1

Measure Theory and Laws of Large Numbers

1.1 Introduction

If you're reading this, you've probably already seen many different types of random variables and have applied the usual theorems and laws of probability to them. We will, however, show you there are some seemingly innocent random variables for which none of the laws of probability apply. Measure theory, as it applies to probability, is a theory that carefully describes the types of random variables the laws of probability apply to. This puts the whole field of probability and statistics on a mathematically rigorous foundation.

You are probably familiar with some proof of the famous strong law of large numbers, which asserts that the long-run average of independent and identically distributed (iid) random variables converges to the expected value. One goal of this chapter is to show you a beautiful and more general alternative proof of this result using the powerful ergodic theorem. In order to do this, we will first take you on a brief tour of measure theory and introduce you to the dominated convergence theorem, one of measure theory's most famous results and the key ingredient we need.

In Section 1.2, we construct an event, called a nonmeasurable event, to which the laws of probability don't apply. In Section 1.3, we introduce the notions of countably and uncountably infinite sets and show you how the elements of some infinite sets cannot be listed in a sequence. In Section 1.4, we define a probability space and the laws of probability that apply to them. In Section 1.5, we introduce the concept of a measurable random variable, and in Section 1.6, we introduce the concepts of convergence and limits. In Section 1.7, we define the expected value in terms of the Lebesgue integral. In Section 1.8, we illustrate and prove the dominated convergence theorem,

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

and Section 1.9, we discuss convergence in probability and distribution. Lastly, in Section 1.10, we prove zero-one laws and the ergodic theorem and use these to obtain the strong law of large numbers.

1.2 A Nonmeasurable Event

Consider a circle that has a radius equal to one. We say that two points on the edge of the circle are in the same family if you can go from one point to the other point by taking steps of length one unit around the edge of the circle. By this we mean each step you take moves you an angle of exactly one radian degree around the circle, and you are allowed to keep looping around the circle in either direction.

Suppose each family elects one of its members to be the head of the family. Here is the question: What is the probability a point X selected uniformly at random along the edge of the circle is the head of its family? It turns out this question has no answer.

The first thing to notice is that each family has an infinite number of family members. Because the circumference of the circle is 2π, you can never get back to your starting point by looping around the circle with steps of length one. If it were possible to start at the top of the circle and get back to the top going a steps clockwise and looping around b times, then you would have $a = b2\pi$ for some integers a, b, and hence $\pi = a/(2b)$. This is impossible because it's well-known that π is an irrational number and can't be written as a ratio of integers.

It may seem to you like the probability should either be zero or one, but we will show you why neither answer could be correct. It doesn't even depend on how the family heads are elected. Define the events $A = \{X$ is the head of its family$\}$, $A_i = \{X$ is i steps clockwise from the head of its family$\}$, and $B_i = \{X$ is i steps counterclockwise from the head of its family$\}$.

Because X was uniformly chosen, we must have $P(A) = P(A_i) = P(B_i)$. But because every family has a head, the sum of these probabilities should equal one, or in other words,

$$1 = P(A) + \sum_{i=1}^{\infty} (P(A_i) + P(B_i)).$$

Thus, if $x = P(A)$ we get $1 = x + \sum_{i=1}^{\infty} 2x$, which has no solution where $0 \leq x \leq 1$. This means it's impossible to compute $P(A)$, and the answer is neither zero nor one, nor any other possible number. The event A is called a non-measurable event, because you can't measure its probability in a consistent way.

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

What's going on here? It turns out that allowing only one head per family, or any finite number of heads, is what makes this event nonmeasurable. If we allowed more than one head per family and gave everyone a 50% chance, independent of all else, of being a head of the family, then we would have no trouble measuring the probability of this event. Or if we let everyone in the top half of the circle be a family head, and again let families have more than one head, the answer would be easy. Later we will give a careful description of what types of events we can actually compute probabilities for.

Being allowed to choose exactly one family head from each family requires a special mathematical assumption called the axiom of choice. This axiom famously can create all sorts of other logical mayhem, such as allowing you to break a sphere into a finite number of pieces and rearrange them into two spheres of the same size (the Banach–Tarski paradox). For this reason, the axiom is controversial and has been the subject of much study by mathematicians.

1.3 Countable and Uncountable Sets

You may now be asking yourself if the existence of a uniform random variable $X \sim U(0,1)$ also contradicts the laws of probability. We know that for all x, $P(X = x) = 0$, but also $P(0 \le X \le 1) = 1$. Doesn't this give a contradiction because

$$P(0 \le X \le 1) = \sum_{x \in [0,1]} P(X = x) = 0?$$

Actually, this is not a contradiction because a summation over an interval of real numbers does not make any sense. Which values of x would you use for the first few terms in the sum? The first term in the sum could use $x = 0$, but it's difficult to decide which value of x to use next.

In fact, infinite sums are defined in terms of a sequence of finite sums:

$$\sum_{i=1}^{\infty} x_i \equiv \lim_{n \to \infty} \sum_{i=1}^{n} x_i,$$

so to have an infinite sum, it must be possible to arrange the terms in a sequence. If an infinite set of items can be arranged in a sequence it is called *countable*; otherwise it is called *uncountable*.

Obviously the integers are countable using the sequence 0, −1, +1, −2, +2, ... The positive rational numbers are also countable if you express them as a ratio of integers and list them in order by the sum of these integers:

$$\frac{1}{1}, \frac{2}{1}, \frac{1}{2}, \frac{3}{1}, \frac{2}{2}, \frac{1}{3}, \frac{4}{1}, \frac{3}{2}, \frac{2}{3}, \frac{1}{4}, \ldots$$

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

The real numbers between zero and one, however, are not countable. Here we will explain why. Suppose somebody thinks they have a method of arranging them into a sequence $x_1, x_2, ...,$ where we express them as $x_j = \sum_{i=1}^{\infty} d_{ij} 10^{-i}$ so that $d_{ij} \in \{0, 1, 2, ..., 9\}$ is the ith digit after the decimal place of the jth number in their sequence. Then you can clearly see that the number

$$y = \sum_{i=1}^{\infty} (1 + I\{d_{ii} = 1\}) 10^{-i},$$

where $I\{A\}$ equals one if A is true and zero otherwise is nowhere to be found in their sequence. This is because y differs from x_i in at least the ith decimal place, so it is different from every number in their sequence. Whenever someone tries to arrange the real numbers into a sequence, this shows that they will always be omitting some of the numbers. This proves that the real numbers in any interval are uncountable and that you can't take a sum over all of them.

So it's true with $X \sim U(0, 1)$ that for any countable set A we have $P(X \in A) = \sum_{x \in A} P(X = x) = 0$, but we can't simply sum up the probabilities like this for an uncountable set. There are, however, some examples of uncountable sets A (the Cantor set, for example) that have $P(X \in A) = 0$.

1.4 Probability Spaces

Let Ω be the set of points in a sample space, and let $\mathcal{F}$ be the collection of subsets of Ω for which we can calculate a probability. These subsets are called events and can be viewed as possible things that could happen. If we let P be the function that gives the probability for any event in $\mathcal{F}$, then the triple $(\Omega, \mathcal{F}, P)$ is called a probability space. The collection $\mathcal{F}$ is usually what is called a sigma field (also called a sigma algebra), which we define next.

Definition 1.1 *The collection of sets $\mathcal{F}$ is a sigma field, or a σ field, if it has the following three properties:*

1. $\Omega \in \mathcal{F}$
2. $A \in \mathcal{F} \rightarrow A^c \in \mathcal{F}$
3. $A_1, A_2, \ldots \in \mathcal{F} \rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

These properties say you can calculate the probability of the whole sample space (Property 1), the complement of any event (Property 2), and

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

the countable union of any sequence of events (Property 3). They also imply that you can calculate the probability of the countable intersection of any sequence of events because $\cap_{i=1}^{\infty} A_i = (\cup_{i=1}^{\infty} A_i^c)^c$.

To specify a σ field, people typically start with a collection of events $\mathcal{A}$ and write $\sigma(\mathcal{A})$ to represent the smallest σ field containing the collection of events $\mathcal{A}$. Thus $\sigma(\mathcal{A})$ is called the σ field "generated" by $\mathcal{A}$. It is uniquely defined as the intersection of all possible sigma fields that contain $\mathcal{A}$, and in Exercise 3 at the end of this chapter, you will show such an intersection is always a sigma field.

Example 1.2 Let $\Omega = \{a, b, c\}$ be the sample space, and let $\mathcal{A} = \{\{a, b\}, \{c\}\}$. Then $\mathcal{A}$ is not a σ field because $\{a, b, c\} \notin \mathcal{A}$, but $\sigma(\mathcal{A}) = \{\{a, b, c\}, \{a, b\}, \{c\}, \phi\}$, where $\phi = \Omega^c$ is the empty set.

Definition 1.3 *A probability measure P is a function, defined on the sets in a sigma field, which has the following three properties:*

1. $P(\Omega) = 1$, *and*
2. $P(A) \geq 0$, *and*
3. $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ *if* $\forall i \neq j$ *we have* $A_i \cap A_j = \phi$.

These properties imply that probabilities must be between zero and one and say that the probability of a countable union of mutually exclusive events is the sum of the probabilities.

Example 1.4 *Dice.* If you roll a pair of dice, the 36 points in the sample space are $\Omega = \{(1, 1), (1, 2), ..., (5, 6), (6, 6)\}$. We can let $\mathcal{F}$ be the collection of all possible subsets of Ω, and it's easy to see that it is a sigma field. Then we can define

$$P(A) = \frac{|A|}{36},$$

where $|A|$ is the number of sample space points in A. Thus, if $A = \{(1, 1), (3, 2)\}$, then $P(A) = 2/36$, and it's easy to see that P is a probability measure.

Example 1.5 *The unit interval.* Suppose we want to pick a uniform random number between zero and one. Then the sample space equals $\Omega = [0, 1]$, the set of all real numbers between zero and one. We can let $\mathcal{F}$ be the collection of all possible subsets of Ω, and it's easy to see that it is a sigma field. But it turns out that it's not possible to put a probability measure on this sigma field. Because one of the sets in $\mathcal{F}$ would be similar to the set of heads of the family (from the nonmeasurable event example), this event cannot have a probability assigned to it. So this sigma field is not a good one to use in probability.

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Example 1.6 *The unit interval again.* Again with $\Omega = [0, 1]$, suppose we use the sigma field $\mathcal{F} = \sigma(\{x\}_{x\in\Omega})$, the smallest sigma field generated by all possible sets containing a single real number. This is a nice enough sigma field, but it would never be possible to find the probability for some interval, such as [0.2, 0.4]. You can't take a countable union of single real numbers and expect to get an uncountable interval somehow. So this is not a good sigma field to use.

If we want to put a probability measure on the real numbers between zero and one, what sigma field can we use? The answer is the *Borel sigma field* $\mathcal{B}$, the smallest sigma field generated by all intervals of the form $[x, y)$ of real numbers between zero and one: $\mathcal{B} = \sigma([x, y)_{x<y\in\Omega})$. The sets in this sigma field are called Borel sets. We will see that most reasonable sets you would be interested in are Borel sets, although sets similar to the one in the "heads of the family" example are not Borel sets.

We can then use the special probability measure, which is called a *Lebesgue measure* (named after the French mathematician Henri Lebesgue), defined by $P([x, y)) = y - x$, for $0 \le x \le y \le 1$, to give us a uniform distribution. Defining it for just these intervals is enough to uniquely specify the probability of every set in $\mathcal{B}$. (This fact can be shown to follow from Theorem 1.65, which is discussed later). And actually, you can do almost all of probability starting from just a uniform(0,1) random variable, so this probability measure is pretty much all you need.

Example 1.7 If $\mathcal{B}$ is the Borel sigma field on [0,1], is $\{.5\} \in \mathcal{B}$? Yes, because $\{0.5\} = \cap_{i=1}^{\infty}[0.5, 0.5 + 1/i)$. Also note that $\{1\} = [0, 1)^c \in \mathcal{B}$.

Example 1.8 If $\mathcal{B}$ is the Borel sigma field on [0,1], is the set of rational numbers between zero and one $Q \in \mathcal{B}$? The argument from the previous example shows $\{x\} \in \mathcal{B}$ for all x, so each number by itself is a Borel set, and we then get $Q \in \mathcal{B}$ because Q is countable union of such numbers. Also note that this then means $Q^c \in \mathcal{B}$, so the set of irrational numbers is also a Borel set.

There are some Borel sets that can't directly be written as a countable intersection or union of intervals like the preceding, but you usually don't run into them.

From the definition of probability, we can derive many of the famous formulas you may have seen before such as

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

and extending this by induction,

$$P(\cup_{i=1}^{n} A_i) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) \cdots \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cdots \cap A_n),$$

where the last formula is usually called the *inclusion–exclusion formula*. Next we give a couple of examples applying these. In these examples, the sample space is finite, and in such cases unless otherwise specified, we assume the corresponding sigma field is the set of all possible subsets of the sample space.

Example 1.9 *Cards.* A deck of n cards is well shuffled many times. (a) What's the probability the cards all get back to their initial positions? (b) What's the probability at least one card is back in its initial position?

Solution Because there are $n!$ different ordering for the cards and all are approximately equally likely after shuffling, the answer to Part (a) is approximately $1/n!$. For the answer to Part (b), let $A_i = \{\text{card } i \text{ is back in its initial position}\}$ and let $A = \cup_{i=1}^{\infty} A_i$ be the event at least one card is back in its initial position. Because $P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = (n-k)!/n!$, and because the number of terms in the kth sum of the inclusion–exclusion formula is $\binom{n}{k}$, we have

$$\begin{aligned} P(A) &= \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} \\ &= \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!} \\ &\approx 1 - 1/e \end{aligned}$$

for large n. ■

Example 1.10 *Coins.* If a fair coin is flipped n times, what is the chance of seeing at least k heads in row?

Solution We will show you that the answer is

$$\sum_{m=1}^{(n+1)/(k+1)} (-1)^{m+1} \left[\tbinom{n-mk}{m} 2^{-m(k+1)} + \tbinom{n-mk}{m-1} 2^{-m(k+1)+1} \right].$$

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

When we define the event $A_i = \{$a run of a tail immediately followed by k heads in a row starts at flip $i\}$, and $A_0 = \{$the first k flips are heads$\}$, we can use the inclusion–exclusion formula to get this solution because

$$P(\text{at least } k \text{ heads in row}) = P(\cup_{i=0}^{n-k-1} A_i)$$

and

$$P(A_{i_1} A_{i_2} \cdots A_{i_m}) = \begin{cases} 0 & \text{if flips for any events overlap} \\ 2^{-m(k+1)} & \text{otherwise and } i_1 > 0 \\ 2^{-m(k+1)+1} & \text{otherwise and } i_1 = 0 \end{cases}$$

and the number of sets of indices $i_1 < i_2 < \cdots < i_m$, where the runs that do not overlap equal $\binom{n-mk}{m}$ if $i_1 > 0$ (imagine the k heads in each of the m runs are invisible, so this is the number of ways to arrange m tails in $n - mk$ visible flips) and $\binom{n-mk}{m-1}$ if $i_1 = 0$. ■

An important property of the probability function is that it is a continuous function on the events of the sample space Ω. To make this precise, let $A_n, n \geq 1$ be a sequence of events, and define the event $\liminf A_n$ as

$$\liminf A_n \equiv \cup_{n=1}^{\infty} \cap_{i=n}^{\infty} A_i.$$

Because $\liminf A_n$ consists of all outcomes of the sample space that are contained in $\cap_{i=n}^{\infty} A_i$ for some n, it follows that $\liminf A_n$ consists of all outcomes that are contained in all but a finite number of the events $A_n, n \geq 1$.

Similarly, the event $\limsup A_n$ is defined by

$$\limsup A_n = \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i.$$

Because $\limsup A_n$ consists of all outcomes of the sample space that are contained in $\cup_{i=n}^{\infty} A_i$ for all n, it follows that $\limsup A_n$ consists of all outcomes that are contained in an infinite number of the events $A_n, n \geq 1$. Sometimes the notation $\{A_n \text{ i.o.}\}$ is used to represent $\limsup A_n$, where i.o. stands for infinitely often and means that an infinite number of the events A_n occur.

Note that by their definitions

$$\liminf A_n \subset \limsup A_n.$$

Definition 1.11 *If* $\limsup A_n = \liminf A_n$, *we say that* $\lim_n A_n$ *exists and define it by*

$$\lim_n A_n = \limsup A_n = \liminf A_n.$$

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Example 1.12 (a) Suppose that $A_n, n \geq 1$ is an increasing sequence of events, in that $A_n \subset A_{n+1}$, $n \geq 1$. Then $\cap_{i=n}^{\infty} A_i = A_n$, showing that

$$\liminf A_n = \cup_{n=1}^{\infty} A_n.$$

Also, $\cup_{i=n}^{\infty} A_i = \cup_{i=1}^{\infty} A_i$, showing that

$$\limsup A_n = \cup_{n=1}^{\infty} A_n.$$

Hence,

$$\lim_n A_n = \cup_{i=1}^{\infty} A_i.$$

(b) If $A_n, n \geq 1$ is a decreasing sequence of events, in that $A_{n+1} \subset A_n$, $n \geq 1$, then it similarly follows that

$$\lim_n A_n = \cap_{i=1}^{\infty} A_i. \qquad \blacksquare$$

The following result is known as the *continuity property of probabilities*.

Proposition 1.13 *If* $\lim_n A_n = A$, *then* $\lim_n P(A_n) = P(A)$.

Proof We prove it first for when A_n is either an increasing or decreasing sequence of events. Suppose $A_n \subset A_{n+1}$, $n \geq 1$. Then, with A_0 defined to be the empty set,

$$\begin{aligned}
P(\lim A_n) &= P(\cup_{i=1}^{\infty} A_i) \\
&= P(\cup_{i=1}^{\infty} A_i (\cup_{j=1}^{i-1} A_j)^c) \\
&= P(\cup_{i=1}^{\infty} A_i A_{i-1}^c) \\
&= \sum_{i=1}^{\infty} P(A_i A_{i-1}^c) \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} P(A_i A_{i-1}^c) \\
&= \lim_{n\to\infty} P(\cup_{i=1}^{n} A_i A_{i-1}^c) \\
&= \lim_{n\to\infty} P(\cup_{i=1}^{n} A_i) \\
&= \lim_{n\to\infty} P(A_n).
\end{aligned}$$

Now, suppose that $A_{n+1} \subset A_n$, $n \geq 1$. Because A_n^c is an increasing sequence of events, the preceding implies that

$$P(\cup_{i=1}^{\infty} A_i^c) = \lim_{n\to\infty} P(A_n^c),$$

or equivalently,

$$P((\cap_{i=1}^{\infty} A_i)^c) = 1 - \lim_{n\to\infty} P(A_n)$$

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

or

$$P(\cap_{i=1}^{\infty} A_i) = \lim_{n\to\infty} P(A_n),$$

which completes the proof whenever A_n is a monotone sequence. Now, consider the general case, and let $B_n = \cup_{i=n}^{\infty} A_i$. Noting that $B_{n+1} \subset B_n$, and applying the preceding yields

$$\begin{aligned} P(\limsup A_n) &= P(\cap_{n=1}^{\infty} B_n) \\ &= \lim_{n\to\infty} P(B_n). \end{aligned} \tag{1.1}$$

Also, with $C_n = \cap_{i=n}^{\infty} A_i$,

$$\begin{aligned} P(\liminf A_n) &= P(\cup_{n=1}^{\infty} C_n) \\ &= \lim_{n\to\infty} P(C_n) \end{aligned} \tag{1.2}$$

because $C_n \subset C_{n+1}$. But

$$C_n = \cap_{i=n}^{\infty} A_i \subset A_n \subset \cup_{i=n}^{\infty} A_i = B_n,$$

showing that

$$P(C_n) \leq P(A_n) \leq P(B_n). \tag{1.3}$$

Thus, if $\liminf A_n = \limsup A_n = \lim A_n$, then we obtain from Equations 1.1 and 1.2 that the upper and lower bounds of Equation 1.3 converge to each other in the limit, and this proves the result. ■

1.5 Random Variables

Suppose you have a function X that assigns a real number to each point in the sample space Ω and you also have a sigma field $\mathcal{F}$. We say that X is an $\mathcal{F}$-measurable random variable if you can compute its entire cumulative distribution function using probabilities of events in $\mathcal{F}$ or, equivalently, that you would know the value of X if you were told which events in $\mathcal{F}$ actually happen. We define the notation $\{X \leq x\} \equiv \{\omega \in \Omega : X(\omega) \leq x\}$, so X is $\mathcal{F}$ measurable if $\{X \leq x\} \in \mathcal{F}$ for all x. This is often written in shorthand notation as $X \in \mathcal{F}$.

Example 1.14 $\Omega = \{a, b, c\}$, $\mathcal{A} = \{\{a, b, c\}, \{a, b\}, \{c\}, \phi\}$, and we define three random variables X, Y, Z as follows:

ω	**X**	**Y**	**Z**
a	1	1	1
b	1	2	7
c	2	2	4

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Which of the random variables X, Y, and Z are $\mathcal{A}$ measurable? Because $\{Y \leq 1\} = \{a\} \notin \mathcal{A}$, then Y is not $\mathcal{A}$ measurable. For the same reason, Z is not $\mathcal{A}$ measurable. The variable X is $\mathcal{A}$ measurable because $\{X \leq 1\} = \{a, b\} \in \mathcal{A}$, and $\{X \leq 2\} = \{a, b, c\} \in \mathcal{A}$. In other words, you can always figure out the value of X using just the events in $\mathcal{A}$, but you can't always figure out the values of Y and Z.

Definition 1.15 *For a random variable X we define*

$$\sigma(X) = \sigma(\{X \leq x\}, \forall x)$$

to be the sigma field generated by all events of the type $\{\mathbf{X} \leq x\}$, where $\sigma(X)$ is the sigma field generated by X.

Alternatively, we can define $\sigma(X)$ as the intersection of all possible sigma fields $\mathcal{F}$ where X is $\mathcal{F}$ measurable; such an uncountable intersection is a sigma field, as in Exercise 3 at the end of this chapter. Intuitively, $\sigma(X)$ contains just enough events to know the value of X when you know which of the events occur.

Definition 1.16 *For random variables X, Y we say that X is Y measurable if $X \in \sigma(Y)$.*

Example 1.17 In the previous example, is $Y \in \sigma(Z)$? Yes, because $\sigma(Z) = \{\{a, b, c\}, \{a\}, \{a, b\}, \{b\}, \{b, c\}, \{c\}, \{c, a\}, \phi\}$, the set of all possible subsets of Ω. Is $X \in \sigma(Y)$? No, because $\{X \leq 1\} = \{a, b\} \notin \sigma(Y) = \{\{a, b, c\}, \{b, c\}, \{a\}, \phi\}$.

To see why $\sigma(Z)$ is as given, note that $\{Z \leq 1\} = \{a\}$, $\{Z \leq 4\} = \{a, c\}$, $\{Z \leq 7\} = \{a, b, c\}$, $\{a\}^c = \{b, c\}$, $\{a, b\}^c = \{c\}$, $\{a\} \cup \{c\} = \{a, c\}$, $\{a, b, c\}^c = \phi$, and $\{a, c\}^c = \{b\}$.

Example 1.18 Suppose X and Y are random variables taking values between zero and one and are measurable with respect to the Borel sigma field $\mathcal{B}$. Is $Z = X + Y$ also measurable with respect to $\mathcal{B}$? Well, we must show that $\{Z \leq z\} \in \mathcal{B}$ for all z. We can write

$$\{X + Y > z\} = \cup_{q \in Q}(\{X > q\} \cap \{Y > z - q\}),$$

where Q is the set of rational numbers. Because $\{X > q\} \in \mathcal{B}$, $\{Y > z - q\} \in \mathcal{B}$, and Q is countable, this means that $\{X + Y \leq z\} = \{X + Y > z\}^c \in \mathcal{B}$ and thus Z is measurable with respect to $\mathcal{B}$.

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Example 1.19 The function $F(x) = P(X \leq x)$ is called the distribution function of the random variable X. If $x_n \downarrow x$ then the sequence of events $A_n = \{X \leq x_n\}$, $n \geq 1$, is a decreasing sequence with a limit that is

$$\lim_n A_n = \cap_n A_n = \{X \leq x\}.$$

Consequently, the continuity property of probabilities yields

$$F(x) = \lim_{n \to \infty} F(x_n),$$

showing that a distribution function is always right continuous. On the other hand, if $x_n \uparrow x$, then the sequence of events $A_n = \{X \leq x_n\}$, $n \geq 1$, is an increasing sequence, implying that

$$\lim_{n \to \infty} F(x_n) = P(\cup_n A_n) = P(X < x) = F(x) - P(X = x).$$

Two events are independent if knowing that one occurs does not change the chance that the other occurs. This is formalized in the following definition.

Definition 1.20 *Sigma fields $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are independent if whenever $A_i \in \mathcal{F}_i$ for $i = 1, \ldots, n$, we have $P(\cap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i)$.*

Using this we say that random variables $X_1, \ldots, X_n$ are *independent* if the sigma fields $\sigma(X_1), \ldots, \sigma(X_n)$ are independent, and we say events $A_1, \ldots, A_n$ are independent if $I_{A_1}, \ldots, I_{A_n}$ are independent random variables.

Remark 1.21 One interesting property of independence is that it's possible that events A, B, C are not independent even if each pair of the events are independent. For example, if we make three independent flips of a fair coin and let A represent the event exactly one head comes up in the first two flips, let B represent the event exactly one head comes up in the last two flips, and let C represent the event exactly one head comes up among the first and last flip. Then each event has probability $1/2$, the intersection of each pair of events has probability $1/4$, but we have $P(ABC) = 0$.

In our next example, we derive a formula for the distribution of the convolution of geometric random variables.

Example 1.22 Suppose we have n coins that we toss in sequence, moving from one coin to the next in line each time a head appears. That is, we continue using a coin until it lands heads, and then we switch to the next one. Let X_i denote the number of flips made with coin i. Assuming that all coin flips are independent and that each lands heads with probability p,

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

we know from our first course in probability that X_i is a geometric random variable with parameter p and that the total number of flips made has a negative binomial distribution with probability mass function

$$P(X_1 + \cdots + X_n = m) = \binom{m-1}{n-1} p^n (1-p)^{m-n}, \quad m \geq n.$$

The probability mass function of the total number of flips when each coin has a different probability of landing heads is easily obtained using the following proposition.

Proposition 1.23 *If $X_1, \ldots, X_n$ are independent geometric random variables with parameters $p_1, \ldots, p_n$, where $p_i \neq p_j$ if $i \neq j$, then, with $q_i = 1 - p_i$, for $k \geq n - 1$*

$$P(X_1 + \cdots + X_n > k) = \sum_{i=1}^{n} q_i^k \prod_{j \neq i} \frac{p_j}{p_j - p_i}.$$

Proof We will prove $A_{k,n} = P(X_1 + \cdots + X_n > k)$ is as given using induction on $k + n$. Because $A_{1,1} = q_1$, we will assume as our induction hypothesis that $A_{i,j}$ is as given previously for all $i + j < k + n$. Then, depending on whether or not the event $\{X_n > 1\}$ occurs, we get

$$\begin{aligned} A_{k,n} &= q_n A_{k-1,n} + p_n A_{k-1,n-1} \\ &= q_n \sum_{i=1}^{n} q_i^{k-1} \prod_{j \neq i} \frac{p_j}{p_j - p_i} + p_n \sum_{i=1}^{n-1} q_i^{k-1} \frac{p_n - p_i}{p_n} \prod_{j \neq i} \frac{p_j}{p_j - p_i} \\ &= \sum_{i=1}^{n} q_i^k \prod_{j \neq i} \frac{p_j}{p_j - p_i}, \end{aligned}$$

which completes the proof by induction. ■

1.6 Convergence, Limits, sup, and inf

A sequence of real numbers $x_1, x_2, \ldots$ *converges to a limit* x, and we write this as $\lim_{n\to\infty} x_n = x$ or $\lim_n x_n = x$ or $x_n \to x$ if for any $\epsilon > 0$ the values in the sequence beyond some point are all within ϵ of x. We write $x_n \uparrow x$ if $x_n \to x$ and the sequence is nondecreasing, and we write $x_n \downarrow x$ if $x_n \to x$ and the sequence is nonincreasing.

If X_n is a sequence of random variables and we write $X_n \to X$, we mean that if we observe the sequence and then consider it as a sequence of real numbers, we will always have $X_n \to X$.

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Example 1.24 If $x_n = n/(n+1)$ for $n = 1, 2, \ldots$ then we have $x_n \uparrow 1$. This is because x_n is nondecreasing, and given any $\epsilon > 0$, we can let $n = 1/\epsilon$ and $1 - x_i = 1/(i+1) < \epsilon$ when when $i > n$.

Example 1.25 If $x_n = n/(n+1)$ when n is even and $x_n = 0$ when n is odd, we say that the sequence has no limit. Because for $n \geq 1$ we have $x_{2n} \geq 2/3$ and $x_{2n+1} = 0$, when $\epsilon = 1/3$ we can never find an n such that all the values beyond the nth value are less than ϵ from the same number.

If x_i for $i \in S$ are real numbers with indices in a set S we write

$$x = \sup_{i \in S} x_i$$

if $x_i \leq x$ for all i and for any $y < x$ there is some $i \in S$ such that $x_i > y$. We say that x is the *supremum* of the set $\{x_i : i \in S\}$, which means it is the smallest possible upper bound for the set. Here S may be either a countable or an uncountable set. We also define the infimum of a set as the largest possible lower bound so that if

$$x = \inf_{i \in S} x_i$$

it means $x_i \geq x$ for all i and for any $y > x$ there is some $i \in S$ such that $x_i < y$.

Example 1.26 If $S = \{1, 2, \ldots\}$ and $x_i = i$, we have that $\sup_{i \in S} x_i = \infty$ and $\inf_{i \in S} x_i = 1$. Also note that there is no maximum value of x_i, so $\max_{i \in S} x_i$ does not exist.

Every set of real numbers has a supremum and an infimum, although these may not actually be in the set. Infinite sets may not have a maximum or minimum value within them, although finite sets always do.

1.7 Expected Value

A random variable X is *continuous* if there is a function f, called its density function, so $P(X \leq x) = \int_{-\infty}^{x} f(t)dt$ for all x. A random variable is *discrete* if it can only take a countable number of different values. In elementary textbooks, you usually see two separate definitions for expected value:

$$E[X] = \begin{cases} \sum_i x_i P(X = x_i) & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous with density } f. \end{cases}$$

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

But it's possible to have a random variable that is neither continuous nor discrete. For example, with $U \sim U(0,1)$, the variable $X = UI_{U>0.5}$ is neither continuous nor discrete. It's also possible to have a sequence of continuous random variables that converges to a discrete random variable or vice versa. For example, if $X_n = U/n$, then each X_n is a continuous random variable, but $\lim_{n\to\infty} X_n$ is a discrete random variable (which equals zero). This means it would be better to have a single more general definition that covers all types of random variables. We introduce this next.

A *simple* random variable is one that can take on only a finite number of different possible values, and its expected value is defined as in the first paragraph in this section for discrete random variables. Using these, we next define the expected value of a more general nonnegative random variable. We will later define it for general random variables X by expressing it as the difference of two nonnegative random variables $X = X^+ - X^-$, where $x^+ = \max(0, x)$ and $x^- = \max(-x, 0)$.

Definition 1.27 *If $X \geq 0$, then we define*

$$E[X] \equiv \sup_{\text{all simple variables } Y \leq X} E[Y].$$

We write $Y \leq X$ for random variables X, Y to mean $P(Y \leq X) = 1$; this is sometimes written as "$Y \leq X$ almost surely" and abbreviated "$Y \leq X$ a.s." For example, if X is nonnegative and $a \geq 0$, then $Y = aI_{X \geq a}$ is a simple random variable such that $Y \leq X$. And by taking a supremum over all simple variables, we of course mean the simple random variables must be measurable with respect to some given sigma field. Given a nonnegative random variable X, one concrete choice of simple variables is the sequence $Y_n = \min(\lfloor 2^n X \rfloor / 2^n, n)$, where $\lfloor x \rfloor$ denotes the integer portion of x. In Exercise 18 at the end of this chapter, we ask you to show that $Y_n \uparrow X$ and $E[X] = \lim_n E[Y_n]$.

Another consequence of the definition of expected value is that if $Y \leq X$, then $E[Y] \leq E[X]$.

Example 1.28 *Markov's inequality.* Suppose $X \geq 0$. Then, for any $a > 0$ we have that $a\, I_{X \geq a} \leq X$. Therefore, $E[a\, I_{X \geq a}] \leq E[X]$ or, equivalently,

$$P(X \geq a) \leq E[X]/a,$$

which is known as Markov's inequality.

Example 1.29 *Chebyshev's inequality.* A consequence of Markov's inequality is that for $a > 0$

$$P(|X| \geq a) = P(X^2 \geq a^2) \leq E[X^2]/a^2,$$

a result known as Chebyshev's inequality.

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Given any random variable $X \geq 0$ with $E[X] < \infty$, and any $\epsilon > 0$, we can find a simple random variable Y with $E[X] - \epsilon \leq E[Y] \leq E[X]$. Our definition of the expected value also gives what is called the Lebesgue integral of X with respect to the probability measure P and is sometimes denoted $E[X] = \int X dP$.

So far we have only defined the expected value of a nonnegative random variable. For the general case, we first define $X^+ = XI_{X\geq 0}$ and $X^- = -XI_{X<0}$ so that we can define $E[X] = E[X^+] - E[X^-]$, with the convention that $E[X]$ is undefined if $E[X^+] = E[X^-] = \infty$.

Remark 1.30 The definition of expected value covers random variables that are neither continuous nor discrete, but if X is continuous with density function f, it is equivalent to the familiar definition $E[X] = \int xf(x)dx$. For example, when $0 \leq X \leq 1$ the definition of the Riemann integral in terms of Riemann sums implies, with $\lfloor x \rfloor$ denoting the integer portion of x,

$$\begin{aligned}
\int_0^1 xf(x)dx &= \lim_{n\to\infty} \sum_{i=0}^{n-1} \int_{i/n}^{(i+1)/n} xf(x)dx \\
&\leq \lim_{n\to\infty} \sum_{i=0}^{n-1} \frac{i+1}{n} P\left(i/n \leq X \leq \frac{i+1}{n}\right) \\
&= \lim_{n\to\infty} \sum_{i=0}^{n-1} i/n P\left(i/n \leq X \leq \frac{i+1}{n}\right) \\
&= \lim_{n\to\infty} E[\lfloor nX \rfloor /n] \\
&\leq E[X],
\end{aligned}$$

where the last line follows because $\lfloor nX \rfloor /n \leq X$ is a simple random variable.

Using that the density function g of $1 - X$ is $g(x) = f(1-x)$, we obtain

$$\begin{aligned}
1 - E[X] &= E[1 - X] \\
&\geq \int_0^1 xf(1-x)dx \\
&= \int_0^1 (1-x)f(x)dx \\
&= 1 - \int_0^1 xf(x)dx.
\end{aligned}$$

Remark 1.31 At this point, you may think it might be possible to express any random variable as sums or mixtures of discrete and continuous random variables, but this is not true. Let $X \sim U(0,1)$ be a uniform random

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

variable, and let $d_i \in \{0, 1, 2, ..., 9\}$ be the ith digit in its decimal expansion so that $X = \sum_{i=1}^{\infty} d_i 10^{-i}$. The random variable $Y = \sum_{i=1}^{\infty} \min(1, d_i) 10^{-i}$ is not discrete and has no intervals over which it is continuous. This variable Y can take any value (between zero and one) having a decimal expansion that uses only the digits 0 and 1, which are a set of values C called a *Cantor set*. Because C contains no intervals, Y is not continuous. And Y is not discrete because C is uncountable; every real number between zero and one, using its base two expansion, corresponds to a distinct infinite sequence of binary digits.

Another interesting fact about a Cantor set is, although C is uncountable, $P(X \in C) = 0$. Let C_i be the set of real numbers between zero and one that have a decimal expansion using only the digits 0 and 1 up to the ith decimal place. Then it's easy to see that $P(X \in C_i) = 0.2^i$ and because $P(X \in C) \leq P(X \in C_i) = 0.2^i$ for any i, we must have $P(X \in C) = 0$. The set C is called an uncountable set having measure zero.

Proposition 1.32 *If $E|X|, E|Y| < \infty$ then (a) $E[aX + b] = aE[X] + b$ for constants a, b, and (b) $E[X + Y] = E[X] + E[Y]$.*

Proof In this proof we assume $X, Y \geq 0$, $a > 0$, and $b = 0$. The general cases will follow using $E[X + Y] = E[X^+ + Y^+] - E[X^- + Y^-]$,

$$E[b + X] = \sup_{Y \leq b+X} E[Y] = \sup_{Y \leq X} E[b + Y] = b + \sup_{Y \leq X} E[Y] = b + E[X],$$

and $-aX + b = a(-X) + b$.

For Part (a) if X is simple we have

$$E[aX] = \sum_x axP(X = x) = aE[X],$$

and because for every simple variable $Z \leq X$ there corresponds another simple variable $aZ \leq aX$, and vice versa, we get

$$E[aX] = \sup_{aZ \leq aX} E[aZ] = \sup_{Z \leq X} aE[Z] = aE[X],$$

where the supremums are over simple random variables.

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

For Part (b) if X, Y are simple we have

$$\begin{aligned}
E[X+Y] &= \sum_z zP(X+Y=z) \\
&= \sum_z z \sum_{x,y:x+y=z} P(X=x, Y=y) \\
&= \sum_z \sum_{x,y:x+y=z} (x+y)P(X=x, Y=y) \\
&= \sum_{x,y} (x+y)P(X=x, Y=y) \\
&= \sum_{x,y} xP(X=x, Y=y) + \sum_{x,y} yP(X=x, Y=y) \\
&= \sum_x xP(X=x) + \sum_y yP(Y=y) \\
&= E[X] + E[Y],
\end{aligned}$$

and applying this in the following second line, we get

$$\begin{aligned}
E[X] + E[Y] &= \sup_{A \le X, B \le Y} E[A] + E[B] \\
&= \sup_{A \le X, B \le Y} E[A+B] \\
&\le \sup_{A \le X+Y} E[A] \\
&= E[X+Y],
\end{aligned}$$

where the supremums are over simple random variables. We then use this inequality in the following third line:

$$\begin{aligned}
E[\min(X+Y, n)] &= 2n - E[2n - \min(X+Y, n)] \\
&\le 2n - E[n - \min(X, n) + n - \min(Y, n)] \\
&\le 2n - E[n - \min(X, n)] - E[n - \min(Y, n)] \\
&= E[\min(X, n)] + E[\min(Y, n)] \\
&\le E[X] + E[Y],
\end{aligned}$$

and we use Part (a) in the first and fourth lines and $\min(X+Y, n) \le \min(X, n) + \min(Y, n)$ in the second line.

This means for any given simple $Z \le X + Y$ we can pick n larger than the maximum value of Z so that $E[Z] \le E[\min(X+Y, n)] \le E[X] + E[Y]$, and taking the supremum over all simple $Z \le X + Y$ gives $E[X+Y] \le E[X] + E[Y]$ and the result is proved. ■

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Proposition 1.33 *If X is a nonnegative integer valued random variable, then*

$$E[X] = \sum_{n=0}^{\infty} P(X > n).$$

Proof Because $E[X] = p_1 + 2p_2 + 3p_3 + 4p_4 \ldots$ (see Exercise 7 at the end of this chapter), where $p_i = P(X = i)$, we rewrite this as

$$\begin{array}{rccccccc} E[X] = p_1 & + & p_2 & + & p_3 & + & p_4 & \ldots \\ & + & p_2 & + & p_3 & + & p_4 & \ldots \\ & & & + & p_3 & + & p_4 & \ldots \\ & & & & & + & p_4 & \ldots . \end{array}$$

Notice that the columns equal $p_1, 2p_2, 3p_3, \ldots$, respectively, whereas the rows equal $P(X > 0), P(X > 1), P(X > 2), \ldots$, respectively. ■

Example 1.34 With $X_1, X_2 \ldots$ independent $U(0, 1)$ random variables, compute the expected value of

$$N = \min \left\{ n : \sum_{i=1}^{n} X_i > 1 \right\}.$$

Solution Using $E[N] = \sum_{n=0}^{\infty} P(N > n)$, and noting that

$$P(N > 0) = P(N > 1) = 1,$$

and

$$\begin{aligned} P(N > n) &= \int_0^1 \int_0^{1-x_1} \int_0^{1-x_1-x_2} \cdots \int_0^{1-x_1-x_2-\cdots-x_{n-1}} dx_n \cdots dx_1 \\ &= 1/n!, \end{aligned}$$

we get $E[N] = e$. ■

1.8 Almost Sure Convergence and the Dominated Convergence Theorem

For a sequence of nonrandom real numbers, recall that we write $x_n \to x$ or $\lim_{n\to\infty} x_n = x$ if for any $\varepsilon > 0$ there exists a value n such that $|x_m - x| < \varepsilon$ for all $m > n$. Intuitively, this means eventually the sequence never leaves an arbitrarily small neighborhood around x. It doesn't simply mean that you can always find terms in the sequence that are arbitrarily close to x, but

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

rather it means that eventually *all* terms in the sequence become arbitrarily close to x. When $x_n \to \infty$, it means that for any $k > 0$ there exists a value n such that $x_m > k$ for all $m > n$.

The sequence of random variables $X_n, n \geq 1$, is said to converge *almost surely* to the random variable X, written as $X_n \longrightarrow_{as} X$, or $\lim_{n\to\infty} X_n = X$ a.s., if with

$$\lim_n X_n = X.$$

The following proposition presents an alternative characterization of almost sure convergence.

Proposition 1.35 $X_n \longrightarrow_{as} X$ *if and only if for any* $\epsilon > 0$

$$P(|X_n - X| < \epsilon \text{ for all } n \geq m) \to 1 \quad as \quad m \to \infty.$$

Proof Suppose first that $X_n \longrightarrow_{as} X$. Fix $\epsilon > 0$, and for $m \geq 1$, define the event

$$A_m = \{|X_n - X| < \epsilon \text{ for all } n \geq m\}.$$

Because $A_m, m \geq 1$, is an increasing sequence of events, the continuity property of probabilities yields that

$$\begin{aligned} \lim_m P(A_m) &= P(\lim_m A_m) \\ &= P(|X_n - X| < \epsilon \text{ for all } n \text{ sufficiently large}) \\ &\geq P(\lim_n X_n = X) \\ &= 1. \end{aligned}$$

To go the other way, assume that for any $\epsilon > 0$

$$P(|X_n - X| < \epsilon \text{ for all } n \geq m) \to 1 \quad \text{as} \quad m \to \infty.$$

Let ϵ_i, $i \geq 1$, be a decreasing sequence of positive numbers that converge to 0, and let

$$A_{m.i} = \{|X_n - X| < \epsilon_i \text{ for all } n \geq m\}.$$

Because $A_{m.i} \subset A_{m+1.i}$ and, by assumption, $\lim_m P(A_{m,i}) = 1$, it follows from the continuity property that

$$1 = P(\lim_{m\to\infty} A_{m.i}) = P(B_i),$$

where $B_i = \{|X_n - X| < \epsilon_i \text{ for all } n \text{ sufficiently large}\}$. But B_i, $i \geq 1$, is a decreasing sequence of events, so invoking the continuity property once again yields

$$1 = \lim_{i\to\infty} P(B_i) = P(\lim_i B_i),$$

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

which proves the result because

$$\begin{aligned}\lim_i B_i &= \{\text{for all } i,\ |X_n - X| < \epsilon_i \text{ for all } n \text{ sufficiently large}\} \\ &= \{\lim_n X_n = X\}.\end{aligned}$$

■

Remark 1.36 The reason for the word almost in "almost surely" is because $P(A) = 1$ doesn't necessarily mean that A^c is the empty set. For example, if $X \sim U(0,1)$, we know that $P(X \neq 1/3) = 1$ even though $\{X = 1/3\}$ is a possible outcome.

The dominated convergence theorem is one of the fundamental building blocks of all limit theorems in probability. It tells you something about what happens to the expected value of random variables in a sequence if the random variables are converging almost surely. Many limit theorems in probability involve an almost surely converging sequence, and being able to accurately say something about the expected value of the limiting random variable is important.

Given a sequence of random variables $X_1, X_2, \ldots$, it may seem to you at first thought that $X_n \to X$ a.s. should imply $\lim_{n\to\infty} E[X_n] = E[X]$. This is sometimes called *interchanging limit and expectation*, because $E[X] = E[\lim_{n\to\infty} X_n]$. But this interchange is not always valid, and the next example illustrates this.

Example 1.37 Suppose $U \sim U(0,1)$ and $X_n = nI_{n<1/U}$. Regardless of what U turns out to be, as soon as n gets larger than $1/U$, we see that the terms X_n in the sequence will all equal zero. This means $X_n \to 0$ a.s., but at the same time we have $E[X_n] = nP(U < 1/n) = n/n = 1$ for all n, and thus $\lim_{n\to\infty} E[X_n] = 1$. Interchanging limit and expectation is not valid in this case.

What's going wrong here? In this case, X_n can increase beyond any level as n gets larger and larger, and this can cause problems with the expected value. The dominated convergence theorem says that if X_n is always bounded in absolute value by some other random variable with finite mean, then we can interchange limit and expectation. We will first state the theorem, give some examples, and then give a proof. The proof is a nice illustration of the definition of expected value.

Proposition 1.38 *The dominated convergence theorem. Suppose $X_n \to X$ a.s., and there is a random variable Y with $E[Y] < \infty$ such that $|X_n| < Y$ for all n. Then*

$$E[\lim_{n\to\infty} X_n] = \lim_{n\to\infty} E[X_n].$$

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

This is often used in the form where Y is a nonrandom constant, and then it's called the *bounded convergence theorem*. Before we prove it, we first give a couple of examples and illustrations.

Example 1.39 Suppose $U \sim U(0,1)$ and $X_n = U/n$. It's easy to see that $X_n \to 0$ a.s., and the theorem would tell us that $E[X_n] \to 0$. In fact, in this case we can easily calculate $E[X_n] = \frac{1}{2n} \to 0$. The theorem applies using $Y = 1$ because $|X_n| < 1$.

Example 1.40 With $X \sim N(0,1)$, let $X_n = \min(X, n)$, and notice $X_n \to X$ almost surely. Because $X_n < |X|$, we can apply the theorem using $Y = |X|$ to tell us $E[X_n] \to E[X]$.

Example 1.41 Suppose $X \sim N(0,1)$ and let $X_n = XI_{X \geq -n} - nI_{X<-n}$. Again $X_n \to X$, so using $Y = |X|$ the theorem tells us $E[X_n] \to E[X]$.

Proof *Proof of the dominated convergence theorem.* To be able to directly apply the definition of expected value, in this proof we assume $X_n \geq 0$. To prove the general result, we can apply the same argument to $X_n + Y \geq 0$ with the bound $|X_n + Y| < 2Y$.

Our approach will be to show that for any $\varepsilon > 0$ we have, for all sufficiently large n, both (a) $E[X_n] \geq E[X] - 3\varepsilon$ and (b) $E[X_n] \leq E[X] + 3\varepsilon$. Because ε is arbitrary, this will prove the theorem.

First, let $N_\varepsilon = \min\{n : |X_i - X| < \varepsilon \text{ for all } i \geq n\}$, and note that $X_n \longrightarrow_{as} X$ implies that $P(N_\varepsilon < \infty) = 1$. To Part (a), note first that for any m

$$X_n + \varepsilon \geq \min(X, m) - mI_{N_\varepsilon > n}.$$

The preceding is true when $N_\varepsilon > n$ because in this case the right-hand side is nonpositive; it is also true when $N_\varepsilon \leq n$ because in this case $X_n + \varepsilon \geq X$. Thus,

$$E[X_n] + \varepsilon \geq E[\min(X, m)] - mP(N_\varepsilon > n).$$

Now, $|X| \leq Y$ implies that $E[X] \leq E[Y] < \infty$. Consequently, using the definition of $E[X]$, we can find a simple random variable $Z \leq X$ with $E[Z] \geq E[X] - \varepsilon$. Because Z is simple, we can then pick m large enough so $Z \leq \min(X, m)$, and thus

$$E[\min(X, m)] \geq E[Z] \geq E[X] - \varepsilon.$$

Then $N_\varepsilon < \infty$ implies, by the continuity property, that $mP(N_\varepsilon > n) < \varepsilon$ for sufficiently large n. Combining this with the preceding shows that for sufficiently large n

$$E[X_n] + \varepsilon \geq E[X] - 2\varepsilon,$$

which is Part (a).

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

For Part (b), apply Part (a) to the sequence of nonnegative random variables $Y - X_n$, which converges almost surely to $Y - X$ with a bound $|Y - X_n| < 2Y$. We get $E[Y - X_n] \geq E[Y - X] - 3\varepsilon$, and rearranging and subtracting $E[Y]$ from both sides gives Part (b). ■

Remark 1.42 Part (a) in the proof holds for nonnegative random variables even without the upper bound Y and under the weaker assumption that $\inf_{m>n} X_m \to X$ as $n \to \infty$. This result is usually referred to as *Fatou's lemma*, which states that for any $\epsilon > 0$ we have $E[X_n] \geq E[X] - \epsilon$ for sufficiently large n, or equivalently that $\inf_{m>n} E[X_m] \geq E[X] - \epsilon$ for sufficiently large n. This result is usually denoted as $\liminf_{n\to\infty} E[X_n] \geq E[\liminf_{n\to\infty} X_n]$.

A result called the *monotone convergence theorem* can also be proved.

Proposition 1.43 *The monotone convergence theorem. If*

$$0 \leq X_n \uparrow X,$$

then $E[X_n] \uparrow E[X]$.

Proof If $E[X] < \infty$, we can apply the dominated convergence theorem using the bound $|X_n| < X$.

Consider now the case where $E[X] = \infty$. For any m, we have $\min(X_n, m) \to \min(X, m)$. Because $E[\min(X, m)] < \infty$, it follows by the dominated convergence theorem that

$$\lim_n E[\min(X_n, m)] = E[\min(X, m)].$$

But because $E[X_n] \geq E[\min(X_n, m)]$, this implies

$$\lim_n E[X_n] \geq \lim_{m\to\infty} E[\min(X, m)].$$

Because $E[X] = \infty$, it follows that for any K there is a simple random variable $A \leq X$ such that $E[A] > K$. Because A is simple, $A \leq \min(X, m)$ for sufficiently large m. Thus, for any K

$$\lim_{m\to\infty} E[\min(X, m)] \geq E[A] > K,$$

proving that $\lim_{m\to\infty} E[\min(X, m)] = \infty$ and completing the proof. ■

We now present a couple of corollaries of the monotone convergence theorem.

Corollary 1.44 *If* $X_i \geq 0$, *then* $E[\sum_{i=1}^{\infty} X_i] = \sum_{i=1}^{\infty} E[X_i]$.

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Proof

$$\begin{aligned}\sum_{i=1}^{\infty} E[X_i] &= \lim_n \sum_{i=1}^{n} E[X_i] \\ &= \lim_n E\left[\sum_{i=1}^{n} X_i\right] \\ &= E\left[\sum_{i=1}^{\infty} X_i\right],\end{aligned}$$

where the final equality follows from the monotone convergence theorem because $\sum_{i=1}^{n} X_i \uparrow \sum_{i=1}^{\infty} X_i$. ■

Corollary 1.45 *If X and Y are independent, then*

$$E[XY] = E[X]E[Y].$$

Proof Suppose first that X and Y are simple. Then we can write

$$X = \sum_{i=1}^{n} x_i I_{\{X=x_i\}}, \quad Y = \sum_{j=1}^{m} y_j I_{\{Y=y_j\}}.$$

Thus,

$$\begin{aligned}E[XY] &= E\left[\sum_i \sum_j x_i y_j I_{\{X=x_i, Y=y_j\}}\right] \\ &= \sum_i \sum_j x_i y_j E[I_{\{X=x_i, Y=y_j\}}] \\ &= \sum_i \sum_j x_i y_j P(X = x_i, Y = y_j) \\ &= \sum_i \sum_j x_i y_j P(X = x_i) P(Y = y_j) \\ &= E[X]E[Y].\end{aligned}$$

Next, suppose X, Y are general nonnegative random variables. For any n, define the simple random variables

$$X_n = \begin{cases} k/2^n, & \text{if } \frac{k}{2^n} < X \le \frac{k+1}{2^n},\ k = 0, \ldots, n2^n - 1. \\ n, & \text{if } X > n \end{cases}$$

Define random variables Y_n in a similar fashion, and note that

$$X_n \uparrow X,\ Y_n \uparrow Y,\ X_n Y_n \uparrow XY.$$

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Hence, by the monotone convergence theorem,

$$E[X_nY_n] \to E[XY].$$

But X_n and Y_n are simple, so

$$E[X_nY_n] = E[X_n]E[Y_n] \to E[X]E[Y],$$

with the convergence again following by the monotone convergence theorem. Thus, $E[XY] = E[X]E[Y]$ when X and Y are nonnegative. The general case follows by writing $X = X^+ - X^-$, $Y = Y^+ - Y^-$, using

$$E[XY] = E[X^+Y^+] - E[X^+Y^-] - E[X^-Y^+] + E[X^-Y^-]$$

and applying the result to each of the four preceding expectations. ■

1.9 Convergence in Probability and in Distribution

In this section, we introduce two forms of convergence that are weaker than almost sure convergence. However, before giving their definitions, we will start with a useful result, known as the *Borel–Cantelli lemma*.

Proposition 1.46 *If* $\sum_j P(A_j) < \infty$, *then* $P(\limsup A_k) = 0$.

Proof Suppose $\sum_j P(A_j) < \infty$. Now,

$$P(\limsup A_k) = P(\cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i).$$

Hence, for any n

$$\begin{aligned} P(\limsup A_k) &\leq P(\cup_{i=n}^{\infty} A_i) \\ &\leq \sum_{i=n}^{\infty} P(A_i), \end{aligned}$$

and the result follows by letting $n \to \infty$. ■

Remark Because $\sum_n I_{A_n}$ is the number of events $A_n, n \geq 1$, that occur, the Borel–Cantelli theorem states that if the expected number of events $A_n, n \geq 1$, that occur is finite, then the probability that an infinite number of them occur is zero. Thus, the Borel–Cantelli lemma is equivalent to the rather intuitive result that if there is a positive probability that an infinite number of the events A_n occur, and then the expected number of them that occur is infinite.

The converse of the Borel–Cantelli lemma requires that the indicator variables for each pair of events be negatively correlated.

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Proposition 1.47 *Let the events* $A_i, i \geq 1$, *be such that*

$$\text{Cov}(I_{A_i}, I_{A_j}) = E[I_{A_i} I_{A_j}] - E[I_{A_i}]E[I_{A_j}] \leq 0, \ i \neq j.$$

If $\sum_{i=1}^{\infty} P(A_i) = \infty$, *then* $P(\limsup A_i) = 1$.

Proof Let $N_n = \sum_{i=1}^{n} I_{A_i}$ be the number of the events $A_1, \ldots, A_n$ that occur, and let $N = \sum_{i=1}^{\infty} I_{A_i}$ be the total number of events that occur. Let $m_n = E[N_n] = \sum_{i=1}^{n} P(A_i)$, and note that $\lim_n m_n = \infty$. Using the formula for the variance of a sum of random variables learned in your first course in probability, we have

$$\begin{aligned}
\text{Var}(N_n) &= \sum_{i=1}^{n} \text{Var}(I_{A_i}) + 2 \sum_{i<j} \text{Cov}(I_{A_i}, I_{A_j}) \\
&\leq \sum_{i=1}^{n} \text{Var}(I_{A_i}) \\
&= \sum_{i=1}^{n} P(A_i)[1 - P(A_i)] \\
&\leq m_n.
\end{aligned}$$

Now, by Chebyshev's inequality, for any $x < m_n$

$$\begin{aligned}
P(N_n < x) &= P(m_n - N_n > m_n - x) \\
&\leq P(|N_n - m_n| > m_n - x) \\
&\leq \frac{\text{Var}(N_n)}{(m_n - x)^2} \\
&\leq \frac{m_n}{(m_n - x)^2}.
\end{aligned}$$

Hence, for any x, $\lim_{n\to\infty} P(N_n < x) = 0$. Because $P(N < x) \leq P(N_n < x)$, this implies that

$$P(N < x) = 0.$$

Consequently, by the continuity property of probabilities,

$$\begin{aligned}
0 &= \lim_{k\to\infty} P(N < k) \\
&= P\left(\lim_k \{N < k\}\right) \\
&= P(\cup_k \{N < k\}) \\
&= P(N < \infty).
\end{aligned}$$

Hence, with a probability of one, an infinite number of the events A_i occur. ■

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Example 1.48 Consider independent flips of a coin that lands heads with probability $p > 0$. For fixed k, let B_n be the event that flips $n, n+1, \ldots, n+k-1$ all land heads. Because the events $B_n, n \geq 1$, are positively correlated, we cannot directly apply the converse to the Borel–Cantelli lemma to obtain that, with a probability of 1; an infinite number of them occur. However, by letting A_n be the event that flips $nk+1, \ldots, nk+k$ all land heads, then because the set of flips these events refer to are nonoverlapping, it follows that they are independent. Because $\sum_n P(A_n) = \sum_n p^k = \infty$, we obtain from Borel–Cantelli that $P(\limsup A_n) = 1$. But $\limsup A_n \subset \limsup B_n$, so the preceding yields the result $P(\limsup B_n) = 1$. ■

Remark 1.49 The converse of the Borel–Cantelli lemma is usually stated as requiring the events $A_i, i \geq 1$, to be independent. Our weakening of this condition can be useful, as the next example shows.

Example 1.50 Consider an infinite collection of balls that are numbered $0, 1, \ldots$ and an infinite collection of boxes also numbered $0, 1, \ldots$. Suppose that ball $i, i \geq 0$, is to be put in box $i + X_i$, where $X_i, i \geq 0$, are iid with probability mass function

$$P(X_i = j) = p_j \qquad \sum_{j \geq 0} p_j = 1.$$

Suppose also that the X_i are not deterministic, so $p_j < 1$ for all $j \geq 0$. If A_j denotes the event that box j remains empty, then

$$\begin{aligned} P(A_j) &= P(X_j \neq 0, X_{j-1} \neq 1, \ldots, X_0 \neq j) \\ &= P(X_0 \neq 0, X_1 \neq 1, \ldots, X_j \neq j) \\ &\geq P(X_i \neq i, \text{for all } i \geq 0). \end{aligned}$$

But

$$\begin{aligned} &P(X_i \neq i, \text{ for all } i \geq 0) \\ &\quad = 1 - P(\cup_{i \geq 0}\{X_i = i\}) \\ &\quad = 1 - p_0 - \sum_{i \geq 1} P(X_0 \neq 0, \ldots, X_{i-1} \neq i-1, X_i = i) \\ &\quad = 1 - p_0 - \sum_{i \geq 1} p_i \prod_{j=0}^{i-1} (1 - p_j). \end{aligned}$$

Now, there is at least one pair $k < i$ such that $p_i p_k \equiv p > 0$. Hence, for that pair

$$p_i \prod_{j=0}^{i-1} (1 - p_j) \leq p_i (1 - p_k) = p_i - p,$$

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

implying that

$$P(A_j) \geq P(X_i \neq i, \text{for all } i \geq 0) \geq p > 0.$$

Hence, $\sum_j P(A_j) = \infty$. Conditional on box j being empty, each ball becomes more likely to be put in box $i, i \neq j$, so for $i < j$,

$$\begin{aligned} P(A_i|A_j) &= \prod_{k=0}^{i} P(X_k \neq i-k|A_j) \\ &= \prod_{k=0}^{i} P(X_k \neq i-k|X_k \neq j-k) \\ &\leq \prod_{k=0}^{i} P(X_k \neq i-k) \\ &= P(A_i), \end{aligned}$$

which is equivalent to $\text{Cov}(I_{A_i}, I_{A_j}) \leq 0$. Hence, by the converse of the Borel–Cantelli lemma we can conclude that, with a probability of one, there will be an infinite number of empty boxes.

We say that the sequence of random variables $X_n, n \geq 1$, *converges in probability* to the random variable X, written $X_n \longrightarrow_p X$, if for any $\epsilon > 0$

$$P(|X_n - X| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty.$$

An immediate corollary of Proposition 1.35 is that almost sure convergence implies convergence in probability. The following example shows that the converse is not true.

Example 1.51 Let $X_n, n \geq 1$ be independent random variables such that

$$P(X_n = 1) = 1/n = 1 - P(X_n = 0).$$

For any $\epsilon > 0$, $P(|X_n| > \epsilon) = 1/n \to 0$; hence, $X_n \longrightarrow_p 0$. However, because $\sum_{n=1}^{\infty} P(X_n = 1) = \infty$, it follows from the converse to the Borel–Cantelli lemma that $X_n = 1$ for infinitely many values of n, showing that the sequence does not converge almost surely to zero.

Let F_n be the distribution function of X_n, and let F be the distribution function of X. We say that X_n converges in distribution to X if

$$\lim_{n \to \infty} F_n(x) = F(x)$$

for all x at which F is continuous. (That is, convergence is required at all x for which $P(X = x) = 0$.)

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

To understand why convergence in distribution only requires that $F_n(x) \to F(x)$ at points of continuity of F, rather than at all values x, let X_n be uniformly distributed on $(0, 1/n)$. Then, it seems reasonable to suppose that X_n converges in distribution to the random variable X that is identically zero. However,

$$F_n(x) = \begin{cases} 0, & \text{if } x < 0 \\ nx, & \text{if } 0 \le x \le 1/n, \\ 1, & \text{if } x > 1/n \end{cases}$$

whereas the distribution function of X is

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \ge 0. \end{cases}$$

Thus, $\lim_n F_n(0) = 0 \neq F(0) = 1$. On the other hand, for all points of continuity of F (that is, for all $x \neq 0$), we have that $\lim_n F_n(x) = F(x)$, so with the definition given, it is indeed true that $X_n \longrightarrow_d X$.

We now show that convergence in probability implies convergence in distribution.

Proposition 1.52

$$X_n \longrightarrow_p X \quad \Rightarrow \quad X_n \longrightarrow_d X.$$

Proof Suppose that $X_n \longrightarrow_p X$. Let F_n be the distribution function of $X_n, n \ge 1$, and let F be the distribution function of X. Now, for any $\epsilon > 0$

$$\begin{aligned} F_n(x) &= P(X_n \le x,\ X \le x + \epsilon) + P(X_n \le x,\ X > x + \epsilon) \\ &\le F(x + \epsilon) + P(|X_n - X| > \epsilon), \end{aligned}$$

where the preceding used

$$X_n \le x,\ X > x + \epsilon \ \Rightarrow |X_n - X| > \epsilon.$$

Letting n go to infinity yields, upon using $X_n \longrightarrow_p X$,

$$\limsup_{n \to \infty} F_n(x) \le F(x + \epsilon). \tag{1.4}$$

Similarly,

$$\begin{aligned} F(x - \epsilon) &= P(X \le x - \epsilon,\ X_n \le x) + P(X \le x - \epsilon,\ X_n > x) \\ &\le F_n(x) + P(|X_n - X| > \epsilon). \end{aligned}$$

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Letting $n \to \infty$ gives

$$F(x-\epsilon) \leq \liminf_{n\to\infty} F_n(x). \tag{1.5}$$

Combining Equations 1.4 and 1.5 shows that

$$F(x-\epsilon) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F(x+\epsilon).$$

Letting $\epsilon \to 0$ shows that if x is a continuity point of F then

$$F(x) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F(x),$$

and the result is proved. ■

Proposition 1.53 *If* $X_n \longrightarrow_d X$, *then*

$$E[g(X_n)] \to E[g(X)]$$

for any bounded continuous function g.

To focus on the essentials, we will present a proof of Proposition 1.53 when all the random variables X_n and X are continuous. Before doing so, we will prove a couple of lemmas.

Lemma 1.54 *Let* G *be the distribution function of a continuous random variable, and let* $G^{-1}(x) \equiv \inf\{t : G(t) \geq x\}$, *be its inverse function. If* U *is a uniform* $(0,1)$ *random variable, then* $G^{-1}(U)$ *has distribution function* G.

Proof Because

$$\inf\{t : G(t) \geq U\} \leq x \Leftrightarrow G(x) \geq U$$

implies

$$P(G^{-1}(U) \leq x) = P(G(x) \geq U) = G(x),$$

we get the result. ■

Lemma 1.55 *Let* $X_n \longrightarrow_d X$, *where* X_n *is continuous with distribution function* F_n, $n \geq 1$, *and* X *is continuous with distribution function* F. *If* $F_n(x_n) \to F(x)$, *where* $0 < F(x) < 1$ *then* $x_n \to x$.

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Proof Suppose there is an $\epsilon > 0$ such that $x_n \leq x - \epsilon$ for infinitely many n. If so, then $F_n(x_n) \leq F_n(x - \epsilon)$ for infinitely many n, implying that

$$F(x) = \liminf_n F_n(x_n) \leq \lim_n F_n(x - \epsilon) = F(x - \epsilon),$$

which is a contradiction. We arrive at a similar contradiction upon assuming there is an $\epsilon > 0$ such that $x_n \geq x + \epsilon$ for infinitely many n. Consequently, we can conclude that for any $\epsilon > 0$, $|x_n - x| > \epsilon$ for only a finite number of n, thus proving the lemma. ■

Proof of Proposition 1.53 Let U be a uniform $(0, 1)$ random variable, and set $Y_n = F_n^{-1}(U)$, $n \geq 1$, and $Y = F^{-1}(U)$. Note that from Lemma 1.54 it follows that Y_n has distribution F_n and Y has distribution F. Because

$$F_n(F_n^{-1}(u)) = u = F(F^{-1}(u)),$$

it follows from Lemma 1.55 that $F_n^{-1}(u) \to F^{-1}(u)$ for all u. Thus, $Y_n \longrightarrow_{as} Y$. By continuity, this implies that $g(Y_n) \longrightarrow_{as} g(Y)$, and because g is bounded, the dominated convergence theorem yields that $E[g(Y_n)] \to E[g(Y)]$. But X_n and Y_n both have distribution F_n, whereas X and Y both have distribution F, so $E[g(Y_n)] = E[g(X_n)]$ and $E[g(Y)] = E[g(X)]$. ■

Remark 1.56 The key to our proof of Proposition 1.53 was showing that, if $X_n \longrightarrow_d X$, we can define random variables $Y_n, n \geq 1$, and Y such that Y_n has the same distribution as X_n for each n, and Y has the same distribution as X, and are such that $Y_n \longrightarrow_{as} Y$. This result (which is true without the continuity assumptions we made) is known as *Skorokhod's representation theorem.*

Skorokhod's representation and the dominated convergence theorem immediately yield the following.

Corollary 1.57 *If $X_n \longrightarrow_d X$ and there exists a constant $M < \infty$ such that $|X_n| < M$ for all n, then*

$$\lim_{n \to \infty} E[X_n] = E[X].$$

Proof Let F_n denote the distribution of X_n, $n \geq 1$, and F that of X. Let U be a uniform $(0, 1)$ random variable, and for $n \geq 1$, set $Y_n = F_n^{-1}(U)$, and $Y = F^{-1}(U)$. Note that the hypotheses of the corollary imply that $Y_n \longrightarrow_{as} Y$ and, because $F_n(M) = 1 = 1 - F_n(-M)$, also that $|Y_n| \leq M$. Thus, by the dominated convergence theorem

$$E[Y_n] \to E[Y],$$

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

which proves the result because Y_n has distribution F_n, and Y has distribution F. ■

Proposition 1.53 can also be used to give a simple proof of Weierstrass' approximation theorem.

Corollary 1.58 *Weierstrass' approximation theorem. Any continuous function f defined on the interval $[0,1]$ can be expressed as a limit of polynomial functions. Specifically, if*

$$B_n(t) = \sum_{i=0}^{n} f(i/n)\binom{n}{i} t^i (1-t)^{n-i},$$

then $f(t) = \lim_{n\to\infty} B_n(t)$.

Proof Let $X_i, i \geq 1$, be a sequence of iid random variables such that

$$P(X_i = 1) = t = 1 - P(X_i = 0).$$

Because $E[\frac{X_1+\cdots+X_n}{n}] = t$, it follows from Chebyshev's inequality that for any $\epsilon > 0$

$$P\left(\left|\frac{X_1+\cdots+X_n}{n} - t\right| > \epsilon\right) \leq \frac{\text{Var}([X_1+\cdots+X_n]/n)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2}.$$

Thus, $\frac{X_1+\cdots+X_n}{n} \to_p t$, implying that $\frac{X_1+\cdots+X_n}{n} \to_d t$. Because f is a continuous function on a closed interval, it is bounded and so Proposition 1.53 yields

$$E\left[f\left(\frac{X_1+\cdots+X_n}{n}\right)\right] \to f(t).$$

But $X_1+\cdots+X_n$ is a binomial (n,t) random variable; thus,

$$E\left[f\left(\frac{X_1+\cdots+X_n}{n}\right)\right] = B_n(t),$$

and the proof is complete. ■

1.10 Law of Large Numbers and Ergodic Theorem

Definition 1.59 *For a sequence of random variables $X_1, X_2, \ldots$ the* tail sigma field $\mathcal{T}$ *is defined as*

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \ldots).$$

Events $A \in \mathcal{T}$ are called tail events.

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Although it may seem as though there are no events remaining in the preceding intersection, there are a lot of examples of interesting tail events. Intuitively, with a tail event you can ignore any finite number of the variables and still be able to tell whether or not the event occurs. Next are some examples.

Example 1.60 Consider a sequence of random variables $X_1, X_2, ...$ having tail sigma field $\mathcal{T}$ and satisfying $|X_i| < \infty$ for all i. For the event $A_x = \{\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = x\}$, it's easy to see that $A_x \in \mathcal{T}$ because to determine if A_x happens you can ignore any finite number of the random variables; their contributions end up becoming negligible in the limit.

For the event $B_x = \{\sup_i X_i = x\}$, it's easy to see that $B_\infty \in \mathcal{T}$ because it depends on the long-run behavior of the sequence. Note that $B_7 \notin \mathcal{T}$ because it depends, for example, on whether or not $X_1 \leq 7$.

Example 1.61 Consider a sequence of random variables $X_1, X_2, ...$ having tail sigma field $\mathcal{T}$, but this time let it be possible for $X_i = \infty$ for some i. For the event $A_x = \{\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = x\}$, we now have $A_x \notin \mathcal{T}$ because any variable along the way that equals infinity will affect the limit.

Remark 1.62 The previous two examples also motivate the subtle difference between $X_i < \infty$ and $X_i < \infty$ almost surely. The former means it's impossible to see $X_5 = \infty$, and the latter only says it has probability zero. An event that has probability zero could still be a possible occurrence. For example, if X is a uniform random variable between zero and one, we can write $X \neq 0.2$ almost surely even though it is possible to see $X = 0.2$.

One approach for proving an event always happens is to first prove that its probability must either be zero or one, and then rule out zero as a possibility. This first type of result is called a *zero-one law*, because we are proving the chance must either be zero or one. A nice way to do this is to show an event A is independent of itself, and hence $P(A) = P(A \bigcap A) = P(A)P(A)$, and thus $P(A) = 0$ or 1. We use this approach next to prove a famous zero-one law for independent random variables, and we will use this in our proof of the law of large numbers.

First, we need the following definition. Events with probability either zero or one are called *trivial* events, and a sigma field is called trivial if every event in it is trivial.

Theorem 1.63 *Kolmogorov's Zero-One Law. A sequence of independent random variables has a trivial tail sigma field.*

Before we give a proof we need the following result. To show that a random variable Y is independent of an infinite sequence of random variables

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

$X_1, X_2, ...$, it suffices to show that Y is independent of $X_1, X_2, ..., X_n$ for every finite $n < \infty$. In elementary courses, this result is often given as a definition, but it can be justified using measure theory in the next proposition. We define $\sigma(X_i, i \in A) \equiv \sigma(\cup_{i \in A} \sigma(X_i))$ to be the smallest sigma field generated by the collection of random variables $X_i, i \in A$.

Proposition 1.64 *Consider the random variables Y and $X_1, X_2, ...$, where $\sigma(Y)$ is independent of $\sigma(X_1, X_2, ..., X_n)$ for every $n < \infty$. Then $\sigma(Y)$ is independent of $\sigma(X_1, X_2, ...)$.*

Before we prove this proposition, we show how this implies Kolmogorov's zero-one law

Proof *Proof of Kolmogorov's zero-one law.* We will argue that any event $A \in \mathcal{T}$ is independent of itself, and thus $P(A) = P(A \cap A) = P(A)P(A)$ and so $P(A) = 0$ or 1. Note that the tail sigma field $\mathcal{T}$ is independent of $\sigma(X_1, X_2, ..., X_n)$ for every $n < \infty$ (because $\mathcal{T} \subseteq \sigma(X_{n+1}, X_{n+2}, ...)$), so by the previous proposition, it is also independent of $\sigma(X_1, X_2, ...)$. Thus, because $\mathcal{T} \subseteq \sigma(X_1, X_2, ...)$, it also is independent of $\mathcal{T}$. ■

Now we prove the proposition.

Proof *Proof of Proposition 1.64.* Pick any $A \in \sigma(Y)$. You might at first think that $\mathcal{H} \equiv \cup_{n=1}^{\infty} \sigma(X_1, X_2, ..., X_n)$ is the same as $\mathcal{F} \equiv \sigma(X_1, X_2, ...)$, and then the theorem would follow immediately because by assumption A is independent of any event in $\mathcal{H}$. But it is not true that $\mathcal{H}$ and $\mathcal{F}$ are the same; $\mathcal{H}$ may not even be a sigma field. Also, the tail sigma field $\mathcal{T}$ is a subset of $\mathcal{F}$ but not necessarily of $\mathcal{H}$. It is, however, true that $\mathcal{F} \subseteq \sigma(\mathcal{H})$ (in fact, it turns out that $\sigma(\mathcal{H}) = \mathcal{F}$) because $\sigma(X_1, X_2, ...) \equiv \sigma(\cup_{n=1}^{\infty} \sigma(X_n))$ and $\cup_{n=1}^{\infty} \sigma(X_n) \subseteq \mathcal{H}$. We will use $\mathcal{F} \subseteq \sigma(\mathcal{H})$ later.

Define the collection of events $\mathcal{G}$ to contain any $B \in \mathcal{F}$, where for every $\epsilon > 0$ we can find a corresponding approximating event $C \in \mathcal{H}$ where $P(B \cap C^c) + P(B^c \cap C) \leq \epsilon$. Because A is independent of any event $C \in \mathcal{H}$, we can see that A must also be independent of any event $B \in \mathcal{G}$ because, using the corresponding approximating event C for any desired $\epsilon > 0$,

$$\begin{aligned} P(A \cap B) &= P(A \cap B \cap C) + P(A \cap B \cap C^c) \\ &\leq P(A \cap C) + P(B \cap C^c) \\ &\leq P(A)P(C) + \epsilon \\ &= P(A)(P(C \cap B) + P(C \cap B^c)) + \epsilon \\ &\leq P(A)P(B) + 2\epsilon \end{aligned}$$

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

and

$$\begin{aligned}
1 - P(A \cap B) &= P(A^c \cup B^c) \\
&= P(A^c) + P(A \cap B^c) \\
&= P(A^c) + P(A \cap B^c \cap C) + P(A \cap B^c \cap C^c) \\
&\leq P(A^c) + P(B^c \cap C) + P(A \cap C^c) \\
&\leq P(A^c) + \epsilon + P(A)P(C^c) \\
&= P(A^c) + \epsilon + P(A)(P(C^c \cap B) + P(C^c \cap B^c)) \\
&\leq P(A^c) + 2\epsilon + P(A)P(B^c) \\
&= 1 + 2\epsilon - P(A)P(B),
\end{aligned}$$

which when combined gives

$$P(A)P(B) - 2\epsilon \leq P(A \cap B) \leq P(A)P(B) + 2\epsilon.$$

Because ϵ is arbitrary, this shows $\sigma(Y)$ is independent of $\mathcal{G}$. We obtain the proposition by showing $\mathcal{F} \subseteq \sigma(\mathcal{H}) \subseteq \mathcal{G}$ and thus that $\sigma(Y)$ is independent of $\mathcal{F}$, as follows. First note we immediately have $\mathcal{H} \subseteq \mathcal{G}$, and thus $\sigma(\mathcal{H}) \subseteq \sigma(\mathcal{G})$, and we will be finished if we can show $\sigma(\mathcal{G}) = \mathcal{G}$.

To show that $\mathcal{G}$ is a sigma field, clearly $\Omega \in \mathcal{G}$ and $B^c \in \mathcal{G}$ whenever $B \in \mathcal{G}$. Next let $B_1, B_2, ...$ be events in $\mathcal{G}$. To show that $\cup_{i=1}^{\infty} B_i \in \mathcal{G}$, pick any $\epsilon > 0$ and let C_i be the corresponding approximating events that satisfy $P(B_i \cap C_i^c) + P(B_i^c \cap C_i) < \epsilon/2^{i+1}$. Then pick n so that

$$\sum_{i>n} P(B_i \cap B_{i-1}^c \cap B_{i-2}^c \cap \cdots \cap B_1^c) < \epsilon/2.$$

In the following, we use the approximating event $C \equiv \cup_{i=1}^{n} C_i \in \mathcal{H}$ to get

$$\begin{aligned}
&P(\cup_i B_i \cap C^c) + P((\cup_i B_i)^c \cap C) \\
&\qquad \leq P\left(\bigcup_{i=1}^{n} B_i \cap C^c\right) + \epsilon/2 + P\left(\left(\bigcup_{i=1}^{n} B_i\right)^c \cap C\right) \\
&\qquad \leq \sum_i P(B_i \cap C_i^c) + P(B_i^c \cap C_i) + \epsilon/2 \\
&\qquad \leq \sum_i \epsilon/2^{i+1} + \epsilon/2 \\
&\qquad = \epsilon,
\end{aligned}$$

and thus $\cup_{i=1}^{\infty} B_i \in \mathcal{G}$. ■

A more powerful theorem, called the extension theorem, can be used to prove Kolmogorov's zero-one law. We state it without proof.

Theorem 1.65 *The extension theorem. Suppose you have random variables $X_1, X_2, ...$, and you consistently define probabilities for all events in $\sigma(X_1, X_2, ..., X_n)$ for every n. This implies a unique value of the probability of any event in $\sigma(X_1, X_2, ...)$.*

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Remark 1.66 To see how this implies Kolmogorov's zero-one law, specify probabilities under the assumption that A is independent of any event $B \in \cup_{n=1}^{\infty}\mathcal{F}_n$. The extension theorem will say that A is independent of $\sigma(\cup_{n=1}^{\infty}\mathcal{F}_n)$.

We will prove the law of large numbers using the more powerful ergodic theorem. This means we will show that the long-run average for a sequence of random variables converges to the expected value under more general conditions then just for independent random variables. We will define these more general conditions next.

Given a sequence of random variables $X_1, X_2, \ldots$, suppose (for simplicity and without loss of generality) that there is a one-to-one correspondence between events of the form $\{X_1 = x_1, X_2 = x_2, X_3 = x_3...\}$ and elements of the sample space Ω. An event A is called an *invariant event* if the occurrence of

$$\{X_1 = x_1, X_2 = x_2, X_3 = x_3...\} \in A$$

implies both

$$\{X_1 = x_2, X_2 = x_3, X_3 = x_4...\} \in A$$

and

$$\{X_1 = x_0, X_2 = x_1, X_3 = x_2...\} \in A.$$

In other words, an invariant event is not affected by shifting the sequence of random variables to the left or right. For example, $A = \{\sup_{n\geq 1} X_n = \infty\}$ is an invariant event if $X_n < \infty$ for all n because $\sup_{n\geq 1} X_n = \infty$ implies both $\sup_{n\geq 1} X_{n+1} = \infty$ and $\sup_{n\geq 1} X_{n-1} = \infty$.

On the other hand, the event $A = \{\lim_n X_{2n} = 0\}$ is not invariant because if a sequence $x_2, x_4, ...$ converges to zero it doesn't necessarily mean that $x_1, x_3, ...$ converges to zero. Consider the example where $P(X_1 = 1) = 1/2 = 1 - P(X_1 = 0)$ and $X_n = 1 - X_{n-1}$ for $n > 1$. In this case, either $X_{2n} = 0$ and $X_{2n-1} = 1$ for all $n \geq 1$ or $X_{2n} = 1$ and $X_{2n-1} = 0$ for all $n \geq 1$, so $\{\lim_n X_{2n} = 0\}$ and $A = \{\lim_n X_{2n-1} = 0\}$ cannot occur together.

It can be shown (see Exercise 22 at the end of this chapter) that the set of invariant events makes up a sigma field, called the *invariant sigma field*, and is a subset of the tail sigma field. A sequence of random variables $X_1, X_2, ...$ is called *ergodic* if it has a trivial invariant sigma field and is called *stationary* if the random variables $(X_1, X_2, ..., X_n)$ have the same joint distribution as the random variables $(X_k, X_{k+1}, \ldots, X_{n+k-1})$ for every n, k.

We are now ready to state the ergodic theorem, and an immediate corollary will be the strong law of large numbers.

Theorem 1.67 *The ergodic theorem. If the sequence $X_1, X_2, ...$ is stationary and ergodic with $E|X_1| < \infty$, then $\frac{1}{n}\sum_{i=1}^{n} X_i \to E[X_1]$ almost surely.*

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

Because a sequence of iid random variables is clearly stationary and, by Kolmogorov's zero-one law, ergodic, we get the strong law of large numbers as an immediate corollary.

Corollary 1.68 *The strong law of large numbers. If $X_1, X_2, \ldots$ are iid with $E|X_1| < \infty$, then $\frac{1}{n}\sum_{i=1}^{n} X_i \to E[X_1]$ almost surely.*

Proof *Proof of the ergodic theorem.* Given $\varepsilon > 0$, let $Y_i = X_i - E[X_1] - \varepsilon$ and $M_n = \max(0, Y_1, Y_1 + Y_2, \ldots, Y_1 + Y_2 + \cdots + Y_n)$. Because $\frac{1}{n}\sum_{i=1}^{n} Y_i \le \frac{1}{n} M_n$, we will first show that $M_n/n \to 0$ almost surely, and then the theorem will follow after repeating the whole argument applied instead to $Y_i = -X_i + E[X_1] - \varepsilon$.

Letting $M_n' = \max(0, Y_2, Y_2 + Y_3, \ldots, Y_2 + Y_3 + \cdots + Y_{n+1})$ and using stationarity in the last equality, we have

$$\begin{aligned} E[M_{n+1}] &= E[\max(0, Y_1 + M_n')] \\ &= E[M_n' + \max(-M_n', Y_1)] \\ &= E[M_n] + E[\max(-M_n', Y_1)], \end{aligned}$$

and because $M_n \le M_{n+1}$ implies $E[M_n] \le E[M_{n+1}]$, we can conclude $E[\max(-M_n', Y_1)] \ge 0$ for all n.

Because $\{M_n/n \to 0\}$ is an invariant event, by ergodicity it must have probability either zero or one. If we were to assume the probability is zero, then $M_{n+1} \ge M_n$ would imply $M_n \to \infty$ and also $M_n' \to \infty$, and thus $\max(-M_n', Y_1) \to Y_1$. The dominated convergence theorem using the bound $|\max(-M_n', Y_1)| \le |Y_1|$ would then give $E[\max(-M_n', Y_1)] \to E[Y_1] = -\varepsilon$, which would then contradict the previous conclusion that $E[\max(-M_n', Y_1)] \ge 0$ for all n. This contradiction means we must have $M_n/n \to 0$ almost surely, and the theorem is proved. ■

1.11 Exercises

1. For $n = 1, 2, \ldots$, let $x_n = (-n)^{-n}$. What can you say about $\sup_n x_n$, $\inf_n x_n$, $\max_n x_n$, $\min x_n$, and $\lim_n x_n$?

2. Given a sigma field $\mathcal{F}$, if $A_i \in \mathcal{F}$ for all $1 \le i \le n$, is $\cap_{i=1}^{n} A_i \in \mathcal{F}$?

3. Suppose $\mathcal{F}_i, i = 1, 2, 3, \ldots$ are sigma fields. (a) Is $\cap_{i=1}^{\infty} \mathcal{F}_i$ necessarily always a sigma field? Explain. (b) Does your reasoning in (a) also apply to the intersection of an uncountable number of sigma fields? (c) Is $\cup_{i=1}^{\infty} \mathcal{F}_i$ necessarily always a σ field? Explain.

4. (a) Suppose $\Omega = \{1, 2, \ldots, n\}$. How many different sets will there be in the sigma field generated by starting with the individual elements in Ω? (b) Is it possible for a sigma field to have a countably infinite number of different sets in it? Explain.

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

5. Show that if X and Y are real-valued random variables measurable with respect to some given sigma field, then so is XY with respect to the same sigma field.

6. If X is a random variable, is it possible for the cumulative distribution function (CDF) $F(x) = P(X \leq x)$ to be discontinuous at a countably infinite number of values of x? Is it possible for it to be discontinuous at an uncountably infinite number of values of x? Explain.

7. Show that $E[X] = \sum_i x_i P(X = x_i)$ if X can only take a countably infinite number of different possible values.

8. Prove that if $X \geq 0$ and $E[X] < \infty$, then $\lim_{n\to\infty} E[XI_{X>n}] = 0$.

9. Assume $X \geq 0$ is a random variable, but don't necessarily assume that $E[1/X] < \infty$. Show that $\lim_{n\to\infty} E[\frac{n}{X} I_{X>n}] = 0$ and $\lim_{n\to\infty} E[\frac{1}{nX} I_{X>n}] = 0$.

10. Use the definition of expected value in terms of simple variables to prove that if $X \geq 0$ and $E[X] = 0$ then $X = 0$ almost surely.

11. Show that if $X_n \longrightarrow_d c$ then $X_n \longrightarrow_p c$.

12. Show that if $E[g(X_n)] \to E[g(X)]$ for all bounded, continuous functions g then $X_n \longrightarrow_d X$.

13. If $X_1, X_2, \ldots$ are nonnegative random variables with the same distribution (but the variables are not necessarily independent) and $E[X_1] < \infty$, prove that $\lim_{n\to\infty} E[\max_{i<n} X_i/n] = 0$.

14. For random variables $X_1, X_2, ...$, let $\mathcal{T}$ be the tail sigma field, and let $S_n = \sum_{i=1}^n X_i$. (a) Is $\{\lim_{n\to\infty} S_n/n > 0\} \in \mathcal{T}$? (b) Is $\{\lim_{n\to\infty} S_n > 0\} \in \mathcal{T}$?

15. If $X_1, X_2, ...$ are nonnegative iid random variables with $P(X_i > 0) > 0$, show that $P(\sum_{i=1}^\infty X_i = \infty) = 1$.

16. Suppose $X_1, X_2, ...$ are continuous iid random variables and

$$Y_n = I_{\{X_n > \max_{i<n} X_i\}}.$$

(a) Argue that Y_i is independent of Y_j for $i \neq j$. (b) What is $P(\sum_{i=1}^\infty Y_i < \infty)$? (c) What is $P(\sum_{i=1}^\infty Y_i Y_{i+1} < \infty)$?

17. Suppose there is a single server and the ith customer to arrive requires the server spend U_i time serving them, the time between their arrival and the next customer's arrival is V_i, and $X_i = U_i - V_i$ are iid with mean μ. (a) If Q_{n+1} is the amount of time the $(n+1)$ customer must wait before being served, explain why $Q_{n+1} = \max(Q_n + X_n, 0)$ $= \max(0, X_n, X_n + X_{n-1}, ..., X_n + \cdots + X_1)$. (b) Show $P(Q_n \to \infty) = 1$ if $\mu > 0$.

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press

18. Given a nonnegative random variable X, define the sequence of random variables $Y_n = \min(\lfloor 2^n X \rfloor / 2^n, n)$, where $\lfloor x \rfloor$ denotes the integer portion of x. Show that $Y_n \uparrow X$ and $E[X] = \lim_n E[Y_n]$.

19. Show that for any monotone functions f and g if X, Y are independent random variables then so are $f(X), g(Y)$.

20. Let $X_1, X_2, \ldots$ be random variables with $X_i < \infty$ and suppose $\sum_n P(X_n > 1) < \infty$. Compute $P(\sup_n X_n < \infty)$.

21. Suppose $X_n \to_p X$ and that there is a random variable Y with $E[Y] < \infty$ such that $|X_n| < Y$ for all n. Show $E[\lim_{n\to\infty} X_n] = \lim_{n\to\infty} E[X_n]$.

22. For random variables $X_1, X_2, \ldots$, let $\mathcal{T}$ and $\mathcal{I}$ be the set of tail events and the set of invariant events, respectively. Show that $\mathcal{I}$ and $\mathcal{T}$ are both sigma fields.

23. A ring is hanging from the ceiling by a string. Someone will cut the ring in two positions chosen uniformly at random on the circumference, and this will break the ring into two pieces. Player I gets the piece that falls to the floor, and player II gets the piece that stays attached to the string. Whoever gets the bigger piece wins. Does either player have an advantage here? Explain.

24. A box contains four marbles. One marble is red, and each of the other three marbles is either yellow or green, but you have no idea exactly how many of each color there are or if the other three marbles are all the same color or not. (a) Someone chooses one marble at random from the box, and if you can correctly guess the color, you will win \$1,000. What color would you guess? Explain. (b) If this game is to be played four times using the same box of marbles (and the marble drawn each time is placed back in the box), what guesses would you make if you had to make all four guesses ahead of time? Explain.

25. For a sequence of iid continuous random variables $X_1, X_2, \ldots$, let $N = \inf\{n \geq 2 : X_{n+1} > X_n\}$ be the first time the next variable is larger than its immediate predecessor. Compute $E[N]$.

26. Is it possible to pick a random positive integer uniformly at random? Is it possible to pick a positive real number uniformly at random? Explain why or why not.

27. In a group of n people, what is the expected number of distinct birthdays?

28. If a fair coin is flipped n times, what is the expected number of runs of k heads in a row if overlapping runs are each counted separately? What is the expected number of times a run of at least k heads appears in n flips, without counting overlapping runs?

https://doi.org/10.1017/9781009179928.002 Published online by Cambridge University Press