
Conditional Expectation

Conditional expectation is much more important than one might at first think, both for finance and probability theory. Indeed, conditional expectation is at the core of modern probability theory because it provides the basic way of incorporating known information into a probability measure. In finance, we treat asset price movements as stochastic processes when determining prices of derivatives, conditional expectations appear in these dynamic settings. Therefore, if you do not have a solid understanding of conditional expectation, the essential tool to process evolving information, you will never have a penetrating understanding of modern finance.

Conditional expectation is presented in almost every textbook using Lebesgue integral, however, Lebesgue integral is somewhat abstract, and most people are not familiar with it. Thus, I prepare a step by step version that makes use of only unconditional expectation, which is much more accessible.

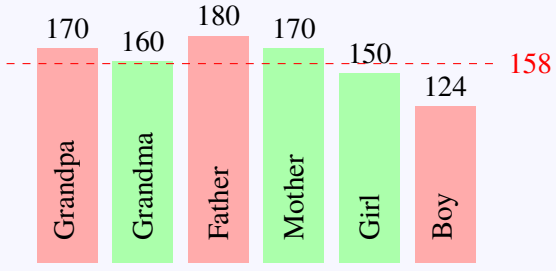
In this chapter, we will assume that random variables come from $L^1(\Omega, \mathcal{F}, P)$ (collection of random variables with $E(|X|) < \infty$ for any random variable X), such that all expected values that are mentioned exist (as real numbers).

§ 3.1 Discrete Case

The definition of conditional expectation may appear somewhat abstract at first. This section is designed to help the beginner step by step through several special cases, which become increasingly involved. The first and simplest case to consider is that of the conditional expectation $E(Y | E)$ of a random variable Y given an event E . For example, the average score of female students is an easy to understand case of conditional expectation: find the average score given the sex is female.

3.1.1 Conditioning on an Event

Example 3.1.1: In a family of six, the heights are as follows (old to young)



Suppose one person is randomly picked up, and is a male, what is the expected height?

The average height of males is

$$\frac{170 + 180 + 124}{3} = 158$$

which equals to the expected height given that the gender is male: Let's define the random variables as follows

	1	2	3	4	5	6	$E(\cdot)$
Height Y	170	160	180	170	150	124	159
Gender S	1	0	1	0	0	1	
$Y I_E$	170	0	180	0	0	124	79

where $E = \{S = 1\}$ means male. In elementary probability course, we have known that

$$\begin{aligned} E(Y | E) &= E(Y | S = 1) = \sum_i y_i P(Y = y_i | S = 1) \\ &= 170 \cdot \frac{1}{3} + 180 \cdot \frac{1}{3} + 124 \cdot \frac{1}{3} = 158 \end{aligned}$$

Note that

$$E(Y I_E) = \frac{1}{6} (170 + 0 + 180 + 0 + 0 + 120) = 79 \quad P(E) = \frac{1}{2}$$

we find that

$$\frac{E(Y I_E)}{E(I_E)} = \frac{E(Y I_E)}{P(E)} = \frac{79}{1/2} = 158$$

In example 3.1.1, we have $E(Y | E) = \frac{E(Y I_E)}{E(I_E)}$, this observation motivates the following definition.

Definition 3.1: For any random variable Y on (Ω, \mathcal{F}) and any event $E \in \mathcal{F}$ such that $P(E) > 0$, the conditional expectation of Y given E is defined by

$$E(Y | E) = \frac{E(Y I_E)}{E(I_E)} = \frac{E(Y I_E)}{P(E)} \tag{3.1}$$

Conditional expectation $E(Y | E)$ is a constant (real number). Thus

$$E(I_E E(Y | E)) = E(Y | E) E(I_E) = E(I_E Y)$$

Conditional expectation $E(Y | E)$ is just a partial averaging: In Example 3.1.1, the average height of male is $E(Y | E) = 158$, it is not the average height of the whole family $E(Y) = 159$, that is why it is called partial averaging. Conditional expectation $E(Y | E)$ is a local mean, which refines the idea of the expectation $E(Y)$ as a mean. Furthermore, since

$$E(Y) = E(Y(I_E + I_{E'})) = E(Y I_E) + E(Y I_{E'}) = E(Y | E)P(E) + E(Y | E')P(E')$$

the expectation $E(Y)$ is a probability weighted average of conditional expectations.

Compare to $E(Y) = \int_{\Omega} Y(w) dP(w)$, since $E(I_E \cdot Y) = \int_{\Omega} I_E Y(w) dP(w) = \int_E Y(w) dP(w)$, which is restricted to a partial territory of the whole sample space, $E(I_E \cdot Y)$ is a *partial expectation* of Y . Analogously, $E(Y) = \frac{E(I_{\Omega} \cdot Y)}{P(\Omega)}$ is the average of Y , $E(Y | E) = \frac{E(I_E \cdot Y)}{P(E)}$ is the *partial average* of Y over event E .

A: Computation

In the definition of conditional expectation of Y given an event E , we do not restrict Y to be discrete, it can be continuous.

- When Y is discrete

$$E(Y I_E) = \int_{\Omega} Y(w) I_E dP(w) = \sum_{w \in E} Y(w) P(w) = \sum_i y_i P(Y = y_i, E) \tag{3.2}$$

The last equality means we collect the atoms in $Y(w) = y_i$ (grouping), which is left as an exercise (Exercise 3.2).

- When Y is continuous

$$E(Y I_E) = \int_{\Omega} Y(w) I_E dP(w) = \int_E Y(w) dP(w) = \int_{Y(E)} y dF_Y(y)$$

- When $Y = c$ is a constant, $E(Y | E) = \frac{E(Y I_E)}{E(I_E)} = \frac{Y E(I_E)}{E(I_E)} = Y$

Example 3.1.2: Three fair coins, 10, 20 and 50 cent coins are tossed. Let Y be the total amount shown by these three coins (sum of the values of those coins that land heads up). What is the expectation of Y given that two coins have landed heads up?

Let E denote the event that two coins have landed heads up. We want to find $E(Y | E)$. Clearly, E consists of three elements (H stands for heads and L for tails as in Example 2.1.1)

$$E = \{HHL, HLL, LHH\}$$

each having the same probability $\frac{1}{8}$. The corresponding values of Y are

$$Y(HHL) = 10 + 20 = 30$$

$$Y(HLH) = 10 + 50 = 60$$

$$Y(LHH) = 20 + 50 = 70$$

Therefore

$$E(Y \mathbb{1}_E) = \sum_{w \in E} Y(w) P(w) = \frac{30}{8} + \frac{60}{8} + \frac{70}{8} = 20$$

and

$$E(Y | E) = \frac{E(Y \mathbb{1}_E)}{P(E)} = \frac{20}{3/8} = \frac{160}{3}$$

B: Conditional Probability

Similarly to Equation (2.18), $P(E) = E(\mathbb{1}_E)$, we have the following result.

Proposition 3.2: If $P(F) > 0$

$$P(E|F) = E(\mathbb{1}_E|F) \tag{3.3}$$

Proof. By $\mathbb{1}_{EF} = \mathbb{1}_E \mathbb{1}_F$, we have

$$E(\mathbb{1}_E|F) = \frac{E(\mathbb{1}_E \mathbb{1}_F)}{P(F)} = \frac{E(\mathbb{1}_{EF})}{P(F)} = \frac{P(EF)}{P(F)} = P(E|F) \quad \square$$

In elementary probability theory, conditional probabilities are used in the computation of conditional expectation: Given an event E , if Y is a discrete random variable

$$E(Y | E) \equiv \sum_i y_i P(Y = y_i | E) \tag{3.4}$$

Which is consistent with Definition 3.1, for (by Eq 3.2)

$$\sum_i y_i P(Y = y_i | E) = \frac{1}{P(E)} \sum_i y_i P(Y = y_i, E) = \frac{E(Y \mathbb{1}_E)}{P(E)}$$

If X is a discrete random variable and $P(X = x) > 0$ for some real number x , then $\{X = x\}$ is an event. In elementary probability theory: $E(Y | X = x) \equiv E(Y | \{X = x\})$

- If Y is discrete, let the conditional probability mass function be $f(y|x) = P(Y = y | X = x)$, then

$$E(Y | X = x) = \sum_y y f(y|x) = \sum_y y P(Y = y | X = x)$$

Which is a specific form of Equation (3.4)

- If Y is continuous

$$E(Y | X = x) = \int_{-\infty}^{+\infty} y f(y|x) dy$$

where $f(y|x) = f(x,y)/P(X = x)$ is the hybrid conditional density function, and $f(x,y)$ is the hybrid density function.

Let X the number of heads in Example 3.1.2, then the expectation of Y given that two coins have landed heads up is $E(Y | X = 2)$. Since

$$P(Y = 30 | X = 2) = \frac{P(Y = 30, X = 2)}{P(X = 2)} = \frac{P(\{HHL\} \cap \{HHL, HLL, LHH\})}{P(\{HHL, HLL, LHH\})} = \frac{1}{3}$$

Similarly, $P(Y = 60 | X = 2) = P(Y = 70 | X = 2) = \frac{1}{3}$, and

$$P(Y = y | X = 2) = 0 \quad y = 0, 10, 20, 50, 80$$

thus

$$E(Y | X = 2) = \sum_y yP(Y = y | X = 2) = \frac{1}{3}(30 + 60 + 70) = \frac{160}{3}$$

Undoubtedly, we can compute $E(Y | X = 0)$, $E(Y | X = 1)$, and $E(Y | X = 3)$, thus conditional expectations on all possible values of X are computed.

3.1.2 Conditioning on a Discrete Random Variable

Let X be a discrete random variable with possible values in $\{x_1, x_2, \dots\}$ such that $P\{X = x_i\} > 0$ for each i . Because we do not know in advance which of events $E_i = \{X = x_i\}$ will occur, we need to consider all possibilities, involving a sequence of conditional expectations

$$E(Y | X = x_i) \equiv E(Y | \{X = x_i\}) = E(Y | E_i) \quad i = 1, 2, \dots$$

Obviously, $E(Y | X = x)$ is a function of $x \in X(\Omega) = \{x_1, x_2, \dots\}$. Let's record this function by

$$r(x) = E(Y | X = x)$$

and denote

$$E(Y | X) \equiv r(X)$$

then $E(Y | X)$ is a random variable, which is called the conditional expectation of Y given X . In more details, $R = E(Y | X)$ is a new discrete random variable such that

$$R(w) = E(Y | X)(w) = \begin{cases} E(Y | X = x_1) & w \in \{X = x_1\} \\ E(Y | X = x_2) & w \in \{X = x_2\} \\ \vdots & \vdots \\ E(Y | X = x_i) & w \in \{X = x_i\} \\ \vdots & \vdots \end{cases} \quad (3.5)$$

In fact, we are computing conditional expectation on an event many times.

Definition 3.3: Let Y be a random variable and let X be a discrete random variable. Then the conditional expectation of Y given X is defined by Equation (3.5).

Equation (3.5) says that $E(Y | X)$ is a discrete random variable if X is discrete, $E(Y | X)$ takes values of the sequence of conditional expectation on $\{X = x_i\}$. Because of $E(Y | X) = r(X)$, the conditional expectation $E(Y | X)$ is a function of X . As a random variable, $E(Y | X)$ has an expectation, we will see that the expectation of $E(Y | X)$ is $E(Y)$ in Proposition 3.4.

Let's define the *conditional random variable* $Y_{|X=x}$ to be the restriction of Y to the event $\{X = x\}$

$$Y_{|X=x} = \begin{cases} Y(w) & w \in \{X = x\} \\ \text{undefined} & \text{otherwise} \end{cases} \tag{3.6}$$

Then, $Y_{|X=x}$ has its distribution, the **conditional distribution¹** of Y given $\{X = x\}$. When Y is discrete or continuous, because the conditional PMF/PDF $f(y | x_i) = 0$ if $y \notin Y(X = x_i)$, the expectation of $Y_{|X=x_i}$ is equal to the conditional expectation of Y on event $\{X = x_i\}$

$$E(Y_{|X=x_i}) = E(Y | X = x_i) = r(x_i)$$

For this reason, the conditional random variable $Y_{|X=x}$ is also written as $(Y | X = x)$.

The conditional expectation $R = E(Y | X)$ is a constant over $\{X = x_i\}$

$$R(w) = r(x_i) = E(Y_{|X=x_i}) \quad \forall w \in \{X = x_i\}$$

We see that the conditional expectation R is a partial average, an average over each partial sample space $\{X = x_i\}$. See figure 3.1 for an illustration.

Example 3.1.3: Three fair coins, 10, 20 and 50 cent coins are tossed as in Example 3.1.2. Let Y be the total amount shown by these three coins, and X be the total amount shown by the 10 and 20 cent coins only. What is the conditional expectation of Y on X ?

Clearly, X is a discrete random variable with four possible values: 0, 10, 20 and 30 cents. We find the four corresponding conditional expectations in a similar way as in Example 3.1.2: For

$$\{X = 0\} = \{LLH, LLL\}$$

and

$$Y(LLH) = 50 \quad Y(LLL) = 0$$

thus

$$E(Y | X = 0) = \frac{E(Y 1_{X=0})}{P(X = 0)} = \frac{\sum_{w \in \{X=0\}} Y(w) P(w)}{2/8} = \frac{50/8 + 0/8}{2/8} = 25$$

similarly $E(Y | X = 10) = 35$, $E(Y | X = 20) = 45$ and $E(Y | X = 30) = 55$. Therefore

$$E(Y | X) = \begin{cases} 25 & w \in \{LLH, LLL\} = \{X = 0\} \\ 35 & w \in \{HLH, HLL\} = \{X = 10\} \\ 45 & w \in \{LHH, LHL\} = \{X = 20\} \\ 55 & w \in \{HHH, HHL\} = \{X = 30\} \end{cases}$$

Which shows that $E(Y | X)$ is a random variable (a function of $w \in \Omega$), and it is a function of X .

Example 3.1.4: Let $\Omega = [0, 1]$ with P the uniform measure ($P(w \leq a) = a$ for $a \in [0, 1]$). Find

¹Recall that $\{X = x\}$ can be thought of as the new sample space for the conditional probabilities here, in this sense, the restriction version of Y , $Y_{|X=x}$, can be thought of as the random variable define on the partial sample space $\{X = x\}$.

$E(Y | X)$ for

$$Y(w) = 2w^2 \quad X(w) = \begin{cases} 1 & w \in [0, \frac{1}{3}] \\ 2 & w \in (\frac{1}{3}, \frac{2}{3}] \\ 0 & w \in (\frac{2}{3}, 1] \end{cases}$$

Clearly, X is discrete with three possible values 0, 1 and 2. The corresponding events are

$$\{X = 0\} = (\frac{2}{3}, 1] \quad \{X = 1\} = [0, \frac{1}{3}] \quad \{X = 2\} = (\frac{1}{3}, \frac{2}{3}]$$

The CDF of Y is

$$F_Y(y) = P(Y(w) \leq y) = P(2w^2 \leq y) = P(w \leq \sqrt{y/2}) = \sqrt{y/2} \quad y \in [0, 2]$$

If $w \in \{X = 0\}$, then $Y(w) \in (\frac{8}{9}, 2]$. We obtain

$$\{X = 0\} = \{8/9 < Y \leq 2\}$$

thus $1_{X=0} = 1_{8/9 < Y \leq 2}$

$$E(Y 1_{X=0}) = E(Y 1_{8/9 < Y \leq 2}) = \int_{8/9}^2 y dF_Y(y) = \int_{8/9}^2 y \cdot \frac{1}{4} \sqrt{\frac{2}{y}} dy = \frac{38}{81}$$

and

$$E(Y | X = 0) = \frac{E(Y 1_{X=0})}{P(X = 0)} = \frac{38}{27}$$

similarly

$$E(Y | X = 1) = \frac{2}{27} \quad E(Y | X = 2) = \frac{14}{27}$$

We see that $E(Y | X)$ is a discrete random variable even Y is continuous. The graph of $E(Y | X)$ is shown in Figure 3.1 with more interpretations.

On computation² of $E(Y 1_{X=0})$, we can use mixed joint CDF: First find out $F(x, y)$ with $x \in \{0, 1, 3\}$, for example when $x = 0$

$$\begin{aligned} F(0, y) &= P(X = 0, Y(w) \leq y) = P(2/3 < w \leq 1, 2w^2 \leq y) \\ &= P(2/3 < w \leq \sqrt{y/2}) = \sqrt{y/2} - 2/3 \quad 8/9 < y \leq 2 \end{aligned}$$

and the *hybrid/mixed density function* $f(x, y) = \frac{\partial F(x, y)}{\partial y}$, then $f(0, y) = \frac{1}{4} \sqrt{\frac{2}{y}}$

$$E(Y 1_{X=0}) = \int_{-\infty}^{+\infty} \sum_{i=1}^{\infty} (y 1_{x_i=0}) \cdot f(x_i, y) dy = \int_{-\infty}^{+\infty} y f(0, y) dy = \int_{8/9}^2 y \cdot \frac{1}{4} \sqrt{\frac{2}{y}} dy = \frac{38}{81}$$

²Using Lebesgue integral, we have the following ways

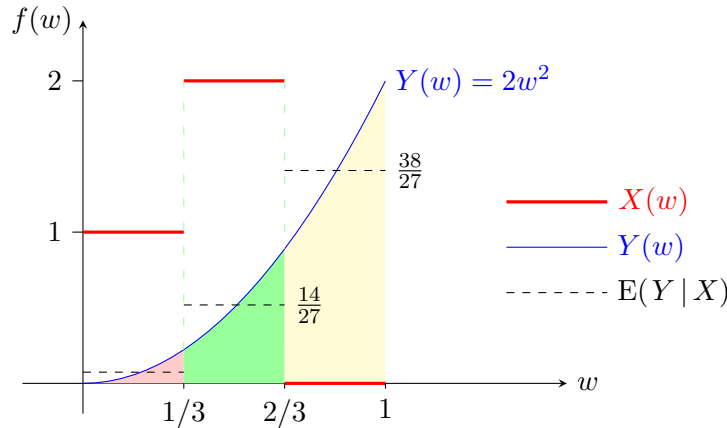
- $\int_{\mathbb{B}} Y(w) dP(w) = \int_{\mathbb{B}} Y(w) dw$, ($P(w)$ is Lebesgue measure on $[0, 1]$, and Riemann and Lebesgue integrals agree)

$$E(Y 1_{X=0}) = \int_{X=0} Y(w) dP(w) = \int_{2/3}^1 2w^2 dw = \frac{38}{81}$$

- $\int_{\mathbb{E}} h(Y(w)) dP(w) = \int_{Y(\mathbb{E})} h(y) dF_Y(y)$. For $Y(X = 0) = (8/9, 2]$

$$E(Y 1_{X=0}) = \int_{X=0} Y(w) dP(w) = \int_{Y(X=0)} y dF_Y(y) = \int_{8/9}^2 y \cdot \frac{1}{4} \sqrt{\frac{2}{y}} dy = \frac{38}{81}$$

Figure 3.1: Conditioning on a Discrete Random Variable In Example 3.1.4, we see that conditional expectation $E(Y | X)$ is a *partial average*: $E(Y | X = 0) = \frac{38}{27}$ is the average height of Y over the interval $(\frac{2}{3}, 1] = \{X = 0\}$. Similarly, $E(Y | X = 1) = \frac{2}{27}$ and $E(Y | X = 2) = \frac{14}{27}$ are the average height over $[0, \frac{1}{3}]$ and $(\frac{1}{3}, \frac{2}{3}]$ respectively. Because $E(Y 1_{X=0}) = \int_{X(w)=0} Y(w) dw = \int_{2/3}^1 2w^2 dw = \frac{38}{81}$ is the area under the graph $Y = 2w^2$ in the interval $(\frac{2}{3}, 1] = \{X = 0\}$, and $P(X = 0) = \frac{1}{3}$ is the length of the corresponding interval. Thus, let's define a rectangle having same area as the region under the graph, whose width is the length of the underlying interval, then $E(Y | X = 0) = \frac{E(Y 1_{X=0})}{P(X=0)}$ is the height of this rectangle.



note that given hybrid density function $f(x, y)$

$$E(h(X, Y)) = \int_{-\infty}^{+\infty} \sum_{i=1}^{\infty} h(x_i, y) f(x_i, y) dy$$

where the discrete portion is compute by summation, and the continuous portion is compute by integral.

Besides, the conditional PDF of Y given $X = x_i$ is $f(y | x_i) = f(x_i, y) / P(X = x_i)$, thus

$$E(Y | X = 0) = \int_{-\infty}^{+\infty} y f(y | 0) dy = \int_{8/9}^2 y \cdot \frac{1}{4} \sqrt{\frac{2}{y}} \frac{1}{1/3} dy = \frac{38}{27}$$

3.1.3 Properties

Toward a general definition and properties of conditional expectation, we explain some of the most important properties of conditional expectation of Y given a discrete random variable X . We will show that, what matters is the information contain in X , not the values taken by X .

A: Law of Total Expectation

Proposition 3.4: If X is a discrete random variable

$$E(E(Y | X)) = E(Y)$$

Proof. If X is discrete, $E(Y | X)$ is a discrete random variable, thus

$$\begin{aligned} E(E(Y | X)) &= \sum_i E(Y | X = x_i)P(X = x_i) = \sum_i E(Y 1_{X=x_i}) \\ &= E\left(\sum_i Y 1_{X=x_i}\right) = E\left(Y \sum_i 1_{X=x_i}\right) = E(Y) \quad \square \end{aligned}$$

We do not assume that Y is a discrete random variable. The law of total expectation is an extremely useful result that often enables us to compute unconditional expectations easily by first conditioning on some appropriate random variable. The following examples demonstrate its use.

Example 3.1.5: A lucky wheel gives equal chance for \$3, \$5 and \$7. You keep spinning until you land on \$3. What is the expected prize for this game?

Let Y be the amount of prize, and X be the prize on the first spin. Obviously, $Y_{|X=3} = 3$ and

$$E(Y | X = 3) = 3$$

Note that if $X = 5$, we return to the wheel again, conditional random variable $Y_{|X=5}$ has the same distribution as $Y + 5$, say, $Y_{|X=5} \sim Y + 5$. Similarly, $Y_{|X=7} \sim Y + 7$. We have

$$E(Y | X = 5) = 5 + E(Y)$$

$$E(Y | X = 7) = 7 + E(Y)$$

and

$$E(Y) = \sum_{i \in \{3,5,7\}} E(Y | X = i)P(X = i) = \frac{1}{3}[3 + 5 + E(Y) + 7 + E(Y)]$$

hence $E(Y) = 15$.

Remark: If the player hits the second item, she obtains \$5 and return to the wheel. But once she return to the wheel, the problem is as before. Hence $E(Y | X = 5) = 5 + E(Y)$.

The divide and conquer is a key in the application of the law of total expectation.

Example 3.1.6: The game of craps is begun by rolling an ordinary pair of dice.

- If the sum of the dice is 2, 3, or 12, the player loses (denoted as $Z = 0$).
- If it is 7 or 11, the player wins ($Z = 1$).
- If it is any other number $i \in I = \{4, 5, 6, 8, 9, 10\}$, the player goes on rolling the dice until the sum is either 7 (the player loses) or i (the player wins).

Let Y be the number of rolls of the dice in a game of craps. Find the expected number of rolls $E(Y)$, the probability that the player wins $P(Z = 1)$, and the expected number of rolls given that the player wins, $E(Y | Z = 1)$.

Let S_n be the sum of dice on the n th roll, and $P_i = P(S_n = i)$, then reading from the antidiagonals

in Example 2.1.2

$$P_i = P_{14-i} = \frac{i-1}{36} \quad i = 2, 3, \dots, 7$$

Condition on the initial sum S_1

$$E(Y | S_1 = i) = \begin{cases} 1 & i = 2, 3, 7, 11, 12 \\ 1 + \frac{1}{P_i + P_7} & i \in I \end{cases}$$

for if $S_1 = i \in I$ that does not end the game, then the dice will continue to be rolled until the sum is either i or 7, and the number of rolls until this occurs is a geometric random variable with parameter $P_i + P_7$.

Therefore

$$\begin{aligned} E(Y) &= E(E(Y | S_1)) = \sum_{i=2}^{12} E(Y | S_1 = i)P(S_1 = i) = \sum_{i=2}^{12} E(Y | S_1 = i)P_i \\ &= 1 + \sum_{i \in I} \frac{1}{P_i + P_7} P_i = 1 + 2 \sum_{i=4}^6 \frac{1}{P_i + P_7} P_i \quad P_i = P_{14-i} \\ &= 1 + 2 \left(\frac{1}{3} + \frac{2}{5} + \frac{5}{11} \right) = \frac{557}{165} \approx 3.3758 \end{aligned}$$

Now we compute $P(Z = 1)$, since $\{Z = 1\} \cap \{S_1 = 2\} = \emptyset$ and $\{S_1 = 7\} \subset \{Z = 1\}$, we see that condition on the initial sum S_1

$$P(Z = 1 | S_1 = i) = \begin{cases} 0 & i = 2, 3, 12 \\ 1 & i = 7, 11 \end{cases}$$

If $i \in I$, let $q_i = 1 - (P_i + P_7)$, and (let $(E, F) \equiv E \cap F$)

$$V_{2i} = (S_2 = i, S_1 = i) = \{S_2 = i\} \cap \{S_1 = i\}$$

$$V_{ni} = (S_n = i, \{S_{n-1}, \dots, S_3, S_2\} \cap \{i, 7\} = \emptyset, S_1 = i) \quad n > 2$$

be the event that the player wins on the n th roll. Then $P(V_{ni}) = P_i^2 q_i^{n-2}$,

$$\{Z = 1\} = \{S_1 = 7\} + \{S_1 = 11\} + \sum_{i \in I} \sum_{n=2}^{\infty} V_{ni}$$

and

$$\begin{aligned} P(Z = 1 | S_1 = i) &= \frac{P(Z = 1, S_1 = i)}{P(S_1 = i)} = \frac{P(\sum_{n=2}^{\infty} V_{ni})}{P_i} \\ &= \frac{1}{P_i} \sum_{n=2}^{\infty} P_i^2 q_i^{n-2} = P_i \sum_{k=0}^{\infty} q_i^k = \frac{P_i}{P_i + P_7} \quad i \in I \end{aligned}$$

which confirms Equation (2.16). Thus, by Eq (2.3)

$$\begin{aligned} p &= P(Z = 1) = \sum_{i=2}^{12} P(Z = 1 | S_1 = i)P_i \\ &= (P_7 + P_{11}) + \sum_{i \in I} \frac{P_i^2}{P_i + P_7} = \frac{244}{495} \approx 0.4929 \end{aligned}$$

Finally, we compute $E(Y | Z = 1)$: It is easy to find that (Exercise 3.7)

$$P(Y = n | Z = 1) = \frac{1}{p} \sum_{i \in I} P_i^2 q_i^{n-2} \quad n > 1 \tag{3.7}$$

and

$$\sum_{n=2}^{\infty} n q_i^{n-2} = \frac{1 + P_i + P_7}{(P_i + P_7)^2} \quad i \in I \tag{3.8}$$

thus

$$\begin{aligned}
 E(Y | Z = 1) &= \sum_{n=1}^{\infty} nP(Y = n | Z = 1) = P(Y = 1 | Z = 1) + \sum_{n=2}^{\infty} nP(Y = n | Z = 1) \\
 &= \frac{P_7 + P_{11}}{p} + \sum_{n=2}^{\infty} n \left(\frac{1}{p} \sum_{i \in I} P_i^2 q_i^{n-2} \right) = \frac{P_7 + P_{11}}{p} + \frac{1}{p} \sum_{i \in I} P_i^2 \left(\sum_{n=2}^{\infty} n q_i^{n-2} \right) \\
 &= \frac{P_7 + P_{11}}{p} + \frac{1}{p} \sum_{i \in I} P_i^2 \frac{1 + P_i + P_7}{(P_i + P_7)^2} = \frac{9858}{3355} \approx 2.9383
 \end{aligned}$$

B: Substitution Rule

Proposition 3.5 (Substitution Rule): Let $h(x, y)$ be a function of x and y

$$E(h(X, Y) | X = x) = E(h(x, Y) | X = x)$$

Proof. Given $X = x$, we have $w \in \{X = x\}$ and

$$h(X, Y) = h(X(w), Y(w)) = h(x, Y(w))$$

which gives $h(X, Y)1_{X=x} = h(x, Y)1_{X=x}$, thus

$$E(h(X, Y) | X = x) = \frac{E(h(X, Y)1_{X=x})}{P(X = x)} = \frac{E(h(x, Y)1_{X=x})}{P(X = x)} = E(h(x, Y) | X = x) \quad \square$$

We can think of $h(X, Y)$ as $h(x, Y)$ when compute conditional expectation given $X = x$. That is, if we have known that $X = x$, we can replace every appearance of X to the left of the conditioning bar by x , other random variables are unchanged.

Remark: The substitution rule is valid for conditional expectation not only for discrete case, but for general case. We are treating X as a constant x , since the sample space is restricted to $\{X = x\}$.

Proposition 3.6: If Y is $\sigma(X)$ -measurable and X is a discrete random variable

$$E(Y | X) = Y$$

Proof. By Doob-Dynkin lemma, there exists a function $h(x)$ such that $Y = h(X)$. Following the substitution rule, $E(h(X) | X = x) = E(h(x) | X = x) = h(x)$. Which means $h(X) = E(h(X) | X)$. Accordingly

$$E(Y | X) = E(h(X) | X) = h(X) = Y \quad \square$$

C: Conditional Information

What if conditioning on a independent random variable? The conditional information is useless, for we do not able to improve the prediction, the conditional mean is equal to unconditional mean.

Proposition 3.7: If random variable Y and discrete random variable X are independent, then

$$E(Y | X) = E(Y)$$

Proof. Given $Y \perp X$, we have $Y \perp 1_{X=x_i}$. By Equation (3.1)

$$E(Y | X = x_i) = \frac{E(Y 1_{X=x_i})}{P(X = x_i)} = \frac{E(Y) E(1_{X=x_i})}{P(X = x_i)} = E(Y)$$

thus, $E(Y | X) = E(Y)$ follows from Equation (3.5). □

Remark: If $Y \perp X$, we have $E(Y | X = x_i) = E(Y)$, knowing the realizations of X has no effect on the prediction of Y .

Example 3.1.7: Let $Y_i \sim N(i, 2i)$, $Y_i \perp X$, $i = 0, 1$. If $P(X = 2) = P(X = 1) = 1/2$, compute $E(Y_X^2 | X = 1)$. By substitution rule

$$E(Y_X^2 | X = 1) = E(Y_1^2 | X = 1) = E(Y_1^2) = [E(Y_1)]^2 + \text{var}(Y_1) = 3$$

Note that $U = Y_X^2$ is a random variable, say

$$U(w) = \begin{cases} Y_1^2(w) & w \in \{X = 1\} \\ Y_2^2(w) & w \in \{X = 2\} \end{cases}$$

Suppose that X indicates day or night, and Y_i is the position of a pollen grain in water, then $U|_{X=1}$ must be Y_1^2 , the square of particle position at night ($X = 1$).

If $E(Y | X) = E(Z | X) = 0$, we have $E(Y) = E(Z) = 0$, but we do not have $E(YZ | X) = 0$.

$$E(Y | X) = E(Z | X) = 0 \not\Rightarrow E(YZ | X) = 0$$

Furthermore (If $r(x, z) = E(Y | X = x, Z = z)$, then $E(Y | X, Z) = r(X, Z)$)

$$E(Y | X) = E(Y | Z) = 0 \not\Rightarrow E(Y | X, Z) = 0$$

The following examples illustrate these points.

Example 3.1.8: Toss a fair dice $\Omega = \{1, 2, 3, 4, 5, 6\}$, define

$$X = \begin{cases} 1 & w \in \{1, 6\} \\ 0 & w \in \{2, 3, 4, 5\} \end{cases} \quad Y = 7 - 2w \quad Z = \begin{cases} 1 & w \in \{1, 3, 5\} \\ -1 & w \in \{2, 4, 6\} \end{cases}$$

then $(f(x, z) = f(x)f(z) \implies X \perp Z)$

$$E(Y | X) = E(Z | X) = 0$$

however

$$E(YZ | X) = \begin{cases} \frac{5 \cdot 1 + (-5) \cdot (-1)}{6} / \frac{2}{6} = 5 & X = 1 \\ \frac{3 \cdot (-1) + 1 \cdot 1 + (-1) \cdot (-1) + (-3) \cdot 1}{6} / \frac{4}{6} = -1 & X = 0 \end{cases}$$

Example 3.1.9: Toss a fair dice $\Omega = \{1, 2, 3, 4, 5, 6\}$, let

$$X = \begin{cases} 1 & w \in \{1, 6\} \\ 0 & w \in \{2, 3, 4, 5\} \end{cases} \quad Y = \begin{cases} 1 & w \in \{1, 3, 5\} \\ -1 & w \in \{2, 4, 6\} \end{cases} \quad Z = 1_{w>4}$$

then $(Y \perp X$ and $Y \perp Z)$

$$E(Y | Z) = E(Y | X) = 0$$

however

$$E(Y | X, Z) = \begin{cases} \frac{-1+1-1}{6} / \frac{3}{6} = -\frac{1}{3} & (X, Z) = (0, 0) & w \in \{2, 3, 4\} \\ \frac{1}{6} / \frac{1}{6} = 1 & (X, Z) = (1, 0) & w \in \{1\} \\ \frac{1}{6} / \frac{1}{6} = 1 & (X, Z) = (0, 1) & w \in \{5\} \\ \frac{-1}{6} / \frac{1}{6} = -1 & (X, Z) = (1, 1) & w \in \{6\} \end{cases}$$

D: Measurability and Partial Averaging

The essentials of the conditional expectation are the measurability and partial averaging.

Proposition 3.8: If Y is a random variable and X is a discrete random variable, let $R = E(Y | X)$, then

1. Measurability: R is a $\sigma(X)$ -measurable random variable
2. Partial averaging: For any $E \in \sigma(X)$

$$E(1_E \cdot R) = E(1_E \cdot Y)$$

Remark: Please note that Y is not necessary $\sigma(X)$ -measurable. But if $Y \in \sigma(X)$, by Proposition 3.6, we have $E(Y | X) = Y$

- For any simple event $\{X = x_i\}$ in $\sigma(X)$, by partial averaging property, we have

$$E(Y | X = x_i) = E(R | X = x_i)$$

which states that, $E(Y | X = x_i)$, the partial average of Y over event $\{X = x_i\}$, is equal to $E(R | X = x_i)$, where R is $\sigma(X)$ -measurable.

- We have known that $E(Y | X)$ is the partial average over each simple event $\{X = x_i\}$, the partial averaging property states that this can be extended to any event $E \in \sigma(X)$, thus for any $E \in \sigma(X)$ with $P(E) > 0$

$$E(Y | E) = E(R | E)$$

the partial average of Y is always equal to the partial average of R . (thus called the partial averaging property)

- If $E \notin \sigma(X)$, the partial averaging property may fail. For an instance, in Example 3.1.4, let

$E = [0, \frac{1}{2}] \notin \sigma(X)$, then $E(I_E Y) \neq E(I_E R)$, because

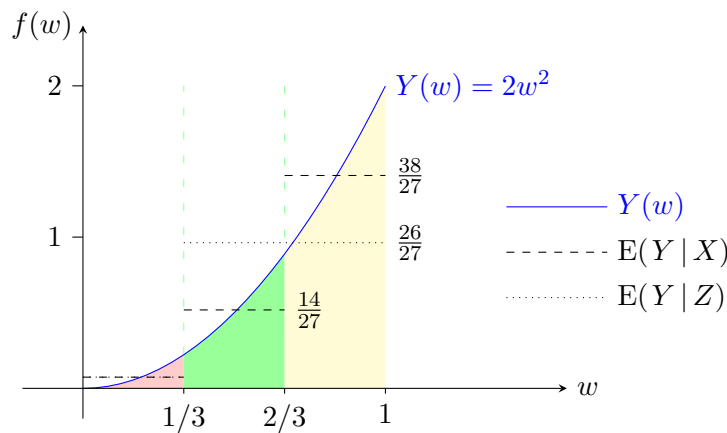
$$E(I_E Y) = \int_E Y(w) dP(w) = \int_0^{1/2} Y(w) dw = \int_0^{1/2} 2w^2 dw = \frac{1}{12}$$

$$E(I_E R) = \int_E R(w) dP(w) = \int_0^{1/2} R(w) dw = \int_0^{1/3} \frac{2}{27} dw + \int_{1/3}^{1/2} \frac{14}{27} dw = \frac{1}{9}$$

where the computation of $E(I_E R)$ is better done by way of modern probability theory.

Figure 3.2: Prediction and Finer Information In Example 3.1.4, let $Z(w) = 1_{X=1}$, then $\sigma(Z) < \sigma(X)$.

As an estimation of Y , $E(Y | X)$ is better than $E(Y | Z)$: If $X = 2$ and $Z = 0$ are observed, then the unobserved true state w must be in $(\frac{1}{3}, \frac{2}{3}]$, hence $Y \in (\frac{2}{9}, \frac{8}{9}]$. We see that $E(Y | Z = 0) = \frac{26}{27}$ is out of range while $E(Y | X = 2) = \frac{14}{27}$ is within range, undoubtedly, $E(Y | X = 2)$ is a better estimation.



The partial averaging property ensures that $E(Y | X)$ is indeed an estimate of Y , which is a valuable prediction if Y is time and money costly to observe or even Y is not observable. In Example 3.1.4, let $Z(w) = 1_{X=1}$, then $E(Y | Z = 0) = \frac{26}{27}$ and $E(Y | Z = 1) = E(Y | X = 1) = \frac{2}{27}$. In Figure 3.2, we see that $\{X = 2\} = (\frac{1}{3}, \frac{2}{3}]$ and $\{X = 0\} = (\frac{2}{3}, 1]$ provide a finer resolution of the uncertainty of the world than $\{Z = 0\} = (\frac{1}{3}, 1] = X^{-1}(\{0, 2\})$ does. As estimations of Y , the partial averaging property says that $E(Y | X = 2) = \frac{14}{27}$ and $E(Y | X = 0) = \frac{38}{27}$ are better than $E(Y | Z = 0) = \frac{26}{27}$.

E: Information Matters

Read carefully from Equation (3.5), we like to note that the conditional expectation $E(Y | X)$ does not depend on the actual values of X but just on the partition generated by X , or equivalently on the event space $\sigma(X)$. For example: toss a fair dice $\Omega = \{1, 2, 3, 4, 5, 6\}$, define

$$Y = 7 - 2w \quad X = \begin{cases} 1 & w \in \{1, 3, 5\} \\ 0 & w \in \{2, 4, 6\} \end{cases} \quad Z = \begin{cases} 135 & w \in \{1, 3, 5\} \\ 246 & w \in \{2, 4, 6\} \end{cases}$$

then $\sigma(X) = \sigma(Z)$, and

$$E(Y | X) = E(Y | Z) = \begin{cases} \frac{5+1-3}{6}/\frac{3}{6} = 1 & w \in \{1, 3, 5\} = X^{-1}(1) = Z^{-1}(135) \\ \frac{3-1-5}{6}/\frac{3}{6} = -1 & w \in \{2, 4, 6\} = X^{-1}(0) = Z^{-1}(246) \end{cases}$$

What matters is the information revealed by the random variables conditioned on, not the values taken by them. When $\sigma(X) = \sigma(Z)$, both random variables provide the identical amount of information. Observing the realization of X or Z , we learn the underlying state of world. We care more on information than the actual values, for information is vital for financial market, most of the decision-making are based on specific information. Needless to say, conditional expectation is an indispensable tool for financial decisions.

§ 3.2 General Case

When conditioning on a discrete random variable, conditional expectation has the following fundamental properties:

1. **Measurability:** $E(Y | X)$ is a function of X , and thus it is a random variable and measurable to $\sigma(X)$. That is to say, the value of $E(Y | X)$ can be determined from the information in X .
2. **Partial averaging:** $E(Y | X)$ as a “best approximation”, the expected value of Y given the information from X . Because $\sigma(X)$ contains all the information from X , for discrete random variable Z , if Z delivers less information such that $\sigma(Z) < \sigma(X)$, then $E(Y | X)$ provides better estimations of Y than $E(Y | Z)$ does.

Which culminate at the general definition of conditional expectation for general case.

3.2.1 Conditioning on an Arbitrary Random Variable

Let X be a uniformly distributed random variable on $[0, 1]$. Then the event $\{X = x\}$ has probability $P(X = x) = 0$ for every $x \in [0, 1]$. In such situations the former definition of $E(Y | X = x)$ no longer makes sense even when $f_X(x) > 0$. We need a new style, a new approach to defining it by means of certain properties which follow from the special case of conditioning with respect to a discrete random variable. Proposition 3.8 provides the key to the definition of the conditional expectation given an arbitrary random variable X .

Definition 3.9: Let Y be an integrable random variable and let X be an arbitrary random variable. Then the conditional expectation of Y given X , denoted as $E(Y | X)$, is any random variable R that satisfies

1. **Measurability:** R is $\sigma(X)$ -measurable
2. **Partial averaging:** For any $E \in \sigma(X)$

$$E(I_E \cdot R) = E(I_E \cdot Y) \quad (3.9)$$

We are merely given the required property that a conditional expectation must satisfy. The existence of $E(Y | X)$ will be discussed later.

- **Measurability:** By Doob-Dynkin lemma, the measurability condition ensures that $R = E(Y | X)$ is a function of X . For example, if (X, Y) follows bivariate normal, Equation (2.27) says that $E(Y | X) = a + bX$, which is a linear function of X .
- **Partial averaging:** If $P(E) > 0$, Eq (3.9) is equivalent to

$$E(R | E) = E(Y | E)$$

$E(Y | E)$ is the partial average of Y over event E . $E(Y | E)$ is indeed an estimate of Y , conditioning on $E \in \sigma(X)$, and it is equal to $E(R | E)$ with $R = E(Y | X) \in \sigma(X)$.

- The partial averaging condition is equivalent to (Lebesgue integral is needed for a proof)

$$E(UR) = E(UY)$$

for any random variable $U \in \sigma(X)$.

- We can also define the conditional probability of an event $E \in \mathcal{F}$ given X by

$$P(E | X) \equiv E(1_E | X)$$

which is equivalent to the following equality in elementary probability

$$P(E | X = x) = E(1_E | X = x) \quad (3.10)$$

When compute conditional expectation, often the Equation (3.9) is not used, for it does not give explicit formula for $E(Y | X)$. However, it is used to establish the fundamental properties given in §3.2.3. As shown in Exercise 3.6, the Equation (2.22) introduced in elementary probability is consistent with our Definition 3.9. Thus, once $f(y | x)$ is known, we compute $r(x) = E(Y | X = x)$, and then easily obtain $E(Y | X) = r(X)$.

Remark: $E(Y | X) = r(X)$ is a random variable, the distribution of $E(Y | X)$ is the distribution of $r(X)$, not the distribution of conditional random variable $Y_{|X=x}$. For example, in the bivariate normal case, from Equation (2.27)

$$E(Y | X) = r(X) = \mu_Y + (X - \mu_X)\rho\sigma_Y/\sigma_X \sim N(\mu_Y, \sigma_Y^2\rho^2)$$

with expectation $E(E(Y | X)) = E(Y) = \mu_Y$. Which is different from the distribution of $Y_{|X=x}$, the conditional distribution of Y given $\{X = x\}$ as in Equation (2.26)

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

note that the expectation of $Y_{|X=x}$ equals the conditional expectation of Y on $X = x$

$$E(Y_{|X=x}) = r(x) = E(Y | X = x)$$

Lemma 3.10: Let (Ω, \mathcal{F}, P) be a probability space and let \mathcal{G} be a σ -algebra contained in \mathcal{F} . If Y is a \mathcal{G} -measurable random variable and for any $G \in \mathcal{G}$

$$E(1_G Y) = 0$$

then $P(Y = 0) = 1$.

Following Lemma 3.10, conditional expectation is unique in the sense of almost surely. If R and R_* both satisfy the two conditions in Definition 3.9, then $E(1_E \cdot (R - R_*)) = 0$ for any $E \in \sigma(X)$, and thus

$$P(R = R_*) = 1$$

or $R = R_*$ a.s. Which means, versions of conditional expectation of Y given X will only differ on null sets (event with zero probability).

According to the following proposition, we also note that, when conditioning on an arbitrary random variable, the conditional expectation $E(Y | X)$ is $\sigma(X)$ -measurable and does not depend on the actual values of X but just on the event space $\sigma(X)$.

Proposition 3.11: If $\sigma(X) = \sigma(X_*)$, then $E(Y | X) = E(Y | X_*)$ a.s.

Proof. Let $R = E(Y | X)$ and $R_* = E(Y | X_*)$, then for any $E \in \sigma(X) = \sigma(X_*)$, we have

$$E(I_E R) = E(I_E Y) = E(I_E R_*)$$

or

$$E(I_E \cdot (R - R_*)) = 0$$

as an immediate consequence of Lemma 3.10, $R = R_*$ a.s., that is $E(Y | X) = E(Y | X_*)$ a.s. \square

3.2.2 Conditioning on an Event Space

Based on the observation that $E(Y | X)$ depends only on the event space $\sigma(X)$, the information revealed by X , rather than on the actual values of X . It is reasonable to talk of conditional expectation given an event space. We are now in a position to make the final step towards the general definition of conditional expectation.

Definition 3.12: Let Y be a random variable on a probability space (Ω, \mathcal{F}, P) , and let \mathcal{G} be an event space contained in \mathcal{F} . Then the conditional expectation of Y given \mathcal{G} , denoted as $E(Y | \mathcal{G})$, is any random variable R that satisfies

1. Measurability: R is \mathcal{G} -measurable
2. Partial averaging: For any $G \in \mathcal{G}$

$$E(I_G R) = E(I_G Y)$$

Definition 3.12 is the general definition of conditional expectation: When conditioning on an arbitrary random variable X , we can take $\mathcal{G} = \sigma(X)$. Particularly, if $\mathcal{G} = \{\emptyset, \Omega\}$, \mathcal{G} is the smallest event space and contains no information, thus

$$E(Y | \mathcal{G}) = E(Y)$$

the conditional expectation becomes an unconditional expectation. When conditioning on an event E with $P(E) > 0$ and $P(E') > 0$, let $\mathcal{G} = \{\emptyset, \Omega, E, E'\}$, it is easy to verify that

$$R(w) = E(Y | \mathcal{G})(w) = \begin{cases} \frac{E(Y I_E)}{P(E)} & w \in E \\ \frac{E(Y I_{E'})}{P(E')} & w \in E' \end{cases}$$

We see that this explicit solution is used by Equation (3.1) in Definition 3.1. Similarly, Equation (3.5) of Definition 3.3 is the explicit solution when X is a discrete random variable. When X is continuous, we do not have an explicit formula for $E(Y | X)$ in general.

- Measurability guarantees that, although the estimate of Y based on the information in \mathcal{G} is itself a random variable, the value of the estimate $E(Y | \mathcal{G})$ can be determined from the information in \mathcal{G} .
- The following property might seem stronger but in fact it is equivalent to the partial averaging property

$$E(UR) = E(UY) \tag{3.11}$$

for any random variable³ $U \in \mathcal{G}$.

- By the definition, $E(Y | \mathcal{G}) \in \mathcal{G}$, and for any $G \in \mathcal{G}$

$$E(I_G \cdot E(Y | \mathcal{G})) = E(I_G \cdot Y) \quad (3.12)$$

- If \mathcal{G} is the event space generated by some other random variable X , i.e., $\mathcal{G} = \sigma(X)$, we generally write $E(Y | X)$ rather than $E(Y | \sigma(X))$.
- It is true that given σ -algebras \mathcal{G} in \mathcal{F} , there exists random variable X , such that $\mathcal{G} = \sigma(X)$ in the sense that they are same up to null sub-sets (For any $G \in \mathcal{G}$, there is $E \in \sigma(X)$, such that $P(G \setminus E) + P(E \setminus G) = 0$ and vice versa).

A: Existence and Uniqueness

The existence and uniqueness of the $E(Y | \mathcal{G})$ come from the following proposition.

Proposition 3.13: $E(Y | \mathcal{G})$ exists and is unique (in the almost surely sense).

Proof. By Corollary 3.20, there exists a random variable R

$$E(I_G Y) = E(I_G R) \quad \forall G \in \mathcal{G}$$

The uniqueness follows from Lemma 3.10, or from the uniqueness of the Radon-Nikodým derivative, up to equivalence. Recall that real-valued random variables Y and X are equivalent if $P(Y = X) = 1$, or $X = Y$ almost surely (a.s.). \square

B: Conditional Probability

The *conditional probability* of an event $E \in \mathcal{F}$ given a σ -algebra \mathcal{G} is defined by

$$P(E | \mathcal{G}) \equiv E(I_E | \mathcal{G}) \quad (3.13)$$

Then, Equation (2.18) is a special case when $\mathcal{G} = \{\emptyset, \Omega\}$ is the smallest event space. Needless to say, Equation (3.13) implies that for any $F \in \mathcal{G}$

$$P(E | F) = E(I_E | F)$$

which incorporates Equation (3.3) and (3.10).

For the conditional probability of E given F , $P(E | F)$ defined in Equation (2.1), the restriction $P(F) > 0$ is needed. However, in the generalized form of Equation (3.13), this restriction is removed, **we can compute $P(E | F)$ even $P(F) = 0$** . For example, let $(X, Y) \sim N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, and

$$E = \{Y \leq \mu_Y + (x - \mu_X)\rho\sigma_Y/\sigma_X\}$$

$$F = \{X = x\}$$

then $P(F) = 0$. However, by Equation (2.26), $P(E | F) = \frac{1}{2}$.

³A random variable is approximated by a simple function. And a simple function is a finite sum of indicator functions times constants.

By partial averaging, for any $G \in \mathcal{G}$

$$E(I_G P(E | \mathcal{G})) = E(I_G I_E) = P(GE) \tag{3.14}$$

As a consequence, for any discrete random variable Z

$$E(Z | \mathcal{G}) = \sum_i z_i P(Z = z_i | \mathcal{G}) \tag{3.15}$$

if all $P(Z = z_i | \mathcal{G})$ are known.

Proposition 3.14: If the conditional probability of an event $E \in \mathcal{F}$ given a σ -algebra $\mathcal{G} \subset \mathcal{F}$ is a constant, $P(E | \mathcal{G}) = c$, then

1. $E \perp \mathcal{G}$ (for any $G \in \mathcal{G}$, there is $E \perp G$)
2. $I_E \perp \mathcal{G}$ (which means $\sigma(I_E) \perp \mathcal{G}$)

3.2.3 General Properties

Assume that a, b are arbitrary real numbers, X, Y are integrable random variables on a probability space (Ω, \mathcal{F}, P) and \mathcal{G}, \mathcal{H} are σ -algebra on Ω contained in \mathcal{F} . Then conditional expectation has the following properties: All equalities and the inequalities hold almost surely under probability measure P .

1. Linearity

$$E(aX + bY | \mathcal{G}) = aE(X | \mathcal{G}) + bE(Y | \mathcal{G})$$

2. Taking out what is known (conditional determinism): If X is \mathcal{G} -measurable

$$E(XY | \mathcal{G}) = X E(Y | \mathcal{G}) \quad X \in \mathcal{G}$$

3. Dropping independent information (“eat independence”): If Y is independent of \mathcal{G} ($\sigma(Y) \perp \mathcal{G}$)

$$E(Y | \mathcal{G}) = E(Y) \quad Y \perp \mathcal{G}$$

4. Tower property (iterated conditioning, “small eat large”): If $\mathcal{H} \subset \mathcal{G}$

$$E(E(Y | \mathcal{G}) | \mathcal{H}) = E(Y | \mathcal{H}) \quad \mathcal{H} \subset \mathcal{G}$$

and for $E(Y | \mathcal{H})$ is \mathcal{H} -measurable, thus \mathcal{G} -measurable, by taking out what is known

$$E(E(Y | \mathcal{H}) | \mathcal{G}) = E(Y | \mathcal{H}) \quad \mathcal{H} \subset \mathcal{G}$$

5. Positivity: If $Y \geq 0$, then $E(Y | \mathcal{G}) \geq 0$. And if $Y > 0$, then $E(Y | \mathcal{G}) > 0$. Note that $Y \geq 0 \implies E(Y) > 0$, but $Y \geq 0 \implies E(Y | \mathcal{G}) \geq 0$.

Remark: The linearity and positivity make the operator of conditional expectation the natural choice for the pricing function in financial market.

- Taking out what is known: For $X \in \sigma(X)$, we have

$$E(XY | X) = X E(Y | X)$$

The idea here is that we are treating X as a constant, since the information of X is given.

- If X is \mathcal{G} -measurable, then $h(X) \in \mathcal{G}$. Since $h(X)$ is determined by information in \mathcal{G}

$$E(h(X)Y | \mathcal{G}) = h(X) E(Y | \mathcal{G})$$

thus conditioning on the event space \mathcal{G} , $h(X)$ can be treated as a constant.

- There are some special and useful forms: If Y is \mathcal{G} -measurable, then (Exercise 3.15)

$$E(Y | \mathcal{G}) = Y \quad Y \in \mathcal{G}$$

In particular, $E(c | \mathcal{G}) = c$ for any constant c .

- Dropping independent information: If $X \perp Y$, then $E(Y | X) = E(Y)$. What is more, see Exercise 3.17:

$$X \perp (Y, Z) \implies E(Y | X, Z) = E(Y | Z)$$

Where the information of X is independent and is dropped out.

- When there are $Y \perp X$ and $Y \perp Z$, there is nothing to drop out in $E(Y | X, Z)$. In Example ??, there are $Y \perp X$ and $Y \perp Z$, however

$$E(Y | X, Z) \neq E(Y) = E(Y | Z) = E(Y | X) = 0$$

- Note that if $X \perp (Y, Z)$, then $X \perp Y | Z$ and thus (Exercise 3.18)

$$X \perp (Y, Z) \implies X \perp Y | Z \implies E(XY | Z) = E(X)E(Y | Z)$$

- Positivity: If $Y \geq 0$, we have $E(Y | \mathcal{G}) \geq 0$, not $E(Y | \mathcal{G}) > 0$. For example, tossing a fair dice and let $Y = 1_{w>4}$ and $X = 1_{w<4}$, there are $E(Y | X = 1) = 0$ and $E(Y | X = 0) = 2/3$. Moreover, $E(Y | Y = 0) = 0$ by partial averaging or substitution rule. Clearly, for any non-null $G \in \mathcal{G}$, if $Y(G) \geq 0$, then $E(Y | \mathcal{G}) > 0$.

We have known that $E(Y | \mathcal{G})$ is a random variable defined on the coarser probability space (Ω, \mathcal{G}, P) and thus on (Ω, \mathcal{F}, P) . Because of the randomness, conditional Jensen's inequality holds almost surely, while Jensen's inequality holds exactly.

Theorem 3.15 (Conditional Jensen's Inequality): *Let $h : \mathbb{R} \rightarrow \mathbb{R}$ be a convex function and let Y be an integrable random variable on a probability space (Ω, \mathcal{F}, P) such that $h(Y)$ is also integrable. Then*

$$h(E(Y | \mathcal{G})) \leq E(h(Y) | \mathcal{G}) \quad a.s.$$

for any σ -algebra \mathcal{G} on Ω contained in \mathcal{F} .

Proof. For $R = E(Y | \mathcal{G})$ takes values in \mathbb{R} , let $y = h(R) + k(x - R)$ be a random supporting line for $h(\cdot)$ at R , then $(X \geq Y$ means $X(w) \geq Y(w)$ for each $w \in \Omega$)

$$h(Y) \geq h(R) + k(Y - R)$$

and $k \in \mathcal{G}$ since k is depend on R . Taking conditional expectations given \mathcal{G} through the inequality gives

$$\begin{aligned} E(h(Y) | \mathcal{G}) &\geq h(R) + E(k(Y - R) | \mathcal{G}) = h(R) + k E(Y - R | \mathcal{G}) \\ &= h(R) = h(E(Y | \mathcal{G})) \end{aligned} \quad \square$$

A: Tower Property

The tower property is also known as smoothing property of conditional expectation, or consistency conditions. As a special case of tower property when $\mathcal{H} = \{0, \Omega\}$, we have the *law of total expectation*

$$E(E(Y | \mathcal{G})) = E(Y)$$

Which also follows by putting $E = \Omega$ in Eq (3.12).

When conditioning twice, with respect to nested σ -algebras, the smaller one (representing the smaller amount of information) always prevails. As seen in Figure 3.2:

$$E(Y | X) = \begin{cases} \frac{2}{27} & w \in [0, \frac{1}{3}] = \{X = 1\} \\ \frac{14}{27} & w \in (\frac{1}{3}, \frac{2}{3}] = \{X = 2\} \\ \frac{38}{27} & w \in (\frac{2}{3}, 1] = \{X = 0\} \end{cases}$$

and

$$E(E(Y | Z) | X) = E(E(Y | X) | Z) = E(Y | Z) = \begin{cases} \frac{2}{27} & w \in [0, \frac{1}{3}] = \{Z = 1\} \\ \frac{26}{27} & w \in (\frac{1}{3}, 1] = \{Z = 0\} \end{cases}$$

Which confirms the tower property as $\sigma(Z) < \sigma(X)$. In the region of $(\frac{1}{3}, 1] = Z^{-1}(0) = X^{-1}\{0, 2\}$

1. If it is first smoothed by Z , $E(Y | Z = 0) = \frac{26}{27}$ is a constant, a second finer-region smooth by X is invisible, due to the uneven has been levelled
2. If it is first smoothed by X , we have $E(Y | X = 0) = \frac{38}{27}$ and $E(Y | X = 2) = \frac{14}{27}$. A second full-region smooth by Z further irons the region out

the smaller information set applies a larger “iron”, the work of a smaller iron was concealed or wiped out by the larger iron. Thus, when nested conditioning, only the result from the smaller information set is revealed. It is now apparent that the tower property results from the partial averaging property of conditional expectation.

The tower property is not only useful in theorem proofs, but also helpful in applications. For discrete random variables X, Y and Z , the specific form of

$$E(Y | X) = E(E(Y | X, Z) | X) \tag{3.16}$$

is (recall Equation 2.6)

$$E(Y | X = x) = \sum_{k \in K} E(Y | X = x, Z = z_k)P(Z = z_k | X = x) \tag{3.17}$$

where $K = \{k : P(X = x, Z = z_k) > 0\}$. For brevity, we write

$$E(Y | X) = \sum_k E(Y | X, Z = z_k)P(Z = z_k | X)$$

Furthermore, if $X \perp Z$

$$E(Y | X) = \sum_k E(Y | X, Z = z_k)P(Z = z_k)$$

For continuous random variables X, Y and Z , with joint density $f(x, y, z)$, Eq (3.16) leads to

$$\int yf(y | x) dy = \int E(Y | X = x, Z = z)f(z | x) dz$$

Example 3.2.1: Let us consider independent Bernoulli trials each of which is a success with probability p . We repeat these trials up to the point where n consecutive successes appear for the first time. We are looking for the expected number of necessary trials for that.

We define the random variable Y_n to denote the number of necessary trials to obtain n consecutive successes. Let Z_k be the Bernoulli random variables for k -th trial and $m = Y_{n-1} + 1$. Then

$$(Y_n | Y_{n-1} = y, Z_m = 1) = y + 1$$

we have $E(Y_n | Y_{n-1} = y, Z_m = 1) = y + 1$ and

$$E(Y_n | Y_{n-1}, Z_m = 1) = Y_{n-1} + 1$$

For $(Y_n | Y_{n-1} = y, Z_m = 0) \sim Y_n + y + 1$, we obtain

$$E(Y_n | Y_{n-1}, Z_m = 0) = E(Y_n) + Y_{n-1} + 1$$

Thus, by $E(Y_n | Y_{n-1}) = E(E(Y_n | Y_{n-1}, Z_m) | Y_{n-1})$, and $Z_m \perp Y_{n-1}$

$$\begin{aligned} E(Y_n | Y_{n-1}) &= \sum_{i=0}^1 E(Y_n | Y_{n-1}, Z_m = i) P(Z_m = i | Y_{n-1}) \\ &= E(Y_n | Y_{n-1}, Z_m = 0)(1 - p) + E(Y_n | Y_{n-1}, Z_m = 1)p \\ &= (E(Y_n) + Y_{n-1} + 1)(1 - p) + (Y_{n-1} + 1)p \\ &= E(Y_n)(1 - p) + Y_{n-1} + 1 \end{aligned}$$

By law of total expectation

$$E(Y_n) = E(E(Y_n | Y_{n-1})) = E(Y_n)(1 - p) + E(Y_{n-1}) + 1$$

which yields

$$E(Y_n) = E(Y_{n-1})/p + 1/p$$

Clearly, the random variable Y_1 represents the number of Bernoulli trials up to the first success, and consequently, has the geometric distribution with parameter p , thus

$$E(Y_1) = 1/p$$

and recursively, we arrive at

$$E(Y_n) = 1/p^n + \dots + 1/p^2 + 1/p = \frac{1 - p^n}{p^n(1 - p)}$$

B: Change of Measure

For any random variable Y , we have $E^Q(Y) = E(GY)$ if $dQ = GdP$ where G is the Radon-Nikodým derivative of Q with respect to P . However, $E^Q(Y | \mathcal{G}) \neq E(GY | \mathcal{G})$, we have the following Equation (3.18) for conditional expectation under change of measure.

Theorem 3.16: Let \mathcal{G} be a sub- σ -algebra of \mathcal{F} on which two probability measures Q and P are defined. If $dQ = GdP$ and Y is Q -integrable, then GY is P -integrable and Q -a.s

$$E^Q(Y | \mathcal{G}) = \frac{E(GY | \mathcal{G})}{E(G | \mathcal{G})} \tag{3.18}$$

Remark: Here $G \geq 0$, Q and P are not necessarily equivalent. If $G > 0$, then $Q \sim P$. As a special case, let $\mathcal{G} = \{\emptyset, \Omega\}$, we have the unconditional version $E^Q(Y) = E(GY)$.

We have learned that conditional probability, $P_F(E) = P(E | F)$, is a probability measure. In Equation (2.5), since $Q(E) \equiv P_{FG}(E) = P(E | FG)$ is a probability measure different from P , given event F and G with $P(FG) > 0$, there is $Q(EG) = Q(E)$, $Q(G) = 1$, and

$$Q(E | G) = \frac{Q(EG)}{Q(G)} = Q(E) = P(E | FG)$$

Let $G = \frac{1_{FG}}{P(FG)}$, $Y = 1_E$ and $\mathcal{G} = \sigma(1_G)$, then $dQ = GdP$

$$E^Q(Y | \mathcal{G}) = E^Q(1_E | 1_G) = \begin{cases} Q(E | G) & w \in G \\ Q(E | G') & w \in G' \end{cases}$$

and

$$E(GY | \mathcal{G}) = E(1_{FG}1_E | 1_G)/P(FG) = \begin{cases} P(EF | G)/P(FG) & w \in G \\ P(EF | G')/P(FG) & w \in G' \end{cases}$$

$$E(G | \mathcal{G}) = E(1_{FG} | 1_G)/P(FG) = \begin{cases} P(F | G)/P(FG) & w \in G \\ P(F | G')/P(FG) & w \in G' \end{cases}$$

By Equation (3.18), when $w \in G$, there is $Q(E | G) = \frac{P(EF | G)}{P(F | G)}$. Thus

$$P(E | FG) = Q(E | G) = \frac{P(EF | G)}{P(F | G)}$$

We have shown that Equation (2.5) is a special case of Equation (3.18).

Remark: Since $G' < (FG)'$, $Q(G') = 0$, we can not compute $Q(E | G')$ in elementary probability context. However, as an implication of Equation (3.18), when $w \in G'$, there is

$$Q(E | G') = \frac{P(EF | G')}{P(F | G')} \tag{3.19}$$

Which shows that we can define a conditional probability on a null event. Note that $Q(E | G')$ depends on event F in Equation (3.19), where event F can be arbitrary as long as $0 < P(FG) < 1$, thus it is not unique. However, since $Q(G') = 0$, there causes no trouble for Equation (3.18) in the sense of Q -a.s.

Corollary 3.17: If Y_t is an adapted stochastic process (i.e. $Y_t \in \mathbb{I}_t$), let $0 \leq t \leq u \leq T$, then

$$E_t^Q(Y_u) = \frac{1}{G_t} E_t(G_u Y_u) \tag{3.20}$$

where $G_t = E(G | \mathbb{I}_t)$ is called the Radon-Nikodým derivative process.

C: Pricing Function

For continuous time or T -step discrete time model, if the market is free of arbitrage, there is an equivalent martingale measure (risk-neutral measure) such that⁴

$$X_t = \wp_{t,u}(X_u) = B_t E_t^Q(X_u/B_u) = \wp_{t,T}(X_T) = B_t E_t^Q(X_T/B_T) \quad 0 \leq t \leq u \leq T \quad (3.21)$$

for any primary asset when the number of primary assets is finite. Given any attainable payoff $X_T \in \mathbb{I}_T$ at time T , the pricing function is represented by (Equation 1.17)

$$X_0 = \wp(X_T) = E(\Psi X_T)$$

where the SDF $\Psi \in \mathbb{I}_T$ is a positive random variable. Let $G = \Psi B_T/B_0 > 0$, then $dQ = GdP$, and

$$X_t = \wp_{t,u}(X_u) = B_t E_t^Q(X_u/B_u) = B_t \frac{E_t(GX_u/B_u)}{E_t(G)} = E_t(G_u X_u/B_u) B_t/G_t$$

where $G_t = E(G | \mathbb{I}_t) = E_t(\Psi B_T)$ is the Radon-Nikodým derivative process. For the forward measure, since $dQ = B_T D_{0,T} dO$, we have

$$X_t = \wp_{t,u}(X_u) = D_{t,T} E_t^O(X_u/D_{u,T}) = \wp_{t,T}(X_T) = D_{t,T} E_t^O(X_T) \quad 0 \leq t \leq u \leq T \quad (3.22)$$

3.2.4 The Best Predictor

We have said that $E(Y | X)$ is a random variable measurable to $\sigma(X)$, thus it is a function of X . In this subsection, we will pursue a deeper interpretation, $E(Y | X)$ as a “best approximation” of Y by a function of X . For this purpose, we assume that the random variables have finite variance.

If Y is a random variable then the ordinary expected value $E(Y)$ represents our best guess of the value of Y if we have no prior information. But now suppose that X is a random variable that is not independent of Y and that we can observe the value of X . Then one might expect that we can use that extra information to our advantage in guessing where Y will end up.

A: Best Mean Square Predictor

One can show that the conditional expectation $E(Y | \mathcal{G})$ is the best mean square \mathcal{G} -measurable predictor of Y :

$$E([Y - E(Y | \mathcal{G})]^2) = \min_{Z \in \mathcal{G}} E((Y - Z)^2)$$

If $\sigma(X) = \mathcal{G}$, then $E(Y | X)$ is best mean square predictor of Y among all functions of X . This is fundamentally important in econometric problems where the predictor X can be observed but not the response variable Y . Thus, the regression function of Y on X is defined by the conditional expectation function, $r(x) = E(Y | X = x)$.

⁴Like multi-step binomial model and Black-Scholes model, we can setup a model such that the market is free of arbitrage. For discrete models with finite number of primary assets and finite number of steps, Equation (3.21) is read from Harrison and Pliska (1981) and Dalang et al. (1990). For continuous time model with finite number of primary assets, it is straightforward from Girsanov's Theorem.

B: Constant and Linear Predictor

For random variable Y , since

$$E([Y - E(Y)]^2) = \min_{y \in \mathbb{R}} E((Y - y)^2)$$

the best constant predictor of Y is $E(Y)$ in the mean square sense, with mean square error $\text{var}(Y)$. The best predictor of Y among the linear functions of X is the *linear projection*⁵ of Y on 1 and X

$$\begin{aligned} P_j(Y | 1, X) &\equiv E \left(Y \begin{bmatrix} 1 \\ X \end{bmatrix}' \right) \left[E \left(\begin{bmatrix} 1 \\ X \end{bmatrix} \begin{bmatrix} 1 \\ X \end{bmatrix}' \right) \right]^{-1} \begin{bmatrix} 1 \\ X \end{bmatrix} \\ &= E(Y) + \frac{\text{cov}(X, Y)}{\text{var}(X)}(X - E(X)) \end{aligned}$$

for which minimizes

$$\min_{a, b \in \mathbb{R}} E([Y - (a + bX)]^2)$$

with mean square error $\text{var}(Y)(1 - \rho_{XY}^2)$. If X and Y follow a bivariate normal distribution, from Eq (2.27), we see that the best linear predictor equals the conditional expectation $E(Y | X)$, the linear regression. Please note that if the regression function $E(Y | X)$ is not linear, then the linear regression model $Y = \alpha + \beta X + \epsilon$ is misspecified, and the OLS (ordinary least square) recovers just the linear projection coefficients⁶ in $P_j(Y | 1, X) = a + bX$ as the sample size grows.

⁵The linear projection of Y on 1 and X satisfies the orthogonality conditions

$$E(1 \cdot \epsilon) = E(X \cdot \epsilon) = 0$$

where $\epsilon = Y - P_j(Y | 1, X)$. If we think of $P_j(Y | 1, X)$ as an approximation of Y , then ϵ is the error in that approximation. The orthogonality conditions state that the error is orthogonal to the plane spanned by 1 and X .

⁶Which could still be useful, because it is the mean square error minimising linear approximation of the conditional expectation function. Let $Y = a + bX + u$ and $E(u | X) \neq 0$, but $E(1 \cdot u) = E(X \cdot u) = 0$ (orthogonality), then OLS estimators are consistent.

§ 3.3 Conditional Information

conditional independence, conditional correlation

In bull market, the returns of some assets are seen to be independent, however, in bear market, they are usually correlated, and moving down in steps.

students play football together, and go home individually after school.

The more you know, the better your decision

3.3.1 Conditional Independence

The definition of conditional probability in Eq (3.13) generalizes the definition in Eq (2.1), thus, we are able to compute the conditional probability when the given event has zero probability. Similarly, conditional independence can be generalized by **simply drop the requirement that the given event has positive probability**. We say events G and H are conditionally independent given K , denoted by $G \perp H | K$, if

$$P(GH | K) = P(G | K)P(H | K)$$

We say events G and H are *conditionally independent* given σ -algebra \mathcal{K} , denoted by $G \perp H | \mathcal{K}$, if

$$P(GH | \mathcal{K}) = P(G | \mathcal{K})P(H | \mathcal{K})$$

that is, for any $K \in \mathcal{K}$, we have $G \perp H | K$.

A: Definition

We are ready to modify the Definition 2.6 to incorporate the conditional information.

Definition 3.18: We say that σ -algebras \mathcal{G} and \mathcal{H} in \mathcal{F} are conditionally independent given σ -algebra \mathcal{K} , denoted by $\mathcal{G} \perp \mathcal{H} | \mathcal{K}$, if for any $G \in \mathcal{G}$, $H \in \mathcal{H}$, and $K \in \mathcal{K}$, the events G and H are conditionally independent given K .

We can now say that random variables X and Y are conditionally independent given Z , denoted by $X \perp Y | Z$, if and only if the σ -algebras $\sigma(X)$ and $\sigma(Y)$ are conditionally independent given $\sigma(Z)$.

This definition says that

$$X \perp Y | Z \iff \sigma(X) \perp \sigma(Y) | \sigma(Z) \iff E \perp F | G$$

for any $E \in \sigma(X)$, $F \in \sigma(Y)$, and $G \in \sigma(Z)$. Or more practically, some textbooks use the following definition

$$X \perp Y | Z \iff P(EF | Z) = P(E | Z)P(F | Z)$$

for any $E \in \sigma(X)$ and $F \in \sigma(Y)$.

Remark: Intuitively, if Y and X are conditionally independent given Z , then knowing Z renders X statistically irrelevant for predicting Y .

- Conditional independence is symmetric, that is

$$Y \perp X | Z \iff X \perp Y | Z$$

- Random variables Y and X are independent given Z , if and only if

$$E(g(X)h(Y) | Z) = E(g(X) | Z) E(h(Y) | Z) \tag{3.23}$$

for all choices of bounded (Borel measurable) functions $g, h : \mathbb{R} \rightarrow \mathbb{R}$. Thus, if $Y \perp X | Z$, we have

$$E(YX | Z) = E(Y | Z) E(X | Z)$$

B: Conditional Probability Density Function

If X, Y and Z are jointly continuous with density $f(x, y, z)$, then random variables Y and X are said to be conditionally independent given Z , written $X \perp Y | Z$, if and only if, for all x, y and z with $f_Z(z) > 0$

$$f(x, y | z) = f(x | z)f(y | z)$$

where functions $f(\cdot | z)$ are conditional probability density functions. If the random variables are jointly discrete with probability mass function $f(x, y, z) = P(X = x, Y = y, Z = z)$, the formulas are same as the continuous case, with $f(\cdot | \cdot)$ representing conditional probability mass function. For example

$$f(y | x, z) = P(Y = y | X = x, Z = z)$$

The other useful ways to define conditional independence are

$$X \perp Y | Z \iff f(x | y, z) = f(x | z) \iff f(y | x, z) = f(y | z)$$

These forms are directly related to the widely used notion of Markov property: If we interpret Y as the “future,” X as the “past,” and Z as the “present,” $Y \perp X | Z$ says that, given the present the future is independent of the past; this is known as Markov property.

Remark: X and Y are conditionally independent given Z if and only if, given any value of Z , the probability distribution of X is the same for all values of Y and the probability distribution of Y is the same for all values of X .

- For conditional PDF and PMF, we have

$$f(x, y | z) = f(x | z)f(y | z) \iff f(y | x, z) = f(y | z)$$

However, if $Y \perp X | Z$, we have $[f(x, y | z) \rightarrow E(XY | Z)]$

$$E(XY | Z) = E(X | Z) E(Y | Z)$$

and

$$E(Y | X, Z) = E(Y | Z)$$

Note that $E(Y | X, Z)$ is not affected by X , for $E(Y | X, Z) = E(Y | Z)$ is a function of Z only. In particular,

$$E(Y | X, Z = z) = E(Y | Z = z)$$

is constant over X , the information of X is dropped out.

- Without reference to conditional distributions

$$X \perp Y | Z \iff f(x, y, z)f_Z(z) = f(x, z)f(y, z)$$

- Using conditional cumulative distribution function

$$F(x, y | z) = P(X \leq x, Y \leq y | Z = z)$$

$$F(y | x, z) = P(Y \leq y | X = x, Z = z)$$

the following version is seemingly seemingly weaker than the general definition but equivalent definition

$$Y \perp X | Z \iff F(x, y | z) = F(x | z)F(y | z)$$

$$\iff F(x | y, z) = F(x | z) \iff F(y | x, z) = F(y | z)$$

for any x, y and z .

For events E, F and G , the examples 2.1.8 and 2.1.9 show that $E \perp F$ neither implies nor is implied by $E \perp F | G$. For random variables X, Y and Z , it follows that $X \perp Y$ neither implies nor is implied by $X \perp Y | Z$.

Example 3.3.1: For discrete random variables X, Y and Z , assume $P(Z = 1) = P(Z = 0) = 1/2$, and $f(x, y | z)$ set by the following table

		$Z = 1$			$Z = 0$				
		$Y = 0$	$Y = 1$	$f_{X Z}$	$Y = 0$	$Y = 1$	$f_{X Z}$		
X	0	1/12	1/4	1/3	X	0	3/10	9/20	3/4
	1	1/6	1/2	2/3		1	1/10	3/20	1/4
$f_{Y Z}$		1/4	3/4		$f_{Y Z}$		2/5	3/5	

Show that $X \perp Y | Z$, and verify $E(Y | Z = 1) = E(Y | X, Z = 1)$; However, $X \perp Y$ is not true.

For any $z \in \{0, 1\}$, it is trivial to check that $f(x, y | z) = f(x | z)f(y | z)$, thus $X \perp Y | Z$.

As a partial verification of $f(y | x, z) = f(y | z)$ (X is dropped out if $X \perp Y | Z$), we see

$$P(Y = 1 | X = 0, Z = 1) = \frac{f(x = 0, y = 1 | z = 1)}{f(x = 0 | z = 1)} = \frac{1/4}{1/3} = \frac{3}{4}$$

$$P(Y = 1 | X = 1, Z = 1) = \frac{f(x = 1, y = 1 | z = 1)}{f(x = 1 | z = 1)} = \frac{1/2}{2/3} = \frac{3}{4}$$

$$P(Y = 1 | Z = 1) = f(y = 1 | z = 1) = \frac{3}{4}$$

thus

$$\begin{aligned} E(Y | X = 0, Z = 1) &= \sum_{y=0,1} yP(Y = y | X = 0, Z = 1) \\ &= P(Y = 1 | X = 0, Z = 1) = \frac{3}{4} \end{aligned}$$

§ 3.4 Exercise

3.1 Show that

$$E(Y | \Omega) = E(Y)$$

3.2 Prove Eq (3.2).

3.3 Show that if X is a constant function, then $E(Y | X)$ is constant and equal to $E(Y)$.

3.4 Show that if $0 < P(F) < 1$

$$E(I_E | I_F) = \begin{cases} P(E | F) & w \in F \\ P(E | F') & w \notin F \end{cases}$$

3.5 If X is a discrete random variable, is it true that the random variable $E(Y | X)$ takes the value $E(Y | X = x_i)$ with probability $P(X = x_i)$?

3.6 In elementary probability theory, we know that for continuous random variables X and Y

$$\begin{aligned} r(x) &= E(Y | X = x) \\ &= \int_{-\infty}^{+\infty} yf(y | x) dy \end{aligned}$$

and $E(Y | X) = r(X)$. Please show that this definition is consistent with our Definition 3.9 of conditional expectation.

3.7 Prove Equation (3.7) and (3.8) in Example 3.1.6.

3.8 Given random variables X, Y and Z . For any $G \in \sigma(Z)$

$$\begin{aligned} &E(I_G \cdot E(X | Z) E(Y | Z)) \\ &= E(I_G E(X | Z) \cdot E(Y | Z)) \\ &= E(I_G E(X | Z) \cdot Y) \\ &= E(I_G \cdot Y E(X | Z)) \end{aligned}$$

thus, by the definition of conditional expectation

$$\begin{aligned} Y E(X | Z) &= E(E(X | Z) E(Y | Z) | Z) \\ &= E(X | Z) E(Y | Z) \end{aligned}$$

Is it correct or not? Why?

3.9 Show that if $\mathcal{G} = \{\emptyset, \Omega\}$, then $E(Y | \mathcal{G}) = E(Y)$.

3.10 Prove Equation (3.15).

3.11 Show that if $E \in \mathcal{G}$, then $E(E(Y | \mathcal{G}) | E) = E(Y | E)$.

3.12 Show that if $Y \geq 0$, then $E(Y | \mathcal{G}) \geq 0$.

3.13 Show that for any event $E \in \mathcal{G}$, $Y - E(Y | \mathcal{G})$ is orthogonal to I_E .

3.14 Show that for any random variable $X \in \mathcal{G}$, $Y - E(Y | \mathcal{G})$ is orthogonal to X .

3.15 If Y is \mathcal{G} -measurable, then $E(Y | \mathcal{G}) = Y$.

3.16 If X and Y are independent, \mathcal{G} and \mathcal{H} are independent. Then

$$\begin{aligned} E(E(XY | \mathcal{G}) | \mathcal{H}) &= E(E(XY | \mathcal{H}) | \mathcal{G}) \\ &= E(X) E(Y) \end{aligned}$$

3.17 If $X \perp (Y, Z)$, then $E(Y | X, Z) = E(Y | Z)$.

3.18 If $X \perp (Y, Z)$, then $E(XY | Z) = E(X) E(Y | Z)$.

3.19 In Example 3.1.5, if we let N denote the number of spins in a game, X_i the prize in i th spin, and $Y_n = \sum_{i=1}^n X_i$. Then Y_N is a random sum of random variables, it is the total prize for the game. Find $E(N)$, $E(Y_N | N = n)$ and show that $E(Y_N) = E(N) E(X_i)$.

3.20 In Example 3.2.1, let N denote the number of trials until the first occurrence of failure. Find $E(Y_n | N = i)$ and then compute $E(Y_n)$ by $E(E(Y_n | N))$.

3.21 Prove Equation (3.17).

3.22 Show that $E(Y | X)$ is best mean square predictor of Y among all functions of X .

3.23 Independent but not conditionally independent

dent: Suppose $X \perp Y$, each taking the values 0 and 1 with probability 0.5. Let $Z = XY$, then $X \perp Y | Z$ is false.

- 3.24 Conditionally independent but not independent: Suppose Z is 0 with probability 0.5 and 1 otherwise. When $Z = 0$ take X and Y to be independent, each having the value

0 with probability 0.9 and the value 1 otherwise. When $Z = 1$, X and Y are again independent, but this time they take the value 1 with probability 0.9 and the value 0 otherwise. Show that X and Y are conditionally dependent given Z , but X and Y are not independent.

§ 3.5 Appendix

We study the conditional expectation of Y given X , which is a concept of fundamental importance in probability, and a vital key to understand the theory of finance.

3.5.1 Review

We will use the two envelopes problem, also known as the exchange paradox, to review concepts and methods in probability theory.

Exchange Paradox (Wikipedia's basic setup): You are given two indistinguishable envelopes, each of which contains a positive sum of money. One envelope contains twice as much as the other. You may pick one envelope and keep whatever amount it contains. You pick one envelope at random but before you open it you are given the chance to take the other envelope instead.

The switching argument: Now suppose you reason as follows:

1. I denote by X the amount in my selected envelope.
2. The probability that X is the smaller amount is $1/2$, and that it is the larger amount is also $1/2$.
3. The other envelope may contain either $2X$ or $X/2$.
4. If X is the smaller amount, then the other envelope contains $2X$.
5. If X is the larger amount, then the other envelope contains $X/2$.
6. Thus the other envelope contains $2X$ with probability $1/2$ and $X/2$ with probability $1/2$.
7. So the expected value of the money in the other envelope is:

$$2X \cdot \frac{1}{2} + \frac{X}{2} \cdot \frac{1}{2} = \frac{5}{4}X$$

8. This is greater than X , so I gain on average by swapping.
9. After the switch, I can denote that content by Y and reason in exactly the same manner as above.

I will conclude that the most rational thing to do is to swap back again. To be rational, I will thus end up swapping envelopes indefinitely. As it seems more rational to open just any envelope than to swap indefinitely, we have a contradiction. What has gone wrong?

The fallacy is the improper use of a symbol to denote at the same time a random variable and its realizations, say, X for random variable $X(w)$ and numbers like $X(H)$ or $X(L)$. In step 1 and 2, X is a random variable, from step 3 to 6 take X as a realization. Let $X(L) = a$ be the lower value of the two amounts, then $X(H) = 2a$, $Y(L) = 2a$ and $Y(H) = a$.

- Step 3 switch the meaning of X from a random variable to a realization. The expression $Y = 2X$ should be $Y(L) = 2X(L) = 2a$, and $Y = X/2$ should be $Y(H) = X(H)/2 = a$. If X is a random variable, step 3 should be: The other envelope may contain $Y = 3a - X$.
- If the $2X$ in step 4 is read as conditional mean, say, $E(Y | X < Y) = 2X$, it makes no sense. For the conditional expectation of Y given an event $\{X < Y\}$ is a number, not a random variable. In fact, $E(Y | X < Y) = E(Y | X = a) = 2a$.

- Step 6 should be: the other envelope contains $Y(L) = 2X(L) = 2a$ with probability $1/2$ and $Y(H) = X(H)/2 = a$ with probability $1/2$.
- If X is a random variable, $\frac{5}{4}X$ is a random variable in step 7, can not be an expected value, for expected value is a number. If X is a real number (realization), the first X is lower value $X(L) = a$ and the second X is higher value $X(H) = 2a$. They do not refer to the same thing, we are adding apples and oranges.

Based on the above analysis, the expected value of the money in the other envelope is:

$$Y(H) \cdot \frac{1}{2} + Y(L) \cdot \frac{1}{2} = 2X(L) \cdot \frac{1}{2} + \frac{X(H)}{2} \cdot \frac{1}{2} = \frac{3}{2}a = E(X)$$

If the amount in the other envelope on step 4 and 5 is interpreted as conditional expectation, the corrections should be: Let the amount be a and $2a$, for $X + Y = 3a$, step 4 reads

$$E(Y | X = a) = E(3a - X | X = a) = E(3a - a | X = a) = 2a$$

Similarly, step 5 reads $E(Y | X = 2a) = a$. Thus step 7 says

$$\begin{aligned} E(Y) &= E(E(Y | X)) \\ &= E(Y | X = a)P(X = a) + E(Y | X = 2a)P(X = 2a) \\ &= 2a \cdot \frac{1}{2} + a \cdot \frac{1}{2} = \frac{3}{2}a = E(X) \end{aligned}$$

Discuss: If you open the envelope, and find that the amount is x , then the expected value of the money in the other envelope is:

$$2x \cdot \frac{1}{2} + \frac{x}{2} \cdot \frac{1}{2} = \frac{5}{4}x > x$$

This is greater than x , so you gain on average by swapping. False! Let the total amount be $c = 3a$, then if x is the lower value, $x = a$, and if x is the higher value, $x = 2a$. Both a and $2a$ are denoted by x , the first x in $2x \cdot \frac{1}{2} + \frac{x}{2} \cdot \frac{1}{2}$ is equal to a , and the second x is $2a$. Thus, the expression $2x \cdot \frac{1}{2} + \frac{x}{2} \cdot \frac{1}{2}$ exactly means

$$(2 \cdot a) \cdot \frac{1}{2} + \left(\frac{1}{2} \cdot 2a\right) \cdot \frac{1}{2} = \frac{3}{2}a = \frac{1}{2}c$$

The expected value $\frac{1}{2} \cdot 2a + \frac{1}{2} \cdot a = \frac{3}{2}a = \frac{1}{2}c$ is the same for both envelopes.

If you open the selected envelope and know the amount in the envelope, no information is added, the expected money of the other envelope is still an unconditional expectation. For you still do not know the state, say, the amount is a lower value or a higher value. It is the information (state of world) matters not the values taken by random variable in the switching decision. No matter what the amount is, once the state is known, your action is clear: If you know it is the lower value, sure you should switch, and if it is the higher value, you should not switch.

To end this exchange paradox, I would like to emphasize the following points:

1. The key message is the state of world revealed by the random variables, not the values taken by them.
2. Variables should be clearly defined, otherwise, we may easily get lost.

then

$$E(1_{X=x_i}R) = E(1_{X=x_i} E(Y | X = x_i)) = E(1_{X=x_i}) E(Y | X = x_i) = E(1_{X=x_i}Y)$$

For each non-empty $E \in \sigma(X)$ is a countable union of $\{X = x_i\}$, which are pairwise disjoint. It follows that for any $E \in \sigma(X)$, $E(1_E \cdot R) = E(1_E \cdot Y)$. \square

Lemma 3.10

Observe that $P\{Y \geq \epsilon\} = 0$ for any $\epsilon > 0$ because

$$0 \leq \epsilon P(Y \geq \epsilon) = \epsilon E(1_{Y \geq \epsilon}) = E(\epsilon 1_{Y \geq \epsilon}) \leq E(Y 1_{Y \geq \epsilon}) = 0$$

The last equality holds, since $\{Y \geq \epsilon\} \in \mathcal{G}$. Similarly, $P\{Y \leq -\epsilon\} = 0$ for any $\epsilon > 0$. As a consequence

$$P(-\epsilon < Y < \epsilon) = 1$$

for any $\epsilon > 0$. Let us put

$$E_n = \{-1/n < Y < 1/n\}$$

Then $P(E_n) = 1$ and

$$\{Y = 0\} = \bigcap_{n=1}^{\infty} E_n$$

Because the E_n form a contracting sequence of events, it follows that

$$P(Y = 0) = \lim_{n \rightarrow \infty} P(E_n) = 1$$

as required.

Proposition 3.14

For any $G \in \mathcal{G}$

$$\begin{aligned} P(EG) &= E(1_{EG}) \\ &= E(1_G 1_E) = E(1_G E(1_E | \mathcal{G})) \quad \text{partial average} \\ &= E(1_G P(E | \mathcal{G})) = E(1_G c) = c E(1_G) = cP(G) \end{aligned}$$

Let $G = \Omega$, $c = \frac{P(E\Omega)}{P(\Omega)} = P(E)$. Or $P(E) = c$ by

$$P(E) = E(1_E) = E(E(1_E | \mathcal{G})) = E(P(E | \mathcal{G})) = E(c) = c$$

Thus for any $G \in \mathcal{G}$, $P(EG) = cP(G) = P(E)P(G)$

$$\begin{aligned} P(E'G) &= P(G) - P(EG) \quad G = EG + E'G \\ &= P(G) - P(E)P(G) \\ &= (1 - P(E))P(G) = P(E')P(G) \end{aligned}$$

thus $\sigma(1_E) \perp \mathcal{G}$. $\sigma(1_E) = \{\emptyset, \Omega, E, E'\}$

A: General Properties

1. Linearity: For any $E \in \mathcal{G}$

$$\begin{aligned} & \mathbb{E}(I_E [a \mathbb{E}(X | \mathcal{G}) + b \mathbb{E}(Y | \mathcal{G})]) \\ &= \mathbb{E}(I_E a \mathbb{E}(X | \mathcal{G})) + \mathbb{E}(I_E b \mathbb{E}(Y | \mathcal{G})) = a \mathbb{E}(I_E \mathbb{E}(X | \mathcal{G})) + b \mathbb{E}(I_E \mathbb{E}(Y | \mathcal{G})) \\ &= a \mathbb{E}(I_E X) + b \mathbb{E}(I_E Y) = \mathbb{E}(a I_E X) + \mathbb{E}(b I_E Y) = \mathbb{E}(I_E (aX + bY)) \end{aligned}$$

By uniqueness this proves the desired equality.

2. Taking out what is known: For any $G \in \mathcal{G}$, if X is \mathcal{G} -measurable, then $U = I_G X \in \mathcal{G}$, and by Eq (3.11)

$$\mathbb{E}(I_G \cdot XY) = \mathbb{E}(UY) = \mathbb{E}(U \mathbb{E}(Y | \mathcal{G})) = \mathbb{E}(I_G \cdot X \mathbb{E}(Y | \mathcal{G}))$$

By uniqueness, $\mathbb{E}(XY | \mathcal{G}) = X \mathbb{E}(Y | \mathcal{G})$.

Traditional method: We first verify the result for $X = I_G$, where $G \in \mathcal{G}$. In this case

$$\mathbb{E}(I_E \cdot I_G \mathbb{E}(Y | \mathcal{G})) = \mathbb{E}(I_{EG} \mathbb{E}(Y | \mathcal{G})) = \mathbb{E}(I_{EG} Y) = \mathbb{E}(I_E \cdot I_G Y)$$

for any $E \in \mathcal{G}$, which implies that

$$\mathbb{E}(I_G Y | \mathcal{G}) = I_G \mathbb{E}(Y | \mathcal{G})$$

by uniqueness. In a similar way we obtain the result if X is a \mathcal{G} -measurable simple function

$$X = \sum_{k=1}^m a_k I_{G_k}$$

where $G_k \in \mathcal{G}$ for $k = 1, 2, \dots, m$. Finally, the result in the general case follows by approximating X by \mathcal{G} -measurable simple functions.

3. Dropping independent information: Since Y is independent of \mathcal{G} , the random variables Y and I_E are independent for any $E \in \mathcal{G}$. It follows that

$$\begin{aligned} \mathbb{E}(I_E Y) &= \mathbb{E}(I_E) \mathbb{E}(Y) && \text{independent} \\ &= \mathbb{E}(I_E \mathbb{E}(Y)) && \mathbb{E}(Y) \text{ is constant} \end{aligned}$$

which proves the assertion $\mathbb{E}(Y | \mathcal{G}) = \mathbb{E}(Y)$.

4. Tower property: Denote

$$R = \mathbb{E}(Y | \mathcal{G}) \in \mathcal{G}$$

then for every $E \in \mathcal{H}$, because $\mathcal{H} \subset \mathcal{G}$, it follows that $E \in \mathcal{G}$, thus

$$\begin{aligned} \mathbb{E}(I_E Y) &= \mathbb{E}(I_E \mathbb{E}(Y | \mathcal{G})) && \text{Eq (3.12)} \\ &= \mathbb{E}(I_E R) && R = \mathbb{E}(Y | \mathcal{G}) \\ &= \mathbb{E}(I_E \mathbb{E}(R | \mathcal{H})) && \text{Eq (3.12)} \end{aligned}$$

$\mathbb{E}(R | \mathcal{H}) \in \mathcal{H}$, by uniqueness

$$\mathbb{E}(Y | \mathcal{H}) = \mathbb{E}(R | \mathcal{H}) = \mathbb{E}(\mathbb{E}(Y | \mathcal{G}) | \mathcal{H})$$

5. Positivity: For any n we put

$$G_n = \{\mathbb{E}(Y | \mathcal{G}) \leq -1/n\}$$

for the random variable $E(Y | \mathcal{G}) \in \mathcal{G}$, we have $G_n \in \mathcal{G}$. If $Y \geq 0$ a.s., then $1_{G_n} Y \geq 0$

$$0 \leq E(1_{G_n} Y) = E(1_{G_n} E(Y | \mathcal{G})) \leq E(1_{G_n} (-1/n)) = -\frac{1}{n} E(1_{G_n}) = -\frac{1}{n} P(G_n)$$

which means that $P(G_n) = 0$. Because

$$\{E(Y | \mathcal{G}) < 0\} = \sum_{n=1}^{\infty} G_n$$

it follows that (G_n increasing)

$$P(E(Y | \mathcal{G}) < 0) = P\left(\sum_{n=1}^{\infty} G_n\right) = \lim_{n \rightarrow \infty} P(G_n) = 0$$

□

B: Change of Measure

Equation (3.18): Let $X = \frac{E(GY | \mathcal{G})}{E(G | \mathcal{G})}$, by the definition of conditional expectation given \mathcal{G} , for any $E \in \mathcal{G}$

$$\begin{aligned} E^Q(1_E X) &= E(G 1_E X) & E^Q(X) &= E(GX) \\ &= E(E(G 1_E X | \mathcal{G})) & &= E(1_E X E(G | \mathcal{G})) \\ &= E(1_E E(GY | \mathcal{G})) & &= E(E(1_E GY | \mathcal{G})) \\ &= E(1_E GY) & &= E^Q(1_E Y) \end{aligned}$$

thus $X = E^Q(Y | \mathcal{G})$.

Equation (3.20): For $G_t = E(G | \mathbb{I}_t)$

$$E(GY_u | \mathbb{I}_t) = E(E(GY_u | \mathbb{I}_u) | \mathbb{I}_t) = E(E(G | \mathbb{I}_u) Y_u | \mathbb{I}_t) = E(G_u Y_u | \mathbb{I}_t) = E_t(G_u Y_u)$$

by Equation (3.18)

$$E_t^Q(Y_u) = \frac{E(GY_u | \mathbb{I}_t)}{E(G | \mathbb{I}_t)} = \frac{E_t(G_u Y_u)}{G_t}$$

Equation (3.21): from Harrison and Pliska (1981) and Dalang et al. (1990), there is an EMM Q such that

$$\frac{P_t}{B_t} = E_t^Q\left(\frac{P_u}{B_u}\right) \quad 0 \leq t \leq u \leq T$$

We see that $P_t = B_t E_t^Q(P_u/B_u)$ works for all the primary assets, thus

$$\wp_{t,u}(X_u) = B_t E_t^Q(X_u/B_u)$$

is a pricing function.

Equation (3.22): Since $dQ = D_{0,T} B_T dO$

$$X_t = \wp_{t,u}(X_u) = B_t E_t^Q(X_u/B_u) = B_t \frac{E_t^O(D_{0,T} B_T X_u/B_u)}{E_t^O(D_{0,T} B_T)} = \frac{E_t^O(X_u B_T/B_u)}{E_t^O(B_T/B_t)}$$

In particular

$$X_t = \wp_{t,T}(X_T) = \frac{E_t^O(X_T)}{E_t^O(B_T/B_t)}$$

Let $X_T = 1$, there is $D_{t,T} = \frac{1}{E_t^O(B_T/B_t)}$ and $D_{u,T} = \frac{1}{E_u^O(B_T/B_u)}$. Thus

$$\begin{aligned} X_t &= \frac{E_t^O(X_u B_T/B_u)}{E_t^O(B_T/B_t)} = D_{t,T} E_t^O(E_u^O(X_u B_T/B_u)) \\ &= D_{t,T} E_t^O(X_u E_u^O(B_T/B_u)) = D_{t,T} E_t^O(X_u/D_{u,T}) \end{aligned}$$

Bibliography

Dalang, Robert C., Andrew Morton, and Walter Willinger, 1990. Equivalent Martingale Measures and No-arbitrage in Stochastic Securities Market Models. *Stochastics and Stochastic Reports*, 29(2): 185–201

Harrison, J. Michael and Stanley R. Pliska, 1981. Martingales and Stochastic Integrals in the Theory of Continuous Trading. *Stochastic Processes and their Applications*, 11(3):215–260