# Math 6810
# (Probability and Fractals)

## Spring 2016

## Lecture notes

Pieter Allaart
University of North Texas

February 2, 2016

**Recommended reading:** (Do not purchase these books before consulting with your instructor!)

1. *Real Analysis* by H. L. Royden (4th edition), Prentice Hall.
2. *Probability and Measure* by P. Billingsley (3rd edition), Wiley.
3. *Probability with Martingales* by D. Williams, Cambridge University Press.
4. *Fractal Geometry: Foundations and Applications* by K. Falconer (2nd edition), Wiley.

# Chapter 1

# Preliminaries

## 1.1 Finite probability spaces

**Definition 1.1.** A *sample space* is a finite set $\Omega$ of objects (thought of as possible outcomes of an experiment). An *event* is a subset $A$ of $\Omega$.

**Example 1.2.** (1) Rolling a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$

(2) Flipping a coin 3 times: $\Omega = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}$

**Definition 1.3.** Let $2^\Omega$ denote the power set of $\Omega$ (set of all subsets of $\Omega$). A *probability measure* on $\Omega$ is a function $\mathrm{P} : 2^\Omega \to [0, 1]$ satisfying:

(i) $\mathrm{P}(\Omega) = 1$, and

(ii) $\mathrm{P}(\bigcup_{k=1}^n A_k) = \sum_{k=1}^n \mathrm{P}(A_k)$ whenever $A_1, \ldots, A_n$ are disjoint subsets of $\Omega$.

The tuple $(\Omega, \mathrm{P})$ is called a (finite) *probability space*.

**Example 1.4.** Most common: equally likely outcomes

$$\mathrm{P}(A) = \frac{\#A}{\#\Omega}, \qquad A \subset \Omega$$

In Example 1.2,(1):

$$\mathrm{P}(\text{roll an even number}) = 3/6 = 1/2$$

In Example 1.2,(2): assuming the coin is fair,

$$\mathrm{P}(\text{exactly two of the coin flips land heads}) = \frac{\#\{hht, hth, thh\}}{\#\Omega} = \frac{3}{8}$$

**Definition 1.5.** Let $A$ and $B$ be events in a probability space $(\Omega, \mathrm{P})$ with $\mathrm{P}(B) > 0$. The *conditional probability of $A$ given $B$* is defined by

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

**Definition 1.6.** Events $A$ and $B$ in a probability space $(\Omega, \mathrm{P})$ are *independent* if $P(A|B) = P(A)$.

Easy to check:

$$\mathrm{P}(A|B) = P(A) \quad \Leftrightarrow \quad \mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B) \quad \Leftrightarrow \quad \mathrm{P}(B|A) = \mathrm{P}(B).$$

**Definition 1.7.** Events $A_1, \ldots, A_n$ are independent iff

$$\mathrm{P}(C_1 \cap \cdots \cap C_n) = \mathrm{P}(C_1) \cdots \mathrm{P}(C_n)$$

for any choice of $C_1, \ldots, C_n$ where for each $i$, $C_i$ is either $A_i$ or $A_i^c$.

**Definition 1.8.** Events $A_1, \ldots, A_n$ are *pairwise independent* iff $\mathrm{P}(A_i \cap A_j) = \mathrm{P}(A_i)\,\mathrm{P}(A_j)$ for all $i \neq j$.

**Example 1.9.** Let $\Omega = \{1, 2, 3, 4\}$ with all outcomes equally likely, let $A = \{1, 2\}$, $B = \{1, 3\}$, and $C = \{1, 4\}$. What does this example illustrate?

**Definition 1.10.** Events $B_1, \ldots, B_n$ are said to be a *partition* of $\Omega$ if $\bigcup_{i=1}^n B_i = \Omega$ and $B_i \cap B_j = \emptyset$ for all $i \neq j$.

**Proposition 1.11** (Principle of conditioning)**.** *If $B_1, \ldots, B_n$ is a partition of $\Omega$ with $\mathrm{P}(B_i) > 0$ for each $i$, then for any $A \subset \Omega$,*

$$\mathrm{P}(A) = \sum_{i=1}^n \mathrm{P}(A|B_i)\,\mathrm{P}(B_i).$$

*Proof.*

$$\begin{aligned}\mathrm{P}(A) &= \mathrm{P}\left(A \cap \left(\bigcup_{i=1}^n B_i\right)\right) = \mathrm{P}\left(\bigcup_{i=1}^n (A \cap B_i)\right) \\ &= \sum_{i=1}^n \mathrm{P}(A \cap B_i) = \sum_{i=1}^n \mathrm{P}(A|B_i)\,\mathrm{P}(B_i).\end{aligned}$$

□

**Definition 1.12.** A *random variable* is a function $X : \Omega \to \mathbb{R}$. Instead of $\mathrm{P}(\{\omega \in \Omega : X(\omega) = x\})$ we write simply: $\mathrm{P}(X = x)$. The *(probability) distribution* of $X$ is the list of values $\mathrm{P}(X = x)$, where $x \in X(\Omega)$. When $X$ and $Y$ have the same distribution, we express this by $X \stackrel{d}{=} Y$.

**Example 1.13.** (1) Roll two dice, and let $X$ be the sum:

$$\Omega = \{(i, j) : i, j \in \{1, 2, \ldots, 6\}\} = \{1, 2, \ldots, 6\}^2, \qquad X(i, j) = i + j.$$

Then, for instance, $\mathrm{P}(X = 9) = \mathrm{P}(\{(4, 5), (5, 4)\}) = 2/36 = 1/18$, etc.

(2) Flip a fair coin 3 times, and let $X$ denote the number of heads:

$$\mathrm{P}(X=i)=\begin{cases}1/8, & i=0\\ 3/8, & i=1\\ 3/8, & i=2\\ 1/8, & i=3\end{cases}$$

**Remark 1.14.** Note that the r.v. $X$ in the last example determines another probability space $(\tilde{\Omega},\tilde{P})$, where $\tilde{\Omega}=\{0,1,2,3\}$ and $\tilde{P}(\omega)=1/8$ for $\omega\in\{0,3\}$, $\tilde{P}(\omega)=3/8$ for $\omega\in\{1,2\}$. This probability space, however, contains "less information" than the space $\Omega$ of Example 1.2,(2).

If on this space we define $\tilde{X}(\omega)=\omega$, then $\tilde{X}$ has the same distribution as $X$, denoted $\tilde{X}\overset{d}{=}X$.

**Definition 1.15.** Random variables $X_1,\dots,X_n$ are *independent* if for any $n$-tuple $(x_1,\dots,x_n)$ in $\mathbb{R}^n$, the events $\{X_1=x_1\},\dots,\{X_n=x_n\}$ are independent.

**Definition 1.16.** The *expectation* or *expected value* of a r.v. $X$ is defined by

$$\mathrm{E}(X)=\sum_{x\in\mathbb{R}} x\,\mathrm{P}(X=x).$$

It is easy to see that

$$\mathrm{E}(X)=\sum_{\omega\in\Omega} X(\omega)\,\mathrm{P}(\omega). \tag{1.1}$$

**Example 1.17.** The expectations of the random variables in Example 1.13 are 7 and 3/2, respectively.

**Definition 1.18.** Let $\mathrm{P}(B)>0$. The *conditional expectation of $X$ given $B$* is

$$\mathrm{E}(X|B)=\sum_{x\in\mathbb{R}} x\,\mathrm{P}(X=x|B).$$

**Proposition 1.19** (Computing expectation by conditioning)**.** *If $B_1,\dots,B_n$ is a partition of $\Omega$ with $\mathrm{P}(B_i)>0$ for each $i$, then*

$$\mathrm{E}(X)=\sum_{i=1}^{n}\mathrm{E}(X|B_i)\,\mathrm{P}(B_i). \tag{1.2}$$

**Example 1.20.** *A fair die is rolled, and whichever number comes up, a fair coin is then flipped that many times. Let $N$ be the outcome of the die roll, and $X$ the number of heads obtained. Then*

$$\mathrm{E}(X)=\sum_{i=1}^{6}\mathrm{E}(X|N=i)\,\mathrm{P}(N=i)=\sum_{i=1}^{6}\frac{i}{2}\cdot\frac{1}{6}=\frac{21}{12}=\frac{7}{4}.$$

**Proposition 1.21.** *(i) For a function $h : \mathbb{R} \to \mathbb{R}$,*

$$\mathrm{E}(h(X)) = \sum_{x \in \mathbb{R}} h(x)\, \mathrm{P}(X = x).$$

*(ii) For any two r.v.'s $X$ and $Y$ and constants $a$ and $b$,*

$$\mathrm{E}(aX + bY) = a\, \mathrm{E}(X) + b\, \mathrm{E}(Y). \tag{1.3}$$

*Proof.* Both statements follow directly from (1.1). □

**Definition 1.22.** The *variance* of a r.v. $X$ is

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

Using (1.3), we can derive the shortcut formula:

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2.$$

Note that $\mathrm{Var}(X) \geq 0$.

**Proposition 1.23.** *(i) For any function $h : \mathbb{R}^2 \to \mathbb{R}$ and r.v.'s $X$ and $Y$,*

$$\mathrm{E}[h(X, Y)] = \sum_x \sum_y h(x, y)\, \mathrm{P}(X = x, Y = y).$$

*(ii) If $X$ and $Y$ are independent, then $\mathrm{E}(XY) = \mathrm{E}(X)\, \mathrm{E}(Y)$.*

*(iii) If $X_1, \ldots, X_n$ are independent r.v.'s, then*

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n).$$

*Proof.* (i) Exercise
(ii) By (i),

$$\begin{aligned}
\mathrm{E}(XY) &= \sum_x \sum_y xy\, \mathrm{P}(X = x, Y = y) \\
&= \sum_x \sum_y xy\, \mathrm{P}(X = x)\, \mathrm{P}(Y = y) \\
&= \left( \sum_x x\, \mathrm{P}(X = x) \right) \left( \sum_y y\, \mathrm{P}(Y = y) \right) \\
&= \mathrm{E}(X)\, \mathrm{E}(Y).
\end{aligned}$$

We prove (iii) for $n = 2$. The general result then follows by induction. Let $X$ and $Y$ be independent r.v.'s. Then by (ii),

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathrm{E}[(X+Y)^2] - [\mathrm{E}(X+Y)]^2 \\ &= \mathrm{E}(X^2 + 2XY + Y^2) - [E(X) + E(Y)]^2 \\ &= \mathrm{E}(X^2) + 2\,\mathrm{E}(X)\,\mathrm{E}(Y) + \mathrm{E}(Y^2) - [(\mathrm{E}(X))^2 + 2\,\mathrm{E}(X)\,\mathrm{E}(Y) + (\mathrm{E}(Y))^2] \\ &= \mathrm{E}(X^2) - (\mathrm{E}(X))^2 + \mathrm{E}(Y^2) - (\mathrm{E}(Y))^2 \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y).\end{aligned}$$

□

**Example 1.24** (Binomial distribution)**.** A r.v. $X$ is said to have a *binomial*$(n, p)$ distribution if

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

where $0 \le p \le 1$. We can construct a binomial r.v. on the probability space $\{0,1\}^n$ with the probability measure P determined by

$$\mathrm{P}(\omega) = \mathrm{P}(\{(\omega_1, \ldots, \omega_n)\}) = p^{k(\omega)}(1-p)^{n-k(\omega)},$$

where $k(\omega) = \#\{i : 1 \le i \le n, \omega_i = 1\}$. Then put $X(\omega) = \sum_{i=1}^n \omega_i$.

This construction has the advantage that we can easily calculate the expectation and variance of $X$, as follows. Let $X_i(\omega) = \omega_i$ so that $X = X_1 + \cdots + X_n$. One checks that

$$\mathrm{P}(X_i = 1) = p, \qquad \mathrm{P}(X_i = 0) = 1 - p,$$

so that $\mathrm{E}(X_i) = p$ and $\mathrm{E}(X_i^2) = p$, whence $\mathrm{Var}(X_i) = p(1-p)$, for $i = 1, \ldots, n$. Thus, by (1.3) and part (iii) of the last proposition,

$$\mathrm{E}(X) = \mathrm{E}(X_1 + \cdots + X_n) = \sum_{i=1}^n \mathrm{E}(X_i) = np,$$

and

$$\mathrm{Var}(X) = \mathrm{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^n \mathrm{Var}(X_i) = np(1-p).$$

Compare this with calculating the summations

$$\mathrm{E}(X) = \sum_{k=1}^n k \binom{n}{k} p^k (1-p)^{n-k}, \qquad \mathrm{E}(X^2) = \sum_{k=1}^n k^2 \binom{n}{k} p^k (1-p)^{n-k}.$$

Note that the distribution, and hence the expectation and variance, of a random variable do not depend on the underlying probability space. Constructing a random variable (or a whole process) on a carefully chosen probability space can greatly facilitate proving distributional properties, as we'll see again later.

## 1.2 Infinite probability spaces

**Note:** Finite probability spaces can only describe finite experiments, and are inadequate for modern probability. It is necessary to employ infinite sample spaces such as $[0,1]$, $\{0,1\}^{\mathbb{N}}$ (the space of all sequences of 0's and 1's), $C([0,1])$ (the space of continuous functions on $[0,1]$), etc. One technical problem with infinite sample spaces is, that it is impossible to define a meaningful probability measure on them which assigns a probability to *every* subset of $\Omega$. Therefore, it is necessary to restrict attention to certain subcollections of sets. These subcollections are called $\sigma$-algebras.

**Definition 1.25.** Let $X$ be any set. A *$\sigma$-algebra* on $X$ is a collection $\mathcal{F}$ of subsets of $X$ such that:

(i) $\emptyset \in \mathcal{F}$;

(ii) if $A \in \mathcal{F}$, then $A^c = X \backslash A \in \mathcal{F}$; and

(iii) if $A_1, A_2, \ldots$ are in $\mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Thus, a $\sigma$-algebra is a collection of subsets of $X$ which includes the empty set, and is closed under complements and countable unions.

**Example 1.26.** (1) $\mathcal{F} = 2^{\Omega}$, the power set of $\Omega$, is a $\sigma$-algebra.

(2) $\mathcal{F} = \{\emptyset, \Omega\}$ is a $\sigma$-algebra, called the *trivial $\sigma$-algebra*.

(3) Let $X$ be a topological space, and $\mathcal{O}$ the collection of all open sets in $X$. The smallest $\sigma$-algebra that contains $\mathcal{O}$ is called the *Borel $\sigma$-algebra* on $X$, denoted $\mathcal{B}(X)$, and its members are called *Borel sets* in $X$. (Note: $\mathcal{B}(X)$ is well defined, as it is easy to check that the intersection of any family of $\sigma$-algebras is again a $\sigma$-algebra, so we can define $\mathcal{B}(X)$ as the intersection of all those $\sigma$-algebras on $X$ which contain $\mathcal{O}$.)

**Definition 1.27.** Let $\mathcal{F}$ be a $\sigma$-algebra on a set $\Omega$. A *probability measure* on $(\Omega, \mathcal{F})$ is a set function $\mathrm{P} : \mathcal{F} \to [0,1]$ satisfying:

(i) $\mathrm{P}(\Omega) = 1$, and

(ii) $\mathrm{P}(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} \mathrm{P}(A_k)$ whenever $A_1, A_2, \ldots$ is a sequence of disjoint subsets of $\Omega$.

The triple $(\Omega, \mathcal{F}, \mathrm{P})$ is called a *probability space*. By an *event* we mean a set $A \in \mathcal{F}$.

Note this definition encompasses the definition of probability measure in the finite case, where we simply took the $\sigma$-algebra $\mathcal{F} = 2^{\Omega}$. This is however no longer possible in the case of $\Omega = [0,1]$, say, because there does not exist any probability measure which assigns a probability to every subset of $[0,1]$ and which gives measure zero to singletons. (This is a consequence of the existence of "nonmeasurable sets"!)

### 1.2.1 Lebesgue measure

One of the most useful probability spaces is the unit interval $[0,1]$ with Borel sets and Lebesgue measure. Lebesgue measure generalizes the concept of length of an interval. Carathéodory's extension theorem tells us that this can be done in a unique way to obtain a probability measure on the Borel sets of $[0,1]$ that assigns to each interval its length. First, we need two definitions.

**Definition 1.28.** An *algebra* in a set $\Omega$ is a collection $\mathcal{A}$ of subsets of $\Omega$ satisfying:

(i) $\emptyset \in \mathcal{A}$;

(ii) if $A \in \mathcal{A}$, then $A^c = X \backslash A \in \mathcal{A}$; and

(iii) if $A_1, \ldots, A_n$ are in $\mathcal{A}$, then $\bigcup_{i=1}^n A_i \in \mathcal{A}$.

Thus, a $\sigma$-algebra is a collection of subsets of $X$ which includes the empty set, and is closed under complements and *finite* unions.

**Definition 1.29.** A *probability measure on an algebra* $\mathcal{A}$ is a set function $\mathrm{P} : \mathcal{A} \to [0,1]$ such that

(i) $\mathrm{P}(\Omega) = 1$; and

(ii) if $A_1, A_2, \ldots$ are disjoint sets in $\mathcal{A}$ for which $\bigcup_{k=1}^\infty A_k \in \mathcal{A}$, then $\mathrm{P}(\bigcup_{k=1}^\infty A_k) = \sum_{k=1}^\infty \mathrm{P}(A_k)$.

**Example 1.30.** Let $\Omega = [0,1]$ and let $\mathcal{A}$ be the collection of all finite unions of subintervals of $\Omega$, where we interpret an interval of the form $[a,a]$ as a single point $\{a\}$. Then $\mathcal{A}$ is an algebra, and we can define a probability measure P on $\mathcal{A}$, called *pre-Lebesgue measure*, by

$$\mathrm{P}\left(\bigcup_{i=1}^n I_i\right) = \sum_{i=1}^n \ell(I_i),$$

for any finite collection $\{I_i\}$ of disjoint intervals in $[0,1]$, where $\ell(I)$ denotes the length of the interval $I$.

**Theorem 1.31** (Carathéodory's extension theorem)**.** *Let* P *be a probability measure on an algebra* $\mathcal{A}$*. Then* P *has a unique extension to a probability measure on the smallest* $\sigma$*-algebra containing* $\mathcal{A}$ *(which we again denote by* P*).*

**Corollary 1.32.** *There is a unique probability measure* P *on* $([0,1], \mathcal{B}([0,1]))$ *such that* $\mathrm{P}(I) = \ell(I)$ *for any subinterval* $I$ *of* $[0,1]$.

**Definition 1.33.** The probability measure of Corollary 1.32 is called *Lebesgue measure* on $[0,1]$.

**Theorem 1.34.** *Lebesgue measure* P *has the following properties:*

*(1) (Monotonicity) For any Borel sets $A$ and $B$ with $A \subset B$,* $\mathrm{P}(A) \leq \mathrm{P}(B)$.

*(2) (Regularity) For each Borel set $A \subset [0,1]$,*

$$\mathrm{P}(A) = \inf\left\{\sum_{i=1}^{\infty} \ell(I_i) : \{I_i\} \text{ are intervals with } A \subset \bigcup_{i=1}^{\infty} I_i\right\},$$
$$\mathrm{P}(A) = \inf\{\mathrm{P}(O) : O \text{ is an open set in } [0,1] \text{ and } A \subset O\},$$
$$\mathrm{P}(A) = \sup\{\mathrm{P}(F) : F \text{ is a closed set in } [0,1] \text{ and } F \subset A\}.$$

*(3) (Continuity)*

*(i) If $\{A_n\}$ is a sequence of Borel sets in $[0,1]$ and $A_{n+1} \subset A_n$ for each $n$, then*

$$\mathrm{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathrm{P}(A_n).$$

*(ii) If $\{A_n\}$ is a sequence of Borel sets in $[0,1]$ and $A_{n+1} \supset A_n$ for each $n$, then*

$$\mathrm{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathrm{P}(A_n).$$

*In fact, monotonicity and continuity hold for* any *probability measure.*

Note that we can define Lebesgue measure on any subinterval of $\mathbb{R}$, including $\mathbb{R}$ itself, by the same process. Lebesgue measure on $\mathbb{R}$ will be denoted by $\lambda$. It is of course not a probability measure.

### 1.2.2 Random variables

**Definition 1.35.** A function $f : \Omega \to \mathbb{R}$ is *$\mathcal{F}$-measurable* if for each $a \in \mathbb{R}$, the set $\{\omega \in \Omega : f(\omega) \leq a\}$ belongs to $\mathcal{F}$. A *random variable* on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is an $\mathcal{F}$-measurable function $X : \Omega \to \mathbb{R}$.

A standard exercise in measure theory shows that if $X$ is a random variable, the set $\{\omega : X(\omega) \in B\}$ is in $\mathcal{F}$ for every Borel set $B$. Instead of $\mathrm{P}(\{\omega : X(\omega) \in B\})$ we will write simply $\mathrm{P}(X \in B)$.

**Example 1.36.** Let $\Omega = [0,1]$ and $\mathcal{F} = \mathcal{B}([0,1])$, and put $X(\omega) = \omega$. Then $\mathrm{P}(X \in I) = \ell(I)$ for every interval $I \subset [0,1]$. We say $X$ has the *standard uniform distribution.* If $a < b$, we can define a new r.v. $Y$ by $Y = a + (b-a)X$. Then

$$\mathrm{P}(c < Y \leq d) = \mathrm{P}\left(\frac{c-a}{b-a} < X \leq \frac{d-a}{b-a}\right) = \frac{d-c}{b-a}$$

whenever $a \leq c < d \leq b$. We say $Y$ has the *uniform$(a,b)$ distribution.*

Independence of events is defined in the same way as in the finite case. Independence of random variables, however, requires a somewhat more careful definition.

**Definition 1.37.** Random variables $X_1, \ldots, X_n$ are *independent* if for any choice of Borel sets $B_1, \ldots, B_n$ in $\mathbb{R}$, the events $\{X_1 \in B_1\}, \ldots, \{X_n \in B_n\}$ are independent. An infinite sequence of random variables $X_1, X_2, \ldots$ is said to be independent if $X_1, \ldots, X_n$ are independent for every $n$.

**Proposition 1.38.** *Random variables $X_1, \ldots, X_n$ are independent if and only if the sets $\{X_1 \le x_1\}, \ldots, \{X_n \le x_n\}$ are independent for any choice of real numbers $x_1, \ldots, x_n$.*

**Example 1.39** (Independent coin tosses)**.** On the space $[0, 1]$ with Lebesgue measure, we can construct independent $\{0, 1\}$-valued r.v.'s as follows. Each number $\omega \in [0, 1]$ has a binary expansion

$$\omega = 0.d_1 d_2 \ldots d_n \cdots = \sum_{i=1}^{\infty} 2^{-i} d_i,$$

where $d_i := d_i(\omega) \in \{0, 1\}$ for each $i$. Numbers of the form $k/2^n$, such as $1/2, 1/4, 3/4$, etc., have two binary expansions, e.g.

$$1/2 = 0.1000 \cdots = 0.0111 \cdots .$$

For such numbers, for definiteness, we choose the expansion ending in all zeros. Exception: for $\omega = 1$ we write $1 = 0.111 \cdots$.

Now the $d_i$'s are measurable functions because, for each $i$, the set $\{\omega : d_i(\omega) = 0\}$ is a finite union of intervals, e.g. $\{\omega : d_2(\omega) = 0\} = [0, 1/4) \cup [1/2, 3/4)$. Thus each $d_i$ is a random variable, and $\mathrm{P}(d_i = 1) = \mathrm{P}(d_i = 0) = 1/2$ (check!). Therefore we can think of $d_i$ as the outcome of a fair coin toss, where 1 represents heads, and 0 represents tails. The sequence $\{d_i\}$ is independent: for any given $n$, let $b_1, \ldots, b_n$ be an arbitrary finite sequence of 0's and 1's, and put

$$x_0 := \sum_{i=1}^{n} 2^{-i} b_i.$$

Then

$$\{\omega : d_1(\omega) = b_1, \ldots, d_n(\omega) = b_n\} = [x_0, x_0 + 2^{-n}],$$

and hence,

$$\mathrm{P}(d_1 = b_1, \ldots, d_n = b_n) = \ell([x_0, x_0 + 2^{-n}]) = 2^{-n} = \mathrm{P}(d_1 = b_1) \cdots \mathrm{P}(d_n = b_n).$$

**Definition 1.40.** The *(cumulative) distribution function* (c.d.f. for short) of a r.v. $X$ is the function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) := \mathrm{P}(X \le x), \qquad x \in \mathbb{R}.$$

**Proposition 1.41.** *The c.d.f. $F$ of any r.v. $X$ has the following properties:*

*(i)* $F$ *is nondecreasing.*

*(ii)* $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$.

*(iii)* $F$ *is* right-continuous*:* $\lim_{x\downarrow x_0} F(x) = F(x_0)$ *for every* $x_0 \in \mathbb{R}$.

*Vice versa, any function* $F$ *with properties (i)-(iii) above is the c.d.f. of some r.v.* $X$.

*Proof.* Statement (i) follows immediately from the monotonicity of P. Statements (ii) and (iii) are consequences of the continuity of P. For the first limit in (ii), set $A_n = \{\omega : X(\omega) \le -n\}$. Then $A_{n+1} \subset A_n$, and $\bigcap_{n=1}^{\infty} A_n = \emptyset$, so

$$\lim_{n\to\infty} F(-n) = \lim_{n\to\infty} \mathrm{P}(A_n) = \mathrm{P}(\emptyset) = 0,$$

and then, since $F$ is nondecreasing, $\lim_{x\to-\infty} F(x) = 0$ as well. The second limit in (ii) follows similarly. For (iii), let $A_n = \{\omega : X(\omega) \le x_0 + 1/n\}$. Then again $A_{n+1} \subset A_n$, and $\bigcap_{n=1}^{\infty} A_n = \{\omega : X(\omega) \le x_0\}$, from which the result follows.

The last statement of the proposition is proved in the next theorem. □

**Notation:** For a c.d.f. $F$, let

$$F^{-1}(y) := \inf\{x : F(x) \ge y\}, \qquad y \in [0, 1). \tag{1.4}$$

We call $F^{-1}$ the *generalized inverse* of $F$.

**Theorem 1.42** (Construction of a r.v. with given c.d.f.)**.** *Let* $F : \mathbb{R} \to [0, 1]$ *satisfy (i)-(iii) of Proposition 1.41. On the Lebesgue space* $([0, 1], \mathcal{B}([0, 1]), \mathrm{P})$*, put* $X(\omega) = F^{-1}(\omega)$*. Then* $X$ *is a r.v. with c.d.f.* $F$.

*Proof.* We show that

$$F^{-1}(\omega) \le x \quad \Longleftrightarrow \quad \omega \le F(x). \tag{1.5}$$

Suppose $F^{-1}(\omega) \le x$. Since $F$ is nondecreasing, this means that for every $\varepsilon > 0$, $F(x+\varepsilon) \ge \omega$. But then $F(x) \ge \omega$, since $F$ is right continuous. The other direction is obvious. By (1.5), we have

$$\mathrm{P}(\omega : X(\omega) \le x) = \mathrm{P}(\omega : F^{-1}(\omega) \le x) = \mathrm{P}(\omega : \omega \le F(x)) = F(x),$$

as required. □

Note that any random variable $X$ determines a unique Borel probability measure $\mu$ on $\mathbb{R}$ which satisfies

$$\mu(B) = \mathrm{P}(X \in B), \qquad B \in \mathcal{B}(\mathbb{R}).$$

Taking $B = (a, b]$ we have in particular,

$$\mu((a, b]) = \mathrm{P}(a < X \le b) = F(b) - F(a), \qquad a < b.$$

We call $\mu$ the *distribution* of $X$, and sometimes write $\mu_X$ when the r.v. $X$ needs to be made explicit. We also sometimes write $\mu = \mathrm{P}\, X^{-1}$.

Observe that an alternative way to construct a random variable $X$ with given distribution $\mu_X$ is to take $X(\omega) = \omega$ on the space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_X)$.

**Theorem 1.43.** *Given any sequence $\mu_1, \mu_2, \ldots$ of Borel probability measures on $\mathbb{R}$, there exists, on the probability space $([0,1], \mathcal{B}([0,1]), \mathrm{P})$, a sequence of independent random variables $X_1, X_2, \ldots$, such that for each $i$ the distribution of $X_i$ is $\mu_i$.*

*Proof.* Let $d_1, d_2, \ldots$ be the independent $\{0,1\}$ random variables from Example 1.39. We define random variables $U_i$ ($i \in \mathbb{N}$) via their binary expansions as follows:

$$\begin{aligned} U_1 &:= 0.d_2 d_4 d_8 \cdots \\ U_2 &:= 0.d_3 d_6 d_{12} \cdots \\ U_3 &:= 0.d_5 d_{10} d_{20} \cdots \\ &\vdots \end{aligned}$$

Precisely, let $p_i$ denote the $i$th prime number, and set

$$U_i := \sum_{j=1}^{\infty} 2^{-j} d_{2^{j-1}p_i}, \qquad i = 1, 2, \ldots.$$

Now if $i \neq j$, then $U_i$ and $U_j$ depend on disjoint subsequences of $\{d_k\}$. It follows that $U_1, U_2, \ldots$ are independent. We next show that they have the standard uniform distribution. We show it for $U_1$; the same argument works for any $U_i$. The binary digits of $U_1$ are $d_2, d_4, d_8, \ldots$, which are independent r.v.'s all with the same distribution as $d_1$, so the sequence $(d_2, d_4, d_8, \ldots)$ has the same joint distribution as the sequence $(d_1, d_2, d_3, \ldots)$. Hence, for any dyadic interval

$$I = [j/2^k, (j+1)/2^k), \tag{1.6}$$

we have

$$\mathrm{P}(U_1 \in I) = \ell(I) = 2^{-k},$$

as in Example 1.39. But any interval $[0, x)$ in $[0,1]$ can be written as a countably union of disjoint intervals of the form (1.6), so by countable additivity of P, $\mathrm{P}(U_1 \in [0,x)) = x$ for $x \in [0,1]$. Hence, $U_1$ is standard uniform.

We now have an infinite sequence $U_1, U_2, \ldots$ of independent, standard uniform random variables. Let $F_i$ be the c.d.f. corresponding to $\mu_i$; that is, $F_i(x) := \mu_i((-\infty, x])$, for $i \in \mathbb{N}$. Set $X_i := F_i^{-1}(U_i)$. Then, as in the proof of Theorem 1.42, $X_i$ has distribution $\mu_i$, and since the $U_i$ are independent, so are the $X_i$. □

**Definition 1.44.** The distribution $\mu$ of a r.v. $X$ is *absolutely continuous* if there is a nonnegative function $f$ on $\mathbb{R}$ such that

$$\mu((a,b]) = \int_a^b f(x)\, dx \qquad \text{for all } a < b.$$

If this is the case, we call $f$ the *(probability) density* of $X$. We also say that $X$ is absolutely continuous.

Note that any such density $f$ must satisfy

$$\int_{-\infty}^{\infty} f(x)\,dx = 1. \tag{1.7}$$

Note: absolute continuity of $\mu_X$ clearly implies that the c.d.f. $F$ of $X$ is continuous. In fact, it is equivalent to absolute continuity of the c.d.f. $F$ of $X$. In particular, $F$ is differentiable almost everywhere, and $F'(x) = f(x)$ at each point where $F'$ exists. It is also equivalent to absolute continuity of $\mu$ with respect to Lebesgue measure $\lambda$ on $\mathbb{R}$.

**Example 1.45.** A r.v. $X$ has the *normal*$(\mu, \sigma^2)$ distribution if it has density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \qquad x \in \mathbb{R},$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$. The normal$(0, 1)$ distribution is called the *standard normal* distribution; it has density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

A straightforward calculus exercise shows that, if $X$ is standard normal and $Y = \mu + \sigma X$ with $\sigma > 0$, then $Y$ is normal$(\mu, \sigma^2)$. A less straightforward exercise (using double integrals!) shows that $\phi$ satisfies (1.7).

**Example 1.46.** The *exponential*$(\lambda)$ distribution, with parameter $\lambda > 0$, is the distribution with density

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \le 0. \end{cases}$$

The exponential distribution has the *memoryless property*: If $X$ is an exponential r.v., then for any $s > 0$ and $t > 0$,

$$\mathrm{P}(X > s + t | X > t) = \mathrm{P}(X > s).$$

**Example 1.47.** The *gamma*$(\alpha, \lambda)$ distribution, with parameters $\lambda > 0$ and $\alpha > 0$, has density

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \frac{(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}, & x > 0 \\ 0, & x \le 0 \end{cases}$$

where $\Gamma$ is the *gamma function*:

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x}\,dx, \qquad \alpha > 0.$$

In case $\alpha = n$, an integer, it can be shown (by repeated integration by parts) that $\Gamma(n) = (n-1)!$. Also, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. (Proof?) The gamma distribution generalizes the exponential distribution, and is further related to it as follows.

**Proposition 1.48.** *If $X_1, \ldots, X_n$ are independent exponential($\lambda$) r.v.'s, then the sum $S := X_1 + \cdots + X_n$ has the gamma($n, \lambda$) distribution.*

**Definition 1.49.** The *joint cumulative distribution function* or *joint c.d.f.* for short, of a pair $(X, Y)$ of random variables, is defined by

$$F(x, y) = \mathrm{P}(X \le x, Y \le y).$$

**Definition 1.50.** Random variables $X$ and $Y$ are *jointly absolutely continuous* if there is a nonnegative function $f : \mathbb{R}^2 \to \mathbb{R}$ such that the joint c.d.f. of $X$ and $Y$ satisfies

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\, dv\, du, \qquad \text{for all } x \text{ and } y.$$

In that case we call $f$ the *joint density* of $X$ and $Y$. Note that

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}.$$

Joint c.d.f.'s and joint densities of three or more random variables are defined similarly.

**Definition 1.51.** A probability distribution $\mu$ on $\mathbb{R}$ is *discrete* if there is a countable subset $C \subset \mathbb{R}$ such that $\mu(C) = 1$. A random variable $X$ is discrete if its distribution is discrete.

We have already seen an example of a discrete distribution, namely the binomial distribution of Example 1.24. More examples follow below.

**Example 1.52** (Bernoulli distribution)**.** A r.v. $X$ has the *Bernoulli*($p$) distribution if $\mathrm{P}(X = 1) = p$ and $\mathrm{P}(X = 0) = 1 - p$, where $0 < p < 1$. We can construct such an $X$ on $[0, 1]$ with Lebesgue measure by putting $X(\omega) = \chi_{[0,p]}(\omega)$. The c.d.f. of a Bernoulli($p$) r.v. is

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - p, & 0 \le x < 1 \\ 1, & x \ge 1. \end{cases}$$

Using the method of Theorem 1.43 we can construct on $\Omega = [0, 1]$ an infinite sequence of independent Bernoulli($p$) r.v.'s. We can interpret this sequence as a sequence of unfair coin tosses, or more generally, a sequence of *Bernoulli trials*, where 1 stands for success, and 0 for failure.

**Example 1.53** (Geometric distribution)**.** Consider a sequence of independent Bernoulli($p$) trials as in the last example. Let $N$ be the number of the trial at which the first success occurs. More precisely, if $X_1, X_2, \ldots$ are Bernoulli($p$) r.v.'s, set $N = \inf\{i : X_i = 1\}$. Then $N$ takes possible values $1, 2, \ldots$, and conceivably, $\infty$. Now

$$\mathrm{P}(N = n) = \mathrm{P}(X_1 = 0, \ldots, X_{n-1} = 0, X_n = 1) = (1 - p)^{n-1} p, \qquad n \in \mathbb{N}.$$

Since these probabilities add to 1, it follows that $\mathrm{P}(N = \infty) = 0$. We say $N$ has the *geometric*$(p)$ distribution. Note the simple formula

$$\mathrm{P}(N > n) = (1-p)^n,$$

which can be derived probabilistically, without summing a geometric series!

**Example 1.54** (Negative binomial distribution)**.** Consider again a sequence of independent Bernoulli$(p)$ trials, but now, for a fixed $r \in \mathbb{N}$, let $N$ be the number of the trial at which the $r$th success occurs. Then $N$ takes possible values $r, r+1, \ldots$, and

$$\mathrm{P}(N = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}, \qquad n = r, r+1, \ldots.$$

(Why?) We say $N$ has the *negative binomial*$(r, p)$ distribution.

**Example 1.55** (Poisson distribution)**.** A r.v. $X$ has the *Poisson*$(\mu)$ distribution (with parameter $\mu > 0$) if it satisfies

$$\mathrm{P}(X = k) = e^{-\mu} \frac{\mu^k}{k!}, \qquad k = 0, 1, 2, \ldots.$$

The Poisson distribution is often used as a simpler model in place of the binomial distribution if $n$ is very large and $p$ is very small. The justification is as follows. Consider the binomial$(n, p)$ distribution with probabilities

$$P_k = \binom{n}{k} p^k q^{n-k},$$

where $q = 1 - p$. Put $\mu = np$. Now let $n \to \infty$ and $p \to 0$ in such a way that $np = \mu$ remains constant. Then

$$P_0 = q^n = (1-p)^n = \left(1 - \frac{\mu}{n}\right)^n \to e^{-\mu},$$

and for $k \geq 0$,

$$\frac{P_{k+1}}{P_k} = \frac{n-k}{k+1} \cdot \frac{p}{q} \to \frac{\mu}{k+1}.$$

Thus, for each fixed $k$,

$$P_k = P_0 \prod_{j=0}^{k-1} \frac{P_{j+1}}{P_j} \to e^{-\mu} \prod_{j=0}^{k-1} \frac{\mu}{j+1} = e^{-\mu} \frac{\mu^k}{k!}.$$

### 1.2.3 Expectation and the Lebesgue integral

**Definition 1.56.** Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space. A measurable function $f$ on $\Omega$ which takes only finitely many values is called a *simple function*. Thus, simple functions are functions of the form

$$\varphi(\omega) = \sum_{i=1}^{n} c_i \chi_{A_i}(\omega), \tag{1.8}$$

where $c_i$ are constants and $A_i$ are $\mathcal{F}$-measurable sets. If the sets $A_i$ are disjoint and $c_i \neq c_j$ for $i \neq j$ we call (1.8) the *canonical representation* of $\varphi$.

**Definition 1.57.** The *integral* of a simple function $\varphi$ with canonical representation (1.8) is defined as

$$\int \varphi(\omega)\, d\,\mathrm{P}(\omega) := \sum_{i=1}^{n} c_i\, \mathrm{P}(A_i). \tag{1.9}$$

(It can be shown that the value of the right-hand side of (1.9) does not depend on the representation of $\varphi$. Thus, it is not necessary to put a simple function in canonical representation before determining its integral.)

**Definition 1.58.** Let $f$ be a nonnegative measurable function on $\Omega$. The integral of $f$ is defined by

$$\int f(\omega)\, d\,\mathrm{P}(\omega) := \sup_{\varphi} \int \varphi(\omega)\, d\,\mathrm{P}(\omega),$$

where the supremum is taken over all simple functions $\varphi$ with $0 \leq \varphi \leq f$ everywhere on $\Omega$.

(Note that the integral of $f$ could take the value $+\infty$.)

Finally, if $f$ is an arbitrary measurable function on $\Omega$, define the positive and negative parts of $f$ respectively by

$$f^+ = \max\{f, 0\}, \qquad f^- = \max\{-f, 0\}.$$

Then $f = f^+ - f^-$, and $f^+ + f^- = |f|$. Note that both $f^+$ and $f^-$ are nonnegative.

**Definition 1.59.** The integral of a measurable function $f$ on $\Omega$ is defined by

$$\int f(\omega)\, d\,\mathrm{P}(\omega) := \int f^+(\omega)\, d\,\mathrm{P}(\omega) - \int f^-(\omega)\, d\,\mathrm{P}(\omega),$$

unless both integrals on the right equal $+\infty$, in which case the integral of $f$ does not exist.

If both $\int f^+\, d\,\mathrm{P}$ and $\int f^-\, d\,\mathrm{P}$ are finite, we say $f$ is *integrable* (with respect to P).

**Note:** We can define the integral of $f$ with respect to a nonfinite measure in the same way. If $\Omega = \mathbb{R}$ with Lebesgue measure $\lambda$, we write $\int f(x)\, dx$ instead of $\int f(x)\, d\lambda(x)$.

**Theorem 1.60** (Properties of the integral)**.** *Assume the integrals of $f$ and $g$ exist.*

(i) *If $f = g$ a.e. (that is, $\mathrm{P}(\omega : f(\omega) = g(\omega)) = 1$), then $\int f \, d\mathrm{P} = \int g \, d\mathrm{P}$.*

(ii) *(Monotonicity) If $f \leq g$, then $\int f \, d\mathrm{P} \leq \int g \, d\mathrm{P}$. In particular, if $f \geq 0$, then $\int f \, d\mathrm{P} \geq 0$.*

(iii) *(Linearity) For real $a$ and $b$,*

$$\int (af + bg) \, d\mathrm{P} = a \int f \, d\mathrm{P} + b \int g \, d\mathrm{P} \, .$$

**Definition 1.61.** We define the integral of $f$ over a set $A \in \mathcal{F}$ by

$$\int_A f \, d\mathrm{P} := \int f \chi_A \, d\mathrm{P} \, .$$

**Definition 1.62.** The *expectation* of a random variable $X$ on $(\Omega, \mathcal{F}, \mathrm{P})$ is the value of

$$\mathrm{E}(X) := \int X(\omega) \, d\mathrm{P}(\omega),$$

provided the integral exist. If $\mathrm{E}|X| < \infty$, we say $X$ is *integrable*.

**Definition 1.63.** Let $X$ and $Y$ be random variables on a common probability space $(\Omega, \mathcal{F} \mathrm{P})$. We say $X = Y$ *almost surely* (a.s.) if $\mathrm{P}(X = Y) = \mathrm{P}(\omega : X(\omega) = Y(\omega)) = 1$.

**Proposition 1.64.** *If $X = Y$ a.s., then $\mathrm{E}(X) = \mathrm{E}(Y)$.*

*Proof.* This is just a restatement of Theorem 1.60 (i). □

**Theorem 1.65** (Change of variable)**.** *Let $X$ be a random variable with distribution $\mu$. Let $g$ be a real function for which $g(X)$ is either nonnegative or integrable with respect to* P*. Then*

$$\int_\Omega g(X(\omega)) \, d\mathrm{P}(\omega) = \int_{\mathbb{R}} g(x) \, d\mu(x).$$

*Proof.* This is a standard exercise in measure theory. We will not prove it here. □

**Theorem 1.66** (Expectation formulas)**.** *Let $X$ be a random variable and $g$ a real function.*

(i) *If $X$ is discrete, then*

$$\mathrm{E}(g(X)) = \sum_x g(x) \, \mathrm{P}(X = x). \tag{1.10}$$

(ii) *If $X$ is absolutely continuous with density $f$, then*

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) \, dx. \tag{1.11}$$

*Proof.* (i) If $X$ is discrete, then $g(X)$ takes only countably many values, say $c_1, c_2, \ldots$. Assume first that $g(X)$ is nonegative. (If not, consider positive part and negative part separately.) Let $\{B_i\}$ be a partition of $\Omega$ such that $g(X) = c_i$ on $B_i$, so that we can write $g(X(\omega)) = \sum_{i=1}^{\infty} c_i \chi_{B_i}(\omega)$. By the definition of $\int g(X)\, d\,\mathrm{P}$ as a supremum over simple functions, we have

$$\int g(X)\, d\,\mathrm{P} \geq \sum_{i=1}^{n} c_i\, \mathrm{P}(B_i), \qquad \text{for each } n,$$

and so

$$\int g(X)\, d\,\mathrm{P} \geq \sum_{i=1}^{\infty} c_i\, \mathrm{P}(B_i).$$

One checks easily that any simple function $0 \leq \varphi \leq g(X)$ must satisfy

$$\int \varphi\, d\,\mathrm{P} \leq \sum_{i=1}^{\infty} c_i\, \mathrm{P}(B_i),$$

and hence,

$$\int g(X)\, d\,\mathrm{P} = \sum_{i=1}^{\infty} c_i\, \mathrm{P}(B_i).$$

(ii) is a consequence of Theorem 1.65 and another standard exercise in measure theory (Royden, Exer. 11.22): if

$$\mu(A) = \int_A f\, d\lambda, \qquad A \in \mathcal{B}(\mathbb{R}),$$

then

$$\int g\, d\mu = \int g f\, d\lambda.$$

□

Taking $g(x) = x^k$ in the above theorem we obtain the $k^{\text{th}}$ *moment* of $X$, $\mathrm{E}(X^k)$. The *variance* of $X$, denoted $\mathrm{Var}(X)$, is defined the same way as before, so

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

The expectations and variances in the examples below can all be obtained using basic calculus techniques.

**Example 1.67** (Discrete distributions)**.**

(1) Let $X \sim$ geometric$(p)$. Then $\mathrm{E}(X) = 1/p$ and $\mathrm{Var}(X) = (1-p)/p^2$.

(2) Let $X \sim$ negative binomial$(r, p)$. Then $\mathrm{E}(X) = r/p$ and $\mathrm{Var}(X) = r(1-p)/p^2$. (To see this probabilistically, think of $X$ as a sum of $r$ independent geometric$(p)$ r.v.'s!)

(3) Let $X \sim \text{Poisson}(\mu)$. Then $\text{E}(X) = \text{Var}(X) = \mu$.

**Example 1.68** (Absolutely continuous distributions)**.**

(1) Let $X \sim \text{uniform}(0, 1)$. Then $\text{E}(X) = 1/2$ and $\text{Var}(X) = 1/12$.

(2) Let $X \sim \text{normal}(\mu, \sigma^2)$. Then $\text{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$.

(3) Let $X \sim \text{exponential}(\lambda)$. Then $\text{E}(X) = 1/\lambda$ and $\text{Var}(X) = 1/\lambda^2$.

(4) Let $X \sim \text{gamma}(\alpha, \lambda)$. Then $\text{E}(X) = \alpha/\lambda$ and $\text{Var}(X) = \alpha/\lambda^2$. (For $\alpha \in \mathbb{N}$, this follows from Proposition 1.48.)

**Definition 1.69.** Let $\text{P}$ and $\tilde{\text{P}}$ be probability measures on the same space $(\Omega, \mathcal{F})$. We say $\tilde{\text{P}}$ is *absolutely continuous* with respect to $\text{P}$, and write $\tilde{\text{P}} \ll \text{P}$, if $\text{P}(A) = 0$ implies $\tilde{\text{P}}(A) = 0$. If both $\tilde{\text{P}} \ll \text{P}$ and $\text{P} \ll \tilde{\text{P}}$, we say $\text{P}$ and $\tilde{\text{P}}$ are *equivalent*.

Losely speaking, equivalent measures agree on what is "possible", but may disagree on the likelihood of possible things.

**Theorem 1.70** (Radon-Nikodym)**.** *If* $\text{P}$ *and* $\tilde{\text{P}}$ *are probability measures on a* $(\Omega, \mathcal{F})$ *and* $\tilde{\text{P}} \ll \text{P}$*, then there is a nonnegative random variable* $Z$ *on* $(\Omega, \mathcal{F})$ *such that*

$$\tilde{\text{P}}(A) = \text{E}[Z\chi_A] = \int_A Z \, d\,\text{P}, \qquad A \in \mathcal{F}.$$

*Furthermore, if we let* $\tilde{\text{E}}$ *denote expectation with respect to* $\tilde{\text{P}}$*, that is,*

$$\tilde{\text{E}}(X) = \int X \, d\,\tilde{\text{P}},$$

*then*

$$\tilde{\text{E}}(X) = \text{E}[XZ].$$

*In particular,* $\text{E}(Z) = 1$.

We call $Z$ the *Radon-Nikodym derivative* of $\tilde{\text{P}}$ with respect to $\text{P}$, and write

$$Z = \frac{d\,\tilde{\text{P}}}{d\,\text{P}}.$$

*Proof.* The first part of the theorem is just the Radon-Nikodym theorem from measure theory (see Royden, Thm 11.23). The second part follows from the observations at the end of the proof of Theorem 1.66. □

### 1.2.4 Useful inequalities

We end this section with some useful inequalities. First, let $X$ be a nonnegative random variable, and $a > 0$. Then

$$\mathrm{E}(X) = \int X \, d\,\mathrm{P} = \int_{\{X<a\}} X \, d\,\mathrm{P} + \int_{\{X\geq a\}} X \, d\,\mathrm{P} \geq 0 + \int_{\{X\geq a\}} a \, d\,\mathrm{P} = a\,\mathrm{P}(X \geq a).$$

Hence,

$$\mathrm{P}(X \geq a) \leq \frac{\mathrm{E}(X)}{a}.$$

This is called *Markov's inequality*. For an arbitrary random variable $X$ and $k \in \mathbb{N}$ we thus have

$$\mathrm{P}(|X| \geq a) = \mathrm{P}(|X|^k \geq a^k) \leq \frac{\mathrm{E}(|X|^k)}{a^k}.$$

Now suppose that $X$ is a random variable with mean $\mathrm{E}(X) = m$ and finite variance. Taking $k = 2$ in the last inequality above and replacing $X$ with $X - m$, we obtain *Chebyshev's inequality*:

$$\mathrm{P}(|X - m| \geq a) \leq \frac{\mathrm{Var}\, X}{a^2}.$$

Another useful inequality is *Jensen's inequality*: If $X$ is a random variable and $\varphi$ a convex real function, then

$$\mathrm{E}[\varphi(X)] \geq \varphi(\mathrm{E}(X)).$$

(See Royden, sec. 5.5.)

## 1.3 Convergence of random variables

**Definition 1.71.** A sequence of random variables $\{X_n\}$ defined on a common probability space $(\Omega, \mathcal{F}, \mathrm{P})$ *converges almost surely* to a random variable $X$ (defined on the same space) if

$$\mathrm{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Notation: $X_n \to X$ a.s.

We often need conditions under which $X_n \to X$ a.s. implies that $\mathrm{E}(X_n) \to \mathrm{E}(X)$. We state the following convergence theorems from measure theory in their probability contexts without proof. (See Royden, sec. 4.3 or sec. 11.3.)

**Theorem 1.72** (Bounded Convergence Theorem)**.** *If $\{X_n\}$ is uniformly bounded, i.e. there is $K > 0$ such that $|X_n(\omega)| \leq K$ for all $n$ and all $\omega$, then $X_n \to X$ a.s. implies that* $\mathrm{E}(X) = \lim_{n\to\infty} \mathrm{E}(X_n)$.

**Theorem 1.73** (Fatou's Lemma)**.** *If $X_n$ is nonnegative for each $n$ and $X_n \to X$ a.s., then*

$$\mathrm{E}(X) \leq \liminf_{n\to\infty} \mathrm{E}(X_n).$$

**Theorem 1.74** (Monotone Convergence Theorem)**.** *If $X_n$ is nonnegative for each $n$, $X_{n+1} \geq X_n$ a.s. for each $n$, and $X_n \to X$ a.s., then $\mathrm{E}(X) = \lim_{n\to\infty} \mathrm{E}(X_n)$.*

**Theorem 1.75** (Dominated Convergence Theorem)**.** *Suppose $X_n$ are integrable random variables, and there is an integrable random variable $Y$ such that $|X_n| \leq Y$ a.s. for each $n$. Then, if $X_n \to X$ a.s., we have $\mathrm{E}(X) = \lim_{n\to\infty} \mathrm{E}(X_n)$.*

As a typical example of how these theorems may be used, we prove the following expectation formula for nonnegative random variables.

**Theorem 1.76.** *If $X$ is a nonnegative r.v., then*

$$\mathrm{E}(X) = \int_0^\infty \mathrm{P}(X > x)\, dx. \tag{1.12}$$

*Proof.* Assume first that $X$ is a simple r.v. (that is, a r.v. taking only finitely many values). Say the possible values of $X$ are $x_1 < x_2 < \cdots < x_m$. Put $x_0 := 0$, and let $y_i = x_i - x_{i-1}$ for $i = 1, 2, \ldots, m$. Then $x_i = \sum_{\nu=1}^{i} y_\nu$ for each $i \geq 1$. We now calculate

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{i=1}^{m} x_i \,\mathrm{P}(X = x_i) = \sum_{i=1}^{m} \sum_{\nu=1}^{i} y_\nu \,\mathrm{P}(X = x_i) \\
&= \sum_{\nu=1}^{m} \sum_{i=\nu}^{m} y_\nu \,\mathrm{P}(X = x_i) = \sum_{\nu=1}^{m} (x_\nu - x_{\nu-1}) \,\mathrm{P}(X \geq x_\nu) \\
&= \sum_{\nu=1}^{m} \int_{x_{\nu-1}}^{x_\nu} \mathrm{P}(X > x)\, dx = \int_0^\infty \mathrm{P}(X > x)\, dx,
\end{aligned}$$

since $\mathrm{P}(X > x) = 0$ for $x > x_m$. Thus, (1.12) holds for simple random variables. If $X$ is an arbitrary nonnegative r.v., let $\{X_n\}$ be an increasing sequence of nonnegative simple r.v.'s such that $X_n \to X$ a.s. By the Monotone Convergence Theorem (MCT), $\mathrm{E}(X_n) \uparrow \mathrm{E}(X)$. Furthermore, for each $x > 0$, $\mathrm{P}(X_n > x) \uparrow \mathrm{P}(X > x)$. So again by the MCT (applied this time to $[0, \infty)$ with Lebesgue measure $\lambda$),

$$\int_0^\infty \mathrm{P}(X_n > x)\, dx \quad \uparrow \quad \int_0^\infty \mathrm{P}(X > x)\, dx.$$

Since (1.12) holds for each $X_n$, it thus holds for $X$ as well. □

Similarly, the MCT is used to prove the analog of Proposition 1.23(ii).

**Theorem 1.77.** *If $X$ and $Y$ are independent, then $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$.*

*Proof.* Assume first that $X$ and $Y$ are both nonnegative. There are sequences $\{X_n\}$ and $\{Y_n\}$ of nonnegative simple r.v.'s such that for each $n$, $X_n$ and $Y_n$ are independent, $X_n \uparrow X$ a.s. and $Y_n \uparrow Y$ a.s. By Proposition 1.23, $\mathrm{E}(X_n Y_n) = \mathrm{E}(X_n)\,\mathrm{E}(Y_n)$. Applying the MCT to both sides gives $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$, since $X_n Y_n \uparrow XY$.

For general $X$ and $Y$, we have $XY = (X^+ - X^-)(Y^+ - Y^-)$, and hence,

$$\begin{aligned}
\mathrm{E}(XY) &= \mathrm{E}[(X^+ - X^-)(Y^+ - Y^-)] \\
&= \mathrm{E}(X^+Y^+) - \mathrm{E}(X^+Y^-) - \mathrm{E}(X^-Y^+) + \mathrm{E}(X^-Y^-) \\
&= \mathrm{E}(X^+)\,\mathrm{E}(Y^+) - \mathrm{E}(X^+)\,\mathrm{E}(Y^-) - \mathrm{E}(X^-)\,\mathrm{E}(Y^+) + \mathrm{E}(X^-)\,\mathrm{E}(Y^-) \\
&= \big(\mathrm{E}(X^+) - \mathrm{E}(X^-)\big)\big(\mathrm{E}(Y^+) - \mathrm{E}(Y^-)\big) \\
&= \mathrm{E}(X)\,\mathrm{E}(Y).
\end{aligned}$$

□

We now turn to the question of proving almost sure convergence. Recalling the definition of a limit, we have

$$\{X_n \not\to X\} = \bigcup_{\varepsilon>0} \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} \{|X_n - X| \geq \varepsilon\} = \bigcup_{\varepsilon>0,\varepsilon\in\mathbb{Q}} \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} \{|X_n - X| \geq \varepsilon\}.$$

Thus, if we want to show that $X_n \to X$ a.s., it suffices to show (in view of countable additivity) that for each $\varepsilon > 0$,

$$\mathrm{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} \{|X_n - X| \geq \varepsilon\}\right) = 0. \tag{1.13}$$

For this, the *first Borel-Cantelli lemma* is useful. Given a sequence $\{A_n\}$ of events, define

$$\limsup A_n = \{A_n \text{ infinitely often (i.o.)}\} := \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k,$$

$$\liminf A_n = \{A_n \text{ almost always (a.a.)}\} := \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k.$$

**Exercise:** Find several relationships between $\limsup A_n$ and $\liminf A_n$.

**Proposition 1.78** (The first Borel-Cantelli lemma)**.**
*Let $\{A_n\}$ be any sequence of events. If $\sum_{n=1}^{\infty} \mathrm{P}(A_n) < \infty$, then $\mathrm{P}(\limsup A_n) = 0$.*

*Proof.* Note that for each $m$,

$$\mathrm{P}(\limsup A_n) \leq \mathrm{P}\left(\bigcup_{k=m}^{\infty} A_k\right) \leq \sum_{k=m}^{\infty} \mathrm{P}(A_k).$$

By the hypothesis, the last sum tends to zero as $m \to \infty$. Hence, $\mathrm{P}(\limsup A_n) = 0$. □

There exists a partial converse the the first Borel-Cantelli lemma. (It requires the $A_n$ to be independent.)

**Proposition 1.79** (The second Borel-Cantelli lemma)**.**
*Let $\{A_n\}$ be a sequence of independent events. If $\sum_{n=1}^{\infty} \mathrm{P}(A_n) = \infty$, then $\mathrm{P}(\limsup A_n) = 1$.*

*Proof.* Note $(\limsup A_n)^c = \liminf A_n^c$. Using independence and the inequality $1-x \le e^{-x}$ we have, for any $j \in \mathbb{N}$,

$$\mathrm{P}\left(\bigcap_{k=n}^{n+j} A_k^c\right) = \prod_{k=n}^{n+j}(1-\mathrm{P}(A_k)) \le \exp\left(-\sum_{k=n}^{n+j}\mathrm{P}(A_k)\right).$$

By hypothesis, the sum inside the last expression tends to $\infty$ as $j \to \infty$, and hence the exponential tends to 0. Therefore,

$$\mathrm{P}\left(\bigcap_{k=n}^{\infty} A_k^c\right) = \lim_{j\to\infty} \mathrm{P}\left(\bigcap_{k=n}^{n+j} A_k^c\right) = 0, \qquad \text{for each } n.$$

By countable additivity, it follows that $\mathrm{P}(\liminf A_n^c) = 0$, and hence, $\mathrm{P}(\limsup A_n) = 1$. □

In fact it can be shown (as a consequence of *Kolmogorov's zero-one law* - see Billingsley, Theorem 4.5) that when the $A_n$ are independent, $\mathrm{P}(\limsup A_n)$ is always either 0 or 1. The Borel-Cantelli lemmas specify which of the two is the case.

Back to almost sure convergence... To prove that $X_n \to X$ a.s., it suffices in view of the first Borel-Cantelli lemma to show that

$$\sum_{n=1}^{\infty} \mathrm{P}(|X_n - X| \ge \varepsilon) < \infty \qquad \text{for every } \varepsilon > 0.$$

For this, Chebyshev's inequality is often a useful tool. This procedure is illustrated in the proof of the following theorem.

**Theorem 1.80** (Strong law of large numbers)**.** *Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with finite mean $m$, and put $S_n := X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \to m \quad \textit{a.s.}$$

*Proof.* The full proof is rather involved (see Billingsley, Thm 22.1), but we prove the statement here under the assumption that $\mathrm{E}(X_1^4) < \infty$. We may assume that $m = 0$, for otherwise we can replace $X_i$ with $X_i - m$. Note that $\mathrm{E}(X_1^4) < \infty$ implies $\mathrm{E}(X_1^2) < \infty$. Our goal is to show that

$$\mathrm{P}(|S_n/n| \ge \varepsilon \text{ i.o.}) = 0 \tag{1.14}$$

for every $\varepsilon > 0$. Let $\mathrm{E}(X_1^2) = \sigma^2$ and $\mathrm{E}(X_1^4) = \xi^4$. Fix $\varepsilon > 0$. By Markov's inequality,

$$\mathrm{P}(|S_n/n| \ge \varepsilon) = \mathrm{P}(S_n^4 \ge (\varepsilon n)^4) \le \varepsilon^{-4} n^{-4}\,\mathrm{E}(S_n^4). \tag{1.15}$$

Now

$$\mathrm{E}(S_n^4) = \sum_{i,j,k,l} \mathrm{E}(X_i X_j X_k X_l),$$

where the four indices range independently over $1, \ldots, n$. If any of $i, j, k$ and $l$ is different from the other three, then $\mathrm{E}(X_i X_j X_k X_l) = 0$ by independence and the fact that $m = 0$. This leaves terms of the form $\mathrm{E}(X_i^2 X_j^2) = \mathrm{E}(X_i^2)\,\mathrm{E}(X_j^2) = \sigma^4$ with $i \neq j$, of which there are $\binom{4}{2}\binom{n}{2} = 3n(n-1)$, and terms of the form $\mathrm{E}(X_i^4) = \xi^4$, of which there are $n$. Thus,

$$\mathrm{E}(S_n^4) = 3n(n-1)\sigma^4 + n\xi^4 \leq 3\sigma^4 n^2, \quad \text{for all large enough } n.$$

It thus follows by (1.15) that $\sum_{n=1}^{\infty} \mathrm{P}(|S_n/n| \geq \varepsilon) < \infty$, and the first Borel-Cantelli lemma gives (1.14). As explained earlier, this establishes that $S_n/n \to 0$ a.s. □

**Note:** The above proof is easily modified to show that if the r.v.'s $\{X_n\}$ have uniformly bounded fourth moments, they need not be identically distributed, as long as they have identical expectation $m$.

**Definition 1.81.** A sequence of random variables $\{X_n\}$ defined on a common probability space $(\Omega, \mathcal{F}, \mathrm{P})$ *converges in probability* to a random variable $X$ (defined on the same space) if for every $\varepsilon > 0$,

$$\mathrm{P}(|X_n - X| \geq \varepsilon) \to 0, \quad \text{as } n \to \infty. \tag{1.16}$$

Notation: $X_n \to_P X$.

(Note: convergence in probability is the same concept as *convergence in measure* - see Royden sec. 4.5.)

Limits in probability are unique in the following sense.

**Proposition 1.82.** *If $X_n \to_P X$ and $X_n \to_P Y$, then $\mathrm{P}(X = Y) = 1$.*

*Proof.* Let $\varepsilon > 0$. Then

$$\mathrm{P}(|X - Y| \geq \varepsilon) \leq \mathrm{P}(|X_n - X| \geq \varepsilon/2) + \mathrm{P}(|X_n - Y| \geq \varepsilon/2) \to 0.$$

Hence $\mathrm{P}(|X - Y| \geq \varepsilon) = 0$. Since $\varepsilon$ was arbitrary, $\mathrm{P}(X = Y) = 1$. □

**Proposition 1.83.** *Let $\{X_n\}$ be a sequence of r.v.'s which converges almost surely to a r.v. $X$. Then $X_n \to_P X$.*

*Proof.* Fix $\varepsilon > 0$, and let $A_n = \{|X_n - X| \geq \varepsilon\}$. Since $X_n \to X$ a.s., we have in particular that $\mathrm{P}(\limsup A_n) = 0$. Since $\bigcup_{k=n}^{\infty} A_k$ is a decreasing sequence, it follows that

$$\lim_{n\to\infty} \mathrm{P}\left(\bigcup_{k=n}^{\infty} A_k\right) = 0.$$

But then $\lim_{n\to\infty} \mathrm{P}(A_n) = 0$, as $A_n \subset \bigcup_{k=n}^{\infty} A_k$. Hence, $X_n \to_P X$. □

**Exercise 1.84.** Give an example to show that the converse of Proposition 1.83 is false.

The strong law of large numbers and Proposition 1.83 together imply the *weak law of large numbers*, which says that, under the hypotheses of the strong law,

$$S_n/n \to_P m.$$

In case the $\{X_n\}$ have finite variance $\sigma^2$, this last result has a much simpler direct proof using Chebyshev's inequality:

$$\mathrm{P}(|S_n/n - m| \geq \varepsilon) = \mathrm{P}(|S_n - nm| \geq \varepsilon n) \leq \frac{\mathrm{Var}(S_n)}{\varepsilon^2 n^2} = \frac{n\sigma^2}{\varepsilon^2 n^2} = \frac{\sigma^2}{\varepsilon^2 n} \to 0.$$

**Theorem 1.85.** *Let $\{X_n\}$ be a sequence of r.v.'s and $X$ a r.v., all defined on the same probability space. Then $X_n \to_P X$ if and only if every subsequence $\{X_{n_k}\}$ has a further subsequence $\{X_{n_{k(i)}}\}$ which converges to $X$ almost surely.*

*Proof.* Suppose $X_n \to_P X$. Given $\{n_k\}$, we can choose a subsequence $\{n_{k(i)}\}$ such that

$$k \geq k(i) \quad \Longrightarrow \quad \mathrm{P}(|X_{n_k} - X| \geq 1/i) < 2^{-i}.$$

This implies by the first Borel-Cantelli lemma that

$$\mathrm{P}(|X_{n_{k(i)}} - X| \geq 1/i \text{ i.o.}) = 0.$$

Hence, with probability one, $|X_{n_{k(i)}} - X| < 1/i$ for all but finitely many $i$. But this means $\lim_{i\to\infty} X_{n_{k(i)}} = X$ a.s.

Conversely, suppose $X_n$ does not converge to $X$ in probability. Then there is some $\varepsilon > 0$ and some subsequence $\{n_k\}$ such that

$$\mathrm{P}(|X_{n_k} - X| \geq \varepsilon) \geq \varepsilon, \qquad \text{for all } k.$$

No subsequence of $\{X_{n_k}\}$ can converge to $X$ in probability, and hence, by Proposition 1.83, no subsequence can converge to $X$ almost surely. □

**Corollary 1.86.** *Let $X_n \to_P X$ and let $f$ be a continuous real function. Then $f(X_n) \to_P f(X)$.*

*Proof.* Exercise. □

**Definition 1.87.** Let $r \geq 1$, and let $\{X_n\}$ be random variables on a common probability space with $\mathrm{E}(|X_n|^r) < \infty$. Let $X$ be a r.v. defined on the same space as the $X_n$ with $\mathrm{E}(|X|^r) < \infty$. We say $X_n \to X$ *in* $L^r$ if

$$\lim_{n\to\infty} \mathrm{E}(|X_n - X|^r) = 0.$$

**Proposition 1.88.** *If $X_n \to X$ in $L^r$, then $X_n \to_P X$.*

*Proof.* This follows immediately from Markov's inequality. □

**Exercise 1.89.** Give an example (for given $r \geq 1$) of a sequence $\{X_n\}$ and a r.v. $X$ such that:

(1) $X_n$ converges to $X$ almost surely, but not in $L^r$.

(2) $X_n$ converges to $X$ in $L^r$, but not almost surely.

**Definition 1.90.** Let $\{X_n\}$ be a sequence of random variables, not necessarily defined on the same probability space, with c.d.f.'s $\{F_n\}$. Let $X$ be a random variable with c.d.f. $F$. We say $X_n$ *converges to $X$ in distribution*, or *converges weakly to $X$*, if $F_n(x) \to F(x)$ for every continuity point $x$ of $F$. We denote this by $X_n \Rightarrow X$. We also say in this case that $F_n$ converges weakly to $F$, denoted $F_n \Rightarrow F$.

**Theorem 1.91.** *Let $X$, $\{X_n\}$ be random variables defined on a common probability space. If $X_n \to_P X$, then $X_n \Rightarrow X$.*

*Proof.* Let $x$ be a continuity point of the c.d.f. $F$ of $X$; that is, $\mathrm{P}(X = x) = 0$. Verify that for $\varepsilon > 0$,

$$\mathrm{P}(X \leq x - \varepsilon) - \mathrm{P}(|X_n - X| \geq \varepsilon) \leq \mathrm{P}(X_n \leq x) \leq \mathrm{P}(X \leq x + \varepsilon) + \mathrm{P}(|X_n - X| \geq \varepsilon).$$

Let $n \to \infty$ and $\varepsilon \downarrow 0$ to obtain, by (1.16),

$$\mathrm{P}(X \leq x) = \mathrm{P}(X < x) \leq \liminf_{n\to\infty} \mathrm{P}(X_n \leq x) \leq \limsup_{n\to\infty} \mathrm{P}(X_n \leq x) \leq \mathrm{P}(X \leq x).$$

Thus, $\mathrm{P}(X_n \leq x) \to \mathrm{P}(X \leq x)$, as required. □

**Exercise 1.92.** Give an example to show that the converse of Theorem 1.91 fails.

An important tool for proving theorems about convergence in distribution is the following.

**Theorem 1.93** (Skorohod's theorem)**.** *Suppose $X_n \Rightarrow X$. Then there exist random variables $\{Y_n\}$ and $Y$ on a common probability space $(\Omega, \mathcal{F}, \mathrm{P})$ such that $Y_n$ has the same distribution as $X_n$ for each $n$, $Y$ has the same distribution as $X$, and $Y_n(\omega) \to Y(\omega)$ for each $\omega \in \Omega$.*

*Proof.* Take $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}([0, 1])$, and P Lebesgue measure on $[0, 1]$. Let $F_n$ be the c.d.f. of $X_n$, and $F$ the c.d.f. of $X$. For each $\omega \in [0, 1]$, put $Y_n(\omega) = F_n^{-1}(\omega)$ and $Y(\omega) = F^{-1}(\omega)$. (Recall the definition (1.4).) As in Theorem 1.42, $Y_n$ has c.d.f. $F_n$ and $Y$ has c.d.f. $F$. We need to show that $Y_n(\omega) \to Y(\omega)$. The basic idea is that, if $F_n(x) \to F(x)$ for each continuity point $x$ of $F$, then $F_n^{-1}(u) \to F^{-1}(u)$ at each continuity point $u$ of $F^{-1}$. In other words, $Y_n(\omega) \to Y(\omega)$ at each continuity point $\omega$ of $Y$. The technical details of the argument can be found in Billingsley, Thm. 25.6. Since $Y$ is nondecreasing on $[0, 1]$, it has at most countably many discontinuity points, so the set of these discontinuity points has measure 0. At each such point $\omega$, redefine $Y_n(\omega) = Y(\omega) = 0$. Then $Y_n(\omega) \to Y(\omega)$ for all $\omega$, and $Y_n$ and $Y$ still have c.d.f. $F_n$ and $F$, respectively. □

The following theorem, which characterizes weak convergence in terms of expectations, illustrates the power of the Skorohod theorem.

**Theorem 1.94.** *Let $X$, $\{X_n\}$ be random variables. Then $X_n \Rightarrow X$ if and only if*

$$\mathrm{E}[f(X_n)] \to \mathrm{E}[f(X)] \tag{1.17}$$

*for any bounded, continuous real function $f$.*

*Proof.* Suppose $X_n \Rightarrow X$. Let $f$ be bounded and continuous. Construct random variables $Y_n$ and $Y$ as in Theorem 1.93. Then $\mathrm{E}[f(X_n)] = \mathrm{E}[f(Y_n)]$ and $\mathrm{E}[f(X)] = \mathrm{E}[f(Y)]$, and by the Dominated (or the Bounded) Convergence Theorem, $\mathrm{E}[f(Y_n)] \to \mathrm{E}[f(Y)]$. Thus, we have (1.17).

Conversely, suppose (1.17) holds for each bounded and continuous $f$. Let $F_n$ be the c.d.f. of $X_n$, and $F$ the c.d.f. of $X$. To show $X_n \Rightarrow X$, we would like to take $f = \chi_{(-\infty,x]}$ for each continuity point $x$ of $F$, but this $f$ is not continuous. Hence, approximate it by a continuous function as follows. Fix $y > x$, and put

$$f(t) = \begin{cases} 1, & t \le x \\ \frac{y-t}{y-x}, & x \le t \le y \\ 0, & t \ge y. \end{cases}$$

Then $f \ge \chi_{(-\infty,x]}$, so $F_n(x) = \mathrm{E}[\chi_{(-\infty,x]}(X_n)] \le \mathrm{E}[f(X_n)]$. On the other hand, $f \le \chi_{(-\infty,y]}$, and so $\mathrm{E}[f(X)] \le \mathrm{E}[\chi_{(-\infty,y]}(X)] = F(y)$. It follows from (1.17) that

$$\limsup_{n\to\infty} F_n(x) \le F(y).$$

Letting $y \downarrow x$ and using that $F$ is right-continuous gives

$$\limsup_{n\to\infty} F_n(x) \le F(x).$$

By a similar argument (exercise!) we can show that

$$F(x-) \le \liminf_{n\to\infty} F_n(x).$$

Hence, $F_n(x) \to F(x)$ at all continuity points $x$ of $F$. □

The most important statement about convergence in distribution is, of course, the *Central Limit Theorem*. For its proof and several generalizations, see Billingsley, sec. 27.

**Theorem 1.95** (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with mean $m$ and finite positive variance $\sigma^2$, and let $S_n = X_1 + \cdots + X_n$. Then*

$$\frac{S_n - nm}{\sigma\sqrt{n}} \Rightarrow N,$$

*where $N$ is a standard normal random variable.*