# 2

# Information and Conditioning

## 2.1 Information and $\sigma$-algebras

The no-arbitrage theory of derivative security pricing is based on contingency plans. In order to price a derivative security, we determine the initial wealth we would need to set up a hedge of a short position in the derivative security. The hedge must specify what position we will take in the underlying security at each future time contingent on how the uncertainty between the present time and that future time is resolved. In order to make these contingency plans, we need a way to mathematically model the information on which our future decisions can be based. In the binomial model, that information was knowledge of the coin tosses between the initial time and the future time. For the continuous-time model, we need to develop somewhat more sophisticated machinery to capture this concept of information.

We imagine as always that some random experiment is performed, and the outcome is a particular $\omega$ in the set of all possible outcomes $\Omega$. We might then be given some information, not enough to know the precise value of $\omega$, but enough to narrow down the possibilities. For example, the true $\omega$ might be the result of three coin tosses, and we are told only the first one. Or perhaps we are told the stock price at time two, without being told any of the coin tosses. In such a situation, although we do not know the true $\omega$ precisely, we can make a list of sets that are sure to contain it and other sets that are sure not to contain it. These are the sets that are *resolved by the information.*.

Indeed, suppose $\Omega$ is the set of eight possible outcomes of three coin tosses. If we are told the outcome of the first coin toss only, the sets

$$A_H = \{HHH, HHT, HTH, HTT\}, \quad A_T = \{THH, THT, TTH, TTT\} \tag{2.1.1}$$

are resolved. For each of these sets, once we are told the first coin toss, we know if the true $\omega$ is a member. The empty set $\emptyset$ and the whole space $\Omega$ are always resolved, even without any information; the true $\omega$ does not belong to $\emptyset$ and does belong to $\Omega$. The four sets that are resolved by the first coin toss

form the $\sigma$-algebra

$$\mathcal{F}_1 = \big\{\emptyset, \Omega, A_H, A_T\big\}.$$

We shall think of this $\sigma$-algebra as containing the information learned by observing the first coin toss. More precisely, if instead of being told the first coin toss, we are told, for each set in $\mathcal{F}_1$, whether or not the true $\omega$ belongs to the set, we know the outcome of the first coin toss and nothing more.

If we are told the first two coin tosses, we obtain a finer resolution. In particular, the four sets

$$\begin{array}{ll} A_{HH} = \{HHH, HHT\} & A_{HT} = \{HTH, HTT\} \\ A_{TH} = \{THH, THT\} & A_{TT} = \{TTH, TTT\} \end{array} \tag{2.1.2}$$

are resolved. Of course, the sets in $\mathcal{F}_1$ are still resolved. Whenever a set is resolved, so is its complement, which means that $A_{HH}^c$, $A_{HT}^c$, $A_{TH}^c$ and $A_{TT}^c$ are resolved. Whenever two sets are resolved, so is their union, which means that $A_{HH} \cup A_{TH}$, $A_{HH} \cup A_{TT}$, $A_{HT} \cup A_{TH}$ and $A_{HT} \cup A_{TT}$ are resolved. We have already noted that the other two pairwise unions, $A_H = A_{HH} \cup A_{HT}$ and $A_T = A_{TH} \cup A_{TT}$ are resolved. The triple unions are also resolved, and these are the complements already mentioned, e.g.,

$$A_{HH} \cup A_{HT} \cup A_{TH} = A_{TT}^c.$$

In all, we have 16 resolved sets, which together form a $\sigma$-algebra we call $\mathcal{F}_2$, i.e.,

$$\mathcal{F}_2 = \left\{ \begin{array}{l} \emptyset, \Omega, A_H, A_T, A_{HH}, A_{HT}, A_{TH}, A_{TT}, A_{HH}^c, A_{HT}^c, A_{TH}^c, A_{TT}^c, \\ A_{HH} \cup A_{TH},\ A_{HH} \cup A_{TT},\ A_{HT} \cup A_{TH},\ A_{HT} \cup A_{TT} \end{array} \right\}. \tag{2.1.3}$$

We shall think of this $\sigma$-algebra as containing the information learned by observing the first two coin tosses.

If we are told all three coin tosses, we know the true $\omega$ and every subset of $\Omega$ is resolved. There are 256 subsets of $\Omega$, and taken all together, they constitute the $\sigma$-algebra $\mathcal{F}_3$:

$$\mathcal{F}_3 = \text{ The set of all subsets of } \Omega.$$

If we are told nothing about the coin tosses, the only resolved sets are $\emptyset$ and $\Omega$. We form the so-called *trivial $\sigma$-field* $\mathcal{F}_0$ with these two sets:

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

We have then four $\sigma$-algebras, $\mathcal{F}_0$, $\mathcal{F}_1$, $\mathcal{F}_2$ and $\mathcal{F}_3$, indexed by time. As time moves forward, we obtain finer resolution. In other words, if $n < m$, then $\mathcal{F}_m$ contains every set in $\mathcal{F}_n$ and even more. This means that $\mathcal{F}_m$ contains more information than $\mathcal{F}_n$. The collection of $\sigma$-algebras $\mathcal{F}_0$, $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$ is an example of a *filtration*. We give the continuous-time formulation of this situation in the following definition.

**Definition 2.1.1.** *Let $\Omega$ be a nonempty set. Let $T$ be a fixed positive number, and assume that for each $t \in [0,T]$ there is a $\sigma$-algebra $\mathcal{F}(t)$. Assume further that if $s \le t$, then every set in $\mathcal{F}(s)$ is also in $\mathcal{F}(t)$. Then we call the collection of $\sigma$-algebras $\mathcal{F}(t)$, $0 \le t \le T$, a* filtration.

A filtration tells us the information we will have at future times. More precisely, we know that when we get to time $t$, we will know for each set in $\mathcal{F}(t)$ whether the true $\omega$ lies in that set.

*Example 2.1.2.* Suppose our sample space is $\Omega = C_0[0,T]$, the set of continuous functions defined on $[0,T]$ taking the value zero at time zero. Suppose one of these functions $\overline{\omega}$ is chosen at random and we get to observe it up to time $t$, where $0 \le t \le T$. That is to say, we know the value of $\overline{\omega}(s)$ for $0 \le s \le t$, but we do not know the value of $\overline{\omega}(s)$ for $t < s \le T$. Certain subsets of $\Omega$ are resolved. For example, the set $\{\omega \in \Omega; \max_{0\le s\le t} \omega(s) \le 1\}$ is resolved. We would put this in the $\sigma$-algebra $\mathcal{F}(t)$. Other subsets of $\Omega$ are not resolved by time $t$. For example, if $t < T$, the set $\{\omega \in \Omega; \omega(T) > 0\}$ is not resolved by time $t$. Indeed, the sets that are resolved by time $t$ are just those sets that can be described in terms of the path of $\omega$ up to time $t$.[1] Every reasonable[2] subset of $\Omega = C_0[0,T]$ is resolved by time $T$. By contrast, at time zero we see only the value of $\overline{\omega}(0)$, which is equal to zero by the definition of $\Omega$. We learn nothing about the outcome of the random experiment of choosing $\overline{\omega}$ by observing this. The only sets resolved at time zero are $\emptyset$ and $\Omega$, and consequently $\mathcal{F}(0) = \{\emptyset, \Omega\}$. □

Example 2.1.2 provides the simplest setting in which we may construct a Brownian motion. It remains only to assign probability to the sets in $\mathcal{F} = \mathcal{F}(T)$, and then the paths $\omega \in C_0[0,T]$ will be the paths of the Brownian motion.

The discussion preceding Definition 2.1.1 suggests that the $\sigma$-algebras in a filtration can be built by taking unions and complements of certain fundamental sets, in the way $\mathcal{F}_2$ was constructed from the four sets $A_{HH}$, $A_{HT}$, $A_{TH}$ and $A_{TT}$. If this were the case, it would be enough to work with these so-called *atoms* (indivisible sets in the $\sigma$-algebra) and not consider all the other sets. In uncountable sample spaces, however, there are sets that cannot be constructed as countable unions of atoms (and uncountable unions are forbidden because we cannot add up probabilities of such unions). For example, let us fix $t \in (0,T)$ in Example 2.1.2. Now choose a continuous function $f(u)$, defined only for $0 \le u \le t$ and satisfying $f(0) = 0$. The set of continuous functions $\omega \in C_0[0,T]$ that agree with $f$ on $[0,t]$ and that are free to take any values on $(t,T]$ form an atom in $\mathcal{F}_t$. In symbols, this atom is

[1] For technical reasons, we would not include in $\mathcal{F}(t)$ sets such as $\{\omega \in \Omega; \max_{0\le s\le t} \omega(s) \in B\}$ if $B$ is a subset of $\mathbb{R}$ that is not Borel measurable. This technical issue can safely be ignored.

[2] Once again, there are pathological sets such as $\{\omega \in \Omega; \omega(T) \in B\}$, where $B$ is a subset of $\mathbb{R}$ that is not Borel measurable. These are not included in $\mathcal{F}(T)$, but that shall not concern us.

$$\{\omega \in C_0[0,T]; \omega(u) = f(u) \text{ for all } u \in [0,t]\}.$$

Each time we choose a new function $f(u)$, defined for $0 \leq u \leq t$, we get a new atom. However, there is no way to obtain the important set $\{\omega \in \Omega; \omega(t) > 0\}$ by taking countable unions of these atoms. Moreover, it is usually the case that the atoms have zero probability. Consequently, in what follows we work with all the sets of $\mathcal{F}(t)$, especially those with positive probability, not with just the atoms.

Besides observing the evolution of an economy over time, which is the idea behind Example 2.1.2, there is a second way we might acquire information about the value of $\omega$. Let $X$ be a random variable. We assume throughout that there is a "formula" for $X$ and we know this formula even before the random experiment is performed. Because we already know this formula, we are waiting only to learn the value of $\omega$ to plug into the formula so we can evaluate $X(\omega)$. But suppose, rather than being told the value of $\omega$, we are told only the value of $X(\omega)$. This resolves certain sets. For example, if we know the value of $X(\omega)$, then we know if $\omega$ is in the set $\{X \leq 1\}$ (yes, if $X(\omega) \leq 1$ and no if $X(\omega) > 1\}$. Indeed, every set of the form $\{X \in B\}$ where $B$ is a subset of $\mathbb{R}$, is resolved. Again, for technical reasons, we restrict attention to subsets $B$ that are Borel measurable.

**Definition 2.1.3.** *Let $X$ be a random variable, defined on a nonempty sample space $\Omega$. The* $\sigma$-algebra generated by $X$, *denoted $\sigma(X)$, is the collection of all subsets of $\Omega$ of the form $\{X \in B\}$,*[3] *where $B$ ranges over the Borel subsets of $\mathbb{R}$.*

*Example 2.1.4.* We return to the three-period model of Example 1.2.1 of Chapter 1. In that model, $\Omega$ is the set of eight possible outcomes of three coin tosses, and

$$\begin{aligned} &S_2(HHH) = S_2(HHT) = 16, \\ &S_2(HTH) = S_2(HTT) = S_2(THH) = S_2(THT) = 4, \\ &S_2(TTH) = S_2(TTT) = 1. \end{aligned}$$

In Figure 1.2.2 of Chapter 1, we wrote $S_2$ as a function of the first two coin tosses alone, but now we include the irrelevant third toss in the argument to get the full picture. If we take $B$ to be the set containing the single number 16, then $\{S_2 \in B\} = \{HHH, HHT\} = A_{HH}$, where we are using the notation of (2.1.2). It follows that $A_{HH}$ belongs to the $\sigma$-algebra $\sigma(S_2)$. Similarly, we can take $B$ to contain the single number 4 and conclude that $A_{HT} \cup A_{TH}$ belongs to $\sigma(S_2)$, and we can take $B$ to contain the single number 1 to see that $A_{TT}$ belongs to $\sigma(S_2)$. Taking $B = \emptyset$, we obtain $\emptyset$. Taking $B = \mathbb{R}$, we obtain $\Omega$. Taking $B = [4, 16]$, we obtain the set $A_{HH} \cup A_{HT} \cup A_{TH}$. In short, as $B$ ranges over the Borel subsets of $\mathbb{R}$, we will obtain the list of sets

[3] We recall that $\{X \in B\}$ is shorthand notation for the subset $\{\omega \in \Omega; X(\omega) \in B\}$ of $\Omega$.

$$\emptyset, \Omega, A_{HH}, A_{HT} \cup A_{TH}, A_{TT}$$

and all unions and complements of these. This is the $\sigma$-algebra $\sigma(S_2)$.

Every set in $\sigma(S_2)$ is in the $\sigma$-algebra $\mathcal{F}_2$ of (2.1.3), the information contained in the first two coin tosses. On the other hand, $A_{HT}$ and $A_{TH}$ appear separately in $\mathcal{F}_2$ and only their union appears in $\sigma(S_2)$. This is because seeing the first two coin tosses allows us to distinguish an initial head followed by a tail from an initial tail followed by a head, but knowing only the value of $S_2$ does not permit this. There is enough information in $\mathcal{F}_2$ to determine the value of $S_2$ and even more. We say that $S_2$ is $\mathcal{F}_2$-*measurable.* □

**Definition 2.1.5.** *Let $X$ be a random variable, defined on a nonempty sample space $\Omega$. Let $\mathcal{G}$ be a $\sigma$-algebra of subsets of $\Omega$. If every set in $\sigma(X)$ is also in $\mathcal{G}$, we say that $X$ is* $\mathcal{G}$-measurable.

A random variable $X$ is $\mathcal{G}$-measurable if and only if the information in $\mathcal{G}$ is sufficient to determine the value of $X$. If $X$ is $\mathcal{G}$-measurable, than $f(X)$ is also $\mathcal{G}$-measurable for any Borel-measurable function $f$; if the information in $\mathcal{G}$ is sufficient to determine the value of $X$, it will also determine the value of $f(X)$. If $X$ and $Y$ are $\mathcal{G}$-measurable, then $f(X, Y)$ is $\mathcal{G}$-measurable for any Borel-measurable function $f(x, y)$ of two variables. In particular, $X + Y$ and $XY$ are $\mathcal{G}$-measurable.

A portfolio position $\Delta(t)$ taken at time $t$ must be $\mathcal{F}(t)$-measurable (i.e., must depend only on information available to the investor at time $t$). We revisit a concept first encountered in Definition 2.4.1 of Chapter 2 of Volume I.

**Definition 2.1.6.** *Let $\Omega$ be a nonempty sample space, equipped with a filtration $\mathcal{F}(t)$, $0 \leq t \leq T$. Let $X(t)$ be a collection of random variables, indexed by $t \in [0, T]$. We say this collection of random variables is an* adapted stochastic process *if, for each $t$, the random variable $X(t)$ is $\mathcal{F}(t)$-measurable.*

In the continuous-time models of this text, asset prices, portfolio processes (i.e., positions) and wealth processes (i.e., values of portfolio processes) will all be adapted to a filtration that we regard as a model of the flow of public information.

## 2.2 Independence

When a random variable is measurable with respect to a $\sigma$-algebra $\mathcal{G}$, the information contained in $\mathcal{G}$ is sufficient to determine the value of the random variable. The other extreme is when a random variable is independent of a $\sigma$-algebra. In this case, the information contained in the $\sigma$-algebra gives no clue about the value of the random variable. Independence is the subject of the present section. In the more common case, when we have a $\sigma$-algebra $\mathcal{G}$ and a random variable $X$ that is neither measurable with respect to $\mathcal{G}$ nor

independent of $\mathcal{G}$, the information in $\mathcal{G}$ is not sufficient to evaluate $X$, but we can estimate $X$ based on the information in $\mathcal{G}$. We take up this case in the next section.

In contrast to the concept of measurability, we need a probability measure in order to talk about independence. Consequently, independence can be affected by changes of probability measure; measurability is not.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We say that two sets $A$ and $B$ in $\mathcal{F}$ are independent if

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

For example, in $\Omega = \{HH, HT, TH, TT\}$ with $0 \leq p \leq 1$, $q = 1 - p$ and

$$\mathbb{P}(HH) = p^2,\ \mathbb{P}(HT) = pq,\ \mathbb{P}(TH) = pq,\ \mathbb{P}(TT) = q^2,$$

the sets

$$A = \{\text{Head on first toss}\} = \{HH, HT\}$$

and

$$B = \{\text{Head on the second toss}\} = \{HH, TH\}$$

are independent. Indeed,

$$\mathbb{P}(A \cap B) = \mathbb{P}(HH) = p^2 \text{ and } \mathbb{P}(A)\mathbb{P}(B) = (p^2 + pq)(p^2 + pq) = p^2.$$

Independence of sets $A$ and $B$ means that knowing that the outcome $\omega$ of a random experiment is in $A$ does not change our estimation of the probability that it is in $B$. If we know the first toss results in head, we still have probability $p$ for a head on the second toss.

In a similar way, we want to define independence of two random variables $X$ and $Y$ to mean that if $\omega$ occurs and we know the value of $X(\omega)$ (without actually knowing $\omega$), then our estimation of the distribution of $Y$ is the same as when we did not know the value of $X(\omega)$. The formal definitions are the following.

**Definition 2.2.1.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}$ and $\mathcal{H}$ be sub-$\sigma$-algebras of $\mathcal{F}$ (i.e., the sets in $\mathcal{G}$ and the sets in $\mathcal{H}$ are also in $\mathcal{F}$). We say these two* $\sigma$-algebras are independent *if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B) \text{ for all } A \in \mathcal{G},\ B \in \mathcal{H}.$$

*Let $X$ and $Y$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We say these two* random variables are independent *if the $\sigma$-algebras they generate $\sigma(X)$ and $\sigma(Y)$ are independent. We say that the* random variable $X$ is independent of the $\sigma$-algebra $\mathcal{G}$ *if $\sigma(X)$ and $\mathcal{G}$ are independent.*

Recall that $\sigma(X)$ is the collection of all sets of the form $\{X \in C\}$, where $C$ ranges over the Borel subsets of $\mathbb{R}$. Similarly, every set in $\sigma(Y)$ is of the form $\{Y \in D\}$. Definition 2.2.1 says that $X$ and $Y$ are independent if and only if

$$\mathbb{P}\{X \in C \text{ and } Y \in D\} = \mathbb{P}\{X \in C\} \cdot \mathbb{P}\{Y \in D\}$$

for all Borel subsets $C$ and $D$ of $\mathbb{R}$.

*Example 2.2.2.* Recall the space $\Omega$ of three independent coin tosses, on which the stock price random variables of Figure 1.2.2 of Chapter 1 are constructed. Let the probability measure $\mathbb{P}$ be given by

$$\begin{array}{c}\mathbb{P}(HHH) = p^3, \;\; \mathbb{P}(HHT) = p^2q, \, \mathbb{P}(HTH) = p^2q, \, \mathbb{P}(HTT) = pq^2, \\ \mathbb{P}(THH) = p^2q, \; \mathbb{P}(THT) = pq^2, \; \mathbb{P}(TTH) = pq^2, \; \mathbb{P}(TTT) = q^3.\end{array}$$

Intuitively, the random variables $S_2$ and $S_3$ are not independent because if we know that $S_2$ takes the value 16, then we know that $S_3$ is either 8 or 32, and is not 2 or .50. To formalize this, we consider the sets $\{S_3 = 32\} = \{HHH\}$ and $\{S_2 = 16\} = \{HHH, HHT\}$ whose probabilities are $\mathbb{P}\{S_3 = 32\} = p^3$ and $\mathbb{P}\{S_2 = 16\} = p^2$. In order to have independence, we must have

$$\mathbb{P}\{S_2 = 16 \text{ and } S_3 = 32\} = \mathbb{P}\{S_2 = 16\} \cdot \mathbb{P}\{S_3 = 32\} = p^5.$$

But $\mathbb{P}\{S_2 = 16 \text{ and } S_3 = 32\} = \mathbb{P}\{HHH\} = p^3$, so independence requires $p = 1$ or $p = 0$. Indeed, if $p = 1$, then after learning that $S_2 = 16$, we do not revise our estimate of the distribution of $S_3$; we already knew it would be 32. If $p = 0$, then $S_2$ cannot be 16, and we do not have to worry about revising our estimate of the distribution of $S_3$ if this occurs because it will not occur.

As the previous discussion shows, in the interesting cases of $0 < p < 1$, the random variables $S_2$ and $S_3$ are not independent. However, the random variables $S_2$ and $\frac{S_3}{S_2}$ are independent. Intuitively, this is because $S_2$ depends on the first two tosses, and $\frac{S_3}{S_2}$ depends on the third toss only. The $\sigma$-algebra generated by $S_2$ comprises $\emptyset$, $\Omega_3$, the atoms (fundamental sets)

$$\begin{array}{l}\{S_2 = 16\} = \{HHH, HHT\}, \\ \{S_2 = 4\} = \{HTH, HTT, THH, THT\}, \\ \{S_2 = 1\} = \{TTH, TTT\},\end{array}$$

and their unions. The $\sigma$-algebra generated by $\frac{S_3}{S_2}$ comprises $\emptyset$, $\Omega_3$ and the atoms

$$\begin{array}{l}\left\{\dfrac{S_3}{S_2} = 2\right\} = \{HHH, HTH, THH, TTH\}, \\ \left\{\dfrac{S_3}{S_2} = \dfrac{1}{2}\right\} = \{HHT, HTT, THT, TTT\}.\end{array}$$

To verify independence, we can conduct a series of checks of the form

$$\mathbb{P}\left\{S_2 = 16 \text{ and } \frac{S_3}{S_2} = 2\right\} = \mathbb{P}\{S_2 = 16\} \cdot \mathbb{P}\left\{\frac{S_3}{S_2} = 2\right\}.$$

The left-hand side of this equality is

$$\mathbb{P}\left\{S_2 = 16 \text{ and } \frac{S_3}{S_2} = 2\right\} = \mathbb{P}\{HHH\} = p^3,$$

and the right-hand side is

$$\begin{aligned}\mathbb{P}\{S_2 = 16\} \cdot \mathbb{P}\left\{\frac{S_3}{S_2} = 2\right\} \\ &= \mathbb{P}\{HHH, HHT\} \cdot \mathbb{P}\{HHH, HTH, THH, TTH\} \\ &= p^2 \cdot p.\end{aligned}$$

Indeed, for every $A \in \sigma(S_2)$ and every $B \in \sigma\left(\frac{S_3}{S_2}\right)$, we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B). \qquad \square$$

We shall often need independence of more than two random variables. We make the following definition.

**Definition 2.2.3.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \ldots$ be a sequence of sub-$\sigma$-algebras of $\mathcal{F}$. For a fixed positive integer $n$, we say that the $n$* $\sigma$-algebras $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n$ are independent *if*

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) \cdot \cdots \cdot \mathbb{P}(A_n) \quad \text{for all } A_1 \in \mathcal{G}_1, A_2 \in \mathcal{G}_2, \ldots, A_n \in \mathcal{G}_n.$$

*Let $X_1, X_2, X_3, \ldots$ be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We say the* $n$ random variables $X_1, X_2, \ldots, X_n$ are independent *if the $\sigma$-algebras $\sigma(X_1)$, $\sigma(X_2), \ldots, \sigma(X_n)$ are independent. We say* the full sequence of $\sigma$-algebras $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \ldots$ is independent *if for every positive integer $n$, the $n$ $\sigma$-algebras $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n$ are independent. We say the* full sequence of random variables $X_1, X_2, X_3, \ldots$ is independent *if for every positive integer $n$, the $n$ random variables $X_1, X_2, \ldots, X_n$ are independent.*

*Example 2.2.4.* The infinite independent coin-toss space $(\Omega_\infty, \mathcal{F}, \mathbb{P})$ of Example 1.1.4 of Chapter 1 exhibits the kind of independence described in Definition 2.2.3. Let $\mathcal{G}_k$ be the $\sigma$-algebra of information associated with the $k$-th toss. In other words, $\mathcal{G}_k$ comprises the sets $\emptyset$, $\Omega_\infty$ and the atoms

$$\{\omega \in \Omega_\infty; \omega_k = H\} \text{ and } \{\omega \cdots \in \Omega_\infty; \omega_k = T\}.$$

Note that $\mathcal{G}_k$ is different from $\mathcal{F}_k$ in Example 1.1.4 of Chapter 1, the $\sigma$-algebra associated with the first $k$ tosses. Under the probability measure constructed in Example 1.1.4 of Chapter 1, the full sequence of $\sigma$-algebras $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \ldots$ is independent. Now recall the sequence of the random variables of (1.2.8) of Chapter 1.:

$$Y_k(\omega) = \begin{cases} 1, \text{ if } \omega_k = H, \\ 0, \text{ if } \omega_k = T. \end{cases}$$

The full sequence of random variables $Y_1, Y_2, Y_3, \ldots$ is likewise independent. $\square$

The definition of independence of random variables, which was given in terms of independence of $\sigma$-algebras that they generate, is a strong condition that is conceptually useful but difficult to check in practice. We illustrate the first point with the following theorem, and thereafter give a second theorem that simplifies the verification that two random variables are independent. Although this and the next section treat only the case of a pair of random variables, there are analogues of these results for $n$ random variables.

**Theorem 2.2.5.** *Let $X$ and $Y$ be independent random variables and let $f$ and $g$ be Borel-measurable functions on $\mathbb{R}$. Then $f(X)$ and $g(Y)$ are independent random variables.*

PROOF: Let $A$ be in the $\sigma$-algebra generated by $f(X)$. This $\sigma$-algebra is a sub-$\sigma$-algebra of $\sigma(X)$. To see this, recall that, by definition, every set $A$ in $\sigma(f(X))$ is of the form $\{\omega \in \Omega; f(X(\omega)) \in C\}$, where $C$ is a Borel subset of $\mathbb{R}$. We define $D = \{x \in \mathbb{R}; f(x) \in C\}$, and then have

$$A = \{\omega \in \Omega; f(X(\omega)) \in C\} = \{\omega \in \Omega, X(\omega) \in D\}. \tag{2.2.1}$$

The set on the right-hand side of (2.2.1) is in $\sigma(X)$, so $A \in \sigma(X)$.

Let $B$ be in the $\sigma$-algebra generated by $g(Y)$. This $\sigma$-algebra is a sub-$\sigma$-algebra of $\sigma(Y)$, so $B \in \sigma(Y)$. Since $X$ and $Y$ are independent, we have $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$. □

**Definition 2.2.6.** *Let $X$ and $Y$ be random variables. The pair of random variables $(X, Y)$ takes values in the plane $\mathbb{R}^2$, and the* joint distribution measure of $(X, Y)$ *is given by*[4]

$$\mu_{X,Y}(C) = \mathbb{P}\big\{(X, Y) \in C\big\} \text{ for all Borel sets } C \subset \mathbb{R}^2. \tag{2.2.2}$$

*This is a probability measure (i.e., a way of assigning measure between 0 and 1 to subsets of $\mathbb{R}^2$ so that $\mu_{X,Y}(\mathbb{R}^2) = 1$ and the countable additivity property is satisfied). The* joint cumulative distribution function of $(X, Y)$ *is*

$$F_{X,Y}(a, b) = \mu_{X,Y}\big((-\infty, a] \times (-\infty, b]\big) = \mathbb{P}\{X \le a, Y \le b\},\ a \in \mathbb{R}, b \in \mathbb{R}. \tag{2.2.3}$$

*We say that a nonnegative, Borel-measurable function $f_{X,Y}(x, y)$ is a* joint density *for the pair of random variables $(X, Y)$ if*

$$\mu_{X,Y}(C) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{I}_C(x, y) f_{X,Y}(x, y)\, dy\, dx \text{ for all Borel sets } C \subset \mathbb{R}^2. \tag{2.2.4}$$

[4] One way to generate the $\sigma$-algebra of Borel subsets of $\mathbb{R}^2$ is to start with the collection of closed rectangles $[a_1, b_1] \times [a_2, b_2]$ and then add all other sets necessary in order to have a $\sigma$-algebra. Any set in this resulting $\sigma$-algebra is called a *Borel subset of* $\mathbb{R}^2$. All subsets of $\mathbb{R}^2$ normally encountered belong to this $\sigma$-algebra.

Condition (2.2.4) holds if and only if

$$F_{X,Y}(a,b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f_{X,Y}(x,y)\,dy\,dx \text{ for all } a \in \mathbb{R}, b \in \mathbb{R}. \tag{2.2.5}$$

The *distribution measures* (generally called the *marginal distribution measures* in this context) of $X$ and $Y$ are

$$\begin{aligned}\mu_X(A) = \mathbb{P}\{X \in A\} &= \mu_{X,Y}(A \times \mathbb{R}) \text{ for all Borel subsets } A \subset \mathbb{R},\\ \mu_Y(B) = \mathbb{P}\{Y \in B\} &= \mu_{X,Y}(\mathbb{R} \times B) \text{ for all Borel subsets } B \subset \mathbb{R}.\end{aligned}$$

The *(marginal) cumulative distribution functions* are

$$\begin{aligned}F_X(a) = \mu_X(-\infty, a] &= \mathbb{P}\{X \le a\} \text{ for all } a \in \mathbb{R},\\ F_Y(b) = \mu_Y(-\infty, b] &= \mathbb{P}\{Y \le b\} \text{ for all } b \in \mathbb{R}.\end{aligned}$$

If the joint density $f_{X,Y}$ exists, then the marginal densities exist and are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \text{ and } f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx.$$

The *marginal densities*, if they exist, are nonnegative, Borel-measurable functions that satisfy

$$\begin{aligned}\mu_X(A) &= \int_A f_X(x)\,dx \text{ for all Borel subsets } A \subset \mathbb{R},\\ \mu_Y(B) &= \int_B f_Y(y)\,dy \text{ for all Borel subsets } B \subset \mathbb{R}.\end{aligned}$$

These last conditions hold if and only if

$$F_X(a) = \int_{-\infty}^{a} f_X(x)\,dx \text{ for all } a \in \mathbb{R}, \tag{2.2.6}$$

$$F_Y(b) = \int_{-\infty}^{b} f_Y(y)\,dy \text{ for all } b \in \mathbb{R},. \tag{2.2.7}$$

**Theorem 2.2.7.** *Let $X$ and $Y$ be random variables. The following conditions are equivalent:*

*(i) $X$ and $Y$ are independent;*
*(ii) The joint distribution measure factors:*

$$\mu_{X,Y}(A \times B) = \mu_X(A) \cdot \mu_Y(B) \text{ for all Borel subsets } A \subset \mathbb{R}, B \subset \mathbb{R}; \tag{2.2.8}$$

*(iii) The joint cumulative distribution function factors:*

$$F_{X,Y}(a,b) = F_X(a) \cdot F_Y(b) \text{ for all } a \in \mathbb{R}, b \in \mathbb{R}; \tag{2.2.9}$$

*(iv) The joint moment generating function factors:*

$$\mathbb{E}e^{uX+vY} = \mathbb{E}e^{uX} \cdot \mathbb{E}e^{vY} \tag{2.2.10}$$

*for all $u \in \mathbb{R}$, $v \in \mathbb{R}$ for which the expectations are finite;*
*(v) If there is a joint density, the joint density factors:*

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) \text{ for almost every } x \in \mathbb{R}, y \in \mathbb{R}. \tag{2.2.11}$$

*The conditions above imply, but are not equivalent to:*

*(vi) The expectation factors:*

$$\mathbb{E}[XY] = \mathbb{E}X \cdot \mathbb{E}Y, \tag{2.2.12}$$

*provided $\mathbb{E}|XY| < \infty$.*

Outline of Proof: We sketch the various steps that constitute the proof of this theorem.
**(i)⇒(ii)** Assume that $X$ and $Y$ are independent. Then

$$\begin{aligned}\mu_{X,Y}(A \times B) &= \mathbb{P}\{X \in A \text{ and } Y \in B\} \\ &= \mathbb{P}\big(\{X \in A\} \cap \{Y \in B\}\big) \\ &= \mathbb{P}\{X \in A\} \cdot \mathbb{P}\{Y \in B\}, \\ &= \mu_X(A) \cdot \mu_Y(B).\end{aligned}$$

**(ii)⇒(i)** A typical set in $\sigma(X)$ is of the form $\{X \in A\}$, and a typical set in $\sigma(Y)$ is of the form $\{Y \in B\}$. Assume (ii). Then

$$\begin{aligned}\mathbb{P}\big(\{X \in A\} \cap \{Y \in B\}\big) &= \mathbb{P}\{X \in A \text{ and } Y \in B\} \\ &= \mu_{X,Y}(A \times B) \\ &= \mu_X(A) \cdot \mu_Y(B) \\ &= \mathbb{P}\{X \in A\} \cdot \mathbb{P}\{Y \in B\}.\end{aligned}$$

This shows that every set in $\sigma(X)$ is independent of every set in $\sigma(Y)$.
**(ii)⇒(iii)** Assume (2.2.8). Then

$$\begin{aligned}F_{X,Y}(a,b) &= \mu_{X,Y}\big((-\infty,a] \times (-\infty,b]\big) \\ &= \mu_X(-\infty,a] \cdot \mu_Y(-\infty,b] \\ &= F_X(a) \cdot F_Y(b).\end{aligned}$$

**(iii)⇒(ii)** Equation (2.2.9) implies that (2.2.8) holds whenever $A$ is of the form $A = (-\infty,a]$ and $B$ is of the form $B = (-\infty,b]$. This is enough to establish (2.2.8) for all Borel sets $A$ and $B$, but the details of this are beyond the scope of the text.
**(iii)⇒(v)** If there is a joint density, then (iii) implies

$$\int_{-\infty}^{a}\int_{-\infty}^{b} f_{X,Y}(x,y)\,dy\,dx = \int_{-\infty}^{a} f_X(x)\,dx \cdot \int_{\infty}^{b} f_Y(y)\,dy$$

Differentiating first with respect to $a$ and then with respect to $b$, we obtain

$$f_{X,Y}(a,b) = f_X(a)\cdot f_Y(b),$$

which is just (2.2.11) with different dummy variables.
**(v)⇒(iii)** Assume there is a joint density. If we also assume (2.2.11), we can integrate both sides to get

$$\begin{aligned} F_{X,Y}(a,b) &= \int_{-\infty}^{a}\int_{-\infty}^{b} f_{X,Y}(x,y)\,dy\,dx \\ &= \int_{-\infty}^{a}\int_{-\infty}^{b} f_X(x)\cdot f_Y(y)\,dy\,dx \\ &= \int_{-\infty}^{a} f_X(x)\,dx \cdot \int_{-\infty}^{b} f_Y(y)\,dy \\ &= F_X(a)\cdot F_Y(b). \end{aligned}$$

**(i)⇒(iv)** We first use the "standard machine" as in the proof of Theorem 1.5.1 of Chapter 1, starting with the case that $h$ is the indicator function of a Borel subset of $\mathbb{R}^2$, to show that for every real-valued, Borel measurable function $h(x,y)$ on $\mathbb{R}^2$, we have

$$\mathbb{E}|h(X,Y)| = \int_{\mathbb{R}^2} |h(x,y)|\,d\mu_{X,Y}(x,y),$$

and if this quantity is finite, then

$$\mathbb{E}h(X,Y) = \int_{\mathbb{R}^2} h(x,y)\,d\mu_{X,Y}(x,y). \tag{2.2.13}$$

This is true for any pair of random variables $X$ and $Y$, whether or not they are independent. If $X$ and $Y$ are independent, then the joint distribution $\mu_{X,Y}$ is a product of marginal distributions, and this permits us to rewrite (2.2.13) as

$$\mathbb{E}h(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)\,d\mu_Y(y)\,d\mu_X(x). \tag{2.2.14}$$

We now fix numbers $u$ and $v$ and take $h(x,y) = e^{ux+vy}$. Equation (2.2.14) reduces to

$$\begin{aligned} \mathbb{E}e^{uX+vY} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{ux+vy}\,d\mu_Y(x)\,d\mu_X(x) \\ &= \int_{-\infty}^{\infty} e^{ux}\,d\mu_X(x)\cdot\int_{-\infty}^{\infty} e^{vy}\,d\mu_Y(x) \\ &= \mathbb{E}e^{uX}\cdot\mathbb{E}e^{vY}, \end{aligned}$$

where we have used Theorem 1.5.1 of Chapter 1 for the last step.
The proof **(iv)⇒(i)** is beyond the scope of this text.
**(i)⇒(vi)** In the special case that $h(x, y) = xy$, (2.2.14) reduces to

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty} x \, d\mu_X(x) \cdot \int_{-\infty}^{\infty} y \, d\mu_Y(y) = \mathbb{E}X \cdot \mathbb{E}Y,$$

where again we have used Theorem 1.5.1 of Chapter 1 for the last step. □

*Example 2.2.8 (Independent normal random variables).* Random variables $X$ and $Y$ are independent and standard normal if they have the joint density

$$f_{X,Y}(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} \text{ for all } x \in \mathbb{R}, y \in \mathbb{R}.$$

This is the product of the marginal densities

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \text{ and } f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}.$$

We use the notation

$$N(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{1}{2}x^2} \, dx \tag{2.2.15}$$

for the standard normal cumulative distribution function. The joint cumulative distribution function for $(X, Y)$ factors:

$$\begin{aligned} F_{X,Y}(a, b) &= \int_{-\infty}^{a} \int_{-\infty}^{b} f_X(x) f_Y(y) \, dy \, dx \\ &= \int_{-\infty}^{a} f_X(x) \, dx \cdot \int_{-\infty}^{b} f_Y(y) \, dy \\ &= N(a) \cdot N(b). \end{aligned}$$

The joint distribution $\mu_X$ is the probability measure on $\mathbb{R}^2$ that assigns measure to each Borel set $C \subset \mathbb{R}^2$ equal to the integral of $f_{X,Y}(x, y)$ over $C$. If $C = A \times B$, where $A \in \mathcal{B}(\mathbb{R})$ and $B \in \mathcal{B}(\mathbb{R})$, then $\mu_{X,Y}$ factors:

$$\begin{aligned} \mu_{X,Y}(A \times B) &= \int_A \int_B f_X(x) f_Y(y) \, dy \, dx \\ &= \int_A f_X(x) \, dx \cdot \int_B f_Y(y) \, dy \\ &= \mu_X(A) \cdot \mu_Y(B). \end{aligned}$$

□

We give an example to show that property (vi) of Theorem 2.2.7 does not imply independence. We precede this with a definition.

**Definition 2.2.9.** *Let $X$ be a random variable whose expected value is defined. The* variance of $X$, *denoted Var$(X)$, is*

$$Var(X) = \mathbb{E}\left[(X - \mathbb{E}X)^2\right].$$

*Because* $(X - \mathbb{E}X)^2$ *is nonnegative,* $Var(X)$ *is always defined, although it may be infinite. The* standard deviation of $X$ *is* $\sqrt{Var(X)}$*. The linearity of expectations shows that*

$$Var(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}X)^2.$$

*Let* $Y$ *be another random variable and assume that* $\mathbb{E}X$*,* $Var(X)$*,* $\mathbb{E}Y$ *and* $Var(Y)$ *are all finite. The* covariance of $X$ and $Y$ *is*

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

*The linearity of expectations shows that*

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

*In particular,* $\mathbb{E}[XY] = \mathbb{E}X \cdot \mathbb{E}Y$ *if and only if* $Cov(X, Y) = 0$*. Assume, in addition to the finiteness of expectations and variances, that* $Var(X) > 0$ *and* $Var(Y) > 0$*. The* correlation coefficient of $X$ and $Y$ *is*

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)\, Var(Y)}}.$$

*If* $\rho(X, Y) = 0$ *(or equivalently,* $Cov(X, Y) = 0$*), we say that* $X$ and $Y$ are uncorrelated .

Property (vi) of Theorem 2.2.7 implies that independent random variables are uncorrelated. The converse is not true, even for normal random variables, although it is true of *jointly normal* random variables (see Definition 2.2.11 below).

*Example 2.2.10 (Uncorrelated, dependent normal random variables).* Let $X$ be a standard normal random variable and let $Z$ be independent of $X$ and satisfy[5]

$$\mathbb{P}\{Z = 1\} = \frac{1}{2} \text{ and } \mathbb{P}\{Z = -1\} = \frac{1}{2}. \tag{2.2.16}$$

Define $Y = ZX$. We show below that, like $X$, the random variable $Y$ is standard normal. Furthermore, $X$ and $Y$ are uncorrelated, but they are not independent. The pair $(X, Y)$ does not have a joint density.

Let us first determine the distribution of $Y$. We compute

[5] To construct such random variables, we can choose $\Omega = \{(\omega_1, \omega_2); 0 \le \omega_1 \le 1,\ 0 \le \omega_2 \le 1\}$ to be the unit square and choose $\mathbb{P}$ to be two-dimensional Lebesgue measure, according to which $\mathbb{P}(A)$ is equal to the area of $A$ for every Borel subset of $\Omega$. We then set $X(\omega_1, \omega_2) = N^{-1}(\omega_1)$, which is a standard normal random variable under $\mathbb{P}$ (see Example 1.2.6, for a discussion of this probability integral transform). We set $Z(\omega_1, \omega_2)$ to be $-1$ if $0 \le \omega_2 \le \frac{1}{2}$ and to be 1 if $\frac{1}{2} < \omega_2 \le 1$.

$$\begin{aligned} F_Y(b) &= \mathbb{P}\{Y \le b\} \\ &= \mathbb{P}\{Y \le b \text{ and } Z = 1\} + \mathbb{P}\{Y \le b \text{ and } Z = -1\} \\ &= \mathbb{P}\{X \le b \text{ and } Z = 1\} + \mathbb{P}\{-X \le b \text{ and } Z = -1\}. \end{aligned}$$

Because $X$ and $Z$ are independent, we have

$$\begin{aligned} &\mathbb{P}\{X \le b \text{ and } Z = 1\} + \mathbb{P}\{-X \le b \text{ and } Z = -1\} \\ &\quad = \mathbb{P}\{Z = 1\} \cdot \mathbb{P}\{X \le b\} + \mathbb{P}\{Z = -1\} \cdot \mathbb{P}\{-X \le b\} \\ &\quad = \frac{1}{2} \cdot \mathbb{P}\{X \le b\} + \frac{1}{2} \cdot \mathbb{P}\{-X \le b\} \end{aligned}$$

Because $X$ is a standard normal random variable, so is $-X$. Therefore, $\mathbb{P}\{X \le b\} = \mathbb{P}\{-X \le b\} = N(b)$. It follows that $F_Y(b) = N(b)$; in other words, $Y$ is a standard normal random variable.

Since $\mathbb{E}X = \mathbb{E}Y = 0$, the covariance of $X$ and $Y$ is

$$\text{Cov}(X, Y) = \mathbb{E}[XY] = \mathbb{E}\big[ZX^2\big].$$

Because $Z$ and $X$ are independent, so are $Z$ and $X^2$, and we may use Theorem 2.2.7(vi) to write

$$\mathbb{E}\big[ZX^2\big] = \mathbb{E}Z \cdot \mathbb{E}\big[X^2\big] = 0 \cdot 1 = 0.$$

Therefore, $X$ and $Y$ are uncorrelated.

The random variables $X$ and $Y$ cannot be independent, for if they were, then $|X|$ and $|Y|$ would also be independent (Theorem 2.2.5). But $|X| = |Y|$. In particular,

$$\mathbb{P}\{|X| \le 1, |Y| \le 1\} = \mathbb{P}\{|X| \le 1\} = N(1) - N(-1),$$

and

$$\mathbb{P}\{|X| \le 1\} \cdot \mathbb{P}\{|Y| \le 1\} = \big(N(1) - N(-1)\big)^2.$$

These two expressions are not equal, as they would be for independent random variables.

Finally, we want to examine the joint distribution measure $\mu_{X,Y}$ of $(X, Y)$. Since $|X| = |Y|$, the pair $(X, Y)$ takes values only in the set

$$C = \{(x, y); x = \pm y\}.$$

In other words, $\mu_{X,Y}(C) = 1$ and $\mu_{X,Y}(C^c) = 0$. But $C$ has zero area. It follows that for any nonnegative function $f$, we must have

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{I}_C(x, y) f(x, y)\, dy\, dx = 0.$$

One way of thinking about this is to observe that if we want to integrate a function $\mathbb{I}_C(x, y) f(x, y)$ over the plane $\mathbb{R}^2$, we could first fix $x$ and integrate

out the $y$-variable, but since $f(x,y)\mathbb{I}_C(x,y)$ is zero except when $y = x$ and $y = -x$, we will get zero. When we next integrate out the $x$-variable, we will be integrating the zero function, and the end result will be zero. There cannot be a joint density for $(X,Y)$ because with this choice of the set $C$, the left-hand side of (2.2.4) is one, but the right-hand side is zero. Of course, $X$ and $Y$ have marginal densities, because they are both standard normal. Moreover, the joint cumulative distribution function exists (as it always does). In this case it is

$$\begin{aligned}
&F_{X,Y}(a,b)\\
&= \mathbb{P}\{X \le a \text{ and } Y \le b\}\\
&= \mathbb{P}\{X \le a, X \le b \text{ and } Z = 1\} + \mathbb{P}\{X \le a, -X \le b \text{ and } Z = -1\}\\
&= \mathbb{P}\{Z = 1\} \cdot \mathbb{P}\{X \le \min(a,b)\} + \mathbb{P}\{Z = -1\} \cdot \mathbb{P}\{-b \le X \le a\}\\
&= \frac{1}{2}N(\min(a,b)) + \frac{1}{2}\max\{N(a) - N(-b), 0\}.
\end{aligned}$$

There is no joint density $f_{X,Y}(x,y)$ that permits us to write this function in the form (2.2.5). □

**Definition 2.2.11.** *Two random variables $X$ and $Y$ are said to be* jointly normal *if they have the joint density*

$$\begin{aligned}
&f_{X,Y}(x,y)\\
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right.\right.\\
&\qquad\qquad \left.\left. + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\}, \qquad (2.2.17)
\end{aligned}$$

*where $\sigma_1 > 0$, $\sigma_2 > 0$, $|\rho| < 1$, and $\mu_1$, $\mu_2$ are real numbers. More generally, a random column vector $\mathbf{X} = (X_1, \ldots, X_n)^{tr}$, where the superscript tr denotes transpose, is* jointly normal *if it has joint density*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \, det(C)}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)C^{-1}(\mathbf{x}-\mu)^{tr}\right\}. \qquad (2.2.18)$$

*In equation (2.2.18), $\mathbf{x} = (x_1, \ldots, x_n)$ is a row vector of dummy variables, $\mu = (\mu_1, \ldots, \mu_n)$ is the row vector of expectations, and $C$ is the positive definite matrix of covariances.*

In the case of (2.2.17), $X$ is normal with expectation $\mu_1$ and variance $\sigma_1^2$, $Y$ is normal with expectation $\mu_2$ and variance $\sigma_2^2$, and the correlation between $X$ and $Y$ is $\rho$. The density factors (equivalently, $X$ and $Y$ are independent) if and only if $\rho = 0$. In the case (2.2.18), the density factors into the product of $n$ normal densities (equivalently, the components of $\mathbf{X}$ are independent) if and only if $C$ is a diagonal matrix (all the covariances are zero).

Linear combinations of jointly normal random variables (i.e., sums of constants times the random variables) are jointly normal. Since independent normal random variables are jointly normal, a general method for creating jointly normal random variables is to begin with a set of independent normal random variables and take linear combinations. Conversely, any set of jointly normal random variables can be reduced to linear combinations of independent normal random variables. We do this reduction for a pair of correlated normal random variables in Example 2.2.12 below.

Since the distribution of jointly normal random variables is characterized in terms of means and covariances, and joint normality is preserved under linear combinations, it is not necessary to deal directly with the density when making linear changes of variables. The following example illustrates this point.

*Example 2.2.12.* Let $(X, Y)$ be jointly normal with the density (2.2.17). Define $W = Y - \frac{\rho\sigma_2}{\sigma_1}X$. Then $X$ and $W$ are independent. To verify this, it suffices to show that $X$ and $W$ have covariance zero, since they are jointly normal. We compute

$$\begin{aligned}\mathrm{Cov}(X, W) &= \mathbb{E}\big[(X - \mathbb{E}X)(W - \mathbb{E}W)\big] \\ &= \mathbb{E}\big[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\big] - \mathbb{E}\left[\frac{\rho\sigma_2}{\sigma_1}(X - \mathbb{E}X)^2\right] \\ &= \mathrm{Cov}(X, Y) - \frac{\rho\sigma_2}{\sigma_1}\sigma_1^2 \\ &= 0.\end{aligned}$$

The expectation of $W$ is $\mu_3 = \mu_2 - \frac{\rho\sigma_2\mu_1}{\sigma_1}$, and the variance is

$$\begin{aligned}\sigma_3^2 &= \mathbb{E}\big[(W - \mathbb{E}W)^2\big] \\ &= \mathbb{E}\big[(Y - \mathbb{E}Y)^2\big] - \frac{2\rho\sigma_2}{\sigma_1}\mathbb{E}\big[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\big] + \frac{\rho^2\sigma_2^2}{\sigma_1^2}\mathbb{E}\big[(X - \mathbb{E}X)^2\big] \\ &= (1 - \rho^2)\sigma_2^2.\end{aligned}$$

The joint density of $X$ and $W$ is

$$f_{X,W}(x, w) = \frac{1}{2\pi\sigma_1\sigma_3}\exp\left\{-\frac{(x - \mu_1)^2}{2\sigma_1^2} - \frac{(w - \mu_3)^2}{2\sigma_3^2}\right\}.$$

Note finally that we have decomposed $Y$ into the linear combination

$$Y = \frac{\rho\sigma_2}{\sigma_1}X + W \tag{2.2.19}$$

of a pair of independent normal random variables $X$ and $W$. □

## 2.3 General Conditional Expectations

We consider a random variable $X$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sub-$\sigma$-algebra $\mathcal{G}$ of $\mathcal{F}$. If $X$ is $\mathcal{G}$-measurable, then the information in $\mathcal{G}$

is sufficient to determine the value of $X$. If $X$ is independent of $\mathcal{G}$, then the information in $\mathcal{G}$ provides no help in determining the value of $X$. In the intermediate case, we can use the information in $\mathcal{G}$ to estimate but not precisely evaluate $X$. The *conditional expectation of $X$ given $\mathcal{G}$* is such an estimate.

We have already discussed conditional expectations in the binomial model. Let $\Omega$ be the set of all possible outcomes of $N$ coin tosses, and assume these coin tosses are independent with probability $p$ for head and probability $q = 1 - p$ for tail on each toss. Let $\mathbb{P}(\omega)$ denote the probability of a sequence of coin tosses under these assumptions. Let $n$ be an integer, $1 \leq n \leq N-1$, and let $X$ be a random variable. Then the conditional expectation of $X$ under $\mathbb{P}$, based on the information at time $n$, is (see Definition 2.3.1 of Chapter 2):

$$\begin{aligned}&\mathbb{E}_n[X](\omega_1 \dots \omega_n)\\ &= \sum_{\omega_{n+1}\dots\omega_N} p^{\#H(\omega_{n+1}\dots\omega_N)} q^{\#T(\omega_{n+1}\dots\omega_N)} X(\omega_1 \dots \omega_n \omega_{n+1} \dots \omega_N). \quad (2.3.1)\end{aligned}$$

In the special cases $n = 0$ and $n = N$, we define

$$\mathbb{E}_0 X = \sum_{\omega_0\dots\omega_N} p^{\#H(\omega_0\dots\omega_N)} q^{\#T(\omega_0\dots\omega_N)} X(\omega_0 \dots \omega_N) = \mathbb{E}X, \quad (2.3.2)$$

$$E_N[X](\omega_0 \dots \omega_N) = X(\omega_0 \dots \omega_N). \quad (2.3.3)$$

In (2.3.2) we have the estimate of $X$ based on no information, and in (2.3.3) we have the estimate based on full information.

We need to generalize (2.3.1)–(2.3.3) in a way suitable for a continuous-time model. Toward that end, we examine (2.3.1) within the context of a three-period example. Consider the general three-period model of Figure 2.3.1. We assume the probability of head on each toss is $p$ and the probability of tail is $q = 1 - p$, and we compute

$$\mathbb{E}_2[S_3](HH) = pS_3(HHH) + qS_3(HHT), \quad (2.3.4)$$

$$\mathbb{E}_2[S_3](HT) = pS_3(HTH) + qS_3(HTT), \quad (2.3.5)$$

$$\mathbb{E}_2[S_3](TH) = pS_3(THH) + qS_3(THT), \quad (2.3.6)$$

$$\mathbb{E}_2[S_3](TT) = pS_3(TTH) + qS_3(TTT). \quad (2.3.7)$$

Recall the $\sigma$-algebra $\mathcal{F}_2$ of (2.1.3), which is built up from the four fundamental sets (we call them *atoms* because they are indivisible within the $\sigma$-algebra) $A_{HH}$, $A_{HT}$, $A_{TH}$ and $A_{TT}$ of (2.1.2). We multiply (2.3.4) by $\mathbb{P}(A_{HH}) = p^2$, multiply (2.3.5) by $\mathbb{P}(A_{HT}) = pq$, multiply (2.3.6) by $\mathbb{P}(A_{TH}) = pq$, and multiply (2.3.7) by $\mathbb{P}(A_{TT}) = q^2$. The resulting equations may be written as

$$E_2[S_3](HH)\mathbb{P}(A_{HH}) = \sum_{\omega \in A_{HH}} S_3(\omega)\mathbb{P}(\omega), \quad (2.3.8)$$

$$E_2[S_3](HT)\mathbb{P}(A_{HT}) = \sum_{\omega \in A_{HT}} S_3(\omega)\mathbb{P}(\omega), \quad (2.3.9)$$

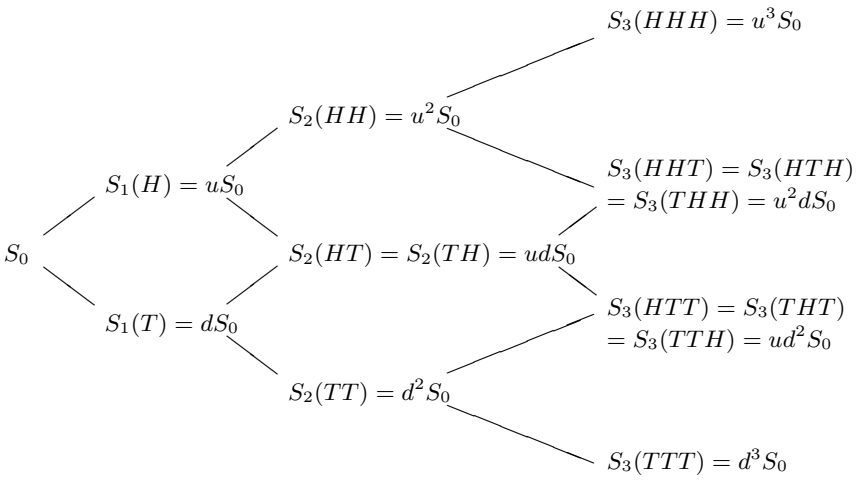


**Fig. 2.3.1.** General three-period model.

$$E_2[S_3](TH)\mathbb{P}(A_{TH}) = \sum_{\omega \in A_{TH}} S_3(\omega)\mathbb{P}(\omega), \tag{2.3.10}$$

$$E_2[S_3](TT)\mathbb{P}(A_{TT}) = \sum_{\omega \in A_{TT}} S_3(\omega)\mathbb{P}(\omega). \tag{2.3.11}$$

We could divide each of these equations by the probability of the atom appearing as the second factor on the left-hand sides, and thereby recover the formulas (2.3.4)–(2.3.7) for the conditional expectations. However, in the continuous-time model, atoms typically have probability zero, and such a step cannot be performed. We therefore take an alternate route here to lay the groundwork for the continuous-time model.

On each of the atoms of $\mathcal{F}_2$, the conditional expectation $\mathbb{E}_2[S_3]$ is constant because the conditional expectation does not depend on the third toss and the atom is created by holding the first two tosses fixed. It follows that the left-hand sides of (2.3.8)–(2.3.11) may be written as integrals of the integrand $\mathbb{E}_2[S_3]$ over the atom. For this purpose, we shall write $\mathbb{E}_2[S_3](\omega) = \mathbb{E}_2[S_3](\omega_1\omega_2\omega_3)$, including the third toss in the argument, even though it is irrelevant. The right-hand sides of these equations are sums, which are Lebesgue integrals on a finite probability space. Using Lebesgue integral notation, we rewrite (2.3.8)–(2.3.11) as

$$\int_{A_{HH}} E_2[S_3](\omega)\, d\mathbb{P}(\omega) = \int_{A_{HH}} S_3(\omega)\, d\mathbb{P}(\omega), \tag{2.3.12}$$

$$\int_{A_{HT}} E_2[S_3](\omega)\, d\mathbb{P}(\omega) = \int_{A_{HT}} S_3(\omega)\, d\mathbb{P}(\omega), \tag{2.3.13}$$

$$\int_{A_{TH}} E_2[S_3](\omega)\, d\mathbb{P}(\omega) = \int_{A_{TH}} S_3(\omega)\, d\mathbb{P}(\omega), \tag{2.3.14}$$

$$\int_{A_{TT}} E_2[S_3](\omega)\, d\mathbb{P}(\omega) = \int_{A_{TT}} S_3(\omega)\, d\mathbb{P}(\omega). \tag{2.3.15}$$

In other words, on each of the atoms the value of the conditional expectation has been chosen to be that constant that yields the same average over the atom as the random variable $S_3$ being estimated.

We turn our attention now to the other sets in $\mathcal{F}_2$. The full list appears in (2.1.3), and every set on the list, except for the empty set, is a finite union of atoms. If we add equations (2.3.12) and (2.3.13), we obtain

$$\int_{A_H} E_2[S_3](\omega)\, d\mathbb{P}(\omega) = \int_{A_H} S_3(\omega)\, d\mathbb{P}(\omega).$$

Similarly, but adding various combinations of (2.3.12)–(2.3.15), we see that

$$\int_{A} \mathbb{E}_2[S_3](\omega)\, d\mathbb{P}(\omega) = \int_{A} S_3(\omega)\, d\mathbb{P}(\omega) \tag{2.3.16}$$

for every set $A \in \mathcal{F}_2$, except possibly for $A = \emptyset$. However, if $A = \emptyset$, equation (2.3.16) still holds, with both sides equal to zero. We call (2.3.16) the *partial averaging property* of conditional expectations, because it says that the conditional expectation and the random variable being estimated give the same value when averaged over "parts" of $\Omega$ (those "parts" that are sets in the conditioning $\sigma$-algebra $\mathcal{F}_2$).

We take (2.3.16) as the defining property of conditional expectations. The precise definition is the following.

**Definition 2.3.1.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{G}$ be a sub-$\sigma$-algebra of $\mathcal{F}$, and let $X$ be a random variable that is either nonnegative or integrable. The* conditional expectation of $X$ given $\mathcal{G}$, *denoted $\mathbb{E}[X|\mathcal{G}]$, is any random variable that satisfies*

*(i)* **(Measurability)** *$\mathbb{E}[X|\mathcal{G}]$ is $\mathcal{G}$-measurable, and*
*(ii)* **(Partial averaging)**

$$\int_{A} \mathbb{E}[X|\mathcal{G}](\omega)\, d\mathbb{P}(\omega) = \int_{A} X(\omega)\, d\mathbb{P}(\omega) \text{ for all } A \in \mathcal{G}. \tag{2.3.17}$$

*If $\mathcal{G}$ is the $\sigma$-algebra generated by some other random variable $W$ (i.e., $\mathcal{G} = \sigma(W)$), we generally write $\mathbb{E}[X|W]$ rather than $\mathbb{E}[X|\sigma(W)]$.*

Property (i) in Definition 2.3.1 guarantees that, although the estimate of $X$ based on the information in $\mathcal{G}$ is itself a random variable, the value of the estimate $\mathbb{E}[X|\mathcal{G}]$ can be determined from the information in $\mathcal{G}$. Property (i) captures the fact that the estimate $\mathbb{E}[X|\mathcal{G}]$ of $X$ is *based on the information in* $\mathcal{G}$. Note in (2.3.4)–(2.3.7) that the conditional expectation $\mathbb{E}_2[S_3]$ is constant on the atoms of $\mathcal{F}_2$; this is property (i) for this case.

The second property ensures that $\mathbb{E}[X|\mathcal{G}]$ is indeed an estimate of $X$. It gives the same averages as $X$ over all the sets in $\mathcal{G}$. If $\mathcal{G}$ has many sets,

which provide a fine resolution of the uncertainty inherent in $\omega$, then this partial averaging property over the "small" sets in $\mathcal{G}$ says that $\mathbb{E}[X|\mathcal{G}]$ is a good estimator of $X$. If $\mathcal{G}$ has only a few sets, this partial averaging property guarantees only that $\mathbb{E}[X|\mathcal{G}]$ is a crude estimate of $X$.

Definition 2.3.1 raises two immediate question. First, does there always exist a random variable $\mathbb{E}[X|\mathcal{G}]$ satisfying properties (i) and (ii)? Second, if there is a random variable satisfying these properties, is it unique? The answer to the first question is yes, and the proof of the existence of $\mathbb{E}[X|\mathcal{G}]$ is based on the Radon-Nikodým Theorem 1.6.7 (see Appendix B). The answer to the second question is a qualified yes, as we now explain. Suppose $Y$ and $Z$ both satisfy conditions (i) and (ii) of Definition 2.3.1. Because both $Y$ and $Z$ are $\mathcal{G}$-measurable, their difference $Y-Z$ is as well, and thus the set $A = \{Y-Z > 0\}$ is in $\mathcal{G}$. From (2.3.17), we have

$$\int_A Y(\omega)\, d\mathbb{P}(\omega) = \int_A X(\omega)\, d\mathbb{P}(\omega) = \int_A Z(\omega)\, d\mathbb{P}(\omega),$$

and thus

$$\int_A \big(Y(\omega) - Z(\omega)\big)\, d\mathbb{P}(\omega) = 0.$$

The integrand is strictly positive on the set $A$, so the only way this equation can hold is for $A$ to have probability zero (i.e, $Y \le Z$ almost surely). We can reverse the roles of $Y$ and $Z$ in this argument and conclude that $Z \le Y$ almost surely. Hence $Y = Z$ almost surely. This means that although different procedures might result in different random variables when determining $\mathbb{E}[X|\mathcal{G}]$, these different random variables will agree almost surely. The set of $\omega$ for which the random variables are different has zero probability.

In this more general context, conditional expectations still have the five fundamental properties developed in Theorem 2.3.2 of Chapter 2 of Volume I. We restate them in the present context.

**Theorem 2.3.2.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G}$ be a sub-$\sigma$-algebra of $\mathcal{F}$.*

*(i)* **(Linearity of conditional expectations)** *If $X$ and $Y$ are integrable random variables and $c_1$ and $c_2$ are constants, then*

$$\mathbb{E}[c_1X + c_2Y|\mathcal{G}] = c_1\mathbb{E}[X|\mathcal{G}] + c_2\mathbb{E}[Y|\mathcal{G}]. \tag{2.3.18}$$

*This equation also holds if we assume that $X$ and $Y$ are nonnegative (rather than integrable) and $c_1$ and $c_2$ are positive, although both sides may be $+\infty$.*

*(ii)* **(Taking out what is known)** *If $X$ and $Y$ are integrable random variables, $Y$ and $XY$ are integrable, and $X$ is $\mathcal{G}$-measurable, then*

$$\mathbb{E}[XY|\mathcal{G}] = X\mathbb{E}[Y|\mathcal{G}]. \tag{2.3.19}$$

*This equation also holds if we assume that $X$ is positive and $Y$ is nonnegative (rather than integrable), although both sides may be $+\infty$.*

*(iii)* **(Iterated conditioning)** *If $\mathcal{H}$ is a sub-$\sigma$ algebra of $\mathcal{G}$ ($\mathcal{H}$ contains less information than $\mathcal{G}$) and $X$ is an integrable random variable, then*

$$\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big] = \mathbb{E}\big[X\big|\mathcal{H}\big]. \tag{2.3.20}$$

*This equation also holds if we assume that $X$ is nonnegative (rather than integrable), although both sides may be $+\infty$.*

*(iv)* **(Independence)** *If $X$ is integrable and independent of $\mathcal{G}$, then*

$$\mathbb{E}[X|\mathcal{G}] = \mathbb{E}X. \tag{2.3.21}$$

*This equation also holds if we assume that $X$ is nonnegative (rather than integrable), although both sides may be $+\infty$.*

*(v)* **(Conditional Jensen's inequality)** *If $\varphi(x)$ is a convex function of a dummy variable $x$ and $X$ is integrable, then*

$$\mathbb{E}\big[\varphi(X)|\mathcal{G}\big] \geq \varphi\big(E[X|\mathcal{G}]\big). \tag{2.3.22}$$

Discussion and sketch of proof: We take each of these properties in turn.

**(i)** Linearity allows us to separate the estimation of random variables into estimation of separate pieces, and then add the estimates of the pieces to estimate the whole. To verify that $\mathbb{E}[c_1X + c_2Y|\mathcal{G}]$ is given by the right-hand side of (2.3.18), we observe that the right-hand side is $\mathcal{G}$-measurable because $\mathbb{E}[X|\mathcal{G}]$ and $\mathbb{E}[Y|\mathcal{G}]$ are $\mathcal{G}$ measurable, and then must check the partial averaging property (ii) of Definition 2.3.1. Using the fact that $E[X|\mathcal{G}]$ and $E[Y|\mathcal{G}]$ themselves satisfy the partial averaging property, we have for every $A \in \mathcal{G}$ that

$$\begin{aligned}
&\int_A \big(c_1\mathbb{E}[X|\mathcal{G}](\omega) + c_2\mathbb{E}[Y|\mathcal{G}](\omega)\big)\, d\mathbb{P}(\omega) \\
&= c_1 \int_A \mathbb{E}[X|\mathcal{G}](\omega)\, d\mathbb{P}(\omega) + c_2 \int_A \mathbb{E}[Y|\mathcal{G}](\omega)\, d\mathbb{P}(\omega) \\
&= c_1 \int_A X(\omega)\, d\mathbb{P}(\omega) + c_2 \int_A Y(\omega)\, d\mathbb{P}(\omega) \\
&= \int_A \big(c_1X(\omega) + c_2Y(\omega)\big)\, d\mathbb{P}(\omega),
\end{aligned}$$

which shows that $c_1\mathbb{E}[X|\mathcal{G}] + c_2\mathbb{E}[Y|\mathcal{G}]$ satisfies the partial averaging property that characterizes $\mathbb{E}[c_1X + c_2Y|\mathcal{G}]$ and hence is $\mathbb{E}[c_1X + c_2Y|\mathcal{G}]$.

**(ii)** Taking out what is known permits us to remove $X$ from the estimation problem if its value can be determined from the information in $\mathcal{G}$. To estimate $XY$, it suffices to estimate $Y$ alone, and then multiply the estimate by $X$. To prove (2.3.19), we observe first that $X\mathbb{E}[Y|\mathcal{G}]$ is $\mathcal{G}$-measurable because both $X$ and $\mathbb{E}[Y|\mathcal{G}]$ are $\mathcal{G}$-measurable. We must check the partial averaging property.

Let us first consider the case that $X$ is a $\mathcal{G}$-measurable indicator random variable (i.e., $X = \mathbb{I}_B$, where $B$ is a set in $\mathcal{G}$). Using the fact that $\mathbb{E}[Y|\mathcal{G}]$ itself satisfies the partial averaging property, we have for every set $A \in \mathcal{G}$ that

$$\begin{aligned}\int_A X(\omega)\mathbb{E}[Y|\mathcal{G}](\omega)\,d\mathbb{P}(\omega) &= \int_{A\cap B} \mathbb{E}[Y|\mathcal{G}](\omega)\,d\mathbb{P}(\omega)\\ &= \int_{A\cap B} Y(\omega)\,d\mathbb{P}(\omega)\\ &= \int_A X(\omega)Y(\omega)\,d\mathbb{P}(\omega). \qquad (2.3.23)\end{aligned}$$

Having proved (2.3.23) for $\mathcal{G}$-measurable indicator random variables $X$, we may use the standard machine developed in the proof of Theorem 1.5.1 of Chapter 1 to obtain this equation for all $\mathcal{G}$-measurable random variables $X$ for which $XY$ is integrable. This shows that $X\mathbb{E}[Y|\mathcal{G}]$ satisfies the partial averaging condition that characterizes $\mathbb{E}[XY|\mathcal{G}]$, and hence $X\mathbb{E}[Y|\mathcal{G}]$ is the conditional expectation $\mathbb{E}[XY|\mathcal{G}]$.

**(iii)** If we estimate $X$ based on the information in $\mathcal{G}$ and then estimate the estimate based on the less information in $\mathcal{H}$, we obtain the random variable we would have gotten by estimating $X$ directly based on the smaller amount of information in $\mathcal{H}$. To prove this, we observe first that $\mathbb{E}[X|\mathcal{H}]$ is $\mathcal{H}$-measurable, by definition. The partial averaging property that characterizes $\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big]$ is

$$\int_A \mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big](\omega)\,d\mathbb{P}(\omega) = \int_A \mathbb{E}[X|\mathcal{G}](\omega)\,\mathbb{P}(\omega) \text{ for all } A\in\mathcal{H}.$$

In order to prove (2.3.20), we must show that we can replace $\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big]$ on the left-hand side of this equation by $\mathbb{E}[X|\mathcal{H}]$. But when $A\in\mathcal{H}$, it is also in $\mathcal{G}$, and the partial averaging properties for $\mathbb{E}[X|\mathcal{H}]$ and $\mathbb{E}[X|\mathcal{G}]$ imply

$$\int_A \mathbb{E}[X|\mathcal{H}](\omega)\,d\mathbb{P}(\omega) = \int_A X(\omega)\,d\mathbb{P}(\omega) = \int_A \mathbb{E}[X|\mathcal{G}](\omega)\,d\mathbb{P}(\omega).$$

This shows that $\mathbb{E}[X|\mathcal{H}]$ satisfies the partial averaging property that characterizes $\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big]$, and hence $\mathbb{E}[X|\mathcal{H}]$ is $\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big]$.

**(iv)** If $X$ is independent of the information in $\mathcal{G}$, then the best estimate we can give of $X$ is its expected value. This is also the estimate we would give based on no information. To prove this, we observe first that $\mathbb{E}X$ is $\mathcal{G}$-measurable. Indeed, $\mathbb{E}X$ is not random and so is measurable with respect to every $\sigma$-algebra. We need to verify that $\mathbb{E}X$ satisfies the partial averaging property that characterizes $\mathbb{E}[X|\mathcal{G}]$, i.e.,

$$\int_A \mathbb{E}X\,d\mathbb{P}(\omega) = \int_A X(\omega)\,d\mathbb{P}(\omega) \text{ for all } A\in\mathcal{G}. \qquad (2.3.24)$$

Let us consider first the case that $X$ is an indicator random variable independent of $\mathcal{G}$ (i.e., $X=\mathbb{I}_B$ where the set $B$ is independent of $\mathcal{G}$). For all $A\in\mathcal{G}$, we have then

$$\int_A X(\omega)\,d\mathbb{P}(\omega) = \mathbb{P}(A\cap B) = \mathbb{P}(A)\cdot\mathbb{P}(B) = \mathbb{P}(A)\mathbb{E}X = \int_A \mathbb{E}X\,d\mathbb{P}(\omega),$$

and (2.3.24) holds. We complete the proof using the standard machine developed in the proof of Theorem 1.5.1 of Chapter 1.
**(v)** Using the linearity of conditional expectations, we can repeat the proof of Theorem 2.2.5 of Chapter 2 to prove the conditional Jensen's inequality. □

We note that $\mathbb{E}[X|\mathcal{G}]$ is an unbiased estimator of $X$:

$$\mathbb{E}\big(\mathbb{E}[X|\mathcal{G}]\big) = \mathbb{E}X. \tag{2.3.25}$$

This equality is just the partial averaging property (2.3.17) with $A = \Omega$.

*Example 2.3.3.* Let $X$ and $Y$ be a pair of jointly normal random variables, with joint density (2.2.17). As in Example 2.2.12, define $W = Y - \frac{\rho\sigma_2}{\sigma_1}X$ so that $X$ and $W$ are independent and (2.2.19) holds:

$$Y = \frac{\rho\sigma_2}{\sigma_1}X + W. \tag{2.2.19}$$

In Example 2.2.12, we saw that $W$ is normal with mean $\mu_3 = \mu_2 - \frac{\rho\sigma_2\mu_1}{\sigma_1}$ and variance $\sigma_3^2 = (1-\rho^2)\sigma_2^2$. Let us take the conditioning $\sigma$-algebra to be $\mathcal{G} = \sigma(X)$. (When $\mathcal{G}$ is generated by a random variable $X$, it is customary to write $\mathbb{E}[\cdots|X]$ rather than $\mathbb{E}[\cdots|\sigma(X)]$.) We estimate $Y$, based on $X$, using (2.2.19) above and properties (i) (Linearity) and (iv) (Independence) from Theorem 2.3.2 to get the linear regression equation

$$\mathbb{E}[Y|X] = \frac{\rho\sigma_2}{\sigma_1}X + \mathbb{E}W = \frac{\rho\sigma_2}{\sigma_1}(X-\mu_1) + \mu_2. \tag{2.3.26}$$

Note that the right-hand side of (2.3.26) is random but is $\sigma(X)$-measurable (i.e., if we know the information in $\sigma(X)$, which is the same as knowing the value of $X$, then we can evaluate $\mathbb{E}[Y|X]$). Subtracting (2.3.26) from (2.2.19), we see that the error made by the estimator is,

$$Y - \mathbb{E}[Y|X] = W - \mathbb{E}W.$$

The error is random, with expected value zero (the estimator is unbiased), and is independent of the estimate $\mathbb{E}[Y|X]$ (because $\mathbb{E}[Y|X]$ is $\sigma(X)$-measurable and $W$ is independent of $\sigma(X)$). The independence between the error and the conditioning random variable $X$ is a consequence of the joint normality in the example. In general, the error and the conditioning random variable are uncorrelated, but not necessarily independent; see Exercise 2.6.8. □

The Independence Lemma 2.5.3 of Chapter 2 of Volume I now takes the following more general form.

**Lemma 2.3.4 (Independence).** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{G}$ be a sub-$\sigma$-algebra of $\mathcal{F}$. Suppose the random variables $X_1, \ldots, X_K$ are $\mathcal{G}$-measurable and the random variables $Y_1, \ldots, Y_L$ are independent of $\mathcal{G}$. Let*

*$f(x_1, \ldots, x_K, y_1, \ldots, y_L)$ be a function of the dummy variables $x_1, \ldots, x_K$ and $y_1, \ldots, y_L$, and define*

$$g(x_1, \ldots, x_K) = \mathbb{E} f(x_1, \ldots, x_K, Y_1, \ldots, Y_L). \tag{2.3.27}$$

*Then*

$$\mathbb{E}\big[f(X_1, \ldots, X_K, Y_1, \ldots, Y_L)\big|\mathcal{G}\big] = g(X_1, \ldots, X_K). \tag{2.3.28}$$

As with Lemma 2.5.3 of Volume I, the idea here is that since the information in $\mathcal{G}$ is sufficient to determine the values of $X_1, \ldots, X_K$, we should hold these random variables constant when estimating $f(X_1, \ldots, X_K, Y_1, \ldots, Y_K)$. The other random variables, $Y_1, \ldots, Y_L$, are independent of $\mathcal{G}$, and so we should integrate them out without regard to the information in $\mathcal{G}$. These two steps, holding $X_1, \ldots, X_K$ constant and integrating out $Y_1, \ldots, Y_L$, are accomplished by (2.3.27). We get an estimate that depends on the values of $X_1, \ldots, X_K$, and to capture this fact, we replaced the dummy (constant) variables $x_1, \ldots, x_K$ by the random variables $X_1, \ldots, X_K$ at the last step. Although Lemma 2.5.3 of Volume I has a relatively simple proof, the proof of Lemma 2.3.4 requires some measure-theoretic ideas beyond the scope of this text, and will not be given.

*Example 2.3.3 continued.* Continuing with the notation of Example 2.3.3, suppose we want to estimate some function $f(x, y)$ of the random variables $X$ and $Y$, based on knowledge of $X$. We cannot use the Independence Lemma directly because $X$ and $Y$ are not independent. However, we can write $Y$ as $Y = \frac{\rho\sigma_2}{\sigma_1}X + W$. Because $X$ is $\sigma(X)$ measurable, $W$ is independent of $\sigma(X)$ and $W$ is normal with mean $\mu_3$ and variance $\sigma_3^2$, the Independence Lemma tells us how to compute $\mathbb{E}[f(X, Y)|X]$. We should first replace the random variable $X$ by a dummy variable $x$ and then take the expectation (i.e, integrate with respect to the distribution of $W$). Thus, we define

$$\begin{aligned} g(x) &= \mathbb{E} f\left(x, \frac{\rho\sigma_1}{\sigma_1}x + W\right) \\ &= \frac{1}{\sigma_3\sqrt{2\pi}} \int_{-\infty}^{\infty} f\left(x, \frac{\rho\sigma_1}{\sigma_2}x + w\right) \exp\left\{-\frac{(w-\mu_3)^2}{2\sigma_3^2}\right\} dw. \end{aligned}$$

Then

$$\mathbb{E}\big[f(X, Y)\big|X\big] = g(X).$$

Our final answer is random, but $\sigma(X)$-measurable, as it should be. □

We have all the tools required to introduce martingales and Markov processes in a continuous-time framework. The definitions are provided below. Examples will be given after we construct Brownian motion and Itô integrals in the next chapters.

**Definition 2.3.5.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $T$ be a fixed positive number, and let $\mathcal{F}(t)$, $0 \le t \le T$, be a filtration of sub-$\sigma$-algebras of $\mathcal{F}$. Consider an adapted stochastic process $M(t)$, $0 \le t \le T$.*

*(i) If*

$$\mathbb{E}[M(t)|\mathcal{F}(s)] = M(s) \text{ for all } 0 \le s \le t \le T,$$

*we say this process is a* martingale. *It has no tendency to rise or fall.*

*(ii) If*

$$\mathbb{E}[M(t)|\mathcal{F}(s)] \ge M(s) \text{ for all } 0 \le s \le t \le T,$$

*we say this process is a* submartingale. *It has no tendency to fall; it may have a tendency to rise.*

*(iii) If*

$$\mathbb{E}[M(t)|\mathcal{F}(s)] \le M(s) \text{ for all } 0 \le s \le t \le T,$$

*we say this process is a* supermartingale. *It has no tendency to rise; it may have a tendency to fall.*

**Definition 2.3.6.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $T$ be a fixed positive number, and let $\mathcal{F}(t)$, $0 \le t \le T$, be a filtration of sub-$\sigma$-algebras of $\mathcal{F}$. Consider an adapted stochastic process $X(t)$, $0 \le t \le T$. Assume that for all $0 \le s \le t \le T$ and for every nonnegative, Borel-measurable function $f$, there is another Borel-measurable function $g$ such that*

$$\mathbb{E}[f(X(t))|\mathcal{F}(s)] = g(X(s)). \tag{2.3.29}$$

*Then we say that the $X$ is a* Markov process.

*Remark 2.3.7.* In Definition 2.3.6, the function $f$ is permitted to depend on $t$, and the function $g$ will depend on $s$. These dependencies are not indicated in (2.3.29) because we wish there to emphasize how the dependence on the sample point $\omega$ works (i.e, the right-hand side depends on $\omega$ only through the random variable $X(s)$). If we indicate the dependence on time by writing $f(t,x)$ rather than $f(x)$, we can write $f(s,x)$ rather than $g(x)$ (we do not need different symbols $f$ and $g$ because the time variables $t$ and $s$ indicate we are dealing with different functions of $x$ at the different times), and can rewrite (2.3.29) as

$$\mathbb{E}[f(t,X(t))|\mathcal{F}(s)] = f(s,X(s)),\ 0 \le s \le t \le T. \tag{2.3.30}$$

Ultimately, we shall see that when we regard $f(t,x)$ as a function of two variables this way, (2.3.30) implies that it satisfies a partial differential equation. This partial differential equation gives us a way to determine $f(s,x)$ if we know $f(t,x)$. The Black-Scholes-Merton partial differential equation is a special case of this. □

## 2.4 Summary

In measure-theoretic probability, information is modeled using $\sigma$-algebras. The information associated with a $\sigma$-algebra $\mathcal{G}$ can be thought of as follows. A random experiment is performed, an outcome $\omega$ is determined, but the value of $\omega$ is not revealed. Instead, for each set in the $\sigma$-algebra $\mathcal{G}$, we are told whether $\omega$ is in the set. The more sets there are on $\mathcal{G}$, the more information this provides. If $\mathcal{G}$ is the trivial $\sigma$-algebra containing only $\emptyset$ and $\Omega$, this provides no information.

A random variable $X$ is *$\mathcal{G}$-measurable* if and only if the set $\{X \in B\} = \{\omega \in \Omega; X(\omega) \in B\}$ is in $\mathcal{G}$ for every Borel subset of $\mathbb{R}$. In this case, the information in $\mathcal{G}$ is enough to determine the value of the random variable $X(\omega)$, even though it may not be enough to determine the value $\omega$ of the outcome of the random experiment.

At the other extreme, the information in a $\sigma$-algebra $\mathcal{G}$ may be irrelevant to the determination of the value of $X$. In this case, we say that $\mathcal{G}$ and $X$ are *independent*. This idea is captured mathematically by Definition 2.2.3, which says that $X$ and $\mathcal{G}$ are independent if for every set $A \in \mathcal{G}$ and every Borel subset $B$ of $\mathbb{R}$, we have

$$\mathbb{P}\{\omega \in \Omega; \omega \in A \text{ and } X(\omega) \in B\} = \mathbb{P}(A) \cdot \mathbb{P}\{\omega \in \Omega; X(\omega) \in B\}.$$

Two random variables $X$ and $Y$ are independent if and only if the *$\sigma$ algebra generated by $X$*, defined to be the collection of sets of the form $\{X \in B\}$, is independent of the $\sigma$-algebra generated by $Y$. In other words, $X$ and $Y$ are independent if and only if

$$\mathbb{P}\{X \in B \text{ and } Y \in C\} = \mathbb{P}\{X \in B\} \cdot \mathbb{P}\{X \in C\} \text{ for all } B \in \mathcal{B}(\mathbb{R}), C \in \mathcal{B}(\mathbb{R}),$$

where $\mathcal{B}(\mathbb{R})$ denotes the $\sigma$-algebra of Borel subsets of $\mathbb{R}$. There are several equivalent ways to describe independence between two random variables, having to do with factoring the joint cumulative distribution function, factoring the joint moment generating function, and factoring the joint density (if there is a joint density). These are set out in Theorem 2.2.7. Independence implies uncorrelatedness, but uncorrelated random variables do not need to be independent. Jointly normally distributed random variables (Definition 2.2.11) are uncorrelated if and only if they are independent, but normally distributed random variables do not need to be *jointly* normal.

Often we find ourselves between the two extremes of random variables $X$ that are $\mathcal{G}$-measurable and random variables $X$ that are independent of $\mathcal{G}$. In such a case, the information in $\mathcal{G}$ is relevant to the determination of the value of $X$, but is not sufficient to completely determine it. We then want to use the information in $\mathcal{G}$ to estimate $X$. We denote our estimate by $\mathbb{E}[X|\mathcal{G}]$, and call this the *conditional expectation of $X$ given $\mathcal{G}$*. This is itself a random variable, but one which is $\mathcal{G}$-measurable (i.e., one which we can evaluate using only the information in $\mathcal{G}$). To be sure this is a good estimate of $X$, we require that it satisfy the *partial averaging property* (see Definition 2.3.1(ii)):

$$\int_A \mathbb{E}[X|\mathcal{G}](\omega)\, d\mathbb{P}(\omega) = \int_A X(\omega)\, d\mathbb{P}(\omega) \text{ for every } A \in \mathcal{G}.$$

Condition expectations behave in many ways like expectations, except that expectations do not depend on $\omega$, and conditional expectations do. The principal properties of conditional expectations are provided in Theorem 2.3.2, and these are reported briefly here.

**Linearity:** $\mathbb{E}[c_1X + c_2Y|\mathcal{G}] = c_1\mathbb{E}[X|\mathcal{G}] + c_2\mathbb{E}[Y|\mathcal{G}]$.
**Taking out what is known:** $\mathbb{E}[XY|\mathcal{G}] = X\mathbb{E}[Y|\mathcal{G}]$ if $X$ is $\mathcal{G}$-measurable.
**Iterated conditioning:** $\mathbb{E}\big[\mathbb{E}[X|\mathcal{G}]\big|\mathcal{H}\big] = \mathbb{E}[X|\mathcal{H}]$ if $\mathcal{H}$ is a sub-$\sigma$-algebra of $\mathcal{G}$.
**Independence:** $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}X$ if $X$ is independent of $\mathcal{G}$.
**Jensen's inequality:** $\mathbb{E}[\varphi(X)|\mathcal{G}] \geq \varphi(\mathbb{E}[X|\mathcal{G}])$ if $\varphi$ is convex.

In continuous-time finance, we work within the framework of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We normally have a fixed final time $T$, and then have a *filtration*, which is a collection of $\sigma$-algebras $\{\mathcal{F}(t); 0 \leq t \leq T\}$, indexed by the time variable $t$. We interpret $\mathcal{F}(t)$ as the information available at time $t$. For $0 \leq s \leq t \leq T$, every set in $\mathcal{F}(s)$ is also in $\mathcal{F}(t)$. In other words, information increases over time. Within this context, an *adapted stochastic process* is a collection of random variables $\{X(t); 0 \leq t \leq T\}$, also indexed by time, such that for every $t$, $X(t)$ is $\mathcal{F}(t)$-measurable; the information at time $t$ is sufficient to evaluate the random variable $X(t)$. We think of $X(t)$ as the price of some asset at time $t$ and $\mathcal{F}(t)$ as the information obtained by watching all the prices in the market up to time $t$.

Two important classes of adapted stochastic processes are *martingales* and *Markov processes*. These are defined in Definitions 2.3.5 and 2.3.6, respectively. A martingale has the property that

$$\mathbb{E}[M(t)|\mathcal{F}(s)] = M(s) \text{ for all } 0 \leq s \leq t \leq T.$$

If $\mathbb{E}[M(t)|\mathcal{F}(s)] \geq M(s)$ when $0 \leq s \leq t \leq T$, we have a *submartingale*. If the inequality is reversed, we have a *supermartingale*. A Markov process has the property that whenever $0 \leq s \leq t \leq T$ and we are given a function $f$, there is another function $g$ such that

$$\mathbb{E}[f(X(t))|\mathcal{F}(s)] = g(X(s)).$$

The important feature here is that the estimate of $f(X(t))$ made at time $s$ depends only on the process value $X(s)$ at time $s$, and not on the path of the process before time $s$.

A useful tool for establishing that a process is Markov is the *Independence Lemma*, Lemma 2.3.4. The simplest version of this says that if $X$ is a $\mathcal{G}$-measurable random variable and $Y$ is independent of $\mathcal{G}$, then

$$\mathbb{E}[f(X,Y)|\mathcal{G}] = g(X),$$

where $g(x) = \mathbb{E}f(x,Y)$.

## 2.5 Notes

In the measure-theoretic view of probability theory, a conditional expectation is itself a random variable, measurable with respect to the conditioning $\sigma$-algebra. This point of view is indispensable for treating the rather complicated conditional expectations that arise in martingale theory. It was invented by Kolmogorov [97]. The term *martingale* was apparently first used by Ville [146], who assigned the name to a betting strategy. The concept dates back to 1934 work of Lévy. The first systematic treatment of martingales was provided by Doob [50].

## 2.6 Exercises

**Exercise 2.6.1.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a general probability space, and suppose a random variable $X$ on this space is measurable with respect to the trivial $\sigma$-algebra $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Show that $X$ is not random (i.e., there is a constant $c$ such that $X(\omega) = c$ for all $\omega \in \Omega$). Such a random variable is called *degenerate*.

**Exercise 2.6.2.** Independence of random variables can be affected by changes of measure. To illustrate this point, consider the space of two coin tosses $\Omega_2 = \{HH, HT, TH, TT\}$, and let stock prices be given by

$$S_0 = 4,\ S_1(H) = 8,\ S_1(T) = 2,$$
$$S_2(HH) = 16,\ S_2(HT) = S_2(TH) = 4,\ S_2(TT) = 1.$$

Consider two probability measures given by

$$\widetilde{\mathbb{P}}(HH) = \tfrac{1}{4},\ \widetilde{\mathbb{P}}(HT) = \tfrac{1}{4},\ \widetilde{\mathbb{P}}(TH) = \tfrac{1}{4},\ \widetilde{\mathbb{P}}(TT) = \tfrac{1}{4},$$
$$\mathbb{P}(HH) = \tfrac{4}{9},\ \mathbb{P}(HT) = \tfrac{2}{9},\ \mathbb{P}(TH) = \tfrac{2}{9},\ \mathbb{P}(TT) = \tfrac{1}{9}.$$

Define the random variable

$$X = \begin{cases} 1, \text{ if } S_2 = 4, \\ 0, \text{ if } S_2 \neq 4. \end{cases}$$

(i) List all the sets in $\sigma(X)$.
(ii) List all the sets in $\sigma(S_1)$.
(iii) Show that $\sigma(X)$ and $\sigma(S_1)$ are independent under the probability measure $\widetilde{\mathbb{P}}$.
(iv) Show that $\sigma(X)$ and $\sigma(S_1)$ are not independent under the probability measure $\mathbb{P}$.
(v) Under $\mathbb{P}$, we have $\mathbb{P}\{S_1 = 8\} = \frac{2}{3}$ and $\mathbb{P}\{S_1 = 2\} = \frac{1}{3}$. Explain intuitively why, if you are told that $X = 1$, you would want to revise your estimate of the distribution of $S_1$.

**Exercise 2.6.3 (Rotating the axes).** Let $X$ and $Y$ be independent standard normal random variables. Let $\theta$ be a constant, and define random variables

$$V = X\cos\theta + Y\sin\theta \text{ and } W = -X\sin\theta + Y\cos\theta.$$

Show that $V$ and $W$ are independent standard normal random variables.

**Exercise 2.6.4.** In Example 2.2.8, $X$ is a standard normal random variable and $Z$ is an independent random variable satisfying

$$\mathbb{P}\{Z=1\} = \mathbb{P}\{Z=-1\} = \frac{1}{2}.$$

We defined $Y = XZ$ and showed the $Y$ is standard normal. We established that although $X$ and $Y$ are uncorrelated, they are not independent. In this exercise, we use moment generating functions to show that $Y$ is standard normal and $X$ and $Y$ are not independent.

(i) Establish the joint moment generating function formula

$$\mathbb{E}e^{uX+vY} = e^{\frac{1}{2}(u^2+v^2)} \cdot \frac{e^{uv}+e^{-uv}}{2}.$$

(ii) Use the formula above to show that $\mathbb{E}e^{vY} = e^{\frac{1}{2}v^2}$. This is the moment generating formula for a standard normal random variable, and thus $Y$ must be a standard normal random variable.

(iii) Use the formula in (i) and Theorem 2.2.7(iv) to show that $X$ and $Y$ are not independent.

**Exercise 2.6.5.** Let $(X,Y)$ be a pair of random variables with joint density function

$$f_{X,Y}(x,y) = \begin{cases} \frac{2|x|+y}{\sqrt{2\pi}} \exp\left\{-\frac{(2|x|+y)^2}{2}\right\} & \text{if } y \geq -|x|, \\ 0 & \text{if } y < -|x|. \end{cases}$$

Show that $X$ and $Y$ are standard normal random variables and that they are uncorrelated but not independent.

**Exercise 2.6.6.** Consider a probability space $\Omega$ with four elements, which we call $a$, $b$, $c$ and $d$ (i.e., $\Omega = \{a,b,c,d\}$). The $\sigma$-algebra $\mathcal{F}$ is the collection of all subsets of $\Omega$ (i.e., the sets in $\mathcal{F}$) are

$$\begin{aligned} &\Omega,\ \{a,b,c\},\ \{a,b,d\},\ \{a,c,d\},\ \{b,c,d\} \\ &\{a,b\},\ \{a,c\},\ \{a,d\},\ \{b,c\},\ \{b,d\},\ \{c,d\}, \\ &\{a\},\ \{b\},\ \{c\},\ \{d\},\ \emptyset. \end{aligned}$$

We define a probability measure $\mathbb{P}$ by specifying that

$$\mathbb{P}\{a\} = \frac{1}{6},\ \mathbb{P}\{b\} = \frac{1}{3},\ \mathbb{P}\{c\} = \frac{1}{4},\ \mathbb{P}\{d\} = \frac{1}{4},$$

and, as usual, the probability of every other set in $\mathcal{F}$ is the sum of the probabilities of the elements in the set, e.g., $\mathbb{P}\{a,b,c\} = \mathbb{P}\{a\} + \mathbb{P}\{b\} + \mathbb{P}\{c\} = \frac{3}{4}$.

We next define two random variables, $X$ and $Y$, by the formulas

$$X(a) = 1, X(b) = 1, X(c) = -1, X(d) = -1$$
$$Y(a) = 1, Y(b) = -1, Y(c) = 1, Y(d) = -1.$$

We then define $Z = X + Y$.

(i) List the sets in $\sigma(X)$.
(ii) Determine $\mathbb{E}[Y|X]$ (i.e, specify the values of this random variable for $a$, $b$, $c$ and $d$). Verify that the partial averaging property is satisfied.
(iii) Determine $\mathbb{E}[Z|X]$. Again, verify the partial averaging property.
(iv) Compute $\mathbb{E}[Z|X] - \mathbb{E}[Y|X]$. Citing the appropriate properties of conditional expectation from Theorem 2.3.2, explain why you get $X$.

**Exercise 2.6.7.** Let $Y$ be an integrable random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G}$ be a sub-$\sigma$-algebra of $\mathcal{F}$. Based on the information in $\mathcal{G}$, we can form the estimate $\mathbb{E}[Y|\mathcal{G}]$ of $Y$ and define the error of the estimation $\text{Err} = Y - \mathbb{E}[Y|\mathcal{G}]$. This is a random variable with expectation zero and some variance $\text{Var}(\text{Err})$. Let $X$ be some other $\mathcal{G}$-measurable random variable, which we can regard as another estimate of $Y$. Show that

$$\text{Var}(Y - X) \leq \text{Var}(\text{Err}).$$

In other words, the estimate $\mathbb{E}[Y|\mathcal{G}]$ minimizes the variance of the error among all estimates based on the information in $\mathcal{G}$. (Hint: Let $\mu = \mathbb{E}(Y - X)$. Compute the variance of $Y - X$ as

$$\mathbb{E}\left[(Y - X - \mu)^2\right] = \mathbb{E}\left[\big((Y - \mathbb{E}[Y|\mathcal{G}]) + (\mathbb{E}[Y|\mathcal{G}] - X - \mu)\big)^2\right].$$

Multiply out the right-hand side and use iterated conditioning to show the cross-term is zero.)

**Exercise 2.6.8.** Let $X$ and $Y$ be integrable random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then $Y = Y_1 + Y_2$, where $Y_1 = \mathbb{E}[Y|X]$ is $\sigma(X)$-measurable and $Y_2 = Y - E[Y|X]$. Show that $Y_2$ and $X$ are uncorrelated. More generally, show that $Y_2$ is uncorrelated with every $\sigma(X)$-measurable random variable.

**Exercise 2.6.9.** Let $X$ be a random variable.

(i) Give an example of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable $X$ defined on this probability space, and a function $f$ so that the $\sigma$-algebra generated by $f(X)$ is not the trivial $\sigma$-algebra $\{\emptyset, \Omega\}$ but is strictly smaller than the $\sigma$-algebra generated by $X$.
(ii) Can the $\sigma$-algebra generated by $f(X)$ ever be strictly larger than the $\sigma$-algebra generated by $X$?

**Exercise 2.6.10.** Let $X$ and $Y$ be random variables (on some unspecified probability space $(\Omega, \mathcal{F}, \mathbb{P})$), and assume they have a joint density $f_{X,Y}(x,y)$. In particular, for every Borel subset $C$ of $\mathbb{R}^2$, we have

$$\mathbb{P}\{(X,Y) \in C\} = \int_C f_{X,Y}(x,y)\,dx\,dy.$$

In elementary probability, one learns to compute $\mathbb{E}[Y|X=x]$, which is a *non-random* function of the *dummy variable* $x$, by the formula

$$\mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy, \tag{2.6.1}$$

where $f_{Y|X}(y|x)$ is the *conditional density* defined by

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

The denominator in this expression, $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,\eta)d\eta$, is the *marginal density* of $X$, and we must assume it is strictly positive for every $x$. We introduce the symbol $g(x)$ for the function $\mathbb{E}[Y|X=x]$ defined by (2.6.1), i.e,

$$g(x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy = \int_{-\infty}^{\infty} \frac{y f_{X,Y}(x,y)}{f_X(x)}\,dy.$$

In measure-theoretic probability, we talk about the *random variable* $\mathbb{E}[Y|X]$. This exercise is to show that when there is a joint density for $(X,Y)$, this random variable can be obtained by substituting the random variable $X$ in place of the dummy variable $x$ in the function $g(x)$. In other words, this exercise is to show that

$$\mathbb{E}[Y|X] = g(X).$$

(We introduced the symbol $g(x)$ in order to avoid the mathematically confusing expression $E[Y|X=X]$.)

Since $g(X)$ is obviously $\sigma(X)$-measurable, to verify that $\mathbb{E}[Y|X] = g(X)$, we need only check that the partial averaging property is satisfied. For every Borel measurable function $h$ mapping $\mathbb{R}$ to $\mathbb{R}$ and satisfying $\mathbb{E}|h(X)| < \infty$, we have

$$\mathbb{E}h(X) = \int_{-\infty}^{\infty} h(x) f_X(x)dx. \tag{2.6.2}$$

This is Theorem 1.5.2 in Chapter 1. Similarly, if $h$ is a function of both $x$ and $y$, then

$$\mathbb{E}h(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) f_{X,Y}(x,y)dxdy, \tag{2.6.3}$$

whenever $(X,Y)$ has a joint density $f_{X,Y}(x,y)$. You may use both (2.6.2) and (2.6.3) in your solution to this problem.

Let $A$ be a set in $\sigma(X)$. By the definition of $\sigma(X)$, there is a Borel subset $B$ of $\mathbb{R}$ such that $A = \{\omega \in \Omega; X(\omega) \in B\}$, or more simply, $A = \{X \in B\}$. Show the partial averaging property

$$\int_A g(X) d\mathbb{P} = \int_A Y d\mathbb{P}.$$

**Exercise 2.6.11.** (i) Let $X$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $W$ be a nonnegative $\sigma(X)$-measurable random variable. Show there exists a function $g$ such that $W = g(X)$. (Hint: Recall that every set in $\sigma(X)$ is of the form $\{X \in B\}$ for some Borel set $B \subset \mathbb{R}$. Suppose first that $W$ is the indicator of such a set, and then use the standard machine.)

(ii) Let $X$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $Y$ be a nonnegative random variable on this space. We do not assume that $X$ and $Y$ have a joint density. Nonetheless, show there is a function $g$ such that $\mathbb{E}[Y|X] = g(X)$.

http://www.springer.com/978-0-387-40100-3

Stochastic Calculus for Finance I
The Binomial Asset Pricing Model
Shreve, S.
2004, XV, 187 p., Hardcover
ISBN: 978-0-387-40100-3