Probability Theory

ST318, academic year 2013-2014

Wilfrid Kendall

Department of Statistics, University of Warwick

17th March 2014

Overall table of contents

1. Introduction
2. Probability as measure: the basic model (prerequisite)
3. Random Variables (prerequisite)
4. Independence
5. Expectation and Modes of Convergence
6. A Tale of Two Limit Theorems
7. Conditional Expectation
8. Martingales

Administrative details (I)

Times: This module takes place in Term 2.
Timetable in 2013/2014:
Mon 1300 (S0.21), Tue 1100 (MS.03), Thu 0900 (MS.03).

Content: Independence, conditioning, zero-one laws, modes of convergence, basic martingale theory.

Aims: To deliver a rigorous presentation of some fundamental results in measure theoretic probability and introduction to the theory of discrete-time martingales, hence providing a firm basis for advanced work on probability and its applications.

Administrative details (II)

Assumes: ST213 Mathematics of Random Events or MA359 Measure Theory (3rd year).

Objectives: The objectives of the course are as follows: at the end of the course a successful student will:

- Understand the ideas relating to independence and zero-one laws and be able to apply these ideas in simple contexts;
- Understand the different modes of convergence for sequences of random variables (more generally random elements) and the relationship between these different modes;
- Be able to state and prove the Central Limit Theorem and understand how this result can be applied;
- Understand some basic results on discrete time martingales and (time permitting) how this theory can be used to prove Kolmogorov's Strong Law of Large Numbers.

Assessment: 100 % by 2-hour exam.

Reading and other resources

1. Books:
   1. Williams (1991) is a famous text on advanced probability; we will from time to time refer to major proofs in there, and it would be a very useful supplementary source.
   2. Jacod and Protter (2003) is an alternative to Williams (1991) written in a rather different style.
2. Online notes: The notes which you are reading now! Prerequisite material is summarized in the sections 2 and 3 – or read the old online lecture notes for ST213 at
   `www2.warwick.ac.uk/fac/sci/statistics/courses/modules/year2/st213/st213.pdf`

1: Explaining terms in the module objectives

Important thoughts:

If I have seen farther than others, it is because I was standing on the shoulders of giants.
Isaac Newton, 1642-1727

Mathematicians stand on each other's shoulders.
Karl Friedrich Gauss, 1777-1855

In the sciences, we are now uniquely privileged to sit side by side with the giants on whose shoulders we stand.
Gerald Holton, Harvard University

If I have not seen as far as others, it is because giants were standing on my shoulders.
Hal Abelson, 1909-1994

Explaining terms in the module objectives

Table of contents of this section

1 Introduction
- Independence
- Zero-one laws
- Modes of convergence
- Conditional expectation
- Martingales

We start by taking a quick peek at some major themes of the lectures, to get a sense of where everything is going.

Independence

- Notion of independence of two events, or of two random variables, is fairly easy. (2 dice throws.)
- Independence of given random variable from large, possibly infinite, number of random variables? (3 cartesian coordinates of location of random object.)
- Independence of more exotic objects? (Rotations in 3-space.)
- Independence or lack of independence raises subtle issues.
  Puzzle: Savage pirate takes 100 prisoners.
  Sets a task: 100 numbered boxes concealing prisoner names.
  Without being able to communicate results, each prisoner may open 50 of 100 boxes in turn. If every prisoner's choice includes the box with their name in, then all prisoners go free.
  If just one fails, then all are executed.
  How can prisoners arrange among themselves to obtain a non-negligible probability of freedom?

Zero-one laws

- Take $X_1, X_2, \ldots,$ independent and identically distributed Bernoulli random variables (values 0 or 1), with $\mathbb{P}[X_i = 1] = \frac{1}{2}$. Suppose I win £$a_i$ when $X_i = 1$, lose £a_i when $X_i = 0$.
- What can I say about long-run profit/loss $Y_n = a_1(2X_1 - 1) + \ldots + a_n(2X_n - 1)$ for large n?
- A zero-one law says either Y_n converges to a limit with probability one, or it does not with probability one. No half-way houses!
- Actually one can show convergence happens if and only if $\sum a_i^2 < \infty$: consider $a_i = 1/i$: convergence can happen even if $\sum_i a_i$ does not converge!

Modes of convergence

- Let $X_1, X_2, \ldots$ be independent random variables with common mean μ and finite variance σ^2. Then (Weak Law of Large Numbers) the sample mean converges in probability:

$$Y_n \quad = \quad \frac{1}{n}(X_1 + \ldots + X_n) \quad \to \quad \mu \qquad \text{in probability}.$$

- Often more is true: the sequence $(Y_n : n \geq 1)$ converges to μ with probability 1 (almost-sure convergence: Strong Law of Large Numbers).
- Convergence is about approximation , in the sense of some penalty being small. Convergence in probability fails to penalize size of penalty, so use convergence in L^p: $Y_n \to Z$ in L^p if $\mathbb{E}[|Y_n - Z|^p] \to 0$.
- Sometimes we just want probabilities to be close, as in Central Limit Theorem; then we need weak convergence.

Conditional expectation

- Given a sequence $X_1, X_2, \ldots, X_n$ (for example, my cumulative winnings/losses at the Casino Royale), how to predict my winnings X_{n+1} at play $n+1$? If prediction penalty is $|m_{n+1} - X_{n+1}|^2$ then "best" prediction is the conditional expectation

$$m_{n+1} \quad = \quad \mathbb{E}[X_{n+1} \mid X_n, \ldots, X_1].$$

- Notice that the prediction m_{n+1} is obviously random (it depends on the outcomes $X_1, X_2, \ldots, X_n$).
- Conventionally I have to use two different formulae for m_{n+1}, depending on whether I work in a discrete or a continuous setting. Can these formulae be unified?

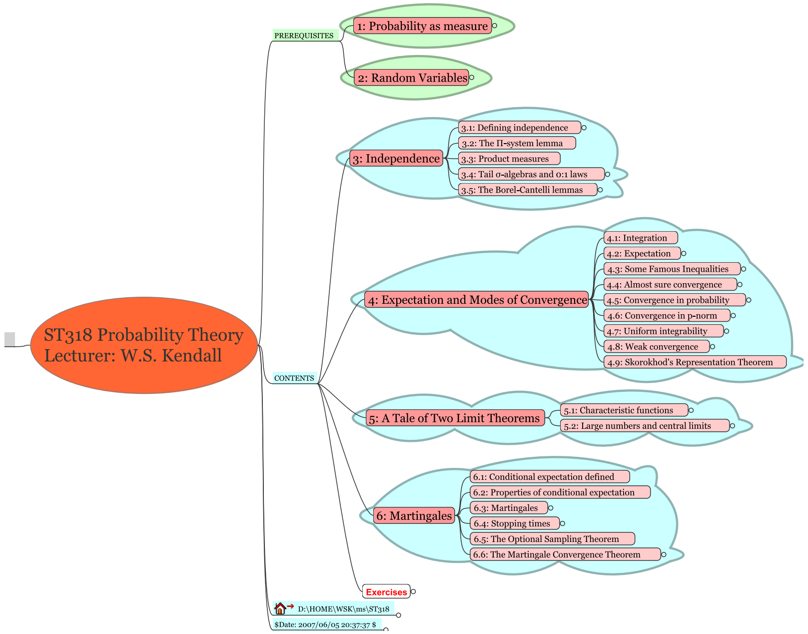
Martingales

- Suppose in the previous setting that

$$m_{n+1} \quad = \quad \mathbb{E}[X_{n+1} \mid X_n, \ldots, X_1] \quad = \quad X_n.$$

- (Probably never holds for games played at the Casino Royale.)
- Such sequences $X_1, X_2, \ldots, X_n$ are called **martingales**, and are central in probability theory.
- Example: successive locations of a simple symmetric random walk.
- Example: successive conditional probabilities of a future event as more and more information is released.

Mindmap of module:



2: Probability as measure: the basic model

One way to make your old car run better is to look up the price of a new model.

The basic model of probability theory

Table of contents of this section

2 Probability as measure: the basic model (prerequisite)
- Introduction
- Measurable space
- Measure
- Uniqueness, extensions and measure
- Model for experiment
- Almost Surely (a.s.)

We now run quickly through some important measure-theoretic fundamentals from ST213 (see also MA359). If the pace seems too fast **then revise the lecture notes for these modules** .

Introduction: The fundamental scenario

We perform an experiment with random outcomes:

- tossing a coin;
- rolling a die;
- measuring the length of a journey between two locations;
- ...

The result of the experiment depends on unknown circumstances, so we consider the chances of obtaining a chosen outcome, rather than compute exactly what the outcome should be.

Some basic probability notation:

- Ω is the set of all possible outcomes of an experiment;
- Take the finite case to start with: $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$.
- The probability of occurrence of any subset of Ω is equal to the sum of probabilities of occurrence of each outcome contained in that subset.

In other words, for $A \in \Omega$

$$\mathbb{P}[A] = \sum_{\omega_i \in A} \mathbb{P}[\{\omega_i\}] .$$

However, if Ω consists of a non-countable infinite number of points then the definition of an event as any subset of Ω becomes unhelpful and the formula for $\mathbb{P}[A]$ above becomes unworkable .

Example 1

Consider an experiment which chooses a point at random on the surface of the unit sphere. Can we define, for all sets F, the probability of the chosen point lying in F as the area of F divided by the area of the sphere?
The answer is no, because there are subsets F which are not "Lebesgue measurable" (one cannot assign areas to such a subset in a manner consistent with the usual additivity requirements for areas).
Sketch of proof using infinite group theory:
http://en.wikipedia.org/wiki/Banach%E2%80%93Tarski_paradox

We need to clarify exactly which sets can serve as events without leading to such troubles, and the probability rules which are to be applied to them.

Measurable space

The eternal silence of these infinite spaces fills me with dread.
Blaise Pascal (1623 - 1662)

Let S be a set, such as the sample space for a probability experiment, or the surface of a sphere, or . . . .

Definition 2.1 (σ-algebra on S)

A collection Σ of subsets of S is called a σ-algebra[a] on S if

1. $S \in \Sigma$
2. $A \in \Sigma$ implies $A^c \in \Sigma$,
3. if $A_1, A_2, \ldots$ is a finite or countably infinite sequence of sets in Σ, then $\bigcup_i A_i \in \Sigma$.

[a] Variant notation: some call this a σ-field.

In property (3) the sets $A_1, A_2, \ldots$ may or may not overlap, may or may not be empty or the whole set S, but there must be only countably many of them!

We occasionally use the following terminology:
An algebra has the same properties with ‘countable’ replaced by ‘finite’ in Lemma 2.1 property (3).
Prove these simple consequences of Definition 2.1 for yourself:

Exercise 2.2 (σ-algebra basics)

Suppose $A_i \in \Sigma$ for $i = 1, 2, \ldots$.

- *Use properties (1) and (2) to show that $\emptyset \in \Sigma$.*
- *Use properties (2) and (3) to show that $\bigcap_i A_i = \left(\bigcup_i A_i^c\right)^c \in \Sigma$.*

Exercise 2.3 (intersections of σ-algebras)

Show that the intersection of an arbitrary collection of σ-algebras is a σ-algebra.

A pair (S, Σ) where S is a set and Σ is a σ-algebra on S is called a measurable space. An element of Σ is called a Σ-measurable subset of S.

Definition 2.4 (σ-algebra generated by a class of subsets)

Let $\mathfrak{C}$ be a class of subsets of S. Then $\sigma(\mathfrak{C})$, the σ-algebra generated by $\mathfrak{C}$, is the intersection of all σ-algebras on S which have $\mathfrak{C}$ as a subclass.

Exercise 2.5 (generating σ-algebras works as intended)

Show that $\sigma(\mathfrak{C})$ is indeed a σ-algebra.

Every topological space (real line, plane, space, sphere, ...) has a σ-algebra which relates naturally to the space.

Definition 2.6 (Borel σ-algebra)

Let S be a topological space. Then the Borel σ-algebra on S is the σ-algebra $\mathfrak{B}(S)$ generated by the family of all open subsets of S.

$$\mathfrak{B}(S) \quad = \quad \sigma(\textit{open sets of } S)\,.$$

It is standard to use the shorthand notation $\mathfrak{B} = \mathfrak{B}(\mathbb{R})$.

Exercise 2.7 (Intervals generate the Borel σ-algebra)

Prove the following equivalences:

$$\begin{aligned}\sigma(\{(-\infty, x) : x \in \mathbb{R}\}) \;&=\; \sigma(\{(-\infty, x] : x \in \mathbb{R}\}) \\ &=\; \sigma(\{(a, b) : a, b \in \mathbb{R}\}) \;=\; \sigma(\{[a, b) : a, b \in \mathbb{R}\})\,.\end{aligned}$$

The σ-algebra $\mathfrak{B}$ is the most important of all σ-algebras. Almost every subset of $\mathbb{R}$ that is in everyday use is in $\mathfrak{B}$.

However with some effort one can construct subsets of $\mathbb{R}$ which are not in $\mathfrak{B}$! The key fact is the so-called “image catastrophe”: if A is a Borel subset of the plane then its projection onto the x-axis need *not* be Borel. Hoffmann-Jørgensen (1994, §1.47) provides an interesting brief discussion of the history of this.

Elements of $\mathfrak{B}$ can be quite complicated. However, and fortunately, much of the time all we need to know about $\mathfrak{B}$ is contained in the statement that

$$\mathfrak{B} \quad = \quad \sigma(\{(-\infty, x] : x \in \mathbb{R}\})\,.$$

In words, $\mathfrak{B}$ is the smallest σ-algebra to contain all closed intervals which are half-infinite on the left.

Measure

Keep five yards from a carriage, ten yards from a horse,
and a hundred yards from an elephant;
but the distance one should keep from a wicked man cannot be measured.
Indian Proverb

Let S be a set, let Σ_0 be an algebra on S and let μ_o be a non-negative set function $\mu_0 : \Sigma_0 \to [0, \infty]$. Then μ_0 is called additive if $\mu_0(\emptyset) = 0$ and for $F, G \in \Sigma_0$,

$$F \cap G = \emptyset \quad \text{implies} \quad \mu_0(F \cup G) = \mu_0(F) + \mu_0(G) .$$

Exercise 2.8 (Total additivity for additive set functions)

Show $\mu_0(F \cup G) = \mu_0(F) + \mu_0(G) - \mu_0(F \cap G)$ *for* $F, G \in \Sigma_0$.

The map μ_0 is called countably additive if $\mu_0(\emptyset) = 0$ and whenever $(F_n, n \in \mathbb{N}$ is a sequence of disjoint sets in Σ_0 with union $F = \bigcup F_n$ in Σ_0 (needed if Σ_0 is not a σ-algebra!), then

$$\mu_0(F) \quad = \quad \sum_n \mu_0(F_n) .$$

Definition 2.9 (measure, measure space)

Let (S, Σ) be a measurable space, that is Σ is a σ-algebra on S. A map

$$\mu : \Sigma \to [0, \infty]$$

is called a measure on (S, Σ) if μ is countably additive. The triple (S, Σ, μ) is called a measure space.

Let (S, Σ, μ) be a measure space. Then μ (or indeed the measure space (S, Σ, μ)) is called:

- finite if $\mu(S) < \infty$
- σ-finite if there is a sequence $(S_n : n \in \mathbb{N})$ of elements of Σ such that $\mu(S_n) < \infty$, $n \in \mathbb{N}$, and $\bigcup S_n = S$.

A measure μ is called a probability measure if $\mu(S) = 1$, and (S, Σ, μ) is then called a probability triple or a probability space. An element F of Σ is called μ-null if $\mu(F) = 0$. A statement $\mathcal{S}$ about points s of S is said to hold μ-almost everywhere if

1. it is measurable: $F = \{s : \mathcal{S}(s) \text{ is false}\} \in \Sigma$ and
2. it is μ-null: $\mu(F) = 0$.

Exercise 2.10

Suppose an experiment involves picking a number from $\mathbb{R}$ *such that the probability of getting any particular number is* 0 *(e.g.: using normal distribution). Consider the event of getting a number which in decimal expansion terminates after 32 decimal places. Show this is a null event.*

Π-systems

Have no friends not equal to yourself.
Confucius

It is usually impossible to write down the typical element of a σ-algebra. However, as we will see, Π-systems are much easier to work with.

Definition 2.11 (Π-system)

Let S be a set. A family $\mathcal{I}$ of subsets of S is called a Π-system if it is stable under finite intersections:

$$I_1, I_2 \in \mathcal{I} \quad \text{implies} \quad I_1 \cap I_2 \in \mathcal{I}\,.$$

Exercise 2.12 (An important Π-system)

Consider the family $\{(-\infty, u] \text{ for } u \in \mathbb{R}\}$ of closed intervals infinite to the left. Show this is a Π-system.

Lemma 2.13 (Uniqueness Lemma for Π-systems)

Let S be a set, let $\mathcal{I}$ be a Π-system on S and let $\Sigma = \sigma(\mathcal{I})$. Suppose that μ_1 and μ_2 are measures on (S, Σ) such that $\mu_1(S) = \mu_2(S) < \infty$ and $\mu_1 = \mu_2$ on $\mathcal{I}$. Then

$$\mu_1 \quad = \quad \mu_2 \qquad \text{on} \quad \Sigma\,.$$

Proof.

Omitted: see Williams (1991, Section 1.6 and Appendix A). □

Corollary 2.14

If two probability measures agree on a Π-system, then they agree on the σ-algebra generated by that Π-system.

Theorem 2.15 (Carathéodory's Extension Theorem)

Let S be a set, Σ_0 an algebra on S and $\Sigma = \sigma(\Sigma_0)$. If μ_0 is a countably additive map $\mu_0 : \Sigma_0 \to [0, \infty]$, then there exists a measure μ on (S, Σ), the Carathéodory extension, such that

$$\mu \quad = \quad \mu_0 \qquad \text{on } \Sigma_0\,.$$

If $\mu_0(S) < \infty$ then this extension is unique.

Proof.

Omitted: see Williams (1991, Section 1.7 and Appendix A). □

Example 2 (Lebesgue measure on the unit interval)

Let $S = (0,1]$ and let Σ_0 be a collection of subsets of the form

$$F = (a_1, b_1] \cup \cdots \cup (a_r, b_r],$$

where $r \in \mathbb{N}$ and $0 \le a_1 \le b_1 \le a_2 \le \cdots \le a_r \le b_r \le 1$. Then Σ_0 is an algebra on $(0,1]$ and $\sigma(\Sigma_0) = \mathfrak{B}(0,1]$. For $F \in \Sigma_0$ define measure $\mu_0(F) = \sum_{k \le r}(b_k - a_k)$. Then μ_0 is well-defined and additive on Σ_0. Also, μ_0 is countably additive on Σ_0 (Williams 1991, Appendix A). Therefore, by Carathéodory's Extension Theorem 2.15 there exists a unique measure μ on $((0,1], \mathfrak{B}(0,1])$ extending μ_0 on Σ_0. This μ is called Lebesgue measure on $((0,1], \mathfrak{B}(0,1])$ and is denoted by Leb. It makes precise the concept of length, that is $\text{Leb}[a,b] = b - a$.

Lemma 2.16 (Elementary inequalities)

Let (S, Σ, μ) be a measure space. Then

1. $\mu(A \cup B) \le \mu(A) + \mu(B)$, *for* $A, B \in \Sigma$,
2. $\mu(\bigcup_{i \le n} F_i) \le \sum_{i \le n} \mu(F_i)$, *for* $F_1, F_2, \ldots, F_n \in \Sigma$.

   In addition, if $\mu(S) < \infty$,
3. $\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$, *for* $A, B \in \Sigma$,
4. *(inclusion-exclusion formula): for* $F_1, F_2, \ldots, F_n \in \Sigma$,

$$\begin{aligned}\mu\Big(\bigcup_{i \le n} F_i\Big) &= \sum_{i \le n} \mu(F_i) - \sum\sum_{i<j \le n} \mu(F_i \cap F_j) \\ &+ \sum\sum\sum_{i<j<k \le n} \mu(F_i \cap F_j \cap F_k) - \cdots + (-1)^{n-1} \mu(F_1 \cap F_2 \cap \cdots \cap F_n).\end{aligned}$$

Exercise 2.17

Prove Lemma 2.16.

Lemma 2.18 (Monotone-convergence properties of measures)

Let (S, Σ, μ) be a measure space. Then

(a) *if $F_n \in \Sigma$, $n \in \mathbb{N}$, and $F_n \uparrow F$, then $\mu(F_n) \uparrow \mu(F)$.*

(b) *if $G_n \in \Sigma$, $n \in \mathbb{N}$, $G_n \downarrow G$ and $\mu(G_k) < \infty$ for some k, then $\mu(G_n) \downarrow \mu(G)$.*

(c) *the union of a countable number of μ-null sets is a μ-null set.*

Exercise 2.19

Prove Lemma 2.18.

Example 3 (what can go wrong with monotonicity ...)

For $n \in \mathbb{N}$, let $H_n = (n, \infty)$.
Then $\text{Leb}(H_n) = \infty$ for all $n \in \mathbb{N}$, but $H_n \downarrow \emptyset$.

Model for experiment

Why are many scientists using lawyers for medical experiments instead of rats?

a) *There are more lawyers than rats.*
b) *The scientists don't become as emotionally attached to them.*
c) *There are some things that even rats won't do for money.*

A model for an experiment involving randomness is a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. Ω is called a sample space, a point $\omega \in \Omega$ is called a sample point. The σ-algebra $\mathcal{F}$ of subsets of Ω is called the family of events so that an *event* is an $\mathcal{F}$-measurable subset of Ω, i.e. an element of $\mathcal{F}$. Finally, by the definition of probability triple, $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$.
If $F \in \mathcal{F}$ then $\mathbb{P}[F]$ represents the probability that a randomly chosen point ω of Ω belongs to F.

Example 4

Consider the experiment of tossing a coin twice:

$$\Omega = \{HH, HT, TH, TT\}, \qquad \mathcal{F} = \mathfrak{P}(\Omega) = \text{set of all subsets of } \Omega.$$

The event 'at most one tail is obtained' is described by $\{HH, HT, TH\}$.

Example 5

Consider the experiment of tossing a coin infinitely many times. Then $\Omega = \{H, T\}^{\mathbb{N}}$, so that $\omega \in \Omega$ is given by

$$\omega = (\omega_1, \omega_2, \ldots), \qquad \omega_n \in \{H, T\}.$$

Let

$$\mathcal{F} = \sigma(\{\omega \in \Omega : \omega_n = H\} : n \in \mathbb{N}\}).$$

Now $\mathcal{F} \neq \mathfrak{P}(\Omega)$: nevertheless $\mathcal{F}$ is big enough to contain the set

$$F = \left\{\omega : \frac{\sharp \text{ of heads in } n \text{ tosses}}{n} \to \frac{1}{2}\right\}.$$

(We shall show this in the next chapter: see Example 9).
This model can be used in Example 4 *via* the map $\omega \mapsto (\omega_1, \omega_2)$.

Example 6

Consider the experiment of choosing a point between 0 and 1 uniformly at random. Then we can take $\Omega = [0, 1]$, $\mathcal{F} = \mathfrak{B}([0, 1])$ and for the probability measure we can take the Lebesgue measure Leb.

Remarkably, it is possible to construct Lebesgue measure using the randomness in Example 5. A finite version of this is used by binary computers to simulate Example 6 (approximately).

Almost Surely

Definition: Secretary's Revenge:
Filing almost everything under "the".

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a statement S about outcomes of a certain experiment is is said to be true almost surely (a.s.) or with probability 1 if

$$\mathbb{P}\left[F\right] \quad = \quad 1\,,$$

where $F = \{\omega : S(\omega) \text{ is true}\}$ is also required to be in $\mathcal{F}$.

Exercise 2.20

Consider the experiment of Example 6. Show that almost surely the outcome is not a rational number.

Example 7

The set of algebraic numbers is the set of real roots of polynomials. Almost surely the outcome of Example 6 is transcendental (i.e., not algebraic).

3: Random Variables (survey of prerequisites)

The generation of random numbers is too important to be left to chance.

Random Variables

Table of contents of this section

3 Random Variables (prerequisite)
- Measurable functions
- Borel functions
- Random objects
- Laws and distribution functions
- Types of distribution functions
- Random vectors

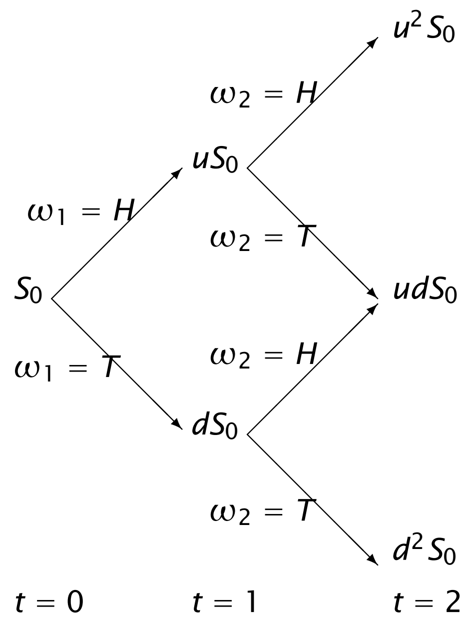
Random variables measure random events.
How does measure theory apply?

This is further important measure-theoretic fundamental material from ST213 (see also MA359). If the pace seems too fast **then revise the lecture notes for these modules** .

Example 8 (Binomial model for stock prices)

- The coin-tossing model of Example 4 can model stock prices. Let S_0 be stock price at time $t = 0$. Toss a coin twice. On H, multiply stock price by a factor of $u > 1$; on T, multiply stock price by a factor of $d < 1$. Let S_i, $i = 1, 2$, be stock prices at times $t = i$, $i = 1, 2$.
- $S_1 = uS_0$ or $S_1 = dS_0$, depending on ω_1. Similarly, S_2 depends on both ω_1 and ω_2. Consider the probability of the event $\{S_2 = u^2 S_0\}$. We have to define a σ-algebra such that the set $\{S_2 = u^2 S_0\}$ is measurable and then assign a probability measure to it.
- In this simple example $\{S_2 = u^2 S_0\}$ is equal to $\{\omega_1 = H, \omega_2 = H\}$, so we can calculate the probability of the event $\{S_2 = u^2 S_0\}$ by using the coin tossing model.

Stock price events



Measurable functions

Let S and T be sets and let h be a map $h : S \to T$. For $A \subseteq T$, define

$$h^{-1}(A) \quad = \quad \{s \in S : h(s) \in A\}.$$

Remark 3.1

In general $h(h^{-1}(A)) \neq A$ and $h^{-1}(h(B)) \neq B$ where $A \in T$ and $B \in S$.

Exercise 3.2

Find examples of these inequalities.

Clue: consider $S = \{0, 1\}$ and $T = \{0, 1\}$. One map will work for both inequalities, and there are only 4 different maps to consider!

Lemma 3.3

Let $h : S \to T$.

(a) *the map h^{-1} preserves all set operations:*

$$h^{-1}\Big(\bigcup_\alpha A_\alpha\Big) = \bigcup_\alpha h^{-1}(A_\alpha), \qquad h^{-1}(A^c) = (h^{-1}(A))^c.$$

(b) *if C is a collection of subsets of T, then*

$$h^{-1}(\sigma(C)) = \sigma(h^{-1}(C)).$$

Proof.

(a) prove as an exercise;

(b) $\Rightarrow$: RHS $\subseteq$ LHS: $\mathcal{C}$ is contained in $\sigma(\mathcal{C})$ which implies that $h^{-1}(\mathcal{C})$ is contained in $h^{-1}(\sigma(\mathcal{C}))$ which implies that $\sigma(h^{-1}(\mathcal{C}))$ is contained in $\sigma(h^{-1}(\sigma(\mathcal{C}))) = h^{-1}(\sigma(\mathcal{C}))$. Last equality holds since $h^{-1}(\sigma(\mathcal{C}))$ is a σ-algebra (check this point!).

$\Leftarrow$: LHS $\subseteq$ RHS: Check that $\mathcal{G} = \{B : h^{-1}(B) \in \sigma(h^{-1}(\mathcal{C}))\}$ is a σ-algebra. Since $\mathcal{C}$ is contained in $\mathcal{G}$ by definition, we have $\sigma(\mathcal{C}) \subseteq \{B : h^{-1}(B) \in \sigma(h^{-1}(\mathcal{C}))\}$ and therefore $h^{-1}(\sigma(\mathcal{C})) \subseteq \sigma(h^{-1}(\mathcal{C}))$ as required. □

Definition 3.4 (Measurable function)

If (S, Σ) and $(T, \mathcal{G})$ are measurable spaces and $h : S \to T$, then h is called $(\Sigma/\mathcal{G})$-measurable if $h^{-1}(A) \in \Sigma$ for all $A \in \mathcal{G}$.

The slash / in $(\Sigma/\mathcal{G})$ is just a separator: no mathematical meaning is intended.
We write $h^{-1} : \mathcal{G} \to \Sigma$.
If $T = \mathbb{R}$ and $\mathcal{G} = \mathfrak{B}$ (recall $\mathfrak{B} = \mathfrak{B}(\mathbb{R})$) then instead of saying that h is $(\Sigma/\mathfrak{B})$-measurable, we say that h is Σ-measurable.

Definition 3.5 (The class of Σ-measurable functions)

The class of Σ-measurable functions on S is denoted by $m\Sigma$.

If either S or T is a subset of $\mathbb{R}^n$, we will usually not make explicit reference to the σ-algebra; it will always be assumed to be the Borel σ-algebra.

Lemma 3.6

Let (S, Σ) and $(T, \mathcal{G})$ be measurable spaces and consider the function $h : S \to T$. If $\mathcal{C} \subseteq \mathcal{G}$ and $\sigma(\mathcal{C}) = \mathcal{G}$, then

$$h^{-1}(\mathcal{C}) \subseteq \Sigma \quad \text{implies} \quad h \text{ is } (\Sigma/\mathcal{G})\text{-measurable}.$$

Proof.

Let $\mathcal{G}'$ be the class of elements $G \in \mathcal{G}$ such that $h^{-1}(G) \in \Sigma$.
By Lemma 3.3 (a), $\mathcal{G}'$ is a σ-algebra.
By one of the hypotheses of the Lemma, $\mathcal{C} \subseteq \mathcal{G}'$, which implies $\sigma(\mathcal{C}) \subseteq \mathcal{G}'$.
Therefore, $\mathcal{G} = \sigma(\mathcal{C}) \subseteq \mathcal{G}' \subseteq \mathcal{G}$, and so $\mathcal{G}' = \mathcal{G}$. □

Lemma 3.7

For any measurable space (S, Σ) a function $h : S \to \mathbb{R}$ is Σ-measurable if

$$\{h \leq a\} \quad = \quad \{s \in S : h(s) \leq a\} = h^{-1}((-\infty, a]) \in \Sigma, \quad a \in \mathbb{R}.$$

Proof.

Let $\mathcal{C}$ be a family of sets given by $\mathcal{C} = \{(-\infty, a] : a \in \mathbb{R}\}$. Then $\sigma(\mathcal{C}) = \mathfrak{B}$ and the result follows by Lemma 3.6. □

Remark 3.8

Or you can check whether $h^{-1}((a, b)) \in \Sigma$ for each real $a < b$,
or $h^{-1}((-\infty, a)) \in \Sigma$ for all real a,
or $h^{-1}([a, \infty)) \in \Sigma$ for all real a.

Lemma 3.9

$m\Sigma$ is an algebra over $\mathbb{R}$, which is to say that it is closed under addition, multiplication and constant multiplication: if $\lambda \in \mathbb{R}$ and $h, h_1, h_2, \in m\Sigma$, then $h_1 + h_2, h_1 h_2, \lambda h \in m\Sigma$.

Proof.

We will prove that the sum of measurable functions is a measurable function. Let $a \in \mathbb{R}$. Then
$\{s : h_1(s) + h_2(s) > a\} = \bigcup_{q\in\mathbb{Q}} \left(\{s : h_1(s) > q\} \cap \{s : h_2(s) > a - q\}\right)$.
(RHS⊆LHS is immediate. If $h_1(s) + h_2(s) > a$ then $h_1(s) > a - h_2(s)$. Pick rational q with $h_1(s) > q > a - h_2(s)$. Then $h_1(s) > q$ and also $h_2(s) > a - q$.)
Since $\{h_1 > q\} \in \Sigma$ and $\{h_2 > a - q\} \in \Sigma$, it follows that $\{h_1 + h_2 > a\} \in \Sigma$, which by Lemma 3.7 implies that $h_1 + h_2$ is in $m\Sigma$. □

Lemma 3.10 (Composition Lemma)

If (S_1, Σ_1), (S_2, Σ_2), (S_3, Σ_3) are measurable spaces and $h : S_1 \to S_2$ is (Σ_1/Σ_2)-measurable, and $g : S_2 \to S_3$ is (Σ_2/Σ_3)-measurable, then $g \circ h : S_1 \to S_3$ is (Σ_1/Σ_3)-measurable.

Proof.

Consider the compositions:

$$S_1 \xrightarrow{h} S_2 \xrightarrow{g} S_3$$

$$S_1 \xleftarrow{h^{-1}} S_2 \xleftarrow{g^{-1}} S_3$$

□

Lemma 3.11 (Measurability of infs, lim infs of functions)

Let (S, Σ) be a measurable space and let $(h_n, n \in \mathbb{N})$ be a sequence of elements of $m\Sigma$. Then

(i) $\inf h_n$

(ii) $\liminf h_n$

(iii) $\limsup h_n$

are all Σ-measurable (into $([-\infty, \infty], \mathfrak{B}[-\infty - \infty])$), but we shall still write $\inf h_n \in m\Sigma$ (for example). Furthermore,

(iv) *$\{s : \lim h_n(s)$ exists in $\mathbb{R}\} \in \Sigma$.*

Proof.

(i) $\{s : \inf h_n(s) \geq a\} = \bigcap_n \{s : h_n(s) \geq a\} \in \Sigma$ if $a \in \mathbb{R}$;

(ii) Define $L_n(s) = \inf_{r\geq n} h_r(s)$. Since $L_n(s)$ is an increasing sequence, define $L(s) = \liminf h_n(s) = \lim L_n(s) = \sup L_n(s)$. By (i) each L_n is Σ-measurable, and so $\{L \leq a\} = \bigcap_n \{L_n \leq a\} \in \Sigma$;

(iii) This is treated in the same way as (ii);

(iv) The set $\{s : \lim h_n(s)$ exists in $\mathbb{R}\}$ is an intersection of Σ-measurable sets:

$$\{\limsup h_n < \infty\} \cap \{\liminf h_n > -\infty\} \cap \\ \cap \{\limsup h_n - \liminf h_n \in \{0\}\},$$

thus $\{s : \lim h_n(s)$ exists in $\mathbb{R}\} \in \Sigma$.

□

Borel functions

- Measures and σ-algebras on their own do not capture all the important notions of probability theory. We need a notion of closeness: the most immediately useful kind is provided by topology.
- A topology on Ω is a collection of subsets of Ω containing Ω and the empty set, which is closed under finite intersection and arbitrary union. A topological space is a pair $(\Omega, \mathcal{T})$, where Ω is a set and $\mathcal{T}$ is a topology on Ω. A set $A \subseteq \Omega$ is an open set in Ω if and only if it is an element of the topology $\mathcal{T}$. If S and T are topological spaces and if h is a map $h : S \to T$ then h is said to be continuous provided $h^{-1}(A)$ is an open set in S for every open set A in T.
- Thus there is an analogy between a measurable space, a measurable set and a measurable function, and a topological space, an open set and a continuous function, respectively.

Definition 3.12 (Borel function)

Let S be a topological space and let $\mathfrak{B}(S)$ be the Borel *sigma*-algebra on S. A function $h : S \to \mathbb{R}$ is called Borel if h is $\mathfrak{B}(S)$-measurable.

Lemma 3.13 (continuous functions are Borel)

If S is a topological space and $h : S \to \mathbb{R}$ is continuous, then h is Borel.

Proof.

Let $\mathcal{C} = \{\textit{open subsets of } \mathbb{R}\}$, and use Lemma 3.6. □

Exercise 3.14

Any continuous function from $\mathbb{R}$ to $\mathbb{R}$ is Borel. Show the converse is false.

Random objects

We now consider random variables and more general "random objects". We need a general definition because it is frequently convenient to bundle together a collection of random variables to be treated as a single random object (eg. coordinates of a random point).

Definition 3.15 (random object)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A random object X in a measurable space (S, Σ) is a $(\mathcal{F}/\Sigma)$-measurable function. Thus,

$$\begin{aligned} X : \Omega &\to S \\ X^{-1} : \Sigma &\to \mathcal{F}. \end{aligned}$$

If the measurable space (S, Σ) above is $(\mathbb{R}, \mathfrak{B})$, then the random object X is called a random variable. Thus, if X is a random variable,

$$X : \Omega \to \mathbb{R}, \qquad X^{-1} : \mathfrak{B} \to \mathcal{F}.$$

Remark 3.16

Since $\{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{F}$ for all $a \in \mathbb{R}$, we can talk about $\mathbb{P}[X \leq a]$ where $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$.

Definition 3.17 (Bernoulli random variable)

Let X be a random variable which takes two values, 0 and 1, such that $\mathbb{P}[X = 1] = p$ and $\mathbb{P}[X = 0] = 1 - p$, $0 < p < 1$. Then X is said to be a Bernoulli random variable.

Definition 3.18 (Indicator random variable)

Let $A \in \mathcal{F}$ be an event and let $I_A : \Omega \to \mathbb{R}$ be the indicator function

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \in A^c. \end{cases}$$

Then I_A is a random variable taking values 1, 0 with probabilities $\mathbb{P}[A]$, $\mathbb{P}[A^c]$. It is the indicator random variable for set A.

Example 9 (Coin tossing: continuation of Example 5)

Let $\Omega = \{H, T\}^{\mathbb{N}}$, $\omega = (\omega_1, \omega_2, \dots)$, $\omega_n \in \{H, T\}$, $n \in \mathbb{N}$, and take

$$\mathcal{F} = \sigma(\{\omega \in \Omega : \omega_n = \widetilde{\omega}\} : n \in \mathbb{N}, \widetilde{\omega} \in \{H, T\}) .$$

Let X_n be the indicator random variable for the event $[\omega_n = H]$, and set $S_n = X_1 + X_2 + \cdots + X_n$. (Note that S_n is a number of heads in first n tosses.) By Lemma 3.9 the sum of random variables S_n is itself a random variable. For $p \in [0, 1]$ consider

$$\Lambda \quad = \quad \{\omega : \frac{\text{number of heads}}{\text{number of tosses}} \to p\} = \{\omega : \frac{S_n}{n} \to p\} .$$

The set Λ can be written as

$$\lambda = \{\omega : \limsup \frac{S_n}{n} = p\} \cap \{\omega : \liminf \frac{S_n}{n} = p\} .$$

Thus $\Lambda \in \mathcal{F}$, by Lemma 3.11.

Remark 3.19

Since $\Lambda \in \mathcal{F}$, the next step would be to find $\mathbb{P}[\Lambda]$. This would lead us to the Strong Law of Large Numbers: in fact $\mathbb{P}[\Lambda] = 1$, and this generalizes to a vast extent.

In the previous example we have a set Ω and a family of maps $X_n : \Omega \to \mathbb{R}$. Thus X_n is $\mathcal{F}$-measurable for all n. One way to think about $\mathcal{F}$ is that it is the smallest σ-algebra all maps X_n, $n \in \mathbb{N}$, are measurable with respect to it. We write

$$\mathcal{F} \quad = \quad \sigma(X_n : n \in \mathbb{N}) .$$

More precisely,

Definition 3.20 (σ-algebra generated by a collection of functions)

If $(Y_\gamma, \gamma \in C)$ is a collection of maps $Y_\gamma : \Omega \to \mathbb{R}$, then

$$\mathfrak{Y} \quad = \quad \sigma(Y_\gamma, \gamma \in C)$$

is defined to be the smallest σ-algebra $\mathfrak{Y}$ on Ω which makes the maps Y_γ measurable (for $\gamma \in C$).

Therefore,

$$\sigma(Y_\gamma, \gamma \in C) = \sigma(\{Y_\gamma \in B\} : \gamma \in C, B \in \mathfrak{B}) .$$

If X is a random variable on $(\Omega, \mathcal{F})$, then $\sigma(X) \subseteq \mathcal{F}$.

Example 10

Let X be a random variable. Then:

(i) $\sigma(X) = X^{-1}(\mathcal{B}) = (\{\omega : X(\omega) \in B\} : B \in \mathcal{B})$;

(ii) $\sigma(X)$ is generated by the Π-system

$$\Pi(X) \quad = \quad (\{\omega : X(\omega) \leq a\}, a \in \mathbb{R}) .$$

Think about this example; you should be able to recognize how it relates to the notion of a distribution function.

Proof.

(i) $X : \Omega \to \mathbb{R}$, $X^{-1} : \mathcal{B} \to X^{-1}\mathcal{B}$. X is measurable relative to $X^{-1}\mathcal{B}$ which is a σ-algebra (by Lemma 3.3 (i)). Thus, $\sigma(X) \subseteq X^{-1}\mathcal{B}$. On the other hand, since X is $\sigma(X)$-measurable then for, $B \in \mathcal{B}$, $X^{-1}(B) \in \sigma(X)$. Thus, $X^{-1}\mathcal{B} \subseteq \sigma(X)$. It follows that $\sigma(X) = X^{-1}(\mathcal{B})$.

(ii) Note that $\Pi(X) = X^{-1}(\Pi(\mathbb{R}))$, where $\Pi(\mathbb{R}) = \{(-\infty, x] : x \in \mathbb{R}\}$ and recall that $\mathcal{B} = \sigma(\Pi(\mathbb{R}))$. By (i) and by Lemma 3.3 (ii)

$$\sigma(X) = X^{-1}(\mathcal{B}) = X^{-1}(\sigma(\Pi(\mathbb{R})) = \sigma(X^{-1}(\Pi(\mathbb{R})) = \sigma(\Pi(X)).$$

□

Example 11

If $(X_n, n \in \mathbb{N})$ is a collection of functions on Ω, and $\mathcal{H}_n = \sigma(X_k : k \leq n)$, then $\bigcup \mathcal{H}_n$ is a Π-system (actually, an algebra) and it generates $\sigma(X_n, n \in \mathbb{N})$.

Example 12 (Toss a coin three times with results $\omega_1, \omega_2, \omega_3$)

Let S_k be random variables such that S_0 is non-random and

$$S_k = \begin{cases} uS_{k-1} & \text{if } \omega_k = H \\ dS_{k-1} & \text{if } \omega_k = T, \end{cases} \qquad k = 1, 2, 3.$$

For instance, S_2 takes values $u^2 S_0$, $udS_0, d^2 S_0$. The σ-algebra generated by S_2 consists of:

(1) $A_{HH} = \{HHT, HHH\} = \{S_2 = u^2 S_0\} = \{w_1 = H, w_2 = H\}$;

(2) $A_{TT} = \{TTT, TTH\} = \{S_2 = d^2 S_0\} = \{w_1 = T, w_2 = T\}$;

(3) $A_{TH} \cup A_{TH} = \{THT, THH, HTH, HTT\} = \{S_2 = udS_0\} = \{w_1 = T, w_2 = H\} \cup \{w_1 = H, w_2 = T\}$;

(4) complements of above sets;

(5) union of the above sets (including (4));

(6) $\emptyset$ and Ω.

Let $(Z_\delta, \delta \in D)$ be a family of random variables for some $(\Omega, \mathcal{F}, \mathbb{P})$ so that each Z_δ is $\mathcal{F}$-measurable. Suppose that someone gives you the observed values of $(Z_\delta, \delta \in D)$. Then you would be able to tell whether an event $F \in \sigma\{Z_\delta, \delta \in D\}$ has occurred or not. In other words, the σ-algebra $\sigma\{Z_\delta, \delta \in D\}$ consists precisely of those events F for which you can decide whether or not F has occurred by knowing the values of $(Z_\delta, \delta \in D)$. For instance, if in Example 12 as an outcome of the experiment you would get $S_1 = uS_0$, $S_2 = u^2 S_0$ and $S_3 = u^2 dS_0$, then you would know that only the event $\{HHT\}$ could have occurred.

Laws and distribution functions

Definition 3.21 (distribution of a random object)

If X is a random object defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in (S, Σ), then the distribution or law of X is the probability measure $\mathcal{L}_X$ on (S, Σ) given by

$$\mathcal{L}_X \quad = \quad \mathbb{P} \circ X^{-1} .$$

Exercise 3.22

Show that $\mathcal{L}_X$ is a probability measure on (S, Σ).

We work mainly with random variables $X : \Omega \to \mathbb{R}$. Then $\mathcal{L}_X$ is a probability measure on $(\mathbb{R}, \mathcal{B})$. By the Uniqueness Lemma 2.13, $\mathcal{L}_X$ is determined by its values on the Π-system of sets $\{(-\infty, x]; x \in \mathbb{R}\} = \Pi(\mathbb{R})$. In other words, the measure $\mathcal{L}_X$ is determined by the distribution function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) \quad = \quad \mathcal{L}_X((-\infty, x]) = \mathbb{P}[X \leq x] .$$

Lemma 3.23 (Properties of the distribution function)

Suppose that F is the distribution function $F = F_X$ of some random variable X. Then

(i) *$F : \mathbb{R} \to [0, 1]$, F is non-decreasing;*
(ii) *F is right-continuous;*
(iii) $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$.

Proof.

We focus on (ii): by Lemma 2.18,

$$\mathbb{P}\left[X \leq x + \frac{1}{n}\right] \quad \downarrow \quad \mathbb{P}[X \leq x] ,$$

which together with the monotonicity of F shows that F is right-continuous. □

Lemma 3.24

Existence of a random variable with given distribution function
Suppose that F has properties $(i) - (iii)$ of Lemma 3.23. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $F_X = F$.

Proof (Sketch).

Example 2 shows how to construct a probability space $([0, 1], \mathfrak{B}, \mathrm{Leb})$ such that $Y(x) = x$ defines a random variable Y for which $\mathbb{P}[Y \leq x] = x$. Given a distribution function F, define

$$F^{-}(y) \quad = \quad \inf\{x : F(x) \geq y\}$$

and show[1] that $X = F^{-}(Y)$ has distribution function F. □

[1] You can find a short proof in Ripley (1987, Theorem 3.1).

Lemma 3.25

There is a "one-to-one" correspondence between probability measures on $(\mathbb{R},\mathfrak{B})$ and distribution functions of random variables.

Proof.

The statement follows directly from the definition of the distribution function and from Lemma 3.24. $\square$

Types of distribution functions

Let $\widetilde{\mathbb{P}}$ be a probability measure on $(\mathbb{R},\mathcal{B})$, let F denote the corresponding distribution function and let X be a random variable with law $\widetilde{\mathbb{P}}$.

Definition 3.26 (discrete random variable)

A random variable X and its corresponding probability measure $\widetilde{\mathbb{P}}$ on $(\mathbb{R},\mathcal{B})$ are said to be discrete if there is a countable subset C of $\mathbb{R}$ with $\mathbb{P}\left[X \in C\right] = 1$.

Therefore, if X is a discrete random variable and takes values in a countable set C, then $\widetilde{\mathbb{P}}[C] = \mathbb{P}\left[X \in C\right] = 1$ and

$$F(x) \quad = \quad \widetilde{\mathbb{P}}\left[(-\infty, x]\right] = \sum_{x_i \le x,\ x_i \in C} \mathbb{P}\left[X = x_i\right], \qquad x \in \mathbb{R}\,.$$

It follows that the distribution function of a discrete random variable is a step function. Its jumps are given by the mass function.

Definition 3.27 (probability mass function of a discrete random variable)

The probability mass function of a discrete random variable X which takes values in a countable set C is the function $f : \mathbb{R} \to [0,1]$ given by

$$f(x) \quad = \quad \begin{cases} \widetilde{\mathbb{P}}[\{x\}] = \mathbb{P}\left[X = x\right] & \text{for } x \in C\,, \\ 0 & \text{for } x \in \mathbb{R} \setminus C\,. \end{cases}$$

Conversely, let F be a right-continuous non-decreasing step function such that $F(-\infty) = 0$, and $F(\infty) = 1$, and that F has jumps at $C = \{x_1, x_2, \ldots\}$ of size $f(x_i)$, $i = 1, 2, \ldots$ (C is at most countable). Then, there exists a random variable X such that $\mathbb{P}\left[X \in C\right] = 1$ and that $\mathbb{P}\left[X = x_i\right] = f(x_i)$, $i = 1, 2, \ldots$.

Definition 3.28 (probability density function of an (absolutely) continuous random variable)

A random variable X and its corresponding probability measure $\widetilde{\mathbb{P}}$ on $(\mathbb{R},\mathcal{B})$ are said to be (absolutely) continuous if there is an integrable function $f(x) : \mathbb{R} \to [0,\infty)$ such that the distribution function F of X can be written as

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

The function f is called a probability density function of the absolutely continuous random variable X.

The distribution function F of an absolutely continuous random variable is, by the definition, continuous, and $F'(x) = f(x)$ except on a set of Lebesgue measure zero. In addition, $\int_{-\infty}^{\infty} f(t)\, dt = 1$. If f is a function satisfying $f \geq 0$ and $\int_{-\infty}^{\infty} f(t)\, dt = 1$, then there exists a random variable X with the law given by

$$\mathbb{P}[x \leq x] = \int_{\infty}^{x} f(t)\, dt, \quad x \in \mathbb{R}.$$

However, it can happen that a continuous distribution function is not absolutely continuous.

Definition 3.29 (singular random variable)

A random variable X and its corresponding probability measure $\widetilde{\mathbb{P}}$ on $(\mathbb{R}, \mathcal{B})$ are said to be singular (with respect to Lebesgue measure) if the distribution function F of X is continuous and there is a measurable set A for which $\widetilde{\mathbb{P}}(A) = 1$ but $\text{Leb}(A) = 0$.

Note that a continuous distribution function can be either singular or absolutely continuous.
The Cantor “middle-thirds” construction can be adapted to produce a singular distribution function.
More generally, it can be shown that any distribution function can be written as a convex combination

$$F = \alpha F_d + \beta F_s + \gamma F_{ac}, \quad \alpha, \beta, \gamma \geq 0,\ \alpha + \beta + \gamma = 1\,,$$

where F_d is a discrete, F_s is a singular and F_{ac} is an absolutely continuous distribution function.

Example 13

Let $0 < a < b$ and $0 < q < r < 1$ and let $F(x)$ be a distribution function given by

$$F(x) \quad = \quad \begin{cases} 0 & \text{if} \quad x < 0 \\ q & \text{if} \quad 0 \leq x < a \\ \frac{r-q}{b-a}x + \frac{qb-ar}{b-a} & \text{if} \quad a \leq x < b \\ 1 & \text{if} \quad b \leq x. \end{cases}$$

Then $F(x) = \alpha F_d(x) + (1-\alpha) F_{ac}(x)$ where

$$F_d(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ \frac{q}{1-(r-q)} & \text{if} \quad 0 \leq x < b \\ 1 & \text{if} \quad b \leq x, \end{cases}, \quad F_{ac}(x) = \begin{cases} 0 & \text{if} \quad x < a \\ \frac{x-a}{b-a} & \text{if} \quad a \leq x < b \\ 1 & \text{if} \quad b \leq x. \end{cases}$$

and $\alpha = 1 - r + q$.

Random vectors

Definition 3.30

Let (S_k, Σ_k), $k = 1, 2, \ldots, n$, be measurable spaces. Define the product σ-algebra

$$\Sigma_1 \star \Sigma_2 \star \cdots \star \Sigma_n \quad = \quad \sigma(\{A_1 \times A_2 \times \cdots \times A_n : A_1 \in \Sigma_1, A_2 \in \Sigma_2, \ldots, A_n \in \Sigma_n\})\,.$$

Several random objects can be defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We have

Lemma 3.31

Let $(\Omega, \mathcal{F})$ and (S_i, Σ_i), $i = 1, 2, \ldots, n$, be measurable spaces. The maps $\{X_i : \Omega \to S_i : i = 1, \ldots, n\}$ are all random objects if and only if the vector

$$Z = (X_1, X_2, \ldots, X_n) : \Omega \to S_1 \times S_2 \times \cdots \times S_n$$

is a random object in $(S_1 \times S_2 \times \ldots, \times S_n, \Sigma_1 \star \Sigma_2 \star \cdots \star \Sigma_n)$.

Proof.

Suppose that Z is a random object. For a fixed i, define a projection map Π_i by $\Pi_i : (S_1 \times S_2 \times \ldots, \times S_n, \Sigma_1 \star \Sigma_2 \star \cdots \star \Sigma_n) \to (S_i, \Sigma_i)$. Then $X_i = \Pi_i \circ Z$. Show that Π_i is a random object. Then, by Lemma 3.10 it follows that X_i is a random object.

Suppose now that $A = A_1 \times A_2 \times \cdots \times A_n \in \Sigma_1 \star \Sigma_2 \star \cdots \star \Sigma_n$. Then

$$\begin{aligned} Z^{-1}(A) &= \{\omega : Z(\omega) \in A\} \\ &= \{\omega : X_1(\omega) \in A_1, X_2(\omega) \in A_2, \ldots, X_n(\omega) \in A_n\} \\ &= \{\omega : \omega \in X_1^{-1}(A_1), \omega \in X_2^{-1}(A_2), \ldots, \omega \in X_n^{-1}(A_n)\} \\ &= X_1^{-1}(A_1) \cap X_2^{-1}(A_2) \cap \cdots \cap X_n^{-1}(A_n) \end{aligned}$$

Since X_i is $(\mathcal{F}/\Sigma_i)$-measurable for every $i = 1, 2 \ldots, n$, it follows that $X_i^{-1}(A_i) \in \mathcal{F}$, $i = 1, 2, \ldots, n$, and finally that $Z^{-1}(A) \in \mathcal{F}$. Therefore, Z is $(\mathcal{F}/(\Sigma_1 \star \Sigma_2 \star \cdots \star \Sigma_n))$-measurable and thus a random object. □

Definition 3.32

Suppose $X_i : (\Omega, \mathcal{F}) \to (S_i, \Sigma_i)$, $i = 1, 2, \ldots, n$, are random objects. Then the distribution of the random object $Z = (X_1, X_2, \ldots, X_n) : \Omega \to S_1 \times S_2 \cdots \times S_n$ is called the joint distribution of $X_1, X_2, \ldots, X_n$.

Now we are ready to proceed to the body of knowledge to be discussed in this module.

4: Independence

Mad, adj.:

Affected with a high degree of intellectual independence

...

Ambrose Bierce, "The Devil's Dictionary"

Independence

Table of contents of this section

4 Independence
- Definitions of independence
- Pi-systems
- Product measures and independence
- Tail sigma-algebra and Kolmogorov's zero-one law
- Borel-Cantelli lemmas

Independence is a characteristic concept of probability theory.

List the notions of independence already encountered . . . .

Rule of thumb: under independence, $\mathbb{P}[\bigcap \cdots] = \prod \mathbb{P}[\cdots]$ (products and intersections over countable sequences only!).

Definitions of independence

The rule-of-thumb works best at highest level: σ-algebras!

Definition 4.1 (independent σ-algebras)

Sub-σ-algebras $\mathcal{G}_1, \mathcal{G}_2, \ldots$ of $\mathcal{G}$ are called independent if, whenever $G_i \in \mathcal{G}_i$, $i \in \mathbb{N}$, and $i_1, i_2, \ldots, i_n$ are distinct, then

$$\mathbb{P}[G_{i_1} \cap G_{i_2} \cap \cdots \cap G_{i_n}] = \prod_{k=1}^{n} \mathbb{P}[G_{i_k}] .$$

Definition 4.2 (independent random variables)

Random variables $X_1, X_2, \ldots$ are called independent if the σ-algebras $\sigma(X_1), \sigma(X_2), \ldots$ are independent.

Definition 4.3 (independent events)

Events $E_1, E_2, \ldots$ are independent if the σ-algebras $\mathcal{E}_1, \mathcal{E}_2, \ldots$ are independent, where

$$\mathcal{E}_n \text{ is the } \sigma\text{-algebra } \{\emptyset,\ E_n,\ \Omega \setminus E_n,\ \Omega\}.$$

Since $\mathcal{E}_n = \sigma(I_{E_n})$, it follows that events $E_1, E_2, \ldots$ are independent if and only if the random variables $I_{E_1}, I_{E_1}, \ldots$ are independent.

The more familiar definitions of independence follow easily: for example events $E_1, E_2, \ldots$ are independent if whenever $n \in \mathbb{N}$ and $i_1, i_2, \ldots, i_n$ are distinct, then

$$\mathbb{P}[E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_n}] \quad = \quad \prod_{k=1}^{n} \mathbb{P}[E_{i_k}] .$$

(This follows directly from Definition 4.3.)

Π-systems

Recall the definition of Π-systems and the basic Lemma 2.13. The use of the lemma permits study of independence via (simple) Π-systems rather than (complicated) σ-algebras.

Lemma 4.4 (The Π-system lemma)

Suppose that $\mathcal{G}$ and $\mathcal{H}$ are sub-σ-algebras of $\mathcal{F}$, and that $\mathcal{I}$ and $\mathcal{J}$ are Π-systems such that $\sigma(\mathcal{I}) = \mathcal{G}$ and $\sigma(\mathcal{J}) = \mathcal{H}$. Then $\mathcal{G}$ and $\mathcal{H}$ are independent if and only if $\mathcal{I}$ and $\mathcal{J}$ are independent in the special sense that

$$\mathbb{P}[I \cap J] \quad = \quad \mathbb{P}[I]\,\mathbb{P}[J], \qquad \text{for } I \in \mathcal{I},\ J \in \mathcal{J} .$$

Proof.

If $\mathcal{G}$ and $\mathcal{H}$ are independent, then so are $\mathcal{I}$ and $\mathcal{J}$ (use Definition 4.1). Suppose $\mathcal{I}$ and $\mathcal{J}$ are independent. For fixed $I \in \mathcal{I}$, the measures (check that they are measures!)

$$H \mapsto \mathbb{P}[I \cap H] \quad \text{and} \quad H \mapsto \mathbb{P}[I]\,\mathbb{P}[H]$$

on $(\Omega, \mathcal{H})$ have the same total mass $\mathbb{P}[I]$ and agree on $\mathcal{J}$. By Lemma 2.13 they agree on $\sigma(\mathcal{J}) = \mathcal{H}$. Thus $\mathbb{P}[I \cap H] = \mathbb{P}[I]\,\mathbb{P}[H]$ for $I \in \mathcal{I},\ H \in \mathcal{H}$. Furthermore, for fixed $H \in \mathcal{H}$, the measures

$$G \mapsto \mathbb{P}[G \cap H] \quad \text{and} \quad G \mapsto \mathbb{P}[G]\,\mathbb{P}[H]$$

on $(\Omega, \mathcal{G})$ have the same total mass $\mathbb{P}[H]$ and agree on $\mathcal{I}$. Thus they agree on $\sigma(\mathcal{I}) = \mathcal{G}$ and therefore $\mathbb{P}[G \cap H] = \mathbb{P}[G]\,\mathbb{P}[H]$ for $G \in \mathcal{G},\ H \in \mathcal{H}$, which by Definition 4.1 means that σ-algebras $\mathcal{G}$ and $\mathcal{H}$ are independent. □

Let X and Y be two random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that whenever $x, y \in \mathbb{R}$,

$$\mathbb{P}[X \leq x, Y \leq y] \quad = \quad \mathbb{P}[X \leq x]\,\mathbb{P}[Y \leq y]\,.$$

It follows that the Π-systems $\Pi(X)$ and $\Pi(Y)$ (see Example 10(ii)) are independent. Hence Lemma 4.4 shows that $\sigma(X)$ and $\sigma(Y)$ are independent; thus X and Y are independent (using Definition 4.2).

Product measures and independence

Let (S_1, Σ_1, μ_1) and (S_2, Σ_2, μ_2) be σ-finite measure spaces and put

$$(S, \Sigma) = (S_1 \times S_2, \Sigma_1 \star \Sigma_2).$$

There exists a unique σ-finite product measure $\mu = \mu_1 \otimes \mu_2$ on (S, Σ) for which $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ for $A_1 \in \Sigma_1, A_2 \in \Sigma_2$.
The concept of independence can be formulated by using a product measure.
Let X and Y be random variables on some $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{L}_{X,Y}$ denote the joint law of X and Y (on $(\mathbb{R}^2, \mathcal{B} \star \mathcal{B})$). Then X and Y are independent if and only if

$$\mathcal{L}_{X,Y} = \mathcal{L}_X \otimes \mathcal{L}_Y$$

where $\mathcal{L}_X$ (respectively $\mathcal{L}_Y$) denotes the law of X (respectively Y) on $(\mathbb{R}, \mathcal{B})$.

The product space construction builds independent random variables with given marginal laws, ensuring physically independent variables are genuinely stochastically independent.

Example 14

Let $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$ (respectively $(\Omega_Y, \mathcal{F}_Y, \mathbb{P}_Y)$) be a probability triple and let X (respectively Y) be a random variable defined on $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$ (respectively $(\Omega_Y, \mathcal{F}_Y, \mathbb{P}_Y)$) with distribution function F_X (respectively F_Y). Take

$$(\Omega, \mathcal{F}) \quad = \quad (\Omega_X \times \Omega_Y, \mathcal{F}_X \star \mathcal{F}_Y)$$

and define for $\omega = (\omega_X, \omega_Y) \in \Omega$

$$\tilde{X}(\omega) = X(\omega_X), \qquad \tilde{Y}(\omega) = Y(\omega_Y).$$

Then, $\tilde{X}$ and $\tilde{Y}$ are independent on $(\Omega, \mathcal{F}, \mathbb{P}_X \otimes \mathbb{P}_Y)$ and have the same law as X and Y respectively.

Tail σ-algebra and Kolmogorov's zero-one law

The next definition captures the idea of the "indefinite future".

Definition 4.5 (tail σ-algebra)

For a sequence $\{\mathcal{F}_n : n \geq 1\}$ of σ-algebras, the tail σ-algebra is

$$\mathcal{T} \quad = \quad \bigcap_{n\geq 1} \sigma(\bigcup_{k\geq n} \mathcal{F}_k) \quad = \quad \bigcap_{n\geq 1} \sigma(\mathcal{F}_n, \mathcal{F}_{n+1}, \ldots) .$$

Events in a tail σ-algebra are called tail events.

Remark 4.6

Think of $\mathcal{F}_n$ as the collection of events determined by the n^{th} of a series of experiments. Then $\mathcal{T}$ is the ensemble of events determined by the 'tail' run of experiments, however far in the future the tail might be. Note that the notion depends on the choice of the sequence $\{\mathcal{F}_n : n \geq 1\}$ of σ-algebras.

Let $X_1, X_2, \ldots$ be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. The tail σ-algebra for this sequence is obtained by putting $\mathcal{F}_n = \sigma(X_n)$, $n = 1, 2, \ldots$ in the above. Writing

$$\mathcal{T}_n \quad = \quad \sigma\Big(\bigcup_{k\geq n} \sigma(X_k)\Big) \quad = \quad \sigma(X_n, X_{n+1}, \ldots) ,$$

we have

$$\mathcal{T} \quad = \quad \bigcap_n \mathcal{T}_n .$$

Note that $\mathcal{T}$ contains many important events, for example:

1. $F_1 = [\lim_{k\to\infty} X_k \text{ exists}] = \{\omega : \lim_k X_k(\omega) \text{ exists}\}$
2. $F_2 = [\sum_k X_k \text{ converges}]$
3. $F_3 = \left[\lim \frac{X_1+X_2+\cdots+X_k}{k} \text{ exists}\right]$

Warwick Statistics

1. Why F_1? because whether $\lim_{k\to\infty} X_k$ exists is determined entirely by X_n, X_{n+1}, ..., hence is in $\mathcal{T}_n = \sigma\{X_n, X_{n+1}, \ldots\}$ for all n (NB: $\sigma(X_n, X_{n+1}, \ldots) = \sigma\{\sigma(X_n), \sigma(X_{n+1}), \ldots\}$, etc.)
2. Why F_2? because whether $\sum_k X_k$ exists is determined entirely by X_n, X_{n+1}, ..., hence is in $\mathcal{T}_n = \sigma(X_n, X_{n+1}, \ldots)$ for all n;
3. Why F_3? We *could* argue similarly, at the price of more detail. Alternatively, notice the following characterization of F_3 then use the following example 15:

$$F_3 \quad = \quad \left[\limsup_k \frac{X_1 + X_2 + \cdots + X_k}{k} = \liminf_k \frac{X_1 + X_2 + \cdots + X_k}{k}\right]$$

Example 15

An important example of a $\mathcal{T}$-measurable random variable is

$$\xi \quad = \quad \limsup_k \frac{X_1 + X_2 + \cdots + X_k}{k} .$$

Proof of Example 15.

We know that ξ as given above is $\mathcal{F}$-measurable (sums of random variables are measurable, lim sup of random variables is measurable). To prove this we only need that each X_n, $n \in \mathbb{N}$, is $\mathcal{F}$-measurable.

In fact ξ will be $\mathcal{G}$-measurable for any sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$ such that X_n is $\mathcal{G}$-measurable for each $n \in \mathbb{N}$,. The intersection of all sub-σ-algebras $\mathcal{G}$ of $\mathcal{F}$ such that each X_n is $\mathcal{G}$-measurable is $\sigma(X_1, X_2, \ldots)$, so this is actually the smallest σ-algebra $\mathcal{G}$ such that all X_n are $\mathcal{G}$-measurable.

Fix $k \in \mathbb{N}$. By properties of lim sup, we can write $\xi = \limsup \frac{\tilde{S}_n}{n}$ where $\tilde{S}_n = S_{k+n} - S_k = X_{k+1} + X_{k+2} + \cdots + X_{k+n}$. Thus $\tilde{S}_n$ is obtained from $(X_{k+1}, X_{k+2}, \ldots)$ in the same way as S_n is obtained from $(X_1, X_2, \ldots)$. Therefore, in the same way as we concluded above that ξ is $\mathcal{F}$-measurable, we conclude now that ξ is $\mathcal{T}_k$-measurable where $\mathcal{T}_k = \sigma(X_{k+1}, X_{k+1}, \ldots)$. Hence, ξ is $\mathcal{T}$-measurable where $\mathcal{T} = \bigcap_k \mathcal{T}_k$, the tail σ-algebra of $(X_n : n \in \mathbb{N})$. □

Given a sequence of random variables, does the corresponding tail σ-algebra contain *all* events depending on the whole sequence?

No.

Consider the event

$$[\sum_n X_n \text{ exists and is equal to } 0].$$

In general (for example, if the X_n are independent and if X_1 is genuinely random, not constant) this event depends inescapably on X_1. Therefore it will not belong to the tail σ-algebra.

The following result of Kolmogorov shows that sometimes probability questions can have radically simply answers!

Theorem 4.7 (Kolmogorov's zero-one law)

Suppose that $\{\mathcal{F}_n : n \geq 1\}$ is an independent sequence of σ-algebras based on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then the tail σ-algebra $\mathcal{T}$ of this sequence is $\mathbb{P}$-trivial, that is
(i) $F \in \mathcal{T}$ implies $\mathbb{P}[F] = 0$ or $\mathbb{P}[F] = 1$.
(ii) if ξ is a $\mathcal{T}$-measurable random variable then ξ is almost surely deterministic in that for some constant $c \in (-\infty, +\infty)$,

$$\mathbb{P}[\xi = c] = 1.$$

Proof of part (i).

Set $\mathcal{T} = \bigcap_n \mathcal{T}_n$, $\mathcal{T}_n = \sigma(\bigcup_{k \geq n+1} \mathcal{F}_k)$ for $n = 0, 1, 2, \ldots$. Set $\mathcal{H}_n = \sigma(\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n)$ for $n \in \mathbb{N}$. Pick $A \in \mathcal{T}$.
By the independence of $\{\mathcal{F}_n : n \geq 1\}$, $\mathcal{T}_n$ and $\mathcal{H}_n$, $n \geq 1$, are independent (use Lemma 4.4). Since $\mathcal{T} \subseteq \mathcal{T}_n$, $n \geq 1$, it follows that $\mathcal{T}$ and $\mathcal{H}_n$ are independent. Since $A \in \mathcal{T}$, A is independent of $\mathcal{H}_n$, and hence independent of any event in $\bigcup_{n \geq 1} \mathcal{H}_n$.
Now $\bigcup_{n \geq 1} \mathcal{H}_n$ is a Π-system so Lemma 4.4 can be used to show that $\sigma(A)$ is independent of $\sigma(\bigcup_{n \geq 1} \mathcal{H}_n) = \sigma(\bigcup_{n \geq 1} \mathcal{F}_n)$.
But $\mathcal{T} \subseteq \sigma(\bigcup_{n \geq 1} \mathcal{F}_n) = \mathcal{T}_0$ and so $\sigma(A)$ is independent of $\mathcal{T}$. It follows that A is independent of itself, that is

$$\mathbb{P}[A] = \mathbb{P}[A \cap A] = \mathbb{P}[A]\ \mathbb{P}[A],$$

and therefore $\mathbb{P}[A] = 0$ or 1. □

Proof of part (ii).

By (i), $\mathbb{P}[\xi \leq x] = 0$ or 1 for every x. Set $c = \sup\{x : \mathbb{P}[\xi \leq x] = 0\}$.
There are three possible cases, $c = -\infty$, $c = +\infty$ and $|c| < \infty$.
Now $c = -\infty$ implies $\mathbb{P}[\xi \leq x] = 1$ for all x, which is not possible.
Also $c = +\infty$ implies $\mathbb{P}[\xi \leq x] = 0$ for all x, which is not possible.
If $|c| < \infty$, then $\mathbb{P}\left[\xi \leq c - \frac{1}{n}\right] = 0$ for all $n > 0$, so

$$0 = \mathbb{P}\left[\bigcup_{n=1}^{\infty} \{\xi \leq c - \frac{1}{n}\}\right] = \mathbb{P}[\xi < c].$$

On the other hand, since $\mathbb{P}\left[\xi \leq c + \frac{1}{n}\right] = 1$ for all $n > 0$,

$$1 = \mathbb{P}\left[\bigcap_{n=1}^{\infty} \{\xi \leq c + \frac{1}{n}\}\right] = \mathbb{P}[\xi \leq c].$$

Thus, $\mathbb{P}[\xi = c] = \mathbb{P}[\xi \leq c] - \mathbb{P}[\xi < c] = 1 - 0 = 1$. □

The Borel-Cantelli lemmas

Usually Kolmogorov's zero-one law is easy to apply and the only tricky part lies in deciding for a particular event whether the probability is zero or one. An important tool, both in its own right and as a way of doing this tricky part, is the twinned pair of results known as the Borel-Cantelli lemmas.

Before describing and proving these lemmas, we need to introduce some ideas to do with "limits of sets", and to deal with a technical result known as "Fatou's lemma".

Definition 4.8 (lim sups and lim infs of events)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple and $(E_n : n \in \mathbb{N})$ a sequence of events. Define:

(i) $$(E_n \text{ i.o.}) = (E_n \text{ infinitely often}) = \limsup E_n = \bigcap_m \bigcup_{n \geq m} E_n$$
$$= \{\omega : \text{ for every } m, \text{ there exists } n(\omega) \geq m \text{ such that } \omega \in E_{n(\omega)}\}$$
$$= \{\omega : \omega \in E_n \text{ for infinitely many } n\}.$$

(ii) $$(E_n \text{ ev.}) = (E_n \text{ eventually}) = \liminf E_n = \bigcup_m \bigcap_{n \geq m} E_n$$
$$= \{\omega : \text{ for some } m(\omega), \omega \in E_n \text{ for all } n \geq m(\omega)\}$$
$$= \{\omega : \omega \in E_n \text{ for all large } n\}.$$

Remark 4.9

$$(\liminf E_n)^c = \limsup E_n^c ; \qquad (E_n \text{ ev.})^c = (E_n^c \text{ i.o.})$$

Note that E_n i.o. (hence E_n ev.) can be considered as a tail event. Set $\mathcal{F}_n = \sigma(E_n) = \{E_n, E_n^c, \emptyset, \Omega\}$, so $\mathcal{T} = \bigcap_n \sigma\{\mathcal{F}_n, \mathcal{F}_{n+1}, \ldots\}$. We know that

$$[E_n \text{ i.o.}] = \limsup E_n = \bigcap_n \bigcup_{r \geq n} E_r = \bigcap_n G_n ,$$

where

$$G_n = \bigcup_{r \geq n} E_r, \quad n \in \mathbb{N}.$$

Since, for fixed $n \in \mathbb{N}$, $E_n \in \mathcal{T}_n = \sigma(E_n, E_{n+1}, \ldots)$, it follows that $G_n \in \mathcal{T}_n$. Since the G_n are decreasing,

$$[E_n \text{ i.o.}] = \bigcap_k G_k = \bigcap_{k \geq n} G_k \in \bigcap_n \mathcal{T}_n = \mathcal{T} .$$

Lemma 4.10 (Fatou's lemma)

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $(E_n : n \in \mathbb{N})$ a sequence of elements of $\mathcal{F}$.

(i) *Fatou's lemma for sets: (true for all measure spaces)*

$$\mu(\liminf E_n) \leq \liminf \mu(E_n)$$

(ii) *reverse Fatou's lemma for sets: if $\mu(\Omega) < \infty$*

$$\mu(\limsup E_n) \geq \limsup \mu(E_n) .$$

Proof.

(i) Let $F_m = \bigcap_{n\geq m} E_n$. Then $F_m \uparrow F = \liminf E_n$. Since $F_m \subseteq E_n$, $n \geq m$, it follows that $\mu(F_m) \leq \mu(E_n)$ for all $n \geq m$, and therefore $\mu(F_m) \leq \inf_{n\geq m} \mu(E_n)$.
By Lemma 2.18 $\mu(F) = \lim_m \mu(F_m)$. Thus,

$$\mu(F) = \lim_m \mu(F_m) \leq \lim_m \inf_{n\geq m} \mu(E_n)\,.$$

(ii) Similarly to (i), by using complements (*possible* since $\mu(\Omega) < \infty$!). Since $(\limsup E_n) = (\liminf E_n^c)^c$, by using (i) we have

$$\begin{aligned}\mu(\limsup E_n) &= \mu(\Omega) - \mu(\liminf E_n^c)\\ &\geq \mu(\Omega) - \liminf \mu(E_n^c) = \mu(\Omega) - \liminf(\mu(\Omega) - \mu(E_n))\\ &= \mu(\Omega) - (\mu(\Omega) + \liminf(-\mu(E_n))) = \limsup \mu(E_n),\end{aligned}$$

because $\liminf(-\mu(E_n)) = -\limsup \mu(E_n)$. □

Lemma 4.11 (Borel-Cantelli lemmas)

Let $(E_n : n \in \mathbb{N})$ be a sequence of events from a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then

- *(BC1) $\sum_n \mathbb{P}[E_n] < \infty$ implies $\mathbb{P}[\limsup E_n] = \mathbb{P}[E_n \text{ i.o.}] = 0$;*
- *(BC2) if the events $(E_n : n \in \mathbb{N})$ are independent then*

$$\sum_n \mathbb{P}[E_n] = \infty \quad \textit{implies} \quad \mathbb{P}[\limsup E_n] = \mathbb{P}[E_n \text{ i.o.}] = 1\,.$$

Warwick Statistics

Proof.

(BC1) Set $G_m = \bigcup_{n\geq m} E_n$, $m \in \mathbb{N}$. Since, for every $k \in \mathbb{N}$, $\bigcap_m G_m \subseteq G_k$, we have, for fixed $k \in \mathbb{N}$,

$$\mathbb{P}[\limsup E_n] = \mathbb{P}\left[\bigcap_m G_m\right] \leq \mathbb{P}[G_k] \leq \sum_{n\geq k} \mathbb{P}[E_n]\,.$$

Now let $k \to \infty$. Because $\Sigma_n \mathbb{P}[E_n] < \infty$, we know that $\Sigma_{n\geq k} \mathbb{P}[E_n] \to 0$ as $k \to \infty$. Thus the result follows. □

Warwick Statistics

Proof.

(BC2) By the independence of $(E_n : n \in \mathbb{N})$, we have that

$$\mathbb{P}\left[\bigcap_{m\leq n\leq r} E_n^c\right] = \prod_{m\leq n\leq r} \mathbb{P}[E_n^c] = \prod_{m\leq n\leq r} (1 - \mathbb{P}[E_n])\,.$$

For $x \geq 0$, $1 - x \leq e^{-x}$, which implies that
$\mathbb{P}[\bigcap_{m\leq n\leq r} E_n^c] = \prod_{m\leq n\leq r}(1 - \mathbb{P}[E_n]) \leq e^{-(\sum_{m\leq n\leq r} \mathbb{P}[E_n])}$.
Now let $r \to \infty$: since $\sum_n \mathbb{P}[E_n] = \infty$,
$\mathbb{P}[\bigcap_{m\leq n} E_n^c] \leq \lim_{r\to\infty} e^{-(\sum_{m\leq n\leq r} \mathbb{P}[E_n])} = 0$
Since $(\limsup E_n)^c = \liminf E_n^c = \bigcup_m \bigcap_{n\geq m} E_n^c$, and $\mathbb{P}[\bigcap_{n\geq m} E_n^c] = 0$, it follows that $(\limsup E_n)^c$ is a union of null-sets and therefore, by Lemma 2.18(c), a null-set itself. Thus, $\mathbb{P}[(\limsup E_n]^c) = 0$ or in other words, $\mathbb{P}[\limsup E_n] = 1$. □

Warwick Statistics

Remark 4.12

(BC1) and (BC2) combine to say that if $(E_n : n \in \mathbb{N})$ are independent then $\mathbb{P}[E_n \text{ i.o.}] = 0$ or 1 according to whether $\Sigma_n \mathbb{P}[E_n]$ converges or not.

Example 16

Let $(X_n : n \in \mathbb{N})$ be an i.i.d. sequence of random variables with

$$\mathbb{P}[X_n > x] = e^{-x}, \quad x \geq 0\,.$$

For which $\alpha > 0$ is it almost surely the case that $X_n > \alpha \log n$ infinitely often? Moreover, if $L = \limsup(X_n / \log n)$ then show $\mathbb{P}[L = 1] = 1$.

Answer:
For $\alpha > 0$,

$$\mathbb{P}[X_n > \alpha \log n] = n^{-\alpha}, \quad n \in \mathbb{N}\,.$$

By (BC1) and (BC2), considering convergence or otherwise of $\sum_{n=1}^{\infty} \mathbb{P}[X_n > \alpha \log n]$ and independence of the X_n's,

$$\mathbb{P}[X_n > \alpha \log n \text{ i.o.}] = \begin{cases} 0 & \text{if } \alpha > 1 \\ 1 & \text{if } \alpha \leq 1\,. \end{cases}$$

Let $L = \limsup(\frac{X_n}{\log n})$. We will show that $\mathbb{P}[L = 1] = 1$. Notice that L is measurable with respect to the tail σ-algebra for $\sigma(X_1)$, $\sigma(X_2)$, ... Therefore we know from Kolmogorov's zero-one law (Theorem 4.7) that L certainly must be almost surely deterministic! So the issue is, to establish what *is* the almost sure deterministic value of L? Set

$$L_n = \sup_{r \geq n} \frac{X_r}{\log r}, \quad n \in \mathbb{N}\,.$$

Then $L = \lim L_n$ (decreasing limit!) and $[L \geq 1] = \bigcap_n [L_n \geq 1]$.
We will show that $[X_n > \log n \text{ i.o.}] \subseteq [L \geq 1]$.

Since

$$\begin{aligned} [L_n \geq 1] &= \left[\sup_{r \geq n} \frac{X_r}{\log r} \geq 1\right] \\ &= \bigcap_k \bigcup_{r \geq n} \left[\frac{X_r}{\log r} > 1 - \frac{1}{k}\right], \end{aligned}$$

we have that

$$\begin{aligned} [L \geq 1] &= \bigcap_n \bigcap_k \bigcup_{r \geq n} \left[\frac{X_r}{\log r} > 1 - \frac{1}{k}\right] \\ &= \bigcap_k \bigcap_n \bigcup_{r \geq n} \left[\frac{X_r}{\log r} > 1 - \frac{1}{k}\right] \\ &= \bigcap_k \left[\frac{X_n}{\log n} > 1 - \frac{1}{k} \text{ i.o.}\right] \supseteq \left[\frac{X_n}{\log n} > 1 \text{ i.o.}\right]. \end{aligned}$$

Hence, by above,

$$\mathbb{P}[L \geq 1] \geq \mathbb{P}\left[\frac{X_n}{\log n} > 1 \text{ i.o.}\right] = \mathbb{P}[X_n > \log n \text{ i.o.}] = 1\,.$$

Next we need to show that $\mathbb{P}[L > 1] = 0$. Since, for a fixed $k \in \mathbb{N}$,

$$\begin{aligned} [L > 1 + \tfrac{2}{k}] &\subseteq [L \geq 1 + \tfrac{2}{k}] = \bigcap_m \left[\frac{X_n}{\log n} > 1 + \tfrac{2}{k} - \tfrac{1}{m} \text{ i.o.}\right] \\ &\subseteq \left[\frac{X_n}{\log n} > 1 + \tfrac{2}{k} - \tfrac{1}{k} \text{ i.o.}\right] = \left[\frac{X_n}{\log n} > 1 + \tfrac{1}{k} \text{ i.o.}\right], \end{aligned}$$

we have that, by above,

$$\mathbb{P}\left[L > 1 + \tfrac{2}{k}\right] \leq \mathbb{P}\left[X_n > (1 + \tfrac{1}{k}) \log n \text{ i.o.}\right] = 0\,.$$

Thus, $[L > 1] = \bigcup_k [L > 1 + \frac{2}{k}]$ is a $\mathbb{P}$-null event (being a countable union of $\mathbb{P}$-null events), that is $\mathbb{P}[L > 1] = 0$. It follows that $\mathbb{P}[L = 1] = 1$.

Example 17 (Coin tossing)

Let $\Omega = \{\omega : \omega = (\omega_1, \omega_2, \ldots), \omega_i \in \{H, T\}\}$. Let $(X_n, n \in \mathbb{N})$ be a sequence of random variables such that

$$X_n(\omega) = \begin{cases} 1 & \text{if } \omega_n = H \\ 0 & \text{if } \omega_n = T, \end{cases} \qquad \omega \in \Omega,\ n \in \mathbb{N}\,.$$

For a fixed $k \in \mathbb{N}$, consider events

$$A_n = \{\omega : (\omega_n, \omega_{n+1}, \ldots, \omega_{n+k-1}) = S\},\ n \in \mathbb{N}\,,$$

where S is any fixed k-long sequence of H and T. Suppose $\mathbb{P}[S] = 2^{-k}$ (fair coins). We will show that

(1) $[A_n \text{ i.o.}]$ is a tail event for the sequence $(X_n, n \in \mathbb{N})$.

(2) $\mathbb{P}[A_n \text{ i.o.}] = 1$.

Proof.

(1) follows directly from the comments made after Definition 4.8. For (2), first notice that $\mathbb{P}[A_n] = (\frac{1}{2})^k$, for all $n \in \mathbb{N}$. Thus, if the events $\{A_n, n \in \mathbb{N}\}$ were independent, we could apply (BC2). However, $\{A_n, n \in \mathbb{N}\}$ are not independent because of ω_n-overlaps. So let

$$B_n = \{\omega : (\omega_{(n-1)k+1}, \omega_{(n-1)k+2}, \ldots, \omega_{nk} = S\},\ n \in \mathbb{N}\,.$$

Then, $\mathbb{P}[B_n] = (\frac{1}{2})^k$, and $\{B_n \text{ i.o.}\} \subset \{A_n \text{ i.o.}\}$, and, more importantly, $\{B_n, n \in \mathbb{N}\}$ are independent. Thus, we can apply (BC2) to the sequence $\{B_n, n \in \mathbb{N}\}$. Since

$$\sum_{n=1}^{\infty} \mathbb{P}[B_n] = \infty\,,$$

we have that $\mathbb{P}[B_n \text{ i.o.}] = 1$, which implies that $\mathbb{P}[A_n \text{ i.o.}] = 1$. □

Runs in coin-tossing (I)

Suppose in the above we seek the probability of a run of heads, but require the run to be longer depending on how long we have been waiting. For example, if we have had to wait for about m coin tosses then the run must be of length of order constant $\times \log m$.
Fix $\alpha > 0$ and set $h_n = \lfloor 2^\alpha + 2^{2\alpha} + \ldots + 2^{\alpha n} \rfloor$, for $\lfloor x \rfloor$ the greatest integer less than x. We ask whether we see infinitely many runs, which must be of length n if contained in the time interval between the h_{n-1}^{th} coin toss and the h_n^{th} coin toss.
Let A_n be the event of a run of length n commencing with a coin toss in the range h_{n-1} up to h_n. There are $h_n - h_{n-1} \leq 2^{\alpha n} + 1$ possible starts, and each start has probability 2^{-n}. Using subadditivity of probability,

$$\mathbb{P}[A_n] \leq (2^{\alpha n} + 1) 2^{-n}\,.$$

Hence the first Borel-Cantelli lemma tells us $\mathbb{P}[A_n i.o.] = 0$ if $\alpha < 1$.

Runs in coin-tossing (II)

What if $\alpha \geq 1$?
Consider events $B_{n,r}$ corresponding to a run of n heads commencing at $h_{n-1} + (r-1)n + 1$. One can now see that the events $B_{n,1}, \ldots, B_{n,s_n}$ are independent (s_n is largest integer smaller than $(h_n - h_{n-1} - 1)/n$), and hence deduce the second Borel-Cantelli lemma applies to the sequence of independent events $B_{1,1}, \ldots, B_{1,s_1}, B_{2,1}, \ldots, B_{2,s_2}, \ldots$.
Now $\mathbb{P}[B_{n,r}] = 2^{-n}$ as before, and so we can calculate

$$\sum_n \sum_{nr \leq h_n - h_{n-1} - 2} \mathbb{P}[B_{n,r}] \geq \sum_n \frac{h_n - h_{n-1} - 2}{n} 2^{-n} \geq \sum_n \frac{2^{\alpha n} - 3}{n} 2^{-n}.$$

This sum diverges when $\alpha \geq 1$, and so the second Borel-Cantelli lemma tells us that in this case almost surely we will see infinitely many of the $B_{n,r}$, hence infinitely many of the A_n, hence infinitely many of such runs.

Lessons learned

1. Definitions of independence;
2. Use of Π-systems;
3. Product measures;
4. Tail σ-algebra and Kolmogorov zero-one law;
5. Language of i.o., ev., lim sup, lim inf for events;
6. Fatou and Borel-Cantelli lemmas;
7. Examples.

5: Expectation, Convergence and Uniform Integrability

"I suppose you expect me to talk."
"No, Mr. Bond. I expect you to die."
Goldfinger

Expectation, Convergence and Uniform Integrability

Table of contents of this section

5 Expectation and Modes of Convergence
- Integration
- Expectation
- Inequalities
- Almost sure convergence
- Convergence in probability
- Convergence in p-norm
- Uniform integrability
- SLLN
- Weak convergence
- Skorokhod Representation Theorem

Expectation is the probabilist's equivalent of both integration and summation.

Integration

Let (S, Σ, μ) be a σ-finite measure space. For suitable elements f of $m\Sigma$ we denote the Lebesgue integral of f with respect to μ by any of

$$\mu(f) \quad = \quad \int_S f(s)\mu(ds) \quad = \quad \int_S f d\mu \,.$$

Let $(m\Sigma)^+$ denote the set of non-negative Σ-measurable functions.
Example: Let the set S be countable. Suppose that $(S, \Sigma) = (\mathbb{N}, \mathcal{P}(\mathbb{N}))$ (where $\mathcal{P}(\mathbb{N})$ denotes the power set of $\mathbb{N}$) and suppose that μ is a measure on (S, Σ) such that $\mu(\{k\}) = 1$ for every $k \in \mathbb{N}$. Let $f \in m\Sigma$. Then,

$$\mu(f) = \int_S f(s)\mu(ds) = \int_S f d\mu = \sum_n f(n)\,.$$

Warwick Statistics

We define the Lebesgue integral of $f \in m\Sigma$ using a 4-point plan:

(1) let $A \in \Sigma$. The integral for the indicator function $\mathbb{I}[A]$ of a measurable set A is defined by

$$\mu(\mathbb{I}[A]) = \int_S \mathbb{I}[A]\, d\mu \quad = \quad \mu(A)\,, \qquad A \in \Sigma\,;$$

(2) an element f of $(m\Sigma)^+$ is called simple if f can be written as a finite sum

$$f = \sum_{i=1}^n a_i \mathbb{I}[A_i]\,,$$

where $a_i \in [0, \infty]$ and and $A_i \in \Sigma$, $i = 1, \ldots, n$; For a simple $f \in (m\Sigma)^+$ as given above, the integral is defined by

$$\mu(f) = \sum_{i=1}^n a_i\, \mu(A_i)\,.$$

Warwick Statistics

(3) given $f \in (m\Sigma)^+$, construct an increasing sequence of simple functions $(f_n : n \in \mathbb{N})$ such that $f_n \in (m\Sigma)^+$ for $n \in \mathbb{N}$, and $f_n \uparrow f$. The integral $\mu(f)$ is then defined by

$$\mu(f) \quad = \quad \int_S f d\mu \quad = \quad \lim_{n\to\infty} \mu(f_n) \quad = \quad \lim_{n\to\infty} \int_S f_n d\mu\,.$$

It can be shown that this definition of $\mu(f)$ for $f \in (m\Sigma)^+$ is unique regardless of the choice of the sequence $(f_n : n \in \mathbb{N})$. (This uses the idea of the monotone convergence theorem.) Note that this immediately implies that the integral of a function in $(m\Sigma)^+$ is non-negative.

(4) For an arbitrary measurable function f write $f = f^+ - f^-$ where

$$f^+(s) \quad = \quad \max(f(s), 0)\,, \qquad f^-(s) \quad = \quad \max(-f(s), 0)\,.$$

Then f^+ and f^- are both in $(m\Sigma)^+$ and $|f| = f^+ + f^-$.

Warwick Statistics

Definition 5.1 (integrable function, $\mathcal{L}^1(S, \Sigma, \mu)$)

For $f \in m\Sigma$ we say that f is integrable and write $f \in \mathcal{L}^1(S, \Sigma, \mu)$ if

$$\mu(|f|) = \mu(f^+) + \mu(f^-) < \infty\,,$$

and we *define* $\mu(f)$ by

$$\mu(f) = \int_S f d\mu = \mu(f^+) - \mu(f^-)\,.$$

Remark 5.2

Note that, for $f \in \mathcal{L}^1(S, \Sigma, \mu)$, $|\mu(f)| \leq \mu(|f|)$.

Note one of the purposes of the 4-point plan is to avoid having to consider $\infty - \infty$.
Note important results: Fatou lemma, MCT, DCT.

Warwick Statistics

Fubini theorem

Suppose we have a product measure $\mu_1 \otimes \mu_2$ on a product measure space $(S_1 \times S_2, \Sigma_1 \star \Sigma_2)$. The following is very useful:

Theorem 5.3 (Fubini theorem)

(i) *Suppose* $f \in \mathcal{L}^1(S_1 \times S_2, \Sigma_1 \star \Sigma_2, \mu_1 \otimes \mu_2)$. *Then*

$$\begin{aligned}(\mu_1 \otimes \mu_2)(f) &= \int \left(\int f(x_1, x_2) \mu_2(\mathrm{d}\, x_2) \right) \mu_1(\mathrm{d}\, x_1) \\ &= \int \left(\int f(x_1, x_2) \mu_1(\mathrm{d}\, x_1) \right) \mu_2(\mathrm{d}\, x_2)\end{aligned}$$

(ii) *Alternatively, suppose f is non-negative and $\Sigma_1 \star \Sigma_2$-measurable. Then the above is true in the extended sense, that if one of the double integrals is infinite then so are they all, and otherwise they must be equal.*

You should compare Theorem 5.3 to the discussion of product measures and independence in Example 14.

Remark 5.4

Let $(S, \Sigma) = (\mathbb{N}, \mathcal{P}(\mathbb{N}))$ and let μ be a measure on (S, Σ) such that $\mu(\{k\}) = 1$ for every $k \in \mathbb{N}$. Then, for $f \in m\Sigma$ we say that f is integrable and write $f \in \mathcal{L}^1(S, \Sigma, \mu)$ if

$$\mu(|f|) = \sum_n |f(n)| < \infty .$$

Expectation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. The expectation of a random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ is just the Lebesgue integral of X with respect to measure $\mathbb{P}$.

Definition 5.5 (expectation of a random variable)

For a random variable $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, the expectation $\mathbb{E}[X]$ is defined (if the corresponding integral exists) by

$$\mathbb{E}[X] = \int_\Omega X d\mathbb{P} = \int_\Omega X(\omega)\, \mathbb{P}[d\omega] .$$

The expectation is defined in the same way for $X \in (m\mathcal{F})^+$. In short, $\mathbb{E}[X] = \mathbb{P}[X]$.

Remark 5.6

We often write $\mathcal{L}^1$ for $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$.

From the definition of the expectation of a random variable, it follows that expectation is linear (over constants):

(i) $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

(ii) $\mathbb{E}[cX] = c\ \mathbb{E}[X]$, for $c \in \mathbb{R}$.

Lemma 5.7 (Exercise sheet 5.1)

Let X and Y be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ (for which expectations exist). Show that

(i) *if $X \geq 0$, then $\mathbb{E}[X] \geq 0$;*

(ii) *if $X \geq Y$, then $\mathbb{E}[X] \geq \mathbb{E}[Y]$;*

(iii) *if $X \geq 0$ and $X > 0$ with positive probability, then $\mathbb{E}[X] > 0$.*

Lemma 5.8 (expectation and distribution, the elementary formula for expectation)

Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and write $\mathcal{L}_X$ for the law of X on $(\mathbb{R}, \mathcal{B})$:

$$\mathcal{L}_X(B) \quad = \quad \mathbb{P}[X \in B], \quad B \in \mathcal{B}.$$

Suppose that $h : \mathbb{R} \to \mathbb{R}$ is a Borel measurable function. Then,

$$h(X) \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}) \quad \textit{if and only if} \quad h \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}, \mathcal{L}_X),$$

and then

$$\mathbb{E}[h(X)] = \mathcal{L}_X(h) = \int_{\mathbb{R}} h(x)\ \mathcal{L}_X(dx).$$

In particular, if it exists then

$$\mathbb{E}[X] = \int_{\mathbb{R}} x\ \mathcal{L}_X(dx).$$

Proof.

Let $h = \mathbb{I}[B]$ for some $B \in \mathcal{B}$. Then $h(X) = \mathbb{I}[B](X)$ is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and its expectation is, by the definition, given by

$$\mathbb{E}[h(X)] = \mathbb{E}[\mathbb{I}[B](X)] = \int_\Omega \mathbb{I}[B](X)(\omega)\ \mathbb{P}[d\omega]$$
$$= \int_\Omega I(X(\omega) \in B)\ \mathbb{P}[d\omega] = \mathbb{P}[X \in B] = \mathcal{L}_X(B) = \mathcal{L}_X(\mathbb{I}[B]) = \mathcal{L}_X(h).$$

(Recall discussion of $\mathcal{L}_X$ in and around Definition 3.21). Hence,

$$\mathbb{E}[h(X)] = \mathcal{L}_X(h) = \int_{\mathbb{R}} h(x)\ \mathcal{L}_X(dx)$$

is true if $h = \mathbb{I}[B]$ for some $B \in \mathcal{B}$. Linearity shows that it is also true if h is a simple function on $(\mathbb{R}, \mathcal{B})$. Monotone convergence then establishes the result for h a non-negative function, and finally linearity again completes the argument for any Borel measurable function h. □

Remark 5.9

If $F(x)$ is the distribution function of X then we write

$$\mathbb{E}[h(X)] \quad = \quad \int_{\mathbb{R}} h(x)\ dF(x).$$

If X is a discrete random variable, then

$$\mathbb{E}[h(X)] = \sum_k h(k)\,\mathbb{P}[X = k].$$

Lemma 5.10

If X is a non-negative random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ then

$$\mathbb{E}[X] = 0 \quad \textit{implies} \quad \mathbb{P}[X > 0] = 0.$$

Proof

We will prove the statement by contradiction. Suppose that $\mathbb{E}[X] = 0$ and $\mathbb{P}[X > 0] > 0$.
We know that

$$[X > \tfrac{1}{n}] \quad \uparrow \quad [X > 0] \quad = \quad \bigcup_n [X > \tfrac{1}{n}].$$

Thus, by Lemma 2.18 about monotone convergence properties of measures,

$$\mathbb{P}[X > 0] \quad = \quad \lim_{n\to\infty} \mathbb{P}\left[X > \tfrac{1}{n}\right].$$

If $\mathbb{P}[X > 0] > 0$ then $\mathbb{P}\left[X > \frac{1}{n}\right] > 0$ for at least one $n \in \mathbb{N}$.
(Otherwise, if $\mathbb{P}\left[X > \frac{1}{n}\right] = 0$ for all $n \in \mathbb{N}$ then we would have $\mathbb{P}[X > 0] = 0$.)

Pick a positive integer n such that $\mathbb{P}\left[X > \frac{1}{n}\right] > 0$. Then

$$\begin{aligned}
\mathbb{E}[X] &= \int_\Omega X d\mathbb{P} = \int_\Omega \mathbb{I}\left[X \leq \tfrac{1}{n}\right] X d\mathbb{P} + \int_\Omega \mathbb{I}\left[X > \tfrac{1}{n}\right] X d\mathbb{P} \\
&\geq \int_\Omega \mathbb{I}\left[X > \tfrac{1}{n}\right] X d\mathbb{P} \qquad \text{because } X \text{ is non-negative} \\
&\geq \frac{1}{n}\int_\Omega \mathbb{I}\left[X > \tfrac{1}{n}\right] d\mathbb{P} = \tfrac{1}{n}\ \mathbb{P}\left[X > \tfrac{1}{n}\right] > 0.
\end{aligned}$$

It follows that $\mathbb{E}[X] > 0$ which is a contradiction. Hence, $\mathbb{P}[X > 0] = 0$.

□

Given a product probability space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \star \mathcal{F}_2, \mathbb{P}_1 \otimes \mathbb{P}_2)$:

Theorem 5.11 (Fubini theorem for expectations)

(i) *Suppose $X = X(\omega_1, \omega_2)$ is a random variable on $\Omega_1 \times \Omega_2$ of finite expectation.*

$$\begin{aligned}
\mathbb{E}[X] &= \int\left(\int X(\omega_1,\omega_2) P_2(\mathrm{d}\,\omega_2)\right)\mathbb{P}_1(\mathrm{d}\,\omega_1) \\
&= \int\left(\int X(\omega_1,\omega_2)\mathbb{P}_1(\mathrm{d}\,\omega_1)\right)\mathbb{P}_2(\mathrm{d}\,\omega_2)
\end{aligned}$$

(ii) *Alternatively, suppose X is a non-negative random variable on $\Omega_1 \times \Omega_2$. Then the above is true in the extended sense, that if one of the double integrals is infinite then so are they all, and otherwise they must be equal.*

This applies when X is a function of two independent random variables (work with joint distribution measure).

Inequalities

Notation: for $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ (or $(m\mathcal{F})^+$) and $F \in \mathcal{F}$, define

$$\mathbb{E}[X; F] = \int_\Omega \mathbb{I}[F] X(\omega)\,\mathbb{P}[d\omega] = \int_F X(\omega)\,\mathbb{P}[d\omega] = \mathbb{E}[X\,\mathbb{I}[F]],$$

where $\mathbb{I}[F]$ is the indicator random variable for set F.
The following inequality, named after Markov, is simple but extraordinarily useful.

Lemma 5.12 (Markov's inequality)

Suppose that $Z \in m\mathcal{F}$ and that $g : \mathbb{R} \to [0, \infty]$ is $\mathcal{B}$-measurable and non-decreasing. Then for all real c

$$\mathbb{E}[g(Z)] \geq \mathbb{E}[g(Z); Z \geq c] \geq g(c)\,\mathbb{P}[Z \geq c].$$

Proof.

Since $Z \in m\mathcal{F}$ and that $g : \mathbb{R} \to [0, \infty]$ is $\mathcal{B}$-measurable, it follows from Lemma 3.10 that $g(Z) = g \circ Z \in (m\mathcal{F})^+$. Thus, the expectation $\mathbb{E}[g(Z)]$ is well-defined and we have that

$$\begin{aligned}
\mathbb{E}[g(Z)] &= \int_{Z<c} g(Z)\ d\mathbb{P} + \int_{Z\geq c} g(Z)\ d\mathbb{P} \\
&\geq \int_{Z\geq c} g(Z)\ d\mathbb{P} \qquad \text{because } g \geq 0 \\
&\geq \int_{Z\geq c} g(c)\ d\mathbb{P} \qquad \text{because } g \text{ is non-decreasing} \\
&= g(c)\int_{Z\geq c} d\mathbb{P} = g(c)\ \mathbb{P}[Z \geq c].
\end{aligned}$$

□

Example 18

For $Z \in (m\mathcal{F})^+$, $\quad c\ \mathbb{P}[Z \geq c] \leq \mathbb{E}[Z], \quad c > 0.$

Example 19

For $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, $\quad c\ \mathbb{P}[|X| \geq c] \leq \mathbb{E}[|X|], \quad c > 0.$

Suppose $Z \geq 0$ and $\mathbb{E}[Z] = \mu < \infty$. Then Markov's inequality shows that

$$\mathbb{P}[Z \geq z] \quad \leq \quad \frac{\mu}{z} \quad \text{for } z > 0\,,$$

and hence the distribution of Z is "smaller" than the Pareto distribution of index 1, with distribution function

$$F(z) \quad = \quad 1 - \frac{\mu}{z} \qquad \text{for } z \geq \mu\,.$$

We can now control deviations from the mean using Chebyshev's inequality.

Lemma 5.13 (Chebyshev's inequality)

Suppose that a random variable X has a finite variance, that is

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

Then, for all $\varepsilon > 0$,

$$\varepsilon^2\ \mathbb{P}[|X - \mathbb{E}[X]| \geq \varepsilon] \leq Var(X)\,.$$

Proof.

Apply Markov's inequality to $Z = (X - \mathbb{E}[X])^2$ with $c = \varepsilon^2$ and $g(c) = c$, $c \geq 0$, otherwise $g(c) = 0$. $\square$

Prototype of Law of Large Numbers

Suppose $X_1, \ldots, X_n$ are independent with common mean μ and common variance σ^2. Then

$$\overline{X}_n \quad = \quad \frac{X_1 + \ldots + X_n}{n}$$

has mean μ and variance σ^2/n, and therefore by Chebyshev's inequality, for any $\varepsilon > 0$,

$$\mathbb{P}[|\overline{X}_n - \mu| > \varepsilon] \quad \leq \quad \frac{\sigma^2}{n\varepsilon^2}\,.$$

Note that the assumptions can be weakened (no need for the distribution to be the same, so long as common mean and variance. No need for independence, so long as not correlated.)
On the other hand, with the above assumptions much more can be said (central limit theorem, as discussed later).

Definition 5.14 (convex function)

A function $f : G \to \mathbb{R}$, where G is an open subinterval of $\mathbb{R}$, is called convex on G if its graph lies below any of its chords, that is for $x, y \in G$ and $0 \leq p, q \leq 1$, $p + q = 1$,

$$f(px + qy) \leq p\,f(x) + q\,f(y)\,.$$

It can be shown that:
(i) a convex function is continuous;
(ii) if a function f is twice differentiable, then f is convex if and only if $f'' \geq 0$.
Important examples of convex functions: $|x|$, x^2, $e^{\theta x}$ ($\theta \in \mathbb{R}$), $|x|^p$ for $p \geq 1$.

The fundamental inequality for convex functions is named after Jensen. It has a whole host of implications.

Theorem 5.15 (Jensen's inequality)

Suppose that $\mathbb{E}[|X|] < \infty$, f is convex on range of X and $\mathbb{E}[|f(X)|) < \infty$. Then

$$\mathbb{E}[f(X)] \quad \geq \quad f(\mathbb{E}[X]) .$$

We will find this of great use later on. Note that the property of convexity itself is just Jensen's inequality in the special case when X takes on only two values.

Proof.

Let f be a convex function. Then it can be shown that there exist real sequences $(a_n, n \in \mathbb{N})$ and $(b_n, n \in \mathbb{N})$ such that $f(x) = \sup_n(a_n x + b_n)$ (Williams 1991, Section 6.6) when x is in the interior of the domain of definition of f.
Thus, for fixed $n \in \mathbb{N}$,

$$f(X) \quad \geq \quad a_n X + b_n ,$$

which implies that

$$\mathbb{E}[f(X)] \quad \geq \quad a_n \mathbb{E}[X] + b_n ,$$

and therefore,

$$\mathbb{E}[f(X)] \quad \geq \quad \sup_n (a_n \mathbb{E}[X] + b_n) \quad = \quad f\Big(\mathbb{E}[X]\Big) .$$

□

Almost sure convergence

There is much much more to be learnt about inequalities (see for example Steele 2004). However we will now consider approximation for random variables, using various ideas of *convergence*.

Definition 5.16 (convergence almost surely)

Let $(X_n : n \in \mathbb{N})$ be a sequence of random variables and let X be a random variable all on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. It is said that the sequence $(X_n : n \in \mathbb{N})$ converges almost surely to a random variable X if

$$\mathbb{P}[\{\omega : X_n(\omega) \to X(\omega)\}] \quad = \quad 1 .$$

We write $X_n \to X$ a.s. or $X_n \xrightarrow{a.s.} X$ or $\mathbb{P}[X_n \to X] = 1$.

Remark 5.17

Almost sure limits are unique only up to equality almost surely. We say that $X = Y$ a.s. if $\mathbb{P}[\omega : X(\omega) = Y(\omega)] = 1$. Thus, if $X_n \to X$ a.s. and $X_n \to Y$ a.s., then $\mathbb{P}[X = Y] = 1$.

Remark 5.18

Note that

$$\begin{aligned}\mathbb{P}[X_n \to X] \quad &= \quad \mathbb{P}[\{\omega : X_n(\omega) \to X(\omega)\}] \\ &= \quad \mathbb{P}\left[\bigcap_{k=1}^{\infty} \bigcup_m \bigcap_{n \geq m} \{|X_n(\omega) - X(\omega)| \leq \tfrac{1}{k}\}\right] .\end{aligned}$$

Hence often we specify random variables only within the *equivalence class* corresponding to almost sure equality.

In the following definition we define the Cauchy criterion for convergence which is useful when no limit is explicitly specified.

Definition 5.19 (Cauchy criterion for convergence)

A sequence $(X_n : n \in \mathbb{N})$ Cauchy converges almost surely if there exists a null set N such that for every $\omega \in \Omega \setminus N$ and every $\varepsilon > 0$, there exists $m(\omega, \varepsilon) \in \mathbb{N}$ such that

$$n' > n \geq m(\omega, \varepsilon) \implies |X_n(\omega) - X_{n'}(\omega)| \leq \varepsilon\,,$$

that is if

$$\begin{aligned} 1 &= \mathbb{P}\left[\omega : (X_n(\omega) : n \in \mathbb{N}) \text{ is Cauchy}\right] \\ &= \mathbb{P}\left[\bigcap_{k=1}^{\infty} \bigcup_{m} \bigcap_{n > n' \geq m} \{|X_n(\omega) - X_{n'}(\omega)| \leq \tfrac{1}{k}\}\right]. \end{aligned}$$

It can be shown that the sequence $(X_n : n \in \mathbb{N})$ converges almost surely if and only if the sequence Cauchy converges almost surely.

Example 20

If $X_n = X + x_n$ for x_n non-random then $(X_n : n \in \mathbb{N})$ converges almost surely if and only if $(x_n : n \in \mathbb{N})$ converges.

The following result is important and very useful. It shows how to translate almost sure convergence into a condition which resembles convergence in probability (see Definition 5.23) and so will help to relate the two kinds of convergence.

Lemma 5.20

The sequence $(X_n : n \in \mathbb{N})$ converges almost surely to X (a) if and only if for all $\varepsilon > 0$ we have

$$\lim_{m\to\infty} \mathbb{P}\left[|X_n - X| > \varepsilon \text{ for some } n \geq m\right] \quad = \quad 0\,,$$

(b) if and only if for all $\varepsilon > 0$ we have

$$\mathbb{P}\left[|X_n - X| > \varepsilon \;\; i.o.\right] \quad = \quad 0\,.$$

Proof.

Exercise sheet 6.1 and 6.2. □

Remark 5.21

By Lemma 5.20, $\mathbb{P}\left[|X_n - X| \leq \frac{1}{k} \text{ ev.}\right] = 1$ for all $k \in \mathbb{N}$ if and only if $X_n \to X$ a.s.

(Just take complements . . . .)

Suppose that $(X_n : n \in \mathbb{N})$ is a sequence of random variables and that X is a random variable. We rephrase the convergence theorems for integrable functions in the notation of expectation:

monotone convergence theorem (MON): if $0 \le X_n \uparrow X$, then $\mathbb{E}[X_n] \uparrow \mathbb{E}[X] \le \infty$;

Fatou's lemma: if $X_n \ge 0$ and $X_n \to X$ a.s. then $\mathbb{E}[X] \le \liminf \mathbb{E}[X_n]$;

dominating convergence theorem (DOM): if $X_n \to X$ a.s. and $|X_n(\omega)| \le Y(\omega)$ for all $n \in \mathbb{N}$ and $\omega \in \Omega$ where $\mathbb{E}[Y] < \infty$, then

$$\mathbb{E}[|X_n - X|] \longrightarrow 0 \quad \text{as } n \to \infty,$$

so that

$$\mathbb{E}[X_n] \to \mathbb{E}[X] \quad \text{as } n \to \infty;$$

Scheffé's lemma: if $X_n \to X$ a.s. and $\mathbb{E}[|X_n|] \to \mathbb{E}[|X|]$ as $n \to \infty$, then

$$\mathbb{E}[|X_n - X|] \to 0 \quad \text{as } n \to \infty\,;$$

bounded convergence theorem (BDD): if $X_n \to X$ a.s. and for some finite constant K, $|X_n(\omega)| \le K$ for all $n \in \mathbb{N}$ and $\omega \in \Omega$, then

$$\mathbb{E}[|X_n - X|] \to 0 \quad \text{as } n \to \infty\,.$$

Remark 5.22

It is easy to see that (BDD) is an immediate consequence of (DOM) obtained by taking $Y(\omega) = K$ (since $\mathbb{P}[\Omega] = 1$ we have that $\mathbb{E}[Y] < \infty$).

Convergence in probability

Almost sure convergence is a rather natural notion hitching on the back of the usual ideas of convergence from mathematical analysis. However, it does not appear to tell much about probabilities, and relates to a whole sequence of random variables rather than just to one random variable at a time. Convergence in probability moves us one step on the way towards information about probabilities alone.

Definition 5.23 (convergence in probability)

Let $(X_n : n \in \mathbb{N})$ be a sequence of random variables and let X be a random variable all on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then X_n converges in probability to X if

$$\mathbb{P}[|X_n - X| > \varepsilon] \to 0 \qquad \text{as } n \to \infty \text{ for all } \varepsilon > 0\,.$$

We write $X_n \to X$ in probability or $X_n \overset{prob}{\longrightarrow} X$.

Sometimes we have to deal with questions of convergence when no limit is in evidence. For almost sure convergence we saw that this is immediately reducible to the numerical case where Cauchy criterion is applicable.
The obvious analogue here is

Definition 5.24 (Cauchy convergence in probability)

The sequence of random variables $(X_n : n \in \mathbb{N})$ on $(\Omega, \mathcal{F}, \mathbb{P})$ Cauchy converges in probability if for all $\varepsilon > 0$

$$\lim_{n,m\to\infty} \mathbb{P}[|X_n - X_m| > \varepsilon] = 0\,.$$

It can be shown that if a sequence $(X_n : n \in \mathbb{N})$ Cauchy converges in probability then there exist a finite random variable X such that $X_n \to X$ in probability.

In fact,

Lemma 5.25

Cauchy convergence in probability and convergence to a limit in probability are equivalent.

Suppose $(X_n : n \in \mathbb{N})$ converges in probability to X. Then, by the triangle inequality, $|X_n - X_m| \leq |X_n - X| + |X - X_m|$. Thus,

$$[|X_n - X_m| > \varepsilon] \quad \subseteq \quad [|X_n - X| > \tfrac{\varepsilon}{2}] \cup [|X - X_m| > \tfrac{\varepsilon}{2}],$$

which implies that

$$\mathbb{P}[|X_n - X_m| > \varepsilon] \quad \leq \quad \mathbb{P}\left[|X_n - X| > \tfrac{\varepsilon}{2}\right] + \mathbb{P}\left[|X - X_m| > \tfrac{\varepsilon}{2}\right]$$

and that

$$\lim_{m,n\to\infty} \mathbb{P}[|X_n - X_m| > \varepsilon] \quad \leq \quad \lim_{n\to\infty} \mathbb{P}\left[|X_n - X| > \tfrac{\varepsilon}{2}\right] + \lim_{m\to\infty} \mathbb{P}\left[|X - X_m| > \tfrac{\varepsilon}{2}\right] = 0\,.$$

Hence, we get Cauchy convergence in probability.
Converse: Exercise. (HINT: Choose a subsequence $(n_k, k \in \mathbb{N})$ of naturals such that

$$\sum_k \mathbb{P}\left[|X_{n_{k+1}} - X_{n_k}| > \left(\frac{1}{2}\right)^k\right] \quad < \quad \infty\,.$$

Show that the subsequence $(X_{n_k}, k \in \mathbb{N})$ converges almost surely to a finite limit by checking the Cauchy criterion. Finally show that the whole sequence $(X_n : n \in \mathbb{N})$ converges in probability.)

Lemma 5.26

If $X_n \to X$ a.s. then $X_n \to X$ in probability.

Proof.

Suppose that $X_n \to X$ a.s. Then, by Lemma 5.20, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[|X_m - X| > \varepsilon \text{ for some } m \geq n] = 0\,.$$

Furthermore, for every $\varepsilon > 0$,

$$\{|X_n - X| > \varepsilon\} \subseteq \{|X_m - X| > \varepsilon \text{ for some } m \geq n\}\,.$$

Thus,

$$0 \leq \lim_{n\to\infty} \mathbb{P}[|X_n - X| > \varepsilon] \leq \lim_{n\to\infty} \mathbb{P}[|X_k - X| > \varepsilon \text{ for some } k \geq n] = 0,$$

and the lemma is proved. □

However, convergence in probability does not imply convergence almost surely. A counterexample is given in the following example.

Example 21 (convergence in probability does not imply convergence almost surely)

Let $(X_n : n \in \mathbb{N})$ be a sequence of independent random variables such that

$$X_n = \begin{cases} 1 & \text{with probability } \frac{1}{n} \\ 0 & \text{with probability } 1 - \frac{1}{n}, \end{cases} \qquad n \in \mathbb{N}\,.$$

Then, for $\varepsilon \geq 1$, $\mathbb{P}[|X_n| > \varepsilon] = 0$ and for $0 < \varepsilon < 1$, $\mathbb{P}[|X_n| > \varepsilon] = \frac{1}{n}$. Thus, for every $\varepsilon > 0$,

$$\mathbb{P}[|X_n| > \varepsilon] \to 0 \quad \text{as } n \to \infty\,,$$

which by the definition means that $X_n \to 0$ in probability. However, for $0 < \varepsilon < 1$,

$$\sum_n \mathbb{P}\left[|X_n| > \varepsilon\right] = \sum_n \frac{1}{n} = +\infty\,,$$

and because the random variables $\{X_n : n \in \mathbb{N}\}$ are independent we can apply (BC2) and conclude that

$$\mathbb{P}\left[|X_n| > \varepsilon \text{ i.o.}\right] = 1\,.$$

Therefore, by Lemma 5.20, X_n does not converge to 0 almost surely. In addition, since $X_n \to 0$ in probability, it follows by Lemma 5.26 that $(X_n : n \in \mathbb{N})$ cannot converge almost surely to any other limit than 0. Thus, $(X_n : n \in \mathbb{N})$ does not converge almost surely.[2]

[2] See Kendall (1993) for an example from empirical Bayesian statistics where this matters.

If we know that convergence in probability is happening "quickly" in the sense stated in the following lemma then we can conclude that we have almost sure convergence.

Lemma 5.27

Let $(X_n : n \in \mathbb{N})$ be a sequence of random variables and let X be a random variable all on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $\sum_n \mathbb{P}\left[|X_n - X| > \varepsilon\right] < +\infty$ for all $\varepsilon > 0$, then $X_n \to X$ a.s.

Proof.

Suppose that $\sum_n \mathbb{P}\left[|X_n - X| > \varepsilon\right] < \infty$ for all $\varepsilon > 0$. Then, by (BC1),

$$\mathbb{P}\left[|X_n - X| > \varepsilon \text{ i.o.}\right] = 0 \text{ for all } \varepsilon > 0.$$

It now follows by Lemma 5.20 that $X_n \to X$ a.s. □

Lemma 5.28

If $X_n \to X$ in probability then there exists a sequence $\{n_k\}$ of integers increasing to infinity such that $X_{n_k} \to X$ a.s.

Suppose that $X_n \to X$ in probability. Then $\lim_{n\to\infty} \mathbb{P}\left[|X_n - X| > \frac{1}{k}\right] = 0$ for all $k \in \mathbb{N}$. It follows that for each $k \in \mathbb{N}$ we can find $n_k \in \mathbb{N}$ such that $\mathbb{P}\left[|X_{n_k} - X| > \frac{1}{k}\right] \le \frac{1}{2^k}$, and consequently,

$$\sum_k \mathbb{P}\left[|X_{n_k} - X| > \frac{1}{k}\right] \le \sum_k \frac{1}{2^k} < +\infty\,.$$

Thus, for any $\varepsilon > 0$,

$$\sum_{k>[\frac{1}{\varepsilon}]}^{\infty} \mathbb{P}\left[|X_{n_k} - X| > \varepsilon\right] \le \sum_{k>[\frac{1}{\varepsilon}]}^{\infty} \mathbb{P}\left[|X_{n_k} - X| > \frac{1}{k}\right] \le \sum_{k>[\frac{1}{\varepsilon}]}^{\infty} \frac{1}{2^k} < +\infty\,.$$

Hence, by Lemma 5.27 $X_{n_k} \to X$ a.s.

In fact

Lemma 5.29

$X_n \to X$ in probability if and only if every subsequence of $(X_n : n \in \mathbb{N})$ contains a further subsequence along which we have almost sure convergence to X.

Proof.

The proof is contained in Exercise sheet 7.7. □

Example 22 (another example showing that convergence in probability does not imply convergence almost surely)

Let $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}[0,1], Leb)$ and define the sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ by

$$X_{n,k}(\omega) = \mathbb{I}\left[\{\omega : \omega \in (\frac{k-1}{n}, \frac{k}{n}]\}\right], \quad \omega \in \Omega,\ 1 \leq k \leq n,\ n \in \mathbb{N}.$$

Then, for any $\varepsilon > 0$ and any fixed $n \in \mathbb{N}$ and $1 \leq k \leq n$,

$$\mathbb{P}\left[|X_{n,k}| > \varepsilon\right] = \mathbb{P}\left[X_{n,k} = 1\right] = \mathbb{P}\left[\{\omega : \omega \in (\frac{k-1}{n}, \frac{k}{n}]\}\right] = Leb\Big((\frac{k-1}{n}, \frac{k}{n}]\Big) = \frac{1}{n}.$$

Thus the sequence $X_{1,1}, X_{2,1}, X_{2,2}, X_{3,1}, \ldots$ converges to 0 in probability as $n \to \infty$.

However, for any fixed $n \in \mathbb{N}$,

$$\mathbb{P}\left[X_{n,k} = 1 \text{ for some } 1 \leq k \leq n\right] = 1.$$

Therefore, $\mathbb{P}\left[X_{n,k} = 1 \text{ i.o.}\right] = 1$ which implies that the sequence $X_{1,1}, X_{2,1}, X_{2,2}, X_{3,1}, \ldots$ does not converge almost surely. Furthermore, fix $k \in \mathbb{N}$ and $\omega \in \Omega$. Then $X_{n,k}(\omega) = 0$ for all $n > [\frac{k}{\omega}] + 1$ which implies that for every fixed $k \in \mathbb{N}$ and $\omega \in \Omega$, $X_{n,k}(\omega) \to 0$ as $n \to \infty$. It follows that for fixed $k \in \mathbb{N}$, the sequence $(X_{n,k}, n \geq k)$ converges almost surely to 0. Therefore, $(X_{n,k}, n \geq k)$ is a subsequence of $(X_{n,k}, n \in \mathbb{N}, 1 \leq k \leq n)$ which converges almost surely.

We restate the Bounded Convergence Theorem (BDD) under the weaker hypothesis of "convergence in probability" rather than "almost sure convergence".
The following is a very useful result.

Theorem 5.30 (Bounded Convergence Theorem (BDD))

Let $(X_n : n \in \mathbb{N})$ be a sequence of random variables and let X be a random variable all on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that $X_n \to X$ in probability and that for some $K \in [0, \infty)$ we have (for every $n \in \mathbb{N}$ and $\omega \in \Omega$) $|X_n(\omega)| \leq K$. Then

$$\mathbb{E}\left[|X_n - X|\right] \to 0 \quad \text{as } n \to \infty,$$

and in particular $\mathbb{E}\left[X_n\right] \to \mathbb{E}\left[X\right]$ as $n \to \infty$.

First we will show that $\mathbb{P}\left[|X| \leq K\right] = 1$. Since $|X_n| \leq K$ for every $n \in \mathbb{N}$, we have that for every $k \in \mathbb{N}$ and $n \in \mathbb{N}$,

$$\left[|X| > K + \tfrac{1}{k}\right] \quad \leq \quad \left(|X - X_n| > \tfrac{1}{k}\right).$$

Thus, for every $k \in \mathbb{N}$ and $n \in \mathbb{N}$,

$$\mathbb{P}\left[|X| > K + \tfrac{1}{k}\right] \quad \leq \quad \mathbb{P}\left[|X - X_n| > \tfrac{1}{k}\right],$$

which together with the hypothesis $X_n \to X$ in probability implies that $\mathbb{P}\left[|X| > K + \frac{1}{k}\right] = 0$ for every $k \in \mathbb{N}$. Hence,

$$\mathbb{P}\left[|X| > K\right] \quad = \quad \mathbb{P}\left[\bigcup_k \left\{|X| > K + \tfrac{1}{k}\right\}\right] = 0,$$

and therefore, $\mathbb{P}\left[|X| \leq K\right] = 1$.

Let $\varepsilon > 0$ be given. Then use Lemma 5.7:

$$\begin{aligned}
&\mathbb{E}\left[|X_n - X|\right] = \\
&\quad \mathbb{E}\left[|X_n - X|\,;\,|X_n - X| > \tfrac{\varepsilon}{3}\right] + \mathbb{E}\left[|X_n - X|\,;\,|X_n - X| \le \tfrac{\varepsilon}{3}\right] \\
&\quad\quad \le 2K\,\mathbb{P}\left[|X_n - X| > \tfrac{\varepsilon}{3}\right] + \tfrac{\varepsilon}{3}\;\mathbb{P}\left[|X_n - X| \le \tfrac{\varepsilon}{3}\right] \\
&\quad\quad\quad \le 2K\,\mathbb{P}\left[|X_n - X| > \tfrac{\varepsilon}{3}\right] + \tfrac{\varepsilon}{3}.
\end{aligned}$$

Since $X_n \to X$ in probability, we can choose $n_0 \in \mathbb{N}$ such that for all $n \ge n_0$

$$\mathbb{P}\left[|X_n - X| > \tfrac{\varepsilon}{3}\right] < \frac{\varepsilon}{3K}.$$

Hence, for given $\varepsilon > 0$ we can find $n_0 \in \mathbb{N}$ such that for all $n \ge n_0$

$$\mathbb{E}\left[|X_n - X|\right] \quad < \quad 2K\,\frac{\varepsilon}{3K} + \frac{1}{3}\,\varepsilon \quad = \quad \varepsilon\,.$$

It follows that $\mathbb{E}\left[|X_n - X|\right] \to 0$ as $n \to \infty$.

Furthermore, we can use $\mathbb{E}[Y] = \mathbb{E}[\max\{Y,0\}] - \mathbb{E}[-\min\{Y,0\}]$, $\mathbb{E}[|Y|] = \mathbb{E}[\max\{Y,0\}] + \mathbb{E}[-\min\{Y,0\}]$, to show

$$0 \le |\,\mathbb{E}[X_n] - \mathbb{E}[X]\,| = |\,\mathbb{E}[X_n - X]\,| \le \mathbb{E}[|X_n - X|]\,.$$

Thus, if $\mathbb{E}[|X_n - X|] \to 0$ as $n \to \infty$ then $|\,\mathbb{E}[X_n] - \mathbb{E}[X]\,| \to 0$ as $n \to \infty$, that is $\mathbb{E}[X_n] \to \mathbb{E}[X]$ as $n \to \infty$.

Convergence in p-norm

Convergence in probability is related to the notion of approximation which says two things are close if they are likely to be not far apart, but which does not control how far apart they might get in bad cases. If you might be penalized more for a bad miss, then you will want something more detailed.
Recall the definition of $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ (Definitions 5.1 and 5.5): we say that a random variable $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ if $\mathbb{E}[|X|] < \infty$. In a similar way we define spaces $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ for $1 < p < \infty$.

Definition 5.31 (the $\mathcal{L}^p$ spaces)

For $1 \le p < \infty$ we say that $X \in \mathcal{L}^p = \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ if $\mathbb{E}[|X|^p] < \infty$, and we define a norm in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ by

$$\|X\|_p \quad = \quad \left(\mathbb{E}[|X|^p]\right)^{\frac{1}{p}}.$$

Definition 5.32 (convergence in $\mathcal{L}^p$, $1 \le p < \infty$)

Let $(X_n : n \in \mathbb{N})$ be a sequence of random variables and let X be a random variable all on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $1 \le p < \infty$. It is said that the sequence $(X_n : n \in \mathbb{N})$ converges in $\mathcal{L}^p$ to a random variable X if each X_n, $n \in \mathbb{N}$, is in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$, $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ and

$$\|X_n - X\|_p \to 0 \qquad \text{as} \quad n \to \infty\,,$$

equivalently,

$$\mathbb{E}[|X_n - X|^p] \to 0 \qquad \text{as } n \to \infty\,.$$

We write $X_n \to X$ in $\mathcal{L}^p$ or $X_n \xrightarrow{\mathcal{L}^p} X$.

L^2 convergence is well-adapted to Chebyshev's inequality 5.13.

The triangle inequality generalizes usefully to the L^p case (note we need $p \geq 1$ here!)

Remark 5.33

If the random variables X_n are in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{E}[|X_n - X|^p] \to 0$ as $n \to \infty$ then X is also in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$. This follows from Minkowski's inequality ($\|U + V\|_p \leq \|U\|_p + \|V\|_p$) (here is where we need $p \geq 1$): hence

$$\|X\|_p \quad \leq \quad \|X - X_n\|_p + \|X_n\|_p \,.$$

For proof of Minkowski's inequality, see Exercise Sheet 8.5.

Lemma 5.34 (convergence to a limit in $\mathcal{L}^p$ if and only if Cauchy convergence in $\mathcal{L}^p$)

Let $1 \leq p < \infty$ and let $(X_n, n \in \mathbb{N})$ be a sequence in $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$. Then, there exists a random variable $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ such that

$$\mathbb{E}[|X_n - X|^p] \to 0 \qquad \text{as } n \to \infty$$

if and only if

$$\mathbb{E}[|X_n - X_m|^p] \to 0 \qquad \text{as } n, m \to \infty\,,$$

(i.e. if $\sup_{n,m \geq k} \|X_n - X_m\|_p \to 0$ as $k \to \infty$.)

Proof.

Left as exercise (Williams 1991, Section 6.10). □

Lemma 5.35

For $r \geq p \geq 1$, if $X \in \mathcal{L}^r(\Omega, \mathcal{F}, \mathbb{P})$, then $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ and

$$\|X\|_p \leq \|X\|_r \,.$$

Consequently,

$$X_n \to X \text{ in } \mathcal{L}^r \implies X_n \to X \text{ in } \mathcal{L}^p \,.$$

Proof.

Let $p \geq 1$. Define a sequence $(Y_n, n \in \mathbb{N})$ by $Y_n(\omega) = \left(\min(|X(\omega)|, n)\right)^p$ for $n \in \mathbb{N}$. Then, for fixed $n \in \mathbb{N}$, Y_n is bounded which implies that Y_n and $Y_n^{\frac{r}{p}}$ are both in $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Let $r \geq p$. Take $\phi(x) = x^{\frac{r}{p}}$ for $x \geq 0$, $\phi(x) = 0$, $x < 0$. Then ϕ is a convex function and we apply Jensen's inequality to obtain

$$(\mathbb{E}[Y_n])^{\frac{r}{p}} \leq \mathbb{E}\left[Y_n^{\frac{r}{p}}\right] = \mathbb{E}[(\min(|X(\omega)|, n))^r] \leq \mathbb{E}[|X|^r] \,.$$

Now, since $Y_n \uparrow |X|^p$, we can use (MON) to get that

$$(\mathbb{E}[|X|^p])^{\frac{r}{p}} \leq \mathbb{E}[|X|^r] \,,$$

which is equivalent to saying that $\|X\|_p \leq \|X\|_r$. □

Theorem 5.36

For $p \geq 1$, convergence in $\mathcal{L}^p$ implies convergence in probability.

Proof.

By Markov's inequality with $g(x) = x^p$ for $x \geq 0$, $g(x) = 0$ otherwise, and $Z = |X_n - X|$,

$$\mathbb{P}[|X_n - X| > \varepsilon] \leq \frac{\mathbb{E}[|X_n - X|^p]}{\varepsilon^p}.$$

□

Example 23 (convergence in $\mathcal{L}^s$ does not imply convergence in $\mathcal{L}^r$ for $r > s \geq 1$)

Let $r > s \geq 1$ and let $(X_n, n \in \mathbb{N})$ be a sequence of random variables such that

$$X_n = \begin{cases} n & \text{with probability } n^{-\frac{r+s}{2}} \\ 0 & \text{with probability } 1 - n^{-\frac{r+s}{2}}. \end{cases} \qquad n \in \mathbb{N}.$$

Then, $\mathbb{E}[|X_n|^s] = \mathbb{E}[X_n^s] = n^s\, n^{-\frac{r+s}{2}} = n^{\frac{s-r}{2}} \longrightarrow 0$ as $n \to \infty$. However, $\mathbb{E}[|X_n|^r] = \mathbb{E}[X_n^s] = n^r\, n^{-\frac{r+s}{2}} = n^{\frac{r-s}{2}} \longrightarrow \infty$ as $n \to \infty$. Hence, $(X_n, n \in \mathbb{N})$ converges to 0 in $\mathcal{L}^s$ but it does not converge in $\mathcal{L}^r$.

Example 24 (convergence in probability does not imply convergence in $\mathcal{L}^p$)

We need random variables increasingly likely to be zero but possibly huge. Let $1 \leq p < \infty$. Define independent random variables by

$$X_n = \begin{cases} n^{\frac{1}{p}} & \text{if } \; U < \frac{1}{n} \\ 0 & \text{if } \; U \geq \frac{1}{n}, \end{cases} \qquad n \in \mathbb{N}.$$

Then $X_n \to 0$ in probability, as for every $\varepsilon > 0$,

$$\mathbb{P}[|X_n| > \varepsilon] = \mathbb{P}\left[X_n = n^{\frac{1}{p}}\right] = \mathbb{P}\left[U < \tfrac{1}{n}\right] = \tfrac{1}{n} \to 0 \quad \text{as } n \to \infty.$$

However $(X_n, n \in \mathbb{N})$ does not converge in $\mathcal{L}^p$, as

$$\mathbb{E}[|X_n|^p] = \mathbb{E}\left[X_n^p\right] = (n^{\frac{1}{p}})^p\ \mathbb{P}\left[U < \tfrac{1}{n}\right] = n\,\tfrac{1}{n} = 1.$$

Example 25 (convergence almost surely does not imply convergence in $\mathcal{L}^p$)

Let $(X_n, n \in \mathbb{N})$ be a sequence of random variables such that

$$X_n = \begin{cases} n^3 & \text{with probability } \frac{1}{n^2} \\ 0 & \text{with probability } 1 - \frac{1}{n^2}. \end{cases} \qquad n \in \mathbb{N}.$$

Then, for every $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}[|X_n| > \varepsilon] = \sum_{n=1}^{\infty} \mathbb{P}\left[X_n = n^3\right] = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty,$$

which by Lemma 5.27 implies that $X_n \to 0$ almost surely.
Let $1 \leq p < \infty$. Then,

$$\mathbb{E}[|X_n|^p] = \mathbb{E}\left[X_n^p\right] = (n^{3p})\,\frac{1}{n^2} = n^{3p-2} \longrightarrow \infty \text{ as } n \to \infty.$$

Hence, $(X_n, n \in \mathbb{N})$ does not converge to 0 in $\mathcal{L}^p$.

Example 26 (convergence in $\mathcal{L}^p$ does not imply convergence almost surely)

Let the independent sequence $(X_n, n \in \mathbb{N})$ be as given in Example 21 (so X_n is Bernoulli and $\mathbb{P}[X_n = 1] = \frac{1}{n}$). Then, as shown in Example 21, $(X_n, n \in \mathbb{N})$ does not converge almost surely. However, for any $1 \leq p < \infty$,

$$\mathbb{E}[|X_n|^p] = \frac{1}{n} \longrightarrow 0 \text{ as } n \to \infty,$$

which implies that $(X_n, n \in \mathbb{N})$ converges to 0 in $\mathcal{L}^p$.

Uniform integrability

We have seen in the previous section that convergence in $\mathcal{L}^p$ implies convergence in probability and that the reverse is not necessarily true. However, under certain circumstances convergence in probability *does* imply convergence in $\mathcal{L}^1$. (In effect, sometimes but not always we can have $\mathbb{E}\lim = \lim \mathbb{E}$.)

In order to determine when this is true, we need a concept of uniform integrability: which means we need to re-phrase the notion of an integrable random variable to involve limits (so that we can then require the limit to be uniform).

Lemma 5.37 (Integrability condition)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then, a random variable $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ if and only if

$$\begin{aligned}\lim_{K\to\infty} \int_{\{\omega:|X(\omega)|>K\}} |X| \, d\mathbb{P} &= \lim_{K\to\infty} \mathbb{E}[|X| \, \mathbb{I}[|X| > K]] \\ &= \lim_{K\to\infty} \mathbb{E}[|X|; |X| > K] = 0 .\end{aligned}$$

Equivalently, if for any given $\varepsilon > 0$, there exists $K \in [0, \infty)$ such that

$$\mathbb{E}[|X|; |X| > K] < \varepsilon .$$

Proof

Suppose that $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Look at

$$\int_{\{\omega:|X(\omega)|>K\}} |X| \, d\mathbb{P} = \mathbb{E}[|X|; |X| > K] = \mathbb{E}[|X| \, \mathbb{I}[|X| > K]] .$$

The random variable $|X| \, \mathbb{I}[|X| > K]$ converges almost surely to zero and is dominated by $|X|$. Since $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, by the DOM theorem $\mathbb{E}[|X|; |X| > K]$ converges to zero as $K \to \infty$, that is

$$\lim_{K\to\infty} \int_{\{\omega:|X(\omega)|>K\}} |X| \, d\mathbb{P} = \lim_{K\to\infty} \mathbb{E}[|X|; |X| > K] = 0 .$$

Conversely suppose that

$$\lim_{K\to\infty}\int_{\{\omega:|X(\omega)|>K\}}|X|\,d\mathbb{P}=0\,.$$

Take K such that $\int_{\{|X|>K\}}|X|\,d\mathbb{P}\leq 1$. Then

$$\mathbb{E}\left[|X|\right]\;=\;\int_{\{|X|>K\}}|X|\,d\mathbb{P}+\int_{\{|X|\leq K\}}|X|\,d\mathbb{P}\;\leq\;1+K\;<\;\infty$$

which is equivalent to saying that $X\in\mathcal{L}^1(\Omega,\mathcal{F},\mathbb{P})$.

□

Now we can talk about uniformity:

Definition 5.38 (uniform integrability)

A set $\mathcal{C}$ of random variables all on the probability space $(\Omega,\mathcal{F},\mathbb{P})$ is said to be uniformly integrable if

$$\lim_{K\to\infty}\sup_{X\in\mathcal{C}}\int_{|X|>K}|X|\,d\mathbb{P}=\lim_{K\to\infty}\sup_{X\in\mathcal{C}}\mathbb{E}\left[|X|\,;\,|X|>K\right]=0\,.$$

Equivalently, if for any given $\varepsilon>0$, there exist $K\in[0,\infty)$ such that

$$\mathbb{E}\left[|X|\,;\,|X|>K\right]<\varepsilon,\quad\text{for all } X\in\mathcal{C}.$$

Remark 5.39

The sequence of random variables $(X_n:n\in\mathbb{N})$ on $(\Omega,\mathcal{F},\mathbb{P})$ is said to be uniformly integrable if

$$\lim_{K\to\infty}\sup_{n\in\mathbb{N}}\int_{|X_n|>K}|X_n|\,d\mathbb{P}=0\,.$$

Equivalently if for any given $\varepsilon>0$, there exist $K\in[0,\infty)$ such that

$$\mathbb{E}\left[|X_n|\,;\,|X_n|>K\right]<\varepsilon,\quad\text{for all } n\in\mathbb{N},$$

that is if $\sup_{n\in\mathbb{N}}\mathbb{E}\left[|X_n|\,;\,|X_n|>K\right]<\varepsilon$.

Example 27

Any finite collection of integrable random variables is uniformly integrable.

Lemma 5.40 (Two sufficient conditions for uniform integrability)

Let $\mathcal{C}$ be a set of random variables from $m\mathcal{F}$.

(a) *Suppose the random variables in $\mathcal{C}$ are bounded in $\mathcal{L}^p$ for some $p>1$, that is, for some constant $A\geq 0$, $\mathbb{E}\left[|X|^p\right]<A$, for all $X\in\mathcal{C}$. Then, $\mathcal{C}$ is uniformly integrable.*

(b) *Suppose the random variables in $\mathcal{C}$ are dominated by an integrable non-negative random variable Y, that is $\mathbb{E}\left[Y\right]<\infty$ and $|X|\leq|Y|$ for all $X\in\mathcal{C}$. Then, $\mathcal{C}$ is uniformly integrable.*

Proof of (a).

We look at the limit

$$\lim_{K\to\infty}\sup_{X\in\mathcal{C}}\int_{|X|>K}|X|\,d\mathbb{P} \quad=\quad \lim_{K\to\infty}\sup_{X\in\mathcal{C}}\mathbb{E}\left[|X|\,;\,|X|>K\right].$$

For $p>1$ we have that if $|X|\geq K>0$ then $|X|^p\geq K^{p-1}|X|$. Thus,

$$|X|\ \mathbb{I}\left[|X|>K\right] \;\leq\; |X|^p\ \mathbb{I}\left[|X|>K\right]/K^{p-1} \;\leq\; |X|^p/K^{p-1},$$

which implies that

$$\mathbb{E}\left[|X|;|X|>K\right]\leq\mathbb{E}\left[|X|^p;|X|>K\right]/K^{p-1}\leq\mathbb{E}\left[|X|^p\right]/K^{p-1}\leq A/K^{p-1}.$$

It follows that $\sup_{X\in\mathcal{C}}\mathbb{E}\left[|X|\,;\,|X|>K\right]\leq A/K^{p-1}$, and finally that $\lim_{K\to\infty}\sup_{X\in\mathcal{C}}\mathbb{E}\left[|X|\,;a,|X|>K\right]\leq\lim_{K\to\infty}A/K^{p-1}=0$. □

Proof of (b).

Since $|X|\leq Y$ for all $X\in\mathcal{C}$, it follows that $\mathbb{I}\left[|X|>K\right]\leq\mathbb{I}\left[Y>K\right]$ for any fixed $X\in\mathcal{C}$ and any $K>0$. Hence, for any $X\in\mathcal{C}$ and any $K>0$, $|X|\ \mathbb{I}\left[|X|>K\right]\leq Y\ \mathbb{I}\left[|X|>K\right]\leq Y\ \mathbb{I}\left[Y>K\right]$; which implies that, for any $X\in\mathcal{C}$ and any $K>0$,

$$\begin{aligned}\mathbb{E}\left[|X|\,;\,|X|>K\right] &\leq \mathbb{E}\left[Y\,;\,Y>K\right],\\ \sup_{X\in\mathcal{C}}\mathbb{E}\left[|X|\,;\,|X|>K\right] &\leq \mathbb{E}\left[Y\,;\,Y>K\right].\end{aligned}$$

Since Y is an integrable random variable it follows from Lemma 5.37 that $\mathbb{E}\left[Y\,;\,Y>K\right]\to0$ as $K\to\infty$. Therefore,

$$\lim_{K\to\infty}\sup_{X\in\mathcal{C}}\mathbb{E}\left[|X|\,;\,|X|>K\right] \quad\leq\quad \lim_{K\to\infty}\mathbb{E}\left[Y\,;\,Y>K\right]=0.$$

□

Example 28 (a family bounded in $\mathcal{L}^1(\Omega,\mathcal{F},\mathbb{P})$ is not necessarily uniformly integrable)

Condition given in part (a) in Lemma 5.40 is not sufficient if $p=1$, which is to say that a family bounded in $\mathcal{L}^1(\Omega,\mathcal{F},\mathbb{P})$ is not necessarily uniformly integrable.
For example, let $(\Omega,\mathcal{F},\mathbb{P})=([0,1],\mathcal{B}([0,1]),Leb)$ and let $E_n=(0,\frac{1}{n})$, $n\in\mathbb{N}$, be a sequence of events in $\mathcal{F}$. Let $(X_n:n\in\mathbb{N})$ be a sequence of random variables on $(\Omega,\mathcal{F},\mathbb{P})$ defined by $X_n=n\mathbb{I}\left[E_n\right]$, $n\in\mathbb{N}$. Then, $\mathbb{E}\left[|X_n|\right]=n\mathbb{P}\left[E_n\right]=1$ for all $n\in\mathbb{N}$ which means that the sequence $(X_n:n\in\mathbb{N})$ is bounded in $\mathcal{L}^1(\Omega,\mathcal{F},\mathbb{P})$.
However it is not uniformly integrable.

Proof.

Take $K>0$. Then, for any $n\in\mathbb{N}$,

$$|X_n|\ \mathbb{I}\left[|X_n|>K\right]=\begin{cases}0 & \text{if}\quad n\leq K\\ n\mathbb{I}\left[E_n\right] & \text{if}\quad n>K.\end{cases}$$

Thus, for $K>0$ and $n>K$

$$\mathbb{E}\left[|X_n|\,;\,|X_n|>K\right]=\mathbb{E}\left[|X_n|\ \mathbb{I}\left[|X_n|>K\right]\right]=\mathbb{E}\left[n\mathbb{I}\left[E_n\right]\right]=n\mathbb{P}\left[E_n\right]=1.$$

It follows that the sequence $(X_n:n\in\mathbb{N})$ is not uniformly integrable. ($\mathbb{E}\left[|X_n|;|X_n|>K\right]=1$ if $n\geq K$.) □

Example 29 (a uniformly integrable family is bounded in $\mathcal{L}^1$)

Let $\mathcal{C}$ be a set of uniformly integrable random variables. Then, by definition of uniform integrability, there exists $K \in [0, \infty)$ such that $\mathbb{E}[|X|; |X| > K] < 1$, for all $X \in \mathcal{C}$ (take $\varepsilon = 1$). Hence, for any $X \in \mathcal{C}$,

$$\mathbb{E}[|X|] = \mathbb{E}[|X|; |X| > K] + \mathbb{E}[|X|; |X| \leq K] \leq 1 + K < \infty,$$

which implies that $\mathcal{C}$ is bounded in $\mathcal{L}^1$.

We are now ready to answer the question from the beginning of this section regarding the circumstances under which convergence in probability implies convergence in $\mathcal{L}^1$.

Theorem 5.41 (A necessary and sufficient condition for $\mathcal{L}^1$ convergence)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(X_n : n \in \mathbb{N})$ be a sequence of random variables in $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ be a random variable. Then $X_n \to X$ in $\mathcal{L}^1$ if and only if the following two conditions are satisfied:

(i) *$X_n \to X$ in probability,*

(ii) *the sequence $(X_n : n \in \mathbb{N})$ is uniformly integrable.*

Proof

Suppose that conditions (i) and (ii) hold. For fixed $K \in [0, \infty)$ define a function $\varphi_K : \mathbb{R} \to [-K, K]$ by

$$\varphi_K(x) = \begin{cases} -K & \text{if} \quad x < -K \\ x & \text{if} \quad -K \leq x \leq K \\ K & \text{if} \quad x > K. \end{cases}$$

Then for all n

$$|\varphi_K(X_n) - X_n| = |K - |X_n|| \times \mathbb{I}[|X_n| > K] \quad \leq \quad 2|X_n| \, \mathbb{I}[|X_n| > K].$$

(In fact $|X_n| \, \mathbb{I}[|X_n| > K]$ would suffice as upper bound.) Let $\varepsilon > 0$. Since the sequence $(X_n : n \in \mathbb{N})$ is uniformly integrable, there exists $K_1 \in [0, \infty)$ such that, for all n, all $K \geq K_1$,

$$\mathbb{E}[|\varphi_K(X_n) - X_n|] \leq \frac{\varepsilon}{3}, \quad n \in \mathbb{N}.$$

Similarly, since $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, by Lemma 5.37 there exists $K_2 \in [0, \infty)$ such that, for all n, for all $K \geq K_2$,

$$\mathbb{E}[|\varphi_K(X) - X|] \quad \leq \quad \frac{\varepsilon}{3}.$$

Let $K = \max\{K_1, K_2\}$. Since $|\varphi_K(x) - \varphi_K(y)| \leq |x - y|$ for every $x, y \in \mathbb{R}$, it follows that for every $a > 0$,

$$\mathbb{P}[|\varphi_K(X_n) - \varphi_K(X)| > a] \quad \leq \quad \mathbb{P}[|X_n - X| > a],$$

and because $X_n \to X$ in probability, we conclude that $\varphi_K(X_n) \to \varphi_K(X)$ in probability. Furthermore, $|\varphi_K(X_n)| \leq K$ for every $n \in \mathbb{N}$. Thus, by (BDD) (Theorem 5.30) for given $\varepsilon > 0$ we can choose $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$,

$$\mathbb{E}[|\varphi_K(X_n) - \varphi_K(X)|] \quad < \quad \frac{\varepsilon}{3}.$$

Therefore, by the triangle inequality, for given $\varepsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$,

$$\mathbb{E}[|X_n - X|] \leq \mathbb{E}[|\varphi_K(X_n) - X_n|] + \mathbb{E}[|\varphi_K(X_n) - \varphi_K(X)|] + \mathbb{E}[|\varphi_K(X) - X|] < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

This means that $\mathbb{E}[|X_n - X|] \to 0$ as $n \to \infty$, or in other words, that $X_n \to X$ in $\mathcal{L}^1$.
The proof of the reverse statement is left as an exercise. See Williams (1991, Section 13.7).

□

Remark 5.42

Suppose $X_n, X \in \mathcal{L}^1$ and that the sequence $(X_n : n \in \mathbb{N})$ converges in probability to X and that there exists an integrable non-negative random variable Y such that $|X_n| \leq Y$, $n \in \mathbb{N}$. Then, by Lemma 5.40, the sequence $(X_n : n \in \mathbb{N})$ is uniformly integrable. It follows by Theorem 5.41 that the sequence $(X_n : n \in \mathbb{N})$ converges in $\mathcal{L}^1$. Hence this is a mild improvement of the Dominated Convergence Theorem (DOM), since we have a weaker condition imposed on the sequence $(X_n : n \in \mathbb{N})$, that is instead of convergence almost surely we only need convergence in probability.

Strong Law of Large Numbers (SLLN)

Theorem 5.43

Suppose $X_1, X_2, \ldots$ are independent and identically distributed, with finite mean μ. Then

$$\frac{1}{n}(X_1 + \ldots + X_n) \quad \to \quad \mu \quad \text{almost surely.}$$

Pairwise independence suffices!

Etemadi's proof of SLLN (I) [non-examinable in 2013-2014]

Proof:
Simplifications:

- We can suppose $X_i \geq 0$ (or simply split $X_i = X_i^+ - X_i^-$ and deal with each part separately);
- Set $Y_n = X_n \mathbb{I}[X_n \leq n]$: then $\sum_n \mathbb{P}[X_n > n] = \sum_n \mathbb{P}[X_1 > n] \leq \sum_{n=0}^\infty \sum_{m \geq n} \mathbb{P}[m < X_1 \leq m+1] \leq \sum_{m=0}^\infty \sum_{0 \leq n \leq m} \mathbb{P}[m < X_1 \leq m+1] = \sum_{m=0}^\infty (m+1)\mathbb{P}[m < X_1 \leq m+1] \leq \mathbb{E}[X_1 + 1] < \infty$, so (Borel-Cantelli I) almost surely $X_n = Y_n$ eventually, and it suffices to show

$$\frac{1}{n}(Y_1 + \ldots + Y_n) \quad \to \quad \mu \quad \text{almost surely.}$$

Etemadi's proof of SLLN (II) [non-examinable in 2013-2014]

Fix $\varepsilon > 0$, $\alpha > 1$, $k(n) = \lfloor \alpha^n \rfloor$. Set $S_n = Y_1 + \ldots + Y_n$.
Use Chebyshev's inequality:

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}\left[|S_{k(n)} - \mathbb{E}\left[S_{k(n)}\right]| > \varepsilon k(n)\right] &\leq \varepsilon^{-2} \sum_{n=1}^{\infty} \operatorname{Var}\left[S_{k(n)}\right]/k(n)^2 \\ = \varepsilon^{-2} \sum_{n=1}^{\infty} k(n)^{-2} \sum_{m=1}^{k(n)} \operatorname{Var}\left[Y_m\right] &= \varepsilon^{-2} \sum_{m=1}^{\infty} \operatorname{Var}\left[Y_m\right] \sum_{n:k(n)\geq m} k(n)^{-2} \\ &\leq \frac{4\alpha^2}{\alpha^2 - 1}\varepsilon^{-2} \sum_{m=1}^{\infty} \mathbb{E}\left[Y_m^2\right] m^{-2}.\end{aligned}$$

(Use $\alpha^2 \geq k(n) = \lfloor \alpha^n \rfloor \geq \alpha^n/2$ for $n \geq 1$, and $\operatorname{Var}\left[Y_m\right] \leq \mathbb{E}\left[Y_m^2\right]$.)

Etemadi's proof of SLLN (III) [non-examinable in 2013-2014]]

Now observe

$$\begin{aligned}\sum_{m=1}^{\infty} \mathbb{E}\left[Y_m^2\right] m^{-2} &= \sum_{m=1}^{\infty} m^{-2}\, \mathbb{E}\left[X_1^2; X_1 \leq m\right] \\ &= \mathbb{E}\left[X_1^2 \sum_{m\geq\lceil X_1\rceil}^{\infty} m^{-2}\right] \leq 1 + \mathbb{E}\left[X_1^2 \int_{\lceil X_1 \rceil}^{\infty} x^{-2}\,\mathrm{d}x\right] \\ &= 1 + \mathbb{E}\left[\frac{X_1^2}{\lceil X_1 \rceil}\right] \leq 1 + \mathbb{E}\left[X_1\right] < \infty.\end{aligned}$$

It follows that almost surely $|S_{k(n)} - \mathbb{E}\left[S_{k(n)}\right]| > \varepsilon k(n)$ for only finitely many n. But $\frac{1}{k(n)}(Y_1 + \ldots + Y_{k(n)}) = S_{k(n)}/k(n)$ and so $S_{k(n)}/k(n) \to \mu$ almost surely.
Handle intermediate values as follows. Suppose $k(n) \leq m \leq k(n+1)$. Then $S_{k(n)}/k(n+1) \leq S_m/m \leq S_{k(n+1)}/k(n)$, and use $\alpha > 1$ close to 1 with $(k(n+1))/\alpha - 1 \leq k(n) \leq k(n+1)$. $\square$

Weak convergence

So far we have introduced three modes of convergence for random variables. Almost sure convergence is closely tied to convergence in the familiar sense of analysis. Convergence in probability has better links with probabilities for individual random variables. $\mathcal{L}^p$ convergence measures some kind of mean size of approximation. There is a fourth mode of convergence, which is strongly linked to the idea of the actual calculated probabilities being close. This is the notion of weak convergence.
For now we restrict our attention to probability measures on $(\mathbb{R}, \mathcal{B})$. Note that $\mathbb{R}$ is a Polish space (complete, separable, metric space). We write $Prob(\mathbb{R})$ for the space of probability measures on $(\mathbb{R}, \mathcal{B})$, and $C_b(\mathbb{R})$ for the space of bounded continuous functions on $\mathbb{R}$. Let $\mu \in Prob(\mathbb{R})$ and $h \in C_b(\mathbb{R})$. Recall that $\mu(h) = \int_{\mathbb{R}} h\, d\mu$.

Some examples

- Consider maximum M_n of n independent Exponential random variables with unit mean. Then $\mathbb{P}\left[M_n \leq t + \log n\right] \to \exp\left(-e^{-t}\right)$.
- Consider X_n drawn uniformly from $\frac{1}{n}, \frac{2}{n}, \ldots, 1$. Then in some sense X_n converges to the Uniform distribution on $[0, 1]$.
- Consider n points (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) drawn independently and uniformly from the square $[-1, 1]^2$. If $R_n = \min\{\sqrt{X_i^2 + Y_i^2} : i = 1, \ldots, n\}$ then $\mathbb{P}\left[R_n > r\right] = (1 - \pi r^2/4)^n$ (use the fact that (X_1, Y_1) has probability area$(A)/4$ of falling in a Borel set $A \subseteq [-1, 1]^2$.) Hence $\sqrt{n} R_n$ has a limiting distribution.
- Consider the Central Limit Theorem. If X_1, X_2, ..., are independent and identically distributed, with finite common mean μ and variance σ^2, then $(X_1 + \ldots + X_n - n\mu)/(\sqrt{n}\sigma)$ has a limiting standard normal distribution.

What common framework can capture all these cases of limiting distribution?

Definition 5.44 (weak convergence of probability measures)

Let $(\mu_n : n \in \mathbb{N})$ be a sequence of probability measures in $Prob(\mathbb{R})$ and let $\mu \in Prob(\mathbb{R})$. We say that $(\mu_n : n \in \mathbb{N})$ converges weakly to μ if $\mu_n(h) \to \mu(h)$ as $n \to \infty$ for all $h \in C_b(\mathbb{R})$ and then we write $\mu_n \xrightarrow{w} \mu$.

Let $\mu \in Prob(\mathbb{R})$. Introduce the associated distribution function $F(x) = \mu((-\infty, x])$, $x \in \mathbb{R}$. Thus, there is a one-to-one correspondence between elements of $Prob(\mathbb{R})$ and distribution functions (see Lemma 3.25).

Definition 5.45 (weak convergence of distribution functions)

Let $(F_n : n \in \mathbb{N})$ be a sequence of distribution functions and F be a distribution function, and let $(\mu_n : n \in \mathbb{N})$ and μ refer respectively to their associated probability measures on $(\mathbb{R}, \mathcal{B})$. Then we say that $(F_n : n \in \mathbb{N})$ converges weakly to F if and only if $(\mu_n : n \in \mathbb{N})$ converges weakly to μ, and write $F_n \xrightarrow{w} F$, that is

$$F_n \xrightarrow{w} F \quad \text{if and only if} \quad \mu_n \xrightarrow{w} \mu\,.$$

We are interested in the case when F_n is the distribution function for some random variable X_n. Then we have for every $h \in C_b(\mathbb{R})$,

$$\mu_n(h) \quad = \quad \int_{\mathbb{R}} h(x)\, dF_n(x) \quad = \quad \mathbb{E}\left[h(X_n)\right].$$

Note that the statement $F_n \xrightarrow{w} F$ is meaningful even if the random variables X_n are defined on different probability spaces (in which case "$X_n \to X$ a.s." and "$X_n \to X$ in probability" are meaningless).

Definition 5.46 (weak convergence of random variables)

Consider random variables X_n and X. Let F_n and F be their corresponding distribution functions. We say that the sequence $(X_n : n \in \mathbb{N})$ converges weakly to X if $F_n \xrightarrow{w} F$, and write $X_n \xrightarrow{w} X$.

Example 30 (weak convergence of distribution functions)

Let $(X_n : n \in \mathbb{N})$ be a sequence of random variables such that $X_n = \frac{1}{n}$, $n \in \mathbb{N}$, and let $X = 0$. Let μ_n, $n \in \mathbb{N}$, be the law of X_n, and let μ be the law of X. Then, μ_n, $n \in \mathbb{N}$, has unit mass at $\frac{1}{n}$ $(\mu(\{\frac{1}{n}\}) = 1)$ and μ has unit mass at zero. Thus, for any $h \in C_b(\mathbb{R})$,

$$\mu_n(h) = h(\tfrac{1}{n}) \to h(0) = \mu(h), \quad n \to \infty\,,$$

which means that $\mu_n \xrightarrow{w} \mu$, and therefore, $F_n \xrightarrow{w} F$. However,

$$\lim_{n\to\infty} F_n(0) = 0 \neq 1 = F(0)\,.$$

The previous example shows $F_n \xrightarrow{w} F$ is possible even when $\lim_{n\to\infty} F_n(x) \neq F(x)$, for some $x \in \mathbb{R}$. We therefore introduce a notion of convergence for distribution functions which avoids considering x where the limiting distribution function is discontinuous:

Definition 5.47 (Convergence in distribution)

We say X_n *converges in distribution* to X if the corresponding distribution function $F_n(x)$ converges to $F_X(x)$ whenever x is a continuity point of F_X. We write $X_n \xrightarrow{d} X$, or $F_n \xrightarrow{d} F_X$.

This definition works well for real-valued random variables (all we are considering here). It does not generalize well to random variables taking values in other spaces (vector spaces, spheres, ...).

Examples of convergence of distribution functions:

- As noted above, if $X_1, X_2, \ldots$ are independent Exponential(1) (so $\mathbb{P}[X_n \leq t] = 1 - e^{-t}$) then $\mathbb{P}[\max\{X_1, \ldots, X_n\} - \log n \leq t] = (1 - e^{-t-\log n})^n = (1 - e^{-t}/n)^n \to \exp(-e^{-t})$. Here $\exp(-e^{-t})$ is the *extreme value distribution function*.
- Consider X_n with a Binomial$(n, \mu/n)$ distribution. Analysis shows that $\mathbb{P}[X_n = k] \to \mathbb{P}[\text{Poisson}(\mu) = k]$. It follows by summation that the Binomial$(n, \mu/n)$ distribution function converges in distribution to the Poisson(μ) distribution function.
- Consider X_n with a Binomial(n, p) distribution and study $(X_n - np)/\sqrt{np(1-p)}$. Similar techniques can be used to show that the distribution function of Y_n converges in distribution to the standard normal distribution function. (We will discuss more general cases in the next section.)

Convergence of distribution functions can often be checked very directly as above. However the theory of weak convergence is often easier.

Theorem 5.48 (Continuous mapping theorem)

Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is continuous, and $X_1, X_2, \ldots$ converge weakly to X. Then $Y_1 = \phi(X_1), Y_2 = \phi(X_2), \ldots$ converge weakly to $Y = \phi(X)$.

Proof.

Suppose $f : \mathbb{R} \to \mathbb{R}$ is bounded and continuous. Then $h = f \circ \phi$ is also bounded and continuous. Therefore
$\mathbb{E}[f(Y_n)] = \mathbb{E}[h(X_n)] \to \mathbb{E}[h(X)] = \mathbb{E}[f(Y)]$. □

Theorem 5.48 can be extended to the case when ϕ has discontinuities, so long as ϕ is almost surely continuous at the value Y. We now prove a special case of this.

Theorem 5.49 (A necessary and sufficient condition linking weak convergence to convergence of distribution functions)

Let $(F_n : n \in \mathbb{N})$ be a sequence of distribution functions on $\mathbb{R}$ and let F be a distribution function on $\mathbb{R}$. Let μ_n, μ be the corresponding probability measures. Then $F_n \xrightarrow{d} F$ if and only if the corresponding probability measures converge weakly: $\mu_n \xrightarrow{w} \mu$.

Proof

Suppose first that $F_n \xrightarrow{w} F$. Then $\mu_n(g) \to \mu(g)$ as $n \to \infty$ for every $g \in C_b(\mathbb{R})$.
For fixed $x \in \mathbb{R}$ and $\delta > 0$ define a function $h \in C_b(\mathbb{R})$ by

$$h(y) = \begin{cases} 1 & \text{if } y \leq x\,, \\ 1 - \frac{y-x}{\delta} & \text{if } x < y < x+\delta\,, \\ 0 & \text{if } y \geq x+\delta\,. \end{cases}$$

Then by assumption $\mu_n(h) \to \mu(h)$ as $n \to \infty$.
Observe that

$$F_n(x) \leq \mu_n(h) \quad \text{and} \quad \mu(h) \leq F(x+\delta)\,,$$

which implies that

$$\limsup_n F_n(x) \leq \limsup_n \mu_n(h) = \mu(h) \leq F(x+\delta)\,.$$

Since F is right-continuous, we can let $\delta \to 0$ to obtain

$$\limsup_n F_n(x) \leq F(x), \quad x \in \mathbb{R}\,. \tag{5.1}$$

Similarly, for fixed $x \in \mathbb{R}$ and $\delta > 0$ define a function $\tilde{h} \in C_b(\mathbb{R})$ by

$$\tilde{h}(y) = \begin{cases} 1 & \text{if } y \leq x-\delta\,, \\ 1 - \frac{y-(x-\delta)}{\delta} & \text{if } x-\delta < y < x\,, \\ 0 & \text{if } y \geq x\,. \end{cases}$$

Then $\mu_n(\tilde{h}) \to \mu(\tilde{h})$ as $n \to \infty$, $F_n(x) \geq \mu_n(\tilde{h})$ and $\mu(\tilde{h}) \geq F(x-\delta)$. It follows that

$$\liminf_n F_n(x) \geq \liminf_n \mu_n(\tilde{h}) = \mu(\tilde{h}) \geq F(x-\delta)\,.$$

So

$$\liminf_n F_n(x) \geq F(x-\delta), \quad x \in \mathbb{R}\,. \tag{5.2}$$

If F is continuous at x then $\lim_{\delta\downarrow 0} F(x-\delta) = F(x)$, and the inequalities (5.1) and (5.2) therefore imply the desired result.
The proof of the reverse statement is a consequence of the Skorokhod Representation Theorem 5.53 given in the next section.

□

Lemma 5.50

Let X_n and X be random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then

(i) *if $X_n \to X$ in probability then $X_n \xrightarrow{w} X$,*
and thus

(ii) *if $X_n \to X$ a.s. then $X_n \xrightarrow{w} X$.*

Proof of (i)

Suppose that $X_n \to X$ in probability. Let F_n and F be the distribution functions of X_n and X, respectively, so, for every $x \in \mathbb{R}$,

$$F_n(x) = \mathbb{P}[X_n \le x], \qquad F(x) = \mathbb{P}[X \le x].$$

Use Theorem 5.41, *or* let $\varepsilon > 0$ be fixed. Then

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

Hence,

$$F(x-\varepsilon) - \mathbb{P}[|X_n - X| > \varepsilon] \le F_n(x) \le F(x+\varepsilon) + \mathbb{P}[|X_n - X| > \varepsilon].$$

Letting $n \to \infty$ we obtain, for all $\varepsilon > 0$,

$$F(x-\varepsilon) \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le F(x+\varepsilon). \qquad (5.3)$$

Suppose that F is continuous at x. Then $F(x-\varepsilon) \uparrow F(x)$ and $F(x+\varepsilon) \downarrow F(x)$ as $\varepsilon \downarrow 0$. Setting $\varepsilon \downarrow 0$ in (5.3), we obtain

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

which implies that

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

Thus, at every point of continuity x of F we have that $\lim_{n\to\infty} F_n(x) = F(x)$, which means that $X_n \xrightarrow{w} X$.

□

Proof of (ii)

The statement follows from part (i) since convergence almost surely implies convergence in probability.

□

Exercise 5.51

Prove that convergence almost surely implies convergence in distribution (part (ii) of Lemma 5.50) by using (BDD).

The converse to the statement (i) in Lemma 5.50 is false unless the limiting random variable X is a constant:

Lemma 5.52

Let $(X_n : n \in \mathbb{N})$ be a sequence of random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $X_n \to a$ in distribution (where $a \in \mathbb{R}$ is a constant) then $X_n \to a$ in probability.

Proof.

Exercise sheet 8.14. □

Example 31 (convergence in distribution does not imply convergence in probability)

Let X be a Bernoulli random variable taking values 0 and 1 with equal probabilities $\frac{1}{2}$. Let $(X_n : n \in \mathbb{N})$ be a sequence of identical random variables given by

$$X_n = X, \qquad n \in \mathbb{N}.$$

Then, $X_n \xrightarrow{w} X$. Let $Y = 1 - X$. It follows that $X_n \xrightarrow{w} Y$ as well since X and Y have the same distribution. However, $(X_n : n \in \mathbb{N})$ cannot converge to Y in any other mode of convergence because $|X_n - Y| = 1$ for all $n \in \mathbb{N}$. So $X_n \xrightarrow{prob} X$ is impossible!

Before we close the section we make the summary of interrelationships between the four modes of convergence:

$$\begin{array}{ccccc} X_n \xrightarrow{a.s.} X & & & & \\ & \searrow & & & \\ & & X_n \xrightarrow{prob} X & \Longrightarrow & X_n \xrightarrow{w} X \\ & \nearrow & & & \\ X_n \xrightarrow{\mathcal{L}^p} X & & & & \end{array}$$

for any $p \geq 1$. Also, if $r > s \geq 1$ then

$$X_n \xrightarrow{\mathcal{L}^r} X \quad \Longrightarrow \quad X_n \xrightarrow{\mathcal{L}^s} X \,.$$

No other implications hold in general.

Skorokhod Representation Theorem

Theorem 5.53 (Skorokhod Representation Theorem)

Let $(F_n : n \in \mathbb{N})$ be a sequence of distribution functions on $\mathbb{R}$, let F be a distribution function on $\mathbb{R}$ and suppose that $F_n \xrightarrow{d} F$. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ carrying a sequence $(X_n : n \in \mathbb{N})$ of random variables and also a random variable X such that $F_n = F_{X_n}$, $F = F_X$, and

$$X_n \to X \quad a.s.$$

Proof.

Omitted: see Williams (1991, Section 17.3). □

Suppose that $(Y_n : n \in \mathbb{N})$ is a sequence of random variables and Y is a random variable such that $Y_n \xrightarrow{w} Y$. It may happen that $(Y_n : n \in \mathbb{N})$ fails to converge to Y in any other mode of convergence. However, by the Skorokhod Representation Theorem 5.53 there exists a sequence $(X_n : n \in \mathbb{N})$ distributed identically to $(Y_n : n \in \mathbb{N})$ and a random variable X distributed identically to Y such that $(X_n : n \in \mathbb{N})$ converges almost surely to X.
Easy examples:

$$\begin{aligned} \text{Exponential}(\lambda_n) \quad &\to \quad \text{Exponential}(\lambda) \qquad \text{if } \lambda_n \to \lambda\,, \\ \frac{1}{n}\text{Uniform}(\{1, 2, \ldots, n\}) \quad &\to \quad \text{Uniform}([0,1]) \qquad \text{if } n \to \infty\,. \end{aligned}$$

6: Characteristic Functions and the Central Limit Theorem

You are confused; but this is your normal state.

Characteristic Functions and the Central Limit Theorem

Table of contents of this section

6 A Tale of Two Limit Theorems
- Characteristic functions
- Central Limit Theorem
- Weak Law of Large Numbers
- Typical work-flow

The Central Limit Theorem is the classic heart of probability theory. It is also an extraordinarily effective computational tool.

Characteristic functions

The theory of characteristic functions forms an important tool for investigating probability distributions.

Definition 6.1 (characteristic function of a random variable)

Let X be a random variable. Let $F = F_X$ be the distribution function of X and let $\mu = \mathcal{L}_X$ denote the law of X. Then the characteristic function φ of X (or of F or of μ) is the map

$$\varphi : \mathbb{R} \to \mathbb{C}$$

defined by

$$\begin{aligned}\varphi(\theta) &= \mathbb{E}\left[e^{i\theta X}\right] = \mathbb{E}\left[\cos(\theta X)\right] + i\ \mathbb{E}\left[\sin(\theta X)\right] \\ &= \int_{\mathbb{R}} e^{i\theta x}\ \mu(dx) = \int_{\mathbb{R}} e^{i\theta x}\ dF(x).\end{aligned}$$

In effect, we just treat i as a constant which happens to satisfy $i^2 = -1$. (In more advanced work one has to be more sophisticated, but this suffices for our purposes.)

Remark 6.2

Since $|e^{i\theta x}| = 1$, the characteristic function $\varphi(\theta)$ is defined for all $\theta \in \mathbb{R}$. This is an important difference between characteristic functions and moment generating functions. (The moment generating function of a random variable X is given by $M_X(t) = \mathbb{E}\left[e^{tX}\right]$ and is defined when this expectation exists.)

Lemma 6.3 (properties of characteristic functions)

Let $\varphi = \varphi_X$ be the characteristic function of a random variable X. Then

(i) $\varphi(0) = 1$;

(ii) $|\varphi(\theta)| \leq 1$ *for all* $\theta \in \mathbb{R}$;

(iii) $\theta \mapsto \varphi(\theta)$ *is continuous on* $\mathbb{R}$;

(iv) $\varphi_{-X}(\theta) = \varphi_X(-\theta) = \overline{\varphi(\theta)}$ *for all* $\theta \in \mathbb{R}$;

(v) $\varphi_{aX+b}(\theta) = e^{i\theta b}\varphi_X(a\theta)$ *for all* $a, b, \theta \in \mathbb{R}$;

(vi) *if* $\mathbb{E}\left[|X|^n\right] < \infty$ *for some* $n \in \mathbb{N}$, *then* $\varphi_X^{(n)}(0) = i^n\,\mathbb{E}\left[X^n\right]$.

Proof.

Exercise sheet 8.16. □

We note three theoretical results:

(1) If X and Y are independent random variables, then

$$\varphi_{X+Y}(\theta) = \varphi_X(\theta)\varphi_Y(\theta), \qquad \theta \in \mathbb{R}.$$

(2) There is a one-to-one correspondence between probability measures on $(\mathbb{R}, \mathcal{B})$ and characteristic functions. Lévy's inversion formula (given below) shows explicitly how a distribution function can be reconstructed from a characteristic function.

(3) Weak convergence of distribution functions corresponds to convergence of the associated characteristic functions. Lévy's convergence theorem (given below) shows this.

In the following, $F(a-)$ denotes the left limit of function $F(x)$ at a, so $F(a-) = \lim_{x \uparrow a} F(x)$.

Theorem 6.4 (Lévy's Inversion Formula)

Let φ be the characteristic function of a random variable X which has law μ and distribution function F. Then, for $a, b \in \mathbb{R}$ such that $a < b$,

$$\begin{aligned}\lim_{T \uparrow \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-i\theta a} - e^{-i\theta b}}{i\theta} \varphi(\theta)\, d\theta &= \frac{1}{2}\mu(\{a\}) + \mu((a,b)) + \frac{1}{2}\mu(\{b\}) \\ &= \frac{1}{2}\left(F(b) + F(b-)\right) - \frac{1}{2}\left(F(a) + F(a-)\right).\end{aligned}$$

Moreover, if $\int_{\mathbb{R}} |\varphi(\theta)|\, d\theta < \infty$, then X has continuous probability density function f with $f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\theta x}\varphi(\theta)\, d\theta$.

Proof.

Omitted: see Williams (1991, Section 16.6). □

By Lévy's Inversion Formula, random variables X and Y have the same characteristic function if and only if they have the same distribution function.

Theorem 6.5 (Lévy's Convergence Theorem)

Let $(F_n : n \in \mathbb{N})$ be a sequence of distribution functions and let for every $n \in \mathbb{N}$, φ_n denote the characteristic function of F_n. Suppose that

(i) *$g(\theta) = \lim_{n\to\infty} \varphi_n(\theta)$ exists for all $\theta \in \mathbb{R}$. and that*

(ii) *$g(\cdot)$ is continuous at 0.*

Then g is a characteristic function for some distribution function F (that is $g = \varphi_F$), moreover $F_n \xrightarrow{w} F$.

Proof.

Omitted: see Williams (1991, Section 18.1). □

Central Limit Theorem

The Central Limit Theorem is one of the greatest results of mathematics. Here we derive it as a corollary of Lévy's Convergence Theorem.

First we need Taylor expansion estimates on characteristic functions. Let X be a random variable such that for some $k \in \mathbb{N}$, $\mathbb{E}\left[|X|^k\right] < \infty$. Then

$$\varphi(\theta) = \sum_{j=0}^{k} \frac{\mathbb{E}\left[X^j\right]}{j!}(i\theta)^j + o(\theta^k) \quad \text{as } \theta \to 0, \tag{6.1}$$

where $f(\theta) = o(g(\theta))$ as $\theta \to L$ means that $\frac{f(\theta)}{g(\theta)} \to 0$ as $\theta \to L$.

For a detailed proof of this, see Williams (1991, Section 18.3).

Theorem 6.6 (Central Limit Theorem (CLT))

Let $(X_n : n \in \mathbb{N})$ be a sequence of independent identically distributed random variables all with zero mean and finite non-zero variance σ^2. Define

$$S_n = X_1 + X_2 + \cdots + X_n, \quad n \in \mathbb{N},$$

and set

$$G_n = \frac{S_n}{\sigma\sqrt{n}}, \quad n \in \mathbb{N}.$$

Then, the sequence $(G_n : n \in \mathbb{N})$ converges in distribution to a random variable with $N(0,1)$ distribution, that is, for $x \in \mathbb{R}$,

$$\mathbb{P}[G_n \le x] \longrightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}}\, dy \qquad \text{as } n \to \infty.$$

Sketch Proof

Let $\theta \in \mathbb{R}$ be fixed. Then, by (6.1),

$$\begin{aligned} \varphi_{G_n}(\theta) &= \varphi_{S_n}\left(\frac{\theta}{\sigma\sqrt{n}}\right) = \left(\varphi_{X_1}\left(\frac{\theta}{\sigma\sqrt{n}}\right)\right)^n \\ &= \left(1 - \frac{1}{2}\frac{\theta^2}{n} + o\left(\frac{\theta^2}{\sigma^2 n}\right)\right)^n \quad \text{as } n \to \infty. \end{aligned}$$

(Justifying this asymptotic expansion is the real work of the proof!)

Let $z_n = -\frac{1}{2}\frac{\theta^2}{n} + o\left(\frac{\theta^2}{\sigma^2 n}\right)$, $n \in \mathbb{N}$. Then $x_n = nz_n \to -\theta^2/2$ so

$$\lim_{n\to\infty} \varphi_{G_n}(\theta) = \lim_{n\to\infty} \left(1 + x_n/n\right)^n = e^{-\frac{\theta^2}{2}}.$$

We know that $\varphi(\theta) = e^{-\frac{\theta^2}{2}}$ is the characteristic function of the $N(0,1)$ distribution (see table below). Hence, by Lévy's Convergence Theorem, $(G_n : n \in \mathbb{N})$ converges in distribution to a $N(0,1)$ random variable, which implies that

$\mathbb{P}[G_n \le x] \longrightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}}\, dy \qquad$ as $n \to \infty$ for every $x \in \mathbb{R}$.

$\square$

Theorem 6.7 (Weak Law of Large Numbers (WLLN))

Let $(X_n : n \in \mathbb{N})$ be a sequence of independent identically distributed random variables each with the same distribution as a random variable X. Suppose that $X \in \mathcal{L}^1$ and that $\mathbb{E}[X] = \mu$. Let $S_n = X_1 + X_2 + \cdots + X_n$, $n \in \mathbb{N}$ and show that

$$\frac{S_n}{n} \to \mu \quad \textit{in probability}.$$

Remark 6.8

Notice that we already proved a special version of this when the X_i's have finite variance: see the discussion following Chebyshev's inequality (Lemma 5.13).

Remark 6.9

In fact convergence holds almost surely: see Theorem 5.43 on the Strong Law of Large Numbers (SLLN).

Proof

By Lemmas (5.50) and (5.52) convergence in probability to a constant is equivalent to convergence in distribution to a constant. Thus, it is sufficient to prove that $\frac{S_n}{n} \to \mu$ in distribution.
We will prove this by showing that the sequence of characteristic functions of the sequence $(\frac{S_n}{n} : n \in \mathbb{N})$ converges to $\varphi_\mu(\theta) = e^{i\mu\theta}$ which is the characteristic function of the constant random variable μ.
By (6.1), we have that

$$\begin{aligned}\varphi_{\frac{S_n}{n}}(\theta) &= \mathbb{E}\left[e^{i\theta\frac{S_n}{n}}\right] = \prod_{k=1}^{n} \mathbb{E}\left[e^{i\theta\frac{X_k}{n}}\right] = \left(\varphi_{\frac{X}{n}}(\theta)\right)^n = \left(\varphi_X\left(\frac{\theta}{n}\right)\right)^n \\ &= \left(1 + i\mu\frac{\theta}{n} + o\left(\frac{\theta}{n}\right)\right)^n \quad \text{as } n \to \infty.\end{aligned}$$

Similarly to the proof of Central Limit Theorem, let $x_n = i\mu\theta + no\left(\frac{\theta}{n}\right)$, $n \in \mathbb{N}$. Then

$$\lim_{n\to\infty} \varphi_{G_n}(\theta) = \lim_{n\to\infty} (1 + x_n/n)^n = e^{i\mu\theta}.$$

Hence, for every $\theta \in \mathbb{R}$, the sequence $(\varphi_{\frac{S_n}{n}}(\theta) : n \in \mathbb{N})$ converges to $\varphi_\mu(\theta) = e^{i\mu\theta}$ which is the characteristic function of the constant random variable μ. It follows by Lévy's Convergence Theorem that $(\frac{S_n}{n} : n \in \mathbb{N})$ converges to μ in distribution, and, by Lemma (5.52) also in probability.

□

Typical work-flow when using characteristic functions

- Given: a problem involving sums of independent random variables;
- Convert to: a problem involving products of characteristic functions;
- Solve with a little algebra and analysis, and with a little help from Lévy's Convergence Theorem 6.5, to obtain the characteristic function of the answer;
- Wanted: the answer as a random variable.

Because of the one-to-one correspondence implied by Lévy's Inversion Formula, Theorem 6.4, if we can spot a random variable which has the given characteristic function then we know that random variable has the correct distribution.

Table of characteristic functions:

Distribution	pmf/pdf	Support	Char. fn.
Constant μ	1	$\{\mu\}$	$e^{i\theta\mu}$
Bernoulli	$p^k q^{1-k}$	$\{0,1\}$	$q + pe^{i\theta}$
Bin(n,p)	$\binom{n}{k} p^k q^{1-k}$	$\{0,1,\ldots,n\}$	$(q + pe^{i\theta})^n$
Poisson(λ)	$e^{-\lambda}\frac{\lambda^k}{k!}$	$\{0,1,2,\ldots\}$	$e^{\lambda(e^{i\theta}-1)}$
$N(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\mathbb{R}$	$e^{i\mu\theta - \frac{1}{2}\sigma^2\theta^2}$
$U[0,1]$	1	$[0,1]$	$\frac{e^{i\theta}-1}{i\theta}$

Distribution	pmf/pdf	Support	Char. fn.
$U[-1,1]$	$\frac{1}{2}$	$[-1,1]$	$\frac{\sin\theta}{\theta}$
$U[a,b]$	$\frac{1}{b-a}$	$[a,b]$	$\frac{1}{b-a}\frac{e^{i\theta b}-e^{i\theta a}}{i\theta}$
Exponential(λ)	$\lambda\ e^{-\lambda x}$	$[0,\infty)$	$\frac{\lambda}{\lambda-i\theta}$
Dbl. exponential	$\frac{1}{2}\ e^{-\lvert x\rvert}$	$\mathbb{R}$	$\frac{1}{1+\theta^2}$
Cauchy	$\frac{1}{\pi(1+x^2)}$	$\mathbb{R}$	$e^{-\lvert\theta\rvert}$
Triangular	$1-\lvert x\rvert$	$[-1,1]$	$2\left(\frac{1-\cos\theta}{\theta^2}\right)$
	$\frac{1-\cos x}{\pi x^2}$	$\mathbb{R}$	$(1-\lvert\theta\rvert)\,\mathbb{I}[[-1,1]](\theta)$

Note the way in which Double exponential and Cauchy distributions, and also Triangular and unnamed distribution in the last row, are paired via

$$\varphi(\theta) = \int_{\mathbb{R}} f(x)\ e^{i\theta x}\ dx, \qquad f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\theta x}\varphi(\theta)\ d\theta$$

(if $\int_{\mathbb{R}}|\varphi(\theta)|\ d\theta < \infty$ and $f(x)$ is a probability density function).

7: Conditional Expectation

Envy, n.:
Wishing you'd been born with an unfair advantage, instead of having to try and acquire one.

QOTD:
All I want is more than my fair share.

Conditional expectation

Table of contents of this section

7 Conditional Expectation
- Conditional expectation
- Properties of conditional expectation

Conditional expectation

Suppose that X and Y are random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with X taking distinct values $x_1, x_2, \ldots x_n$ and Y taking distinct values $y_1, y_2, \ldots y_m$.
Then we know, from First Year Probability, that

$$\mathbb{P}\left[X = x_i | Y = y_j\right] = \frac{\mathbb{P}\left[X = x_i, Y = y_j\right]}{\mathbb{P}\left[Y = y_j\right]}, \qquad \text{for } 1 \leq i \leq n, 1 \leq j \leq m,$$

while

$$\mathbb{E}\left[X | Y = y_j\right] = \sum_{i=1}^{n} x_i \ \mathbb{P}\left[X = x_i | Y = y_j\right], \qquad \text{for } 1 \leq j \leq m.$$

We shall now look more closely at the random variable $\mathbb{E}[X|Y]$ called the conditional expectation of X given Y.
Let $Z = \mathbb{E}[X|Y]$ and let $z_j = \mathbb{E}\left[X|Y = y_j\right]$ for $1 \leq j \leq m$. Then,

$$Z = \sum_{j=1}^{m} z_j \mathbb{I}\left[Y = y_j\right].$$

Properties of the conditional expectation:
It follows that Z is $\sigma(Y)$-measurable, since Z is determined by the value of Y. This is a very important property of the random variable $Z = \mathbb{E}[X|Y]$.

Furthermore, fix j with $1 \leq j \leq m$. Then Z takes the constant value z_j on the set $\{Y = y_j\}$, so

$$\begin{aligned} \mathbb{E}\left[Z; Y = y_j\right] &= z_j \ \mathbb{P}\left[Y = y_j\right] = \mathbb{E}\left[X|Y = y_j\right] \ \mathbb{P}\left[Y = y_j\right] \\ &= \sum_{i=1}^{n} x_i \ \mathbb{P}\left[X = x_i | Y = y_j\right] \ \mathbb{P}\left[Y = y_j\right] \\ &= \sum_{i=1}^{n} x_i \ \mathbb{P}\left[X = x_i, Y = y_j\right] = \mathbb{E}\left[X; Y = y_j\right]. \end{aligned}$$

Finally, consider another random variable $\tilde{Y}$ with $\sigma(\tilde{Y}) = \sigma(Y)$. Then

$$\mathbb{E}[X|Y] = \mathbb{E}\left[X|\tilde{Y}\right],$$

since the (finite) σ-algebra $\sigma(Y)$ is generated by the family of events $\{[Y = y_j], 1 \leq j \leq m\}$. Indeed any event $G \in \sigma(Y)$ can be expressed *via* $\mathbb{I}[G] = \sum_{j \in \tilde{G}} \mathbb{I}\left[Y = y_j\right]$ for some subset $\tilde{G}$ of $\{1, \ldots, m\}$. Thus

$$\begin{aligned} \mathbb{E}[Z; G] &= \mathbb{E}\left[Z \sum_{j \in \tilde{G}} \mathbb{I}\left[Y = y_j\right]\right] = \sum_{j \in \tilde{G}} \mathbb{E}\left[Z; Y = y_j\right] \\ &= \sum_{j \in \tilde{G}} \mathbb{E}\left[X; Y = y_j\right] = \mathbb{E}\left[X \sum_{j \in \tilde{G}} \mathbb{I}\left[Y = y_j\right]\right] = \mathbb{E}[X; G]. \end{aligned}$$

These measurability properties of $Z = \mathbb{E}[X|Y]$ point the way to a more general definition of conditional expectation.

Let $(\Omega, \mathcal{F})$ be a measurable space. Recall that $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$ (we sometimes write, $\mathcal{G} \leq \mathcal{F}$) if $\mathcal{G}$ is a σ-algebra on Ω and $\mathcal{G} \subseteq \mathcal{F}$.

Theorem 7.1 (conditional expectation after Kolmogorov, 1933)

Suppose $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Then there exists a random variable Z on $(\Omega, \mathcal{F}, \mathbb{P})$ such that

(i) *Z is $\mathcal{G}$-measurable*

(ii) $\mathbb{E}[|Z|] < \infty$

(iii) *$\int_G Z \, d\mathbb{P} = \int_G X \, d\mathbb{P}$, for every $G \in \mathcal{G}$, that is $\mathbb{E}[Z; G] = \mathbb{E}[X; G]$ for every $G \in \mathcal{G}$.*

Moreover, if $\tilde{Z}$ is another random variable with these properties then

$$Z = \tilde{Z} \quad \text{a.s., that is } \mathbb{P}\left[Z = \tilde{Z}\right] = 1 .$$

Definition 7.2 (Kolmogorov's definition of conditional expectation)

If Z has properties (i)-(iii) of Theorem 7.1 above then it is called a version of the conditional expectation $\mathbb{E}[X|\mathcal{G}]$ of X given $\mathcal{G}$, and we write $Z = \mathbb{E}[X|\mathcal{G}]$ a.s.

Notice that Theorem 7.1 makes it clear, $\mathbb{E}[X|\mathcal{G}]$ is undefined unless $\mathbb{E}[|X|] < \infty$.
What does this definition mean?
An experiment has been performed. Suppose you don't know exactly which outcome ω from the sample set Ω has occurred, but you know the value of $Y(\omega)$ for every $\mathcal{G}$-measurable random variable Y. Then $Z = \mathbb{E}[X|\mathcal{G}]$ is the expected value of X given this information.
Notice, if $\mathcal{G}$ is the trivial σ-algebra $\{\emptyset, \Omega\}$ (i.e. contains no information) then $\mathbb{E}[X|\mathcal{G}]$ must be constant, and in fact $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$.

Remark 7.3

We often use the abbreviated notation $\mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)]$, $\mathbb{E}[X|Y_1, Y_2, \ldots] = \mathbb{E}[X|\sigma(Y_1, Y_2, \ldots)]$.

Example 32 (Binomial and Poisson distributions)

Suppose the random variables X and N are such that N has the Poisson(λ) distribution, and X has the Binomial(N, p) distribution. Then

$$\mathbb{P}[X = k | N = n] \quad = \quad \binom{n}{k} p^k (1-p)^{n-k}, \qquad \text{whenever } 0 \leq k \leq n,$$

and so $\mathbb{E}[X|N = n] = np$. It follows that $\mathbb{E}[X|N] = Np$. Since $\mathbb{E}[Np] = p\lambda < \infty$ and Np is $\sigma(N)$-measurable, it follows from $\mathbb{E}[Np; N = n] = np\,\mathbb{P}[N = n] = \mathbb{E}[X; N = n]$ that Np is a version of the conditional expectation $\mathbb{E}[X|\sigma(N)]$.

Existence of conditional expectation:

We indicate how the existence part of Theorem 7.1 arises from a famous theorem in measure theory. Consider first the case of a *finite* σ-algebra.
Suppose that $\mathcal{G}$ is a σ-algebra generated by a finite partition $\{A_1, A_2, \ldots, A_n\}$ of Ω such that $\mathbb{P}[A_i] > 0$ for $1 \leq i \leq n$.
It is then the case that every $\mathcal{G}$-measurable random variable Y takes constant values on the sets A_i (for $1 \leq i \leq n$) and is of the form

$$Y = \sum_{i=1}^{n} \alpha_i \; \mathbb{I}[A_i] .$$

Similarly, every event $G \in \mathcal{G}$ is a union of some of the A_i.
If Z is a version of the conditional expectation $\mathbb{E}[X|\mathcal{G}]$, then it must be the case that $\mathbb{E}[Z; G] = \mathbb{E}[X; G]$ for every $G \in \mathcal{G}$.

We can deduce, for each disjoint union $G = A_{j_1} \cup \ldots \cup A_{j_r} \in \mathcal{G}$,

$$\begin{aligned}\mathbb{E}[Z;G] &= \mathbb{E}\left[\sum_{i=1}^n \alpha_i \ \mathbb{I}[A_i];G\right] = \sum_{i=1}^n \alpha_i \ \mathbb{E}[\mathbb{I}[A_i];G] \\ &= \sum_{i=1}^n \sum_{s=1}^r \alpha_i \ \mathbb{E}\left[\mathbb{I}[A_i]\,\mathbb{I}\left[A_{j_s}\right]\right] = \sum_{s=1}^r \alpha_{j_s} \ \mathbb{E}\left[\mathbb{I}\left[A_{j_s}\right]\right] \\ &= \sum_{s=1}^r \alpha_{j_s} \ \mathbb{P}\left[A_{j_s}\right].\end{aligned}$$

On the other hand, $\mathbb{E}[X;G] = \sum_{s=1}^r \mathbb{E}\left[X;A_{j_s}\right]$. Hence, by putting $\alpha_i = \frac{\mathbb{E}[X;A_i]}{\mathbb{P}[A_i]}$, we can ensure that $\mathbb{E}[Z;G] = \mathbb{E}[X;G]$ for every $G \in \mathcal{G}$. Therefore we have shown existence of the conditional expectation $\mathbb{E}[X|\mathcal{G}]$ when the σ-algebra $\mathcal{G}$ is finite.

To deal with the existence of $\mathbb{E}[X|\mathcal{G}]$ for a general σ-algebra $\mathcal{G}$, we need the following form of the Radon-Nikodym theorem.

Theorem 7.4

Let (S,Σ) be a measurable space. If μ and λ are σ-finite measures on (S,Σ) such that $\mu(F) = 0$ implies $\lambda(F) = 0$ for every $F \in \Sigma$, then there exists $f \in (m\Sigma)^+$ such that $\lambda(F) = \int_F f\, d\mu$ for every $F \in \Sigma$.

We say that f is the *density* of the measure λ with respect to the measure μ. Note that this theorem actually holds for *signed* σ-finite measures λ (as might arise, for example, in electrostatics).

Now we can prove the existence of the conditional expectation in the general case. Without loss of generality, suppose that $X \geq 0$ (otherwise split X as $X = X^+ - X^-$). Define a measure μ on $(\Omega, \mathcal{G})$ by

$$\mu(A) \quad = \quad \mathbb{E}[X;A], \qquad \text{for all } A \in \mathcal{G},$$

(μ is indeed a measure: see Exercise sheet 9.9.) Since $\mathbb{P}[A] = 0$ implies that $\mathbb{E}[X;A] = \int_A X\, d\mathbb{P} = 0$, we have that $\mathbb{P}[A] = 0$ implies $\mu(A) = 0$ for every $A \in \mathcal{G}$. Hence by the Radon-Nikodym theorem there exists a $\mathcal{G}$-measurable random variable $\mathbb{E}[X|\mathcal{G}]$ such that

$$\mu(A) \quad = \quad \int_A \mathbb{E}[X|\mathcal{G}] \ d\mathbb{P} \qquad \text{for all } A \in \mathcal{G}.$$

Uniqueness of conditional expectation (I)

We do not need to quote theorems from measure theory to establish (almost-sure) uniqueness. It suffices to employ an argument which occurs in several places in this theory (for example, Lemma 5.10). Let Z_1 and Z_2 be two versions of the conditional expectation of X given $\mathcal{G}$. Then, for every $\varepsilon > 0$, the event $A_\varepsilon = \{Z_1 - Z_2 > \varepsilon\} \in \mathcal{G}$ (because Z_1 and Z_2 are $\mathcal{G}$-measurable).
Therefore, we have that

$$\mathbb{E}[X;A_\varepsilon] \quad = \quad \mathbb{E}[Z_1;A_\varepsilon] \quad = \quad \mathbb{E}[Z_2;A_\varepsilon]\,,$$

which implies that $0 = \mathbb{E}[(Z_1 - Z_2;A_\varepsilon] \geq \varepsilon \ \mathbb{P}[A_\varepsilon]$, because $Z_1 - Z_2 > \varepsilon$ on the $\mathcal{G}$-measurable event A_ε.

Uniqueness of conditional expectation (II)

Thus, $\mathbb{P}[A_\varepsilon] = \mathbb{P}[Z_1 - Z_2 > \varepsilon] = 0$ for every $\varepsilon > 0$. In the same way we conclude that $\mathbb{P}[Z_2 - Z_1 > \varepsilon] = 0$ for every $\varepsilon > 0$. Therefore, for any $\varepsilon > 0$, $\mathbb{P}[|Z_1 - Z_2| > \varepsilon] = 0$. It follows that

$$\mathbb{P}[Z_1 = Z_2] = \mathbb{P}\left[\bigcap_{n=1}^{\infty}\left[|Z_1 - Z_2| \leq \tfrac{1}{n}\right]\right] = \lim_{n\to\infty} \mathbb{P}\left[|Z_1 - Z_2| \leq \tfrac{1}{n}\right] = 1\,,$$

because $\mathbb{P}\left[|Z_1 - Z_2| \leq \frac{1}{n}\right] = 1$ for every positive integer n (since $\mathbb{P}[|Z_1 - Z_2| \leq \varepsilon] = 1$ for every $\varepsilon > 0$). Hence $\mathbb{P}[Z_1 = Z_2] = 1$. Thus Definition 7.2 “defines conditional expectation almost surely.”

Example 33 (Conditional expectation for probability densities)

If X and Y are random variables with joint probability density $f_{X,Y}(x,y)$ and $\mathbb{E}[X] < \infty$, then

$$\mathbb{E}[X|\sigma(Y)] \quad = \quad \int_{\mathbb{R}} x\, f_{X|Y}(x|Y)\,dx\,,$$

where

$$f_{X|Y}(x|Y=y) \quad = \quad \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} & \text{if } f_Y(y) \neq 0 \\ 0 & \text{otherwise,} \end{cases}$$

(where $f_{X|Y}(x|Y=y)$ is the value of $f_{X|Y}(x|Y)$ when $Y = y$) and

$$f_Y(y) \quad = \quad \int_{\mathbb{R}} f_{X,Y}(x,y)\,dx\,.$$

Properties of conditional expectation (I)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. All X’s in the following list of properties of conditional expectation belong to $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Also, $\mathcal{G}$ and $\mathcal{H}$ are sub-σ-algebras of $\mathcal{F}$.

(a) If Y is a version of $\mathbb{E}[X|\mathcal{G}]$ then $\mathbb{E}[Y] = \mathbb{E}[X]$.

(b) If X is $\mathcal{G}$-measurable, then $\mathbb{E}[X|\mathcal{G}] = X$ a.s.

(c) (Linearity) $\mathbb{E}[a_1 X_1 + a_2 X_2|\mathcal{G}] = a_1\ \mathbb{E}[X_1|\mathcal{G}] + a_2\ \mathbb{E}[X_2|\mathcal{G}]$ a.s. (If Y_1 is a version of $\mathbb{E}[X_1|\mathcal{G}]$ and Y_2 is a version of $\mathbb{E}[X_2|\mathcal{G}]$, then $a_1 Y_1 + a_2 Y_2$ is a version of $\mathbb{E}[a_1 X_1 + a_2 X_2|\mathcal{G}]$).

(d) (Positivity) If $X \geq 0$ then $\mathbb{E}[X|\mathcal{G}] \geq 0$.

(e) (cMON) If $0 \leq X_n \uparrow X$ (nb: $\mathbb{E}[X] < \infty$) then $\mathbb{E}[X_n|\mathcal{G}] \uparrow \mathbb{E}[X|\mathcal{G}]$ a.s.

(f) (cFatou) If all $X_n \geq 0$ then $\mathbb{E}[\liminf_n X_n|\mathcal{G}] \leq \liminf_n \mathbb{E}[X_n|\mathcal{G}]$ a.s., and, similarly, if $X_n \leq Y$ for some Y with $\mathbb{E}[|Y|] < \infty$, then $\mathbb{E}[\limsup_n X_n|\mathcal{G}] \geq \limsup_n \mathbb{E}[X_n|\mathcal{G}]$ a.s.

(g) (cDOM) If all $|X_n| \leq Y$ a.s., where $\mathbb{E}[Y] < \infty$, and if $X_n \to X$ a.s., then $\mathbb{E}[X_n|\mathcal{G}] \to \mathbb{E}[X|\mathcal{G}]$ a.s.

Properties of conditional expectation (II)

(h) (cJensen) If $\phi : \mathbb{R} \to \mathbb{R}$ is convex and $\mathbb{E}[|\phi(X)|] < \infty$, then

$$\phi\Big(\mathbb{E}[X|\mathcal{G}]\Big) \leq \mathbb{E}[\phi(X)|\mathcal{G}] \quad \text{a.s.}$$

Corollary: $\|\,\mathbb{E}[X|\mathcal{G}]\,\|_p \leq \|X\|_p$ for $p \geq 1$.

(i) (Tower property) If $\mathcal{H}$ is a sub-σ-algebra of $\mathcal{G}$, then

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]\,|\mathcal{H}] = \mathbb{E}[X|\mathcal{H}] \quad \text{a.s.}$$

Properties of conditional expectation (III)

(j) ("Taking out what is known") If Z is $\mathcal{G}$-measurable and bounded then

$$(*) \qquad \mathbb{E}[ZX|\mathcal{G}] = Z\ \mathbb{E}[X|\mathcal{G}] \quad \text{a.s.}$$

If $p>1$, $\frac{1}{p}+\frac{1}{q}=1$, $X \in \mathcal{L}^p(\Omega,\mathcal{F},\mathbb{P})$ and $Z \in \mathcal{L}^q(\Omega,\mathcal{G},\mathbb{P})$, then $(*)$ holds. If $X \in (m\mathcal{F})^+$, $Z \in (m\mathcal{G})^+$, $\mathbb{E}[X]<\infty$ and $\mathbb{E}[XZ]<\infty$, then $(*)$ holds.

(k) (Role of independence) If $\mathcal{H}$ is independent of $\sigma(\sigma(X),\mathcal{G})$, then

$$\mathbb{E}[X|\mathcal{G},\mathcal{H}] = \mathbb{E}[X|\mathcal{G}] \quad \text{a.s.}$$

In particular, if X is independent of $\mathcal{H}$, then $\mathbb{E}[X|\mathcal{H}] = \mathbb{E}[X]$ a.s.

Proofs of the properties of conditional expectation:

(a) If Y is a version of $\mathbb{E}[X|\mathcal{G}]$ then $\mathbb{E}[Y;G]=\mathbb{E}[X;G]$ for every $G \in \mathcal{G}$. Since $\Omega \in \mathcal{G}$ and $\mathbb{I}[\Omega](\omega)=1$ for every $\omega \in \Omega$,

$$\mathbb{E}[Y] = \mathbb{E}[Y;\Omega] = \mathbb{E}[X;\Omega] = \mathbb{E}[X].$$

(b) If X is $\mathcal{G}$-measurable, then it satisfies the conditions of Definition 7.2. Thus, it is a version of the conditional expectation $\mathbb{E}[X|\mathcal{G}]$, that is $\mathbb{E}[X|\mathcal{G}]=X$ a.s.

(c) (Linearity) Since Y_1 is a version of $\mathbb{E}[X_1|\mathcal{G}]$ and Y_2 is a version of $\mathbb{E}[X_2|\mathcal{G}]$, then Y_1 and Y_2 are $\mathcal{G}$-measurable, $Y_1, Y_2 \in \mathcal{L}^1(\Omega,\mathcal{G},\mathbb{P})$ and $\mathbb{E}[Y_i;G]=\mathbb{E}[X_i;G]$ for every $G \in \mathcal{G}$. Hence $a_1Y_1+a_2Y_2$ is $\mathcal{G}$-measurable; the triangle inequality shows that $a_1Y_1+a_2Y_2 \in \mathcal{L}^1(\Omega,\mathcal{G},\mathbb{P})$, and linearity of expectation shows that $\mathbb{E}[(a_1Y_1+a_2Y_2);G]=\mathbb{E}[(a_1X_1+a_2X_2);G]$ for every $G \in \mathcal{G}$. Hence, by Definition 7.2, $a_1Y_1+a_2Y_2$ is a version of $\mathbb{E}[a_1X_1+a_2X_2|\mathcal{G}]$.

Proofs of properties . . . (II)

(d) (Positivity) Suppose that $X \geq 0$. Let Y be a version of $\mathbb{E}[X|\mathcal{G}]$ and suppose that $\mathbb{P}[Y<0]>0$. Then, by the monotonicity of measures

$$0 < \mathbb{P}[Y<0] = \mathbb{P}\left[\bigcup_{n\geq 1}\{Y<-\tfrac{1}{n}\}\right] = \lim_{n\to\infty}\mathbb{P}\left[Y<-\tfrac{1}{n}\right].$$

It follows that there is a postive integer n_0 such that $\mathbb{P}\left[Y<-\frac{1}{n_0}\right]>0$. Since $[Y<-\frac{1}{n_0}] \in \mathcal{G}$, we have that

$$0 \leq \mathbb{E}\left[X; Y<-\tfrac{1}{n_0}\right] = \mathbb{E}\left[Y; Y<-\tfrac{1}{n_0}\right] < -\tfrac{1}{n_0}\,\mathbb{P}\left[Y<-\tfrac{1}{n_0}\right] \leq 0.$$

But this is a contradiction because $X \geq 0$ implies that $\mathbb{E}[X;F] \geq 0$ for every $F \in \mathcal{F}$.
Thus, $\mathbb{P}[Y<0]=0$ which implies that $Y \geq 0$ a.s.

Proofs of properties . . . (III)

(e) For every positive integer n, let Y_n be a version of the conditional expectation $\mathbb{E}[X_n|\mathcal{G}]$. If $0 \leq X_n \uparrow X$ then by (d) $0 \leq Y_n \uparrow Y = \sup Y_n$. We need to prove Y is a version of $\mathbb{E}[X|\mathcal{G}]$.

Fix $G \in \mathcal{G}$. Then, $0 \leq X_n \uparrow X$ implies $0 \leq X_n\mathbb{I}[G] \uparrow X\mathbb{I}[G]$, and therefore by (MON) $\mathbb{E}[X_n;G] \uparrow \mathbb{E}[X;G]$. Similarly, by (MON), $\mathbb{E}[Y_n;G] \uparrow \mathbb{E}[Y;G]$. Since Y_n is a version of $\mathbb{E}[X_n|\mathcal{G}]$, it must be that $\mathbb{E}[X_n;G]=\mathbb{E}[Y_n;G]$ for every $G \in \mathcal{G}$. Thus, for every $G \in \mathcal{G}$,

$$\mathbb{E}[Y;G] = \lim_{n\to\infty}\mathbb{E}[Y_n;G] = \lim_{n\to\infty}\mathbb{E}[X_n;G] = \mathbb{E}[X;G].$$

Proofs of properties ... (IV)

In addition,

$$\mathbb{E}[|Y|] = \mathbb{E}[Y] = \mathbb{E}[Y;\Omega] = \mathbb{E}[X;\Omega] = \mathbb{E}[X] = \mathbb{E}[|X|] < \infty\,,$$

and, since Y_n is $\mathcal{G}$-measurable for every positive integer n, it follows from Lemma 3.11 that Y is $\mathcal{G}$-measurable.

Thus, Y satisfies all the conditions of Definition 7.2, and so Y is a version of the conditional expectation $\mathbb{E}[X|\mathcal{G}]$.

Proofs of properties ... (V)

(f) Let $L_n = \inf_{m\geq n} X_m$ for positive integers n. Then

$$L_n \uparrow L \quad = \quad \lim_{n\to\infty} L_n = \liminf_n X_n\,.$$

Since $L_n \leq X_m$ for every $m \geq n$, it follows that $\mathbb{E}[L_n] \leq \mathbb{E}[X_n] < \infty$ for every positive integer n. Thus, $L_n \in \mathcal{L}^1$ for every $n \in \mathbb{N}$, and by property (d), $E(L_n|\mathcal{G}) \leq E(X_m|\mathcal{G})$ a.s. for every $m \geq n$.
It follows that $\mathbb{E}[L_n|\mathcal{G}] \leq \inf_{m\geq n}\mathbb{E}[X_m|\mathcal{G}]$ a.s.

By (cMON), $E(L_n|\mathcal{G}) \uparrow E(L|\mathcal{G})$ a.s. Hence, by letting $n\to\infty$,

$$\mathbb{E}[L|\mathcal{G}] = \lim_{n\to\infty}\mathbb{E}[L_n|\mathcal{G}] \leq \lim_{n\to\infty}\inf_{m\geq n}\mathbb{E}[X_m|\mathcal{G}] = \liminf_n \mathbb{E}[X_n|\mathcal{G}] \quad \text{a.s.}$$

For the second part, consider
$L_n = \inf_{m\geq n}(Y - X_m) = L_n = Y - \sup_{m\geq n} X_m$ and note
$\mathbb{E}[L_n] \leq \mathbb{E}[Y] - \mathbb{E}[X_n] \leq \mathbb{E}[|Y|] + \mathbb{E}[|X_n|] < \infty$.

Proofs of properties ... (VI)

(g) If $X_n \to X$ a.s. then $|X_n - X| \to 0$ a.s. Since $|X_n - X| \leq 2Y$, it follows that $\mathbb{E}[|X_n - X|] \leq 2\,\mathbb{E}[Y] < \infty$. Thus, $\mathbb{E}[|X_n - X||\mathcal{G}]$ is well-defined. By (cFatou),

$$0 \leq \limsup_n \mathbb{E}[|X_n - X||\mathcal{G}] \leq \mathbb{E}\left[\limsup_n |X_n - X||\mathcal{G}\right] = 0 \quad \text{a.s.}$$

It follows that $\lim_{n\to\infty}\mathbb{E}[|X_n - X||\mathcal{G}] = 0$ a.s. Finally, since[3]

$$|\,\mathbb{E}[X_n - X|\mathcal{G}]\,| \leq \mathbb{E}[|X_n - X||\mathcal{G}]\,,$$

we deduce that $\mathbb{E}[X_n|\mathcal{G}] \to \mathbb{E}[X|\mathcal{G}]$ a.s.

[3] Actually this follows most easily from (cJensen) below, since $|x|$ is a convex function!

Proofs of properties ... (VII)

(h) Let ϕ be a convex function. Then, as in the proof of Theorem 5.15, there exist real sequences $(a_n, n\in\mathbb{N})$ and $(b_n, n\in\mathbb{N})$ such that

$$\phi(x) = \sup_n (a_n x + b_n)\,.$$

Hence, for fixed $n \in \mathbb{N}$, $\phi(X) \geq a_n X + b_n$, which, by (d), implies that

$$\mathbb{E}[\phi(X)|\mathcal{G}] \geq a_n\,\mathbb{E}[X|\mathcal{G}] + b_n \quad \text{a.s.}$$

Therefore,

$$\mathbb{E}[\phi(X)|\mathcal{G}] \;\geq\; \sup_n\left(a_n\,\mathbb{E}[X|\mathcal{G}] + b_n\right) \;=\; \phi\left(\mathbb{E}[X|\mathcal{G}]\right) \quad \text{a.s.}$$

Proofs of properties . . . (VIII)

Proof of Corollary: Let $p \geq 1$ and let $\phi(x) = |x|^p$. Since ϕ is a convex function, by (cJensen),

$$|\,\mathbb{E}[X|\mathcal{G}]\,|^p \leq \mathbb{E}[|X|^p|\mathcal{G}] \quad \text{a.s.}$$

Taking the expectation of both sides, it follows by (a) that

$$\mathbb{E}[|\,\mathbb{E}[X|\mathcal{G}]\,|^p] \leq \mathbb{E}[|X|^p]\,,$$

that is

$$\|\,\mathbb{E}[X|\mathcal{G}]\,\|_p = \left(\mathbb{E}[|\,\mathbb{E}[X|\mathcal{G}]\,|^p]\right)^{\frac{1}{p}} \leq \left(\mathbb{E}[|X|^p]\right)^{\frac{1}{p}} = \|X\|_p\,.$$

Proofs of properties . . . (IX)

(i) Let $\mathcal{H} \subseteq \mathcal{G}$. The "Tower property" states that Y is a version of $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]\,|\mathcal{H}]$ if and only if Y is a version of $\mathbb{E}[X|\mathcal{H}]$.
By the definition of conditional expectation, for $A \in \mathcal{G}$

$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]\,;A] = \mathbb{E}[X;A] \tag{7.1}$$

Suppose that Y is a version of $\mathbb{E}[X|\mathcal{H}]$. Then, Y is $\mathcal{H}$-measurable, $\mathbb{E}[|Y|] < \infty$ and $\mathbb{E}[Y;A] = \mathbb{E}[X;A]$ for every $A \in \mathcal{H}$. It follows from (7.1) and because $\mathcal{H} \subseteq \mathcal{G}$ that

$$\mathbb{E}[Y;A] = \mathbb{E}[X;A] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\,;A] \quad A \in \mathcal{H}\,.$$

Thus, by Definition 7.2, $Y = \mathbb{E}[X|\mathcal{H}]$ is a version of $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]\,|\mathcal{H}]$.

Proofs of properties . . . (X)

(j) We discuss the case $X \geq 0$ and $Z \geq 0$. Other cases can be deduced from this one by linearity (consider $X = X^+ - X^-$ and $Z = Z^+ - Z^-$).
Let $Y = \mathbb{E}[X|\mathcal{G}]$ a.s. We need to show that YZ is a version of the conditional expectation $\mathbb{E}[ZX|\mathcal{G}]$. We will prove the statement first for $\mathcal{G}$-measurable indicator random variables, then for simple random variables and finally for any $\mathcal{G}$-measurable non-negative random variable Z.
Let $Z = \mathbb{I}[A]$ for some $A \in \mathcal{G}$. Then, by the definition of the conditional expectation Y, for any $G \in \mathcal{G}$,

$$\mathbb{E}[ZY;G] = \mathbb{E}[Y;A\cap G] = \mathbb{E}[X;A\cap G] = \mathbb{E}[XZ;G]\,.$$

Proofs of properties . . . (XI)

Suppose now that Z is a simple non-negative $\mathcal{G}$-measurable random variable, that is $Z = \sum_{i=1}^n \alpha_i \mathbb{I}[A_i]$ for some $\alpha_i \geq 0$, $i = 1, 2, \ldots, n$ and some $\mathcal{G}$-measurable sets A_i, $i = 1, 2, \ldots, n$.
Then, by linearity of conditional expectation,

$$\mathbb{E}[ZY;G] = \mathbb{E}[XZ;G] \qquad \text{for any } G \in \mathcal{G}. \tag{7.2}$$

Note that the simple random variable Z is automatically bounded.

Proofs of properties . . . (XII)

If Z is a non-negative $\mathcal{G}$-measurable random variable, then there exists a sequence of simple non-negative $\mathcal{G}$-measurable random variables that increases to Z. Thus, by (MON), (7.2) holds.
In order to prove that YZ is a version of $\mathbb{E}[ZX|\mathcal{G}]$ we also need to show that YZ is $\mathcal{G}$-measurable, that $\mathbb{E}[|YZ|] < \infty$ and that $\mathbb{E}[|XZ|] < \infty$. By the definition, YZ is $\mathcal{G}$-measurable. In addition, since $X \geq 0$ and $Z \geq 0$, by (d), $Y \geq 0$ a.s., and by (7.2), $\mathbb{E}[|XZ|] = \mathbb{E}[XZ] = \mathbb{E}[YZ] = \mathbb{E}[|YZ|]$.
Finally, under each of the conditions given in the statement, $\mathbb{E}[XZ] < \infty$. More precisely, if Z is bounded, then $\mathbb{E}[X] < \infty$ implies $\mathbb{E}[XZ] < \infty$. If $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$ and $Z \in \mathcal{L}^q(\Omega, \mathcal{G}, \mathbb{P})$, then by Hölder's inequality,

$$\mathbb{E}[|XZ|] \leq \|X\|_p \|Z\|_q < \infty\,.$$

Proofs of properties . . . (XIII)

(k) Without loss of generality, assume $X \geq 0$. For any $G \in \mathcal{G}$ and $H \in \mathcal{H}$, $X\mathbb{I}[G]$ and H are independent, thus

$$\begin{aligned}\mathbb{E}[X; G \cap H] &= \mathbb{E}[X\mathbb{I}[G]\mathbb{I}[H]] \\ &= \mathbb{E}[X\mathbb{I}[G]]\,\mathbb{E}[\mathbb{I}[H]] = \mathbb{E}[X\mathbb{I}[G]]\,\mathbb{P}[H] = \mathbb{E}[X; G]\,\mathbb{P}[H]\,.\end{aligned}$$

Let $Y = \mathbb{E}[X|\mathcal{G}]$ a.s.. Then Y is $\mathcal{G}$-measurable which implies that, for any $G \in \mathcal{G}$, $Y\mathbb{I}[G]$ is independent of $\mathcal{H}$ and that for any $G \in \mathcal{G}$ and $H \in \mathcal{H}$, $\mathbb{E}[Y; G \cap H] = \mathbb{E}[Y\mathbb{I}[G]\mathbb{I}[H]] = \mathbb{E}[Y\mathbb{I}[G]]\,\mathbb{E}[\mathbb{I}[H]] = \mathbb{E}[Y; G]\,\mathbb{P}[H]$.
It follows that $\mathbb{E}[X; G \cap H] = \mathbb{E}[Y; G \cap H]$ for any $G \in \mathcal{G}$ and $H \in \mathcal{H}$.

Proofs of properties . . . (XIV)

Therefore, the measures $F \mapsto \mathbb{E}[X; F]$ and $F \mapsto \mathbb{E}[Y; F]$ on $\sigma(\mathcal{G}, \mathcal{H})$ have the same finite total mass and agree on the Π-system $\{G \cap H : G \in \mathcal{G}, H \in \mathcal{H}\}$. It follows by the Uniqueness Lemma 2.13 that they then agree everywhere on $\sigma(\mathcal{G}, \mathcal{H})$. In addition, $\mathbb{E}[Y] < \infty$ and Y is $\mathcal{G}$-measurable by the definition of Y, and because $\mathcal{G} \subseteq \sigma(\mathcal{G}, \mathcal{H})$, Y is also $\sigma(\mathcal{G}, \mathcal{H})$-measurable.
Thus, the statement is proved.
For the final part take $\mathcal{G} = \{\emptyset, \Omega\}$.

8: Martingales

Ecclesiastes 9:11:
I returned, and saw under the sun, that the race is not to the swift,
nor the battle to the strong, neither yet bread to the wise, nor yet riches to men of understanding,
nor yet favour to men of skill;
but time and chance happeneth to them all.

Martingales

Table of contents of this section

8 Martingales
- Martingales
- Stopping times
- Optional Stopping Theorem
- Martingale Convergence Theorem

Martingales are a potent way of calculating chance and expectation.

Martingales

Definition 8.1 (filtration of σ-algebras)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A filtration of σ-algebras $\{\mathcal{F}_n : n \geq 0\}$ is an increasing family of sub-σ-algebras of $\mathcal{F}$

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \mathcal{F}_{n+1} \subseteq \ldots \mathcal{F}\,.$$

We also define $\mathcal{F}_\infty$ by

$$\mathcal{F}_\infty \quad = \quad \sigma(\bigcup_{n=0}^{\infty} \mathcal{F}_n) \subseteq \mathcal{F}\,.$$

The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with the filtration $\{\mathcal{F}_n : n \geq 0\}$ is called a filtered space and is denoted by $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, \mathbb{P})$.

Intuitively, what we know about the choice of $\omega \in \Omega$ at time n is expressed by the σ-algebra $\mathcal{F}_n$, that is we know the value of $Z(\omega)$ for every $\mathcal{F}_n$-measurable random variable Z.

A random process is a family $(X_t : t \in T)$ of random variables indexed by some set T. If $T = \{0, 1, 2, \ldots\}$ the random process is a 'discrete-time' process. If $T = [0, \infty)$ then the process is a 'continuous-time' process.
Usually, $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$ for some process $(X_n : n \geq 0)$ and in that case, what we know about the choice of $\omega \in \Omega$ at time n is determined by the values of $X_i(\omega)$, $i = 0, 1, 2, \ldots n$.

Definition 8.2 (process adapted to a filtration)

A process $(K_n : n \geq 0)$ is said to be adapted to a filtration $\{\mathcal{F}_n : n \geq 0\}$ if for each $n \in \mathbb{N}$, K_n is $\mathcal{F}_n$-measurable.

If a process $(X_n : n \geq 0)$ is adapted to a filtration $\{\mathcal{F}_n : n \geq 0\}$ then the value $X_n(\omega)$, $\omega \in \Omega$, is known at time n.

Doob[4] introduced the fundamental notion of martingales:

Definition 8.3 (martingale, supermartingale, submartingale)

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, \mathbb{P})$ be a filtered space. A process $(M_n : n \geq 0)$ is called a martingale relative to $(\{\mathcal{F}_n\}, \mathbb{P})$ if

(i) $(M_n : n \geq 0)$ is adapted;

(ii) $\mathbb{E}[|M_n|] < \infty$ for all $n \geq 0$;

(iii) $\mathbb{E}[M_n | \mathcal{F}_{n-1}] = M_{n-1}$ almost surely for all $n \geq 1$.

A supermartingale is defined in the same way except that (iii) is replaced with $\mathbb{E}[M_n | \mathcal{F}_{n-1}] \leq M_{n-1}$ a.s. for all $n \geq 1$, and a submartingale is defined with (iii) replaced with $\mathbb{E}[M_n | \mathcal{F}_{n-1}] \geq M_{n-1}$ a.s. for all $n \geq 1$.

[4] An interview with Doob is to be found at http://www.dartmouth.edu/~chance/Doob/conversation.html

If $(M_n : n \geq 0)$ is a supermartingale then, by property (a) of conditional expectation, $\mathbb{E}[M_n] \leq \mathbb{E}[M_{n-1}]$ for every $n \geq 1$. If $(M_n : n \geq 0)$ is a submartingale then $\mathbb{E}[M_n] \geq \mathbb{E}[M_{n-1}]$ for every $n \geq 1$. Hence, a supermartingale 'decreases on average', and a submartingale 'increases on average'. If $(M_n : n \geq 0)$ is a martingale, then it is also a supermartingale and a submartingale and $\mathbb{E}[M_n] = \mathbb{E}[M_0]$ for every $n \geq 1$.

Note that we *must* specify that the process is adapted to the filtration $\{\mathcal{F}_n\}$ when defining a supermartingale or a submartingale. In the case of a martingale this follows automatically from the properties of a conditional expectation.

Coin-tossing and simple symmetric random walk

Suppose $X_1, X_2, \ldots$ are independent random variables, with $\mathbb{P}[X_i = \pm 1] = \frac{1}{2}$. Thus $X_n = 1$ models the outcome that the n^{th} of a sequence of fair coin tosses comes up heads.
If $\mathcal{F}_n = \sigma\{X_1, \ldots, X_n\}$ then $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots$ provides a filtration of σ-algebras.
The random process $Y = (Y_0, Y_1, Y_2, \ldots)$ defined by $Y_n = X_1 + \ldots + X_n$ determines an adapted process which is in fact a martingale. It would measure the evolution of your fortune if you gambled 1 unit of currency on heads at a casino at each coin-toss.
Suppose you had to pay 0.01 units per gamble; then your fortune at time n would be $Y_n - 0.01n$, forming a supermartingale.
The casino's fortune would be $-Y_n + 0.01n$, forming a submartingale.

Thackeray's martingale system (1854)

Bet 1 unit of currency on $X_1 = 1$. If you win, stop gambling.
Bet 2 units of currency on $X_2 = 1$. If you win, stop gambling.
Bet 4 units of currency on $X_3 = 1$. If you win, stop gambling.
...
Bet 2^{n-1} units of currency on $X_n = 1$. If you win, stop gambling.
...
Eventually, if allowed to play for ever, you will stop gambling and then will have fortune 1 unit of currency.

Example 34

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, \mathbb{P})$ be a filtered space and let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Define

$$M_n = \mathbb{E}[X|\mathcal{F}_n], \qquad n \geq 0.$$

Then $(M_n : n \geq 0)$ is a martingale.

Proof.

Let $n \geq 0$ be fixed. Then, M_n is $\mathcal{F}_n$-measurable and, by (cJensen),

$$\mathbb{E}[|M_n|] = \mathbb{E}[|\mathbb{E}[X|\mathcal{F}_n]|] \leq \mathbb{E}[\mathbb{E}[|X||\mathcal{F}_n]] = \mathbb{E}[|X|] < \infty.$$

Furthermore, by 'tower property',

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

□

Variation on the simple symmetric random walk example

Example 35

Let $(X_n : n \in \mathbb{N})$ be a sequence of independent random variables such that for every $n \in \mathbb{N}$, $\mathbb{E}[|X_n|] < \infty$ and $\mathbb{E}[X_n] = 0$. Define the process $(S_n : n \geq 0)$ by $S_0 = 0$, $S_n = X_1 + X_2 + \cdots + X_n$, for $n \in \mathbb{N}$. Define the sequence of σ-algebras $\{\mathcal{F}_n : n \geq 0\}$ by

$$\mathcal{F}_0 = \{\emptyset, \Omega\}, \qquad \mathcal{F}_n = \sigma(X_1, X_2, \ldots, X_n), \qquad n \geq 1.$$

Then $(S_n : n \geq 0)$ is a martingale relative to $\{\mathcal{F}_n\}$.

Warwick Statistics

Proof.

First note that, by the definition, $\{\mathcal{F}_n : n \geq 0\}$ is a filtration. Furthermore, S_n is $\mathcal{F}_n$-measurable for every $n \geq 0$. In addition, for fixed $n \geq 1$, by the triangle inequality,

$$\mathbb{E}[|S_n|] \leq \mathbb{E}[|X_1|] + \mathbb{E}[|X_2|] + \cdots + \mathbb{E}[|X_n|] < \infty.$$

Finally, for $n \geq 1$,

$$\begin{aligned}
\mathbb{E}[S_n|\mathcal{F}_{n-1}] &= \mathbb{E}[S_{n-1}|\mathcal{F}_{n-1}] + \mathbb{E}[X_n|\mathcal{F}_{n-1}] \quad \text{(by linearity)} \\
&= S_{n-1} + \mathbb{E}[X_n] \quad (\text{because } S_{n-1} \text{ is } \mathcal{F}_{n-1}\text{-measurable} \\
&\qquad \text{and } X_n \text{ is independent of } \mathcal{F}_{n-1}) \\
&= S_{n-1}. \qquad (\text{because } \mathbb{E}[X_n] = 0)
\end{aligned}$$

Hence, $(S_n : n \geq 0)$ is a martingale. □

Warwick Statistics

Stopping times

Definition 8.4 (stopping time)

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, \mathbb{P})$ be a filtered space. A map $T : \Omega \to \{0, 1, 2, \ldots, \infty\}$ is called a stopping time if

$$[T \leq n] = \{\omega : T(\omega) \leq n\} \in \mathcal{F}_n, \quad \text{for all } n \in \{0, 1, 2, \ldots, \infty\}.$$

Equivalently, if

$$[T = n] = \{\omega : T(\omega) = n\} \in \mathcal{F}_n, \quad \text{for all } n \in \{0, 1, 2, \ldots, \infty\}.$$

Note that T can be ∞.

- The time when Thackeray's martingale strategy stops playing: this *is* a stopping time.
- The time $n - 1$ when n is the first time that $X_n - X_{n-1} = 1$: this *is not* a stopping time.

Warwick Statistics

The equivalence of statements in the previous definition follows from equalities holding for $n \geq 0$:

$$\begin{aligned} [T \leq n] &= \bigcup_{k=0}^{n} [T = k], \\ [T = n] &= [T \leq n] \setminus [T \leq n-1], \end{aligned}$$

valid for any random variable T taking only non-negative integer values. Note also that if T is a stopping time then $[T < n] = [T \leq n-1] \in \mathcal{F}_n$, $n \in \mathbb{N}$.

Suppose that T is the time at which a player decides to stop the game. The player bases his decision whether to stop immediately after the n^{th} game only on what has happened in the game up to (and including) time n. Thus, $[T = n] \in \mathcal{F}_n$.

Examples of stopping times:

1) The constant time n is a stopping time.
2) If S and T are stopping times then so are $S \wedge T = \min\{S, T\}$ and $S \vee T = \max\{S, T\}$.
3) If $\{T_n : n \geq 0\}$ are stopping times, then so are $\sup_n T_n$ and $\inf_n T_n$.

Proof for case of sup.

$$[\sup_m T_m \leq n] = \bigcap_{m=1}^{\infty} [T_m \leq n] \in \mathcal{F}_n, \qquad n \in \mathbb{N}.$$

□

4) If $(X_n : n \geq 0)$ is adapted and $A \in \mathcal{B}$, then $T = \inf\{k \geq 0 : X_k \in A\} = \{$time of first entry of $(X_n : n \geq 0)$ into $A\}$ is a stopping time. By convention, $\inf \emptyset = \infty$, so $T = \infty$ if $(X_n : n \geq 0)$ never enters A.

Proof.

Let $n \geq 0$ be fixed. Then $[T \leq n] = [\text{for some } k \leq n, X_k \in A] = \bigcup_{k=0}^{n} [X_k \in A] \in \mathcal{F}_n$ because for every $k = 0, 1, \ldots, n$, $[X_k \in A] \in \mathcal{F}_k \subseteq \mathcal{F}_n$. □

Note that the time you stop gambling in Thackeray's martingale system is an example of such a random time.

5) If $(X_n : n \geq 0)$ is adapted, T is a stopping time and $A \in \mathcal{B}$, then

$$S = \inf\{k \geq T : X_k \in A\}$$

is also a stopping time.

Proof.

For fixed $n \geq 0$,

$$[S \leq n] = \bigcup_{i=0}^{n} \Big([T = i] \cap \big(\bigcup_{k=i}^{n} [X_k \in A] \big) \Big) \in \mathcal{F}_n.$$

□

Example 36

If $(X_n : n \geq 0)$ is adapted and $A \in \mathcal{B}$, then

$$L = \sup\{n : n \leq 10, X_n \in A\}$$

is not in general a stopping time.

Definition 8.5

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, \mathbb{P})$ be a filtered space and let T be a stopping time. The pre-T σ-algebra $\mathcal{F}_T$ is defined by
$\mathcal{F}_T = \{A \in \mathcal{F} : A \cap [T \leq n] \in \mathcal{F}_n \text{ for every } n \geq 0\}$.

Remark 8.6

1) If $T = n$ then $\mathcal{F}_T = \mathcal{F}_n$ (the two definitions agree).
2) $\mathcal{F}_T$ is a σ-algebra because T is a stopping time.

Some properties of $\mathcal{F}_T$:

(1) If S and T are stopping times and $S \leq T$, then $\mathcal{F}_S \subseteq \mathcal{F}_T$.

Proof.

Let $A \in \mathcal{F}_S$. Fix $n \geq 0$. Then, since $[T \leq n] \subseteq [S \leq n]$,

$$A \cap [T \leq n] = A \cap [S \leq n] \cap [T \leq n] \in \mathcal{F}_n ,$$

because $A \cap [S \leq n] \in \mathcal{F}_n$ (since $A \in \mathcal{F}_S$) and $[T \leq n] \in \mathcal{F}_n$ (since T is a stopping time). □

(2) T is $\mathcal{F}_T$-measurable.

Proof.

We need to show that $[T \leq k] \in \mathcal{F}_T$ for every $k \geq 0$. Let $k \geq 0$ be fixed. Then, for every $n \geq 0$,
$[T \leq k] \cap [T \leq n] = [T \leq \min\{k, n\}] \in \mathcal{F}_{\min\{k,n\}} \subseteq \mathcal{F}_n$. □

(3) If S and T are stopping times and $A \in \mathcal{F}_S$, then $A \cap [S \leq T] \in \mathcal{F}_T$.

Proof.

Let $n \geq 0$ be fixed. Then,

$$\begin{aligned} A \cap [S \leq T] \cap [T \leq n] &= A \cap \left(\bigcup_{i=0}^{n} ([T = i] \cap [S \leq i]) \right) \\ &= \bigcup_{i=0}^{n} \left(A \cap [T = i] \cap [S \leq i] \right) \in \mathcal{F}_n, \end{aligned}$$

because for every $i = 0, 1, \ldots, n$, $A \cap [S \leq i] \in \mathcal{F}_i$ (since $A \in \mathcal{F}_S$), $[T = i] \in \mathcal{F}_i$ (since T is a stopping time) and $\mathcal{F}_i \subseteq \mathcal{F}_n$. □

(4) If S and T are stopping times and $A \in \mathcal{F}_S$, then $A \cap [S < T] \in \mathcal{F}_T$.

(5) If S and T are stopping times then $[S \leq T]$, $[S < T]$, $[S = T] \in \mathcal{F}_T \cap \mathcal{F}_S$.

(6) If $(X_n : n \geq 0)$ is adapted and T is a stopping time which is almost surely finite, then X_T is $\mathcal{F}_T$-measurable.

Proof.

We have to show that $[X_T \leq x] \in \mathcal{F}_T$ for every $x \in \mathbb{R}$. Let $x \in \mathbb{R}$ and $n \geq 0$ be fixed. Then

$$[X_T \leq x] \cap [T \leq n] = \bigcup_{i=0}^{n} \left([T = i] \cap [X_i \leq x] \right) \in \mathcal{F}_n ,$$

because $[X_i \leq x] \in \mathcal{F}_i$ (since $(X_n : n \geq 0)$ is adapted), $[T = i] \in \mathcal{F}_i$ (since T is a stopping time) and $\mathcal{F}_i \subseteq \mathcal{F}_n$. □

The proofs of all these properties follow a set pattern rather closely, indicating that Definition 8.5 is the "right" definition for $\mathcal{F}_T$.

The following example is relevant to gambling and investment.

Example 37 (Stopped martingales)

Let $(M_n : n \geq 0)$ be a martingale and let T be a stopping time. Then the stopped process $(M_{T\wedge n} : n \geq 0)$ is a martingale.

Proof: Since $M_{n-1} = \mathbb{E}[M_n | \mathcal{F}_{n-1}]$ for every $n \geq 1$, we know $\mathbb{E}[M_n; A] = \mathbb{E}[M_{n-1}; A]$ and $A \in \mathcal{F}_{n-1}$. So

$$\mathbb{E}[(M_n - M_{n-1}); A] = 0, \qquad n \geq 1,\ A \in \mathcal{F}_{n-1}.$$

Therefore, for every $n \geq 1$ and $A \in \mathcal{F}_{n-1}$,

$$\begin{aligned}\mathbb{E}[(M_{T\wedge n} - M_{T\wedge(n-1)}); A]) &= \mathbb{E}[\mathbb{I}[T \geq n](M_n - M_{n-1}); A] \\ &= \mathbb{E}[(M_n - M_{n-1}); A \cap [T \geq n]] = 0,\end{aligned}$$

because $A \in \mathcal{F}_{n-1}$ and $[T \geq n] \in \mathcal{F}_{n-1}$ (since T is a stopping time), thus $A \cap [T \geq n] \in \mathcal{F}_{n-1}$.

It follows that

$$\mathbb{E}[M_{T\wedge n}; A] = \mathbb{E}[M_{T\wedge(n-1)}; A], \qquad n \geq 1,\ A \in \mathcal{F}_{n-1}.$$

In addition, for every $n \geq 0$, $\mathbb{E}[|M_{T\wedge n}|] \leq \sup_{0\leq i\leq n} \mathbb{E}[|M_i|] < \infty$. Furthermore, by property (6) of $\mathcal{F}_T$ given above, $M_{T\wedge n}$ is $\mathcal{F}_{T\wedge n}$-measurable for every $n \geq 0$, and because $\mathcal{F}_{T\wedge n} \subseteq \mathcal{F}_n$, it is also $\mathcal{F}_n$-measurable. Hence, the process $(M_{T\wedge n} : n \geq 0)$ is adapted which completes the proof that it is a martingale. □

In fact this is a special case of Theorem 8.7 following.

Optional Stopping

Theorem 8.7 (Optional Stopping Theorem for Bounded Stopping Times)

Let $(X_n, \mathcal{F}_n)_{n\geq 0}$ be a supermartingale and let S and T be stopping times such that $0 \leq S \leq T \leq N$ for some $N \geq 0$. Then

$$X_S \quad\geq\quad \mathbb{E}[X_T | \mathcal{F}_S] \quad a.s.$$

If $(X_n, \mathcal{F}_n)_{n\geq 0}$ is a martingale then $X_S = \mathbb{E}[X_T | \mathcal{F}_S]$ a.s.

Corollary 8.8 (Optional Stopping Theorem for Uniformly Integrable Case)

Suppose in the above that $(X_n, \mathcal{F}_n)_{n\geq 0}$ is a uniformly integrable martingale. Then S and T need not be bounded above by N, so long as they are almost surely finite.

Before we prove the theorem we prove the following lemma, which is again in the spirit of Lemma 5.10:

Lemma 8.9

Let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Then $\mathbb{E}[X; A] \geq 0$ for every $A \in \mathcal{G}$ if and only if $\mathbb{E}[X|\mathcal{G}] \geq 0$ a.s..

Proof: (1) Suppose that $\mathbb{E}[X|\mathcal{G}] \geq 0$ a.s. Use the defining properties of conditional expectation:

$$\mathbb{E}[X; A] \quad=\quad \mathbb{E}[\mathbb{I}[A]\,\mathbb{E}[X|\mathcal{G}]] \quad\geq\quad 0, \qquad \text{for } A \in \mathcal{G}.$$

(2) Suppose now that $\mathbb{E}[X;A] \geq 0$ for every $A \in \mathcal{G}$. Consider $A = B_n = [\mathbb{E}[X|\mathcal{G}] < -1/n]$. Then (using the defining properties of conditional expectation)

$$\begin{aligned}\mathbb{E}[X;B_n] &= \mathbb{E}[\mathbb{E}[X|\mathcal{G}];B_n] \\ &= \mathbb{E}[\mathbb{E}[X|\mathcal{G}];\mathbb{E}[X|\mathcal{G}] < -1/n] \\ &< \mathbb{E}\left[-\frac{1}{n};\mathbb{E}[X|\mathcal{G}] < -1/n\right] \\ &= -\frac{1}{n} \times \mathbb{P}[\mathbb{E}[X|\mathcal{G}] < -1/n]\,.\end{aligned}$$

We already know that $\mathbb{E}[X;B_n]$ must be non-negative, since $B_n \in \mathcal{G}$; and the only way that $-\frac{1}{n} \times \mathbb{P}[\mathbb{E}[X|\mathcal{G}] < -1/n]$ can be non-negative is if $\mathbb{P}[B_n] = 0$.
We deduce $\mathbb{P}[B_n] = 0$ for $n = 1, 2, \ldots$, and hence $\mathbb{P}[\mathbb{E}[X|\mathcal{G}] < 0] = 0$.
Hence $\mathbb{E}[X|\mathcal{G}] \geq 0$ almost surely, as required. □

Warwick Statistics

Proof of Theorem 8.7:

Since $(X_n : n \geq 0)$ is a supermartingale, we know that $\mathbb{E}[|X_n|] < \infty$ for every $n \geq 0$. Hence, because T is a bounded stopping time and

$$X_T = \sum_{i=0}^{N} X_i \ \mathbb{I}[T = i]\,,$$

we find that $\mathbb{E}[|X_T|] < \infty$. Similarly $\mathbb{E}[|X_S|] < \infty$. Thus, conditional expectation $\mathbb{E}[X_T|\mathcal{F}_S]$ is well defined.

Since X_S is $\mathcal{F}_S$-measurable and $\mathbb{E}[X_S|\mathcal{F}_S] = X_S$, it is enough to show that

$$\mathbb{E}[X_S|\mathcal{F}_S] \geq \mathbb{E}[X_T|\mathcal{F}_S]\,.$$

Warwick Statistics

We will show that $\mathbb{E}[(X_S - X_T);A] \geq 0$ for any $A \in \mathcal{F}_S$. Let $A \in \mathcal{F}_S$ be fixed. Then (collapsing sum!)

$$\begin{aligned}\mathbb{E}[(X_S - X_T);A] &= \mathbb{E}\left[\sum_{j=1}^{N}(X_{j-1} - X_j) \times \mathbb{I}[S < j \leq T];A\right] \\ &= \sum_{j=1}^{N} \mathbb{E}\left[(X_{j-1} - X_j)\ \mathbb{I}[A]\ \mathbb{I}[S \leq j-1]\ \mathbb{I}[T > j-1]\right] \\ &= \sum_{j=1}^{N} \mathbb{E}\left[(X_{j-1} - X_j)\ \mathbb{I}[A \cap \{S \leq j-1\} \cap \{T > j-1\}]\right]\end{aligned}$$

Since $A \in \mathcal{F}_S$, $A \cap \{S \leq j-1\} \in \mathcal{F}_{j-1}$, and also $[T \geq j] = [T \leq j-1]^c \in \mathcal{F}_{j-1}$ because S, T are stopping times. Thus, $A \cap \{S \leq j-1\} \cap \{T > j-1\} \in \mathcal{F}_{j-1}$.

Warwick Statistics

The supermartingale property of the process $(X_n : n \geq 0)$ implies that, for every $j \geq 1$,

$$\mathbb{E}\left[X_{j-1} - X_j|\mathcal{F}_{j-1}\right] = X_{j-1} - \mathbb{E}\left[X_j|\mathcal{F}_{j-1}\right] \geq 0 \quad \text{a.s.}$$

Thus, by Lemma 8.9, $\mathbb{E}\left[(X_{j-1} - X_j);B\right] \geq 0$ for every $B \in \mathcal{F}_{j-1}$, which, because $A \cap \{S \leq j-1\} \cap \{T > j-1\} \in \mathcal{F}_{j-1}$, implies that for every $j \geq 1$,

$$\mathbb{E}\left[(X_{j-1} - X_j);A \cap [S \leq j-1] \cap [T > j-1]\right] \geq 0\,.$$

It follows that $\mathbb{E}[(X_S - X_T);A] \geq 0$ for every $A \in \mathcal{F}_S$. Finally, we conclude by Lemma 8.9 that $\mathbb{E}[X_S - X_T|\mathcal{F}_S] \geq 0$ a.s. □

Warwick Statistics

Proof of Optional Stopping Theorem for Uniformly Integrable Case

Sketch.

We assume without proof a known result: if the family $\{X_0, X_1, \ldots\}$ is uniformly integrable then so is the family $\{X_S : \text{ stopping times } S\}$. The result then follows rapidly by noting

(i) The result holds for $\min\{S, n\} \leq \min\{T, n\}$ for any finite n;

(ii) So $\mathbb{E}\left[X_{\min\{S,n\}}; A\right] = \mathbb{E}\left[X_{\min\{T,n\}}; A\right]$ for any A in $\mathcal{F}_{\min\{S,n\}}$;

(iii) Now apply Theorem 5.41: $X_{\min\{S,n\}} \to X_S$, $X_{\min\{Y,n\}} \to X_T$ a.s., therefore by uniform integrability $\mathbb{E}\left[X_S; A\right] = \mathbb{E}\left[X_T; A\right]$ for any A in $\mathcal{F}_{\min\{S,m\}}$ (some fixed m);

(iv) Finally, extend to any A in $\mathcal{F}_S$ by a Π-system argument. □

Example 38 (Optional stopping for symmetric simple random walks)

Suppose $\{X_n\}$ is a simple symmetric random walk begun at 0, and $T = \inf\{n : X_n = -a \text{ or } X_n = b\}$ for $a, b \geq 0$. Then $\{X_n\}$ is a martingale. What is more, the stopped martingale $\{X_{\min\{n,T\}}$ is bounded hence uniformly integrable. Finally, $T < \infty$ a.s., by an argument showing that eventually a run of $a + b$ consecutive $+1$ jumps must occur.

So by Corollary 8.8 we see $\mathbb{E}\left[X_T\right] = 0$. But $\mathbb{E}\left[X_T\right] = -a\,\mathbb{P}\left[X_t = -a\right] + b\,\mathbb{P}\left[X_T = b\right]$ and $\mathbb{P}\left[X_t = -a\right] + \mathbb{P}\left[X_T = b\right] = 1$. We can solve to obtain

$$\mathbb{P}\left[X_t = -a\right] \quad = \quad \frac{b}{a+b}\,.$$

Example 39 (Mean escape times for symmetric simple random walks)

In the previous example, the runs argument also shows $\mathbb{E}\left[T\right] < \infty$. (Consider $\mathbb{E}\left[T/(a+b)\right] \leq \sum_k (1 - 2^{-a-b})^k$.)
Consider $f(x) = \mathbb{E}\left[T | X_0 = x\right]$. Now $Y_n = \mathbb{E}\left[T | \mathcal{F}_n\right]$ determines a martingale (Example 34). So

$$f(x) \;=\; Y_0 \;=\; \mathbb{E}\left[Y_1 | \mathcal{F}_0, X_0 = x\right] \;=\; \mathbb{E}\left[\mathbb{E}\left[T | \mathcal{F}_1\right] | \mathcal{F}_0, X_0 = x\right].$$

Hence we can deduce (independent jumps!)

$$\begin{aligned} f(x) = \frac{1}{2}\mathbb{E}\left[T | \mathcal{F}_1, X_1 = x+1, X_0 = x\right] + \frac{1}{2}\mathbb{E}\left[T | \mathcal{F}_1, X_1 = x-1, X_0 = x\right] \\ = \frac{1}{2}\mathbb{E}\left[1 + T | \mathcal{F}_0, X_0 = x+1\right] + \frac{1}{2}\mathbb{E}\left[1 + T | \mathcal{F}_0, X_0 = x-1\right] \end{aligned}$$

and so $2f(x) = 2 + f(x+1) + f(x-1)$. Since $f(-a) = f(b) = 0$, we can solve this recurrence relation: $f(x) = (x+a)(b-x)$.

We can also work with asymmetric simple random walks $\{X_n\}$: these are not martingales, but $Y_n = (q/p)^{X_n}$ will be for suitable p and q. This allows us to find a formula for $\mathbb{P}\left[X_t = -a\right]$.

We can also find a formula for the *probability generating function* of T as a function of the starting point x.

These ideas generalize considerably. For example, they eventually lead to the *Foster-Lyapunov criteria* for recurrence and rate of convergence to equilibrium for Markov chains.

Martingale Convergence

Let $(X_n : n \geq 0)$ be a supermartingale and let $a, b \in \mathbb{R}$, $a < b$, be fixed. Define stopping times $T'_0 = 0$, $S'_{n+1} = \inf\{k > T'_n : X_k \leq a\}$, $T'_{n+1} = \inf\{k > S'_{n+1} : X_k \geq b\}$, for $n \geq 0$.
Fix $N \in \mathbb{N}$ and define $T_k = T'_k \wedge N$ and $S_k = S'_k \wedge N$, $k \geq 1$.
Let

$$\begin{aligned} U_N[a,b] &= \sup\{k \geq 0 : T_k < N\} \\ &= \text{number of upcrossings of } [a,b] \text{ completed strictly before } N. \end{aligned}$$

Lemma 8.10 (Doob's Upcrossing Lemma)

$$\mathbb{E}\left[U_N[a,b]\right] \leq \frac{\mathbb{E}\left[(a - X_N)^+\right]}{b-a},$$

where $(a - X_N)^+ = (a - X_N); a - X_N \geq 0$.

Proof T_k, $k \geq 0$, and S_k, $k \geq 1$, are bounded stopping times, so apply Optional Stopping Theorem 8.7; $X_{S_k} \geq \mathbb{E}\left[X_{T_k} | \mathcal{F}_{S_k}\right]$ a.s. for every $k \geq 1$.
Then, for $k \geq 1$,

$$\begin{aligned} 0 \geq (X_{S_k} - a)\, \mathbb{I}\left[S_k < N\right] &\geq \mathbb{E}\left[X_{T_k} - a\, \mathbb{I}\left[S_k < N\right] | \mathcal{F}_{S_k}\right] \\ &= \mathbb{E}\left[X_{T_k} - a\, \mathbb{I}\left[S_k < T_k < N\right] + (X_N - a)\, \mathbb{I}\left[S_k < N = T_k\right] | \mathcal{F}_{S_k}\right] \\ &\geq \mathbb{E}\left[(b-a)\, \mathbb{I}\left[S_k < T_k < N\right] + (X_N - a)\, \mathbb{I}\left[S_k < N = T_k\right] | \mathcal{F}_{S_k}\right]. \end{aligned}$$

By taking expectations of the above, and re-arranging, we obtain

$$\begin{aligned} \mathbb{E}\left[(a - X_N); S_k < N = T_k\right] &\geq \\ &(b-a)\ \mathbb{P}\left[T_k < N\right] = (b-a)\, \mathbb{P}\left[U_N[a,b] \geq k\right]. \end{aligned}$$

But

$$\begin{aligned} \mathbb{E}\left[a - X_N; S_k < N = T_k\right] &\leq \mathbb{E}\left[(a - X_N)^+; S_k < N = T_k\right] \\ &\leq \mathbb{E}\left[(a - X_N)^+; U_N[a,b] = k-1\right] \end{aligned}$$

(because $[S_k < N = T_k] \subseteq [U_N[a,b] = k-1]$).
Therefore,

$$\mathbb{E}\left[(a - X_N)^+; U_N[a,b] = k-1\right] \geq (b-a)\, \mathbb{P}\left[U_N[a,b] \geq k\right] \geq 0.$$

Summing over $k \geq 1$ yields

$$\mathbb{E}\left[(a - X_N)^+\right] \geq (b-a) \sum_{k=1}^{\infty} \mathbb{P}\left[U_N[a,b] \geq k\right] = (b-a)\ \mathbb{E}\left[U_N[a,b]\right]$$

(because if Z is a non-negative random variable taking integer values, then $\mathbb{E}[Z] = \sum_{k=1}^{\infty} \mathbb{P}[Z \geq k]$). □

Theorem 8.11 (Doob's Supermartingale Convergence Theorem)

Let $(X_n : n \geq 0)$ be a supermartingale bounded in $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ ($\sup_n \mathbb{E}[|X_n|] < \infty$). Then $X_n \to X_\infty$ a.s. and $X_\infty \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$.

Proof: Since

$$\mathbb{E}\left[U_N[a,b]\right] \leq \frac{\mathbb{E}\left[(a - X_N)^+\right]}{b-a} \leq \frac{\mathbb{E}\left[|a| + |X_N|\right]}{b-a} \leq \frac{|a| + \sup_n \mathbb{E}\left[|X_n|\right]}{(b-a)},$$

by letting $N \to \infty$ we deduce by (MON) that

$$\mathbb{E}\left[U_\infty[a,b]\right] \leq \frac{|a| + \sup_n \mathbb{E}\left[|X_n|\right]}{(b-a)},$$

thus,

$$\mathbb{P}\left[U_\infty[a,b] < \infty\right] = 1. \qquad (8.1)$$

Let

$$\begin{aligned}\Lambda &= \{\omega : X_n(\omega) \text{ does not converge to a limit in } [-\infty,\infty]\}\\ &= \{\omega : \liminf_n X_n(\omega) < \limsup_n X_n(\omega)\}\\ &= \bigcup_{\{a,b\in\mathbb{Q}:a<b\}} \Lambda_{a,b},\end{aligned}$$

where

$$\Lambda_{a,b} = \{\omega : \liminf_n X_n(\omega) < a < b < \limsup_n X_n(\omega)\}\,.$$

But $\Lambda_{a,b} \subseteq \{\omega : U_\infty[a,b](\omega) = \infty\}$. Therefore, by (8.1), for any $a, b \in \mathbb{Q}$, $a < b$, $\mathbb{P}[\Lambda_{a,b}] = 0$, which implies that $\mathbb{P}[\Lambda] = 0$ (because a countable union of null-sets is a null-set).
Hence, $X_\infty = \lim_{n\to\infty} X_n$ exists a.s. in $[-\infty,\infty]$.

But

$$\begin{aligned}\mathbb{E}[|X_\infty|] &= \mathbb{E}\left[\liminf_n |X_n|\right] && (\text{because } |X_\infty| = \lim_{n\to\infty}|X_n|)\\ &\leq \liminf_n \mathbb{E}[|X_n|] && (\text{by Fatou})\\ &\leq \sup_n \mathbb{E}[|X_n|] < \infty,\end{aligned}$$

so that

$$\mathbb{P}[X_\infty < \infty] = 1 \quad\text{and}\quad X_\infty \in \mathcal{L}^1(\Omega,\mathcal{F},\mathbb{P})$$

(because $\mathbb{E}[|Z|] < \infty$ implies $\mathbb{P}[Z < \infty] = 1$). □

Lemma 8.12

A non-negative supermartingale converges almost surely.

Proof.

Let $(X_n : n \geq 0)$ be a non-negative supermartingale. Then, for every $n \geq 0$

$$\mathbb{E}[|X_{n+1}|] = \mathbb{E}[X_{n+1}] \leq \mathbb{E}[X_0] = \mathbb{E}[|X_0|] < \infty\,.$$

Hence the process $(X_n : n \geq 0)$ is bounded in $\mathcal{L}^1(\Omega,\mathcal{F},\mathbb{P})$ which by Doob's Supermartingale Convergence Theorem 8.11 implies that $(X_n : n \geq 0)$ converges almost surely. □

Example 40 (Simple symmetric random walk stopped at zero)

Suppose (S_n) is a simple symmetric random walk but at some positive integer $S_0 > 0$.
Let $T = \inf\{k \geq 0 : S_k = 0\}$. Then T is a stopping time and, by Example 37, the process $(S_{T\wedge n} : n \geq 0)$ is a non-negative martingale.
Thus, by Lemma 8.12, $(S_{T\wedge n} : n \geq 0)$ converges almost surely.
However, $(S_{T\wedge n} : n \geq 0)$ does not converge in $\mathcal{L}^1$: in fact it can be shown that $\mathbb{P}[T < \infty] = 1$ which implies that $S_{T\wedge n} \to S_T$ a.s.. Hence, $\mathbb{E}[S_{T\wedge n}] = \mathbb{E}[S_1] = 1$ for every $n \geq 0$ but $\mathbb{E}[S_T] = 0$.

The last result of these lectures can be thought of as a gorwn-up version of the reflection principle for simple symmetric random walks, and produces much insight into the fluctuation of martingales.

Theorem 8.13 (Doob's Submartingale Inequality)

Let $(X_n : n \geq 0)$ be a non-negative submartingale. Then, for fixed $c > 0$ and $n \geq 0$,

$$c \; \mathbb{P}\left[\sup_{k \leq n} X_k \geq c\right] \quad \leq \quad \mathbb{E}\left[X_n ; \sup_{k \leq n}\{X_k \geq c\}\right] \leq \mathbb{E}\left[X_n\right] .$$

Proof: Let $F = [\sup_{k \leq n} X_k \geq c]$. Express F as a disjoint union $F = F_0 \cup F_1 \cup \cdots \cup F_n$, where

$$\begin{aligned} F_0 &= [X_0 \geq c] \\ F_k &= [X_0 < c] \cap [X_1 < c] \cap \cdots \cap [X_{k-1} < c] \cup [X_k \geq c] , \\ &\qquad \text{for } 1 \leq k \leq n . \end{aligned}$$

Then $F_k \in \mathcal{F}_k$ for $0 \leq k \leq n$, where $\mathcal{F}_k = \sigma(X_0, X_1, \ldots, X_k)$, $k \geq 0$. Hence (using Lemma 8.9),

$$\begin{aligned} \mathbb{E}\left[X_n ; F_k\right] &\geq \mathbb{E}\left[X_k ; F_k\right] \\ &\geq c\,\mathbb{P}\left[F_k\right] \quad (\text{because } X_k \geq c \text{ on } F_k). \end{aligned}$$

Summing over k yields the result. □

Bibliography

This is a rich hypertext bibliography. Journals are linked to their homepages, and stable URL links (as provided for example by JSTOR or Project Euclid) have been added where known. Access to such URLs is not universal: in case of difficulty you should check whether you are registered (directly or indirectly) with the relevant provider. In the case of preprints, icons , , , linking to homepage locations are inserted where available: note that these are less stable than journal links!.

Hoffmann-Jørgensen, J. (1994).
Probability with a view toward statistics. Vol. I.
Chapman & Hall Probability Series. New York: Chapman & Hall.

Jacod, J. and P. Protter (2003).
Probability essentials (Second ed.).
Universitext. Berlin: Springer-Verlag.
University library has 6 copies under QA 273.J2.

Kendall, W. S. (1993).
On the empty cells of Poisson histograms.
Journal of Applied Probability 30, 561–574.

Ripley, B. D. (1987).
Stochastic simulation.
Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc.

Steele, J. M. (2004).
The Cauchy-Schwarz master class.
MAA Problem Books Series. Washington, DC: Mathematical Association of America.
An introduction to the art of mathematical inequalities.

Williams, D. (1991).
Probability with martingales.
Cambridge Mathematical Textbooks. Cambridge: Cambridge University Press.
University library has 8 copies under QA273.W4.