
CONTENTS

I INTRODUCTION TO PROBABILITY THEORY

1	PROLOGUE	5
2	INTRODUCTION	7
2.1	History	8
3	MATHEMATICAL PRELIMINARIES	11
3.1	The Completeness Axiom	11
3.2	Countability and Algebra of Sets	13
3.3	Functions of Sets	15
3.4	Sequences in \mathbb{R}	18
3.5	Limits Supremum and Infimum	23
3.6	Convergence of a Sequence of Functions	25
4	THE PROBABILITY MEASURE SPACE	33
4.1	The Probability Triple	37
4.2	The Lebesgue Measure on the Unit Interval	41
5	LEBESGUE INTEGRATION	45
6	RANDOM VARIABLES AND THEIR EXPECTATIONS	49
6.1	Random Variable as a Set Function	49
6.2	Expectation of a Random Variable	51
6.3	Independence	54
6.4	Useful Inequalities	55
7	IMPORTANT LIMIT THEOREMS IN PROBABILITY THEORY	59
7.1	The Borel-Cantelli Lemmas	59
7.2	Strong Law of Large Numbers	60
7.3	Weak Law of Large Numbers	62
7.4	The Central Limit Theorem	62

7.5	The Law of Iterated Logarithms	65
8	CHARACTERISTIC FUNCTION	67

Part I

INTRODUCTION TO PROBABILITY
THEORY

PROLOGUE

Many good books about probability theory have been written. By usual standards of "good" in mathematics, this means the materials are rigorous. The implication is that the level of preparation to follow a "good" text is quite demanding - solid calculus and exposure to real analysis. For modern statistics majors who need to spend (more and more) time learning how to program and work with large data sets, it is inevitable that they enter the study of probability theory with some background deficiencies. Consequently, traditional material on learning probability theory may not be quite helpful.

The set of notes here have been prepared to help statistics majors learn something about the foundations of probability theory. The general layout consists of some motivational material, followed by definitions, theorems, and examples. The proof of some theorems are skipped, whereas those that are illustrative are given.

Happy learning!

T.F. Khang
Oct 2020 - Jan 2021

INTRODUCTION

The simplest laws of nature are statements that declare that some event of interest will certainly occur or certainly not occur, when a set of conditions are fulfilled. For example, if we take a fish out of water long enough (condition), it will die (event of interest) for sure. If a white-coloured substance does not contain starch (condition), then when iodine is added to it, the substance will not turn blue (event of interest) for sure.

An event E is said to be random with respect to a set of conditions, if it sometimes occurs, and sometimes does not. For example, a patient may receive a heart bypass operation (condition). He or she may or may not be alive (outcome) six months after the operation (condition). We cannot be certain about the outcome in a single case. But if we consider many similar instances and observe the relative frequency of survival after six months, then perhaps this value can give us some idea about how risky the operation is in general.

The study of probability is concerned with quantifying the degree of uncertainty that we have regarding the occurrence of E with respect to a set of conditions. To many, probability is understood intuitively as a kind of measure of how likely an event of interest will occur. Unfortunately, intuition is an unreliable companion. It is necessary that an under-

standing of probability be based on a formal mathematical framework.

The following example, known as the Monty Hall problem, illustrates that intuition often misleads in probability calculations.

Example 2.0.1The Monty Hall problem

You are on a game show and shown the choice of three doors. Behind one of the doors is a car; behind the others, goats. You pick a door, say No. 1. The host, who can see what is behind the doors, shows you another door (say, No. 3) which has a goat. You are then asked whether you want to pick door No. 2. Is your chance of getting the car (assuming you find cars more attractive than goats) better if you change your mind, or it does not matter?



Try solving the Monty Hall problem yourself before consulting the internet. Many well-known scientists, including the famous mathematician Paul Erdős, gave the wrong answer to this problem. Interestingly, Erdős, who discovered many new results using his sharp mathematical intuitions, was reported to have remained sceptical of the correct solution, until he was shown the result of a computer simulation for the Monty Hall problem!

2.1 HISTORY

Early humans believed that every aspect of their life, or that of nature, is determined by an abstract force called *fate* (producing certainty). Yet, unknown to them, they used mechanisms that we recognise today as essentially random

(producing uncertainty) to have a "sneak peek" of what is in store for them in the future.

Example 2.1.1 When faced with uncertain outcomes of some important event (e.g. a forthcoming battle, succession), kings in the past would frequently summon a spiritual person to "seek guidance from the spirits". This practice is known as divination. Often, the spiritual person (a shaman, priest, etc.) would perform some rituals. For example, in ancient Rome, animals were sacrificed, disembowelled, and their entrails inspected for specific patterns. This is known as the practice of "haruspex". In ancient China, shamans burned the plastron of turtles or shoulder bones of oxen, and then examined the resulting cracked patterns. Certain patterns were considered auspicious, whereas others were considered ill-omen. ♣

While probability remained entangled with mystical pursuits, it made little progress. A more sophisticated understanding of probability eventually came from the more secular aspect of living - that of gaming. When humans discovered how to gamble, they spent much time thinking about how to discover some kind of "sure-win" rule by observing the outcomes of gambling games over a long period of time. But every person had a different opinion! For a simple game of guessing heads or tails of a tossed coin, one gambler might feel that if there is a run of heads, the next toss is almost sure to be a tail. Another might use the strategy of betting the opposite of whatever that came up in the most recent toss. The "sure-win" strategy depended on whom one was talking to.

Eventually, analysis of games of chance caught the attention of mathematicians. The first book that discussed elements of probability - *Liber de Ludo Aleae* (Book on Games

of Chance) - was written by Gerolamo Cardano (written in 1564, but published in 1663). For the first time, the sample spaces of the toss of a single die and two dice were properly enumerated. Then, Blaise Pascal and Pierre de Fermat - two French mathematicians, solved the problem of fair division of stakes in an interrupted game in 1654.

Two players throw a pair of dice 24 times. Each player puts up equal stakes. The first player to get double-six for a certain number of times will win all stakes. If the game is stopped before either player has won, how should the players divide the stakes so that it is fair to both of them?

The publication of the *Doctrine of Chance*, in 1718 by Abraham de Moivre - another mathematician, is considered a milestone in the development of probability theory. The book contains calculated probabilities to various gambling scenarios commonly encountered in Europe, in the form of ratio of the number of occurrences of the outcome of interest to the total number of possible occurrences. Later in 1812, Pierre de Laplace, in his book *Théorie Analytique des Probabilités*, showed that probabilistic ideas can be applied to science and practical problems.

Gradually, contributions from real analysis and increasing understanding of the importance of randomness in natural phenomena culminated in the development of modern probability theory by the Russian mathematician Andrey Kolmogorov, who laid down the axioms of probability theory in 1933. By then, probability had become an established branch of mathematics. If you find the theory of probability hard, it is a normal reaction. After all, its formal structure had only been around for 90 years!

MATHEMATICAL PRELIMINARIES

Here, we cover some materials from real analysis as background preparation for the coming chapters.

3.1 THE COMPLETENESS AXIOM

Definition 3.1.1 (Supremum of a set of real numbers). Let E be a nonempty set of real numbers.

- (i) We say that E is bounded above if and only if there exists a number $M \in \mathbb{R}$, called the upper bound of E , such that $a \leq M$ for all $a \in E$.
- (ii) We call a number s the supremum of E , and write $s = \sup E$, if and only if s is an upper bound of E , and $s \leq M$ for all upper bounds M of E . ♣

Example 3.1.1 If $E = [0, 1]$, then $\sup E = 1$. The reasoning is as follows. An upper bound of E is just 1 (requirement (i) of Definition 3.1.1). Of course, 2, 3, or any real number greater than 1 is also an upper bound of E . However, only 1 is the *smallest* of all possible upper bounds. This satisfies requirement (ii) of Definition 3.1.1. Hence $\sup E = 1$. In this case, $\max E = \sup E = 1$.

Remark. If $E = [0, 1)$, then we still have $\sup E = 1$, but $\max E$ does not exist, since we can always find some element in E that is arbitrarily close to 1.

Theorem 3.1.1 (Approximation property for suprema). If E has a supremum and $\varepsilon > 0$ is any positive number, then there exists some point $a \in E$ such that

$$\sup E - \varepsilon < a \leq \sup E.$$



The completeness axiom is stated simply as follows: If E is a nonempty subset of \mathbb{R} that is bounded above, then E has a supremum.

Theorem 3.1.2 (Density of rationals). Let $a, b \in \mathbb{R}$ and $a < b$. There exists some $q \in \mathbb{Q}$ such that $a < q < b$.



Definition 3.1.2 (Infimum of a set of real numbers). Let E be a nonempty set of real numbers.

(i) We say that E is bounded below if and only if there exists an a number $M \in \mathbb{R}$ called the lower bound of E , such that $a \geq M$ for all $a \in E$.

(ii) We call a number t the infimum of E , and write $t = \inf E$, if and only if t is a lower bound of E , and $t \geq M$ for all lower bounds M of E .



Example 3.1.2 In Example 3.1.1, we have $\inf E = 0$. The reasoning is as follows. A lower bound of E is 0 (requirement (i) of Definition 3.1.2). Other possible lower bounds are -1 , -2.71 , etc. However, only 0 is the largest of all possible lower bounds. This satisfies requirement (ii) of Definition 3.1.2. In this case too, $\min E = \inf E = 0$.

Remark. If $E = (0, 1)$, then $\inf E = 0$, but $\min E$ does not exist.

Definition 3.1.3 (Boundedness of a set). A set E is bounded if and only if it is bounded below and above. ♣

Example 3.1.3 Consider the set $E = \{x \in \mathbb{R} : x^2 - 6x + 5 = 0\}$. Factorising $x^2 - 6x + 5 = (x - 5)(x - 1)$, we find that the roots of the equation are $E = \{1, 5\}$. Hence, $\sup E = 5$ and $\inf E = 1$.

Theorem 3.1.3. Let $E \subset \mathbb{R}$ be nonempty. Then,

- (i) E has a supremum if and only if $-E$ has an infimum, in which case, $\inf(-E) = -\sup E$.
- (ii) E has an infimum if and only if $-E$ has a supremum, in which case, $\sup(-E) = -\inf E$.

Theorem 3.1.4 (Monotone property). Let A and B be two nonempty subsets of \mathbb{R} , and that $A \subseteq B$. Then,

- (i) If B has a supremum, then $\sup A \leq \sup B$.
- (ii) If B has an infimum, then $\inf A \geq \inf B$.

3.2 COUNTABILITY AND ALGEBRA OF SETS

Definition 3.2.1. Let E be a set of real numbers. We say that

- (i) E is finite if and only if either $E = \emptyset$ or there exists an $n \in \mathbb{N}$ and a one-to-one function from $\{1, 2, \dots, n\}$ onto E .
- (ii) E is countable if and only if there is a one-to-one function from \mathbb{N} onto E .
- (iii) E is at most countable if and only if E is finite or countable.
- (iv) E is uncountable if and only if E is neither finite nor countable.

By Definition 3.2.1, a set is countable if it has the same number of elements as \mathbb{N} . It is finite if it has less, and uncountable if it has more. An example of an infinite set that is countable is $E = \{2, 4, 8, 16, \dots, 2^n, \dots\}$. This is so because there is a one-to-one mapping from the set of integers $\{1, 2, 3, \dots, n, \dots\}$ to $\{2, 2^2, 2^3, \dots, 2^n, \dots\}$.

Theorem 3.2.1. The set of open interval $E = (0, 1)$ is uncountable.

Proof. We will prove this statement by contradiction. Suppose $E = (0, 1)$ is countable. Then, by definition of countability, there exists some function f on \mathbb{N} such that $f(1), f(2), \dots$ map to every element of $(0, 1)$. We can write down the numbers $f(i), i \in \mathbb{N}$ in decimal notation:

$$\begin{aligned} f(1) &= 0.\alpha_{11}\alpha_{12}\dots, \\ f(2) &= 0.\alpha_{21}\alpha_{22}\dots, \\ f(3) &= 0.\alpha_{31}\alpha_{32}\dots, \\ &\vdots \end{aligned}$$

where α_{ij} are the j th digit in the decimal expansion of $f(i)$ and not one of these expansions end in 9's. Consider x , the number whose decimal expansion is given by $0.\beta_1\beta_2\dots$, where $\beta_k = \alpha_{kk} + 1$ if $\alpha_{kk} \leq 5$, and $\alpha_{kk} - 1$ if $\alpha_{kk} > 5$. By this construction, x is a number in $(0, 1)$ whose decimal expansion does not contain 9. By hypothesis, there is some $j \in \mathbb{N}$ such that $f(j) = x$, that is, $f(j) = 0.\alpha_{j1}\alpha_{j2}\dots = 0.\beta_1\beta_2\dots$, implying that $\beta_k = \alpha_{kk} = \alpha_{kk} \pm 1$, which gives $0 = \pm 1$, a contradiction. \square

Theorem 3.2.2 (Countability of sets of real numbers). Let A and B be two nonempty sets of real numbers.

(i) If $A \subseteq B$, and B is at most countable, then A is at most

countable.

(ii) If $A \subseteq B$, and A is uncountable, then B is uncountable.

(iii) \mathbb{R} is uncountable.

Theorem 3.2.3 (De Morgan's Laws). Let A_i , $i = 1, 2, \dots, n$ be a collection of subsets of some nonempty set X . Then

$$(i) (\bigcup_{i=1}^n A_i)^c = \bigcap_{i=1}^n A_i^c$$

$$(ii) (\bigcap_{i=1}^n A_i)^c = \bigcup_{i=1}^n A_i^c$$

Proof. (i) For some x in the left-hand side set, we have $x \in X$ and $x \notin \bigcup_{i=1}^n A_i$. This means $x \in A_i^c$ for all $i = 1, 2, \dots, n$, that is, $x \in \bigcap_{i=1}^n A_i^c$.

(ii) The result of (i) holds as well for $(\bigcup_{i=1}^n A_i^c)^c = \bigcap_{i=1}^n (A_i^c)^c$. The proof is complete by complementing both sides. \square

3.3 FUNCTIONS OF SETS

Definition 3.3.1. Let X be Y be sets, and f a function that maps elements in X onto Y , i.e. $f : X \rightarrow Y$. The *image* of a set $E \subseteq X$ under f is defined as

$$f(E) = \{y \in Y : y = f(x) \text{ for some } x \in E\}.$$

The *inverse image* of a set $E \subseteq Y$ under f is the set

$$f^{-1}(E) = \{x \in X : f(x) = y \text{ for some } y \in E\}.$$

Note that f need not be a one-to-one function for $f^{-1}(E)$ to be defined. In general, $f^{-1}(E)$ is the inverse image of E under f , and not the image of E under the inverse function f^{-1} , unless f is one-to-one.

Example 3.3.1 Consider the set $E = \{-2, -1, 1, 2\}$. Let $x \in E$, and define f to be the one-to-one linear function

$f : x \rightarrow x + 1$. Then, $f(E) = \{-1, 0, 2, 3\}$. For the set $\{-2, -1, 1, 2\}$ in Y , its inverse image under f is $\{-3, -2, 0, 1\}$, since these are the values in X such that $f(x) = x + 1 = y$. Since f is one-to-one in this case, if we had used the inverse function of f , which is $f^{-1}(x) = x - 1$, $E = \{-2, -1, 1, 2\}$ would have been mapped to $\{-3, -2, 0, 1\}$ as well.

For the same set E , consider a many-to-one mapping such as $f : x \rightarrow x^2$. We have $f(E) = \{1, 4\}$. Then, $f^{-1}(E) = \{-\sqrt{2}, -1, 1, \sqrt{2}\}$, since these are the values in X such that $f(x) = x^2 = y$, for some $y \in E$. We cannot use the inverse function of f in this case, because it does not exist as $f(x) = x^2$ is many-to-one.

Example 3.3.2 Let $E = (-1, 5]$. For the function $f(x) = x^2$, we have $f(E) = [0, 25]$. Then, $f^{-1}(E) = [-\sqrt{5}, \sqrt{5}]$.

Example 3.3.3 Let $E = (1, 2]$. For the function $f(x) = \log x$, we have $f(E) = (0, \log 2]$. Then, $f^{-1}(E) = (e, e^2]$ (Fig. 3.3.1). Note that in this example, since $f(x) = \log x$ is one-to-one, treating $f^{-1}(E)$ as the image of E under the the inverse function $f^{-1}(x) = e^x$ works as well.

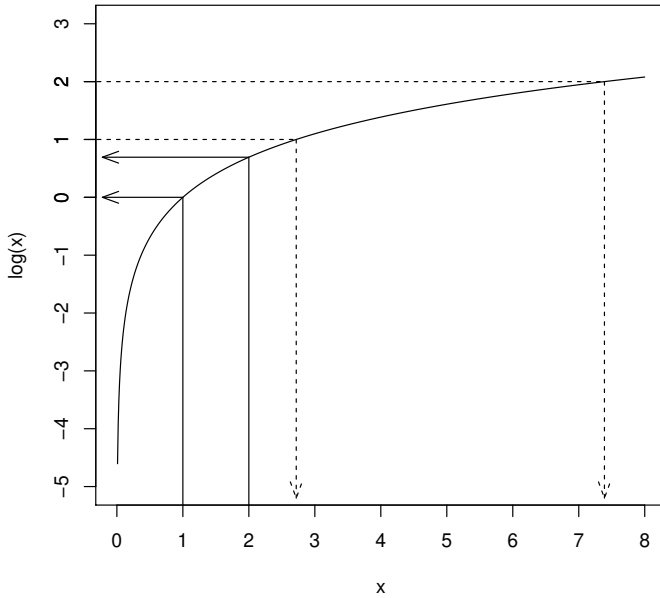


Figure 3.3.1: The solid, arrowed lines define the boundaries of the set $(0, \log 2]$, which is the image of $(1, 2]$ via $f(x) = \log x$. The broken, arrowed lines define the boundaries of the set $(e, e^2]$, which is the set of the inverse image of $(1, 2]$.

Theorem 3.3.1 (Algebra of functions of sets). Let X and Y be sets, and define $f : X \rightarrow Y$. If A_i , $i = 1, 2, \dots, n$ is a

collection of subsets of X , then

(i)

$$\begin{aligned} f\left(\bigcup_{i=1}^n A_i\right) &= \bigcup_{i=1}^n f(A_i) \\ f\left(\bigcap_{i=1}^n A_i\right) &\subseteq \bigcap_{i=1}^n f(A_i) \end{aligned}$$

(ii) If B and C are subsets of X , then

$$f(C \setminus B) \supseteq \bigcap_{i=1}^n f(C) \setminus f(B)$$

(iii) If $A_i, i = 1, 2, \dots, n$ is a collection of subsets of Y , then

$$\begin{aligned} f^{-1}\left(\bigcup_{i=1}^n A_i\right) &= \bigcup_{i=1}^n f^{-1}(A_i) \\ f^{-1}\left(\bigcap_{i=1}^n A_i\right) &= \bigcap_{i=1}^n f^{-1}(A_i) \end{aligned}$$

(iv) If B and C are subsets of Y , then

$$f^{-1}(C \setminus B) = f^{-1}(C) \setminus f^{-1}(B).$$

(v) If $E \subseteq f(X)$, then $f(f^{-1}(E)) = E$; if $E \subseteq X$, then $f^{-1}(f(E)) \supseteq E$. ♣

3.4 SEQUENCES IN \mathbb{R}

A sequence of real numbers consists of a list of ordered numbers x_1, x_2, \dots . Thus, sequence x_1, x_2, \dots is identical to sequence y_1, y_2, \dots if and only if $x_i = y_i$ of all $i = 1, 2, \dots$. Generally, we denote a sequence as $\{x_n\}_{n \in \mathbb{N}}$.

Example 3.4.1 The sequence of numbers $\{1, 1/2, 1/3, 1/4, \dots\}$ is represented as $\{1/n\}_{n \in \mathbb{N}}$. The sequence $\{\cos k\pi\}_{k \in \mathbb{N}}$ is $\{-1, 1, -1, 1, \dots\}$.

A set $\{x_n : n \in \mathbb{N}\}$ is different from a sequence $\{x_n\}_{n \in \mathbb{N}}$. For example, the sequence of natural numbers $\{1, 2, 3, 4, \dots\}$ is not the same as the sequence $\{2, 3, 1, 4, \dots\}$. However, as a set, $\{1, 2, 3, 4, \dots\}$ is identical to $\{2, 3, 1, 4, \dots\}$.

The concept of limit is fundamental to the study of behaviour of sequences.

Definition 3.4.1 (Limit of a sequence of real numbers). We say that a sequence of real numbers $\{x_n\}$ *converges* to some real number $a \in \mathbb{R}$, if and only if for every $\varepsilon > 0$ there exists an $N \in \mathbb{N}$ such that $|x_n - a| < \varepsilon$ if $n \geq N$. ♣

Example 3.4.2 The sequence of numbers $\{\frac{1}{n}\}_{n \in \mathbb{N}}$ converges to 0 as n tends to infinity. We write $\lim_{n \rightarrow \infty} \frac{1}{n} = 0$.

Proof. Here $a = 0$. Our proof is done if we can show that $\frac{1}{n} < \varepsilon$ for every $\varepsilon > 0$, when n is "large enough", i.e. equal to or larger than some threshold value N . To obtain this inequality, it would be sufficient to show that

$$\frac{1}{n} \leq \frac{1}{N} < \varepsilon.$$

Choosing $N > 1/\varepsilon$ completes the proof. \square

Example 3.4.3 Show that the sequence of numbers $\{(2 + n)/(5 + n)\}_{n \in \mathbb{N}}$ converges to 1 as $n \rightarrow \infty$.

Proof. Here $a = 1$. Our proof is done if we can show that $|\frac{2+n}{5+n} - 1| < \varepsilon$ for every $\varepsilon > 0$, when $n \geq N$ (i.e. n is "large enough"). Equivalently, we just need to show that $\frac{3}{5+n} < \varepsilon$.

To obtain this inequality, it would be sufficient to show that

$$\frac{3}{5+n} \leq \frac{3}{5+N} < \varepsilon.$$

This suggests that we choose $N > \frac{3}{\varepsilon} - 5$.

Remark. The preceding inequality can be further simplified as

$$\frac{3}{5+n} \leq \frac{3}{5+N} < \frac{3}{N} < \varepsilon,$$

so taking $N > \frac{3}{\varepsilon}$ will also work. The point is to show that we can find some N that is "large enough" for the deviation from a to be restricted by an arbitrary ε .

□

Some sequences of real numbers do not have a limit. Here is an example.

Example 3.4.4 The sequence of numbers $\{(-1)^n\}_{n \in \mathbb{N}}$ has no limit.

Proof. We prove this statement by establishing a contradiction. Suppose $\{(-1)^n\}_{n \in \mathbb{N}}$ has a limit which is a . Then for every ε , say 1, there exists some $n \geq N$ such that $|(-1)^n - a| < 1$. Now, if n is odd, we have $|-1 - a| < 1$, that is, $|1 + a| < 1$. If n is even, we have $|1 - a| < 1$. Since $2 = |1 + 1| = |1 - a + 1 + a|$, by the triangle inequality, we have $|1 - a + 1 + a| \leq |1 - a| + |1 + a| < 1 + 1 = 2$, which is an absurdity (i.e. $2 < 2$). Therefore our hypothesis that a exists must be false, implying that the contrary (i.e. a does not exist) is true.

□

Definition 3.4.2 (Subsequence of a sequence). A subsequence of a sequence $\{x_n\}_{n \in \mathbb{N}}$ is a sequence of the form $\{x_{n_k}\}_{k \in \mathbb{N}}$, where each $n_k \in \mathbb{N}$, and $n_1 < n_2 < \dots$ ♣

We can construct subsequences easily by deleting from $\{x_1, x_2, \dots\}$ all x_n except those for which $n = n_k$ for some k . For example, for the sequence $\{1/(2n)\}_{n \in \mathbb{N}}$ we can delete those x_n that cannot be expressed in the form of 2^k , $k = 1, 2, \dots$. This leads to the subsequence $\{1/2, 1/4, 1/8, \dots\}$.

Theorem 3.4.1. If $\{x_n\}_{n \in \mathbb{N}}$ converges to a and $\{x_{n_k}\}_{k \in \mathbb{N}}$ is any subsequence of $\{x_n\}_{n \in \mathbb{N}}$, then x_{n_k} converges to a as $k \rightarrow \infty$.

By Theorem 3.4.1, since the sequence $\{1/(2n)\}_{n \in \mathbb{N}}$ converges to 0, its subsequence $\{1/2^{n-1}\}_{n \in \mathbb{N}}$ converges to 0 as well.

We now define boundedness of sequence of real numbers.

Definition 3.4.3 (Boundedness of sequences). Let $\{x_n\}$ be a sequence of real numbers. We say that:

- (i) $\{x_n\}$ is bounded above if and only if there is an $M \in \mathbb{R}$ such that $x_n \leq M$ for all $n \in \mathbb{N}$.
- (ii) $\{x_n\}$ is bounded below if and only if there is an $m \in \mathbb{R}$ such that $x_n \geq m$ for all $n \in \mathbb{N}$.
- (iii) $\{x_n\}$ is bounded if and only if it is bounded above and bounded below. ♣

Theorem 3.4.2. Every convergent sequence is bounded.

Proof. Without loss of generality, we may suppose $\varepsilon = 1$. If a sequence $\{x_n\}$ converges to $a \in \mathbb{R}$, then there is some threshold $N \in \mathbb{N}$ such that $|x_n - a| < 1$ when $n \geq N$. Now, since $|x_n| = |x_n - a + a|$, applying the triangle inequality, we obtain $|x_n| \leq |x_n - a| + |a|$. Thus, $|x_n - a| < 1$

implies that $|x_n| < 1 + |a|$, that is, x_n is bounded (above by $1 + |a|$, below by $-1 - |a|$). For $1 \leq n \leq N$, clearly $|x_n| \leq \max\{|x_1|, |x_2|, \dots, |x_N|\} = M$. The sequence $\{x_n\}$ is dominated by $\max\{M, 1 + |a|\}$. \square

Figure 3.4.1 gives a specific example where Theorem 3.4.2 is applied.

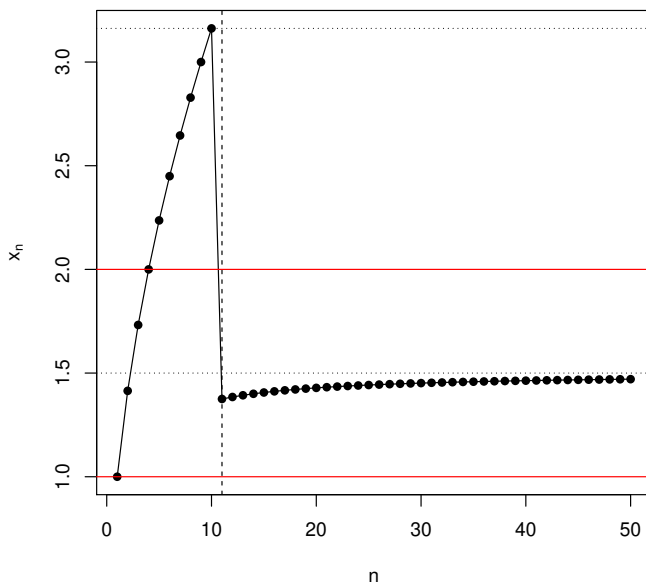


Figure 3.4.1: An example of a convergent sequence that is bounded. For $n \leq 10$, $x_n \leq \sqrt{10}$; for $n \geq 11$, with $\varepsilon = 0.5$, x_n is bounded above by 2 (upper red line). The whole sequence, which converges to 1.5, is therefore bounded by $\max\{\sqrt{10}, 2\} = \sqrt{10}$.

Definition 3.4.4 (Monotone Convergence theorem). If $\{x_n\}$ is increasing and bounded above, or if it is decreasing and bounded below, then $\{x_n\}$ has a finite limit. ♣

Example 3.4.5 If $a > 0$, then $a^{1/n} \rightarrow 1$ as $n \rightarrow \infty$.

Proof. The proof is shown by covering three possible cases of $a > 0$: (i) $a = 1$; (ii) $a > 1$; (iii) $0 < a < 1$.

Case (i): Clearly when $a = 1$, $a^{1/n} = 1$ for all $n \in \mathbb{N}$, and hence $a^{1/n} \rightarrow 1$ as $n \rightarrow \infty$.

Case (ii): For $a > 1$, $\{a^{1/n}\}$ is a decreasing sequence, since $\log a^{1/n} = (\log a)/n$, which decreases as n increases. It is also bounded below because $a^{1/n} > 0$. Therefore, applying the Monotone Convergence Theorem, $\lim_{n \rightarrow \infty} a^{1/n}$ exists. Now, we just need to show that for every $\varepsilon > 0$, $|a^{1/n} - 1| < \varepsilon$ for some $n \geq N$. Equivalently, we show that $|\frac{\log a}{n} - 0| < \varepsilon$. We set $\frac{\log a}{n} \leq \frac{\log a}{N} < \varepsilon$. Thus, $N > (\log a)/\varepsilon$.

Case (iii): If $0 < a < 1$, then $1/a > 1$. We have

$$\lim_{n \rightarrow \infty} a^{1/n} = \lim_{n \rightarrow \infty} \frac{1}{1/a^{1/n}} = \frac{1}{\lim_{n \rightarrow \infty} (1/a)^{1/n}} = 1,$$

using the result of Case (ii). □


3.5 LIMITS SUPREMUM AND INFIMUM

Definition 3.5.1. Let $\{x_n\}$ be a sequence of real numbers. The limit supremum of $\{x_n\}$ is the extended real number

$$\limsup_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} (\sup_{k \geq n} x_k),$$

and the limit infimum of $\{x_n\}$ is the extended real number

$$\liminf_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} (\inf_{k \geq n} x_k).$$

Remark: An extended real number x means either $x \in \mathbb{R}$, $x = \infty$, or $x = -\infty$. 

Example 3.5.1 For the sequence $\{(-1)^n\}$, $\sup_{k \geq n} (-1)^k = 1$, for all $n \in \mathbb{N}$, whereas $\inf_{k \geq n} (-1)^k = -1$ for all $n \in \mathbb{N}$. Therefore $\limsup_{n \rightarrow \infty} (-1)^n = 1$, and $\liminf_{n \rightarrow \infty} (-1)^n = -1$.

Theorem 3.5.1. Let $\{x_n\}$ be a sequence of real numbers and x be an extended real number. Then, $x_n \rightarrow x$ as $n \rightarrow \infty$ if and only if

$$\limsup_{n \rightarrow \infty} x_n = \liminf_{n \rightarrow \infty} x_n = x.$$

Theorem 3.5.1 implies that, if a sequence $\{x_n\}$ converges to some value x , then its limit supremum and limit infimum are finite and the same. Conversely, checking the latter condition can be used to establish sequence convergence. If $\{x_n\}$ diverges to $\pm\infty$, then $\limsup_{n \rightarrow \infty} x_n = \liminf_{n \rightarrow \infty} x_n = \pm\infty$.

Example 3.5.2 For the sequence $\{\frac{n}{n^2+1}\}$, $\sup_{k \geq n} \frac{k}{k^2+1} = \frac{n}{n^2+1}$, so that $\limsup_{n \rightarrow \infty} \frac{n}{n^2+1} = 0$; then, $\inf_{k \geq n} \frac{k}{k^2+1} = 0$, so that $\liminf_{n \rightarrow \infty} \frac{n}{n^2+1} = 0$. We conclude that $\{\frac{n}{n^2+1}\}$ converges to 0 by Theorem 3.5.1.

There is an alternative way to evaluate the limit supremum and limit infimum, which is sometimes more convenient than applying Definition 3.5.1.

Theorem 3.5.2. Let $\{x_n\}$ be a sequence of real numbers. The limit supremum and limit infimum of $\{x_n\}$ can be expressed as

$$\begin{aligned}\limsup_{n \rightarrow \infty} x_n &= \inf_{n \in \mathbb{N}} (\sup_{k \geq n} x_k), \\ \liminf_{n \rightarrow \infty} x_n &= \sup_{n \in \mathbb{N}} (\inf_{k \geq n} x_k).\end{aligned}$$

Example 3.5.3 For the sequence $\{n/(2n+3)\}$, we have $\inf_{k \geq n} k/(2k+3) = n/(2n+3)$, and $\sup_{k \geq n} k/(2k+3) = 1/2$. Therefore,

$$\begin{aligned}\limsup_{n \rightarrow \infty} x_n &= \inf_{n \in \mathbb{N}} (\sup_{k \geq n} x_k) = \inf_{n \in \mathbb{N}} \{1/2\} = 1/2 \\ \liminf_{n \rightarrow \infty} x_n &= \sup_{n \in \mathbb{N}} (\inf_{k \geq n} x_k) = \sup_{n \in \mathbb{N}} \{n/(2n+3)\} = 1/2\end{aligned}$$

We conclude that the sequence $\{n/(2n+3)\}$ converges to $1/2$.

3.6 CONVERGENCE OF A SEQUENCE OF FUNCTIONS

By now, we are more familiar with the idea of convergence of a sequence of numbers. The same concept can be applied to study the convergence of a sequence of functions, $\{f_n(x)\}_{n \in \mathbb{N}}$, where $x \in \mathbb{R}$. An example of a sequence of functions is the following:

$$\{(1 - x/n)^n\} = \{(1 - x), (1 - x/2)^2, (1 - x/3)^3, \dots\}$$

with the function defined over $x > 0$. If we graph this sequence of functions (Fig. 3.6.1), we suspect that it behaves more and more like some function that depends only on x as $n \rightarrow \infty$. Can you recognise what function this is?

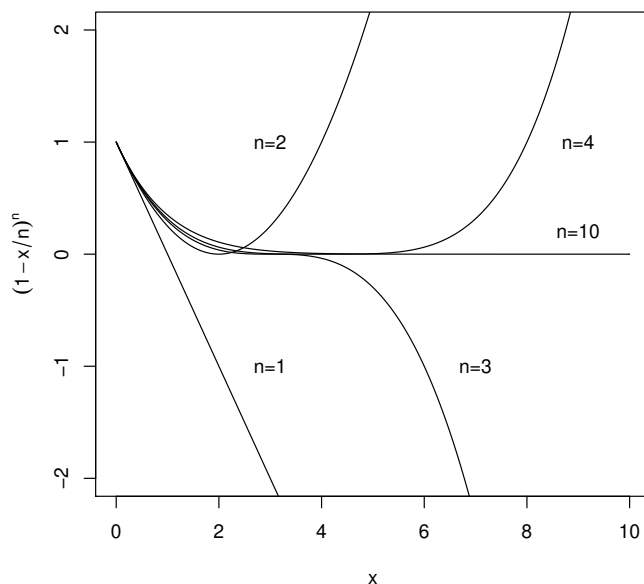


Figure 3.6.1: Sequence of functions $\{(1 - x/n)^n\}$, for $n = 1, 2, 3, 4, 10$, over $x > 0$.

When we discuss convergence of a sequence of functions, we are concerned about two main modes: pointwise convergence, and uniform convergence.

Definition 3.6.1 (Pointwise convergence of sequence of functions). Let E be a nonempty subset of \mathbb{R} . A sequence of functions $f_n : E \rightarrow \mathbb{R}$ is said to converge pointwise on E ($f_n \rightarrow f$ pointwise on E as $n \rightarrow \infty$) if and only if $f(x) = \lim_{n \rightarrow \infty} f_n(x)$ exists for each $x \in E$. ♣

Example 3.6.1 For the sequence of functions $\{(1 - x/n)^n\}$, where $x \geq 0$, we know that $\lim_{n \rightarrow \infty} (1 - x/n)^n = e^{-x}$ for all $x \in \mathbb{R}$ (hence $x \geq 0$) from elementary calculus. We can then say that $\{(1 - x/n)^n\}$ converges pointwise to e^{-x} over $x \geq 0$.

Example 3.6.2 For the sequence of functions $\{x^n\}$, where $0 \leq x \leq 1$, we can easily see that x^n converges pointwise to $f(x)$ on $[0, 1]$, where $f(x) = 0$, for $0 \leq x < 1$, and $f(x) = 1$ for $x = 1$. This example shows two important properties of pointwise convergence: (i) that $\lim_{n \rightarrow \infty} f_n(x) = f(x)$ depends on x ; (ii) that the property of continuity of $f(x)$ is not guaranteed to be preserved under operation of taking pointwise limit.

It turns out that the concept of uniform convergence has more desirable mathematical properties that makes it useful. The main difference between pointwise convergence and uniform convergence is that the threshold integer N cannot depend on x .

Definition 3.6.2 (Uniform convergence of sequence of functions). Let E be a nonempty subset of \mathbb{R} . A sequence of functions $f_n : E \rightarrow \mathbb{R}$ is said to converge uniformly on E ($f_n \rightarrow f$ uniformly on E as $n \rightarrow \infty$) if and only if for every $\varepsilon > 0$ there is an $N \in \mathbb{N}$ such that $n \geq N$ implies that $|f_n(x) - f(x)| < \varepsilon$, for all $x \in E$. ♣

Example 3.6.3 Prove that $\{1/(n(1 + x^2))\}$ converges uniformly to 0 for $x \in \mathbb{R}$.

Proof. We need to show that there is some N such that

$$\left| \frac{1}{n(1 + x^2)} - 0 \right| < \varepsilon,$$

when $n \geq N$. In our case this is straightforward:

$$\left| \frac{1}{n(1+x^2)} - 0 \right| = \frac{1}{n(1+x^2)} \leq \frac{1}{n} \leq \frac{1}{N} < \varepsilon.$$

Taking $N > 1/\varepsilon$ completes the proof. \square

Example 3.6.4 Prove that $\{(\sin nx)/n\}$ converges uniformly to 0 for $x \in \mathbb{R}$.


Proof. We need to show that there is some N such that

$$\left| \frac{\sin nx}{n} - 0 \right| < \varepsilon,$$

when $n \geq N$. Since $|\sin nx| \leq 1$, we have

$$\left| \frac{\sin nx}{n} \right| \leq \frac{1}{n} \leq \frac{1}{N} < \varepsilon$$

Taking $N > 1/\varepsilon$ completes the proof. \square

For illustration, see Fig. 3.6.2. 

Example 3.6.5 Let $\{f_n(x) = x/(x^2 + n^2)\}$, $x \geq 0$ be a sequence of functions. We claim that f_n converges uniformly to 0. To prove this, we need to show that, for all $\varepsilon > 0$, there is some N such that when $n \geq N$, then

$$\left| \frac{x}{x^2 + n^2} - 0 \right| < \varepsilon.$$

To show this, we first note that $f_n(x)$ is bounded above by $\sup_{x \geq 0} |f_n(x)|$. Since $f_n(0) = 0$, $f_n(a) > 0$ for $a \geq 0$ and $\lim_{n \rightarrow \infty} f_n(x) = 0$, there exists a supremum for $f_n(x)$. We can find it using differential calculus:

$$\frac{d}{dx} f_n(x) = \frac{(x^2 + n^2) \cdot 1 - x(2x)}{(x^2 + n^2)^2} = \frac{n^2 - x^2}{(x^2 + n^2)^2} = 0,$$

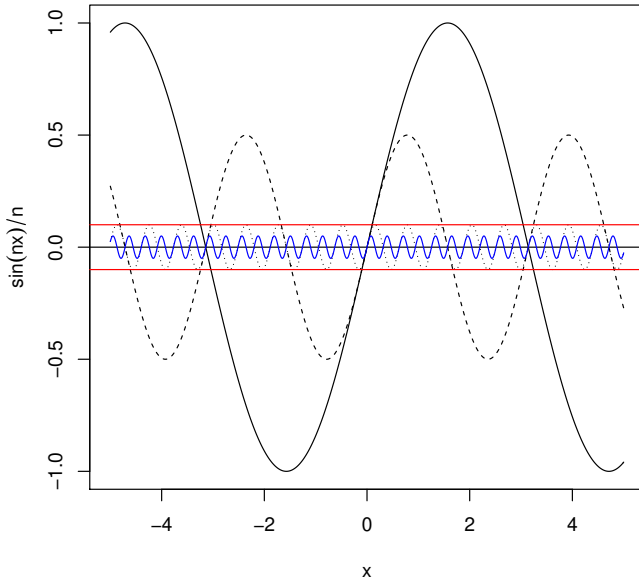


Figure 3.6.2: Visualisation of uniform convergence of sequence of $\sin(nx)/n$ for $n = 1, 2, 10, 20$ (blue). Here $\varepsilon = 0.1$ (red lines), and we see that for $|\sin nx/n| < 0.1$, we have to choose $N > 1/0.1 = 10$. Let's say we choose $N = 11$. Then for $n = 20 > 11$ (blue curve), it is clear that $\sin(20x)/20$ deviates from 0 by an amount below $\varepsilon = 0.1$. In fact, starting from $n \geq 11$, all f_n will deviate from 0 by an amount less than $\varepsilon = 0.1$.

which occurs at $x = n$. Therefore

$$\sup_{x \geq 0} |f_n(x)| = \frac{n}{n^2 + n^2} = \frac{1}{2n},$$

which leads to

$$\left| \frac{x}{x^2 + n^2} - 0 \right| \leq \frac{1}{2n} \leq \frac{1}{2N} < \varepsilon.$$

Choosing $N > 1/(2\varepsilon)$ completes the proof.

Example 3.6.6 The sequence of functions $\{f_n(x) = x/(x + n)\}$ converges pointwise to 0, for $x \geq 0$, since $\lim_{n \rightarrow \infty} x/(x + n) = 0$. However, it fails to converge uniformly to 0. To understand why, we first note that $|f_n(x)|$ is bounded above by $\sup_{x \geq 0} |f_n|$. Since $0 \leq x/(x + n) < 1$, clearly $\sup_{x \geq 0} |f_n| = 1$. Therefore,

$$\left| \frac{x}{x + n} - 0 \right| < 1,$$

that is, it can only be bounded for $\varepsilon \geq 1$, but not $0 < \varepsilon < 1$.

A uniformly convergent sequence of functions has several properties that it useful for analysis. The first one involves integrability of $f(x)$.

Theorem 3.6.1 (Uniform convergence of sequence of Riemann - integrable functions $f_n(x)$ implies integrability of $f(x)$). Suppose $f_n \rightarrow f$ uniformly on a closed interval $[a, b]$. If each $f_n(x)$ is Riemann-integrable on $[a, b]$, then so is $f(x)$, and we have

$$\lim_{n \rightarrow \infty} \int_a^b f_n(x) dx = \int_a^b [\lim_{n \rightarrow \infty} f_n(x)] dx.$$

Theorem 3.6.1 makes it legitimate to perform interchange of limit and integration operations, which is a useful re-

sult that we will use when dealing with convergence of a distribution function.

Example 3.6.7 Since $f_n(x) = x/(x^2 + n^2)$ converges uniformly to 0 over $x \geq 0$, by Theorem 3.6.1,

$$\lim_{n \rightarrow \infty} \int_0^1 \frac{x}{x^2 + n^2} dx = \int_0^1 \lim_{n \rightarrow \infty} [x^2/(x^2 + n^2)] dx = 0.$$

Remark: If we have no concept of uniform convergence, we would first have to evaluate the integrate on the left-hand side:

$$\int_0^1 \frac{x}{x^2 + n^2} dx = \frac{1}{2} \log(x^2 + n^2) \Big|_0^1 = \frac{1}{2} \log \left(1 + \frac{1}{n^2} \right),$$

and take limit accordingly. In general, the approach of evaluating integrals first and then taking limits should only be tried if the sequence of functions considered cannot be shown to be uniformly convergent, since the integrals may be very hard to evaluate, or fail to yield any closed form.

Example 3.6.8 Referring to Example 3.6.6, $f_n(x) = x/(x + n)$ is not uniformly convergence over $x \geq 0$, so we cannot justify the interchange of limit and integration operation by Theorem 3.6.1 to evaluate

$$\lim_{n \rightarrow \infty} \int_0^1 \frac{x}{x + n} dx.$$

Doing it the standard way by evaluating the integral first, we get

$$\int_0^1 \frac{x}{x + n} dx = x - n \log(n + x) \Big|_0^1 = 1 + n \log n - n \log(n + 1).$$

Then, taking limit and using L'Hôpital's Rule,

$$\begin{aligned}\lim_{n \rightarrow \infty} \int_0^1 \frac{x}{x+n} dx &= 1 + \lim_{n \rightarrow \infty} \left[\frac{\log\left(\frac{n}{n+1}\right)}{1/n} \right] \\ &= 1 + \lim_{n \rightarrow \infty} \left(\frac{1/n - 1/(n+1)}{-1/n^2} \right) = 0.\end{aligned}$$

Remark: If we ignored the condition of uniform convergence and went ahead with interchange of limit and integration operation anyway, we would obtain the correct answer of 0 as well, but that is just being lucky.

Besides the interesting property shown in Theorem 3.6.1, here are other properties of uniformly convergence sequence of functions.

Theorem 3.6.2 (Other properties of uniformly convergent sequence of functions). Let $\{f_n(x)\}$ be a sequence of functions over $x \in [a, b]$, $a, b \in \mathbb{R}$. If $f_n(x)$ converges uniformly to $f(x)$, then

- (i) $f_n(x)$ converges pointwise to $f(x)$.
- (ii) $f(x)$ is also continuous, if $f_n(x)$ is continuous for each $n \in \mathbb{N}$.
- (iii) $f(x)$ is also bounded on the interval $[a, b]$, if $f_n(x)$ is bounded on the same interval.

THE PROBABILITY MEASURE SPACE

Suppose we have a chance mechanism that produces certain outcomes. The set consisting of all such possible outcomes is called the sample space, generally denoted as Ω .

The general idea of a probability measure space consists of defining a space consisting of collections of subsets of Ω , and assigning a measure between 0 and 1 to such subsets.

Example 4.0.1 Suppose we have a discrete space $\Omega = \{1, 2, 3\}$. One possible collection of subsets of Ω (generally denoted using script font, e.g. \mathcal{A}) is the collection of singletons: $\{\{1\}, \{2\}, \{3\}\}$. We may also be interested in a collection of subsets consisting of pairs of numbers: $\{\{1, 2\}, \{2, 3\}\}$. The smallest possible collection of subsets of Ω is the empty set $\{\emptyset\}$, while the largest possible collection of subsets of Ω is the power set

$$\{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \Omega\}.$$

In order make sure that set operations performed on elements in \mathcal{A} produce sets that stay within some well-defined space, we need the notion of a *field*.

Definition 4.0.1 (Field). A collection of subsets of Ω , denoted \mathcal{F} , is called a field if it contains Ω , and is closed under the formation of complements and finite unions, that is:

- (i) $\Omega \in \mathcal{F}$;
- (ii) $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$;
- (iii) $A, B \in \mathcal{F}$ implies $A \cup B \in \mathcal{F}$.



Definition 4.0.1 gives rise to additional elementary facts as follows.

Theorem 4.0.1. If \mathcal{F} is a field, then

- (a) $A, B \in \mathcal{F}$ implies that $A \setminus B \in \mathcal{F}$ (the backslash symbol denotes set difference).
- (b) $A, B \in \mathcal{F}$ implies that $A \triangle B \in \mathcal{F}$ (the triangle set symbol denotes symmetric difference).
- (c) $A_1, A_2, \dots, A_n \in \mathcal{F}$ implies that $\bigcup_{i=1}^n A_i \in \mathcal{F}$.

Proof. (a) $A \setminus B = A \cap B^c \in \mathcal{F}$ property (iii).

(b) $A \triangle B = (A \setminus B) \cup (B \setminus A) \in \mathcal{F}$ by part (a) and property (iii).

(c) Proof by induction: Assume to be true for $n = k$:

$A_1, A_2, \dots, A_k \in \mathcal{F}$ implies that $\bigcup_{i=1}^k A_i \in \mathcal{F}$. For $k = 2$, this is just property (iii). For the case of $n = k + 1$, $A_1, A_2, \dots, A_k, A_{k+1} \in \mathcal{F}$, then $\bigcup_{i=1}^{k+1} A_i = \bigcup_{i=1}^k A_i \cup A_{k+1}$. Since $\bigcup_{i=1}^k A_i \in \mathcal{F}$ by hypothesis, and $A_{k+1} \in \mathcal{F}$, it follows from property (iii) that $\bigcup_{i=1}^{k+1} A_i \in \mathcal{F}$. \square

Property (iii) of Definition 4.0.1 implies closure under finite intersections as follows. Let $A, B \in \mathcal{F}$. Then, $A^c, B^c \in \mathcal{F}$ (by property (ii)), and using De Morgan's law and property (iii), $A^c \cup B^c = (A \cap B)^c \in \mathcal{F}$. Then, applying property (ii) again, we have $A \cap B \in \mathcal{F}$. Thus, property (iii), that of

closure under finite union, is equivalent to closure under finite intersection, that is,

$$(iii') A, B \in \mathcal{F} \quad \text{implies} \quad A \cap B \in \mathcal{F}.$$

Example 4.0.2 Let $\Omega = \{0, 1, 2\}$. The collection of subsets $\mathcal{A} = \{\{0\}, \{1, 2\}\}$ is not a field, because property (i) is not satisfied. If we add Ω and \emptyset to \mathcal{A} , then it is a field. Trivially, \emptyset is the complement of Ω , and the complement of $\{0\}$ is $\{1, 2\}$ which falls in \mathcal{A} , and vice versa (property (ii)). Also, $\{0\} \cup \{1, 2\} = \Omega \in \mathcal{A}$, so property (iii) is satisfied. We do not need to check property (iii'), since this is already implied by property (iii). Nevertheless, we can check it anyway - clearly, $\{0\} \cap \{1, 2\} = \emptyset \in \mathcal{A}$.

To be useful for the study of probability, the notion of a field is extended to a σ -field (sigma-field) as follows:

Definition 4.0.2 (σ -field). The field defined in Definition 4.0.1 is called a σ -field if property (iii) is extended to countable unions:

$$(iv) A_1, A_2, \dots \in \mathcal{F} \text{ implies } \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

Property (iv) is equivalent to closure under countable intersection by application of the infinite form of De Morgan's law:

$$(iv') A_1, A_2, \dots \in \mathcal{F} \text{ implies } \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}. \quad \clubsuit$$

More generally, Ω can be an interval. We can still define a field as shown in the next example.

Example 4.0.3 Consider finite disjoint subintervals of $\Omega = (0, 1]$, that is, $(a_i, b_i]$, $i = 1, 2, \dots, n$, where $0 < a_i < b_i \leq 1$. We can form a field \mathcal{B}_0 by collecting such disjoint intervals and augmenting the collection with \emptyset . Here is an example for you to verify:

$$\mathcal{B} = \left\{ (\emptyset, 0, \frac{1}{4}], (\frac{1}{4}, \frac{3}{4}], (\frac{3}{4}, 1], (0, \frac{3}{4}], (\frac{1}{4}, 1], (0, \frac{1}{4}] \cup (\frac{3}{4}, 1], \Omega \right\}.$$

Example 4.0.4 More generally, let us check that the collection of finite disjoint unions of subintervals of $\Omega = (0, 1]$ that is augmented by the empty set is a field, \mathcal{B}_0 . Define $A = (a_1, a'_1] \cup (a_2, a'_2] \cdots \cup (a_m, a'_m] = \bigcup_{i=1}^m (a_i, a'_i]$, where $a_1 \leq a_2 \leq \cdots \leq a_m$. If $(a_i, a'_i]$ are disjoint, then $A^c = (0, a_1] \cup (a'_1, a_2] \cup \cdots \cup (a'_{m-1}, a_m] \cup (a_m, 1]$, which also lies in \mathcal{B}_0 . If $B = (b_1, b'_1] \cup (b_2, b'_2] \cdots \cup (b_n, b'_n] = \bigcup_{j=1}^n (b_j, b'_j]$ with $(b_i, b'_i]$ again disjoint, then

$$A \cap B = \bigcup_{i=1}^m \bigcup_{j=1}^n \{(a_i, a'_i] \cap (b_j, b'_j]\}.$$

Each intersection is again an interval or else the empty set, and the union is disjoint. Therefore $A \cap B \in \mathcal{B}_0$. In this way we verify that \mathcal{B}_0 satisfies properties (i), (ii) and (iii'). In this example, it is much easier to check property (iii') than (iii). \mathcal{B}_0 is not a σ -field because it does not contain the singletons $\{x\}$, which arise as countable intersection $\bigcap_{n=1}^{\infty} (x - 1/n, x]$ of elements of \mathcal{B}_0 .

It is possible that for a particular collection of subsets \mathcal{A} , application of finite and countable operations on elements of \mathcal{A} can produce sets that are outside of \mathcal{A} . To ensure that we work with a σ -field that contains \mathcal{A} and is the smallest among all possible σ -fields, we have the notion of a σ -field generated by \mathcal{A} .

Definition 4.0.3 (σ -field generated by \mathcal{A}). Let \mathcal{A} be a collection of subsets of Ω . The intersection of all σ -fields containing \mathcal{A} is called the σ -field generated by \mathcal{A} , and written as $\sigma(\mathcal{A})$. It has three properties:

- (i) $\mathcal{A} \subseteq \sigma(\mathcal{A})$;
- (ii) $\sigma(\mathcal{A})$ is a σ -field;
- (iii) if $\mathcal{A} \subseteq \mathcal{G}$ and \mathcal{G} is a σ -field, then $\sigma(\mathcal{A}) \subseteq \mathcal{G}$.



Example 4.0.5 Let $\Omega = \{1, 2, 3, 4, 5, \dots\}$. Consider the collection of subsets $\mathcal{A} = \{\emptyset, \{1\}, \Omega\}$. We see that \mathcal{A} is not a σ -field because it is not closed under complementation with respect to $\{1\}$. However, $\sigma(\mathcal{A}) = \{\emptyset, \{1\}, \{2, 3, 4, \dots\}, \Omega\}$ is definitely a σ -field as it is closed under complementation and countable union. We can check that property (i) and (ii) in Definition 4.0.3 are indeed true. Now, suppose we have a σ -field:

$$\begin{aligned} \mathcal{G} = \{ & \emptyset, \{1\}, \{2, 3, 4, \dots\}, \{1, 2, 3\}, \{4, 5, 6, \dots\}, \\ & \{1, 4, 5, 6, \dots\}, \{2, 3\}, \Omega \}, \end{aligned}$$

in which $\mathcal{A} \subset \mathcal{G}$, clearly. Then, $\sigma(\mathcal{A}) \subset \mathcal{G}$ verifies property (iii).

4.1 THE PROBABILITY TRIPLE

Definition 4.1.1 (Probability measures). A set function is a real-valued function defined on a collection of subsets of Ω . We call a set function P on a field \mathcal{F} a probability measure if the following three conditions are satisfied:

- (i) $A \in \mathcal{F}$ implies that $0 \leq P(A) \leq 1$;
- (ii) $P(\emptyset) = 0$, and $P(\Omega) = 1$;
- (iii) [countable additivity] If A_1, A_2, \dots is a disjoint sequence of sets in \mathcal{F} , and if $\bigcup_{k=1}^{\infty} A_k \in \mathcal{F}$, then

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k).$$



Note that in property (iii) of Definition 4.1.1, we assume that $\bigcup_{k=1}^{\infty} A_k \in \mathcal{F}$ since \mathcal{F} is just a field, and not a σ -field, so closure under countable union is not guaranteed. It turns

out that this assumption can be removed because a probability measure on a field has a unique extension to the generated σ -field (The Carathéodory extension theorem). The proof of this theorem is rather challenging and is omitted.

From Definition 4.1.1, we obtain the following results:

Theorem 4.1.1. Consider the probability triple (Ω, \mathcal{F}, P) .

(i) [monotonicity of P] If $A, B \in \mathcal{F}$ and $A \subseteq B$, then $P(A) \leq P(B)$.

(ii) [inclusion-exclusion formula for $n = 2$]: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

(iii) [inclusion-exclusion formula: general form]:

If $A_1, A_2, \dots, A_n \in \mathcal{F}$, then

$$\begin{aligned} P\left(\bigcup_{k=1}^n A_k\right) &= \sum_{i=1}^n P(A_i) - \sum_{i < j} P(A_i \cap A_j) \\ &\quad + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) + \dots \\ &\quad + (-1)^{n+1} P(A_1 \cap A_2 \dots \cap A_n). \end{aligned}$$

(iv) [Boole's inequality (finite subadditivity of P)]:

$$P\left(\bigcup_{k=1}^n A_k\right) \leq \sum_{k=1}^n P(A_k).$$

Proof. (i) $B \setminus A$ and A are disjoint, therefore $P[(B \setminus A) \cup A] = P(B \setminus A) + P(A) = P(B)$, implying that $P(A) \leq P(B)$.

(ii) $A \cup B = A \cap B^c \cup B$. Therefore, $P(A \cup B) = P(A \cap B^c) + P(B)$. But $A = A \cap B^c \cup A \cap B$, so $P(A) = P(A \cap B^c) + P(A \cap B)$. Rearranging and substituting $P(A \cap B^c)$ into the preceding equation completes the proof.

(iii) Proof by induction: For $n = 2$, it is just the result in (ii). Assume to be true for $k = n$. Then for $k = n + 1$,

$$P\left(\bigcup_{k=1}^{n+1} A_k\right) = P\left(\bigcup_{k=1}^n A_k\right) + P(A_{n+1}) - P\left(\bigcup_{k=1}^n A_k \cap A_{n+1}\right)$$

Applying the inclusion-exclusion formula for $n = k$ to the first term and the third term completes the proof (work out the details).

(iv) Let $B_1 = A_1$, $B_k = A_k \cap A_1^c \cap \cdots \cap A_{k-1}^c$. Then, the B_k are disjoint, so $\bigcup_{k=1}^n A_k = \bigcup_{k=1}^n B_k$, hence $P(\bigcup_{k=1}^n A_k) = \sum_{k=1}^n P(B_k)$. By monotonicity of P , $P(B_k) \leq P(A_k)$, hence $P(\bigcup_{k=1}^n A_k) \leq \sum_{k=1}^n P(A_k)$.

□

Countable additivity implies countable subadditivity (see finite subadditivity in Theorem 4.1.1), that is, if A_1, A_2, \dots is a sequence of sets (not necessarily disjoint) in \mathcal{F} , and if $\bigcup_{k=1}^{\infty} A_k \in \mathcal{F}$, then

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} P(A_k).$$

In order to talk meaningfully about the probability of an "event", we must first define an appropriate probability space that consists of the triple (Ω, \mathcal{F}, P) . In fact, if the "event" that we are interested in is not in \mathcal{F} , we cannot assign any probability measure to it, since by definition, P is a set function that operates on elements of \mathcal{F} only. In fact, we will use the term "event" to refer to only elements

in \mathcal{F} that can be assigned a value through the probability measure P .

Example 4.1.1 Suppose we toss a fair coin two times. Then, $\Omega = \{HH, HT, TH, TT\}$. The power set is a σ -field:

$$\begin{aligned}\mathcal{F} = \{ & \emptyset, \\ & \{HH\}, \{HT\}, \{TH\}, \{TT\}, \\ & \{HH, HT\}, \{HH, TH\}, \dots, \{TH, TT\}, \\ & \{HH, HT, TH\}, \{HH, HT, TT\}, \dots, \{HT, TH, TT\}, \\ & \Omega\}\end{aligned}$$

Then, for the probability triple (Ω, \mathcal{F}, P) , the elements of \mathcal{F} are all events because we can always assign a probability measure to each of them through the set function P . For example, one possibility is $P(\{HH\}) = 1/4, P(\{HT\}) = 1/4, P(\{TH\}) = 1/4, P(\{TT\}) = 1/4$. Thus, $P(\{HH, HT\}) = P(\{HH\}) + P(\{HT\}) = 1/2$, and we can consider the probability measures of events such as $\{HH, HT\} \cup \{HT, TT\} = \{HH, HT, TT\} \in \mathcal{F}$. When we ask what is the probability of getting one head in two tosses, this then translates to $P(\{HT, TH\}) = P(\{HT\}) + P(\{TH\}) = 1/2$.

Example 4.1.2 For Example 4.1.1, we can consider another σ -field:

$$\mathcal{F}^* = \{\emptyset, \{HH\}, \{HT, TH, TT\}, \Omega\}$$

Then, for the probability triple $(\Omega, \mathcal{F}^*, P)$, we can assign probability measures to $\{HH\}$, but not to $\{HT\}, \{TH\}, \{TT\}$, since the latter are not in \mathcal{F}^* . We can also just assign a probability measure to $\{HT, TH, TT\}$ by complementation, but nothing can be said about $P(\{HT\}), P(\{TH\}), P(\{TT\})$. For this probability triple, it is meaningless to ask about the

probability of getting one tail, or the probability of getting two tails.

The preceding example shows that we have to choose our \mathcal{F} carefully so that we can assign probability measures to sets that we are interested in. Fortunately, for a finite discrete space, the power set will include all possible sets of interest.

Example 4.1.3 One may casually ask the following question that involves probability:

What is the probability that I see the sun rising from the east tomorrow, if I remain at this spot?

Our sample space Ω consists of all the possible directions. We can construct $\mathcal{F} = \{\emptyset, \text{East}, \text{Not east}, \Omega\}$. Then, on the probability triple (Ω, \mathcal{F}, P) , it would be meaningful to talk about $P(\text{East})$, and certainly you could assign any value between 0 and 1 (inclusive of the end points) which is a reflection of your degree of belief that the sun is going to rise from the east tomorrow. However, for this probability triple, it would be meaningless to talk about the probability that the sun will rise from the west (It is certainly not $1 - P(\text{East})$). There are of course many other σ -fields much larger than \mathcal{F} in the current example that will contain "East", but if all you want is to answer that particular question, then constructing \mathcal{F} as in the preceding example is sufficient.

4.2 THE LEBESGUE MEASURE ON THE UNIT INTERVAL

Consider \mathcal{I} , the collection of subintervals $(a, b]$ of $(0, 1]$. For some element $I = (a, b] \in \mathcal{I}$, define the length of I

as $\lambda(I) = b - a$, and $\lambda(\emptyset) = 0$. If $A = \cup_{i=1}^n I_i$, and I_i are disjoint, then

$$\lambda(A) = \sum_{i=1}^n \lambda(I_i) = \sum_{i=1}^n (b_i - a_i).$$

The function λ thus defines a set function on the Borel field \mathcal{B}_0 . We call λ the Lebesgue (pronounced as "Luh-Beg") measure.

It turns out that λ can be extended from \mathcal{B}_0 to the Borel σ -field $\mathcal{B} = \sigma(\mathcal{B}_0)$, which contains the singletons (recall that these arise from countable intersections of intervals of $(x - 1/n, x]$). With this extension, we can then find the Lebesgue measure of a singleton. It can be shown that $\lambda(\{x\}) = 0$ (Tutorial 3).

Example 4.2.1 Consider $A = [a, b] \in \mathcal{B}$. We can express it as the union of two disjoint sets $A = \{a\} \cup (a, b]$. Therefore $\lambda(A) = \lambda(\{a\}) + \lambda((a, b]) = b - a$, since $\lambda(\{a\}) = 0$. Similarly, we can show that $\lambda((a, b]) = 0$. Thus, the Lebesgue measure of open, closed, or half-open intervals is just the length of that interval.

Example 4.2.2 (Lebesgue measure of a countable set is 0) Consider $\mathbb{Q}_{(0,1)}$, the set of rational numbers in $(0, 1)$. This set is infinitely large, but countable (by the diagonal method). Since \mathbb{Q} can be expressed as a countable union of singletons, by countable additivity, it follows that $\lambda(\mathbb{Q}) = \sum_{i=1}^{\infty} \lambda(\{q_i\}) = 0$, where q_i is the sequence of rational numbers in $[0, 1]$. In the language of probability measure, if we randomly pick a value in $[0, 1]$, the probability that it is a rational number (i.e. being an element in $\mathbb{Q}_{[0,1]}$) is 0.

It may come as a surprise that there are uncountable sets that have Lebesgue measure of 0. The most famous example is the Cantor set C .

Example 4.2.3 (The Cantor set) We can construct the Cantor set as follows. Start with the unit interval $[0, 1]$. Then, remove the middle third $(1/3, 2/3)$. For the remaining two intervals: $[0, 1/3]$ and $[2/3, 1]$, again remove the middle third $(1/9, 2/9)$ for $[0, 1/3]$; $(7/9, 8/9)$ for $[2/3, 1]$. Continue doing so inductively. Then, at the n th stage, we would have removed the 2^{n-1} middle thirds of all remaining sub-intervals, each of length $1/3^n$. What is left over after such removal is the Cantor set C .

The complement of the Cantor set is a countable union of open intervals, hence it is a Borel set (i.e. it belongs to the Borel σ -field, \mathcal{B}). Hence, C is also a Borel set - it contains the set of singletons $\{1/3, 2/3, 1/9, 2/9, 7/9, 8/9, \dots\}$ as well as some other intervals. This makes C uncountable.

Since the complement of the Cantor set has Lebesgue measure equal to $\sum_{n=1}^{\infty} 2^{n-1}/3^n = 1$, it follows that $\lambda(C) = 0$.

LEBESGUE INTEGRATION

In elementary calculus, we learn that the indefinite integral of some regular function $f(x)$ produces another function $g(x)$, that, when differentiated, allows us to recover $f(x)$. For example $\int x dx = x^2/2 + C$, and the derivative of $x^2/2 + C$ produces x , which is the integrand. Later, we learn to view the integral geometrically, as the area under the curve (or hypervolumes, for multiple variables) obtained by approximating the area under curve using infinitely many rectangular bars with infinitesimal width. This is the Riemann integral.

The Riemann integral is useful when $f(x)$ is regular. However, non-regular $f(x)$ may not be Riemann-integrable. The simplest example is the function $f(x)$ over $(0, 1)$ that takes the value 1 if x is irrational, and 0 if x is rational. This function is not Riemann-integrable over $(0, 1)$, because there is no clear way of making partitions with infinitesimal widths along $(0, 1)$.

The Lebesgue integral solves the problem of integrating such non-regular functions. Moreover, where $f(x)$ is regular, it is equal to the Riemann integral. Thus, the Lebesgue integral is an appropriate framework for working with integrals involving probability measures.

Definition 5.0.1 (Lebesgue integration). The Lebesgue integral of a non-negative function f on a measure space $(\Omega, \mathcal{F}, \mu)$ is defined as

$$\int_{\Omega} f d\mu = \sup \sum_i \{ \inf_{\omega \in A_i} f(\omega) \} \mu(A_i).$$

For general f , write f^+ and f^- as the positive and the negative parts. Thus, $f = f^+ - f^-$, and

$$\int_{\Omega} f d\mu = \int_{\Omega} f^+ d\mu - \int_{\Omega} f^- d\mu.$$



Geometrically, the Lebesgue integral involves partitioning the domain of integration to sets A_1, A_2, \dots, A_n using n horizontal cut-offs on f . The quantity $\inf_{\omega \in A_i} f(\omega)$ is the height of a rectangular bar, while $\mu(A_i)$, the measure of set A_i , is the width of the rectangular bar. Thus, the product $\{ \inf_{\omega \in A_i} f(\omega) \} \mu(A_i)$ is just the area of the i th rectangular bar. Summing over a set of such rectangular bars gives an approximation to the area under the curve (always an underestimation, since \inf is used). The Lebesgue integral is then the supremum of the sum of these rectangular bars.

Technically, for general f , Definition 5.0.1 does not say much about how the value of the integral can be operationally obtained. However, if f is a step function (piecewise constant), then the Lebesgue integral is just the sum of area of rectangular bars. For regular f , the Lebesgue integral is equal to the Riemann integral, so one may switch to evaluating the Lebesgue integral of a regular function using the Riemann integral.

Example 5.0.1 Define $(\Omega, \mathcal{F}, \mu)$, where μ is a Lebesgue measure. Let $A = \{ \omega : \omega \in (0, 1] \setminus (0, 1] \cap \mathbb{Q} \}$. Denote

$\mathbb{1}_A$ the indicator function that takes value 1, when ω is an irrational number in $(0, 1)$, and 0 if it is a rational number in $(0, 1)$. Since $(0, 1] = (0, 1] \setminus (0, 1] \cap \mathbb{Q} \cup (0, 1] \cap \mathbb{Q}$ We have

$$\int_{(0,1]} \mathbb{1}_A d\mu = \int_{(0,1] \setminus (0,1] \cap \mathbb{Q}} 1 d\mu + \int_{(0,1] \cap \mathbb{Q}} 0 d\mu.$$

Clearly the second integral is 0, since the integrand is 0 and the Lebesgue measure of the set of rational numbers in $(0, 1)$ is also 0. For the first integral, we note that

$$\int_{(0,1]} 1 d\mu = \int_{(0,1] \setminus (0,1] \cap \mathbb{Q}} 1 d\mu + \int_{(0,1] \cap \mathbb{Q}} 1 d\mu.$$

Since the Lebesgue measure of $(0, 1]$ is 1, and $1 \times 1 = 1$, the left-hand side is equal to 1. The second integral is 0 since the Lebesgue measure of the set of rational numbers in $(0, 1)$ is 0. This means the first integral integrates to 1.

Example 5.0.2 Suppose $\Omega = (0, \infty)$, and let μ define the Lebesgue measure on $(0, \infty)$. Let $f(x) = 2/\{\pi(1+x^2)\}$. Consider \mathbb{Q} , the set of rational numbers, which is infinitely large. Perhaps $\int_{\mathbb{Q}} f d\mu$ integrates to some value? To find out, First, we note that $(0, \infty) = (0, \infty) \setminus \mathbb{Q} \cup \mathbb{Q}$. Therefore

$$\int_{(0,\infty)} f d\mu = \int_{(0,\infty) \setminus \mathbb{Q}} f d\mu + \int_{\mathbb{Q}} f d\mu.$$

Since \mathbb{Q} is countable, it has Lebesgue measure 0. Therefore the second integral on the right-hand side is 0. This means that integrals of continuous functions over countable sets (or finite sets) is 0. The Lebesgue integral on the left-hand side can be evaluated using the Riemann integral, since f is regular. It is equal to 1. This means the first integral on the right-hand side is 1.

Example 5.0.3 There exists some regular functions that are not Lebesgue-integrable. Here is an example. The function $f(x) = (\sin x)/x$ is Riemann-integrable over $(0, \infty)$:

$$\int_0^\infty \frac{\sin x}{x} dx = \frac{\pi}{2}.$$

However, f is not Lebesgue-integrable because each of its positive part and negative part integrates to ∞ . It is sufficient to consider the positive part. Now, $f^+ = (\sin x)/x$ over the intervals $(0, \pi), (2\pi, 3\pi), (3\pi, 4\pi), \dots$, and $f^+ = 0$ elsewhere. Disregarding the first interval, The midpoints of these intervals occur at $5\pi/2, 9\pi/2, \dots$. Thus, the k th ($k = 1, 2, \dots$) interval has midpoint at $[(2k+1) + 2k]\pi/2 = (4k+1)\pi/2$. We can form triangles centered at these midpoints for these intervals. Such triangles have height equal to $\sin((4k+1)\pi/2)/[(4k+1)\pi/2] = 2/[(4k+1)\pi]$. Hence, they have area equal to $1/(4k+1)\pi$. Now, since

$$\int_{(0, \infty)} f^+ d\mu > \sum_{k=1}^{\infty} \frac{1}{(4k+1)\pi},$$

it follows that the integral does not converge, since the infinite sum on the right-hand side is divergent (p -series with $p = 1$).

RANDOM VARIABLES AND THEIR EXPECTATIONS

We call a variable that can take particular values with certain probability a random variable in elementary probability. For example, we conduct an experiment where we count the number of tosses of a fair coin until we get a head. This quantity is then a variable, because it can take values $1, 2, 3, \dots$, each with a different probability. If we denote this quantity by X , then the first few probabilities can be written as $P(X = 1) = 1/2$, $P(X = 2) = 1/4$, $P(X = 3) = 1/8$, etc. Since probabilities are involved, it makes sense to call X a “random variable”. However, as we have studied the probability triple (Ω, \mathcal{F}, P) , we know that only sets in \mathcal{F} can be assigned probabilities. Hence, a statement like “ $X = 1$ ” must be referring to a set in \mathcal{F} .

6.1 RANDOM VARIABLE AS A SET FUNCTION

Definition 6.1.1 (Random variable). Given a probability triple (Ω, \mathcal{F}, P) , a random variable is a set function X from Ω to the set of real numbers \mathbb{R} , such that

$$\{\omega \in \Omega; X(\omega) \leq x\} \in \mathcal{F},$$

that is, $X^{-1}((-\infty, x]) \in \mathcal{F}$.



Example 6.1.1 If we toss a fair coin twice, the sample space $\Omega = \{HH, HT, TH, TT\}$. The number of heads is a random quantity that can take the value of 0, 1, 2. What does it mean when we call the number of heads a random variable X ? At the elementary level, we would write $P(X = 1)$ to express the probability of getting one head in two tosses of coin. How does this tie up with X being a set function? From what we know about the probability triple, $X = 1$ must be some set in \mathcal{F} (the power set in this case) for us to assign a probability measure to it. This set must be $\{HT, TH\}$; thus $X = 1$ is equivalent to $\{HT, TH\}$. Hence, when we write $P(X = 1)$, we mean $P(\{HT, TH\})$. We can get to $\{HT, TH\}$ in \mathcal{F} from Ω by mapping $\{HT\}$ and $\{TH\}$ in Ω to 1 using the set function X , that is, $X\{HT\} = 1$, and $X\{TH\} = 1$. Then, using inverse mapping, map 1 to $\{HT, TH\}$ in \mathcal{F} , that is $X^{-1}(1) = \{HT, TH\} \in \mathcal{F}$. It follows that $P(X = 1)$ means $P(X^{-1}(1))$, which in turn can be evaluated as $P(\{HT, TH\}) = P(\{HT\}) + P(\{TH\}) = 1/2$, since by assumption of fair coin, each of the four outcomes is equiprobable.

Example 6.1.2 Suppose $\Omega = (0, 1]$. What does it mean when we call X a uniform random variable in $(0, 1]$? From elementary probability, we would have no problem writing something like $P(X \in (0, 0.2]) = 0.2$, and it would be right. To understand it from first principles, we know we need to have a probability triple (Ω, \mathcal{F}, P) , and that $X \in (0, 0.2]$ must be some set in \mathcal{F} (here, the Borel σ -field) for it to be assigned a probability measure. We can map $(0, 0.2]$ in Ω using X to the interval $(0, 0.2]$ in \mathbb{R} , and then apply inverse mapping X^{-1} to map $(0, 0.2]$ to $(0, 0.2]$ in \mathcal{F} . If we

take the probability measure P here as the Lebesgue measure, then the Lebesgue measure of $(0, 0.2]$ is just its length, 0.2. Thus, $P(X \in (0, 0.2]) = P(\{\omega : X(\omega) \in (0, 0.2]\}) = P(X^{-1}(0, 0.2]) = P((0, 0.2] \in \mathcal{F}) = 0.2$.

Example 6.1.3 Suppose $\Omega = \mathbb{R}$. What does it mean when we say X is a normal random variable with mean 0 and variance 1, and we compute something like $P(X \in (-2, 2))$? For starters, we define a probability triple (Ω, \mathcal{F}, P) , where \mathcal{F} is the Borel σ -field consisting of subsets of intervals of \mathbb{R} (this set is very large). We map $(-2, 2)$ in Ω to $(-2, 2)$ in \mathbb{R} , and then apply inverse mapping $X^{-1}((-2, 2))$ to $(-2, 2) \in \mathcal{F}$. The probability measure P on $(-2, 2)$ is defined as the Lebesgue integral $\int_{(-2, 2)} f dP$, where f is the pdf of $N(0, 1)$. Since f is regular, the Lebesgue integral is just the Riemann integral $\int_{-2}^2 e^{-x^2/2} / \sqrt{2\pi} dx$, from which numerical integration gives the value ≈ 0.95 .

6.2 EXPECTATION OF A RANDOM VARIABLE

Definition 6.2.1 (Expectation of a random variable). Let X be a random variable with probability triple (Ω, \mathcal{F}, P) . Its expectation is defined as the Lebesgue integral

$$\mathbb{E}(X) = \int_{\Omega} X dP.$$



Example 6.2.1 Let $\Omega = (0, 1]$. Define X to be a random variable with probability triple (Ω, \mathcal{F}, P) , where P is the

Lebesgue measure on $(0, 1]$. Suppose $X(\omega) = 1, \omega \in (0, 0.3)$; $X(\omega) = 2, \omega = 0.3$; $X(\omega) = -1, \omega \in (0.3, 1]$. Hence,

$$\begin{aligned}\mathbb{E}(X) &= \int_{\Omega} X^+ dP - \int_{\Omega} X^- dP \\ &= 1(0.3) + 2(0) - 1(0.7) = -0.4.\end{aligned}$$

Theorem 6.2.1. Let X be a non-negative random variable. Then

$$\mathbb{E}(X) = \int_0^{\infty} P(X \geq x) dx.$$

Proof. By definition,

$$\mathbb{E}(X) = \int_{\Omega} X dP = \int_{\Omega} \int_0^{\infty} \mathbb{1}_{x \leq X} dx dP.$$

Since the region of integration is a rectangle, by Fubini's Theorem, we can interchange the order of integration. Hence,

$$\mathbb{E}(X) = \int_0^{\infty} \int_{\Omega} \mathbb{1}_{x \leq X} dP dx = \int_0^{\infty} P(X \geq x) dx.$$

□

Theorem 6.2.2. If X is a non-negative discrete random variable. Then

$$\mathbb{E}(X) = \sum_{x=1}^{\infty} P(X \geq x).$$

Proof. By definition,

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} kP(X = k).$$

Note that $k = \sum_{x=1}^k 1$, hence

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} \sum_{x=1}^k P(X = k).$$

Interchanging order of summation,

$$\mathbb{E}(X) = \sum_{x=1}^{\infty} \sum_{k=x}^{\infty} P(X = k) = \sum_{x=1}^{\infty} P(X \geq x).$$

□

Example 6.2.2 Let $\Omega = (0, \infty)$. Define X to be an exponential(1) random variable with probability triple (Ω, \mathcal{F}, P) . The tail probability is given by $P(X \geq x) = \int_x^{\infty} e^{-t} dt = e^{-x}$. By Theorem 6.2.2,

$$\mathbb{E}(X) = \int_0^{\infty} e^{-x} = 1.$$

If X is a gamma random variable with shape parameter α and scale parameter β , then

$$P(X \geq x) = \int_x^{\infty} \frac{1}{\Gamma(\alpha)\beta^\alpha} t^{\alpha-1} e^{-t/\beta} dt.$$

Using Theorem 6.2.2,

$$\begin{aligned} \mathbb{E}(X) &= \int_0^{\infty} \int_x^{\infty} \frac{1}{\Gamma(\alpha)\beta^\alpha} t^{\alpha-1} e^{-t/\beta} dt dx \\ &= \int_0^{\infty} \int_0^t \frac{1}{\Gamma(\alpha)\beta^\alpha} t^{\alpha-1} e^{-t/\beta} dx dt \\ &= \alpha\beta \int_0^{\infty} \frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}} t^{\alpha+1-1} e^{-t/\beta} dt = \alpha\beta. \end{aligned}$$


Note that the region of integration is the upper half of the line $t = x$, so interchange of order of integration in the

second step gives the rays entering x from 0 and exiting at t , with t running from 0 to ∞ .

Example 6.2.3 Let X be a discrete uniform random variable over the set of integers from 1 to n . Clearly, we have $P(X \geq x) = (n - x + 1)/n = 1 + \frac{1}{n} - \frac{x}{n}$. Hence,

$$\mathbb{E}(X) = \sum_{x=1}^n \left(1 + \frac{1}{n} - \frac{x}{n}\right) = n + 1 - \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}.$$

6.3 INDEPENDENCE

Definition 6.3.1 (Independence of events). Consider the sets $A_1, A_2, \dots, A_n \in \mathcal{F}$. We say that A_1, A_2, \dots, A_n are independent if and only if the joint probability of all possible $n, n-1, n-2, \dots, 2$ events can be written as the product of the probability of each of the $n, n-1, n-2, \dots, 2$ events, respectively. 

Example 6.3.1 Let $A_1, A_2, A_3 \in \mathcal{F}$. Assign $P(A_1) = 1/2$, $P(A_2) = 1/4$, $P(A_3) = 2/3$. Suppose the joint probabilities are given by $P(A_1 \cap A_2 \cap A_3) = 1/12$; $P(A_1 \cap A_2) = 1/8$; $P(A_1 \cap A_3) = 1/3$; $P(A_2 \cap A_3) = 1/10$. We can verify that $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$, $P(A_1 \cap A_2) = P(A_1)P(A_2)$, $P(A_1 \cap A_3) = P(A_1)P(A_3)$. However, $P(A_2)P(A_3) = 1/6 \neq 1/10 = P(A_2 \cap A_3)$. Hence, A_1, A_2, A_3 are not independent.

The concept of independence of a set of events can be extended to a set of random variables. For example, if X_1, X_2 are independent random variables, then $\mathbb{E}(X_1 X_2) =$

$\mathbb{E}(X_1)\mathbb{E}(X_2)$. Thus, if $\mathbb{E}(X_1) = \mathbb{E}(X_2) = 0$, and the second moment exists for X_1 and X_2 then

$$\mathbb{E}(X_1 + X_2)^2 = \mathbb{E}(X_1^2 + 2X_1X_2 + X_2^2) = \mathbb{E}(X_1^2) + \mathbb{E}(X_2^2).$$

Similarly,

$$\mathbb{E}(e^{t(X_1+X_2)}) = \mathbb{E}(e^{tX_1} \cdot e^{tX_2}) = \mathbb{E}(e^{tX_1})\mathbb{E}(e^{tX_2}).$$

6.4 USEFUL INEQUALITIES

In this section, we look at several inequalities in probability.

Theorem 6.4.1. [Markov's inequality] Let X be a random variable with mean μ . If the k th moment of $|X|$ exists for $k = 1, 2, \dots$, then

$$P(|X| \geq \alpha) \leq \frac{\mathbb{E}(|X|^k)}{\alpha^k}.$$

Proof. Let $\mathbb{1}$ denote the indicator function. It is true that

$$|X|^k \geq |X|^k \mathbb{1}_{|X| \geq \alpha} \geq \alpha^k \mathbb{1}_{|X| \geq \alpha}.$$

Taking expectation on both sides, noting that $\mathbb{E}(\mathbb{1}_{|X| \geq \alpha}) = P(|X| \geq \alpha)$, and then rearranging completes the proof. \square

Theorem 6.4.2. [Chebyshev's inequality] Let X be a random variable with mean μ and variance σ^2 . Then

$$P(|X - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}.$$

Proof. We apply Markov's inequality (Theorem 6.4.1) with $k = 2$ and replacing X with its centered form $X - \mu$. Then,

$$P(|X - \mu| \geq \alpha) \leq \frac{\mathbb{E}(X - \mu)^2}{\alpha^2} = \frac{\sigma^2}{\alpha^2},$$

since $\mathbb{E}(X - \mu)^2 = \sigma^2$ by definition. \square

Theorem 6.4.3. [Cantelli's inequality (one-sided)] Let X be a random variable with mean μ and variance σ^2 . Then

$$P(X - \mu \geq \alpha) \leq \frac{\sigma^2}{\sigma^2 + \alpha^2}.$$

Proof. Equivalently, we prove that

$$P(Y \geq \alpha) \leq \frac{\sigma^2}{\sigma^2 + \alpha^2},$$

where $Y = X - \mu$, $\mathbb{E}(Y) = 0$ and $\text{Var}(Y) = \sigma^2$. By Markov's inequality,

$$P(Y \geq \alpha) \leq \frac{\mathbb{E}(Y^2)}{\alpha^2}.$$

Applying the extended form of Markov's inequality with $k = 2$, for some $\beta > 0$,

$$\begin{aligned} P(Y \geq \alpha) = P(Y + \beta \geq \alpha + \beta) &\leq \frac{\mathbb{E}(Y + \beta)^2}{(\alpha + \beta)^2} \\ &= \frac{\sigma^2 + \beta^2}{(\alpha + \beta)^2}, \end{aligned}$$

since $\sigma^2 = \text{Var}(Y + \beta) = \mathbb{E}(Y + \beta)^2 - (\mathbb{E}(Y + \beta))^2 = \mathbb{E}(Y + \beta)^2 - \beta^2$. The upper bound can be sharpened by minimising it with respect to β :

$$\frac{d}{d\beta} \left[\frac{\sigma^2 + \beta^2}{(\alpha + \beta)^2} \right] = \frac{(\alpha + \beta)^2 \cdot 2\beta - (\sigma^2 + \beta^2) \cdot 2(\alpha + \beta)}{(\alpha + \beta)^4} = 0.$$

Factorising, we obtain $(\alpha + \beta)[(\alpha + \beta)\beta - (\sigma^2 + \beta^2)] = (\alpha + \beta)(\alpha\beta - \sigma^2) = 0$. Since $\beta > 0$, the solution is $\beta = \sigma^2/\alpha$. Substituting $\beta = \sigma^2/\alpha$ into the right-hand side of the inequality completes the proof. \square

Theorem 6.4.4. [Cantelli's inequality (two-sided)] Let X be a random variable with mean μ and variance σ^2 . Then

$$P(|X - \mu| \geq \alpha) \leq \frac{2\sigma^2}{\sigma^2 + \alpha^2}.$$

Theorem 6.4.5. [Jensen's inequality] Let X be a random variable with finite moments. If f is a convex function, then

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)).$$

If f is a concave function, then the inequality sign reverses.

Example 6.4.1 Suppose X is a gamma random variable with shape parameter α and scale parameter β . The logarithm function is concave. By Jensen's inequality,

$$\begin{aligned} \log(\mathbb{E}(X)) &\geq \mathbb{E}(\log(X)) \\ \therefore \mathbb{E}(\log(X)) &\leq \log \alpha + \log \beta, \end{aligned}$$

since the gamma distribution with shape parameter α and scale parameter β has mean $\alpha\beta$.

Theorem 6.4.6. [Schwarz's inequality] Let X and Y be random variables with finite moments. Then

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

Example 6.4.2 Let $Y = 1/X$. By Schwarz's inequality,

$$\begin{aligned} 1 &\leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(1/X^2)} \\ \mathbb{E}(X^2) &\geq \frac{1}{\mathbb{E}(1/X^2)} \\ \therefore \mathbb{E}(1/X^2) &\geq \frac{1}{\mathbb{E}(X^2)}. \end{aligned}$$

Since $f(x) = 1/x^2$ is convex, the result of Schwarz's inequality is consistent with those of Jensen's inequality.

IMPORTANT LIMIT THEOREMS IN PROBABILITY THEORY

If we are able to obtain a random sample of size n from the population of interest, then the sample mean is an intuitive and reasonable estimator of the population mean μ . As sample size increases, we demand that the sample mean be "close" in some sense to μ , otherwise our confidence in using the sample mean as an estimator of μ would not have any rational basis. The Law of Large Numbers is the main result that establishes the notion of "closeness" of the sample mean to μ probabilistically. There are two versions of the Law of Large Numbers. The strong version establishes almost sure convergence of the sample mean to μ ; the weak version establishes convergence in probability of the sample mean to μ .

7.1 THE BOREL-CANTELLI LEMMAS

We first look at two theoretical results.

Theorem 7.1.1. [First Borel-Cantelli Lemma] Let $A_1, A_2, \dots \in \mathcal{F}$. If $\sum_n P(A_n)$ converges, then $P(\limsup_n A_n) = 0$.

Proof. Since $\limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \subset \bigcup_{k=m}^{\infty} A_k$, where m is some integer, then

$$0 \leq P(\limsup_n A_n) \leq P(\bigcup_{k=m}^{\infty} A_k) \leq \sum_{k=m}^{\infty} P(A_k).$$

By hypothesis, $\sum_n P(A_n)$ converges, so the tail sum must sum to 0 as $m \rightarrow \infty$. \square

Theorem 7.1.2. [Second Borel-Cantelli Lemma]

Let $A_1, A_2, \dots \in \mathcal{F}$. If $\{A_n\}$ is an independent sequence of events and $\sum_n P(A_n)$ diverges, then $P(\limsup_n A_n) = 1$.

Proof. $P(\limsup_n A_n) = 1$ implies $P(\limsup_n A_n)^c = 0$. Since $(\limsup_n A_n)^c = \liminf_n A_n^c$, we just need to prove that $P(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c) = 0$. It is enough to show that $P(\bigcap_{k=n}^{\infty} A_k^c) = 0$ for all n (alternatively, we can proceed using Boole's inequality). Now, since $1 - x \leq e^{-x}$,

$$0 \leq P\left(\bigcap_{k=n}^{n+j} A_k^c\right) = \prod_{k=n}^{n+j} (1 - P(A_k)) \leq \exp\left(-\sum_{k=n}^{n+j} P(A_k)\right).$$

By hypothesis of divergence of $\sum_n P(A_n)$, the right-hand side tends to 0 as $j \rightarrow \infty$, hence

$$P(\bigcap_{k=n}^{\infty} A_k^c) = \lim_{j \rightarrow \infty} P(\bigcap_{k=n}^{n+j} A_k^c) = 0.$$

\square

7.2 STRONG LAW OF LARGE NUMBERS

Theorem 7.2.1. [The Strong Law of Large Numbers] Let random variables X_1, X_2, \dots, X_n be independent and identically

distributed with finite mean μ and finite variance σ^2 . Denote $S_n = \sum_{i=1}^n X_i$. The sample mean S_n/n converges with probability 1 to μ , that is

$$P(\lim_{n \rightarrow \infty} S_n/n = \mu) = 1.$$

Proof. Without loss of generality, assume $\mu = 0$. We aim to show that $P(|S_n/n| \geq \varepsilon \text{ i.o.}) = 0$, that is, it is impossible that S_n/n escapes from its mean 0 infinitely often. Assume the existence of the second and the fourth moments: $\mathbb{E}(X_i^2) = \sigma^2$, $\mathbb{E}(X_i^4) = \zeta^4$. Consider $\mathbb{E}(S_n^4)$. Expanding S_n^4 , we have $\mathbb{E}(S_n^4) = n\mathbb{E}(X_1^4) + \sum_{i \neq j} \mathbb{E}(X_i^2 X_j^2)$. We can ignore products of random variables where there is one index different from three others (e.g. $X_1 X_2^3$), since for these cases, the expectation will be zero (e.g. $\mathbb{E}(X_1 X_2^3) = \mathbb{E}(X_1)\mathbb{E}(X_2^3) = 0$). Now, $\mathbb{E}(X_i^2 X_j^2) = \mathbb{E}(X_i^2)\mathbb{E}(X_j^2) = \sigma^4$, and the number of such combinations of i, j pairs is equal to $\binom{4}{2} \times \binom{n}{2} = 3n(n-1)$. Therefore

$$\mathbb{E}(S_n^4) = n\zeta^4 + 3n(n-1)\sigma^4 = 3\sigma^2 n^2 + n(\zeta^4 - 3\sigma^4) \leq Kn^2,$$

where K is some constant that does not depend on n . By Markov's inequality for $k = 4$,

$$P(|S_n/n| \geq \varepsilon) = P(|S_n| \geq n\varepsilon) \leq \frac{\mathbb{E}(S_n^4)}{n^4 \varepsilon^4} \leq \frac{Kn^2}{n^4 \varepsilon^4}.$$

Since the sum

$$\sum_{n=1}^{\infty} P(|S_n/n| \geq \varepsilon) \leq \sum_{n=1}^{\infty} \frac{K}{n^2 \varepsilon^4} = \frac{K}{\varepsilon^4} \frac{\pi^2}{6}$$

is convergent, by the First Borel-Cantelli Lemma,

$$P(|S_n/n| \geq \varepsilon \text{ i.o.}) = P(\limsup_n |S_n/n| \geq \varepsilon) = 0.$$

□

7.3 WEAK LAW OF LARGE NUMBERS

Theorem 7.3.1. [The Weak Law of Large Numbers] Let random variables X_1, X_2, \dots, X_n be independent and identically distributed with finite mean μ and finite variance σ^2 . Denote $S_n = \sum_{i=1}^n X_i$. The sample mean S_n/n converges in probability to μ , that is

$$\lim_{n \rightarrow \infty} P(|S_n/n - \mu| \geq \varepsilon) = 0.$$

Proof. Since $\mathbb{E}(S_n/n) = \mu$, and $\text{Var}(S_n/n) = \sigma^2/n$, we apply Chebyshev's inequality:

$$0 \leq P(|S_n/n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0,$$

as $n \rightarrow \infty$. □

7.4 THE CENTRAL LIMIT THEOREM

The Central Limit Theorem (CLT) is one of the most celebrated theorems in probability. Essentially, the result guarantees that, given n independent and identically distributed random variables, the normalised sample mean converges in distribution to $N(0, 1)$, regardless of the underlying distribution of the random variables, as long as they have finite mean and finite variance.

Theorem 7.4.1. [The Central Limit Theorem] Let random variables X_1, X_2, \dots, X_n be independent and identically distributed with finite mean μ and finite variance σ^2 . Denote $S_n = \sum_{i=1}^n X_i$. The normalised sample mean

$$\frac{S_n/n - \mu}{\sigma/\sqrt{n}}$$

converges in distribution to $N(0, 1)$.

It turns out that it is only necessary to check either one of two conditions to be sure that the normalised sample mean converges in distribution to $N(0, 1)$. These two conditions are the Lindeberg condition, and the Lyapunov condition.

Theorem 7.4.2. [The Lindeberg condition for iid random variables] Consider the random variables X_1, X_2, \dots, X_n , which are independent and identically distributed with mean μ and variance $\sigma^2 > 0$. Let $Y_i = X_i - \mu$, and denote $S_n = \sum_{i=1}^n X_i$. If

$$\lim_{n \rightarrow \infty} \frac{1}{\sigma^2} \mathbb{E}(Y_1^2 \mathbb{1}_{|Y_1| \geq \varepsilon \sigma \sqrt{n}}) = 0,$$

then $\sqrt{n}(S_n/n - \mu)/\sigma$ converges in distribution to $N(0, 1)$.

The Lindeberg condition involves checking that the truncated second moment of the centered random variable converges to 0 as $n \rightarrow \infty$, since the denominator is always a positive constant.

Theorem 7.4.3. [The Lyapunov condition for iid random variables] Consider the random variables X_1, X_2, \dots, X_n , which are independent and identically distributed with mean μ and variance $\sigma^2 > 0$. Let $Y_i = X_i - \mu$, and denote $S_n = \sum_{i=1}^n X_i$. If there exists some $\delta > 0$ such that

$$\lim_{n \rightarrow \infty} \frac{1}{n^{\delta/2} (\sigma^2)^{\delta/2+1}} \mathbb{E}|Y_1|^{2+\delta} = 0,$$

then $\sqrt{n}(S_n/n - \mu)/\sigma$ converges in distribution to $N(0, 1)$.

The Lyapunov condition involves checking that the $(2 + \delta)$ th moment of the absolute value of the centered random variable is finite for some $\delta > 0$, since the denominator is

a function of n . For simplicity, we usually choose $\delta = 2$ so that $2 + \delta = 4$ is even, and $|Y_1|^4 = Y_1^4$.

Example 7.4.1 Consider n iid exponential(1) random variables. Denote $S_n = \sum_{i=1}^n X_i$. Since the mean of an exponential(1) random variable is 1, and the variance is also 1. By the Central Limit Theorem, $\sqrt{n}(S_n/n - 1)$ converges in distribution to $N(0, 1)$. Either the Lindeberg condition or the Lyapunov condition guarantees this result. As an exercise, let us check both of these conditions.

Lindeberg condition:

$$\begin{aligned} \mathbb{E}(Y_1^2 \mathbb{1}_{|Y_1| \geq \varepsilon \sigma \sqrt{n}}) &= \int_{|x-1| \geq \varepsilon \sigma \sqrt{n}} (x-1)^2 e^{-x} \mathbb{1}_{x>0} dx \\ &= \int_{x \geq 1 + \varepsilon \sigma \sqrt{n} \cup x \leq 1 - \varepsilon \sigma \sqrt{n}} (x-1)^2 e^{-x} \mathbb{1}_{x>0} dx \\ &= \int_{1 + \varepsilon \sigma \sqrt{n}}^{\infty} (x-1)^2 e^{-x} \mathbb{1}_{x>0} dx + \\ &\quad \int_0^{1 - \varepsilon \sigma \sqrt{n}} (x-1)^2 e^{-x} \mathbb{1}_{x>0} dx. \end{aligned}$$

If $n \rightarrow \infty$, then the lower limit of integration of the first integral (on the right-hand side) tends to ∞ , thus the first integral vanishes (evaluates to 0). The upper limit of integration of the second integral tends to $-\infty$ as $n \rightarrow \infty$. But the exponential pdf is 0 if x is negative, so the second integral also vanishes. Thus, we show that the expectation tends to 0 as $n \rightarrow \infty$.

Lyapunov condition: Choose $\delta = 2$. We just need to show that $\mathbb{E}(Y_1^4) < \infty$.

$$\begin{aligned} \mathbb{E}(Y_1^4) &= \int_0^{\infty} (x-1)^4 e^{-x} dx \\ &= \int_0^2 (x-1)^4 e^{-x} dx + \int_2^{\infty} (x-1)^4 e^{-x} dx. \end{aligned}$$

Since $(x-1)^4 < 1$ when $0 < x < 2$, the first integral on the right-hand side is bounded by

$$\int_0^2 (x-1)^4 e^{-x} dx < \int_0^2 1 \cdot e^{-x} dx = 1 - e^{-2}.$$

When $x \geq 2$, $(x-1)^4 < x^4$. Therefore the second integral on the right-hand side is bounded by

$$\int_2^\infty (x-1)^4 e^{-x} dx < \int_2^\infty x^4 \cdot e^{-x} dx < \int_0^\infty x^4 \cdot e^{-x} dx = 4!,$$

because the right-most integrand is the pdf of the gamma distribution with $\alpha = 5$ and $\beta = 1$. Since both integrals have been shown to be bounded, it follows that $\mathbb{E}(Y_1^4) < \infty$.

7.5 THE LAW OF ITERATED LOGARITHMS

For simplicity of illustration we shall assume that we are dealing with random variables with mean 0 and variance 1. We have seen that the Law of Large Numbers concerns the behaviour of S_n/n . The Strong Law of Large Numbers establishes convergence of S_n/n to 0 with probability 1, and the Weak Law of Large Numbers establishes convergence of S_n/n in probability to 0. Then, the Central Limit Theorem concerns the behaviour of S_n/\sqrt{n} as $n \rightarrow \infty$. It establishes that S_n/\sqrt{n} converges in distribution to $N(0,1)$.

The following result, known as the law of the iterated logarithm, concerns the behaviour of $S_n/\sqrt{2n \log \log n}$ as $n \rightarrow \infty$. Note that the order of magnitude of the denominator $\sqrt{n \log \log n}$ is between \sqrt{n} and n . Essentially, the law states that $S_n/\sqrt{2n \log \log n}$ converges to 0 with probability 1.

Theorem 7.5.1. [The Law of the Iterated Logarithm] Consider the random variables X_1, X_2, \dots, X_n , which are independent and identically distributed with mean 0 and variance 1. Denote $S_n = \sum_{i=1}^n X_i$. Then

$$P \left(\limsup_n \frac{S_n}{\sqrt{2n \log \log n}} = 1 \right) = 1,$$

or equivalently, for $\varepsilon > 0$,

$$\begin{aligned} P(S_n/\sqrt{n} \geq (1 + \varepsilon)\sqrt{2 \log \log n} \text{ i.o.}) &= 0, \\ P(S_n/\sqrt{n} \geq (1 - \varepsilon)\sqrt{2 \log \log n} \text{ i.o.}) &= 1. \end{aligned}$$

CHARACTERISTIC FUNCTION

In elementary probability and statistics, we learn about the moment generating function (mgf) of a probability distribution. For example, the mgf of a discrete distribution such as the Poisson distribution with mean λ can be found as

$$\begin{aligned}
 M(t) = \mathbb{E}(e^{tX}) &= \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda} \lambda^x}{x!} \\
 &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\
 &= e^{-\lambda + \lambda e^t}.
 \end{aligned}$$

For the well known normal distribution with mean μ and variance σ^2 ,

$$\begin{aligned}
 \mathbb{E}(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} dx \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x^2 - 2\mu x - 2t\sigma^2 x + \mu^2)/(2\sigma^2)} dx \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x^2 - (\mu + t\sigma^2))^2 + \mu^2 - (\mu + t\sigma^2)]/(2\sigma^2)} dx \\
 &= e^{(2\mu t\sigma^2 + t^2\sigma^2)/(2\sigma^2)} = e^{\mu t + t^2\sigma^2/2}.
 \end{aligned}$$

The mgf has the property that $M(0) = 1$, and the n th order moment can be found by differentiating $M(t)$ n times, and evaluating it at $t = 0$. It would be nice if every probability distribution has a unique mgf, because this would enable us to identify the distribution of a random variable by simply checking its mgf. However, it turns out that the mgf does not exist for some distributions.

Example 8.0.1 (The Cauchy Distribution does not have an mgf) The Cauchy distribution has pdf given by

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

where $x \in \mathbb{R}$. Now,

$$M(t) = \int_{-\infty}^{\infty} \frac{e^{tx}}{\pi(1+x^2)} dx.$$

Since $e^{tx} = 1 + tx + (tx)^2/2! + \dots$, it is true that $e^{tx} > tx$. Therefore

$$\begin{aligned} M(t) &> \int_{-\infty}^{\infty} \frac{tx}{\pi(1+x^2)} dx \\ &= \frac{t}{2\pi} \int_{-\infty}^{\infty} \frac{2x}{1+x^2} dx \\ &= \frac{t}{2\pi} \lim_{m \rightarrow \infty} \log(1+x^2) \Big|_{-m}^m = \infty. \end{aligned}$$

It turns out that if t is replaced by it , where i is the complex number $\sqrt{-1}$ then $\mathbb{E}(e^{itX})$ is guaranteed to exist for all probability distributions.

Theorem 8.0.1 (The characteristic function). Let X be a random variable with pdf $f(x)$ (or pmf, for discrete distributions). The characteristic function (cf) is defined as $\phi(t) = \mathbb{E}(e^{itX})$ and always exists.

Proof. We just need to show that $\phi(t)$ is always bounded. Using Euler's formula: $e^{ix} = \cos x + i \sin x$, and the fact that $|a + ib| = \sqrt{a^2 + b^2}$,

$$\begin{aligned} \mathbb{E}(e^{itX}) &= \int_{-\infty}^{\infty} e^{itx} f(x) dx \\ &= \int_{-\infty}^{\infty} \{\cos(tx) + i \sin(tx)\} f(x) dx \\ &\leq \int_{-\infty}^{\infty} |\{\cos(tx) + i \sin(tx)\}| f(x) dx \\ &\leq 1, \end{aligned}$$

since $\cos^2(tx) + \sin^2(tx) = 1$. \square

Strictly speaking, the evaluation of the cf requires the application of complex integration. However, to acquire this skill, a semester of study in complex analysis is usually required. Thus, for practical use in this course, we treat i as a real constant and proceed with real integration. In general, this heuristic produces results that are identical to those obtained using proper complex integration. Another approach to evaluating cf is to use the mgf, if it exists. We first find the mgf, and then replace i with it .

Example 8.0.2 The cf of the uniform distribution in $(0, 1)$ is given by

$$\phi(t) = \int_0^1 e^{itx} \cdot 1 dx = \left. \frac{e^{itx}}{it} \right|_0^1 = \frac{e^{it} - 1}{it}.$$

Using Euler's formula, we can express $\phi(t)$ equivalently as

$$\begin{aligned} \phi(t) &= \frac{\cos t - 1 + i \sin t}{it} \times \frac{i}{i} \\ &= \frac{\sin t}{t} + i \left(\frac{1 - \cos t}{t} \right). \end{aligned}$$

Example 8.0.3 The mgf of the standard normal distribution is given by $e^{t^2/2}$. Replacing t with it , then we obtain the cf $\phi(t) = e^{-t^2/2}$. Suppose we use the machinery of real integration. Treating i as a constant, we get

$$\begin{aligned}\phi(t) &= \int_{-\infty}^{\infty} \frac{e^{itx}}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-[(x-it)^2+t^2]/2} \\ &= e^{-t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-it)^2/2} \\ &= e^{-t^2/2},\end{aligned}$$

since the integrand is a "normal pdf with mean it and variance 1".

The n th moment of a random variable X can be found by differentiating its cf n times, dividing it by i^n , and then setting $t = 0$.

Theorem 8.0.2. Let X be a random variable with cf $\phi(t)$. The n th moment of X is given by $\phi^{(n)}(0)$, where $\phi^{(n)}(t) = \frac{d^n}{dt^n} \phi(t)$.

Proof. Writing e^{itX} using its Maclaurin series representation and then taking expectation, we get

$$\mathbb{E}(e^{itX}) = \mathbb{E} \left(1 + (itX) + \frac{(itX)^2}{2!} + \frac{(itX)^3}{3!} + \cdots \right).$$

Differentiating both sides once, we get

$$\begin{aligned}\phi^{(1)}(t) &= \mathbb{E} (iX + t(iX)^2 + t(iX)^3/2! + \cdots) \\ \phi^{(1)}(0) &= i\mathbb{E}(X) \\ \therefore \frac{\phi^{(1)}(0)}{i} &= \mathbb{E}(X).\end{aligned}$$

Differentiating n times, the first term of the infinite series on the right-hand side is $(iX)^n$, whereas the other terms are functions of t . Therefore,

$$\frac{\phi^{(n)}(0)}{i^n} = \mathbb{E}(X^n).$$

□

Example 8.0.4 The cf of the standard normal distribution is given by $\phi(t) = e^{-t^2/2}$. Since $\phi^{(1)}(t) = -t\phi(t)$, the mean (first moment) is $\phi^{(1)}(0)/i = 0$. Differentiating another time, we have $\phi^{(2)}(t) = -(t\phi^{(1)}(t) + \phi(t) \cdot 1)$. The second moment is therefore equal to $\phi^{(2)}(0)/i^2 = -(0 + 1)/(-1) = 1$, which means the variance is $1 - 0^2 = 1$, as expected.

Example 8.0.5 The modified Bernoulli random variable X has pmf given by $P(X = -1) = 1/2$ and $P(X = 1) = 1/2$. Its cf is then $\phi(t) = \frac{1}{2}e^{-it} + \frac{1}{2}e^{it} = \cosh(it)$. Alternatively, using Euler's formula, the numerator is $\cos(-t) + i\sin(-t) + \cos t + i\sin t = 2\cos t$, since $\cos t$ is an even function and $\sin t$ is an odd function. Therefore, $\phi(t) = \cos t$. We then have $\phi^{(1)}(t) = -\sin t$, $\phi^{(2)}(t) = -\cos t$, $\phi^{(3)}(t) = \sin t$, $\phi^{(4)}(t) = \cos t = \phi(t)$. Hence by induction, we have $\phi^{2k-1}(0) = 0$, and $\phi^{2k}(0) = (-1)^k$. We can conclude that the odd moments of X are 0, and the even moments are 1, since $(-1)^k/i^{2k} = 1$.

For a continuous distribution, its pdf can be recovered from knowledge of cf using a result known as the inversion formula.

Theorem 8.0.3. Let X be a continuous random variable with cf $\phi(t)$. The pdf of X , $f(x)$, can be obtained from its cf through the use of the inversion formula:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(t) dt.$$

Example 8.0.6 The Cauchy distribution has cf $\phi(t) = e^{-|t|}$. Using the inversion formula, we should recover $f(x) = 1/\{\pi(1+x^2)\}$. Let's try out the inversion formula. Putting aside the constant $1/2\pi$, we have

$$\int_{-\infty}^{\infty} e^{-itx} e^{-|t|} dt = \int_0^{\infty} e^{-itx} e^{-t} dt + \int_{-\infty}^0 e^{-itx} e^t dt.$$

The first integral evaluates to

$$\int_0^{\infty} e^{-t(1+ix)} dt = \left. \frac{-e^{-t(1+ix)}}{1+ix} \right|_0^{\infty} = \frac{1}{1+ix}.$$

For the second integral, let $t = -u$. Then,

$$\begin{aligned} \int_{-\infty}^0 e^{-itx} e^t dt &= - \int_{\infty}^0 e^{iux} e^{-u} du \\ &= \int_0^{\infty} e^{-u(1-ix)} du \\ &= \left. \frac{-e^{-u(1-ix)}}{1-ix} \right|_0^{\infty} \\ &= \frac{1}{1-ix}. \end{aligned}$$

Summing both integrals,

$$\frac{1}{1+ix} + \frac{1}{1-ix} = \frac{2}{1+x^2}.$$

Multiplying by the constant $1/(2\pi)$ recovers the Cauchy pdf.

It is also possible to use cf to compute probabilities of continuous distributions. However, they are generally useful for theoretical situations (e.g. finding bounds for probabilities) as their evaluations require complex integration and may be complicated.

Theorem 8.0.4. If the probability measure μ has characteristic function $\phi(t)$, and if $\mu\{a\} = \mu\{b\} = 0$, then

$$\mu(a, b] = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} \phi(t) dt.$$

Example 8.0.7 Let X be a standard normal random variable. Suppose we wish to find $P(-1 < X < 1)$. Using Theorem 8.0.4, we have

$$\begin{aligned} \mu(-1, 1] &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{it} - e^{-it}}{it} e^{-t^2/2} dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{(\cos t + i \sin t) - (\cos(-t) + i \sin(-t))}{it} e^{-t^2/2} dt \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sin t}{t} e^{-t^2/2} dt \\ &\leq \frac{1}{\pi} \int_{-\infty}^{\infty} e^{-t^2/2} dt \\ &= \frac{\sqrt{2\pi}}{\pi} = \sqrt{\frac{2}{\pi}} \approx 0.8, \end{aligned}$$

since $\sin t/t \leq 1$. From numerical integration, the actual value of $\mu(-1, 1]$ is 0.682. For $\mu(-2, 2]$, we know that the actual value is about 0.95. Using Theorem 8.0.4,

$$\begin{aligned} &\mu(-2, 2] \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{2it} - e^{-2it}}{it} e^{-t^2/2} dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{(\cos 2t + i \sin 2t) - (\cos(-2t) + i \sin(-2t))}{it} e^{-t^2/2} dt \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sin 2t}{t} e^{-t^2/2} dt \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{2 \sin t \cos t}{t} e^{-t^2/2} dt \\ &\leq \frac{2}{\pi} \int_{-\infty}^{\infty} e^{-t^2/2} \cos t dt. \end{aligned}$$

Now,

$$\int_{-\infty}^{\infty} e^{it} e^{-t^2/2} dt = \int_{-\infty}^{\infty} e^{-t^2/2} \cos t \, dt + i \int_{-\infty}^{\infty} e^{-t^2/2} \sin t \, dt.$$

But the left-hand side is equal to

$$\begin{aligned} \int_{-\infty}^{\infty} e^{it} e^{-t^2/2} dt &= \int_{-\infty}^{\infty} e^{-(t^2 - 2it)/2} dt \\ &= \int_{-\infty}^{\infty} e^{-(t-i)^2/2} e^{-1/2} dt \\ &= \sqrt{\frac{2\pi}{e}}. \end{aligned}$$

Since the integrand of the imaginary part is an odd function, the integral integrates to 0. Thus, the integral of the real part is equal to the left hand side. Thus, $\mu(-2, 2] \leq 2\sqrt{2}/\sqrt{\pi e} \approx 0.968$.