ACTEX Learning

INV 201

Quantitative Finance Study Manual

1st Edition

Yiping Guo, Ph.D.

An SOA Exam

ACTEX Learning

Learn Today. Lead Tomorrow.

Actuarial & Financial Risk Resource Materials

Since 1972

Printed in the United States of America.

PREFACE

About INV 201

This study manual has been developed for candidates preparing for the SOA's new Fellowship exam, **INV 201: Quantitative Finance**, first administered in Fall 2025. INV 201 replaces the former QFIQF exam in the Quantitative Finance and Investment track, with a revised syllabus that reflects both continuity and modernization. While the broad spirit of the two exams overlaps, INV 201 has been redesigned with different focuses and coverage on the mathematical foundations of derivative pricing and risk management.

As the title suggests, this exam is not merely about finance in a general sense but about quantitative finance, which is a discipline grounded in mathematics. The word "quantitative" captures the essence of the exam far more precisely than "finance." To perform well, candidates must have strong mathematical knowledge and practice. There is almost no question that can be solved by memorization alone. Among the FSA exams, INV 201 stands out for its heavy emphasis on mathematics, and this should be clearly understood from the outset.

The syllabus is organized into three broad areas, though as a new exam, it remains subject to adjustments by the SOA. Should major changes occur, every effort will be made to update the manual accordingly.

- **Topic I – Key Types of Derivatives (5%–15%)**:

  This topic covers the fundamental properties and payoff functions of instruments such as forwards, futures, swaps and options. While this section has lighter weight in the exam, mastering it is crucial, as it provides the foundation for the valuation and applications that follow.

- **Topic II – Valuation of Derivatives (40%–60%)**:

  This topic forms the heart of the exam. Here, candidates will encounter the principles of no-arbitrage and replication, stochastic calculus, martingale methods, and risk-neutral valuation. This is the central area of modern quantitative finance, and candidates should expect multiple questions spanning from mathematical foundation to pricing methodology of a wide range of derivatives.

- **Topic III – Applications and Risks of Derivatives (30%–50%)**:

  This topic builds naturally on valuation. It emphasizes how derivatives are used for hedging and risk management, requiring a clear understanding of Greeks, embedded options, and variable annuities.

About the Study Manual

Because of the mathematical intensity of INV 201, this manual takes a different approach than is typical for earlier exams such as FM or FAM. At the Associate level, the required knowledge base is relatively narrow, and often working through past exam questions is sufficient for preparation. INV 201 is different: it is built on modern probability theory, much of which goes beyond standard undergraduate education. As such, this manual starts from first principles, building the necessary mathematical framework before progressing to exam-level problems. The treatment of mathematics here is serious but not rigorously formal. We will not pursue general proofs for their own sake, but rather develop a semi-formal, exam-ready understanding of the tools.

The manual follows the exam syllabus closely, with careful attention to building both financial and mathematical intuition. Concepts are explained step by step, with visualizations introduced where they strengthen understanding. Worked examples appear throughout the chapters, and practice problems are provided at the end of each chapter to help consolidate learning.

Although SOA's course strategy guide often places INV 201 after "INV 101: Portfolio Management", the two exams differ greatly in style and substance. INV 101 focuses more on investment concepts and applications, whereas INV 201 is a stand-alone, mathematics-intensive subject. Knowledge of INV 101 is not required to succeed in INV 201.

Before beginning preparation for INV 201, candidates should already have a solid base in undergraduate mathematics and probability, as well as a basic knowledge of finance. Specifically:

- **Calculus**: Single-variable and multivariable differentiation/integration, limits, Taylor series, change of variables, and basic ODEs.
- **Linear algebra**: Vectors, matrices, rank, and column spaces.
- **Probability**: Random variables, expectations and variances, common distributions (binomial, Poisson, uniform, exponential, normal, multivariate normal), the central limit theorem, and conditional expectations.
- **Statistics**: Maximum likelihood estimation, linear regression, and basic knowledge of `R` programming.
- **Finance**: Some familiarity with cash, bonds, stocks, forwards, and options, together with the concepts of time value of money and interest rates.

INV 201 is an advanced quantitative exam that demands both mathematical knowledge and financial intuition. This manual is designed to be a self-contained resource to help you bridge the gap from fundamental concepts to exam-level applications, giving you the depth of understanding required not just to pass, but to excel.

About the Author

Yiping Guo is a senior quantitative analyst at a leading Canadian bank, specializing in option pricing. He holds a Ph.D. in Actuarial Science from the University of Waterloo. His academic background and professional experience span a wide range of topics in quantitative finance, actuarial science and machine learning.

Acknowledgment

I am deeply grateful to ACTEX, and in particular Joana Amorim, Yijia Liu, and Caitlyn Darnell, for their consistent support throughout the development of this manual. I would also like to thank the SOA for providing the preliminary version of the exam syllabus at the early stages of writing, which was of great help in shaping the structure of this work.

Yiping Guo, Ph.D.

This page is intentionally left blank.

Contents

Chapter 3

MODERN PROBABILITY THEORY

Learning Outcomes and Chapter Overview

*****FROM THE INV 201 EXAM SYLLABUS*****

Topic II: Valuation of Derivatives

Learning Outcomes:

d) Understand Stochastic Calculus theory and technique used in pricing derivatives.

Note: *The material covered in this chapter does not directly correspond to learning outcome d), except for discrete-time martingales. However, it is foundational to the Stochastic Calculus theory later. Candidates are strongly recommended to study this chapter seriously.*

Topic II is about derivative pricing. Starting from here, the style of the manual becomes much more mathematical, significantly different from Topic I. Chapter 3 builds the probabilistic foundation for the pricing task, such as how to formalize information and how to condition. Later chapters will build models on this base.

Pricing derivatives is, at heart, a problem about expectations under the right probability measure. To take those expectations properly, we need a modern toolkit that goes slightly beyond undergraduate probability: probability spaces and σ-algebras to encode uncertainty, general conditional expectations to update information as time evolves. We will conclude this chapter by a brief introduction to discrete-time stochastic processes.

Our approach is mathematically serious but not fully rigorous: proofs are brief, statements precise, and we will not discuss the most general forms for the concepts involved. Every tool is motivated by how it will be used and why it is necessary. However, we do assume undergraduate-level knowledge of probability theory, although important concepts will be reviewed.

3.1 Foundations of General Probability Theory

3.1.1 Probability Space

We begin by fixing the three "probability attributes" that every valuation model quietly relies on:

- Ω: A universe of possible outcomes;
- $\mathcal{F}$: A collection of events of interest;
- $\mathbb{P}$: A probability measure that assigns weights to those events.

The set Ω is the collection of all elementary outcomes. In a toy example with a single coin toss, $\Omega = \{\text{H,T}\}$. For a single real-valued measurement X, a convenient choice is $\Omega = \mathbb{R}$ with each $\omega \in \Omega$ viewed as a possible realized value. Nothing probabilistic has been decided yet. We have only fixed the universe of possibilities.

Next we decide which questions about outcomes are legitimate events. An *event* is a subset $A \subseteq \Omega$ that can be answered by "Yes, the outcome fell in A" or "No, it did not." We do not want an arbitrary list of events. We want a family that is stable under the basic logical operations we naturally perform:

- If we can ask "Did A occur?", we must also be able to ask "Did A not occur?" which is the complement A^c.
- If we can ask "Did A occur?" and also "Did B occur?", we must be able to ask "Did A or B occur?", which is the union $A \cup B$. The same reasoning extends to "A or B or C or $\cdots$" which is a countable union.

These closures are exactly what a σ-algebra guarantees.

Definition 3.1.1 ($\boldsymbol{\sigma}$-Algebra of Events). *A collection $\mathcal{F}$ is a σ-*algebra *if*

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

(ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$;

(iii) $A_1, A_2, \cdots \in \mathcal{F} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

A short example shows why these closures are not optional. Consider repeated coin tosses and the question "Did we ever see heads by time n?" For each n, that is an event A_n. The question "Did we ever see heads at some time?" is the yes–no question "Did A_1 or A_2 or A_3 or $\cdots$ occur?", which is exactly the countable union $\bigcup_{n=1}^{\infty} A_n$. Without closure under countable unions we could ask each finite question but not the natural "ever?" question. The σ-algebra requirement makes all of these composite yes–no questions legitimate events.

Only after $(\Omega, \mathcal{F})$ has been fixed do we attach probabilities to events. For that, we need the probability measure $\mathbb{P}$.

Definition 3.1.2 (Probability Measure and Probability Space). *A probability measure on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0,1]$ such that*

(i) Normalization: $\mathbb{P}(\Omega) = 1$;

(ii) Countable additivity: For any sequence of pairwise disjoint events $A_1, A_2, \cdots \in \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_{i\geq 1}^{\infty} A_i\right) = \sum_{i\geq 1}^{\infty} \mathbb{P}(A_i). \tag{3.1}$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ *is called a probability space.*

This definition encodes a few immediate and very useful properties of probabilities.

- Monotonicity: If $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
- Complements: For any event A, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
- Finite additivity: For any pairwise disjoint events $A_1, \cdots, A_n \in \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_{i\geq 1}^{n} A_i\right) = \sum_{i\geq 1}^{n} \mathbb{P}(A_i). \tag{3.2}$$

- Subadditivity (Union bounds): For any finite collection $A_1, \cdots, A_n \in \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i). \tag{3.3}$$

  In particular, $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

A few examples help connect these properties with ordinary probability statements. Suppose X is a single real-valued measurement. Events such as $X \leq 0$ or $X \in (a, b]$ are subsets of the sample space, because each statement corresponds to a set of outcomes for which the statement is true. The role of $\mathcal{F}$ is to make sure that such sets are legitimate events and can be assigned probabilities.

The same idea applies when we ask questions over many times. For example, suppose A_i is the event that a threshold is exceeded at time t_i. Then the event that the threshold is exceeded at least once is

$$\bigcup_{i\geq 1} A_i.$$

This is why closure under countable unions matters. It allows us to move from each individual event A_i to the combined event that at least one of them occurs. The subadditivity property (3.3) says that the probability of at least one alarm is no larger than the sum of the individual alarm probabilities.

Let us summarize the key ideas.

- Ω records what could happen.
- $\mathcal{F}$ records the yes–no questions we will allow.
- $\mathbb{P}$ attaches consistent weights to those answers.

In Section 3.1.2, we will place *random variables* on top of $(\Omega, \mathcal{F})$ and then develop expectations and inequalities. Measurability will be mentioned there, as the minimal condition that makes questions about a random variable into events in $\mathcal{F}$.

3.1.2 Random Variables and Distributions

We now attach numbers to outcomes. The outcome $\omega \in \Omega$ is abstract; what we work with is a real number computed from it, such as a measurement or a payoff at a fixed time. A random variable is the device that turns outcomes into numbers.

Definition 3.1.3 (Random Variable). *Let* $(\Omega,\mathcal{F},\mathbb{P})$ *be a probability space. A random variable is a real-valued function* $X : \Omega \to \mathbb{R}$ *such that for any set* $B \in \mathcal{B}(\mathbb{R})$*,*

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}, \tag{3.4}$$

where $\mathcal{B}(\mathbb{R})$ *denotes the Borel* σ*-algebra on* $\mathbb{R}$*.*[1] *It can also be said that* X *is* $\mathcal{F}$*-measurable.*

The word "measurable" in the definition is the crucial part. It guarantees that whenever we ask a natural question about the value of X, for example, "Is $X \le 3$?" or "Is $X \in (a,b]$?", the set of outcomes where this is true is an event in $\mathcal{F}$, and therefore has a well-defined probability. In short, measurability ensures that random variables fit consistently into the probability structure.

Once X is fixed, it determines how probability on $(\Omega,\mathcal{F},\mathbb{P})$ is transported to the real line. The resulting probability on $\mathbb{R}$ is the distribution of X.

Definition 3.1.4 (Distribution (Law) of X). *For any set* $B \subset \mathbb{R}$*,*

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(\omega : X(\omega) \in B). \tag{3.5}$$

The function $\mathbb{P}_X$ *is a probability measure on* $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ *and is called the* distribution *(or law) of* X*.*

A convenient summary of $\mathbb{P}_X$ is the cumulative distribution function

$$F_X(x) = \mathbb{P}(X \le x),\ x \in \mathbb{R}. \tag{3.6}$$

It is non-decreasing and right-continuous, with $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to+\infty} F_X(x) = 1$. You can think of $F_X(x)$ as the total probability that has "accumulated" up to the threshold x. Where F_X jumps, there is point mass; where F_X is continuous, probability is spread over intervals.

When X takes values in a countable set $\{x_1,x_2,\dots\}$, we call X a *discrete* random variable. The distribution is captured by the probability mass function (PMF)

$$p(x_i) = \mathbb{P}(X = x_i),\ i = 1,2,\cdots, \tag{3.7}$$

with $p(x_i) \ge 0$ and $\sum_i p(x_i) = 1$. The CDF is a step function,

$$F_X(x) = \sum_{x_i \le x} p(x_i), \tag{3.8}$$

[1] A "Borel set" is simply any subset of the real line that can be built from intervals by taking unions, intersections, and complements. This covers virtually all sets of numbers one would naturally want to assign probabilities to, such as single points, intervals, or countable unions of intervals. For the purpose of this manual, the formal definition is omitted here.

and the jump of F_X at x_i equals $p_X(x_i)$.

When probability is spread continuously over intervals, we say X is a *continuous* random variable, and we often describe the distribution by a probability density function (PDF) $f(x)$ with

$$F_X(x) = \int_{-\infty}^{x} f(t)dt, \quad \mathbb{P}(a < X \leq b) = \int_a^b f(t)dt, \tag{3.9}$$

where we have $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(t)dt = 1$. Individual points carry no probability in this case: $\mathbb{P}(X = x) = 0$ for each fixed x, even if $f(x) > 0$. Where F_X is differentiable, $f(x) = F'(x)$.

Many variables are *mixed*, meaning that their distributions have both point masses and continuous parts. For instance, a simple mixed distribution may put probability p at 0 and spread the remaining probability $1 - p$ over $(0,\infty)$ according to a density f. In this case, f describes only the continuous part, so

$$\int_0^{\infty} f(t)\, dt = 1 - p. \tag{3.10}$$

The distribution is then described by

$$\mathbb{P}(X = 0) = p, \quad \mathbb{P}(a < X \leq b) = \int_a^b f(t)\, dt, \tag{3.11}$$

for intervals $(a,b] \subset (0,\infty)$. The CDF has a jump of size p at 0 and rises smoothly elsewhere. In general, jumps of F_X correspond to point masses, while smooth parts of F_X correspond to density.

In short, a random variable is a function $X : \Omega \to \mathbb{R}$ that turns abstract outcomes into numbers, and its distribution, described by CDF, PMF, or PDF, is the complete probabilistic description of those numbers.

3.2 General Expectations and Limiting Theorems

3.2.1 General Definition of Mathematical Expectations

We use expectation to turn a random variable into a single average under $\mathbb{P}$. In elementary probability, this average is written as a sum in the discrete case and as an integral against a density in the continuous case. The *Lebesgue integral* is a more general notation that covers both cases at once. Instead of integrating with respect to length dx, it integrates with respect to probability $d\mathbb{P}$. Thus outcomes with larger probability receive larger weight in the average.

In this manual, expectations are understood as Lebesgue integrals, which treat discrete, continuous, and mixed distributions in one line.

Definition 3.2.1 (**Expectation (Lebesgue)**). *Let $X : (\Omega,\mathcal{F},\mathbb{P}) \to \mathbb{R}$ be a random variable. Define*

$$\mathbb{E}[X] = \int_\Omega X(\omega) d\mathbb{P}(\omega) = \int_\Omega X d\mathbb{P}. \tag{3.12}$$

The definition above of expectation is only valid when X is *integrable*, defined as follows.

Definition 3.2.2 (Integrability). *If*

$$\mathbb{E}|X| = \int_\Omega |X(\omega)| d\mathbb{P}(\omega) < \infty, \tag{3.13}$$

we say X is integrable, the Lebesgue integral $\int_\Omega X \, d\mathbb{P}$ exists and we denote it by $\mathbb{E}[X]$.

More generally, we say $X \in L^p$ for $p \geq 1$ if

$$\mathbb{E}|X|^p = \int_\Omega |X(\omega)|^p d\mathbb{P}(\omega) < \infty. \tag{3.14}$$

When $p = 1$, $X \in L^1$ and is integrable; when $p = 2$, $X \in L^2$ and has finite variance (also called square-integrable).

At first sight the integral in (3.12), (3.13) and (3.14) look unusual compared to the familiar Riemann integrals

$$\int f(x)dx.$$

In a Riemann integral, dx means cutting the real line into small equal pieces and adding up the function values weighted by length. Here, for a Lebesgue integral

$$\int X d\mathbb{P},$$

we integrate with respect to the probability measure $\mathbb{P}$, which assigns weight according to the likelihood of outcomes. One way to think of it is that instead of slicing the domain into equal-length intervals, we slice according to probability mass and weight each possible value by how likely it is. This is the main idea of the Lebesgue integral, which provides a unified framework for discrete, continuous, and mixed distributions alike.

From this general definition we can recover the classical formulas.

$$\mathbb{E}[X] = \int_\mathbb{R} x \, dF_X(x). \tag{3.15}$$

And more specifically,

- For a discrete random variable X with PMF $p(x)$, this reduces to

$$\mathbb{E}[X] = \sum_x xp(x). \tag{3.16}$$

- For a continuous random variable with density $f(x)$,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x) \, dx. \tag{3.17}$$

To make the general definition and its connection to the classical definition clearer, we use the following example to illustrate.

Example 3.2.1. Let $\Omega = [0,1]$ with uniform probability measure, that is, $\omega \sim U(0,1)$. Define

$$X(\omega) = -\log(\omega).$$

(a) Use the general definition of expectation to compute $\mathbb{E}[X]$ directly.

(b) First derive the distribution function of X and use it to calculate $\mathbb{E}[X]$. Compare with the result with (a).

Solution.

(a) Since $\omega \sim U(0,1)$, the probability measure on $[0,1]$ has density 1 with respect to ordinary length measure. In other words, for a small interval $d\omega$, the probability mass is

$$d\mathbb{P}(\omega) = 1 \cdot d\omega = d\omega.$$

Therefore,

$$\mathbb{E}[X] = \int_\Omega X d\mathbb{P} = \int_0^1 X(\omega) d\omega = \int_0^1 -\log(\omega) d\omega = \omega - \omega\log(\omega)\Big|_0^1 = 1.$$

(b) We first calculate the distribution function of X:

$$F(x) = \mathbb{P}(X \le x) = \mathbb{P}(X(\omega) \le x) = \mathbb{P}(-\log(\omega) \le x) = \mathbb{P}(e^{-x} \le \omega \le 1) = 1 - e^{-x}.$$

This implies that X follows an exponential distribution with mean 1. Therefore, $\mathbb{E}[X] = 1$, which is the same as the answer in (a).

3.2.2 Common Properties and Inequalities of Expectations

After getting familiar with the new abstract definition of expectations, now let us discuss some common properties and inequalities of expectations, which will be used more frequently in practice.

- **Linearity of Expectations**:

  Theorem 3.2.1. *For integrable random variables X, Y and scalars $a, b \in \mathbb{R}$,*

  $$\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]. \tag{3.18}$$

  This property is fundamental: expectation is an averaging operator, and averages are linear. In practice this means we can pull constants outside and decompose complicated expressions into simpler parts. For instance, if X is a random payoff and c is a risk-free interest rate, then $\mathbb{E}[X - c] = \mathbb{E}[X] - c$.

- **Fubini's Theorem**[2]:

  Theorem 3.2.2. *For a set of integrable random variables* $\{X_s\}_{0\le s\le t}$,

$$\mathbb{E}\left[\int_0^t X_s ds\right] = \int_0^t \mathbb{E}[X_s]ds. \tag{3.19}$$

  This result is basically the continuous version of the linearity property discussed above. It allows us to exchange the notions of the expectation operator and integration sign as long as the integrand X_s is integrable (which means $\mathbb{E}|X_s| < \infty$ for all $s \in [0,t]$).

- **Law of the Unconscious Statistician (LOTUS)**:

  Theorem 3.2.3. *Let X be an integrable random variable with distribution F_X, and define a transformation function $g : \mathbb{R} \to \mathbb{R}$. If $Y = g(X)$ is integrable, then*

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)\, dF_X(x). \tag{3.20}$$

  For a discrete random variable with PMF $p(x)$, this becomes $\mathbb{E}[g(X)] = \sum_x g(x)p(x)$. For a continuous random variable with density $f(x)$, we have $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\, dx$.

  This result is powerful. Without it, we would need to calculate the distribution of $Y = g(X)$ in order to compute $\mathbb{E}[g(X)] = \int y dF_Y(y)$, and this is non-trivial in many cases. By LOTUS, we only need the distribution of X.

- **Connection between Expectations and Probabilities**:

  The general definition of expectation is such an important concept in modern probability theory, since it bridges a wide range of key components. One of the most important result is to represent any probability using expectation. For that, we first need to introduce an important concept, the *indicator function*:

  Definition 3.2.3 (Indicator Function). *Let A be an event. Then the indicator function for A is defined as*[3]

$$\mathbf{1}_{\{A\}} = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases} \tag{3.21}$$

  Then, we have the following.

  Theorem 3.2.4. *For a random variable X and an event $\{X \in B\}$,*

$$\mathbb{E}[\mathbf{1}_{\{X\in B\}}] = \mathbb{P}(X \in B). \tag{3.22}$$

[2] In general, Fubini's theorem allows the order of two integrals to be exchanged under suitable integrability conditions. In the present setting, this means exchanging integration over time with integration over probability:

$$\int_\Omega \left(\int_0^t X_s(\omega)\, ds\right) d\mathbb{P}(\omega) = \int_0^t \left(\int_\Omega X_s(\omega) d\mathbb{P}(\omega)\right) ds.$$

This is exactly the statement in (3.19).

[3] Equivalent notations to $\mathbf{1}_{\{A\}}$ include $\mathbf{1}(A)$, $\mathbb{I}_{\{A\}}$, $\mathbb{I}(A)$.

This result is particularly useful when proving results when the underlying type of random variable is unknown (discrete, continuous, or mixed). We will illustrate this by using this result to prove the Theorem 3.2.5 below.

- **Expectation via Tail Distribution**:

Theorem 3.2.5. *For a non-negative random variable X,*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)\, dx = \int_0^\infty (1 - F(x))\, dx. \tag{3.23}$$

Proof. By Theorem 3.2.4, we have

$$\begin{aligned}\int_0^\infty \mathbb{P}(X > x)\, dx &= \int_0^\infty \mathbb{E}[\mathbf{1}_{\{X>x\}}]dx \\ &= \mathbb{E}\left[\int_0^\infty \mathbf{1}_{\{X>x\}}dx\right] \qquad ((*)) \\ &= \mathbb{E}\left[\int_0^X 1dx + \int_X^\infty 0dx\right] \\ &= \mathbb{E}\left[\int_0^X 1dx\right] \\ &= \mathbb{E}[X],\end{aligned}$$

where in $(*)$, we used Fubini's theorem to change the order of expectation and integration.[4]

□

This identity expresses the expectation as the area under the survival function (the "tail probability" curve). It provides a useful alternative formula, particularly for random variables whose distribution is easier to describe in terms of tail probabilities. For instance, in Example 3.2.1, we can calculate by the expectation of X using the derived distribution function:

$$\mathbb{E}[X] = \int_0^\infty (1 - F(x))dx = \int_0^\infty e^{-x}dx = 1$$

In addition to the properties discussed above, there are a few common useful inequalities regarding expectations and moments.

- **Triangle Inequalities for Expectations**:

Theorem 3.2.6. *For integrable random variables X and Y,*

$$\mathbb{E}|X + Y| \le \mathbb{E}|X| + \mathbb{E}|Y|, \tag{3.24}$$

and

$$|\mathbb{E}[X]| \le \mathbb{E}|X|. \tag{3.25}$$

[4] Here, Fubini's theorem is valid given that the integrability condition is met: $\mathbb{E}|\mathbf{1}_{\{X>x\}}| \le 1 < \infty$.

The first statement is the probabilistic analogue of the triangle inequality in Euclidean geometry: the "length" (absolute value) of a sum is no greater than the sum of lengths. The second inequality simply says that the absolute value of the average cannot exceed the average of the absolute values. Both are useful for bounding expectations when exact calculation is difficult.

- **Jensen's Inequality**:

Definition 3.2.4 (**Convex and Concave Function**). *A function* $\varphi : \mathbb{R} \to \mathbb{R}$ *is called convex if for all* $0 < t < 1$ *and* $x_1, x_2 \in \mathbb{R}$,

$$\varphi(tx_1 + (1-t)x_2) \le t\varphi(x_1) + (1-t)\varphi(x_2). \tag{3.26}$$

It is called concave if for all $0 < t < 1$ *and* $x_1, x_2 \in \mathbb{R}$,

$$\varphi(tx_1 + (1-t)x_2) \ge t\varphi(x_1) + (1-t)\varphi(x_2). \tag{3.27}$$

In the special case where $\varphi(x)$ *is second-order differentiable, it is convex if and only if* $\varphi''(x) \ge 0$ *for all* x*; it is concave if and only if* $\varphi''(x) \le 0$ *for all* x.

Theorem 3.2.7. *Let* X *be an integrable random variable.*

- *If* $\varphi(x)$ *is a convex function, then*

$$\mathbb{E}[\varphi(X)] \ge \varphi(\mathbb{E}[X]). \tag{3.28}$$

- *If* $\varphi(x)$ *is a concave function, then*

$$\mathbb{E}[\varphi(X)] \le \varphi(\mathbb{E}[X]). \tag{3.29}$$

The following two plots illustrate the idea of Jensen's inequality geometrically. Being able to draft these plots is very helpful in the exam to remind the correct form of Jensen's inequality.

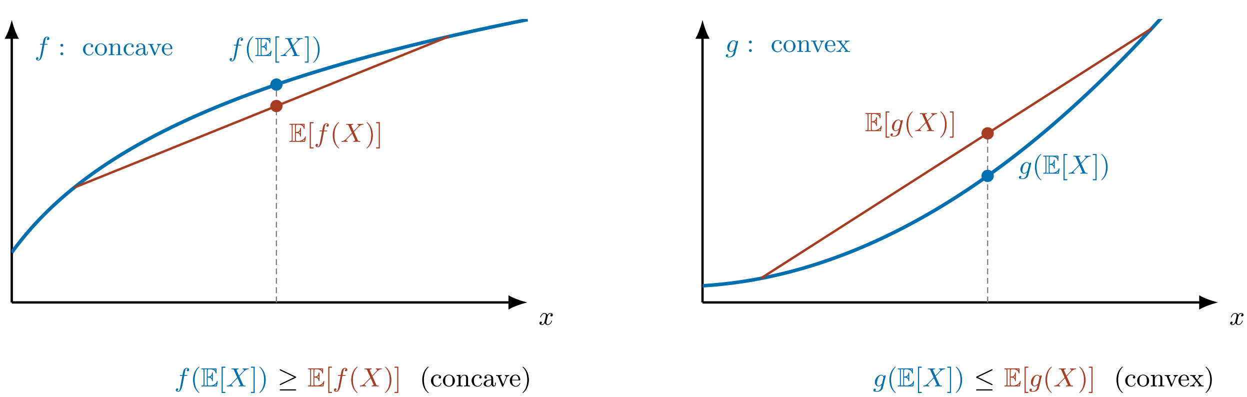

Figure 3.1: Geometric illustration of Jensen's inequalities.

- **Relationship of the First and Second Moment**:

  Theorem 3.2.8. *If $X \in L^2$, then $X \in L^1$ and*

  $$\mathbb{E}|X| \leq \sqrt{\mathbb{E}[X^2]}. \tag{3.30}$$

  This result is a direct application of the Jensen's inequality with $\varphi(x) = \sqrt{x}$ applied to random variable X^2.

  It has two important implications:

  - If a random variable is square-integrable ($X \in L^2$), then it is integrable ($X \in L^1$). Said differently: finite variance implies finite mean.
  - Conversely, if a random variable fails to have a finite mean ($X \notin L^1$), then it cannot possibly have a finite variance ($X \notin L^2$). For example, a Cauchy random variable with the PDF

    $$f(x) = \frac{1}{\pi(1+x^2)}, \tag{3.31}$$

    does not have a finite mean , since

    $$\int_{-\infty}^{\infty} |x| f(x)\, dx = \frac{2}{\pi} \int_0^{\infty} \frac{x}{1+x^2}\, dx = \infty. \tag{3.32}$$

    Hence it does not have a finite variance either.

Among the three results above, Jensen's inequality is particularly useful and provides us with a powerful tool in probability theory and in option pricing. It allows us to derive useful bounds for expectations even though the explicit form of the formula is not available. We use two examples to illustrate.

Example 3.2.2. For a random variable X with $\mathbb{E}[X] = 1$, use Jensen's inequality to find a lower bound or upper bound of its third moment $\mathbb{E}[X^3]$.

Solution. We first notice that the transformation $\varphi(X) = X^3$ is a convex function since

$$\varphi''(x) = 3x^2 \geq 0.$$

Therefore, we can apply the Jensen's inequality to find a lower bound to $\mathbb{E}[X^3]$:

$$\mathbb{E}[X^3] \geq (\mathbb{E}[X])^3 = 1.$$

The upper bound is not available from Jensen's inequality.

Example 3.2.3. Let $\varphi(x) = (x)^+ = \max(x,0)$ be a real-valued function that represent the positive part of x.

(a) Show that $\varphi(x) = (x)^+$ is a convex function.

(b) Use Jensen's inequality to show that

$$\mathbb{E}[(S_T - K)^+] \geq (\mathbb{E}[S_T] - K)^+.$$

Solution.

(a) Here, $\varphi(x) = (x)^+$ is not a second-order differentiable function since it has a discontinuity point at $x = 0$. Therefore, we have to prove the convexity by definition.

We want to show that, for any $0 < t < 1$ and x_1,$x_2 \in \mathbb{R}$,

$$\varphi(tx_1 + (1-t)x_2) \leq t\varphi(x_1) + (1-t)\varphi(x_2).$$

We discuss in two cases:

– If $tx_1 + (1-t)x_2 \leq 0$, then

$$(tx_1 + (1-t)x_2)^+ = 0 \leq tx^+ + (1-t)\, y^+.$$

– If $tx_1 + (1-t)x_2 > 0$, then

$$\Big(tx + (1-t)y\Big)^+ = tx + (1-t)y \leq tx^+ + (1-t)y^+.$$

Therefore, $\varphi(x) = (x)^+$ is convex.

(b) Since $\varphi(x) = (x)^+$ is convex, by Jensen's inequality:

$$\mathbb{E}[(S_T - K)^+] \geq (\mathbb{E}[S_T - K])^+ = (\mathbb{E}[S_T] - K)^+,$$

where the last step is by the linearity of expectations.

To conclude this part, we review an important concept related to the expectation, the moment generating function (MGF):

Definition 3.2.5 (Moment Generating Function). *For a random variable X, its moment generating function is defined as:*

$$M_X(t) = \mathbb{E}[e^{tX}],\ \ t \in \mathbb{R}. \tag{3.33}$$

Assume that the expectation $M_X(t)$ exists.[5]

Essentially, the MGF is the Taylor expansion of all the moments of a random variable X, which carries all the information of the distribution. Each distribution will have a **unique** MGF, therefore, if two random variables share the same MGF, they must be identically distributed. In

[5] In this scope of INV 201, we always assume that the MGF exists. More generally, one can adopt the so-called *characteristic function*, defined as $\varphi(X) = \mathbb{E}[e^{itX}]$, which is based on Fourier transform. It plays the same role as MGF, but can be applied to any random variables, regardless of the existence of their moments.

INV 201, the most important MGF is the one for the normal distribution $\mathcal{N}(\mu,\sigma^2)$:

$$M(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}. \tag{3.34}$$

3.2.3 Limiting Theorems

In many scenarios in probability and statistics we care about the sample average of a sequence of random variables and how it behaves. Limiting theorems tell us what happens as we aggregate more data. The *Law of Large Numbers (LLN)* captures the first-moment information, where the average is heading. The *Central Limit Theorem (CLT)* captures the second-moment information, how the residual fluctuations look and scale.

Theorem 3.2.9 (**Law of Large Numbers**). *Let $X_1,X_2,\ldots X_n \sim X$ be independent and identically distributed (i.i.d.) with finite mean $\mu = \mathbb{E}[X] < \infty$. We have*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu, \tag{3.35}$$

where the notation $\to$ means "converges to". [6]

The intuition of LLN is rather simple. It means that averaging cancels noise, so the empirical mean stabilizes at the true mean. For example, when tossing the same fair coin many many times, the proportion of heads obtained will eventually converge to 0.5. Note the LLN is qualitative and it identifies the limit but not the speed.

To quantify the uncertainty of the sample mean, we need the CLT, stated as follows.

Theorem 3.2.10 (**Central Limit Theorem**). *Let $X_1,X_2,\ldots X_n \sim X$ be i.i.d. with finite mean $\mu = \mathbb{E}[X] < \infty$ and finite variance $\sigma^2 = Var(X) < \infty$.*[7] *We have*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0,1), \text{ or equivalently, } \bar{X}_n \to \mathcal{N}\left(\mu,\frac{\sigma^2}{n}\right), \tag{3.36}$$

where the $\mathcal{N}(0,1)$ means the standard normal (Gaussian) distribution, with density function

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

In addition,

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \to \mathcal{N}(0,1), \text{ or equivalently, } S_n \to \mathcal{N}\left(n\mu,n\sigma^2\right), \tag{3.37}$$

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

[6] In fact, there exist two versions of LLN, named strong law of large numbers (SLLN) and weak law of large numbers (WLLN). In each of the SLLN and WLLN, the type of convergence is more specifically defined (convergence almost surely for SLLN and convergence in probability for WLLN). For the purpose of this manual, we omit the technical detail here.

[7] Both the finite variance and the i.i.d. assumption can be weakened to obtain a similar form of CLT or another type of limiting law, but only if we control tail contributions of each individual terms and prevent any single summand from carrying too much variance. We only discuss this simplest form of CLT in this manual.

The CLT quantifies the uncertainty of $\bar{X}$ as the sample size n grows. The typical error of $\bar{X}_n$ is of order $1/\sqrt{n}$, and its shape is approximately Gaussian. The finite variance assumption is crucial here, which rules out tails so heavy that a few terms dominate.

The example below summarizes some commonly applications of CLT.

Example 3.2.4.

(a) Let $Y_n \sim$ Binomial(n,p). Apply the CLT to show that Y_n converges to a normal distribution.

(b) Let $Y_n \sim$ Poisson(n). Apply the CLT to show that Y_n converges to a normal distribution.

(c) Let $Y_n \sim$ Gamma(n,λ), with PDF

$$f(y) = \frac{\lambda^n}{\Gamma(n)} e^{-\lambda y} y^{n-1}.$$

Apply the CLT to show that Y_n converges to a normal distribution.

Solution.

(a) Notice that $Y_n \sim \text{Binomial}(n,p) = X_1 + \cdots + X_n$, where $X_1, \cdots ,X_n \overset{i.i.d.}{\sim} \text{Bernoulli}(p)$. Also, $\mathbb{E}[Y_n] = np$ and $\text{Var}(Y_n) = np(1-p)$, therefore, by CLT,

$$Y_n \to \mathcal{N}\left(np,np(1-p)\right).$$

In addition, the sample proportion $\bar{X}_n = Y_n/n$ also follows a normal distribution:

$$\bar{X}_n = \frac{Y_n}{n} \to \mathcal{N}\left(p,\frac{p(1-p)}{n}\right)$$

(b) Notice that $Y_n \sim \text{Poisson}(n) = X_1 + \cdots + X_n$, where $X_1, \cdots ,X_n \overset{i.i.d.}{\sim} \text{Poisson}(1)$. Also, $\mathbb{E}[Y_n] = n$ and $\text{Var}(Y_n) = n$, therefore, by CLT,

$$Y_n \to \mathcal{N}\left(n,n\right).$$

(c) Notice that $Y_n \sim \text{Gamma}(n,\lambda) = X_1 + \cdots + X_n$, where $X_1, \cdots ,X_n \overset{i.i.d.}{\sim} \text{Exponential}(\lambda)$. Also, $\mathbb{E}[Y_n] = n/\lambda$ and $\text{Var}(Y_n) = n/\lambda^2$, therefore, by CLT,

$$Y_n \to \mathcal{N}\left(\frac{n}{\lambda},\frac{n}{\lambda^2}\right).$$

3.3 Conditioning and Information Sets

In probability, conditioning is one of the most important tools to understand randomness, particularly for stochastic processes.It formalizes how probabilities and expectations update once new information is revealed. In quantitative finance, it also plays a fundamental role: pricing relies on conditional expectations of discounted future payoffs given today's information.

These ideas are essential before moving to more advanced topics. They are standard in undergraduate probability courses, which we assume the reader has some background in, but our treatment here will go slightly further for the need of future chapters.

3.3.1 Independence

We first review the definition of independence between two random variables in elementary probability theory.

Definition 3.3.1 (Independence via Events). *Random variables X and Y are independent if for all measurable sets A,B,*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B). \tag{3.38}$$

Equivalently, with joint CDF $F_{X,Y}$ and marginals F_X,F_Y, for all x,$y \in \mathbb{R}$,

$$F_{X,Y}(x,y) = F_X(x)\,F_Y(y). \tag{3.39}$$

Some common special cases include:

- If X, Y are both discrete with joint PMF $p_{X,Y}$, they are independent if and only if for all x,$y \in \mathbb{Z}$

$$p_{X,Y}(x,y) = p_X(x)p_Y(y). \tag{3.40}$$

- If X, Y are both continuous with joint PDF $f_{X,Y}$, they are independent if and only if for all x,$y \in \mathbb{R}$

$$f_{X,Y}(x,y) = f_X(x)f_Y(y). \tag{3.41}$$

The same idea can also be expressed through expectations of functions of X and Y.

Definition 3.3.2 (Independence via Test Functions). *X and Y are independent if and only if, for all bounded measurable f,g,*

$$\mathbb{E}\Big[f(X)g(Y)\Big] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)]. \tag{3.42}$$

This statement is infeasible to use in practice to prove independence, since it is impossible to go over all functions f and g. Instead, it could be used to demonstrated dependence: X and Y are dependent if there exist bounded measurable f,g such that

$$\mathbb{E}\Big[f(X)g(Y)\Big] \neq \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)]. \tag{3.43}$$

In what follows, we discuss some immediate corollaries regarding independence:

- If X and Y are independent random variables, and f, g are measurable functions, then $f(X)$ and $g(Y)$ are also independent. This result is fairly intuitive. applying f to X and g to Y is just "relabeling" outcomes already determined by X and Y. Since Y tells you nothing about X, it also tells you nothing about any function of X (and vice versa).
- If X and Y are independent random variables, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]. \tag{3.44}$$

  This result can be directly obtained from (3.42) when setting $f(X) = X$ and $g(Y) = Y$.

  Recall that the covariance of two random variables is

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y], \tag{3.45}$$

  and (X,Y) are *uncorrelated* if $\mathrm{Cov}(X,Y) = 0$. When X and Y are independent, $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$. This result can therefore also be read as "**Independence implies uncorrelation**".

It should be noted that, in general, "uncorrelated" need not mean "independent." We will illustrate this using a simple example.

Example 3.3.1. Let $X \sim U[-1,1]$ and $Y = X^2$.

(a) Are X and Y independent? Explain.

(b) Are X and Y uncorrelated? What can you say combining your finding in (a)?

Solution.

(a) X and $Y = X^2$ are clearly dependent, since Y is a non-degenerate function of X. (Here we say Y is non-degenerate if it is not deterministic).

(b) We first have $\mathbb{E}[X] = 0$, thus $\mathbb{E}[X]\mathbb{E}[Y] = 0$. In addition,

$$\mathbb{E}[XY] = \mathbb{E}[X^3] = 0,$$

where $\mathbb{E}[X^3] = 0$ since X is symmetric around $X = 0$. Therefore, we have $\mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] = 0$. X and Y are uncorrelated, but not independent.

A key exception where "uncorrelated" **does** imply "independence" is the jointly normal case. If (X,Y) is bivariate normal[8] with means μ_X,μ_Y, variances σ_X^2,σ_Y^2, and correlation $\rho = 0$, the joint density is

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y}\exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\} = f_X(x)f_Y(y).$$

[8] Here, bivariate normal does not merely mean that both X and Y are individually normal. It means the vector (X,Y) has a joint distribution that can be represented as a linear transformation of two independent standard normals (Z_1,Z_2). Equivalently, (X,Y) should have a joint PDF defined for bivariate normal.

3.3.2 General Conditional Expectations and Properties

When we talk about "information sets" we mean what is currently known about the outcome. Often we do not observe the whole world but only a summary Y or, more abstractly, a collection of events $\mathcal{G}$ that we can distinguish. The right question is then:

"What is the best average of a target X that respects only what is known?"

Conditional expectation answers this. You can picture it in two equivalent ways.

- Partition view: The value of Y carves the sample space into slices, and within each slice we average X; the resulting number depends on the slice, so it is itself a random variable.
- Projection view: Among all functions that depend only on the known information, $\mathbb{E}[X|\mathcal{G}]$ is the one closest to X in mean-square error. We will revisit this idea in greater detail later in this subsection.

There are two basic definitions that capture this idea, depending on what the information set looks like. Conditioning on an event is the most elementary case; conditioning on a random variable generalizes the notion and is the form most used in probability and finance. The two are linked, since any event A is naturally associated with a random variable, namely the indicator function 1_A.

Definition 3.3.3 (Conditional Expectation on an Event). *For an event A with $\mathbb{P}(A) > 0$ and a random variable X, the conditional expectation of X on A is defined as:*

$$\mathbb{E}[X|A] = \frac{\mathbb{E}[X\,\mathbf{1}_{\{A\}}]}{\mathbb{P}(A)}. \tag{3.46}$$

This is simply the average of X restricted to the slice of outcomes where A occurs.

Definition 3.3.4 (Conditional Expectation on a Random Variable). *For random variables X and Y, the conditional expectation of X given Y is defined as the $\sigma(Y)$-measurable random variable $\mathbb{E}[X|Y]$ such that for all $B \in \mathcal{B}(\mathbb{R})$*

$$\mathbb{E}\left[X\mathbf{1}_{\{Y\in B\}}\right] = \mathbb{E}\left[\mathbb{E}[X|Y]\cdot\mathbf{1}_{\{Y\in B\}}\right]. \tag{3.47}$$

Here $\sigma(Y)$ is the σ-algebra generated by the random variable Y. For the sake of this manual, we will not discuss its formal definition. It simply refers to the collection of events that can be described entirely in terms of the value of Y. In other words, it represents all the information that observing Y gives us about the underlying sample space. Saying that $\mathbb{E}[X|Y]$ is $\sigma(Y)$-measurable just means it can be written as some function of Y, and it only depends on Y.

How to intuitively understand the definition (3.47)? $\mathbb{E}[X|Y]$ tells us the average of X within each slice of outcomes where Y takes the same value. Once we observe $Y = y$ the conditional expectation gives the best prediction of X based only on that information. It can be viewed as a "moving average" of X, moving along with the value of Y.

In fact, the event-conditioning is a special case of the random variable-conditioning: if $Y = \mathbf{1}_{\{A\}}$, then

$$\mathbb{E}[X|Y] = \begin{cases} \mathbb{E}[X|A], & A \text{ occurs.} \\ \mathbb{E}[X|A^c], & \text{otherwise.} \end{cases} \tag{3.48}$$

A conditional expectation $\mathbb{E}[X|Y]$ is a random variable and it hence varies with the observed information. Despite this, it shares some of the similar rules as the ordinary expectation $\mathbb{E}[X]$ (which is a constant), now "inside" the information set:

- **Linearity of Conditional Expectations**:

  Theorem 3.3.1. *For integrable random variables X, Y, Z and scalars $a, b \in \mathbb{R}$,*

  $$\mathbb{E}[aX + bY|Z] = a\mathbb{E}[X|Z] + b\mathbb{E}[Y|Z]. \tag{3.49}$$

- **Law of the Unconscious Statistician (LOTUS) for Conditional Expectations:**:

  Theorem 3.3.2. *Let X, Z be integrable random variables, and define a transformation function $g : \mathbb{R} \to \mathbb{R}$. If $Y = g(X)$ is integrable, then*

  $$\mathbb{E}[g(X)|Z] = \int_{\mathbb{R}} g(x)\, dF_{X|Z}(x), \tag{3.50}$$

  where $F_{X|Z}(x)$ is the conditional distribution of X given Z.

  Jensen's Inequality for Conditional Expectations:

  Theorem 3.3.3. *Let X and Z be integrable random variables.*

  – *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function, then*

  $$\mathbb{E}[\varphi(X)|Z] \geq \varphi(\mathbb{E}[X|Z]). \tag{3.51}$$

  – *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a concave function, then*

  $$\mathbb{E}[\varphi(X)|Z] \leq \varphi(\mathbb{E}[X|Z]). \tag{3.52}$$

Apart from the similar properties we have already seen from the ordinary expectations, conditional expectations also have some unique properties which can greatly help practical computations. In undergraduate probability, when coming to conditional expectations, one might usually apply formulas like

$$\mathbb{E}[X|Y = y] = \int_{\mathbb{R}} f_{X|Y}(x|y)dx, \tag{3.53}$$

where $f_{X|Y}(x|y)$ is the conditional PDF of X given $Y = y$. However, in practice and the area of quantitative finance, it is very unlikely (or something impossible) to derive the conditional distribution. Instead, we usually apply the following rules to help simplifying computations.

- **Invariance under one-to-one transformations**:

  Theorem 3.3.4. *If g is a one-to-one measurable function, then*

  $$\mathbb{E}[X|Y] = \mathbb{E}[X|g(Y)] \tag{3.54}$$

  almost surely (a.s.).[9] *This result is intuitive to understand, as a 1-1 transformation of Y loses no information, so the best predictor given Y or given $g(Y)$ must coincide.*

- **Conditioning under independence**:

  Theorem 3.3.5. *If X is independent of Y, then for any measurable g,*

  $$\mathbb{E}[X|g(Y)] = \mathbb{E}[X] \tag{3.55}$$

  almost surely (a.s.).

  This result is also very intuitive. If the Y contain no information about X, the best predictor of X using Y is just the ordinary expectation of X.

- **Conditioning on known quantities**:

  Theorem 3.3.6. *If $h(X)$ is a measurable function,*

  $$\mathbb{E}[h(X)|X] = h(X) \quad \textit{and} \quad \mathbb{E}[h(X)Y|X] = h(X)\mathbb{E}[Y|X]. \tag{3.56}$$

  These rules follow from $\sigma(X)$-measurability. They match our intuition: once X is known, any measurable function of X is also known.

The following example will illustrate how to use those properties in a nutshell.

Example 3.3.2. Let $X \sim \mathcal{N}(0,1)$, $Y = X^2$, $Z = X^3$. Calculate the following conditional expectations:

(a) $\mathbb{E}[Y|X]$

(b) $\mathbb{E}[X|Y]$

(c) $\mathbb{E}[Y|Z]$

Solution.

(a) Using the "Conditioning on known quantities" law:

$$\mathbb{E}[Y|X] = \mathbb{E}[X^2|X] = X^2$$

[9] Here, since the conditional expectations are random variables, the "=" sign has a different interpretation from the standard notation when comparing numbers. The term "almost surely" means both random variables have the same value with probability 1.

(b) We want to find $\mathbb{E}[X|Y] = \mathbb{E}[X|X^2]$. By the symmetry of $X \sim \mathcal{N}(0,1)$, we notice that when observing $X^2 = t \geq 0$, we have

$$X = \begin{cases} \sqrt{t}, & \text{with probability } \frac{1}{2}, \\ -\sqrt{t}, & \text{with probability } \frac{1}{2}. \end{cases}$$

It implies that $\mathbb{E}[X|X^2 = t] = 0$ for all $t \geq 0$. Therefore,

$$\mathbb{E}[X|Y] = \mathbb{E}[X|X^2] = 0$$

(c) Since $Z = X^3$ is a 1-1 function, thus

$$\mathbb{E}[Y|Z] = \mathbb{E}[X^2|X^3] = \mathbb{E}[X^2|X] = X^2 = Y.$$

We finally discuss two important theorems connecting unconditional expectations/variances to conditional expectations/variances.

- **Law of Total Expectation**:[10]
  - **Basic form**:

    Theorem 3.3.7. *For random variables X and Y,*

    $$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]. \tag{3.57}$$

    To understand this result, think of $\mathbb{E}[X|Y]$ as the "best predictor of X" once you know Y. If you then average these predictors over all possible values of Y, you get back the unconditional average of X. For example, suppose X is a person's exam score and Y is the class they belong to. Within each class, the conditional expectation $\mathbb{E}[X|Y]$ is the class average. If we then average these class averages, weighted by the proportion of students in each class, we recover the overall average score.

  - **General three-variable form**:

    Theorem 3.3.8. *For random variables X, Y and Z,*

    $$\mathbb{E}[X|Y] = \mathbb{E}[\mathbb{E}[X|Y,Z]|Z]. \tag{3.58}$$

    The intuition is similar to the basic two-variable form. If we know both Y and Z, we can form a finer predictor $\mathbb{E}[X|Y,Z]$. But if we later "forget" Z and keep only Y, the best predictor collapses back to $\mathbb{E}[X|Y]$. Conditioning on more information refines the predictor; removing that information makes the prediction rough again.

[10] This result is also called "double expectation rule" or "tower law" of conditional expectations.

We can interpret using the same exam score example. Suppose X is exam score, Y is the school, and Z is the class within the school. The conditional expectation $\mathbb{E}[X|Y,Z]$ is the class average. If we then average across classes within each school, we get the school average, which is exactly $\mathbb{E}[X|Y]$.

- **Law of Total Variance**:

Theorem 3.3.9. *For random variables X and Y,*

$$\mathrm{Var}(X) = \mathbb{E}\left[\mathrm{Var}(X|Y)\right] + \mathrm{Var}\left(\mathbb{E}[X|Y]\right). \tag{3.59}$$

*This is also called the "**EVVE**" rule.*

To understand this intuitively, we can think of it as splitting the total variability $\mathrm{Var}(X)$ of a quantity into two parts:

I. $\mathbb{E}\left[\mathrm{Var}(X|Y)\right]$: Average "within-group" variability;

II. $\mathrm{Var}\left(\mathbb{E}[X|Y]\right)$: Variability of the "between-group" errors (difference averaged values in different classes).

More concretely, suppose a population is partitioned by Y into subgroups with different means and internal spreads. The first term averages subgroup spreads; the second term measures how far subgroup means are from the grand mean. If subgroup means differ a lot, the second term dominates; if subgroups are very noisy internally, the first term dominates.

3.3.3 Geometric Interpretation of Conditional Expectations

We already know the working idea: if all you are allowed to use is the information in Y, then the **best** forecast of X is the conditional expectation $\mathbb{E}[X|Y]$. "Best" means it has the smallest average squared error among all candidates that depend only on Y. Concretely, the competitors look like $g(Y)$ for some function g with $\mathbb{E}[g(Y)^2] < \infty$. The set of all such competitors is the space $\{g(Y)\}$ (formally this is $L^2(\sigma(Y))$).

Why it is the best in mean-square error? Let $\hat{X} = \mathbb{E}[X|Y]$. Take any other candidate $Z = g(Y)$. Expand the square just like in least squares:

$$\mathbb{E}[(X-Z)^2] = \mathbb{E}[(X-\hat{X})^2] + 2\mathbb{E}[(X-\hat{X})(\hat{X}-Z)] + \mathbb{E}[(\hat{X}-Z)^2]. \tag{3.60}$$

The middle term vanishes. Indeed,

$$\mathbb{E}[(X-\hat{X})(\hat{X}-Z)] = \mathbb{E}[\mathbb{E}[(X-\hat{X})\cdot \underbrace{(\hat{X}-Z)}_{\text{Function of } Y} |Y]] = \mathbb{E}[(\hat{X}-Z)\cdot \underbrace{\mathbb{E}[X-\hat{X}|Y]}_{=\mathbb{E}[X]-\mathbb{E}[\mathbb{E}[X|Y]]=0}] = 0, \tag{3.61}$$

Therefore,

$$\mathbb{E}[(X-Z)^2] = \underbrace{\mathbb{E}[(X-\hat{X})^2]}_{\geq 0} + \underbrace{\mathbb{E}[(\hat{X}-Z)^2]}_{=0 \text{ only if } Z=\hat{X}} \geq \mathbb{E}[(X-\hat{X})^2]. \tag{3.62}$$

So $\hat{X} = \mathbb{E}[X|Y]$ is the unique minimizer.

Therefore we can think of $\mathbb{E}[XY]$ as an orthogonal projection of X onto the "space" of Y, which contains the subspace of all functions of Y. This is illustrated in the Figure 3.2.

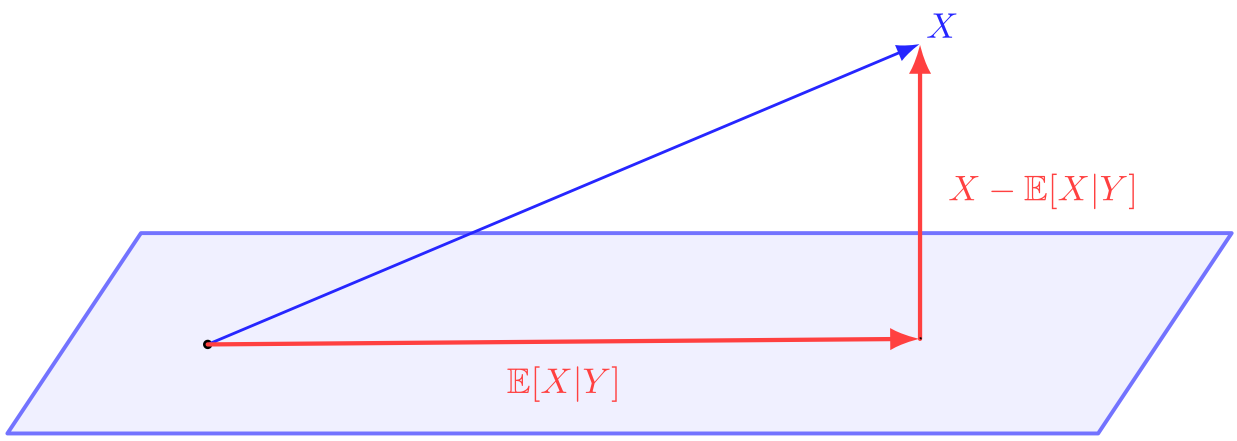

Figure 3.2: Geometric interpretation of conditional expectations.

More precisely, we equip all square-integrable random variables with the inner product $\langle U,V\rangle = \mathbb{E}[UV]$ in this space. Here, the second moment $\mathbb{E}[U^2] = \langle U,U\rangle$ denotes the L^2-norm of X. The picture clearly tells us a few things:

- $\mathbb{E}[X|Y]$ lies in the space of all functions of Y, therefore it's also a function of Y.
- The residual term $X - \mathbb{E}[X|Y]$ is orthogonal to **every** function of Y, say $g(Y)$. This result can be easily shown using the law of total expectation:

$$\mathbb{E}[(X - \mathbb{E}[X|Y])g(Y)] = \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[X|Y])g(Y)]|Y] = \mathbb{E}[g(Y)\underbrace{\mathbb{E}[(X - \mathbb{E}[X|Y])]}_{=0}|Y] = 0 \tag{3.63}$$

Therefore $\mathbb{E}[X|Y]$ is exactly the orthogonal projection of X onto the subspace $\{g(Y)\}$. Furthermore, we can use a seemingly trivial but surprisingly elegant "regression decomposition" to summarize the result:

$$X = \mathbb{E}[X|Y] + (X - \mathbb{E}[X|Y]), \tag{3.64}$$

where the conditional expectation $\mathbb{E}[X|Y]$ is the best predictor fitted to X using Y, and $(X - \mathbb{E}[X|Y])$ is the estimation error which is orthogonal to the fitted value $\mathbb{E}[X|Y]$. In the language of probability, the orthogonality means uncorrelation.

Let us then discuss how this geometric interpretation can help us understand the two important results of conditional expectations: the law of total expectation and the law of total variance.

- The law of total expectation states that for random variables X, Y and Z,

$$\mathbb{E}[X|Y] = \mathbb{E}[\mathbb{E}[X|Y,Z]|Z] \tag{3.65}$$

Think of two nested "information spaces": the larger $\{g(Y,Z)\}$ and the smaller $\{h(Z)\}$. In the picture, this should mean a big surface containing a thinner line. Geometrically, we first drop X to the big surface (use Y,Z), then slide within it to the line (use only Z); you land where you would have landed by projecting straight to the line.

The basic form of the theorem,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]], \tag{3.66}$$

is the special case where Z carries no information (take $\sigma(Z)$ to be the space of all constants). Then $\{h(Z)\}$ is just the constant line, "projection onto $\{h(Z)\}$" means take ordinary expectations. Intuitively, averaging after using Y is the same as averaging directly, because "project then project" equals "project once" when the spaces are nested.

- The law of total variance states that for random variables X and Y,

$$\text{Var}(X) = \mathbb{E}\left[\text{Var}(X|Y)\right] + \text{Var}\left(\mathbb{E}[X|Y]\right). \tag{3.67}$$

In fact, there is an elegant way to understand this mathematically using the argument of Pythagorean theorem. Using the geometric interpretation, we can decompose any random variable X into:

$$X = \mathbb{E}[X|Y] + (X - \mathbb{E}[X|Y]), \tag{3.68}$$

We can define a centered inner product in this space such that $\langle U,V\rangle_c = \text{Cov}(U,V)$. Now, the L^2-norm becomes $\|U\|^2 = \langle U,U\rangle_c = \text{Var}(U)$. Therefore, by the Pythagorean theorem based on the new norm, we have

$$\|X\|^2 = \|\mathbb{E}[X|Y]\|^2 + \|X - \mathbb{E}[X|Y]\|^2 \tag{3.69}$$

$$\Longrightarrow \text{Var}(X) = \text{Var}\left(\mathbb{E}[X|Y]\right) + \text{Var}(X - \mathbb{E}[X|Y]). \tag{3.70}$$

Intuitively, the residual $X - \mathbb{E}[X|Y]$ is the part of X that remains unexplained after observing Y. If we fix a value of Y, the conditional mean of this residual is zero, and its conditional variance is exactly the conditional variance of X given Y. Averaging these conditional variances over all possible values of Y gives

$$\text{Var}(X - \mathbb{E}[X|Y]) = \mathbb{E}[\text{Var}(X|Y)]. \tag{3.71}$$

A formal verification of this last identity is left as an end-of-chapter practice in Question 6.

3.4 Introduction to Discrete-Time Stochastic Processes

Up to now, a random variable X has been a single snapshot of uncertainty. Finance needs evolution, so we index uncertainty by time. A discrete-time *stochastic process* (or random process) is a sequence $(X_1,X_2,\cdots)$, equivalently denoted as $\{X_n\}_{n\geq 0}$, on a common probability space $(\Omega,\mathcal{F},\mathbb{P})$; each $\omega\in\Omega$ yields a sample path $n\mapsto X_n(\omega)$.

We begin in discrete time $n=0,1,2,\cdots$ because it is more intuitive and mathematically simple, and it also connects directly to the binomial-tree pricing we use in Chapter 4, the binomial tree pricing models. Continuous-time models that dominate INV 201 will later arise as limits of these stepwise constructions.

3.4.1 Discrete-Time Filtrations

To speak precisely about what is known as time evolves, we formalize the flow of information using *filtrations*. At each time n, $\mathcal{F}_n$ represents the information available up to that time. As time moves forward, information can increase, but it should not disappear.

Definition 3.4.1 (Discrete-Time Filtration). *A filtration is an increasing collection of σ-algebras $\{\mathcal{F}_n\}_{n\geq 0}$ such that $\mathcal{F}_0\subseteq\mathcal{F}_1\subseteq\cdots\subseteq\mathcal{F}$.*

For a stochastic process $\{X_n\}_{n\geq 0}$, the most natural choice of information at time n is all information generated by the observations $X_0,\ldots,X_n$. For random variables $X_0,\ldots,X_n$, the notation

$$\sigma(X_0,\ldots,X_n)$$

represents the σ-algebra generated by random variables $(X_0,\ldots,X_n)$. In plain words, it contains exactly the events that can be decided once the values of $X_0,\ldots,X_n$ are observed. This motivates the definition of *natural filtrations*.

Definition 3.4.2 (Natural Filtration in Discrete Time). *The natural filtration generated by a stochastic process $\{X_n\}_{n\geq 0}$ is defined by*

$$\mathcal{F}_n=\sigma(X_0,X_1,\ldots,X_n),\quad n\geq 0.$$

Thus, $\mathcal{F}_n$ encodes the information available up to and including time n. Events in $\mathcal{F}_n$ are exactly the events that can be determined from the observed history $X_0,X_1,\ldots,X_n$.

With information formalized, we demonstrate the ideas in the simplest discrete-time model of uncertainty, the simple random walk.

Definition 3.4.3 (Simple Random Walk). *Let $\{X_n\}_{n\geq 0}$ be i.i.d. random variables with $\mathbb{P}(X_n=1)=p$ and $\mathbb{P}(X_n=-1)=1-p$. Let $S_0\in\mathbb{R}$ and*

$$S_n=S_0+\sum_{t=1}^{n}X_t,\quad n\geq 1. \tag{3.72}$$

Then $\{S_n\}_{n\geq 0}$ is a simple random walk. When $p=1/2$, $\{S_n\}_{n\geq 0}$ is called a symmetric random walk.

The natural way to describe the information revealed by the random walk up to time n is through its natural filtration

$$\mathcal{F}_n = \sigma(S_1, \cdots, S_n).$$

This means that at time n we know the entire path of positions taken by the process up to that point. An important fact is that this coincides with the filtration generated by the increments:

$$\sigma(S_1, \ldots, S_n) = \sigma(X_1, \cdots, X_n).$$

The reason is straightforward: from the increments we can reconstruct the sums via $S_n = S_0 + X_1 + \cdots + X_n$, while from the sums we can recover each increment as $X_n = S_n - S_{n-1}$. Since each family determines the other, the generated σ-algebras are identical.

This illustrates what a filtration does. All past values are fully determined by $\mathcal{F}_n$: for example, S_{n-1} is measurable with respect to $\mathcal{F}_n$. In contrast, the future remains uncertain. The next increment X_{n+1} is independent of $\mathcal{F}_n$, so it represents genuinely new information.

3.4.2 Discrete-Time Martingales

The word "*martingale*" arose from gambling in history. In 18th and 19th century France, a martingale was referred to the double-or-nothing strategy: keep doubling your bet after each loss so that the first win recovers everything. In theory, if the game is fair and the gambler has unlimited capital, this strategy appears attractive because a win will eventually erase all previous losses. Mathematically it fails (you will eventually run out of capital), but it left us with the right intuition: a fair game has no predictable drift given what you currently know.

In modern finance this idea is central. For most pricing models, once you discount appropriately and adopt the right probability measure (risk-neutral measure $\mathbb{Q}$, which will be introduced later), asset prices behave like martingales. This means that, given current information, they have no systematic tendency to rise or fall in expectation.

Definition 3.4.4 (Discrete-Time Martingale). *Let $\{\mathcal{F}_n\}_{n\geq 0}$ be a filtration on $(\Omega, \mathcal{F}, \mathbb{P})$. A process $\{X_n\}_{n\geq 0}$ is a $\mathbb{P}$-martingale if for all $n \geq 1$:*

*I. **Measurability**: X_n is $\mathcal{F}_n$-measurable.*[11]

*II. **Integrability**: $\mathbb{E}|X_n| < \infty$.*

*III. **Martingale property**:*

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n. \tag{3.73}$$

Intuitively, a martingale is a "fair game": given today's information, the conditional expectation of the next value equals today's value.. Let us explain each of these three properties intuitively.

- Property 1, the measurability is just a fundamental assumption to make sure we are discussing in the correct probability space, and no future information should be used to determine the current value of the process.

[11] Some other texts might frame this property as "X_n is $\mathcal{F}_n$-adapted." These two terms are used interchangeably in INV 201.

- Property 2, the integrability simply guarantees that all expectations $\mathbb{E}[X_n]$ and conditional expectations $\mathbb{E}[X_n|\mathcal{F}_m]$ exist.
- Property 3, the martingale property, means that the process has no predictable systematic drift. Given today's information, the best forecast of the next value is the current value.

The simplest example of discrete martingale is the symmetric random walk, as we will show below. In the exam, verifying a process being a martingale is very commonly seen, so candidates are suggested to obtain a high level of proficiency of problem solving.

Example 3.4.1. Let $\{X_n\}_{n\geq 0}$ be i.i.d. random variables with $\mathbb{P}(X_n = 1) = \mathbb{P}(X_n = -1) = 1/2$. Let $\{S_n\}_{n\geq 0}$ be a symmetric random walk defined as

$$S_n = \sum_{t=1}^{n} X_t, \quad n \geq 1.$$

Show that $\{S_n\}_{n\geq 0}$ is a martingale.

Solution. We show $\{S_n\}_{n\geq 0}$ is a martingale by verifying the three properties:

I. Measurability:

Clearly, $S_n = X_1 + \cdots + X_n$ and it only depends on information up to time n, therefore, S_n is $\mathcal{F}_n$-measurable.

II. Integrability:

First notice that:

$$\mathbb{E}|X_n| = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1.$$

Then, by triangle inequality:

$$\mathbb{E}|S_n| = \mathbb{E}|X_1 + \cdots + X_n| \leq \mathbb{E}|X_1| + \cdots \mathbb{E}|X_n| = n < \infty.$$

III. Martingale property:

$$\begin{aligned} \mathbb{E}[S_{n+1}|\mathcal{F}_n] &= \mathbb{E}[S_n + X_{n+1}|\mathcal{F}_n] \\ &= \mathbb{E}[S_n|\mathcal{F}_n] + \mathbb{E}[X_{n+1}|\mathcal{F}_n] && \text{(Linearity)} \\ &= S_n + \mathbb{E}[X_{n+1}|\mathcal{F}_n] && (\mathcal{F}_n\text{-measurability of } S_n) \\ &= S_n, \end{aligned}$$

where the final step is because X_{n+1} is independent of $\mathcal{F}_n$, and by symmetry,

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] = \mathbb{E}[X_{n+1}] = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot -1 = 0.$$

The symmetry assumption is crucial here. If instead $\mathbb{P}(X_n = 1) = p \neq 1/2$, then

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] = 2p - 1,$$

so the random walk would have a predictable drift and would not be a martingale.

Another important example of discrete-time martingales in option price (particularly the risk-neutral pricing technique) is called the Doob martingale, as we will discuss below.

Theorem 3.4.1 (Doob Martingale). *Let Z be an integrable random variable and $\{\mathcal{F}_n\}_{n\geq 0}$ be a filtration. Let*

$$X_n := \mathbb{E}[Z|\mathcal{F}_n], \quad n \geq 0.$$

Then the process $\{X_n\}_{n\geq 0}$ is a martingale, called the Doob martingale.

Proof. We show $\{X_n\}_{n\geq 0}$ is a martingale by verifying the three properties:

I. Measurability:

By the definition of conditional expectations, $X_n = \mathbb{E}[Z|\mathcal{F}_n]$ is $\mathcal{F}_n$-measurable. Therefore, X_n is $\mathcal{F}_n$-measurable.

II. Integrability:

By the triangle inequality and the law of total expectation,

$$\mathbb{E}|X_n| = \mathbb{E}|\mathbb{E}[Z|\mathcal{F}_n]| \leq \mathbb{E}[\mathbb{E}[|Z||\mathcal{F}_n]] = \mathbb{E}|Z| < \infty,$$

where the final step is by the assumption that Z is an integrable random variable.

III. Martingale property:

$$\begin{aligned}
\mathbb{E}[X_{n+1}|\mathcal{F}_n] &= \mathbb{E}[\mathbb{E}[Z|\mathcal{F}_{n+1}]|\mathcal{F}_n] \\
&= \mathbb{E}[\mathbb{E}[Z|\mathcal{F}_{n+1},\mathcal{F}_n]|\mathcal{F}_n] && (\mathcal{F}_n \subset \mathcal{F}_{n+1}) \\
&= \mathbb{E}[Z|\mathcal{F}_n] = X_n. && \text{(Law of total expectation)}
\end{aligned}$$

□

We discuss some intuition behind the Doob martingale. Think of Z as the random quantity you ultimately care about. Information arrives sequentially: at date n the market knows $\mathcal{F}_n = \sigma(Y_0, \ldots, Y_n)$. Your "best guess" of Z given what the first n messengers told you is exactly $X_n = \mathbb{E}[Z|\mathcal{F}_n]$. As new information arrives, the guess updates in an unbiased way, which is exactly the martingale property.

Last, we introduce two concepts which are closely related to martingales. As we have just discussed, martingales represent processes that do not drift up or down as time evolves. Similarly, when we do have such a drift, the process becomes the **submartingales** or **supermartingales**.

Definition 3.4.5 (Discrete-Time Submartingale and Supermartingale). *Let $\{\mathcal{F}_n\}_{n\geq 0}$ be a filtration on $(\Omega,\mathcal{F},\mathbb{P})$. Assume that for a process $\{X_n\}_{n\geq 0}$, X_n is $\mathcal{F}_n$-measurable and $\mathbb{E}|X_n| < \infty$ hold for all $n \geq 1$. Then, $\{X_n\}_{n\geq 0}$ is a $\mathbb{P}$-submartingale if*

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] \geq X_n. \tag{3.74}$$

Also, $\{X_n\}_{n\geq 0}$ is a $\mathbb{P}$-supermartingale if

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] \leq X_n. \tag{3.75}$$

3.4.3 Discrete-Time Markov Processes

Apart from martingales, another fundamental class is Markov processes. Whereas martingales encode "no predictable drift," Markov processes encode "no hidden memory": conditional on the present state, the future does not depend on the past. In discrete-time modeling this is often the right simplification because the state today summarizes all relevant information for tomorrow.

Definition 3.4.6 (Discrete-Time Markov process). *Let $\{X_n\}_{n\geq 0}$ be a process on $(\Omega,\mathcal{F},\mathbb{P})$ with natural filtration $\{\mathcal{F}_n\}_{n\geq 0}$. We say X is Markov if for every $n \geq 1$ and every Borel set A,*

$$\mathbb{P}\left(X_{n+1} \in A|\mathcal{F}_n\right) = \mathbb{P}\Big(X_{n+1} \in A|X_n\Big). \tag{3.76}$$

Equivalently, for every $n > m \geq 0$ and every measurable function φ, there exists a measurable function g such that

$$\mathbb{E}\left[\varphi(X_n)|\mathcal{F}_m\right] = g(X_m). \tag{3.77}$$

Markovness is useful in computation because it collapses conditioning from the whole history to the current state. In discrete pricing models, this means value functions can be computed backward from maturity using only today's state variable.

Martingales and Markov processes address different questions and neither property implies the other. Some simple processes satisfy both. For the symmetric random walk $S_n = S_0 + \sum_{t=1}^n X_t$ with i.i.d. mean-zero increments X_t, we already know that $\{S_n\}_{n\geq 0}$ is a martingale satisfying $\mathbb{E}[S_{n+1}|\mathcal{F}_n] = S_n$. It is very easy to show that it is also a Markov process:

$$\mathbb{P}\left(S_{n+1} = M|\mathcal{F}_n\right) = \mathbb{P}\left(X_{n+1} = M - S_n|\mathcal{F}_n\right) = \mathbb{P}\left(X_{n+1} = M - S_n\right),$$

which depends on the past only through S_n.

In the below, we discuss two counterexamples to show that martingales and Markov processes do not imply each other.

Example 3.4.2. Let $\{S_n\}_{n\geq 0}$ be an asymmetric random walk defined as

$$S_n = \sum_{t=1}^{n} X_t, \quad n \geq 1,$$

where $\{X_n\}_{n\geq 0}$ be i.i.d. random variables with $\mathbb{P}(X_n = 1) = p \neq 1/2$, $\mathbb{P}(X_n = -1) = 1 - p$. Show that $\{S_n\}_{n\geq 0}$ is a Markov process but not a martingale.

Solution. The proof of $\{S_n\}_{n\geq 0}$ being Markov is the same as the symmetric random walk example, so we do not repeat the details.

To show $\{S_n\}_{n\geq 0}$ is not a martingale, we have

$$\begin{aligned}
\mathbb{E}[S_{n+1}|\mathcal{F}_n] &= \mathbb{E}[S_n + X_{n+1}|\mathcal{F}_n] \\
&= \mathbb{E}[S_n|\mathcal{F}_n] + \mathbb{E}[X_{n+1}|\mathcal{F}_n] && \text{(Linearity)} \\
&= S_n + \mathbb{E}[X_{n+1}|\mathcal{F}_n] && (\mathcal{F}_n\text{-measurability of } S_n) \\
&= S_n + (2p - 1) \neq S_n,
\end{aligned}$$

where the final step is by $\mathbb{E}[X] = p \cdot 1 - (1-p) = 2p - 1 \neq 0$.

More specifically, when $p > 1/2$, $\{S_n\}_{n\geq 0}$ is a submartingale drifting up over time, which makes intuitive sense since each increment X_n has a larger probability to be positive than negative. Similar arguments apply to the case when $p < 1/2$, and $\{S_n\}_{n\geq 0}$ is a supermartingale.

Example 3.4.3. At time 0, draw a step size $\theta \in \{1,2\}$ with $\mathbb{P}(\theta = 1) = \mathbb{P}(\theta = 2) = 1/2$, and keep it fixed. Let $\{X_n\}_{n\geq 1}$ be i.i.d. with $\mathbb{P}(X_n = \pm 1) = 1/2$, independent of θ. Define $S_0 = 0$ and for $n \geq 1$

$$S_{n+1} = S_n + \theta X_{n+1} = \theta(X_1 + \cdots + X_{n+1}),$$

with natural filtration $\mathcal{F}_n = \sigma(\theta, X_1, \cdots, X_n)$.

$\{S_n\}_{n\geq 0}$ is a martingale but not a Markov process.

Solution. To show that $\{S_n\}_{n\geq 0}$ is a martingale, we verify the three properties:

I. Measurability:

Clearly, since $S_n = \theta(X_1 + \cdots + X_n)$ and it only depends on information up to time t (the information of θ is available at time 0), S_n is $\mathcal{F}_n$-measurable.

II. Integrability:

First notice that $\mathbb{E}|X_n| = 1$ and $\mathbb{E}|\theta| = 3/2$. Then, by the independence between θ and $(X_1, \cdots, X_n)$ and the triangle inequality:

$$\mathbb{E}|S_n| = \mathbb{E}|\theta(X_1+\cdots+X_n)| = \mathbb{E}|\theta| \cdot \mathbb{E}|X_1+\cdots+X_n| \leq \frac{3}{2} \underbrace{(\mathbb{E}|X_1| + \cdots \mathbb{E}|X_n|)}_{=n} = \frac{3n}{2} < \infty.$$

III. Martingale property:

$$\begin{aligned}
\mathbb{E}[S_{n+1}|\mathcal{F}_n] &= \mathbb{E}[S_n + \theta X_{n+1}|\mathcal{F}_n] \\
&= \mathbb{E}[S_n|\mathcal{F}_n] + \mathbb{E}[\theta X_{n+1}|\mathcal{F}_n] && \text{(Linearity)} \\
&= S_n + \theta \cdot \underbrace{\mathbb{E}[X_{n+1}|\mathcal{F}_n]}_{=\mathbb{E}[X_{n+1}]=0} && (\mathcal{F}_n\text{-measurability of } S_n \text{ and } \theta) \\
&= S_n.
\end{aligned}$$

Next, to show that $\{S_n\}_{n\geq 0}$ is not a Markov process, consider two histories with the same current state $S_2 = 0$ at time 2:

- History 1: At time 0, we drew $\theta = 1$; at time 1, we got $X_1 = 1$ and $S_1 = \theta X_1 = 1$; at time 2, we got $X_2 = -1$ and $S_2 = S_1 + \theta X_2 = 1 - 1 = 0$.
- History 2: At time 0, we drew $\theta = 2$; at time 1, we got $X_1 = 1$ and $S_1 = \theta X_1 = 2$; at time 2, we got $X_2 = -1$ and $S_2 = S_1 + \theta X_2 = 2 - 2 = 0$.

In History 1 where $\theta = 1$, we have

$$\mathbb{P}(S_3 = 1) = \mathbb{P}(S_2 + \theta X_3 = 1) = \mathbb{P}(X_3 = 1) = \frac{1}{2}.$$

However, in History 2 where $\theta = 2$, we have

$$\mathbb{P}(S_3 = 1) = \mathbb{P}(S_2 + \theta X_3 = 1) = \mathbb{P}(X_3 = 1/2) = 0.$$

This says that the conditional law of S_3 depends on the full history (through θ), not just on the current value $S_2 = 0$; therefore $\{S_n\}_{n\geq 0}$ is not Markov.

3.5 Practice Questions

1. Let $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}([0,1])$, and $\mathbb{P}$ be the uniform probability measure on $[0,1]$. Define

$$X(\omega) = (\omega - 0.8)^+.$$

   Calculate the following:

   (a) $\mathbb{P}(X < 0)$

   (b) $\mathbb{P}(X \leq 0)$

   (c) $\mathbb{P}(0.1 \leq X \leq 0.3)$.

2. Assume that each year has 365 days (omitting leap years). In a group of n people, what is the expected number of distinct birthdays among the n people, i.e., the expected number of days on which at least one of the people was born?

3. Let $X \sim \mathcal{N}(\mu,\sigma^2)$ with $\sigma > 0$. Compare the value of $\mathbb{E}[e^X]$ and $e^{\mathbb{E}[X]}$ by using each of the two approaches:

   (a) Comparing $\mathbb{E}[e^X]$ and $e^{\mathbb{E}[X]}$ directly.

   (b) Applying Jensen's inequality.

4. Let X and Y be two independent random variables on $(\Omega,\mathcal{F},\mathbb{P})$ with $\mathbb{P}(Y = 1) = p$ and $\mathbb{P}(Y = 2) = 1 - p$. Assume that $\mathbb{E}[X] = \mu$ and $\text{Var}(X) = \sigma^2$. Derive $\mathbb{E}[Z]$, where $Z = X^Y$.

5. Let $\Omega = [-1,1]$, $\mathcal{F} = \mathcal{B}([-1,1])$, and $\mathbb{P}$ be the uniform probability measure on $[-1,1]$. Say X and Y are random variables on $(\Omega,\mathcal{F},\mathbb{P})$.

   (a) Find $\mathbb{E}[X|Y]$ if $Y(\omega) = \omega^3$.

   (b) Find $\mathbb{E}[X|Y]$ if $Y(\omega) = \omega^2$.

6. In Section 3.3, we have discussed the law of total variance. For random variables X and Y, we have

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X|Y)] + \text{Var}(\mathbb{E}[X|Y]).$$

   We have also introduced that geometric interpretation of conditional expectations, where we can use the Pythagorean theorem to interpret the the law of total variance, as follows:

$$\text{Var}(X) = \text{Var}\left(\mathbb{E}[X|Y]\right) + \text{Var}(X - \mathbb{E}[X|Y]).$$

   This question guides you to complete the proof. Let $\varepsilon = X - \mathbb{E}[X|Y]$ be the residual of the conditional expectation.

   (a) Show that $\mathbb{E}[\varepsilon] = \mathbb{E}[\varepsilon|Y] = 0$.

(b) Show that $\mathrm{Var}(\varepsilon|Y) = \mathrm{Var}(\mathbb{E}[X|Y])$.

(c) Using the results from (a) and (b) to show that

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

which recovers the formula of the law of total variance from the Pythagorean theorem.

7. Let $\{X_n\}_{n\geq 0}$ be a martingale with a natural filtration $\{\mathcal{F}_n\}_{n\geq 0}$. For an integrable convex function φ, show that $\{\varphi(X_n)\}_{n\geq 0}$ is a submartingale.

8. Let $\{S_n\}_{n\geq 0}$ be an asymmetric random walk defined as

$$S_n = \sum_{t=1}^{n} X_t, \quad n \geq 1,$$

where $\{X_n\}_{n\geq 0}$ be i.i.d. random variables with $\mathbb{P}(X_n = 1) = p \neq 1/2$, $\mathbb{P}(X_n = -1) = 1 - p$.

(a) Find a constant c such that $\{Y_n\}_{n\geq 0}$ is a martingale with

$$Y_n = S_n - cn.$$

(b) Show that the process $\{Z_n\}_{n\geq 0}$ with

$$Z_n = \left(\frac{1-p}{p}\right)^{S_n}$$

is a martingale.

3.6 Solution to Practice Questions

1. We use $\mathcal{L}$ to denote the uniform measure on [0,1].
   (a) Notice that $X \geq 0$ holds for all $\omega \in \Omega$, thus
   $$\mathbb{P}(X < 0) = \mathcal{L}(\emptyset) = 0.$$
   (b) Notice that $\omega \in \Omega = [0,1]$, we have
   $$\mathbb{P}(X \leq 0) = \mathbb{P}(\underbrace{(\omega - 0.8)^+}_{\geq 0} \leq 0) = \mathbb{P}((\omega - 0.8)^+ = 0) = \mathbb{P}(0 \leq \omega \leq 0.8) = \mathcal{L}([0,0.8]) = 0.8.$$
   (c) Notice that $\omega \in \Omega = [0,1]$, we have
   $$\mathbb{P}(0.1 \leq X \leq 0.3) = \mathbb{P}(0.1 \leq (\omega - 0.8)^+ \leq 0.3) = \mathbb{P}(0.9 \leq \omega \leq 1) = \mathcal{L}([0.9,1]) = 0.1.$$

2. Let X be the number of distinct birthdays, and we can write it as
   $$X = \sum_{i=1}^{n} \mathbf{1}_i,$$
   where $\mathbf{1}_i = 1$ if the i-th day is a distinct birthday. Then, we have
   $$\mathbb{E}[\mathbf{1}_i] = \mathbb{P}(\text{At least 1 people were born on day } i) = 1 - \left(\frac{364}{365}\right)^n.$$
   Last, by the linearity of expectations,
   $$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{1}_i\right] = \sum_{i=1}^{n} \mathbb{E}[\mathbf{1}_i] = n\left(1 - \left(\frac{364}{365}\right)^n\right).$$

3. (a) Recall that the MGF of $\mathcal{N}(\mu,\sigma^2)$ is
   $$M(t) = \mathbb{E}[e^{tX}] = e^{\mu t + \frac{1}{2}\sigma^2 t^2}.$$
   Letting $t = 1$ gives
   $$\mathbb{E}[e^X] = e^{\mu + \frac{1}{2}\sigma^2} > e^{\mu} = e^{\mathbb{E}[X]}.$$
   (b) Let $\varphi(x) = e^x$. Clearly $\varphi(x)$ is a convex function since $\varphi''(x) = e^x \geq 0$. Therefore, $\mathbb{E}[\varphi(X)] \geq \varphi(\mathbb{E}[X])$, as shown in (a).

4. We first derive $\mathbb{E}[Z|Y]$:
   $$\mathbb{E}[Z|Y] = \mathbb{E}[X^Y] = \begin{cases} \mathbb{E}[X] = \mu, & Y = 1 \text{ (with probability } p\text{)}, \\ \mathbb{E}[X^2] = \mu^2 + \sigma^2, & Y = 2 \text{ (with probability } 1 - p\text{)}. \end{cases}$$
   Then, applying the law of total expectation,
   $$\mathbb{E}[Z] = \mathbb{E}[\mathbb{E}[Z|Y]] = p \cdot \mathbb{E}[Z|Y = 1] + (1 - p) \cdot \mathbb{E}[Z|Y = 2] = p\mu + (1 - p)(\mu^2 + \sigma^2).$$

Who We Are

A Benefit Corporation Experienced at Teaching Actuaries!

EXPERIENCED

More than 50 years experience helping students prepare and pass actuarial exams! We are an eLearning technology and education company leveraging experts in the field to constantly update our learning content in a format that works for you.

TRUSTWORTHY

ACTEX Learning is a leading US based provider of study materials for actuarial exams. Our authors and content contributors are renowned academics and Actuaries that are proud to have their names on the cover of our manuals and textbooks!

MISSION FOCUSED

We are a Benefit Corporation focusing on the mission of accessible high quality actuarial education. We're dedicated to empowering actuarial students by offering test prep materials that are not just effective and efficient but also tailored to suit every type of student.

Become an ACTEX Champion

Join our Global ACTEX Champion Program and bring the benefits to your Actuarial Club. To learn more about the program, scan the QR code on the right. If you have any questions or would like to speak to a Champion Coordinator, please do not hesitate to reach out to us at champions@actexlearning.com.

ACTEX Learning

ACTEX Has the Solutions to Help You with Exam Prep

Study Guide

Our study guides offer the most recommended actuarial prep program. Search our interactive manuals for different topics and toggle easily between concepts and study materials.

Available for P, FM, FAM, ALTAM, ASTAM, PA. ATPA, MAS-I, MAS-II, CAS 5, CAS 6 US & CAN, CAS 7, CAS 8, CAS 9

GOAL

Want to know you're prepared for your exam? Practice efficiently with our robust database of questions and solutions and predict your success through GOAL's innovative scoring system. GOAL also features dedicated instructor support so you can get help where you need it and pass the exam with confidence!

Available for P, FM, FAM, ALTAM, ASTAM, MAS-I, MAS-II, CAS 5, CAS 6 US & CAN

Flashcards

Master key topics and formulas with our flashcards, which allow you to filter by topic. To help you direct your focus, each card is rated to indicate its importance on the exam.

Available for P, FM, FAM, ALTAM, ASTAM, PA. ATPA, MAS-I, MAS-II, CAS 5, CAS 6 US & CAN, CAS 7, CAS 8, CAS 9

Videos

Studies have shown video learning can lead to better retention. We offer hours of video instruction to aid you in your studies. They're a great way to deepen your learning on challenging topics and gain a variety of perspectives from our expert instructors.

Available for P, FM, FAM, ALTAM, ASTAM, PA. ATPA, MAS-I, MAS-II, CAS 5, CAS 6 US & CAN, CAS 7, CAS 8, CAS 9

VEE Courses

ACTEX offers convenient online courses approved for CAS VEE credits. All courses are available on-demand. Students complete the courses at their own pace and take the final exams on the date of their choosing.

Available for Accounting & Finance, Mathematical Statistics, and Economics

Study Materials are Available for the Following:

SOA: P, FM, FAM, ALTAM, ASTAM, SRM, PA, ATPA, CFE, GI, GH, ILA, RET

CAS: MAS-I, MAS-II, CAS 5, CAS 6C, CAS 6US, CAS 7, CAS 8, CAS 9

Graded Mock Exams

The ACTEX Graded Mock Exam is a great way to predict your exam outcome! Before you take the official exam - take the new ACTEX Graded Mock Exam and get feedback from an expert. The ACTEX Graded Mock Exam has all the typical elements your SOA exam will have. This can help you evaluate your progress. The questions and format are set up just like the SOA exam.

Available for ALTAM, ASTAM, PA

Bootcamp

ACTEX Bootcamps provide a more individualized approach, allow you to ask questions in real time, and boost your last-minute learning. You'll review the harder topics on the exam, as well as common errors and exam strategy. All classes are recorded for future on-demand viewing.

Available for P, FM, FAM, SRM

Online Courses

Alongside our P & FM study guide, this course is comparable to a one-semester college class. This course offers SOA Exam practice problems, video solutions, timed practice tests, sample questions, and more. You'll also have 1:1 email support from an instructor for 180 days after purchase.

The Advanced topics in Predictive Analytics video course is designed to help you more easily climb the steep ATPA learning curve. This module-focused video course for Topic 3 in the syllabus, includes videos, end of module assessments and lecture slides. This video course is a deep dive into the three modules. Access to an instructor during the duration of the course as well as participation in a discussion forum.

Available for P, FM, and ATPA

Formula Sheets

This at-a-glance tool helps you memorize and recall key formulas and information. It covers important formulas needed to prepare your exam. Also, it's an easy-to-print format you can study with, no matter where you are.

Available for P, FM, FAM, ALTAM, ASTAM, PA. ATPA, MAS-I, MAS-II, CAS 5

Textbooks

Looking for Extra Preparation?

Explore our range of textbooks designed to support your studies. From recommended to required readings, ACTEX offers exceptional materials to help you succeed.

ACTEX Learning

Use GOAL to Practice What You've Learned

- Over 22,000 exam-style problems with detailed solutions
- 3 learning modes (Practice, Quiz, Simulated Exams)
- 3 levels of difficulty (Core, Advanced and Mastery)
- You control your topics and sub-topics
- Dedicated instructor support

GOAL is currently available for the following SOA & CAS Exams:

Exam P	1,050+ Questions
Exam FM	1,500+ Questions
Exam FAM-L	1,300+ Questions
Exam FAM-S	900+ Questions
Exam FAM	2,300+ Questions
Exam ALTAM	1,400+ Questions
Exam ASTAM	1,300+ Questions
Exam SRM	1,150+ Questions
Exam MAS-I	1,050+ Questions
Exam MAS-II	850+ Questions
Exam CAS 5C	500+ Questions
Exam CAS 6US	550+ Questions
Exam CAS 6CA	650+ Questions

Use GOAL Score to Gauge Your Exam Readiness

Measure how prepared you are to pass your exam with a tool that suits any study approach. A GOAL Score of 70 or above indicates readiness.

Your score is broken into categories, allowing you to study efficiently by concentrating on problem areas. GOAL Score quantifies your exam readiness by measuring both your performance and the consistency of your performance. Your GOAL Score also analyzes your strengths and weaknesses by category, topic, and level of difficulty.

Scan to Learn More