High Dimensional Probability Notes

Nong Minh Hieu[1]

[1] School of Physical and Mathematical Sciences, Nanyang Technological University (NTU - Singapore)

Contents

1 Random variables

1.1 Basic Inequalities

First, we revisit the definition of a random variable as well as some basic inequalities that we learned in introductory statistics.

Definition 1.1 (Random variable).

Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space. A random variable X is defined as a mapping from the sample space Ω to $\mathbb{R}$:

$$X : \Omega \to \mathbb{R} \tag{1}$$

Σ is the σ-algebra containing the possible events (collection of subsets of Ω) and $\mathbb{P}$ is a probability measure that assigns events with probabilities:

$$\mathbb{P} : \Sigma \to [0, 1] \tag{2}$$

For a given probability space $(\Omega, \Sigma, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$, we will use the following basic notations throughout this note:

- $\|X\|_{L^p}$ - The p^{th} root of the p^{th} moment of the random variable X.

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p}, \ p \in (0, \infty) \tag{3}$$

$$\|X\|_{L^\infty} = \operatorname{ess\,sup} |X| \tag{4}$$

- $L^p(\Omega, \Sigma, \mathbb{P})$ - The space of random variables X satisfying:

$$L^p(\Omega, \Sigma, \mathbb{P}) = \left\{ X : \Omega \to \mathbb{R} \middle| \|X\|_{L^p} < \infty \right\} \tag{5}$$

Some basic inequalities and identities:

- **1. Jensen's Inequality** - For a random variable X and a convex function $\varphi : \mathbb{R} \to \mathbb{R}$, we have:

$$\varphi(\mathbb{E}X) \leqslant \mathbb{E}\varphi(X) \tag{6}$$

- **2. Monotonicity of L^p norm** - For a random variable X:

$$\|X\|_{L^p} \leqslant \|X\|_{L^q}, \ 0 \leqslant p \leqslant q \leqslant \infty. \tag{7}$$

- **3. Minkowski's Inequality** - For $1 \leqslant p \leqslant \infty$ and two random variables X, Y in $L^p(\Omega, \Sigma, \mathbb{P})$ space:

$$\|X + Y\|_{L^p} \leqslant \|X\|_{L^p} + \|Y\|_{L^p}. \tag{8}$$

- **4. Holder's Inequality** - For $p, q \in [1, \infty]$ such that $1/p + 1/q = 1$. Then, for random variables $X \in L^p(\Omega, \Sigma, \mathbb{P})$ and $Y \in L^q(\Omega, \Sigma, \mathbb{P})$, we have:

$$|\mathbb{E}XY| \leqslant \|X\|_{L^p} \cdot \|Y\|_{L^q}. \tag{9}$$

- **5. Markov's Inequality** - For a non-negative random variable X and $t > 0$, we have:

$$\mathbb{P}(X \geqslant t) \leqslant \frac{\mathbb{E}X}{t}. \tag{10}$$

  We can also generalize Markov's Inequality for p^{th} moment:

$$\mathbb{P}(|X| \geqslant t) \leqslant \frac{\mathbb{E}[|X|^p]}{t^p}, \forall t > 0, k \in [2, \infty). \tag{11}$$

- **6. Chebyshev's Inequality** - For a random variable X with mean μ and variance σ^2. Then, for any $t > 0$, we have:

$$\mathbb{P}(|X-\mu| \geqslant t) \leqslant \frac{\sigma^2}{t^2}. \tag{12}$$

- **7. Integral Identity** - Let X be a non-negative random variable, we have:

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > t)dt. \tag{13}$$

Exercises

Exercise 1.1.1: Generalized Integral Identity

Let X be a random variable (not necessarily non-negative). Prove the following identity:

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > t)dt - \int_{-\infty}^0 \mathbb{P}(X < t)dt. \tag{14}$$

Solution (Exercise 1.1.1).

For $x \in \mathbb{R}$, using the basic integral indentity, we have:

$$|x| = \int_0^\infty \mathbf{1}\{t < |x|\}dt$$

We consider the following cases:

- *When $x < 0 \implies x = -|x|$:*

$$x = -\int_0^\infty \mathbf{1}\{t < |x|\}dt = -\int_0^\infty \mathbf{1}\{t < -x\}dt = -\int_0^\infty \mathbf{1}\{-t > x\}dt = -\int_{-\infty}^0 \mathbf{1}\{t > x\}dt.$$

- *When $x \geqslant 0 \implies x = |x|$:*

$$x = \int_0^\infty \mathbf{1}\{t < |x|\}dt = \int_0^\infty \mathbf{1}\{t < x\}dt.$$

Therefore, for $x \in \mathbb{R}$, we can write:

$$x = \int_0^\infty \mathbf{1}\{t < x\}dt - \int_{-\infty}^0 \mathbf{1}\{t > x\}dt.$$

Therefore, for a random variable X not necessarily non-negative, we have:

$$\begin{aligned}
\mathbb{E}X &= \mathbb{E}\left[\int_0^\infty \mathbf{1}\{t < X\}dt - \int_{-\infty}^0 \mathbf{1}\{t > X\}dt\right] \\
&= \mathbb{E}\int_0^\infty \mathbf{1}\{t < X\}dt - \mathbb{E}\int_{-\infty}^0 \mathbf{1}\{t > X\}dt \\
&= \int_0^\infty \mathbb{E}\mathbf{1}\{t < X\}dt - \int_{-\infty}^0 \mathbb{E}\mathbf{1}\{t > X\}dt \\
&= \int_0^\infty \mathbb{P}(t < X)dt - \int_{-\infty}^0 \mathbb{P}(t > X)dt.
\end{aligned}$$

$\square$.

Exercise 1.1.2: p^{th}-moments via tails

Let X be a random variable and $p \in (0, \infty)$. Show that:

$$\mathbb{E}|X|^p = \int_0^\infty p t^{p-1} \mathbb{P}(|X| > t) dt. \tag{15}$$

Solution (Exercise 1.1.2).

Let X be a random variable that is not necessarily non-negative. Using the integral identity, we have:

$$\mathbb{E}|X|^p = \int_0^\infty \mathbb{P}(u < |X|^p) du.$$

Let $t^p = u \implies pt^{p-1}dt = du$. Since we integrate u from $0 \to \infty$, we also integrate t from $0 \to \infty$ when changing the variables. Hence, we have:

$$\mathbb{E}|X|^p = \int_0^\infty \mathbb{P}(t^p < |X|^p) pt^{p-1} dt = \int_0^\infty \mathbb{P}(t < |X|) pt^{p-1} dt.$$

Hence, we obtained the desired identity. □.

1.2 Limit Theorems

1.2.1 Weak Law of Large Numbers

Theorem 1.1: Weak Law of Large Numbers (WLLN)

Let $X_1, \ldots, X_N$ be $i.i.d$ random variables with mean μ. Consider the sum:

$$S_N = X_1 + \cdots + X_N$$

Then, the sample mean **converges to μ in probability** ($S_N/N \xrightarrow{p} \mu$):

$$\lim_{N \to \infty} \mathbb{P}\Big(|S_N/N - \mu| > \epsilon\Big) = 0, \ \forall \epsilon > 0 \tag{16}$$

Proof (Weak Law of Large Numbers (**WLLN**)).

We split the proof into two sections corresponding to the assumptions of finite variance and non-finite variance.

1. ***Finite variance case**: Suppose that $\mathrm{Var} X_i = \sigma^2 < \infty$ for all $1 \leqslant i \leqslant N$. Let $\bar{X} = S_N/N$. Then, $\bar{X}$ is a random variable with the following mean and variance:*

   $$\mathbb{E}\bar{X} = \mu \quad \textit{and} \quad \mathrm{Var}\bar{X} = \frac{\sigma^2}{N}.$$

   Hence, by the Chebyshev's inequality, we have:

   $$\mathbb{P}\Big(|S_N/N - \mu| > \epsilon\Big) = \mathbb{P}\Big(|\bar{X} - \mu| > \epsilon\Big) \leqslant \frac{\sigma^2}{N\epsilon^2}.$$

   Therefore, we have:

   $$\lim_{N \to \infty} \mathbb{P}\Big(|S_N/N - \mu| > \epsilon\Big) \leqslant \lim_{N \to \infty} \frac{\sigma^2}{N\epsilon^2} = 0.$$

   *Hence, we have $\lim_{N \to \infty} \mathbb{P}\Big(|S_N/N - \mu| > \epsilon\Big) = 0$ and we obtained (**WLLN**).*

2. ***Non-finite variance case***: *In this case, we rely on the Levy Continuity Theorem* (**LCT**), *which relies on the convergence of the characteristic function. For $n \geqslant 1$, define the sequence of random variable $Y_n = S_n/n$. Hence, we have:*

$$\begin{aligned} \varphi_{Y_n}(t) &= \varphi_{S_n/n}(t) \\ &= \varphi_{S_n}(t/n) \\ &= \prod_{i=1}^{n} \varphi_{X_i}(t/n) = \Big[\varphi_X(t/n)\Big]^n, \end{aligned}$$

Where $X = X_1 = \cdots = X_n$. By Taylor's expansion, we have:

$$\varphi_X(t/n) = 1 + \frac{it\mathbb{E}[X]}{n} + \mathcal{O}(1/n^2) = 1 + \frac{it\mu}{n} + \mathcal{O}(1/n^2).$$

Hence, we have:

$$\lim_{n\to\infty} \varphi_{Y_n}(t) = \lim_{n\to\infty} \left(1 + \frac{it\mu}{n} + \mathcal{O}(1/n^2)\right)^n = e^{it\mu}.$$

Therefore, by (**LCT**), *we have $Y_n \xrightarrow{p} \mu$.*

Remark 1.1 (Taylor expansion of Moment Generating and Characteristic Functions).
Given a random variable X. For reference, the following are the Taylor expansions of the Moment Generating Function $M_X(t)$ and the Characteristic Function $\varphi_X(t)$:

$$\begin{aligned} M_X(t) &= \mathbb{E}[e^{tX}] = 1 + \sum_{n=1}^{\infty} \frac{t^n}{n!}\mathbb{E}[X^n], \\ \varphi_X(t) &= \mathbb{E}[e^{itX}] = 1 + \sum_{n=1}^{\infty} \frac{(it)^n}{n!}\mathbb{E}[X^n]. \end{aligned} \tag{17}$$

For the sake of my laziness, here are the Taylor expansion for the first three terms of both the MGF and the CF:

$$\begin{aligned} M_X(t) &= 1 + t\mathbb{E}[X] + \frac{t^2}{2}\mathbb{E}[X^2] + \mathcal{O}(t^3), \\ \varphi_X(t) &= 1 + it\mathbb{E}[X] - \frac{t^2}{2}\mathbb{E}[X^2] + \mathcal{O}(t^3). \end{aligned} \tag{18}$$

□.

Theorem 1.2: Levy Continuity Theorem (LCT)

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

$$\forall t \in \mathbb{R} : \lim_{n\to\infty} \varphi_{X_n}(t) = \varphi_X(t) \iff X_n \xrightarrow{d} X, \tag{19}$$

for some random variable X. In a special case where $X = c$ for some $c \in \mathbb{R}$, we have:

$$\forall t \in \mathbb{R} : \lim_{n\to\infty} \varphi_{X_n}(t) = e^{itc} \iff X_n \xrightarrow{p} c. \tag{20}$$

Proof (Levy Continuity Theorem (**LCT**)).
The proof for (**LCT**) *can be found in Gut 2004, Section 9.1, Theorem 9.1 and Collorary 9.1* □.

1.2.2 Strong Law of Large Numbers

Theorem 1.3: Strong Law of Large Numbers (SLLN)

Let $X_1, \ldots, X_N$ be *i.i.d* random variables with mean $\mu < \infty$. Consider the sum:

$$S_N = X_1 + \cdots + X_N$$

Then, the sample mean **converges to μ almost surely** ($S_N/N \xrightarrow{a.s} \mu$):

$$\mathbb{P}\left(\limsup_{N\to\infty} |S_N/N - \mu| > \epsilon\right) = 0, \ \forall \epsilon > 0 \tag{21}$$

Proof (Strong Law of Large Numbers **(SLLN)**).

For the sake of simplicity, we will present the proof for **(SLLN)** *with an additional assumption that* $\mathbb{E}[|X_n|^4] < \infty, \forall n \geqslant 1$. *The proof for the general case of* **(SLLN)** *(also called the Kolmogorov Strong Law) can be found in Gut 2004, Section 6, Theorem 6.1. For convenience, we assume the following:*

1. $\mathbb{E}[|X_n|^4] = K < \infty$.
2. $\mathbb{E}[X_n] = 0$. *For non-zero mean case, we can set* $Y_n = X_n - \mu$ *and repeat the same arguments made below.*

We aim to prove that $\mathbb{P}\left(\limsup_{N\to\infty} |S_N/N| > \epsilon\right) = 0$ *for any* $\epsilon > 0$. *Firstly, use the Multinomial formula to expand* $\mathbb{E}[S_n]$. *The expansion will contain the terms in the following forms:*

$$X_i^2, X_i^3 X_j, X_i^2 X_j^2, X_i^2 X_j X_k, X_i X_j X_k X_\ell,$$

where i, j, k, ℓ *are distinct indices. By independence, we have:*

$$\mathbb{E}[X_i^3 X_j] = \mathbb{E}[X_i^2 X_j X_k] = \mathbb{E}[X_i X_l X_k X_\ell] = 0.$$

As a result, we have the following remaining terms by the Multinomial formula:

$$\begin{aligned}\mathbb{E}[S_n^4] &= \sum_{i=1}^n \mathbb{E}[X_i^4] + \binom{4}{2} \sum_{1\leqslant i<j\leqslant n} \mathbb{E}[X_i^2 X_j^2] \\ &= \sum_{i=1}^n \mathbb{E}[X_i^4] + 6 \underbrace{\sum_{1\leqslant i<j\leqslant n} \mathbb{E}[X_i^2 X_j^2]}_{n(n-1)/2 \text{ terms}} \\ &= nK + 3n(n-1)\mathbb{E}[X_i^2 X_j^2].\end{aligned}$$

By independence, we have $\mathbb{E}[X_i^2 X_j^2] = \mathbb{E}[X_i^2]\mathbb{E}[X_j^2]$ *and for any* $1 \leqslant i \leqslant n$. *Furthermore, we have* $\mathbb{E}[X_i^2] = \mathrm{Var}(X_i) + \mu^2 = \sigma^2 + \mu^2$. *Therefore:*

$$\mathbb{E}[S_n^4] = nK + 3n(n-1)(\sigma^2 + \mu^2) < nK + 3n^2(\sigma^2 + \mu^2).$$

Applying Markov's Inequality with the fourth moment, we have:

$$\begin{aligned}\mathbb{P}(|S_n/n| \geqslant \epsilon) &= \mathbb{P}(|S_n| \geqslant n\epsilon) \\ &\leqslant \frac{\mathbb{E}[S_n^4]}{n^4\epsilon^4} \\ &< \frac{K}{n^3\epsilon^4} + \frac{3(\sigma^2+\mu^2)}{n^2}.\end{aligned}$$

Therefore, we have:

$$\sum_{n=1}^{\infty} \mathbb{P}(|S_n/n| \geqslant \epsilon) < \frac{K}{\epsilon^4} \sum_{n=1}^{\infty} n^{-3} + 3(\sigma^2 + \mu^2) \sum_{n=1}^{\infty} n^{-2} < \infty \tag{22}$$

Finally, by the Borel-Cantelli Lemma (**BCL**)*, we have:*

$$\mathbb{P}\left(\limsup_{n\to\infty} |S_n/n| \geqslant \epsilon \right) = 0, \quad \forall \epsilon > 0.$$

□.

Theorem 1.4: Borel-Cantelli Lemma (BCL)

1. **First Borel-Cantelli Lemma**: Given a probability space $(X, \mathcal{S}, \mathbb{P})$ and a sequence $\{A_n\}_{n=1}^{\infty} \subset \mathcal{S}$. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, we have:

$$\mathbb{P}\left(\limsup_{n\to\infty} A_n \right) = 0. \tag{23}$$

2. **Second Borel-Cantelli Lemma**: On the other hand, if $\sum_{n=1}^{\infty} \mathbb{P}(E_n) = \infty$, we have:

$$\mathbb{P}\left(\limsup_{n\to\infty} A_n \right) = 1. \tag{24}$$

Proof (Borel-Cantelli Lemma (**BCL**)).

We focus on proving the first Borel-Cantelli lemma. We define another sequence of $\mathcal{S}$-measurable sets $\{B_n\}_{n=1}^{\infty}$ such that:

$$B_n = \bigcup_{k=n}^{\infty} A_n.$$

Hence, we have $B_{\ell+1} \subset B_\ell$ for every $\ell \geqslant 1$. In other words, B_n is a decreasing sequence of $\mathcal{S}$-measurable sets. By continuity of measure, we have:

$$\begin{aligned}
\mathbb{P}\left(\lim_{n\to\infty} B_n \right) &= \lim_{n\to\infty} \mathbb{P}(B_n) \\
&= \lim_{n\to\infty} \sum_{k=n}^{\infty} \mathbb{P}(A_n) \quad \textit{(By additivity)} \\
&= \sum_{i=1}^{\infty} \mathbb{P}(A_i) - \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(A_n) \\
&= 0.
\end{aligned}$$

Furthermore, we have:

$$\mathbb{P}\left(\lim_{n\to\infty} B_n \right) = \mathbb{P}\left(\lim_{n\to\infty} \bigcup_{k=n}^{\infty} \right) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_n \right) = \mathbb{P}\left(\limsup_{n\to\infty} A_n \right).$$

Hence proved the first Borel-Cantelli Lemma. To prove the second Borel-Cantelli Lemma, we prove the following:

$$\begin{aligned}
1 - \mathbb{P}\left(\limsup_{n\to\infty} A_n \right) &= \mathbb{P}\left(\left\{ \limsup_{n\to\infty} A_n \right\}^c \right) \\
&= \mathbb{P}\left(\liminf_{n\to\infty} A_n^c \right) = 0.
\end{aligned}$$

□.

1.2.3 Uniform Law of Large Numbers

The Uniform Law of Large Numbers (**ULLN**) provides a convergence result for collection of estimators where the convergence is uniform in the parameters space.

Theorem 1.5: Uniform Law of Large Numbers (ULLN)

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution $p(.;\boldsymbol{\theta})$ over $\mathcal{X}$ that depends on the true parameters $\boldsymbol{\theta}$. Let Θ be the parameters space and $f : \Theta \times \mathcal{X} \to \mathbb{R}$ be a function that satisfies the following conditions:

1. Θ is compact.
2. $f(\theta, x)$ is continuous in θ for almost all $x \in \mathcal{X}$.
3. There exists a function K such that $\mathbb{E}_{\boldsymbol{\theta}}[K(X)] < \infty$ and $|f(\theta, x)| \leqslant K(x)$ for all $\theta \in \Theta$.

Then, we have:

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_1, \theta) - \mathbb{E}_{\boldsymbol{\theta}}[f(X, \theta)] \right| \xrightarrow{p} 0. \tag{25}$$

Proof (Theorem 1.5).

I am too lazy to include the proof. Hence, a nice proof is provided in Ferguson 1996, Theorem 16, Page 111. The (**ULLN**) *will be used later to prove the consistency of the Maximum Likelihood Estimator.* $\square$.

1.2.4 Central Limit Theorem

Theorem 1.6: Central Limit Theorem (CLT)

Let $X_1, \ldots X_n$ be a sequence of *i.i.d* random variables with expected value μ and finite variance σ^2. Then, we have:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \quad as \quad n \to \infty, \tag{26}$$

where $\bar{X}_n = S_n/n$ and $\mathcal{N}(0,1)$ is the standard normal distribution.

Proof (Central Limit Theorem (**CLT**)).

We prove this via the Characteristic Function. Let $\bar{Z}_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$*, notice that:*

$$\bar{Z}_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma},$$

Let $Z_i = X_i - \mu$ *for* $1 \leqslant i \leqslant n$ *and suppose* $Z = Z_1 = \cdots = Z_n$*, we have:*

$$\begin{aligned}
\varphi_{\bar{Z}_n}(t) = \varphi_{\sum_{i=1}^{n} Z_i}\left(\frac{t}{\sqrt{n}}\right) &= \left[\varphi_Z\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&= \left[1 + \frac{it\mathbb{E}[Z]}{\sqrt{n}} - \frac{t^2}{2n}\mathbb{E}[Z^2] + \mathcal{O}(1/n)\right]^n \quad \textit{(Taylor's Expansion)} \\
&= \left[1 - \frac{t^2}{2n} + \mathcal{O}(1/n)\right]^n.
\end{aligned}$$

The final equality comes from the fact that $\mathbb{E}[Z] = 0$ *and* $\mathbb{E}[Z^2] = \mathbb{E}[Z]^2 + \mathrm{Var}(Z) = 1$. *Finally, we have:*

$$\lim_{n\to\infty} \varphi_{\bar{Z}_n}(t) = \lim_{n\to\infty} \left[1 - \frac{t^2}{2n} + \mathcal{O}(1/n)\right]^n = e^{-t^2/2}.$$

Since $e^{-t^2/2}$ *is the Characteristic Function of the standard normal distribution, by* (**LCT**), *we have* $\bar{Z}_n \xrightarrow{d} \mathcal{N}(0,1)$. □.

1.3 Convergence of Random Variables

In this section, we revise the modes of convergence in random variables.

1.3.1 Convergence in Distribution

Definition 1.2 (Convergence in Distribution)**.**
Given a sequence of real-valued random variables $X_1, X_2, \dots$ *with CDFs* $F_1, F_2, \dots$. *We say that the sequence converges in distribution to a random variable* X *with CDF* F, *denoted* $X_n \xrightarrow{d} X$ *if:*

$$\lim_{n\to\infty} F_n(x) = F(x), \tag{27}$$

for all $x \in \mathbb{R}$ *at which* F *is continuous. Convergence in distribution can also be referred to as* *weak convergence* *in measure theory.*

1.3.2 Convergence in Probability

Definition 1.3 (Convergence in Probability)**.**
Given a sequence of real-valued random variables $X_1, X_2, \dots$. *We say that the sequence converges in probability to a random variable* X, *denoted* $X_n \xrightarrow{p} X$ *if:*

$$\lim_{n\to\infty} \mathbb{P}\Big(|X_n - X| \geqslant \epsilon\Big) = 0, \quad \forall \epsilon > 0. \tag{28}$$

We also refer to convergence in probability as *convergence in measure* *in measure theory.*

Proposition 1.1: $X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$

Let X and the sequence $X_1, X_2, \dots$ be real-valued random variables. If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$.

Proof (Proposition 1.1)**.**
We first prove the following claim: Let X, Y *be random variables,* $a \in \mathbb{R}$ *and* $\epsilon > 0$, *the inequality* $\mathbb{P}(Y \leqslant a) \leqslant \mathbb{P}(X \leqslant a + \epsilon) + \mathbb{P}(|Y - X| \geqslant \epsilon)$ *holds. We have:*

$$\begin{aligned}
\mathbb{P}(Y \leqslant a) &= \mathbb{P}(Y \leqslant a, X \leqslant a + \epsilon) + \mathbb{P}(Y \leqslant a, X \geqslant a + \epsilon) \\
&\leqslant \mathbb{P}(X \leqslant a + \epsilon) + \mathbb{P}(Y - X \leqslant a - X, a - X \leqslant -\epsilon) \\
&\leqslant \mathbb{P}(X \leqslant a + \epsilon) + \mathbb{P}(Y - X \leqslant -\epsilon) \\
&\leqslant \mathbb{P}(X \leqslant a + \epsilon) + \mathbb{P}(Y - X \leqslant -\epsilon) + \mathbb{P}(Y - X \geqslant \epsilon) \\
&= \mathbb{P}(X \leqslant a + \epsilon) + \mathbb{P}(|Y - X| \geqslant \epsilon).
\end{aligned}$$

Using the above inequality, we have:

$$\mathbb{P}(X \leqslant a-\epsilon) - \mathbb{P}(|X_n - X| \geqslant \epsilon) \leqslant \mathbb{P}(X_n \leqslant a) \leqslant \mathbb{P}(X \leqslant a+\epsilon) + \mathbb{P}(|X_n - X| \geqslant \epsilon).$$

Taking limits as $n \to \infty$ from both sides, we have:

$$F_X(a-\epsilon) \leqslant \lim_{n\to\infty} F_{X_n}(a) \leqslant F_X(a+\epsilon).$$

Taking $\epsilon \to 0^+$, we have $\lim_{n\to\infty} F_{X_n}(a) = F_X(a)$. □.

Proposition 1.2: $X_n \xrightarrow{d} c \iff X_n \xrightarrow{p} c$

Let $c \in \mathbb{R}$ be a constant and $X_1, X_2, \ldots$ be a sequence of real-valued random variables. Then, $X_n \xrightarrow{d} c \iff X_n \xrightarrow{p} c$.

Proof (Proposition 1.2, Pishro-Nik 2014).
Since $X_n \xrightarrow{d} c$, we immediately have the following:

$$\lim_{n\to\infty} F_{X_n}(c-\epsilon) = 0,$$
$$\lim_{n\to\infty} F_{X_n}(c+\epsilon/2) = 1.$$

Then, for any $\epsilon > 0$, we have:

$$\begin{aligned}
\lim_{n\to\infty} (|X_n - c| \geqslant \epsilon) &= \lim_{n\to\infty} \mathbb{P}\Big[\mathbb{P}(X_n \leqslant c-\epsilon) + \mathbb{P}(X_n \geqslant c+\epsilon)\Big] \\
&= \underbrace{\lim_{n\to\infty} F_{X_n}(c-\epsilon)}_{=0} + \lim_{n\to\infty} \mathbb{P}(X_n \geqslant c+\epsilon) \\
&\leqslant \lim_{n\to\infty} \mathbb{P}(X_n \geqslant c+\epsilon/2) \\
&= 1 - \underbrace{\lim_{n\to\infty} F_{X_n}(c+\epsilon/2)}_{=1} \\
&= 0.
\end{aligned}$$

From the above, we have $\lim_{n\to\infty} \mathbb{P}(|X_n - c| \geqslant \epsilon) = 0$ and $X_n \xrightarrow{p} c$. □.

1.3.3 Convergence in L^p norm

Definition 1.4 (Convergence in L^p norm).
Given a sequence of random variables $X_1, X_2, \ldots$ and a real number $p \in [1, \infty)$. We say that the sequence converges in L^pnorm to a random variable X, denoted as $X_n \xrightarrow{L^p} X$ if:

$$\lim_{n\to\infty} \mathbb{E}|X_n - X|^p = 0. \tag{29}$$

Proposition 1.3: $X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{p} X$

Let $p \geqslant 1$ and $X_1, X_2, \ldots$ be a sequence of real-valued random variables. Let X be a random variable, then, $X_n \xrightarrow{L^p} X \implies X_n \xrightarrow{p} X$.

Proof (Proposition 1.3).
Let $\epsilon > 0$, *we have:*

$$\begin{aligned}\mathbb{P}(|X_n - X| \geqslant \epsilon) &= \mathbb{P}(|X_n - X|^p \geqslant e^p) \quad (p \geqslant 1)\\ &\leqslant \frac{\mathbb{E}|X_n - X|^p}{\epsilon^p}. \quad (\textit{Markov's Inequality})\end{aligned}$$

Taking the limits from both sides, we have $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geqslant \epsilon) = 0$ *and* $X_n \xrightarrow{p} X$. □.

1.3.4 Almost-sure Convergence

Definition 1.5 (Convergence almost-surely).
Let $X_1, X_2, \ldots$ *be a sequence of real-valued random variables that map from a sample space* Ω. *Let* X *also be a real-valued random variable. We say that* X_n *converges almost surely to* X, *denoted as* $X_n \xrightarrow{a.s} X$, *if:*

$$\mathbb{P}\Big(\limsup_{n\to\infty} E_n\Big) = 0 \quad \textit{where} \quad E_n = \Big\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geqslant \epsilon\Big\}.$$

Remark 1.2 (Consequence of (**BCL**)).

$$\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty \implies X_n \xrightarrow{a.s} X. \tag{30}$$

Proposition 1.4: $X_n \xrightarrow{a.s} X \implies X_n \xrightarrow{p} X$

Let $X_1, X_2, \ldots$ be a sequence of real-valued random variables and also let X be a real valued random variables. If $X_n \xrightarrow{a.s} X$ then $X_n \xrightarrow{p} X$.

Proof (Proposition 1.4).
Let $f_n : \Omega \to \mathbb{R}_+$ *be a sequence of nonnegative Borel-measurable functions such that* $f_n(\omega) = |X_n(\omega) - X(\omega)|$. *By Fatou's Lemma (reverse), we have:*

$$\begin{aligned}\underbrace{\mathbb{P}\Big(\limsup_{n\to\infty}\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geqslant \epsilon\}\Big)}_{=0} &= \int f_n d\mathbb{P}\\ &\geqslant \limsup_{n\to\infty} \mathbb{P}(|X_n - X| \geqslant \epsilon)\\ &\geqslant \lim_{n\to\infty} \mathbb{P}(|X_n - X| \geqslant \epsilon).\end{aligned}$$

Hence, we have $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geqslant \epsilon) = 0$ *and* $X_n \xrightarrow{p} X$. □.

Theorem 1.7: Continuous Mapping Theorem (CMT)

Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function and $X_1, X_2, \ldots$ be a sequence of real-valued random variables. Then, the following statements hold true:

1. $X_n \xrightarrow{d} X \implies f(X_n) \xrightarrow{d} f(X)$.
2. $X_n \xrightarrow{p} X \implies f(X_n) \xrightarrow{p} f(X)$.
3. $X_n \xrightarrow{a.s} X \implies f(X_n) \xrightarrow{a.s} f(X)$.

Proof (Continuous Mapping Theorem (**CMT**)).

Since almost-sure convergence implies the other two modes of convergence, we only have to handle the almost-sure convergence case. Since h is continuous, for any $\omega \in \Omega$ such that $X_n(\omega) \to X(\omega)$, we have $f(X_n(\omega)) \to f(X(\omega))$. Therefore, we have:

$$\Big\{\omega \in \Omega : X_n(\omega) \to X(\omega)\Big\} \subseteq \Big\{\omega \in \Omega : f(X_n(\omega)) \to f(X(\omega))\Big\}.$$

Therefore, we have:

$$\begin{aligned}&\mathbb{P}\Big(\limsup_{n\to\infty}\Big\{\omega \in \Omega : |f(X_n(\omega)) - f(X(\omega))| \leqslant \epsilon\Big\}\Big)\\ &\geqslant \mathbb{P}\Big(\limsup_{n\to\infty}\Big\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \leqslant \epsilon\Big\}\Big) = 1,\end{aligned}$$

for all $\epsilon > 0$. Therefore, we have $f(X_n) \xrightarrow{a.s} f(X)$. □.

2 Statistical Inference

2.1 Sufficiency & Likelihood

2.1.1 Sufficiency

Definition 2.1 (Sufficient Statistics).

Let $\mathbf{X} = (X_1, \dots, X_n) \sim p(.;\boldsymbol{\theta})$ be a random sample drawn i.i.d from a distribution with parameters $\boldsymbol{\theta}$. Let $\mathbf{U} = T(\mathbf{X})$ be a statistic, then it is called a sufficient statistic if the conditional distribution $p_{\mathbf{X}|\mathbf{U}}$ does not depend on $\boldsymbol{\theta}$.

Example 2.1 (Bernoulli random variables).

Let $\mathbf{X} = (X_1, \dots, X_n) \sim \text{Bernoulli}(\theta)$ be a random sample from the Bernoulli distribution. Let $\mathbf{U} = \frac{1}{n}\sum_{i=1}^n X_i$, then $\mathbf{U}$ is a sufficient statistic of θ. To illustrate this, suppose that $\mathbf{x} = (x_1, \dots, x_n)$ is an observation of the random sample $\mathbf{X}$ and $\mathbf{u} = \frac{1}{n}\sum_{i=1}^n x_i$. We have:

$$\begin{aligned}
\mathbb{P}(\mathbf{X} = \mathbf{x}|\mathbf{U} = \mathbf{u}) &= \frac{\mathbb{P}(\mathbf{X} = \mathbf{x}, \mathbf{U} = \mathbf{u})}{\mathbb{P}(\mathbf{U} = \mathbf{u})} \\
&= \frac{\mathbb{P}(X_1 = x_1, \dots, X_n = x_n, \sum_{i=1}^n X_i = \sum_{i=1}^n x_i)}{\mathbb{P}(\sum_{i=1}^n X_i = \sum_{i=1}^n x_i)} \\
&= \frac{\mathbb{P}(X_1 = x_1, \dots, X_n = x_n)}{\mathbb{P}(\sum_{i=1}^n X_i = \sum_{i=1}^n x_i)} \\
&= \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\mathbb{P}(\sum_{i=1}^n X_i = \sum_{i=1}^n x_i)}.
\end{aligned}$$

Now, setting $k = \sum_{i=1}^n x_i$, The denominator is basically the probability that the Bernoulli variables sums up to k. Hence, we can calculate the denominator as follows:

$$\mathbb{P}\left(\sum_{i=1}^n X_i = k\right) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

Therefore, we have:

$$\mathbb{P}(\mathbf{X} = \mathbf{x}|\mathbf{U} = \mathbf{u}) = \frac{\theta^k(1-\theta)^{n-k}}{\binom{n}{k}\theta^k(1-\theta)^{n-k}} = \frac{1}{\binom{n}{k}}.$$

Therefore, the conditional distribution does not depend on θ and $\mathbf{U}$ is a sufficient statistic.

Definition 2.2 (Sufficiency Principle).

*If $\mathbf{U} = T(\mathbf{X})$ is a sufficient statistic for $\boldsymbol{\theta}$, then any inference about $\boldsymbol{\theta}$ should only depend on the sample $\mathbf{X}$ through $\mathbf{U}$. In other words, if we estimate $\boldsymbol{\theta}$ using an estimator $\hat{\boldsymbol{\theta}}$, only $\mathbf{U}$ shows up in the formula of $\hat{\boldsymbol{\theta}}$, not the sample $\mathbf{X}$ itself. We will see why this is the case in the Factorisation Theorem (**FacT**), which states that we can factorise the density function into a function of $\mathbf{U}, \boldsymbol{\theta}$ and a function of the observations $\mathbf{x}$ and thus, the inference about $\boldsymbol{\theta}$ is independent of the observations $\mathbf{x}$.*

Theorem 2.1: Factorisation Theorem (FacT)

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample with joint density function $p(\mathbf{x}; \boldsymbol{\theta})$ over $\mathcal{X}^n$. The statistic $\mathbf{U} = T(\mathbf{X})$ is sufficient for the parameters $\boldsymbol{\theta}$ if and only if we can find functions h, g such that:

$$p(\mathbf{x}; \boldsymbol{\theta}) = g(T(\mathbf{x}), \boldsymbol{\theta})h(\mathbf{x}),$$

for all $\mathbf{x} \in \mathbb{R}^n$ and $\boldsymbol{\theta} \in \Theta$.

Proof (Factorisation Theorem (**FacT**)).
We have to conduct the proof in both directions.

- $T(\mathbf{X})$ *is sufficient* $\Longrightarrow$ *Factorisation exists: Let* $\mathbf{U} = T(\mathbf{X})$ *be a sufficient statistics and* $\mathbf{u} = T(\mathbf{x})$ *be the statistics evaluated on the observations* $\mathbf{x}$*. Then, we have:*

$$\begin{aligned} p(\mathbf{x}; \boldsymbol{\theta}) &= \mathbb{P}(\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}) \\ &= \mathbb{P}(\mathbf{X} = \mathbf{x} | \mathbf{U} = \mathbf{u}; \boldsymbol{\theta})\mathbb{P}(\mathbf{U} = \mathbf{u}; \boldsymbol{\theta}). \end{aligned}$$

  Since $\mathbf{U} = T(\mathbf{X})$ *is a sufficient statistics,* $\mathbb{P}(= \mathbf{x} | \mathbf{U} = \mathbf{u}; \boldsymbol{\theta})$ *does not depend on* $\boldsymbol{\theta}$*. Hence, we denote* $h(\mathbf{x}) = \mathbb{P}(\mathbf{X} = \mathbf{x} | \mathbf{U} = \mathbf{u}; \boldsymbol{\theta})$*. Furthermore,* $\mathbb{P}(\mathbf{U} = \mathbf{u}; \boldsymbol{\theta})$ *is a function of* $\mathbf{u}$ *and* $\boldsymbol{\theta}$*. We denote this function as* $g(\mathbf{u}, \boldsymbol{\theta})$ *and conclude that the factorisation* $p(\mathbf{x}; \boldsymbol{\theta}) = h(\mathbf{x})g(T(\mathbf{x}), \boldsymbol{\theta})$ *indeed exists.*

- *Factorisation exists* $\Longrightarrow$ $T(\mathbf{X})$ *is sufficient: Suppose that there exists* g, h *such that* $p(\mathbf{x}; \boldsymbol{\theta}) = g(T(\mathbf{x}), \boldsymbol{\theta})h(\mathbf{x})$*. We then have:*

$$\mathbb{P}(\mathbf{X} = \mathbf{x} | \mathbf{U} = \mathbf{u}; \boldsymbol{\theta}) = \frac{p(\mathbf{x}; \boldsymbol{\theta})}{\mathbb{P}(\mathbf{U} = \mathbf{u}; \boldsymbol{\theta})} = \frac{g(\mathbf{u}, \boldsymbol{\theta})h(\mathbf{x})}{\mathbb{P}(\mathbf{U} = \mathbf{u}; \boldsymbol{\theta})}.$$

  We denote $A_{\mathbf{u}} = \left\{\tilde{\mathbf{x}} \in \mathcal{X}^n : T(\tilde{\mathbf{x}}) = \mathbf{u}\right\}$*. We have:*

$$\begin{aligned} \mathbb{P}(\mathbf{U} = \mathbf{u}; \boldsymbol{\theta}) &= \sum_{\tilde{\mathbf{x}} \in A_{\mathbf{u}}} \mathbb{P}(\mathbf{X} = \tilde{\mathbf{x}}) \\ &= \sum_{\tilde{\mathbf{x}} \in A_{\mathbf{u}}} p(\tilde{\mathbf{x}}; \boldsymbol{\theta}) = \sum_{\tilde{\mathbf{x}} \in A_{\mathbf{u}}} g(T(\tilde{\mathbf{x}}), \boldsymbol{\theta})h(\tilde{\mathbf{x}}) \\ &= g(\mathbf{u}, \boldsymbol{\theta}) \sum_{\tilde{\mathbf{x}} \in A_{\mathbf{u}}} h(\tilde{\mathbf{x}}). \end{aligned}$$

  From the above, we have:

$$\mathbb{P}(\mathbf{X} = \mathbf{x} | \mathbf{U} = \mathbf{u}; \boldsymbol{\theta}) = \frac{h(\mathbf{x})}{\sum_{\tilde{\mathbf{x}} \in A_{\mathbf{u}}} h(\tilde{\mathbf{x}})},$$

  and the above expression does not depend on $\boldsymbol{\theta}$*. Hence,* $T(\mathbf{X})$ *is a sufficient statistics.*

2.1.2 Likelihood

Definition 2.3 (Likelihood Function).
Let $\mathbf{X} = (X_1, \ldots, X_n)$ *be a random sample from a distribution* $p(.|\theta)$ *that depends on parameters* $\theta \in \Theta$*. Let* $\mathbf{x} = (x_1, \ldots, x_n)$ *be an observation of the random sample* $\mathbf{X}$*. Then, the likelihood function* $L(\theta; \mathbf{x})$ *is defined as follows:*

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} p(x_i; \theta), \quad \theta \in \Theta. \tag{31}$$

In some cases, we also use the log-likelihood function:

$$\ell(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}) = \sum_{i=1}^{n} \log p(x_i; \theta), \quad \theta \in \Theta. \tag{32}$$

Essentially, $L(\theta; \mathbf{x})$ quantifies the likelihood that θ generates the observations $\mathbf{x}$. In a way, it is the inverse of probability density (mass) functions, we can see the contrast as follows:

- ***Probability Density Function**: The parameters are fixed but the observations are random.*
- ***Likelihood Function**: The observations are fixed but the parameters are variable.*

Definition 2.4 (Maximum Likelihood Estimator)**.**
Given $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution $p(.; \theta)$ that depends on $\theta \in \Theta$ and $\mathbf{x} = (x_1, \ldots, x_n)$ be an observation of $\mathbf{X}$. The Maximum Likelihood Estimator $\theta_{MLE} \in \Theta$ is the parameter that maximizes the likelihood function:

$$\theta_{MLE} = \arg\max_{\theta \in \Theta} L(\theta; \mathbf{x}). \tag{33}$$

In the subsequent propositions, we will discuss some of the key properties of MLE.

Proposition 2.1: Consistency of MLE

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution $p(.; \boldsymbol{\theta})$ dependent on a true set of parameters $\boldsymbol{\theta}$. Let Θ be the parameters space. Then, the Maximum Likelihood Estimator $\theta_{MLE} = \arg\max_{\theta \in \Theta} L(\theta; \mathbf{X})$, which is a random variable, is consistent:

$$\lim_{n \to \infty} \mathbb{P}(|\theta_{MLE} - \boldsymbol{\theta}| \geqslant \epsilon) = 0, \quad \forall \epsilon > 0, \tag{34}$$

provided that the following conditions are met:

1. *Θ is a compact space.*
2. *$p(x; \theta)$ is continuous in θ for almost all x.*
3. *There exists a function K such that $\mathbb{E}_{\boldsymbol{\theta}}[K(X)] < \infty$ and $|\log p(x; \theta)| < K(x)$.*

Furthermore, we can also show that θ_{MLE} is asymptotically unbiased. In other words, $\lim_{n \to \infty} \mathbb{E}[\theta_{MLE}] = \boldsymbol{\theta}$.

Proof (Proposition 2.1)**.**
For properties (1) and (2), we can just prove the consistency of MLE because consistency implies asymptotic unbiasedness (We will provide a proof for this claim as a lemma below).

1. *Consistency: By (**SLLN**), for any $\theta \in \Theta$, we have the following*

$$n^{-1} \log L(\theta; \mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} \log p(X_i; \theta) \xrightarrow{a.s} \mathbb{E}_{\boldsymbol{\theta}}[\log p(X; \theta)], \tag{35}$$

where $\mathbb{E}_{\boldsymbol{\theta}}$ denotes the expectation taken over the distribution corresponding to the true parameters $\boldsymbol{\theta}$. We denote $q(\theta_2|\theta_1)$ as $\mathbb{E}_{\theta_1}[\log p(X; \theta_2)]$. Then, we want to prove that $\mathbb{E}_{\boldsymbol{\theta}}[\log p(X; \boldsymbol{\theta})] \geqslant \mathbb{E}_{\boldsymbol{\theta}}[\log p(X; \theta)]$ for all $\theta \in \Theta$ and $X \sim p(.|\boldsymbol{\theta})$. To do so, we prove the *following inequality:*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\log p(X;\theta)}{\log p(X;\boldsymbol{\theta})}\right] \leqslant 0, \quad \forall \theta \in \Theta. \tag{36}$$

□.

Proposition 2.2: Asymptotic Normality of MLE

2.2 Point Estimation

2.2.1 Bias, Variance, Consistency and MSE

2.2.2 Sufficient Statistics & Rao-Blackwell Theorem

Theorem 2.2: Rao-Blackwell Theorem (RB)

2.2.3 Estimator Variance & Cramer-Rao Lower Bound

Definition 2.5 (Fisher Information)**.**
Let $\mathbf{X} = (X_1, \ldots, X_n) \sim p(.;\boldsymbol{\theta})$ be a random sample from a distribution parameterized by $\boldsymbol{\theta}$. The (total) Fisher Information about θ in the random sample $\mathbf{X}$ is defined as follows:

$$\mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{X}}\left[\left(\frac{\partial}{\partial\theta}\log L(\theta;\mathbf{X})\right)^2 \;\middle|\; \boldsymbol{\theta}\right]. \tag{37}$$

The Fisher Informaton is the total information about $\boldsymbol{\theta}$ contained in the sample $\mathbf{X}$.

Theorem 2.3: Cramer-Rao Lower Bound (CRLB)

2.2.4 Maximum Likelihood Estimation (MLE)

A List of Definitions

B Important Theorems

C Important Propositions

D References

References

Durrett, Rick (2010). *Probability: Theory and Examples*. 4th. USA: Cambridge University Press. ISBN: 0521765390.

Ferguson, Thomas S. (1996). *A Course in Large Sample Theory*. Chapman & Hall.

Gut, Allan (2004). *A Graduate Course in Probability*. Graduate Text in Mathematics.

Pishro-Nik, Hossein (2014). *Introduction to Probability, Statistics and Random Processes*. Kappa Research, LLC.

undefinedinlar, Erhan (2011). *Probability and Stochastics*. Springer New York. ISBN: 9780387878591. URL: http://dx.doi.org/10.1007/978-0-387-87859-1.

Wikipedia (2023). *Vitali set — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Vitali%20set&oldid=1187241923. [Online; accessed 24-December-2023].