Chapter 2
Distribution and conditional expectation

Following the same spirit as in the previous chapter, we continue to explore the implications of applying the results from measure theory to the field of probability theory. This chapter introduces in particular concepts that will be key later to studying various stochastic processes.

In the first section, we state the Radon–Nikodým theorem [78] without proof, and again refer the reader to Rudin [88] for the details. This result basically says that, under certain assumptions and having two positive measures, there is a unique measurable function ϕ such that the measure of a set for the first measure can be expressed using the function ϕ and the second measure. The function ϕ is called to the Radon–Nikodým derivative of the first with respect to the second measure. A nice consequence of the theorem is that positive measures can be expressed using a measurable function and a well-known reference measure such as the Lebesgue measure. This result is important in probability theory because it is the main ingredient to define more rigorously key probability concepts seen at the undergraduate level. The theorem is used in particular in this chapter to define the distribution of a random variable and the concept of conditional expectation.

Distribution — In the second section, we define the distribution of a random variable and prove the change of variable formula which shows how quantities related to the random variable can be computed using the distribution rather than the more mysterious underlying probability measure. The random variable is said to be continuous or discrete depending on whether its distribution satisfies a property called absolute continuity with respect to the Lebesgue or the counting measure. In particular, the Radon–Nikodým theorem implies that the density function of a continuous random variable and the probability mass function of a discrete random variable can both be seen as the Radon–Nikodým derivative of their distribution with respect to a reference measure. This not only unifies discrete and continuous random variables under the same umbrella, but also, together with the change of variable formula, this gives a powerful computational tool to study random variables.

N. Lanchier, *Stochastic Modeling*, Universitext,
DOI 10.1007/978-3-319-50038-6_2

Conditional expectation — In the third section, we introduce the important concept of conditional expectation which will be important later for defining and studying martingales and Markov chains. Roughly speaking, the conditional expectation of a random variable with respect to a σ-algebra is the best guess we can make about the random variable given the information contained in the σ-algebra. The existence and uniqueness of the conditional expectation follows from the Radon–Nikodým theorem. We show how important formulas seen at the undergraduate level can be derived from this more abstract concept of conditional expectation, and also how to compute the conditional expectation in practice by breaking down the random variable under consideration into pieces that are either measurable with respect to or independent of the conditioning σ-algebra.

2.1 Radon–Nikodým theorem

This theorem is a general result in measure theory that has interesting applications in probability theory discussed in the next sections. To motivate the theorem, note that, given a positive measurable function ϕ on $(\Omega, \mathscr{F}, \mu)$,

$$\nu(A) = \int_A \phi \, d\mu = \int \phi \, \mathbf{1}_A \, d\mu \quad \text{for all} \quad A \in \mathscr{F}$$

defines a new measure ν on $(\Omega, \mathscr{F})$. Indeed, for each sequence (A_n) of mutually exclusive measurable sets, we have

$$\nu\left(\bigcup_{i=1}^{\infty} A_j\right) = \int \lim_{n\to\infty} \sum_{i=1}^{n} (\phi \, \mathbf{1}_{A_i}) \, d\mu = \lim_{n\to\infty} \sum_{i=1}^{n} \int \phi \, \mathbf{1}_{A_i} \, d\mu = \sum_{i=1}^{\infty} \nu(A_i)$$

according to the monotone convergence theorem. The Radon–Nikodým theorem is in some sense the converse of the previous statement as it gives the existence and uniqueness of ϕ under certain conditions on the measure ν.

Definition 2.1 (Absolute continuity). The measure ν is said to be absolutely continuous with respect to μ, which we write $\nu \ll \mu$, whenever

$$\text{for all } A \in \mathscr{F}, \;\; \mu(A) = 0 \;\text{ implies that }\; \nu(A) = 0. \tag{2.1}$$

This definition is motivated by the fact that

$$\mu(A) = 0 \quad \text{implies that} \quad \nu(A) = \int_A \phi \, d\mu = 0.$$

In particular, given two positive measures μ and ν, the absolute continuity of ν with respect to μ is a necessary condition for the existence of the function ϕ above. The Radon–Nikodým theorem states that this condition is also sufficient.

Theorem 2.1 (Radon–Nikodým). *Let μ and ν be two σ-finite measures such that ν is absolutely continuous with respect to μ. Then,*

- *There is $\phi : (\Omega, \mathscr{F}) \to (\mathbb{R}, \mathscr{B})$ measurable positive such that*

$$\nu(A) = \int_A \phi \, d\mu \quad \textit{for all} \quad A \in \mathscr{F}.$$

  The function ϕ is written $\phi = d\nu/d\mu$ and is called the **Radon–Nikodým derivative** *of the measure ν with respect to the measure μ.*
- *It is unique in the sense that two such derivatives are equal μ-almost everywhere: the set where both functions differ has measure zero for μ.*

This version of the theorem was proved by Nikodým in [78] and extends an earlier result due to Radon who proved the theorem in the special case where the underlying space is $\mathbb{R}^n$ rather than a general measure space. To understand the assumption of the theorem, assume for a moment that the measures ν and μ are simply nonnegative functions defined on the real line. Then, there exists a function ϕ that satisfies $\nu = \phi\,\mu$ if and only if

$$\mu(x) = 0 \quad \text{implies} \quad \nu(x) = 0 \quad \text{for all} \quad x \in \mathbb{R}.$$

This last condition can be viewed as the analog of the absolute continuity for positive measures (2.1). For a proof of the theorem, we refer to [88, Chapter 6]. In the next two sections, we use this theorem to redefine more rigorously various key concepts of probability theory.

2.2 Induced measure and distribution

Having a real random variable X on a probability space, one can define a probability measure ν_X on the Borel σ-algebra by setting

$$\nu_X(B) = P(X \in B) = \int_\Omega \mathbf{1}_{X^{-1}(B)} \, dP \quad \text{for all} \quad B \in \mathscr{B}.$$

It is called the **measure induced** by X in measure theory and **distribution** of X in probability theory. To study a random variable in practice, probabilists do not work with the probability measure P but with its distribution by using the following result called **change of variables formula**.

Theorem 2.2. *Let $X : (\Omega, \mathscr{F}, P) \to (\mathbb{R}, \mathscr{B})$ and $h : \mathbb{R} \to \mathbb{R}$ be measurable. Then, whenever h is positive or integrable,*

$$E(h(X)) = \int_{\mathbb{R}} h \, d\nu_X = \int_{\mathbb{R}} h(x) \, \nu_X(dx). \tag{2.2}$$

Proof. The steps of the proof follow the construction of the integral.

Step 1 — Assume first that $h = \mathbf{1}_B$ for some Borel set B. Then,

$$E(\mathbf{1}_B(X)) = E(\mathbf{1}\{X \in B\}) = P(X \in B) = \nu_X(B) = \int_{\mathbb{R}} \mathbf{1}_B \, d\nu_X.$$

Step 2 — In case $h = a_1 \mathbf{1}_{B_1} + \cdots + a_n \mathbf{1}_{B_n}$ is a simple measurable function, we use the previous step and the linearity of the integral to obtain

$$E(h(X)) = \sum_{i=1}^{n} a_i E(\mathbf{1}_{B_i}(X)) = \sum_{i=1}^{n} a_j \int_{\mathbb{R}} \mathbf{1}_{B_i} \, d\nu_X \; = \int_{\mathbb{R}} h \, d\nu_X.$$

Step 3 — Assume that h is positive. Recall from (1.6) that

$$s_n(x) = \min\left(n, 2^{-n} \lfloor 2^n h(x) \rfloor\right) \quad \text{for all} \quad x \in \mathbb{R}$$

defines a nondecreasing sequence of simple measurable functions with pointwise limit h. In particular, by the monotone convergence theorem,

$$E(h(X)) = \lim_{n\to\infty} E(s_n(X)) = \lim_{n\to\infty} \int_{\mathbb{R}} s_n \, d\nu_X = \int_{\mathbb{R}} h \, d\nu_X.$$

Step 4 — Finally, when h is integrable, we write $h = h^+ - h^-$. Since h^+ and h^- are both positive (so the previous step applies) and integrable,

$$E(h(X)) = E(h^+(X)) - E(h^-(X)) \; = \int_{\mathbb{R}} h^+ \, d\nu_X \; - \int_{\mathbb{R}} h^- \, d\nu_X \; = \int_{\mathbb{R}} h \, d\nu_X.$$

This completes the proof. □

Note that the probability that $X \in A$ and the expected value of X can be computed by taking $h = \mathbf{1}_A$ and $h = \mathrm{id}$ respectively, i.e.,

$$P(X \in A) = \int_{\mathbb{R}} \mathbf{1}_A \, d\nu_X = \int_A d\nu_X \quad \text{and} \quad E(X) = \int_{\mathbb{R}} x \, d\nu_X.$$

In practice, it is convenient to express the distribution ν_X as a measurable function times a standard measure such as the Lebesgue measure. This idea is again related to the Radon–Nikodým theorem, which we now use to show the connection between the theory we have presented so far and undergraduate probability classes.

Probability density function. A random variable X is said to be **continuous** whenever its distribution ν_X is absolutely continuous with respect to the Lebesgue measure. Since the Lebesgue measure is σ-finite, the Radon–Nikodým theorem applies and gives the existence of a measurable function ϕ_X such that $d\nu_X = \phi_X \, d\lambda$. The derivative ϕ_X is then called the probability density function of the random variable and the change of variable formula (2.2) becomes

$$E(h(X)) = \int h \, \phi_X \, d\lambda = \int_{\mathbb{R}} h(x) \, \phi_X(x) \, dx. \tag{2.3}$$

Probability mass function. A random variable X is said to be **discrete** whenever its range S is either finite or countable, in which case its distribution ν_X is absolutely continuous with respect to the counting measure on the set S, that is,

$$\nu_X \ll \mu_S \quad \text{where} \quad \mu_S(A) = \operatorname{card}(A) \ \text{ for all } \ A \subset S.$$

Using once more the Radon–Nikodým theorem gives the existence of a measurable function ϕ_X such that $d\nu_X = \phi_X \, d\mu_S$. In this case, the derivative ϕ_X is called the probability mass function and the change of variable formula (2.2) becomes

$$E(h(X)) = \int h \, \phi_X \, d\mu_S = \sum_{x \in S} h(x) \, \phi_X(x). \tag{2.4}$$

In conclusion, random variables are characterized by their distributions which, in turn, are characterized by their Radon–Nikodým derivative ϕ_X with respect to some standard measures, either the Lebesgue measure or the counting measure. To this extent, measure theory and the abstract integral unify both discrete and continuous random variables by interpreting both probability density and probability mass functions as Radon–Nikodým derivatives. In practice, we deal with the integral with respect to the Lebesgue measure (2.3) or with respect to the counting measure (2.4) instead of the somewhat mysterious probability measure P. For a list of the most common distributions along with their interpretation and probability mass/density functions, we refer to Figure 2.1.

2.3 Conditional expectation

To study stochastic processes later, the next step is to define conditional expectation since this is a key concept to express certain dependency relationships among random variables and define martingales and Markov chains. This concept is introduced in the following definition. The fact that the conditional expectation exists and is unique follows from the Radon–Nikodým theorem.

Definition 2.2. Let $X \in L^1(\Omega, \mathscr{F}, P)$ and let $\mathscr{G} \subset \mathscr{F}$ be a σ-algebra.

- The **conditional expectation of X given $\mathscr{G}$** is any

$$Z \in \mathscr{M}(\Omega, \mathscr{G}) \quad \text{such that} \quad E(X \mathbf{1}_A) = E(Z \mathbf{1}_A) \quad \text{for all} \quad A \in \mathscr{G}.$$

  The variable Z is called a **version** of $E(X \,|\, \mathscr{G})$.
- Having a second random variable Y, we let $E(X \,|\, Y) = E(X \,|\, \sigma(Y))$.
- Also, we define the **conditional probability** as

$$P(X \in B \,|\, \mathscr{G}) = E(\mathbf{1}\{X \in B\} \,|\, \mathscr{G}).$$

Name/parameters	Interpretation/origin	Probability mass/density function	Mean	Variance
Uniform (S), S finite	Equally likely outcomes.	$\phi(x) = \dfrac{1}{\text{card}(S)}$ for all $x \in S$		
Uniform (S), $0 < \lambda(S) < \infty$	Equally likely outcomes.	$\phi(x) = \dfrac{1}{\lambda(S)}$ for all $x \in S$		
Bernoulli (p)	Coin flip - success/failure.	$\phi(1) = 1 - \phi(0) = p$	p	$p(1-p)$
Binomial (n,p)	Number of successes in a sequence of n independent Bernoulli trials. Discrete-time analog of Poisson.	$\phi(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for all $x \in \{0,1,\ldots,n\}$	np	$np(1-p)$
Poisson (μ)	Number of Poisson events in a time interval of length one. Continuous-time analog of binomial.	$\phi(x) = \dfrac{\mu^x}{x!} e^{-\mu}$ for all $x \in \mathbb{N}$	μ	μ
Geometric (p)	Time to the first/next success in a sequence of independent Bernoulli trials. Discrete-time analog of exponential.	$\phi(x) = (1-p)^{x-1} p$ for all $x \in \mathbb{N}^*$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
Exponential (μ)	Time to the first/next Poisson event. Continuous-time analog of the geometric random variable.	$\phi(x) = \mu e^{-\mu x}$ for all $x \in \mathbb{R}_+$	$\dfrac{1}{\mu}$	$\dfrac{1}{\mu^2}$
Negative Binomial (n,p)	Time to the nth success in a sequence of independent Bernoulli trials. Discrete-time analog of gamma.	$\phi(x) = \binom{x-1}{n-1} p^n (1-p)^{x-n}$ for all $x \in \{n, n+1, \ldots\}$	$\dfrac{n}{p}$	$\dfrac{n(1-p)}{p^2}$
Gamma (n,μ)	Time to the nth Poisson event. Continuous-time analog of the negative binomial.	$\phi(x) = \dfrac{\mu e^{-\mu x} (\mu x)^{n-1}}{n!}$ for all $x \in \mathbb{R}_+$	$\dfrac{n}{\mu}$	$\dfrac{n}{\mu^2}$
Normal (μ, σ^2)	Universal limit in the central limit theorem.	$\phi(x) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$ for all $x \in \mathbb{R}$	μ	σ^2

Fig. 2.1 Most common distributions.

Theorem 2.3. *The conditional expectation exists and is unique in the sense that two versions of the conditional expectation are equal P-almost surely.*

Proof. Assuming first that X is positive, since

$$\nu(A) = E(X\,\mathbf{1}_A) = \int_A X\,dP \quad \text{for all} \quad A \in \mathscr{G}$$

defines a finite measure $\nu \ll P$ on the space $(\Omega, \mathscr{G})$, there exists Z that satisfies the statement of the theorem, namely the Radon–Nikodým derivative $d\nu/dP$, and two such random variables are equal P-almost surely. In the general case where the random variable X is integrable, the first part of the proof applies to its positive part and its negative part. In particular, there exist two random variables Z_+ and Z_- measurable with respect to $\mathscr{G}$ such that

$$\begin{aligned} E(X\,\mathbf{1}_A) &= E(X^+\,\mathbf{1}_A) - E(X^-\,\mathbf{1}_A) \\ &= E(Z_+\,\mathbf{1}_A) - E(Z_-\,\mathbf{1}_A) = E((Z_+ - Z_-)\,\mathbf{1}_A) \end{aligned}$$

for all $A \in \mathscr{G}$. The uniqueness P-almost surely again follows from the uniqueness of the Radon–Nikodým derivative. □

The conditional expectation has several interesting properties. For instance, it can be proved that the conditional expectation inherits the following properties from the unconditional expectation:

- The function $X \mapsto E(X\,|\,\mathscr{G})$ is linear and nondecreasing.
- Jensen's inequality:

$$\phi(E(X\,|\,\mathscr{G})) \le E(\phi(X)\,|\,\mathscr{G}) \quad \text{for all convex functions } \phi.$$

- Monotone convergence:

$$\lim_{n\to\infty} E(X_n\,|\,\mathscr{G}) = E(X\,|\,\mathscr{G}) \quad \text{whenever} \quad X_n \uparrow X.$$

- Dominated convergence:

$$\lim_{n\to\infty} E(X_n\,|\,\mathscr{G}) = E(X\,|\,\mathscr{G}) \quad \text{whenever} \quad X_n \to X \ \text{ and } \ |X_n| \le Y \in L^1.$$

The conditional expectation will be used later in two contexts.

1. From the concept of conditional expectation, we can recover important formulas seen in undergraduate probability classes. These formulas show how to compute the probability of an event or the expected value of a random variable by conditioning on a partition of the sample space or on another random variable.
2. We also show how to compute the conditional expectation of a random variable in practice by breaking down this variable into pieces that are measurable with respect to the σ-algebra and pieces that are independent of the σ-algebra. These results will be our main tools to study martingales.

We now explore these two aspects.

Computing by conditioning. The most trivial but also one of the most useful properties, obtained by taking $A = \Omega$ in Definition 2.2, states that

$$E(E(X\,|\,Y)) = E(X) \quad \text{for all} \quad X \in L^1(\Omega, \mathscr{F}, P). \tag{2.5}$$

This equation is most useful looking at it backward, namely, it is used in practice to compute the expected value of a random variable, i.e., the right-hand side of (2.5), by conditioning on another random variable, i.e., the left-hand side. To deduce how to compute probability and expected value by conditioning, recall that the conditional probability of an event and the conditional expectation of a discrete random variable given an event are defined as

$$P(A\,|\,B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad E(X\,|\,B) = \frac{E(X\,\mathbf{1}_B)}{P(B)}. \tag{2.6}$$

Now, assume that we have a σ-algebra $\mathscr{G}$ generated by a partition (B_n) of events that all have a strictly positive probability. The $\mathscr{G}$-measurability of the conditional expectation implies that

$$Z = E(X\,|\,\mathscr{G}) = a_1\,\mathbf{1}_{B_1} + \cdots + a_n\,\mathbf{1}_{B_n} + \cdots \quad \text{for some} \quad a_n \in \mathbb{R},$$

but since $B_n \in \mathscr{G}$, by definition of the conditional expectation,

$$E(X\,\mathbf{1}_{B_n}) = E(Z\,\mathbf{1}_{B_n}) = a_n\,P(B_n).$$

In particular, we obtain the following expression:

$$E(X\,|\,\mathscr{G}) = \sum_{n=1}^{\infty} \frac{E(X\,\mathbf{1}_{B_n})}{P(B_n)}\,\mathbf{1}_{B_n}. \tag{2.7}$$

Recall from our convention that, even if it is not specified, this equation only holds almost surely. Setting $X = \mathbf{1}_A$ in (2.7), taking the expected value, using (2.5) and recalling the definition of conditional probability from (2.6), we deduce

$$P(A) = E(E(\mathbf{1}_A\,|\,\mathscr{G})) = \sum_{n=1}^{\infty} \frac{E(\mathbf{1}_A\,\mathbf{1}_{B_n})}{P(B_n)}\,E(\mathbf{1}_{B_n}) = \sum_{n=1}^{\infty} P(A\,|\,B_n)\,P(B_n). \tag{2.8}$$

Now, let Y be a discrete random variable. Since the σ-algebra $\sigma(Y)$ is generated by a countable partition, we deduce from (2.7) that

$$E(X\,|\,Y) = \sum_{y} \frac{E(X\,\mathbf{1}\{Y = y\})}{P(Y = y)}\,\mathbf{1}\{Y = y\}$$

where the sum is over the range of Y. Taking the expected value, using (2.5) and recalling the definition of conditional expectation from (2.6), we get

$$E(X) = E(E(X\,|\,Y)) = \sum_y E(X\,|\,Y=y)\,P(Y=y). \tag{2.9}$$

Equations (2.8)–(2.9) are two important formulas seen at the undergraduate level. They are useful in practice to compute unconditional probability and expected value provided one is able to find natural partitions or random variables that make the conditional objects easier to compute than their unconditional counterparts. The exercises at the end of this chapter give a wide variety of examples of application of these equations. Note that these two formulas can be proved more easily by simply using the σ-additivity of the probability. The approach we used is to show how they can also be derived from the general definition of conditional expectation.

Computing conditional expectations. Finally, we give some tools that will be useful later to prove that a stochastic process is a martingale and/or a Markov chain. To better understand the real meaning of the conditional expectation hidden behind its somewhat mysterious definition, one can think of the conditional expectation as the best possible approximation of X given the information contained in the σ-algebra $\mathscr{G}$, namely, the best possible approximation by a random variable which is $\mathscr{G}$-measurable. The larger the σ-algebra, the better the approximation. With this in mind, it is clear intuitively and easy to prove that

$$\begin{array}{rlll} \textbf{perfect information} & E(X\,|\,\mathscr{G}) = X & \text{when} & \mathscr{G} = \mathscr{F} \\ \textbf{no information} & E(X\,|\,\mathscr{G}) = E(X) & \text{when} & \mathscr{G} = \{\varnothing, \Omega\}. \end{array} \tag{2.10}$$

Following along these lines, we also prove that, when $\mathscr{H} \subset \mathscr{G}$,

$$E(E(X\,|\,\mathscr{G})\,|\,\mathscr{H}) = E(E(X\,|\,\mathscr{H})\,|\,\mathscr{G}) = E(X\,|\,\mathscr{H})$$

indicating that, after the double conditioning, the only information available comes from the smaller σ-algebra. It is convenient to call this property the **projection rule** since one can think of the conditional expectation as the projection of a random variable onto a subset of measurable functions, so the property above simply says that projecting twice has the same result as projecting onto the smaller subset.

To compute the conditional expectation of a random variable in practice, the trick is to break it down into pieces for which the information is perfect and pieces for which there is no information and then use (2.10). To make this precise, we prove a couple of lemmas showing basically that the pieces that are perfectly known under the conditioning can be moved outside the conditional expectation.

Lemma 2.1. *Let $X, Y \in L^1(\Omega, \mathscr{F}, P)$. Then,*

$$E(X+Y\,|\,\mathscr{G}) = X + E(Y\,|\,\mathscr{G}) \quad \textit{when } X \textit{ is } \mathscr{G}\textit{-measurable}.$$

Proof. This follows from (2.10) and the linearity of the conditional expectation. □

Lemma 2.2. *Let $Y, XY \in L^1(\Omega, \mathscr{F}, P)$. Then,*

$$E(XY\,|\,\mathscr{G}) = X\,E(Y\,|\,\mathscr{G}) \quad \textit{when } X \textit{ is } \mathscr{G}\textit{-measurable}.$$

Proof. Let $A, B \in \mathscr{G}$ and set $X = \mathbf{1}_A$. Since $A \cap B \in \mathscr{G}$,

$$E(\mathbf{1}_B E(Y \,|\, \mathscr{G})\, \mathbf{1}_A) = E(E(Y \,|\, \mathscr{G})\, \mathbf{1}_{A\cap B}) = E(Y\, \mathbf{1}_{A\cap B}) = E(\mathbf{1}_B Y\, \mathbf{1}_A)$$

which shows the result when X is an indicator function. We conclude following the same steps and using the same ingredients as in the proof of Theorem 2.2. More precisely, we extend the result to simple random variables using that the conditional expectation is linear, then to positive random variables using the monotone convergence theorem, and finally to integrable random variables by looking at negative and positive parts separately. □

To study branching processes later, we will also need the following result.

Lemma 2.3. *Let $(X_n) \subset L^1(\Omega, \mathscr{F}, P)$ be a sequence of identically distributed random variables and let T be an integer-valued random variable. Then,*

$$E(X_1 + X_2 + \cdots + X_T \,|\, T) = T\, E(X_1 \,|\, T).$$

Proof. Letting $A_n = \{T = n\}$ for all $n \in \mathbb{N}$, we have

$$\begin{aligned} E((X_1 + \cdots + X_T)\, \mathbf{1}_{A_n}) &= E((X_1 + \cdots + X_n)\, \mathbf{1}_{A_n}) \\ &= E(nX_1\, \mathbf{1}_{A_n}) = E(n\, E(X_1 \,|\, T)\, \mathbf{1}_{A_n}) = E(T\, E(X_1 \,|\, T)\, \mathbf{1}_{A_n}). \end{aligned}$$

Since $\sigma(T)$ is generated by the partition (A_n), the previous set of equations remains true replacing A_n by any $A \in \sigma(T)$, which proves the result. □

The previous three lemmas show how to deal in practice with the pieces that are measurable with respect to the conditioning. We now look at the other extreme: the pieces of the random variable for which we have no information. In the second part of (2.10), we have no information about the random variable because the σ-algebra does not give any information. More generally, we have no information about the random variable whenever it is independent of the σ-algebra, so we define independence and show that the second statement in (2.10) extends to this case. The reader should be already familiar with the independence of events. In contrast, the next definition is about independence of σ-algebras and random variables.

Definition 2.3. Two σ-algebras $\mathscr{G}, \mathscr{H} \subset \mathscr{F}$ are said to be **independent** if each event in $\mathscr{G}$ and each event in $\mathscr{H}$ are pairwise independent, i.e.,

$$P(A \cap B) = P(A)\, P(B) \quad \text{for all} \quad A \in \mathscr{F} \text{ and } B \in \mathscr{G}.$$

Two random variables X and Y are independent when

$$P(X \in A, Y \in B) = P(X \in A)\, P(Y \in B) \quad \text{for all} \quad A, B \in \mathscr{B}(\mathbb{R})$$

meaning that $\sigma(X)$ and $\sigma(Y)$ are independent.

The definition will become clear later when dealing with concrete examples of random variables independent from a σ-algebra in the context of stochastic processes.

Before we state our next result about conditional expectation, recall that, having a larger collection, finite or countable, of events (A_n), these events are said to be independent whenever

$$P\left(\bigcap_{i\in I} A_i\right) = \prod_{i\in I} P(A_i) \quad \text{for all} \quad I \subset \mathbb{N}^* \text{ finite.}$$

Recall also that a collection of events may be pairwise independent but not independent, showing that independence is a somewhat subtle concept. To give a simple counterexample, flip a fair coin twice and let X_1 and X_2 be respectively the outcome of the first and second flip. Both flips are assumed to result in independent outcomes. Then, define the events

$$A_1 = \{X_1 = H\} \qquad A_2 = \{X_2 = H\} \qquad A_3 = \{X_1 \neq X_2\}.$$

These events are pairwise independent. Indeed, they all have probability one-half, and the intersection of any two has probability one-fourth since

$$A_1 \cap A_2 = \{HH\} \qquad A_1 \cap A_3 = \{HT\} \qquad A_2 \cap A_3 = \{TH\}.$$

However, they are not independent because

$$P(A_1 \cap A_2 \cap A_3) = P(\varnothing) = 0 \neq (1/2)^3 = P(A_1)\,P(A_2)\,P(A_3).$$

Returning to conditional expectation, we now prove that, when the random variable and the conditioning σ-algebra are independent, suggesting that the σ-algebra does not provide any information about the random variable, our best guess is again the unconditional expectation, i.e., we can remove the conditioning. This is proved in the following lemma.

Lemma 2.4. *Let X and $\mathscr{G}$ be independent. Then, $E(X\,|\,\mathscr{G}) = E(X)$.*

Proof. The joint distribution of two independent random variables is equal to the product of their distributions. Therefore, if X and Y are independent, by applying Fubini's theorem to the function $h(x,y) = xy$, we obtain

$$E(XY) = \int h\,d\nu_{X,Y} = \int x\,\nu_X(dx) \int y\,\nu_Y(dy) = E(X)\,E(Y).$$

In particular, since X and $Y = \mathbf{1}_A$ are independent for all $A \in \mathscr{G}$,

$$E(X\,\mathbf{1}_A) = E(XY) = E(X)\,E(Y) = E(X)\,E(\mathbf{1}_A) = E(E(X)\,\mathbf{1}_A).$$

Moreover, since $Z = E(X)$ is constant, it is $\mathscr{G}$-measurable. □

Lemmas 2.1–2.4 are the main tools to compute the conditional expectation. As previously explained, in order to be able to apply these lemmas in practice, the first step is to break down the random variable into pieces that are measurable with respect to the σ-algebra and pieces that are independent of the σ-algebra.

2.4 Exercises

Probability and conditional probability

Exercise 2.1. Let A and B be two events.

1. Prove that the occurrence of A makes the occurrence of B more likely if and only if the occurrence of B makes the occurrence of A more likely.
2. Prove that A and B are independent if and only if A and B^c are independent.

Exercise 2.2. Suppose that two balls numbered 1 and 2 are independently black or white with the same probability of one-half.

1. Given that ball 2 is black, find the probability that the other ball is black.
2. Given that at least one of the two balls is black, find the probability that the other ball is also black.

Exercise 2.3. An urn contains three white and three black balls. A fair die is rolled and that number of balls is randomly chosen from the urn. Find the probability that all the selected balls are white.

Exercise 2.4. An urn contains a white and b black balls. Take one ball at random, paint it black in case it is white, and put the ball back in the urn. Then, take again a ball at random. By what factor does the probability that the second ball is white decreases in comparison with the probability that the first ball is white?

Exercise 2.5. An urn contains nine white balls and six black balls. Take three balls at random, paint the white ones black, and put the balls back in the urn. Then, take again three balls at random. Find the probability that these last three balls are all white.

Exercise 2.6. Assume that a parallel system with n components is operational if and only if at least one of its component is working. Compute the conditional probability that component 1 is working given that the system is operational under the condition that the components work independently with probability p.

Exercise 2.7. Consider $n \geq 2$ individuals labeled $1, 2, \ldots, n$. Individual 1 creates a rumor that she tells individual 2. Then, each individual independently tells the next individual either the information she learned with probability p or the opposite of the information she learned with probability $q = 1 - p$.

1. Condition on whether or not the information told by individual 2 is the one she learned to express p_n as a function of p_{n-1} where p_n be the probability that the information received by individual n is the rumor created by individual 1.
2. Deduce an explicit expression for p_n.
3. Compute p_n when $p = 0$, and in the limit as $n \to \infty$ when $p \in (0, 1)$.

Exercise 2.8 (Craps). At this game, the player throws two fair dice.

- If the sum is 7 or 11 then she wins.
- If the sum is 2, 3 or 12 then she loses.
- If the sum is anything else, she continues throwing the dice until the sum is either that number again, in which case she wins, or 7, in which case she loses.

Find the probability of winning at the game of craps.

Hint: Compute first the probability that the sum i appears before the sum 7 by conditioning on the value of the first roll.

Exercise 2.9. Players A and B play until they are two points apart. Find the probability that A is the first player to have two more points than the other player if each point is independently won by A with probability p.

Exercise 2.10. Let A and B be two tennis players. Assume that each point is scored independently by player A with probability $p = 3/5$. Compute the probability that A wins a game, where the winner of a game is the first player with four points and two more points than the other player.

Exercise 2.11. Two players A and B take turn playing a game until one of the two players wins. Independently at each step, player A wins with probability p while player B wins with probability q. Find the values of p and q for which the game is fair, i.e., each player is equally likely to be the first one to win.

Hint: Condition on whether A wins the first game or not.

Exercise 2.12. Three evenly matched players A, B and C play a series of games. The winner of each game plays the next game with the waiting player until a player wins two games in a row and is declared the overall winner. Find the probability of each of the players being the overall winner when A and B play the first game.

Exercise 2.13. A player plays alternatively with two opponents until she wins twice in a row. Assuming that she wins independently each game with probability p against opponent 1 and with probability $q > p$ against opponent 2, should she start playing with opponent 1 or with opponent 2 if her objective is to minimize the expected number of games she has to play?

Exercise 2.14. Consider a contest with 2^n evenly matched players. The players are randomly paired off against each other, then the 2^{n-1} winners are again paired off, and so on, until a single winner remains. Find the probability that two randomly chosen contestants play each other.

Exercise 2.15 (The ballot problem). In an election, candidates A and B receive respectively a and b votes with $a > b$. Let $p(a,b)$ be the probability that candidate A is always ahead of B when all the orderings of the votes are equally likely.

1. Express $p(a,b)$ as a function of $p(a-1,b)$ and $p(a,b-1)$ by conditioning on the candidate who receives the last vote.
2. Deduce that $p(a,b) = (a-b)/(a+b)$.

Exercise 2.16 (The best prize problem). Consider the game where n distinct prizes are presented one by one to a player. All $n!$ possible permutations are assumed to be equally likely. Each time a prize is revealed, the player learns about the relative rank of that prize compared to the ones already seen and, based on this information, must decide to either leave with that prize or wait for a hopefully better prize. A natural strategy is to reject the first m prizes and then accept the first one that is better than the first m prizes provided there is one.

1. Letting B be the event that the best prize is selected and P_m be the probability under the strategy described above, prove that
$$P_m(B) = \frac{m}{n}\left(\frac{1}{m} + \frac{1}{m+1} + \cdots + \frac{1}{n-1}\right).$$
2. Deduce that, as $n \to \infty$, the maximal probability of selecting the best prize under the strategy described above converges to $1/e \approx 0.368$.

Exercise 2.17. Let $m \leq n$ and let
$$U_1, \ldots, U_m \sim \text{Uniform}\{1,2,\ldots,n\}$$
be independent. Condition on the event that the random variables are all distinct to compute the probability of the event
$$A = \{U_1 < U_2 < \cdots < U_m\}.$$

Exercise 2.18. Compute $P(A \subset B)$ where A and B are chosen uniformly at random among the 2^n possible subsets of a set with n elements.

Hint: Condition on the cardinal of B.

Exercise 2.19 (Euler ϕ function). For all $n \in \mathbb{N}^*$, let $\phi(n)$ be the number of integers less than or equal to n which are prime to n. We want to show that
$$\phi(n) = n \prod_{p \in P_n} \left(1 - \frac{1}{p}\right) \quad \text{where} \quad P_n = \{p : p \text{ is prime and divides } n\}.$$
Let $X \sim \text{Uniform}\{1,2,\ldots,n\}$ and $A_p = \{p \text{ divides } X\}$ for each $p \leq n$.

1. Compute $P(A_p)$ when p divides n.
2. Let $p_1, \ldots, p_k$ be distinct prime divisors of n. Prove that
$$A_{p_1}, A_{p_2}, \ldots, A_{p_k} \quad \text{are independent.}$$
3. Use the previous two questions to conclude.

Expected value and conditional expectation

Exercise 2.20. Players $1,2,\ldots,n$ take turns flipping a coin having probability p of turning up heads, with the successive flips being independent.

1. Thinking of the geometric random variable with success probability p as the first flip resulting in heads, compute its expected value by conditioning on the outcome of the first flip.
2. Use the same idea to find the probability mass function of the random variable X referring to the first player who gets heads.

Exercise 2.21. Let (X_n) be a sequence of independent and identically distributed discrete random variables with finite range, say $\{1,2,\ldots,m\}$. Find the expected value of the total number T of random variables one needs to observe until the first outcome appears again.

Hint: Condition on the outcome of the first random variable.

Exercise 2.22 (Matching rounds problem). Referring to Exercise 1.25,

1. Find $E(X)$ and $\mathrm{Var}(X)$ where

$$X = \text{number of husbands paired with their wife.}$$

Now, assume that the couples for which a match occurs depart while the others are again randomly paired. This continues until there is no couple left.

2. Prove that there are in average n rounds in this process.

Exercise 2.23. Let (X_i) be a sequence of independent and identically distributed discrete random variables and fix a pattern $(x_1,x_2,\ldots,x_n)$ such that,

$$(x_{n-k+1},x_{n-k+2},\ldots,x_n) \neq (x_1,x_2,\ldots,x_k) \quad \text{for all} \quad k<n.$$

Let T be the number of random variables until the pattern appears.

1. Prove that $T = i+n$ if and only if

$$T > i \quad \text{and} \quad (X_{i+1},X_{i+2},\ldots,X_{i+n}) = (x_1,x_2,\ldots,x_n).$$

2. Deduce that $E(T) = (p_{x_1}\, p_{x_2} \cdots p_{x_n})^{-1}$ where $p_x = P(X_i = x)$.

Exercise 2.24. Assume that an *a priori* biased coin whose probability of landing on heads is given by p is continually flipped.

1. Find the expected value of the time T_{HT} until the pattern HT appears.
2. Find the expected value of the time T_{HH} until the pattern HH appears.
3. More generally, find the expected value of the time T_n until H appears n times in a row by conditioning on T_{n-1}.

Exercise 2.25 (Compound random variable). Let (X_n) be a sequence of identically distributed random variables with

$$\mu = E(X_n) < \infty \quad \text{and} \quad \sigma^2 = \operatorname{Var}(X_n) < \infty$$

and let T be an independent nonnegative integer-valued random variable also with finite mean and finite variance.

1. Prove that $E(X_1 + X_2 + \cdots + X_T) = \mu\, E(T)$.
2. Assuming in addition that the X_n are independent, show that

$$\operatorname{Var}(X_1 + X_2 + \cdots + X_T) = \sigma^2\, E(T) + \mu^2 \operatorname{Var}(T).$$

Hint: For both parts, condition on the random variable T.

Exercise 2.26. Let $p \in (0,1)$ and $X, Y \sim \text{Bernoulli}(p)$ be independent.

1. Compute $E(X \mid Z)$ where $Z = \mathbf{1}\{X + Y = 0\}$.
2. Deduce that $E(X \mid Z)$ and $E(Y \mid Z)$ are not independent.

http://www.springer.com/978-3-319-50037-9

Stochastic Modeling

Lanchier, N.

2017, XIII, 303 p. 63 illus., 6 illus. in color., Softcover
ISBN: 978-3-319-50037-9