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* (樣本空間) , 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).

第二章 機率論基礎

第一節 基礎定義

在這個章節，我們將定義並且討論在機率論中最基本的概念，包含機率空間、隨機變數、期望值、特徵函數等等。

機率論可以被視為測度論的特例，也就是說當測度總質量為有限或者為 1 的情況，所以在第一章中複習過的很多概念以及定理，在機率論中還是有效的。此外，從機率論的角度來看，機率空間本身並不是這麼重要，我們感興趣的是隨機變數可以取的值，以及這些值出現的頻率。

在機率論以及測度論中的不同概念有下列對應：

機率空間	$\longleftrightarrow$	測度空間
事件	$\longleftrightarrow$	σ 代數中的元素
隨機變數	$\longleftrightarrow$	可測函數
期望值	$\longleftrightarrow$	積分

再來我們會將上述不同名詞給出明確的定義，以及提供不同例子。

第一小節 機率空間

我們在第 1.1 節中回顧了測度空間的定義，在機率論中，我們需要定義的機率空間其實只是它的一個特例。

定義 2.1.1 ： 令 $(\Omega, \mathcal{A})$ 為一個可測空間，且令 $\mathbb{P}$ 為一個在 $(\Omega, \mathcal{A})$ 之上總質量為 1 的測度，稱作機率測度 (probability measure) ，那麼我們將 $(\Omega, \mathcal{A}, \mathbb{P})$ 稱為一個機率空間 (probability space) 。

- 集合 Ω 稱作樣本空間 (sample space) ，可以視為隨機試驗中，描述「隨機單位」的空間。
- 集合 $\mathcal{A}$ 包含了所有的可測事件，也可以簡稱做事件，也就是所有 Ω 可以被「機率」所測量的子集合。換句話說，元素 $A \in \mathcal{A}$ 是個 Ω 的子集合，其中包含了所有滿足特定條件的「隨機單位」。
- 對於任何 $A \in \mathcal{A}$，$\mathbb{P}(A)$ 描述可測事件 A 發生的機率。
- 若樣本空間 Ω 為離散（有限或可數無限），則我們可以考慮 $\mathcal{A} = \mathcal{P}(\Omega)$，並將任何定義在可測空間 $(\Omega, \mathcal{A})$ 上的機率測度 $\mathbb{P}$ 稱作離散機率 (discrete probability)，機率空間 $(\Omega, \mathcal{A}, \mathbb{P})$ 則稱作離散

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 $\Omega = \{1, \ldots, 6\}^2$ 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,

$$\begin{aligned}\Omega = \{&(\text{Head}, \text{Head}, \text{Head}), (\text{Head}, \text{Head}, \text{Tail}), (\text{Head}, \text{Tail}, \text{Head}), (\text{Head}, \text{Tail}, \text{Tail}),\\ &(\text{Tail}, \text{Head}, \text{Head}), (\text{Tail}, \text{Head}, \text{Tail}), (\text{Tail}, \text{Tail}, \text{Head}), (\text{Tail}, \text{Tail}, \text{Tail})\},\end{aligned}$$

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.

機率空間 (discrete probability space)。

範例 2.1.2 【離散機率空間】： 給定一顆六面數字分別為 1 至 6 的公正骰子。

(1) 考慮骰子丟兩次的機率實驗，那麼我們可以令機率空間為 $\Omega = \{1, \ldots, 6\}^2$，可以測量的事件集合為 $\mathcal{A} = \mathcal{P}(\Omega)$，機率測度 $\mathbb{P}$ 則有下列性質：對於任何 $A \in \mathcal{A}$，我們有 $\mathbb{P}(A) = \frac{|A|}{36}$。

(2) 考慮一顆不公正的銅板，其出現正面的機率是反面的兩倍，考慮投擲此銅板三次的實驗，那麼我們可以令機率空間為

$$\begin{aligned}\Omega = \{&(\text{正}, \text{正}, \text{正}), (\text{正}, \text{正}, \text{反}), (\text{正}, \text{反}, \text{正}), (\text{正}, \text{反}, \text{反}),\\ &(\text{反}, \text{正}, \text{正}), (\text{反}, \text{正}, \text{反}), (\text{反}, \text{反}, \text{正}), (\text{反}, \text{反}, \text{反})\},\end{aligned}$$

可以測量的事件集合為 $\mathcal{A} = \mathcal{P}(\Omega)$，機率測度 $\mathbb{P}$ 則有下列性質：若對於任何 $\omega \in \Omega$，我們記正面及反面出現次數為 $H(\omega)$ 及 $T(\omega)$，那麼我們有 $\mathbb{P}(\omega) = (\frac{2}{3})^{H(\omega)}(\frac{1}{3})^{T(\omega)}$。

範例 2.1.3 【一般機率空間】： 平面上的單位圓可以記作

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

且可以詮釋為二維空間中的單位方向。若我們考慮可測空間 $(\mathbb{S}^1, \mathcal{B}(\mathbb{S}^1))$ 及機率測度

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

則 $\mathbb{P}$ 會是個在 $\mathbb{S}^1 = \mathbb{R}/(2\pi\mathbb{Z})$ 上的均匀機率。這也可以視為由在 $\mathbb{R}$ 上的勒貝格測度「引導」出來在商空間 $\mathbb{S}^1$ 上的（機率）測度。

第二小節 隨機變數

在機率論中，其實機率空間本身顯得並沒有這麼重要，因為我們感興趣的是在隨機試驗中，可以觀察且測量的量，以及這些不同結果發生的頻率。因此，我們在下面定義隨機變數的概念，並將它視為「在一個隨機試驗中，隨機單位事件給出的結果」。

定義 2.1.4 ： 令 $(\Omega, \mathcal{A}, \mathbb{P})$ 為機率空間及 $(E, \mathcal{E})$ 為可測空間，則我們稱可測函數 $X : \Omega \longrightarrow E$ 為取值在 E 中的隨機變數 (random variable)。

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 $\{2, \ldots, 12\}$.

(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

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

$$\begin{tikzcd}
(\Omega,\mathcal{A},\mathbb{P}) \arrow[dd, "X"'] \arrow[dr, "\mathbb{P}"] & \\
& {[0,1]} \\
(E,\mathcal{E}) \arrow[ur, "\mathbb{P}_X"] &
\end{tikzcd}$$

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 $\mathbb{P}$.

範例 2.1.5 ： 我們取上章節範例 2.1.2 中的機率空間，並考慮在上面定義的隨機變數。

(1) 定義 $X((i,j)) = i + j$，那麼 X 是個值域為 $\{2, \ldots, 12\}$ 的隨機變數。

(2) 定義 $H(\omega)$ 為「在 ω 之中，正面出現的次數」，那麼 H 是個值域為 $\{0, 1, 2, 3\}$ 的隨機變數。

範例 2.1.6 ： 考慮範例 2.1.3 定義的機率空間，並考慮隨機變數

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

則隨機變數 X 及 Y 取值在 $[-1, 1]$ 之中並可被視為均勻隨機方向在 x 軸及 y 軸上的投影。

隨機變數 $X : \Omega \longrightarrow E$ 的定義域（也就是機率空間）本身並沒這麼重要，我們比較在意的是它的值域，也就是說，他取不同值的機率。換句話說，我們想要理解機率測度 $\mathbb{P}$ 在 X 函數作用下的影像測度。

定義 2.1.7 ： 定義 $\mathbb{P}_X$ 為測度 $\mathbb{P}$ 在隨機變數 X 之下的影像測度 (image measure) ，我們稱之為隨機變數 X 的分佈 (distribution) 或是律 (law)。換句話說，$\mathbb{P}_X$ 是個在可測空間 $(E, \mathcal{E})$ 之上的機率測度，可以寫作

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

在機率論的語言中，通常我們也會將上述縮寫為

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

$$\begin{tikzcd}
(\Omega,\mathcal{A},\mathbb{P}) \arrow[dd, "X"'] \arrow[dr, "\mathbb{P}"] & \\
& {[0,1]} \\
(E,\mathcal{E}) \arrow[ur, "\mathbb{P}_X"] &
\end{tikzcd}$$

圖 2.1: 描述影像測度 $\mathbb{P}_X$ 的示意圖，此影像測度也可以看作為機率測度 $\mathbb{P}$ 透過隨機變數 X 向前推進 (pushforward) 所得到的測度，也記作 $X_*\mathbb{P}$。我們真正在意的會是機率測度 $\mathbb{P}_X$ 而不是 $\mathbb{P}$。

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.

我們應該以下面方式來理解影像測度：對於任意機率空間中的一點 $\omega \in \Omega$，$X(\omega)$ 是個在 E 中的點，那麼 $\mathbb{P}_X(B)$ 則表示這個隨機點在 B 中的機率。

範例 2.1.8 ： 令 $(\Omega, \mathcal{A}, \mathbb{P})$ 為機率空間，其中 $\Omega = [0, \pi]$、$\mathcal{A}$ 是伯雷爾 σ 代數，以及 $\mathbb{P} = \frac{1}{\pi}\lambda$。考慮隨機變數 $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}$ 定義做：

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

影像測度 $\mathbb{P}_X$ 是定義在可測空間 $([0,1], \mathcal{B}([0,1]))$ 上。我們可以透過 $\mathbb{P}_X([a,b])$ 對於所有 $0 \leqslant a \leqslant b \leqslant 1$ 的值，來刻劃這個影像測度。對於 $0 \leqslant a \leqslant b \leqslant 1$，我們有

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

註解 2.1.9 ： 若隨機變數 $X : (\Omega_1, \mathcal{A}_1, \mathbb{P}_1) \longrightarrow (E, \mathcal{E})$ 及 $Y : (\Omega_2, \mathcal{A}_2, \mathbb{P}_2) \longrightarrow (E, \mathcal{E})$ 有相同的分佈，也就是說 $\mathbb{P}_{1,X} := X_*\mathbb{P}_1 = Y_*\mathbb{P}_2 =: \mathbb{P}_{2,Y}$，則我們可以記作

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

若我們要說隨機變數 X 的分佈是 μ，也就是說 $\mathbb{P}_{1,X} = \mu$，則我們也可以記作

$$X \sim \mu,$$

並說 X 遵從 μ 的分佈。

註解 2.1.10 ： 在 $(E, \mathcal{E}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ 的情況下，$\mathbb{P}_X$ 是個在 $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ 上的機率測度，若他對 $\mathbb{R}^d$ 上的勒貝格測度 λ 絕對連續，則根據 Radon–Nikodym 定理（定理 1.3.16 ），我們知道存在密度函數 $g : \mathbb{R}^d \longrightarrow \mathbb{R}_+$ 使得我們可以將 $\mathbb{P}_X$ 寫作 $\mathbb{P}_X = g \cdot \lambda$。我們稱 g 為隨機變數 X 的機率密度函數 (probability density function)。

註解 2.1.11 【隨機變數的正則構造】 **：** 若 μ 是個在 $\mathbb{R}^d$ 上的機率測度，我們可以用下列方式來構造一個分佈為 μ 的隨機變數：取 $\Omega = \mathbb{R}^d$，$\mathcal{A} = \mathcal{B}(\mathbb{R}^d)$，$\mathbb{P} = \mu$ 並且令 $X(\omega) = \omega$，我們可以檢查 X 的分佈是 μ。稍後在命題 2.1.23 中，當我們給定 $\mathbb{R}$ 上的機率測度 μ 時，我們會看到該如何從均勻測度出發，來構造分佈為 μ 的實數隨機變數。

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.

第三小節 期望值

有了隨機變數之後，第一個重要的量是其期望值。

定義 2.1.12 ： 令 X 為一值域為 $\mathbb{R}$ 的隨機變數（也稱為實隨機變數），在下列情況下：

- X 非負；
- X 可取正負值，且 $\int |X| \, \mathrm{d}\mathbb{P} < \infty$，

我們可以定義其期望值 (expectation) ，並記作

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

若 $X = (X_1, \ldots, X_d)$ 是個 d 維的實隨機變數（也就是其值域為 $\mathbb{R}^d$），若所有的期望值 $\mathbb{E}[X_i]$ 都是有定義的，我們可以定義 X 的期望值為 $\mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])$。

註解 2.1.13 ： 令 B 為一可測子集以及 $X = \mathbb{1}_B$，則 $\mathbb{E}[X] = \mathbb{P}(B)$。一般來講，我們可以將 $\mathbb{E}[X]$ 視為隨機變數 X 的「平均值」，例如在 Ω 為有限集合，且 $\mathbb{P}$ 為均勻分佈的情況下，$\mathbb{E}[X]$ 代表的是 X 所有可能取的值的（加權）平均值。

命題 2.1.14 ： 令 X 為一值域為 $(E, \mathcal{E})$ 的隨機變數，則對於任何可測函數 $f : E \longrightarrow [0, \infty]$，我們有

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

證明： 我們首先證明對於任意可測集合 $B \in \mathcal{E}$，指標函數 $f = \mathbb{1}_B$ 滿足此式，接著透過線性組合，證明對於所有的非負簡單函數皆為真。再來，由於非負可測函數可以寫作非負簡單函數序列的非遞減極限（命題 1.2.14），根據單調收斂定理，得證。 □

註解 2.1.15 ： 若 f 不恆正，只要期望值 $\mathbb{E}[|f(X)|]$ 是有限的，上述命題依然為真。不要忘記可取正負值的一般函數，他的積分也是以相同方式定義的，請見定義 1.2.18 。

註解 2.1.16 ： 機率分佈 $\mathbb{P}_X$ 可以讓我們輕易計算 $f(X)$ 類型的隨機變數的期望值。反過來看，對於任意可測集合 $B \in \mathcal{E}$，若我們考慮 $f = \mathbb{1}_B$，則 $\mathbb{E}[f(X)] = \mathbb{P}(X \in B) = \mathbb{P}_X(B)$。不要忘記，機率測度 $\mathbb{P}_X$ 是定義在 $\mathcal{E}$ 上的，也就是說，如果我們知道 $(\mathbb{P}_X(B))_{B \in \mathcal{E}}$，則此機率測度是被唯一決定的。換句

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

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

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

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

話說，若是我們有辦法對「夠多的」可測函數 f 寫出下列式子

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

則透過上面的討論，我們知道隨機變數 X 的分佈會是被機率測度 ν 所描述的。在實際操作上，我們可以對所有非負可測函數，或是在 $\mathcal{C}_c(\mathbb{R}^d)$ 中所有的函數來計算式 (2.2)。

範例 2.1.17 ： 我們考慮與範例 2.1.8 中相同的範例，並且比較兩邊的計算。回顧 $(\Omega, \mathcal{A}, \mathbb{P})$ 是個機率空間，其中 $\Omega = [0, \pi]$、$\mathcal{A}$ 是伯雷爾 σ 代數，以及 $\mathbb{P} = \frac{1}{\pi}\lambda$。隨機變數 $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}$ 定義做：

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

我們可以透過對於所有非負可測函數 $f : \mathbb{R} \to \mathbb{R}$，計算 $\mathbb{E}[f(X)]$ 的值，來刻劃 X 的分佈。對於非負可測函數 $f : \mathbb{R} \to \mathbb{R}$，我們記

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

其中在第二行，我們使用積分項在 $\omega \mapsto \pi - \omega$ 下的對稱性；在第三行，我們使用變數變換 $x = \sin\omega$。這代表著，影像測度 $\mathbb{P}_X$ 寫做：

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

這是個對於勒貝格測度來說，是絕對連續的測度，他的密度函數為 $x \mapsto \frac{2}{\pi}\frac{1}{\sqrt{1-x^2}}$。

下列命題給我們一個重要的例子，來幫助理解這樣的概念。

命題 2.1.18 ： 令 $(X_1, \ldots, X_d)$ 為一值域為 $\mathbb{R}^d$ 的隨機變數。假設其機率密度函數為 $p(x_1, \ldots, x_d)$，則對於任何 $1 \leqslant j \leqslant d$，隨機變數 X_j 的機率密度函數也存在，且可以寫作

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

註解 2.1.19 ： 考慮一個二維隨機變數的情況下（$d = 2$），我們有

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

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 $X = (X_1, \dots, X_d)$ 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 $\mathbb{P}_X$.

證明： 令 π_j 為對於 j 座標的投影：$\pi_j(x_1, \dots, x_d) = x_j$，對於所有非負可測函數 $f : \mathbb{R} \longrightarrow \mathbb{R}_+$，我們可以使用富比尼定理：

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

□

註解 2.1.20 ： 若 $X = (X_1, \dots, X_d)$ 是個值域 d 維的實隨機變數，我們稱 X_j 的分佈為邊緣分佈 (marginal distribution)，並記為 $\mathbb{P}_{X_j}$。根據上述命題，我們可以輕易得到 $\mathbb{P}_{X_j} = (\pi_j)_* \mathbb{P}_X$，也就是說，從 X 的分佈我們可以輕易求得任意 X_j 的分佈。需要知道的是，若我們知道所有的邊緣分佈 $\mathbb{P}_{X_j}$，我們是無法找回 X 的分佈的。

問題 2.1.21： 試著構造兩個二維的實隨機變數 $X = (X_1, X_2)$ 及 $X' = (X_1', X_2')$ 使得對於 $j = 1, 2$，X_j 及 X_j' 有著相同的邊緣分佈，但 X 及 X' 的分佈卻不相同。

第四小節 累積分佈函數

定義 2.1.22 ： 令 X 為一實隨機變數，我們定義函數 $F_X : \mathbb{R} \longrightarrow [0, 1]$ 為

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

我們稱他為隨機變數 X 的累積分佈函數 (cumulative distribution function)，記作 c.d.f.，也可以稱作分佈函數 (distribution function)。

我們提醒在定理 1.2.29 中有看到，F_X 是個非遞減、右連續，且在 $-\infty$ 及 $+\infty$ 的極限分別為 0 及 1 的函數。此外我們知道，逆命題也成立，也就是說如果 F 是個滿足上列條件的函數，則我們可以找到唯一的機率測度 μ 使得對於所有 $t \in \mathbb{R}$，我們有 $F(t) = \mu((-\infty, t])$。也就是說 F 是個實隨機變數的累積分佈函數。

此外，累積分佈函數 F_X 也可以完整描述隨機變數 X 的分佈 $\mathbb{P}_X$，其中我們有：

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

且 F_X 中的不連續點為 $\mathbb{P}_X$ 的原子。

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

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

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.

若 $X := (X_1, \dots, X_d)$ 是個取值在 $\mathbb{R}^d$ 中的隨機變數，我們也能夠以類似的方式來討論上面的概念。對於所有 $(t_1, \dots, t_d) \in \mathbb{R}^d$，我們可以定義

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

命題 2.1.23 ： 令 $F : \mathbb{R} \longrightarrow [0, 1]$ 是個非遞減、右連續，且在 $-\infty$ 及 $+\infty$ 的極限分別為 0 及 1 的函數。定義函數 $h : (0, 1) \longrightarrow \mathbb{R}$ 為

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

若 Y 是個隨機變數使得 $\mathbb{P}_Y$ 的分佈與 $[0, 1]$ 上勒貝格測度相同，則隨機變數 $h(Y)$ 的分配函數會是 F，也就是說 $F_{h(Y)} = F$。

註解 2.1.24 ： 在此命題的敘述中，我們注意到，若 F 是個雙射函數（也就是嚴格遞增），則我們有 $h = F^{-1}$。

證明： 固定 $y \in (0, 1)$ 及 $x \in \mathbb{R}$。若 $F(x) \geqslant y$，根據 h 的定義，我們可以得到 $x \geqslant h(y)$。反過來，若 $F(x) < y$，使用 F 的右連續性，我們可以找到 $\varepsilon > 0$ 使得 $F(x + \varepsilon) < y$，再根據單調性，對於所有 $z \leqslant x + \varepsilon$，我們會有 $F(z) < y$，因此 $h(y) \geqslant x + \varepsilon > x$。因此，我們有下列的等價關係：

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

令 Y 為隨機變數使得 $\mathbb{P}_Y$ 是在 $[0, 1]$ 上的勒貝格測度，以及 $Z := h(Y)$。我們注意到，根據命題 1.2.2，Z 也會是一個隨機變數。由於 $\mathbb{P}(Y \in (0, 1)) = 1$，我們有

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

□

註解 2.1.25 ： 從命題 2.1.23，我們得知若均勻隨機變數存在（這是我們的假設，也是測度論會去構造的東西，這裡不深入討論），則對於任何在 $\mathbb{R}$ 上給定的分佈，我們可以透過他的分佈函數來構造有此分佈的實隨機變數。

第二節 隨機變數的生成 σ 代數

令 $(E, \mathcal{E})$ 為一可測空間。

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,

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

定義 2.2.1 ： 若 $X : (\Omega, \mathcal{A}) \longrightarrow (E, \mathcal{E})$ 為隨機變數，我們可以定義 $\sigma(X)$ 為 $\mathcal{A}$ 最小的子 σ 代數，使得 X 是可測的，我們稱之為<u>由 X 生成的 σ 代數</u>：

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

註解 2.2.2 ： 我們可以將 $\sigma(X)$ 視為需要能夠正確描述隨機變數 X 的最小的 σ 代數。

範例 2.2.3 ： 固定機率空間 $(\Omega, \mathcal{A}, \mathbb{P})$ 並考慮隨機變數 $X : (\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$。

(1) 若隨機變數 X 取值恆為常數，則 $\sigma(X) = \{\varnothing, \Omega\}$ 為平凡 σ 代數。

(2) 給定可測集合 $A \in \mathcal{A}$，若隨機變數 X 寫作 $X = \mathbb{1}_A$，則 $\sigma(X) = \{\varnothing, A, A^c, \Omega\}$。

(3) 若隨機變數 X 只會取 0 及 1 兩個值，則存在 $A \in \mathcal{A}$ 使得 $X = \mathbb{1}_A$，因此 $\sigma(X)$ 的描述與上面情況像同。

(4) 我們不難驗證，對於任意 $A, B \in \mathcal{A}$ 滿足 $A \notin \{\varnothing, B, B^c, \Omega\}$，隨機變數 $\mathbb{1}_A$ 並不會是 $\sigma(\mathbb{1}_B)$ 可測的，我們看到在樣本空間上選擇不同的 σ 代數，也會造成隨機變數的概念改變。

範例 2.2.4 ： 考慮機率空間 $(\Omega, \mathcal{A}, \mathbb{P}) := ([0,2), \mathcal{B}([0,2)), \frac{1}{2}\lambda)$，其中 λ 為勒貝格測度。我們定義隨機變數 $X, Y : (\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow ([0,2), \mathcal{B}([0,2)))$ 做 $X(\omega) = \omega$ 及 $Y(\omega) = \lfloor \omega \rfloor$，則

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

我們可以注意到 $([0,2), \sigma(Y), \mathbb{P})$ 可以被視為一個離散機率空間：定義

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

則 $(\Omega, \mathcal{A}, \mathbb{P})$ 與 $(\Omega', \mathcal{A}', \mathbb{P}')$ 有<u>相同</u>的結構。

註解 2.2.5 ： 我們可以將這個定義拓展到任意集合的隨機變數 $X = (X_i)_{i \in I}$。假設對任意 $i \in I$，X_i 的值域為 $(E_i, \mathcal{E}_i)$，我們可以定義

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

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.

命題 2.2.6 ： 令 X 為值域在 $(E, \mathcal{E})$ 中的隨機變數，Y 為實隨機變數，則下列性質等價：

(1) Y 對 $\sigma(X)$ 可測。

(2) 存在一可測函數 $f : (E, \mathcal{E}) \longrightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ 使得 $Y = f(X)$。

證明：根據命題 1.2.2 ，若 (2) 成立，則 (1) 成立。

假設 Y 對 $\sigma(X)$ 是可測的，我們想要證明 (2). 首先處理當 Y 為簡單函數的狀況：對於 $i \in \{1, \dots, n\}$，令 $\lambda_i \in \mathbb{R}$ 以及 $A_i \in \sigma(X)$，且 Y 寫作

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

對於所有 $i \in \{1, \dots, n\}$，我們可以找到 $B_i \in \mathcal{E}$ 使得 $A_i = X^{-1}(B_i)$，所以 Y 可以重新寫作

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

其中 $f = \sum_{i=1}^n \lambda_i \mathbb{1}_{B_i}$ 是個對於 $\mathcal{E}$ 可測的函數。

在一般情況下，存在對 $\sigma(X)$ 可測的簡單函數序列 (Y_n) 使得 Y_n 簡單收斂至 Y（命題 1.2.14），因此根據此證明前半部所述，對於所有 n，我們可以找到可測函數 $f_n : E \longrightarrow \mathbb{R}$ 使得 $Y_n = f_n(X)$。對於所有 $x \in E$，我們可以設

$$f(x) = \begin{cases} \lim_{n\to\infty} f_n(x) & \text{如果極限存在,} \\ 0 & \text{其他情況.} \end{cases}$$

根據命題 1.2.6 ，函數 f 仍然為可測。此外，我們注意到，對於所有 $\omega \in \Omega$，我們有

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

若 $x = X(\omega)$，則在上面的定義中，$\lim f_n(x)$ 的極限是存在的，而且我們有

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

也就是說 $Y = f(X)$。□

第三節 常用機率分佈

在這章節中，我們介紹一些常用的隨機變數分佈。

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,

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

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,

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

在第 2.3.1 小節中，我們會刻劃兩種類別的機率分佈：離散分佈以及絕對連續分佈。在第 2.3.2 小節中，我們給出常用的離散隨機變數分佈的例子，在第 2.3.3 小節中我們給出常用的絕對連續隨機變數的例子。

第一小節 機率分佈的分類

我們這裡探討兩個不同類型的隨機變數：離散型隨機變數以及絕對連續型隨機變數。

令 $(\Omega, \mathcal{A}, \mathbb{P})$ 為一機率空間且 $(E, \mathcal{E})$ 為一可測空間，考慮一個值域為 E 的隨機變數 $X : \Omega \longrightarrow E$。

定義 2.3.1 【離散型隨機變數】： 當 E 為一可數集合，且 $\mathcal{E} = \mathcal{P}(E)$ 時，X 的分佈可以寫作

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

其中 $p_x = \mathbb{P}(X = x)$，δ_x 代表的是在 x 的狄拉克測度。上述式子可以透過下列式子來理解：

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

從上述性質可以得知，如果想要知道一離散隨機變數 X 的分佈，我們需要知道的是對於所有 $x \in E$，機率 $\mathbb{P}(X = x)$ 的值。

當機率空間不為離散時，我們這裡只討論下面的情況，也就是可以定義密度函數的情況。

定義 2.3.2 【絕對連續型隨機變數】： 假設 $(E, \mathcal{E}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$，若 $\mathbb{P}_X$ 對於勒貝格測度 λ 是絕對連續的話，我們說 $\mathbb{P}_X$ 是個絕對連續機率分佈 (absolutely continuous probability distribution)，而 X 是個絕對連續隨機變數 (absolutely continuous random variable)。在這樣的情況下，我們可以使用定理 1.3.16 中的 Radon–Nikodym 定理，以建構一個非負可測函數 $p : \mathbb{R}^d \to \mathbb{R}_+$ 使得

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

除了在一測度為零的集合之上，此函數 p 的定義是唯一的，我們稱之為 X 的密度函數 (density function) 或機率密度函數 (probability density function)。此外，在一維的情況下 $d = 1$，對於所有 $\alpha \leqslant \beta$，我們有

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

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 (均匀分佈) 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 $\{1, \dots, 6\}$.

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 $E = \{0, \dots, n\}$ 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 $X \sim \mathrm{Bin}(n, p)$.

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

註解 2.3.3 ： 若 $\mathbb{P}_X$ 對於任意 $\mathbb{R}^d$ 中的單點測度皆為零，則我們說 $\mathbb{P}_X$ 是個連續測度或連續分佈，且 X 是個連續隨機變數。我們不難驗證，絕對連續測度（或隨機變數）也會是連續的；但在習題 1.33 中我們看到，Cantor 分佈是個連續但不是絕對連續的（機率）測度。

第二小節 離散型隨機變數

下面我們介紹在未來會常常使用的離散機率分佈。令 $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ 為離散機率空間，考慮值域空間 E 以及隨機變數 $X : \Omega \longrightarrow E$，將對應到機率分佈 $\mathbb{P}_X$ 的質量函數記作 f_X，根據不同的下列狀況我們會有不同的分佈。

均勻分佈 (Uniform distribution) 若 E 為有限集合，$n = |E|$ 為其元素個數，且質量函數 f_X 滿足以下條件：

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

我們稱 X 為在 E 之上的均勻分佈，記作 $X \sim \mathrm{Unif}(E)$。

> 例 ： 投擲一顆六面的公正骰子，點數分別為 $\{1, \dots, 6\}$。

伯努力分佈 (Bernoulli distribution) 給定參數 $p \in [0, 1]$，若 $E = \{0, 1\}$ 且 f_X 滿足下列條件：

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

我們稱 X 為參數為 p 的伯努力分佈，記作 $X \sim \mathrm{Ber}(p)$。

> 例 ： 投擲一顆不公正硬幣，得到正面（記作 1）的機率為 p，反面（記作 0）的機率為 $1 - p$。

二項式分佈 (Binomial distribution) 給定參數 $n \in \mathbb{N}_0$ 及 $p \in [0, 1]$，若 $E = \{0, \dots, n\}$ 且 f_X 滿足下列條件：

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

我們稱 X 為參數為 (n, p) 的二項式分佈，記作 $X \sim \mathrm{Bin}(n, p)$。

> 例 ： 考慮上述不公正硬幣，投擲 n 次，則正面出現的次數滿足二項式分佈。

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 $X \sim \mathrm{Geo}(p)$.

> 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.

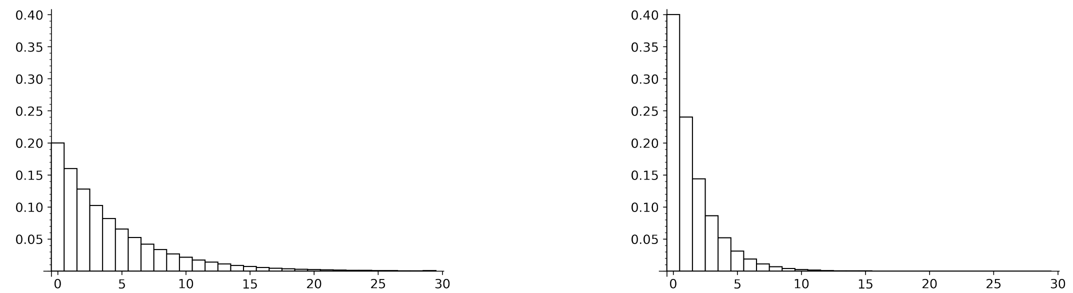

Figure 2.3: Mass functions of the geometric distribution with parameters $n = 30, p = 0.2$ on the left and $n = 30, p = 0.4$ on the right.

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 $X \sim \mathrm{Hypergeo}(N, K, n)$.

> 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 $\mathrm{Hypergeo}(N, K, n)$.

圖 2.2: 二項式分佈的質量函數，左圖的參數為 $n = 30, p = 0.2$，右圖的參數為 $n = 30, p = 0.5$。

幾何分佈 (Geometric distribution) 給定參數 $p \in (0, 1)$，若 $E = \mathbb{N}_0$ 且 f_X 滿足下列條件：

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

我們稱 X 為參數為 p 的幾何分佈，記作 $X \sim \mathrm{Geo}(p)$。

> 例 ： 投擲上述不公正硬幣，在得到第一次正面之前反面出現的次數，滿足幾何分佈。若我們把正面視為「成功」，反面視為「失敗」，則此分佈描述成功前必須要經過的失敗次數。

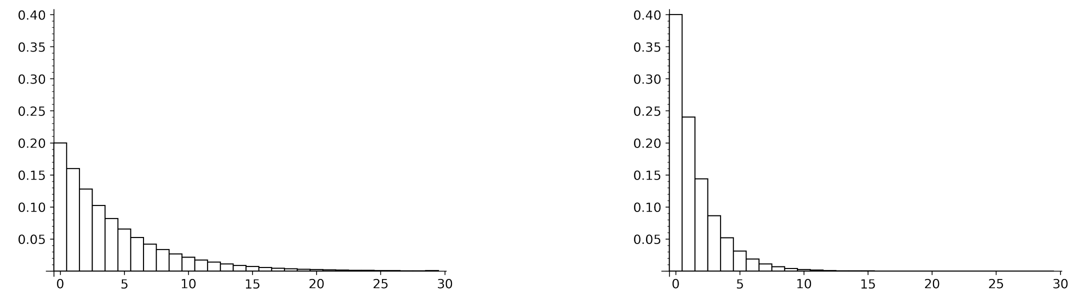

圖 2.3: 幾何分佈的質量函數，左圖的參數為 $n = 30, p = 0.2$，右圖的參數為 $n = 30, p = 0.4$。

超幾何分佈 (Hypergeometric distribution) 給定非負整數 N, K, n 滿足 $0 \leqslant K \leqslant N$，取 $E = \{(n - N + K) \vee 0, \ldots, K \wedge n\}$ 且 f_X 滿足下列條件：

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

我們稱 X 為參數為 (N, K, n) 的超幾何分佈，記作 $X \sim \mathrm{Hypergeo}(N, K, n)$。

> 例 ： 在一個球箱中，有 N 顆球，其中 K 顆是白色的，當我們進行不重複抽樣取 n 顆球時，抽到的白球數量會滿足超幾何分佈 $\mathrm{Hypergeo}(N, K, n)$。

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 $X \sim \mathrm{Pois}(\lambda)$.

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 $\mathrm{Bin}(n, p_n)$, 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 $X \sim \mathrm{Unif}([a, b])$.

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,

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

帕松分佈 (Poisson distribution) 令 $\lambda > 0$ 為一參數，$E = \mathbb{N}_0$ 且 $\mathbb{P}$ 滿足下列條件：

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

我們稱 X 為參數為 λ 的帕松分佈，記作 $X \sim \mathrm{Pois}(\lambda)$。

圖 2.4: 帕松分佈的質量函數，左圖的參數為 $\lambda = 5$，右圖的參數為 $\lambda = 10$。

註解 2.3.4 ： 不管從理論或是應用的角度來看，帕松分佈都具有重要意義。在稍後的章節，當我們探討機率分佈的收斂時，我們可以證明帕松分佈可以被理解為二項式分佈的極限。換句話說，令 X_n 為滿足二項式分佈 $\mathrm{Bin}(n, p_n)$ 的隨機變數，且當 n 趨近無窮大時，我們有 $np_n \to \lambda$，則

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

第三小節 有密度的隨機變數

在這個章節中，我們定義常見的連續隨機變數分佈，並且討論他們的性質。我們將所要討論的連續隨機變數記作 X，並將他的機率密度函數記作 f_X。

均勻分佈 (Uniform distribution) 令 $a < b$，若 p 可寫為

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

則稱 X 為在區間 $[a, b]$ 上的均勻分佈，記作 $X \sim \mathrm{Unif}([a, b])$。

指數分佈 (Exponential distribution) 令 $\lambda > 0$ 為一額外參數，若 f_X 滿足：

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

則稱 X 為參數為 λ 的指數分佈，記作 $X \sim \mathcal{E}(\lambda)$ 或 $X \sim \mathrm{Exp}(\lambda)$。我們不難驗證，若 $X \sim \mathrm{Exp}(\lambda)$，則

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

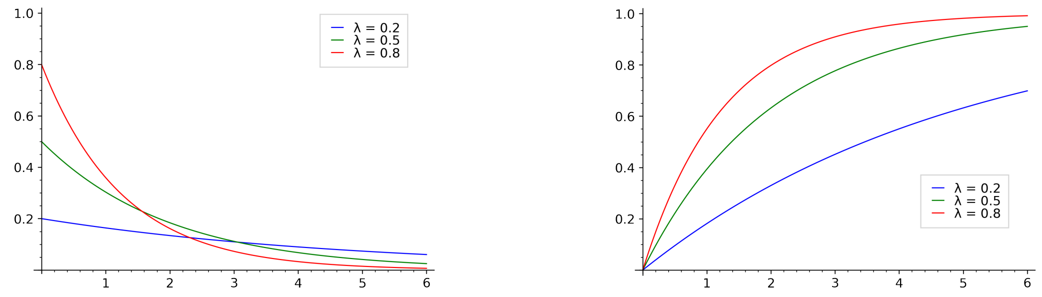

Figure 2.5: The left figure shows the probability density function of the exponential distribution with different parameters, the right figure shows their distribution functions.

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) .

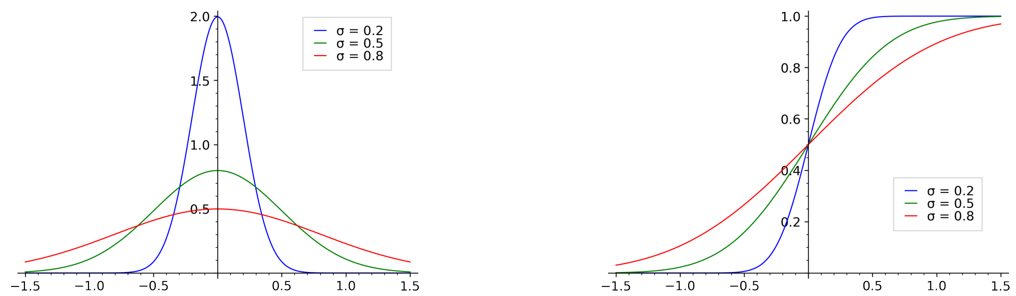

Figure 2.6: The left figure shows the probability density function of the Gaussian distribution with different parameters, the right figure shows their distribution functions. Here we take $\mu = 0$.

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 $X \sim \mathrm{Cauchy}(x_0, \gamma)$. It is not hard to check that the expectation of a Cauchy distribution is not defined. (See Exercise 2.15.)

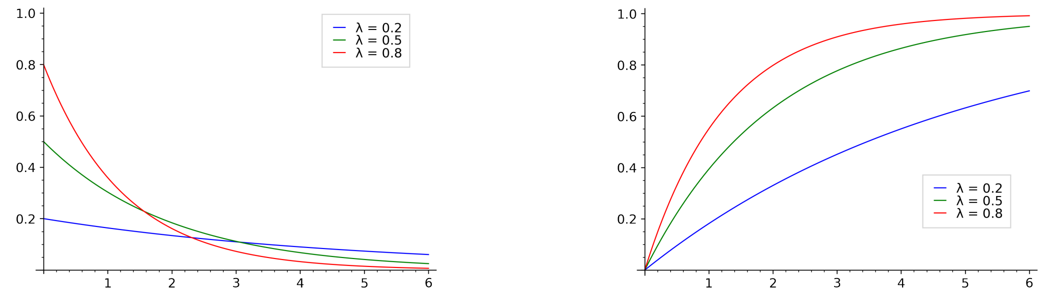

圖 2.5: 左邊是不同參數指數分佈的機率密度函數，右邊是他們的分佈函數。

常態分佈 (Normal distribution) 或高斯分佈 (Gaussian distribution) 令 $m \in \mathbb{R}$ 及 $\sigma > 0$ 為兩額外參數，若 f_X 滿足：

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

則稱 X 為均值為 m 變異數為 σ^2 的高斯分佈，記作 $X \sim \mathcal{N}(m, \sigma^2)$。我們不難驗證，若 $X \sim \mathcal{N}(m, \sigma^2)$，則

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

當 $m = 0$，我們稱 X 置中 (centered)；當 $\sigma = 1$，我們稱 X 約化 (reduced)；當兩者皆成立，我們稱之為標準常態分佈 (standard normal distribution) 。

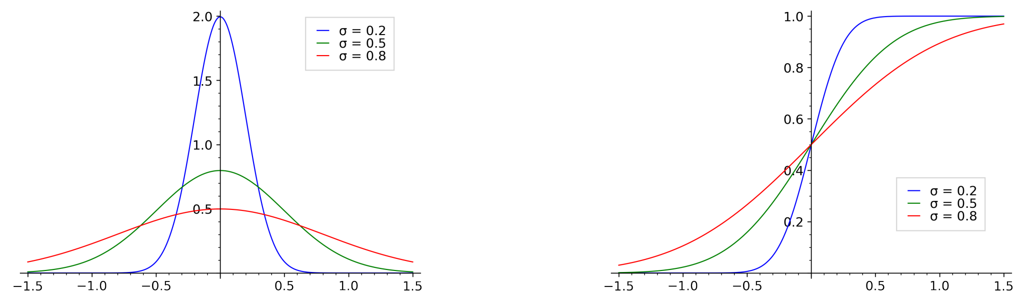

圖 2.6: 左邊是不同高斯分佈的機率密度函數，右邊是他們的分佈函數，這裡我們取 $\mu = 0$。

柯西分佈 (Cauchy distribution) 給定參數 $\gamma > 0$ 及 $x_0 \in \mathbb{R}$，若 f_X 滿足：

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

則稱 X 為參數為 (x_0, γ) 的柯西分佈，記作 $X \sim \mathrm{Cauchy}(x_0, \gamma)$。我們可以檢查，柯西分佈的期望值是沒有定義的。（參照習題 2.15）

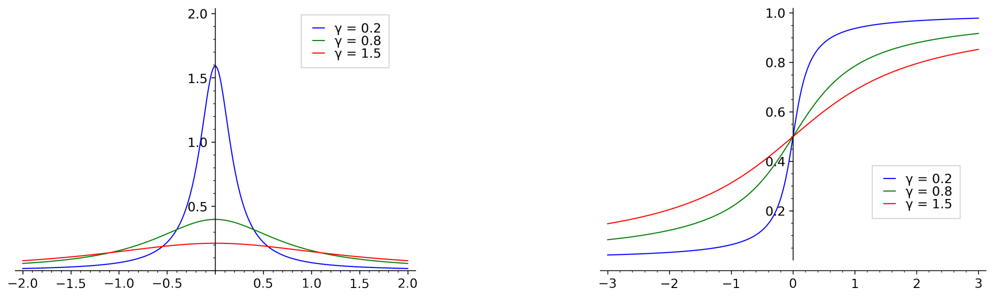

Figure 2.7: The left figure shows the probability density function of the Cauchy distribution with different parameters, the right figure shows their distribution functions. Here we take $x_0 = 0$.

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 $X \sim \mathrm{Gamma}(\alpha, \beta)$. 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 $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

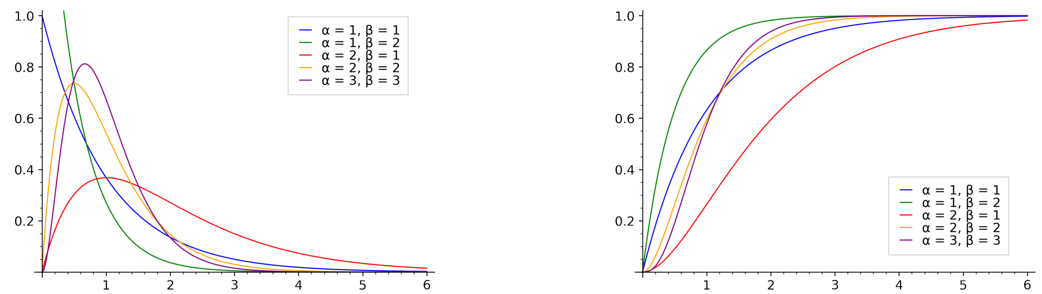

Figure 2.8: The left figure shows the probability density function of the Gamma distribution with different parameters, the right figure shows their distribution functions.

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

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

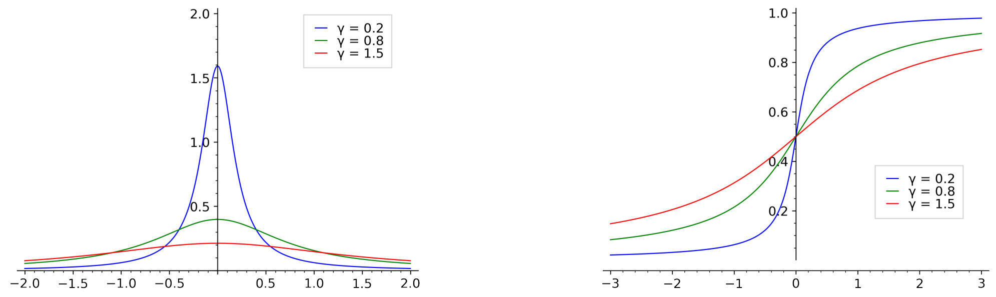

圖 2.7: 左邊是不同柯西分佈的機率密度函數，右邊是他們的分佈函數，這裡我們取 $x_0 = 0$。

珈瑪分佈 (Gamma distribution) 給定參數 $\alpha, \beta > 0$，若 f_X 滿足：

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

其中 $\Gamma(\alpha)$ 為珈瑪函數，定義做

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

則我們稱 X 為參數為 (α, β) 的珈瑪分佈，記作 $X \sim \Gamma(\alpha, \beta)$ 或 $X \sim \mathrm{Gamma}(\alpha, \beta)$。我們回顧珈瑪函數的重要性質：對於 $\alpha > 0$，我們有 $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$（變數變換）、$\Gamma(1) = 1$，以及 $\Gamma(\frac{1}{2}) = \sqrt{\pi}$。

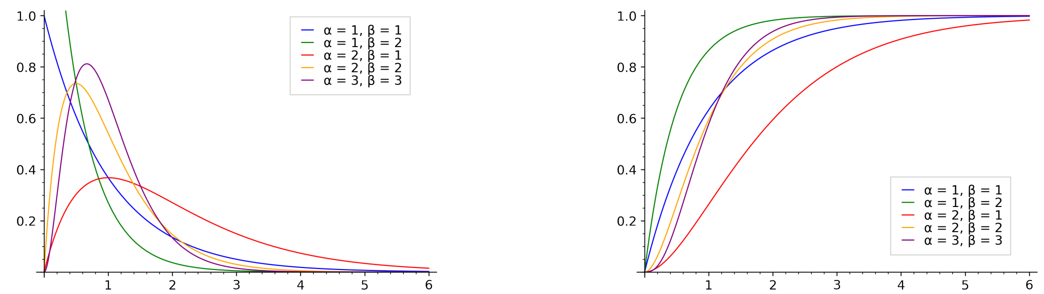

圖 2.8: 左邊是不同珈瑪分佈的機率密度函數，右邊是他們的分佈函數。

貝塔分佈 (Beta distribution) 給定參數 $\alpha, \beta > 0$，若 f_X 滿足：

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

其中 $B(\alpha, \beta)$ 為貝塔函數，定義做

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

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 $X \sim \mathrm{Beta}(\alpha, \beta)$.

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

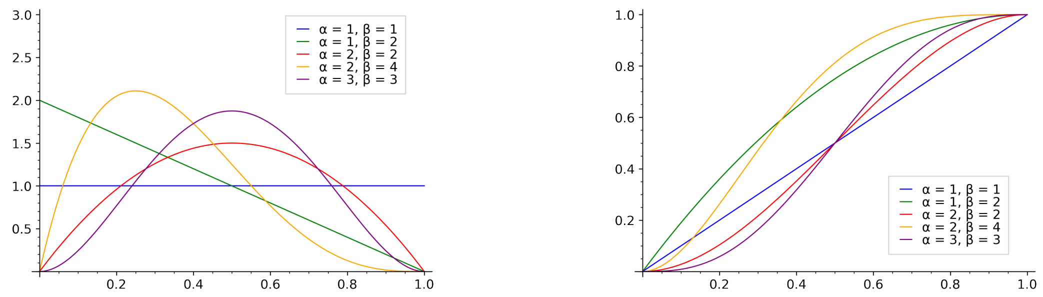

Figure 2.9: The left figure shows the probability density function of the Beta distribution with different parameters, the right figure shows their distribution functions.

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,

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

Its *standard deviation* (標準差) is denoted

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

則我們稱 X 為參數為 (α, β) 的貝塔分佈，記作 $X \sim \mathrm{Beta}(\alpha, \beta)$。

註解 2.3.5 ： 機率分佈 $\mathrm{Beta}(1,1)$ 為在 $[0,1]$ 上的均勻分佈，且我們有下列對稱性：若 $X \sim \mathrm{Beta}(\alpha, \beta)$，則 $1 - X \sim \mathrm{Beta}(\beta, \alpha)$。

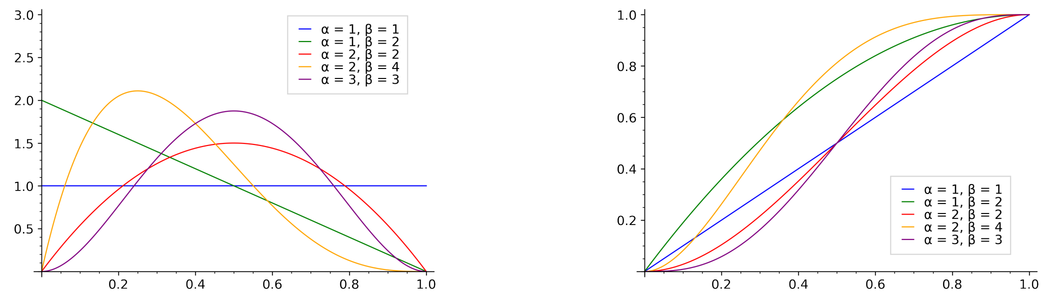

圖 2.9: 左邊是不同貝塔分佈的機率密度函數，右邊是他們的分佈函數。

第四節 隨機變數的動差

動差是隨機變數冪次的期望值，可以讓我們得到一些非線性的特性，例如在統計上有重要意義的變異數。在機率論中，動差方法 (method of moments) 有很多重要的應用，甚至可以拿來證明圖論、數論中各種存在性的問題，可以參見例如習題 2.24 和習題 2.25 。

第一小節 動差及變異數

定義 2.4.1 ： 固定正整數 $k \geqslant 1$，若實隨機變數 X 滿足 $\mathbb{E}[|X|^k] < \infty$，也就是說 $X \in L^k$，則我們說 X 的 k 階動差 (moment) 為有限。我們稱 $\mathbb{E}[X^k]$ 為 X 的 k 階動差。

定義 2.4.2 ： 令 $X \in L^2(\Omega, \mathcal{P}(\Omega), \mathbb{P})$，則 X 的變異數 (variance) 定義為

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

其標準差 (standard deviation) 則記作

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

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

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

註解 2.4.3 ： 我們可以這樣理解上述定義中的變異數 $\mathrm{Var}(X)$：它是一種可以量測隨機變數 X 在其期望值 $\mathbb{E}[X]$ 附近分散程度的數學量。

在式 (2.4) 中，取平方是重要的，不然我們會有

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

而無法正確描述我們的隨機變數 X 到底離他的期望值多遠。

若要避免相消為零的現象發生，我們可以考慮 $\mathbb{E}[|X-\mathbb{E}[X]|]$，他的好處是，我們不需要要求隨機變數 $X \in L^2$，我們只需要 $X \in L^1$ 的假設即足夠，但缺點是，絕對值的規律性比平方來得差，且在計算上比較不容易，因此才會考慮平方。在後面的章節中，我們也會看到取平方在機率理論中的重要性，這與中央極限定理有關係。

我們可以這樣理解上述定義中的變異數 $\mathrm{Var}(X)$：它量測隨機變數 X 在其平均值 $\mathbb{E}[X]$ 附近分散的程度，另外我們可以注意到，若且為若 $\mathrm{Var}(X) = 0$ 則 X 幾乎必然為常數。

命題 2.4.4 ： 變異數 $\mathrm{Var}(X)$ 會滿足下列最佳化問題：

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

證明：我們只需要證明，對於所有實數 a，我們有

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

透過直接的計算，我們有

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

□

我們能利用期望值及變異數估計隨機變數取值的機率：

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

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

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

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

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 $K_Y = AK_XA^T$.

2.4.2 Linear Regression

定義 2.4.5 ： 令 $X, Y \in L^2(\Omega, \mathcal{P}(\Omega), \mathbb{P})$，則它們的共變異數 (covariance) 定義為

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

我們注意到，共變異數的定義是變異數的推廣：$\operatorname{Var}(X) = \operatorname{Cov}(X, X)$。

註解 2.4.6 ： 兩隨機變數 X 及 Y 之間的共變異數測量它們之間的關聯性 (correlation)，這和隨機變數的獨立性是不一樣的。參閱第 3.1.2 小節。

註解 2.4.7 ： 給定離散機率空間 $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$，我們不難檢查共變異數函數

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

是個對稱、雙線性的運算子。換句話說，對於所有 $X, Y, Z \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ 及 $a, b \in \mathbb{R}$，我們有

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

這讓我們可以使用柯西不等式，進而得到

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

此外，若 X 及 Y 其中之一幾乎必然為常數，則 $\operatorname{Cov}(X,Y) = 0$。

定義 2.4.8 ： 若 $X = (X_1, \ldots, X_d)$ 為一值域為 $\mathbb{R}^d$ 的隨機變數，且假設其所有的分量都在 $L^2(\Omega, \mathcal{A}, \mathbb{P})$ 之中，則 X 的共變異數矩陣 (covariance matrix) 定義為

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

問題 2.4.9： 試證明當 X 是個 d 維的實隨機變數時，其共變異數矩陣 K_X 是個半正定 (positive semi-definite) 對稱矩陣，換句話說，證明對於所有的 $\lambda = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^d$，我們有 $\lambda K_X \lambda^T \geqslant 0$。

問題 2.4.10： 若 A 是個 $n \times d$ 維的矩陣，且 X 為 d 維實隨機變數，定義 $Y = AX$，試證明 $K_Y = AK_XA^T$。

第二小節 線性回歸

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

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

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,

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

Proof : Let H be the vector subspace of $L^2(\Omega, \mathcal{A}, \mathbb{P})$ generated by $1, Y_1, \ldots, Y_n$. 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* (特徵函數) $\Phi_X : \mathbb{R}^d \longrightarrow \mathbb{C}$ is defined by,

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

令 $X, Y_1, \ldots, Y_n$ 為在 $L^2(\Omega, \mathcal{A}, \mathbb{P})$ 中的隨機變數，我們想要找出近似 X 而且能被寫為 $1, Y_1, \ldots, Y_n$ 線性組合的隨機變數。換句話說，我們想要找出實數 $\beta_0, \ldots, \beta_n$ 將下列數值最小化：

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

命題 2.4.11 ： 當 (α_j) 為下列線性系統

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

的任意一個解時，設

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

則我們有

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

證明： 令 H 為由 $1, Y_1, \ldots, Y_n$ 生成在 $L^2(\Omega, \mathcal{A}, \mathbb{P})$ 中的子向量空間，由於 Z 會讓泛函函數 $U \in H \mapsto \|X - U\|_2$ 達到最小值，所以 Z 是 X 在 H 之上的正交投影。若我們把 Z 寫作

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

透過正交投影的性質，我們有 $\mathbb{E}[(X - Z) \cdot 1] = 0$，也就是說 $\alpha_0 = \mathbb{E}[X]$。同樣的，對於所有 $1 \leqslant k \leqslant n$，我們有

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

也就是說 $\operatorname{Cov}(X, Y_k) = \operatorname{Cov}(Z, Y_k)$；接著利用式 (2.7) 中 Z 的線性組合，我們可以得到式 (2.5)。

反過來說，若 (α_i) 滿足式 (2.5)，則式 (2.6) 中定義的 Z 是個在 H 中的元素，而且 $X - Z$ 和 H 是正交的，也就是說 Z 是 X 在 H 上的正交投影。 □

第三小節 特徵函數

定義 2.4.12 ： 若 X 是個值域為 $\mathbb{R}^d$ 的隨機變數，則其特徵函數 (characteristic function) $\Phi_X : \mathbb{R}^d \longrightarrow \mathbb{C}$ 定義為

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

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

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

註解 2.4.13 ： 上述定義也可以重新寫作

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

換句話說，我們可以把 Φ_X 視作 $\mathbb{P}_X$ 的傅立葉變換 (Fourier transform) ：$\Phi_X(\xi) = \hat{\mathbb{P}}_X(\xi)$。另外，從勒貝格控制收斂定理可以得知 Φ_X 為一在 $\mathbb{R}^d$ 上的連續有界函數。

我們這裡列出以第 2.3 節中提到的機率分佈為例子，所對應到的特徵函數。

隨機變數 X	特徵函數 $\Phi_X(\xi)$
$\mathrm{Ber}(p)$	$1-p+pe^{\mathrm{i}\,\xi}$
$\mathrm{Bin}(n,p)$	$(1-p+pe^{\mathrm{i}\,\xi})^n$
$\mathrm{Geo}(p)$	$\frac{p}{1-(1-p)e^t}$
$\mathrm{Pois}(\lambda)$	$\exp(\lambda(e^{\mathrm{i}\,\xi}-1))$
$\mathrm{Unif}([a,b])$	$\frac{e^{\mathrm{i}\,\xi b}-e^{\mathrm{i}\,\xi a}}{\mathrm{i}\,\xi(b-a)}$
$\mathcal{E}(\lambda)$	$\frac{\lambda}{\lambda-\mathrm{i}\,\xi}$
$\mathcal{N}(m,\sigma^2)$	$\exp(\mathrm{i}\,m\xi-\frac{1}{2}\sigma^2\xi^2)$

在這小節，我們將會證明特徵函數可以完全決定隨機變數的機率分佈。

首先，我們證明常態分佈在傅立葉變換 $\mathcal{F}:\mathbb{P}_X\mapsto\hat{\mathbb{P}}_X$ 之下的不變性 (invariance)。

引理 2.4.14 ： 若 X 為一分佈為 $\mathcal{N}(0,\sigma^2)$ 的隨機變數，則

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

證明： 透過被積分式的奇偶性，我們可以得到 $\Phi_X(\xi)$ 的虛數部份為零，所以我們要計算

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

我們對上述式子微分，由於可積函數 $x\mapsto |x|e^{-x^2/2}$ 是 $x\mapsto x\sin(\xi x)e^{-x^2/2}$ 的上界，所以微分算

$x \mapsto x\sin(\xi x)e^{-x^2/2}$, the differential operator can be inverted with the integration operator, i.e.,

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

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,

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

子可以移至積分之內，也就是說

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

我們對上述式子進行部份積分，可以得到

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

因此我們證明出來 f 是微分方程 $f'(\xi) = -\xi f(\xi)$ 且初始值為 $f(0) = 1$ 的解，由於唯一性，我們得到 $f(\xi) = \exp(-\xi^2/2)$。 □

定理 2.4.15 ： 任一隨機變數 X 的特徵函數可以完全定義其分佈，換句話說，傅立葉變換 $\mathcal{F} : \mathbb{P}_X \mapsto \hat{\mathbb{P}}_X$ 在 $\mathbb{R}^d$ 之上的分佈所構成的空間是單射 (injective) 的。

證明：首先，假設一維的情況 $d = 1$。對於所有的 $\sigma > 0$，我們可以定義常態分佈 $\mathcal{N}(0, \sigma^2)$ 的密度函數

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

若 μ 為一在 $\mathbb{R}$ 之上的機率測度，我們定義：

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

要證明此定理，我們將分兩步驟：

(1) μ_σ 可以完全被 $\widehat{\mu}$ 所定義；

(2) 對於任何的函數 $\varphi \in \mathcal{C}_b(\mathbb{R})$，當 $\sigma \to 0$ 我們有下列收斂 $\int \varphi(x)\mu_\sigma(\mathrm{d}x) \longrightarrow \int \varphi(x)\mu(\mathrm{d}x)$。

首先，我們來證明第一點。透過引理 2.4.14 ，我們可以得到，對於所有的 $x \in \mathbb{R}$，有

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

接著，我們可以把 $f_\sigma(x)$ 重新寫作

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

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,

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

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 $X = (X_1,\dots,X_n)$ 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 $\xi = (\xi_1,\dots,\xi_n)$ tends to 0, we have,*

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

在上式的倒數第二個等式當中，我們用富比尼定理（定理 1.4.3），因為 μ 是個機率測度且 $g_{1/\sigma}$ 對勒貝格測度來說，是個可積函數。

接著證明第二點，首先注意到，對於所有的 $\varphi \in C_b(\mathbb{R})$，我們有

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

其中我們用到富比尼定理以及 g_σ 是個偶函數的特性。接著，我們使用 g_σ 的其他特性：

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

這讓我們簡單得到

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

由於對於所有的 $\sigma > 0$，$|g_\sigma * \varphi| \leqslant \sup|\varphi|$，透過勒貝格收斂定理，我們有

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

因此此定理在一維的情況下為真。

在 d 維的情況下，證明相似。我們定義下列函數

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

並且利用對於所有 $\xi \in \mathbb{R}^d$，我們有

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

□

問題 2.4.16：將 $C_b(\mathbb{R})$ 中緊緻支撐 (compactly supported) 函數構成的集合記作 $C_c(\mathbb{R})$。證明若 $\varphi \in C_c(\mathbb{R})$，則可以強化式 (2.10) 中的收斂為在 $\mathbb{R}$ 上的均勻收斂；換句話說，證明

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

命題 2.4.17 ：令 $X = (X_1,\dots,X_n)$ 為一 n 維的實隨機變數，假設 $\|X\|_2^2$ 是個可積的隨機變數，則 Φ_X 是個 $\mathcal{C}^2$ 函數，而且當 $\xi = (\xi_1,\dots,\xi_n)$ 趨近於 0 時，我們有

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

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

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

(c) For all $j = 1, \ldots, n$, 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

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

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.

證明：下面三個性質並不難檢查：

(a) 對於所有 $\xi \in \mathbb{R}^n$，函數 $x \mapsto e^{\mathrm{i}\,\xi\cdot x}$ 對於 $\mathbb{P}_X$ 是可積的。

(b) 對於所有 $x \in \mathbb{R}^n$，函數 $\xi \mapsto e^{\mathrm{i}\,\xi\cdot x}$ 對於 ξ 是可微的，而且他在 ξ 的微分寫做：

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

(c) 對於所有 $j = 1, \ldots, n$，偏微分 $x \mapsto \mathrm{i}\,x_j e^{\mathrm{i}\,\xi\cdot x}$ 是連續的，而且可以被 x_j 所控制住，由於 $X_j \in L^2 \subseteq L^1$，這個控制函數對 $\mathbb{P}_X$ 是可積的。

因此，定理 1.2.24 讓我們可以交換微分算子和積分算子的順序，讓我們改寫 Φ_X 的微分，也就是說：

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

此外，定理 1.2.22 告訴我們這些偏微分都是連續的。

對於二階微分，我們可以使用相似的方法。更確切來說，我們會需要用到 $\mathbb{E}[|X_j X_k|] \leqslant \mathbb{E}[X_j]^2\,\mathbb{E}[X_k]^2 < \infty$ 來解釋二階微分和積分可以交換。 □

要小心的是，不同於特徵函數可以完全描述機率分佈，在一般情況下，只知道動差 $(\mathbb{E}[X^k])_{k\geqslant 0}$ 並不足以描述 X 的機率分佈。在習題 2.22 中，我們會看到反例，也就是說，我們會構造出（至少）兩個不同的機率分佈，使得他們的所有動差皆是相等的。接著在習題 2.23 中，我們會討論在什麼樣的情況之下，動差可以唯一決定機率分佈。

第四小節 生成函數

當離散隨機變數的值域為非負整數 $\mathbb{N}_0$ 時，我們可以定義他的生成函數 (generating function) 。

定義 2.4.18 ： 令 X 為一值域為 $\mathbb{N}_0$ 的隨機變數，其生成函數 G_X 定義在下列級數會收斂的 $s \in \mathbb{C}$ 上：

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

反過來說，若我們知道離散隨機變數 X 的生成函數，透過讀取他的係數，我們也可以描述 X 的分佈。

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

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

由式 (2.11) 定義出來的級數會滿足一些基本性質，這部份屬於微積分的先備知識，這裡只做回顧。

收斂性 存在收斂半徑 (radius of convergence) $0 \leqslant R \leqslant \infty$ 使得當 $|s| < R$ 時，級數 $G_X(s)$ 會收斂；當 $|s| > R$ 時，級數 $G_X(s)$ 會發散。此外，對於所有 $R' < R$，級數 $G_X(s)$ 在 $\{s : |s| \leqslant R'\}$ 上會均勻收斂。這裡我們注意到，由於 $G_X(1) = 1$，所以顯然有 $R \geqslant 1$。

微分性 在定義域 $\{s : |s| < R\}$ 上，我們可以一項一項對級數 $G_X(s)$ 微分或積分無限多次。

唯一性 若存在 $0 < R' \leqslant R$ 使得當 $|s| < R'$ 時，等式 $G_X(s) = G_Y(s)$ 成立，則對於所有 $n \in \mathbb{N}_0$，我們有 $\mathbb{P}(X = n) = \mathbb{P}(Y = n)$。此外，我們有

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

連續性 【Abel 定理】由於各項 $\mathbb{P}(X = n)$ 皆為非負，若 $R < \infty$，我們有

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

命題 2.4.19 ： 我們可以透過 G_X 的微分來計算 X 的期望值：

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

更一般來說，對於所有正整數 $p \geqslant 1$，我們有

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

也就是說透過 X 的生成函數，我們可以簡單計算 X 的所有動差。此外，對於 $X \in L^2(\Omega, \mathcal{P}(\Omega), \mathbb{P})$，我們還有

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

證明： 對於所有正整數 $p \geqslant 1$，根據 $G_X^{(p)}(s)$ 在 $\{s : |s| < R\}$ 上的連續性（已知 $R \geqslant 1$），我們有

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

其中 $\varphi(n) = n(n-1)\dots(n-p+1)$，得證。

令 $X \in L^2(\Omega, \mathcal{P}(\Omega), \mathbb{P})$，我們根據前面性質，可以得到

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

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,

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

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 $k \geqslant 0$,*

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

因此 $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = G''_X(1) + G'_X(1) - G'_X(1)^2$，得證。 □

範例 2.4.20 ： 若 X 是滿足二項式分佈 $\mathrm{Bin}(n,p)$ 的隨機變數，則他的生成函數寫作：

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

我們可以計算他的期望值

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

若要計算他的變異數，我們先計算

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

接著帶入

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

當我們的目的是要計算 X 的動差時，我們可以直接定義他的動差生成函數 (moment generating function) ，這可以簡化我們的計算。

定義 2.4.21 ： 給定取值在 $\mathbb{N}_0$ 中的離散隨機變數 X，我們定義他的動差生成函數 (moment generating function) ，也稱作指數生成函數 (exponential generating function) 為

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

其中我們將 G_X 的收斂半徑記作 R，並要求 $e^t < R$。

註解 2.4.22 ： 若我們以 $M_X(t) := \mathbb{E}[e^{tX}]$ 來定義動差生成函數，則我們只需要要求 X 為實數離散隨機變數即可。

命題 2.4.23 ： 離散隨機變數 X 的動差可以透過 M_X 的微分來計算。更確切的說，對於所有非負整數 $k \geqslant 0$，我們有

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

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 $\mathbb{E}[|X|^k] \leqslant (Ck)^{k/2}$.

證明： 我們有

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

□

我們也能夠對 X 做拉普拉斯轉換 (Laplace transform)：

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

要注意的是，與指數生成函數相同，這並不是對於所有 $\lambda \in \mathbb{R}$ 都是有定義的。

第五小節 尾端機率

給定隨機變數 X，$\mathbb{P}(|X| > \lambda)$ 或 $\mathbb{P}(|X| > \lambda)$ 稱作尾端機率 (tail probability) 。一般來講，若我們能夠有效控制 X 越高階的動差，則代表尾端機率會越小，且反之亦然。

定義 2.4.24 ： 根據尾端機率，我們可以將隨機變數分佈歸類為：

(1) 次高斯分佈 (sub-Gaussian distribution) ：存在 $C, c > 0$ 使得

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

(2) 重尾分佈 (heavy-tailed distribution) ：存在 $C, \alpha > 0$ 使得

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

註解 2.4.25 ： 我們可以注意到，柯西分佈會是個重尾分佈，因為如果 $X \sim \text{Cauchy}(0,1)$，則

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

此外，高斯分佈也是個次高斯分佈。

問題 2.4.26： 給定隨機變數 X，則下列三敘述等價：

(1) X 是個次高斯分佈；
(2) 存在 $C, c > 0$ 使得對於所有 $t \in \mathbb{R}$，我們有 $\mathbb{E}[e^{tX}] \leqslant C\exp(ct^2)$；
(3) 存在 $C > 0$ 使得對於所有 $k \geqslant 1$，我們有 $\mathbb{E}[|X|^k] \leqslant (Ck)^{k/2}$。

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

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

問題 2.4.27：令 X 為實隨機變數且 $k > 0$。若 X 在 L^k 中，證明當 $\lambda \longrightarrow \infty$ 時，我們有

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