{R}$. Let $Y_{ni} = \mathbf{1}_{\{X_i \le m + x/\sqrt{n}\}}$ for all $i = 1, \dots, n$, which are iid. Note that we have $\sum_{i=1}^n Y_{ni} \sim \mathrm{Bin}(n, p_n)$ (binomial distribution), with $p_n = F(m + x/\sqrt{n})$. Hence,

$$\begin{aligned} \mathbb{P}\big(\sqrt{n}(\widehat{m} - m) \le x\big) &= \mathbb{P}\left(\widehat{m} \le m + \frac{x}{\sqrt{n}}\right) \\ &= \mathbb{P}\left(\sum_{i=1}^n \mathbf{1}_{\{X_i \le m + x/\sqrt{n}\}} \ge \frac{n+1}{2}\right) \\ &= \mathbb{P}\left(\sum_{i=1}^n Y_{ni} \ge \frac{n+1}{2}\right) \\ &= \mathbb{P}\left(\frac{\sum_{i=1}^n Y_{ni} - np_n}{\sqrt{np_n(1-p_n)}} \ge \frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}}\right). \end{aligned}$$

Since F is differentiable (hence continuous), we have $F(m) = 1/2$ by Theorem 5.2.a. Applying Taylor's theorem, we know that there exists $\delta \in (0,1)$ such that

$$p_n = F(m + x/\sqrt{n}) = F(m) + F'\left(m + \frac{\delta x}{\sqrt{n}}\right)\frac{x}{\sqrt{n}} = \frac{1}{2} + \underbrace{f\left(m + \frac{\delta x}{\sqrt{n}}\right)}_{\substack{(f \text{ continuous at } m) \\ \longrightarrow \\ n\to\infty \\ f(m)}} \frac{x}{\sqrt{n}} \xrightarrow[n\to\infty]{} \frac{1}{2}.$$

Also, we have

$$\sqrt{n}(p_n - 1/2) = f\left(m + \frac{\delta x}{\sqrt{n}}\right) x \xrightarrow[n\to\infty]{} xf(m).$$

It follows that

$$\frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}} = \frac{1/(2\sqrt{n}) - \sqrt{n}(p_n - 1/2)}{\sqrt{p_n(1-p_n)}} \xrightarrow[n\to\infty]{} \frac{-xf(m)}{1/2} = -2xf(m).$$

Hence, by Slutsky's theorem we have

$$\underbrace{\frac{\sum_{i=1}^n Y_{ni} - np_n}{\sqrt{np_n(1-p_n)}}}_{\xrightarrow[n\to\infty]{\text{d}} \mathrm{N}(0,1)} \cdot \underbrace{\frac{-2xf(m)}{\frac{(n+1)/2 - np_n}{\sqrt{np_n(1-p_n)}}}}_{\xrightarrow[n\to\infty]{} 1} \xrightarrow[n\to\infty]{\text{d}} \mathrm{N}(0,1).$$

This implies that

$$\mathbb{P}\big(\sqrt{n}(\widehat{m}-m)\le x\big)=\mathbb{P}\left(\frac{\sum_{i=1}^n Y_{ni}-np_n}{\sqrt{np_n(1-p_n)}}\cdot\frac{-2xf(m)}{\frac{(n+1)/2-np_n}{\sqrt{np_n(1-p_n)}}}\ge -2xf(m)\right)$$
$$\underset{n\to\infty}{\longrightarrow} 1-\Phi(-2xf(m))=\underbrace{\Phi\left(\frac{x}{1/(2f(m))}\right)}_{\text{CDF of }\mathrm{N}(0,\,1/(4f(m)^2))},$$

as desired.

For the case where n is even, the argument is highly similar and we just need to replace "$(n+1)/2$" by "$n/2$"; after the replacement the steps still go through in a similar way. □

5.2.7 **Bahadur representation.** To establish asymptotic normality of sample quantiles in general, the technique of **Bahadur representation** is helpful, which works as follows. Given a statistic X_n which is potentially complex (e.g., nonlinear), we approximate it by a simpler statistic Y_n (e.g., linear or quadratic) probabilistically:

$$X_n-Y_n=o_p(1).$$

Here we attempt to "represent" a complex X_n by a simple Y_n, which may simplify our analysis of the complex X_n. The following lemma provides a sufficient condition for the existence of Bahadur representation (Ghosh, 1971).

Lemma 5.2.f (Ghosh)**.** If $X_n=O_p(1)$ and $\mathbb{P}(X_n\le t,Y_n\ge t+\delta)+\mathbb{P}(X_n\ge t+\delta,Y_n\le t)\underset{n\to\infty}{\longrightarrow}0$ for all $t\in\mathbb{R}$ and $\delta>0$ *(small probability for X_n and Y_n to have "different sizes" for large n)*, then $X_n-Y_n=o_p(1)$.

Proof. Fix any $\varepsilon>0$ and $\delta>0$. Since $X_n=O_p(1)$, there exists $M_\varepsilon>0$ and $N_\varepsilon\in\mathbb{N}$ such that $\mathbb{P}(|X_n|>M_\varepsilon)<\varepsilon$ for all $n\ge N_\varepsilon$.

Now, divide the interval $[-M_\varepsilon,M_\varepsilon]$ into m subintervals $[t_1,t_2],\dots,[t_m,t_{m+1}]$ such that the length of each of them is less than or equal to $\delta/2$ (e.g., dividing it equally into sufficiently many subintervals). Then, we can write $[-M_\varepsilon,M_\varepsilon]=\bigcup_{k=1}^m[t_k,t_{k+1}]$, with $t_{k+1}-t_k\le\delta/2$ for every $k=1,\dots,m$.

Hence, for all $n\ge N_\varepsilon$ we have

$$\begin{aligned}\mathbb{P}(|X_n-Y_n|\ge\delta)&=\mathbb{P}(|X_n-Y_n|\ge\delta,|X_n|\le M_\varepsilon)+\mathbb{P}(|X_n-Y_n|\ge\delta,|X_n|>M_\varepsilon)\\&\le\mathbb{P}(|X_n-Y_n|\ge\delta,X_n\in[-M_\varepsilon,M_\varepsilon])+\underbrace{\mathbb{P}(|X_n|>M_\varepsilon)}_{<\varepsilon}\\&\le\sum_{k=1}^m\mathbb{P}(|X_n-Y_n|\ge\delta,X_n\in[t_k,t_{k+1}])+\varepsilon\\&\le\sum_{k=1}^m\underbrace{\mathbb{P}(X_n-Y_n\ge\delta,X_n\in[t_k,t_{k+1}])}_{=:A_{n,k}}+\sum_{k=1}^m\underbrace{\mathbb{P}(X_n-Y_n\le-\delta,X_n\in[t_k,t_{k+1}])}_{=:B_{n,k}}+\varepsilon\end{aligned}$$

Note that for every $k=1,\dots,m$ we have

$$\begin{aligned}A_{n,k}&=\mathbb{P}(Y_n\le X_n-\delta,X_n\in[t_k,t_{k+1}])\\&\le\mathbb{P}(Y_n\le t_{k+1}-\delta,X_n\ge t_k)\overset{\text{(assumption)}}{\underset{n\to\infty}{\longrightarrow}}0\end{aligned}$$

and

$$\begin{aligned}B_{n,k}&=\mathbb{P}(Y_n\ge X_n+\delta,X_n\in[t_k,t_{k+1}])\\&\le\mathbb{P}(Y_n\ge t_{k+1}+\delta,X_n\le t_{k+1})\overset{\text{(assumption)}}{\underset{n\to\infty}{\longrightarrow}}0.\end{aligned}$$

Letting $\varepsilon \to 0^+$ then yields $\lim_{n\to\infty} \mathbb{P}(|X_n - Y_n| \geq \delta) = 0$. The result then follows since $\delta > 0$ is arbitrary. □

5.2.8 **Bahadur representation for sample quantiles.** Using Lemma 5.2.f, we can establish the following Bahadur representation for sample quantiles.

Proposition 5.2.g (Bahadur representation for sample quantiles)**.** Let $p \in (0,1)$. If $F'(\xi_p) > 0$, then

$$\sqrt{n}(\widehat{\xi}_p - \xi_p) - \sqrt{n}\frac{F(\xi_p) - F_n(\xi_p)}{F'(\xi_p)} = o_p(1).$$

Proof. Let $X_n := \sqrt{n}(\widehat{\xi}_p - \xi_p)$ and $Y_n := \sqrt{n}(F(\xi_p) - F_n(\xi_p))/F'(\xi_p)$. Then, we need to show that $X_n - Y_n = o_p(1)$. Fix any $t \in \mathbb{R}$ and $\delta > 0$. Let $\xi_{n,t} := \xi_p + t/\sqrt{n}$. Now consider:

$$\begin{aligned}
\mathbb{P}(X_n \leq t, Y_n \geq t+\delta) &= \mathbb{P}\Big(\widehat{\xi}_p \leq \xi_{n,t}, Y_n \geq t+\delta\Big) \\
&\overset{\text{(Theorem 5.2.a)}}{=} \mathbb{P}\Big(F_n(\widehat{\xi}_p) \leq F_n(\xi_{n,t}), Y_n \geq t+\delta\Big) \\
&= \mathbb{P}\Big(F(\xi_{n,t}) - F_n(\widehat{\xi}_p) \leq F(\xi_{n,t}) - F_n(\xi_{n,t}), Y_n \geq t+\delta\Big) \\
&= \mathbb{P}\left(\sqrt{n}\frac{F(\xi_{n,t}) - F_n(\xi_{n,t})}{F'(\xi_p)} \leq \sqrt{n}\frac{F(\xi_{n,t}) - F_n(\widehat{\xi}_p)}{F'(\xi_p)}, \sqrt{n}\frac{F(\xi_p) - F_n(\xi_p)}{F'(\xi_p)} \geq t+\delta\right) \\
&= \mathbb{P}(Z_{n,t} \leq W_{n,t}, Z_{n,0} \geq t+\delta)
\end{aligned}$$

where

$$Z_{n,t} := \sqrt{n}\frac{F(\xi_{n,t}) - F_n(\xi_{n,t})}{F'(\xi_p)} \quad \text{and} \quad W_{n,t} := \sqrt{n}\frac{F(\xi_{n,t}) - F_n(\widehat{\xi}_p)}{F'(\xi_p)}.$$

Showing that $W_{n,t} - t = o_p(1)$**.** By Theorem 5.2.a, we have $F_n(\widehat{\xi}_p-) \leq p \leq F_n(\widehat{\xi}_p)$. Hence,

$$|F_n(\widehat{\xi}_p) - p| \leq |F_n(\widehat{\xi}_p) - F_n(\widehat{\xi}_p-)| = \underbrace{1/n}_{\text{jump size}},$$

which implies that

$$\lim_{n\to\infty} \sqrt{n}(p - F_n(\widehat{\xi}_p)) = 0.$$

Also, by the definition of derivative we have

$$\lim_{n\to\infty} \frac{F(\xi_p + t/\sqrt{n}) - p}{t/\sqrt{n}} = \lim_{n\to\infty} \frac{F(\xi_p + t/\sqrt{n}) - F(\xi_p)}{t/\sqrt{n}} = F'(\xi_p);$$

here $F'(\xi_p)$ exists, so F is continuous at ξ_p, which implies by Theorem 5.2.a that $F(\xi_p) = p$. Multiplying both sides by $t/F'(\xi_p)$ gives

$$\lim_{n\to\infty} \sqrt{n}\frac{F(\xi_{n,t}) - p}{F'(\xi_p)} = t.$$

Therefore, we have

$$\lim_{n\to\infty} W_{n,t} = \underbrace{\lim_{n\to\infty} \sqrt{n}\frac{F(\xi_{n,t}) - p}{F'(\xi_p)}}_{t} + \underbrace{\lim_{n\to\infty} \sqrt{n}\frac{p - F_n(\widehat{\xi}_p)}{F'(\xi_p)}}_{0} = t,$$

and hence $W_{n,t} - t = o_p(1)$.

Showing that $Z_{n,t} - Z_{n,0} = o_p(1)$. We have

$$\begin{aligned}
Z_{n,t} - Z_{n,0} &= \sqrt{n}\frac{F(\xi_{n,t}) - F_n(\xi_{n,t})}{F'(\xi_p)} - \sqrt{n}\frac{F(\xi_p) - F_n(\xi_p)}{F'(\xi_p)} \\
&= \sqrt{n}\frac{F(\xi_{n,t}) - F(\xi_p)}{F'(\xi_p)} - \sqrt{n}\frac{F_n(\xi_{n,t}) - F_n(\xi_p)}{F'(\xi_p)} \\
&= \sqrt{n}\frac{\mathbb{E}\big[\mathbf{1}_{\{\xi_p < X_1 \le \xi_{n,t}\}}\big] - \sum_{i=1}^n \mathbf{1}_{\{\xi_p < X_i \le \xi_{n,t}\}}/n}{F'(\xi_p)} \\
&= -\frac{\sum_{i=1}^n \big(\mathbf{1}_{\{\xi_p < X_i \le \xi_{n,t}\}} - \mathbb{E}\big[\mathbf{1}_{\{\xi_p < X_i \le \xi_{n,t}\}}\big]\big)}{\sqrt{n}F'(\xi_p)} \\
&= -\frac{\sum_{i=1}^n (Y_{n,i} - \mathbb{E}[Y_{n,i}])}{\sqrt{n}F'(\xi_p)}
\end{aligned}$$

where $Y_{n,i} := \mathbf{1}_{\{\xi_p < X_i \le \xi_{n,t}\}}$ for all $i = 1, \dots, n$.

Since $\mathbb{E}[Y_{n,1}] = \mathbb{P}(\xi_p < X_1 \le \xi_{n,t}) = F(\xi_p + t/\sqrt{n}) - F(\xi_p) \overset{(F \text{ continuous at } \xi_p)}{\underset{n\to\infty}{\longrightarrow}} 0$, we have

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (Y_{n,i} - \mathbb{E}[Y_{n,i}])\right)^2\right] &\overset{\text{(zero mean)}}{=} \operatorname{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (Y_{n,i} - \mathbb{E}[Y_{n,i}])\right) \\
&= \frac{1}{n}\sum_{i=1}^n \operatorname{Var}(Y_{n,i}) = \operatorname{Var}(Y_{n,1}) \\
&\le \mathbb{E}\big[Y_{n,1}^2\big] = \mathbb{E}[Y_{n,1}] \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

By Chebyshev's inequality, we have $\mathbb{P}(|Z_{n,t} - Z_{n,0}| > \varepsilon) \underset{n\to\infty}{\longrightarrow} 0$ for all $\varepsilon > 0$, and thus $Z_{n,t} - Z_{n,0} = o_p(1)$.

Showing that $\mathbb{P}(X_n \le t, Y_n \ge t + \delta) \underset{n\to\infty}{\longrightarrow} 0$. Using the previous results, we get

$$\begin{aligned}
\mathbb{P}(X_n \le t, Y_n \ge t + \delta) &= \mathbb{P}(Z_{n,t} \le W_{n,t}, Z_{n,0} \ge t + \delta) \\
&\le \mathbb{P}(Z_{n,t} \le W_{n,t}, Z_{n,0} \ge t + \delta, |W_{n,t} - t| \le \delta/2) + \mathbb{P}(|W_{n,t} - t| > \delta/2) \\
&\le \mathbb{P}(|Z_{n,t} - Z_{n,0}| \ge \delta/2) + \mathbb{P}(|W_{n,t} - t| > \delta/2) \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

Completing the proof. Similarly, we have

$$\begin{aligned}
\mathbb{P}(X_n \ge t + \delta, Y_n \le t) &= \mathbb{P}\Big(\widehat{\xi}_p \ge \xi_{n,t+\delta}, Y_n \le t\Big) \\
&= \mathbb{P}(Z_{n,t+\delta} \ge W_{n,t+\delta}, Z_{n,0} \le t) \\
&\le \mathbb{P}(Z_{n,t+\delta} \ge W_{n,t+\delta}, Z_{n,0} \le t, |W_{n,t+\delta} - (t+\delta)| \le \delta/2) + \mathbb{P}(|W_{n,t+\delta} - (t+\delta)| > \delta/2) \\
&\le \mathbb{P}(|Z_{n,t+\delta} - Z_{n,0}| \ge \delta/2) + \mathbb{P}(|W_{n,t+\delta} - (t+\delta)| > \delta/2) \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

Also, it can be shown that $X_n = O_p(1)$ (details omitted). Hence, by Lemma 5.2.f, we have $X_n - Y_n = o_p(1)$. □

5.2.9 Asymptotic normality of sample quantiles.

Theorem 5.2.h (Asymptotic normality of sample quantiles)**.** Let $p \in (0, 1)$. If $F'(\xi_p) > 0$, then

$$\sqrt{n}\Big(\widehat{\xi}_p - \xi_p\Big) \underset{n\to\infty}{\xrightarrow{\text{d}}} \text{N}\left(0, \frac{p(1-p)}{F'(\xi_p)^2}\right).$$

Proof. By Proposition 5.2.g, we can write

$$\sqrt{n}(\widehat{\xi}_p - \xi_p) = \sqrt{n}\frac{F(\xi_p) - F_n(\xi_p)}{F'(\xi_p)} + R_n$$

where $R_n \xrightarrow[n\to\infty]{\text{p}} 0$. Note that

$$\sqrt{n}\frac{F(\xi_p) - F_n(\xi_p)}{F'(\xi_p)} = -\frac{\sqrt{p(1-p)}}{F'(\xi_p)} \underbrace{\sqrt{n}\frac{\sum_{i=1}^n \mathbf{1}_{\{X_i \le \xi_p\}}/n - p}{\sqrt{p(1-p)}}}_{\xrightarrow[n\to\infty]{\text{d}} \text{N}(0,1) \text{ by CLT}} \xrightarrow[n\to\infty]{\text{d}} \text{N}\left(0, \frac{p(1-p)}{F'(\xi_p)^2}\right).$$

So, the result follows by Slutsky's theorem. □

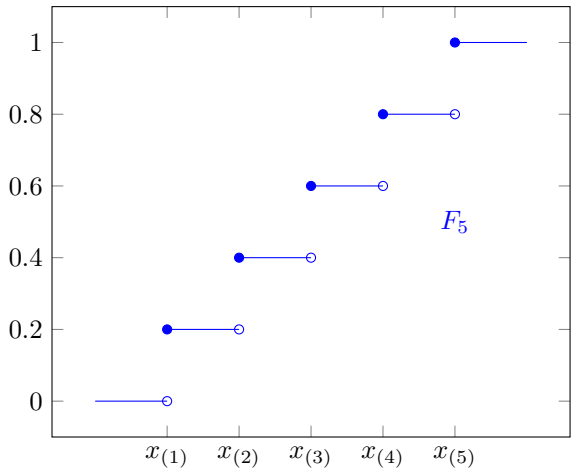
5.2.10 **A more general definition of quantiles.** Recall that we have defined the *pth quantile* by $\xi_p = \inf\{t \in \mathbb{R} : F(t) \ge p\}$, which satisfies $F(\xi_p-) \le p \le F(\xi_p)$ by Theorem 5.2.a. Intuitively, this chain of inequalities captures the "essence" of quantile: F should "jump across" the level p at the pth quantile, if there is a jump there; otherwise, F should equal p exactly at the pth quantile. In view of this, there is a more general definition of quantiles, based on this chain of inequalities.

A **general pth quantile** is any value ξ_p satisfying that $F(\xi_p-) \le p \le F(\xi_p)$, where $p \in (0,1)$.

5.2.11 **General sample quantiles.** Given a random sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, a **general pth sample quantile** is any value $\widehat{\xi}_p$ satisfying that $F_n(\xi_p-) \le p \le F_n(\xi_p)$, where $p \in (0,1)$.

We can observe from the plot of the empirical CDF F_n that such a general pth sample quantile can be expressed as:

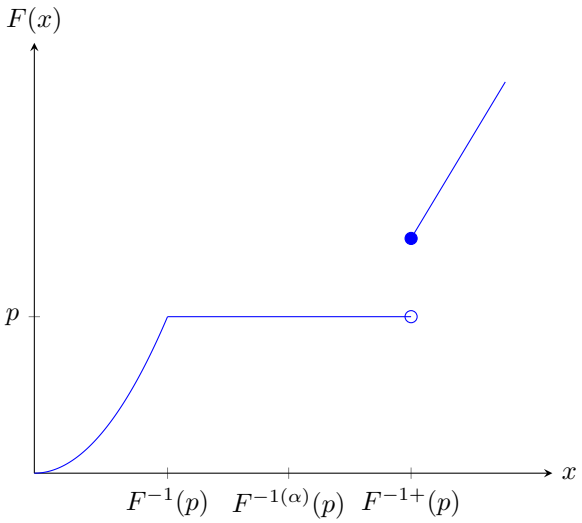
$$\widehat{\xi}_p = \begin{cases} \text{any value in } [X_{(np)}, X_{(np+1)}] & \text{if } np \text{ is an integer,} \\ X_{(\lceil np \rceil)} & \text{if } np \text{ is not an integer.} \end{cases}$$



5.2.12 **Properties of general (sample) quantiles.**

(a) The set of general pth quantiles is given by the closed interval $[F^{-1}(p), F^{-1+}(p)]$, where $F^{-1}(p) := \inf\{t \in \mathbb{R} : F(t) \ge p\}$ and $F^{-1+}(p) := \sup\{t \in \mathbb{R} : F(t) \le p\}$.

[Note: In view of this property, we can express a general pth quantile as a convex combination of $F^{-1}(p)$ and $F^{-1+}(p)$: $F^{-1(\alpha)}(p) := \alpha F^{-1}(p) + (1-\alpha)F^{-1+}(p)$, where $\alpha \in [0,1]$. The function $F^{-1(\alpha)}(\cdot)$ is sometimes known as the **α-mixed generalized inverse** of F.



]

(b) If F is continuous at a general pth quantile ξ_p, then $F(\xi_p) = p$.

(c) Let $X \sim F$ and ξ_q be a general qth quantile of X for every $q \in (0,1)$. Then, $-\xi_{1-p}$ is a general pth quantile of $-X$ for every $p \in (0,1)$.

[Note: With $X \sim F$, a general pth quantile of X is a value ξ_p that satisfies $F(\xi_p-) \le \xi_p \le F(\xi_p)$; the chain of inequalities depends on the CDF of X.]

(d) Let $X \sim F$. The value $p\mathbb{E}[(X-c)_+] + (1-p)\mathbb{E}[(X-c)_-]$ is minimized at $c = \xi_p$, where ξ_p is a general pth quantile of X, and $(X-c)_+ := \max\{X-c, 0\}$ and $(X-c)_- := \max\{-(X-c), 0\}$ denote the *positive part* and *negative part* of $X-c$, respectively.

(e) Given an iid random sample $X_1, \ldots, X_n$, the value $(1/n)\sum_{i=1}^n (p(X_i-c)_+ + (1-p)(X_i-c)_-)$ is minimized at $c = \widehat{\xi}_p$, where $\widehat{\xi}_p$ is a general pth sample quantile.

Proof.

(a) First note that if a and b are general pth quantiles, then every $c \in (a,b)$ is also a general pth quantile, because with $F(a-) \le p \le F(a)$ and $F(b-) \le p \le F(b)$, we have

$$F(c-) \overset{(c\le b)}{\le} F(b-) \le p \le F(a) \overset{(a\le c)}{\le} F(c).$$

It then remains to show that $F^{-1}(p)$ is the smallest general pth quantile and $F^{-1+}(p)$ is the largest general pth quantile. Fix any general pth quantile ξ_p.

Consider first $F^{-1}(p) = \inf\{t \in \mathbb{R} : F(t) \ge p\}$. It is a general pth quantile as shown in Theorem 5.2.a. Also, by definition we have $p \le F(\xi_p)$, and hence $\xi_p \in \{t \in \mathbb{R} : F(t) \ge p\}$. It follows that $F^{-1}(p) \le \xi_p$, and so $F^{-1}(p)$ is the smallest general pth quantile.

Now consider $F^{-1+}(p) = \sup\{t \in \mathbb{R} : F(t) \le p\}$. Note that every $s < F^{-1+}(p)$ must satisfy $F(s) \le p$; otherwise, $F^{-1+}(p)$ would not be the *least* upper bound. Hence, $F(F^{-1+}(p)-) \le p$. Also, every $s > F^{-1+}(p)$ must satisfy $F(s) > p$; otherwise, $F^{-1+}(p)$ would not even be an upper bound. Hence, by the right continuity of F, letting $s \to (F^{-1+}(p))^+$ gives $F(F^{-1+}(p)) \ge p$. Therefore, $F^{-1+}(p)$ is a general pth quantile. It remains to show that $\xi_p \le F^{-1+}(p)$. Assume to the contrary that $\xi_p > F^{-1+}(p)$. Then, we must have $F(\xi_p) > p$ as mentioned above, and so $F(\xi_p-) \ge p$. Together with $F(\xi_p-) \le p$, we get $F(\xi_p-) = p$. Hence, there exists $s \in (F^{-1+}(p), \xi_p)$ such that $F(s) \le p$, contradicting to the upper bound property of $F^{-1+}(p)$.

(b) By assumption, we have $F(\xi_p-) = F(\xi_p)$, which implies that $F(\xi_p) = F(\xi_p-) \le p \le F(\xi_p)$. Hence, $F(\xi_p) = p$.

(c) By assumption, we have $F(\xi_q-) \le q \le F(\xi_q)$ for every $q \in (0,1)$. Note that the CDF of $-X$ is given by

$$F_{-X}(x) = \mathbb{P}(-X \le x) = \mathbb{P}(X \ge -x) = 1 - \mathbb{P}(X < -x) = 1 - F((-x)-),$$

and we have

$$F_{-X}(x-) = \mathbb{P}(-X < x) = \mathbb{P}(X > -x) = 1 - \mathbb{P}(X \le -x) = 1 - F(-x).$$

Also, letting $q = 1-p$ gives

$$\begin{aligned} & F(\xi_{1-p}-) \le 1-p \le F(\xi_{1-p}) \\ \implies & -F(\xi_{1-p}-) \ge p-1 \ge -F(\xi_{1-p}) \\ \implies & 1-F(\xi_{1-p}-) \ge p \ge 1-F(\xi_{1-p}) \\ \implies & F_{-X}(-\xi_{1-p}) \ge p \ge F_{-X}(-\xi_{1-p}-), \end{aligned}$$

which means that $-\xi_{1-p}$ is a general pth quantile of $-X$.

(d) For all $c > \xi_p$, we have

$$\begin{aligned} & p\mathbb{E}[(X-c)_+] + (1-p)\mathbb{E}[(X-c)_-] \\ &= p\mathbb{E}\big[(X-c)\mathbf{1}_{\{X>c\}}\big] + (1-p)\mathbb{E}\big[(c-X)\mathbf{1}_{\{X\le c\}}\big] \\ &= p\big(\mathbb{E}\big[(X-c)\mathbf{1}_{\{X>\xi_p\}}\big] - \mathbb{E}\big[(X-c)\mathbf{1}_{\{\xi_p<X\le c\}}\big]\big) \\ &\quad + (1-p)\big(\mathbb{E}\big[(c-X)\mathbf{1}_{\{X\le\xi_p\}}\big] + \mathbb{E}\big[(c-X)\mathbf{1}_{\{\xi_p<X\le c\}}\big]\big) \\ &= p\left(\underbrace{\mathbb{E}\big[(X-\xi_p)\mathbf{1}_{\{X>\xi_p\}}\big]}_{\mathbb{E}[(X-\xi_p)_+]} + \mathbb{E}\big[(\xi_p-c)\mathbf{1}_{\{X>\xi_p\}}\big] - \mathbb{E}\big[(X-c)\mathbf{1}_{\{\xi_p<X\le c\}}\big]\right) \\ &\quad + (1-p)\left(\mathbb{E}\big[(c-\xi_p)\mathbf{1}_{\{X\le\xi_p\}}\big] + \underbrace{\mathbb{E}\big[(\xi_p-X)\mathbf{1}_{\{X\le\xi_p\}}\big]}_{\mathbb{E}[(X-\xi_p)_-]} + \mathbb{E}\big[(c-X)\mathbf{1}_{\{\xi_p<X\le c\}}\big]\right) \\ &= p\mathbb{E}[(X-\xi_p)_+] + (1-p)\mathbb{E}[(X-\xi_p)_-] \\ &\quad + p\mathbb{E}\big[(\xi_p-c)\mathbf{1}_{\{X>\xi_p\}}\big] + (1-p)\mathbb{E}\big[(c-\xi_p)\mathbf{1}_{\{X\le\xi_p\}}\big] + \underbrace{\mathbb{E}\big[(c-X)\mathbf{1}_{\{\xi_p<X\le c\}}\big]}_{\ge 0} \\ &\ge p\mathbb{E}[(X-\xi_p)_+] + (1-p)\mathbb{E}[(X-\xi_p)_-] + p(c-\xi_p)(F(\xi_p)-1) + (1-p)(c-\xi_p)F(\xi_p) \\ &= p\mathbb{E}[(X-\xi_p)_+] + (1-p)\mathbb{E}[(X-\xi_p)_-] + \underbrace{(c-\xi_p)(F(\xi_p)-p)}_{\ge 0} \\ &\ge p\mathbb{E}[(X-\xi_p)_+] + (1-p)\mathbb{E}[(X-\xi_p)_-]. \end{aligned}$$

Now fix any $c < \xi_p$. Then, we have $-c > -\xi_p$. By (c), we know that $-\xi_p$ is a general $(1-p)$th quantile of $-X$. Hence, using the case established above, and the property that $(-x)_+ = x_-$ and $(-x)_- = x_+$, we get

$$\begin{aligned} & (1-p)\mathbb{E}[(-X-(-c))_+] + p\mathbb{E}[(-X-(-c))_-] \ge (1-p)\mathbb{E}[(-X-(-\xi_p))_+] + p\mathbb{E}[(-X-(-\xi_p))_-] \\ & \quad \implies (1-p)\mathbb{E}[(X-c)_-] + p\mathbb{E}[(X-c)_+] \ge (1-p)\mathbb{E}[(X-\xi_p)_-] + p\mathbb{E}[(X-\xi_p)_+], \end{aligned}$$

as desired.

(e) It follows directly by setting "F" in (c) to be the empirical CDF F_n.

□

Remarks:

- Parts (d) and (e) are indeed closely related to *quantile regression*, which is still an active area of research. The basic idea is as follows. For the famous *linear regression*, our objective is to minimize some "mean squares". In contrast, for *quantile regression*, we are trying to minimizing objective functions like the ones in (d) and (e), which are some weighted averages of positive and negative parts. As demonstrated here, quantiles serve as optimal solutions (explaining the name "quantile regression").
- For the special case $p = 1/2$, the objective functions in (d) and (e) become $(1/2)\mathbb{E}[(X-c)_+ + (X-c)_-] = (1/2)\mathbb{E}[|X-c|]$ and $(1/2n)\sum_{i=1}^n |X_i - c|$, respectively. As we can see here, the sample median (sample 0.5th quantile) is indeed closely related to minimizing *mean absolute deviation*.

5.3 U-Statistics

5.3.1 In many statistical methodologies, estimators are often constructed by performing some kinds of "averages" over the iid random variables contained in the random sample. To describe this general framework, the concept of *U-statistics* is introduced, which includes many commonly used estimators (statistics) as special cases.

5.3.2 **Definitions.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ and $h(x_1, \ldots, x_m)$ be a symmetric function in its arguments (i.e., the order of the inputs can be changed arbitrarily without affecting the output). Let $\theta = \mathbb{E}[h(X_1, \ldots, X_m)]$. [Note: Due to the iid property, we have $\theta = \mathbb{E}[h(X_{i_1}, \ldots, X_{i_m})]$ for all $1 \le i_1 < \cdots < i_m \le n$.]

The **U-statistic** and **V-statistic** of order m with kernel h are defined as

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < \cdots < i_m \le n} h(X_{i_1}, \ldots, X_{i_m}) = \frac{1}{n(n-1)\cdots(n-m+1)} \sum_{i_1, \ldots, i_m \text{ all distinct}} h(X_{i_1}, \ldots, X_{i_m})$$

and

$$V_n = \frac{1}{n^m} \sum_{i_1=1}^n \cdots \sum_{i_m=1}^n h(X_{i_1}, \ldots, X_{i_m}),$$

respectively.

As we can see here, both U-statistic and V-statistic are some kinds of "average". For the former, the averaging takes place over h's evaluated at inputs collected from the sample $X_1, \ldots, X_n$ *with all distinct indices*, while the latter has averaging over h's evaluated at all possible inputs from the sample $X_1, \ldots, X_n$, possibly with repetitions. While it may seem more natural to follow the *V-statistic* approach and average over every possibility, the repetition of inputs involved makes it *biased* for θ in general, since $\mathbb{E}[h(X_{i_1}, \ldots, X_{i_m})]$ may no longer be θ if some inputs of h are identical. On the other hand, U-statistic is always *unbiased* for θ since the indices involved are always *all distinct*:

$$\mathbb{E}[U_n] = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < \cdots < i_m \le n} \underbrace{\mathbb{E}[h(X_{i_1}, \ldots, X_{i_m})]}_{=\theta} = \frac{1}{\binom{n}{m}} \binom{n}{m} \theta = \theta.$$

Moreover, intuitively, as n gets sufficiently large, the terms where some indices are the same should only have "little" contribution, so U-statistic and V-statistic should be "close" *asymptotically*.

In view of these, U-statistics receive more attention than V-statistics, and there have been more developments on the theory of *U-statistics*. Here, we will focus on U-statistics only, and our main goal is to establish its *asymptotic normality* (which implies consistency as mentioned before, hence satisfying the "CAN" property).

5.3.3 **Examples of U-statistics.**

(a) *(Sample mean)* By taking $m = 1$ and h to be the identity function, the U-statistic reduces to the sample mean: $U_n = (1/n)\sum_{i=1}^n X_i$.
[Note: Here $\theta = \mathbb{E}[X_1]$ is the mean.]

(b) *(Sample variance)* By taking $m = 2$ and $h(x_1, x_2) = (x_1 - x_2)^2/2$ (so $\theta = \mathbb{E}\big[(X_1 - X_2)^2/2\big] = (\mathrm{Var}(X_1) + \mathrm{Var}(X_2))/2 = \mathrm{Var}(X_1)$ is the variance), the U-statistic becomes

$$U_n = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \frac{(X_i - X_j)^2}{2} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(X_i - X_j)^2}{2}$$

$$= \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i^2 + X_j^2 - 2X_iX_j) = \frac{1}{2n(n-1)} \left(n \sum_{i=1}^{n} X_i^2 + n \sum_{j=1}^{n} X_j^2 - 2 \left(\sum_{k=1}^{n} X_k \right)^2 \right)$$

$$= \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2 \right) = S_n^2,$$

which is the sample variance.

[Note: Here $\theta = \mathbb{E}\big[(X_1 - X_2)^2/2\big] = (\mathbb{E}\big[X_1^2\big] - 2\mathbb{E}[X_1]\mathbb{E}[X_2] + \mathbb{E}\big[X_2^2\big])/2 = (\mathrm{Var}(X_1) + \mathrm{Var}(X_2))/2 = \mathrm{Var}(X_1)$ is the variance.]

5.3.4 **Hoeffding decomposition.** A major result about U-statistics is the *Hoeffding decomposition*, which is also known as the *projection method*.

The **Hoeffding decomposition** of a statistic $T = T(X_1, \ldots, X_n)$ is as follows:

$$T - \mathbb{E}[T] = \sum_{i=1}^{n} T_i + \sum_{1 \le i_1 < i_2 \le n} T_{i_1, i_2} + \cdots + \sum_{1 \le i_1 < i_2 < \cdots < i_{n-1} \le n} T_{i_1, i_2, \ldots, i_{m-1}} + T_{1, \ldots, n}.$$

where $T_i = \mathbb{E}[T - \mathbb{E}[T] | X_i]$ *(main effects)*, $T_{i_1,i_2} = \mathbb{E}[T - \mathbb{E}[T] | X_{i_1}, X_{i_2}] - T_{i_1} - T_{i_2}$ *(2-way interaction effects)*, $T_{i_1,i_2,i_3} = \mathbb{E}[T - \mathbb{E}[T] | X_{i_1}, X_{i_2}, X_{i_3}] - T_{i_1} - T_{i_2} - T_{i_3} - T_{i_1,i_2} - T_{i_2,i_3} - T_{i_1,i_3}$ *(3-way interaction effects)*, etc.

[Note: The decomposition above can be obtained by rearranging the following equality:

$$T_{1,\ldots,n} \overset{\text{(definition)}}{=} \underbrace{\mathbb{E}[T - \mathbb{E}[T] | X_1, \ldots, X_m]}_{T - \mathbb{E}[T] \text{ by TOWIK property}} - \left(\sum_{i=1}^{n} T_i + \sum_{1 \le i_1 < i_2 \le n} T_{i_1, i_2} + \cdots + \sum_{1 \le i_1 < i_2 < \cdots < i_{n-1} \le n} T_{i_1, i_2, \ldots, i_{m-1}} \right)$$

]

Intuitively, this equality decomposes different "sources" of contributions to $T - \mathbb{E}[T]$. As this term involves n random variables $X_1, \ldots, X_n$, it can be expressed as a sum of "effects" up to n-way interaction effects.

A nice feature of Hoeffding decomposition is the uncorrelatedness of the terms involved.

Theorem 5.3.a. In the Hoeffding decomposition, all the random variables T_i's, T_{i_1,i_2}'s, ..., $T_{1,\ldots,n}$ are uncorrelated.

Proof. First note that all of those random variables have zero mean (by the law of total expectation), so it suffices to show that the expectation of the product of any two of them is 0. Now consider:

$$\mathbb{E}[T_1 T_2] = \mathbb{E}[\mathbb{E}[T_1 T_2 | X_1]] = \mathbb{E}[T_1 \mathbb{E}[T_2 | X_1]] \overset{\text{(independence)}}{=} \mathbb{E}[T_1 \mathbb{E}[T_2]] = 0,$$

$$\mathbb{E}[T_1 T_{1,2}] = \mathbb{E}[\mathbb{E}[T_1 T_{1,2} | X_1]] = \mathbb{E}[T_1 \mathbb{E}[T_{1,2} | X_2]] \overset{\text{(independence, tower property)}}{=} \mathbb{E}[T_1 (T_1 - T_1 - \mathbb{E}[T_2])] = 0,$$

and

$$\mathbb{E}[T_{1,2} T_{1,3}] = \mathbb{E}[T_{1,2} \mathbb{E}[T_{1,3} | X_1, X_2]] = \mathbb{E}[T_{1,2} \mathbb{E}[T_{1,3} | X_1]] \overset{\text{(independence, tower property)}}{=} \mathbb{E}[T_{1,2} (T_1 - T_1 - \mathbb{E}[T_3])] = 0.$$

Using a similar argument (conditioning suitably to allow simplifications), we can establish the uncorrelatedness in general. □

5.3.5 **Hoeffding decomposition to U-statistics.** While the Hoeffding decomposition involves n sums for a general statistic $T = T(X_1, \ldots, X_n)$, it turns out that for a U-statistic U_n of order $m = 2$, only the first two sums are needed in its Hoeffding decomposition.

Theorem 5.3.b. The Hoeffding decomposition to U-statistics of order $m = 2$ is given by

$$U_n - \theta = \frac{2}{n}\sum_{i=1}^{n} g(X_i) + \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \psi(X_i, X_j)$$

with $h_{ij} := h(X_i, X_j)$, $g_i := g(X_i) = \mathbb{E}[h_{ij}|X_i] - \theta$ where j is any index different from i, and $\psi(X_i, X_j) = h_{ij} - g_i - g_j - \theta$.

[Note: It can be shown that $\mathbb{E}[h_{ij}|X_i]$ is a function of X_i, which depends only on i but not j, provided that $j \neq i$. Indeed, $g(X_1), \ldots, g(X_n)$ are iid, and $\psi(X_1, X_2), \psi(X_1, X_3), \ldots, \psi(X_{n-1}, X_n)$ are also iid.]

Proof. We have

$$\begin{aligned}
T_1 = \mathbb{E}[U_n - \theta|X_1] &= \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \mathbb{E}[h(X_i, X_j) - \theta|X_1] \\
&= \frac{2}{n(n-1)} \left(\sum_{j=2}^{n} \mathbb{E}[h(X_1, X_j) - \theta|X_1] + \sum_{2\le i<j\le n} \underbrace{\mathbb{E}[h(X_i, X_j) - \theta|X_1]}_{\mathbb{E}[h(X_i,X_j)-\theta]=0 \text{ by independence}} \right) \\
&= \frac{2}{n(n-1)}(n-1)\mathbb{E}[h(X_1, X_2) - \theta|X_1] = \frac{2}{n} g_1,
\end{aligned}$$

and similarly,

$$T_i = \frac{2}{n} g_i$$

for all $i = 1, \ldots, n$. Now, consider

$$\begin{aligned}
\mathbb{E}[U_n - \theta|X_1, X_2] &= \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \mathbb{E}[h(X_i, X_j) - \theta|X_1, X_2] \\
&= \frac{2}{n(n-1)} \Big(\underbrace{\mathbb{E}[h(X_1, X_2) - \theta|X_1, X_2]}_{h_{12}-\theta \text{ by TOWIK property}} + \sum_{j=3}^{n} \underbrace{\mathbb{E}[h(X_1, X_j) - \theta|X_1, X_2]}_{\mathbb{E}[h(X_1,X_j)-\theta|X_1]=g_1} + \sum_{j=3}^{n} \underbrace{\mathbb{E}[h(X_2, X_j)|X_1, X_2]}_{\mathbb{E}[h(X_2,X_j)-\theta|X_2]=g_2} \\
&\quad + \sum_{3\le i<j\le n} \underbrace{\mathbb{E}[h(X_i, X_j) - \theta|X_1, X_2]}_{\mathbb{E}[h(X_i,X_j)-\theta]=0 \text{ by independence}} \Big) \\
&= \frac{2}{n(n-1)}(h_{12} - \theta + (n-2)g_1 + (n-2)g_2).
\end{aligned}$$

Therefore, we get

$$\begin{aligned}
T_{1,2} &= \mathbb{E}[U_n - \theta|X_1, X_2] - T_1 - T_2 \\
&= \frac{2}{n(n-1)}(h_{12} - \theta + (n-2)g_1 + (n-2)g_2) - \frac{2}{n} g_1 - \frac{2}{n} g_2 \\
&= \frac{2}{n(n-1)}(h_{12} - \theta + (n-2)(g_1 + g_2) - (n-1)(g_1 + g_2)) \\
&= \frac{2}{n(n-1)}(h_{12} - g_1 - g_2 - \theta) = \frac{2}{n(n-1)}\psi(X_1, X_2),
\end{aligned}$$

and similarly,

$$T_{i,j} = \frac{2}{n(n-1)}\psi(X_i, X_j)$$

for all $1 \le i < j \le n$.

We have shown that the first two terms in the Hoeffding decomposition to a U-statistic U_n with order $m = 2$ are as specified in the result. It then remains to show that these two terms are enough, i.e.,

$$\frac{2}{n}\sum_{i=1}^{n} g_i + \frac{2}{n(n-1)} \sum_{1\le i<j\le n} (h_{ij} - g_i - g_j - \theta) = U_n - \theta.$$

Noting that

$$\frac{2}{n(n-1)} \sum_{1\le i<j\le n} (h_{ij} - \theta) = \frac{2}{n(n-1)} \sum_{1\le i<j\le n} h_{ij} - \frac{2}{n(n-1)} \cdot \frac{n(n-1)}{2}\theta = U_n - \theta,$$

it suffices to show that

$$\frac{2}{n}\sum_{i=1}^{n} g_i - \frac{2}{n(n-1)} \sum_{1\le i<j\le n} (g_i + g_j) = 0.$$

To this end, we observe that in the sum $\sum_{1\le i<j\le n}(g_i + g_j)$, each of $g_1, \dots, g_n$ appears exactly $n-1$ times. Thus, we can express the sum as $(n-1)\sum_{i=1}^{n} g_i$, and so

$$\frac{2}{n}\sum_{i=1}^{n} g_i - \frac{2}{n(n-1)} \sum_{1\le i<j\le n} (g_i + g_j) = \frac{2}{n}\sum_{i=1}^{n} g_i - \frac{2}{n}\sum_{i=1}^{n} g_i = 0.$$

□

5.3.6 **Asymptotic normality of U-statistics.** Using the Hoeffding decomposition to U-statistics of order $m = 2$ derived in Theorem 5.3.b, we can establish the asymptotic normality in a straightforward way. Since many estimators can be expressed as U-statistics, this result allows us to justify the asymptotic normality of many estimators *at once* and is therefore quite significant.

Theorem 5.3.c (Asymptotic normality of U-statistics)**.** Let U_n be a U-statistic of order $m = 2$. If $\mathbb{E}\big[h(X_1, X_2)^2\big] < \infty$ and $0 < \sigma_g^2 := \mathrm{Var}(g(X_1)) < \infty$, then

$$\sqrt{n}(U_n - \theta) \xrightarrow[n\to\infty]{\mathrm{d}} \mathrm{N}\big(0, 4\sigma_g^2\big).$$

Proof. By Theorem 5.3.b, we have

$$\sqrt{n}(U_n - \theta) = \frac{2}{\sqrt{n}}\sum_{i=1}^{n} g(X_i) + \sqrt{n}R_n$$

where $R_n = 2/(n(n-1)) \sum_{1\le i<j\le n} \psi(X_i, X_j)$.

Since

$$\begin{aligned}\mathbb{E}\big[(\sqrt{n}R_n)^2\big] &\overset{\text{(zero mean)}}{=} \mathrm{Var}\big(\sqrt{n}R_n\big) \overset{\text{(Theorem 5.3.a)}}{=} \frac{4n}{n^2(n-1)^2} \sum_{1\le i<j\le n} \mathrm{Var}(\psi(X_i, X_j)) \\ &= \frac{4n}{n^2(n-1)^2}\frac{n(n-1)}{2}\mathrm{Var}(\psi(X_1, X_2)) = \frac{2n}{n(n-1)}\mathrm{Var}(\psi(X_1, X_2)) \xrightarrow[n\to\infty]{} 0,\end{aligned}$$

by Chebyshev's inequality we have $\sqrt{n}R_n \xrightarrow[n\to\infty]{\mathrm{p}} 0$. (The assumptions ensure that $\mathrm{Var}(\psi(X_1, X_2))$ is finite.)

Also, by CLT, we have

$$\frac{2}{\sqrt{n}}\sum_{i=1}^{n} g(X_i) = \sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^{n} 2\underbrace{g(X_i)}_{\text{zero mean}} \xrightarrow[n\to\infty]{\mathrm{d}} \mathrm{N}\big(0, 4\sigma_g^2\big).$$

Hence, by Slutsky's theorem, we have

$$\sqrt{n}(U_n - \theta) \xrightarrow[n\to\infty]{\mathrm{d}} \mathrm{N}\big(0, 4\sigma_g^2\big).$$

□

[Note: Actually, the asymptotic normality of U-statistics is not limited to the order $m = 2$, and holds for general order. However, as you may expect, the proof would be more involved and so we just focus on the order $m = 2$ case, which is enough for STAT7609.]

References

Bickel, P. J., & Doksum, K. A. (2015). *Mathematical statistics: Basic ideas and selected topics* (2nd ed., Vol. 1). Chapman and Hall/CRC.

Chung, K. L. (2007). *A course in probability theory* (3rd ed.). Academic Press.

Ghosh, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. *The Annals of Mathematical Statistics*, *42*(6), 1957–1961.

Shao, J. (2003). *Mathematical statistics*. Springer New York.

Vaart, A. W. V. D. (1998). *Asymptotic statistics* (1st ed.). Cambridge University Press.

Links to Definitions

Results

Section 1

- [1.0.2]a: first Borel-Cantelli lemma
- [1.0.2]b: second Borel-Cantelli lemma
- [1.0.2]c: Borel 0-1 law
- [1.0.2]d: tail probability bound

- Proposition 1.1.a: equivalent criteria for almost sure convergence
- Proposition 1.1.b: Cauchy criteria for almost sure convergence
- Proposition 1.1.c: Cauchy criteria for convergence in probability
- Theorem 1.2.a: implications between modes of convergence
- [1.2.2]: equivalence between convergence to constant in probability and in distribution
- Lemma 1.2.b: passage of almost sure boundedness to the limit for convergence in probability
- Lemma 1.2.c: limiting behaviour of expectation taken over sets with probabilities going to zero
- Theorem 1.2.d: dominated convergence in probability implies convergence in L^p
- Corollary 1.2.e: bounded convergence in probability implies convergence in L^p
- [1.2.5]: convergence in probability or L^p sufficiently fast implies almost sure convergence
- [1.2.6]: convergence in probability implies almost sure convergence for a subsequence
- Theorem 1.2.f: Skorokhod's representation theorem
- Theorem 1.3.a: characterization of uniform integrability
- [1.3.2]: properties of uniform integrability
- Lemma 1.3.b: C_r-inequality
- [1.3.4]: convergence in L^p implies convergence of pth absolute moments
- Theorem 1.3.c: Vitali's theorem
- [1.3.6]: convergence in distribution with uniform integrability implies convergence of pth absolute moments
- [1.4.1]: convergence in probability, a.s., and L^p are closed under sums/differences
- [1.4.2]: convergence probability and a.s. are closed under products
- Theorem 1.4.a: continuous mapping theorem
- Theorem 1.4.b: Slutsky's theorem

Section 2

- Proposition 2.1.a: same convergence behaviour for equivalent sequences
- Theorem 2.1.b: WLLN under pairwise independence assumption
- Theorem 2.1.c: Kolmogorov-Feller WLLN
- Proposition 2.2.a: Hajek-Renyi maximal inequality
- Proposition 2.2.b: Kolmogorov maximal inequality
- Theorem 2.2.c: mean criterion for a.s. convergence of series
- Theorem 2.2.d: variance criterion for a.s. convergence of series
- Lemma 2.2.e: Kronecker lemma
- Theorem 2.2.f: Kolmogorov SLLN

- Theorem 2.2.g: Kolmogorov three-series theorem
- Corollary 2.2.h: Kolmogorov two series thoerem for a.s. absolute convergence
- Proposition 2.2.i: SLLN for independent random variables
- Corollary 2.2.j: SLLN for independent random variables with sufficient moment conditions
- Lemma 2.2.k: bounds on $\mathbb{E}[|X|]$ in terms of probability sums
- Theorem 2.2.l: Kolmogorov SLLN for iid random variables
- Corollary 2.2.m: SLLN for iid random variables with necessary and sufficient moment condition
- Proposition 2.2.n: Marcinkiewicz SLLN for iid random variables
- Theorem 2.3.a: classical CLT
- Proposition 2.3.b: Lindeberg-Feller CLT for triangular arrays
- Theorem 2.3.c: Lindeberg-Feller CLT
- Proposition 2.3.d: Lyapunov CLT for triangular arrays
- Theorem 2.3.e: Lyapunov CLT
- [2.3.5]: equivalent form of Lindeberg condition
- Theorem 2.3.f: Polya's theorem

Section 3

- Proposition 3.2.a: moment formulas for exponential family in the canonical form
- Proposition 3.4.a: bias-variance decomposition
- Theorem 3.4.b: Cramér-Rao lower bound
- [3.4.6]: properties of Fisher information matrix
- [3.4.7]: properties of CRLB
- [3.4.8]: properties about CRLB for exponential family in the canonical form
- [3.4.9]: Fisher information matrix for location-scale family

Section 4

- [4.1.4]: general steps for finding MLE
- Proposition 4.2.a: invariance property of MLE under bijective transformation
- Theorem 4.2.b: invariance property of MLE
- Theorem 4.2.c: asymptotic efficiency and strong consistency of MLE

Section 5

- Theorem 5.1.a: univariate delta method
- Proposition 5.1.b: univariate delta method with some zero derivatives
- Theorem 5.1.c: multivariate delta method
- Theorem 5.2.a: properties of quantiles
- Proposition 5.2.b: quantile transformation
- Lemma 5.2.c: probability bound for establishing strong consistency of sample quantiles
- Theorem 5.2.d: strong consistency of sample quantiles
- Proposition 5.2.e: asymptotic normality of sample median
- Lemma 5.2.f: Ghosh's lemma
- Proposition 5.2.g: Bahadur representation for sample quantiles
- Theorem 5.2.h: asymptotic normality of sample quantiles
- [5.2.12]: properties of general (sample) quantiles
- Theorem 5.3.a: uncorrelatedness of terms in Hoeffding decomposition
- Theorem 5.3.b: Hoeffding decomposition to U-statistics of order 2
- Theorem 5.3.c: asymptotic normality of U-statistics of order 2