CHAPTER 2

Probability

Probability theory is applied to situations where uncertainties exist, where we can identify a fixed set that includes all possible outcomes, and we can assign a probability describing the chance or likelihood for each outcomes. However, the outcome of a random experiment cannot be predicted with certainty before it occurs. Thus, probability theory is the right tool for solving problems of *known unknowns* (known uncertainties), such as the games in casinos. Admittedly, probability theory, the mathematical language we use to describe randomness, is the only game in the town to guess *unknown unknowns* (unknown uncertainties) in financial market.

Undergraduate students learning finance find that the probability theory in the literature is different from what they have learned in colleges and universities. A general review on probability theory is beyond the author's ability and the scope of this book. In this chapter, limited topics closely related to finance market are reviewed, and presented in the viewpoint of financial studies.

§ 2.1 Probability Spaces

A probability space consists of three parts:

1. A sample space, which is the set of all possible outcomes.
2. An event space, which is a collection of all the events.
3. The assignment of probabilities to the events.

Understanding these three concepts is the first step toward modern probability theory. Determining the independence of events is important because knowledge of the occurrence of one event has no effect on the probability of another event. Because financial markets are driven by information, independence, conditional probability and conditional independence are extremely useful for financial analysis. We should have a clear concept that conditional probability is a probability measure. Consequently, we understand well that independence and conditional independence are not implied by each other for they are defined in different probability spaces.

2.1.1 Sample Space

The non-empty set of all possible outcomes will be denoted by Ω and called the *sample space*. The elements of Ω will be denoted by w. If there are J possible outcomes of the experiment, we find it convenient to number them 1 through J, and so

$$\Omega = \{1, 2, ..., J\}$$

Example 2.1.1: If we toss a fair coin three times, the sample space is

$$\Omega = \{HHH, HHL, HLH, HLL, LHH, LHL, LLH, LLL\}$$

where an element is denoted by $w = w_1w_2w_3$ with $w_i = H$ standing for head, and $w_i = L$ standing for tail, $i = 1, 2, 3$.

Example 2.1.2: If the experiment consists of rolling a pair of dice — with the outcome being the pair (w_1, w_2), where w_1 is the value that appears on the first dice and w_2 the value on the second — then the sample space consists of the following 36 outcomes:

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

or

$$\Omega = \{(w_1, w_2) : w_1, w_2 = 1, 2, 3, 4, 5, 6\}$$

Any set of possible outcomes of the experiment is called an *event*. That is, a subset E of Ω, written by $\mathsf{E} \subseteq \Omega$ or $\mathsf{E} \leqslant \Omega$, is called an event. Intuitively, we should think of an event as a meaningful statement about the experiment. If an outcome is in the subset E, we say that event E *occurs*. In Example 2.1.2, if

$$\mathsf{E} = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$$

then E is the event that the sum of the dice equals 8. If the outcome is $(4,4)$, then the sum of the dice equals 8, we conclude that event E occurs.

In particular, Ω is called the certain event. An *elementary event* (also called an *atomic event*, a *simple event*, or a *point*, and often called a *state* in finance literature) is an event which contains only a single outcome in the sample space.

If $\{\mathsf{E}_i\}$ is a sequence of events, then the union of these events, denoted by $\sum_{i=1}^{\infty} \mathsf{E}_i$, or $\bigcup_{i=1}^{\infty} \mathsf{E}_i$, is defined to be the event which consists of all outcomes that are in E_i for at least one value of $i \in \mathbb{N}$. Similarly, the intersection of the events $\{\mathsf{E}_i\}$, denoted by $\prod_{i=1}^{\infty} \mathsf{E}_i$, or $\bigcap_{i=1}^{\infty} \mathsf{E}_i$, is defined to be the event consisting of those outcomes which are in all of the events E_i, $i \in \mathbb{N}$.

2.1.2 Event Space

Definition 2.1: A collection of events is called an *event space*, or *event field*, and denoted by $\mathcal{F}$, if it satisfies

(a) non-empty: $\Omega \in \mathcal{F}$;

(b) closed under complement: if $\mathsf{E} \in \mathcal{F}$, then $\mathsf{E}' \in \mathcal{F}$;

(c) closed under countable unions: if $\mathsf{E}_i \in \mathcal{F}$ for $i = 1, 2, \cdots$, then $\sum_{i=1}^{\infty} \mathsf{E}_i$ is in $\mathcal{F}$.

It is known from measure theory that a collection of sets that satisfy the above conditions is called a σ-algebra[1] (or a σ-field) on Ω. In finance literature, an event space is often called a *information set*. The pair $(\Omega, \mathcal{F})$ is often called a *measurable space*. It is sometimes convenient to call an event E is $\mathcal{F}$*-measurable* if $\mathsf{E} \in \mathcal{F}$. (Remind: E an event if $\mathsf{E} \subseteq \Omega$)

The smallest event space is trivial

$$\mathcal{F}_0 = \{\emptyset, \Omega\}$$

In Example 2.1.1, the event space that identify the first toss is

$$\mathcal{F}_1 = \{\emptyset, \Omega, \mathsf{E}_H, \mathsf{E}_L\}$$

where E_H is the event "the first toss is a head", and $\mathsf{E}_L = \mathsf{E}'_H$ is the event "the first toss is a tail"

$$\mathsf{E}_H = \{w : w_1 = H\} = \{HHH, HHL, HLH, HLL\}$$
$$\mathsf{E}_L = \{w : w_1 = L\} = \{LHH, LHL, LLH, LLL\} = \mathsf{E}'_H$$

[1] Greek letters σ and δ are often used when countable unions and countable intersections are involved. I am thinking that σ-algebra is an ugly name. It is an algebra, not a field in modern sense of mathematics.

we next define four sets

$$\mathsf{E}_{HH} = \{w : w_1 = H, w_2 = H\} = \{HHH, HHL\}$$
$$\mathsf{E}_{HL} = \{w : w_1 = H, w_2 = L\} = \{HLH, HLL\}$$
$$\mathsf{E}_{LH} = \{w : w_1 = L, w_2 = H\} = \{LHH, LHL\}$$
$$\mathsf{E}_{LL} = \{w : w_1 = L, w_2 = L\} = \{LLH, LLL\}$$

then the event space that describes the first two tosses is

$$\mathcal{F}_2 = \{\emptyset, \Omega, \mathsf{E}_{HH}, \mathsf{E}_{HL}, \mathsf{E}_{LH}, \mathsf{E}_{LL}, \mathsf{E}'_{HH}, \mathsf{E}'_{HL}, \mathsf{E}'_{LH}, \mathsf{E}'_{LL}\}$$
$$+ \{\mathsf{E}_H, \mathsf{E}_L, \mathsf{E}_{HH} + \mathsf{E}_{LH}, \mathsf{E}_{HH} + \mathsf{E}_{LL}, \mathsf{E}_{HL} + \mathsf{E}_{LH}, \mathsf{E}_{HL} + \mathsf{E}_{LL}\}$$

For Ω is a finite set of 8 elements, then the number of subsets of Ω is $2^8 = 256$. Let $\mathcal{F}_3$ be the set of all subsets of Ω, which is called the power set of Ω, then $\mathcal{F}_3$ is the largest possible event space. Event spaces depict the structure of information. The progressive event spaces $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$ reveal evolving detailed information.

We require that event space is closed under countable unions, why? Consider tossing a fair coin until "heads" occurs, where the sample space $\Omega = \mathbb{N}$ consists of an infinite number of points. Let $\mathsf{E}_i = \{2i\}$ for $i = 1, 2, \cdots$, then the event that heads show up on an even number toss should be

$$\mathsf{E} = \sum_{i=1}^{\infty} \mathsf{E}_i = \{2n : n \in \mathbb{N}\}$$

A suitable event space should contain such event.

Example 2.1.3: Let $\mathcal{F}$ be the collection of all subsets S of $\mathbb{N}$, such that either S or S' is finite. Is it an event space?

$\mathcal{F}$ is not close under the formation of countable unions, so it is not an event space. Let

$$\mathsf{S}_n = \{2n\} \qquad n = 1, 2, 3, \cdots$$

be the set of single even number, $\mathsf{S}_n \in \mathcal{F}$. However

$$\mathsf{E} = \sum_{n=1}^{\infty} \mathsf{S}_n = \{2n : n \in \mathbb{N}\} \notin \mathcal{F}$$

is the set of all even number, both E and E' are infinite sets.

2.1.3 Probability Measure

Probability assigns numbers to events. In Example 2.1.2, all elementary events of the sample space have the same chance to occur. Let E be the event that the sum of the dice equals 8, then we define the probability of the event E and denote it by $\mathrm{P}(\mathsf{E})$ as follows

$$\mathrm{P}(\mathsf{E}) = \frac{\#(\mathsf{E})}{\#(\Omega)} = \frac{5}{36}$$

where $\#(\mathsf{E})$ is the number of elements (simple events) in E.

Kolmogorov laid the foundations of modern probability theory in 1933. Who combined the notion of sample space and measure theory, and presented his axiom system for probability theory.

Definition 2.2: Probability is a set function, defined on event space $\mathcal{F}$ with values on $[0,1]$, satisfying the following properties:

(a) $0 \leqslant \mathrm{P}(\mathsf{E}) \leqslant 1$ for any event $\mathsf{E} \in \mathcal{F}$

(b) $\mathrm{P}(\Omega) = 1$

(c) For mutually exclusive events $\mathsf{E}_i \in \mathcal{F}$ for $i = 1, 2, \cdots$ (that is, events for which $\mathsf{E}_i\mathsf{E}_j = \emptyset$ when $i \neq j$), we have

$$\mathrm{P}\left(\sum_{i=1}^{\infty} \mathsf{E}_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(\mathsf{E}_i)$$

Any function $\mathrm{P} : \mathcal{F} \to [0,1]$ satisfying these conditions is called a *probability measure* (or simply *probability*).

If $\mathrm{P}(\mathsf{E}) = 1$ then we say E occurs with probability 1, or *almost surely* (a.s.). There are many equivalent definitions (at least 10) for probability measure, for example, replace the property (a) by: If $\mathsf{E} \in \mathcal{F}$, then $\mathrm{P}(\mathsf{E}') = 1 - \mathrm{P}(\mathsf{E})$.

In Definition 2.2, property (a) states that the probability of any event is some number between 0 and 1. Property (b) states that, with probability 1, the outcome will be a point in the sample space Ω. Property (c) is called *countable additivity*, which states that, for any sequence of mutually exclusive events, the probability of at least one of these events occurring is just the sum of their respective probabilities.

When the sample space Ω has finite J points, finite additivity

$$\mathrm{P}\left(\sum_{i=1}^{J} \mathsf{E}_i\right) = \sum_{i=1}^{J} \mathrm{P}(\mathsf{E}_i)$$

is equivalent to countable additivity. However, the generality to countable additivity is necessary when the sample space consists of an infinite number of points. Consider tossing a fair coin until "heads" occurs. A suitable sample space is $\Omega = \mathbb{N}$, let $\mathsf{E}_i = \{2i\}$, then $\mathsf{E} = \{2n : n \in \mathbb{N}\} = \sum_{i=1}^{\infty} \mathsf{E}_i$ is the event that the game ends at an even toss. For every positive integer J

$$\mathrm{P}(\mathsf{E}) > \mathrm{P}(\{2, 4, \cdots, 2J\}) = \sum_{i=1}^{J} 2^{-2i}$$

Hence (If $a_n > b_n$, with $\lim a_n = a$ and $\lim b_n = b$, we have $a \geqslant b$, not $a > b$. For example, $a_n = 1/n$, $b_n = 0$)

$$\mathrm{P}(\mathsf{E}) \geqslant \lim_{J\to\infty} \sum_{i=1}^{J} 2^{-2i} = \frac{1}{3}$$

Similarly

$$\mathrm{P}(\mathsf{E}) < \mathrm{P}(\{1, 3, \cdots, 2J-1\}') = 1 - \mathrm{P}(\{1, 3, \cdots, 2J-1\}) = 1 - \sum_{i=1}^{J} 2^{-(2i-1)}$$

Then

$$\mathrm{P}(\mathsf{E}) \leqslant 1 - \lim_{J\to\infty} \sum_{i=1}^{J} 2^{-(2i-1)} = \frac{1}{3}$$

Thus, the only "sensible" value for $\mathrm{P}(\mathsf{E})$ is $1/3$. Which is consistent with countable additivity

$$\mathrm{P}(\mathsf{E}) = \mathrm{P}\left(\sum_{i=1}^{\infty} \mathsf{E}_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(\mathsf{E}_i) = \sum_{i=1}^{\infty} 2^{-2i} = \frac{1}{3}$$

Proposition 2.3: If $\{\mathsf{E}_i\}$ is an increasing sequence of events ($\mathsf{E}_i \leqslant \mathsf{E}_{i+1}$), then

$$\lim_{n\to\infty} \mathrm{P}(\mathsf{E}_n) = \mathrm{P}\left(\sum_{i=1}^{\infty} \mathsf{E}_i\right)$$

and if $\{\mathsf{E}_i\}$ is a decreasing sequence of events ($\mathsf{E}_i \geqslant \mathsf{E}_{i+1}$), then

$$\lim_{n\to\infty} \mathrm{P}(\mathsf{E}_n) = \mathrm{P}\left(\prod_{i=1}^{\infty} \mathsf{E}_i\right) = \mathrm{P}\left(\bigcap_{i=1}^{\infty} \mathsf{E}_i\right)$$

Note that although the union of a sequence of events is an event, even knowing the probabilities of each of the events in this sequence does not usually permit computation of the probability of their union. More informations usually needed, such as that these events are mutually exclusive.

2.1.4 Conditional Probability

In Example 2.1.2, the probability of the event that the sum of the dice equals 8 is $\mathrm{P}(\mathsf{E}) = \frac{5}{36}$. If we get a tip that the number of the first dice is not greater than 4, say, $\mathsf{F} = \{(w_1, w_2) : w_1 \leqslant 4\}$: Given this "inside information", the elements with $w_1 > 4$ are excluded, the number of possible outcomes is reduced to 24. Clearly, the probability of the event E is now revised by

$$\frac{\#(\mathsf{FE})}{\#(\mathsf{F})} = \frac{\mathrm{P}(\mathsf{EF})}{\mathrm{P}(\mathsf{F})} = \frac{3}{24} = \frac{1}{8}$$

rather than $\frac{\#(\mathsf{E})}{\#(\Omega)} = \frac{5}{36}$. This motivates the following definition.

Let E and F be events in a random experiment with $\mathrm{P}(\mathsf{F}) > 0$. The *conditional probability* of E given F is defined to be

$$\mathrm{P}(\mathsf{E} \,|\, \mathsf{F}) = \frac{\mathrm{P}(\mathsf{EF})}{\mathrm{P}(\mathsf{F})} \tag{2.1}$$

Namely, suppose that we know that an event F has occurred, if E is another event then E occurs if and only if both event E and F occur. It follows that F can be thought of as the new sample space, and hence the probability that the event EF occurs (in the reduced sample space F) will equal the probability of EF relative to the probability of F.

Example 2.1.4: Let's play a game of Russian roulette. You are tied to a chair. Heres a gun, a revolver. Heres the barrel of the gun, six chambers, all empty. Now watch me as I put two bullets into the barrel, into two adjacent chambers. I close the barrel and spin it. I put the gun to your head and pull the trigger. Click. Lucky you! Now I'm going to pull the trigger one more time. Which would you prefer: that I spin the barrel first or that I just pull the trigger?

This is a question once asked on Wall Street job interviews. You are given the choice between an unconditional and a conditional probability of death.

- In the case the barrel is spun again, the probability of death is the (unconditional) probability that chamber contains a bullet, which is $2/6 = 1/3$.
- Given that the first shot is blank, we know that the former chamber is among the 4 empty ones. There is exactly one empty chamber followed by a nonempty one, hence if the trigger is pulled without the extra spin, the probability of death is $1/4$, the (conditional) probability of current chamber containing a bullet.

Thus, you should ask to pull the trigger again without randomly spinning the barrel first.

Remark: The more you know, the better your decision. We should make full use of all available information.

Let's try to define sample space: If the chamber is empty, it is denoted by 0, and otherwise by 1. Then for the state of the former and current chamber (first and second shot), a proper sample space is

$$\Omega = \{00, 01, 10, 11\} \tag{2.2}$$

Because there are three chambers out of six such that the former and current slot is empty, $\#(\{00\}) = 3$, $\mathrm{P}(00) = 3/6$, similarly, $\mathrm{P}(01) = \mathrm{P}(10) = \mathrm{P}(11) = 1/6$.

- In the case the barrel is spun again, the probability of death is $\mathrm{P}(\{01, 11\}) = 1/6 + 1/6 = 1/3$
- If the barrel is not spun again, given that the first shot is empty, the sample space is reduced to (the former chamber is one of the four empty slots)

  $$\Omega_* = \{00, 01\}$$

  and $\#(\Omega_*) = 4$. The probability of death is $\mathrm{P}_*(01) = 1/4$ (one chamber out of four such that the next chamber is loaded). Which is consistent with Equation (2.1): Let $\mathsf{E} = \{01, 11\}$ be the event that the second shot is not empty, and $\mathsf{F} = \{00, 01\}$ be the event that the first shot is empty, then the probability of death is

  $$\mathrm{P}(\mathsf{E} \,|\, \mathsf{F}) = \frac{\mathrm{P}(\mathsf{EF})}{\mathrm{P}(\mathsf{F})} = \frac{\mathrm{P}(01)}{\mathrm{P}(\{00, 01\})} = \frac{1/6}{3/6 + 1/6} = \frac{1}{4}$$

Remark: Other than the sample space define in (2.2), for the same problem, we can setup multiple sample spaces. Let us number the chamber by 1 to 6, and remember that the next of chamber 6 is 1 (in a circle), and define

$$\Omega = \{000011, 000110, 001100, 011000, 110000, 100001\}$$

Because of symmetry, let the first shot comes from chamber 1, then $\mathsf{F} = \{000011, 000110, 001100, 011000\}$ is the event that the first shot is empty, and $\mathsf{E} = \{011000, 110000\}$ is the event that the second shot (without spinning) is not empty. Without the extra spin, the probability of death at second shot is $\mathrm{P}(\mathsf{E} \,|\, \mathsf{F}) = \frac{\mathrm{P}(\mathsf{EF})}{\mathrm{P}(\mathsf{F})} = \frac{1}{4}$.

A: Conditional Probability is a Probability Measure

Given $\mathrm{P}(\mathsf{F}) > 0$, let $\mathrm{P}_{\mathsf{F}}(\mathsf{E}) = \mathrm{P}(\mathsf{E} \,|\, \mathsf{F})$ in Equation (2.1) for any event E, then $\mathrm{P}_{\mathsf{F}}(\cdot)$ is a probability measure defined on $\mathcal{F}$. Because $\mathrm{P}_{\mathsf{F}}(\cdot)$ satisfies all the conditions in Definition 2.2, in more details:

1. $0 \leqslant \mathrm{P}_{\mathsf{F}}(\mathsf{E}) \leqslant 1$ for any event $\mathsf{E} \in \mathcal{F}$
2. $\mathrm{P}_{\mathsf{F}}(\Omega) = 1$ (there is $\mathrm{P}_{\mathsf{F}}(\mathsf{F}) = 1, \mathrm{P}_{\mathsf{F}}(\Omega \backslash \mathsf{F}) = \mathrm{P}_{\mathsf{F}}(\mathsf{F}') = 0$)
3. For mutually exclusive events $\mathsf{E}_i \in \mathcal{F}$ for $i = 1, 2, \cdots$ (that is, events for which $\mathsf{E}_i\mathsf{E}_j = \emptyset$ when $i \neq j$), we have

$$\mathrm{P}_{\mathsf{F}}\left(\sum_{i=1}^{\infty} \mathsf{E}_i\right) = \sum_{i=1}^{\infty} \mathrm{P}_{\mathsf{F}}(\mathsf{E}_i)$$

Apparently, $\mathrm{P}_{\mathsf{F}}(\mathsf{Z}) = 0$ for any event $\mathsf{Z} \leqslant \mathsf{F}'$.

Remark: When $\mathrm{P}(\mathsf{F}) = 0$, we can define $\mathrm{P}_{\mathsf{F}}(\mathsf{E})$ for a specific application. An example can be found in Equation (3.19).

B: Multiplication Rule

Sometimes conditional probabilities are known and can be used to find the probabilities of other events. It follows from Equation (2.1) that there is a *multiplication rule*:

$$\mathrm{P}(\mathsf{EF}) = \mathrm{P}(\mathsf{E} \,|\, \mathsf{F})\mathrm{P}(\mathsf{F})$$

Which can be extended to the intersection of an arbitrary number of events

$$\mathrm{P}(\mathsf{E}_1\mathsf{E}_2\mathsf{E}_3 \cdots \mathsf{E}_J) = \mathrm{P}(\mathsf{E}_1)\mathrm{P}(\mathsf{E}_2 \,|\, \mathsf{E}_1)\mathrm{P}(\mathsf{E}_3 \,|\, \mathsf{E}_1\mathsf{E}_2) \cdots \mathrm{P}(\mathsf{E}_J \,|\, \mathsf{E}_1\mathsf{E}_2 \cdots \mathsf{E}_{J-1})$$

For partition $\{\mathsf{F}_i\}$ (which means $\sum_i \mathsf{F}_i = \Omega$ and $\mathsf{F}_i\mathsf{F}_j = \emptyset$ for $i \neq j$) with $\mathrm{P}(\mathsf{F}_i) > 0$ for any i, we have

$$\mathrm{P}(\mathsf{E}) = \sum_i \mathrm{P}(\mathsf{EF}_i) = \sum_i \mathrm{P}(\mathsf{E} \,|\, \mathsf{F}_i)\mathrm{P}(\mathsf{F}_i) \tag{2.3}$$

By the definition of conditional probability from Equation (2.1), if $\mathrm{P}(\mathsf{FG}) > 0$

$$\mathrm{P}(\mathsf{EF} \,|\, \mathsf{G}) = \mathrm{P}(\mathsf{E} \,|\, \mathsf{FG})\mathrm{P}(\mathsf{F} \,|\, \mathsf{G}) \tag{2.4}$$

or

$$\mathrm{P}(\mathsf{E} \,|\, \mathsf{FG}) = \frac{\mathrm{P}(\mathsf{EF} \,|\, \mathsf{G})}{\mathrm{P}(\mathsf{F} \,|\, \mathsf{G})} \tag{2.5}$$

Equation (2.3) has the following conditional version

$$\mathrm{P}\left(\mathsf{E} \,|\, \mathsf{G}\right) = \sum_i \mathrm{P}(\mathsf{EF}_i \,|\, \mathsf{G}) = \sum_{i:\mathrm{P}(\mathsf{F}_i\mathsf{G})>0} \mathrm{P}(\mathsf{E} \,|\, \mathsf{F}_i\mathsf{G})\mathrm{P}(\mathsf{F}_i \,|\, \mathsf{G}) \tag{2.6}$$

Note that in the last summation of Equation (2.6), if $\mathrm{P}(\mathsf{F}_i\mathsf{G}) = 0$, the term is excluded. The summation counts only the events with positive probabilities in the sequence of event $\mathsf{F}_i\mathsf{G}$.

2.1.5 Independence and Conditional Independence

Two events E and F are said to be *statistically independent* (or simply *independent*), denoted by $\mathsf{E} \perp \mathsf{F}$, if

$$\mathrm{P}(\mathsf{EF}) = \mathrm{P}(\mathsf{E})\mathrm{P}(\mathsf{F}) \tag{2.7}$$

We see the fact that Equation (2.7) is symmetric in E and F, which gives

$$\mathsf{E} \perp \mathsf{F} \iff \mathsf{F} \perp \mathsf{E}$$

Whenever E is independent of F, F is also independent of E. Thus, if E and F are independent

$$\mathrm{P}(\mathsf{E}\,|\,\mathsf{F}) = \mathrm{P}(\mathsf{E}) \qquad \mathrm{P}(\mathsf{F}\,|\,\mathsf{E}) = \mathrm{P}(\mathsf{F})$$

that is, the prior occurrence of F does not affect the probability of E, and vice versa. Besides, for

$$\mathsf{E} \perp \mathsf{F} \iff \mathsf{E}' \perp \mathsf{F} \tag{2.8}$$

we have $\mathsf{E} \perp \mathsf{F} \iff \mathsf{E}' \perp \mathsf{F} \iff \mathsf{E} \perp \mathsf{F}' \iff \mathsf{E}' \perp \mathsf{F}'$.

The terms *independent* and *disjoint* sound vaguely similar but they are actually very different. Disjointedness is purely a set-theory concept while independence is a probability (measure-theoretic) concept. Note that two disjoint events can never be independent, except in the trivial case that one of the events is null (zero probability).

Remark: Independence concerns only the relation on probability, it does not care other connections/relevance. Only probability matters! Suppose that we toss 2 fair dice as in Example 2.1.2: Let F denote the event that the first dice equals 4, and S be the event that the sum of the dice equals 7. Is F independent of S?

$$\mathrm{P}(\mathsf{FS}) = \mathrm{P}((4,3)) = \frac{1}{36} = \frac{1}{6}\cdot\frac{1}{6} = \mathrm{P}(\mathsf{F})\mathrm{P}(\mathsf{S}) \tag{2.9}$$

we leave it for the reader to present the intuitive argument why the event that the sum of the dice equals seven is independent of the outcome on the first dice. What about the total is 8? Let E be the event of getting a sum of 8

$$\mathsf{E} = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$$

we see that

$$\mathrm{P}(\mathsf{FE}) = \mathrm{P}((4,4)) = \frac{1}{36} \neq \mathrm{P}(\mathsf{F})\mathrm{P}(\mathsf{E}) = \frac{1}{6}\cdot\frac{5}{36} = \frac{5}{216}$$

If the first dice lands on 1, no chance of getting a total of 8. In other words, our chance of getting 8 depends on the outcome of the first dice. Hence F and E cannot be independent.

A: Pair-wise Independent

A family of events $\{\mathsf{E}_i\}$ is *pair-wise independent* if

$$\mathrm{P}(\mathsf{E}_i\mathsf{E}_j) = \mathrm{P}(\mathsf{E}_i)\mathrm{P}(\mathsf{E}_j) \qquad \forall i \neq j$$

Extending the definition of independence to more than two events requires some care. Three events E, F, and G are said to be independent if they are pair-wise independent and

$$\mathrm{P}(\mathsf{EFG}) = \mathrm{P}(\mathsf{E})\mathrm{P}(\mathsf{F})\mathrm{P}(\mathsf{G}) \tag{2.10}$$

We will show in the following examples that pair-wise independent does not imply Eq (2.10), and vise versa.

Example 2.1.5: In Example 2.1.2, let E be the event that the first throw is odd, F the event that the second throw is odd, and G be the event that the sum is 3. Then each pair of these events is

independent

$$\mathrm{P}(\mathsf{EF}) = \mathrm{P}(\mathsf{E})\mathrm{P}(\mathsf{F}) = \frac{1}{4}$$
$$\mathrm{P}(\mathsf{EG}) = \mathrm{P}(\mathsf{E})\mathrm{P}(\mathsf{G}) = \frac{1}{36}$$
$$\mathrm{P}(\mathsf{FG}) = \mathrm{P}(\mathsf{F})\mathrm{P}(\mathsf{G}) = \frac{1}{36}$$

however

$$\mathrm{P}(\mathsf{EFG}) = 0 \neq \mathrm{P}(\mathsf{E})\mathrm{P}(\mathsf{F})\mathrm{P}(\mathsf{G}) = \frac{1}{72}$$

This example gives three events that are pairwise independent, but not (mutually) independent. The following example gives three events with the property that the probability of the intersection is the product of the probabilities, such that Eq (2.10) is true, but the events are not pairwise independent.

Example 2.1.6: Suppose that we throw a fair dice one time. Let $\mathsf{E} = \{1, 2, 3, 4\}$, $\mathsf{F} = \{2, 3, 4\}$, $\mathsf{G} = \{4, 5, 6\}$. Then

$$\mathrm{P}(\mathsf{EFG}) = \frac{1}{6} = \mathrm{P}(\mathsf{E})\mathrm{P}(\mathsf{F})\mathrm{P}(\mathsf{G}) = \frac{4}{6}\frac{3}{6}\frac{3}{6}$$

However

$$\mathrm{P}(\mathsf{EF}) = \frac{1}{2} \neq \mathrm{P}(\mathsf{E})\mathrm{P}(\mathsf{F}) = \frac{4}{6}\frac{3}{6} = \frac{1}{3}$$
$$\mathrm{P}(\mathsf{EG}) = \frac{1}{6} \neq \mathrm{P}(\mathsf{E})\mathrm{P}(\mathsf{G}) = \frac{4}{6}\frac{3}{6} = \frac{1}{3}$$
$$\mathrm{P}(\mathsf{FG}) = \frac{1}{6} \neq \mathrm{P}(\mathsf{F})\mathrm{P}(\mathsf{G}) = \frac{3}{6}\frac{3}{6} = \frac{1}{4}$$

Of course, we may also extend the definition of independence to more than three events.

Definition 2.4: A finite family of events $\{\mathsf{E}_i\}_{i=1}^n$ are said to be (mutually) independent if

$$\mathrm{P}\left(\mathsf{E}_{i_1}\mathsf{E}_{i_2}\cdots\mathsf{E}_{i_k}\right) = \mathrm{P}\left(\mathsf{E}_{i_1}\right)\mathrm{P}\left(\mathsf{E}_{i_2}\right)\cdots\mathrm{P}\left(\mathsf{E}_{i_k}\right) \tag{2.11}$$

for any $k = 2, 3, \cdots, n$ and for any $1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant n$. An arbitrary family of events is defined to be independent if each of its finite subfamilies is independent.

A finite set of events is mutually independent if and only if every event is independent of any intersection of the other events. Intuitively, the events $\{\mathsf{E}_i\}_{i=1}^n$ are independent if knowledge of the occurrence of any subset of these events has no effect on the probability of any other event. That is, every subset of these events satisfies Eq (2.11).

B: Conditionally Independent

Events E and F are *conditionally independent* given G with $\mathrm{P}(\mathsf{G}) > 0$, denoted by $\mathsf{E} \perp \mathsf{F} \,|\, \mathsf{G}$, if

$$\mathrm{P}\left(\mathsf{EF} \,|\, \mathsf{G}\right) = \mathrm{P}\left(\mathsf{E} \,|\, \mathsf{G}\right)\mathrm{P}\left(\mathsf{F} \,|\, \mathsf{G}\right)$$

When $\mathrm{P}(\mathsf{FG}) > 0$

$$\mathsf{E} \perp \mathsf{F} \,|\, \mathsf{G} \iff \mathrm{P}(\mathsf{E} \,|\, \mathsf{FG}) = \mathrm{P}(\mathsf{E} \,|\, \mathsf{G}) \tag{2.12}$$

say, given G, joining the occurrence of F has no effect. Similarly to Eq (2.8), we have

$$\mathsf{E} \perp \mathsf{F} \,|\, \mathsf{G} \iff \mathsf{E}' \perp \mathsf{F} \,\big|\, \mathsf{G} \tag{2.13}$$

However, please note that

$$\mathsf{E} \perp \mathsf{F} \,|\, \mathsf{G} \not\Longrightarrow \mathsf{E} \perp \mathsf{F} \,|\, \mathsf{G}'$$

for the given event (information) has changed. The Chinese idiom goes: sweet oranges becomes sour and bitter when growing in north. When the environment or regime changes, our life and behaviors are inclined to change.

Example 2.1.7: In Example 2.1.2, let $\mathsf{E} = \{w_1 = 1\}$, $\mathsf{F} = \{w_2 = 1\}$ and $\mathsf{G} = \{w_1 \leqslant 3, w_2 \leqslant 3\}$. Then $\mathsf{E} \perp \mathsf{F} \,|\, \mathsf{G}$ for

$$\mathrm{P}(\mathsf{E} \,|\, \mathsf{FG}) = \frac{1}{3} = \frac{3}{9} = \mathrm{P}(\mathsf{E} \,|\, \mathsf{G})$$

However, given G', E and F are not conditionally independent, for

$$\mathrm{P}\left(\mathsf{E} \,|\, \mathsf{FG}'\right) = 0 \neq \frac{3}{27} = \mathrm{P}\left(\mathsf{E} \,|\, \mathsf{G}'\right)$$

It is worth noting that the conditional independence of E and F does not imply the independence of E and F, nor it is implied by the independence of E and F:

$$\mathsf{E} \perp \mathsf{F} \,|\, \mathsf{G} \not\Longleftrightarrow \mathsf{E} \perp \mathsf{F}$$

or

$$\mathrm{P}(\mathsf{E} \,|\, \mathsf{FG}) = \mathrm{P}(\mathsf{E} \,|\, \mathsf{G}) \not\Longleftrightarrow \mathrm{P}(\mathsf{E} \,|\, \mathsf{F}) = \mathrm{P}(\mathsf{E})$$

The following two examples show that $\mathsf{E} \perp \mathsf{F}$ neither implies nor is implied by $\mathsf{E} \perp \mathsf{F} \,|\, \mathsf{G}$.

Example 2.1.8 (independent but not conditionally independent): Let's flip two fair coins, and define the following events

- E: first coin comes up heads
- F: second coin comes up heads
- G: two flips were the same

E and F here are independent. However, E and F are not conditionally independent given G, since if you know G then your first coin flip will inform the other one. As an exercise

$$\mathrm{P}(\mathsf{E} \,|\, \mathsf{FG}) = 1 \neq \mathrm{P}(\mathsf{E} \,|\, \mathsf{G}) = 1/2$$

thus E and F are *conditionally dependent* given G.

Example 2.1.9 (conditionally independent but not independent): there are one fair coin and one coin biased towards heads in a box. I choose a coin at random and toss it twice. Define the following events

- E: first flip resulting in heads.
- F: second flip resulting in heads.
- G: the fair coin has been selected.

If we know E has occurred, we would guess that it is more likely that we have chosen a biased coin than a fair coin. Which in turn increases the conditional probability that F occurs, say, $\mathrm{P}(\mathsf{F} \mid \mathsf{E}) > \mathrm{P}(\mathsf{F})$ (as an exercise). This suggests that E and F are not independent. On the other hand, given G, E and F are independent.

The notion of conditional independence can easily be extended to more than two events: the conditional version of Equation (2.11) is

$$\mathrm{P}\left(\mathsf{E}_{i_1}\mathsf{E}_{i_2}\cdots\mathsf{E}_{i_k} \mid \mathsf{F}\right) = \mathrm{P}\left(\mathsf{E}_{i_1} \mid \mathsf{F}\right)\mathrm{P}\left(\mathsf{E}_{i_2} \mid \mathsf{F}\right)\cdots\mathrm{P}\left(\mathsf{E}_{i_k} \mid \mathsf{F}\right)$$

which is similar to Definition 2.4, with the modification that all probabilities in the definition are now conditional on F. As a note, even if we assume that events $\{\mathsf{E}_i\}_{i=1}^n$ are conditionally independent given F, it is not necessary that they be conditionally independent given F'. For example, in bull market, the returns of some assets are seen to be independent, however, in bear market, they are usually dropping in steps, thus highly correlated.

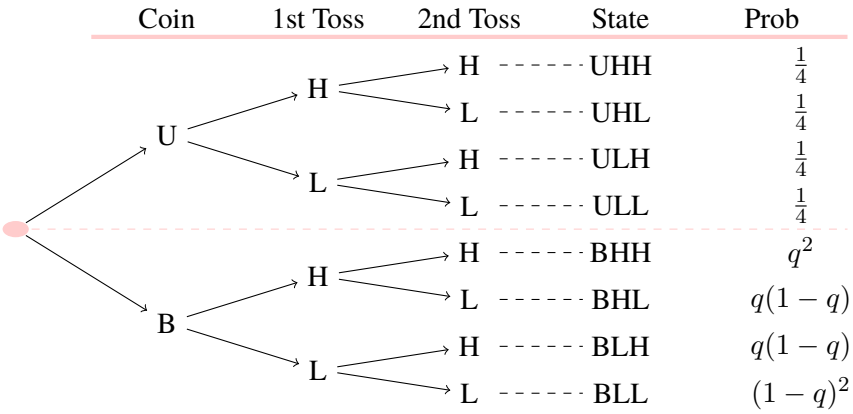
2.1.6 Triplet

Given a sample space Ω and an event space $\mathcal{F}$ of its subsets, if a set function $\mathrm{P} : \mathcal{F} \to [0,1]$ is a probability measure. Then the triplet $(\Omega, \mathcal{F}, \mathrm{P})$ is called a *probability space*.

Given a sample space Ω, we may interest in different event spaces. Furthermore, given the same measurable space $(\Omega, \mathcal{F})$, we may have many probability measures. For example, given a measurable space $(\Omega, \mathcal{F})$ and a probability measure P, for any event F with $\mathrm{P}(\mathsf{F}) > 0$, there is a probability measure $\mathrm{P}_{\mathsf{F}}(\cdot)$. In real world, subjective probabilities express a person's degree of belief, we might use experience, intuition, or a hunch to arrive at a value. Thus, different people can be expected to assign different probabilities to the same event. In finance, except the real world probability measure P, there is a imaginary world with probability measure Q, where investors are risk-neutral.

In an elementary approach to probability, any subset of the sample space is usually called an event. However, this gives rise to problems when the sample space is infinite, so that a more precise definition of an event is necessary. Under this definition only measurable subsets of the sample space, constituting a σ-algebra over the sample space itself, are considered events. However, this has essentially only theoretical significance, since in general the σ-algebra can always be defined to include all subsets of interest in applications.

Figure 2.1: Product Sample Space A 3-step experiment: (1) choosing a coin, U represents fair coin and B represents biased coin with probability $q > \frac{1}{2}$ for heads. (2) first toss, H standing for head, and L standing for tail. (3) second toss.



A: Product Sample Space

The *Cartesian product* of two sets X and Y is

$$\mathsf{X} \times \mathsf{Y} = \{(x, y) : x \in \mathsf{X}, y \in \mathsf{Y}\}$$

that is, the set of order pair (x, y) with $x \in \mathsf{X}, y \in \mathsf{Y}$. In a similar way we define the Cartesian product of $J \in \mathbb{N}$ sets. The sample space for a coin toss is $\{H, L\}$, for two tosses is

$$\{H, L\} \times \{H, L\} = \{(H, H), (H, L), (L, H), (L, L)\}$$

or simply $\{HH, HL, LH, LL\}$. When drawing a card from a standard deck of fifty-two playing cards, we could specify both the denomination and the suit, and a sample space describing each individual card can be constructed as the Cartesian product of the two sample spaces: suits (clubs, diamonds, hearts, spades) and ranks (Ace through King)

$$\{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\} \times \{A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K\}$$

For product of multiple sample spaces for a sequence of experiments, we often generate the sample space for the problem by making a tree diagram. For example, the sample space for Example 2.1.9 is shown in Figure 2.1.

B: Almost Surely

In elementary probability theory, we say if an event has zero probability, it is an impossible event. However, when the sample space is infinite, there are possible events with zero probability. For an instance, random points on unit square landing on the diagonal line is possible, but with probability zero. When the sample space is infinite, there will be more complicated and intrigued, amazing things may arise, as seen in the coming Example 2.1.10.

Example 2.1.10: Suppose that we possess an infinitely large urn and an infinite collection of balls labeled ball number 1, 2, and so on. Consider an experiment as follows.

1. At 1 hour to 0 a.m. balls numbered 1 through 10 are placed in the urn, and ball number 10 is withdrawn. At $1/2$ hour to 0 a.m., balls numbered 11 through 20 are placed in the urn, and ball number 20 is withdrawn. At $1/4$ hour to 0 a.m., balls numbered 21 to 30 are placed in the urn, and ball number 30 is withdrawn, and so on. The question of interest is, how many balls are in the urn at 0 a.m.?
2. Let us now change the experiment and suppose that at 1 hour to 0 a.m., balls numbered 1 through 10 are placed in the urn, and ball numbered 1 is withdrawn, at $1/2$ hour to 0 a.m., balls numbered 11 through 20 are place in the urn, and ball number 2 is withdrawn. At $1/4$ hour to 0 a.m., balls number 21 through 30 are placed in the urn, and ball number 3 is withdrawn, and so on. For this new experiment how many balls are in the urn at 0 a.m.?
3. Let us now suppose that whenever a ball is to be withdrawn that ball is randomly selected from among those present. That is, suppose that at 1 hour to 0 a.m. balls numbered 1 through 10 are placed in the urn, and a ball is randomly selected and withdrawn, and so on. In this case how many balls are in the urn at 0 a.m.?

In the first case only balls numbered $10n$, $n \geqslant 1$, are ever withdrawn; whereas in the second case all the balls are eventually withdrawn. Thus we see that the manner in which the withdrawn balls are selected makes a difference. In the third case, let's define F_n to be the event that ball number 1 is still in the urn after the first n withdrawals have been made. Clearly

$$\mathrm{P}(\mathsf{F}_n) = \frac{9}{10}\frac{18}{19}\cdots\frac{9n}{9n+1} = \prod_{i=1}^{n}\frac{9i}{9i+1}$$

We just note that if ball number 1 is still to be in the urn after the first n withdrawals, the first ball withdrawn can be any one of 9 out of 10, the second any one of 18 out of 19, and so on. Now, the event that ball number 1 is in the urn at 0 a.m. is just the event $\mathsf{E}_1 = \prod_{n=1}^{\infty}\mathsf{F}_n$, and as shown in Exercise 2.10

$$\mathrm{P}(\mathsf{E}_1) = 0$$

Thus, letting E_i denote the event that ball number i is in the urn at 0 a.m., we have shown that $\mathrm{P}(\mathsf{E}_1) = 0$. Similarly, we can show that $\mathrm{P}(\mathsf{E}_i) = 0$ for any ball number i. Finally, the probability that the urn is not empty at 0 a.m. should be

$$\mathrm{P}\left(\sum_{i=1}^{\infty}\mathsf{E}_i\right) \leqslant \sum_{i=1}^{\infty}\mathrm{P}(\mathsf{E}_i) = 0$$

by Boole's inequality (see Exercise 2.11). Thus, the urn will be empty at 0 a.m. almost surely (with probability 1).

C: Information and Decision-making

In Example 2.1.10, we see that the manners how the balls are withdrawn have an impact on the occurrence of the event that the urn is empty at 0 a.m. What's more, as a meaningful statement about the experiment, the constituent of an event may be changed by human activities. The actions or behaviors of participants in a game may reveal additional information, you should re-evaluate your decisions as new information emerges. In other words, information matters, you should re-evaluate the conditional probabilities as new information is added.

Example 2.1.11 (Monty Hall): Suppose you're on a game show, and you're given the choice of three doors, A, B and C: Behind one door is a car; behind the others, goats. You pick a door, say door A, you're hoping for the car of course. Monty Hall, the game show host, who knows what's behind the doors, opens another door, say door C, which has a goat. For simplicity, Assume that the car is initially hidden randomly, and Monty chooses a goat-hiding door to open completely at random. Now, having eliminated one of the choices, Monty asks you, "Do you want to keep your original guess door A? Or change it to door B?"

Typically, people will answer, "There's a 1 in 2 chance that I'll get it right, so it doesn't make any difference if I change my guess or not, the probability is 1/2 either way!" Surprisingly, the odds aren't 50-50 in this game. If you switch doors you'll win 2/3 of the time, while you have only a 1/3 chance if you stick to your initial choice.

Let W be the event that you win the car, and H be the event that you guess correctly at the first step, then $\mathrm{P}(\mathsf{H}) = \frac{1}{3}$. If you stick to your original choice

$$\mathrm{P}(\mathsf{W}) = \mathrm{P}(\mathsf{W}\,|\,\mathsf{H})\mathrm{P}(\mathsf{H}) + \mathrm{P}(\mathsf{W}\,|\,\mathsf{H}')\mathrm{P}(\mathsf{H}') = 1 \cdot \frac{1}{3} + 0 \cdot \frac{2}{3} = \frac{1}{3}$$

If you change to the leftover door

$$\mathrm{P}(\mathsf{W}) = \mathrm{P}(\mathsf{W}\,|\,\mathsf{H})\mathrm{P}(\mathsf{H}) + \mathrm{P}(\mathsf{W}\,|\,\mathsf{H}')\mathrm{P}(\mathsf{H}') = 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3}$$

It is to your advantage to switch your choice.

Let's write down the sample space explicitly, we shall see that the winning event W is changed by switching doors, your action (switching door) has an effect on the event W. Due to symmetry, we assume that your initial choice is door A, let $\{ik\}$ be the event that the car is behind door i and Monty opens door k, then a proper sample space is

$$\Omega_A = \{AB, AC, BC, CB\} \tag{2.14}$$

If you keep your original guess, $\mathsf{W} = \{AB, AC\}$, you win if and only if your initial guess is correct, there is $\mathrm{P}(\mathsf{W}) = \mathrm{P}(\mathsf{H}) = \frac{1}{3}$. If you switch doors, $\mathsf{W} = \{BC, CB\}$, you win if and only if your initial guess is incorrect, there is $\mathrm{P}(\mathsf{W}) = \mathrm{P}(\mathsf{H}') = \frac{2}{3}$. Evidently, the switching door action changes what the set of outcomes referred to by event W (variable change).

The fatal flaw in the Monty Hall paradox is not taking the new information revealed by Monty into account, thinking the chances are the same before and after. When you initially pick your door, the chance that you picked the correct one is $\frac{1}{3}$. Thus, the probability that you chose incorrectly is $\frac{2}{3}$, the two unselected doors as a whole has a chance of $\frac{2}{3}$. Undoubtedly, Monty's elimination transfers the whole chance of $\frac{2}{3}$ to the leftover door. Monty does improve the chance of the leftover door: now the weeds are pulled out, the leftover door contains the chance of the two unselected doors. Monty asks you to make a decision, not to try your luck or to follow your intuition that the odds are 50-50. You should exercise your logical and mathematical reasoning to your own interest.

In the view of finance, when you are offer a free right to make choice, you are given a free financial option: Monty offers you a free financial option — switching doors is a right but not an obligation, which has a non-negative value. You should evaluate this option and try to exercise it if it improves your chance.

Switching doors is a different action than choosing between the two remaining doors at random, as the first action uses the conditional probability and the latter does not. As a demonstration, let's compute the conditional probability that the car is behind door B, given that your initial choice is door A and Monty opens door C: Let $\mathsf{N} = \{A, B, C\}$, for any $i, j, k \in \mathsf{N}$, define $\{ijk\}$ to be the event that the car is behind door i, your initial choice is door j and Monty opens door k. Then the sample space of Monty Hall problem can be denoted by

$$\Omega = \{A, B, C\} \times \{A, B, C\} \times \{A, B, C\} \tag{2.15}$$

note that the probabilities of events $\{iji\}$ and $\{ijj\}$ are all zero. Let $\mathsf{D}_i = \{ijk : j, k \in \mathsf{N}\}$, $\mathsf{Y}_j = \{ijk : i, k \in \mathsf{N}\}$ and $\mathsf{M}_k = \{ijk : i, j \in \mathsf{N}\}$. Then event D_i indicates that the car is behind door i; Y_j is the event that you choose door j, and M_k is the event that Monty opens door k. Since the car is initially hidden randomly, then

$$\mathrm{P}(\mathsf{D}_i \,|\, \mathsf{Y}_j) = \mathrm{P}(\mathsf{D}_i) = \frac{1}{3} \qquad i, j \in \mathsf{N}$$

From the settings of the game we have

$$\mathrm{P}(\mathsf{M}_C \,|\, \mathsf{D}_A\mathsf{Y}_A) = \frac{1}{2} \qquad \mathrm{P}(\mathsf{M}_C \,|\, \mathsf{D}_B\mathsf{Y}_A) = 1 \qquad \mathrm{P}(\mathsf{M}_C \,|\, \mathsf{D}_C\mathsf{Y}_A) = 0$$

By Equation (2.6)

$$\mathrm{P}(\mathsf{M}_C \,|\, \mathsf{Y}_A) = \sum_{i \in \mathsf{N}} \mathrm{P}(\mathsf{M}_C \,|\, \mathsf{D}_i\mathsf{Y}_A)\mathrm{P}(\mathsf{D}_i \,|\, \mathsf{Y}_A) = \frac{1}{3}\left(\frac{1}{2} + 1 + 0\right) = \frac{1}{2}$$

Given that your initial choice is door A and Monty opens door C, by Eq (2.5) and (2.4), the conditional probability of D_B (the car is behind door B) is

$$\mathrm{P}(\mathsf{D}_B \,|\, \mathsf{M}_C\mathsf{Y}_A) = \frac{\mathrm{P}(\mathsf{D}_B\mathsf{M}_C \,|\, \mathsf{Y}_A)}{\mathrm{P}(\mathsf{M}_C \,|\, \mathsf{Y}_A)} = \frac{\mathrm{P}(\mathsf{M}_C \,|\, \mathsf{D}_B\mathsf{Y}_A)\mathrm{P}(\mathsf{D}_B \,|\, \mathsf{Y}_A)}{\frac{1}{2}} = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

Similarly, the conditional probability of D_A is

$$\mathrm{P}(\mathsf{D}_A \,|\, \mathsf{M}_C\mathsf{Y}_A) = \frac{\mathrm{P}(\mathsf{D}_A\mathsf{M}_C \,|\, \mathsf{Y}_A)}{\mathrm{P}(\mathsf{M}_C \,|\, \mathsf{Y}_A)} = \frac{\mathrm{P}(\mathsf{M}_C \,|\, \mathsf{D}_A\mathsf{Y}_A)\mathrm{P}(\mathsf{D}_A \,|\, \mathsf{Y}_A)}{\frac{1}{2}} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

Alternatively

$$\mathrm{P}(\mathsf{D}_A \,|\, \mathsf{M}_C\mathsf{Y}_A) = 1 - \mathrm{P}(\mathsf{D}_B \,|\, \mathsf{M}_C\mathsf{Y}_A) - \mathrm{P}(\mathsf{D}_C \,|\, \mathsf{M}_C\mathsf{Y}_A) = 1 - \frac{2}{3} - 0 = \frac{1}{3}$$

Thus, in the light of current situation, switching is the winning strategy.

When a behavior is not random, it may reveal additional different information. For example, assume you know that Monty does not open the door randomly among all legal alternatives but instead, Monty will open the losing doors on the right. In this situation, if you choose door A and Monty opens door B, then the car must be hidden behind door C.

Our goal in this example of Monty Hall paradox is not to understand the puzzle — it's to emphasize that conditional probability plays a key role in our decision-making. The more you know, the better your decision. Whenever there is new information, you should recompute the *conditional probability* (*conditional expectation*) and optimize your decision accordingly. To incorporate known information, the popular mathematical tool is conditional expectation, which is studied in Chapter 3.

D: Symmetry

Symmetry is a powerful guiding principle. In probability theory, symmetry refers to a situation in which multiple viewpoints will behave exactly the same. When a probability problem exhibits symmetry, all relevant probabilities are invariant under certain transformations, such as the interchange of two label A and B in the model, or reverse the positive and negative sign. Thus, in most case, it suffice to consider only one of the viewpoints to reduce complexity.

Symmetry usually means that the events are obviously exclusive, mutually exhaustive, and interchangeable in all relevant ways with respect to probabilities. In Example 2.1.11, doors A, B and C are symmetry, in the sense that your initial choice is door A, or B, or C, makes no real difference, they behave exactly the same. Thus, without loss of generality, we assume that your initial choice is door A, and denote the sample space to be the Ω_A in Eq (2.14). Obviously, Ω_A is a reduced version of Ω in Eq (2.15), by setting $j = A$ and trimming all null sets. Under this reduction, $\mathsf{D}_A = \{AB, AC\}$, $\mathsf{D}_B = \{BC\}$ and $\mathsf{D}_C = \{CB\}$. Now, if you keep your original guess, $\mathsf{W} = \{AB, AC\} = \mathsf{D}_A$, there is

$$\mathrm{P}(\mathsf{W}) = \mathrm{P}(\mathsf{D}_A) = \frac{1}{3}$$

Similarly, if you switch doors, $\mathsf{W} = \{BC, CB\} = \mathsf{D}_B + \mathsf{D}_C$, there is

$$\mathrm{P}(\mathsf{W}) = \mathrm{P}(\mathsf{D}_B) + \mathrm{P}(\mathsf{D}_C) = \frac{2}{3}$$

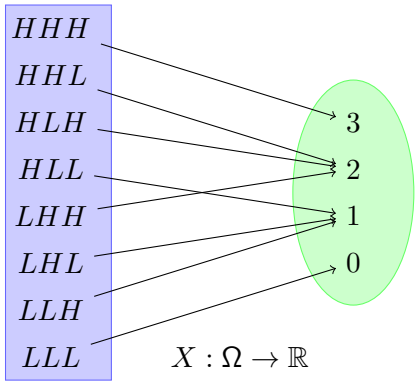
The results agree with our analysis in previous text.

§ 2.2 Random Variable and Distribution

In the course of elementary probability, we know that a random variable has a whole set of values, and it could take on any of those values, randomly. In modern probability theory, a random variable is understood as a numerical value function defined on a sample space. A random variable is a procedure for assigning numbers to random outcomes, and, contrary to its name, this procedure itself is neither random nor variable. A random variable digitizes the outcomes, the correspondence between the sample space and real number is defined according to various interests.

2.2.1 Random Variable

As opposed to the actual outcome itself in the random experiment, we are interested mainly in some function of the outcome. In flipping a coin, we may be interested in the total number of heads that occur and not care at all about the actual head-tail sequence that results. Let us define by X the number of heads in Example 2.1.1. Then, we mark the outcomes as numbers



If the outcome of the experiment was $w = HHH \in \Omega$, then $X = 3$. Actually

$$X(HHH) = 3$$

furthermore

$$X(HHL) = X(HLH) = X(LHH) = 2$$
$$X(HLL) = X(LHL) = X(LLH) = 1$$
$$X(LLL) = 0$$

In this way, we define a function $X : \Omega \to \mathbb{R}$ with domain in the finite sample space Ω and with values in $\mathbb{R}$. Such a function is called a *random variable*. Intuitively, you should think of a random variable X as a measurement of interest in the context of the random experiment. A random variable X is random in the sense that its value depends on the outcome of the experiment, which cannot be predicted with certainty before the experiment is run. Each time the experiment is run, an outcome (state) $w \in \Omega$ occurs, and a given random variable X takes on the value $X(w)$.

It is possible to define more than one random variable in a sample space. For example, we could

define the random variable Y to be the number of reversals, then

$$Y(HHH) = Y(LLL) = 0$$
$$Y(HHL) = Y(LHH) = Y(HLL) = Y(LLH) = 1$$
$$Y(HLH) = Y(LHL) = 2$$

The important point is simply that a random variable is a function defined on the sample space Ω.

A: Information

For random variable X, we see that if $X = 2$, then the actual outcome was HHL, HLH, or LHH. If $X \in \{0, 3\}$, then the actual outcome was LLL or HHH. Thus, a statement about the random variable defines an event. For simplicity, we write

$$X^{-1}(2) = \{HHL, HLH, LHH\}$$

and

$$X^{-1}(\{0, 3\}) = \{LLL, HHH\}$$

In general, for any Borel set B of $\mathbb{R}$ ($\mathsf{B} \subset \mathbb{R}$) we denote by $X^{-1}(\mathsf{B})$ or by $\{X \in \mathsf{B}\}$ the *inverse image* of B under X, that is

$$X^{-1}(\mathsf{B}) = \{X \in \mathsf{B}\} \equiv \{w \in \Omega : X(w) \in \mathsf{B}\}$$

For example, if $\mathsf{B} = (-\infty, 3]$, $X^{-1}(\mathsf{B}) = \{X \leqslant 3\} < \Omega$. If B contains only one real number, say, $\mathsf{B} = \{2\}$, we write $X^{-1}(2) \equiv X^{-1}(\{2\}) = X^{-1}(\mathsf{B})$, or

$$X^{-1}(2) = \{X = 2\}$$

The family of all sets of the form $\{X \in \mathsf{B}\}$ is denoted by $\sigma(X)$ and is called the *σ-algebra generated by X*. For example

$$\sigma(Y) = \{\emptyset, \Omega, Y^{-1}(0), Y^{-1}(1), Y^{-1}(2),\ Y^{-1}(\{0, 1\}),\ Y^{-1}(\{0, 2\}),\ Y^{-1}(\{1, 2\})\}$$

which has fewer events than $\mathcal{F}_3$ (with 256 events), thus representing a coarser type of information.

Remark: In finance, the value of a random variable X is not the only concern, in most situations, the information reveal by X, the inverse image, and the σ-algebra generated by X, is more important. When X takes a value, it uncovers the undergoing event, and the true state of world.

The *image* of event E under X, written $X(\mathsf{E})$, is defined by

$$X(\mathsf{E}) = \{X(w) : w \in \mathsf{E}\}$$

Remark: The image of event E under X, $X(\mathsf{E})$, is a Borel set. The inverse image of Borel set B under X, $X^{-1}(\mathsf{B})$, is an event.

- For any event $\mathsf{E} \in \sigma(X)$, there is a Borel set $\mathsf{B} = X(\mathsf{E})$ and $\mathsf{E} = X^{-1}(\mathsf{B}) = X^{-1}(X(\mathsf{E}))$. For example, let $\mathsf{E} = \{HHH, LLL\} \in \sigma(X)$, then $X(\mathsf{E}) = \{0, 3\}$ is a Borel set, and $X^{-1}(\{0, 3\}) = \mathsf{E}$ is an event.
- If $\mathsf{E} \notin \sigma(X)$, then $X^{-1}(X(\mathsf{E})) \in \sigma(X)$ as $X(\mathsf{E})$ is a Borel set, but $X^{-1}(X(\mathsf{E})) \neq \mathsf{E}$. For example, let $\mathsf{E} = Y^{-1}(2) \notin \sigma(X)$, then $X^{-1}(X(\mathsf{E})) = X^{-1}(\{1, 2\}) \neq \mathsf{E}$.

When the sample space Ω is infinite, there may exist a subset that cannot be assigned a meaningful "probability" (called non-measurable set). Therefore, a more precise definition of a random variable is necessary. As a function, a random variable is required to be measurable, which allows for probabilities to be assigned to sets of its potential values.

Definition 2.5: A random variable on $(\Omega, \mathcal{F})$ is a function $X : \Omega \to \mathbb{R}$ such that for each real number x

$$\{X \leqslant x\} \in \mathcal{F}$$

Remark: X is a random variable on $(\Omega, \mathcal{F})$ if and only if $X^{-1}(\mathsf{B}) \in \mathcal{F}$ for any Borel set B of $\mathbb{R}$. Although the latter is seemingly stronger.

- Random variables are defined on a measurable space $(\Omega, \mathcal{F})$, not on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. The probability measure is not involved in the definition of a random variable.
- Random variables can be discrete, that is, taking any of a specified finite or countable list of values; or continuous, taking any numerical value in an interval or collection of intervals; or a mixture of both types.
- For random variables X and Y, each element in $\sigma(X, Y)$ can be formed by sets of the following form

  $$(X \leqslant x, Y \leqslant y) \equiv \{X \leqslant x\} \cap \{Y \leqslant y\}$$

  for certain real number x and y.

For any event E, $\mathit{1}_{\mathsf{E}} \equiv \mathit{1}_{\mathsf{E}}(w) = \mathit{1}\ (w \in \mathsf{E})$ is called the *indicator variable* of E. $\mathit{1}_{\mathsf{E}}$ is a random variable

$$\mathit{1}_{\mathsf{E}}(w) = \begin{cases} 1 & w \in \mathsf{E} \\ 0 & w \in \mathsf{E}' \end{cases}$$

$\mathit{1}_{\mathsf{E}} = 1$ if and only if E occurs, and $\mathit{1}_{\mathsf{E}} = 0$ if and only if E' occurs. Let $Z = \mathit{1}_{\mathsf{E}}$, then

$$\{Z = 1\} = Z^{-1}(1) = \mathsf{E}$$

Let $\mathcal{G}$ be an event space, a random variable Z is said to be *$\mathcal{G}$-measurable*, and denoted by $Z \in \mathcal{G}$, if

$$\{Z \leqslant z\} \in \mathcal{G}$$

for all $z \in \mathbb{R}$, that is, $\sigma(Z) < \mathcal{G}$. Thus, a random variable Y being $\sigma(X)$-measurable, $Y \in \sigma(X)$, is the same as $Y^{-1}(\mathsf{B}) \in \sigma(X)$ for any Borel set B, which is the same as $\sigma(Y) < \sigma(X)$, thus $\sigma(X)$ is a finer event space. Let X be the number of heads in Example 2.1.1, and Y be the number of reversals, then Y is not $\sigma(X)$ measurable, for $Y^{-1}(2) = \{HLH, LHL\} \notin \sigma(X)$. Conversely, X is not $\sigma(Y)$ measurable, for $X^{-1}(0) = \{LLL\} \notin \sigma(Y)$. To check the measurability of random variable to an event space, there is an easy way by Doob-Dynkin lemma: for some function $h(\cdot)$

$$Y = h(X) \iff Y \in \sigma(X) \iff \sigma(Y) < \sigma(X)$$

In Example 2.1.1, $Y(X = 1) = \{1, 2\}$, and $X(Y = 0) = \{0, 3\}$, because of multi-value, Y can not be a function of X, and X is not a function of Y, thus, $Y \notin \sigma(X)$ and $X \notin \sigma(Y)$.

B: Independence

A random variable's possible values might represent the possible events. Thus, the random variables X and Y on $(\Omega, \mathcal{F}, \mathrm{P})$ are said to be independent, and denoted by $X \perp Y$, if any events in $\sigma(X)$ are independent of any events in $\sigma(Y)$.

Definition 2.6: We say that σ-algebras $\mathcal{G}$ and $\mathcal{H}$ in $\mathcal{F}$ are independent, denoted by $\mathcal{G} \perp \mathcal{H}$, if for any $\mathsf{G} \in \mathcal{G}$ and $\mathsf{H} \in \mathcal{H}$ the events G and H are independent. We can now say that random variables X and Y are independent if and only if the σ-algebras $\sigma(X)$ and $\sigma(Y)$ are independent.

Remark: By the definition of independence, if $X \perp Y$, then $\{Y < c\} \perp \{X = c\}$ for any constant c. As a quick demonstration, in tossing a pair of dice, let X be the points of first dice and Y the second's, apparently, when $c = 3$, $Y < 3$ is independent of $X = 3$.

Example 2.2.1: Let's keep rolling a pair of dice, and end the game either the sum is 4 (win, denoted as $Z = 1$) or 7 (lose, $Z = 0$). Let the number of rolls be Y, are Y and Z independent?

Let S_n be the sum of nth roll, define $\mathsf{V}_1 = \{S_1 = 4\}$ and (let $(\mathsf{E}, \mathsf{F}) \equiv \mathsf{E} \cap \mathsf{F}$)

$$\mathsf{V}_n = (S_n = 4, \{S_{n-1}, \cdots, S_2, S_1\} \cap \{4, 7\} = \emptyset) \qquad n > 1$$

then V_n is the event that the player wins at the nth roll. Let $\mathsf{R}_1 = \{S_1 \in \{4, 7\}\}$ and

$$\mathsf{R}_n = (S_n \in \{4, 7\}, \{S_{n-1}, \cdots, S_2, S_1\} \cap \{4, 7\} = \emptyset) \qquad n > 1$$

then R_n is the event that the number of rolls is n. Clearly, the event that the player wins is $\mathsf{V} = \sum_n \mathsf{V}_n$, and there is $\mathsf{R}_n\mathsf{V} = \mathsf{V}_n$. Let $P_4 = 1/12$, $P_7 = 1/6$ and $q = 1 - (P_4 + P_7)$, then $\mathrm{P}(\mathsf{V}_n) = P_4 q^{n-1}$, $\mathrm{P}(\mathsf{R}_n) = (P_4 + P_7)q^{n-1}$. For events V_n are mutually exclusive

$$\mathrm{P}(\mathsf{V}) = \sum_n \mathrm{P}(\mathsf{V}_n) = \sum_{n=1}^{\infty} P_4 q^{n-1} = \frac{P_4}{P_4 + P_7} \tag{2.16}$$

Thus

$$\mathrm{P}(\mathsf{R}_n\mathsf{V}) = \mathrm{P}(\mathsf{V}_n) = P_4 q^{n-1} = \mathrm{P}(\mathsf{R}_n)\mathrm{P}(\mathsf{V})$$

which states that $\{Y = n\} = \mathsf{R}_n$ is independent of $\{Z = 1\} = \mathsf{V}$, $n = 1, 2, \cdots$. By Eq (2.8), we have $\{Y = n\} \perp \{Z = 0\}$. Therefore, $\sigma(Y)$ and $\sigma(Z)$ are independent, or equivalently, probability mass function $f(y, z) = f_Y(y) f_Z(z)$, say, $Y \perp Z$.

Remark: Let n be the value of the number of final throw

$$\mathrm{P}(Z = 1 \,|\, Y = n) = \mathrm{P}(\mathsf{V} \,|\, \mathsf{R}_n) = \frac{\mathrm{P}(\mathsf{V}\mathsf{R}_n)}{\mathrm{P}(\mathsf{R}_n)} = \frac{\mathrm{P}(\mathsf{V}_n)}{\mathrm{P}(\mathsf{R}_n)} = \frac{P_4 q^{n-1}}{(P_4 + P_7)q^{n-1}} = \frac{P_4}{P_4 + P_7}$$

Which is equal to the probability that the final sum is 4 out of a sum of 4 or 7. We see that given the number n of rolls, the conditional probability of winning is a constant, knowing the value $Y = n$ does not affect the probability of $\{Z = 1\}$. Thus, the number of rolls is independent of the event of winning. More formally, by Eq (2.3)

$$\mathrm{P}(Z = 1) = \sum_i \mathrm{P}(Z = 1 \,|\, Y = i)\mathrm{P}(Y = i) = \mathrm{P}(Z = 1 \,|\, Y = n)$$

which indicates $\{Y = n\} \perp \{Z = 1\}$.

Definition 2.7: For random variables X, Y and Z, we say Y and (X, Z) are independent, denoted by $Y \perp (X, Z)$, if and only if $\sigma(Y) \perp \sigma(X, Z)$, where $\sigma(X, Z)$ is the family of all events of the form

$$(X^{-1}(\mathsf{B}), Z^{-1}(\mathsf{C})) \equiv \{X \in \mathsf{B}\} \cap \{Z \in \mathsf{C}\}$$

for any Borel sets B and C.

A set of random variables $\mathsf{X} = \{X_1, X_2, \ldots, X_n\}$ is mutually independent if and only if for any finite subset of X are mutually independent.

2.2.2 Distribution Function

A random variable that can take on at most a countable number of possible values is said to be discrete. For a discrete random variable X, we define the *probability mass function* (PMF) $f_X(x)$ by

$$f_X(x) = \mathrm{P}(X = x) \equiv \mathrm{P}(\{X = x\})$$

The probability mass function $f_X(x)$ is positive for at most a countable number of values. That is, if X must assume one of the values $x_1, x_2, \cdots$, then

$$f_X(x_i) > 0 \qquad i = 1, 2, \cdots$$

and $f_X(x) = 0$ for $x \in \mathbb{R} \backslash \{x_1, x_2, \cdots\}$. Since X must take on one of the values $x_1, x_2, \cdots$, we have

$$\sum_i f_X(x_i) = 1$$

The *cumulative distribution function* (CDF), or simply *distribution function*, $F_X(x)$ can be expressed in terms of $f_X(x)$ by

$$F_X(x) = \mathrm{P}(X \leqslant x) \equiv \mathrm{P}(\{X \leqslant x\}) = \sum_{i: x_i \leqslant x} f_X(x_i)$$

Let X be a continuous random variable. Then we define a cumulative distribution function of X, similar to the discrete random variable case, to be the function

$$F_X(x) = \mathrm{P}(X \leqslant x)$$

If there exists[2] a function $f_X(x)$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t \qquad -\infty < x < +\infty$$

Then the function $f_X(x)$ is called *probability density function* (PDF) of the random variable X.

A: Joint Distribution

We are often interested in joint behavior concerning two or more random variables. To describe probabilities of such behaviors, we define, for any two random variables X and Y, the *joint probability*

[2] Not every probability distribution has a density function: for example, the Cantor distribution has neither a probability density function nor a probability mass function.

distribution function of X and Y by

$$F(x,y) = \mathrm{P}(X \leqslant x, Y \leqslant y) \qquad -\infty < x, y < +\infty$$

For

$$F_X(x) = \mathrm{P}(X \leqslant x) = F(x, +\infty)$$
$$F_Y(y) = \mathrm{P}(Y \leqslant y) = F(+\infty, y)$$

We call $F_X(x)$ and $F_Y(y)$ the *marginal distributions* of $F(x,y)$.

B: Conditional Distribution

For continuous random variables X and Y, if they have a joint probability density function $f(x,y)$

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,\mathrm{d}v\mathrm{d}u$$

then the *conditional probability density function* of Y given that $X = x$ is defined, for all values of x such that $f_X(x) > 0$, by

$$f\left(y \,|\, x\right) = \frac{f\left(x,y\right)}{f_X\left(x\right)}$$

Similarly, if X and Y are both discrete random variables, it is useful to define the *joint probability mass function* of X and Y by

$$f(x,y) = \mathrm{P}(X = x, Y = y)$$

and the *conditional probability mass function* of Y given that $X = x$ is defined, for all values of x such that $\mathrm{P}\left(X = x\right) > 0$, by

$$f\left(y \,|\, x\right) = \mathrm{P}\left(Y = y \,|\, X = x\right) = \frac{\mathrm{P}\left(X = x, Y = y\right)}{\mathrm{P}\left(X = x\right)} = \frac{f\left(x,y\right)}{f_X\left(x\right)}$$

C: Independent

Random variables X and Y (whatever their type) are independent, denoted by $X \perp Y$, if and only if their joint distribution function is the product of their individual distribution functions, that is

$$F(x,y) = F_X(x)F_Y(y)$$

If X and Y are jointly continuous with density $f(x,y)$, or they are both discrete with probability mass function $f(x,y)$, then they are independent if and only if

$$f(x,y) = f_X(x)f_Y(y)$$

2.2.3 Expectation

One of the most important concepts in probability theory is that of the expectation of a random variable. For discrete random variables, the expected value is a weighted average of its values with weights of the respective probabilities.

Definition 2.8: Let X be a random variable on $(\Omega, \mathcal{F}, \mathrm{P})$, then the *expectation* of X, denoted by $\mathrm{E}(X)$, is defined as the Lebesgue integral

$$\mathrm{E}(X) = \int_{\Omega} X(w)\,\mathrm{d}P(w)$$

In a more familiar way:

- If X is a discrete random variable, then the expectation of X is

$$\mathrm{E}(X) = \sum_{w\in\Omega} X(w)\mathrm{P}(w) = \sum_{i} x_i \mathrm{P}(X = x_i) \tag{2.17}$$

  when $\sum_i |x_i|\,\mathrm{P}(X = x_i) < \infty$

- If X is a continuous random variable with probability density function $f_X(x)$, then

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} x\,\mathrm{d}F_X(x) = \int_{-\infty}^{+\infty} x f_X(x)\,\mathrm{d}x$$

  when $\int_{-\infty}^{+\infty} |x|\ \mathrm{d}F_X(x) < \infty$

Alternative names for $\mathrm{E}(X)$ are the *expected value* or the *mean* of X. In particular, in the case of a finite $\Omega = \{x_1, x_2, ..., x_N\}$ with uniform probability, we obtain the arithmetic average

$$\mathrm{E}(X) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

For indicator variable 1_{E}, by Eq (2.17) we have

$$\mathrm{E}(1_{\mathsf{E}}) = \mathrm{P}(\mathsf{E}) \tag{2.18}$$

The expectation operator has the following properties: For random variables X and Y

1. Linearity: $\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y), \forall a, b \in \mathbb{R}$
2. Increasing: if $X \geqslant Y$, then $\mathrm{E}(X) \geqslant \mathrm{E}(Y)$; And if $X \gneqq Y$, then $\mathrm{E}(X) > \mathrm{E}(Y)$
3. X and Y are independent if and only if

$$\mathrm{E}(g(X)h(Y)) = \mathrm{E}(g(X))\,\mathrm{E}(h(Y)) \tag{2.19}$$

   for all choices of bounded (Borel measurable) functions $g, h : \mathbb{R} \to \mathbb{R}$.

A: Law of the Unconscious Statistician

The expectation of a function $h(X)$ of a random variable X with CDF $F_X(x)$ is given by

$$\mathrm{E}(h(X)) = \int_{\Omega} h(X(w))\,\mathrm{d}P(w) = \int_{\mathbb{R}} h(x)\,\mathrm{d}F_X(x)$$

In elementary probability theory, if X has a probability density function $f_X(x)$, one does not explicitly know the distribution of $h(X)$, there is

$$\mathrm{E}(h(X)) = \int_{-\infty}^{+\infty} h(x) f_X(x)\,\mathrm{d}x$$

and if X is a discrete random variable, then

$$\mathrm{E}(h(X)) = \sum_{i} h(x_i)\mathrm{P}(X = x_i)$$

For a (measurable) function h of several random variables: If X and Y have a joint probability mass function $f(x, y)$, then

$$\mathrm{E}(h(X,Y)) = \sum_i \sum_j h(x_i, y_j) f(x_i, y_j)$$

If X and Y have a joint probability density function $f(x, y)$, then

$$\mathrm{E}(h(X,Y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(x,y) f(x,y)\,\mathrm{d}x\mathrm{d}y$$

In general, the expectation operator and functions of random variables do not commute; that is

$$\mathrm{E}(h(X)) \neq h(\mathrm{E}(X))$$

Luckily, when the function is convex, we have a notable inequality called *Jensen's inequality*: If X takes values in an interval I and $h : \mathsf{I} \to \mathbb{R}$ is convex on I, then

$$\mathrm{E}(h(X)) \geqslant h(\mathrm{E}(X))$$

B: Variance and Covariance

The *variance* of X, denoted by $\mathrm{var}(X)$, is defined by

$$\mathrm{var}(X) \equiv \mathrm{E}([X - \mathrm{E}(X)]^2) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$$

In other words, the variance measures the average square of the difference between X and its expected value. Variance measures the dispersion, how far a set of numbers is spread out. The standard deviation of a random variable X is the square root of its variance, and is denoted by

$$\sigma_X = \sqrt{\mathrm{var}(X)}$$

unlike the variance, it is expressed in the same units as the data.

The *covariance* of any two random variables X and Y, denoted by $\mathrm{cov}(X, Y)$, is defined by

$$\mathrm{cov}(X,Y) \equiv \mathrm{E}([X - \mathrm{E}(X)][Y - \mathrm{E}(Y)]) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)$$

If, in addition, $\sigma_X > 0$ and $\sigma_Y > 0$, the Pearson *correlation coefficient* of X and Y is defined as

$$\rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

It can be shown that

$$-1 \leqslant \rho_{XY} \leqslant 1 \tag{2.20}$$

In fact, the inequalities in (2.20) is an application of the Cauchy-Schwarz inequality

$$[\mathrm{E}(XY)]^2 \leqslant \mathrm{E}(X^2)\,\mathrm{E}(Y^2)$$

Remark: We have $\rho_{XY} = \pm 1 \iff Y = c \pm \frac{\sigma_Y}{\sigma_X} X$ for some constant c. However, when Y is a nonlinear function of X, the value of ρ may give no hint on the relationship. For example, let $X \sim \mathrm{N}(0, 1)$ and $Y = \pm X^3$, then $\mathrm{cov}(X, Y) = \pm 3$ and $\mathrm{var}(Y) = \mathrm{E}(X^6) = 15$, we have $\rho_{XY} = \pm\frac{3}{\sqrt{15}} = \pm 0.774$. Furthermore, if X has a symmetric PDF and $Y = X^{2n}$ for some natural number n, then

$$\mathrm{cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = \mathrm{E}(X^{2n+1}) - 0 = 0$$

there is $\rho_{XY} = 0$ for any n. In particular, $\rho_{XY} = 0$ when $Y = X^2$.

C: Moment-generating Function

The moment-generating function of a random variable X is

$$M_X(t) \equiv \mathrm{E}(e^{tX}) \qquad -\infty < x < +\infty$$

wherever this expectation exists. Note that $M_X(0)$ always exists and is equal to 1. The derivatives of the moment generating function at 0 determine the moments of the random variable (hence the name)

$$M_X^{(n)}(0) = \mathrm{E}(X^n)$$

Differentiating $M_X(t)$ n times with respect to t and setting $t = 0$ we hence obtain the nth moment about the origin. A key problem with moment-generating functions is that moments and the moment-generating function may not exist[3]. By contrast, the characteristic function ($\mathrm{E}(e^{\mathrm{i}tX})$, where i is the imaginary unit) always exists, and thus may be used instead.

D: Conditional Expectation

If X and Y are jointly discrete random variables, we compute the *conditional expectation* of Y given that $X = x$, for all values of x such that $\mathrm{P}(X = x) > 0$, by

$$\mathrm{E}(Y \mid X = x) = \sum_i y_i \mathrm{P}(Y = y_i \mid X = x) \tag{2.21}$$

If X and Y are jointly continuous with conditional probability density $f(y \mid x)$, the *conditional expectation* of Y given that $X = x$, is computed by

$$\mathrm{E}(Y \mid X = x) = \int_{-\infty}^{+\infty} y f(y \mid x)\, \mathrm{d}y \tag{2.22}$$

Clearly, $\mathrm{E}(Y \mid X = x)$ is a function of x. Let's record this function by

$$r(x) = \mathrm{E}(Y \mid X = x)$$

then, $r(X)$ is a random variable, and this random variable is chronically written by

$$\mathrm{E}(Y \mid X) \equiv r(X)$$

and called the conditional expectation of Y given X (see Chapter 3). An extremely important property of conditional expectations is given by

$$\mathrm{E}(Y) = \mathrm{E}(\mathrm{E}(Y \mid X)) \tag{2.23}$$

which is known as the *law of total expectation*, the *law of iterated expectations*, or the *tower rule*.

2.2.4 Probability Distribution

We give the "cheat sheet" of some important distribution used in asset pricing: Binomial distribution gives the probability of upward movements in binomial model, and lognormal random variables are used in modelling asset price.

[3] The lognormal random variable has the moments of any order, but does not have a moment generating function.

A: Bernoulli Distribution

The Bernoulli distribution is a discrete probability distribution, which takes value 1 with success probability p and value 0 with failure probability $1-p$. So if X is a random variable with this distribution, we have:

$$\mathrm{P}(X=1)=p \qquad \mathrm{P}(X=0)=1-p$$

The probability mass function f of this distribution is

$$f(k;p)=\begin{cases} p & k=1 \\ 1-p & k=0 \end{cases}$$

This can also be expressed as

$$f(k;p)=p^k(1-p)^{1-k} \qquad k\in\{0,1\}$$

The expected value and variance of a Bernoulli random variable X are

$$\mathrm{E}(X)=p\cdot 1+(1-p)\cdot 0=p$$
$$\mathrm{var}(X)=\mathrm{E}\left(X^2\right)-[\mathrm{E}(X)]^2=p-p^2=p(1-p)$$

B: Geometric Distribution

The geometric distribution is useful to model the number of Bernoulli trials needed to get one success, or the number of failures before the first success. Suppose that independent Bernoulli trials with success probability p, are performed until a success occurs. If we let X equal the number of trials required, then

$$\mathrm{P}(X=k)=f(k;p)=(1-p)^{k-1}p \qquad k=1,2,\cdots$$

The expected value and variance of a geometrically distributed random variable X are

$$\mathrm{E}(X)=1/p$$
$$\mathrm{var}(X)=(1-p)/p^2$$

C: Binomial Distribution

The probability of getting exactly k successes in N independent Bernoulli trials is given by the probability mass function:

$$\mathrm{P}(X=k)=f(k;N,p)=\binom{N}{k}p^k(1-p)^{N-k} \qquad k=0,1,2,\cdots,N$$

where the N choose k

$$\binom{N}{k}=\frac{N!}{k!(N-k)!}$$

is the binomial coefficient (hence the name of the distribution). The formula can be understood as follows: we want k successes and $N-k$ failures. However, the k successes can occur anywhere among the N trials, and there are $\binom{N}{k}$ different ways of distributing k successes in a sequence of N trials.

We write $X \sim \mathrm{B}(N,p)$ if the random variable X follows the binomial distribution with parameters N and p. By $X = \sum_i^N X_i$, where X_i is iid Bernoulli with success probability p, we have

$$\mathrm{E}(X) = Np$$
$$\mathrm{var}(X) = Np(1-p)$$

D: Normal distribution

The probability density function of a normal random variable X is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The parameter μ in this definition is the mean of the distribution, and the parameter σ is its standard deviation. Thus, a random variable follows normal distribution is denoted by $X \sim \mathrm{N}(\mu, \sigma^2)$, and

$$\mathrm{E}(X) = \mu \qquad \mathrm{var}(X) = \sigma^2$$

Let

$$Z = \frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1)$$

Then, Z is a normal random variable having mean 0 and variance 1, which is called a *standard normal random variable*. The probability density function of Z is the *standard normal probability density function*

$$\phi(z) \equiv \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

The function

$$N(z) \equiv \mathrm{P}(Z \leqslant z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\,\mathrm{d}x$$

is called the *standard normal distribution function*. Probabilities for $Z \geqslant z$ can be obtained by using the symmetry of the standard normal density about 0 to conclude that

$$\mathrm{P}(Z \geqslant z) = \mathrm{P}(Z \leqslant -z)$$

or equivalently

$$1 - N(z) = N(-z) \tag{2.24}$$

Consequently (or replacing z by $-z$), $-Z \sim Z$, i.e., $-Z$ and Z are identical distributed

$$\mathrm{P}(Z \leqslant z) = N(z) = 1 - N(-z) = \mathrm{P}(Z \geqslant -z) = \mathrm{P}(-Z \leqslant z)$$

If $\log(Y) \sim \mathrm{N}(\mu, \sigma^2)$ is a normal random variable, then the random variable $Y = e^X$ is said to be a *lognormal* random variable, denoted by $Y \sim \mathrm{LN}\left(\mu, \sigma^2\right)$, with

$$\mathrm{E}(Y) = e^{\mu+\sigma^2/2} \tag{2.25}$$
$$\mathrm{var}(Y) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$$

thus

$$Y \sim \mathrm{LN}\left(\mu, \sigma^2\right) \iff \ln(Y) \sim \mathrm{N}\left(\mu, \sigma^2\right)$$

The normal distribution is immensely useful because of the central limit theorem, which states that, under mild conditions, the mean of many random variables independently drawn from the same distribution is distributed approximately normally, irrespective of the form of the original distribution.

E: Bivariate Normal

Random variables X and Y follow a bivariate normal distribution, denoted by

$$(X, Y) \sim \mathrm{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$$

when they have joint probability density function

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{t}{2(1-\rho^2)}\right)$$

where

$$t = \frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}$$

and ρ is the correlation coefficient between X and Y. The followings are some important properties:

- The marginal distributions are

$$X \sim \mathrm{N}(\mu_X, \sigma_X^2)$$
$$Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$$

- The conditional distribution of Y given that $X = x$ is (the distribution of *conditional random variable* $Y_{|X=x}$, defined in Eq 3.6)

$$Y_{|X=x} \sim \mathrm{N}(\mu_Y + (x-\mu_X)\rho\sigma_Y/\sigma_X,\ \sigma_Y^2(1-\rho^2)) \tag{2.26}$$

- The conditional expectation (regression function of Y on X)

$$\mathrm{E}(Y \mid X) = \mu_Y + (X-\mu_X)\rho\sigma_Y/\sigma_X = a + bX \tag{2.27}$$

with

$$b = \rho\sigma_Y/\sigma_X = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)}$$
$$a = \mu_Y - \mu_X\rho\sigma_Y/\sigma_X = \mu_Y - \mu_X b$$

- Suppose two random variables X and Y are jointly normally distributed, if X and Y are uncorrelated, then they are independent.

Remark: The multivariate normal distribution is written by $\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$, with mean vector $\boldsymbol{\mu} = \mathrm{E}(\mathbf{x})$ and variance matrix $\mathrm{var}(\mathbf{x}) = \mathbf{V}$. If $\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$, then $\mathbf{Ax} \sim \mathrm{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{AVA}')$ for any conforming matrix $\mathbf{A}$. Thus, let $(X, Y) \sim \mathrm{N}\left(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho\right)$ and $Z = Y - X\rho\sigma_Y/\sigma_X$. Then

$$(X, Z) \sim \mathrm{N}\left(\mu_X,\ \mu_Y - \mu_X\rho\sigma_Y/\sigma_X,\ \sigma_X^2,\ \sigma_Y^2\left(1-\rho^2\right),\ 0\right)$$

We see that Z is a function of X, but Z is independent (mere concern of probability) of X.

2.2.5 Central Limit Theorem

The ubiquity of normal random variables is explained by the central limit theorem, probably the most important theoretical result in probability.

Theorem 2.9 (Lindeberg-Lévy CLT): *Suppose* $\{X_i\}_{i=1}^{\infty}$ *is a sequence of i.i.d. random variables with* $\mathrm{E}(X_i) = \mu$ *and* $\mathrm{var}(X_i) = \sigma^2 < \infty$. *Let* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, *then as* n *approaches infinity, the*

asymptotic distribution (converge in distribution) of random variables $\frac{1}{\sigma}\sqrt{n} \cdot \left(\bar{X}_n - \mu\right)$ *is standard normal* $\mathrm{N}(0,1)$

$$\frac{1}{\sigma}\sqrt{n} \cdot \left(\bar{X}_n - \mu\right) \overset{a}{\sim} \mathrm{N}(0,1)$$

The strong *law of large numbers* (LLN) states that the sample average converges almost surely to the expected value, $\bar{X}_n \to \mu$ a.s. We see that the CLT explores the distribution of the standardized version of $\bar{X}_n$ (demeaned and scaled).

§ 2.3 Change of Measure

A measure is a generalization of the concepts of length, area, and volume. Intuitively, a measure is a meaningful "size". A distance can be measure by kilometer or mile, 3.2 kilometers have the same length with 2 miles. A probability is a measurement of odds, we can assign different values to the same event just like the change of units of length from kilometer to mile. However, there is a bit more complex in probability, for a probability measure must assign value 1 to the entire probability space.

2.3.1 Probability as Measure

We know that a probability measure representing chance or likelihood is a measure with total measure one. In Definition 2.2, any function $\mathrm{P} : \mathcal{F} \to [0, 1]$ satisfying those three conditions is called a probability measure. Thus, if we use different method to represent chance or likelihood, such as in the coming Example 2.3.1, by using length of subinterval in $[0, 1]$, or area of a region in a unit circle, we may have different probabilities assign to the same event.

Example 2.3.1: Take a fixed circle of radius 1, and draw at random a chord. What is the probability that the length of the chord will be greater than the side of the equilateral triangle inscribed in that circle?

The length of the side of the inscribed equilateral triangle is $\sqrt{3}$, thus, we should compute the probability that the random chord has length greater than $\sqrt{3}$. To give meaning to "random chord", we shall reformulate the problem in the following three distinct ways:

1. Take all random lines through a fixed point C on the bottom edge of the circle. Apart from the horizontal tangent (zero probability), all such lines will intersect the circle. For the length of the chord to be greater than $\sqrt{3}$, the line must lie between $60°$ and $120°$, within an angle of $120° - 60° = 60°$ out of a total $180°$. Hence, the probability is $1/3$.
2. Take all random lines perpendicular to a fixed radius. The length of the chord is greater than $\sqrt{3}$ when the point of intersection lies on the inner half of the radius. Hence, the probability is $1/2$.
3. For a chord to have length greater than $\sqrt{3}$, its center must lie at a distance less than $1/2$ from point O, the center of the circle. The area of a circle of radius $1/2$ is a quarter of that of the original circle. Hence, the probability is $1/4$.

Note that random experiments could be performed in such ways that $1/3$, $1/2$ or $1/4$ would be the correct probability. Since each solution is based upon a different assumption about probability measures: measured by angle, length and area respectively.

The probability is $p = 1/2$, if the probability is measured by length, and the probability is $q = 1/4$, if the probability is measured by area. The change of measure from length to area gives arise to the distinct probability from $p = 1/2$ to $q = 1/4$. Let us take all random chords perpendicular to the fixed radius

OC, and random variable X be the distance from the center of the chord to the center of the circle. Then our sample space is $\Omega = [0, 1]$, and $\{X \leqslant 1/2\}$ is the event that a random chord has length greater than $\sqrt{3}$. When the probability is measured by length, the CDF is a uniform distribution over $[0, 1]$

$$P(x) = \mathrm{P}(X \leqslant x) = \frac{x - 0}{1 - 0} = x \qquad 0 \leqslant x \leqslant 1$$

and thus

$$\mathrm{P}(X \leqslant 1/2) = 1/2$$

When the probability is measured by area, the CDF is

$$Q(x) = \mathrm{Q}(X \leqslant x) = \frac{\pi \cdot x^2}{\pi \cdot 1^2} = x^2 \qquad 0 \leqslant x \leqslant 1$$

and thus

$$\mathrm{Q}(X \leqslant 1/2) = (1/2)^2 = 1/4$$

2.3.2 Transforming the Probability Measure

In example 2.3.1, probability measures P and Q disagree on the probability distributions. However, what is the linkage between the P measure and Q measure? How to transform the underlying probability distributions? To address these problems, we start by the simplest case, a finite sample space.

A: Discrete Sample Space

Definition 2.10: Let Ω be a discrete sample space on which we have two probability measures P and Q. Assume that P and Q both give positive probability to every element of Ω, let

$$G(w) = \frac{\mathrm{Q}(w)}{\mathrm{P}(w)} \tag{2.28}$$

the random variable G is called the *Radon-Nikodým derivative* of Q with respect to P.

In a finite sample space Ω, it is really a *quotient* rather than a derivative. Probability measures P and Q give different weights to the simple events.

Theorem 2.11: *Let* P *and* Q *be probability measures on a finite sample space* Ω*, assume that* $\mathrm{P}(w) > 0$ *and* $\mathrm{Q}(w) > 0$ *for every* $w \in \Omega$*, and define the random variable* G *by Eq (2.28). Then we have*

(a) $\mathrm{P}(G > 0) = 1$

(b) $\mathrm{E}(G) = 1$*, where* $\mathrm{E}(\cdot) = \mathrm{E}^P(\cdot)$

(c) For any random variable Y*, the expectation under* Q *is*

$$\mathrm{E}^Q(Y) = \mathrm{E}(GY)$$

and

$$\mathrm{E}(Y) = \mathrm{E}^Q(Y/G)$$

This theorem is illustrated and validated by the following example.

Example 2.3.2: Suppose that we throw a dice one time. Let

$$\mathrm{P}(w) = 1/6 \qquad \mathrm{Q}(w) = \begin{cases} 1/3 & w = 4 \\ 1/6 & w = 1 \\ 1/8 & w \in \{2,3,5,6\} \end{cases}$$

then

$$G(w) = \frac{\mathrm{Q}(w)}{\mathrm{P}(w)} = \begin{cases} 2 & w = 4 \\ 1 & w = 1 \\ 3/4 & w \in \{2,3,5,6\} \end{cases}$$

and

$$\mathrm{E}(G) = 2 \cdot \frac{1}{6} + 1 \cdot \frac{1}{6} + \frac{3}{4} \cdot \frac{4}{6} = 1$$

Besides, if $Y = 1_{w>4}$, then

$$GY = \begin{cases} 0 & w \in \{1,2,3,4\} \\ 3/4 & w \in \{5,6\} \end{cases}$$

thus

$$\mathrm{E}(GY) = 0 \cdot \frac{4}{6} + \frac{3}{4} \cdot \frac{2}{6} = \frac{1}{4}$$

Which is equal to

$$\mathrm{E}^Q(Y) = \mathrm{E}^Q(1_{w>4}) = \mathrm{Q}(w > 4) = \mathrm{Q}(w \in \{5,6\}) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$$

or compute directly by the definition of expectation

$$\mathrm{E}^Q(Y) = \sum_i y_i \mathrm{Q}(Y = x_i) = 0 \cdot \left(\frac{1}{6} + \frac{1}{8} + \frac{1}{8} + \frac{1}{3}\right) + 1 \cdot \left(\frac{1}{8} + \frac{1}{8}\right) = \frac{1}{4}$$

B: Equivalent Probability Measures

Suppose that random variable X has a CDF

$$P(x) = \mathrm{P}(X \leqslant x)$$

under probability measure P. Let $G = g(X) > 0$ be a random variable. If $\mathrm{E}(G) = 1$, we define measure Q by

$$\mathrm{Q}(\mathsf{E}) = \mathrm{E}(G 1_{\mathsf{E}}) \tag{2.29}$$

for any event E (please verify that Q is a probability measure). For any $x \in \mathbb{R}$, let $\mathsf{E} = \{X \leqslant x\}$, then

$$\mathrm{Q}(X \leqslant x) = \mathrm{Q}(\mathsf{E}) = \mathrm{E}(G 1_{X \leqslant x}) = \int_{-\infty}^{+\infty} g(u)\, 1(u \leqslant x)\, \mathrm{d}P(u) = \int_{-\infty}^{x} g(u)\, \mathrm{d}P(u)$$

thus the CDF for X under measure Q is

$$Q(x) = \mathrm{Q}(X \leqslant x) = \int_{-\infty}^{x} g(u)\, \mathrm{d}P(u)$$

Given the probability measure Q defined by Eq (2.29), we find that (proof in appendix)

$$\mathrm{Q}(\mathsf{E}) = 0 \iff \mathrm{P}(\mathsf{E}) = 0 \tag{2.30}$$

If this condition is satisfied, we say Q and P are *equivalent probability measures*, or Q is equivalent to P, denoted by $\mathrm{Q} \sim \mathrm{P}$. Q and P are called "equivalent" because, they assign positive probabilities to the same domains, although the two probability distributions are different. They agree on which simple events are possible; they disagree only on what these positive probabilities are.

For any $x \in \mathbb{R}$, if

$$\int_{-\infty}^{x} \mathrm{d}Q(u) = Q(x) = \int_{-\infty}^{x} g(u)\ \mathrm{d}P(u)$$

we write in differential notation $\mathrm{d}Q = g(X)\,\mathrm{d}P$, or

$$\frac{\mathrm{d}Q}{\mathrm{d}P} = g(X) = G$$

and $G = g(X)$ is called *Radon-Nikodým derivative* of Q with respect to P.

Remark: When refer to probability measure we write $\mathrm{d}Q = G\,\mathrm{d}P$ for $\mathrm{d}Q(w) = g(X(w))\,\mathrm{d}P(w)$. When refer to CDF, we write $\mathrm{d}Q(x) = g(x)\,\mathrm{d}P(x)$.

If $\mathrm{Q} \sim \mathrm{P}$ and $\mathrm{d}Q = G\,\mathrm{d}P$, then for any random variable Y, we have

$$\mathrm{E}^{Q}(Y) = \mathrm{E}(GY)$$

and

$$\mathrm{E}(Y) = \mathrm{E}^{Q}(Y/G)$$

C: Normal Random Variables

Theorem 2.12: *Let* $Z \sim \mathrm{N}(0,1)$ *under* P, *and*

$$G = g(Z) = e^{\mu Z - \mu^2/2}$$

Define measure Q *by Eq (2.29), then* Q *is an equivalent probability measure, and* $Z \sim \mathrm{N}^{Q}(\mu, 1)$ *(*$\mathrm{N}(\mu,1)$ *distribution under* Q*).*

Proof. By Eq (2.25), $\mathrm{E}(G) = 1$, thus Q is an equivalent probability measure. For

$$\mathrm{Q}(Z \leqslant z) = \int_{-\infty}^{z} g(x)\ \mathrm{d}P(x) = \int_{-\infty}^{z} g(x)\,\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{(x-\mu)^2}{2}}\,\mathrm{d}x$$

which states that $Z \sim \mathrm{N}^{Q}(\mu, 1)$. □

What we have just shown is that there exists a function $g(z)$ such that if we multiply a normal probability measure by this function, we get a new probability measure. The resulting random variable is again normal but has a different mean. Change of probability measure is useful in simulation and estimation of probabilities of rare events.

Example 2.3.3 (simulations of rare events): Let $X \sim \mathrm{N}(6,1)$, consider estimation of $\mathrm{P}(X < 0)$ by simulations. This probability is about 10^{-10}. Direct simulation is done by the expression

$$\mathrm{P}(X < 0) \approx \frac{1}{N} \sum_{i=1}^{N} 1_{x_i < 0}$$

Note that in a million runs of $\mathrm{N}(6,1)$, $N = 10^6$, we should expect no values below zero, and the

estimate is 0.

Let $g(X) = e^{-\mu X + \mu^2/2} = e^{-6X+18}$, then under new measure Q, $X \sim \mathrm{N}^Q(0,1)$, and

$$\mathrm{P}(X < 0) = \mathrm{E}(1_{X<0}) = \mathrm{E}^Q\left(\frac{1}{g} 1_{X<0}\right) \approx \frac{1}{N} e^{-18} \sum_{i=1}^{N} e^{6x_i} 1_{x_i<0}$$

Now, about half of the observations of $\mathrm{N}^Q(0,1)$ will be negative, resulting in a more precise estimate, even for a small number of runs.

Remark: If P is the probability such that X is $\mathrm{N}(6,1)$, then $X-6$ has $\mathrm{N}(0,1)$ distribution under P. This is an operation on the values of X. When we change the probability measure P to Q, we leave the values as they are, but assign different probabilities to them (more precisely to sets of outcomes). Under the new measure the same X has $\mathrm{N}(0,1)$ distribution.

D: Independence

Indeed, two events can be independent relative to one probability measure and dependent relative to another. In Example 2.1.2, suppose that we toss 2 biased dice, both with $\mathrm{Q}(w_1 = 4) = \mathrm{Q}(w_2 = 4) = 2/7$ and $1/7$ for other points. Let F denote the event that the first dice equals 4, and S be the event that the sum of the dice equals 7. Form Eq (2.9), we know that $\mathsf{F} \perp \mathsf{S}$ under measure P. However, is F still independent of S under Q? Since

$$\mathsf{F} = \{(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\}$$
$$\mathsf{S} = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$$

We have $\mathrm{Q}(\mathsf{F}) = \frac{2}{7}$ and (the two biased dice are independent under Q)

$$\mathrm{Q}(\mathsf{S}) = \frac{1}{7}\cdot\frac{1}{7} + \frac{1}{7}\cdot\frac{1}{7} + \frac{1}{7}\cdot\frac{2}{7} + \frac{2}{7}\cdot\frac{1}{7} + \frac{1}{7}\cdot\frac{1}{7} + \frac{1}{7}\cdot\frac{1}{7} = \frac{8}{49}$$

For

$$\mathrm{Q}(\mathsf{F})\mathrm{Q}(\mathsf{S}) = \frac{2}{7}\cdot\frac{8}{49} = \frac{2}{49}\cdot\frac{8}{7} > \mathrm{Q}(\mathsf{FS}) = \mathrm{Q}((4,3)) = \frac{2}{7}\cdot\frac{1}{7} = \frac{2}{49}$$

F is not independent of S under Q.

Remark: Independence concerns only the relationship on probability. In a new set of probability measure, the equality in Equation (2.7) may be broken. Independence may fail by way of change of measure, where the value of probability may vary, thus independence is a relative concept.

Conditional independence is defined on a given conditional probability. By conditioning, conditional probability is a special way of change of measure (but not equivalent to original probability measure). Thus, the equality in the joint probability of two events and the product of their probabilities may not be retained. As seen in Example 2.1.7 and 2.1.8, conditional independence neither implies nor is implied by independence.

E: Stochastic Dominance

Random variable X is *first-order stochastic dominance*[4] (FSD) over Y, denoted by $X \succsim Y$, if

$$\mathrm{P}(X > x) \geqslant \mathrm{P}(Y > x)$$

for all x, with strict inequality at some x. In finance, the payoff X gives at least as high a probability of receiving larger than x as does Y, and for some x, X gives a higher probability of receiving greater than x. In terms of the cumulative distribution functions, $X \succsim Y$ means that $F_X(x) \leqslant F_Y(x)$ for all x, and $F_X(x) < F_Y(x)$ at some x (X and Y are not identically distributed). We see that

$$X \gtrsim Y \implies X \succsim Y$$

However, $X \succsim Y \not\Longrightarrow X \gtrsim Y$, for the FSD cares solely the CDF, not state by state.

Example 2.3.4: In tossing a fair coin, let $X(H) = 4$, $X(L) = 2$, $Y(H) = 1$, and $Y(L) = 3$, there is $X \succsim Y$, but in state L, $X(L) < Y(L)$.

If $X \gtrsim Y$, because of the state by state dominance, we have

$$\mathrm{E}(X) > \mathrm{E}(Y) \text{ and } \mathrm{E}^Q(X) > \mathrm{E}^Q(Y)$$

the statewise dominance, $X \gtrsim Y$, is irrelevant to probability measure. However, for the first-order stochastic dominance

$$X \succsim Y \implies \mathrm{E}(X) > \mathrm{E}(Y)$$

but $X \succsim Y$ under P does not imply $\mathrm{E}^Q(X) > \mathrm{E}^Q(Y)$. FSD is relevant to probability measure, $X \succsim Y$ in P world does not imply that $X \succsim Y$ in Q world. In Example 2.3.4, if $\mathrm{Q}(H) = 0.1$, then the CDF under Q has

$$F_X^Q(2) = 0.9 > F_Y^Q(2) = 0.1$$

Thus, X is not FSD over Y under Q. Furthermore, there is $\mathrm{E}(X) = 3 > \mathrm{E}(Y) = 2$ under P, but under Q

$$\mathrm{E}^Q(X) = 2.2 < \mathrm{E}^Q(Y) = 2.8$$

2.3.3 Risk-Neutral and Forward Probability Measure

In Finance, in addition to the actual probability measure P, there are two probability measures that merit our attention

1. The risk-neutral probability measure Q
2. The forward probability measure O

The actual probability measure, is the one that generates the data, and is used in the empirical estimation of model parameters. However, the measure Q and O are employed for the easier computation of derivatives' price.

[4]Note that some textbooks define first-order stochastic dominance by $F_X(x) \leqslant F_Y(x)$ for all x, and denoted by $X \succeq Y$, where X and Y following the identical distribution are not excluded.

Let B_t be the value of risk-free bond at future time t, define

$$G = \Psi B_t / B_0 > 0$$

where Ψ is SDF defined in Eq (1.17), then $\mathrm{E}(G) = \mathrm{E}(\Psi B_t)/B_0 = \wp(B_t)/B_0 = 1$. We can define an equivalent probability measure Q by $\mathrm{d}Q = G\,\mathrm{d}P$, and under probability measure Q, for any payoff X at future time t

$$\mathrm{E}^Q\left(\frac{X}{B_t}\right) = \mathrm{E}\left(G \cdot \frac{X}{B_t}\right) = \mathrm{E}\left(\Psi \frac{B_t}{B_0} \cdot \frac{X}{B_t}\right) = \frac{1}{B_0}\,\mathrm{E}\left(\Psi X\right) = \frac{X_0}{B_0} \tag{2.31}$$

Equation (2.31) is a generalization of the risk-neutral pricing formula (1.19), which simplifies the computation of asset price. Probability measure Q is called *risk-neutral probability measure*. In pricing of financial derivatives, if the risk-free interest rate is a constant r, $B_t = e^{rt}$, the computation of

$$X_0 = B_0\,\mathrm{E}\left(\frac{X}{B_t}\right) = \frac{B_0}{B_t}\,\mathrm{E}^Q(X) = e^{-rt}\,\mathrm{E}^Q(X)$$

is often easier than that of $X_0 = \mathrm{E}\left(\Psi X\right)$, for the latter requires the modeling of the joint distribution of Ψ and X, which is more involved. Besides, it is very interesting that the expected growth rate of every risky asset equals risk-free rate, investors are risk-neutral in Q world.

Let $D_{0,t}$ be the price at time 0 for a discount bond with payoff 1 at future time t, then $\Psi/D_{0,t} > 0$, and $\mathrm{E}(\Psi/D_{0,t}) = \mathrm{E}(\Psi)/D_{0,t} = \wp(1)/D_{0,t} = 1$. We can define an equivalent probability measure O by $\mathrm{d}O = (\Psi/D_{0,t})\,\mathrm{d}P$, and under probability measure O, for any payoff X at future time t

$$\mathrm{E}^O\left(\frac{X}{D_{t,t}}\right) = \mathrm{E}\left(\frac{\Psi}{D_{0,t}} \cdot \frac{X}{D_{t,t}}\right) = \frac{1}{D_{0,t}}\,\mathrm{E}\left(\Psi X\right) = \frac{X_0}{D_{0,t}}$$

Which means

$$X_0 = D_{0,t}\,\mathrm{E}^O(X) \tag{2.32}$$

Probability measure O is called *forward probability measure*. If the world of interest rate is non-stochastic, then $B_t = 1/D_{0,t}$ and

$$\mathrm{d}Q = \Psi B_t \mathrm{d}P = \Psi B_t \cdot (D_{0,t}/\Psi)\mathrm{d}O = B_t D_{0,t}\mathrm{d}O = \mathrm{d}O$$

the forward probability measure coincides with the risk-neutral probability measure. However, when the interest rates are random, Equation (2.32) is more convenient than Equation (2.31), because of its using the discount factor implied by the market rather than the bivariate dynamics.

If we have to calculate an expectation (for example, the pricing function in Eq 1.17), and if this expectation is easier to calculate with an equivalent measure, then it may be worth switching probability measures. Although the new measure may not be the one that governs the true states of nature, it is only an auxiliary or a shortcut. After all, the purpose is not to make a statement about the odds of various states of nature. The purpose is to calculate a quantity in a convenient fashion.

§ 2.4 Exercise

2.1 Two cards are randomly selected from a deck of 52 playing cards. What is the conditional probability they are both aces, given that they are of different suits?

2.2 Let $\Omega = [0,1]$ with uniform probability, consider the following closed intervals: $\mathsf{E} = [\frac{1}{8}, \frac{5}{8}]$, $\mathsf{F} = [\frac{1}{4}, \frac{3}{4}]$, and $\mathsf{G} = [\frac{1}{2}, 1]$. Show that $\mathrm{P}(\mathsf{EFG}) = \mathrm{P}(\mathsf{E})\mathrm{P}(\mathsf{F})\mathrm{P}(\mathsf{G})$, but $\mathrm{P}(\mathsf{EF}) \neq \mathrm{P}(\mathsf{E})\mathrm{P}(\mathsf{F})$.

2.3 Prove Equation (2.4) and (2.6).

2.4 Let event $\mathsf{F} = \mathsf{G} + \mathsf{H}$, and $\mathsf{GH} = \emptyset$. Given event E, with $\mathrm{P}(\mathsf{EG}) > 0$, $\mathrm{P}(\mathsf{EH}) > 0$, show that $\mathrm{P}(\mathsf{E} \,|\, \mathsf{G}) + \mathrm{P}(\mathsf{E} \,|\, \mathsf{H}) > \mathrm{P}(\mathsf{E} \,|\, \mathsf{F})$

2.5 Prove Equation (2.8).

2.6 Prove Equation (2.12) and (2.13).

2.7 In Example 2.1.2, why the event that the sum of the dice equals seven is independent of the outcome on the first dice, please explain intuitively.

2.8 In Example 2.1.8, show that $\mathrm{P}(\mathsf{E} \,|\, \mathsf{FG}) = 1$, and $\mathrm{P}(\mathsf{E} \,|\, \mathsf{G}) = 1/2$.

2.9 In Example 2.1.9, define a suitable sample space, write out the events E, F and G, and show that $\mathrm{P}(\mathsf{F} \,|\, \mathsf{E}) > \mathrm{P}(\mathsf{F})$.

2.10 In Example 2.1.10, show that $\mathrm{P}(\mathsf{E}_1) = 0$.

2.11 Prove Boole's inequality

$$\mathrm{P}\left(\sum_{i=1}^{\infty} \mathsf{E}_i\right) \leqslant \sum_{i=1}^{\infty} \mathrm{P}(\mathsf{E}_i)$$

2.12 Let X be the number of heads in Example 2.1.1, and Y be the number of reversals, show that Y is not $\sigma(X)$ measurable.

2.13 For random variables X, Y and Z, if $Y \perp (X, Z)$, show that $Y \perp X$, $Y \perp Z$, $X \perp Y \,|\, Z$, and $Y \perp Z \,|\, X$.

2.14 Given $0 < u < v$, then for random variable X, if $\mathrm{E}(|X|^v) < \infty$, then $\mathrm{E}(|X|^u) < \infty$.

2.15 Prove Equation (2.18).

2.16 Prove Equation (2.20).

2.17 Show that $|\mathrm{E}(X)| \leqslant \mathrm{E}(|X|)$.

2.18 Prove the Cauchy-Schwarz inequality

$$[\mathrm{E}(XY)]^2 \leqslant \mathrm{E}(X^2)\,\mathrm{E}(Y^2)$$

2.19 Show that if X is a nonnegative random variable, then $\mathrm{E}(X) = \int_0^{+\infty} \mathrm{P}(X > x)\,\mathrm{d}x$

2.20 Show that $\int_{-\infty}^{x} 1_{X \leqslant z}\,\mathrm{d}z = (x - X)^+$ and

$$\mathrm{E}((x - X)^+) = \int_{-\infty}^{x} F_X(z)\,\mathrm{d}z$$

2.21 Prove Equation (2.24).

2.22 Prove Equation (2.25).

2.23 If $X \sim \mathrm{N}(\mu, \sigma^2)$, given number a, let $t = (a - \mu)/\sigma$, show that $\mathrm{E}(X 1_{X \geqslant a}) = \mu N(-t) + \sigma\phi(t)$, and $\mathrm{E}(X^2 1_{X \geqslant a}) = (\mu^2 + \sigma^2)N(-t) + \sigma(a + \mu)\phi(t)$.

2.24 Prove Equation (2.26) and (2.27). Then compute $\mathrm{E}(XY)$, and show that ρ is the correlation coefficient between X and Y.

2.25 In Example 2.3.2, if $Y = 1_{w>3}$, very that $\mathrm{E}^Q(Y) = \mathrm{E}(GY)$.

2.26 Verify that Q in Equation (2.29) is a probability measure.

2.27 Removal of the mean: If Z has $\mathrm{N}(\mu, 1)$ distribution under P. Find the Radon-Nikodým derivative $g(Z)$, of Q with respect to P, such that Z has $\mathrm{N}(0, 1)$ distribution under Q.

2.28 Change mean and variance: If Z has $\mathrm{N}(\mu, \sigma^2)$ distribution under P. Find the Radon-Nikodým derivative $g(Z)$, of Q with respect to P, such that Z has $\mathrm{N}(\mu_*, \sigma_*^2)$ distribution under Q.

§ 2.5 Appendix

For those who are interest in the analysis of finance, probability textbooks at the level like Walpole et al. (2012), Ross (2010) or Schwarzlander (2011) are not adequate, textbooks like Dineen (2013), Kopp et al. (2013) or Rosenthal (2006) are at the very beginning and a must. We need modern probability theory, where concepts such as Borel sets and Lebesgue integral are essential.

2.5.1 Jensen's Inequality

A function $f(x)$ of one real variable is said to be *convex* if, for all x, y and $0 \leqslant l \leqslant 1$

$$f(lx + (1-l)y) \leqslant lf(x) + (1-l)f(y)$$

A function is convex if the line segment between any two points on the graph of the function lies above or on the graph. A convex function f defined on some open interval I is continuous on I. Let

$$k(x,y) = \frac{f(x) - f(y)}{x - y}$$

$k(x,y) = k(y,x)$ is symmetric. We have the following useful characterization of convexity.

Proposition 2.13: A function $f(x)$ is convex if and only if $k(x,y)$ is increasing in x, for every fixed y.

Proof. $\Longrightarrow$: For all x, y and $0 \leqslant l \leqslant 1$, without lost of generality, let $x < y$, and $0 < l < 1$, define $t = lx + (1-l)y$, then $k(t,x) \leqslant k(y,x)$ shows that $f(lx + (1-l)y) \leqslant lf(x) + (1-l)f(y)$, thus $f(x)$ is convex.

$\Longleftarrow$: Let f be a convex function on an interval I in $\mathbb{R}$, and $x < t < y$ in I, we will show that

$$k(t,x) \leqslant k(y,x) \leqslant k(y,t)$$

Define $l = \frac{y-t}{y-x} \in (0,1)$, then $t = lx + (1-l)y$ and

$$(y-x)[f(t) - f(x)] \leqslant (y-x)[lf(x) + (1-l)f(y) - f(x)] = (t-x)[f(y) - f(x)]$$

which results in $k(t,x) \leqslant k(y,x)$, say, $k(x,y)$ is increasing in x, for every fixed $y < x$. Similarly

$$(y-x)[f(y) - f(t)] \geqslant (y-x)[f(y) - lf(x) - (1-l)f(y)] = (y-t)[f(y) - f(x)]$$

which gives $k(x,y) \leqslant k(t,y)$, say, $k(x,y)$ is increasing in x, for every fixed $y > x$. □

A: Supporting Line

A differentiable function is convex on an interval if and only if the function lies above all of its tangents

$$f(x) \geqslant f(t) + f'(t)(x-t)$$

for all x and t in the interval. The tangent line is a special case of supporting line: Let f be a function defined on an interval $\mathsf{I} < \mathbb{R}$, given $t \in \mathsf{I}$, if there exist a number k (that may depend on t), such that

$$f(x) \geqslant f(t) + k(x-t)$$

for all $x \in \mathsf{I}$. The graph of

$$y = f(t) + k(x - t)$$

is called a *supporting line* for f at t. A supporting line is a straight line passing through the point and situating under the graph.

Proposition 2.14: A function is convex if and only if it has at least one supporting line at each point in the domain.

Proof. $\Longrightarrow$: For any x, y in I and $0 \leqslant l \leqslant 1$, then $t = lx + (1-l)y$ is in I. Let the slope of supporting line for f at t be k, then

$$f(t) + k(x - t) \leqslant f(x)$$

$$f(t) + k(y - t) \leqslant f(y)$$

thus $lf(x) + (1-l)f(y) \geqslant f(t) = f(lx + (1-l)y)$, f is convex.

$\Longleftarrow$: For a convex function f defined on an interval $\mathsf{I} = (a, b)$, given $t \in \mathsf{I}$, then $k(h, t)$ is decreasing as $h \to t + 0$ and bounded below by $k((t + a)/2, t)$, thus

$$f'(t + 0) = \lim_{h \to t+0} \frac{f(h) - f(t)}{h - t} = \lim_{h \to t+0} k(h, t)$$

exist. Similarly, $f'(t - 0) = \lim_{h \to t-0} k(h, t)$ exist. We see that both left and right derivatives exist and

$$k(x, t) \leqslant f'(t - 0) \leqslant f'(t + 0) \leqslant k(y, t)$$

for any $x < t < y$ in I (by $k(x, y)$ is increasing, and $a_n \leqslant b_n \implies \lim a_n \leqslant \lim b_n$). Thus, for any k with $f'(t - 0) \leqslant k \leqslant f'(t + 0)$, for example, $k = (f'(t - 0) + f'(t + 0))/2$, we have

$$f(t) + k(x - t) \leqslant f(x)$$

for all x in I.

If the interval is closed, $\mathsf{I} = [a, b]$, the slopes of supporting line at end points maybe $\pm\infty$. For example, the lower part of unit circle, the slopes of supporting line are $\pm\infty$ at end points ± 1. □

B: Proof to Jensen's Inequality

For $t = \mathrm{E}(X) \in \mathsf{I}$, let $y = h(t) + k(x - t)$ be a supporting line for h at t, then

$$h(X) \geqslant h(t) + k(X - t)$$

Taking expected values through the inequality gives

$$\mathrm{E}(h(X)) \geqslant h(t) = h(\mathrm{E}(X))$$

2.5.2 Sets

We reserve some symbols for special sets:

- $\mathbb{P}$: denoting the set of all primes, $\mathbb{P} = \{2, 3, 5, 7, 11, 13, 17, ...\}$

- $\mathbb{N}$: denoting the set of all natural numbers, $\mathbb{N} = \{1, 2, 3, ...\}$, and $\mathbb{N}_0 = \mathbb{N} + \{0\}$
- $\mathbb{Z}$: denoting the set of all integers (whether positive, negative or zero)
- $\mathbb{Q}$: denoting the set of all rational numbers
- $\mathbb{R}$: denoting the set of all real numbers
- $\mathbb{C}$: denoting the set of all complex numbers, $\mathbb{C} = \{a + b\mathrm{i} : a, b \in \mathbb{R}\}$, where i is the imaginary unit, with $\mathrm{i}^2 = -1$ (see 7.6.1 for more detail)

A: Basic Operations

The basic operations for sets are union, intersection and complement.

- Union: Two sets can be "added" together. The union of A and B, denoted by $\mathsf{A} + \mathsf{B}$, or $\mathsf{A} \cup \mathsf{B}$, is the set of all elements that are in either A or B. In symbols
$$\mathsf{A} + \mathsf{B} = \{x : x \in \mathsf{A} \text{ or } x \in \mathsf{B}\}$$
- Intersection: The intersection of A and B, denoted by AB, or $\mathsf{A} \cap \mathsf{B}$, is the set of elements that are in both A and B. Formally
$$\mathsf{AB} = \{x : x \in \mathsf{A} \text{ and } x \in \mathsf{B}\}$$
If $\mathsf{AB} = \emptyset$, then A and B are said to be disjoint.
- Complement: Two sets can also be "subtracted". The relative complement of A in B, denoted by $\mathsf{B}\backslash\mathsf{A}$, or $\mathsf{B} - \mathsf{A}$, is the set of elements in B, but not in A. Formally
$$\mathsf{B} - \mathsf{A} = \{x : x \notin \mathsf{A} \text{ and } x \in \mathsf{B}\}$$
If a universe U is defined, then the relative complement of A in U is called the absolute complement (or simply complement) of A, and is denoted by $\mathsf{A}' = \mathsf{U} - \mathsf{A}$ or sometimes A^{c}.

B: Order of Precedence

Note that
$$\mathsf{AB}' = \mathsf{A}\backslash\mathsf{B} = \mathsf{A} - \mathsf{B} \qquad \mathsf{A} \cap \mathsf{B} = \mathsf{AB} \qquad \mathsf{A} \cup \mathsf{B} = \mathsf{A} + \mathsf{B}$$
and thus $\sum_{i=1}^{\infty} \mathsf{A}_i = \bigcup_{i=1}^{\infty} \mathsf{A}_i$. For clarity, we observe the following order of precedence: prime (complement), backslash (complement), intersection, symmetric difference ($\mathsf{A} \veebar \mathsf{B} = \{x : (x \in \mathsf{A}) \veebar (x \in \mathsf{B})\}$), minus (complement), and plus (union). For example
$$\mathsf{A} - \mathsf{BC} = \mathsf{A}\,(\mathsf{BC})'$$
and $\mathsf{A} + \mathsf{B} - \mathsf{C} \neq (\mathsf{A} + \mathsf{B})\mathsf{C}' = \mathsf{AC}' + \mathsf{BC}'$, but $\mathsf{A} + \mathsf{B} - \mathsf{C} = \mathsf{A} + \mathsf{BC}'$.

C: Laws of Set Algebra

The union and intersection of sets may be seen as analogous to the addition and multiplication of real numbers. Like addition and multiplication, the operations of union and intersection are commutative
$$\mathsf{A} + \mathsf{B} = \mathsf{B} + \mathsf{A} \qquad \mathsf{AB} = \mathsf{BA}$$

and associative

$$(\mathsf{A}+\mathsf{B})+\mathsf{C}=\mathsf{A}+(\mathsf{B}+\mathsf{C}) \qquad (\mathsf{AB})\mathsf{C}=\mathsf{A}(\mathsf{BC})$$

For distributive laws, intersection distributes over union

$$\mathsf{A}(\mathsf{B}+\mathsf{C})=\mathsf{AB}+\mathsf{AC}$$

However, unlike addition and multiplication, union also distributes over intersection

$$\mathsf{A}+(\mathsf{BC})=(\mathsf{A}+\mathsf{B})(\mathsf{A}+\mathsf{C})$$

For the absorption (big eat small) in the union of sets: $\mathsf{A}+\mathsf{B}=\mathsf{A}$ if B is a subset of A ($\mathsf{B}\leqslant\mathsf{A}$ or $\mathsf{B}\subseteq\mathsf{A}$), thus $\mathsf{AA}+\mathsf{AC}+\mathsf{BA}=\mathsf{A}$.

D: Sequence of Sets

For a sequence of real number, we know that

$$a_n=\frac{1}{n}>0 \implies \lim_{n\to\infty} a_n=0$$

However

$$\mathsf{B}_n=\left[\frac{1}{n},3-\frac{1}{n}\right] \implies \mathsf{B}=\lim_{n\to\infty}\mathsf{B}_n=\bigcup_{n=1}^{\infty}\mathsf{B}_n=(0,3)$$

Note that $0\notin\mathsf{B}$. Otherwise, if $0\in\mathsf{B}$, there is at least one n such that $0\in\mathsf{B}_n$, but $0\notin\mathsf{B}_n$ for any n.

2.5.3 Borel Set

By definition, every open interval (a,b) is a Borel set, and the union, intersection or complement of Borel sets are Borel. Hence, virtually every subset of $\mathbb{R}$ is a Borel set. For example, for every real number a, the open half-line

$$(a,\infty)=\sum_{n=1}^{\infty}(a,a+n) \text{ and } (-\infty,a)=\sum_{n=1}^{\infty}(a-n,a)$$

are Borel sets. For real numbers a and b, the union

$$(-\infty,a)+(b,\infty)$$

is Borel. Every complement of a Borel set is Borel, so every closed interval

$$[a,b]=((-\infty,a)+(b,\infty))'$$

is a Borel set. By way of intersections, a closed interval can also be written as

$$[a,b]=\bigcap_{n=1}^{\infty}\left(a-\frac{1}{n},b+\frac{1}{n}\right)$$

In addition, since

$$[a,\infty)=\sum_{n=1}^{\infty}[a,a+n] \text{ and } (-\infty,a]=\sum_{n=1}^{\infty}[a-n,a]$$

the closed half-lines are Borel.

Half open and half closed intervals are also Borel. For Example

$$(a,b]=\bigcap_{n=1}^{\infty}\left(a,b+\frac{1}{n}\right) \text{ and } [a,b)=\bigcap_{n=1}^{\infty}\left(a-\frac{1}{n},b\right)$$

Alternatively, they can be written as intersections of open half-lines and closed half-lines

$$(a, b] = (-\infty, b] \cap (a, \infty) \text{ and } [a, b) = (-\infty, b) \cap [a, \infty)$$

Every set which contains only one real number is Borel. Indeed, if a is real number, then

$$\{a\} = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, a + \frac{1}{n}\right)$$

this means that every set containing finitely many real numbers is Borel. In fact, every set containing countably infinitely many numbers is Borel. Thus, the set of natural numbers $\mathbb{N}$ is a Borel set, the set of rational numbers $\mathbb{Q}$, and as is its complement, the set of irrational numbers $\mathbb{Q}'$ is Borel.

A Borel set is any set that can be formed from open sets (or, equivalently, from closed sets) through the operations of countable union, countable intersection, and relative complement. There are, however, sets which are not Borel. We have just seen that any non-Borel set must have uncountably many points.

The most important σ-algebra on $\mathbb{R}$ is the Borel σ-algebra, defined as follows.

Definition 2.15: The Borel σ-algebra on $\mathbb{R}$, denoted by $\mathcal{B}(\mathbb{R})$, is the smallest σ-algebra on $\mathbb{R}$ that contains all open intervals (a, b) in $\mathbb{R}$. An element of the Borel σ-algebra is called a Borel set.

Let X be a set, and let $\mathcal{F}$ be a family of subsets of X. Then there is a smallest σ-algebra on X that includes $\mathcal{F}$. The smallest σ-algebra is called the σ-algebra generated by $\mathcal{F}$.

Proposition 2.16: The σ-algebra $\mathcal{B}(\mathbb{R})$ of Borel sets of $\mathbb{R}$ is generated by each of the following collections of subsets in $\mathbb{R}$: (a) all open intervals, (b) all closed intervals, (c) all intervals of the form $(a, b]$ (or of the form $[a, b)$), (d) all intervals of the form (a, ∞) (or of the form $[a, \infty)$, $(-\infty, b]$, or $(-\infty, b)$), (e) all open sets, (f) all closed sets, or (g) all intervals of the form $(-\infty, b]$ for which the endpoint b is a rational number.

Because the rational number is dense in $\mathbb{R}$, $\mathcal{B}(\mathbb{R})$ is generated by the collection of intervals $(-\infty, b]$ for which the endpoint b is a rational number. The Borel σ-algebra on $\mathbb{R}^N$ ($N \geqslant 2$), denoted by $\mathcal{B}(\mathbb{R}^N)$, is the smallest σ-algebra on $\mathbb{R}^N$ that contains all open subsets in $\mathbb{R}^N$. Thus, $\mathcal{B}(\mathbb{R}^N)$ is generated by each of the following collections of sets: (a) the collection of all closed subsets of $\mathbb{R}^N$, or (b) the collection of all rectangles in $\mathbb{R}^N$ that have the form $\{(x_1, x_2, \cdots, x_N) : a_i < x_i \leqslant b_i\}$.

2.5.4 Normally Distributed and Uncorrelated does not Imply Independent

When two random variables are normally distributed, it is sometimes mistakenly thought that uncorrelatedness implies independence. However, this is incorrect if the variables are merely marginally normally distributed but not jointly normally distributed.

Let $X \sim \mathrm{N}(0, 1)$, let Z be independent of X, and $Z = 1$ or -1, each with probability $1/2$. Define

$Y = XZ$, we first show that $Y \sim \mathrm{N}(0,1)$: For

$$\begin{aligned}
&\mathrm{P}(Y \leqslant y) = \mathrm{P}(XZ \leqslant y) \\
&= \mathrm{P}(XZ \leqslant y, Z = 1) + \mathrm{P}(XZ \leqslant y, Z = -1) \qquad \text{by } \mathrm{P}(\mathsf{E}) = \sum \mathrm{P}(\mathsf{EF}_i) \\
&= \mathrm{P}(X \leqslant y, Z = 1) + \mathrm{P}(-X \leqslant y, Z = -1) \\
&= \mathrm{P}(X \leqslant y)\mathrm{P}(Z = 1) + \mathrm{P}(-X \leqslant y)\mathrm{P}(Z = -1) \qquad \text{by } X \perp Z \\
&= N(y)/2 + N(y)/2 = N(y) \qquad\qquad \text{by } -X \sim X
\end{aligned}$$

However, X and Y are not correlated, but not independent: For

$$\mathrm{cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = \mathrm{E}(XXZ) - 0 = \mathrm{E}(X^2)\,\mathrm{E}(Z) = 0$$

thus X and Y are uncorrelated. To see that X and Y are not independent, observe that when $-1 < X < 1$, there is $Y \leqslant |Y| = |X| < 1$, and

$$\mathrm{P}(Y > 2 \,|\, -1 < X < 1) = \frac{\mathrm{P}(Y > 2, -1 < X < 1)}{\mathrm{P}(-1 < X < 1)} = 0 \neq \mathrm{P}(Y > 2) = N(-2)$$

The joint CDF (always exists) of (X, Y) is

$$\begin{aligned}
&\mathrm{P}(X \leqslant x, Y \leqslant y) = \mathrm{P}(X \leqslant x, XZ \leqslant y) \\
&= \mathrm{P}(X \leqslant x, XZ \leqslant y, Z = 1) + \mathrm{P}(X \leqslant x, XZ \leqslant y, Z = -1) \qquad \text{by } \mathrm{P}(\mathsf{E}) = \sum \mathrm{P}(\mathsf{EF}_i) \\
&= \mathrm{P}(X \leqslant x, X \leqslant y, Z = 1) + \mathrm{P}(X \leqslant x, -X \leqslant y, Z = -1) \\
&= \mathrm{P}(X \leqslant \min(x, y), Z = 1) + \mathrm{P}(-y \leqslant X \leqslant x, Z = -1) \\
&= \mathrm{P}(X \leqslant \min(x, y))\mathrm{P}(Z = 1) + \mathrm{P}(-y \leqslant X \leqslant x)\mathrm{P}(Z = -1) \\
&= \frac{1}{2}N(\min(x, y)) + \frac{1}{2}\max(N(x) - N(-y), 0)
\end{aligned}$$

It is not a bivariate normal, X and Y are not jointly normally distributed. Note that the distribution of the simple linear combination $X + Y$ concentrates positive probability at 0:

$$\mathrm{P}(X + Y = 0) = \mathrm{P}((1 + Z)X = 0) = 1/2$$

and so it is not normally distributed.

Although X and Y have marginal density, they do not have joint density. Since $|X| = |Y|$, the pair (X, Y) takes values only in the set

$$\mathsf{B} = \{(x, y) : y = \pm x\}$$

which has zero area, thus for any nonnegative function $f(x, y)$, we must have

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} 1_{y=\pm x} f(x, y)\, \mathrm{d}x\mathrm{d}y = 0$$

Observe that $\mathrm{P}(Y = \pm X) = 1$, if there exists joint probability density, then

$$0 = \int\int 1_{y=\pm x} f(x, y)\, \mathrm{d}x\mathrm{d}y = \mathrm{E}(1_{Y=\pm X}) = \mathrm{P}(Y = \pm X) = 1$$

leads to a contradiction.

Remark: For normal random variables X and Y, even they have joint density, they may not jointly normally distributed. For example, let X and Y be standard Gaussian random variables with joint PDF

$$f(x, y) = \begin{cases} 2 f_X(x) f_Y(y) & xy > 0 \\ 0 & \text{otherwise} \end{cases}$$

2.5.5 Proof

We show that conditional probability is a probability measure. For mathematical expectation on discrete random variables, it is necessary to understand that $\sum_{w\in\Omega} X(w)\mathrm{P}(w) = \sum_i x_i \mathrm{P}(X = x_i)$ and the law of the unconscious statistician. On Eq (2.30), we provide keys to reach the conclusion: $G > 0 \iff \mathrm{Q} \sim \mathrm{P}$.

A: Sequence of events

Proof to Proposition 2.3: When $\{\mathsf{E}_i\}$ is increasing, let (progressive additional part)

$$\mathsf{F}_1 = \mathsf{E}_1$$
$$\mathsf{F}_i = \mathsf{E}_i \left(\sum_{n=1}^{i-1} \mathsf{E}_n\right)' = \mathsf{E}_i \mathsf{E}_{i-1}' \qquad i > 1$$

F_i are mutually exclusive events such that

$$\sum_{i=1}^{\infty} \mathsf{E}_i = \sum_{i=1}^{\infty} \mathsf{F}_i \text{ and } \sum_{i=1}^{n} \mathsf{E}_i = \sum_{i=1}^{n} \mathsf{F}_i \qquad \forall n$$

Thus

$$\mathrm{P}\left(\sum_{i=1}^{\infty} \mathsf{E}_i\right) = \mathrm{P}\left(\sum_{i=1}^{\infty} \mathsf{F}_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(\mathsf{F}_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{P}(\mathsf{F}_i)$$
$$= \lim_{n\to\infty} \mathrm{P}\left(\sum_{i=1}^{n} \mathsf{F}_i\right) = \lim_{n\to\infty} \mathrm{P}\left(\sum_{i=1}^{n} \mathsf{E}_i\right) = \lim_{n\to\infty} \mathrm{P}(\mathsf{E}_n)$$

Which proves the result when $\{\mathsf{E}_i\}$ is increasing.

If $\{\mathsf{E}_i\}$ is a decreasing sequence, then $\{\mathsf{E}_i'\}$ is an increasing sequence

B: Conditional Probability

Conditional probability is a probability measure.

The first two properties is easy to verify: For $\mathrm{P}(\mathsf{F}) > 0$ and $0 \leqslant \mathrm{P}(\mathsf{EF}) \leqslant \mathrm{P}(\mathsf{F})$, we have

$$0 \leqslant \frac{\mathrm{P}(\mathsf{EF})}{\mathrm{P}(\mathsf{F})} = \mathrm{P}_{\mathsf{F}}(\mathsf{E}) \leqslant \frac{\mathrm{P}(\mathsf{F})}{\mathrm{P}(\mathsf{F})} = 1$$

And $\mathrm{P}_{\mathsf{F}}(\Omega) = \frac{\mathrm{P}(\Omega\mathsf{F})}{\mathrm{P}(\mathsf{F})} = \frac{\mathrm{P}(\mathsf{F})}{\mathrm{P}(\mathsf{F})} = 1$.

For the countable additivity, note that $\mathsf{E}_i\mathsf{E}_j = \emptyset$ implies $(\mathsf{E}_i\mathsf{F})(\mathsf{E}_j\mathsf{F}) = \emptyset$, thus

$$\mathrm{P}_{\mathsf{F}}\left(\sum_{i=1}^{\infty} \mathsf{E}_i\right) = \mathrm{P}\left(\sum_{i=1}^{\infty} \mathsf{E}_i \middle| \mathsf{F}\right) = \frac{\mathrm{P}\left(\left(\sum_{i=1}^{\infty} \mathsf{E}_i\right)\mathsf{F}\right)}{\mathrm{P}(\mathsf{F})} = \frac{\mathrm{P}\left(\sum_{i=1}^{\infty} \mathsf{E}_i\mathsf{F}\right)}{\mathrm{P}(\mathsf{F})}$$
$$= \frac{\sum_{i=1}^{\infty} \mathrm{P}(\mathsf{E}_i\mathsf{F})}{\mathrm{P}(\mathsf{F})} = \sum_{i=1}^{\infty} \mathrm{P}(\mathsf{E}_i \,|\, \mathsf{F}) = \sum_{i=1}^{\infty} \mathrm{P}_{\mathsf{F}}(\mathsf{E}_i)$$

C: Expectation

Proof to Equation (2.17): $\{X^{-1}(x_i)\}$ is a partition, and by $\sum_{w\in\mathsf{E}} \mathrm{P}(w) = \mathrm{P}(\mathsf{E})$

$$\begin{aligned}\mathrm{E}(X) &= \sum_{w\in\Omega} X(w)\mathrm{P}(w) = \sum_i \sum_{w\in X^{-1}(x_i)} X(w)\mathrm{P}(w) \qquad \text{grouping } X^{-1}(x_i)\\ &= \sum_i \sum_{w\in X^{-1}(x_i)} x_i\mathrm{P}(w) = \sum_i x_i \sum_{w\in X^{-1}(x_i)} \mathrm{P}(w)\\ &= \sum_i x_i\mathrm{P}(X^{-1}(x_i)) = \sum_i x_i\mathrm{P}(X = x_i)\end{aligned}$$

If $X \gneq Y$, then $\mathrm{E}(X) > \mathrm{E}(Y)$

Proof: $X \gneq Y \implies X - Y \gneq 0 \implies \mathrm{E}(X - Y) > 0 \implies \mathrm{E}(X) > \mathrm{E}(Y)$

D: Law of the Unconscious Statistician

Given that random variable X takes N distinctive values $x_1, x_2, \cdots, x_N$, and $Y = h(X)$ has $K \leqslant N$ distinctive values $y_1, y_2, \cdots, y_K$. By law of the unconscious statistician, we have $\mathrm{E}(Y) = \mathrm{E}(h(X)) = \sum_{i=1}^N h(x_i)\mathrm{P}(X = x_i)$. However, following the definition of expectation, for random variable Y, $\mathrm{E}(Y) = \sum_{k=1}^K y_k\mathrm{P}(Y = y_k)$. Verify that the law of the unconscious statistician is valid.

Verification: For $Y^{-1}(y_k)$ is a partition, and $X^{-1}(x_i)$ is a finer partition, with

$$\sum_{i:h(x_i)=y_k} X^{-1}(x_i) = Y^{-1}(y_k)$$

there is $\sum_{i:h(x_i)=y_k} \mathrm{P}(X^{-1}(x_i)) = \mathrm{P}\left(\sum_{i:h(x_i)=y_k} X^{-1}(x_i)\right) = \mathrm{P}(Y^{-1}(y_k))$. Thus

$$\begin{aligned}\mathrm{E}(h(X)) &= \sum_{i=1}^N h(x_i)\mathrm{P}(X = x_i) = \sum_{k=1}^K \sum_{i:h(x_i)=y_k} h(x_i)\mathrm{P}(X = x_i) \qquad \text{grouping } i : h(x_i) = y_k\\ &= \sum_{k=1}^K \sum_{i:h(x_i)=y_k} y_k\mathrm{P}(X^{-1}(x_i)) = \sum_{k=1}^K y_k \sum_{i:h(x_i)=y_k} \mathrm{P}(X^{-1}(x_i))\\ &= \sum_{k=1}^K y_k\mathrm{P}(Y^{-1}(y_k)) = \sum_{k=1}^K y_k\mathrm{P}(Y = y_k) = \mathrm{E}(Y)\end{aligned}$$

E: Discrete Sample Space

Proof to Proposition 2.11: Property (a) follows immediately

Property (b)

$$\mathrm{E}\left(G\right) = \sum_{w\in\Omega} G\left(w\right)\mathrm{P}\left(w\right) = \sum_{w\in\Omega} \frac{\mathrm{Q}\left(w\right)}{\mathrm{P}\left(w\right)}\mathrm{P}\left(w\right) = \sum_{w\in\Omega} \mathrm{Q}\left(w\right) = 1$$

Property (c)

$$\mathrm{E}\left(GY\right) = \sum_{w\in\Omega} G\left(w\right)Y\left(w\right)\mathrm{P}\left(w\right) = \sum_{w\in\Omega} \frac{\mathrm{Q}\left(w\right)}{\mathrm{P}\left(w\right)}Y\left(w\right)\mathrm{P}\left(w\right) = \sum_{w\in\Omega} Y\left(w\right)\mathrm{Q}\left(w\right) = \mathrm{E}^{Q}\left(Y\right)$$

F: Equivalent Probability Measures

On Eq (2.30), a rigid proof requires Lebesgue integral. See items in Absolute Continuity and Density Functions for a detail proof.

- Item 11: G is a probability density function of Q relative to P ($\mathrm{d}Q = G\mathrm{d}P$), and $1/G$ is a probability density function of P relative to Q.
- Item 12: density functions are essentially unique.

In fact, $G > 0 \iff \mathrm{Q} \sim \mathrm{P}$.

G: Stochastic Dominance

$$X \succsim Y \implies \mathrm{E}(X) > \mathrm{E}(Y).$$

Proof. Let

$$X^+ = \max(X, 0) \qquad X^- = \max(-X, 0)$$

then $X = X^+ - X^-$. For any $x \geqslant 0$

$$\mathrm{P}(X^+ > x) = \mathrm{P}(X > x) \geqslant \mathrm{P}(Y > x) = \mathrm{P}(Y^+ > x)$$

and $\mathrm{P}(X^+ > x_0) > \mathrm{P}(Y^+ > x_0)$ for some x_0 (right-continuous, for $\mathrm{P}(X > x) = 1 - F_X(x)$), By Exercise 2.19

$$\mathrm{E}(X^+) - \mathrm{E}(Y^+) = \int_0^{+\infty} \mathrm{P}(X^+ > x) - \mathrm{P}(Y^+ > x)\,\mathrm{d}x > 0$$

Similarly, by

$$\begin{aligned}\mathrm{P}(X^- > x) &= \mathrm{P}(-X > x) = \mathrm{P}(X < -x) = 1 - \mathrm{P}(X \geqslant -x) \\ &\leqslant 1 - \mathrm{P}(Y \geqslant -x) = \mathrm{P}(Y < -x) = \mathrm{P}(-Y < x) = \mathrm{P}(Y^- > x)\end{aligned}$$

there is $\mathrm{E}(X^-) - \mathrm{E}(Y^-) < 0$. Thus

$$\mathrm{E}(X) - \mathrm{E}(Y) = \big(\mathrm{E}(X^+) - \mathrm{E}(Y^+)\big) - \big(\mathrm{E}(X^-) - \mathrm{E}(Y^-)\big) > 0$$ □

Bibliography

Dineen, Seán, 2013. *Probability Theory in Finance: A Mathematical Guide to the Black-Scholes Formula*, 2/e. American Mathematical Society

Kopp, Ekkehard, Jan Malczak, and Tomasz Zastawniak, 2013. *Probability for Finance*. Cambridge University Press, New York

Rosenthal, Jeffrey S., 2006. *A First Look at Rigorous Probability Theory*, 2/e. World Scientific, Singapore

Ross, Sheldon M., 2010. *A First Course in Probability*, 8/e. Pearson Education, Upper Saddle River, NJ

Schwarzlander, Harry, 2011. *Probability Concepts and Theory for Engineers*. John Wiley & Sons

Walpole, Ronald E., Raymond H. Myers, Sharon L. Myers, and Keying Ye, 2012. *Probability & Statistics for Engineers & Scientists*, 9/e. Prentice Hall