Lorenzo Naranjo October 28, 2026

Conditioning and Independence

Stochastic processes describe how uncertainty unfolds over time. At each date some information has been revealed and some has not, and the quantities we care about, such as the probability of a default or the expected value of a future payoff, must be computed given what is known at that date. *Conditioning* is the operation that updates probabilities and expectations in light of new information. *Independence* describes the opposite situation, in which learning one thing tells us nothing about another.

These notes continue the notes on probability basics and use the same setting: a discrete probability space (Ω, P) with a finite or countably infinite set of outcomes, where every subset of Ω is an event. We first define conditional probability and use it to describe independence. We then define the conditional expectation of a random variable given another, describe the information carried by a random variable, and finally define the conditional expectation given an information set. The last notion, together with its properties, is the tool we will use to define martingales.

For more complete treatments, see Blitzstein and Hwang (2019) and Grimmett and Stirzaker (2001).

Conditional Probability

For two events A and B with $\mathrm{P}(B) > 0$, the *conditional probability* of A given B is

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}.$$

Once we learn that B occurred, the outcomes outside B are no longer possible. Conditioning keeps the outcomes in B and rescales their probabilities by $1/\mathrm{P}(B)$ so that they add up to one.

Example 1. Suppose we roll a fair six-sided die and define the events

$$A = \{2, 4, 6\} \quad \text{and} \quad B = \{4, 5, 6\}.$$

Then

$$A \cap B = \{4, 6\}, \qquad \mathrm{P}(B) = \frac{3}{6}, \qquad \mathrm{P}(A \cap B) = \frac{2}{6}.$$

Therefore, the conditional probability of A given B is

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{2/6}{3/6} = \frac{2}{3}.$$

Learning that the roll is at least four raises the probability that it is even from $1/2$ to $2/3$.□

Rearranging the definition gives the *multiplication rule* $\mathrm{P}(A \cap B) = \mathrm{P}(A \mid B)\mathrm{P}(B)$. If the events $B_1, B_2, \ldots$ form a *partition* of Ω, meaning that they are pairwise disjoint and their union is Ω, then countable additivity and the multiplication rule give the *law of total probability*,

$$\mathrm{P}(A) = \sum_k \mathrm{P}(A \cap B_k) = \sum_k \mathrm{P}(A \mid B_k)\mathrm{P}(B_k),$$

where the sum runs over the events with $\mathrm{P}(B_k) > 0$. The probability of A is a weighted average of its conditional probabilities, with weights given by the probabilities of the events in the partition. We will meet this idea again when we average conditional expectations.

Independence

Independent Events

Two events A and B are **independent** if

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B).$$

When $\mathrm{P}(B) > 0$, dividing by $\mathrm{P}(B)$ shows that independence is equivalent to $\mathrm{P}(A \mid B) = \mathrm{P}(A)$: learning that B occurred does not change the probability of A. The definition in terms of the product has the advantage of being symmetric in A and B and of making sense even when one of the events has probability zero.

Example 2. Suppose that the weather tomorrow can be either sunny, fair or rainy. In addition, a certain stock tomorrow can either go up or down in price.

We can define

$$W = \{\text{sunny}, \text{fair}, \text{rainy}\}$$

and

$$S = \{\text{up}, \text{down}\}.$$

The set of outcomes can be described as all possible pairwise combinations of weather tomorrow and the stock price movement. The sample space Ω is then the Cartesian product of W and S, i.e., $\Omega = W \times S$.

We can then define the weather events

$$\begin{aligned} \text{Sunny} &= \{(\text{sunny}, \text{up}), (\text{sunny}, \text{down})\}, \\ \text{Fair} &= \{(\text{fair}, \text{up}), (\text{fair}, \text{down})\}, \\ \text{Rainy} &= \{(\text{rainy}, \text{up}), (\text{rainy}, \text{down})\}. \end{aligned}$$

The table below describes the probabilities for tomorrow's weather.

Weather	Sunny	Fair	Rainy
Probability	0.3	0.5	0.2

Similarly, the stock events can be defined as

$$\begin{aligned} \text{Up} &= \{(\text{sunny}, \text{up}), (\text{fair}, \text{up}), (\text{rainy}, \text{up})\}, \\ \text{Down} &= \{(\text{sunny}, \text{down}), (\text{fair}, \text{down}), (\text{rainy}, \text{down})\}. \end{aligned}$$

The probabilities of the stock price going up or down are described in the table below.

Stock	Up	Down
Probability	0.6	0.4

If the weather does not affect the likelihood of the stock going up or down, the conditional probability that the stock goes up should be the same whatever the weather, so that $\mathrm{P}(\text{Up} \mid \text{Sunny}) = \mathrm{P}(\text{Up}) = 0.6$, and similarly for the other combinations. By the multiplication rule, this means that

$$\mathrm{P}(\text{Stock} \cap \text{Weather}) = \mathrm{P}(\text{Stock})\mathrm{P}(\text{Weather}),$$

where Stock is either Up or Down, and Weather is either Sunny, Fair, or Rainy.

The table below describes the combined probabilities of the stock price movement and the weather tomorrow that are consistent with the independence of those events.

Stock\Weather	Sunny	Fair	Rainy
Up	0.18	0.30	0.12
Down	0.12	0.20	0.08

In each column, the stock goes up 60% of the time and down 40% of the time. The weather does not change the relative proportions of the probabilities for the stock price. □

The previous example shows how to generate independent events out of two discrete probability spaces (Ω_1, P_1) and (Ω_2, P_2). If we define $\Omega = \Omega_1 \times \Omega_2$ and let $\mathrm{P}(\omega_1, \omega_2) = \mathrm{P}_1(\omega_1)\mathrm{P}_2(\omega_2)$ for each $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$, the pair (Ω, P) is a well-defined discrete probability space. In this new probability space, the events $A = \{\omega_1\} \times \Omega_2$ and $B = \Omega_1 \times \{\omega_2\}$ are independent for any $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$.

Proof

We have that

$$\begin{aligned}\mathrm{P}(A) &= \sum_{\omega_2' \in \Omega_2} \mathrm{P}(\omega_1, \omega_2') \\ &= \sum_{\omega_2' \in \Omega_2} \mathrm{P}_1(\omega_1)\mathrm{P}_2(\omega_2') \\ &= \mathrm{P}_1(\omega_1) \sum_{\omega_2' \in \Omega_2} \mathrm{P}_2(\omega_2') \\ &= \mathrm{P}_1(\omega_1).\end{aligned}$$

Similarly, $\mathrm{P}(B) = \mathrm{P}_2(\omega_2)$. Since $A \cap B = \{(\omega_1, \omega_2)\}$, we have that $\mathrm{P}(A \cap B) = \mathrm{P}_1(\omega_1)\mathrm{P}_2(\omega_2) = \mathrm{P}(A)\mathrm{P}(B)$, proving that A and B are independent. □

This product construction is how we build models of repeated experiments, such as a sequence of independent coin flips, where the outcome of each flip is unaffected by the others.

Example 3. The sample space Ω is independent of any event $A \subset \Omega$, since $\mathrm{P}(A \cap \Omega) = \mathrm{P}(A) = \mathrm{P}(A)\mathrm{P}(\Omega)$. Learning that Ω occurred, which we knew all along, does not change the probability of A. For the same reason, the empty set is independent of every event. □

Independent Random Variables

Two random variables X and Y are **independent** if the events $\{X = x\}$ and $\{Y = y\}$ are independent for all values x and y. Thus, if X and Y are independent we have that

$$\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\mathrm{P}(Y = y) \quad \text{for all } x, y,$$

or equivalently, the joint probability mass function factors into the product of the marginals,

$$p_{X,Y}(x, y) = p_X(x)\, p_Y(y).$$

More generally, the random variables $X_1, X_2, \ldots, X_n$ are independent if their joint probability mass function factors as

$$\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathrm{P}(X_1 = x_1) \cdots \mathrm{P}(X_n = x_n) \quad \text{for all } x_1, \ldots, x_n.$$

An infinite sequence $X_1, X_2, \ldots$ is independent if every finite subcollection is independent. The coin flips that drive a random walk are the leading example.

If X and Y are independent, anything we compute from X alone is also independent of anything we compute from Y alone. This fact, and its consequence for expectations, is used constantly in the rest of the course.

Functions of Independent Random Variables

If X and Y are independent and f and g are real-valued functions, then $f(X)$ and $g(Y)$ are independent and

$$\mathrm{E}(f(X)g(Y)) = \mathrm{E}(f(X))\mathrm{E}(g(Y)), \tag{1}$$

provided the expectations exist. In particular, $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$.

Proof

For any values a and b, the event $\{f(X) = a\}$ is the disjoint union of the events $\{X = x\}$ over the values x with $f(x) = a$, and similarly for $\{g(Y) = b\}$. Therefore,

$$\begin{aligned}
\mathrm{P}(f(X) = a, g(Y) = b) &= \sum_{x:f(x)=a} \sum_{y:g(y)=b} p_X(x)\, p_Y(y) \\
&= \sum_{x:f(x)=a} p_X(x) \sum_{y:g(y)=b} p_Y(y) \\
&= \mathrm{P}(f(X) = a)\mathrm{P}(g(Y) = b),
\end{aligned}$$

so $f(X)$ and $g(Y)$ are independent. For the expectation,

$$\begin{aligned}
\mathrm{E}(f(X)g(Y)) &= \sum_{x \in R_X} \sum_{y \in R_Y} f(x)g(y)\, p_X(x)\, p_Y(y) \\
&= \sum_{x \in R_X} f(x)\, p_X(x) \sum_{y \in R_Y} g(y)\, p_Y(y) \\
&= \mathrm{E}(f(X))\mathrm{E}(g(Y)).
\end{aligned}$$

Taking $f(x) = x$ and $g(y) = y$ gives $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$. □

The notes on generating functions rely on this result with $f(x) = z^x$ and $g(y) = z^y$, which shows that the generating function of a sum of independent random variables is the product of their generating functions.

Equation (1) also implies that independent random variables are uncorrelated. Indeed,

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \mathrm{E}(X)\mathrm{E}(Y) - \mathrm{E}(X)\mathrm{E}(Y) = 0.$$

As a consequence, the variance of a sum of independent random variables is the sum of their variances, as shown for uncorrelated random variables in the notes on probability basics. The converse is not true: random variables can be uncorrelated without being independent.

Example 4. Consider a random variable X that takes the values $\{1, 0, -1\}$, each with probability $1/3$. We compute:

$$\mathrm{E}(X) = \frac{1}{3}(1) + \frac{1}{3}(0) + \frac{1}{3}(-1) = 0,$$

and

$$\mathrm{E}(X^3) = \frac{1}{3}(1^3) + \frac{1}{3}(0^3) + \frac{1}{3}((-1)^3) = 0.$$

Now, define $Y = X^2$. The covariance between X and Y is

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \mathrm{E}(X^3) - \mathrm{E}(X)\mathrm{E}(X^2) = 0.$$

However, X and Y are not independent. Since $Y = 0$ exactly when $X = 0$,

$$\mathrm{P}(X = 0, Y = 0) = \frac{1}{3} \neq \frac{1}{3} \times \frac{1}{3} = \mathrm{P}(X = 0)\mathrm{P}(Y = 0).$$

The random variables are uncorrelated, yet Y is completely determined by X. □

Covariance measures only the *linear* association between two random variables. In the example, Y depends on X through the symmetric function x^2, and positive and negative values of X cancel in the covariance.

Conditional Expectation Given a Random Variable

For x such that $p_X(x) > 0$, the conditional probability mass function of Y given $X = x$ is

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x, y)}{p_X(x)} = \mathrm{P}(Y = y \mid X = x).$$

The *conditional expectation* of Y given $X = x$ is the expectation of Y computed with this conditional pmf,

$$\mathrm{E}(Y \mid X = x) = \sum_{y \in R_Y} y \, p_{Y|X}(y \mid x).$$

This is a number that depends on x. Evaluating it at the observed value of X gives a random variable, denoted $\mathrm{E}(Y \mid X)$, which assigns to each outcome ω the value $\mathrm{E}(Y \mid X = X(\omega))$. In words, $\mathrm{E}(Y \mid X)$ is our best forecast of Y once we have observed X, and it is random because it depends on what we observe.

When X and Y are independent the conditional pmf equals the marginal pmf, $p_{Y|X}(y \mid x) = p_Y(y)$, and therefore $\mathrm{E}(Y \mid X) = \mathrm{E}(Y)$. Observing X does not change our forecast of Y.

A key result in probability theory is the *law of iterated expectations*:

$$\begin{aligned}
\mathrm{E}(\mathrm{E}(Y \mid X)) &= \sum_{x \in R_X} \mathrm{E}(Y \mid X = x) \, p_X(x) \\
&= \sum_{x \in R_X} \sum_{y \in R_Y} y \, p_{Y|X}(y \mid x) \, p_X(x) \\
&= \sum_{x \in R_X} \sum_{y \in R_Y} y \, p_{X,Y}(x, y) \\
&= \sum_{y \in R_Y} y \, p_Y(y) \\
&= \mathrm{E}(Y).
\end{aligned}$$

The expected value of the conditional expectation equals the expected value of Y itself. Averaging our forecasts over everything we might observe gives back the unconditional

forecast, in the same way that the law of total probability averages conditional probabilities.

Example 5. The notes on probability basics used the following joint pmf of a random variable X taking values in $\{-1, 0, 1\}$ and a random variable Y taking values in $\{0, 1, 2, 3\}$:

$X \setminus Y$	0	1	2	3
-1	0.12500	0.09375	0.06250	0.03125
0	0.06250	0.12500	0.12500	0.06250
1	0.03125	0.06250	0.09375	0.12500

We compute the conditional expectation of Y given $X = 1$. Adding up the row $X = 1$,

$$p_X(1) = 0.03125 + 0.06250 + 0.09375 + 0.12500 = 0.3125.$$

Dividing each entry of the row by $p_X(1)$, the conditional pmf is

$$p_{Y|X}(y \mid 1) = 0.1,\ 0.2,\ 0.3,\ 0.4 \quad \text{for } y = 0, 1, 2, 3.$$

Therefore,

$$\mathrm{E}(Y \mid X = 1) = 0(0.1) + 1(0.2) + 2(0.3) + 3(0.4) = 2.$$

The same calculation for the other rows gives $\mathrm{E}(Y \mid X = -1) = 1$ and $\mathrm{E}(Y \mid X = 0) = 1.5$, with $p_X(-1) = 0.3125$ and $p_X(0) = 0.375$. The law of iterated expectations then yields

$$\mathrm{E}(Y) = 1(0.3125) + 1.5(0.375) + 2(0.3125) = 1.5.$$

□

Information

Information Generated by a Random Variable

To describe how information arrives over time, we need a way to say what we know after observing a random variable. Consider a sample space with four possible outcomes $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ and three random variables X, Y, and Z with the following values:

Outcome	Probability	X	Y	Z
ω_1	0.10	−10	20	15
ω_2	0.30	−5	10	−10
ω_3	0.40	5	0	15
ω_4	0.20	10	0	−10

Observing X reveals which outcome occurred, since X takes a different value at each outcome. Observing Y does not: if $Y = 0$ we only know that either ω_3 or ω_4 occurred. Observing Z leaves two possible outcomes whatever its value.

Each random variable splits the sample space into the sets on which it takes each of its values. For Y, these sets are

$$\{Y = 20\} = \{\omega_1\}, \qquad \{Y = 10\} = \{\omega_2\}, \qquad \{Y = 0\} = \{\omega_3, \omega_4\},$$

and for Z,

$$\{Z = 15\} = \{\omega_1, \omega_3\}, \qquad \{Z = -10\} = \{\omega_2, \omega_4\}.$$

These sets form a partition of Ω, called the partition *generated* by the random variable. Observing the random variable tells us which cell of the partition contains the true outcome, and nothing more. We can then decide whether any event made of whole cells has occurred, but not events that split a cell. The collection of all events that are unions of cells is the **information set** generated by the random variable. We denote it by $\mathcal{F}_Y$ for Y, and similarly for other random variables:

$$\mathcal{F}_Y = \{\emptyset, \{\omega_1\}, \{\omega_2\}, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \{\omega_1, \omega_3, \omega_4\}, \{\omega_2, \omega_3, \omega_4\}, \Omega\},$$

and

$$\mathcal{F}_Z = \{\emptyset, \{\omega_1, \omega_3\}, \{\omega_2, \omega_4\}, \Omega\}.$$

Since X separates all four outcomes, $\mathcal{F}_X$ contains every subset of Ω. A finer partition corresponds to more information, and in this example $\mathcal{F}_Z \subset \mathcal{F}_X$ and $\mathcal{F}_Y \subset \mathcal{F}_X$. Neither $\mathcal{F}_Y$ nor $\mathcal{F}_Z$ contains the other: knowing Y does not tell us whether $\{\omega_1, \omega_3\}$ occurred, and

knowing Z does not tell us whether $\{\omega_1\}$ occurred. In the general theory, the information set generated by a random variable is called the σ-algebra generated by it.

Observing several random variables $X_1, \ldots, X_n$ together generates the partition whose cells are the sets $\{X_1 = x_1, \ldots, X_n = x_n\}$, and the corresponding information set is denoted by $\mathcal{F}_{X_1,\ldots,X_n}$. For simplicity, in what follows we write $\mathcal{F}$ for an information set generated by a partition of Ω into cells $C_1, C_2, \ldots$ with positive probabilities.

A random variable W is *known* given $\mathcal{F}$, or *$\mathcal{F}$-measurable*, if it is constant on each cell of the partition, so that observing the information in $\mathcal{F}$ reveals the value of W. In the example, Y^2 is known given $\mathcal{F}_Y$ but Z is not, since Z takes the values 15 and -10 on the cell $\{Y = 0\} = \{\omega_3, \omega_4\}$. A random variable is known given $\mathcal{F}_Y$ exactly when it can be written as a function of Y.

Conditional Expectation Given Information

The conditional expectation $\mathrm{E}(Y \mid X)$ averages Y over the outcomes that are consistent with the observed value of X. The same idea works for any information set. Given $\mathcal{F}$, we know which cell of the partition contains the outcome, so our best forecast of a random variable Y is its average over that cell.

Conditional Expectation Given an Information Set

Let $\mathcal{F}$ be the information set generated by a partition $C_1, C_2, \ldots$ of Ω with $\mathrm{P}(C_k) > 0$ for all k. The *conditional expectation* of Y given $\mathcal{F}$ is the random variable $\mathrm{E}(Y \mid \mathcal{F})$ defined by

$$\mathrm{E}(Y \mid \mathcal{F})(\omega) = \frac{\mathrm{E}(Y\mathbf{1}_{C_k})}{\mathrm{P}(C_k)} = \frac{1}{\mathrm{P}(C_k)} \sum_{\omega' \in C_k} Y(\omega')\mathrm{P}(\omega') \qquad \text{for } \omega \in C_k.$$

On each cell, $\mathrm{E}(Y \mid \mathcal{F})$ equals the expectation of Y computed with the probabilities conditional on that cell. By construction, $\mathrm{E}(Y \mid \mathcal{F})$ is constant on each cell, so it is known given $\mathcal{F}$. When $\mathcal{F} = \mathcal{F}_X$, the cells are the sets $\{X = x\}$ and the definition reduces to $\mathrm{E}(Y \mid X)$. When $\mathcal{F} = \{\emptyset, \Omega\}$ contains no information, the only cell is Ω and $\mathrm{E}(Y \mid \mathcal{F}) = \mathrm{E}(Y)$.

Example 6. Using the table above, we compute $\mathrm{E}(X \mid \mathcal{F}_Z)$. The partition generated by Z has two cells, $C_1 = \{\omega_1, \omega_3\}$ and $C_2 = \{\omega_2, \omega_4\}$, each with probability 0.5. On each cell we average X:

$$\mathrm{E}(X \mid \mathcal{F}_Z) = \frac{(-10)(0.10) + 5(0.40)}{0.5} = 2 \qquad \text{on } C_1,$$

$$\mathrm{E}(X \mid \mathcal{F}_Z) = \frac{(-5)(0.30) + 10(0.20)}{0.5} = 1 \qquad \text{on } C_2.$$

If we observe $Z = 15$ our forecast of X is 2, and if we observe $Z = -10$ it is 1. Averaging the forecasts over the two cells gives $2(0.5) + 1(0.5) = 1.5$, which is $\mathrm{E}(X) = -1 - 1.5 + 2 + 2 = 1.5$, as the law of iterated expectations requires. □

The following properties are all we need to work with conditional expectations in the rest of the course. They hold for general information sets, but we prove them only for information sets generated by partitions.

Properties of Conditional Expectation

Let $\mathcal{F}$ be an information set, and let Y and W be random variables with finite expectations.

i. **Linearity.** For constants a and b, $\mathrm{E}(aY + bW \mid \mathcal{F}) = a\,\mathrm{E}(Y \mid \mathcal{F}) + b\,\mathrm{E}(W \mid \mathcal{F})$.

ii. **Taking out what is known.** If W is known given $\mathcal{F}$, then $\mathrm{E}(WY \mid \mathcal{F}) = W\,\mathrm{E}(Y \mid \mathcal{F})$. In particular, $\mathrm{E}(W \mid \mathcal{F}) = W$.

iii. **Independence.** If Y is independent of every event in $\mathcal{F}$, then $\mathrm{E}(Y \mid \mathcal{F}) = \mathrm{E}(Y)$.

iv. **Tower property.** If $\mathcal{G} \subset \mathcal{F}$ is a smaller information set, then $\mathrm{E}(\mathrm{E}(Y \mid \mathcal{F}) \mid \mathcal{G}) = \mathrm{E}(Y \mid \mathcal{G})$. In particular, $\mathrm{E}(\mathrm{E}(Y \mid \mathcal{F})) = \mathrm{E}(Y)$.

Proof

Fix a cell C of the partition that generates $\mathcal{F}$. Linearity follows from the linearity of $\mathrm{E}(\cdot\,\mathbf{1}_C)$.

If W is known given $\mathcal{F}$, it takes a constant value w on C, so $WY\mathbf{1}_C = wY\mathbf{1}_C$ and

$$\frac{\mathrm{E}(WY\mathbf{1}_C)}{\mathrm{P}(C)} = w\frac{\mathrm{E}(Y\mathbf{1}_C)}{\mathrm{P}(C)},$$

which is $W\,\mathrm{E}(Y \mid \mathcal{F})$ on C. Taking $Y = 1$ gives $\mathrm{E}(W \mid \mathcal{F}) = W$.

If Y is independent of the event C, then $\mathrm{P}(Y = y, C) = \mathrm{P}(Y = y)\mathrm{P}(C)$ for every y, so that

$$\mathrm{E}(Y\mathbf{1}_C) = \sum_y y\,\mathrm{P}(Y = y, C) = \mathrm{P}(C)\sum_y y\,\mathrm{P}(Y = y) = \mathrm{P}(C)\mathrm{E}(Y),$$

and dividing by $\mathrm{P}(C)$ gives $\mathrm{E}(Y \mid \mathcal{F}) = \mathrm{E}(Y)$ on C.

For the tower property, since $\mathcal{G}$ contains less information than $\mathcal{F}$, each cell D of the partition generating $\mathcal{G}$ is a union of cells C of the partition generating $\mathcal{F}$. On each such C, $\mathrm{E}(Y \mid \mathcal{F})$ equals $\mathrm{E}(Y\mathbf{1}_C)/\mathrm{P}(C)$, so that

$$\mathrm{E}\,(\mathrm{E}\,(Y \mid \mathcal{F})\mathbf{1}_D) = \sum_{C \subset D} \frac{\mathrm{E}(Y\mathbf{1}_C)}{\mathrm{P}(C)}\,\mathrm{P}(C) = \sum_{C \subset D} \mathrm{E}(Y\mathbf{1}_C) = \mathrm{E}(Y\mathbf{1}_D).$$

Dividing by $\mathrm{P}(D)$ shows that $\mathrm{E}(\mathrm{E}(Y \mid \mathcal{F}) \mid \mathcal{G}) = \mathrm{E}(Y \mid \mathcal{G})$ on D. Taking $\mathcal{G} = \{\emptyset, \Omega\}$ gives $\mathrm{E}(\mathrm{E}(Y \mid \mathcal{F})) = \mathrm{E}(Y)$. □

Each property has a simple interpretation. Taking out what is known says that quantities we have already observed behave like constants when we forecast. The independence property says that information unrelated to Y does not change our forecast. The tower property says that if we first forecast with more information and then average that forecast using less information, we obtain the forecast based on less information: today's forecast of tomorrow's forecast is today's forecast.

Filtrations

Information accumulates over time. A sequence of information sets $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$, where $\mathcal{F}_n$ describes what is known at time n, is called a **filtration**. The inclusions express that nothing is forgotten: every event whose occurrence is known at time m is still known at any later time $n > m$.

Example 7. Let $X_1, X_2, \ldots$ be independent coin flips, each equal to $+1$ or -1 with probability $1/2$, and let $\mathcal{F}_n = \mathcal{F}_{X_1,\ldots,X_n}$ be the information generated by the first n flips, with $\mathcal{F}_0 = \{\emptyset, \Omega\}$. An infinite sequence of flips has uncountably many outcomes, but any statement about the first n flips only needs the finite sample space $\{-1, +1\}^n$. The cells

of the partition generating $\mathcal{F}_n$ are the 2^n possible sequences of the first n flips. Each flip splits every cell in two, so the partitions become finer and $\mathcal{F}_n \subset \mathcal{F}_{n+1}$.

The position of the random walk $S_n = X_1 + \cdots + X_n$ is known given $\mathcal{F}_n$, whereas the next flip X_{n+1} is independent of every event in $\mathcal{F}_n$. The properties of conditional expectation then give

$$\mathrm{E}(S_{n+1} \mid \mathcal{F}_n) = \mathrm{E}(S_n \mid \mathcal{F}_n) + \mathrm{E}(X_{n+1} \mid \mathcal{F}_n) = S_n + \mathrm{E}(X_{n+1}) = S_n.$$

Given everything observed up to time n, the best forecast of the position at time $n + 1$ is the current position. □

A process with this property is called a *martingale*, and it is the subject of the notes on discrete-time martingales. The calculation in the example uses exactly three of the properties above: linearity, taking out what is known, and independence. The tower property enters when we forecast more than one period ahead.

Practice Problems

Problem 1. Let A and B be events with positive probabilities.

a. Use the definition of conditional probability and the law of total probability to derive *Bayes' rule*,

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(A \mid B)\mathrm{P}(B)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A \mid B)\mathrm{P}(B)}{\mathrm{P}(A \mid B)\mathrm{P}(B) + \mathrm{P}(A \mid B^C)\mathrm{P}(B^C)}.$$

b. A credit model flags 90% of the firms that will default within a year and 10% of the firms that will not. Two percent of firms default. Given that a firm is flagged, what is the probability that it defaults? Given that it is not flagged?

c. Explain why the first answer is so much smaller than 90%.

Problem 2. Over each of three periods a stock moves up with probability $p = 0.6$ and down with probability $q = 0.4$, independently of the other periods. Let U be the number of up moves.

a. Describe the sample space and the probability of each outcome. How many outcomes have exactly k up moves?

b. Find the probability mass function of U.
c. Write $U = I_1 + I_2 + I_3$, where $I_j = 1$ if the stock moves up in period j and $I_j = 0$ otherwise, and use it to compute $\mathrm{E}(U)$ and $\mathrm{V}(U)$.

Problem 3. Independence is easy to misjudge. The following three facts help.

a. Show that two disjoint events with positive probabilities are never independent.
b. Show that if A and B are independent, then so are A and B^C.
c. Flip two fair coins, and let A be the event that the first coin lands heads, B the event that the second lands heads, and C the event that both coins land the same way. Show that each pair of these events is independent, but that

$$\mathrm{P}(A \cap B \cap C) \neq \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C).$$

What does knowing A and B tell you about C?

Problem 4 (★). A fund's gross return in each of two years is 1.2 or 0.9 with probability $1/2$ each, independently across years. Let R_1 and R_2 be the two gross returns.

a. Compute $\mathrm{E}(R_1)$ and $\mathrm{E}(R_1 R_2)$.
b. Compute $\mathrm{V}(R_1 R_2)$ using $\mathrm{E}\left((R_1 R_2)^2\right) = \mathrm{E}(R_1^2)\mathrm{E}(R_2^2)$. Why does this equation hold?
c. Find the probability mass function of $R_1 R_2$. What is the most likely value, and how does it compare with $\mathrm{E}(R_1 R_2)$?

Problem 5 (★). A stock trades at \$100 and over each of two periods moves up by 20% with probability $p = 0.6$ or down by 10% with probability 0.4, independently across periods. Let $\mathcal{F}_1$ be the information generated by the first move, and let $C = \max(S_2 - 100, 0)$ be the payoff of a call at time 2.

a. List the outcomes, their probabilities, and the values of S_1, S_2, and C.
b. Compute $\mathrm{E}(C \mid \mathcal{F}_1)$ on each cell of the partition that generates $\mathcal{F}_1$.
c. Compute $\mathrm{E}(C)$ by averaging $\mathrm{E}(C \mid \mathcal{F}_1)$, and check the answer against the direct calculation from the pmf of C.

Problem 6 (★). Say that Y is *mean independent* of X if $\mathrm{E}(Y \mid X) = \mathrm{E}(Y)$.

a. Explain why independence implies mean independence.
b. Use the properties of conditional expectation to show that mean independence implies $\mathrm{Cov}(X, Y) = 0$.
c. Let X take the values $-1, 0, 1$ with probability $1/3$ each, and $Y = X^2$. Show that X and Y are uncorrelated but Y is not mean independent of X.
d. Let X be 0 or 1 with probability $1/2$ each. When $X = 0$, let $Y = 0$, and when $X = 1$, let Y be $+1$ or -1 with probability $1/2$ each. Show that Y is mean independent of X but not independent of it. Interpret X as a calm or volatile market and Y as a return.

Problem 7 (★★). Let $X_1, X_2, \ldots$ be independent, each equal to $+1$ with probability p and -1 with probability $q = 1 - p$. Let $S_n = X_1 + \cdots + X_n$, with $S_0 = 0$, and let $\mathcal{F}_n$ be the information generated by the first n steps.

a. Show that $\mathrm{E}(S_{n+1} \mid \mathcal{F}_n) = S_n + (p - q)$, and use the tower property to show that $\mathrm{E}(S_{n+k} \mid \mathcal{F}_n) = S_n + k(p - q)$.
b. Show that $\mathrm{E}(S_{n+1}^2 \mid \mathcal{F}_n) = S_n^2 + 2(p - q)S_n + 1$. Conclude that when $p = 1/2$,

$$\mathrm{E}(S_{n+1}^2 - (n+1) \mid \mathcal{F}_n) = S_n^2 - n.$$

c. For general p, let $M_n = S_n - n(p - q)$ and $\sigma^2 = \mathrm{V}(X_1) = 4pq$. Show that

$$\mathrm{E}(M_{n+1}^2 - (n+1)\sigma^2 \mid \mathcal{F}_n) = M_n^2 - n\sigma^2.$$

Problem 8 (★★). Let X and Y be random variables, and write $m(X) = \mathrm{E}(Y \mid X)$.

a. Show that for any function g,

$$\mathrm{E}\left((Y - g(X))^2\right) = \mathrm{E}\left((Y - m(X))^2\right) + \mathrm{E}\left((m(X) - g(X))^2\right),$$

so that $m(X)$ is the best forecast of Y based on X in the mean-square sense. Hint: compute the expectation of the cross term by first conditioning on X.
b. Show that

$$\mathrm{V}(Y) = \mathrm{E}(\mathrm{V}(Y \mid X)) + \mathrm{V}(\mathrm{E}(Y \mid X)), \qquad \text{where } \mathrm{V}(Y \mid X) = \mathrm{E}\left((Y - m(X))^2 \mid X\right).$$

c. Verify the variance decomposition for the joint pmf below.

$X \setminus Y$	0	1	2
0	0.20	0.15	0.05
1	0.10	0.20	0.30

References

Blitzstein, Joseph K., and Jessica Hwang. 2019. *Introduction to Probability*. 2nd ed. Chapman, Hall/CRC.

Grimmett, Geoffrey, and David Stirzaker. 2001. *Probability and Random Processes*. 3rd ed. Oxford University Press.