2 Basics of Probability Theory

2.1 Basic Definitions

In this section, we will define and discuss the most basic notions in Probability Theory, including probability spaces, random variables, expectation, characteristic function, and so on.

From the point of view of Analysis, Probability Theory can be seen as a particular case where the total mass of the measure is equal to 1. This means that the most of the results and theorems we reviewed in Chapter 1 are still valid. Moreover, from the point of view of Probability Theory, the probability space itself is not so much important; instead, one is more interested in the values taken by a *random variable* along with their *frequencies*.

We have the following correspondance between different notions:

probability space	$\longleftrightarrow$	measure space
events	$\longleftrightarrow$	elements of a σ-algebra
random variable	$\longleftrightarrow$	measurable function
expectation	$\longleftrightarrow$	integral

We are going to give more precise definitions to the aforementioned notions with examples.

2.1.1 Probability Spaces

We reviewed the definition of measurable spaces in Section 1.1. In Probability Theory, the probability spaces we will define are just particular cases of such spaces.

Definition 2.1.1 : Let $\mathbb{P}$ be a measure of total mass 1 on the measurable space $(\Omega, \mathcal{A})$. We call $\mathbb{P}$ a *probability measure* (機率測度) and $(\Omega, \mathcal{A}, \mathbb{P})$ a *probability space* (機率空間) 。

- The set Ω is caclled *sample space* (様本空間) , which can be regarded as the space of "randomness units" in our random experiment.
- The set $\mathcal{A}$ contains all the *measurable events*, also called *events*, which are subsets of Ω whose "probability" can be *measured*. In other words, an element $A \in \mathcal{A}$ is a subset of Ω, including "random units" satisfying some particular conditions.
- For any $A \in \mathcal{A}$, the quantity $\mathbb{P}(A)$ describes the *probability* that the measurable event A occurs.
- If the sample space Ω is discrete (finite or countably infinite), then we may consider $\mathcal{A} = \mathcal{P}(\Omega)$ and any probability measure $\mathbb{P}$ defined on the measurable space $(\Omega, \mathcal{A})$ is called a discrete probability (離散機率), and the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is called a discrete probability space (離散機率空間).

Example 2.1.2 (Discrete probability spaces) **:** Given a fair dice with numbers 1 to 6.

(1) Consider a random experiment where the dice is drawn twice. We can set the probability space to be $\Omega = \{1, \dots, 6\}^2$ and the set of measurable events $\mathcal{A} = \mathcal{P}(\Omega)$. Then, the probability measure $\mathbb{P}$ satisfies $\mathbb{P}(A) = \frac{|A|}{36}$ for all $A \in \mathcal{A}$.

(2) Consider an unfair coin whose Head appears with twice the probability of tail. We toss the coin three times. The probability space is equal to,

$$\Omega = \{(\text{Head}, \text{Head}, \text{Head}), (\text{Head}, \text{Head}, \text{Tail}), (\text{Head}, \text{Tail}, \text{Head}), (\text{Head}, \text{Tail}, \text{Tail}),$$
$$(\text{Tail}, \text{Head}, \text{Head}), (\text{Tail}, \text{Head}, \text{Tail}), (\text{Tail}, \text{Tail}, \text{Head}), (\text{Tail}, \text{Tail}, \text{Tail})\},$$

The set of measurable events is $\mathcal{A} = \mathcal{P}(\Omega)$ and the probability measure $\mathbb{P}$ is defined as $\mathbb{P}(\omega) = (\frac{2}{3})^{H(\omega)}(\frac{1}{3})^{T(\omega)}$, where for any $\omega \in \Omega$, we denote $H(\omega)$ and $T(\omega)$ the number of times that head and tail appear.

Example 2.1.3 (General probability space) **:** The unit circle in the plane is denoted by

$$\mathbb{S}^1 := \{z \in \mathbb{C} : |z| = 1\} \simeq \mathbb{R}/(2\pi\mathbb{Z}),$$

and can be interpreted as unit directions in the two-dimensional space. If we consider the measurable space $(\mathbb{S}^1, \mathcal{B}(\mathbb{S}^1))$ and the probability measure

$$\mathbb{P}([a, b]) = \frac{b-a}{2\pi}, \quad a \leqslant b, |b-a| \leqslant 2\pi$$

then $\mathbb{P}$ is a *uniform measure* defined on $\mathbb{S}^1 = \mathbb{R}/(2\pi\mathbb{Z})$. This may also be seen as the probability measure on the quotient space $\mathbb{S}^1$ "induced" by the Lebesgue measure on $\mathbb{R}$.

2.1.2 Random Variables

In Probability Theory, the probability space is not so important eventually. What we are interested in is what we can observe and measure in a random experiment, and the frequencies of different outcomes. Hence, below we will define the notion of *random variables* and see it as "the outcome of the unit random event in a random experiment".

Definition 2.1.4 : Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(E, \mathcal{E})$ be a measurable space. A measurable function $X : \Omega \longrightarrow E$ is called a *random variable* (隨機變數) with values in E.

Example 2.1.5 : We take the examples from Example 2.1.2 in the previous section, and consider random variables defined above.

(1) Let $X((i, j)) = i + j$. Then X is a random variable with values in $\{2, \dots, 12\}$.

(2) Let $H(\omega)$ be "the number of heads in ω". Then, H is a random variable with values in $\{0, 1, 2, 3\}$.

Example 2.1.6 : Consider the probability space defined as in Example 2.1.3 and the following random variables,

$$X(\omega) = \cos(\omega), \qquad Y(\omega) = \sin(\omega), \qquad \forall \omega \in \mathbb{S}^1.$$

Then, the random variables X and Y take values in $[-1, 1]$ can be regarded as the projections of the uniform random direction on the x-axis and the y-axis.

The domain of definition (i.e. probability space) of a random variable $X : \Omega \longrightarrow E$ is not so relevant, we care more about its domain of arrival. More precisely, we want to know the probability that a random variable takes a (some) particular value(s). In other words, we want to understand the *image measure* of $\mathbb{P}$ under X.

Definition 2.1.7 : Let $\mathbb{P}_X$ be the *image measure* (影像測度) of $\mathbb{P}$ under the random variable X. It is called the *distribution* (分佈) or the *law* (律) of X. In other words, $\mathbb{P}_X$ is a probability measure on the measurable space $(E, \mathcal{E})$ and can be written as,

$$\mathbb{P}_X(B) := \mathbb{P}(X^{-1}(B)), \qquad \forall B \in \mathcal{E}.$$

In the language of Probability Theory, the above quantity is also abbreviated as

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(\omega \in \Omega : X(\omega) \in B), \qquad \forall B \in \mathcal{E}.$$

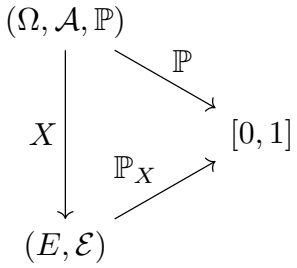

Figure 2.1: The diagram illustrating the image measure $\mathbb{P}_X$, or the pushforward (向前推進) measure of $\mathbb{P}$ by the random variable X, also denoted $X_*\mathbb{P}$. What we will care about is the probability measure $\mathbb{P}_X$ instead of $\mathbb{P}$.

One should understand the image measure as follows. Given any point ω in the probability space Ω, the image $X(\omega)$ can be seen as a point in E and $\mathbb{P}_X(B)$ the probability that this point is in B.

Example 2.1.8 : Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, where $\Omega = [0, \pi]$, $\mathcal{A}$ is the Borel σ-algebra, and $\mathbb{P} = \frac{1}{\pi}\lambda$. Consider the random variable $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}$ defined by

$$X(\omega) = \sin(\omega), \quad \omega \in \Omega = [0, \pi].$$

The image measure $\mathbb{P}_X$ is defined on the measurable space $([0, 1], \mathcal{B}([0, 1]))$. To characterize it, it is

enough to know $\mathbb{P}_X([a,b])$ for all $0 \leqslant a \leqslant b \leqslant 1$. For $0 \leqslant a \leqslant b \leqslant 1$, we have

$$\begin{aligned}\mathbb{P}_X([a,b]) &= \mathbb{P}(\omega \in \Omega : \sin(\omega) \in [a,b]) \\ &= \mathbb{P}(\omega \in \Omega : \omega \in [\sin^{-1} a, \sin^{-1} b]) + \mathbb{P}(\omega \in \Omega : \omega \in [\pi - \sin^{-1} b, \pi - \sin^{-1} a]) \\ &= \frac{2}{\pi}(\sin^{-1} b - \sin^{-1} a).\end{aligned}$$

Remark 2.1.9 : If two random variables $X : (\Omega_1, \mathcal{A}_1, \mathbb{P}_1) \longrightarrow (E, \mathcal{E})$ and $Y : (\Omega_2, \mathcal{A}_2, \mathbb{P}_2) \longrightarrow (E, \mathcal{E})$ have the same distribution, that is, $\mathbb{P}_{1,X} := X_* \mathbb{P}_1 = Y_* \mathbb{P}_2 =: \mathbb{P}_{2,Y}$, then we may write

$$X \overset{(\mathrm{d})}{=} Y \qquad \text{or} \qquad X \sim Y \qquad \text{or} \qquad X \sim \mathbb{P}_{2,Y}.$$

If we want to say that X has distribution μ, that is $\mathbb{P}_{1,X} = \mu$, then we may write

$$X \sim \mu,$$

and say that X follows the distribution of μ.

Remark 2.1.10 : In the case that $(E, \mathcal{E}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $\mathbb{P}_X$ is a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. If it is absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}^d$, then by the Radon–Nikodym theorem (Theorem 1.3.16), we can find a density function $g : \mathbb{R}^d \longrightarrow \mathbb{R}_+$ such that $\mathbb{P}_X$ can be written as $\mathbb{P}_X = g \cdot \lambda$. We call g the probability density function (機率密度函數) of the random variable X.

Remark 2.1.11 (Canonical construction of a random variable) **:** If μ is a probability measure defined on $\mathbb{R}^d$, we can construct a random variable whose distribution is μ. Consider $\Omega = \mathbb{R}^d$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^d)$, $\mathbb{P} = \mu$ and let $X(\omega) = \omega$. We can easily check that the distribution of X is μ. Later in Proposition 2.1.23, when a probability distribution μ in $\mathbb{R}$ is given, we will see how to construct a real-valued random variable with the distribution μ starting from the uniform one.

2.1.3 Expectation

The first quantity of interest, after we define the notion of random variables, is *expectation*.

Definition 2.1.12 : Let X be an real-valued random variable (also called a real random variable). Then we can define its *expectation* (期望值) if one of the following conditions is satisfied,

- X is non-negative,
- X can have either sign and $\int |X| \, \mathrm{d}\mathbb{P} < \infty$.

In this case, we write

$$\mathbb{E}[X] = \int_\Omega X(\omega) \, \mathbb{P}(\mathrm{d}\omega). \tag{2.1}$$

Let $X = (X_1, \ldots, X_d)$ be a d-dimensional real random variable (or $\mathbb{R}^d$-valued). If all the expecations $\mathbb{E}[X_i]$ are well-defined, then we define the expectation of X to be $\mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])$.

Remark 2.1.13 : Let B be a measurable subset and $X = \mathbb{1}_B$. Then, $\mathbb{E}[X] = \mathbb{P}(B)$. Generally speaking, we can view $\mathbb{E}[X]$ as the "average" of the random variable X. For example, in the case that Ω is a finite set and $\mathbb{P}$ is the uniform distribution, the quantity $\mathbb{E}[X]$ represents the (weighted) average of all the possible values taken by X.

Proposition 2.1.14 : *Let X be a random variable with values in $(E, \mathcal{E})$. Then, for any measurable function $f : E \longrightarrow [0, \infty]$, we have*

$$\mathbb{E}[f(X)] = \int_E f(x)\mathbb{P}_X(\mathrm{d}x).$$

Proof : First, we show the statement for indicator functions $f = \mathbb{1}_B$ where $B \in \mathcal{E}$ is any measurable set. Then using linear combinations, the statement also holds for any non-negative simple function. To conclude, we use the fact that a non-negative measurable function is the non-decreasing limit of a sequence of non-negative simple functions (Proposition 1.2.14), then by the monotone convergence theorem, the statement is true. □

Remark 2.1.15 : Even if f is not non-negative, as long as the expectation $\mathbb{E}[|f(X)|]$ is finite, the above statement still holds. We recall that the integral of a general function without sign constraints is defined in the same way, see Definition 1.2.18.

Remark 2.1.16 : From above, we know that the probability distribution $\mathbb{P}_X$ allows us to compute the expectation of any random variable of type $f(X)$ easily. For its converse, we may consider $f = \mathbb{1}_B$ for any measurable set $B \in \mathcal{E}$, then we find $\mathbb{E}[f(X)] = \mathbb{P}(X \in B) = \mathbb{P}_X(B)$. We should not forget that the probability measure $\mathbb{P}_X$ is defined on $\mathcal{E}$, meaning that $(\mathbb{P}_X(B))_{B\in\mathcal{E}}$ uniquely determines the probability measure. Alternatively speaking, if we can write

$$\mathbb{E}[f(X)] = \int f \,\mathrm{d}\nu, \tag{2.2}$$

for "enough number" of measurable functions f, then the above discussion allows us to recover the distribution of the random variable X, which is given by the probability measure ν. In practice, we may for instance compute Eq. (2.2) for all non-negative measurable functions or all functions in $\mathcal{C}_c(\mathbb{R}^d)$.

Example 2.1.17 : We look at the same example as in Example 2.1.8 and compare their computations. Recall that $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space, where $\Omega = [0, \pi]$, $\mathcal{A}$ is the Borel σ-algebra, and $\mathbb{P} = \frac{1}{\pi}\lambda$. The random variable $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}$ is defined by

$$X(\omega) = \sin(\omega), \quad \omega \in \Omega = [0, \pi].$$

We may characterize the distribution of X by computing $\mathbb{E}[f(X)]$ for all non-negative measurable

functions $f : \mathbb{R} \to \mathbb{R}$. For a non-negative measurable function $f : \mathbb{R} \to \mathbb{R}$, we write

$$\begin{aligned}\mathbb{E}[f(X)] = \mathbb{E}[f(X(\omega))] &= \frac{1}{\pi}\int_0^\pi f(\sin\omega)\,\mathrm{d}\omega \\ &= \frac{2}{\pi}\int_0^{\pi/2} f(\sin\omega)\,\mathrm{d}\omega \\ &= \frac{2}{\pi}\int_0^1 f(x)\frac{\mathrm{d}x}{\sqrt{1-x^2}},\end{aligned}$$

where in the second line, we use the symmetry $\omega \mapsto \pi - \omega$ of the integrand; and in the third line, we apply the change of variables $x = \sin\omega$. This means that the image measure $\mathbb{P}_X$ writes

$$\mathbb{P}_X(\mathrm{d}x) = \frac{2}{\pi}\frac{\mathrm{d}x}{\sqrt{1-x^2}}.$$

It is an absolutely continuous measure with respect to the Lebesgue measure, with density given by $x \mapsto \frac{2}{\pi}\frac{1}{\sqrt{1-x^2}}$.

The following proposition gives an important example that illustrates such an idea.

Proposition 2.1.18 : *Let $(X_1, \dots, X_d)$ be a random variable with values in $\mathbb{R}^d$. Assume that it has a probability density function $p(x_1, \dots, x_d)$. Then, for any $1 \leqslant j \leqslant d$, the random variable X_j also has a probability density function which is written as,*

$$p_j(x) = \int_{\mathbb{R}^{d-1}} p(x_1, \dots, x_{j-1}, x, x_{j+1}, \dots, x_d)\,\mathrm{d}x_1 \dots \mathrm{d}x_{j-1}\,\mathrm{d}x_{j+1} \dots \mathrm{d}x_d.$$

Remark 2.1.19 : In the case of a bi-dimensional random variable ($d = 2$), we have,

$$p_1(x) = \int_{\mathbb{R}} p(x, y)\,\mathrm{d}y, \qquad p_2(y) = \int_{\mathbb{R}} p(x, y)\,\mathrm{d}x.$$

Proof : Let π_j be the projection on the j-th coordinate, $\pi_j(x_1, \dots, x_d) = x_j$. Then, for any non-negative measurable function $f : \mathbb{R} \longrightarrow \mathbb{R}_+$, we can apply Fubini's theorem,

$$\begin{aligned}\mathbb{E}[f(X_j)] = \mathbb{E}[f(\pi_j(X))] &= \int_{\mathbb{R}^d} f(x_j)p(x_1, \dots, x_d)\,\mathrm{d}x_1 \dots \mathrm{d}x_d \\ &= \int_{\mathbb{R}} f(x_j)\Big(\int_{\mathbb{R}^{d-1}} p(x_1, \dots, x_d)\,\mathrm{d}x_1 \dots \mathrm{d}x_{j-1}\,\mathrm{d}x_{j+1} \dots \mathrm{d}x_d\Big)\,\mathrm{d}x_j \\ &= \int_{\mathbb{R}} f(x_j)p_j(x_j)\,\mathrm{d}x_j.\end{aligned}$$

□

Remark 2.1.20 : If $X = (X_1, \ldots, X_d)$ is a d-dimensional real random variable, the distribution of X_j is called the *marginal distribution* (**邊緣分佈**) , denoted $\mathbb{P}_{X_j}$. From the above proposition, we can easily show that $\mathbb{P}_{X_j} = (\pi_j)_* \mathbb{P}_X$, meaning that we can recover the law of X_j from the law of X. However, it is important to note that knowing all the marginal distributions $\mathbb{P}_{X_j}$ *does not allow us* to recover the law of X.

Question 2.1.21: Construct two bi-dimensional real random variables $X = (X_1, X_2)$ and $X' = (X_1', X_2')$ such that for $j = 1, 2$, X_j and X_j' have the same marginal distribution, where as the distributions of X and X' are not equal.

2.1.4 Cumulative Distribution Function

Definition 2.1.22 : Let X be a real-valued random variable. We define the function $F_X : \mathbb{R} \longrightarrow [0, 1]$ by

$$F_X(t) = \mathbb{P}(X \leqslant t) = \mathbb{P}_X((-\infty, t]), \qquad t \in \mathbb{R}.$$

It is called the *cumulative distribution function* (**累積分佈函數**) of the random variable X, denoted c.d.f., or *distribution function* (**分佈函數**) .

We recall that from Theorem 1.2.29, we know that F_X is a non-descreasing and right-continuous function with limits equal to 0 and 1 at $-\infty$ and $+\infty$. Moreover, we know that the converse also holds. In other words, if F is a function satisfying the above properties, then we can find a unique probability measure μ such that for all $t \in \mathbb{R}$, we have $F(t) = \mu((-\infty, t])$. This means that F is the cumulative distribution function of a real-valued random variable.

Additionally, the cumulative distribution function F_X characterizes the distribution $\mathbb{P}_X$ of the random variable X. In particular, we have,

$$\begin{aligned} \mathbb{P}(a \leqslant X \leqslant b) &= F_X(b) - F_X(a-), && a \leqslant b, \\ \mathbb{P}(a < X < b) &= F_X(b-) - F_X(a), && a < b, \end{aligned}$$

where the discontinuities of F_X correspond to atoms of $\mathbb{P}_X$.

If $X := (X_1, \ldots, X_d)$ is a random variable with values in $\mathbb{R}^d$, then we may define the above notion in a similar way. For any $(t_1, \ldots, t_d) \in \mathbb{R}^d$, let us define

$$F_X(t_1, \ldots, t_d) := \mathbb{P}(X_1 \leqslant t_1, \ldots, X_d \leqslant t_d) = \mathbb{P}_X\big((-\infty, t_1] \times \ldots \times (-\infty, t_d]\big).$$

Proposition 2.1.23 : *Let $F : \mathbb{R} \longrightarrow [0, 1]$ be non-decreasing and right-continuous function with limits equal to 0 and 1 at $-\infty$ and $+\infty$. Define the function $h : (0, 1) \longrightarrow \mathbb{R}$ by*

$$h(y) := \inf\{z \in \mathbb{R} : F(z) \geqslant y\}, \qquad \forall y \in (0, 1). \tag{2.3}$$

If Y is a random variable such that $\mathbb{P}_Y$ follows the Lebesgue distribution on $[0, 1]$, then the distribution function of the random variable $h(Y)$ is F, that is $F_{h(Y)} = F$.

Remark 2.1.24 : In the statement of this proposition, if F is bijective (i.e. strictly increasing), then we have $h = F^{-1}$.

Proof : Fix $y \in (0,1)$ and $x \in \mathbb{R}$. If $F(x) \geqslant y$, we find $x \geqslant h(y)$ by the definition of h. Conversely, if $F(x) < y$, using the right continuity of F, we may find $\varepsilon > 0$ such that $F(x+\varepsilon) < y$. Then, by the monotonicity, for any $z \leqslant x + \varepsilon$, we have $F(z) < y$, so $h(y) \geqslant x + \varepsilon > x$. Therefore, the following equivalence relation holds,

$$x \geqslant h(y) \Leftrightarrow F(x) \geqslant y.$$

Let Y be a random variable such that $\mathbb{P}_Y$ is the Lebesgue measure on $[0,1]$ and $Z := h(Y)$. We note that Z is also a random variable by Proposition 1.2.2. Since $\mathbb{P}(Y \in (0,1)) = 1$, we find

$$F_Z(x) = \mathbb{P}(h(Y) \leqslant x) = \mathbb{P}(Y \leqslant F(x)) = F_Y(F(x)) = F(x).$$

□

Remark 2.1.25 : By Proposition 2.1.23, we know that if the uniform random variable exists (which is our assumption and its construction will be done in the course of Measure theory), then for any given distribution on $\mathbb{R}$, we may construct a real random variable with the given distribution using its distribution function.

2.2 σ-algebra Generated by a Random Variable

Let $(E, \mathcal{E})$ be a measurable space.

Definition 2.2.1 : For a random variable $X : (\Omega, \mathcal{A}) \longrightarrow (E, \mathcal{E})$, we define $\sigma(X)$ to be the smallest sub-σ-algebra of $\mathcal{A}$ such that X is measurable, called the *σ-algebra generated by X*,

$$\sigma(X) = \{A = X^{-1}(B) : B \in \mathcal{E}\}.$$

Remark 2.2.2 : We may interprete $\sigma(X)$ as the smallest σ-algebra allowing us to correctly describe the random variable X.

Example 2.2.3 : Fix a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and consider a random variable $X : (\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

(1) If the random variable X is a constant, then $\sigma(X) = \{\varnothing, \Omega\}$ is the trivial σ-algebra.

(2) Given a measurable set $A \in \mathcal{A}$, if the random variable X writes $X = \mathbb{1}_A$, then $\sigma(X) = \{\varnothing, A, A^c, \Omega\}$.

(3) If a random variable X only takes values 0 and 1, then there exists $A \in \mathcal{A}$ such that $X = \mathbb{1}_A$, so $\sigma(X)$ is as described in the previous case.

(4) It is not hard to check that, for any $A, B \in \mathcal{A}$ with $A \notin \{\varnothing, B, B^c, \Omega\}$, the random variable $\mathbb{1}_A$ is not $\sigma(\mathbb{1}_B)$-measurable. We see that the notion of random variables depends on the σ-algebra with which we equip the sample space.

Example 2.2.4 : Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P}) := ([0,2), \mathcal{B}([0,2)), \frac{1}{2}\lambda)$, where λ is the Lebesgue measure. We define the random variables $X, Y : (\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow ([0,2), \mathcal{B}([0,2)))$ by $X(\omega) = \omega$ and $Y(\omega) = \lfloor \omega \rfloor$. Then,

$$\sigma(X) = \mathcal{B}([0,2)), \qquad \text{and} \qquad \sigma(Y) = \{\varnothing, [0,1), [1,2), [0,2)\}.$$

We may note that $([0,2), \sigma(Y), \mathbb{P})$ is actually a discrete probability space. In fact, define

$$\Omega' := \{0,1\}, \quad \mathcal{A}' := \{\varnothing, \{0\}, \{1\}, \{0,1\}\}, \quad \text{and} \quad \mathbb{P}' := \tfrac{1}{2}\delta_0 + \tfrac{1}{2}\delta_1,$$

then $(\Omega, \mathcal{A}, \mathbb{P})$ and $(\Omega', \mathcal{A}', \mathbb{P}')$ have the *same* structure.

Remark 2.2.5 : We can generalize this definition to any collection of random variables $X = (X_i)_{i \in I}$. Assume that for any $i \in I$, the random variable X_i takes its values in $(E_i, \mathcal{E}_i)$. Then, we define,

$$\sigma(X) = \sigma(X_i^{-1}(B_i) : B_i \in \mathcal{E}_i, i \in I).$$

Proposition 2.2.6 : *Let X be a random variable with values in $(E, \mathcal{E})$ and Y be another real-valued random variable. The following properties are equivalent.*

(1) Y is measurable with respect to $\sigma(X)$.

(2) There exists a measurable function $f : (E, \mathcal{E}) \longrightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $Y = f(X)$.

Proof : According to Proposition 1.2.2, if (2) holds, then (1) also holds.

Assume that Y is measurable with respect to $\sigma(X)$, we want to show (2). First, we deal with the case where Y is a simple function. For $i \in \{1, \dots, n\}$, let $\lambda_i \in \mathbb{R}$ and $A_i \in \sigma(X)$ and write Y as,

$$Y = \sum_{i=1}^{n} \lambda_i \mathbb{1}_{A_i}.$$

For all $i \in \{1, \dots, n\}$, we can find $B_i \in \mathcal{E}$ such that $A_i = X^{-1}(B_i)$. Hence, Y can be rewritten as,

$$Y = \sum_{i=1}^{n} \lambda_i \mathbb{1}_{A_i} = \sum_{i=1}^{n} \lambda_i \mathbb{1}_{B_i} \circ X = f \circ X,$$

where $f = \sum_{i=1}^{n} \lambda_i \mathbb{1}_{B_i}$ is a $\mathcal{E}$-measurable function.

In general, there exists a sequence (Y_n) of simple functions that converges simply to Y (Proposition 1.2.14). From what we said above, for all n, there exists a measurable function $f_n : E \longrightarrow \mathbb{R}$ such that $Y_n = f_n(X)$. For all $x \in E$, set

$$f(x) = \begin{cases} \lim_{n \to \infty} f_n(x) & \text{if the limit exists,} \\ 0 & \text{otherwise.} \end{cases}$$

According to Proposition 1.2.6, the function f is still measurable. Moreover, note that for all $\omega \in \Omega$,

$$\lim_{n\to\infty} f_n(X(\omega)) = \lim_{n\to\infty} Y_n(\omega) = Y(\omega).$$

If we write $x = X(\omega)$, then in the above definition, the quantity $\lim f_n(x)$ exists and we have,

$$f(X(\omega)) = \lim_{n\to\infty} f_n(X(\omega)) = Y(\omega),$$

meaning that $Y = f(X)$. □

2.3 Common Probability Distributions

In this section, we will introduce some common probability distributions.

In Section 2.3.1, we characterize two two categories of probability distributions, *discrete distributions* and *absolutely continuous distributions*. In Section 2.3.2, we give examples of common discrete distributions and in Section 2.3.3, those of absolutely continuous distributions.

2.3.1 Categories of Probability Distributions

We discuss two types of random variables here: *discrete random variables* and *absolutely continuous random variables*.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(E, \mathcal{E})$ be a measurable space. Consider a random variable $X : \Omega \longrightarrow E$ with values in E.

Definition 2.3.1 (discrete random variables) **:** When E is a countable set and $\mathcal{E} = \mathcal{P}(E)$, the distribution of X writes,

$$\mathbb{P}_X = \sum_{x\in E} p_x \delta_x,$$

where $p_x = \mathbb{P}(X = x)$, δ_x represents the Dirac measure at x. We can understand the above formula as,

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(\bigcup_{x\in B} \{X = x\}) = \sum_{x\in B} \mathbb{P}(X = x) = \sum_{x\in E} p_x \delta_x(B).$$

In consequence, to know the distribution of a random variable X, it is sufficient to know the value of $\mathbb{P}(X = x)$ for all $x \in E$.

When the probability space E is not discrete, we only consider the following case where the *density function* exists.

Definition 2.3.2 (absolutely continuous random variables) **:** Assume that $(E, \mathcal{E}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. If $\mathbb{P}_X$ is absolutely continuous with respect to the Lebesgue measure λ, then we say that $\mathbb{P}_X$ is an absolutely continuous probability distribution (絕對連續機率分佈), and X is an absolutely continuous random variable (絕對連續隨機變數). In such a circumstance, we can apply the Radon–Nikodym theorem, see Theorem 1.3.16, to obtain a non-negative measurable function $p : \mathbb{R}^d \to \mathbb{R}_+$ such that

$$\mathbb{P}_X(B) = \int_B p(x)\,\mathrm{d}x.$$

This function p is unique up to a set of measure zero and it is called the *density function* (密度函數) or the *probability density function* (機率密度函數) of X. Additionally, in the one-dimensional case $d = 1$, for all $\alpha \leqslant \beta$, we have,

$$\mathbb{P}(\alpha \leqslant X \leqslant \beta) = \int_{\alpha}^{\beta} p(x)\,\mathrm{d}x.$$

Remark 2.3.3 : If $\mathbb{P}_X$ has measure zero for any singleton set of $\mathbb{R}^d$, then we say that $\mathbb{P}_X$ is a continuous measure or a continuous distribution, and that X is a continuous random variable. It is not hard to check that, an absolutely continuous measure (or random variable) is also continuous. However, in Exercise 1.33, we have seen that the Cantor distribution is a (probability) measure that is continuous but not absolutely continuous.

2.3.2 Discrete Random Variables

Below we present some common discrete probability distributions that we will often come across with. Let $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ be a discrete probability space. Write E for the space in which the random variable $X : \Omega \longrightarrow E$ takes value, write f_X for the mass function corresponding to the probability distribution $\mathbb{P}_X$. We are interested in the following different distributions.

Uniform distribution (均勻分佈) Assume E is finite and $n = |E|$ denotes the number of elements in E. If the mass function f_X satisfies,

$$f_X(x) = \mathbb{P}(X = x) = \frac{1}{n}, \qquad \forall x \in E,$$

then we say that X is a random variable with the uniform distribution on E, denoted $X \sim \mathrm{Unif}(E)$.

> Ex. : Draw a fair dice with six faces with numbers in $\{1, \ldots, 6\}$.

Bernoulli distribution (伯努力分佈) Let $p \in [0, 1]$ be an additional parameter. If $E = \{0, 1\}$ and f_X satisfies,

$$f_X(1) = \mathbb{P}(X = 1) = p, \qquad f_X(0) = \mathbb{P}(X = 0) = 1 - p,$$

then we say that X is a random variable with the Bernoulli distribution of parameter p, denoted $X \sim \mathrm{Ber}(p)$.

> Ex. : Toss an unfair coin whose probability of getting head (denoted by 1) is p and probability of getting tail (denoted 0) is $1 - p$.

Binomial distribution (二項式分佈) Let $n \in \mathbb{N}_0$ and $p \in [0, 1]$ be two additional parameters. If $E = \{0, \ldots, n\}$ and f_X satisfies,

$$f_X(k) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \qquad \forall k \in E,$$

then we say that X is a random variable with the Binomial distribution of parameter (n, p), denoted $X \sim \text{Bin}(n, p)$.

> Ex. : Consider the unfair coin above tossed n times, then the number of heads follows the binomial distribution.

Figure 2.2: Mass functions of the binomial distribution with parameters $n = 30, p = 0.2$ on the left and $n = 30, p = 0.5$ on the right.

Geometric distribution（幾何分佈） Let $p \in (0, 1)$ be an additional parameter. If $E = \mathbb{N}_0$ and f_X satisfies,

$$f_X(k) = \mathbb{P}(X = k) = (1 - p)^k p, \qquad \forall k \in \mathbb{N}_0,$$

then we say that X is a random variable with the geometric distribution of paramter p, denoted $X \sim \text{Geo}(p)$.

> Ex. : Toss the unfair coin described above. The number of tails before the first head appears follows the geometric distribution. If we interpret head as "success" and tail as "failure", then this distribution gives the number of failures before the first success.

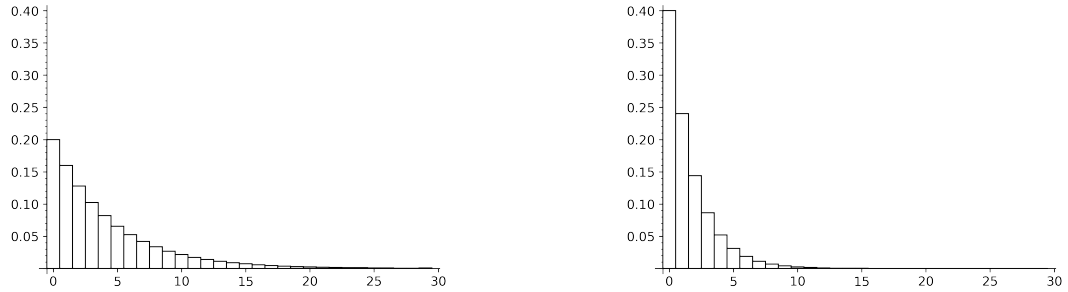

Figure 2.3: Mass functions of the geometric distribution with parameters $n = 30, p = 0.2$ on the left and $n = 30, p = 0.4$ on the right.

Hypergeometric distribution（超幾何分佈） Given nonnegative integers N, K, n with $0 \leqslant K \leqslant N$. Take $E = \{(n - N + K) \vee 0, \ldots, K \wedge n\}$ and f_X satisfying

$$f_X(k) = \mathbb{P}(X = k) = \binom{K}{k}\binom{N-K}{n-k} \Big/ \binom{N}{n}, \qquad \forall k \in E,$$

then we say that X is a random variable with the hypergeometric distribution with parameters (N, K, n), denoted $X \sim \text{Hypergeo}(N, K, n)$.

Ex. : In an urn containing N balls, among which K are white. When we perform a sampling of n balls without replacement, the number of selected white balls satisfies the hypergeometric distribution $\text{Hypergeo}(N, K, n)$.

Poisson distribution (帕松分佈) Let $\lambda > 0$ be an additional parameter. If $E = \mathbb{N}_0$ and $\mathbb{P}$ satisfies,

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad \forall k \in \mathbb{N}_0.$$

then we say that X is a random variable with the Poisson distribution of parameter λ, denoted $X \sim \text{Pois}(\lambda)$.

Figure 2.4: Mass functions of the Poisson distribution with parameter $\lambda = 5$ on the left and $\lambda = 10$ on the right.

Remark 2.3.4 : Either from a theoretical perspective or from applications, Poisson distribution is an important distribution. Later, when we talk about convergence of probability distributions, we will see that the Poisson distribution can be obtained via the limit of Binomial distributions. In other words, let X_n be a random variable with distribution $\text{Bin}(n, p_n)$, then if $np_n \to \lambda$ when n goes to infinity, then

$$\lim_{n\to\infty} \mathbb{P}(X_n = k), \quad \forall k \in \mathbb{Z}_{\geqslant 0}.$$

2.3.3 Random Variables with Density

In this section, we will define some common distributions of continuous random variables and discuss some of their properties. Let us denote by X the continuous random variable of concern and let f_X be its probability density function.

Uniform distribution (均勻分佈) Let $a < b$. If p writes,

$$p(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x).$$

then we say that X is a random variable with the uniform distribution on $[a, b]$, denoted $X \sim \text{Unif}([a, b])$.

Exponential distribution (指數分佈) Let $\lambda > 0$ be an additional parameter. If f_X satisfies

$$f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{\mathbb{R}_{\geqslant 0}}(x), \qquad \forall x \in \mathbb{R},$$

then we say that X is a random variable with the exponential distribution of parameter λ, denoted $X \sim \mathcal{E}(\lambda)$ or $X \sim \mathrm{Exp}(\lambda)$. It is not hard to check that,

$$\mathbb{E}[X] = \frac{1}{\lambda} \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

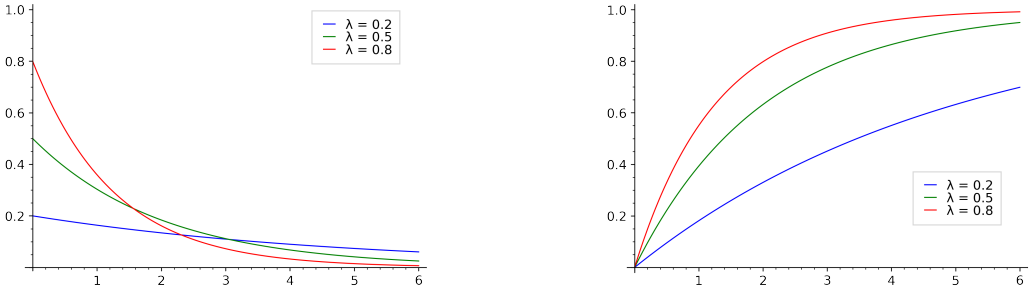

Figure 2.5: The left figure shows the probability density function of the exponential distribution with different parameters, the right figure shows their distribution functions.

Normal distribution（常態分佈）or Gaussian distribution（高斯分佈） Let $m \in \mathbb{R}$ and $\sigma > 0$ be two additional parameters. If f_X satisfies

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-m)^2}{2\sigma^2}\Big),$$

then we say that X is a random variable with the Gaussian (or normal) distribution with mean m and variance σ^2, denoted $X \sim \mathcal{N}(m, \sigma^2)$. It is not hard to check that, if $X \sim \mathcal{N}(m, \sigma^2)$, then

$$\mathbb{E}[X] = m \qquad \text{and} \qquad \mathrm{Var}(X) = \sigma^2.$$

When $m = 0$, we say that X is centered (置中); when $\sigma = 1$, we say that X is reduced (約化); when both are satisfied, we call it the *standard normal distribution* (標準常態分佈) .

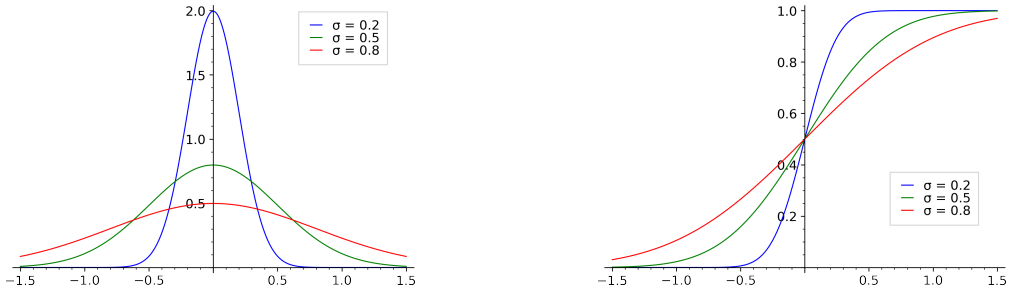

Figure 2.6: The left figure shows the probability density function of the Gaussian distribution with different parameters, the right figure shows their distribution functions. Here we take $\mu = 0$.

Cauchy distribution（柯西分佈） Given parameters $\gamma > 0$ and $x_0 \in \mathbb{R}$. If f_X satisfies

$$f_X(x) = \frac{1}{\pi\gamma(1 + (\frac{x-x_0}{\gamma})^2)} = \frac{1}{\pi}\frac{\gamma}{(x-x_0)^2+\gamma^2}, \qquad \forall x \in \mathbb{R},$$

then we say that X is a Cauchy distribution with parameters (x_0, γ), denoted $X \sim \text{Cauchy}(x_0, \gamma)$. It is not hard to check that the expectation of a Cauchy distribution is not defined. (See Exercise 2.15.)

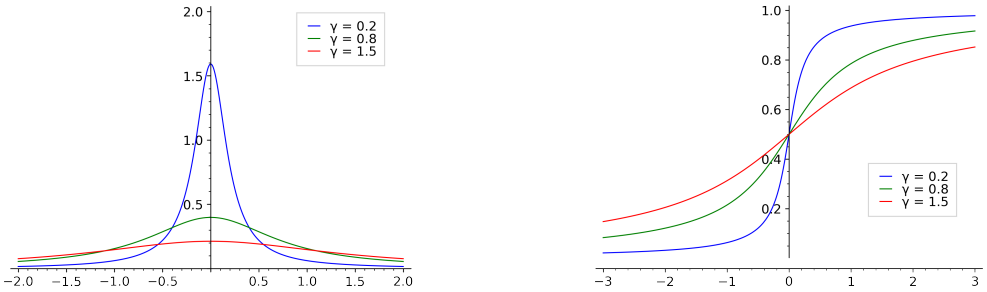

Figure 2.7: The left figure shows the probability density function of the Cauchy distribution with different parameters, the right figure shows their distribution functions. Here we take $x_0 = 0$.

Gamma distribution（珈瑪分佈） Given parameters $\alpha, \beta > 0$. If f_X satisfies

$$f_X(x) = \frac{1}{\Gamma(\alpha)}\beta^\alpha x^{\alpha-1}e^{-\beta x}\mathbb{1}_{x>0}, \qquad \forall x \in \mathbb{R},$$

where $\Gamma(\alpha)$ is the Gamma function defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,\mathrm{d}x,$$

then we say that X is a Gamma distribution with parameters (α, β), denoted $X \sim \Gamma(\alpha, \beta)$ or $X \sim \text{Gamma}(\alpha, \beta)$. We recall some important properties of the Gamma function, $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ for all $\alpha > 0$ (change of variables), $\Gamma(1) = 1$, and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

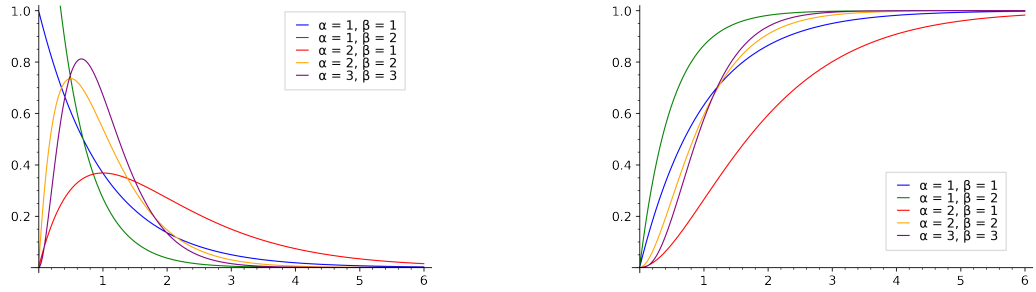

Figure 2.8: The left figure shows the probability density function of the Gamma distribution with different parameters, the right figure shows their distribution functions.

Beta distribution (貝塔分佈) Given parameters $\alpha, \beta > 0$. If f_X satisfies

$$f_X(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}_{x\in(0,1)}, \qquad \forall x \in \mathbb{R},$$

where $B(\alpha,\beta)$ is the Beta function, defined by

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,\mathrm{d}x = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

then we say that X is the Beta distribution with parameters (α,β), denoted $X \sim \text{Beta}(\alpha,\beta)$.

Remark 2.3.5 : The probability distribution $\text{Beta}(1,1)$ is the uniform distribution on $[0,1]$. We also have the following symmetry, for $X \sim \text{Beta}(\alpha,\beta)$, we have $1-X \sim \text{Beta}(\beta,\alpha)$.

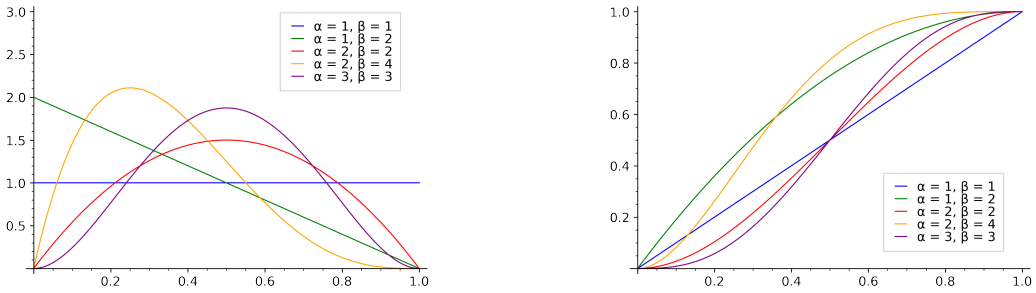

Figure 2.9: The left figure shows the probability density function of the Beta distribution with different parameters, the right figure shows their distribution functions.

2.4 Moments of Random Variables

Moments are expectations of powers of random variables and exhibit some non-linear properties. An important example is the notion of *variance* in both statistics and probability. In Probability Theory, methods of moments (動差方法) also have important applications in showing existence in Graph Theory or Number Theory, see Exercise 2.24 and Exercise 2.25 for instance.

2.4.1 Moments and Variance

Definition 2.4.1 : Fix a positive integer $k \geqslant 1$. We say that the k-th *moment* (動差) of a real random variable X is finite if $\mathbb{E}[|X|^k] < \infty$, or $X \in L^k$. We call $\mathbb{E}[X^k]$ the k-th moment of X.

Definition 2.4.2 : Let $X \in L^2(\Omega, \mathcal{P}(\Omega), \mathbb{P})$. Then, the *variance* (變異數) of X is defined as,

$$\text{Var}(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \geqslant 0. \tag{2.4}$$

Its *standard deviation* (標準差) is denoted

$$\sigma_X = \sqrt{\mathrm{Var}(X)}.$$

Remark 2.4.3 : We can understand the variance $\mathrm{Var}(X)$ in the above definition as *a* quantity measuring how much X deviates from its expectation $\mathbb{E}[X]$.

It is important to square in the definition given by Eq. (2.4), otherwise, we would find

$$\mathbb{E}[X - \mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[X] = 0,$$

which cannot describe correctly how far the random variable X is away from its expectation.

To avoid the cancellation from happening, we may look at the quantity $\mathbb{E}[|X - \mathbb{E}[X]|]$. It has the advantage that we do not need to require the condition $X \in L^2$, we only need $X \in L^1$ instead. However, the absolute value has less regularity than the square function, which is its disadvantage, leading to more complicated computations. Later in another chapter, we will see the importance of the square in probability theory, in particular, its relation with the central limit theorem.

We can understand the variance $\mathrm{Var}(X)$ in the above definition as such: it measures how much X deviates from its expectation $\mathbb{E}[X]$. Moreover, we can notice that, $\mathrm{Var}(X) = 0$ if and only if X is almost surely a constant.

Proposition 2.4.4 : *The variance* $\mathrm{Var}(X)$ *satisfies the following optimization problem,*

$$\mathrm{Var}(X) = \inf_{a \in \mathbb{R}} \mathbb{E}\left[(X - a)^2\right].$$

Proof : We only need to show that, we have, for any real nubmer a that

$$\mathbb{E}[(X - a)^2] = \mathrm{Var}(X) + (\mathbb{E}[X] - a)^2.$$

By a direct computation, we find

$$\begin{aligned}
\mathbb{E}\left[(X - a)^2\right] &= \mathbb{E}\left[\left((X - \mathbb{E}[X]) + (\mathbb{E}[X] - a)\right)^2\right] \\
&= \mathrm{Var}(X) + 2(\mathbb{E}[X] - a)\,\mathbb{E}[X - \mathbb{E}[X]] + (\mathbb{E}[X] - a)^2 \\
&= \mathrm{Var}(X) + (\mathbb{E}[X] - a)^2.
\end{aligned}$$

□

Using the expectation and the variance, we can estimate the following probability,

Markov's inequality (馬可夫不等式) If $X \in L^1_+(\Omega, \mathcal{A}, \mathbb{P})$ and $a > 0$, then

$$\mathbb{P}(X \geqslant a) \leqslant \frac{1}{a}\,\mathbb{E}[X].$$

Bienaymé-Chebyshev inequality (柴比雪夫不等式) If $X \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ and $a > 0$, then

$$\mathbb{P}(|X - \mathbb{E}[X]| \geqslant a) \leqslant \frac{1}{a^2} \operatorname{Var}(X).$$

Definition 2.4.5 : If $X, Y \in L^2(\Omega, \mathcal{P}(\Omega), \mathbb{P})$, then their *covariance* (共變異數) is defined by,

$$\operatorname{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[X(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

We note that, the definition of covariance extends that of variance, $\operatorname{Var}(X) = \operatorname{Cov}(X, X)$.

Remark 2.4.6 : The covariance measures the correlation (關聯性) between two random variables X and Y, which is not the same notion as independence. See Section 3.1.2.

Remark 2.4.7 : Given a discrete probability space $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$, it is not hard to check that the covariance function

$$\operatorname{Cov}(\cdot, \cdot) : L^2(\Omega, \mathcal{A}, \mathbb{P}) \times L^2(\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow \mathbb{R}$$

is a symmetric and bilinear operator. In other words, for all $X, Y, Z \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ and $a, b \in \mathbb{R}$, we have

$$\begin{aligned} \operatorname{Cov}(X, Y) &= \operatorname{Cov}(Y, X), \\ \operatorname{Cov}(aX + bY, Z) &= a \operatorname{Cov}(X, Z) + b \operatorname{Cov}(Y, Z). \end{aligned}$$

This allows us to apply the Cauchy-Schwarz inequality to obtain

$$|\operatorname{Cov}(X, Y)| \leqslant \sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}.$$

Moreoever, if one of X and Y is almost surely a constant, then $\operatorname{Cov}(X, Y) = 0$.

Definition 2.4.8 : Consider a random variable $X = (X_1, \ldots, X_d)$ with values in $\mathbb{R}^d$. Let us assume that all its components ara in $L^2(\Omega, \mathcal{A}, \mathbb{P})$. Then, we may define the *covariance matrix* (共變異數矩陣) of X by

$$K_X = \big(\operatorname{Cov}(X_i, X_j)\big)_{1 \leqslant i,j \leqslant d}.$$

Question 2.4.9: Prove that when X is a d-dimensional real random variable, its covariance matrix K_X is a positive semi-definite (半正定) symmetric matrix. In other words, prove that for all $\lambda = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^d$, we have $\lambda K_X \lambda^T \geqslant 0$.

Question 2.4.10: If A is a matrix of size $n \times d$ and X is a d-dimensional real random variable, define $Y = AX$ and prove that $K_Y = A K_X A^T$.

2.4.2 Linear Regression

Let $X, Y_1, \dots, Y_n$ be random variables in $L^2(\Omega, \mathcal{A}, \mathbb{P})$. We want to find a random variable which is a linear combination of $1, Y_1, \dots, Y_n$ approximating X. In other words, we want to look for real numbers $\beta_0, \dots, \beta_n$ that minimize the following quantity,

$$\mathbb{E}[(X - (\beta_0 + \beta_1 Y_1 + \dots \beta_n Y_n))^2].$$

Proposition 2.4.11 : *When (α_j) is a solution to the following linear system,*

$$\sum_{j=1}^{n} \alpha_j \operatorname{Cov}(Y_j, Y_k) = \operatorname{Cov}(X, Y_k), \qquad 1 \leqslant k \leqslant n, \tag{2.5}$$

we set

$$Z = \mathbb{E}[X] + \sum_{j=1}^{n} \alpha_j (Y_j - \mathbb{E}[Y_j]), \tag{2.6}$$

and we have,

$$\inf_{\beta_0, \dots, \beta_n \in \mathbb{R}} \mathbb{E}\left[\left(X - (\beta_0 + \beta_1 Y_1 + \dots \beta_n Y_n) \right)^2 \right] = \mathbb{E}[(X - Z)^2].$$

Proof : Let H be the vector subspace of $L^2(\Omega, \mathcal{A}, \mathbb{P})$ generated by $1, Y_1, \dots, Y_n$. Since Z reaches the minimum of the functional $U \in H \mapsto \|X - U\|_2$, Z is the orthogonal projection of X on H. If we write Z as,

$$Z = \alpha_0 + \sum_{j=1}^{n} \alpha_j (Y_j - \mathbb{E}[Y_j]), \tag{2.7}$$

we get $\mathbb{E}[(X - Z) \cdot 1] = 0$ from the properties of the orthogonal projection, giving $\alpha_0 = \mathbb{E}[X]$. Similarly, for all $1 \leqslant k \leqslant n$, we have

$$\mathbb{E}\left[(X - Z) \cdot (Y_k - \mathbb{E}[Y_k]) \right] = 0,$$

meaning that $\operatorname{Cov}(X, Y_k) = \operatorname{Cov}(Z, Y_k)$. To conclude, we use the linear combination of Z in Eq. (2.7) to obtain Eq. (2.5).

Conversely, if (α_i) satisfies Eq. (2.5), then Z defined in Eq. (2.6) is an element in H and since $X - Z$ and H are orthogonal, Z is the orthogonal projection of X on H. □

2.4.3 Characteristic Function

Definition 2.4.12 : If X is a random variable with values in $\mathbb{R}^d$, then its *characteristic function* (特徴函数) $\Phi_X : \mathbb{R}^d \longrightarrow \mathbb{C}$ is defined by,

$$\Phi_X(\xi) = \mathbb{E}\left[\exp(\mathrm{i}\, \xi \cdot X) \right], \qquad \xi \in \mathbb{R}^d.$$

Remark 2.4.13 : The above definition rewrites,

$$\Phi_X(\xi) = \int_{\mathbb{R}^d} e^{\mathrm{i}\, \xi \cdot x} \mathbb{P}_X(\mathrm{d}x).$$

In other words, we can see Φ_X as the *Fourier transform* (傅立葉變換) of $\mathbb{P}_X$, i.e., $\Phi_X(\xi) = \hat{\mathbb{P}}_X(\xi)$. Moreover, from the dominated convergence theorem, we know that Φ_X is a bounded continuous function on $\mathbb{R}^d$.

Below we give the list of characteristic functions of the distributions mentioned in Section 2.3.

random variableX	characteristic function $\Phi_X(\xi)$
$\mathrm{Ber}(p)$	$1-p+pe^{\mathrm{i}\,\xi}$
$\mathrm{Bin}(n,p)$	$(1-p+pe^{\mathrm{i}\,\xi})^n$
$\mathrm{Geo}(p)$	$\dfrac{p}{1-(1-p)e^t}$
$\mathrm{Pois}(\lambda)$	$\exp(\lambda(e^{\mathrm{i}\,\xi}-1))$
$\mathrm{Unif}([a,b])$	$\dfrac{e^{\mathrm{i}\,\xi b}-e^{\mathrm{i}\,\xi a}}{\mathrm{i}\,\xi(b-a)}$
$\mathcal{E}(\lambda)$	$\dfrac{\lambda}{\lambda-\mathrm{i}\,\xi}$
$\mathcal{N}(m,\sigma^2)$	$\exp(\mathrm{i}\,m\xi-\frac{1}{2}\sigma^2\xi^2)$

In this section, we will prove that the characteristic function of a random variable *determines entirely* its distribution.

First, we prove the invariance (不變性) of the normal distribution under the map of Fourier transform $\mathcal{F}:\mathbb{P}_X\mapsto\hat{\mathbb{P}}_X$.

Lemma 2.4.14 : *Let X be a random variable with distribution $\mathcal{N}(0,\sigma^2)$, then*

$$\Phi_X(\xi)=\exp\Big(-\frac{\sigma^2\xi^2}{2}\Big),\qquad \xi\in\mathbb{R}.$$

Proof : Using the parity of the integrand, the imaginary part of $\Phi_X(\xi)$ is zero. Thus, we need to compute,

$$f(\xi)=\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\cos(\xi x)\,\mathrm{d}x.$$

We take the derivative of the above formula. Since the integrable function $x\mapsto|x|e^{-x^2/2}$ bounds $x\mapsto x\sin(\xi x)e^{-x^2/2}$, the differential operator can be inverted with the integration operator, i.e.,

$$f'(\xi)=-\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi}}xe^{-x^2/2}\sin(\xi x)\,\mathrm{d}x.$$

We apply the formula of integration by parts and get,

$$f'(\xi) = -\xi \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cos(\xi x)\, \mathrm{d}x = -\xi f(\xi).$$

As a consequence, f is a solution to the differential equation $f'(\xi) = -\xi f(\xi)$ with initial condition $f(0) = 1$. Due to uniqueness, we deduce that $f(\xi) = \exp(-\xi^2/2)$. □

Theorem 2.4.15 : *Given a random variable X. Then, its distribution can be entirely determined by its characteristic function. In other words, the Fourier transform $\mathcal{F} : \mathbb{P}_X \mapsto \hat{\mathbb{P}}_X$ is injective* (單射) *on the space of distributions on $\mathbb{R}^d$.*

Proof : First, we start with the one-dimensional case $d = 1$. For all $\sigma > 0$, the density function of the normal distribution $\mathcal{N}(0, \sigma^2)$ writes,

$$g_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{x^2}{2\sigma^2}\Big).$$

If μ is a probability measure on $\mathbb{R}$, we define,

$$\begin{aligned} f_\sigma(x) &= \int_{\mathbb{R}} g_\sigma(x-y)\mu(\mathrm{d}y) \overset{\text{(def)}}{=} g_\sigma * \mu(x), \\ \mu_\sigma(\mathrm{d}x) &= f_\sigma(x)\, \mathrm{d}x. \end{aligned}$$

We show the statement in two steps,

(1) μ_σ can be entirely defined by $\widehat{\mu}$;

(2) for any function $\varphi \in \mathcal{C}_b(\mathbb{R})$, when $\sigma \to 0$, we have the convergence $\int \varphi(x)\mu_\sigma(\mathrm{d}x) \longrightarrow \int \varphi(x)\mu(\mathrm{d}x)$.

We prove (1) now. From Lemma 2.4.14, we know that for all $x \in \mathbb{R}$, we have,

$$\sqrt{2\pi\sigma^2} g_\sigma(x) = \exp\Big(-\frac{x^2}{2\sigma^2}\Big) = \int_{\mathbb{R}} e^{\mathrm{i}\,\xi\cdot x} g_{1/\sigma}(\xi)\, \mathrm{d}\xi.$$

Then, we can rewrite $f_\sigma(x)$ as,

$$\begin{aligned} f_\sigma(x) = \int_{\mathbb{R}} g_\sigma(x-y)\mu(\mathrm{d}y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} e^{\mathrm{i}\,\xi(x-y)} g_{1/\sigma}(\xi)\, \mathrm{d}\xi \Big) \mu(\mathrm{d}y) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{\mathrm{i}\,\xi x} g_{1/\sigma}(\xi) \Big(\int_{\mathbb{R}} e^{-\mathrm{i}\,\xi y} \mu(\mathrm{d}y) \Big)\, \mathrm{d}\xi \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{\mathrm{i}\,\xi x} g_{1/\sigma}(\xi) \widehat{\mu}(-\xi)\, \mathrm{d}\xi. \end{aligned} \tag{2.8}$$

In the second-last equality, we use Fubini's Theorem (Theorem 1.4.3), because μ is a probability measure and $g_{1/\sigma}$ is integrable with respect to Lebesgue measure.

To prove (2), we note that for all $\varphi \in C_b(\mathbb{R})$, we have,

$$\int \varphi(x)\mu_\sigma(\mathrm{d}x) = \int \varphi(x)\Big(\int g_\sigma(y-x)\mu(\mathrm{d}y)\Big)\,\mathrm{d}x = \int g_\sigma * \varphi(y)\mu(\mathrm{d}y), \tag{2.9}$$

where we use Fubini's Theorem and the property that g_σ is an even function. Next, we need other properties of g_σ,

$$\int g_\sigma(x)\,\mathrm{d}x = 1, \quad \text{and} \quad \lim_{\sigma\to 0}\int_{\{|x|>\varepsilon\}} g_\sigma(x)\,\mathrm{d}x = 0, \quad \forall \varepsilon > 0,$$

which gives us easily,

$$\forall y \in \mathbb{R}, \qquad \lim_{\sigma\to 0} g_\sigma * \varphi(y) = \varphi(y). \tag{2.10}$$

Since for all $\sigma > 0$, $|g_\sigma * \varphi| \leqslant \sup|\varphi|$, from the dominated convergence theorem, we obtain,

$$\lim_{\sigma\to 0}\int \varphi(x)\mu_\sigma(\mathrm{d}x) = \lim_{\sigma\to 0}\int g_\sigma * \varphi(y)\mu(\mathrm{d}y) = \int \varphi(x)\mu(\mathrm{d}x).$$

Hence, the theorem is true when $d = 1$.

For a general value of d, the proof is similar. We define the following function,

$$g_\sigma^{(d)}(x_1,\ldots,x_d) = \prod_{j=1}^d g_\sigma(x_j),$$

and use the fact that for all $\xi \in \mathbb{R}^d$, we have,

$$\int_{\mathbb{R}^d} g_\sigma^{(d)}(x)e^{\mathrm{i}\,\xi\cdot x}\,\mathrm{d}x = \prod_{j=1}^d \int_{\mathbb{R}} g_\sigma(x_j)e^{\mathrm{i}\,\xi_j x_j}\,\mathrm{d}x_j = (2\pi\sigma^2)^{d/2} g_{1/\sigma}^{(d)}(\xi).$$

□

Question 2.4.16: $C_c(\mathbb{R})$ denotes the set of functions in $C_b(\mathbb{R})$ that are compactly supported (緊緻支撐). Prove that, for $\varphi \in C_c(\mathbb{R})$, we can improve the convergence in Eq. (2.10) and replace with the uniform convergence on $\mathbb{R}$. In other words, show that,

$$\forall \varphi \in C_c(\mathbb{R}), \qquad \|g_\sigma * \varphi - \varphi\|_\infty \xrightarrow[\sigma\to 0]{} 0.$$

Proposition 2.4.17 : *Let $X = (X_1,\ldots,X_n)$ be a n-dimensional real random variable. Assume that $\|X\|_2^2$ is integrable, then Φ_X is a $\mathcal{C}^2$ function and when $\xi = (\xi_1,\ldots,\xi_n)$ tends to 0, we have,*

$$\Phi_X(\xi) = 1 + \mathrm{i}\sum_{j=1}^d \xi_j\,\mathbb{E}[X_j] - \frac{1}{2}\sum_{j,k=1}^d \xi_j\xi_k\,\mathbb{E}[X_jX_k] + o(\|\xi\|^2).$$

Proof : It is not hard to check that

(a) For all $\xi \in \mathbb{R}^n$, the function $x \mapsto e^{\mathrm{i}\,\xi\cdot x}$ is integrable with respect to $\mathbb{P}_X$.

(b) For all $x \in \mathbb{R}^n$, the function $\xi \mapsto e^{\mathrm{i}\,\xi\cdot x}$ is differentiable with respect to ξ, and its differential at ξ

writes

$$\eta = (\eta_1, \dots, \eta_n) \mapsto \sum_{i=j}^{n} \mathrm{i}\,\eta_j x_j e^{\mathrm{i}\,\xi\cdot x} \quad \text{or} \quad \sum_{j=1}^{n} \mathrm{i}\,x_j e^{\mathrm{i}\,\xi\cdot x}\,\mathrm{d}\xi_j.$$

(c) For all $j = 1, \dots, n$, the partial derivative $x \mapsto \mathrm{i}\,x_j e^{\mathrm{i}\,\xi\cdot x}$ is continuous, and can be dominated by x_j, which is integrable with respect to $\mathbb{P}_X$, since $X_j \in L^2 \subseteq L^1$.

Therefore, Theorem 1.2.24 allows us to invert the differential operator and the integration operator when we compute the derivative of Φ_X, which is

$$\forall j = 1, \dots, n, \qquad \frac{\partial \Phi_X}{\partial \xi_j}(\xi) = \mathrm{i}\,\mathbb{E}[X_j e^{\mathrm{i}\,\xi\cdot X}].$$

Moreover, Theorem 1.2.22 implies that these partial derivatives are continuous.

For the second partial derivatives, we can proceed in a similar way. In particular, we need to use the fact that $\mathbb{E}[|X_j X_k|] \leqslant \mathbb{E}[X_j]^2\,\mathbb{E}[X_k]^2 < \infty$ to justify that the second derivative and the integral can be inverted. □

It is important to note that, only the characteristic function describes the probability distribution entirely, but not the moments. In general, knowing all the moment $(\mathbb{E}[X^k])_{k\geqslant 0}$ is not enough to deduce the probability distribution of X. In Exercise 2.22, we will see counterexamples. In particular, we will construct (at least) two different probability distributions having the same moments. Then, in Exercise 2.23, we will discuss the situations in which the moments uniquely determine the probability distribution.

2.4.4 Generating Function

When a discrete random variable takes values in the set of nonnegative integers $\mathbb{N}_0$, we may define its *generating function* (生成函數) .

Definition 2.4.18 : Let X be a discrete random variable with values in $\mathbb{N}_0$. Its generating function G_X is defined for $s \in \mathbb{C}$ such that

$$G_X(s) = \mathbb{E}[s^X] = \sum_{n=0}^{\infty} \mathbb{P}(X = n) s^n \tag{2.11}$$

converges. Conversely, if we know the generating function of a discrete random variable X, then by reading its coefficients, we recover the distribution of X.

The series defined by Eq. (2.11) satisfies some properties, which are considered as prerequisites that were studied in the calculus class. Below is a reminder on these results.

Convergence There exists a *radius of convergence* (收斂半徑) $0 \leqslant R \leqslant \infty$ such that the series $G_X(s)$ converges when $|s| < R$; and the series $G_X(s)$ diverges when $|s| > R$. Moreover, for any $R' < R$, the series $G_X(s)$ *converges uniformly* on $\{s : |s| \leqslant R'\}$. Here we note that we clearly have $R \geqslant 1$ because $G_X(1) = 1$.

Differentiability We may differentiate the series $G_X(s)$ term by term infinitely many times on its domain of definition $\{s : |s| < R\}$.

Uniqueness If there exists $0 < R' \leqslant R$ such that the equality $G_X(s) = G_Y(s)$ holds for all $|s| < R'$, then for all $n \in \mathbb{N}_0$, we have $\mathbb{P}(X = n) = \mathbb{P}(Y = n)$. Moreover, we have

$$\mathbb{P}(X = n) = \frac{1}{n!} G_X^{(n)}(0), \qquad \forall n \geqslant 0.$$

Continuity (Abel's theorem) Since all the terms $\mathbb{P}(X = n)$ are nonnegative, if $R < \infty$, we have

$$\lim_{s \uparrow R} G_X(s) = \sum_{n=0}^{\infty} \mathbb{P}(X = n) R^n.$$

Proposition 2.4.19 : *We may compute the expectation of X using the derivative of G_X,*

$$G'_X(1) := \lim_{s \uparrow 1} G_X(s) = \mathbb{E}[X] \in [0, \infty].$$

More generally, for any positive integer $p \geqslant 1$, we have

$$G_X^{(p)} := \lim_{s \uparrow 1} G_X^{(p)}(s) = \mathbb{E}[X(X-1)\dots(X-p+1)],$$

which means that from the generating function of X, we can easily obtain all the moments of X. In particular, for $X \in L^2(\Omega, \mathcal{P}(\Omega), \mathbb{P})$, we have

$$\mathrm{Var}(X) = G''_X(1) + G'_X(1) - G'_X(1)^2.$$

Proof : For all positive integer $p \geqslant 1$, by the continuity of $G_X^{(p)}(s)$ on $\{s : |s| < R\}$ (knowing that $R \geqslant 1$), we find

$$\lim_{s \uparrow 1} G_X^{(p)}(s) = \sum_{n=0}^{\infty} \mathbb{P}(X = n) \cdot n(n-1)\dots(n-p+1) = \sum_{n=0}^{\infty} \mathbb{P}(X = n) \cdot \varphi(n),$$

where $\varphi(n) = n(n-1)\dots(n-p+1)$, as desired.

Let $X \in L^2(\Omega, \mathcal{P}(\Omega), \mathbb{P})$. From the above property, we find

$$\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] = G''_X(1) + G'_X(1),$$

so $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = G''_X(1) + G'_X(1) - G'_X(1)^2$, as desired. □

Example 2.4.20 : If X is a random variable following the binomial distribution $\mathrm{Bin}(n, p)$, then its generating function writes,

$$G_X(s) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} s^k = \big((1-p) + ps\big)^n.$$

We may compute its expectation

$$\mathbb{E}[X] = G'_X(s)_{s=1} = \left[n((1-p)+ps)^{n-1}p\right]_{s=1} = np.$$

If we want to compute its variance, we start with

$$G''_X(1) = \left[n(n-1)((1-p)+ps)^{n-2}p^2\right]_{s=1} = n(n-1)p^2,$$

then find

$$\mathrm{Var}(X) = n(n-1)p^2 + np - (np)^2 = np - np^2 = np(1-p).$$

When our goal is to compute the moments of X, we may directly define its *moment generating function* (動差生成函數) , allowing us to simplify the computations.

Definition 2.4.21 : Given a discrete random variable X with values in $\mathbb{N}_0$, we define its *moment generating function* (動差生成函數) , or *exponential generating function* (指數生成函數) , as

$$M_X(t) := G_X(e^t) = \mathbb{E}[e^{tX}],$$

where we denote the converging radius of G_X by R and require $e^t < R$.

Remark 2.4.22 : If we define the moment generating function by $M_X(t) := \mathbb{E}[e^{tX}]$, then we only need to require that X is a *real* discrete random variable.

Proposition 2.4.23 : *The moments of the discrete random variable X can be computed using the derivatives of M_X. More precisely, we have, for all nonnegative integer $k \geqslant 0$,*

$$\mathbb{E}[X^k] = M_X^{(k)}(0).$$

Proof : We have

$$\begin{aligned} M_X(t) &= \sum_{k=0}^{\infty} e^{tk}\,\mathbb{P}(X=k) = \sum_{k=0}^{\infty}\sum_{n=0}^{\infty} \frac{(tk)^n}{n!}\,\mathbb{P}(X=k) \\ &= \sum_{n=0}^{\infty} \frac{t^n}{n!}\Big(\sum_{k=0}^{\infty} k^n\,\mathbb{P}(X=k)\Big) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\,\mathbb{E}[X^n]. \end{aligned}$$

□

We may also define the Laplace transform (拉普拉斯轉換) of X as follows,

$$L_X(\lambda) := \mathbb{E}[e^{-\lambda X}].$$

We need to note that, similar to the exponential generating function, this is not well defined for all $\lambda \in \mathbb{R}$.

2.4.5 Tail Probability

Given a random variable X, the quantities $\mathbb{P}(|X| > \lambda)$ or $\mathbb{P}(|X| > \lambda)$ are called *tail probability* (尾端機率) . Generally speaking, higher the order of the moment of X we can control, smaller is the tail probability, and vice versa.

Definition 2.4.24 : We can classify the distribution of a random variable into different categories depending on the tail probability.

(1) *sub-Gaussian distribution* (次高斯分佈) : there exist $C, c > 0$ such that,

$$\mathbb{P}(|X| > x) \leqslant C \exp(-cx^2), \qquad x \to \infty.$$

(2) *heavy-tailed distribution* (重尾分佈) : there exists $C, \alpha > 0$ such that,

$$\mathbb{P}(|X| > x) \sim Cx^{-\alpha}, \qquad x \to \infty.$$

Remark 2.4.25 : We may note that the Cauchy distribution is a heavy-tailed distribution. Indeed, if $X \sim \text{Cauchy}(0, 1)$, then

$$\mathbb{P}(|X| > x) \sim \frac{2}{\pi x}, \qquad x \to \infty.$$

Moreover, a Gaussian distribution is also a sub-Gaussian distribution.

Question 2.4.26: Given a random variable X, prove that the three following statements are equivalent.

(1) X is a sub-Gaussian distribution.

(2) There exist $C, c > 0$ such that $\mathbb{E}[e^{tX}] \leqslant C \exp(ct^2)$.

(3) There exists $C > 0$ such that for all $k \geqslant 1$, we have $\mathbb{E}[|X|^k] \leqslant (Ck)^{k/2}$.

Question 2.4.27: Let X be a real random variable and $k > 0$. If X is in L^k, prove that when $\lambda \longrightarrow \infty$, we have that

$$\lambda^k \, \mathbb{P}(|X| > \lambda) \longrightarrow 0.$$