Overview

This is an introduction to the mathematical foundations of probability theory. It is intended as a supplement or follow-up to a graduate course in real analysis. The first two sections assume the knowledge of measure spaces, measurable functions, Lebesgue integral, and notions of convergence of functions; the third assumes Fubini's Theorem; the fifth assumes knowledge of Fourier transform of nice (Schwartz) functions on $\mathbb{R}$; and Section 6 uses the Radon-Nikodym Theorem.

The mathematical foundations of probability theory are exactly the same as those of Lebesgue integration. However, probability adds much intuition and leads to different developments of the area. These notes are only intended to be a brief introduction — this might be considered what every graduate student should know about the theory of probability.

Probability uses some different terminology than that of Lebesgue integration in $\mathbb{R}$. These notes will introduce the terminology and will also relate these ideas to those that would be encountered in an elementary (by which we will mean pre-measure theory) course in probability or statistics. Graduate students encountering probabilty for the first time might want to also read an undergraduate book in probability.

1 Probability spaces

Definition A *probability space* is a measure space with total measure one. The standard notation is $(\Omega, \mathcal{F}, \mathbb{P})$ where:

- Ω is a set (sometimes called a *sample space* in elementary probability). Elements of Ω are denoted ω and are sometimes called *outcomes*.
- $\mathcal{F}$ is a σ-algebra (or σ-field, we will use these terms synonymously) of subsets of Ω. Sets in $\mathcal{F}$ are called *events*.
- $\mathbb{P}$ is a function from $\mathcal{F}$ to $[0,1]$ with $\mathbb{P}(\Omega) = 1$ and such that if $E_1, E_2, \ldots \in \mathcal{F}$ are disjoint,
$$\mathbb{P}\left[\bigcup_{j=1}^{\infty} E_j\right] = \sum_{j=1}^{\infty} \mathbb{P}[E_j].$$
We say "probability of E" for $\mathbb{P}(E)$.

A *discrete probability space* is a probability space such that Ω is finite or countably infinite. In this case we usually choose $\mathcal{F}$ to be all the subsets of Ω (this can be written $\mathcal{F} = 2^{\Omega}$), and the probability measure $\mathbb{P}$ is given by a function $p : \Omega \to [0,1]$ with $\sum_{\omega\in\Omega} p(\omega) = 1$. Then,

$$\mathbb{P}(E) = \sum_{\omega\in E} p(\omega).$$

We will consider another important example here, the probability space associated to an infinite number of flips of a coin. Let

$$\Omega = \{\omega = (\omega_1, \omega_2, \ldots) : \omega_j = 0 \text{ or } 1\}.$$

We think of 0 as "tails" and 1 as "heads". For each positive integer n, let

$$\Omega_n = \{(\omega_1, \omega_2, \ldots, \omega_n) : \omega_j = 0 \text{ or } 1\}.$$

Each Ω_n is a finite set with 2^n elements. We can consider Ω_n as a probability space with σ-algebra 2^{Ω_n} and probability $\mathbb{P}_n$ induced by

$$p_n(\omega) = 2^{-n}, \quad \omega \in \Omega_n.$$

Let $\mathcal{F}_n$ be the σ-algebra on Ω consisting of all events that depend only on the first n flips. More formally, we define $\mathcal{F}_n$ to be the collection of all subsets A of Ω such that there is an $E \in 2^{\Omega_n}$ with

$$A = \{(\omega_1, \omega_2, \ldots) : (\omega_1, \omega_2, \ldots, \omega_n) \in E\}. \tag{1}$$

Note that $\mathcal{F}_n$ is a finite σ-algebra (containing 2^{2^n} subsets) and

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \cdots .$$

If A is of the form (1), we let

$$\mathbb{P}(A) = \mathbb{P}_n(E).$$

One can easily check that this definition is consistent. This gives a function $\mathbb{P}$ on

$$\mathcal{F}^0 := \bigcup_{j=1}^{\infty} \mathcal{F}_j.$$

Recall that an *algebra* of subsets of Ω is a collection of subsets containing Ω, closed under complementation, and closed under *finite* unions. To show closure under finite unions, it suffices to show that if $E_1, E_2 \in \mathcal{F}_0$, then $E_1 \cup E_2 \in \mathcal{F}_0$.

Proposition 1.1. *$\mathcal{F}^0$ is an algebra but not a σ-algebra.*

Proof. $\Omega \in \mathcal{F}^0$ since $\Omega \in \mathcal{F}_1$. Suppose $E \in \mathcal{F}^0$. Then $E \in \mathcal{F}_n$ for some n, and hence $E^c \in \mathcal{F}_n$ and $E^c \in \mathcal{F}$. Also, suppose $E_1, E_2 \in \mathcal{F}$. Then there exists j, k with $E_1 \in \mathcal{F}_j, E_2 \in \mathcal{F}_k$. Let $n = \max\{j, k\}$. Then since the σ-algebras are increasing, $E_1, E_2 \in \mathcal{F}_n$ and hence $E_1 \cup E_2 \in \mathcal{F}_n$. Therefore, $E_1 \cup E_2 \in \mathcal{F}^0$ and $\mathcal{F}^0$ is an algebra.

To see that $\mathcal{F}^0$ is not a σ-algebra consider the singleton set

$$E = \{(1, 1, 1, \ldots)\}.$$

E is not in $\mathcal{F}^0$ but E can be written as a countable intersection of events in $\mathcal{F}^0$,

$$E = \bigcap_{j=1}^{\infty} \{(\omega_1, \omega_2, \ldots) : \omega_1 = \omega_2 = \cdots = \omega_j = 1\}.$$

□

Proposition 1.2. *The function $\mathbb{P}$ is a (countably additive) measure on $\mathcal{F}^0$, i.e., it satisfies $\mathbb{P}[\emptyset] = 0$, and if $E_1, E_2, \ldots \in \mathcal{F}^0$ are disjoint with $\cup_{n=1}^{\infty} E_n \in \mathcal{F}^0$, then*

$$\mathbb{P}\left[\bigcup_{n=1}^{\infty} E_n\right] = \sum_{n=1}^{\infty} \mathbb{P}(E_n).$$

Proof. $\mathbb{P}[\emptyset] = 0$ is immediate. Also it is easy to see that $\mathbb{P}$ is *finitely additive*, i.e., if $E_1, E_2, \ldots, E_n \in \mathcal{F}^0$ are disjoint then

$$\mathbb{P}\left[\bigcup_{j=1}^{n} E_j\right] = \sum_{j=1}^{n} \mathbb{P}(E_j).$$

(To see this, note that there must be an N such that $E_1, \ldots, E_n \in \mathcal{F}_N$ and then we can use the additivity of $\mathbb{P}_N$.)

Showing countable subadditivity is harder. In fact, the following stronger fact holds: suppose $E_1, E_2, \ldots \in \mathcal{F}^0$ are disjoint and $E = \cup_{n=1}^{\infty} E_n \in \mathcal{F}^0$. Then there is an N such that $E_j = \emptyset$ for $j > N$. Once we establish this, countable additivity follows from finite additivity.

To establish this, we consider Ω as the *topological* space

$$\{0,1\} \times \{0,1\} \times \cdots$$

with the product topology where we have given each $\{0,1\}$ the discrete topology (all four subsets are open). The product topology is the smallest topology such that all the sets in $\mathcal{F}^0$ are open. Note also that all sets in $\mathcal{F}^0$ are closed since they are complements of sets in $\mathcal{F}^0$. It follows from Tychonoff's Theorem that Ω is a compact topological space under this topology. Suppose $E_1, E_2, \ldots$ are as above with $E = \cup_{n=1}^{\infty} E_n \in \mathcal{F}^0$. Then $E_1, E_2, \ldots$ is an open cover of the closed (and hence compact) set E. Therefore there is a finite subcover, $E_1, E_2, \ldots, E_N$. Since the $E_1, E_2, \ldots$ are disjoint, this must imply that $E_j = \emptyset$ for $j > N$. $\square$

Let $\mathcal{F}$ be the smallest σ-algebra containing $\mathcal{F}^0$. Then the Carathéodory Extension Theorem tells us that $\mathbb{P}$ can be extended uniquely to a complete measure space $(\mathbb{P}, \bar{\mathcal{F}}, \mathbb{P})$ where $\mathcal{F} \subset \bar{\mathcal{F}}$.

Remark The astute reader will note that the construction we just did is exactly the same as the construction of Lebesgue measure on $[0,1]$. Here we denote a real number $x \in [0,1]$ by its dyadic expansion

$$x = .\omega_1\omega_2\omega_3\cdots = \sum_{j=1}^{\infty} \frac{\omega_j}{2^j}.$$

(There is a slight nuisance with the fact that $.011111\cdots = .100000\cdots$, but this can be handled.) The σ-algebra $\mathcal{F}$ above corresponds to the Borel subsets of $[0,1]$ and the completion $\overline{\mathcal{F}}$ corresponds to the Lebesgue measurable sets. If the most complicated probability space we were interested were the space above, then we could just use Lebesgue measure on $[0,1]$. In fact, for almost all important applications of probability, one *could* choose the measure space to be $[0,1]$ with Lebesgue measure (see Exercise (3)). However, this choice is not always the most convenient or natural.

Remark Continuing on the last remark, one generally does not care what probability space one is working on. What one observes are "random variables" which are discussed in the next section.

2 Random Variables and Expectation

Definition A *random variable* X is a measurable function from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to the reals, i.e., it is a function

$$X : \Omega \to (-\infty, \infty)$$

such that for every Borel set B,

$$X^{-1}(B) = \{X \in B\} \in \mathcal{F}.$$

Here we use the shorthand notation

$$\{X \in B\} = \{\omega \in \Omega : X(\omega) \in B\}.$$

If X is a random variable, then for every Borel subset B of $\mathbb{R}$, $X^{-1}(B) \in \mathcal{F}$. We can define a function on Borel sets by

$$\mu_X(B) = \mathbb{P}\{X \in B\} = \mathbb{P}[X^{-1}(B)].$$

This function is in fact a measure, and $(\mathbb{R}, \mathcal{B}, \mu_X)$ is a probability space. The measure μ_X is called the *distribution* of the random variable. If μ_X gives measure one to a countable set of reals, then X is called a *discrete random variable*. If μ_X gives zero measure to every singleton set, and hence to every countable set, X is called a *continuous random variable*. Every random variable can be written as a sum of a discrete random variable and a continuous random variable. All random varaibles defined on a discrete probability space are discrete.

The distribution μ_X is often given in terms of the *distribution function*[1] defined by

$$F_X(x) = \mathbb{P}\{X \leq x\} = \mu_X(-\infty, x].$$

Note that $F = F_X$ satisfies the following:

- $\lim_{x \to -\infty} F(x) = 0$.
- $\lim_{x \to \infty} F(x) = 1$.
- F is a nondecreasing function.
- F is right continuous, i.e., for every x,

$$F(x+) \doteq \lim_{\epsilon \to 0+} F(x + \epsilon) = F(x).$$

[1] often called cumulative distribution function (cdf) in elementary courses

Conversely, any F satisfying the conditions above is the distribution function of a random variable. The distribution can be obtained from the distribution function by setting

$$\mu_X(-\infty, x] = F_X(x),$$

and extending uniquely to the Borel sets.

For some continuous random variables X, there is a function $f = f_X : \mathbb{R} \to [0, \infty)$ such that

$$\mathbb{P}\{a \le X \le b\} = \int_a^b f(x)\, dx.$$

Such a function, if it exists, is called the *density*[2] of the random variable. If the density exists, then

$$F(x) = \int_{-\infty}^x f(t)\, dt.$$

If f is continuous at t, then the fundamental theorem of calculus implies that

$$f(x) = F'(x).$$

A density f satisfies

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

Conversely, any nonnegative function that integrates to one is the density of a random variable.

Example If E is an event, the *indicator function* of E is the random variable

$$1_E(\omega) = \begin{cases} 1, & \omega \in E, \\ 0, & \omega \notin E. \end{cases}$$

(The corresponding function in analysis is often called the characteristic function and denoted χ_E. Probabilists never use the term characteristic function for the indicator function because the term characteristic function has another meaning. The term indicator function has no ambiguity.)

Example Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space for infinite tossings of a coin as in the previous section. Let

$$X_n(\omega_1, \omega_2, \ldots) = \omega_n = \begin{cases} 1, & \text{if } n\text{th flip heads}, \\ 0, & \text{if } n\text{th flip tails}. \end{cases} \tag{2}$$

$$S_n = X_1 + \cdots + X_n = \#\text{ heads on first } n \text{ flips}. \tag{3}$$

Then $X_1, X_2, \ldots,$ and $S_1, S_2, \ldots$ are discrete random variables. If $\mathcal{F}_n$ denotes the σ-algebra of events that depend only on the first n flips, then S_n is also a random variable on the probability space $(\Omega, \mathcal{F}_n, \mathbb{P})$. However, S_{n+1} is not a random variable on $(\Omega, \mathcal{F}_n, \mathbb{P})$.

[2]More precisely, it is the density or Radon-Nikoydm derivative with respect to Lebesgue measure. In elementary courses, the term probability density function (pdf) is often used.

Example Let μ be any probability measure on $(\mathbb{R}, \mathcal{B})$. Consider the trivial random variable

$$X = x.$$

Then X is a random variable and $\mu_X = \mu$. Hence every probability measure on $\mathbb{R}$ is the distribution of a random variable.

Example A random variable X has a *normal distribution* with mean μ and variance σ^2 if it has density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty.$$

If $\mu = 0$ and $\sigma^2 = 1$, X is said to have a *standard normal distribution.* The distribution function of the standard normal is often denoted Φ,

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

Example If X is a random variable and

$$g : (\mathbb{R}, \mathcal{B}) \to \mathbb{R}$$

is a Borel measurable function, then $Y = g(X)$ is also a random variable.

Example Recall that the Cantor function is a continuous function $F : [0,1] \to [0,1]$ with $F(0) = 0, F(1) = 1$ and such that $F'(x) = 0$ for all $x \in [0,1] \setminus A$ where A denotes the "middle thirds" Cantor set. Extend F to $\mathbb{R}$ by setting $F(x) = 0$ for $x \leq 0$ and $F(x) = 1$ for $x \geq 1$. Then F is a distribution function. A random variable with this distribution function is continuous, since F is continuous. However, such a random variable has no density.

Remark The term random variable is a little misleading but it standard. It is perhaps easier to think of a random variable as a "random number". For example, in the case of coin-flippings we get an infinite sequence of random numbers corresponding to the results of the flips.

Definition If X is a nonnegative random varible, the *expectation* of X, denoted $\mathbb{E}(X)$, is

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

where the integral is the Lebesgue integral. If X is a random variable with $\mathbb{E}(|X|) < \infty$, then we also define the expectation by

$$\mathbb{E}(X) = \int X\, dP.$$

If X takes on positive and negative values, and $\mathbb{E}(|X|) = \infty$, the expectation is not defined.

Other terms used for expectation are *expected value* and *mean* (the term "expectation value" can be found in the physics literature but it is not good English). The letter μ is often used for mean. If X is a discrete random variable taking on the values $a_1, a_2, a_3, \ldots$ we have the standard formula from elementary courses:

$$\mathbb{E}(X) = \sum_{j=1}^{\infty} a_j \mathbb{P}\{X = a_j\}, \tag{4}$$

provided the sum is absolutely convergent.

Lemma 2.1. *Suppose X is a random variable with distribution μ_X. Then,*

$$\mathbb{E}(X) = \int_{\mathbb{R}} x \, d\mu_X.$$

(Either side exists if and only if the other side exists.)

Proof. First assume that $X \geq 0$. If n, k are positive integers let

$$A_{k,n} = \left\{ \omega : \frac{k-1}{n} \leq X(\omega) < \frac{k}{n} \right\}.$$

For every $n < \infty$, consider the discrete random variable X_n taking values in $\{k/n : k \in \mathbb{Z}\}$,

$$X_n = \sum_{k=1}^{\infty} \frac{k}{n} 1_{A_{k,n}}.$$

Then,

$$X_n - \frac{1}{n} \leq X \leq X_n.$$

Hence

$$\mathbb{E}[X_n] - \frac{1}{n} \leq \mathbb{E}[X] \leq \mathbb{E}[X_n].$$

But,

$$\mathbb{E}\left[X_n - \frac{1}{n}\right] = \sum_{k=1}^{\infty} \frac{k-1}{n} \, \mathbb{P}(A_{k,n}),$$

$$\mathbb{E}[X_n] = \sum_{k=1}^{\infty} \frac{k}{n} \, \mathbb{P}(A_{k,n}),$$

and

$$\frac{k-1}{n} \, \mu_X[\frac{k-1}{n}, \frac{k}{n}) \leq \int_{[\frac{k-1}{n}, \frac{k}{n})} x \, d\mu_X \leq \frac{k}{n} \, \mu_X[\frac{k-1}{n}, \frac{k}{n}).$$

By summing we get

$$\mathbb{E}\left[X_n - \frac{1}{n}\right] \leq \int_{[0,\infty)} x \, d\mu_X \leq \mathbb{E}[X_n].$$

By letting n go to infinity we get the result. The general case can be done by writing $X = X^+ - X^-$. We omit the details. □

In particular, the expectation of a random variable depends only on its distribution and not on the probability space on which it is defined. If X has a density f, then the measure μ_X is the same as $f(x)\,dx$ so we can write (as in elementary courses)

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx,$$

where again the expectation exists if and only if

$$\int_{-\infty}^{\infty} |x|\, f(x)\, dx < \infty.$$

Since expectation is defined using the Lebesgue integral, all results about the Lebesgue integral (e.g., linearity, Monotone Convergence Theorem, Fatou's Lemma, Dominated Convergence Theorem) also hold for expectation.

Exercise 2.2. *Use Hölder's inequality to show that if X is a random variable and $q \geq 1$, then $\mathbb{E}[|X|^q] \geq (\mathbb{E}[X])^q$. For which X does equality hold?*

Example Consider the coin flipping random variables (2) and (3). Clearly

$$\mathbb{E}[X_n] = 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{2}.$$

Hence

$$\mathbb{E}[S_n] = \mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = \frac{n}{2}.$$

Also,

$$\mathbb{E}[X_1 + X_1 + \cdots + X_1] = \mathbb{E}[X_1] + \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_1] = \frac{n}{2}.$$

Note that in the above example, $X_1 + \cdots + X_n$ and $X_1 + \cdots + X_1$ do not have the same distribution, yet the linearity rule for expectation show that they have the same expectation. A very powerful (although trivial!) tool in probability is the linearity of the expectation even for random variables that are "correlated". If $X_1, X_2, \ldots$ are nonnegative we can even take countable sums,

$$\mathbb{E}\left[\sum_{n=1}^{\infty} X_n\right] = \sum_{n=1}^{\infty} \mathbb{E}[X_n].$$

This can be derived from the rule for finite sums and the monotone convergence theorem by considering the increasing sequence of random variables

$$Y_n = \sum_{j=1}^{n} X_j.$$

Lemma 2.3. *Suppose X is a random variable with distribution μ_X, and g is a Borel measurable function. Then,*

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)\, d\mu_X.$$

(Either side exists if and only if the other side exists.)

Proof. Assume first that g is a nonnegative function. Then there exists an increasing sequence of nonnegative simple functions with g_n approaching g. Note that $g_n(X)$ is then a sequence of nonnegative simple random variables approaching $g(X)$. For the simple functions g_n it is immediate from the definition that

$$\mathbb{E}[g_n(X)] = \int_{\mathbb{R}} g_n(x)\, d\mu_X.$$

and hence by the monotone convergence theorem,

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)\, d\mu_X.$$

The general case can be done by writing $g = g^+ - g^-$ and $g(X) = g^+(X) - g^-(X)$. □

If X has a density f, then we get the following formula found in elementary courses

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx.$$

Definition The *variance* of a random variable X is defined by

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

provided this expecation exists.

Note that

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

The variance is often denoted $\sigma^2 = \sigma^2(X)$ and the square root of the variance, σ, is called the *standard deviation*[3].

[3]There is a slightly different use of the term standard deviation in statistics; see any book on statistics for a discussion.

Lemma 2.4. *Let X be a random variable.*

(Markov's Inequality) If $a > 0$,

$$\mathbb{P}\{|X| \geq a\} \leq \frac{\mathbb{E}[|X|]}{a}.$$

(Chebyshev's Inequality) If $a > 0$,

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

Proof. Let $X_a = a1_{\{|X| \geq a\}}$. Then $X_a \leq |X|$ and hence

$$a\mathbb{P}\{|X| \geq a\} = \mathbb{E}[X_a] \leq \mathbb{E}[X].$$

This gives Markov's inequality. For Chebyshev's inequality, we apply Markov's inequality to the random variable $[X - \mathbb{E}(X)]^2$ to get

$$\mathbb{P}\{[X - \mathbb{E}(X)]^2 \geq a^2\} \leq \frac{\mathbb{E}([X - \mathbb{E}(X)]^2)}{a^2}.$$

□

Exercise 2.5 (Generalized Chebyshev Inequality)**.** *Let $f : [0, \infty) \to [0, \infty)$ be a nondecreasing Borel function and let X be a nonnegative random variable. Then for all $a > 0$,*

$$\mathbb{P}\{X \geq a\} \leq \frac{\mathbb{E}[f(X)]}{f(a)}.$$

3 Independence

Throughout this section we will assume that there is a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. All σ-algebras will be sub-σ-algebras of $\mathcal{A}$, and all random variables will be defined on this space.

Definition

- Two events A, B are called *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
- A collection of events $\{A_\alpha\}$ is called *independent* if for every distinct $A_{\alpha_1}, \ldots, A_{\alpha_n}$,

$$\mathbb{P}(A_{\alpha_1} \cap \cdots \cap A_{\alpha_n}) = \mathbb{P}(A_{\alpha_1}) \cdots \mathbb{P}(A_{\alpha_n}).$$

- A collection of events $\{A_\alpha\}$ is called *pairwise independent* if for each distinct $A_{\alpha_1}, A_{\alpha_2}$,

$$\mathbb{P}(A_{\alpha_1} \cap A_{\alpha_2}) = \mathbb{P}(A_{\alpha_1})\mathbb{P}(A_{\alpha_2}).$$

Clearly independence implies pairwise independence but the converse is false. An easy example can be seen by rolling two dice and letting $A_1 = \{$ sum of rolls is 7 $\}$, $A_2 = \{$ first roll is 1 $\}$, $A_3 = \{$ second roll is 6 $\}$. Then

$$\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) = \frac{1}{6},$$

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_2 \cap A_3) = \frac{1}{36},$$

but $\mathbb{P}(A_1 \cap A_2 \cap A_3) = 1/36 \neq \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$.

Definition A finite collection of σ-algebras $\mathcal{F}_1, \ldots, \mathcal{F}_n$ is called *independent* if for any $A_1 \in \mathcal{F}_1, \ldots, A_n \in \mathcal{F}_n$,

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n).$$

An infinite collection of σ-algebras is called independent if every finite subcollection is independent.

Definition If X is a random variable and $\mathcal{F}$ is a σ-algebra, we say that X is *$\mathcal{F}$-measurable* if $X^{-1}(B) \in \mathcal{F}$ for every Borel set B, i.e., if X is a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We define $\mathcal{F}(X)$ to be the smallest σ-algebra $\mathcal{F}$ for which X is $\mathcal{F}$-measurable.

In fact,

$$\mathcal{F}(X) = \{X^{-1}(B) : B \text{ Borel } \}.$$

(Clearly $\mathcal{F}(X)$ contains the right hand side, and we can check that the right hand side is a σ-algebra.) If $\{X_\alpha\}$ is a collection of random variables, then $\mathcal{F}(\{\mathcal{X}_\alpha\})$ is the smallest σ-algebra $\mathcal{F}$ such that each X_α is $\mathcal{F}$-measurable.

Remark Probabilists think of a σ-algebra as "information". Roughly speaking, we have the information in the σ-algebra $\mathcal{F}$ if for every event $E \in \mathcal{F}$ we know whether or not the event E has occured. A random variable X is $\mathcal{F}$-measurable if we can determine the value of X by knowing whether or not E has occured for every $\mathcal{F}$-measurable event E. For example suppose we roll two dice and X is the sum of the two dice. To determine the value of X it suffices to know which of the following events occured:

$$E_{j,1} = \{\text{first roll is } j\}, \quad j = 1, \ldots 6,$$

$$E_{j,2} = \{\text{second roll is } j\}, \quad j = 1, \ldots 6.$$

More generally, if X is any random variable, we can determine X by knowing which of the events $E_a := \{X < a\}$ have occured.

Exercise 3.1. *Suppose $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$ is an increasing sequence of σ-algebras and*

$$\mathcal{F}^0 = \bigcup_{n=1}^{\infty} \mathcal{F}_n.$$

Then $\mathcal{F}^0$ is an algebra.

Exercise 3.2. *Suppose X is a random variable and g is a Borel measurable function. Let $Y = g(X)$. Then $\mathcal{F}(Y) \subset \mathcal{F}(X)$. The inclusion can be a strict inclusion.*

The following is a very useful way to check independence.

Lemma 3.3. *Suppose $\mathcal{F}^0$ and $\mathcal{G}^0$ are two algebras of events that are independent, i.e., $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for $A \in \mathcal{F}^0, B \in \mathcal{G}^0$. Then $\mathcal{F} := \sigma(\mathcal{F}^0)$ and $\mathcal{G} := \sigma(\mathcal{G}^0)$ are independent σ-algebras.*

Proof. Let $B \in \mathcal{G}^0$ and define the measure

$$\mu_B(A) = \mathbb{P}(A \cap B), \quad \tilde{\mu}_B(A) = \mathbb{P}(A)\mathbb{P}(B).$$

Note that $\mu_B, \tilde{\mu}_B$ are finite measures and they agree on $\mathcal{F}^0$. Hence (using Cartheodory extension) they must agree on $\mathcal{F}$. Hence $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ for $A \in \mathcal{F}, B \in \mathcal{G}^0$. Now take $A \in F$ and consider the measures

$$\nu_A(B) = \mathbb{P}(A \cap B), \quad \tilde{\nu}_A(B) = \mathbb{P}(A)\mathbb{P}(B).$$

□

If $X_1, \ldots, X_n$ are random variables, we can consider them as a random vector $(X_1, \ldots, X_n)$. The *joint distribution* of the random variables is the measure $\mu = \mu_{X_1,\ldots,X_n}$ on Borel sets in $\mathbb{R}^n$ given by

$$\mu(B) = \mathbb{P}\{(X_1, \ldots, X_n) \in B\}.$$

Each random variable $X_1, \ldots, X_n$ also has a distribution on $\mathbb{R}$, μ_{X_i}; these are called the *marginal distributions*. The *joint distribution function* $F = F_{X_1,\ldots,X_n}$ is defined by

$$F(t_1, \ldots, t_n) = \mathbb{P}\{X_1 \leq t_1, \ldots, X_n \leq t_n\} = \mu[\,(-\infty, t_1] \times \cdots \times (-\infty, t_n]\,].$$

Definition Random variables $X_1, X_2, \ldots, X_n$ are said to be *independent* if any of these (equivalent) conditions hold:

- For all Borel sets $B_1, \ldots, B_n$,

$$\mathbb{P}\{X_1 \in B_1, \ldots, X_n \in B_n\} = \mathbb{P}\{X_1 \in B_1\}\mathbb{P}\{X_2 \in B_2\} \cdots \mathbb{P}\{X_n \in B_n\}.$$

- The σ-algebras $\mathcal{F}(X_1), \ldots, \mathcal{F}(X_n)$ are independent.
- $\mu = \mu_{X_1} \times \mu_{X_2} \times \cdots \times \mu_{X_n}$.
- For all real $t_1, \ldots, t_n$,

$$F(t_1, \ldots, t_n) = F_{X_1}(t_1)F_{X_2}(t_2) \cdots F_{X_n}(t_n).$$

The random variables $X_1, \ldots, X_n$ have a *joint density* (joint probability density function) $f(x_1, \ldots, x_n)$ if for all Borel sets $B \in \mathbb{R}^n$,

$$\mathbb{P}\{(X_1, \ldots, X_n) \in B\} = \int_B f(x_1, \ldots, x_n)\, dx_1\, dx_2\, \cdots\, dx_n.$$

Random variables with a joint density are independent if and only if

- $$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n),$$
  where $f_{X_j}(x_j)$ denotes the density of X_j.

An infinite collection of random variables is called independent if each finite subcollection is independent. The following follows immediately from the definition and Exercise 3.2.

Proposition 3.4. *If $X_1, \ldots, X_n$ are independent random variables and $g_1, \ldots, g_n$ are Borel measurable functions, then $g_1(X_1), \ldots, g_n(X_n)$ are independent.*

Proposition 3.5. *If X, Y are independent random variables with $\mathbb{E}|X| < \infty, \mathbb{E}|Y| < \infty$, then $\mathbb{E}[|XY|] < \infty$ and*

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]. \tag{5}$$

Proof. First consider simple random variables

$$S = \sum_{i=1}^{m} a_i 1_{A_i}, \quad T = \sum_{j=1}^{n} b_j 1_{B_j},$$

where S is $\mathcal{F}(X)$ measurable and T is $\mathcal{F}(Y)$ measurable. Then each A_i is independent of each B_j and (5) can be derived immediately using $\mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i)\,\mathbb{P}(B_j)$. Suppose X, Y are nonnegative random variables. Then there exists simple random variables S_n, measurable with respect to $\mathcal{F}(X)$, that increase to X and simple random variables T_n, measurable with respect to $\mathcal{F}(Y)$, increasing to Y. Note that $S_n T_n$ increases to XY. Hence by the monotone convergence theorem,

$$\mathbb{E}[XY] = \lim_{n\to\infty} \mathbb{E}[S_n T_n] = \lim_{n\to\infty} \mathbb{E}[S_n]\,\mathbb{E}[T_n] = \mathbb{E}[X]\,\mathbb{E}[Y].$$

Finally for general X, Y, write $X = X^+ - X^-, Y^+ - Y^-$. Note that X^+, X^- are independent of Y^+, Y^- (Propostion 3.4) and hence the result follows from linearity of the integral. □

The multiplication rule for expectation can also be viewed as an application of Fubini's Theorem. Let μ, ν denote the distributions of X and Y. If they are independent, then the distribution of (X, Y) is $\mu \times \nu$. Hence

$$\mathbb{E}[XY] = \int_{\mathbb{R}^2} xy\, d(\mu \times \nu) = \int \left[\int xy\, d\mu(x)\right]\, d\nu(y) = \int \mathbb{E}[X] y\, d\nu(y) = \mathbb{E}[X]\,\mathbb{E}[Y].$$

Random variables X, Y that satisfy $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ are called *orthogonal*. The last proposition shows that independent integrable random variables are orthogonal.

Exercise 3.6. *Give an example of orthogonal random variables that are not independent.*

Proposition 3.7. *Suppose $X_1, \ldots, X_n$ are random variables with finite variance. If $X_1, \ldots, X_n$ are pairwise orthogonal, then*

$$\mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n].$$

Proof. Without loss of generality we may assume that all the random variables have zero mean (otherwise, we subtract the mean without changing any of the variances). Then

$$\begin{aligned} \mathrm{Var}\left[\sum_{j=1}^{n} X_j\right] &= \mathbb{E}\left[(\sum_{j=1}^{n} X_j)^2\right] \\ &= \sum_{j=1}^{n} \mathbb{E}[X_j^2] + \sum_{i \neq j} \mathbb{E}[X_i X_j] \\ &= \sum_{j=1}^{n} \mathrm{Var}(X_j) + \sum_{i \neq j} \mathbb{E}[X_i X_j]. \end{aligned}$$

But by orthogonality, if $i \neq j$, $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\mathbb{E}[X_j] = 0$ □

This proposition is most often used when the random variables are independent. It is not true that $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ for all random variables. Two extreme cases are $\mathrm{Var}(X + X) = \mathrm{Var}(2X) = 4\mathrm{Var}(X)$ and $\mathrm{Var}(X - X) = 0$.

Remark The last proposition is really just a generalization of the Pythagorean Theorem. A natural framework for discussing random variables with zero mean and finite variance is the Hilbert space L^2 with the inner product $\langle X, Y \rangle = \mathbb{E}[XY]$. In this case, X, Y are orthogonal iff $\langle X, Y \rangle = 0$.

Exercise 3.8. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the unit interval with Lebesgue measure on the Borel subsets. Follow this outline to show that one can find independent random variables $X_1, X_2, X_3, \ldots$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$, each normal mean zero, variance 1. Let use write $x \in [0, 1]$ in its dyadic expansion*

$$x = .\omega_1\omega_2\omega_3 \cdots, \quad \omega_j \in \{0, 1\}.$$

as in Section 1.

(i) Show that the random variable $X(x) = x$ is a uniform random variable on $[0, 1]$, i.e., has density $f(x) = 1_{[0,1]}$.

(ii) Use a diagonalization process to define $U_1, U_2, \ldots$, independent random variables each with a uniform distribution on $[0, 1]$.

(iii) Show that $X_j = \Phi^{-1}(U_j)$ has a normal distribution with mean zero, variance one. Here Φ is the normal distribution function.

4 Sums of independent random variables

Definition

- A sequence of random variables Y_n converges to Y *in probability* if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\{|Y_n - Y| > \epsilon\} = 0.$$

- A sequence of random variables Y_n converges to Y *almost surely (a.s.)* or *with probability one (w.p.1)* if there is an event E with $P(E) = 1$ and such that for $\omega \in E$, $Y_n(\omega) \to Y(\omega)$.

Remark In other words, "in probability" is the same as "in measure" and "almost surely" is the analogue of "almost everywhere". The term almost surely is not very accurate (something that happens with probability one happens *surely*!) so many people prefer to say "with probability one".

Exercise 4.1. *Suppose $Y_1, Y_2, \ldots$ is a sequence of random variables with $\mathbb{E}[Y_n] \to \mu$ and $\mathrm{Var}[Y_n] \to 0$. Show that $Y_n \to \mu$ in probability. Give an example to show that it is not necessarily true that $Y_n \to \mu$ w.p.1.*

Much of the focus of classical probability is on sums of independent random variables. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and assume that

$$X_1, X_2, \ldots$$

are independent random variables on this probability space. We say that the random variable are *identically distributed* if they all have the same distribution, and we write *i.i.d.* for "independent and identically distributed."

If $X_1, X_2, \ldots$ are independent (or, more generally, orthogonal) random variables each with mean μ and variance σ^2, then

$$\mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{1}{n}\left[\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)\right] = \mu,$$

$$\mathrm{Var}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{1}{n^2}\left[\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)\right] = \frac{\sigma^2}{n}.$$

The equality for the expectations does not require the random variables to be orthogonal, but the expression for the variance requires it. The mean of the weighted average is the same as the mean of one of the random variables, but the standard deviation of the weighted average is $\sigma/\sqrt{n}$ which tends to zero.

Theorem 4.2 (Weak Law of Large Numbers). *If $X_1, X_2, \ldots$ are independent random variables such that $\mathbb{E}[X_n] = \mu$ and $\mathrm{Var}[X_n] \leq \sigma^2$ for each n, then*

$$\frac{X_1 + \cdots + X_n}{n} \to \mu$$

in probability.

Proof. We have

$$\mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \mu, \quad \mathrm{Var}\left[\frac{X_1 + \cdots + X_n}{n}\right] \leq \frac{\sigma^2}{n} \to 0.$$

See Exercise 4.1 □

The *Strong Law of Large Numbers (SLLN)* is a statement about almost sure convergence of the weighted sums. We will prove a version of the strong law here. We start with an easy, but very useful lemma. Recall that the limsup of events is defined

$$\limsup A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m.$$

Probabilists often write $\{A_n \text{ i.o.}\}$ for lim sup; here "i.o." stands for "infinitely often".

Lemma 4.3 (Borel-Cantelli Lemma). *Suppose $A_1, A_2, \ldots$ is a sequence of events.*

(*i*) *If $\sum \mathbb{P}(A_n) < \infty$, then*

$$\mathbb{P}\{A_n \text{ i.o.}\} = \mathbb{P}(\limsup A_n) = 0.$$

(*ii*) *If $\sum \mathbb{P}(A_n) = \infty$, and the $A_1, A_2, \ldots$ are independent, then*

$$\mathbb{P}\{A_n \text{ i.o.}\} = \mathbb{P}(\limsup A_n) = 1.$$

Remark The conclusion in (*ii*) is not necessarily true for dependent events. For example, if A is an event of probability 1/2, and $A_1 = A_2 = \cdots A$, then $\limsup A_n = A$ which has probability 1/2.

Proof.

(*i*) If $\sum \mathbb{P}(A_n) < \infty$,

$$\mathbb{P}(\limsup A_n) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{m=n}^{\infty} A_n\right) \leq \lim_{n\to\infty} \sum_{m=n}^{\infty} \mathbb{P}(A_m) = 0.$$

(*ii*) Assume $\sum \mathbb{P}(A_n) = \infty$ and $A_1, A_2, \ldots$ are independent. We will show that

$$\mathbb{P}[(\limsup A_n)^c] = \mathbb{P}\left[\bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m^c\right] = 0.$$

To show this it suffices to show that for each n

$$\mathbb{P}\left(\bigcap_{m=n}^{\infty} A_m^c\right) = 0,$$

and for this we need to show for each n,

$$\lim_{M\to\infty} \mathbb{P}\left(\bigcap_{m=n}^{M} A_m^c\right) = 0.$$

But by independence,

$$\mathbb{P}\left(\bigcap_{m=n}^{M} A_m^c\right) = \prod_{m=n}^{M} [1 - \mathbb{P}(A_m)] \le \exp\left\{-\sum_{n=m}^{M} \mathbb{P}(A_m)\right\} \to 0.$$

□

Theorem 4.4 (Strong Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be independent random variables each with mean μ. Suppose there exists an $M < \infty$ such that $\mathbb{E}[X_n^4] \le M$ for each n (this holds, for example, when the sequence is i.i.d. and $\mathbb{E}[X_1^4] < \infty$). Then w.p.1*

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu.$$

Proof. Without loss of generality we may assume that $\mu = 0$ for otherwise we can consider the random variables $X_1 - \mu, X_2 - \mu, \ldots$. Our first goal is to show that

$$\mathbb{E}[(X_1 + X_2 + \cdots + X_n)^4] \le 3M\, n^2. \tag{6}$$

To see this, we first note that if $i \notin \{j, k, l\}$,

$$\mathbb{E}[X_i X_j X_k X_l] = \mathbb{E}[X_i]\, \mathbb{E}[X_j X_k X_l] = 0. \tag{7}$$

When we expand $(X_1 + X_2 + \cdots + X_n)^4$ and then use the sum rule for expectation we get n^4 terms. We get n terms of the form $\mathbb{E}[X_i^4]$ and $3n(n-1)$ terms of the form $\mathbb{E}[X_i^2 X_j^2]$ with $i \ne j$. There are also terms of the form $\mathbb{E}[X_i X_j^3], \mathbb{E}[X_i X_j X_k^2]$ and $\mathbb{E}[X_i X_j X_k X_l]$ for distinct i, j, k, l; all of these expectations are zero by (7) so we will not bother to count the exact number. We know that $\mathbb{E}[X_i^4] \le M$. Also (see Exercise 2.2), if $i \ne j$,

$$\mathbb{E}[X_i^2 X_j^2] \le \mathbb{E}[X_i^2]\, \mathbb{E}[X_j^2] \le \left(\mathbb{E}[X_i^4]\, \mathbb{E}[X_j^4]\right)^{1/2} \le M.$$

Since the sum consists of at most $3n^2$ terms each bounded by M, we get (6).

Let A_n be the event

$$A_n = \left\{\left|\frac{X_1 + \cdots + X_n}{n}\right| \ge \frac{1}{n^{1/8}}\right\} = \{|X_1 + X_2 + \cdots + X_n| \ge n^{7/8}\}.$$

By the generalized Chebyshev inequality (Exercise 2.5),

$$\mathbb{P}(A_n) \leq \frac{\mathbb{E}[(X_1 + \cdots + X_n)^4]}{(n^{7/8})^4} \leq \frac{3M}{n^{3/2}}.$$

In particular, $\sum \mathbb{P}(A_n) < \infty$. Hence by the Borel-Cantelli Lemma,

$$\mathbb{P}(\limsup A_n) = 0.$$

Suppose for some ω, $(X_1(\omega) + \cdots + X_n(\omega))/n$ does not converge to zero. Then there is an $\epsilon > 0$ such that

$$X_1(\omega) + X_2(\omega) + \cdots + X_n(\omega) \geq \epsilon n,$$

for infinitely many values of n. This implies that $\omega \in \limsup A_n$. Hence the set of ω at which we do not have convergence has probability zero. □

Remark The strong law of large numbers holds in greater generality than under the conditions given here. In fact, if $X_1, X_2, \ldots$ is any i.i.d. sequence of random variables with $\mathbb{E}(X_1) = \mu$, then the strong law holds

$$\frac{X_1 + \cdots + X_n}{n} \to \mu, \quad \text{w.p.1.}$$

We will not prove this here. However, as the next exercise shows, the strong law does not hold for all independent sequences of random variables with the same mean.

Exercise 4.5. *Let $X_1, X_2, \ldots$ be independent random variables. Suppose that*

$$\mathbb{P}\{X_n = 2^n\} = 1 - \mathbb{P}\{X_n = 0\} = 2^{-n}.$$

Show that that $\mathbb{E}[X_n] = 1$ for each n and that w.p.1,

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \longrightarrow 0.$$

Verify that the hypotheses of Theorem 4.4 do not hold in this case.

Exercise 4.6.

a. Let X be a random variable. Show that $\mathbb{E}[|X|] < \infty$ if and only if

$$\sum_{n=1}^{\infty} \mathbb{P}\{|X| \geq n\} < \infty.$$

b. Let $X_1, X_2, \ldots$ be i.i.d. random variables. Show that

$$\mathbb{P}\{|X_n| \geq n \text{ i.o.}\} = 0$$

if and only if $\mathbb{E}[|X_1|] < \infty$.

We now establish a result known as the Kolomogorov Zero-One Law. Assume $X_1, X_2, \ldots$ are independent. Let $\mathcal{F}_n$ be the σ-algebra generated by $X_1, \ldots, X_n$, and let $\mathcal{G}_n$ be the σ-algebra generated by $X_{n+1}, X_{n+2}, \ldots$. Note that $\mathcal{F}_n$ and $\mathcal{G}_n$ are independent σ-algebras, and

$$F_1 \subset \mathcal{F}_2 \subset \cdots, \quad \mathcal{G}_1 \supset \mathcal{G}_2 \supset \cdots.$$

Let $\mathcal{F}$ be the σ-algebra generated by $X_1, X_2, \ldots$, i.e., the smallest σ-algbra containing the algebra

$$\mathcal{F}^0 \doteq \bigcup_{n=1}^{\infty} \mathcal{F}_n.$$

The *tail σ-algebra* $\mathcal{T}$ is defined by

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \mathcal{G}_n.$$

(The intersection of σ-algberas is a σ-algebra, so $\mathcal{T}$ is a σ-algebra.) An example of an event in $\mathcal{T}$ is

$$\left\{ \lim_{m\to\infty} \frac{X_1 + \cdots + X_m}{m} = 0 \right\}.$$

It is easy to see that this event is in $\mathcal{G}_n$ for each n since

$$\left\{ \lim_{m\to\infty} \frac{X_1 + \cdots + X_m}{m} = 0 \right\} = \left\{ \lim_{m\to\infty} \frac{X_{n+1} + \cdots + X_m}{m} = 0 \right\}.$$

We write $\triangle$ for symmetric difference, $A \triangle B = (A \setminus B) \cup (B \setminus A)$. The next lemma says roughly that $\mathcal{F}_0$ is "dense" in $\mathcal{F}$.

Lemma 4.7. *Suppose $\mathcal{F}^0$ is an algebra of events and $\mathcal{F} = \sigma(\mathcal{F}^0)$. If $A \in \mathcal{F}$, then for every $\epsilon > 0$ there is an $A_0 \in \mathcal{F}^0$ with*

$$\mathbb{P}(A \triangle A_0) < \epsilon.$$

Proof. Let $\mathcal{B}$ be the collection of all events A such that for every $\epsilon > 0$ there is an $A_0 \in \mathcal{F}^0$ such that $\mathbb{P}(A \triangle A_0) < \epsilon$. Trivially, $\mathcal{F}^0 \subset \mathcal{B}$. We will show that $\mathcal{B}$ is a σ-algebra and this will imply that $\mathcal{F} \subset \mathcal{B}$. Obviously, $\Omega \in \mathcal{B}$.

Suppose $A \in \mathcal{B}$ and let $\epsilon > 0$. Find $A_0 \in \mathcal{F}^0$ such that $\mathbb{P}(A \triangle A_0) < \epsilon$. Then $A_0^c \in \mathcal{F}^0$ and

$$\mathbb{P}(A^c \triangle A_0^c) = \mathbb{P}(A \triangle A_0) < \epsilon.$$

Hence $A^c \in \mathcal{B}$.

Suppose $A_1, A_2, \ldots \in \mathcal{B}$ and let $A = \cup A_j$. Let $\epsilon > 0$, and find n such that

$$\mathbb{P}\left[\bigcup_{j=1}^{n} A_j \right] \geq \mathbb{P}[A] - \frac{\epsilon}{2}.$$

For $j = 1, \ldots, n$, let $A_{j,0} \in \mathcal{F}^0$ such that

$$\mathbb{P}[A_j \triangle A_{j,0}] \leq \epsilon 2^{-j-1}.$$

Let $A_0 = A_{1,0} \cup \cdots \cup A_{n,0}$ and note that

$$A \triangle A_0 \subset \left[\bigcup_{j=1}^{n} A_j \triangle A_{j,0}\right] \cup \left[A \setminus \bigcup_{j=1}^{n} A_j\right],$$

and hence $\mathbb{P}(A \triangle A_0) < \epsilon$ and $A \in \mathcal{B}$. □

Theorem 4.8 (Kolomogorov Zero-One Law)**.** *If $A \in \mathcal{T}$ then $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$.*

Proof. Let $A \in \mathcal{T}$ and let $\epsilon > 0$. Find a set $A_0 \in \mathcal{F}^0$ with $\mathbb{P}(A \triangle A_0) < \epsilon$. Then $\mathcal{A}_0 \in \mathcal{F}_n$ for some n. Since $\mathcal{T} \subset \mathcal{G}_n$ and $\mathcal{F}_n$ is independent of $\mathcal{G}_n$, A and A_0 are independent. Therefore $\mathbb{P}(A \cap A_0) = \mathbb{P}(A)\mathbb{P}(A_0)$. Note that $\mathbb{P}(A \cap A_0) \geq \mathbb{P}(A) - \mathbb{P}(A \triangle A_0) \geq \mathbb{P}(A) - \epsilon$. Letting $\epsilon \to 0$, we get $\mathbb{P}(A) = \mathbb{P}(A)\mathbb{P}(A)$. This implies $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$. □

As an example, suppose $X_1, X_2, \ldots$ are independent random variables. Then the probability of the event

$$\left\{\omega : \lim_{n\to\infty} \frac{X_1(\omega) + \cdots + X_n(\omega)}{n} = 0\right\}$$

is zero or one.

5 Central Limit Theorem

In this section, we will use some knowledge of the Fourier transform that we review here. If $g : \mathbb{R} \to \mathbb{R}$, the *Fourier transform* of g is defined by

$$\hat{g}(y) = \int_{-\infty}^{\infty} e^{-ixy}\, g(x)\, dx.$$

We will call g a *Schwartz* function it is C^∞ and all of its derivatives decay at $\pm\infty$ faster than every polynomial. If g is Schwartz, then $\hat{g}$ is Schwartz. (Roughly speaking, the Fourier transform sends smooth functions to bounded functions and bounded functions to smooth functions. Very smooth and very bounded functions are sent to very smooth and very bounded functions.) The inversion formula

$$g(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ixy}\, \hat{g}(y)\, dy \tag{8}$$

holds for Schwartz functions. A straightforward computation shows that

$$\widehat{e^{-x^2/2}} = \sqrt{2\pi} e^{-y^2/2}.$$

We also need

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\,\hat{g}(x)\,dx &= \int_{-\infty}^{\infty} f(x)\left[\int_{-\infty}^{\infty} g(y)\,e^{-ixy}\,dy\right] \\ &= \int_{-\infty}^{\infty} g(y)\left[\int_{-\infty}^{\infty} f(x)\,e^{-ixy}\,dx\right]dy \\ &= \int_{-\infty}^{\infty} \hat{f}(y)\,g(y)\,dy.\end{aligned}$$

This is valid provided that the interchange of integrals can be justified; in particular, it holds for Schwartz f, g. The characteristic function is a version of the Fourier transform.

Definition The *characteristic function* of a a random variable X is the function $\phi = \phi_X : \mathbb{R} \to \mathbb{C}$,

$$\phi(t) = \mathbb{E}[e^{iXt}].$$

Properties of Characteristic Functions

- $\phi(0) = 1$ and for all t
$$|\phi(t)| = \left|\mathbb{E}[e^{iXt}]\right| \le \mathbb{E}\left[|e^{iXt}|\right] = 1.$$

- ϕ is a continuous function of t. To see this note that the collection of random variables $Y_s = e^{isX}$ are dominated by the random variable 1 which has finite expectation. Hence by the dominated convergence theorem,
$$\lim_{s\to t}\phi(s) = \lim_{s\to t}\mathbb{E}[e^{isX}] = \mathbb{E}\left[\lim_{s\to t} e^{isX}\right] = \phi(t).$$
In fact, it is uniformly continuous (Exercise 5.1).

- If $Y = aX + b$ where a, b are constants, then
$$\phi_Y(t) = \mathbb{E}[e^{i(aX+b)t}] = e^{ibt}\mathbb{E}[e^{iX(at)}] = e^{ibt}\phi_X(at).$$

- The function $M_X(t) = e^{tX}$ is often called the *moment generating function* of X. Unlike the characteristic function, the moment generating function does not always exist for all values of t. When it does exist, however, we can use the formal relation $\phi(t) = M(it)$.

- If a random variable X has a density f then
$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\,dx = \hat{f}(-t).$$

- If two random variables have the same characteristic function, then they have the same distribution. This fact is not obvious — this will follow from Proposition 5.6 which gives an inversion formula for the characteristic function similar to (8).

- If $X_1, \ldots, X_n$ are independent random variables, then

$$\begin{aligned} \phi_{X_1+\cdots+X_n}(t) &= \mathbb{E}[e^{it(X_1+\cdots+X_n)}] \\ &= \mathbb{E}[e^{itX_1}]\mathbb{E}[e^{itX_2}]\cdots\mathbb{E}[e^{-tX_n}] = \phi_{X_1}(t)\phi_{X_2}(t)\cdots\phi_{X_n}(t). \end{aligned}$$

- If X is a normal random variable, mean zero, variance 1, then by completing the square in the exponential one can compute

$$\phi(t) = \int_{-\infty}^{\infty} e^{ixt}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,dx = e^{-t^2/2}.$$

  If Y is normal mean μ, variance σ^2, then Y has the same distribution as $\sigma X + \mu$, hence

$$\phi_Y(t) = \phi_{\sigma X+\mu}(t) = e^{i\mu t}\phi_X(\sigma t) = e^{i\mu t}e^{-\sigma^2 t^2/2}.$$

Exercise 5.1. *Show that ϕ is a uniformly continuous function of t.*

Proposition 5.2. *Suppose X is a random variable with characteristic function ϕ. Suppose $\mathbb{E}[|X|] < \infty$. Then ϕ is continuously differentiable and*

$$\phi'(0) = i\mathbb{E}(X).$$

Proof. Note that for real θ

$$|e^{i\theta} - 1| \leq |\theta|.$$

This can be checked geometrically or by noting that

$$|e^{i\theta} - 1| = \left|\int_0^\theta ie^{is}\,ds\right| \leq \int_0^\theta |ie^{is}|\,ds = \theta.$$

We write the difference quotient

$$\frac{\phi(t+\delta) - \phi(t)}{\delta} = \int_{-\infty}^{\infty} \frac{e^{i(t+\delta)x} - e^{itx}}{\delta}\,d\mu(x).$$

Note that

$$\left|\frac{e^{i(t+\delta)x} - e^{itx}}{\delta}\right| \leq \frac{|\delta x|}{|x|} \leq |x|.$$

Saying $\mathbb{E}[|X|] < \infty$ is equivalent to saying that $|x|$ is integrable with respect to the measure μ. By the dominated convergence theorem,

$$\begin{aligned} \lim_{\delta\to 0}\int_{-\infty}^{\infty} \frac{e^{i(t+\delta)x} - e^{itx}}{\delta}\,d\mu(x) &= \int_{-\infty}^{\infty}\left[\lim_{\delta\to 0}\frac{e^{i(t+\delta)x} - e^{itx}}{\delta}\right]\,d\mu(x). \\ &= \int_{-\infty}^{\infty} ixe^{itx}\,d\mu(x). \end{aligned}$$

Hence,

$$\phi'(t) = \int_{-\infty}^{\infty} ixe^{itx}\, d\mu(x),$$

and, in particular,

$$\phi'(0) = i\mathbb{E}[X].$$

□

The following proposition can be proved in the same way.

Proposition 5.3. *Suppose X is a random variable with characteristic function ϕ. Suppose $\mathbb{E}[|X|^k] < \infty$ for some positive integer k. Then ϕ has k continuous derivatives and*

$$\phi^{(j)}(t) = i^j \mathbb{E}[X^j], \quad j = 0, 1 \ldots, k. \tag{9}$$

Formally, (9) can be derived by writing

$$\mathbb{E}[e^{iXt}] = \mathbb{E}\left[\sum_{j=0}^{\infty} \frac{(iX)^j t^j}{j!}\right] = \sum_{j=0}^{\infty} \frac{i^j\, \mathbb{E}(X^j)}{j!} t^j,$$

differentiating both sides, and evaluating at $t = 0$. Proposition 5.3 tells us that the existence of $\mathbb{E}[|X|^k]$ allows one to justify this rigorously up to the kth derivative.

Proposition 5.4. *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with mean μ and variance σ^2 and characteristic function ϕ. Let*

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{\sigma^2 n}},$$

and let ϕ_n be the characteristic function of Z_n. Then for each t,

$$\lim_{n\to\infty} \phi_n(t) = e^{-t^2/2}.$$

Recall that $e^{-t^2/2}$ is the characteristic function of a normal mean zero, variance 1 random variable.

Proof. Without loss of generality we will assume $\mu = 0, \sigma^2 = 1$ for otherwise we can consider $Y_j = (X_j - \mu)/\sigma$. By Proposition 5.3, we have

$$\phi(t) = 1 + \phi'(0)t + \frac{1}{2}\phi''(0)t^2 + \epsilon_t t^2 = 1 - \frac{t^2}{2} + \epsilon_t t^2,$$

where $\epsilon_t \to 0$ as $t \to 0$. Note that

$$\phi_n(t) = \left[\phi(\frac{t}{\sqrt{n}})\right]^n,$$

and hence for fixed t,

$$\lim_{n\to\infty} \phi_n(t) = \lim_{n\to\infty} \left[\phi(\frac{t}{\sqrt{n}})\right]^n = \lim_{n\to\infty} \left[1 - \frac{t^2}{2n} + \frac{\delta_n}{n}\right]^n,$$

where $\delta_n \to 0$ as $n \to \infty$. Once n is sufficiently large that $|\delta_n - (t^2/2)| < n$ we can take logarithms and expand in a Taylor series,

$$\log\left[1 - \frac{t^2}{2n} + \frac{\delta_n}{n}\right] = -\frac{t^2}{2n} + \frac{\rho_n}{n}, \quad n \to \infty$$

where $\rho_n \to 0$. (Since ϕ can take on complex values, we need to take a complex logarithm with $\log 1 = 0$, but there is no problem provided that $|\delta_n - (t^2/2)| < n$.) Therefore

$$\lim_{n\to\infty} \log \phi_n(t) = -\frac{t^2}{2}.$$

□

Theorem 5.5 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables with mean μ and finite variance. Then for every $-\infty < a < b < \infty$,*

$$\lim_{n\to\infty} \mathbb{P}\left\{a \le \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le b\right\} = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$$

Proof. Let ϕ_n be the characteristic function of $n^{-1/2}\sigma^{-1}(X_1 + \cdots + X_n - n\mu)$. We have already shown that for each t, $\phi_n(t) \to e^{-t^2/2}$. It suffices to show that if μ_n is any sequence of distributions such that their characteristic functions ϕ_n converge pointwise to $e^{-t^2/2}$ then for every $a < b$,

$$\lim_{n\to\infty} \mu_n[a, b] = \Phi(b) - \Phi(a),$$

where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy.$$

Fix a, b and let $\epsilon > 0$. We can find a C^∞ function $g = g_{a,b,\epsilon}$ such that

$$0 \le g(x) \le 1, \quad -\infty < x < \infty,$$

$$g(x) = 1, \quad a \le x \le b,$$

$$g(x) = 0, \quad x \notin [a - \epsilon, b + \epsilon].$$

We will show that

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} g(x)\, d\mu_n(x) = \int_{-\infty}^{\infty} g(x) \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx. \tag{10}$$

Since for any distribution μ,

$$\mu[a,b] \le \int_{-\infty}^{\infty} g(x)\, d\mu(x) \le \mu[a-\epsilon, b+\epsilon],$$

we then get the theorem by letting $\epsilon \to 0$.

Since g is a C^∞ function with compact support, it is a Schwartz function. Let

$$\hat{g}(y) = \int_{-\infty}^{\infty} e^{-ixy} g(x)\, dx,$$

be the Fourier transform. Then, using the inversion formula (8),

$$\begin{aligned}
\int_{-\infty}^{\infty} g(x)\, d\mu_n(x) &= \int_{-\infty}^{\infty} \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ixy} \hat{g}(y)\, dy \right] d\mu_n(x) \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} e^{ixy} d\mu_n(x) \right] \hat{g}(y)\, dy \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_n(y) \hat{g}(y)\, dy.
\end{aligned}$$

Since $|\phi_n \hat{g}| \le |\hat{g}|$ and $\hat{g}$ is L^1, we can use the dominated convergence theorem to conclude

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} g(x)\, d\mu_n(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-y^2/2} \hat{g}(y)\, dy.$$

Recalling that $\widehat{e^{-x^2/2}} = \sqrt{2\pi} e^{-y^2/2}$, and that

$$\int \hat{f} g = \int f \hat{g},$$

we get (10). □

The next proposition establishes the uniqueness of characteristic functions by giving an inversion formula. Note that in order to specify a distribution μ, it suffices to give $\mu[a,b]$ at all a, b with $\mu\{a,b\} = 0$.

Proposition 5.6. *Let X be a random variable with distribution μ, distribution function F, and characteristic function ϕ. Then for every $a < b$ such that F is continuous at a and b,*

$$\mu[a,b] = F(b) - F(a) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-iya} - e^{-iyb}}{iy} \phi(y)\, dy.$$

Note that it is tempting to write the right hand side of the equation as

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-iya} - e^{-iyb}}{iy} \phi(y)\, dy, \tag{11}$$

but this integrand is not necessarily integrable on $(-\infty, \infty)$. However, we do know that

$$\left| \frac{e^{-iya} - e^{-iyb}}{iy} \right| |\phi(y)| \leq (b-a), \tag{12}$$

and hence the function is integrable on the bounded interval $[-T, T]$. We can write (11) if we interpret this as the improper Riemann integral and not as the Lebesgue integral on $(-\infty, \infty)$.

Proof. We first fix $T < \infty$. The bound (12) allows us to use Fubini's Theorem:

$$\begin{aligned} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-iya} - e^{-iyb}}{iy} \phi(y)\, dy &= \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-iya} - e^{-iyb}}{iy} \left[\int_{-\infty}^{\infty} e^{iyx}\, d\mu(x) \right] dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_{-T}^{T} \frac{e^{iy(x-a)} - e^{iy(x-b)}}{iy}\, dy \right] d\mu(x) \end{aligned}$$

It is easy to check that for any real c,

$$\int_{-T}^{T} \frac{e^{icx}}{ix}\, dx = 2 \int_{0}^{|c|T} \frac{\sin x}{x}\, dx.$$

We will need the fact that

$$\lim_{T\to\infty} \int_{0}^{T} \frac{\sin x}{x}\, dx = \frac{\pi}{2}.$$

There are two relevant cases.

- If $x - a$ and $x - b$ have the same sign (i.e., $x < a$ or $x > b$),

$$\lim_{T\to\infty} \int_{-T}^{T} \frac{e^{iy(x-a)} - e^{iy(x-b)}}{iy}\, dy = 0.$$

- If $x - a > 0$ and $x - b < 0$,

$$\lim_{T\to\infty} \int_{-T}^{T} \frac{e^{iy(x-a)} - e^{iy(x-b)}}{iy}\, dy = 4 \lim_{T\to\infty} \int_{0}^{T} \frac{\sin x}{x}\, dx = 2\pi.$$

We do not need to worry about what happens at $x = a$ or $x = b$ since μ gives zero measure to these points. Since the function

$$g_T(a, b, x) = \int_{-T}^{T} \frac{e^{iy(x-a)} - e^{iy(x-b)}}{iy}\, dy$$

is uniformly bounded in T, a, b, x, we can use dominated convergence theorem to conclude that

$$\lim_{T\to\infty} \int_{-\infty}^{\infty} \left[\int_{-T}^{T} \frac{e^{iy(x-a)} - e^{iy(x-b)}}{iy}\, dy \right] d\mu(x) = \int 2\pi\, 1_{(a,b)}\, d\mu(x) = 2\pi\mu[a, b].$$

Therefore,

$$\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-iya} - e^{-iyb}}{iy} \phi(y)\, dy = \mu[a, b].$$

□

Exercise 5.7. *A random variable X has a* Poisson distribution *with parameter $\lambda > 0$ if X takes on only nonnegative integer values and*

$$\mathbb{P}\{X = k\} = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots.$$

In this problem, suppose that X, Y are independent Poisson random variables on a probability space with parameters λ_X, λ_Y, respectively.

(i) Find $\mathbb{E}[X], \mathrm{Var}[X], \mathbb{E}[Y], \mathrm{Var}[Y]$.

(ii) Show that $X + Y$ has a Poisson distribution with parameter $\lambda_X + \lambda_Y$ by computing directly

$$\mathbb{P}\{X + Y = k\}.$$

(iii) Find the characteristic function for X, Y and use this to find an alternative derivation of (ii).

(iv) Suppose that for each integer $n > \lambda$ we have independent random variables $Z_{1,n}, Z_{2,n}, \ldots, Z_{n,n}$ with

$$\mathbb{P}\{Z_{j,n} = 1\} = 1 - \mathbb{P}\{Z_{j,n} = 0\} = \frac{\lambda}{n}.$$

Find

$$\lim_{n\to\infty} \mathbb{P}\{Z_{1,n} + Z_{2,n} + \cdots + Z_{n,n} = k\}.$$

Exercise 5.8. *A random variable X is* infinitely divisible *if for each positive integer n, we can find i.i.d. random variables $X_1, \ldots, X_n$ such that $X_1 + \cdots + X_n$ have the same distribution as X.*

(i) Show that Poisson random variables and normal random variables are infinitely divisible.

(ii) Suppose that X has a uniform distribution on $[0, 1]$. Show that X is not infinitely divisible.

(iii) Find all infinitely divisible distributions that take on only a finite number of values.

6 Conditional Expectation

6.1 Definition and properties

Suppose X is a random variable with $\mathbb{E}[|X|] < \infty$. We can think of $\mathbb{E}[X]$ as the best guess for the random variable given no information about the outcome of the random experiment which produces the number X. Conditional expectation deals with the case where one has some, but not all, information about an event.

Consider the example of flipping coins, i.e., the probability space

$$\Omega = \{(\omega_1, \omega_2, \ldots) : \omega_j = 0 \text{ or } 1\}.$$

Let $\mathcal{F}_n$ be the σ-algebra of all events that depend on the first n flips of the coins. Intutitively we think of the σ-algebra $\mathcal{F}_n$ as the "information" contained in viewing the first n flips. Let $X_n(\omega) = \omega_n$ be the indicator function of the event "the nth flip is heads". Let

$$S_n = X_1 + \cdots + X_n$$

be the total number of heads in the first n flips. Clearly $\mathbb{E}[S_n] = n/2$. Suppose we are interested in S_3 but we get to see the flip of the first coin. Then our "best guess" for S_3 depends on the value of the first flip. Using the notation of elementary probability courses we can write

$$\mathbb{E}[S_3 \mid X_1 = 1] = 2, \quad \mathbb{E}[S_3 \mid X_1 = 0] = 1.$$

More formally we can write $\mathbb{E}[S_3 \mid \mathcal{F}_1]$ for the random variable that is measurable with respect to the σ-algebra $\mathcal{F}_1$ that takes on the value 2 on the event $\{X_1 = 1\}$ and 1 on the event $\{X_1 = 0\}$.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{G} \subset \mathcal{F}$ another σ-algebra. Let X be an integrable random variable.

Let us first consider the case when $\mathcal{G}$ is a finite σ-algebra. In this case there exist disjoint, nonempty events $A_1, A_2, \ldots, A_k$ that generate $\mathcal{G}$ in the sense that $\mathcal{G}$ consists precisely of those sets obtained from unions of the events $A_1, \ldots, A_k$. If $\mathbb{P}(A_j) > 0$, we define $\mathbb{E}[X \mid \mathcal{G}]$ on A_j to be the average of X on A_j,

$$\mathbb{E}[X \mid \mathcal{G}] = \frac{\int_{A_j} X \, dP}{\mathbb{P}(A_j)} = \frac{\mathbb{E}[X 1_{A_j}]}{\mathbb{P}(A_j)}, \quad \text{on } A_j. \tag{13}$$

If $\mathbb{P}(A_j) = 0$, this definition does not make sense, and we just let $\mathbb{E}[X \mid \mathcal{G}]$ take on an arbitrary value, say c_j, on A_j. Note that $\mathbb{E}[X \mid \mathcal{G}]$ is a $\mathcal{G}$-measurable random variable and that its definition is unique except perhaps on a set of total probability zero.

Proposition 6.1. *If $\mathcal{G}$ is a finite σ-algebra and $\mathbb{E}[X \mid \mathcal{G}]$ is defined as in (13), then for every event $A \in \mathcal{G}$,*

$$\int_A \mathbb{E}[X \mid \mathcal{G}] \, dP = \int_A X \, dP.$$

Proof. Let $\mathcal{G}$ be generated by $A_1, \ldots, A_k$ and assume that $A = A_1 \cup \cdots \cup A_l$. Then

$$\begin{aligned}
\int_A \mathbb{E}[X \mid \mathcal{G}]\, dP &= \sum_{j=1}^{l} \int_{A_j} \mathbb{E}[X \mid \mathcal{G}]\, dP \\
&= \sum_{j=1}^{l} \mathbb{P}(A_j) \frac{\int_{A_j} X\, dP}{\mathbb{P}(A_j)} \\
&= \sum_{j=1}^{l} \int_{A_j} X\, dP \\
&= \int_A X\, dP.
\end{aligned}$$

This proposition gives us a clue as how to define $\mathbb{E}[X \mid \mathcal{G}]$ for infinite $\mathcal{G}$.

Proposition 6.2. *Suppose X is an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, and $\mathcal{G} \subset \mathcal{F}$ is a σ-algebra. Then there exists a $\mathcal{G}$-measurable random variable Y such that if $A \in \mathcal{G}$,*

$$\int_A Y\, dP = \int_A X\, dP. \tag{14}$$

Moreover, if $\tilde{Y}$ is another $\mathcal{G}$-measurable random variable satisfying (14), then $\tilde{Y} = Y$ a.s.

Proof. The uniqueness is immediate since if $Y, \tilde{Y}$ are $\mathcal{G}$-measurable random variables satisfying (14), then

$$\int_A (Y - \tilde{Y})\, dP = 0.$$

for all $A \in \mathcal{G}$ and hence $Y - \tilde{Y} = 0$ almost surely.

To show existence consider the probability space $(\Omega, \mathcal{G}, \mathbb{P})$. Consider the (signed) measure on $(\Omega, \mathcal{G})$,

$$\nu(A) = \int_A X\, dP.$$

Note that $\nu << P$. Hence by the Radon-Nikodym theorem there is a $\mathcal{G}$-measurable random variable with

$$\nu(A) = \int_A Y\, dP.$$

□

Using the proposition, we can define the conditional expectation $\mathbb{E}[X \mid \mathcal{G}]$ to be the random variable Y in the proposition. It is characterized by the properties

- $\mathbb{E}[X \mid \mathcal{G}]$ is $\mathcal{G}$-measurable.

- For every $A \in \mathcal{G}$,
$$\int_A \mathbb{E}[X \mid \mathcal{G}]\, dP = \int_A X\, dP. \tag{15}$$

The random variable $\mathbb{E}[X \mid \mathcal{G}]$ is only defined up to a set of probability zero.

Exercise 6.3. *This exercise gives an alternative proof of the existence of* $\mathbb{E}[X \mid \mathcal{G}]$ *using Hilbert spaces (but not using Radon-Nikodym theorem). Let* H *denote the real Hilbert space of all square-integrable random variables on* $(\Omega, \mathcal{F}, \mathbb{P})$.

1. *Show that the space of* $\mathcal{G}$*-measurable, square-integrable random varaibles is a closed subspace of* H.

2. *For square-integrable* X, define $\mathbb{E}(X \mid \mathcal{G})$ *to be the Hilbert space projection onto the space of* $\mathcal{G}$*-measurable random variables. Show that* $\mathbb{E}(X \mid \mathcal{G})$ *satisfies* (15).

3. *Show that* $\mathbb{E}(X \mid \mathcal{G})$ *exists for integrable* X. *(Hint: if* $X \geq 0$, *we can approximate* X *from below by simple random variables* X_n. *Define*
$$\mathbb{E}(X \mid \mathcal{G}) = \lim_{n\to\infty} \mathbb{E}(X_n \mid \mathcal{G}).$$
For general X, *write* $X = X^+ - X^-$.*)*

Properties of Conditional Expectation

- $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X]$.

  Proof. This is a special case of (15) where $A = \Omega$.

- If $a, b \in \mathbb{R}$, $\mathbb{E}[aX + bY \mid \mathcal{G}] = a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}]$. (This equality and others below are really equalities up an event of probability zero.)

  Proof. Note that $a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}]$ is $\mathcal{G}$-measurable. Also, if $A \in \mathcal{G}$,
$$\begin{aligned}\int_A (a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}])\, dP &= a\int_A \mathbb{E}[X \mid \mathcal{G}]\, dP + b\int_A \mathbb{E}[Y \mid \mathcal{G}]\, dP \\ &= a\int_A X\, dP + b\int_A Y\, dP \\ &= \int_A (aX + bY)\, dP.\end{aligned}$$
  The result follows by uniqueness of the conditional expectation.

- If X is $\mathcal{G}$ measurable, then $\mathbb{E}[X \mid \mathcal{G}] = X$.

- If $\mathcal{G}$ is independent of X, then $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$. ($\mathcal{G}$ is independent of X if $\mathcal{G}$ is independent of the σ-algebra generated by X. Equivialent, they are indepedent if for every $A \in \mathcal{G}$. 1_A is independent of X.)

  Proof. The constant random variable $\mathbb{E}[X]$ is certainly $\mathcal{G}$-measurable. Also, if $A \in \mathcal{G}$,

$$\int_A \mathbb{E}[X]\, dP = \mathbb{P}(A)\mathbb{E}[X] = \mathbb{E}[1_A]\mathbb{E}[X] = \mathbb{E}[1_A X] = \int_A X\, dP.$$

- If $\mathcal{H} \subset \mathcal{G} \subset F$ are σ-algebras, then

$$\mathbb{E}[X \mid \mathcal{H}] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}]. \tag{16}$$

  Proof. Clearly, $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}]$ is $\mathcal{H}$-measurable. If $A \in \mathcal{H}$, then $A \in \mathcal{G}$, and hence

$$\int_A \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}]\, dP = \int_A \mathbb{E}[X \mid \mathcal{G}]\, dP = \int_A X\, dP.$$

- If Y is $\mathcal{G}$-measurable, then

$$\mathbb{E}[XY \mid \mathcal{G}] = Y\mathbb{E}[X \mid \mathcal{G}]. \tag{17}$$

  Proof. Clearly $Y\mathbb{E}[X \mid \mathcal{G}]$ is $\mathcal{G}$-measurable. We need to show that for every $A \in \mathcal{G}$,

$$\int_A Y\mathbb{E}[X \mid \mathcal{G}]\, dP = \int_A YX\, dP. \tag{18}$$

  We will first consider the case where $X, Y \geq 0$. Note that this implies that $\mathbb{E}[X \mid \mathcal{G}] \geq 0$ a.s., since $\int_A \mathbb{E}[X \mid \mathcal{G}]\, dP = \int_A X\, dP \geq 0$ for every $A \in \mathcal{G}$. Find simple $\mathcal{G}$-measurable random variables $0 \leq Y_1 \leq Y_2 \leq \cdots$ such that $Y_n \to Y$. Then $Y_n X$ converges monotonically to YX and $Y_n\mathbb{E}[X \mid \mathcal{G}]$ converges monotonically to $Y\mathbb{E}[X \mid \mathcal{G}]$. If $A \in \mathcal{G}$ and

$$Z = \sum_{j=1}^{n} c_j 1_{B_j}, \quad B_j \in \mathcal{G},$$

  is a $\mathcal{G}$-measurable simple random variable,

$$\begin{aligned}
\int_A Z\, \mathbb{E}[X \mid \mathcal{G}]\, dP &= \sum_{j=1}^{n} c_j \int_A 1_{B_j}\mathbb{E}[X \mid \mathcal{G}]\, dP \\
&= \sum_{j=1}^{n} c_j \int_{A\cap B_j} \mathbb{E}[X \mid \mathcal{G}]\, dP \\
&= \sum_{j=1}^{n} c_j \int_{A\cap B_j} X\, dP \\
&= \int_A [\sum_{j=1}^{n} c_j 1_{B_j}] X\, dP \\
&= \int_A ZX\, dP.
\end{aligned}$$

Hence (18) holds for simple nonnegative Y and nonnegative X, and hence by the monotone convergence theorem it holds for all nonnegative Y and nonnegative X. For general Y, X, we write $Y = Y^+ - Y^-, X = X^+ - X^-$ and use linearity of expectation and conditional expectation.

If X, Y are random variables and X is integrable we write $\mathbb{E}[X \mid Y]$ for $\mathbb{E}[X \mid \sigma(Y)]$, Here $\sigma(Y)$ denotes the σ-algebra generated by Y. Intuitively, $\mathbb{E}[X \mid Y]$ is the best guess of X given the value of Y. Note that $\mathbb{E}[X \mid Y]$ can be written as $\phi(Y)$ for some function ϕ, i.e., for each possible value of Y there is a value of $\mathbb{E}[X \mid Y]$. Elementary texts often write this as $\mathbb{E}[X \mid Y = y]$ to indicate that for each value of y there is a value of $\mathbb{E}[X \mid Y]$. Similarly, we can define $\mathbb{E}[X \mid Y_1, \ldots Y_n]$.

Example Consider the coin-tossing random variables discussed earlier in the section. What is $\mathbb{E}[X_1 \mid S_n]$? Symmetry tells us that

$$\mathbb{E}[X_1 \mid S_n] = \mathbb{E}[X_2 \mid S_n] = \cdots = \mathbb{E}[X_n \mid S_n].$$

Also, linearity gives

$$\mathbb{E}[X_1 \mid S_n] + \mathbb{E}[X_2 \mid S_n] + \cdots + \mathbb{E}[X_n \mid S_n] = \\ \mathbb{E}[X_1 + \cdots + X_n \mid S_n] = \mathbb{E}[S_n \mid S_n] = S_n.$$

Hence,

$$\mathbb{E}[X_n \mid S_n] = \frac{S_n}{n}.$$

Note that $\mathbb{E}[X_1 \mid S_n]$ is not the same thing at $\mathbb{E}[X_1 \mid X_1, \ldots, X_n] = \mathbb{E}[X_1 \mid \mathcal{F}_n]$ which would be just X_1. The first conditioning is given only the number of heads on the first n flips while the second conditioning is given all the outcomes of the n flips.

Example Suppose X, Y are two random variables with joint density function $f(x, y)$. Let us take our probability space to be $\mathbb{R}^2$, let $\mathcal{F}$ be the Borel sets, and let $\mathbb{P}$ be the measure

$$\mathbb{P}(A) = \int_A f(x, y)\, dxdy.$$

Then the random variables $X(x, y) = x$ and $Y(x, y) = y$ have the joint density $f(x, y)$. Then $\mathbb{E}[Y \mid X]$ is a random varaible with density

$$f(y \mid X) = \frac{f(X, y)}{\int_{-\infty}^{\infty} f(X, z)\, dz}.$$

This formula is valued when the denominator is zero. However, the set of x such that

$$\int_{-\infty}^{\infty} f(x, z)\, dz = 0,$$

is a null set under the measure $\mathbb{P}$ and hence the density $f(y \mid X)$ is well defined almost surely. This density is often called the conditional density of Y given X.

6.2 Martingales

A martingale is a model of a fair game. Suppose we have a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an increasing sequence of σ-algebras

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots$$

Such a sequence is called a *filtration*, and we think of $\mathcal{F}_n$ as representing the amount of knowledge we have at time n. The inclusion of the σ-algebras imply that we do not lose any knowledge that we have gained.

Definition A *martingale (with respect to the filtration* $\{\mathcal{F}_n\}$ is a sequence of integrable random variables M_n such that M_n is $\mathcal{F}_n$ measurable for each n and for all $n < m$,

$$\mathbb{E}[M_m \mid \mathcal{F}_n] = M_n.$$

Remark To prove (11), it suffices to show that for all n,

$$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = M_n.$$

This follows from successive applications of (16). If the filtration is not specified explicitly, then one assumes that one uses the filtration generated by M_n, i.e., $\mathcal{F}_n$ is the σ-algebra generated by $M_1, \ldots, M_n$.

Examples

We list some examples. We leave it as exercises to show that they are martingales.

- If $X_1, X_2, \ldots$ are independent, mean-zero random variables and $S_n = X_1 + \cdots + X_n$, then S_n is a martingale with respect to the filtration generated by X_n.
- If $X_1, X_2, \ldots$ are as above, $\mathbb{E}[X_j^2] = \sigma_j^2 < \infty$ for each j, then

  $$M_n = S_n^2 - \sum_{j=1}^{n} \sigma_j^2$$

  is a martingale with respect to the σ-algebra generated by $X_1, X_2, \ldots$.
- Suppose $X_1, X_2, \ldots$ are as in the first example and $\mathcal{F}_n$ is the filtration generated by $X_1, X_2, \ldots$. We choose $\mathcal{F}_0$ to be the trivial σ-algebra. For each n, let B_n be a bounded random variable that is measurable with repsect to $\mathcal{F}_{n-1}$. We think of B_n as being the "bet" on the game X_n; we can see the results of $X_1, \ldots, X_{n-1}$ before choosing a bet but one cannot see X_n. The total fortune by time n is given by $W_0 = 0$ and

  $$W_n = \sum_{j=1}^{n} B_j \, X_j.$$

  Then W_n is a martingale.

- Suppose $X_1, X_2, \ldots$ are independent coin-flipping random variables, i.e., with distribution
$$\mathbb{P}\{X_j = 1\} = \mathbb{P}\{X_j = -1\} = \frac{1}{2}.$$
A particular case of the last example is where $B_0 = 1$ and $B_n = 2^n$ if $X_1 = X_2 = \cdots = X_{-1} = -1$ and $B_n = 0$ otherwise. In other words, one keeps doubling one's bet until one wins. Note that
$$\mathbb{P}\{W_n = 1\} = 1 - 2^{-n}, \qquad \mathbb{P}\{W_n = 1 - 2^n\} = 2^{-n}. \tag{19}$$
This betting strategy is sometimes called the *martingale betting strategy.*

We want to prove a version of the *optional stopping theorem* which is a mathematical way of saying that one cannot beat a fair game. The last example above shows that we need to be careful — if I play a fair game and keep doubling my bet, I will eventually win and come out ahead (provided that I always have enough money to put down the bet when I lose!)

Definition A *stopping time* with respect to a filtration $\{\mathcal{F}_n\}$ is a random variable taking values in $\{0, 1, 2 \ldots\} \cup \{\infty\}$ such that for each n, the event
$$\{T \le n\} \in \mathcal{F}_n.$$

In other words, we only need the information at time n to know whether we have stopped at time n. Examples of stopping times are:

- Constant random variables are stopping times.
- If M_n is a martingale with respect to $\{\mathcal{F}_n\}$ and $V \subset \mathbb{R}$ is a Borel set, then
$$T = \min\{j : M_j \in V\}$$
is a stopping time.
- If T_1, T_2 are stopping times, then so are $T_1 \wedge T_2$ and $T_1 \vee T_2$.
- In particular, if T is a stopping time then so are $T_n = T \wedge n$. Note that $T_0 \le T_1 \le \ldots$ and $T = \lim T_n$.

Proposition 6.4. *If M_n is a martingale and T is a stopping time with respect to $\{\mathcal{F}_n\}$, then $Y_n = M_{T \wedge n}$ is a martingale with respect to $\{\mathcal{F}_n\}$. In particular, $\mathbb{E}[Y_n] = \mathbb{E}[Y_0] = \mathbb{E}[M_0]$.*

Proof. Note that
$$Y_n = M_n \, 1\{T \ge n\} + \sum_{j=0}^{n-1} M_j \, 1\{T = j\}.$$

From this it is easy to see that Y_n is $\mathcal{F}_n$ measurable and $\mathbb{E}[|Y_n|] < \infty$. Also note that the event $1\{T \geq n+1\}$ is the complement of the event $1\{T \leq n\}$ and hence is $\mathcal{F}_n$-measurable. Therefore, (17) implies

$$\mathbb{E}(M_{n+1}\, 1\{T \geq n+1\} \mid \mathcal{F}_n) = 1\{T \geq n+1\}\, \mathbb{E}(M_{n+1} \mid \mathcal{F}_n) = 1\{T \geq n+1\}\, M_n.$$

Therefore, using linearity,

$$\begin{aligned} \mathbb{E}(Y_{n+1} \mid \mathcal{F}_n) &= 1\{T \geq n+1\}\, M_n + \sum_{j=0}^{n} M_j\, 1\{T = j\} \\ &= 1\{T \geq n\}\, M_n + \sum_{j=0}^{n-1} M_j\, 1\{T = j\} = Y_n. \end{aligned}$$

Examples

- Let $X_1, X_2, \ldots$ be independent, each with $\mathbb{P}\{X_j = 1\} = \mathbb{P}\{X_j = -1\} = 0$. Let $M_n = 1 + X_1 + \cdots + X_n$, N an integer greater than 1, and

  $$T = \min\{j : M_n = 0 \text{ or } M_n = N\}.$$

  T is the first time that a "random walker" starting at 1 reaches 0 or N. Then $M_{n\wedge T}$ is a martingale. In particular,

  $$1 = \mathbb{E}[M_0] = \mathbb{E}[M_{n\wedge T}].$$

  It is easy to check that w.p.1 $T < \infty$. Since $M_{n\wedge T}$ is a sequence of bounded random variables approaching M_t, the dominated convergence theorem implies

  $$\mathbb{E}[M_T] = 1.$$

  Note that

  $$\mathbb{E}[M_T] = N\, \mathbb{P}\{M_T = N\},$$

  and hence we have computed a probability,

  $$\mathbb{P}\{M_T = N\} = \frac{1}{N}.$$

- Let $X_1, X_2, \ldots$ be as above and let

  $$S_n = X_1 + \ldots + X_n.$$

  We have noted that $M_n = S_n^2 - n$ is a martingale. Let N be a positive integer and

  $$T = \min\{n : |S_n| = N\}.$$

Then $\mathbb{E}(M_{n\wedge T})$ is a martingale and hence

$$0 = \mathbb{E}[M_0] = \mathbb{E}[M_{n\wedge T}] = \mathbb{E}[S^2_{n\wedge T} - (n \wedge T)^2].$$

With probability one,

$$\lim_{n\to\infty} M_{n\wedge T} = M_T.$$

We would like to say that

$$\mathbb{E}[M_T] = \lim_{n\to\infty} \mathbb{E}[M_n \wedge T] = 0. \tag{20}$$

We can justify this using the dominated convergence. it is not difficult to show (exercise) that $\mathbb{E}[T] < \infty$, and hence $M_{n\wedge T}$ is dominated by the integrable random variable $N^2 + T$. This justifies (20). Note that this implies that

$$\mathbb{E}[T] = N^2.$$

- Consider the martingale betting strategy as above and let

$$T = \min\{n : X_n = 1\} = \min\{x : W_n = 1\}.$$

Then $W_{T\wedge n}$ is a martingale, and hence $\mathbb{E}[W_{T\wedge n}] = \mathbb{E}[W_0] = 0$. This could be checked easily using (19). Note that $\mathbb{P}\{T < \infty\} = 1$ and

$$\lim_{T\to\infty} W_{T\wedge n} = W_T = 1.$$

Hence,

$$1 = \mathbb{E}[W_T] \neq \lim_{T\to\infty} \mathbb{E}[W_{T\wedge n}] = 0.$$

In this case there is no integrable random variable Y such that $|W_{T\wedge N}| \leq Y$ for all N.

□

Exercise 6.5. *Consider the second example above.*

- *Show that there exists $c < \infty, \rho < 1$ (depending on N) such that for all n,*

$$\mathbb{P}\{T > n\} \leq c\,\rho^n.$$

Use this to conclude that $\mathbb{E}[T] < \infty$.

- *Suppose we changed the problem by assuming that $M_0 = k$ where $k < N$ is a positive integer. In this case, what is $\mathbb{E}[T]$?.*

7 Brownian motion

Brownian motion is a model for random continuous motion. We will consider only the one-dimensional version. Imagine that $B_t = B_t(\omega)$ denotes the position of a random particle at time t. For fixed t, $B_t(\omega)$ is a random variable. For fixed ω, we have a function $t \mapsto B_t(\omega)$. Hence, we can consider the Brownian motion either as a collection of random varibles B_t parametrized by time or as a random function from $[0, \infty)$ into $\mathbb{R}$.

Definition A collection of random variables $B_t, t \in [0, \infty)$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *Brownian motion* if it satisfies the following.

(i) $B_0 = 0$ w.p.1.

(ii) For each $0 \leq s_1 \leq t_1 \leq s_2 \leq t_2 \leq \cdots \leq s_n \leq t_n$, the random variables

$$B_{t_1} - B_{s_1}, \ldots, B_{t_n} - B_{s_n}$$

are independent.

(iii) For each $0 \leq s \leq t$, the random variable $B_t - B_s$ has the same distribution as B_{t-s}.

(iv) W.p.1., $t \mapsto B_t(\omega)$ is a continuous function of t.

It is a nontrivial theorem, which we will not prove here, that states that if B_t is a processes satisfying (i)–(iv) above, then the distribution of $B_t - B_s$ must be normal. (This fact is closely related to but does not immediately follow from the central limit theorem.) Note that

$$B_1 = \sum_{j=1}^{n} \left[B_{\frac{j}{n}} - B_{\frac{j-1}{n}} \right],$$

so a normal distribution is reasonable. However, one cannot conclude normality only from assumptions (i)–(iii), see Exercise 5.8. Since we know Brownian motion must have normal increments (but we do not want to prove it here!), we will modify our definition to include this fact.

Definition A collection of random variables $B_t, t \in [0, \infty)$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *Brownian motion* with drift μ and variance parameter σ^2, if it satisfies the following.

(i) $B_0 = 0$ w.p.1.

(ii) For each $0 \leq s_1 \leq t_1 \leq s_2 \leq t_2 \leq \cdots \leq s_n \leq t_n$, the random variables

$$B_{t_1} - B_{s_1}, \ldots, B_{t_n} - B_{s_n}$$

are independent.

(iii) For each $0 \leq s \leq t$, the random variable $B_t - B_s$ has a normal distribution with mean $\mu(t-s)$ and variance $\sigma^2(t-s)$.

(iv) W.p.1., $t \mapsto B_t(\omega)$ is a continuous function of t.

B_t is called a *standard* Brownian motion if $\mu = 0, \sigma^2 = 1$.

Exercise 7.1. *Suppose B_t is a standard Brownian motion. Show that $\hat{B}_t := \sigma B_t + \mu t$ is a Brownian motion with drift μ and variance parameter σ^2.*

In this section, we will show that Brownian motion exists. We will take any probability space $(\Omega, \mathcal{F}, \mathbb{P})$ sufficiently large that there exist a countable sequence of standard normal random variables $N_1, N_2, \ldots$ defined on the space. Exercise 3.8 shows that the unit interval with Lebesgue measure is large enough. For ease, we will show how to construct a standard Brownian motion $B_t, 0 \le t \le 1$. Let $D_n = \{j/2^n; j = 1, \ldots, 2^n\}, D = \cup_{n=0}^{\infty} D_n$ and suppose that the random variables $\{N_t : t \in D\}$ have been relabeled so they are indexed by the countable set D. The basic strategy is as follows:

- **Step 1**. Use the random variables $N_t, t \in D$ to define random variables $B_t, t \in D$ that satisfy (i)–(iii), restricted to $s, t, s_j, t_j \in D$.
- **Step 2**. Show that w.p.1 the function $t \mapsto B_t, t \in D$ is uniformly continuous.
- **Step 3**. On the event that $t \mapsto B_t, t \in D$ is uniformly continuous, define $B_t, t \in [0, 1]$ by continuity. Show this satisfies the defintion.

7.1 Step 1.

We will recursively define B_t for $t \in D_n$. In our construction we will see that at each stage the random variables

$$\Delta(j, n) := B_{j2^{-n}} - B_{(j-1)2^{-n}}, \quad j = 1, \ldots, 2^n,$$

are independent, normal, variance 2^{-n}. We start with $n = 0$, and set $B_0 = 0, B_1 = N_1$. To do the next step, we use the following lemma.

Lemma 7.2. *Suppose X, Y are independent normal random variables with mean zero, variance σ^2, and let $Z = X + Y, W = X - Y$. Then Z, W are independent random variables with mean zero and variance $2\sigma^2$.*

Proof. For ease, assume $\sigma^2 = 1$. If $V \in \mathbb{R}^2$ is a Borel set, let $\tilde{V} = \{(x, y) : (x+y, x-y) \in V\}$. Then

$$\begin{aligned} \mathbb{P}\{(Z, W) \in V\} &= \mathbb{P}\{(X, Y) \in \tilde{V}\} \\ &= \int_{\tilde{V}} \frac{1}{2\pi} e^{-(x^2+y^2)/2}\, dx\, dy = \int_V \frac{1}{4\pi} e^{-(z^2+w^2)/4}\, dz\, dw. \end{aligned}$$

The second equality uses the change of variables $x = (z + w)/2, y = (z - w)/2$. □

We define

$$B_{1/2} = \frac{1}{2} B_1 + \frac{1}{2} N_{1/2} = \frac{1}{2} N_1 + \frac{1}{2} N_{1/2},$$

and hence

$$B_1 - B_{1/2} = \frac{1}{2} N_1 - \frac{1}{2} N_{1/2}.$$

Since $N_{1/2}, N_1$ are independent, standard normal random variables, the lemma tells us that

$$\Delta(1,1), \Delta(2,1)$$

are independent, mean-zero normals with variance $1/2$. Recursively, if B_t is defined for $t \in D_n$, we define

$$\Delta(2j+1, n+1) = \frac{1}{2}\Delta(j,n) + 2^{-(n+2)/2} N_{(2j+1)2^{-(n+1)}},$$

and hence

$$\Delta(2j+2, n+1) = \frac{1}{2}\Delta(j,n) - 2^{-(n+2)/2} N_{(2j+1)2^{-(n+1)}}.$$

Then continual use of the lemma shows that this works.

7.2 Step 2.

Let

$$M_n = \sup\{|B_s - B_t| : s,t \in D, |s-t| \le 2^{-n}\}.$$

Showing continuity is equivalent to showing that w.p.1, $M_n \to 0$ as $n \to \infty$, or, equivalently, for each $\epsilon > 0$,

$$\mathbb{P}\{M_n > \epsilon \text{ i.o.}\} = 0.$$

It is easier to work with a different quantity:

$$Z(j,n) = \sup\left\{\left|B_t - B_{(j-1)2^{-n}}\right| : t \in D, (j-1)2^{-n} \le t \le j2^{-n}\right\},$$

$$Z_n = \max\{Z(j,n) : j = 1, \ldots, 2^n\}.$$

Note that $M_n \le 3\, Z_n$. Hence, for any $\epsilon > 0$,

$$\mathbb{P}\{M_n > 3\epsilon\} \le \mathbb{P}\{Z_n > \epsilon\} \le \sum_{j=1}^{2^n} \mathbb{P}\{Z(j,n) > \epsilon\} = 2^n\, \mathbb{P}\{Z(1,n) > \epsilon\}.$$

Since B_t has the same distribution as $\sqrt{t}\, B_1$, a little thought will show that the distribution of $Z(1,n)$ is the same as the distribution of

$$2^{-n/2} Z(1,0) = 2^{-n/2} \sup\{|B_t| : t \in D\} = 2^{-n/2} \lim_{m\to\infty} \max\{|B_t| : t \in D_m\}.$$

For any $\delta > 0$, let us consider

$$\mathbb{P}\{\max\{|B_t| : t \in D_m\} > \delta\} \le 2\,\mathbb{P}\{\max\{B_t : t \in D_m\} > \delta\}.$$

We now write

$$\{\max\{B_t : t \in D_m\} > \delta\} = \bigcup_{k=1}^{2^m} A_k,$$

where

$$A_k = A_{k,m} = \{B_{k2^{-m}} > \delta, B_{j2^{-m}} \leq \delta, j < k\}.$$

Note that

$$\mathbb{P}(A_k \cap \{B_1 > \delta\}) \geq \mathbb{P}(A_k \cap \{B_1 - B_{k2^{-m}} \geq 0\}) = \mathbb{P}(A_k)\,\mathbb{P}\{B_1 - B_{k2^{-m}} \geq 0\} = \frac{1}{2}\,\mathbb{P}(A_k).$$

Since the A_k are disjoint, we get

$$\mathbb{P}\{B_1 > \delta\} = \sum_{k=1}^{2^m} \mathbb{P}(A_k \cap \{B_1 > \delta\}) \geq \frac{1}{2} \sum_{k=1}^{2^m} \mathbb{P}(A_k) = \frac{1}{2}\,\mathbb{P}\left\{\max\{B_t : t \in D_m\} > \delta\right\},$$

or, for $\delta \geq \sqrt{\pi/2}$,

$$\mathbb{P}\left\{\max\{B_t : t \in D_m\} > \delta\right\} \leq 2\,\mathbb{P}\{B_1 > \delta\} = \sqrt{\frac{2}{\pi}} \int_\delta^\infty e^{-x^2/2}\,dx \leq \sqrt{\frac{2}{\pi}} \int_\delta^\infty e^{-x\delta/2}\,dx \leq e^{-\delta^2/2}.$$

Since this holds for all m, Putting this all together, we have

$$\mathbb{P}\{M_n > 3\epsilon\} \leq 2^n\,\mathbb{P}\{Z(1,n) > \epsilon\, 2^{-n/2}\} \leq 2^n\, \exp\left\{-\epsilon^2\, 2^{n-1}\right\}.$$

Therefore,

$$\sum_{n=1}^{\infty} \mathbb{P}\{M_n > 3\epsilon\} < \infty,$$

and by the Borel-Cantelli lemma,

$$\mathbb{P}\{M_n > 3\epsilon \text{ i.o.}\} = 0.$$

Exercise 7.3. *Let $\alpha < 1/2$ and*

$$Y = \sup\left\{\frac{|B_t - B_s|}{(t-s)^\alpha} : 0 \leq s < t \leq 1\right\}.$$

Show that w.p,.1., $Y < \infty$. Show that the result is not true for $\alpha = 1/2$.

7.3 Step 3.

This step is straightforward and left as an exercise. One important fact that is used is that if $X = \lim X_n$ and X_n is normal mean zero, variance σ_n^2 and $\sigma_n \to \sigma$, then X is normal mean zero, variance σ^2.

7.4 Wiener measure

Brownian motion $B_t, 0 \leq t \leq 1$ can be considered as a random variable taking values in the metric space $C[0,1]$, the set of continuous functions with the usual sup norm. The distribution of this random variable is a probability measure on $C[0,1]$ and is called *Wiener measure*. If $f \in C[0,1]$ and $\epsilon > 0$, then the Wiener measure of the open ball of radius ϵ about f is

$$\mathbb{P}\{|B_t - f(t)| < \epsilon, 0 \leq t \leq 1\}.$$

Note that $C[0,1]$ with Wiener measure is a probability space. In fact, if we start with this probability space, then it is trivial to define a Brownian motion — we just let $B_t(f) = f(t)$.

8 Elementary probability

In this section we discuss some results that would usually be covered in an undergraduate calculus-based course in probability.

8.1 Discrete distributions

A discrete distribution on $\mathbb{R}$ can be considered as a function $q : \mathbb{R} \to [0,1]$ such that $S_q := \{x : q(x) > 0\}$ is finite or countably infinite, and

$$\sum_{x \in S_q} q(x) = 1.$$

We call S_q the *support* of q. If X is a random variable with distribution q, then

$$\mathbb{P}\{X = x\} = q(x).$$

8.1.1 Bernoulli and Binomial

Let $p \in (0,1)$. Independent repititions of an experiment with probability p of success are called *Bernoulli trials*. The results of the trials can be represented by independent *Bernoulli* random variables

$$X_1, X_2, X_3, \ldots$$

with $\mathbb{P}\{X_j = 1\} = 1 - \mathbb{P}\{X_j = 0\} = p$. If n is a positive integer, then the random variable

$$Y = X_1 + \cdots + X_n$$

is said to have a *binomial distribution* with parameters n and p. Y represents the number of sucesses in n Bernoulli trials each with probability p of success. Note that

$$\mathbb{P}\{Y = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

The binomial coefficient $\binom{n}{k}$ represents the number of ordered n-tuples of 0s and 1s that have exactly k 1s. Clearly $\mathbb{E}[X_j] = p$ and

$$\mathrm{Var}[X_j] = \mathbb{E}[X_j^2] - (\mathbb{E}[X_j])^2 = p - p^2 = p\,(1-p).$$

Hence, if Y has a binomial distribution with parameters n, p,

$$\mathbb{E}[Y] = \sum_{j=1}^{n} \mathbb{E}[X_j] = np, \qquad \mathrm{Var}[Y] = \sum_{j=1}^{n} \mathrm{Var}[X_j] = np(1-p).$$

8.1.2 Poisson distribution

A random variable X has a *Poisson distribution* with parameter $\lambda > 0$, if it has distribution

$$\mathbb{P}\{X = k\} = e^{-\lambda}\,\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

The Taylor series for the exponential shows that

$$\sum_{k=0}^{\infty} \mathbb{P}\{X = k\} = 1.$$

The Poisson distribution arises as a limit of the binomial distribuiton as $n \to \infty$ when the expected number of successes stays fixed at λ. In other words, it is the limit of binomial random variables $Y_n = Y_{n,\lambda}$ with parameters n and λ/n. This limit is easily derived; if k is a nonnegative integer,

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\{Y_n = k\} &= \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n} \left(1 - \frac{\lambda}{n}\right)^{-k} \\
&= \frac{\lambda^k\, e^{-\lambda}}{k!}.
\end{aligned}$$

It is easy to check that

$$\mathbb{E}[X] = \lambda, \quad \mathbb{E}[X(X-1)] = \lambda^2, \quad \mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda.$$

The expectation and variances could be derived from the corresponding quanitities for Y_n,

$$\mathbb{E}[Y_n] = \lambda, \qquad \mathrm{Var}[Y_n] = \lambda\,\left(1 - \frac{\lambda}{n}\right).$$

8.1.3 Geometric distribution

If $p \in (0,1)$, then X has a *geometric distribution* with parameter p if

$$\mathbb{P}\{X = k\} = (1-p)^{k-1}\, p, \quad k = 1, 2, \ldots \tag{21}$$

X represents the number of Bernoulli trials needed until the first success assuming the probability of success on each trial is p. (The event that the kth trial is the first success is the event that the first $k-1$ trials are failures and the last one is a success. From this, (21) follows immediately.) Note that

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k\,(1-p)^{k-1}\, p = -p\,\frac{d}{dp}\sum_{k=1}^{\infty}(1-p)^k = -p\,\frac{d}{dp}\left[\frac{1-p}{p}\right] = \frac{1}{p}.$$

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{k=1}^{\infty} k\,(k-1)\,(1-p)^{k-1}\, p \\ &= p(1-p)\,\frac{d^2}{dp^2}\left[\frac{1-p}{p}\right] = \frac{2(1-p)}{p^2}.\end{aligned}$$

$$\mathbb{E}[X^2] = \frac{2}{p^2} - \frac{1}{p}, \qquad \mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1-p}{p^2}.$$

The geometric distributions can be seen to the only discrete distributions with support $\{1, 2, \ldots\}$ having the *loss of memory property*: for all positive integers j, k,

$$\mathbb{P}\{X \geq j + k \mid X \geq k\} = \mathbb{P}\{X = j\}.$$

The term geometric distribution is sometimes also used for the distribution of $Z = X - 1$,

$$\mathbb{P}\{Z = k\} = (1-p)^k\, p, \quad k = 0, 1, 2, \ldots$$

We can think of Z as the number of failures before the first success in Bernoulli trials. Note that

$$\mathbb{E}[Z] = \mathbb{E}[X] - 1 = \frac{1-p}{p}, \qquad \mathrm{Var}[Z] = \mathrm{Var}[X] = \frac{1-p}{p^2}.$$

8.2 Continuous distributions

The continuous distributions from elementary probability courses are given by specifying the density f or the distribution function F. If X is a random variable with this distribution and $a < b$,

$$\mathbb{P}\{a < X < b\} = \mathbb{P}\{a \leq X \leq b\} = \int_a^b f(x)\,dx = F(b) - F(a).$$

Note that

$$F(b) = \int_{-\infty}^{b} f(x)\,dx,$$

and the fundamental theorem of calculus implies

$$F'(x) = f(x),$$

if f is continuous at x. Hence, it suffices to specify either the density or the distribution.

8.2.1 Normal distribution

A random variable Z has a *standard normal distribution* if it has density

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}, \quad -\infty < x < \infty.$$

To check that this is a density, we can compute

$$\left[\int_0^\infty e^{-x^2/2}\,dx\right]^2 = \int_0^\infty \int_0^\infty e^{-x^2/2}\, e^{-y^2/2}\,dx\,dy = \int_0^\infty \int_0^{2\pi} r e^{-r^2/2}\,d\theta\,dr = 2\pi.$$

It is easy to check that

$$\mathbb{E}[Z] = 0, \qquad \mathrm{Var}[Z] = \mathbb{E}[Z^2] = 1.$$

The distribution function of Z is often denoted by Φ,

$$\Phi(x) = \int_0^x \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}\,dt.$$

The normal distribution with mean μ and variance σ^2, often denoted $N(\mu, \sigma^2)$ is the distribution of $X = \sigma\, Z + \mu$. It is easily checked that the density is

$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}.$$

Indeed, if F denotes the distribution function for X,

$$F(x) = \mathbb{P}\{\sigma Z + \mu \le x\} = \mathbb{P}\left\{Z \le \frac{x-\mu}{\sigma}\right\} = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Hence,

$$F'(x) = \frac{1}{\sigma}\,\Phi'\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma}\,\frac{1}{\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2}.$$

An important property of normal distributions is the following.

- If $X_1, \ldots, X_n$ are independent random variables where X_j has a $N(\mu_j, \sigma_j^2)$ distribution, then $X_1 + \ldots + X_n$ has a $N(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots \sigma_n^2)$ distribution.

8.2.2 Exponential distribution

A random variable X has an *exponential distribution* with parameter $\lambda > 0$ if it has density

$$f(x) = \lambda\, e^{-\lambda x}, \quad x > 0,$$

or, equivalently, if

$$\mathbb{P}\{X \geq t\} = e^{-\lambda t}, \quad t \geq 0.$$

Note that

$$\mathbb{E}[X] = \int_0^\infty \lambda\, x\, e^{-\lambda x}\, dx = \frac{1}{\lambda}, \qquad \mathbb{E}[X^2] = \int_0^\lambda x^2\, e^{-\lambda x}\, dx = \frac{2}{\lambda^2},$$

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{\lambda^2}.$$

The exponential distribution has the *loss of memory property*, i.e., if $G(t) = \mathbb{P}\{X \geq t\}$, then

$$\mathbb{P}\{X \geq s + t \mid X \geq s\} = \mathbb{P}\{X \geq t\}, \quad s, t \geq 0.$$

It is not difficult to show that this property characterizes the exponential distribution. If $G(t) = \mathbb{P}\{X \geq t\}$, then the above equality translates to

$$G(s + t) = G(s)\, G(t),$$

and the only continuous functions on $[0, \infty)$ satisfying the condition above with $G(0) = 1$ and $G(\infty) = 0$ are of the form $G(t) = e^{-\lambda t}$ for some $\lambda > 0$.

8.2.3 Gamma distribution

A random variable X has a *Gamma distribution* with parameters $\lambda > 0$ and $\alpha > 0$ if it has density

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\, x^{\alpha - 1}\, e^{-\lambda x}, \quad x > 0.$$

Here

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

is the usual Gamma function.

Example

- If $\alpha = 1$, then the Gamma distribution is the exponential distribution.
- If $\alpha = n$ is a positive integer, and $X_1, \ldots, X_n$ are independent random variables, each exponential with parameter λ, then $X_1 + \cdots + X_n$ has a Gamma distribution with parameters λ and α. We omit the derivation of this.

- More generally, if $X_1, \ldots, X_n$ are independent random variables, each Gamma with parameters λ and α_j, then $X_1 + \cdots + X_n$ has a Gamma distribution with parameters λ and $\alpha_1 + \cdots + \alpha_n$.

- If Z has a standard normal distribution, then Z^2 has a Gamma distribution with parameters $\lambda = 1/2, \alpha = 1/2$. To see this, let F denote the distribution of Z^2. Then,

$$F(x) = \mathbb{P}\{Z^2 \le x\} = \mathbb{P}\{|Z| \le \sqrt{x}\} = 2\int_0^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

  Hence,

$$f(x) = F'(x) = \frac{1}{\sqrt{2\pi x}}\, e^{-x/2}.$$

If n is a positive integer, then the Gamma distribution with parameters $\lambda = 1/2$ and $\alpha = n/2$ is called the χ^2-distribution with n degrees of freedom. It is the distribution of

$$Z_1^2 + \cdots + Z_n^2$$

where $Z_1, \ldots, Z_n$ are independent standard normal random variables. This distribution is very important in statistics.

8.3 Poisson process

Let $T_1, T_2, \ldots$ be independent random variables, each with an exponential distribution with parameter λ. Let $\tau_0 = 0$ and for $k > 0$, $\tau_k = T_1 + \cdots + T_k$. Define the random variables N_t by

$$N_t = k, \quad \text{if} \quad \tau_k \le t < \tau_{k+1}.$$

Then N_t is a *Poisson process* with parameter λ. We think of τ_k as the time until k events have occured where events happen "randomly" with an average rate of λ events per time unit. The random variable N_t denotes the number of events that have occurred by time t.

The process N_t has the following properties.

- For each $0 \le s < t$, $N_t - N_s$ has a Poisson distribution with mean $\lambda(t - s)$.

- If $0 < t_1 < t_2 < \ldots < t_n$, then the random variables

$$N_{t_1}, \;\; N_{t_2} - N_{t_1}, \;\; \ldots, \;\; N_{t_n} - N_{t_{n-1}}$$

  are independent.

- With probability one $t \mapsto N_t$ is nondecreasing, right continuous, and all jumps are of size one. For every $t > 0$,

$$\mathbb{P}\{N_{t+\Delta t} = k + 1 \mid N_t = k\} = \lambda\, \Delta t + o(\Delta t), \quad \delta t \to 0 + .$$