CONDITIONAL EXPECTATION AND MARTINGALE

by

AVINASH DESAI AND SUBHRAJIT MAITY

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY,
PUNE
IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE OF

MASTER OF SCIENCE

UNDER THE GUIDANCE OF

DR. MOUMANTI PODDER

Assistant Professor, IISER, Pune

MS. MAYA BHOSALE

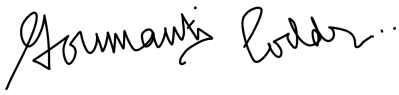
Assistant Professor, SPPU

DEPARTMENT OF MATHEMATICS
SAVITRIBAI PHULE PUNE UNIVERSITY

Certificate of Completion

This is to certify that the project titled **"Conditional Expectation and Martingale"** has been successfully carried out and completed by **Avinash Desai** and **Subhrajit Maity** of the Department of Mathematics, Savitribai Phule Pune University.

This project was carried out under the supervision of **Ms. Maya Bhosale**, SPPU and External expert **Dr. Moumanti Podder**, IISER Pune during the academic year **2024–25** for the partial fulfillment of the M.Sc. degree. In our opinion, the thesis meets the criteria for the award of the degree.



Dr. Moumanti Podder
(External Mentor)
Assistant Professor
Dept. of Mathematics, IISER, Pune

Ms. Maya Bhosale
(Internal Mentor)
Assistant Professor
Dept. of Mathematics, SPPU

Dr. Yashwant Borse
Head of the Department
Dept. of Mathematics, SPPU

Acknowledgement

We take this opportunity to thank our external project guide **Dr. Moumanti Podder**, Assistant Professor of the Department of Mathematics, Indian Institute of Science Education and Research, Pune (IISER Pune) for providing us the opportunity to work in an exciting and challenging field of Conditional Expectation and Martingale. Our interactions with her have been of immense help in defining my project goals and in identifying ways to achieve them.

Our honorable mention goes to **Ms. Maya Bhosale**, Department of Mathematics, Savitribai Phule Pune University, Pune, who was mentoring this project for granting us the opportunity to gain project experience, and for support and encouragement.

This is a great pleasure and immense satisfaction to express our deepest sense of gratitude and thanks to everyone who has directly or indirectly helped us in completing this project successfully.

Mr. Avinash Desai **Mr. Subhrajit Maity**

Abstract

This work offers a rigorous exploration of conditional expectation and martingale theory grounded in the framework of measure-theoretic probability. The work begins by developing the foundational elements of measure theory—algebras, σ-algebras, and the Borel σ-algebra—which form the basis for defining probability spaces and measurable functions. We then construct the Lebesgue integral and define expectations in this formal setting, enabling a precise treatment of random variables and their distributions.

A central focus is the development of conditional expectation, particularly its definition with respect to a σ-field. We extend this to the notion of filtrations—increasing families of σ-fields that model the accumulation of information over time. Filtrations play a crucial role in the study of stochastic processes, allowing us to rigorously define conditional expectations in dynamic settings and laying the foundation for the study of martingales.

In the second part, we delve into martingales, a class of stochastic processes that preserve conditional expectations relative to a filtration. We explore their defining properties, convergence results, and key theoretical tools such as the Optional Skipping Theorem and Optional Sampling Theorems. The framework of filtrations allows us to understand martingales as processes that evolve "fairly" over time with respect to the available information.

Finally, we illustrate the power of these ideas by applying martingale theory to the study of Markov chains, particularly in classifying states as recurrent or transient. Throughout, the work aims to bridge abstract theory with practical insight, providing a comprehensive and accessible treatment of conditional expectation and martingales suitable for advanced study and further research in probability theory, stochastic processes, and related fields.

Contents

Chapter 1
Measure theory and Probability spaces

1.1 Algebras and σ-algebras of Sets

1.1.1 Algebra of Sets

An **algebra** $\mathcal{A}$ over a set Ω (called the sample space) is a collection of subsets of Ω such that:

(i) $\Omega \in \mathcal{A}$,

(ii) If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$ (closed under complements),

(iii) If $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$ (closed under finite unions).

Because an algebra is closed under complements and finite unions, it is also closed under finite intersections.

1.1.2 σ-algebra

A **σ-algebra** $\mathcal{F}$ over a set Ω is an algebra that is also closed under countable unions. Formally, it satisfies:

(i) $\Omega \in \mathcal{F}$,

(ii) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$,

(iii) If $A_1, A_2, A_3, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ (closed under countable unions).

By De Morgan's laws, it also follows that a σ-algebra is closed under countable intersections.

1.1.3 Example

Let $\Omega = \{1, 2, 3\}$.

- An algebra on Ω could be:

$$\mathcal{A} = \{\emptyset, \{1, 2\}, \{3\}, \Omega\}$$

- A σ-algebra on Ω must contain all subsets of Ω, i.e., the power set:

$$\mathcal{F} = \mathcal{P}(\Omega) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \Omega\}$$

1.2 The Borel σ-algebra

The **Borel σ-algebra** on $\mathbb{R}$, denoted by $\mathcal{B}(\mathbb{R})$, is the smallest σ-algebra that contains all open subsets of $\mathbb{R}$.

Equivalently, $\mathcal{B}(\mathbb{R})$ is the σ-algebra generated by the collection of open intervals:

$$\mathcal{B}(\mathbb{R}) = \sigma\left(\{(a, b) \subset \mathbb{R} \mid a < b\}\right)$$

1.2.1 Properties

- $\mathcal{B}(\mathbb{R})$ contains:
  - All open intervals (a, b),
  - All closed intervals $[a, b]$,
  - All half-open intervals $[a, b), (a, b]$,
  - All countable unions and intersections of such intervals,
  - All singletons and countable sets.
- It is the minimal σ-algebra containing all the topology of $\mathbb{R}$.

1.2.2 Use in Probability Theory

In probability theory and measure theory, the Borel σ-algebra is commonly used as the domain of definition for probability measures on $\mathbb{R}$. That is, a probability space over $\mathbb{R}$ is usually written as $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P)$, where P is a probability measure defined on the Borel sets.

1.2.3 Example

Let P be a probability measure defined by the standard normal distribution:

$$P((a, b)) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

This is a valid probability measure on the Borel σ-algebra $\mathcal{B}(\mathbb{R})$.

1.3 Probability Spaces

A **probability space** is a triple $(\Omega, \mathcal{F}, P)$ where Ω is a set of "outcomes," $\mathcal{F}$ is a set of "events," and $P : \mathcal{F} \to [0, 1]$ is a function that assigns probabilities to events. We assume that $\mathcal{F}$ is a **σ-field** (or **σ-algebra**), i.e., a (nonempty) collection of subsets of Ω that satisfy

- if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$, and
- if $A_i \in \mathcal{F}$ is a countable sequence of sets, then $\bigcup_i A_i \in \mathcal{F}$.

Here and in what follows, **countable** means finite or countably infinite. Since $\bigcap_i A_i = (\bigcup_i A_i^c)^c$, it follows that a σ-field is closed under countable intersections. We omit the last property from the definition to make it easier to check.

Without P, $(\Omega, \mathcal{F})$ is called a **measurable space**, i.e., it is a space on which we can put a measure. A **measure** is a nonnegative countably additive set function; that is, a function $\mu : \mathcal{F} \to \mathbb{R}$ with

(i) $\mu(A) \geq \mu(\emptyset) = 0$ for all $A \in \mathcal{F}$, and

(ii) if $A_i \in \mathcal{F}$ is a countable sequence of disjoint sets, then

$$\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i)$$

If $\mu(\Omega) = 1$, we call μ a **probability measure**. Probability measures are usually denoted by P.

The next result gives some consequences of the definition of a measure that we will need later. In all cases, we assume that the sets we mention are in $\mathcal{F}$.

1.3.1 Theorem

Let μ be a measure on $(\Omega, \mathcal{F})$

(i) **monotonicity.** *If $A \subset B$ then $\mu(A) \leq \mu(B)$.*

(ii) **subadditivity.** *If $A \subset \bigcup_{m=1}^{\infty} A_m$ then $\mu(A) \leq \sum_{m=1}^{\infty} \mu(A_m)$.*

(iii) **continuity from below.** *If $A_i \uparrow A$ (i.e., $A_1 \subset A_2 \subset \ldots$ and $\bigcup_i A_i = A$) then $\mu(A_i) \uparrow \mu(A)$.*

(iv) **continuity from above.** *If* $A_i \downarrow A$ *(i.e.,* $A_1 \supset A_2 \supset \ldots$ *and* $\bigcap_i A_i = A$*), with* $\mu(A_1) < \infty$ *then* $\mu(A_i) \downarrow \mu(A)$.

Proof:

(i) Let $B - A = B \cap A^c$ be the **difference** of the two sets. Using + to denote disjoint union, $B = A + (B - A)$ so

$$\mu(B) = \mu(A) + \mu(B - A) \geq \mu(A).$$

(ii) Let $A'_n = A_n \cap A$, $B_1 = A'_1$ and for $n > 1$, $B_n = A'_n - \bigcup_{m=1}^{n-1} A'_m$. Since the B_n are disjoint and have union A, we have using (ii) of the definition of measure, $B_m \subset A_m$, and (i) of this theorem

$$\mu(A) = \sum_{m=1}^{\infty} \mu(B_m) \leq \sum_{m=1}^{\infty} \mu(A_m).$$

(iii) Let $B_n = A_n - A_{n-1}$. Then the B_n are disjoint and have $\bigcup_{m=1}^{\infty} B_m = A$, $\bigcup_{m=1}^{n} B_m = A_n$ so

$$\mu(A) = \sum_{m=1}^{\infty} \mu(B_m) = \lim_{n\to\infty} \sum_{m=1}^{n} \mu(B_m) = \lim_{n\to\infty} \mu(A_n).$$

(iv) $A_1 - A_n \uparrow A_1 - A$ so (iii) implies $\mu(A_1 - A_n) \uparrow \mu(A_1 - A)$. Since $A_1 \supset A_n$ we have $\mu(A_1 - A) = \mu(A_1) - \mu(A)$ and it follows that $\mu(A_n) \downarrow \mu(A)$.

1.3.2 Example: Discrete probability spaces.

Let Ω = a countable set, i.e., finite or countably infinite. Let $\mathcal{F}$ = the set of all subsets of Ω. Let

$$P(A) = \sum_{\omega \in A} p(\omega) \quad \text{where } p(\omega) \geq 0 \quad \text{and} \quad \sum_{\omega \in \Omega} p(\omega) = 1$$

A little thought reveals that this is the most general probability measure on this space. In many cases when Ω is a finite set, we have $p(\omega) = 1/|\Omega|$ where $|\Omega|$ = the number of points in Ω.

For a simple concrete example that requires this level of generality, consider the astragali, dice used in ancient Egypt made from the ankle bones of sheep. This die could come to rest on the top side of the bone for four points or on the bottom for three points. The side of the bone was slightly rounded. The die could come to rest on a flat and narrow piece for six points or somewhere on the rest of the side for one point. There is no reason to think that all four outcomes are equally likely, so we need probabilities p_1, p_3, p_4, and p_6 to describe P.

1.4 Distributions

Probability spaces become a little more interesting when we define random variables on them. A real valued function X defined on Ω is said to be a **random variable** if for every Borel set $B \subset \mathbb{R}$, we have

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

When we need to emphasize the σ-field, we will say that X is $\mathcal{F}$-measurable or write $X \in \mathcal{F}$. If Ω is a discrete probability space (see Example 1.1.2), then any function $X : \Omega \to \mathbb{R}$ is a random variable.

A second trivial, but useful, type of example of a random variable is the **indicator function** of a set $A \in \mathcal{F}$:

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

The notation is supposed to remind you that this function is 1 on A. Analysts call this object the characteristic function of A. In probability, that term is used for something quite different.

If X is a random variable, then X induces a probability measure on $\mathbb{R}$ called its **distribution** by setting

$$\mu(A) = P(X \in A)$$

for Borel sets A. Using the notation introduced above, the right-hand side can be written as

$$P(X^{-1}(A)).$$

In words, we pull $A \in \mathcal{R}$ back to $X^{-1}(A) \in \mathcal{F}$ and then take P of that set.

To check that μ is a probability measure we observe that if the A_i are disjoint then using the definition of μ; the fact that X lands in the union if and only if it lands in one of the A_i; the fact that if the sets $A_i \in \mathcal{R}$ are disjoint then the events $\{X \in A_i\}$ are disjoint; and the definition of μ again; we have:

$$\mu\left(\bigcup_i A_i\right) = P\left(X \in \bigcup_i A_i\right) = P\left(\bigcup_i \{X \in A_i\}\right) = \sum_i P(X \in A_i)$$
$$= \sum_i \mu(A_i).$$

The distribution of a random variable X is usually described by giving its **distribution function**,

$$F(x) = P(X \leq x).$$

1.5 Random Variables

In this section, we will develop some results that will help us later to prove that quantities we define are random variables, i.e., they are measurable. Since most of what we have to say is true for random elements of an arbitrary measurable space $(S, \mathcal{S})$ and the proofs are the same (sometimes easier), we will develop our results in that generality. First, we need a definition.

A function $X : \Omega \to S$ is said to be a **measurable map** from $(\Omega, \mathcal{F})$ to $(S, \mathcal{S})$ if

$$X^{-1}(B) \equiv \{\omega : X(\omega) \in B\} \in \mathcal{F} \quad \text{for all } B \in \mathcal{S}$$

If $(S, \mathcal{S}) = (\mathbb{R}^d, \mathcal{R}^d)$ and $d > 1$, then X is called a **random vector**. Of course, if $d = 1$, X is called a **random variable**, or *r.v.* for short.

The next result is useful for proving that maps are measurable.

1.5.1 Theorem

If $\{\omega : X(\omega) \in A\} \in \mathcal{F}$ for all $A \in \mathcal{A}$ and $\mathcal{A}$ generates $\mathcal{S}$ (i.e., $\mathcal{S}$ is the smallest σ-field that contains $\mathcal{A}$), then X is measurable.

Proof: Writing $\{X \in B\}$ as shorthand for $\{\omega : X(\omega) \in B\}$, we have

$$\{X \in \cup_i B_i\} = \cup_i \{X \in B_i\}$$

$$\{X \in B^c\} = \{X \in B\}^c$$

So the class of sets $\mathcal{B} = \{B : \{X \in B\} \in \mathcal{F}\}$ is a σ-field. Since $\mathcal{B} \supset \mathcal{A}$ and $\mathcal{A}$ generate $\mathcal{S}$, $\mathcal{B} \supset \mathcal{S}$.

It follows from the two equations displayed in the previous proof that if $\mathcal{S}$ is a σ-field, then $\{\{X \in B\} : B \in \mathcal{S}\}$ is a σ-field. It is the smallest σ-field on Ω that makes X a measurable map. It is called the **σ-field generated by** X and denoted $\sigma(X)$. For future reference, we note that

$$\sigma(X) = \{\{X \in B\} : B \in \mathcal{S}\}$$

1.5.2 Theorem

If $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ and $f : (S, \mathcal{S}) \to (T, \mathcal{T})$ are measurable maps, then $f(X)$ is a measurable map from $(\Omega, \mathcal{F})$ to $(T, \mathcal{T})$.

proof: Let $B \in \mathcal{T}$. Then

$$\{\omega : f(X(\omega)) \in B\} = \{\omega : X(\omega) \in f^{-1}(B)\} \in \mathcal{F},$$

since by assumption $f^{-1}(B) \in \mathcal{S}$ and X is measurable.

1.5.3 Theorem

If $X_1, \ldots, X_n$ are random variables and $f : (\mathbb{R}^n, \mathcal{R}^n) \to (\mathbb{R}, \mathcal{R})$ is measurable, then $f(X_1, \ldots, X_n)$ is a random variable.

Proof: In view of Theorem 1.5.2, it suffices to show that $(X_1, \ldots, X_n)$ is a random vector. To do this, we observe that if $A_1, \ldots, A_n$ are Borel sets, then

$$\{(X_1, \ldots, X_n) \in A_1 \times \cdots \times A_n\} = \bigcap_i \{X_i \in A_i\} \in \mathcal{F}$$

Since sets of the form $A_1 \times \cdots \times A_n$ generate $\mathcal{R}^n$, the desired result follows from Theorem 1.3.1.

1.5.4 Theorem

If $X_1, \ldots, X_n$ are random variables, then $X_1 + \cdots + X_n$ is a random variable.

Proof: In view of Theorem 1.5.3, it suffices to show that $f(x_1, \ldots, x_n) = x_1 + \cdots + x_n$ is measurable. Note that the set

$$\{x : x_1 + \cdots + x_n < a\}$$

is an open set and hence is in $\mathcal{R}^n$.

1.5.5 Theorem

If $X_1, X_2, \ldots$ are random variables then so are

$$\inf_n X_n, \quad \sup_n X_n, \quad \limsup_n X_n, \quad \liminf_n X_n.$$

Proof: Since the infimum of a sequence is $< a$ if and only if some term is $< a$ (if all terms are $\geq a$, then the infimum is), we have

$$\left\{\inf_n X_n < a\right\} = \bigcup_n \{X_n < a\} \in \mathcal{F}.$$

A similar argument shows

$$\left\{\sup_n X_n > a\right\} = \bigcup_n \{X_n > a\} \in \mathcal{F}.$$

For the last two, we observe

$$\liminf_{n\to\infty} X_n = \sup_n \left(\inf_{m \geq n} X_m\right),$$

$$\limsup_{n\to\infty} X_n = \inf_n \left(\sup_{m\geq n} X_m \right).$$

To complete the proof in the first case, note that $Y_n = \inf_{m\geq n} X_m$ is a random variable for each n, so $\sup_n Y_n$ is as well.

1.6 Integration

Let μ be a σ-finite measure on $(\Omega, \mathcal{F})$. We will be primarily interested in the special case μ is a probability measure, but we will sometimes need to integrate with respect to infinite measure and it is no harder to develop the results in general.

In this section we will define $\int f\, d\mu$ for a class of measurable functions. This is a four-step procedure:

1. Simple functions
2. Bounded functions
3. Nonnegative functions
4. General functions

This sequence of four steps is also useful in proving integration formulas.

Step 1.φ is said to be a **simple function** if $\varphi(\omega) = \sum_{i=1}^n a_i 1_{A_i}$ and A_i are disjoint sets with $\mu(A_i) < \infty$. If φ is a simple function, we let

$$\int \varphi\, d\mu = \sum_{i=1}^n a_i \mu(A_i)$$

The representation of φ is not unique since we have not supposed that the a_i are distinct. However, it is easy to see that the last definition does not contradict itself.

We will prove the next three conclusions four times, but before we can state them for the first time, we need a definition. $\varphi \geq \psi$ *μ-almost everywhere* (or $\varphi \geq \psi$ μ-a.e.) means $\mu(\{\omega : \varphi(\omega) < \psi(\omega)\}) = 0$. When there is no doubt about what measure we are referring to, we drop the μ.

1.6.1 lemma

Let φ and ψ be simple functions.

(i) If $\varphi \geq 0$ a.e. then $\int \varphi \, d\mu \geq 0$.

(ii) For any $a \in \mathbb{R}$, $\int a\varphi \, d\mu = a \int \varphi \, d\mu$.

(iii) $\int (\varphi + \psi) \, d\mu = \int \varphi \, d\mu + \int \psi \, d\mu$.

Proof: (i) and (ii) are immediate consequences of the definition. To prove (iii), suppose

$$\varphi = \sum_{i=1}^{m} a_i \mathbf{1}_{A_i} \quad \text{and} \quad \psi = \sum_{j=1}^{n} b_j \mathbf{1}_{B_j}$$

To make the supports of the two functions the same, we let $A_0 = \cup_i B_i - \cup_i A_i$, let $B_0 = \cup_i A_i - \cup_i B_i$, and let $a_0 = b_0 = 0$. Now

$$\varphi + \psi = \sum_{i=0}^{m} \sum_{j=0}^{n} (a_i + b_j) \mathbf{1}_{A_i \cap B_j}$$

and the $A_i \cap B_j$ are pairwise disjoint, so

$$\begin{aligned} \int (\varphi + \psi) \, d\mu &= \sum_{i=0}^{m} \sum_{j=0}^{n} (a_i + b_j) \mu(A_i \cap B_j) \\ &= \sum_{i=0}^{m} \sum_{j=0}^{n} a_i \mu(A_i \cap B_j) + \sum_{j=0}^{n} \sum_{i=0}^{m} b_j \mu(A_i \cap B_j) \\ &= \sum_{i=0}^{m} a_i \mu(A_i) + \sum_{j=0}^{n} b_j \mu(B_j) = \int \varphi \, d\mu + \int \psi \, d\mu \end{aligned}$$

In the next-to-last step, we used $A_i = +_j (A_i \cap B_j)$ and $B_j = +_i (A_i \cap B_j)$, where $+$ denotes a disjoint union.

We will prove (i)–(iii) three more times as we generalize our integral. As a consequence of (i)–(iii), we get three more useful properties. To keep from repeating their proofs, which do not change, we will prove

1.6.2 lemma

If (i) and (iii) hold then we have:

(iv) If $\varphi \le \psi$ a.e. then $\int \varphi\, d\mu \le \int \psi\, d\mu$.

(v) If $\varphi = \psi$ a.e. then $\int \varphi\, d\mu = \int \psi\, d\mu$.

If, in addition, (ii) holds when $a = -1$ we have:

(vi) $\left|\int \varphi\, d\mu\right| \le \int |\varphi|\, d\mu$

Proof: By (iii), $\int \psi\, d\mu = \int \varphi\, d\mu + \int (\psi - \varphi)\, d\mu$ and the second integral is ≥ 0 by (i), so (iv) holds. $\varphi = \psi$ a.e. implies $\varphi \le \psi$ a.e. and $\psi \le \varphi$ a.e. so (v) follows from two applications of (iv). To prove (vi) now, notice that $\varphi \le |\varphi|$ so (iv) implies $\int \varphi\, d\mu \le \int |\varphi|\, d\mu$. $-\varphi \le |\varphi|$, so (iv) and (ii) imply $-\int \varphi\, d\mu \le \int |\varphi|\, d\mu$. Since $|\psi| = \max(y, -y)$, the result follows.

Step 2. Let E be a set with $\mu(E) < \infty$ and let f be a bounded function that vanishes on E^c. To define the integral of f, we observe that if φ, ψ are simple functions that have $\varphi \le f \le \psi$, then we want to have

$$\int \varphi\, d\mu \le \int f\, d\mu \le \int \psi\, d\mu$$

so we let

$$\int f\, d\mu = \sup_{\varphi \le f} \int \varphi\, d\mu = \inf_{\psi \ge f} \int \psi\, d\mu \tag{1.6.1}$$

Here and for the rest of Step 2, we assume that φ and ψ vanish on E^c. To justify the definition, we have to prove that the sup and inf are equal. It follows from (iv) in Lemma 1.4.2 that

$$\sup_{\varphi \le f} \int \varphi\, d\mu \le \inf_{\psi \ge f} \int \psi\, d\mu$$

To prove the other inequality, suppose $|f| \le M$ and let

$$E_k = \left\{ x \in E : \frac{kM}{n} \ge f(x) > \frac{(k-1)M}{n} \right\} \quad \text{for } -n \le k \le n$$

$$\psi_n(x) = \sum_{k=-n}^{n} \frac{kM}{n} \mathbf{1}_{E_k}, \quad \varphi_n(x) = \sum_{k=-n}^{n} \frac{(k-1)M}{n} \mathbf{1}_{E_k}$$

By definition, $\psi_n(x) - \varphi_n(x) = (M/n)\mathbf{1}_E$, so

$$\int (\psi_n(x) - \varphi_n(x))\, d\mu = \frac{M}{n}\mu(E)$$

Since $\varphi_n(x) \le f(x) \le \psi_n(x)$, it follows from (iii) in Lemma 1.4.1 that

$$\sup_{\varphi \le f} \int \varphi\, d\mu \ge \int \varphi_n\, d\mu = -\frac{M}{n}\mu(E) + \int \psi_n\, d\mu$$

$$\ge -\frac{M}{n}\mu(E) + \inf_{\psi \ge f} \int \psi\, d\mu$$

The last inequality holds for all n, so the proof is complete.

1.6.3 lemma

Let E be a set with $\mu(E) < \infty$. If f and g are bounded functions that vanish on E^c, then:

(i) If $f \ge 0$ a.e. then $\int f\, d\mu \ge 0$.

(ii) For any $a \in \mathbb{R}$, $\int af\, d\mu = a \int f\, d\mu$.

(iii) $\int f + g\, d\mu = \int f\, d\mu + \int g\, d\mu$.

(iv) If $g \le f$ a.e. then $\int g\, d\mu \le \int f\, d\mu$.

(v) If $g = f$ a.e. then $\int g\, d\mu = \int f\, d\mu$.

(vi) $\left|\int f\, d\mu\right| \le \int |f|\, d\mu$.

Proof: Since we can take $\varphi \equiv 0$, (i) is clear from the definition. To prove (ii), we observe that if $a > 0$, then $a\varphi \le af$ if and only if $\varphi \le f$, so

$$\int af\, d\mu = \sup_{\varphi \le f} \int a\varphi\, d\mu = \sup_{\varphi \le f} a \int \varphi\, d\mu = a \sup_{\varphi \le f} \int \varphi\, d\mu = a \int f\, d\mu.$$

For $a < 0$, we observe that $a\varphi \le af$ if and only if $\varphi \ge f$, so

$$\int af\, d\mu = \sup_{\varphi \ge f} \int a\varphi\, d\mu = \sup_{\varphi \ge f} a \int \varphi\, d\mu = a \inf_{\varphi \ge f} \int \varphi\, d\mu = a \int f\, d\mu.$$

To prove (iii), we observe that if $\psi_1 \ge f$ and $\psi_2 \ge g$, then $\psi_1 + \psi_2 \ge f + g$ so

$$\inf_{\psi \ge f+g} \int \psi\, d\mu \le \inf_{\substack{\psi_1 \ge f \\ \psi_2 \ge g}} \int (\psi_1 + \psi_2)\, d\mu.$$

Using linearity for simple functions, it follows that

$$\int f+g\,d\mu = \inf_{\psi\geq f+g}\int \psi\,d\mu \leq \inf_{\substack{\psi_1\geq f\\ \psi_2\geq g}}\int \psi_1\,d\mu + \int \psi_2\,d\mu = \int f\,d\mu + \int g\,d\mu.$$

To prove the other inequality, observe that the last conclusion applied to $-f$ and $-g$ and (ii) implies

$$-\int f+g\,d\mu \leq -\int f\,d\mu - \int g\,d\mu.$$

(iv)–(vi) follow from (i)–(iii) by Lemma 1.6.2.

Notation. We define the integral of f over the set E:

$$\int_E f\,d\mu \equiv \int f\cdot \mathbf{1}_E\,d\mu$$

Step 3. If $f\geq 0$ then we let

$$\int f\,d\mu = \sup\left\{\int h\,d\mu : 0\leq h\leq f, h \text{ is bounded and } \mu(\{x:h(x)>0\})<\infty\right\}$$

The last definition is nice since it is clear that this is well defined. The next result will help us compute the value of the integral.

1.6.4 lemma

Let $E_n\uparrow\Omega$ have $\mu(E_n)<\infty$ and let $a\wedge b=\min(a,b)$. Then

$$\int_\Omega f\wedge n\,d\mu = \lim_{n\to\infty}\int_{E_n} f\wedge n\,d\mu$$

Proof: It is clear that from (iv) in Lemma 1.6.3 that the left-hand side increases as n does. Since $h=(f\wedge n)\mathbf{1}_{E_n}$ is a possibility in the sup, each term is smaller than the integral on the right. To prove that the limit is $\int f\,d\mu$, observe that if $0\leq h\leq f$, $h\leq M$, and $\mu(\{x:h(x)>0\})<\infty$, then for $n\geq M$ using (i), (ii), and (iii),

$$\int_{E_n} f\wedge n\,d\mu \geq \int_{E_n} h\,d\mu = \int h\,d\mu - \int_{E_n^c} h\,d\mu$$

Now $0\leq\int_{E_n^c} h\,d\mu\leq M\mu(E_n^c\cap\{h(x)>0\})\to 0$ as $n\to\infty$, so

$$\liminf_{n\to\infty}\int_{E_n} f\wedge n\,d\mu \geq \int h\,d\mu$$

which proves the desired result since h is an arbitrary member of the class that defines the integral of f.

1.6.5 lemma

Suppose $f, g \geq 0$.

(i) $\int f\, d\mu \geq 0$

(ii) If $a > 0$ then $\int af\, d\mu = a \int f\, d\mu$

(iii) $\int f + g\, d\mu = \int f\, d\mu + \int g\, d\mu$

(iv) If $0 \leq g \leq f$ a.e. then $\int g\, d\mu \leq \int f\, d\mu$

(v) If $0 \leq g = f$ a.e. then $\int g\, d\mu = \int f\, d\mu$

Here we have dropped (vi) because it is trivial for $f \geq 0$.

Proof: (i) is trivial from the definition. (ii) is clear, since when $a > 0$, $ah \leq af$ if and only if $h \leq f$ and we have $\int ah\, d\mu = a \int h\, d\mu$ for h in the defining class.

For (iii), we observe that if $f \geq h$ and $g \geq k$, then $f + g \geq h + k$ so taking the sup over h and k in the defining classes for f and g gives

$$\int f + g\, d\mu \geq \int f\, d\mu + \int g\, d\mu$$

To prove the other direction, we observe $(a + b) \wedge n \leq (a \wedge n) + (b \wedge n)$ so (iv) from Lemma 1.6.3 and (iii) from Lemma 1.6.4 imply

$$\int_{E_n} (f + g) \wedge n\, d\mu \leq \int_{E_n} f \wedge n\, d\mu + \int_{E_n} g \wedge n\, d\mu$$

Letting $n \to \infty$ and using Lemma 1.6.4 gives (iii). As before, (iv) and (v) follow from (i), (iii), and Lemma 1.6.2.

Step 4. We say f is integrable if $\int |f|\, d\mu < \infty$. Let

$$f^+(x) = f(x) \vee 0 \quad \text{and} \quad f^-(x) = (-f(x)) \vee 0$$

where $a \vee b = \max(a, b)$. Clearly,

$$f(x) = f^+(x) - f^-(x) \quad \text{and} \quad |f(x)| = f^+(x) + f^-(x)$$

We define the integral of f by

$$\int f\, d\mu = \int f^+\, d\mu - \int f^-\, d\mu$$

The right-hand side is well defined since $f^+, f^- \leq |f|$ and we have (iv) in Lemma 1.4.5. For the final time, we will prove our six properties. To do this, it is useful to know:

1.6.6 lemma

If $f = f_1 - f_2$ where $f_1, f_2 \geq 0$ and $\int f_i \, d\mu < \infty$, then

$$\int f \, d\mu = \int f_1 \, d\mu - \int f_2 \, d\mu$$

Proof: $f_1 + f^- = f_2 + f^+$ and all four functions are ≥ 0, so by (iii) of Lemma 1.6.5,

$$\int f_1 \, d\mu + \int f^- \, d\mu = \int f_1 + f^- \, d\mu = \int f_2 + f^+ \, d\mu = \int f_2 \, d\mu + \int f^+ \, d\mu$$

Rearranging gives the desired conclusion.

1.7 Product Measures, Fubini's Theorem

Let $(X, \mathcal{A}, \mu_1)$ and $(Y, \mathcal{B}, \mu_2)$ be two σ-finite measure spaces. Let

$$\Omega = X \times Y = \{(x, y) : x \in X, y \in Y\}$$

$$\mathcal{S} = \{A \times B : A \in \mathcal{A}, B \in \mathcal{B}\}$$

Sets in $\mathcal{S}$ are called *rectangles*. It is easy to see that $\mathcal{S}$ is a semi-algebra:

$$(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$$

$$(A \times B)^c = (A^c \times B) \cup (A \times B^c) \cup (A^c \times B^c)$$

Let $\mathcal{F} = \mathcal{A} \times \mathcal{B}$ be the σ-algebra generated by $\mathcal{S}$.

1.7.1 Theorem

There is a unique measure μ on $\mathcal{F}$ with

$$\mu(A \times B) = \mu_1(A)\mu_2(B)$$

Notation. μ is often denoted by $\mu_1 \times \mu_2$.

Using Theorem 1.7.1 and induction, it follows that if $(\Omega_i, \mathcal{F}_i, \mu_i)$, $i = 1, \ldots, n$, are σ-finite measure spaces and $\Omega = \Omega_1 \times \cdots \times \Omega_n$, there is a unique measure

μ on the σ-algebra $\mathcal{F}$ generated by sets of the form $A_1 \times \cdots \times A_n$, $A_i \in \mathcal{F}_i$, that has

$$\mu(A_1 \times \cdots \times A_n) = \prod_{m=1}^{n} \mu_m(A_m)$$

When $(\Omega_i, \mathcal{F}_i, \mu_i) = (\mathbb{R}, \mathcal{R}, \lambda)$ for all i, the result is Lebesgue measure on the Borel subsets of n-dimensional Euclidean space $\mathbb{R}^n$.

Returning to the case in which $(\Omega, \mathcal{F}, \mu)$ is the product of two measure spaces, $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$, our next goal is to prove:

1.7.2 Theorem, Fubini's Theorem

If $f \geq 0$ or $|f|$ is μ-integrable, then

$$\int_X \left[\int_Y f(x,y)\, \mu_2(dy) \right] \mu_1(dx) = \int_{X \times Y} f \, d\mu = \int_Y \left[\int_X f(x,y)\, \mu_1(dx) \right] \mu_2(dy)$$

Proof: We will prove only the first equality, since the second follows by symmetry. Two technical things that need to be proved before we can assert that the first integral makes sense are:

- When x is fixed, $y \mapsto f(x,y)$ is $\mathcal{B}$-measurable.
- $x \mapsto \int_Y f(x,y)\mu_2(dy)$ is $\mathcal{A}$-measurable.

We begin with the case $f = \mathbf{1}_E$. Let

$$E_x = \{y : (x,y) \in E\}$$

be the **cross-section** at x.

Chapter 2
Laws of Large Numbers

2.1 Independence

Two events A and B are **independent** if $P(A \cap B) = P(A)P(B)$.

Two random variables X and Y are **independent** if for all $C, D \in \mathcal{R}$,

$$P(X \in C, Y \in D) = P(X \in C)P(Y \in D)$$

i.e., the events $A = \{X \in C\}$ and $B = \{Y \in D\}$ are independent.

Two σ-fields $\mathcal{F}$ and $\mathcal{G}$ are **independent** if for all $A \in \mathcal{F}$ and $B \in \mathcal{G}$ the events A and B are independent.
As the next result shows, the second definition is a special case of the third.

2.1.1 Theorem

(i) If X and Y are independent then $\sigma(X)$ and $\sigma(Y)$ are.

(ii) Conversely, if $\mathcal{F}$ and $\mathcal{G}$ are independent, $X \in \mathcal{F}$, and $Y \in \mathcal{G}$, then X and Y are independent.

Proof

(i) If $A \in \sigma(X)$ then it follows from the definition of $\sigma(X)$ that $A = \{X \in C\}$ for some $C \in \mathcal{R}$. Likewise, if $B \in \sigma(Y)$ then $B = \{Y \in D\}$ for some $D \in \mathcal{R}$, so using these facts and the independence of X and Y,

$$P(A \cap B) = P(X \in C, Y \in D) = P(X \in C)P(Y \in D) = P(A)P(B)$$

(ii) Conversely, if $X \in \mathcal{F}$, $Y \in \mathcal{G}$ and $C, D \in \mathcal{R}$ it follows from the definition of measurability that $\{X \in C\} \in \mathcal{F}$, $\{Y \in D\} \in \mathcal{G}$. Since $\mathcal{F}$ and $\mathcal{G}$ are independent, it follows that

$$P(X \in C, Y \in D) = P(X \in C)P(Y \in D)$$

The first definition is, in turn, a special case of the second.

2.1.2 Theorem

(i) If A and B are independent then so are A^c and B, A and B^c, and A^c and B^c. (ii) Conversely, events A and B are independent if and only if their indicator random variables 1_A and 1_B are independent.

Proof

(i) Subtracting $P(A \cap B) = P(A)P(B)$ from $P(B) = P(B)$ shows

$$P(A^c \cap B) = P(A^c)P(B).$$

The second and third conclusions follow by applying the first one to the pairs of independent events (B, A) and (A, B^c).

(ii) If $C, D \in \mathcal{R}$ then $\{1_A \in C\} \in \{\emptyset, A, A^c, \Omega\}$ and $\{1_B \in D\} \in \{\emptyset, B, B^c, \Omega\}$, so there are 16 things to check. When either set involved is $\emptyset$ or Ω, the equality holds, so there are only four cases to worry about and they are all covered by (i).

In view of the fact that the first definition is a special case of the second, which is a special case of the third, we take things in the opposite order when we say what it means for several things to be independent. We begin by reducing to the case of finitely many objects. An infinite collection of objects (σ-fields, random variables, or sets) is said to be independent if every finite subcollection is.

σ-fields $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n$ are **independent** if whenever $A_i \in \mathcal{F}_i$ for $i = 1, \ldots, n$, we have

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i)$$

Random variables $X_1, \ldots, X_n$ are **independent** if whenever $B_i \in \mathcal{R}$ for $i = 1, \ldots, n$ we have

$$P\left(\bigcap_{i=1}^{n} \{X_i \in B_i\}\right) = \prod_{i=1}^{n} P(X_i \in B_i)$$

Sets $A_1, \ldots, A_n$ are **independent** if whenever $I \subset \{1, \ldots, n\}$ we have

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i)$$

At first glance, it might seem that the last definition does not match the other two. However, if you think about it for a minute, you will see that if the indicator variables $1_{A_i},\ 1 \leq i \leq n$ are independent and we take $B_i = \{1\}$ for $i \in I$, and $B_i = \mathbb{R}$ for $i \notin I$, then the condition in the definition results. Conversely,

2.1.3 Theorem

Let $A_1, A_2, \ldots, A_n$ be independent.

(i) $A_i^c, A_2, \ldots, A_n$ are independent.

(ii) $1_{A_1}, \ldots, 1_{A_n}$ are independent.

Proof

(i) Let $B_1 = A_1^c$ and $B_i = A_i$ for $i > 1$. If $I \subset \{1, \ldots, n\}$ does not contain 1 it is clear that

$$P\left(\bigcap_{i \in I} B_i\right) = \prod_{i \in I} P(B_i).$$

Suppose now that $1 \in I$ and let $J = I \setminus \{1\}$. Subtracting $P(\cap_{i \in I} A_i) = \prod_{i \in I} P(A_i)$ from $P(\cap_{i \in J} A_i) = \prod_{i \in J} P(A_i)$ gives $P(A_1 \cap \cap_{i \in J} A_i) = P(A_1^c) \prod_{i \in J} P(A_i)$.

(ii) Iterating (i) we see that if $B_i \in \{A_i, A_i^c\}$ then $B_1, \ldots, B_n$ are independent. Thus if $C_i \in \{A_i, A_i^c, \Omega\}$, $P(\cap_{i=1}^n C_i) = \prod_{i=1}^n P(C_i)$. The last equality holds trivially if some $C_i = \emptyset$, so noting $1_{A_i} \in \{\emptyset, A_i, A_i^c, \Omega\}$, the desired result follows.

One of the first things to understand about the definition of independent events is that it is not enough to assume $P(A_i \cap A_j) = P(A_i)P(A_j)$ for all $i \neq j$. A sequence of events $A_1, \ldots, A_n$ with the last property is called **pairwise independent**. It is clear that independent events are pairwise independent. The next example shows that the converse is not true.

Example: Let X_1, X_2, X_3 be independent random variables with

$$P(X_i = 0) = P(X_i = 1) = \frac{1}{2}$$

Let $A_1 = \{X_2 = X_3\},\ A_2 = \{X_3 = X_1\},\ A_3 = \{X_1 = X_2\}$. These events are pairwise independent since if $i \neq j$, then

$$P(A_i \cap A_j) = P(X_1 = X_2 = X_3) = \frac{1}{4} = P(A_i)P(A_j)$$

but they are not independent since

$$P(A_1 \cap A_2 \cap A_3) = \frac{1}{4} \neq \frac{1}{8} = P(A_1)P(A_2)P(A_3)$$

In order to show that random variables X and Y are independent, we have to check that $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for all Borel sets A and B. Since there are a lot of Borel sets, our next topic is:

2.1.4 Sufficient Conditions for Independence

Our main result is Theorem 2.1.7. To state that result, we need a definition that generalizes all our earlier definitions.

Collections of sets $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n \subset \mathcal{F}$ are said to be **independent** if whenever $A_i \in \mathcal{A}_i$ and $I \subset \{1, \ldots, n\}$ we have

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i)$$

If each collection is a single set i.e., $\mathcal{A}_i = \{A_i\}$, this definition reduces to the one for sets.

2.1.5 lemma

Without loss of generality we can suppose each $\mathcal{A}_i$ contains Ω. In this case the condition is equivalent to

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i) \quad \text{whenever } A_i \in \mathcal{A}_i$$

since we can set $A_i = \Omega$ for $i \notin I$.

2.2 Expectation

Let X be a simple random variable on (Ω, F, P), taking the values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$. If the random experiment is repeated independently N times, X will take the value x_i roughly Np_i times, so the arithmetic average of the values of X in the N observations is roughly

$$\frac{1}{N}[Np_1x_1 + Np_2x_2 + \cdots + Np_nx_n] = \sum_{i=1}^{n} p_i x_i.$$

This is a reasonable figure for the average value of X. If X is represented as $\sum_{i=1}^{n} x_i I_{B_i}$, where the B_i are disjoint sets in F, the average value may be expressed as $\sum_{i=1}^{n} x_i P(B_i)$, which is $\int_\Omega X\, dP$. Since arbitrary random variables are ultimately built up from simple ones, it is reasonable to take $\int_\Omega X\, dP$ as the definition of the average value (henceforth to be called the "expectation") of X.

2.2.1 Definition.

If X is a random variable on (Ω, F, P), the *expectation* of X is defined by

$$E(X) = \int_\Omega X\, dP$$

provided the integral exists. Thus $E(X)$ is the integral of the Borel measurable function X with respect to the probability measure P, so that all the results of integration theory are applicable. The same definition is used if X is an extended random variable.

In many situations it is inconvenient to compute $E(X)$ by integrating over Ω; the following result expresses $E(X)$ as an integral with respect to the induced probability measure P_X, which in turn is determined by the distribution function F.

First, a word about notation. If F is a distribution function on $\mathbb{R}^n$ with corresponding Lebesgue–Stieltjes measure μ, and $g\colon (\mathbb{R}^n, B) \to (\mathbb{R}, B)$, then

$$\int_{\mathbb{R}} g(x)\, dF(x)$$

means $\int_{\mathbb{R}} g\, d\mu$; it is *not* a Riemann–Stieltjes integral.

2.2.2 Theorem.

Let X be a random variable on (Ω, F, P), with distribution function F. Let g be a Borel measurable function from $\mathbb{R}$ to $\mathbb{R}$.
If $Y = g \circ X$, then

$$E(Y) = \int_{\mathbb{R}} g(x)\, dF(x) \qquad \left(= \int_{\mathbb{R}} g\, dP_X \right)$$

in the sense that if either of the two sides exists, so does the other, and the two sides are equal.

Proof. We use the basic technique of starting with indicators and proceeding to more complicated functions.
Let g be an indicator I_B, $B \in B(\mathbb{R})$. Then

$$E(Y) = E(I_B \circ X) = E(I(X \in B)) = P_X(B) = \int_{\mathbb{R}} g\, dP_X$$

so that $E(Y)$ and $\int_{\mathbb{R}} g\, dP_X$ exist and are equal.
Now let g be a nonnegative simple function, say, $g(x) = \sum_{j=1}^{n} x_j I_{B_j}(x)$, the B_j disjoint sets in $B(\mathbb{R})$. Then

$$E(Y) = \sum_{j=1}^{n} x_j E(I_{B_j} \circ X) = \sum_{j=1}^{n} x_j \int_{\mathbb{R}} I_{B_j}\, dP_X \quad \text{by what we have just proved}$$

$$= \int_{\mathbb{R}} \left(\sum_{j=1}^{n} x_j I_{B_j} \right) dP_X \quad \text{since } g \geq 0$$

$$= \int_{\mathbb{R}} g\, dP_X.$$

Again, both integrals exist and are equal.
If g is a nonnegative Borel measurable function, let $g_1, g_2, \ldots$ be nonnegative simple functions with $g_n \uparrow g$. We have just proved that

$$E(g_n \circ X) = \int_{\mathbb{R}} g_n\, dP_X;$$

hence by the monotone convergence theorem,

$$E(g \circ X) = \int_{\mathbb{R}} g\, dP_X,$$

and again both integrals exist and are equal.

Finally, if $g = g^+ - g^-$ is an arbitrary Borel measurable function and $Y = g \circ X$, we have

$$E(Y) = E(Y^+) - E(Y^-) = E(g^+ \circ X) - E(g^- \circ X)$$

$$= \int_{\mathbb{R}} g^+ \, dP_X - \int_{\mathbb{R}} g^- \, dP_X \quad \text{by what we have already proved}$$

$$= \int_{\mathbb{R}} g \, dP_X.$$

If $E(Y)$ exists and, say, $E(Y^-)$ is finite, then $\int_{\mathbb{R}} g^- \, dP_X$ is finite, and hence $\int_{\mathbb{R}} g \, dP_X$ exists; by the same reasoning, the existence of $\int_{\mathbb{R}} g \, dP_X$ implies that of $E(Y)$.

2.2.3 Corollaries and Extensions.

(a) Let X be a random vector on (Ω, F, P), and let g be a Borel measurable function from $\mathbb{R}^n$ to $\mathbb{R}$. Then

$$E(g \circ X) = \int_{\mathbb{R}^n} g(x) \, dF(x)$$

in the sense that if either integral exists, so does the other, and the two are equal.

The proof is exactly as in 1.7.2, with $\mathbb{R}$ replaced by $\mathbb{R}^n$.

(b) More generally, let X be a random object on (Ω, F, P), that is,

$$X : (\Omega, F) \to (\Omega', F')$$

where (Ω', F') is an arbitrary measurable space. Let g be a Borel measurable real (or extended real) valued function on (Ω', F'). Let P_X be the probability measure induced by X:

$$P_X(B) = P(\omega : X(\omega) \in B), \qquad B \in F'.$$

Then

$$E(g \circ X) = \int_{\Omega'} g \, dP_X$$

in the sense that if either integral exists, so does the other, and the two are equal.

Again, the proof is just as in 2.2.2, with $\mathbb{R}$ replaced by Ω' and $B(\mathbb{R})$ by F'.

(c) If X is a random variable (or random vector) with density f, then

$$\int g(x)\,dF(x) = \int g(x)f(x)\,dx$$

(integration over $\mathbb{R}$ in the case of a random variable, and over $\mathbb{R}^n$ in the case of a random vector) in the sense that if either integral exists, then so does the other, and the two are equal.

When g is an indicator I_B, this says that $P_X(B) = \int_B f(x)\,dx$, which holds for any Borel set B. The proof is completed by passing in turn to nonnegative simple functions, nonnegative Borel measurable functions, and arbitrary Borel measurable functions.

(d) If X is a discrete random variable with probability function p, then

$$\int g(x)\,dF(x) = \sum_x g(x)p(x),$$

where the series is interpreted as $\sum_x g^+(x)p(x) - \sum_x g^-(x)p(x)$, and again the interpretation is that the integral exists iff the sum exists (that is,

$$\sum_x g^+(x)p(x) < \infty \qquad \text{or} \qquad \sum_x g^-(x)p(x) < \infty),$$

and in this case the two are equal.

This is proved by starting with indicators as before.

Expectations of certain functions of X are of special interest.

2.2.4 Definition.

Let X be a random variable on (Ω, F, P). If $k > 0$, the number $E(X^k)$ is called the k**th moment** of X; $E[|X|^k]$ is called the k**th absolute moment** of X. $E[(X - E(X))^k]$ is called the k**th central moment**; $E[|X - E(X)|^k]$ the k**th absolute central moment**; central moments are defined only when $E(X)$ is finite.

The first moment ($k = 1$) is $E(X)$, sometimes called the **mean** of X, and the first central moment (if it exists) is always 0. The second central moment

$$\sigma^2 = E[(X - E(X))^2]$$

is called the **variance** of X, sometimes written $\operatorname{Var} X$, and the positive square root σ the **standard deviation**.

Note that $E(X^k)$ is finite iff $E[|X|^k]$ is finite, by 1.6.4(b). Also, finiteness of the kth moment implies finiteness of lower moments, as we now prove.

2.2.5 Lemma.

If $k > 0$ and $E(X^k)$ are finite, then $E(X^j)$ is finite for $0 < j < k$.

First Proof.

$$E(|X|^j) = \int_\Omega |X|^j dP = \int_{\{|X|<1\}} |X|^j dP + \int_{\{|X|\geq 1\}} |X|^j dP$$

$$\leq P(|X| < 1) + \int_\Omega |X|^k dP < \infty. \quad \square$$

Second Proof. We have $\|X\|_j \leq \|X\|_k$ for $0 < j < k$.
Central moments of integral order can be obtained from moments, as follows.

2.2.6 Lemma.

If n is a positive integer greater than 1, $E(X^{n-1})$ is finite, and $E(X^n)$ exists, then

$$E[(X - E(X))^n] = \sum_{k=0}^{n} \binom{n}{k} (-1)^k E(X)^{n-k} E(X^k).$$

In particular, if $E(X)$ is finite ($E(X^2)$ always exists since $X^2 \geq 0$), then

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

Proof. Use the binomial theorem and the additivity theorem for integrals.

A similar formula expresses moments in terms of central moments. (Write $X^n = (X - E(X) + E(X))^n$ and use the binomial theorem.)
We now restate a result required for Measure-theoretic context.

2.2.7 Chebyshev's Inequality.

(a) If X is a nonnegative random variable, $0 < p < \infty$ and $0 < \varepsilon < \infty$,

$$P(X \geq \varepsilon) \leq \frac{E(X^p)}{\varepsilon^p}.$$

(b) If X is a random variable with finite mean m and variance σ^2, and $0 < k < \infty$,

$$P(|X - m| \geq k\sigma) \leq \frac{1}{k^2}.$$

This is a quantitative result to the effect that a random variable with small variance is likely to be close to its mean.

A normally distributed random variable has the useful property that the distribution is completely determined by the mean and variance. Specifically, if X has the normal density, that is,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right],$$

then $m = E(X)$ and $\sigma^2 = \operatorname{Var} X$; the computation is straightforward, using the standard integrals

$$\int_{-\infty}^{\infty} \exp(-x^2)\, dx = \sqrt{\pi}, \quad \text{and} \quad \int_{-\infty}^{\infty} x^2 \exp(-x^2)\, dx = \frac{1}{2}\sqrt{\pi}.$$

The phrase "normal (m, σ^2)" is used for a random variable that is normally distributed with mean m and variance σ^2.

The following result on the expectation of a product of independent random variables is a direct consequence of Fubini's theorem.

2.2.8 Theorem.

Let $X_1, \ldots, X_n$ be independent random variables on (Ω, F, P). If all X_i are nonnegative or if $E(X_i)$ is finite for all i, then $E(X_1 \cdots X_n)$ exists and equals

$$E(X_1)E(X_2)\cdots E(X_n).$$

Proof. If all $X_i \geq 0$, then by 2.2.3(a),

$$E(X_1 \cdots X_n) = \int_{\mathbb{R}} x_1 \cdots x_n \, dP_X(x_1, \ldots, x_n) \quad \text{where } X = (X_1, \ldots, X_n).$$

Since P_X is the product of the P_{X_i}, Fubini's theorem yields

$$E(X_1 \cdots X_n) = \int_{\mathbb{R}} x_1 dP_{X_1}(x_1) \cdots \int_{\mathbb{R}} x_n dP_{X_n}(x_n) = E(X_1) \cdots E(X_n).$$

(This can also be proved without Fubini's theorem by starting with indicators and proceeding to nonnegative simple functions and then nonnegative measurable functions, but the present proof is faster. Note also that the result holds for extended random variables, with the same proof.)

If all $E(X_i)$ are finite, the above argument shows that

$$E(|X_1 \cdots X_n|) = \prod_{i=1}^{n} E(|X_i|) < \infty,$$

and thus Fubini's theorem may be applied just as in the first part of the proof.

2.2.9 Theorem.

If $X_1, \ldots, X_n$ are independent complex-valued random variables and $E(X_i)$ is finite for all i, then $E(X_1 \cdots X_n)$ is finite and equals

$$E(X_1) \cdots E(X_n).$$

Proof. First, let $n = 2$, $X_1 = Y_1 + iZ_1$, $X_2 = Y_2 + iZ_2$. By 4.8.2(d), Y_1 and Y_2 are independent, as are Y_1 and Z_2, Z_1 and Y_2, and Z_1 and Z_2. By 4.10.8,

$$\begin{aligned} E(X_1X_2) &= E(Y_1)E(Y_2) - E(Z_1)E(Z_2) + iE(Y_1)E(Z_2) + iE(Z_1)E(Y_2) \\ &= (E(Y_1) + iE(Z_1))(E(Y_2) + iE(Z_2)) = E(X_1)E(X_2). \end{aligned}$$

Now let $n > 2$, $X_j = Y_j + iZ_j$, $j = 1, \ldots, n$, and assume the result has been established for $n-1$ random variables. If $V = (Y_1, Z_1, Y_2, Z_2, \ldots, Y_{n-1}, Z_{n-1})$ and $W = (Y_n, Z_n)$, we claim that V and W are independent. By independence of $X_1, \ldots, X_n$, $P(V \in A, W \in B) = P(V \in A)P(W \in B)$ when A and B are measurable rectangles. Pass from measurable rectangles to finite disjoint unions of measurable rectangles, and then, by means of the monotone class theorem, to arbitrary Borel sets.

The independence of V and W implies that $X_1 \cdots X_{n-1}$ and X_n are independent, so that

$$E(X_1 \cdots X_n) = E(X_1 \cdots X_{n-1})E(X_n) = E(X_1) \cdots E(X_n)$$

by the induction hypothesis.

2.3 Weak Laws of Large Numbers

In this section, we will prove several "weak laws of large numbers." The first order of business is to define the mode of convergence that appears in the conclusions of the theorems. We say that Y_n converges to Y in **probability** if for all $\epsilon > 0$, $P(|Y_n - Y| > \epsilon) \to 0$ as $n \to \infty$.

2.3.1 L^2 Weak Laws

Our first set of weak laws come from computing variances and using Chebyshev's inequality. Extending a definition given in Example 2.1.14 for two random variables, a family of random variables X_i, $i \in I$ with $EX_i^2 < \infty$ is said to be **uncorrelated** if we have

$$E(X_iX_j) = EX_iEX_j \quad \text{whenever } i \neq j$$

The key to our weak law for uncorrelated random variables, is:

2.3.2 Theorem

Let $X_1, \ldots, X_n$ have $E(X_i^2) < \infty$ and be uncorrelated. Then

$$\text{var}(X_1 + \cdots + X_n) = \text{var}(X_1) + \cdots + \text{var}(X_n)$$

where $\text{var}(Y) =$ the variance of Y.

Proof: Let $\mu_i = EX_i$ and $S_n = \sum_{i=1}^n X_i$. Since $ES_n = \sum_{i=1}^n \mu_i$, using the definition of the variance, writing the square of the sum as the product

$$\begin{aligned}
\text{var}(S_n) = E(S_n - ES_n)^2 &= E\left(\sum_{i=1}^n (X_i - \mu_i)\right)^2 \\
&= E\left(\sum_{i=1}^n \sum_{j=1}^n (X_i - \mu_i)(X_j - \mu_j)\right) \\
&= \sum_{i=1}^n E(X_i - \mu_i)^2 + 2\sum_{i=1}^n \sum_{j=1}^{i-1} E((X_i - \mu_i)(X_j - \mu_j))
\end{aligned}$$

where in the last equality we have separated out the diagonal terms $i = j$ and used the fact that the sum over $1 \le i < j \le n$ is the same as the sum over $1 \le j < i \le n$.

The first sum is $\text{var}(X_1) + \cdots + \text{var}(X_n)$ so we want to show that the second sum is zero. To do this, we observe

$$\begin{aligned}
E((X_i - \mu_i)(X_j - \mu_j)) &= EX_iX_j - \mu_i EX_j - \mu_j EX_i + \mu_i\mu_j \\
&= EX_iX_j - \mu_i\mu_j = 0
\end{aligned}$$

since X_i and X_j are uncorrelated.

In words, Theorem 2.3.2 says that for uncorrelated random variables the variance of the sum is the sum of the variances. The second ingredient in our proof of Theorem 2.2.3 is the following consequence of (1.6.4):

$$\text{var}(cY) = c^2 \,\text{var}(Y)$$

The third and final ingredient is

2.3.3 lemma

If $p > 0$ and $E|Z_n|^p \to 0$ then $Z_n \to 0$ in probability.

Proof: Chebyshev's inequality, with $\varphi(x) = x^p$ and $X = |Z_n|$ implies that if $\varepsilon > 0$ then $P(|Z_n| > \varepsilon) \le \varepsilon^{-p} E|Z_n|^p \to 0$.

We can now easily prove

2.3.4 Theorem L^2 weak law

Let $X_1, X_2, \ldots$ be uncorrelated random variables with $EX_i = \mu$ and $\text{var}(X_i) \leq C < \infty$. If $S_n = X_1 + \cdots + X_n$ then as $n \to \infty$, $S_n/n \to \mu$ in L^2 and in probability.

Proof: To prove L^2 convergence, observe that $E(S_n/n) = \mu$, so

$$E(S_n/n - \mu)^2 = \text{var}(S_n/n) = \frac{1}{n^2}\left(\text{var}(X_1) + \cdots + \text{var}(X_n)\right) \leq \frac{Cn}{n^2} \to 0$$

To conclude there is also convergence in probability, we apply Lemma 2.3.3 to $Z_n = S_n/n - \mu$.

The most important special case of Theorem 2.3.4 occurs when $X_1, X_2, \ldots$ are independent random variables that all have the same distribution. In the jargon, they are **independent and identically distributed** or **i.i.d.** for short. Theorem 2.3.4 tells us in this case that if $EX_i^2 < \infty$ then S_n/n converges to $\mu = EX_i$ in probability as $n \to \infty$.

2.4 Borel-Cantelli Lemmas

If A_n is a sequence of subsets of Ω, we let

$$\limsup A_n = \lim_{m\to\infty} \bigcup_{n=m}^{\infty} A_n = \{\omega \text{ that are in infinitely many } A_n\}$$

(the limit exists since the sequence is decreasing in m) and let

$$\liminf A_n = \lim_{m\to\infty} \bigcap_{n=m}^{\infty} A_n = \{\omega \text{ that are in all but finitely many } A_n\}$$

$$\limsup_{n\to\infty} A_n = \lim_{m\to\infty} \bigcup_{n=m}^{\infty} A_n, \quad \liminf_{n\to\infty} A_n = \lim_{m\to\infty} \bigcap_{n=m}^{\infty} A_n$$

It is common to write $\limsup A_n = \{\omega : \omega \in A_n \text{ i.o.}\}$, where i.o. stands for infinitely often. An example which illustrates the use of this notation is: "$X_n \to 0$ a.s. if and only if for all $\varepsilon > 0$, $P(|X_n| > \varepsilon \text{ i.o.}) = 0$."

2.4.1 Theorem, First Borel-Cantelli Lemma

If $\sum_{n=1}^{\infty} P(A_n) < \infty$ then

$$P\left(\limsup A_n\right) = 0.$$

Proof: Let $N = \sum 1_{A_n}$ be the number of events that occur. Fubini's theorem implies $\mathbb{E}N = \sum P(A_k) < \infty$, so we must have $N < \infty$ a.s.

2.4.2 Theorem, Second Borel–Cantelli Lemma

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $A_1, A_2, \ldots$ be independent events in $\mathcal{F}$. If $\sum_{n=1}^{\infty} P(A_n) = \infty$, then

$$P\left(\limsup A_n\right) = 1.$$

Proof:

$$P\left(\limsup_n A_n\right) = P\left(\bigcap_n \bigcup_{k\geq n} A_k\right) = \lim_{n\to\infty} P\left(\bigcup_{k=n}^{\infty} A_k\right) = \lim_{n\to\infty} \lim_{m\to\infty} P\left(\bigcup_{k=n}^{m} A_k\right).$$

Now

$$P\left(\bigcup_{k=n}^{m} A_k\right)^c = P\left(\bigcap_{k=n}^{m} A_k^c\right) = \prod_{k=n}^{m} P(A_k^c) \quad \text{by independence}$$

$$\leq \prod_{k=n}^{m} \exp[-P(A_k)] \quad \text{since } P(A_k^c) = 1 - P(A_k) \leq \exp[-P(A_k)]$$

$$\to 0 \quad \text{as } m \to \infty \quad \text{since } \sum P(A_k) = \infty.$$

2.5 Convergence Theorems

2.5.1 Theorem.

Let $X_1, X_2, \ldots$ be independent random variables with finite expectation. If $\sum_{n=1}^{\infty} \operatorname{Var} X_n < \infty$, then $\sum_{n=1}^{\infty} [X_n - E(X_n)]$ converges a.e.

[All random variables are assumed to be defined on a fixed probability space $(\Omega, \mathcal{F}, P)$, and "almost everywhere" refers to the probability measure P. Also, throughout this chapter, convergence will always mean to a **finite** limit.]

Proof: We may assume that $E(X_n) \equiv 0$. Let $S_n = X_1 + \cdots + X_n$. Then S_n converges iff $S_j - S_k \to 0$ as $j, k \to \infty$, and this happens a.e. iff for each $\varepsilon > 0$,

$$P\left(\bigcup_{j,k \geq n} \{|S_j - S_k| \geq \varepsilon\}\right) \to 0 \quad \text{as } n \to \infty$$

Equivalently, we must prove that for each $\varepsilon > 0$,

$$P\left(\bigcup_{k=1}^{\infty} \{|S_{m+k} - S_m| \geq \varepsilon\}\right) \to 0 \quad \text{as } m \to \infty.$$

We have

$$P\left(\bigcup_{k=1}^{\infty} \{|S_{m+k} - S_m| \geq \varepsilon\}\right) = \lim_{n \to \infty} P\left(\bigcup_{k=1}^{n} \{|S_{m+k} - S_m| \geq \varepsilon\}\right)$$

$$= \lim_{n \to \infty} P\left(\max_{1 \leq k \leq n} |S_{m+k} - S_m| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} \limsup_{n \to \infty} \operatorname{Var}(S_{m+n} - S_m) \quad \text{by 6.1.4}$$

$$= \frac{1}{\varepsilon^2} \limsup_{n \to \infty} \sum_{j=1}^{n} \operatorname{Var}(X_{m+j}) \quad \to 0 \quad \text{as } m \to \infty \quad \text{since } \sum_{n=1}^{\infty} \operatorname{Var} X_n < \infty.$$

2.5.2 Kolmogorov Strong Law of Large Numbers.

Let $X_1, X_2, \ldots$ be independent random variables, each with finite mean and variance, and let $\{b_n\}$ be an increasing sequence of positive real numbers with $b_n \to \infty$. If

$$\sum_{n=1}^{\infty} \frac{\operatorname{Var} X_n}{b_n^2} < \infty,$$

then (with $S_n = X_1 + \cdots + X_n$)

$$\frac{S_n - E(S_n)}{b_n} \to 0 \quad \text{a.e.}$$

Proof.

$$\sum_{n=1}^{\infty} \operatorname{Var}\left(\frac{X_n - E(X_n)}{b_n}\right) = \sum_{n=1}^{\infty} \frac{\operatorname{Var} X_n}{b_n^2} < \infty \quad \text{by hypothesis.}$$

By 2.5.1, $\sum_{n=1}^{\infty} (X_n - E(X_n))/b_n$ converges a.e. But

$$\frac{S_n - E(S_n)}{b_n} = \frac{1}{b_n} \sum_{k=1}^{n} b_k \left(\frac{X_k - E(X_k)}{b_k}\right),$$

and this approaches zero a.e. by Kronecker lemma.

2.5.3 Lemma

If Y is a nonnegative random variable,

$$\sum_{n=1}^{\infty} P\{Y \geq n\} \leq E(Y) \leq 1 + \sum_{n=1}^{\infty} P\{Y \geq n\}.$$

Proof:

$$\sum_{n=1}^{\infty} P(Y \geq n) = \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} P(k \leq Y < k+1) = \sum_{k=1}^{\infty} \sum_{n=1}^{k} P(k \leq Y < k+1)$$

$$= \sum_{k=1}^{\infty} kP(k \leq Y < k+1) = \sum_{k=0}^{\infty} \int_{\{k \leq Y < k+1\}} k \, dP$$

$$\leq \sum_{k=0}^{\infty} \int_{\{k \leq Y < k+1\}} Y \, dP = E(Y) \leq \sum_{k=0}^{\infty} (k+1) P(k \leq Y < k+1)$$

$$= \sum_{n=1}^{\infty} P(Y \geq n) + \sum_{k=0}^{\infty} P(k \leq Y < k+1) = \sum_{n=1}^{\infty} P(Y \geq n) + 1.$$

2.5.4 Strong Law of Large Numbers, iid Case

Let $X_1, X_2, \ldots$ be iid random variables with finite expectation m, and let $S_n = X_1 + \cdots + X_n$, then

$$\frac{S_n}{n} \to m \quad \text{a.e.}$$

Proof: Since all X_n have the same distribution,

$$\sum_{n=1}^{\infty} P(|X_n| \geq n) = \sum_{n=1}^{\infty} P(|X_1| \geq n);$$

thus

$$\sum_{n=1}^{\infty} P(|X_n| \geq n) \leq E(|X_1|) < \infty \quad \text{by 2.5.3.}$$

By the Borel–Cantelli lemma, $P(|X_n| \geq n \text{ for infinitely many } n) = 0$. Define

$$Y_n = \begin{cases} X_n, & \text{if } |X_n| < n; \\ 0, & \text{if } |X_n| \geq n. \end{cases}$$

Then, except on a set of probability 0, $Y_n = X_n$ for sufficiently large n. Thus, assuming (without loss of generality) that $m = 0$, it suffices to show that

$$\frac{1}{n}\sum_{j=1}^{n} Y_j \to 0 \quad \text{a.e.}$$

Now,

$$E(Y_n) = E(X_n \mathbf{1}_{\{|X_n|<n\}}) = E(X_1 \mathbf{1}_{\{|X_1|<n\}})$$

by the iid hypothesis. Then

$$E(X_1 \mathbf{1}_{\{|X_1|<n\}}) \to E(X_1) = 0 \quad \text{as } n \to \infty$$

by the dominated convergence theorem. Consequently,

$$\frac{1}{n}\sum_{j=1}^{n} E(Y_j) \to 0,$$

and therefore it is sufficient to show that

$$\frac{1}{n}\sum_{j=1}^{n} [Y_j - E(Y_j)] \to 0 \quad \text{a.e.}$$

If we can show that

$$\sum_{n=1}^{\infty} \left[\frac{Y_n - E(Y_n)}{n}\right]$$

converges a.e., then the Kronecker lemma with $b_n = n$ and

$$x_n = \frac{Y_n - E(Y_n)}{n},$$

yields

$$\frac{1}{n}\sum_{j=1}^{n} [Y_j - E(Y_j)] \to 0 \quad \text{a.e.},$$

as desired.

Now the Y_n are functions of the independent random variables X_n, and hence are independent, so by 6.2.1, it suffices to show that

$$V = \sum_{n=1}^{\infty} \text{Var}\left(\frac{Y_n}{n}\right) < \infty.$$

(Note that $|Y_n| < n$, so $\text{Var}(Y_n)$ is finite, although nothing is known about $\text{Var}(X_n)$.)

But

$$V = \sum_{n=1}^{\infty} \frac{1}{n^2} \mathrm{Var}(Y_n) \leq \sum_{n=1}^{\infty} \frac{1}{n^2} E(Y_n^2)$$

(since $\mathrm{Var}(Y_n) = E(Y_n^2) - [E(Y_n)]^2$)

$$= \sum_{n=1}^{\infty} \frac{1}{n^2} E(X_1^2 \mathbf{1}_{\{|X_1|<n\}}) \quad \text{by the iid hypothesis}$$

$$= \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{m=1}^{n} E(X_1^2 \mathbf{1}_{\{m-1\leq|X_1|<m\}}) = \sum_{m=1}^{\infty} E(X_1^2 \mathbf{1}_{\{m-1\leq|X_1|<m\}}) \sum_{n=m}^{\infty} \frac{1}{n^2}.$$

By comparing $\sum(1/n^2)$ with $\int(1/x^2)\,dx$, we find that

$$\sum_{n=m}^{\infty} \frac{1}{n^2} \leq \frac{K}{m}$$

for some fixed positive constant K. Thus

$$V \leq K \sum_{m=1}^{\infty} \frac{1}{m} E(X_1^2 \mathbf{1}_{\{m-1\leq|X_1|<m\}}).$$

If $m - 1 \leq |X_1| < m$, then $X_1^2 = |X_1||X_1| \leq m|X_1|$; hence

$$V \leq K \sum_{m=1}^{\infty} E(|X_1| \mathbf{1}_{\{m-1\leq|X_1|<m\}}) = KE(|X_1|) < \infty.$$

Chapter 3
Conditional Probability and Expectation

3.1 Introduction

We know conditional probability $P(B|A)$ only when $P(A) > 0$. However, conditional probabilities given events of probability zero are in no sense degenerate cases; they occur naturally in many problems. For example, consider the following two-stage random experiment. A random variable X is observed, where X has distribution function F. If X takes the value x, a random variable Y is observed, where the distribution of Y depends on x. (For example, if $0 \leq x \leq 1$, a coin with probability of heads x might be tossed independently n times, with Y the resulting number of heads.) Thus $P(x, B) = P(Y \in B \mid X = x)$ is prescribed in the statement of the problem, although the event $\{X = x\}$ may have probability zero for all values of x.

Let us try to construct a model for the above situation. Let $\Omega = \mathbb{R}^2$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^2)$, $X(x, y) = x$, $Y(x, y) = y$. Instead of specifying the joint distribution function of X and Y, we specify the distribution function of X, and thus the corresponding probability measure P_X; also, for each x we are given a probability measure $P(x, \cdot)$ defined on $\mathcal{B}(\mathbb{R})$; $P(x, B)$ is interpreted (informally for now) as $P(Y \in B \mid X = x)$.

We claim that the probability of any event of the form $\{(X, Y) \in C\}$ is determined. Reasoning intuitively, the probability that X falls into $(x, x+dx]$ is $dF(x)$. Given that this occurs, in other words (roughly), given $X = x$, (X, Y) will lie in C iff Y belongs to the section $C(x) = \{y : (x, y) \in C\}$. The probability of this event is $P(x, C(x))$. The total probability that (X, Y) will belong to C is

$$P(C) = \int_{-\infty}^{\infty} P(x, C(x))\, dF(x). \tag{1}$$

In the special case $C = \{(x, y) : x \in A, y \in B\} = A \times B$, $C(x) = B$ if $x \in A$ and $C(x) = \emptyset$ if $x \notin A$; therefore

$$P(C) = P(A \times B) = \int_A P(x, B)\, dF(x). \tag{2}$$

Now if $P(x, B)$ is Borel measurable in x for each fixed $B \in \mathcal{B}(\mathbb{R})$, then by the product measure theorem, there is a unique (probability) measure on $\mathcal{B}(\mathbb{R}^2)$ satisfying (2) for all $A, B \in \mathcal{B}(\mathbb{R})$, namely, the measure given by (1). Thus in the mathematical formulation of the problem, we take the probability measure P on $\mathcal{F} = \mathcal{B}(\mathbb{R}^2)$ to be the unique measure determined by P_X and the measures $P(x, \cdot)$, $x \in \mathbb{R}$.

Example: Let X be uniformly distributed between 0 and 1. If $X = x$, a coin with probability x of heads is tossed independently n times. If Y is the resulting number of heads, find $P(Y = k)$, $k = 0, 1, \ldots, n$.
Let us translate this into mathematical terms. Let $\Omega_1 = [0, 1]$, $\mathcal{F}_1 = \mathcal{B}[0, 1]$. We have specified $P_X(A) = \int_A dx =$ Lebesgue measure of A, $A \in \mathcal{F}_1$.
For each x, we are given $P(x, B)$, to be interpreted as the conditional probability that $Y \in B$, given $X = x$. We may take $\Omega_2 = \{0, 1, \ldots, n\}$, $\mathcal{F}_2$ the class of all subsets of Ω_2; then $P(x, \{k\}) = \binom{n}{k} x^k (1 - x)^{n-k}$, $k = 0, 1, \ldots, n$ (this is Borel measurable in x). We take $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$, P the unique probability measure determined by P_X and the $P(x, \cdot)$, namely,

$$P(C) = \int_0^1 P(x, C(x))\, dP_X(x) = \int_0^1 P(x, C(x))\, dx.$$

Now let $X(x, y) = x$, $Y(x, y) = y$. Then

$$P(Y = k) = P(\Omega_1 \times \{k\}) = \int_0^1 P(x, \{k\})\, dx$$

$$= \int_0^1 \binom{n}{k} x^k (1 - x)^{n-k}\, dx = \binom{n}{k} \beta(k + 1, n - k + 1),$$

where $\beta(r, s) = \int_0^1 x^{r-1}(1 - x)^{s-1}\, dx$, $r, s > 0$, is the *beta function*. We can express $\beta(r, s)$ as $\Gamma(r)\Gamma(s)/\Gamma(r + s)$, where $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x}\, dx$, $r > 0$, is the *gamma function*. Since $\Gamma(n + 1) = n!$, $n = 0, 1, \ldots$, we have

$$P(Y = k) = \binom{n}{k} \frac{k!(n - k)!}{(n + 1)!} = \frac{1}{n + 1}, \quad k = 0, 1, \ldots, n.$$

In solving a problem of this type, intuitive reasoning serves as a useful check on the formal development. Thus, the probability that X falls near x is dx; given that $X = x$, the probability that k heads will be obtained is

$$\binom{n}{k} x^k (1-x)^{n-k}.$$

Integrate this from 0 to 1 to obtain the total probability.
The next example involves an n-stage random experiment.

Example: Let X_1 be uniformly distributed between 0 and 1. If $X_1 = x_1$, let X_2 be uniformly distributed between 0 and x_1. In general, if $X_1 = x_1, \ldots, X_k = x_k$, let X_{k+1} be uniformly distributed between 0 and x_k ($k = 1, \ldots, n-1$). Find the expectation of X_n.
Here we have $\Omega_j = \mathbb{R}$, $\mathcal{F}_j = \mathcal{B}(\mathbb{R})$, $\Omega = \prod_{j=1}^n \Omega_j$, $\mathcal{F} = \prod_{j=1}^n \mathcal{F}_j$, $X_j(x_1, \ldots, x_n) = x_j$, $j = 1, \ldots, n$.
Set $P_1 =$ Lebesgue measure on $(0,1)$, and for each $x_1 \in (0,1)$,

$$P(x_1, \cdot) = \frac{1}{x_1}[\text{Lebesgue measure on } (0, x_1)],$$

that is,

$$P(x_1, B) = \frac{1}{x_1} \int_{B \cap (0, x_1)} dx_2.$$

In general, for each $x_1, \ldots, x_k \in (0,1)$, $k = 1, \ldots, n-1$, take

$$P(x_1, \ldots, x_k, \cdot) = \frac{1}{x_k}[\text{Lebesgue measure on } (0, x_k)].$$

(We use open intervals to avoid division by zero.)
Let P be the unique measure on $\mathcal{F}$ determined by P_1 and the $P(x_1, \ldots, x_k, \cdot)$. We may find the expectation of a Borel measurable function g from $\mathbb{R}^n$ to $\mathbb{R}$ by Fubini's theorem:

$$\int_\Omega g\, dP = \int_{\Omega_1} P_1(dx_1) \int_{\Omega_2} P(x_1, dx_2) \cdots \int_{\Omega_n} P(x_1, \ldots, x_{n-1}, dx_n)\, g(x_1, \ldots, x_n).$$

In the present case we have $g(x_1, \ldots, x_n) = X_n(x_1, \ldots, x_n) = x_n$. Thus

$$\mathbb{E}(X_n) = \int_0^1 dx_1 \int_0^{x_1} \frac{1}{x_1} dx_2 \cdots \int_0^{x_{n-2}} \frac{1}{x_{n-2}} dx_{n-1} \int_0^{x_{n-1}} \frac{x_n}{x_{n-1}} dx_n$$

$$= \int_0^1 \frac{x_1}{2^{n-1}} dx_1 = 2^{-n}.$$

This example has an alternative interpretation. Let $Y_1, \ldots, Y_n$ be independent random variables, each uniformly distributed between 0 and 1. Let Z_k be the product $Y_1 \cdots Y_k$, $1 \leq k \leq n$. It turns out (see Problem 2, Section 5.6) that $(Z_1, \ldots, Z_n)$ has the same distribution as $(X_1, \ldots, X_n)$; hence $\mathbb{E}(X_n) = 2^{-n}$.

3.2 The General Concept of Conditional Probability and Expectation

We have seen that specification of the distribution of a random variable X, together with $P(x, B)$, x real, $B \in \mathcal{B}(\mathbb{R})$, interpreted intuitively as the conditional probability that $Y \in B$ given $X = x$, determines a unique probability

$$P(\{X \in A\} \cap B) = \int_A P(x, B)\, dP_X(x),$$

where P_X is the probability measure induced by X, namely,

$$P_X(A) = P(\omega : X(\omega) \in A), \quad A \in \mathcal{F}'.$$

3.2.1 Theorem

Let $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ be a random object on $(\Omega, \mathcal{F}, P)$, and let B be a fixed set in $\mathcal{F}$. Then there is a real-valued Borel measurable function g on $(\Omega', \mathcal{F}')$ such that for each $A \in \mathcal{F}'$,

$$P(\{X \in A\} \cap B) = \int_A g(x)\, dP_X(x).$$

Furthermore, if h is another such function then $g = h$ a.e. $[P_X]$. [We define $P(B|X = x)$ as $g(x)$; it is essentially unique for a given B.]

Proof: Let $\lambda(A) = P(\{X \in A\} \cap B), A \in \mathcal{F}'$. Then λ is a finite measure on $\mathcal{F}'$, absolutely continuous with respect to P_X $[P_X(A) = 0 \Rightarrow \lambda(A) = 0]$. The result follows from the Radon–Nikodym theorem.

3.2.2 Examples

(a) Let X take on only countably many values $x_1, x_2, \ldots$, with

$$p_i = P(X = x_i) > 0, \qquad \sum_{i=1}^{\infty} p_i = 1.$$

We claim that

$$g(x_i) = P(B|X = x_i) = \frac{P(B \cap \{X = x_i\})}{P(X = x_i)}, \qquad i = 1, 2, \ldots$$

(Since P_X is concentrated on the x_i, we need not bother to specify $g(x)$ for x unequal to any of the x_i.) Thus the general definition reduces to the elementary definition in the discrete case. To prove this, let $\Omega' = \{x_1, x_2, \ldots\}$, with $\mathcal{F}'$ the collection of all subsets of Ω'. If $A \in \mathcal{F}'$ and g is defined as above, then

$$\begin{aligned}
\int_A g(x)\, dP_X(x) &= \int_{\Omega'} g(x)\mathbf{1}_A(x)\, dP_X(x) \\
&= \sum_{i=1}^{\infty} g(x_i)\mathbf{1}_A(x_i)P_X\{x_i\} \\
&= \sum_{x_i \in A} g(x_i)P(X = x_i) \\
&= \sum_{x_i \in A} P(B \cap \{X = x_i\}) \\
&= P(\{X \in A\} \cap B).
\end{aligned}$$

Since there is essentially only one g satisfying

$$\int_A g\, dP_X = P(\{X \in A\} \cap B), \quad A \in \mathcal{F}',$$

the g we proposed must be correct.

(b) Let X and Y be random variables with joint density f:

$$[\Omega = \mathbb{R}^2, \quad \mathcal{F} = \mathcal{B}(\mathbb{R}^2), \quad X(x, y) = x, \quad Y(x, y) = y,$$
$$P(A) = \iint_A f(x, y)\, dx\, dy, \quad A \in \mathcal{F}].$$

Now $\{X = x\}$ has probability zero for each x, but there is a reasonable approach to the conditional probability $P\{Y \in C \mid X = x\}$, as follows:

$$\begin{aligned}
P\{Y \in C \mid x - h < X < x + h\} &= \frac{P\{x - h < X < x + h, Y \in C\}}{P\{x - h < X < x + h\}} \\
&= \frac{\displaystyle\int_{x-h}^{x+h} \int_C f(u, y)\, dy\, du}{\displaystyle\int_{x-h}^{x+h} f_1(u)\, du},
\end{aligned}$$

where $f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$ is the density of X.

For small h, this is (hopefully) approximately

$$\frac{2h\int_C f(x,y)\,dy}{2hf_1(x)} = \int_C \frac{f(x,y)}{f_1(x)}\,dy.$$

We are led to define

$$h(y \mid x) = \frac{f(x,y)}{f_1(x)}$$

as the conditional density of Y given $X = x$ (or for short, the conditional density of Y given X). Note that h is defined only when $f_1(x) \neq 0$; however, if $S = \{(x,y) : f_1(x) = 0\}$, then $P\{(X,Y) \in S\} = 0$, since

$$\begin{aligned} P\{(X,Y) \in S\} &= \iint_S f(x,y)\,dx\,dy = \int_{\{x:f_1(x)=0\}} \left[\int_{-\infty}^{\infty} f(x,y)\,dy\right] dx \\ &= \int_{\{x:f_1(x)=0\}} f_1(x)\,dx = 0. \end{aligned}$$

Thus we may essentially ignore those (x,y) for which the conditional density is not defined.

We expect that $P\{Y \in C \mid X = x\} = \int_C h(y \mid x)\,dy$. More generally, if $B \in \mathcal{F}$ and $X = x$, then B will occur iff $Y \in B(x)$. To find $P\{Y \in B(x) \mid X = x\}$, we integrate $h(y \mid x)$ over $y \in B(x)$. Thus we propose

$$g(x) = \int_{B(x)} h(y \mid x)\,dy, \quad B \in \mathcal{F}, \quad x \in \mathbb{R},$$

as the conditional probability of B given $X = x$. To prove this, first note that

$$g(x) = \int_{-\infty}^{\infty} \mathbf{1}_{B(x)}(y) h(y \mid x)\,dy;$$

hence g is Borel measurable by Fubini's theorem. Also, if $A \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned} P(\{X \in A\} \cap B) &= \iint_{x\in A,\,(x,y)\in B} f(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \mathbf{1}_{B(x)}(y) h(y \mid x)\,dy\right] \mathbf{1}_A(x) f_1(x)\,dx \\ &= \int_{x\in A} f_1(x) \int_{y\in B(x)} h(y \mid x)\,dy\,dx \\ &= \int_A g(x) f_1(x)\,dx \\ &= \int_A g(x)\,dP_X(x) \quad \text{[2.2.3(c)]}. \end{aligned}$$

Therefore $g(x) = P(B \mid X = x)$.

In this example we may look at the formula $f(x, y) = f_1(x)h(y \mid x)$ in two ways. If (X, Y) has density f, we have a notion of conditional probability:

$$P\{Y \in C \mid X = x\} = \int_C h(y \mid x)\, dy.$$

On the other hand, suppose that we specify that X has density f_1, and whenever $X = x$, we select Y according to the density $h(\cdot \mid x)$; in other words, we specify $P(x, B) = \int_B h(y \mid x)\, dy, B \in \mathcal{B}(\mathbb{R})$. A unique measure P on $\mathcal{B}(\mathbb{R}^2)$ is determined, satisfying, for $A, B \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned} P\{X \in A, Y \in B\} &= \int_A P(x, B) f_1(x)\, dx \\ &= \iint_{x \in A, y \in B} f_1(x) h(y \mid x)\, dx\, dy. \end{aligned}$$

Therefore (X, Y) has density $f(x, y) = f_1(x)h(y \mid x)$.

Thus we have two points of view. We may regard the conditional density of Y given $X = x$ as ultimately derived from the joint density of X and Y. On the other hand, we may regard the observation of X and Y as a two-stage random experiment, where the distribution of Y at stage 2 depends on the value of X at stage 1. The above discussion shows that the assignment of probabilities to events involving (X, Y) is the same in either case.

We may also define conditional densities in higher dimensions. For example, if X, Y, Z, W have joint density f, we define (say) the conditional density of (Z, W) given (X, Y) as

$$h(z, w \mid x, y) = \frac{f(x, y, z, w)}{f_{XY}(x, y)},$$

where $f_{XY}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y, z, w)\, dz\, dw$. If $B \in \mathcal{B}(\mathbb{R}^4)$, then, exactly as before,

$$P\{(Z, W) \in B \mid X = x, Y = y\} = \iint_{B(x,y)} h(z, w \mid x, y)\, dz\, dw.$$

This is verified by proving that, for $A \in \mathcal{B}(\mathbb{R}^2)$,

$$P\{(X, Y) \in A\} \cap B = \int_A \left[\int_B P(B \mid X = x, Y = y) f_{XY}(x, y)\, dx\, dy \right].$$

(c) Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be given, with no probability defined as yet. Take $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$, $X(\omega_1, \omega_2) = \omega_1$, $Y(\omega_1, \omega_2) = \omega_2$. Assume that we are given P_X, a probability measure on $(\Omega_1, \mathcal{F}_1)$, and also that we are given $P(x, B)$, $x \in \Omega_1$, $B \in \mathcal{F}_2$, a probability measure in B for each fixed x, and a Borel measurable function of x for each fixed B. (We are specifying the distribution of X and the conditional distribution of Y, given $X = x$.) By the product measure theorem, there is a unique measure P on $\mathcal{F}$ such that

$$P(X \in A, Y \in B) = P(A \times B) = \int_A P(x, B)\, dP_X(x).$$

It follows that $P(x, B)$ is in fact the conditional probability $P(Y \in B | X = x)$.

We now consider conditional expectation. Let X and Y be random variables on $(\Omega, \mathcal{F}, P)$; we ask for a reasonable definition of the expectation of Y given that $X = x$, written $E(Y|X = x)$. Intuitively, $E(Y|X = x)$ should reflect the long-run average value of Y in a sequence of independent trials when we look only at those observations on which $\{X = x\}$ has occurred.

If X and Y are discrete and we are given that $X = x$, the conditional probability of an event involving Y is governed by the set of conditional probabilities $p(y|x) = P(Y = y|X = x)$. Thus a reasonable figure for $E(Y|X = x)$ is $\sum_y yp(y|x)$. Similarly, if (X, Y) is absolutely continuous, and $h = h(y|x)$ is the conditional density of Y given $X = x$, we expect that $E(Y|X = x)$ should be $\int_{-\infty}^{\infty} yh(y|x)\, dy$. What we need is a general framework that includes these special cases.

Let Y be a random variable (or an extended random variable) on $(\Omega, \mathcal{F}, P)$, and let $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ be a random object. Our general definition of conditional probability hinges on a version of the theorem of total probability:

$$P(\{X \in A\} \cap B) = \int_A P(B|X = x)\, dP_X(x), \quad A \in \mathcal{F}', \quad B \in \mathcal{F}.$$

There is a closely related "theorem of total expectation," which may be developed intuitively as follows. The probability that X falls near x is $dP_X(x)$; given that $X = x$, the average value of Y is what we are looking for, namely, $E(Y|X = x)$. It is reasonable to hope that the total expectation may be found by adding all the contributions:

$$E(Y) = \int_{\Omega'} E(Y|X = x)\, dP_X(x).$$

To develop this further, we replace Y by $YI_{\{X \in A\}}$, where $A \in \mathcal{F}'$. If $x \in A$, we expect that $E(YI_{\{X \in A\}}|X = x) = E(Y|X = x)$ since $X(\omega) = x \in A$

implies $I_{\{X\in A\}}(\omega) = 1$. If $x \notin A$, we expect that $E(YI_{\{X\in A\}}|X = x) = 0$. Replacing Y by $YI_{\{X\in A\}}$ in the above version of the theorem of total expectation, we obtain:

$$E(YI_{\{X\in A\}}) = \int_{\Omega'} E(YI_{\{X\in A\}}|X = x)\, dP_X(x)$$

or

$$\int_{\{X\in A\}} Y\, dP = \int_A E(Y|X = x)\, dP_X(x).$$

In fact, this requirement essentially determines $E(Y|X = x)$.

3.2.3 Theorem

Let Y be an extended random variable on $(\Omega, \mathcal{F}, P)$, and $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$, a random object. If $E(Y)$ exists, there is a function $g : (\Omega', \mathcal{F}') \to (\mathbb{R}, \mathcal{B})$ such that for each $A \in \mathcal{F}'$,

$$\int_{\{X\in A\}} Y\, dP = \int_A g(x)\, dP_X(x).$$

(As usual, $\mathcal{B}$ denotes the class of Borel sets.) Furthermore, if h is another such function, then $g = h$ a.e. $[P_X]$. [We define $E(Y|X = x)$ as $g(x)$; it is essentially unique *for a given* Y.]

Proof. Let

$$\lambda(A) = \int_{\{X\in A\}} Y\, dP = \int_{X^{-1}(A)} Y\, dP, \quad A \in \mathcal{F}'.$$

Then λ is a countably additive set function on $\mathcal{F}'$ by 1.6.1, and is absolutely continuous with respect to P_X since $P_X(A) = P(X \in A)$. The result follows from the Radon–Nikodym theorem.

Conditional expectation includes conditional probability as a special case, as we now prove.

3.2.4 Corollary

If X is a random object on $(\Omega, \mathcal{F}, P)$ and $B \in \mathcal{F}$, then

$$E(I_B|X = x) = P(B|X = x) \quad \text{a.e. } [P_X].$$

Proof. In 3.2.3, set $Y = I_B$; the defining equation for conditional expectation becomes

$$P(\{X \in A\} \cap B) = \int_A E(I_B|X = x)\, dP_X(x).$$

The result now follows from 3.2.1.

Let us compare the general definition with the intuitive concept in special cases.

3.2.5 Examples

(a) Let X take on only countably many values $x_1, x_2, \ldots$ (assume all $P(X = x_i) > 0$). We have seen that

$$P(B|X = x_i) = \frac{P(B \cap \{X = x_i\})}{P(X = x_i)}, \quad B \in \mathcal{F}.$$

Thus we should expect that

$$E(I_B|X = x_i) = \frac{1}{P(X = x_i)} \int_{\{X=x_i\}} I_B\, dP.$$

Proceeding from indicators to nonnegative simple functions to nonnegative measurable functions to arbitrary measurable functions, we should like to believe that if $E(Y)$ exists,

$$E(Y|X = x_i) = \frac{1}{P(X = x_i)} \int_{\{X=x_i\}} Y\, dP, \quad i = 1, 2, \ldots \tag{1}$$

[We are not proving anything here since we do not yet know, for example, that

$$E\left(\sum_{i=1}^{n} Y_i \middle| X = x\right) = \sum_{i=1}^{n} E(Y_i|X = x).]$$

To establish (1), let

$$g(x_i) = \frac{1}{P(X = x_i)} \int_{\{X=x_i\}} Y\, dP, \quad i = 1, 2, \ldots.$$

(We may assume $\Omega' = \{x_1, x_2, \ldots\}$, with $\mathcal{F}'$ the class of all subsets of Ω'.) Then

$$\int_{\{X\in A\}} Y\, dP = \sum_{x_i \in A} P(X = x_i) \cdot \frac{1}{P(X = x_i)} \int_{\{X=x_i\}} Y\, dP$$

$$= \sum_{x_i \in A} P(X = x_i) g(x_i) = \int_A g(x)\, dP_X(x), \quad A \in \mathcal{F}',$$

as desired.

In the special case when Y is discrete, (1) assumes a simpler form. If Y takes on the values $y_1, y_2, \ldots$, we obtain (using countable additivity of the integral)

$$E(Y|X = x_i) = \sum_j y_j \frac{P(X = x_i, Y = y_j)}{P(X = x_i)}$$

$$= \sum_j y_j P(Y = y_j | X = x_i). \tag{2}$$

(b) Let $B \in \mathcal{F}$, and assume $P(B) > 0$. If $E(Y)$ exists, we define the conditional expectation of Y given B as follows. Let $X = I_B$, and set $E(Y|B) = E(Y|X = 1)$. This is a special case of (a); we obtain [see (1)]

$$E(Y|B) = \frac{1}{P(B)} \int_B Y\, dP,$$

in other words,

$$E(Y|B) = \frac{E(Y I_B)}{P(B)}. \tag{3}$$

(c) Let X and Y be random variables having a joint density f, and let $h = h(y|x)$ be the conditional density of Y given X. We claim that if $E(Y)$ exists,

$$E(Y|X = x) = \int_{-\infty}^{\infty} y h(y|x)\, dy. \tag{4}$$

To prove this, note that

$$\int_{\{X \in A\}} Y\, dP = \iint_{\{(x,y): x \in A\}} y f(x,y)\, dx\, dy$$

$$= \int_{x \in A} \left[\int_{-\infty}^{\infty} y h(y|x)\, dy \right] f_1(x)\, dx \quad \text{by Fubini's theorem}$$

$$= \int_A \left[\int_{-\infty}^{\infty} y h(y|x)\, dy \right] dP_X(x),$$

proving (4).

Notice also that if q is a Borel measurable function from $\mathbb{R}$ to $\mathbb{R}$ and $E[q(Y)]$ exists, then

$$E(q(Y)|X = x) = \int_{-\infty}^{\infty} q(y)h(y|x)\,dy \tag{5}$$

by the same argument as above. Similarly, if X and Y are discrete [see (a), (2)] and $E[q(Y)]$ exists, then

$$E(q(Y)|X = x_i) = \sum_j q(y_j)P(Y = y_j|X = x_i). \tag{6}$$

(d) Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be given, with no probability defined as yet. Let $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$, $X(x, y) = x$, $Y(x, y) = y$. Assume that a probability measure P_X on $\mathcal{F}_1$ is given, and also that we are given $P(x, B)$, $x \in \Omega_1$, $B \in \mathcal{F}_2$, a probability measure in B for each fixed x, and a Borel measurable function of x for each fixed B. Let P be the unique measure on $\mathcal{F}$ determined by P_X and the $P(x, \cdot)$.
If $f\colon (\Omega_2, \mathcal{F}_2) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $E[f(Y)]$ exists, we claim that

$$E(f(Y)|X = x) = \int_{\Omega_2} f(y)P(x, dy). \tag{7}$$

To see this, we note, with the aid of Fubini's theorem, that

$$\int_{\{X\in A\}} f(Y)\,dP = \int_{\Omega} f(Y)I_{\{X\in A\}}\,dP$$

$$= \int_{\Omega_1} \left(\int_{\Omega_2} f(Y(x, y))I_A(x)P(x, dy) \right) dP_X(x)$$

$$= \int_A \left(\int_{\Omega_2} f(y)P(x, dy) \right) dP_X(x).$$

3.3 Conditional Expectation Given a σ-Field

It will be very convenient to regard conditional expectations as functions defined on the sample space Ω. Let us first recall the main result of the previous section.

If Y is an extended random variable on $(\Omega, \mathcal{F}, P)$ whose expectation exists, and $X\colon (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ is a random object, then $g(x) = E(Y|X = x)$ is characterized as the a.e. $[P_X]$ unique function:

$$g\colon (\Omega', \mathcal{F}') \to (\mathbb{R}, \mathcal{B})$$

satisfying

$$\int_{\{X\in A\}} Y\, dP = \int_A E(Y|X = x)\, dP_X(x), \qquad A \in \mathcal{F}'. \tag{1}$$

Now let $h(\omega) = g(X(\omega))$; then

$$h\colon (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$$

Thus $h(\omega)$ is the conditional expectation of Y, given that X takes the value $x = X(\omega)$; consequently, h measures the average value of Y given X, but h is defined on Ω rather than Ω'.
It will be useful to have an analog on (1) for h. We claim that

$$\int_{\{X\in A\}} h\, dP = \int_{\{X\in A\}} Y\, dP \qquad \text{for each } A \in \mathcal{F}'. \tag{2}$$

To prove this, note that

$$\begin{aligned}
\int_{\{X\in A\}} h\, dP &= \int_\Omega g(X(\omega)) I_A(X(\omega))\, dP(\omega) \\
&= \int_{\Omega'} g(x) I_A(x)\, dP_X(x) \quad \text{[by 2.2.3(b)]} \\
&= \int_A g(x)\, dP_X(x) \\
&= \int_{\{X\in A\}} Y\, dP \quad \text{[by (1)]}.
\end{aligned}$$

Since $\{X \in A\} = X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$, we may express (2) as follows:

$$\int_C h\, dP = \int_C Y\, dP \quad \text{for each } C \in X^{-1}(\mathcal{F}'), \tag{3}$$

where $X^{-1}(\mathcal{F}') = \{X^{-1}(A) : A \in \mathcal{F}'\}$.

The σ-field $X^{-1}(\mathcal{F}')$ will be very important for us, and we shall look at some of its properties before proceeding.

Definition: Let $X\colon (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ be a random object. The σ-field induced by X is given by

$$\sigma(X) = X^{-1}(\mathcal{F}').$$

Thus a set in $\sigma(X)$ is of the form $\{X \in A\}$ for some $A \in \mathcal{F}'$. In particular, if $X = (X_1, \ldots, X_n)$, a random vector, $\sigma(X)$ consists of all sets $\{X \in B\}$, $B \in \mathcal{B}(\mathbb{R}^n)$.

3.3.1 Theorem

Let $X\colon (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$.

(a) The induced σ-field $\sigma(X)$ is the smallest σ-field $\mathcal{G}$ of subsets of Ω making X measurable relative to $\mathcal{G}$ and $\mathcal{F}'$.

(b) If $\Omega' = \prod_j \Omega_j, \mathcal{F}' = \prod_j \mathcal{F}_j$ so that $X = (X_1, X_2, \ldots)$, where $X_j\colon (\Omega, \mathcal{F}) \to (\Omega_j, \mathcal{F}_j)$ is the jth coordinate of X, then $\sigma(X)$ is the smallest σ-field $\mathcal{G}$ of subsets of Ω making each X_j measurable relative to $\mathcal{G}$ and $\mathcal{F}_j$.

(c) If $Z\colon (\Omega, \sigma(X)) \to (\mathbb{R}, \mathcal{B})$ (or $(\mathbb{R}, \mathcal{B})$), then $Z = f \circ X$ for some $f\colon (\Omega', \mathcal{F}') \to (\mathbb{R}, \mathcal{B})$. Conversely, if $Z = f \circ X$ and $f\colon (\Omega', \mathcal{F}') \to (\mathbb{R}, \mathcal{B})$, then $Z\colon (\Omega, \sigma(X)) \to (\mathbb{R}, \mathcal{B})$.

Proof.

(a) If $A \in \mathcal{F}'$, then $X^{-1}(A) \in \sigma(X)$ by definition of $\sigma(X)$; hence $\sigma(X)$ makes X measurable. If $\mathcal{G}$ is any σ-field making X measurable, $X^{-1}(A) \in \mathcal{G}$ for all $A \in \mathcal{F}'$; hence $\sigma(X) \subseteq \mathcal{G}$.

(b) By 4.11.3, X is measurable relative to $\mathcal{G}$ and $\mathcal{F}'$ iff each X_j is measurable relative to $\mathcal{G}$ and $\mathcal{F}_j$. The result follows from (a).

(c) Assume $Z\colon (\Omega, \sigma(X)) \to (\mathbb{R}, \mathcal{B})$. If Z is an indicator I_C, $C \in \sigma(X)$, then $C = X^{-1}(A)$ for some $A \in \mathcal{F}'$. If $f = I_A$, then $f \circ X = I_{X^{-1}(A)} = I_C = Z$.
If $Z = \sum_{k=1}^n z_k I_{C_k}$ is a finite-valued simple function and $I_{C_k} = f_k \circ X$ as above, then $Z = f \circ X$, where $f = \sum_{k=1}^n z_k f_k$.

In general, let $Z_1, Z_2, \ldots$ be finite-valued simple functions such that $Z_n \to Z$. We can express $Z_n = f_n \circ X$ as above; define $f = \lim_{n\to\infty} f_n$ where the limit exists, and 0 elsewhere. Then

$$Z(\omega) = \lim_{n\to\infty} Z_n(\omega) = \lim_{n\to\infty} f_n(X(\omega)) = f(X(\omega)).$$

The converse holds because a composition of measurable functions is measurable.

Now let us return to Eq. (3) at the beginning of this section:

$$\int_C h\,dP = \int_C Y\,dP, \qquad C \in \sigma(X),$$

where $h = g \circ X, g(x) = E(Y|X = x)$. Since $g\colon (\Omega', \mathcal{F}') \to (\mathbb{R}, \mathcal{B})$, we have

$$h\colon (\Omega, \sigma(X)) \to (\mathbb{R}, \mathcal{B}) \text{ by 3.3.1(c).}$$

This fact, along with (3), characterizes h, and gives us the concept of conditional expectation given a σ-field.

3.3.2 Theorem

Let Y be an extended random variable on $(\Omega, \mathcal{F}, P)$, $\mathcal{G}$ a sub σ-field of $\mathcal{F}$. Assume that $E(Y)$ exists. Then there is a function $E(Y|\mathcal{G})\colon (\Omega, \mathcal{G}) \to (\mathbb{R}, \mathcal{B})$, called the conditional expectation of Y given $\mathcal{G}$, such that

$$\int_C Y\,dP = \int_C E(Y|\mathcal{G})\,dP \qquad \text{for each } C \in \mathcal{G}.$$

Any two such functions must coincide a.e. [P]. [Note that we cannot simply set $E(Y|\mathcal{G}) = Y$, as $E(Y|\mathcal{G})$ is required to be measurable relative to $\mathcal{G}$.]

3.3.3 Comment

If $g(x) = E(Y|X = x)$ and $h(\omega) = g(X(\omega))$, then by 5.4.3,

$$h = E(Y|\mathcal{G}), \quad \text{where } \mathcal{G} = \sigma(X);$$

for convenience we shall usually write

$$h = E(Y|X).$$

[For example, if $E(Y|X = x) = x^2$, then $E(Y|X) = X^2$.]

We have seen that the conditional expectation $g(x) = E(Y|X = x)$, $x \in \Omega'$, can be transferred to Ω by forming $h(\omega) = g(X(\omega))$. Conversely, any conditional expectation $E(Y|\mathcal{G})$, $\mathcal{G}$ an arbitrary sub σ-field of $\mathcal{F}$, arises from a random object X in this way. Simply take $X : (\Omega, \mathcal{F}) \to (\Omega, \mathcal{F})$ to be the identity map: $X(\omega) = \omega, \omega \in \Omega$. Then $X^{-1}(\mathcal{G}) = \mathcal{G}$, so if $g(x) = E(Y|X = x)$, then $h = E(Y|\sigma(X)) = E(Y|\mathcal{G})$.

Now intuitively, $E(Y|\mathcal{G}) = E(Y|X)$ is the average value of Y, given that X is known. But what does it mean to "know" $X\colon (\Omega, \mathcal{F}) \to (\Omega, \mathcal{G})$? The events involving X are sets of the form $\{X \in G\}, G \in \mathcal{G}$, and since X is the identity map, $\{X \in G\} = G$. Since an event corresponds to a question that has a yes or no answer, $E(Y|\mathcal{G})$ may be interpreted as the average value of $Y(\omega)$, given that we know, for each $G \in \mathcal{G}$, whether or not $\omega \in G$. Some examples may help to make this clear.

3.3.4 Examples

(a) Let X be discrete, with values $x_1, x_2, \ldots$; take $\Omega' = \{x_1, x_2, \ldots\}$, with $\mathcal{F}'$ the class of all subsets of Ω', and assume $P(X = x_i) > 0$ for all i. We have seen in 5.3.5(a) that

$$g(x_i) = E(Y|X = x_i) = \frac{1}{P(X = x_i)} \int_{\{X=x_i\}} Y \, dP.$$

Let $h = E(Y|X)$, that is, $h(\omega) = g(X(\omega))$. Then h has the constant value $g(x_i)$ on the set $\{X = x_i\}$, and $\mathcal{G} = X^{-1}(\mathcal{F}')$ consists of all unions of the sets $\{X = x_i\}$. Knowledge of the value of $X(\omega)$ is equivalent to knowledge, for each $G \in \mathcal{G}$, of the membership or nonmembership of $\omega \in G$.

(b) Let X and Y be random variables with a joint density f. Let $\Omega = \mathbb{R}^2, \mathcal{F} = \mathcal{B}(\mathbb{R}^2), P(B) = \iint_B f(x, y)\, dx\, dy, B \in \mathcal{F}$; $X(x, y) = x, Y(x, y) = y$. Take $\Omega' = \mathbb{R}, \mathcal{F}' = \mathcal{B}(\mathbb{R})$. We have seen in 5.3.5(c) that $g(x) = E(Y|X = x) = \int_{-\infty}^{\infty} y h_0(x, y)\, dy$, where h_0 is the conditional density of Y given X. Let $h = E(Y|X)$, that is, $h(\omega) = g(X(\omega))$ or $h(x, y) = g(x)$.
Thus $E(Y|X)$ is constant on vertical strips; also, $X^{-1}(\mathcal{F}')$ consists of all sets $B \subseteq \mathbb{R}, B \in \mathcal{B}(\mathbb{R})$. Since $x \in B$ iff $(x, y) \in B \times \mathbb{R}$, information about $X(\omega)$...

3.4 Properties of Conditional Expectation

The conditional expectation $E(Y|X = x)$ is a more intuitive object than the conditional expectation $E(Y|\mathcal{G})$; however, the intuition cannot easily be pushed beyond the case in which X is a finite-dimensional random vector. Thus in formal arguments in which $\mathcal{G}$ is an arbitrary σ-field, we are forced to use $E(Y|\mathcal{G})$.

For convenience, we develop the basic properties of conditional expectation in pairs, one argument for $E(Y|\mathcal{G})$ and another (usually very similar) for $E(Y|X = x)$. Theorems about conditional probabilities are obtained by replacing Y by I_B, and results concerning $E(Y|X)$ are obtained by setting $\mathcal{G} = \mathcal{F}_X = \sigma(X)$.

In the discussion to follow, $Y, Y_1, Y_2, \ldots$ are extended random variables on $(\Omega, \mathcal{F}, P)$, with all expectations assumed to exist; $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ is a random object, and $\mathcal{G}$ is a sub σ-field of $\mathcal{F}$. The phrase "a.e." with no measure specified will always mean a.e. [P]. If $Z : (\Omega, \mathcal{G}) \to (\mathbb{R}, \mathcal{B})$, we say that Z is $\mathcal{G}$-measurable, and if $g : (\Omega', \mathcal{F}') \to (\mathbb{R}, \mathcal{B})$, we say that g is $\mathcal{F}'$-measurable.

3.4.1 Theorem

If Y is a constant k a.e., then

(a) $E(Y|\mathcal{G}) = k$ a.e.

(a′) $E(Y|X = x) = k$ a.e. $[P_X]$

If $Y_1 \leq Y_2$ a.e., then

(b) $E(Y_1|\mathcal{G}) \leq E(Y_2|\mathcal{G})$ a.e.

(b′) $E(Y_1|X = x) \leq E(Y_2|X = x)$ a.e. $[P_X]$

[A statement such as $E(Y_1|\mathcal{G}) \leq E(Y_2|\mathcal{G})$ a.e. means that if Z_j is a version of $E(Y_j|\mathcal{G})$, in other words, Z_j satisfies the defining requirement 3.3.2, then $Z_1 \leq Z_2$ a.e.]

(c) $E(Y|\mathcal{G}) \leq E(|Y||\mathcal{G})$ a.e.

(c′) $E(Y|X = x)| \leq E(|Y||X = x)$ a.e. $[P_X]$

Proof:
(a) The function constant at k is $\mathcal{G}$-measurable and

$$\int_C Y\,dP = \int_C k\,dP, \quad C \in \mathcal{G}.$$

(a′) If $g(x) \equiv k, x \in \Omega'$, then g is $\mathcal{F}'$-measurable and

$$\int_{\{X\in A\}} Y\,dP = \int_A k\,dP_X.$$

(b) $\int_C Y_1\,dP \leq \int_C Y_2\,dP$; hence

$$\int_C E(Y_1|\mathcal{G})\,dP \leq \int_C E(Y_2|\mathcal{G})\,dP \quad \text{for each } C \in \mathcal{G}.$$

(b′) $\int_{\{X\in A\}} Y_1\,dP \leq \int_{\{X\in A\}} Y_2\,dP$; hence

$$\int_A E(Y_1|X = x)\,dP_X \leq \int_A E(Y_2|X = x)\,dP_X \quad \text{for each } A \in \mathcal{F}'.$$

Parts (c) and (c′) follow from (b) and (b′), along with the observation that

$$-|Y| \leq Y \leq |Y|.$$

We now prove an additivity theorem for conditional expectations.

3.4.2 Theorem

(a) If $a, b \in \mathbb{R}$, and $aE(Y_1)+bE(Y_2)$ is well defined (not of the form $\infty-\infty$), then

$$E(aY_1 + bY_2|\mathcal{G}) = aE(Y_1|\mathcal{G}) + bE(Y_2|\mathcal{G}) \quad \text{a.e.}$$

(a′) If $a, b \in \mathbb{R}$ and $aE(Y_1) + bE(Y_2)$ are well defined, then

$$E(aY_1 + bY_2|X = x) = aE(Y_1|X = x) + bE(Y_2|X = x) \quad \text{a.e. } [P_X].$$

Proof: (a) If $C \in \mathcal{G}$,

$$\int_C (aY_1+bY_2)\,dP = \int_C aY_1\,dP + \int_C bY_2\,dP \quad \text{by the additivity theorem for integrals}$$

$$= \int_C aE(Y_1|\mathcal{G})\,dP + \int_C bE(Y_2|\mathcal{G})\,dP \quad \text{by definition of conditional expectation.}$$

Thus $\int_C aE(Y_1|\mathcal{G})\,dP + \int_C bE(Y_2|\mathcal{G})\,dP$ is well defined, so again by the additivity theorem for integrals,

$$\int_C (aY_1 + bY_2)\,dP = \int_C [aE(Y_1|\mathcal{G}) + bE(Y_2|\mathcal{G})]\,dP,$$

as desired.

(a′) This is done as in (a), with C replaced by $\{X \in A\}$ and

$$\int_C \mathbb{E}(Y_j \mid \mathcal{G})\,dP \quad \text{by} \quad \int_A \mathbb{E}(Y_j \mid X = x)\,dP_X.$$

In the future, we shall dispose of proofs of this type with a phrase such as "same as (a)."

3.4.3 Theorem

If $Y_n \geq 0$ for all n and $Y_n \uparrow Y$ a.e., then

(a) $\mathbb{E}(Y_n \mid \mathcal{G}) \uparrow \mathbb{E}(Y \mid \mathcal{G})$ a.e.

(a′) $\mathbb{E}(Y_n \mid X = x) \uparrow \mathbb{E}(Y \mid X = x)$ a.e. $[P_X]$.

If all $Y_n \geq 0$, then

(b) $\mathbb{E}\left(\sum_{n=1}^{\infty} Y_n \mid \mathcal{G}\right) = \sum_{n=1}^{\infty} \mathbb{E}(Y_n \mid \mathcal{G})$ a.e.

(b′) $\mathbb{E}\left(\sum_{n=1}^{\infty} Y_n \mid X = x\right) = \sum_{n=1}^{\infty} \mathbb{E}(Y_n \mid X = x)$ a.e. $[P_X]$.

In particular, if $B_1, B_2, \ldots$ are disjoint sets in $\mathcal{F}$,

(c) $P\left(\bigcup_{n=1}^{\infty} B_n \mid \mathcal{G}\right) = \sum_{n=1}^{\infty} P(B_n \mid \mathcal{G})$ a.e.

(c′) $P\left(\bigcup_{n=1}^{\infty} B_n \mid X = x\right) = \sum_{n=1}^{\infty} P(B_n \mid X = x)$ a.e. $[P_X]$.

Proof: **(a)** $\int_C Y_n \, dP = \int_C \mathbb{E}(Y_n \mid \mathcal{G}) \, dP$, $C \in \mathcal{G}$; by 3.4.1(b), the $\mathbb{E}(Y_n \mid \mathcal{G}) \uparrow$ increase to a $\mathcal{G}$-measurable function h. By the monotone convergence theorem,

$$\int_C Y \, dP = \int_C h \, dP;$$

hence $h = \mathbb{E}(Y \mid \mathcal{G})$ a.e.

(a′) Same as (a).

(b) By 3.4.2(a), $\mathbb{E}(\sum_{k=1}^{n} Y_k \mid \mathcal{G}) = \sum_{k=1}^{n} \mathbb{E}(Y_k \mid \mathcal{G})$ a.e. Let $n \to \infty$ and apply part (a) to obtain the desired result.

(b′) Same as (b).
Finally, (c) is a special case of (b), and (c′) of (b′).

If we take the expectation of a conditional expectation, the result is the same as if we were to take the expectation directly.

3.4.4 Theorem

(a)

$$\mathbb{E}[\mathbb{E}(Y \mid \mathcal{G})] = \mathbb{E}(Y);$$

hence if Y is integrable, so is $\mathbb{E}(Y \mid \mathcal{G})$.

(a') $\int_{\Omega'} \mathbb{E}(Y \mid X = x) \, dP_X(x) = \mathbb{E}(Y)$.

Proof.

(a)

$$\int_{\Omega} Y \, dP = \int_{\Omega} \mathbb{E}(Y \mid \mathcal{G}) \, dP.$$

(a') $\int_{\Omega'} \mathbb{E}(Y \mid X = x) \, dP_X(x) = \int_{X \in \Omega'} Y \, dP = \int_{\Omega} Y \, dP.$

We now prove the dominated convergence theorem for conditional expectations.

3.4.5 Theorem

If $|Y_n| \le Z$ for all n, with $\mathbb{E}(Z)$ finite, and $Y_n \to Y$ a.e., then

(a) $\mathbb{E}(Y_n \mid \mathcal{G}) \to \mathbb{E}(Y \mid \mathcal{G})$ a.e.

(a') $\mathbb{E}(Y_n \mid X = x) \to \mathbb{E}(Y \mid X = x)$ a.e. $[P_X]$.

The extended monotone convergence theorem and Fatou's lemma may be proved for conditional expectations, as follows.

3.4.6 Theorem

Assume $Y_n \ge Z$ for all n, where $\mathbb{E}(Z) > -\infty$.

(a) If $Y_n \uparrow Y$ a.e., then $\mathbb{E}(Y_n \mid \mathcal{G}) \uparrow \mathbb{E}(Y \mid \mathcal{G})$ a.e.

(a') If $Y_n \uparrow Y$ a.e., then $\mathbb{E}(Y_n \mid X = x) \uparrow \mathbb{E}(Y \mid X = x)$ a.e. $[P_X]$.

(b) $\liminf_{n\to\infty} \mathbb{E}(Y_n \mid \mathcal{G}) \ge \mathbb{E}(\liminf_{n\to\infty} Y_n \mid \mathcal{G})$ a.e.

(b') $\liminf_{n\to\infty} \mathbb{E}(Y_n \mid X = x) \ge \mathbb{E}(\liminf_{n\to\infty} Y_n \mid X = x)$ a.e. $[P_X]$.

Now assume $Y_n \le Z$ for all n, where $\mathbb{E}(Z) < +\infty$.

(c) If $Y_n \downarrow Y$ a.e., then $\mathbb{E}(Y_n \mid \mathcal{G}) \downarrow \mathbb{E}(Y \mid \mathcal{G})$ a.e.

(c') If $Y_n \downarrow Y$ a.e., then $\mathbb{E}(Y_n \mid X = x) \downarrow \mathbb{E}(Y \mid X = x)$ a.e. $[P_X]$.

(d) $\limsup_{n\to\infty} \mathbb{E}(Y_n \mid \mathcal{G}) \le \mathbb{E}(\limsup_{n\to\infty} Y_n \mid \mathcal{G})$ a.e.

(d') $\limsup_{n\to\infty} \mathbb{E}(Y_n \mid X = x) \le \mathbb{E}(\limsup_{n\to\infty} Y_n \mid X = x)$ a.e. $[P_X]$.

Proof.

(a) If $C \in \mathcal{G}$ then $\int_C Y_n \, dP = \int_C \mathbb{E}(Y_n \mid \mathcal{G}) \, dP$, and $\mathbb{E}(Y_n \mid \mathcal{G})$ increases to a limit h a.e. By the extended monotone convergence theorem,

$$\int_C Y \, dP = \int_C h \, dP, \quad \text{and therefore } h = \mathbb{E}(Y \mid \mathcal{G}) \text{ a.e.}$$

(b) Let $Y_n' = \inf_{k \ge n} Y_k$; then have $Y_n' \uparrow Y' = \liminf_{n\to\infty} Y_n$. By (a),

$$\mathbb{E}(Y_n' \mid \mathcal{G}) \uparrow \mathbb{E}(Y' \mid \mathcal{G}) \text{ a.e.}$$

But $Y_n' \le Y_n$, so

$$\mathbb{E}(Y_n' \mid \mathcal{G}) \le \mathbb{E}(Y_n \mid \mathcal{G}),$$

hence

$$\mathbb{E}(Y' \mid \mathcal{G}) = \lim_{n\to\infty} \mathbb{E}(Y_n' \mid \mathcal{G}) \le \liminf_{n\to\infty} \mathbb{E}(Y_n \mid \mathcal{G}) \quad \text{by 5.5.1(b).}$$

(c) This follows from (a) upon replacing all extended random variables by their negatives.

(d)
$$\begin{aligned}\mathbb{E}(\limsup Y_n \mid \mathcal{G}) &= -\mathbb{E}(\liminf(-Y_n) \mid \mathcal{G}) \\ &\geq -\liminf \mathbb{E}(-Y_n \mid \mathcal{G}) \quad \text{by (b)} \\ &= \limsup \mathbb{E}(Y_n \mid \mathcal{G})\end{aligned}$$

The proofs of (a') to (d') are the same as the proofs of (a) to (d). □

Thus far we have examined $\mathbb{E}(Y \mid \mathcal{G})$ and $\mathbb{E}(Y \mid X = x)$ under various hypotheses on Y; now we impose conditions on $\mathcal{G}$ and X.

3.4.7 Theorem

(a) $\mathbb{E}(Y \mid \{\emptyset, \Omega\}) = \mathbb{E}(Y)$ a.e.

(a') If X is a constant b.a.e., then $\mathbb{E}(Y \mid X = x) = \mathbb{E}(Y)$ a.e. $[P_X]$.

(b) $\mathbb{E}(Y \mid \mathcal{F}) = Y$ a.e.

(b') If $X\colon (\Omega, \mathcal{F}) \to (\Omega, \mathcal{F})$ is the identity map, then $\mathbb{E}(Y \mid X = \omega) = Y(\omega)$ a.e. $[P]$.

Proof.

(a) $\int_C Y\, dP = \int_C \mathbb{E}(Y)\, dP$ if $C = \emptyset$ or Ω.

(a') If $b \in A$,
$$\int_{X \in A} X\, dP = \int_\Omega Y\, dP = \mathbb{E}(Y) = \int_A \mathbb{E}(Y)\, dP_X.$$
If $b \notin A$,
$$\int_{\{X \in A\}} Y\, dP = \int_\emptyset Y\, dP = 0 = \int_A \mathbb{E}(Y)\, dP_X.$$

(b) $\displaystyle\int_C Y\, dP = \int_C E(Y)\, dP$, $C \in F$, and Y is F-measurable.

(b') $\displaystyle\int_{\{X \in A\}} Y\, dP = \int_A Y\, dP = E(Y) = \int_A E(Y)\, dP_X$, and Y is F $(=F_X)$-measurable.

The following result is preparatory to the next theorem.

3.4.8 Lemma.

If $f : (\Omega, G) \to (\mathbb{R}, B)$, μ is a measure on G, and B is an atom of G relative to μ; that is, $B \in G$, $\mu(B) > 0$, and if $A \in G$, $A \subset B$, then $\mu(A) = 0$ or $\mu(B - A) = 0$, then f is a.e. constant on B.

Proof: If $x \in \mathbb{R}$ and $\mu(\omega \in B : f(\omega) < x) = 0$, then $\mu(\omega \in B : f(\omega) < y) = 0$ for all $y \leq x$. Let $k = \sup\{x \in \mathbb{R} : \mu(\omega \in B : f(\omega) < x) = 0\}$. Then

$$\mu(\omega \in B : f(\omega) < k) = \mu\left(\bigcup_{\substack{r \text{ rational} \\ r<k}} \{\omega \in B : f(\omega) < r\}\right) = 0.$$

If $x > k$, then $\mu(\omega \in B : f(\omega) < x) > 0$; hence $\mu(\omega \in B : f(\omega) < x) = 0$ since B is an atom. Thus

$$\mu(\omega \in B : f(\omega) > k) = \mu\left(\bigcup_{\substack{r \text{ rational} \\ r>k}} \{\omega \in B : f(\omega) \geq r\}\right) = 0.$$

It follows that $f = k$ a.e. on B.

We now show that conditional expectation is an "averaging" or "smoothing" operation; if B is an atom of G, $E(Y|G) = k$ a.e. on B, where k is the average value of Y on B.

3.4.9 Theorem.

(a) Let B be an atom of G relative to P. Then

$$E(Y|G) = \frac{1}{P(B)} \int_B Y \, dP = \frac{E(Y\mathbb{I}_B)}{P(B)} \quad \text{a.e. on } B.$$

As a special case, let $B_1, B_2, \ldots$ be disjoint sets in F whose union is Ω, with $P(B_n) > 0$ for all n. Let G be the minimal σ-field over the B_n, so that G is the collection of all unions formed from the B_n. Then

$$E(Y|G) = \frac{1}{P(B_n)} \int_{B_n} Y \, dP \quad \text{a.e. on } B_n, \quad n = 1, 2, \ldots$$

(a') If $B = \{X = x_0\}$ and $P(B) > 0$, then

$$E(Y|X = x_0) = \frac{1}{P(B)} \int_B Y \, dP.$$

We now consider successive conditioning relative to two σ-fields, one of which is coarser than (that is, a subset of) the other. The result is that no matter which conditioning operation is applied first, the result is the same as the conditioning with respect to the coarser σ-field alone. This is intuitively reasonable; for example, to find the average value of a real-valued function f defined on $[0,3]$, we may compute a_1, the average of f on $[0,1]$, and a_2, the average of f on $[1,3]$; the average of f on $[0,3]$, namely, $\frac{1}{3}\int_0^3 f(x)\,dx$, is then $\frac{1}{3}a_1 + \frac{2}{3}a_2$.

3.4.10 Theorem

(a) If $G_1 \subset G_2$, then $E(E(Y|G_2)|G_1) = E(Y|G_1)$ a.e.

(a') If $f : (\Omega', J') \to (\Omega'', J'')$, then $E(E(Y|X)|f \circ X) = E(Y|f \circ X)$ a.e.

(b) If $G_1 \subset G_2$, then $E(E(Y|G_1)|G_2) = E(Y|G_1)$ a.e.

(b') If $f : (\Omega', J') \to (\Omega'', J'')$, then $E(E(Y|f \circ X)|X) = E(Y|f \circ X)$ a.e.

Proof: (a) If $C \in G_1$, then

$$\int_C E(Y|G_1)\,dP = \int_C Y\,dP = \int_C E(Y|G_2)\,dP$$

since $C \in G_2$. Thus $E(E(Y|G_2)|G_1) = E(Y|G_1)$ a.e.
Alternatively, if $C \in G_1$, then

$$\int_C E(E(Y|G_2)|G_1)\,dP = \int_C E(Y|G_2)\,dP = \int_C Y\,dP$$

since $C \in G_2$; thus $E(Y|G_1) = E(E(Y|G_2)|G_1)$ a.e.
(a') Let $G_2 = X^{-1}(J')$, $G_1 = [f \circ X]^{-1}(J'') = X^{-1}(f^{-1}(J''))$, and apply (a).
(b) $E(Y|G_1)$ is G_1-measurable, hence G_2-measurable, and

$$\int_C E(Y|G_1)\,dP = \int_C E(Y|G_1)\,dP, \qquad C \in rG_2.$$

(b') Take rG_1 and G_2 as in (a'), and apply (b).

If we take the conditional expectation of a product of two random variables, under certain conditions one of the terms can be factored out, as follows.

3.4.11 Theorem.

(a) If Z is G-measurable and both Y and YZ are integrable, then

$$E(YZ|G) = ZE(Y|G) \quad \text{a.e.}$$

In particular,

$$E(Z|G) = Z \quad \text{a.e.}$$

(a') If $f : (\Omega', J') \to (\mathbb{R}, B)$ and both Y and $Y(f \circ X)$ are integrable, then

$$E(Y(f \circ X) \mid X = x) = f(x)E(Y \mid X = x) \quad \text{a.e. } [P_X].$$

In particular,

$$E(f \circ X \mid X = x) = f(x) \quad \text{a.e. } [P_X].$$

Proof. (a) If Z is an indicator I_B, $B \in G$, and $C \in G$, we have

$$\begin{aligned}\int_C YZ\,dP = \int_C Y\,dP = \int_{C\cap B} Y\,dP = \int_C I_B E(Y|G)\,dP \\ = \int_C ZE(Y|G)\,dP.\end{aligned}$$

and $ZE(Y|G)$ is G-measurable. Thus the result holds for indicators.
Now let Z be simple, say

$$Z = \sum_{j=1}^{n} z_j I_{B_j} \quad \text{with} \quad B_j \in G.$$

By 3.4.2(a),

$$E(YZ|G) = \sum_{j=1}^{n} z_j E(YI_{B_j}|G) = \sum_{j=1}^{n} z_j I_{B_j} E(Y|G) = ZE(Y|G).$$

If Z is an arbitrary G-measurable function, let $Z_1, Z_2, \ldots$ be simple (and G-measurable) with $|Z_n| \le |Z|$ and $Z_n \to Z$. Now $E(YZ_n|G) = Z_n E(Y|G)$ by what we have just proved, and $E(YZ_n|G) \to E(YZ|G)$ by 3.4.5(a). (The integrability of YZ is used here.) Since Y is integrable, so is $E(Y|G)$; hence $E(Y|G)$ is finite a.e., and consequently $Z_n E(Y|G) \to ZE(Y|G)$ a.e. [Note that, for example, $1/n \to 0$ but $(1/n)(\infty) \nrightarrow 0(\infty) = 0$; thus finiteness of $E(Y|G)$ is important.] Therefore,

$$E(YZ|G) = ZE(Y|G).$$

(a') Let $f = I_B$, $B \in F$. Then

$$\int_{\{X\in A\}} Y(f\circ X)\,dP = \int_{\{X\in A\}} Y I_{f\circ X\in B}\,dP = \int_{\{X\in A\cap B\}} Y\,dP$$
$$= \int_{A\cap B} E(Y|X=x)\,dP_X(x) = \int_A f(x)E(Y|X=x)\,dP_X(x).$$

Thus the result holds when f is an indicator. Passage to simple functions and then to arbitrary measurable functions is carried out just as in (a).

Chapter 4
Martingale Theory

4.1 Martingales

Probability theory has its roots in games of chance, and it is often profitable to interpret results in terms of a gambling situation. For example, if $X_1, X_2, \ldots$ is a sequence of random variables, we may think of X_n as our total winnings after n trials in a succession of games. Having survived the first n trials, our expected fortune after trial $n+1$ is

$$\mathbb{E}(X_{n+1} \mid X_1, \ldots, X_n).$$

If this equals X_n, the game is "fair" since the expected gain on trial $n+1$ is

$$\mathbb{E}(X_{n+1} - X_n \mid X_1, \ldots, X_n) = \mathbb{E}(X_{n+1} \mid X_1, \ldots, X_n) - X_n = 0.$$

If $\mathbb{E}(X_{n+1} \mid X_1, \ldots, X_n) \geq X_n$, the game is "favorable," and if $\mathbb{E}(X_{n+1} \mid X_1, \ldots, X_n) \leq X_n$, the game is "unfavorable."

We are going to study sequences of this type; the results to be obtained will have significance outside the casino as well as inside.

Definitions: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\{X_1, X_2, \ldots\}$ a sequence of integrable random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, and $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$ an increasing sequence of sub σ-fields of $\mathcal{F}$; X_n is assumed $\mathcal{F}_n$-measurable, that is,

$$X_n : (\Omega, \mathcal{F}_n) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})).$$

The sequence $\{X_n\}$ is said to be a **martingale** relative to the $\mathcal{F}_n$ (alternatively, we say that $\{X_n, \mathcal{F}_n\}$ is a martingale) iff for all $n = 1, 2, \ldots$,

$$\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) = X_n \quad \text{a.e.},$$

a **submartingale** iff $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \geq X_n$ a.e., a **supermartingale** iff $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \leq X_n$ a.e. (In statements involving conditional expectations, the "a.e." is always understood and will usually be omitted.)

Let $\{\mathcal{F}_n\}$ be a *decreasing* sequence of sub σ-fields of $\mathcal{F}$, with X_n assumed $\mathcal{F}_n$-measurable. If $\mathbb{E}(X_n \mid \mathcal{F}_{n+1}) = X_{n+1}$, we say that $\{X_n, \mathcal{F}_n\}$ is a **reverse martingale**. Similarly, $\mathbb{E}(X_n \mid \mathcal{F}_{n+1}) \geq X_{n+1}$ defines a **reverse submartingale**, and $\mathbb{E}(X_n \mid \mathcal{F}_{n+1}) \leq X_{n+1}$ defines a **reverse supermartingale**.

4.1.1 Properties

(a) If $\{X_n, \mathcal{F}_n\}$ is a martingale, then

$$\mathbb{E}(X_{n+k} \mid \mathcal{F}_n) = X_n, \quad n, k = 1, 2, \ldots$$

(with corresponding statements for sub- and supermartingales). For

$$\mathbb{E}(X_{n+2} \mid \mathcal{F}_n) = \mathbb{E}[\mathbb{E}(X_{n+2} \mid \mathcal{F}_{n+1}) \mid \mathcal{F}_n] \quad \text{by 3.4.10(a)}$$

$$= \mathbb{E}(X_{n+1} \mid \mathcal{F}_n) = X_n.$$

The general statement follows by induction.

(b) If $\{X_n, F_n\}$ is a martingale, then

$$E(X_{n+1} \mid X_1, \ldots, X_n) = X_n, \quad n = 1, 2, \ldots$$

Thus $\{X_n\}$ is automatically a martingale relative to the standard σ-fields $\sigma(X_1, \ldots, X_n)$ (with corresponding statements for sub- and supermartingales).

For $F_1 \subset \cdots \subset F_n$, and thus $X_1, \ldots, X_n$ are all F_n-measurable. Since $\sigma(X_1, \ldots, X_n)$ is the smallest σ-field making $X_1, \ldots, X_n$ measurable [see 3.3.1(b)], we have $\sigma(X_1, \ldots, X_n) \subset F_n$. If, in the defining relation

$$E(X_{n+1} \mid F_n) = X_n,$$

we take conditional expectations with respect to $\sigma(X_1, \ldots, X_n)$, we obtain the desired result by 3.4.10(a) and 3.4.11(a).

If we say that $\{X_n\}$ is a martingale (or sub-, supermartingale) without mentioning the σ-fields F_n, we shall always mean $F_n = \sigma(X_1, \ldots, X_n)$, so that

$$E(X_{n+1} \mid X_1, \ldots, X_n) = X_n.$$

(c) $\{X_n, F_n\}$ is a martingale iff

$$\int_A X_n \, dP = \int_A X_{n+1} \, dP \quad \text{for all } A \in F_n, \quad n = 1, 2, \ldots$$

This follows since the condition $E(X_{n+1} \mid F_n) = X_n$ a.e. $[P]$ is equivalent to

$$\int_A E(X_{n+1} \mid F_n)\, dP = \int_A X_n\, dP \quad \text{for all } A \in F_n$$

also

$$\int_A E(X_{n+1} \mid F_n)\, dP = \int_A X_{n+1}\, dP$$

by definition of conditional expectation.
Similarly, $\{X_n, F_n\}$ is a submartingale iff

$$\int_A X_n\, dP \leq \int_A X_{n+1}\, dP \quad \text{for all } A \in F_n,$$

and a supermartingale iff

$$\int_A X_n\, dP \geq \int_A X_{n+1}\, dP \quad \text{for all } A \in F_n.$$

In particular, $E(X_n)$ is constant in a martingale, increases in a submartingale, and decreases in a supermartingale.

(d) The defining condition for a martingale relative to the σ-fields $\sigma(X_1, \ldots, X_n)$ is equivalent to

$$E(X_{n+1} \mid X_1 = x_1, \ldots, X_n = x_n) = x_n \quad \text{a.e. } [P_{(X_1,\ldots,X_n)}]$$

with similar statements for sub- and supermartingales.
For if $A \in \sigma(X_1, \ldots, X_n)$, then A is of the form $\{(X_1, \ldots, X_n) \in B\}$, $B \in B(\mathbb{R}^n)$. If $X = (X_1, \ldots, X_n)$, then

$$\int_A X_{n+1}\, dP = \int_{X \in B} X_{n+1}\, dP = \int_B E(X_{n+1} \mid X_1 = x_1, \ldots, X_n = x_n)\, dP_X$$

by 3.2.3, and

$$\int_A X_n\, dP = \int_B E(X_n \mid X_1 = x_1, \ldots, X_n = x_n)\, dP_X = \int_B x_n\, dP_X \quad \text{by 3.4.11(a).}$$

The result now follows from (c).

(e) A finite sequence $\{X_k, F_k, k = 1, \ldots, n\}$ is called a martingale iff

$$E(X_{k+1} \mid F_k) = X_k, \quad k = 1, 2, \ldots, n-1;$$

finite sub- and supermartingale sequences are defined similarly.

(f) If $\{X_n, F_n\}$ and $\{Y_n, F_n\}$ are submartingales, so is $\{\max(X_n, Y_n), F_n\}$. For $E(\max(X_{n+1}, Y_{n+1}) \mid F_n) \geq E(X_{n+1} \mid F_n) \geq X_n$, and similarly

$$E(\max(X_{n+1}, Y_{n+1}) \mid F_n) \geq Y_n.$$

The same approach shows that if $\{X_n, F_n\}$ and $\{Y_n, F_n\}$ are supermartingales, so is $\{\min(X_n, Y_n), F_n\}$.

4.1.2 Examples

If $X_n \equiv X$, then $\{X_n\}$ is a martingale; if $X_1 \leq X_2 \leq \cdots$, then $\{X_n\}$ is a submartingale; if $X_1 \geq X_2 \geq \cdots$, then $\{X_n\}$ is a supermartingale (assuming all random variables are integrable).
We give some more substantial examples.

(a) Let $Y_1, Y_2, \ldots$ be independent random variables with zero mean, and set $X_n = \sum_{k=1}^{n} Y_k$, $F_n = \sigma(Y_1, \ldots, Y_n)$. Then $\{X_n, F_n\}$ is a martingale. For

$$E(X_{n+1} \mid F_n) = E(X_n + Y_{n+1} \mid Y_1, \ldots, Y_n) = X_n + E(Y_{n+1} \mid Y_1, \ldots, Y_n)$$

$$\text{since } X_n \text{ is } F_n\text{-measurable} = X_n + E(Y_{n+1}) = X_n \quad \text{(by independence)}$$

$$\text{since } E(Y_j) \equiv 0.$$

(b) Let $Y_1, Y_2, \ldots$ be independent random variables with $E(Y_j) = a_j \neq 0$, and set

$$X_n = \prod_{j=1}^{n} \left(\frac{Y_j}{a_j}\right), \quad F_n = \sigma(Y_1, \ldots, Y_n).$$

Then $\{X_n, F_n\}$ is a martingale. For

$$E(X_{n+1} \mid F_n) = E\left(\frac{X_n Y_{n+1}}{a_{n+1}} \middle| Y_1, \ldots, Y_n\right) = X_n E\left(\frac{Y_{n+1}}{a_{n+1}} \middle| Y_1, \ldots, Y_n\right)$$

$$\text{by 3.4.11(a)} = X_n.$$

(c) Let Y be an integrable random variable on (Ω, F, P).
If $\{F_n\}$ is an *increasing* sequence of sub-σ-fields of F, and $X_n = E(Y \mid F_n)$, then $\{X_n, F_n\}$ is a martingale. For

$$E(X_{n+1} \mid F_n) = E(E(Y \mid F_{n+1}) \mid F_n) = E(Y \mid F_n) \quad \text{since } F_n \subset F_{n+1} = X_n.$$

If $\{F_n\}$ is a decreasing sequence of sub-σ-fields of F, and $X_n = E(Y \mid F_n)$, then $\{X_n, F_n\}$ is a **reverse martingale**. For

$$E(X_n \mid F_{n+1}) = E(E(Y \mid F_n) \mid F_{n+1}) = E(Y \mid F_{n+1}) \quad \text{since } F_{n+1} \subset F_n = X_{n+1}.$$

Note that as in 4.1.1(a), $E(X_n \mid F_{n+k}) = X_{n+k},\ n, k = 1, 2, \ldots$.

(d) (Branching Processes) We define a Markov chain with state space $S = \{0, 1, 2, \ldots\}$. The state at time n, denoted by X_n, is to represent the number of offspring after n generations. $X_n = k, X_{n+1}$ is the sum of k independent, identically distributed, nonnegative integer valued random variables, say $Y_1, \ldots, Y_k$, where $P(Y_i = l) = p_l$, $l = 0, 1, 2, \ldots$. Thus p_l is the probability that a given being will produce exactly l offspring. (Formally, we take $p_{kj} = P\{Y_1 + \cdots + Y_k = j\}, k = 1, 2, \ldots, j = 0, 1, \ldots; p_{00} = 1.$)
Let $m = E(Y_i) = \sum_{l=0}^{\infty} l p_l$. If m is finite and greater than 0, then $\{X_n/m^n\}$ is a martingale relative to

$$\mathcal{F}_n = \sigma(X_0, \ldots, X_n) = \sigma\left(\frac{X_1}{m}, \frac{X_2}{m^2}, \ldots, \frac{X_n}{m^n}\right).$$

For

$$E\left(\frac{X_{n+1}}{m^{n+1}}\middle| X_0 = i_0, \ldots, X_n = i_n\right) = \sum_{j=0}^{\infty} p_{i_n j} \frac{j}{m^{n+1}}$$

(see 3.2.5(a), Eq. (2) and the definition of a Markov chain)

$$= \frac{1}{m^{n+1}} \sum_{j=0}^{\infty} j P\{Y_1 + \cdots + Y_{i_n} = j\}$$

$$= \frac{1}{m^{n+1}} E(Y_1 + \cdots + Y_{i_n})$$

$$= \frac{i_n m}{m^{n+1}} = \frac{i_n}{m^n}.$$

The result now follows from 4.1.1(d).

(e) Consider the branching process of part (d). Let $g(s) = \sum_j p_j s^j, s \geq 0$. If for some r we have $g(r) = r$, then $\{r^{X_n}\}$ is a martingale relative to the σ-fields $\mathcal{F}(X_0, \ldots, X_n)$. For as in (d),

$$E(r^{X_{n+1}} | X_0 = i_0, \ldots, X_n = i_n) = \sum_{j=0}^{\infty} p_{i_n j} r^j$$

$$= \sum_{j=0}^{\infty} r^j P\{Y_1 + \cdots + Y_{i_n} = j\}$$

$$= E[\exp(Y_1 + \cdots + Y_{i_n})] = [E(r^Y)]^{i_n} = [g(r)]^{i_n} = r^{i_n}.$$

4.2 Optional Skipping Theorem

Statement: Let $\{X_n, \mathcal{F}_n\}$ be a submartingale. Let $\varepsilon_1, \varepsilon_2, \ldots$ be random variables defined by

$$\varepsilon_k = \begin{cases} 1 & \text{if } (X_1, \ldots, X_k) \in B_k, \\ 0 & \text{if } (X_1, \ldots, X_k) \notin B_k, \end{cases}$$

where the B_k are arbitrary sets in $\mathcal{B}(\mathbb{R}^n)$. Set

$$\begin{aligned} Y_1 &= X_1, \\ Y_2 &= X_1 + \varepsilon_1(X_2 - X_1), \\ &\vdots \\ Y_n &= X_1 + \varepsilon_1(X_2 - X_1) + \cdots + \varepsilon_{n-1}(X_n - X_{n-1}). \end{aligned}$$

Then $\{Y_n, \mathcal{F}_n\}$ is also a submartingale and $E(Y_n) \leq E(X_n)$ for all n. If $\{X_n, \mathcal{F}_n\}$ is a martingale, so is $\{Y_n, \mathcal{F}_n\}$ and $E(Y_n) = E(X_n)$ for all n.

Interpretation: Let X_n be the gambler's fortune after n trials; then Y_n is our fortune if we follow an optional skipping strategy. After observing $X_1, \ldots, X_k$, we may choose to bet with the gambler at trial $k+1$ [in this case $\varepsilon_k = \varepsilon_k(X_1, \ldots, X_k) = 1$] or we may pass ($\varepsilon_k = 0$). Our gain on trial $k+1$ is $\varepsilon_k(X_{k+1} - X_k)$. The theorem states that whatever strategy we employ, if the game is initially "fair" (a martingale) or "favorable" (a submartingale), it remains fair (or favorable), and no strategy of this type can increase the expected winning.

Proof:

$$\begin{aligned} E(Y_{n+1}|\mathcal{F}_n) &= E(Y_n + \varepsilon_n(X_{n+1} - X_n)|\mathcal{F}_n) \\ &= Y_n + \varepsilon_n E(X_{n+1} - X_n|\mathcal{F}_n) \end{aligned}$$

since ε_n is a Borel measurable function of $X_1, \ldots, X_n$, and hence is

$$\sigma(X_1, \ldots, X_n) \subset \mathcal{F}_n\text{-measurable.}$$

Therefore

$$E(Y_{n+1}|\mathcal{F}_n) = Y_n + \varepsilon_n(X_n - X_n) = Y_n \quad \text{in the martingale case}$$

$$\geq Y_n + \varepsilon_n(X_n - X_n) = Y_n \quad \text{in the submartingale case.}$$

Since $Y_1 = X_1$, we have $E(X_1) = E(Y_1)$. Having shown $E(X_k - Y_k) \geq 0$ (= 0 in the martingale case),

$$\begin{aligned} X_{k+1} - Y_{k+1} &= X_{k+1} - Y_k - \varepsilon_k(X_{k+1} - X_k) \\ &= (1 - \varepsilon_k)(X_{k+1} - X_k) + X_k - Y_k. \end{aligned}$$

Thus

$$\begin{aligned} E(X_{k+1} - Y_{k+1}|\mathcal{F}_k) &= (1 - \varepsilon_k)E(X_{k+1} - X_k|\mathcal{F}_k) + E(X_k - Y_k|\mathcal{F}_k) \\ &\geq E(X_k - Y_k|\mathcal{F}_k) = X_k - Y_k, \end{aligned}$$

with equality in the martingale case. Take expectations and use $E[E(X|\mathcal{G})] = E(X)$ to obtain

$$E(X_{k+1} - Y_{k+1}) \geq E(X_k - Y_k) \geq 0,$$

with equality in the martingale case.

4.3 Optional Sampling Theorems

Let $\{X_n, n = 1, 2, \ldots\}$ be a martingale, with X_n interpreted as a gambler's total capital after n plays of a game of chance. Suppose that after each trial, the gambler decides either to quit or to keep playing. If T is the time of quitting, what can be said about the final capital X_T?

First of all, the random variable T must have the property that if we observe $X_1, \ldots, X_n$, we can come to a definite decision as to whether or not $T = n$. A nonnegative random variable of this type is called a *stopping time*.

Definition. Let $\{F_n, n = 0, 1, \ldots\}$ be an increasing sequence of sub σ-fields of F. A **stopping time** for the F_n is a map $T\colon \Omega \to \{0, 1, \ldots, \infty\}$ such that $\{T \leq n\} \in F_n$ for each nonnegative integer n. Since $\{T = n\} = \{T \leq n\} - \{T \leq n-1\}$ and $\{T \leq n\} = \bigcup_{k=0}^{n} \{T = k\}$, the definition is equivalent to

the requirement that $\{T = n\} \in F_n$ for all $n = 0, 1, \ldots$. If $\{X_n, n = 0, 1, \ldots\}$ is a sequence of random variables, a stopping time for $\{X_n\}$ is, by definition, a stopping time relative to the σ-fields $F_n = \sigma(X_0, \ldots, X_n)$.

If S and T are stopping times, so are $S \vee T = \max(S, T)$ and $S \wedge T = \min(S, T)$. ($\{S \vee T \leq n\} = \{S \leq n\} \cap \{T \leq n\}$, $\{S \wedge T \leq n\} = \{S \leq n\} \cup \{T \leq n\}$). Also, if $T \equiv n$ then T is a stopping time.

By far the most important example of a stopping time is the **hitting time** of a set. If $\{X_n\}$ is a sequence of random variables and $B \in B(\mathbb{R})$, let $T(\omega) = \min\{n\colon X_n(\omega) \in B\}$ if $X_n(\omega) \in B$ for some n; $T(\omega) = \infty$ if $X_n(\omega)$ is never in B. T is a stopping time since $\{T \leq n\} = \bigcup_{k \leq n}\{X_k \in B\} \in F(X_k, k \leq n)$.

If T is a stopping time for $\{X_n\}$, an event A is said to be "prior to T" iff, whenever $T = n$, we can tell by examination of the $X_k, k \leq n$, whether or not A has occurred. The formal definition is as follows.

Definition. Let T be a stopping time for the σ-fields $F_n, n = 0, 1, \ldots$, and let A belong to F. The set A is said to be *prior to* T iff $A \cap \{T \leq n\} \in F_n$ for all $n = 0, 1, \ldots$ [Equivalently, $A \cap \{T = n\} \in F_n$ for all $n = 0, 1, \ldots$]. The collection of all sets prior to T will be denoted by F_T; it follows quickly that F_T is a σ-field. Also, if $T \equiv n$ then F_T is simply F_n.

If S and T are stopping times and $S \leq T$, then $F_S \subset F_T$. For if $A \in F_S$ then

$$A \cap \{T \leq k\} = \bigcup_{i=1}^{k} [A \cap \{S = i\}] \cap \{T \leq k\}.$$

But $A \cap \{S = i\} \in F_i \subset F_k$, and $\{T \leq k\} \in F_k$; hence $A \in F_T$.

If the stopping time T is constant at n, then X_T is F_T-measurable. We would like this idea to carry over to a general stopping time. Formally, let T be a finite stopping time for the σ-fields F_n, and define X_T in the natural way; if $T(\omega) = n$, let $X_T(\omega) = X_n(\omega)$. If $B \in B(\mathbb{R})$, then $\{X_T \in B\} \in F_T$, in other words, X_T is F_T-measurable. (Since $F_T \subset F$ by definition, it follows in particular that X_T is a random variable.) To see this, write

$$\{X_T \in B\} \cap \{T \leq n\} = \bigcup_{k=0}^{n} [\{X_k \in B\} \cap \{T = k\}].$$

Since $\{X_k \in B\} \cap \{T = k\} \in F_k$ for $k \leq n$, we have

$$\{X_T \in B\} \cap \{T \leq n\} \in F_n.$$

Also, as T is finite, we have $\bigcup_{n=0}^{\infty}\{T \le n\} = \Omega$, so that $\{X_T \in B\} \in F_T$, as desired.
If T is not necessarily finite, the same argument shows that $1_{\{T<\infty\}}X_T$ is F_T-measurable.

Now in the gambling situation described at the beginning of the section, a basic quantity of interest is $E(X_T)$, the average accumulation at the quitting time. For example, if $E(X_T)$ turns out to be the same as $E(X_1)[= E(X_n)$ for all n by the martingale property], the gambler's strategy does not offer any improvement over the procedure of stopping at a fixed time. Now in comparing X_1 and X_T we are considering two stopping times S and T ($S \equiv 1$) with $S \le T$, and looking at X_S versus X_T. More generally, if $T_1 \le T_2 \le \cdots$ form an increasing sequence of finite stopping times, we may examine the sequence $X_{T_1}, X_{T_2}, \ldots$. If the sequence forms a martingale, then $E(X_{T_n}) = E(X_{T_1})$ for all n, and if $T_1 \equiv 1$, then $E(X_{T_n}) = E(X_1)$.

Thus if we sample the gambler's fortune at random times $T_1, T_2, \ldots$, the basic question is whether the martingale (or submartingale) property is preserved. This will always be the case when the sequence $\{X_n\}$ is finite.

4.3.1 Theorem.

Let $\{X_n, F_n, n = 1, \ldots, m\}$ be a submartingale, and let $T_1, T_2, \ldots$ be an increasing sequence of stopping times for the F_n. Then $\{X_{T_n}\}$ is a submartingale. [In other words, the T_n take values in $\{1, \ldots, m\}$, and $\{T_n \le k\} \in F_k$, $k = 1, \ldots, m$.] The σ-fields F_{T_n} are defined as before:

$$F_{T_n} = \{A \in F : A \cap \{T_i \le k\} \in F_k,\ k = 1, \ldots, m\}.$$

Then the X_{T_n} form a submartingale relative to the σ-fields F_{T_n}, a martingale if $\{X_n\}$ is a martingale.

Proof: We follow Breiman (1968). Define $Y_n = X_{T_n}$, and note that each Y_n is integrable:

$$\int_\Omega |X_{T_n}|\, dP = \sum_{i=1}^{m} \int_{\{T_n=i\}} |X_i|\, dP \le \sum_{i=1}^{m} E(|X_i|) < \infty.$$

As the T_n increase with n, so do the F_{T_n} .
Now if $A \in F_{T_n}$, we must show that $\int_A Y_{n+1}\, dP \ge \int_A Y_n\, dP$ (with equality in the martingale case). Since $A = \bigcup_j [A \cap \{T_n = j\}]$, it suffices to replace A by $D_j = A \cap \{T_n = j\}$, which belongs to F_j; now if $k > j$, we note that $T_n = j$ implies $T_{n+1} \ge j$, so that

$$\int_{D_j} Y_{n+1}\,dP = \sum_{i=j}^{k} \int_{D_j\cap\{T_{n+1}=i\}} Y_{n+1}\,dP + \int_{D_j\cap\{T_{n+1}>k\}} Y_{n+1}\,dP.$$

Thus

$$\int_{D_j} Y_{n+1}\,dP = \sum_{i=j}^{k} \int_{D_j\cap\{T_{n+1}=i\}} X_i\,dP + \int_{D_j\cap\{T_{n+1}>k\}} X_k\,dP$$

$$- \int_{D_j\cap\{T_{n+1}>k\}} (X_k - Y_{n+1})\,dP. \tag{1}$$

Now combine the $i = k$ term in (1) with the $\int X_k\,dP$ term to obtain

$$\int_{D_j\cap\{T_{n+1}=k\}} X_k\,dP + \int_{D_j\cap\{T_{n+1}>k\}} X_k\,dP$$

$$= \int_{D_j\cap\{T_{n+1}\geq k\}} X_k\,dP \geq \int_{D_j\cap\{T_{n+1}\geq k\}} X_{k-1}\,dP$$

since $\{T_{n+1} \geq k\} = \{T_{n+1} \leq k-1\}^c \in F_{k-1}$ and $D_j \in F_j \subset F_{k-1}$. But

$$\int_{D_j\cap\{T_{n+1}\geq k\}} X_{k-1}\,dP = \int_{D_j\cap\{T_{n+1}>k-1\}} X_{k-1}\,dP,$$

so this term may be combined with the $i = k-1$ term of (1) to obtain

$$\int_{D_j\cap\{T_{n+1}>k-2\}} X_{k-2}\,dP.$$

Proceeding inductively, we find

$$\int_{D_j} Y_{n+1}\,dP \geq \int_{D_j\cap\{T_{n+1}\geq j\}} X_j\,dP - \int_{D_j\cap\{T_{n+1}>k\}} (X_k - Y_{n+1})\,dP. \tag{2}$$

Now $\{T_{n+1} > k\}$ is empty for $k \geq m$. Finally, $D_j \cap \{T_{n+1} \geq j\} = D_j$ since $D_j \subset \{T_n = j\}$, and $X_j = Y_n$ on D_j. Thus

$$\int_{D_j} Y_{n+1}\,dP \geq \int_{D_j} Y_n\,dP$$

as desired. In the martingale case, all inequalities in the proof become equalities.

Theorem 4.3.1 extends immediately to the case of an infinite sequence if each T_n is bounded, that is, for each n there is a positive constant K_n such that $T_n \leq K_n$ a.e. The same proof may be used; the key point is that $\{T_{n+1} > k\}$ is still empty for sufficiently large k.

When $\{X_n\}$ is an infinite sequence, the martingale or submartingale property is not preserved in general, but the following result gives useful sufficient conditions.

4.3.2 Optional Sampling Theorem.

Let $\{X_1, X_2, \ldots\}$ be a submartingale, and let $T_1, T_2, \ldots$ be an increasing sequence of finite stopping times for $\{X_n\}$, with $Y_n = X_{T_n}$, $n = 1, 2, \ldots$. If

(a) $E(|X_n|) < \infty$ for all n, and

(b) $\lim_{k\to\infty} \inf_n \int_{\{T_n > k\}} |X_k|\, dP = 0$ for all n,

then $\{Y_n\}$ is a submartingale relative to the σ-fields F_{T_n}. If $\{X_n\}$ is a martingale, so is $\{Y_n\}$.

Proof. Since integrability of the Y_n is now hypothesis (a), we can follow the proof of 4.3.1 to (2). The first integral on the right-hand side is $\int_{D_j} Y_n\, dP$ as before, but in the second integral, we no longer have $\{T_{n+1} > k\}$ empty for large k. But by hypothesis (B), $\int_{D_j \cap \{T_{n+1} > k\}} X_k\, dP \to 0$ as $k \to \infty$ through an appropriate subsequence, and $\int_{D_j \cap \{T_{n+1} > k\}} Y_{n+1}\, dP \to 0$ as $k \to \infty$ since $\{T_{n+1} > k\}$ decreases to the empty set. Thus

$$\int_{D_j} Y_{n+1}\, dP \geq \int_{D_j} Y_n\, dP$$

as desired. As before, all inequalities become equalities in the martingale case.

If $\{X_n\}$ is a submartingale with a last element X_∞, we can define the random variable X_T for any stopping time T. On the set $\{T = \infty\}$, $X_T = X_\infty$. In this case, the optional sampling theorem holds.

4.4 Applications to Markov Chains

In this section we apply martingale theory to the problem of classifying the states of a Markov chain. We must use a few basic properties of Markov

chains. In particular, a state i is said to be **recurrent** iff starting at i there will be a return to i with probability 1; otherwise the state is *transient*. If C is a set of states such that every state in C can be reached (in a finite number of steps) from every other state, then all states in C are of the same type, recurrent or transient.

4.4.1 Theorem.

Let $[p_{ij}]$ be the transition matrix of a Markov chain such that every state in the state space S is reachable from every other state (sometimes called an **irreducible chain**). Choose a fixed state, and label it 0 for convenience. The states are transient iff there is a nonconstant bounded $f\colon S \to \mathbb{R}$ such that

$$\sum_{j\in S} p_{ij} f(j) = f(i) \quad \text{for all } i \neq 0.$$

Proof: Suppose such an f exists. By adding a constant to f we may assume that $f \geq 0$. Assume the initial state is $i \neq 0$, and let $\{X_n\}$ be the corresponding sequence of random variables. Let T be the time at which 0 is reached, and let $Y_n = X_{T\wedge n},\ n = 0, 1, \ldots$; $\{Y_n\}$ can be realized as a Markov chain with the same initial distribution and transition matrix as $\{X_n\}$, except that 0 is now an absorbing state. In other words, the transition matrix for $\{Y_n\}$ is

$$\hat{p}_{ij} = p_{ij} \quad \text{for all } j \quad \text{if } i \neq 0, \hat{p}_{00} = 1.$$

Thus $\sum_{j\in S} \hat{p}_{ij}^{(n)} f(j) = f(i)$ for all $i \in S$. In matrix form, $\hat{P}^n f = f$; by induction $\hat{P}^n f = f$, that is,

$$\sum_{j\in S} \hat{p}_{ij}^{(n)} f(j) = f(i),$$

where $\hat{p}_{ij}^{(n)} = P(Y_n = j \mid Y_0 = i)$. But this says that $E[f(Y_n) \mid Y_0 = i] = f(i)$. If the states of the original chain are recurrent, then 0 will be visited with probability 1; hence $Y_n \to 0$ a.e. By the dominated convergence theorem, $E[f(Y_n) \mid Y_0 = i] \to f(0)$. We conclude that $f(i) = f(0)$ for all i, contradicting the hypothesis that f is nonconstant.

Conversely, if the states are transient, we define $f\colon S \to \mathbb{R}$ as follows. If $i \neq 0$, let $f(i) = f_i^0$, the probability that, starting from i, 0 will eventually be reached; take $f(0) = 1$. Now in order ultimately to reach 0 from $i \neq 0$, we may either go directly to 0 at step 1, or go to a state $j \neq 0$ and then reach 0 at some time after the first step. It follows that

$$f(i) = \sum_{j\in S} p_{ij} f(j), \quad i \neq 0.$$

(This may be formalized using the Markov property.)
Now f is clearly bounded, and $f_i^0 < 1$ for some $i \neq 0$, otherwise 0 would be a recurrent state. Thus f is nonconstant.
Martingale theory is used in deriving the following sufficient condition for recurrence.

4.4.2 Theorem.

Let $[p_{ij}]$ be the transition matrix of an irreducible Markov chain whose state space S is the set of nonnegative integers. If there is a function $f\colon S \to \mathbb{R}$ such that $f(i) \to \infty$ as $i \to \infty$, and $\sum_j p_{ij} f(j) \leq f(i)$ for all $i \neq 0$, then the chain is recurrent.

Proof: As $f(i) \to \infty$ as $i \to \infty$, f is bounded below, so without loss of generality we may assume $f \geq 0$. Let the initial state be $i \neq 0$, and form the process $\{Y_n\}$ as in 6.9.1. The $\sum_j \hat{p}_{ij} f(j) \leq f(i)$ for all i, which implies that $\{f(Y_n)\}$ is a nonnegative supermartingale, and hence converges a.e. to a finite limit. For

$$E[f(Y_n) \mid Y_0 = i_0, \ldots, Y_{n-1} = i_{n-1}] = E[f(Y_n) \mid Y_{n-1} = i_{n-1}]$$

by the Markov property

$$= \sum_j \hat{p}_{i_{n-1}j} f(j) \leq f(i_{n-1}).$$

Note also that

$$E[f(Y_n) \mid Y_0 = i] = \sum_{j \in S} \hat{p}_{ij}^{(n)} f(j) \leq f(i) < \infty;$$

hence $\{f(Y_n)\}$ is integrable.
Assume the states transient. Then $f_i^0 < 1$ for some $i > 0$. Choose such an i as the initial state. Now $X_n \to \infty$ a.e. since a finite set of transient states cannot be visited infinitely often. This means that with probability 1, $Y_n \to 0$ or ∞. But $P(Y_n \to 0) = f_i^0 < 1$, and hence $P(Y_n \to \infty) > 0$; this implies that $P(f(Y_n) \to \infty) > 0$, a contradiction.
The proof of 4.4.2 shows that if $[p_{ij}]$ is the transition matrix of a Markov chain $\{X_n\}$, f is a real-valued function on the state space, and

$$\sum_j p_{ij} f(j) \leq f(i) \quad \text{for all } i,$$

where the series is assumed to converge absolutely, then, with a fixed initial state i, the sequence $\{f(X_n)\}$ is a supermartingale. Similarly, replacement of "$\leq$" by "$=$" in this equation yields a martingale, and replacement by "$\geq$" yields a submartingale.
We now apply 4.4.1 and 4.4.2 to a queueing process.

4.4.3 Example

Assume that customers are to be served at discrete times $t = 0, 1, \ldots$, and at most one customer can be served at a given time. Say there are X_t customers before the completion of service at time t, and in the interval $[t, t+1)$, Y_t new customers arrive, where $P(Y_t = k) = p_k$, $k = 0, 1, \ldots$. The number of customers before completion of service at time $t+1$ is

$$X_{t+1} = (X_t - 1)^+ + Y_t.$$

The queueing process may be represented as a Markov chain whose state space is the set of nonnegative integers and whose transition matrix is

$$\Pi = \begin{bmatrix} p_0 & p_1 & p_2 & \cdots & & \\ p_0 & p_1 & p_2 & \cdots & & \\ 0 & p_0 & p_1 & p_2 & \cdots & \\ 0 & 0 & p_0 & p_1 & p_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots & & \end{bmatrix}.$$

We assume that $p_0 > 0$ and $p_0 + p_1 < 1$, so that the chain is irreducible. The equations $\sum_{j=0}^{\infty} p_{ij} f(j) = f(i)$, $i > 0$, are

$$p_0 f(i-1) + p_1 f(i) + p_2 f(i+1) + \cdots + p_n f(i+n-1) + \cdots = f(i). \quad (1)$$

Let $m = E(Y) = \sum_{k=1}^{\infty} k p_k$; we show that the states are transient if $m > 1$, recurrent if $m \leq 1$.

First assume $m > 1$; if $f(i) = r^i$, then (1) becomes

$$p_0 r^{i-1} + p_1 r^i + p_2 r^{i+1} + \cdots + p_n r^{i+n-1} + \cdots = r^i,$$

or

$$\sum_{k=0}^{\infty} p_k r^k = r.$$

But this can be satisfied for some $r \in (0, 1)$. Thus $\{r^i\}$ is bounded and nonconstant, so by 4.4.1, the states are transient.

Now assume $m \leq 1$, and let $f(i) = i$. Then if $i > 0$,

$$\begin{aligned}\sum_{j=0}^{\infty} p_{ij} f(j) &= p_0(i-1) + p_1 i + p_2(i+1) + \cdots \\ &= \sum_{k=i-1}^{\infty} k p_{k-i+1} \\ &= \sum_{k=i-1}^{\infty} (k-i+1) p_{k-i+1} + i - 1 \\ &= \sum_{k=0}^{\infty} k p_k + i - 1 \leq 1 + i - 1 = i = f(i).\end{aligned}$$

By 4.4.2, the states are recurrent.

If i is a recurrent state and μ_i is the average length of time required to return to i when the initial state is i, then i is said to be **recurrent null** if $\mu_i = \infty$, **recurrent positive** if $\mu_i < \infty$. It can be shown that the states are recurrent null if $m = 1$, recurrent positive if $m < 1$.

References and Links

1. Probability and Measure by Patrick Billingsley,
2. Probability: Theory and Examples by Rick Durrett,
3. Probability with Martingale by David Williams,
4. Probability and Measure Theory by Robert B. Ash and Catherine A. Doleans-Dade,
5. Measure Theory and Integration by G. de Bara,
6. Video lectures of Prof. Joydeep Dutta, Department of Economic Sciences, Indian Institute of Technology, Kanpur,
7. Martingale in Gambling and Finance by Prakhar Saxena.