Review of Probability Fundamental

Jianing Yao
Department of MSIS-RUTCOR
Rutgers University, the State University of New Jersey
Piscataway, NJ 08854 USA

February 5, 2015

The foundation of formulating continuous-times, risk-neutral pricing theory (main topic of this class) is *measure-theoretic probability*, which at first appears to be overly abstract, and sometimes you may even doubt its relevance to the field of quantitative finance. However, it provides a general and powerful way to express both conceptual ideas and basic formulas of risk-neutral pricing. Indeed, it is the framework employed in the research literature of mathematical finance. Don't be afraid, you **DO NOT** need to learn advanced, measure-theoretic techniques or proof. All you need to do, which is not easy either, is to acquire an intuitive understanding of the measure-theoretic language and be able to translate it into practical computational formulae.

1 Probability Spaces, distribution function and expectation

The basis of measure-theoretic probability is the *probability space*, the triple $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the set called *outcome space*, $\mathcal{F}$ is a collection of subsets of Ω for which a probability can be calculated. These subsets are usually called *events*, they are possible things that can happen. The third element $\mathbb{P}$ is a function (*set function*) which gives the probability for an event in $\mathcal{F}$. The terminology for $\mathcal{F}$ is *sigma field, or σ-algebra*, which we define now:

Definition 1.1 The collection of subsets $\mathcal{F}$ is a σ-algebra, if

- $\Omega, \emptyset \in \mathcal{F}$;
- $A \in \mathcal{F}$ implies $A^C \in \mathcal{F}$;
- If A_1, $A_2 \cdots \in \mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

At first look, those assumptions are rather abstract, but actually it is quite reasonable to do so. It allows you to calculate the probability of the whole sample space (the first property), the complement of any event and the countable union of any sequence of events.

In fact, one can also calculate the probability of countable intersection of any sequence of events, i.e., $A_1, \cdots \in \mathcal{F}$ implies $\cap_{i=1}^{\infty} A_i \in \mathcal{F}$ (verify by yourself!).

In general, $\mathcal{F}$ can be anything as long as those axioms in *definition* 1.1 are satisfied. Nevertheless, we usually start with a collection of events $\mathcal{A}$ and use $\sigma(\mathcal{A})$ to represent the smallest σ-algebra containing the collection of events $\mathcal{A}$. For this reason, $\sigma(\mathcal{A})$ is referred as the σ-algebra *generated by* $\mathcal{A}$. The importance of this notion will be realized in the near future, for now, let's just give an example to illustrate the idea:

Example 1.1 Let $\Omega = \{a, b, c\}$ be the outcome space, let $\mathcal{A} = \{\{a, b\}, \{c\}\}$. Then $\mathcal{A}$ is not a σ-algebra because, at least, $\{a, b, c\} \notin \mathcal{A}$. The σ-algebra generated by $\mathcal{A}$ is: $\sigma(\mathcal{A}) = \{\{a, b, c\}, \{a, b\}, \{c\}, \emptyset\}$.

That's enough discussion for the σ-algebra, now *probability measure* $\mathbb{P}$:

Definition 1.2 A *probability measure* $\mathbb{P}$ is a *set function*: $\mathcal{F} \mapsto [0, 1]$, which satisfies:

- $\mathbb{P}(\Omega) = 1$;
- $\mathbb{P}(A) \geq 0$, $\forall A \in \mathcal{F}$;
- $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$, for any disjoint event A_i and A_j, where $i \neq j$.

Again, nothing counter intuition, the definition basically says that probability must be positive and between zero and one, in addition, the probability of a countable union of mutually exclusive events is the sum of the probabilities.

The major difference between elementary and measure-theoretic probability is the treatment of random variables. In the elementary probability, a random variable is just a representation of the numerical result of a trial with random outcome, and a random variable is modeled by specifying its cumulative distribution function, probability mass function, or density function, depending on the case. Probability space do not enter the definition. In the measure-theoretic approach, the underlying model is always a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and random variables are functions on Ω that are *measurable* with respect to $\mathcal{F}$. The idea is that X assigns a numerical attribute $X(\omega)$ to each outcome $\omega \in \Omega$, the value of X is random because ω is random. Let's formally explain what is the *measurability*: X is *measurable* with respect to $\mathcal{F}$ if and only if $\{\omega; X(\omega) \leq x\}$ belongs to $\mathcal{F}$ for all x, or equivalently, $X^{-1}([-\infty, x]) \in \mathcal{F}$ for every $x \in \mathbb{R}$[1]. The reason for imposing measurability is simple, we want at a minimum to assign probabilities to the event that X be less than or equal to x for any $x \in \mathbb{R}$, but probabilities are only defined for sets in $\mathcal{F}$. The measurability assumption allows us to define the cumulative distribution function of X:

$$F_X(x) = \mathbb{P}(\{\omega; X(\omega) \leq x\}), \quad -\infty < x < +\infty$$

Example 1.2 As in the *example* 1.1, $\Omega = \{a, b, c\}$, $\sigma(\mathcal{A}) = \{\{a, b, c\}, \{a, b\}, \{c\}, \emptyset\}$, and we define three random variables X, Y, Z as follows (in *Figure* 1). Which of the random variables are $\sigma(\mathcal{A})$ measurable? Well since $\{Y \leq 1\} = \{a\} \notin \mathcal{A}$, then Y is not $\sigma(\mathcal{A})$ measurable. For the same reason, Z is not as well. The variable X is $\sigma(\mathcal{A})$ measurable because $\{X \leq 1\} = \{a, b\} \in \sigma(\mathcal{A})$, $\{X \leq 2\} = \{a, b, c\} \in \mathcal{A}$.

[1] Actually, there are other alternatives to define a measurable functions, $X^{-1}(x, +\infty] \in \mathcal{F}$; $X^{-1}[x, +\infty] \in \mathcal{F}$; $X^{-1}[-\infty, x) \in \mathcal{F}$; $X^{-1}\{+\infty\}$, $X^{-1}\{-\infty\}$ and $X^{-1}(A) \in \mathcal{F}$ for every Borel sets on $\mathbb{R}$.

Table 1: Example

ω	X	Y	Z
a	1	1	1
b	1	2	7
a	2	2	4

Another important example of random variable (measurable function) that is both theoretically and practically important is the *indicator function*:

Example 1.3 Let $A \in \mathcal{F}$, define the *indicator function*:

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

This is surely a $\mathcal{F}$-measurable and hence a random variable.

Simple enough, but it is the building block of random variables and leads to the general definition of expectation. Here is the idea: we firstly define so called *simple random variable* which takes on only a finite number of values, i.e., $X = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}$, where $a_i \in \mathbb{R}$ and $A_i \in \mathcal{F}$, $1 \leq i \leq n$. If this is the case, we can define the expectation easily:

$$\mathbb{E}[X] = \sum_{i=1}^{n} a_i \mathbb{P}(A_i) := \int X(\omega)\mathbb{P}(d\omega) \tag{1}$$

Let X, Y be two simple random variables and β a real number. We clearly can write both X and Y in the form above, with the same subsets A_i which form a partition of Ω, i.e., $\cup_{i=1}^{n} A_i = \Omega$, and with a_i for X and b_i for Y. Then βX and $X + Y$ are again in the form (1) with same A_i and with the respective numbers βa_i and $a_i + b_i$. Thus, $\mathbb{E}[\beta X] = \beta \mathbb{E}[X]$ and $\mathbb{E}[X+Y] = \mathbb{E}[X]+\mathbb{E}[Y]$. I guess that's what you expect for the expectation, in mathematical language, it says the expectation is linear on the vector space of all simple random variables. Additionally, if $X \leq Y$, we will have $a_i \leq b_i$ for all i, as a result, $\mathbb{E}[X] \leq \mathbb{E}[Y]$. If you don't like this seems "convoluted" explanation, the bottom line is, the expectation of simple random variable coincide with your intuition of expectation, so we didn't annihilate your previous knowledge.

The next step is supported by a theorem which you don't need to prove, any **non-negative** random variable X (restricted to non-negativity for technical reason) is the limit of a monotonically increasing sequence of **non-negative** simple random variables ${X_n}$[2]. You can imagine that if the partition goes finer and finer it will become a general random variable, but non-negative. More importantly, the expectation for an non-negative random variable is also defined by passing to the limit:

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

[2] $X_n = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}$, where $a_i \in \mathbb{R}$ and $A_i \in \mathcal{F}$, $1 \leq i \leq n$. The subscript n corresponds to the number of partitions.

Luckily, this expectation again enjoys all properties that you think an expectation should have. Lastly, for an arbitrary random variable X, one can decompose it into[3] $X = X^+ - X^-$(can you verify $|X| = X^+ + X^-$?), both terms on the left hand side are non-negative random variables. Correspondingly, by the linearity of expectation on non-negative random variables:

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-] \tag{2}$$

We say a random variable has *a finite expectation* (is "integrable", expectation is an integral) if both $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ are finite. The expectation is the result of the difference in (2), also written as $\int_\Omega X(\omega)\mathbb{P}(d\omega)$. We write $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ to denote the set of all integrable random variables. That is, $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ implies the random variable X has a finite mean.

Above discussion is not the way that we are going to calculate the expectation, in fact, the definition is mostly of theoretic use, to actually compute $\mathbb{E}[X]$ when X has a density f_X, i.e. $F_X(x)$ is differentiable, we use the usual formula:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x) dx$$

In the measure-theoretic framework, this is proved with some efforts, while in elementary probability it is a definition. Another consequence is that

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx$$

2 Partial information, filtration and conditional expectation

In the last section, we are struggling with σ-algebra, one may ask why it's important. For example, why not always assign a probability to every subset of Ω? The first reason is so called *axiom of choice*, it says the collection of all subsets of a set is too rich, and it may not be possible to assign a probability to every subset and satisfy the countable additivity at the same time. Therefore, it is necessary to impose restrictions the class of subsets of Ω to which probabilities are assigned. However, we want this class of subsets to at least have some nice properties, otherwise it would be difficult to do even elementary calculations. Fortunately, it turns out there is a general procedure for constructing countably additive probability measure on σ-algebras, and so a nice theory is available. A second reason for the importance of σ-algebras, is that, in probabilistic analysis, subsets of events often come in σ-algebra packages. This is especially true in modeling *partial information* and defining *conditional expectation*, as shall be described shortly.

The following basic facts are important to know:

- The family of all subsets (including the empty set) of a set is a σ-algebra. When the outcome space Ω of a probability space model is finite or countable, we usually take $\mathcal{F}$ to be this family;

[3] $x^+ = \max(0, x)$ and $x^- = \max(-x, 0)$.

- If $\mathcal{C}$ is any family of subsets of some set Ω, there exists a σ-algebra, denoted by $\sigma(\mathcal{C})$, satisfying (i) $\mathcal{C} \subset \sigma(\mathcal{C})$; (ii) if $\mathcal{G}$ is any σ-algebra containing $\mathcal{C}$, then $\sigma(\mathcal{C}) \subset \mathcal{G}$. We call $\sigma(\mathcal{C})$ the σ-algebra generated by $\mathcal{C}$. It is the smallest σ-algebra containing $\mathcal{C}$.

The *example* 1.1 illustrates the second point above, but here are some general examples:

(i) Let $\mathcal{C} = \{A_1, A_2, \cdots\}$ be a disjoint partition of a set Ω. Then $\sigma(\mathcal{C})$ is the collection of all finite or countable unions of subsets in $\mathcal{C}$ plus the empty set. The simplest example occurs when $\mathcal{C} = \{A, A^C\}$ is a partition of Ω, in this case, $\sigma(\mathcal{C}) = \{\emptyset, A, A^C, \Omega\}$.

(ii)Let $\mathcal{U}$ be the family of all open sub-intervals of the real line $\mathbb{R}$. The σ-algebra $\sigma(\mathcal{U})$ is called the Borel σ-algebra of $\mathbb{R}$. Loosely speaking, most of the thing that you can think of, e.g., open intervals, closed intervals, point, some combinations of them are included in this σ-algebra, but not all.

(iii) Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, define

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

where $\mathcal{B}$ is a Borel subset of $\mathbb{R}$. This family is in fact a σ-algebra (because X is measurable), and is called the σ-algebra generated by X; it is the family of all events defined only in terms of the value of X. Since X is a random variable, $\sigma(X)$ will be contained in $\mathcal{F}$, but in general it will not be all of $\mathcal{F}$.

(iv) Let $\{X_\alpha; \alpha \in \mathcal{I}\}$ be any family of random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$; here $\mathcal{I}$ stands for an arbitrary index set. Let $\mathcal{C}$ be the family of all events U such that $U \in \sigma(X_\alpha)$ for some $\alpha \in \mathcal{I}$. Then $\sigma(\mathcal{C})$ is called the σ-algebra generated by $\{X_\alpha; \alpha \in \mathcal{I}\}$ and is often denoted $\sigma(\{X_\alpha; \alpha \in \mathcal{I}\})$.

To illustrate (i), (iii) and (iv), I will use the multi-period binomial model. It is a toy model of this course, which you will learn soon. **At the first reading, you can skip this part, but do return to it when this model is taught, because this makes above abstract examples more concrete. Maybe you can read it now as well, not that hard. Actually, this is also a model of consecutive coin tosses, with** 1 **representing the head and** -1 **representing the tail.**

- Time periods: $t_0 = 0, t_1 = h, t_2 = 2h, ..., t_n = nh = T$;
- Outcome space: $\Omega = \{\omega = (\omega_1, ..., \omega_n); \omega_i \in \{-1, 1\} \text{ for each } i\}$;
- $\mathcal{F}$ is the collection of all subsets of Ω[4];
- For $1 \leq i \leq n$, let random variable $\xi(t_i)(\omega) = \omega_i$, it is the market movement in period i. $\xi(t_i)(\omega) = \omega_i = 1$ means an 'up' movement, $\xi(t_i)(\omega) = \omega_i = -1$ means a 'down' movement.

(a) $\sigma(\xi(t_i))$*: the sigma algebra generated by market movement in period* i

It is sufficient to write $\sigma(\xi(t_1))$. Define

$$A_1 = \{(\omega_1, ..., \omega_n); \xi(t_1)(\omega) = \omega_1 = 1\},\ A_{-1} = \{(\omega_1, ..., \omega_n); \xi(t_1)(\omega) = \omega_1 = -1\}$$

[4]Since the space is finite, σ-algebra is usually the collection of all subsets of Ω, as pointed out previously

That is to say, A_1 is the set of all sequences in Ω that start with 1. since 1 and -1 are the only two possible values of ξ_1, these two events partition Ω and

$$\sigma(\xi_1) = \{A_1, A_{-1}, \Omega, \emptyset\}$$

Each $\sigma(\xi(t_i))$ can be represented in a similar way.

Remark 2.1 One may wonder this is not the same procedure as described in (iv). Actually, for discrete random variable X that takes values only in the set of distinct numbers $\{c_1, ..., c_K\}$, we can let $B_i = \{\omega; X(\omega) = c_i\}$, $1 \leq i \leq K$. Then $\sigma(X)$ produced by (iv) is the same as that induced by partition $\{B_1, ..., B_K\}$ as being illustrated in (i).

(b) $\sigma(\xi(t_1), \xi(t_2))$*: the sigma algebra generated by consecutive two market movements*

This will be the smallest σ-algebra containing both $\sigma(\xi(t_1))$ and $\sigma(\xi(t_2))$. The former one we have already, for $\sigma(\xi(t_2))$, we define

$$A_{(\cdot,1)} = \{(\omega_1, ..., \omega_n : \xi(t_2)(\omega) = \omega_2 = 1)\},\ A_{(\cdot,-1)} = \{(\omega_1, ..., \omega_n : \xi(t_2)(\omega) = \omega_2 = -1)\}$$

Similarly,

$$\sigma(\xi(t_2)) = \{A_{(\cdot,1)}, A_{(\cdot,-1)}, \emptyset, \Omega\}$$

Based on (i), we can generate σ-algebra from $\sigma(\xi(t_1))$ and $\sigma(\xi(t_2))$. The result is the same as the following construction: we instead define $A_{(1,1)} = \{(\omega_1, ..., \omega_n) : (\omega_1, \omega_2) = (1, 1)\}$, $A_{(1,-1)} = \{(\omega_1, ..., \omega_n) : (\omega_1, \omega_2) = (1, -1)\}$, $A_{(-1,1)} = \{(\omega_1, ..., \omega_n) : (\omega_1, \omega_2) = (-1, 1)\}$ and $A_{(-1,-1)} = \{(\omega_1, ..., \omega_n) : (\omega_1, \omega_2) = (-1, -1)\}$, this is clearly a partition of Ω, then we generate sigma algebra from this collection of subsets. We usually adopt the second approach (I will explain later).

(c) $\sigma(\xi(t_1), ..., \xi(t_n))$*: the sigma algebra generated by a sequence of random variables*

This is the smallest σ-algebra of events containing each of $\sigma(\xi(t_1)), ..., \sigma(\xi(t_k))$. As we did in (b), we partition Ω into:

$$A_{(\eta_1,...,\eta_k)} = \{(\omega_1, ..., \omega_n) : (\omega_1, ..., \omega_k) = (\eta_1, ..., \eta_k)\}$$

where $(\eta_1, ..., \eta_k)$ is any sequence of k 1's and -1's.

2.1 Partial Information and Filtration

This leads to the notion of *partial information*. Using the example above, the complete outcome of a multi-period market model is the evolution of market prices from 0 to T. At any intermediate time t, $0 < t < T$, a market participant has observed the market evolution only up to t. This is partial information about the entire market path over $[0, T]$ and it will change one's assessment of how market might behave for times later than t. Since its importance, we need a uniform and theoretically convenient way to model partial information, we already knew it, the σ-algebra(another reason of caring about σ-algebra). But this time, we only take a *sub-σ-algebra* $\mathcal{G} \subset \mathcal{F}$. We again use the multi-period binomial model, $\mathcal{F}$ is the collection of all subsets of Ω, certainly a σ-algebra. In the meanwhile,

$\mathcal{F}_1 := \sigma(\xi(t_1)) = \{A_{(1)}, A_{(-1)}, \emptyset, \Omega\}$ is a σ-algebra but contained in $\mathcal{F}$, a sub-σ-algebra. By the same reasoning, $\mathcal{F}_1 \subset \mathcal{F}_2 := \sigma(\xi(t_1), \xi(t_2)) \subset \mathcal{F}_k := \sigma(\xi(t_1), ..., \xi(t_k))$, for $k \geq 3$. Now, we can explain why we use the second approach in (b) and also in (c), the sub-σ-algebra $\mathcal{F}_k$ represents the partial information market participants have at the end of k periods from observing the first k market movements $\omega_1, ..., \omega_k$. At period k, we can tell whether any event in the sub-σ-algebra $\mathcal{F}_k$ happened or not.

The following is important and intuitive. Let X be a random variable, observing $\sigma(X)$ is equivalent to knowing the value of $X(\omega)$. Likewise, observing $\sigma(X(s), s \leq t)$ is equivalent to knowing the value of $X(s)(\omega)$ for all $s \leq t$.

A *filtration* is a sequence $\mathcal{F}_1$, $\mathcal{F}_2$,... that is increasing in the sense that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \cdots$. As time goes on, one takes in new observations but never forgets old information, the partial information $\mathcal{F}_n$ accumulated by time t_n increases (Recall, the important filtration associated to the multi-period, binomial market model is defined by $\mathcal{F}_k = \sigma(\xi(t_1), ..., \xi(t_k)), k = 1, ..., n$, it represents the information contained in observation of the first k market movements). A random process $\{X(t_k)\}$ is said to be *adapted to* $\{\mathcal{F}_k\}_{k\geq 0}$ if $X(t_k)$ is $\mathcal{F}_k$-measurable. Intuitively, it say $X(t_k)$ can only depend on its history but not the future.

2.2 Conditional Expectation and conditional probability

2.2.1 Elementary definitions

Let X and Y be two discrete random variables. In elementary probability, the conditional expectation of Y given $X = x$ is defined to be:

$$\mathbb{E}[Y|X = x] := \sum_y y\mathbb{P}(Y = y|X = x) = \sum_y y\frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)} \tag{3}$$

Here the sum is over y in the range of Y, and is defined for x such that $\mathbb{P}(X = x) > 0$. If X and Y are two continuous random variables with joint density $f(x, y)$, the definition is:

$$\mathbb{E}[Y|X = x] = \int y\frac{f(x, y)}{f_X(x)}dy \tag{4}$$

where f_X is the density of X. This formula defines $\mathbb{E}[Y|X = x]$ on the set where $f_X(x) > 0$. When $f_X(x) = 0$, we can define it anyway we like, but it is most convenient just to set the conditional expectation to 0 (we don't have to worry about these values because the probability X takes on a value in the set $\{x : f_X(x) = 0\}$ is 0). In each case, the conditional expectation is a function of the range of values x, that X can take on.

These definitions and their extension are not adequate for our purposes, because partial information can come in much more complicated forms than observation of a single, or even several random variables. For example, in asset pricing, we would like to condition on observing the whole past history of the market. In general, we want to define the conditional expectation of X given $\mathcal{G}$, where $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$, since σ-algebras are the most general way we represent partial information. The general definition will be developed first in the simpler case of a σ-algebra induced by a partition $A_1, ..., A_K$, and then extended to the general case.

2.2.2 Conditioning on a σ-algebra generated by a partition

First, it is necessary to define $\mathbb{E}[Y|A]$, the expected value of Y given A has occurred, where A is an event for which $\mathbb{P}(A) > 0$.

Definition 2.1 If $\mathbb{P}(A) > 0$ and if $\mathbb{E}[Y\mathbf{1}_A]$ is well-defined,

$$\mathbb{E}[Y|A] = \frac{\mathbb{E}[Y\mathbf{1}_A]}{\mathbb{P}(A)}$$

Example 2.2 Let $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. Suppose

$$\mathbb{P}(\{\omega_1\}) = \frac{3}{8},\ \mathbb{P}(\{\omega_2\}) = \frac{1}{8},\ \mathbb{P}(\{\omega_3\}) = \frac{1}{4},\ \mathbb{P}(\{\omega_4\}) = \frac{1}{4}$$

and $Y(\omega_i) = i$. Let $A = \{\omega_1, \omega_2\}$, then $\mathbb{P}(A) = \frac{1}{2}$, and

$$\mathbb{E}[Y\mathbf{1}_A] = \sum_{i=1}^{4} Y(\omega_i)\mathbf{1}_A(\omega_i)\mathbb{P}(\{\omega_i\}) = Y(\omega_1)\mathbb{P}(\{\omega_1\}) + Y(\omega_2)\mathbb{P}(\{\omega_2\}) = \frac{3}{8} + \frac{2}{8} = \frac{5}{8}$$

Thus,

$$\mathbb{E}[Y|A] = \frac{5}{4}$$

Example 2.3 Let X be a discrete random variable and let $A = \{X = z\}$ where $\mathbb{P}(X = z) > 0$. Suppose Y is also discrete, then, since $\mathbf{1}_A(\omega) = \mathbf{1}_{\{z\}}(X(\omega))$. By *definition* 2.1 says

$$\begin{aligned}\mathbb{E}[Y|\{X = z\}] =& \frac{\mathbb{E}[Y\mathbf{1}_{\{z\}}(X)]}{\mathbb{P}(X = z)} = \frac{1}{\mathbb{P}(X = z)} \sum_{x,y} y\mathbf{1}_{\{z\}}(x)\mathbb{P}(X = x, Y = y) \\ &= \frac{1}{\mathbb{P}(X = z)} \sum_{y} y\mathbb{P}(X = z, Y = y)\end{aligned}$$

This is exactly the formula for $\mathbb{E}[Z|X = z]$ given in (3) using the definition of conditional expectation in elementary probability. Thus, *definition* 2.1 is consistent with (3).

As the next step, let $\{A_1, ..., A_K\}$ be a partition of Ω into disjoint events. Let $\mathcal{G}$ be the σ-algebra induced by this partition.

Definition 2.2 Let Y be a random variable for which $\mathbb{E}[Y]$ is defined and finite ($Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$). Then the conditional expectation of Y given $\mathcal{G}$ is

$$\mathbb{E}[Y|\mathcal{G}](\omega) := \sum_{i=1}^{K} \mathbb{E}[Y|A_i]\mathbf{1}_{A_i}(\omega) \tag{5}$$

Unlike the conditioning on a specific event, this version of conditional expectation is a random variable. For each ω, its value is the conditional expectation of Y given the event A_i that actually occurs. In the simplest case, when $\mathcal{G} = \{A, A^C, \emptyset, \Omega\}$,

$$\mathbb{E}[Y|\mathcal{G}](\omega) = \mathbb{E}[Y|A]\mathbf{1}_A(\omega) + \mathbb{E}[Y|A^C]\mathbf{1}_{A^C}(\omega)$$

It returns the value $\mathbb{E}[Y|A]$ if A occurs, and the value $\mathbb{E}[Y|A^C]$ if it does not.

Remark 2.4 It is standard practice to omit the dependence on ω and write simply $\mathbb{E}[Y|\mathcal{G}]$ when using conditional expectations.

Example 2.5 Consider the set-up of *example* 2.2 and let $\mathcal{G} = \{A, A^C, \emptyset, \Omega\}$. Here $A^C = \{\omega_3, \omega_4\}$. By a similar calculation, $\mathbb{E}[Y|A^C] = \frac{7}{2}$. Thus,

$$\mathbb{E}[Y|\mathcal{G}] = \frac{5}{4}\mathbf{1}_A(\omega) + \frac{7}{2}\mathbf{1}_{A^C}(\omega) = \frac{5}{4}\mathbf{1}_{\{\omega_1,\omega_2\}}(\omega) + \frac{7}{2}\mathbf{1}_{\{\omega_3,\omega_4\}}(\omega)$$

2.2.3 Conditioning on a discrete random variable

Let X be a discrete random variable taking on values in the finite set $\{c_1, c_2, ..., c_K\}$. Then $\sigma(X)$ is the σ-algebra induced by the partition $\{X = c_1\}$, ..., $\{X = c_K\}$.

Definition 2.3 In this case, $\mathbb{E}[Y|X] := \mathbb{E}[Y|\sigma(X)]$.

This returns to the previous case:

$$\mathbb{E}[Y|X](\omega) = \sum_{i=1}^{K} \mathbb{E}[Y|\{X = c_i\}]\mathbf{1}_{X=c_i}(\omega) = \sum_{i=1}^{K} \mathbb{E}[Y|X = c_i]\mathbf{1}_{\{X=c_i\}}(\omega)$$

When $X(\omega) = c_i$, $\mathbb{E}[Y|X](\omega) = \mathbb{E}[Y|X = c_i]$. This means we can use the formula (3) to compute conditional expectations when conditioning on a random variable.

Example 2.6 Again consider the set-up of *example* 2.2 and define a second random variable by $X(\omega_1) = X(\omega_2) = 0$ and $X(\omega_3) = X(\omega_4) = 1$. Then $A = \{X = 0\}$ and $A^C = \{X = 1\}$ and so $\sigma(X) = \mathcal{G} = \{A, A^C, \emptyset, \Omega\}$. Thus $\mathbb{E}[Y|X]$ is exactly $\mathbb{E}[Y|\mathcal{G}]$ as computed in *example* 2.5. However, it is usual to express it as follows,

$$\mathbb{E}[Y|X] = \frac{5}{4}\mathbf{1}_{\{0\}}(X) + \frac{7}{2}\mathbf{1}_{\{1\}}(X)$$

This displays the answer clearly as a random variable, since X is a random variable.

Note in *example* 2.6 $\mathbb{E}[Y|X]$ is a function of X. This is a general fact, always true, as will be explained in the next paragraph.

Let's give another, rather important characterization of $\mathbb{E}[Y|\mathcal{G}]$:

Theorem 2.7 Let $\mathcal{G}$ be the σ-algebra induced by a finite disjoint partition of Ω. Suppose $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Then $Z = \mathbb{E}[Y|\mathcal{G}]$ is the unique random variable satisfying:

- Z is $\mathcal{G}$-measurable;
- For every $A \in \mathcal{G}$, $\mathbb{E}[\mathbf{1}_A Z] = \mathbb{E}[\mathbf{1}_A Y]$.

Where do these conditions come from? Intuitively, the value of $\mathbb{E}[X|\mathcal{G}](\omega)$ should be determined from observing $\mathcal{G}$. Observation of $\mathcal{G}$ determines the value of any $\mathcal{G}$-measurable random variables. This is why the first condition is required. The second condition is harder to explain, let's skip for now. Also, we will skip the proof of theorem, but it is not hard, readers who have interests can try to prove it.

Remark 2.8 In this theorem, 'unique' means 'unique up to sets of probability zero'. In other word, if $\bar{Z}$ is another random variable satisfying above conditions, then $\mathbb{P}(Z = \mathbb{E}[Y|\mathcal{G}]) = 1$.

An immediate result follows from above theorem:

$$\mathbb{E}[\mathbf{1}_A \mathbb{E}[Y|\mathcal{G}]] = \mathbb{E}[\mathbf{1}_A Y] \text{ for every } A \in \mathcal{G},$$

Memorize this fact!

On the other hand, it follows in general that if X is a discrete random variable, $\mathbb{E}[Y|X] = \mathbb{E}[X|\sigma(X)]$ is $\sigma(X)$-measurable. There exists a theorem[5] saying any $\sigma(X)$-measurable function can be expressed as a function of X. Thus, there is always a function h such that $\mathbb{E}[Y|X] = h(X)$. We already saw an instance of this in *example* 2.6.

2.2.4 Conditioning on a general σ-algebra

The *definition* 2.2 will not work for general σ-algebras. To see the problem, suppose that X is a continuous random variable. For every x, $\{X = x\}$ is an event of probability 0. How then can one make sense of conditioning on such an event. The key idea is to use the characterization of conditional expectation found in *theorem* 2.7, which makes no explicit reference to an object such as $\mathbb{E}[Y|A]$ requiring to divide by $\mathbb{P}(A)$.

Theorem 2.9 Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G}$ be the sub-σ-algebra of $\mathcal{F}$. Suppose $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Then Z be a random variable satisfying:

- Z is $\mathcal{G}$-measurable;
- For every $A \in \mathcal{G}$, $\mathbb{E}[\mathbf{1}_A Z] = \mathbb{E}[\mathbf{1}_A Y]$.

Then, we say Z is conditional expectation of Y given $\mathcal{G}$ and denote it by $\mathbb{E}[Y|\mathcal{G}]$.

The following result from measure and integration addresses whether such a random variable Z exists:

Theorem 2.10 If $\mathbb{E}[|Y|] < \infty$, then $\mathbb{E}[Y|\mathcal{G}]$ satisfying the conditions in above theorem exists. It is unique in the sense that if Z and Z' both satisfy the conditions, then $\mathbb{P}(Z = Z') = 1$.

Because of this theorem, we can always write down $\mathbb{E}[Y|\mathcal{G}]$ in good conscience, once we have checked $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Unfortunately, this theorem does not tell us how to calculate the conditional expectation. We have to come up with a formula that we think is correct and then check it satisfies the conditions above.

For this sake, we revert to the formulas of elementary probability theory. The following case is extremely important:

$\mathcal{G} = \sigma(X)$. In this case, we use $\mathbb{E}[Y|X]$ to denote $\mathbb{E}[Y|\sigma(X)]$. Since this is $\sigma(X)$-measurable, it follows from the same theorem that we used before: there exists a function ψ such that $\mathbb{E}[Y|\sigma(X)] = \psi(X)$. When Y and X have a joint density $f(x, y)$, $\psi(x)$ is precisely

[5]The theorem says: suppose Y is $\sigma(X_1, ..., X_M)$-measurable. Then there is a function h such that $Y(\omega) = h(X_1(\omega), ..., X_M(\omega))$. h can be chosen to be a Borel function.

the function of y defined by the right hand side of (4). The following formula states this explicitly,

$$\mathbb{E}[Y|X](\omega) = \int_{-\infty}^{\infty} y \frac{f(X(\omega), y)}{f_X(X(\omega))} dy$$

This formula shows that the measure-theoretic definition of conditional expectation does indeed generalize the elementary definition.

In addition, this conditional expectation satisfies linearity and the following two important properties that we will use over and over:

1. For any constant $c_1, c_2 \in \mathbb{R}$, $\mathbb{E}[c_1X + c_2Y|\mathcal{G}] = c_1\mathbb{E}[X|\mathcal{G}] + c_2\mathbb{E}[Y|\mathcal{G}]$;
2. If $\mathbb{E}[|XY|] < +\infty$ and X is $\mathcal{G}$-measurable, then $\mathbb{E}[XY|\mathcal{G}] = X\mathbb{E}[X|\mathcal{G}]$;
3. If Z is $\mathcal{G}$-measurable, $\mathbb{E}[ZX|\mathcal{G}] = Z\mathbb{E}[X|\mathcal{G}]$;
4. Tower property: if $\mathcal{G} \subset \mathcal{H}$, then $\mathbb{E}[\mathbb{E}[X|\mathcal{H}]|\mathcal{G}] = \mathbb{E}[X|\mathcal{G}]$;
5. If X is independent of $\mathcal{G}$ (every $A \in \mathcal{G}$ is independent of every B in $\sigma(X)$), then $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$.

Lastly, let's introduce the conditional probability. Recall the indicator function:

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

for $A \in \mathcal{F}$. The conditional probability is defined as:

$$\mathbb{P}(A|\mathcal{G}) := \mathbb{E}[\mathbf{1}_A|\mathcal{G}]$$

As a result, it is a random variable. All formulas and properties of conditional expectation pass along to conditional expectation. All in all, it is consistent with the one you learn from elementary probability.

3 Modes of Convergence

In elementary mathematic courses, one speaks of the convergence of functions: $f_n : \mathbb{R} \mapsto \mathbb{R}$, then $\lim_{n\to\infty} f_n = f$ if $\lim_{n\to\infty} f_n(x) = f(x)$ for all x in $\mathbb{R}$. This is so called *point-wise convergence of functions*. A random variable is also a function (measurable function), thus the notion of point-wise convergence is inherited. However, this definition is not that useful as one may expect, it is simply too strong. In this sections, we will talk about several modes of convergence that is in common use.

Let's fix the probability space $(\Omega, \mathcal{F}, \mathbb{P})$,

Definition 3.1 (*almost surely convergence*) We say that a sequence of random variables $(X_n)_{n\geq 0}$ converges almost surely to a random variable X if

$$N = \{\omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\} \text{ has } \mathbb{P}(N) = 0.$$

The set N is called a *null set*. We usually abbreviate almost sure convergence by writing:

$$\lim_{n\to\infty} X_n = X \text{ a.s.}$$

Note that we can alternatively use the following criterion:

$$N^C = \{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\} \text{ and then } \mathbb{P}(N^C) = 1$$

Example 3.1 Let X_n be an i.i.d. sequence of random variables with $\mathbb{P}(X_n = 1) = p$ and $\mathbb{P}(X_n = 0) = 1 - p$. For example we can imagine tossing a slightly unbalanced coin ($p > \frac{1}{2}$) repeatedly, and $\{X_n = 1\}$ corresponds to heads on th n-th toss and $\{X_n = 0\}$ the tails. In the "long run", we would expect the proportion of heads to be p, this would justify our model that claims the probability of head is p. Mathematically, we would want

$$\lim_{n\to\infty} \frac{X_1(\omega) + \cdots + X_n(\omega)}{n} = p \text{ for all } \omega \in \Omega.$$

This simply does not happen! For example let $\omega_o = \{T, T, ...\}$, the sequence of fails. For this ω_0,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X_j(\omega_0) = 0$$

More generally, we have the event

$$A = \{\omega : \text{ only a finite number of heads occur}\}$$

Then,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X_j(\omega) = 0 \text{ for all } \omega \in A.$$

We readily admit that the event A is very unlikely to occur. Indeed, we can show that $\mathbb{P}(A) = 0$. Eventually, we can eventually show that:

$$\mathbb{P}(\{\omega : \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X_j(\omega) = p\}) = 1$$

This is *almost surely convergence*, where the limiting random variable is a constant p.

However, this version of convergence can also fail just like point-wise convergence can fail sometimes, we will introduce two more types of convergence.

Definition 3.2 (*Convergence in mean/mean square*) A sequence of random variables $(X_n)_{n\geq 1}$ converges in mean/mean square to X (where $1 < p < \infty$) if $|X_n|$, $|X|$ has finite mean/variance and

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|] = 0 \ , \ \lim_{n\to\infty} \mathbb{E}[|X_n - X|^2] = 0, \text{ respectively.}$$

Remark 3.2 In general, we have the notion of convergence in p-th mean, but the most important case, at least for the course purpose, is when $p = 1$ and $p = 2$, called *convergence in mean* and *convergence in mean square*, respectively. Sometimes, we may denote as convergence in $\mathcal{L}^1$ and convergence in $\mathcal{L}^2$, or simply write

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

In particular, we are interested in $\mathcal{L}^1$ and $\mathcal{L}^2$.

Let's give an example, actually a theorem, where $\mathcal{L}^2$ convergence happens. This is the famous *Strong Law of Large Number*(SLLN),

Theorem 3.3 Let $(X_n)_{n\geq 1}$ be i.i.d and defined on the same space. Let

$$\mathbb{E}[X_j] = \mu \text{ and } \sigma^2 = \sigma_{X_j}^2 < +\infty$$

Let $S_n = \sum_{j=1}^n X_j$. Then

$$\lim_{n\to\infty} \frac{S_n}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n X_j = \mu \text{ a.s. and in } \mathcal{L}^2.$$

Observe, we also have almost surely convergence. Actually, *example* 3.1 is a special case of SLLN, thus has almost surely convergence. The proof for almost surely convergence is a little bit involved, but in $\mathcal{L}^2$ is easy. Let's prove this part.

Proof. Let us note that without loss of generality we can assume $\mu = \mathbb{E}[X_j] = 0$. Indeed, if $\mu \neq 0$, then we can replace X_j with $Z_j = X_j - \mu$. We obtain $\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n Z_j = 0$ and therefore

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n (X_j - \mu) = \lim_{n\to\infty} (\frac{1}{n} \sum_{j=1}^n X_j) - \mu = 0$$

from which we deduce the result. Thus, we can assume $\mu = 0$. Set $Y_n = \frac{S_n}{n}$. Then $\mathbb{E}[Y_n] = \frac{1}{n} \sum_{j=1}^n \mathbb{E}[X_j] = 0$. Moreover, $\mathbb{E}[Y_n^2] = \frac{1}{n^2} \sum_{1\leq j,k\leq n} \mathbb{E}[X_j X_k]$. However, if $j \neq k$,

$$\mathbb{E}[X_j X_k] = \mathbb{E}[X_j]\mathbb{E}[X_k] = 0$$

since X_j and X_k are assumed to be independent. Therefore,

$$\begin{aligned} \mathbb{E}[Y_n^2] &= \frac{1}{n^2} \sum_{j=1}^n \mathbb{E}[X_j^2] \\ &= \frac{1}{n^2} \sum_{j=1}^n \sigma^2 = \frac{1}{n^2} (n\sigma)^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Hence, $\lim_{n\to\infty} \mathbb{E}[Y_n^2] = 0$. □

Definition 3.3 (*convergence in probability*) A sequence of random variables $(X_n)_{n\geq 1}$ converges in probability to X if for any $\epsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}) = 0$$

denoted as:

$$X_n \xrightarrow{P} X$$

Before stating the relationship between those convergences, we first introduce a significant theorem in measure and integration –*Dominant Convergence Theorem*(DCT):

Theorem 3.4 (*Lebesgue Dominant Convergence Theorem*) If the random variables X_n converge almost surely to X and if $|X_n| < Y$ a.s. and $Y \in \mathcal{L}^1$, then $X_n, X \in \mathcal{L}^1$ and

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

We also say $\{X_n\}_{n\geq 1}$ convergence in mean to X.

Although we will not prove it, this theorem will be used quite often throughout this course. For example, it aids to prove the following theorem (proof is not required)

Theorem 3.5 Let X_n be a sequence of random variables:

- If $X_n \xrightarrow{\mathcal{L}^p} X$, then $X_n \xrightarrow{P} X$;
- If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{P} X$;

Overall, convergence in probability is the weakest of the three types of convergence.

4 Moment Generating Function

4.1 Basics about M.G.F

As you may know from elementary probability, expectation and variance contains important information about the distribution function of the random variable. Actually, there are other characterizations of random variable. Moment generating function is one of them, and it serves as a powerful tool for computation and proof. Let's firstly make the definition:

Definition 4.1 If X is a random variable, then its *moment generating function* is

$$\phi_X(t) = \mathbb{E}[e^{tX}] \begin{cases} \sum_x e^{tx}\mathbb{P}(X = x) & \text{in discrete case,} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x)dx & \text{in continuous case.} \end{cases}$$

here $f_X(\cdot)$ is the probability density function of X.

Example 4.1 Assume that X is *exponential*(1) random variable, that is,

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0, \\ 0 & x \leq 0. \end{cases}$$

then

$$\begin{aligned} \phi_X(t) &= \lambda \int_0^\infty e^{tx} e^{-\lambda x} dx \\ &= \frac{\lambda}{1-\lambda} e^{(t-\lambda)x}|_0^{+\infty}, \text{ (for the case } t < \lambda) \\ &= \frac{\lambda}{\lambda - t} \end{aligned}$$

If $t \geq \lambda$, the moment generating function does not exists.

Remark 4.2 Above example reveals an important fact that is moment generating function is only meaningful when the integral/summation converge.

From *Taylor's expansion* we know that

$$e^{tX} = 1 + tX + \frac{1}{2}t^2X^2 + \frac{1}{3!} + \cdots$$

Let's take expectation on both sides,

$$\mathbb{E}[e^{tX}] = 1 + t\mathbb{E}[X] + \frac{1}{2}t^2\mathbb{E}[X^2] + \frac{1}{3!}t^3\mathbb{E}[X^3] + \cdots$$

Note: this is not always true but under some mild conditions, because the right hand side is a infinite sequence, interchange of summation and integration is true when we need some convergence of the summation (*Fubini's Theorem* which we may learn later on). Let's forget about this for the time being. We usually call the expectation of k-th power of X, $m_k = \mathbb{E}[X^k]$, the *kth moment of* X. Note also that

$$\begin{aligned} &\frac{d}{dt}\mathbb{E}[e^{tX}]|_{t=0} = \mathbb{E}[X]; \\ &\frac{d^2}{dt^2}\mathbb{E}[e^{tX}]|_{t=0} = \mathbb{E}[X^2]. \end{aligned}$$

which lets you compute the expectation and variance of a random variable, because $Var[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. In *example* 4.1,

$$\frac{d}{dt}\phi_X(t)|_{t=0} = (\frac{\lambda}{\lambda - t})'|_{t=0} = \frac{\lambda}{(\lambda - t)^2}|_{t=0} = \frac{1}{\lambda}$$

and

$$\frac{d^2}{dt^2}\phi_X(t)|_{t=0} = (\frac{\lambda}{(\lambda - t)^2})'|_{t=0} = \frac{2}{\lambda^2}$$

Then

$$Var[X] = \frac{2}{\lambda^2} - (\frac{1}{\lambda})^2 = \frac{1}{\lambda^2}$$

That's exactly the mean for *exponential(λ)* random variable.

Let's give two more examples of moment generating function of particular distribution random variables:

Example 4.3 Compute the moment generating function for a *Poisson*(λ) random variable ($\mathbb{P}(n = k) = \frac{\lambda^n}{n!}e^{-\lambda}$, for $k = 0, 1, 2, ..., \lambda > 0$).

$$\begin{aligned}
\phi_X(t) &= \sum_{n=0}^{\infty} e^{tn}\frac{\lambda^n}{n!}e^{-\lambda} \\
&= e^{-\lambda}\sum_{n=0}^{\infty}\frac{(e^t\lambda)^n}{n!} \\
&= e^{-\lambda+\lambda e^t} \\
&= e^{\lambda(e^t-1)}
\end{aligned}$$

Example 4.4 Compute the moment generating function for a *standard normal* random variable, $X \sim N(0, 1)$.

$$\begin{aligned}
\phi_X(t) &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}e^{tx}dx \\
&= \frac{1}{\sqrt{2\pi}}e^{\frac{1}{2}t^2}\int_{-\infty}^{\infty}e^{-\frac{1}{2}(x-t)^2}dx \\
&= e^{\frac{1}{2}t^2}
\end{aligned}$$

where from the first to the second line we have used, in the exponent, $tx - \frac{1}{2}x^2 = -\frac{1}{2}(x - t)^2 + \frac{1}{2}t^2$. As an exercise, you may compute the moment generating function of a normal random variable, i.e., $X \sim N(\mu, \sigma^2)$, the answer will be:

$$\phi_X(t) = e^{\mu t+\frac{1}{2}\sigma^2 t^2}$$

Lastly, the moment generating function indeed characterize distribution and convergence of distributions. This is given by the following theorem, which we will not prove.

Theorem 4.5 Assume that the moment generating functions for random variables X, Y, and X_n are finite for all t when moment generating function is applied, then

- If $\phi_X(t) = \phi_Y(t)$ for all t, then $\mathbb{P}(X \leq x) = P(Y \leq x)$ for all x;
- If $\phi_{X_n}(t) \to \phi_X(t)$ for all t, and $\mathbb{P}(X \leq x)$ is continuous in x, then $\mathbb{P}(X_n \leq x) \to P(X \leq x)$ for all x.

4.2 M.G.F for summation of sequence of random variables

Suppose that we have a sequence of random variables $X_1, X_2, ..., X_n$ and define $S_n = \sum_{i=1}^n X_i$, then

Lemma 4.6 If $X_1, X_2, ..., X_n$ are independent, then

$$\phi_{S_n}(t) = \phi_{X_1}(t) \cdots \phi_{X_n}(t)$$

If X_i is identically distributed as X, then

$$\phi_{S_n}(t) = (\phi_X(t))^n$$

Proof. This follows from direct calculation

$$\mathbb{E}[e^{tS_n}] = \mathbb{E}[e^{tX_1} \cdot e^{tX_2} \cdots e^{tX_n}] = \mathbb{E}[e^{tX_1}] \cdot \mathbb{E}[e^{tX_2}] \cdots \mathbb{E}[e^{tX_n}]$$

□

We know that the sum of independent Poisson random variable is Poisson from elementary probability. This can be verified via moment generating function. Suppose we have n independent random variables $X_1, ..., X_n$, such that

$$\phi_{X_i}(t) = e^{\lambda_i(e^t - 1)}, \ \forall i$$

Then

$$\phi_{X_1 + \cdots X_n}(t) = e^{(\lambda_1 + \cdots + \lambda_n)(e^t - 1)}$$

Thus, $X_1 + \cdots + X_n$ is *Poisson* with rate $(\lambda_1 + \cdots \lambda_n)$. Very similarly, one could also prove that the sum of independent normal random variable is normal (please verify by yourself!)

At last, let's apply moment generating function to prove the famous *Central Limit Theorem*:

Theorem 4.7 Assume that X is a random variable with mean μ and variance σ^2. Let $S_n = X_1 + \cdots + X_n$, where $X_1, ..., X_n$ are i.i.d., and distributed as X. Let

$$T_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Then, for every x,

$$\mathbb{P}(T_n \leq x) \to \mathbb{P}(Z \leq x),$$

as $n \to \infty$, where Z is a standard normal random variable.

Proof. Let $Y = \frac{X - \mu}{\sigma}$ and $Y_i = \frac{X_i - \mu}{\sigma}$, then Y_i are independent, distributed as Y, $\mathbb{E}[Y_i] = 0$, $Var[Y_i] = 1$, and

$$T_n = \frac{Y_1 + \cdots + Y_n}{\sqrt{n}}$$

To finish the proof, we show that $\phi_{T_n}(t) \to \phi_Z(t) = e^{t^2/2}$ as $n \to \infty$:

$$\begin{aligned}
\phi_{T_n}(t) &= \mathbb{E}[e^{tT_n}] \\
&= \mathbb{E}[e^{\frac{t}{\sqrt{n}}Y_1 + \cdots + \frac{t}{\sqrt{n}}Y_n}] \\
&= \mathbb{E}[e^{\frac{t}{\sqrt{n}}Y_1}] \cdots \mathbb{E}[e^{\frac{t}{\sqrt{n}}Y_n}] \\
&= \mathbb{E}[e^{\frac{t}{\sqrt{n}}Y}]^n \\
&= \left(1 + \frac{t}{\sqrt{n}}\mathbb{E}[Y] + \frac{1}{2}\frac{t^2}{n}\mathbb{E}[Y^2] + \frac{1}{6}\frac{t^3}{n^{3/2}}\mathbb{E}[Y^3] + \cdots\right)^n \\
&\approx \left(1 + \frac{t^2}{2}\frac{1}{n}\right)^n \\
&\to e^{\frac{t^2}{2}}
\end{aligned}$$

□

5 Reference

1. Jean Jacod, Philip Protter, "Probability Esentials", Springer, 2004;
2. Sheldom,M.Ross, Erol,A. Pekoz, "A second course in probability", ProbabilityBookstore.com, 2007
3. E.Cinlar, "Probability and Stochastics", Springer, 2011
4. Daniel, Ocone, "Notes in mathematical finance 2011"