STA347: Probability Theory Lecture Notes

by

Dayi Li

Department of Statistical Sciences
University of Toronto

Jun 2024

TABLE OF CONTENTS

1 Probability Basics 1

1.1 Probability Measures 1

1.2 Independence and Conditional Probability 5

1.3 Exercises 8

2 Further Probability Basics 9

2.1 Random Variables 9

2.2 Distributions 10

2.3 Conditional Distribution 15

2.4 Exercises 17

3 Convergence 18

3.1 Limit Events 18

3.2 Types of Convergence 20

3.3 Exercises 22

4 Expectation 23

4.1 Basic Definition and Properties 23

4.2 Conditional Expectation 25

4.3 Limit Theorems 26

4.4 Moment-Generating Functions 28

4.5 Exercises 30

5 Properties of Expectation 32

5.1 Concentration of Measure 32

5.2 Various Inequalities 34

5.3 L^p Convergence 35

5.4 Exercises 37

6 Laws of Large Numbers 39

6.1 Familiar Results 39

6.2 Advanced Weak LLN 40

6.3 Advanced Strong LLN 41

6.4 Applications 42

6.5 Exercises 44

7 Central Limit Theorems 45

7.1 Weak Convergence 45

7.2 Characteristic Functions 48

7.3 Central Limit Theorem 50

7.4 Exercises 52

8 Stochastic Process 53

8.1 Poisson Process 53

8.2 Properties of Poisson Process 57

8.3 Exercise 60

LECTURE 1

PROBABILITY BASICS

1.1 Probability Measures

Definition 1.1.1. A *sample space* Ω is any non-empty set.

Remark. A sample space can be anything from a concrete set of objects to a set of highly abstract objects. The sample space contains "all the states of the world".

Example 1.1.2. Real numbers on an interval: $\Omega = [0, 1]$.
Coin flipping: $\Omega = \{H, T\}$.
Weather: $\Omega = \{\text{hot, cold, mild}\} \times \{\text{wet, dry}\}$.
Extreme example: $\Omega = \{\text{all possible locations of every grain of sand in the Dune Sea on Tatooine}\}$.

Definition 1.1.3. For a sample space Ω, an event is a subset $E \subseteq \Omega$.

Remark (Some Heuristics of Probability). As we have seen in introductory probability courses (STA 257), a central theme that we are interested in is to say that a certain event E has some probability that is a real number between 0 and 1. Mathematically, we write this as $\mathbb{P}(E) \in [0, 1]$. As we change the event E, the corresponding $\mathbb{P}(E)$ will change, but every event E only has one value of $\mathbb{P}(E)$ associated with it.

Based on the above observation, modern probability theory formulates the notion of probability as a function $\mathbb{P}(\cdot)$. However, this function is not the typical function that we have encountered in first/second year calculus courses that maps from $\mathbb{R}$ to $\mathbb{R}$. Instead, it maps some events or sets, i.e., $E \subset \Omega$, to some real numbers in $[0, 1]$, thus, probability is a *set function*.

To this end, in order to formally define probability as a function, we need to specify its domain and range. It is obvious that the range of this function is $[0, 1]$. However, what is the domain of this function? As we have mentioned, probability is a function that maps an event (or a set) $E \subset \Omega$ to $[0, 1]$, so certainly the domain of this function has to be related to various subsets (events) of the sample space Ω. Therefore, we require the following notion of σ-algebra to properly define the domain of $\mathbb{P}$.

Definition 1.1.4. (σ-Algebra) For a sample space Ω, let $\mathcal{P}(\Omega)$ be its power set. A subset $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is called a σ-algebra (or σ-field) on Ω if and only if

i) $\Omega \in \mathcal{F}$;
ii) If an event $E \in \mathcal{F}$, then $E^c \in \mathcal{F}$;
iii) If a sequence of events $E_1, E_2, E_3, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} E_i \in \mathcal{F}$.

A σ-algebra on Ω is what we will be using as the domain of $\mathbb{P}$. But you will be asking the question, why do we need to go to the trouble of using σ-algebra? Why can't we just use the power set $\mathcal{P}(\Omega)$? Well, this is because some of the basic properties of $\mathbb{P}(\cdot)$ that we want it to satisfy can sometime be impossible to achieve if we define its domain to be $\mathcal{P}(\Omega)$. Next, we will discuss these basic properties of $\mathbb{P}(\cdot)$ that we want it to satisfy.

Definition 1.1.5. (General Measure) For a set $\mathcal{X}$, and a σ-algebra $\mathcal{F}$ on $\mathcal{X}$, we say μ is a *measure* defined on $\mathcal{F}$ if the following hold:

i) (non-negative) $\forall A \subseteq \mathcal{X}, \mu(A) \geq 0$, that is, $\mu : \mathcal{F} \to [0, \infty)$;
ii) (empty set has zero measure) $\mu(\emptyset) = 0$;
iii) (countably additive) $\forall A_1, A_2, \ldots \subseteq \mathcal{X}$ such that $A_i \cap A_j = \emptyset,\ \forall\ i \neq j$,

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i). \tag{1.1}$$

We call $(\mathcal{X}, \mathcal{F})$ a measurable-space.

Definition 1.1.6. (Probability Measure) For a sample space Ω, and a σ-algebra $\mathcal{F}$ on Ω, we say $\mathbb{P}$ is a *probability measure* defined on $\mathcal{F}$ if the following hold:

i) $\forall E \in \mathcal{F}, 0 \leq \mathbb{P}(E) \leq 1$, that is, $\mathbb{P} : \mathcal{F} \to [0, 1]$;
ii) $\mathbb{P}(\Omega) = 1$;
iii) $\forall E_1, E_2, \ldots \in \mathcal{F}$ such that $E_i \cap E_j = \emptyset,\ \forall\ i \neq j$,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i). \tag{1.2}$$

We call $(\Omega, \mathcal{F}, \mathbb{P})$ a probability triplet or a probability space.

Example 1.1.7. Probabilities of flipping a coin twice. Let $\Omega = \{HH, HT, TH, TT\}$, $\mathcal{F} = \mathcal{P}(\Omega)$. $\mathbb{P}(HH) = \mathbb{P}(HT) = \mathbb{P}(TH) = \mathbb{P}(TT) = 1/4$. Then $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

Remark. For the general measure in definition 1.1.5, it is a mathematically rigorous description and generalization of our typical notion of length/area/volume. A measure is a set function defined on a collection of sets (a σ-algebra) that maps a set to a non-negative real number.

For a probability measure in definition 1.1.6, we are simply restricting the range of our measure to $[0, 1]$.

In this course, we will not dive into the details of σ-algebra and measure, but it is important to know of their existence for a better understanding of probability. In essence, σ-algebra contains information that we are interested in, for example, what are the different possible outcomes of an experiment.

Now go back to our previous question: why not just consider the power set $\mathcal{P}(\Omega)$ upon which to define our $\mathbb{P}$? The reason is that if we want $\mathbb{P}$ to satisfy the above axioms on the power set $\mathcal{P}(\Omega)$, such $\mathbb{P}$ may not exist (see Vitali set for example). σ-algebra instead is a mid-way solution: it is big enough so it encompasses nearly all the events that we are interested in, and it is also a "nice" set upon which we can define a probability measure properly.

Although we will not be covering σ-algebra, we will use the notations such as $(\Omega, \mathcal{F}, \mathbb{P})$ for accuracy.

Example 1.1.8 (Lebesgue/Uniform Measure on an Interval). Let $\Omega = [L, U]$ for some $L < U \in \mathbb{R}$. $\mathcal{F} = \mathcal{B}(\Omega)$ is the Borel σ-algebra on Ω. Define $\mathbb{P}([a, b)) = \frac{b-a}{U-L}$ when $L \leq a \leq b \leq U$. $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

Proof. In this class we can assume this is true. □

Remark. Borel σ-algebra and the formal definition of Lebesgue measure is out of the scope of this course, thus we do not cover the proof. It suffices to know that the Borel σ-algebra is the σ-algebra that contains all the "nice" sets and pretty much everything that we are interested in about the interval $[L, U]$. Borel σ-algebra is generated using intervals in Ω through operations of complement, intersection, and union.

Lebesgue measure is the mathematically rigorous definition of length/area/volume. The standard definition of uniform probability measure on an interval/region is thus defined through Borel σ-algebra on the interval/region and the Lebesgue measure.

Example 1.1.9 (Counting/Uniform Measure on a finite space). Let Ω be a finite sample space, i.e., $\Omega = \{x_i\}_{i=1}^n$. $\mathcal{F} = \mathcal{P}(\Omega)$. If we define $\mathbb{P}$ to assign every element in Ω the same probability measure, i.e., $\mathbb{P}(\{x_i\}) = 1/n$, $\forall x_i \in \Omega$, then $\mathbb{P}$ is the uniform probability measure on $\mathcal{F}$. Moreover, $\mathbb{P}(A) = |A|/|\Omega|$ for $A \subseteq \Omega$ where $|\cdot|$ is the cardinality of a set.

Proposition 1.1.10. The following properties are satisfied by any probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

1. $\forall E \in \mathcal{F}, \mathbb{P}(E^c) = 1 - \mathbb{P}(E)$;
2. $\mathbb{P}(\emptyset) = 0$;
3. $\forall E, F \in \mathcal{F}$ such that $E \subseteq F$, $\mathbb{P}(E) \leq \mathbb{P}(F)$;
4. $\forall E, F \in \mathcal{F}$, $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$;
5. $\forall E, F \in \mathcal{F}$, $\mathbb{P}(E \cap F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cup F)$;
6. $\forall E, F \in \mathcal{F}$, $\mathbb{P}(F \setminus E) = \mathbb{P}(F) - \mathbb{P}(F \cap E)$.

Proof. Exercise (review from STA257). □

Lemma 1.1.11. Consider a sequence of sets E_i. Then,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(E_i) \tag{1.3}$$

Proof. Let $F_1 = E_1$, $F_i = E_i \cap \left(\bigcup_{k=1}^{i-1} F_k\right)^c$ for $i \geq 2$. Observe that F_i's are disjoint and $\bigcup_{i=1}^{\infty} F_i = \bigcup_{i=1}^{\infty} E_i$. Moreover, $F_i \subseteq E_i$ for all i. Thus,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(F_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(E_i). \tag{1.4}$$

□

Lemma 1.1.12. Let $\mathbb{P}$ be the uniform measure on $[0, 1]$. Then $\mathbb{P}(E) = 0$ for any countable set E.

Proof. Let $E = \{x_i\}_{i=1}^{\infty}$. Fix $\varepsilon > 0$ and define $E_i(\varepsilon) = [x_i - 2^{-i}\varepsilon, x_i + 2^{-i}\varepsilon)$. Observe that $E \subseteq$

$\bigcup_{i=1}^{\infty} E_i(\varepsilon)$. Thus,

$$\mathbb{P}(E) \leq \mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i(\varepsilon)\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(E_i(\varepsilon)) = \sum_{i=1}^{\infty}(x_i + 2^{-i}\varepsilon - x_i + 2^{-i}\varepsilon) = 2\varepsilon \sum_{i=1}^{\infty} 2^{-i} = 2\varepsilon. \tag{1.5}$$

Since ε was arbitrary, $\mathbb{P}(E) = 0$. □

Proposition 1.1.13. Let $\Omega = \mathbb{N}$. There is no uniform probability measure on Ω.

Proof. Exercise. □

Lemma 1.1.14. Consider $E_1, \ldots, E_n$. Then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathbb{P}(E_i) - \sum_{i<j} \mathbb{P}(E_i \cap E_j) + \sum_{i<j<k} \mathbb{P}(E_i \cap E_j \cap E_k) + \cdots + (-1)^{n+1}\mathbb{P}\left(\bigcap_{i=1}^{n} E_i\right). \tag{1.6}$$

Proof. Exercise. *Hint: Induction* □

Definition 1.1.15 (Monotone Sequence of Events). We say a sequence of events $\{A_n\}$ is *non-increasing* if $A_{n+1} \subseteq A_n$ for each n, and $\{A_n\}$ is *non-decreasing* if $A_n \subseteq A_{n+1}$ for each n.

Definition 1.1.16. Let $\{A_n\}$ be a monotone sequence of events, i.e., either non-increasing or non-decreasing. We define the limit of $\{A_n\}$ as

$$\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n \tag{1.7}$$

if $\{A_n\}$ is non-increasing, and

$$\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n \tag{1.8}$$

if $\{A_n\}$ is non-decreasing.

If $\{A_n\}$ is non-increasing, we also write $\{A_n\} \searrow A$ where $A = \bigcap_{n=1}^{\infty} A_n$. We say that $\{A_n\}$ converges from *above* to A.

If $\{A_n\}$ is non-decreasing, we also write $\{A_n\} \nearrow A$ where $A = \bigcup_{n=1}^{\infty} A_n$. We say that $\{A_n\}$ converges from *below* to A.

Proposition 1.1.17 (Continuity of Measure.). If $\{A_n\}$ is monotone, then $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(\lim_{n\to\infty} A_n)$.

Proof. Suppose $\{A_n\} \nearrow \lim_{n\to\infty} A_n = A$, and by convention let $A_0 = \emptyset$. Define $B_1 = A_1$ and $B_n = A_n \cap A_{n-1}^c$ for $n \geq 2$. Observe that

$$\begin{aligned}\bigcup_{m=1}^{n} B_m &= \bigcup_{m=1}^{n} (A_m \cap A_{m-1}^c) \\ &= \bigcup_{m=1}^{n} A_m \cap \bigcup_{m=1}^{n} A_{m-1}^c \\ &= A_n \cap A_0^c \\ &= A_n.\end{aligned} \tag{1.9}$$

Then,

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}\left(\bigcup_{m=1}^{\infty} B_m\right) \\ &= \sum_{m=1}^{\infty} \mathbb{P}(B_m) \\ &= \lim_{n\to\infty} \sum_{m=1}^{n} \mathbb{P}(B_m) \\ &= \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{m=1}^{n} B_m\right) \\ &= \lim_{n\to\infty} \mathbb{P}(A_n).\end{aligned} \tag{1.10}$$

If $\{A_n\} \searrow A$, then $\{A_n^c\} \nearrow A^c$, so

$$\begin{aligned}\mathbb{P}(A) &= 1 - \mathbb{P}(A^c) \\ &= 1 - \lim_{n\to\infty} \mathbb{P}(A_n^c) \\ &= 1 - \lim_{n\to\infty} [1 - \mathbb{P}(A_n)] \\ &= \lim_{n\to\infty} \mathbb{P}(A_n).\end{aligned} \tag{1.11}$$

□

Example 1.1.18. Uniform measure and $A_n = [0, 1 - 1/n]$.

Example 1.1.19.

$$A_n = \begin{cases} \Omega, & n \text{ odd} \\ \emptyset, & n \text{ even} \end{cases}. \tag{1.12}$$

1.2 Independence and Conditional Probability

Definition 1.2.1. Events $E_1, \ldots, E_n \subseteq \Omega$ are independent with respect to a probability measure $\mathbb{P}$ if for each $\mathcal{I} \subseteq [n]$

$$\mathbb{P}\left(\bigcap_{i\in\mathcal{I}} E_i\right) = \prod_{i\in\mathcal{I}} \mathbb{P}(E_i). \tag{1.13}$$

Proposition 1.2.2. If $E_1, \ldots, E_n$ are independent, then $E_1^c, \ldots, E_n$ are independent.

Proof. Fix $\mathcal{I} \subseteq [n]$. If $1 \notin \mathcal{I}$, then clearly

$$\mathbb{P}\left(\bigcap_{i\in\mathcal{I}} E_i\right) = \prod_{i\in\mathcal{I}} \mathbb{P}(E_i). \tag{1.14}$$

If $1 \in \mathcal{I}$, then define $\mathcal{I}' = \mathcal{I} \setminus \{1\}$.

$$\begin{aligned}\mathbb{P}\left(E_1^c \cap \bigcap_{i \in \mathcal{I}'} E_i\right) &= \mathbb{P}\left(\bigcap_{i \in \mathcal{I}'} E_i \setminus E_1\right) \\ &= \mathbb{P}\left(\bigcap_{i \in \mathcal{I}'} E_i\right) - \mathbb{P}\left(E_1 \cap \bigcap_{i \in \mathcal{I}'} E_i\right) \\ &= \prod_{i \in \mathcal{I}'} \mathbb{P}(E_i) - \prod_{i \in \mathcal{I}} \mathbb{P}(E_i) \\ &= [1 - \mathbb{P}(E_1)] \prod_{i \in \mathcal{I}'} \mathbb{P}(E_i) \\ &= \mathbb{P}(E_1^c) \prod_{i \in \mathcal{I}'} \mathbb{P}(E_i).\end{aligned} \tag{1.15}$$

□

Definition 1.2.3. An infinite collection of events $\{E_\alpha : \alpha \in \mathcal{I}\}$ are independent if for any finite subset $\mathcal{J} \subseteq \mathcal{I}$, the events $\{E_i : i \in \mathcal{J}\}$ are independent.

Definition 1.2.4. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and an event $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, the conditional probability of an event $A \in \mathcal{F}$ given B is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1.16}$$

Lemma 1.2.5. Two events A and B with non-zero measure are independent if and only if that either

$$\mathbb{P}(A \mid B) = \mathbb{P}(A), \tag{1.17}$$

$$\mathbb{P}(B \mid A) = \mathbb{P}(B). \tag{1.18}$$

Proof. Follows directly from definition. □

Lemma 1.2.6. Given the setting in definition 1.2.4, define

$$\mathbb{P}_B(A) = \mathbb{P}(A \mid B), \forall A \in \mathcal{F}. \tag{1.19}$$

$\mathbb{P}_B$ is also a probability measure.

Proof. Exercise. □

Proposition 1.2.7 (Law of Total Probability). Suppose $E_1, \ldots, E_n \subseteq \Omega$ are non-empty and satisfy $E_i \cap E_j = \emptyset,\ \forall i \neq j$ and $\bigcup_{i=1}^n E_i = \Omega$. Then for an event $A \subseteq \Omega$,

$$\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A \mid E_i)\mathbb{P}(E_i). \tag{1.20}$$

Proof. Observe that for any A, if $i \neq j$, we have $(A \cap E_i) \cap (A \cap E_j) = \emptyset$. Thus,

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \cap \Omega) \\ &= \mathbb{P}\left(A \cap \bigcup_{i=1}^{n} E_i\right) \\ &= \mathbb{P}\left(\bigcup_{i=1}^{n}(A \cap E_i)\right) \\ &= \sum_{i=1}^{n} \mathbb{P}(A \cap E_i) \\ &= \sum_{i=1}^{n} \mathbb{P}(A \mid E_i)\mathbb{P}(E_i).\end{aligned} \tag{1.21}$$

□

Proposition 1.2.8. Law of Total Probability also holds when we have a countably infinite sequence of events $E_1, E_2, \dots$.

Proof. Exercise. □

Theorem 1.2.9 (Bayes' Theorem). Given two non-empty events A, B, we have

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}. \tag{1.22}$$

Proof. Follows from definition. □

Proposition 1.2.10. For an event A, and non-empty $E_1, \dots, E_n \subseteq \Omega$ with $E_i \cap E_j = \emptyset$, $\forall i \neq j$ and $\bigcup_{i=1}^{n} E_i = \Omega$. We have

$$\mathbb{P}(E_i \mid A) = \frac{\mathbb{P}(A \mid E_i)\mathbb{P}(E_i)}{\sum_{j=1}^{n} \mathbb{P}(A \mid E_j)\mathbb{P}(E_j)}, \forall i \in [n] \tag{1.23}$$

Proof. Follows directly from above. □

1.3 Exercises

Exercise 1.1. Prove Proposition 1.1.10.

Exercise 1.2. Prove lemma 1.1.14.

Exercise 1.3. Prove proposition 1.1.13.

Exercise 1.4. Prove proposition 1.2.8.

Exercise 1.5. Suppose there are n people, each assigned a number from $1, \dots, n$. There are also n number of seats assigned the number $1, \dots, n$. Now we assign the n people into the n seats. What is the probability that none of the n people is assigned the seat that match their number. What happens if $n \to \infty$?

Exercise 1.6. Suppose we are flipping a coin. At the n-th flip, the probability of landing a head is m/n where m is the number of times we have flipped heads. Suppose we flip the coin for n times. Given we got a head in the first flip, what is the probability that we get k number of heads for $k \in \{1, \dots, n\}$.

Exercise 1.7. Suppose $E_1, E_2, \dots \subseteq \Omega$. Show that $\mathbb{P}(E_i) = 1$ for all $i \in \mathbb{N}$ if and only if $\mathbb{P}\left(\bigcap_{i=1}^{\infty} E_i\right) = 1$.

Exercise 1.8. Find an example of an Ω, $\mathbb{P}$, and sets A, B, C such that

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$$

but A, B, C are not independent. *Hint: Ω does not need to have more than 4 elements.*

Exercise 1.9. Prove lemma 1.2.6.

Exercise 1.10. Prove that if $\{A_\alpha\}_{\alpha \in \mathcal{I}}$ is independent then so is $\{A_\alpha^c\}_{\alpha \in \mathcal{I}}$.

Exercise 1.11. Suppose Ω is countable. Prove that it is impossible for there to be a sequence of events $A_1, A_2, \dots, \subseteq \Omega$ that are independent and $\mathbb{P}(A_i) = 1/2$ for all i.

Exercise 1.12. Show that the previous question is still the case if $\mathbb{P}(A_i) = p \in (0, 1)$ for all i.

LECTURE 2

FURTHER PROBABILITY BASICS

2.1 Random Variables

Definition 2.1.1. Given a sample space Ω, a *random variable* (R.V.) is a function $X : \Omega \longrightarrow \mathbb{R}$.

Example 2.1.2. Coin flipping. $X(H) = 1, X(T) = 0$.

Remark. A R.V. X is neither random nor a variable. It is just a deterministic mapping from a sample space to the real line. The only random part is the outcome ω arising from the sample space Ω. As in the above example, when we flip a coin, the outcome can be either head or tail, which is random. The sample space of this experiment is then $\Omega = \{H, T\}$. A R.V. X defined on Ω can then be the number of head we get after flipping a coin. Therefore, in this case, we have

$$\begin{aligned} X(\omega) &= 1, \text{ if } \omega = H, \\ X(\omega) &= 0, \text{ if } \omega = T. \end{aligned} \tag{2.1}$$

However, definition 2.1.1 is not exactly sufficient. This is because when talking about R.V., we are almost always interested in quantities such as $\mathbb{P}(X \in A)$ for some $A \subseteq \mathbb{R}$. This means that we need the event $\{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\}$ to be a well-defined set (or measurable), i.e., it needs to be in a σ-algebra $\mathcal{F}$, so that $\mathbb{P}(X \in A)$ is well-defined. Most of A that we are interested in are all contained in the Borel σ-algebra of $\mathbb{R}$, $\mathcal{B}(\mathbb{R})$, which are all the sets generated by intervals in $\mathbb{R}$.

Formally, we require $\{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}$ for any $A \in \mathcal{B}(\mathbb{R})$, and we call X "$\mathcal{F}/\mathcal{B}$-measurable". Thus, the actual definition of a R.V. is *a function $X : \Omega \to \mathbb{R}$ on a measurable space $(\Omega, \mathcal{F})$ that is "$\mathcal{F}/\mathcal{B}$-measurable"*. More succinctly, we say that a R.V. is *a measurable map $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$*. Since we will also equip $(\Omega, \mathcal{F})$ with a probability measure $\mathbb{P}$, when we are talking about a R.V. X, it is always defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

In fact, if we require all the events of the form $\{\omega \in \Omega : X(\omega) \in A\}$ with $A \in \mathcal{B}(\mathbb{R})$ to be in some σ-algebra $\mathcal{F}$, there is a minimal sized σ-algebra that satisfies this condition, and we call this σ-algebra the σ-algebra generated by X, or $\sigma(X)$. That is, $\sigma(X) = \{\{\omega \in \Omega : X(\omega) \in A\} : A \in \mathcal{B}(\mathbb{R})\}$.

However, the details here are out of the scope of the course as it ventures into measure theory. For the remainder of the course, when we are talking about R.V.s, we will just assume that previous details are satisfied.

Definition 2.1.3. Random variables $X_1, \ldots, X_n$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are *independent* if for all sets $A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}\left(\bigcap_{i=1}^{n} \{X_i \in A_i\}\right) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i). \tag{2.2}$$

Theorem 2.1.4. Random variables $X_1, \dots, X_n$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are *independent* if and only if for any $x_1, \dots, x_n \in \mathbb{R}$,

$$\mathbb{P}\left(X_1 \leq x_1, \dots X_n \leq x_n\right) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x_i). \tag{2.3}$$

Proof. The reverse direction is obvious by definition. The forward direction is outside the scope of this class. □

Definition 2.1.5. An infinite collection of random variables $\{X_\alpha : \alpha \in \mathcal{I}\}$ are independent if for any finite subset $\mathcal{J} \subseteq \mathcal{I}$, the random variables $\{X_i : i \in \mathcal{J}\}$ are independent.

2.2 Distributions

Lemma 2.2.1. A random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ induces a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by

$$\mu(A) = \mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}) \tag{2.4}$$

for any $A \in \mathcal{B}(\mathbb{R})$.

Proof. $\mu(A) \in [0,1]$ for any $A \in \mathcal{B}(\mathbb{R})$ is trivial since $\mathbb{P}$ is a probability measure. Also, since $X(\omega) \in \mathbb{R}$ for all $\omega \in \Omega$, $\mu(\mathbb{R}) = \mathbb{P}(\Omega) = 1$. Finally, consider disjoint $A_1, A_2, \dots \in \mathcal{B}(\mathbb{R})$, and define $E_i = \{\omega : X(\omega) \in A_i\}$. Since $A_i \in \mathcal{B}(\mathbb{R})$ so $E_i \in \mathcal{F}$, thus $\mathbb{P}(E_i)$ is defined. Since X is a function, it is impossible for $X(\omega) = a$ and $X(\omega) = b$ when $a \neq b$, so the E_i's are also disjoint. Then,

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(X(\omega) \in \bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i) = \sum_{i=1}^{\infty} \mu(A_i). \tag{2.5}$$

□

Definition 2.2.2. A random variable X and probability measure $\mathbb{P}$ generate the *cumulative distribution function* (CDF) defined by

$$F(x) = \mu((-\infty, x]) = \mathbb{P}(X \leq x). \tag{2.6}$$

Theorem 2.2.3. A distribution function F for a random variable X uniquely defines the measure μ.

Proof. Outside the scope of this class. □

Example 2.2.4. Uniform. $\mathbb{P}$ is uniform on $[0,1]$ and $X(\omega) = \omega$. For $x \in [0,1]$,

$$F(x) = \mu([0,x)) = \mathbb{P}(X(\omega) \in [0,x)) = \mathbb{P}(\omega \in [0,x)) = x. \tag{2.7}$$

Theorem 2.2.5. A distribution function satisfies the following properties:

i) For all $x \leq y \in \mathbb{R}$, $F(x) \leq F(y)$;
ii) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$;

iii) F is right-continuous.

Proof. For i), if $x \leq y$, then $\{\omega : X(\omega) \leq x\} \subseteq \{\omega : X(\omega) \leq y\}$. Thus, $\mathbb{P}(X(\omega) \leq x) \leq \mathbb{P}(X(\omega) \leq y)$, which gives $F(x) \leq F(y)$.

For ii), suppose $x_n \uparrow \infty$ and $y_n \downarrow -\infty$, so that $\{\omega : X(\omega) < x_n\} \nearrow \{\omega : X(\omega) < \infty\}$ and $\{\omega : X(\omega) < y_n\} \searrow \{\omega : X(\omega) < -\infty\}$. By continuity of measure, $F(x_n) \uparrow \mathbb{P}(X(\omega) < \infty) = 1$ and $F(y_n) \downarrow \mathbb{P}(X(\omega) < -\infty) = 0$. Since these were arbitrary sequences, ii) holds.

For iii), fix an arbitrary $x \in \mathbb{R}$ and suppose $x_n \downarrow x$. By the same logic, $\{\omega : X(\omega) < x_n\} \searrow \{\omega : X(\omega) \leq x\}$, so $F(x_n) \downarrow F(x)$, showing iii). □

Definition 2.2.6. The *inverse CDF* of a random variable X is defined by

$$F^{-1}(y) = \sup\{x : F(x) < y\}. \tag{2.8}$$

Theorem 2.2.7. If F satisfies properties i) to iii), it is the CDF of some random variable.

Proof. Let $U \sim \text{Uniform}(0,1)$ and define the random variable $Y(\omega) = F^{-1}(U(\omega))$. Consider arbitrary $x, t \in [0,1]$.

First, suppose $F^{-1}(t) > x$. That is, $\sup\{y : F(y) < t\} > x$, so $F(x) < t$ since F is non-decreasing.

Next, suppose $F^{-1}(t) \leq x$. By the same logic, $F(x+\delta) \geq t$ for any $\delta > 0$. Since F is right continuous, this gives $F(x) \geq t$.

These combined have given that $\{t : F^{-1}(t) \leq x\} = \{t : t \leq F(x)\}$. Thus,

$$\mathbb{P}(Y \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x), \tag{2.9}$$

so Y is a random variable with the CDF F. □

Definition 2.2.8 (Discrete Random Variable). A random variable X is called discrete if its range is finite or countably infinite, i.e., $X(\omega) \in \{x_i\}_{i=1}^n$ or $X(\omega) \in \{x_i\}_{i=1}^\infty$.

Definition 2.2.9 (Probability Mass Function). Let X be a discrete random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $D \subseteq \mathbb{R}$ be the set of possible values that X can take. The *probability mass function* or p.m.f. of X is defined as

$$\begin{aligned} p_X(x) = \mathbb{P}(X = x) > 0 \text{ if } x \in D, \\ p_X(x) = 0 \text{ if } x \notin D. \end{aligned} \tag{2.10}$$

D is called the support of X.

Lemma 2.2.10. A discrete random variable X with p.m.f. $p_X(x)$ and support D has CDF

$$F_X(x) = \sum_{y \in D : y \leq x} p_X(y). \tag{2.11}$$

Proof. Exercise. □

Proposition 2.2.11. A p.m.f. $p_X(x)$ uniquely defines the distribution of a discrete random variable X.

Proof. Exercise. □

Example 2.2.12 (Bernoulli). A Bernoulli random variable X with probability of success p has p.m.f.

$$\begin{aligned} p_X(x) &= p, \text{ if } x = 1, \\ p_X(x) &= 1 - p, \text{ if } x = 0. \end{aligned} \tag{2.12}$$

We denote X as $X \sim \text{Bernoulli}(p)$.

Example 2.2.13 (Binomial). A Binomial random variable X with probability of success p and number of trials n has p.m.f.

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, \dots, n. \tag{2.13}$$

We denote X as $X \sim \text{Bin}(n, p)$.

Example 2.2.14 (Geometric). A geometric random variable X with probability of success p has p.m.f.

$$p_X(x) = (1-p)^{n-1} p, x \in \mathbb{N}^+. \tag{2.14}$$

We denote X as $X \sim \text{Geo}(p)$.

Example 2.2.15 (Poisson). A Poisson random variable X with rate parameter $\lambda > 0$ has p.m.f.

$$p_X(x) = e^{-\lambda} \frac{\lambda^x}{x!}, x \in \mathbb{N}. \tag{2.15}$$

We denote X as $X \sim \text{Pois}(\lambda)$.

Definition 2.2.16 (Continuous Random Variable). A random variable X is called continuous if its CDF is continuous everywhere.

Definition 2.2.17 (Probability Density Function). A random variable X is called *absolutely continuous* if there exists a function $f : \mathbb{R} \to [0, \infty)$, such that for all $x \in \mathbb{R}$,

$$F(x) = \int_{-\infty}^{x} f(y) \mathrm{d}y. \tag{2.16}$$

f is called the *probability density function* (p.d.f.) of X or simply the density. The support D of X is defined as

$$D = \{x \in \mathbb{R} : f(x) > 0\}. \tag{2.17}$$

Remark. Continuous random variable encompasses absolutely continuous random variables. One example of continuous random variable that is not absolutely continuous (does not have a p.d.f.) is the Cantor distribution. Such distribution is called singular. However, singular random variables are rare, and we will just refer to absolutely continuous random variable as continuous random variables.

Lemma 2.2.18. If X has a density function, $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$.

Proof.

$$\mathbb{P}(X = x) = \lim_{\delta \to 0} \mathbb{P}(x - \delta < X \leq x + \delta) = \lim_{\delta \to 0} \int_{x-\delta}^{x+\delta} f(y) \mathrm{d}y = 0. \tag{2.18}$$

□

Proposition 2.2.19. A p.d.f. $f_X(x)$ uniquely defines the distribution of a continuous random variable X.

Proof. Exercise. □

Example 2.2.20 (Uniform). A uniform random variable on the interval (a, b) with $a < b \in \mathbb{R}$ has p.d.f.

$$f_X(x) = \frac{1}{b - a},\ x \in (a, b). \tag{2.19}$$

We denote X as $X \sim \text{Unif}(a, b)$.

Example 2.2.21 (Beta). A Beta random variable with parameters $a, b > 0$ has p.d.f.

$$f_X(x) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1 - x)^{b-1},\ x \in (0, 1). \tag{2.20}$$

We denote X as $X \sim \text{Beta}(a, b)$.

Example 2.2.22 (Exponential). An exponential random variable with rate parameter $\lambda > 0$ has p.d.f.

$$f_X(x) = \lambda e^{-\lambda x},\ x > 0. \tag{2.21}$$

We denote X as $X \sim \text{Exp}(\lambda)$.

Example 2.2.23 (Gamma). An Gamma random variable with shape parameter $\alpha > 0$ and rate parameter β has p.d.f.

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},\ x > 0. \tag{2.22}$$

We denote X as $X \sim \text{Gamma}(\alpha, \beta)$.

Example 2.2.24 (Normal/Gaussian). A normal or a Gaussian random variable with mean μ and standard deviation σ has p.d.f.

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right),\ x \in \mathbb{R}. \tag{2.23}$$

We denote X as $X \sim \mathcal{N}(\mu, \sigma^2)$.

Theorem 2.2.25 (Change of Variable). Suppose X has continuous p.d.f. f_X with support (a, b), and g is a strictly monotonic and differentiable function. Then $Y = g(X)$ has density,

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|. \tag{2.24}$$

Y has support $(g(a), g(b))$ if g is increasing and $(g(b), g(a))$ if g is decreasing. Note that a, b can take infinities, and we define $g(\pm\infty) = \lim_{x\to\pm\infty} g(x)$.

Proof. Suppose g is increasing, then

$$\begin{aligned}\mathbb{P}(Y \leq y) &= \mathbb{P}(g(X) \leq y) \\ &= \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)),\end{aligned} \tag{2.25}$$

and

$$\begin{aligned}f_Y(y) &= \frac{d}{dy} F_X(g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y).\end{aligned} \tag{2.26}$$

For the support of Y, since $f_X(x) > 0$ if and only if $a < x < b$, and that g is strictly increasing, thus $f_Y(y) > 0$ if and only if $g(a) < y < g(b)$.

The proof when g is decreasing is left as an exercise. □

Definition 2.2.26. A random vector $X = (X_1, \ldots, X_n)$ is a function $X : \Omega \to \mathbb{R}^n$.

Remark. In the above, we have omitted the attention on measurability.

Definition 2.2.27. We define the *joint distribution function* of a *random vector* $X = (X_1, \ldots, X_n)$ by

$$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n), \tag{2.27}$$

using the notation $\mathbb{P}(A, B) \overset{\text{def}}{=} \mathbb{P}(A \cap B)$.

Theorem 2.2.28. A joint distribution function $F : \mathbb{R}^n \to \mathbb{R}$ satisfies the following properties:

i) If $(x_1, \ldots, x_n)$ and $(x_1', \ldots, x_n')$ satisfy $x_i \leq x_i'$ for all $i \in [n]$, $F(x_1, \ldots, x_n) \leq F(x_1', \ldots, x_n')$;
ii) $\lim_{x_i\to-\infty} F(x_1, \ldots, x_n) = 0$ for all $i \in [n]$ and $\lim_{x_1\to\infty,\ldots,x_n\to\infty} F(x_1, \ldots, x_n) = 1$;
iii) $\lim_{h\to0^+} F(x_1 + h, \ldots, x_n) = \cdots = \lim_{h\to0^+} F(x_1, \ldots, x_n + h) = F(x_1, \ldots, x_n)$.

Proof. Analogous to the proof of Theorem 2.2.5. □

Definition 2.2.29. A random vector has a *joint probability mass function* $p : \mathbb{R}^n \to \mathbb{R}_+$ if $\forall (x_1, \ldots, x_n) \in D \subseteq \mathbb{R}^n$

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = p(x_1, \ldots, x_n), \tag{2.28}$$

where D is the support of $(X_1, \ldots, X_n)$.

Definition 2.2.30. A random vector has a *joint density function* $f : \mathbb{R}^n \to \mathbb{R}_+$ if $\forall (x_1, \ldots, x_n) \in \mathbb{R}^n$

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(y_1, \ldots, y_n) \mathrm{d}y_1 \cdots \mathrm{d}y_n. \tag{2.29}$$

Example 2.2.31. A multivariate Gaussian random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and covariance matrix Σ with entry $\Sigma_{i,j} = \rho_{i,j}\sigma_i\sigma_j$ has joint p.d.f.

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-n/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right), \ \boldsymbol{x} \in \mathbb{R}^n. \tag{2.30}$$

$\rho_{i,j}$ is the correlation coefficient of X_i and X_j, σ_i and σ_j are their standard deviation. Specifically, if $n = 2$, then (X, Y) has joint p.d.f.

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right\} \tag{2.31}$$

Lemma 2.2.32 (Marginal distribution). Suppose a random vector has joint p.m.f. p and support D. Suppose that X_i has support D_i, then the marginal p.m.f. of X_i is

$$p_{X_i}(x) = \sum_{x_1 \in D_1} \cdots \sum_{x_{i-1} \in D_{i-1}} \sum_{x_{i+1} \in D_{i+1}} \cdots \sum_{x_n \in D_n} p(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n),\ x \in D_i. \tag{2.32}$$

Proof. Exercise. □

Lemma 2.2.33. Suppose a random vector has joint p.d.f. f and support D. Suppose that X_i has support D_i, then the marginal p.d.f. of X_i is

$$f_{X_i}(x) = \int_{D_1} \cdots \int_{D_{i-1}} \int_{D_{i+1}} \cdots \int_{D_n} f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)\mathrm{d}x_1 \ldots \mathrm{d}x_{i-1}\mathrm{d}x_{i+1} \ldots \mathrm{d}x_n,\ x \in D_i. \tag{2.33}$$

Proof. Exercise. □

Example 2.2.34. For a multivariate Gaussian random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and covariance matrix Σ, the marginal distribution of X_i's are all Gaussian.

Proof. Exercise. You only need to try it for $n = 2$. Higher value of n is too much work. □

2.3 Conditional Distribution

Definition 2.3.1. Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ with joint p.m.f. p and support D. Furthermore, for each random variable X_i, assume its p.m.f. is p_i. The *conditional probability mass function* of $\boldsymbol{X}_{-i} \mid X_i$ is defined as

$$p_{\boldsymbol{X}_{-i}|X_i}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n \mid x_i) = \frac{p(x_1, \ldots, x_n)}{p_i(x_i)},\ \forall (x_1, \ldots, x_n) \in D. \tag{2.34}$$

$\boldsymbol{X}_{-i}$ denotes the random vector without the i-th element, i.e., removing X_i.

Definition 2.3.2. Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ with joint p.d.f. f and support D. Furthermore, for each random variable X_i, assume its p.d.f. is f_i. The *conditional probability density function* of $\boldsymbol{X}_{-i} \mid X_i$ is defined as

$$f_{\boldsymbol{X}_{-i}|X_i}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n \mid x_i) = \frac{f(x_1, \ldots, x_n)}{f_i(x_i)},\ \forall (x_1, \ldots, x_n) \in D. \tag{2.35}$$

Lemma 2.3.3. Suppose (X, Y) is a random vector either with a joint p.m.f. or a joint p.d.f. Their conditional p.m.f. or the conditional p.d.f. are themselves a proper p.m.f. or p.d.f. Thus, $X \mid Y$ and $Y \mid X$ are also random variables.

Proof. Exercise. □

Example 2.3.4. Suppose (X, Y) is jointly Gaussian with $\boldsymbol{\mu} = (\mu_X, \mu_Y)$ is the mean vector, and

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}. \tag{2.36}$$

Then $X \mid Y$ and $Y \mid X$ are also Gaussian random variables.

Proof. Exercise. □

Definition 2.3.5 (Bayes' Theorem). Suppose (X, Y) is a random vector with joint p.d.f. f, and X and Y have p.d.f. f_X and f_Y. Then

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{\int f_{Y|X}(y \mid x) f_Y(y) dy} = \frac{f_{X|Y}(x \mid y) f_Y(y)}{\int f_{X|Y}(x \mid y) f_Y(y) dy}. \tag{2.37}$$

Proof. Obviously,

$$f(x,y) = f_{X|Y}(x \mid y) f_Y(y). \tag{2.38}$$

We just need to show that

$$\int f_{X|Y}(x \mid y) f_Y(y) dy = f_X(x). \tag{2.39}$$

Clearly,

$$\begin{aligned} \int f_{X|Y}(x \mid y) f_Y(y) dy &= \int \frac{f(x,y)}{f_Y(y)} f_Y(y) dy \\ &= \int f(x,y) dy \\ &= f_X(x). \end{aligned} \tag{2.40}$$

□

Example 2.3.6. Suppose $X \mid Y = y \sim \mathcal{N}(y, \sigma^2)$ and $Y \sim \mathcal{N}(\mu, \tau^2)$. Then $Y \mid X = x \sim \mathcal{N}\left(\frac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)$.

Proof. Will do this in class. □

2.4 Exercises

Exercise 2.1. Prove all left over exercises in this chapter.

Exercise 2.2. Consider the joint Gaussian random vector $\boldsymbol{X} = (X_1, \dots, X_n)$. Prove that X_i's are independent if and only if all the off-diagonal entry of the covariance matrix Σ is zero, i.e, they are uncorrelated.

Exercise 2.3. Suppose $X \mid Y = y \sim \text{Bin}(n, y)$ and $Y \sim \text{Beta}(a, b)$ where $n \in \mathbb{N}$ and $a, b \in \mathbb{R}^+$. Find the conditional p.d.f. of $Y \mid X$.

Exercise 2.4. Suppose $X \mid Y = y \sim \text{Gamma}(\alpha, y)$ and $Y \sim \text{Exp}(\lambda)$ where $\alpha, \lambda \in \mathbb{R}^+$. Find the conditional p.d.f. of $Y \mid X$.

Exercise 2.5. Let $X_1 = \text{Unif}(0, 1)$. For all $n \geq 2$, set $X_n \mid X_{n-1} \sim \text{Unif}(X_{n-1}, 1)$. Find the p.d.f. of X_n

Exercise 2.6. Suppose $\mathbb{P}(Z = 0) = \mathbb{P}(Z = 1) = 1/2$, $Y \sim \mathcal{N}(0, 1)$, and that Y and Z are independent. Let $X = YZ$. What is the CDF of X?

Exercise 2.7. Suppose $\mathbb{P}(Z = 1) = \mathbb{P}(Z = -1) = 1/2$, $Y \sim \mathcal{N}(0, 1)$, and that Y and Z are independent. Let $X = YZ$.

a) Prove that $X \sim \mathcal{N}(0, 1)$.
b) Prove that $\mathbb{P}(|X| = |Y|) = 1$.
c) Prove that X and Y are not independent.

Exercise 2.8. Let X be a random variable such that $\mathbb{P}(X > 0) > 0$. Prove that there exists a $\delta > 0$ such that $\mathbb{P}(X \geq \delta) > 0$.

Exercise 2.9. If X and Y are independent on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then $f(X)$ and $g(Y)$ are independent on $(\Omega, \mathcal{F}, \mathbb{P})$ for any functions $f, g : \mathbb{R} \to \mathbb{R}$ (You can assume that f, g are measurable, i.e., $\{x \in \mathbb{R} : f(x) \in A\} \in \mathcal{B}(\mathbb{R})$ for any $A \in \mathcal{B}(\mathbb{R})$).

Exercise 2.10. Let $\mathbb{P}$ be the uniform measure on $[0, 1]$. Define $A = (a, b)$ and $B = (c, d)$, with $a < c$. State necessary and sufficient conditions for A and B to be independent.

Exercise 2.11. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $A, B \in \mathcal{F}$ are independent. Define $X(\omega) = \mathbb{I}\{\omega \in A\}$ and $Y(\omega) = \mathbb{I}\{\omega \in B\}$. Show that X and Y are random variables and they are independent.

Exercise 2.12. Review the exponential family of distributions (note this is not just the exponential distribution, but the exponential *family*).

LECTURE 3

CONVERGENCE

3.1 Limit Events

Definition 3.1.1. Consider a sequence $A_1, A_2, \ldots \subseteq \Omega$. Define the *limit events* by

$$\limsup_{n\longrightarrow\infty} A_n = \{A_n \text{ i.o.}\} = \bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k \tag{3.1}$$

and

$$\liminf_{n\longrightarrow\infty} A_n = \{A_n \text{ a.a.}\} = \bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k. \tag{3.2}$$

Corollary 3.1.2. $\mathbb{P}(A_n \text{ i.o.}) = 1 - \mathbb{P}(A_n^c \text{ a.a.})$.

Remark. We require the introduction and discussion of limit events to talk about convergence of random variables later. To better understand the limit events, we use $\{A_n \text{ i.o.}\}$ as an example.

For the sequence of events $\{A_n\}$ to occur infinitely often (i.o.), it means that for any $n \in \mathbb{N}^+$, there exists a $k \geq n$ such that A_k occurs. To translate the above statement to set operations, notice the quantifier "for any" and "there exists" have one-to-one correspondence to intersection and union. Thus, the statement "for any $n \in \mathbb{N}^+$" translates to "$\bigcap_{n=1}^{\infty}$" as it needs to happen for all $n \in \mathbb{N}^+$, while "there exists $k \geq n$ such that A_k occurs" translates to "$\bigcup_{k=n}^{\infty} A_k$" since as long as one A_k occurs where $k \geq n$ would suffice. Combining things together, we have $\{A_n \text{ i.o.}\} = \bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k$.

Additionally, lim sup and lim inf are used here for set operations in a similar fashion as for real sequence. Notice that $\bigcup_{k=n}^{\infty} A_k$ is the biggest set we can construct using $A_n, A_{n+1}, \ldots$, so in a sense, it is the "supremum" of the sets $A_n, A_{n+1}, \ldots$. Now since $\{\bigcup_{k=n}^{\infty} A_k\}_{n=1}^{\infty}$ is a decreasing sequence of events, its limit defined as $\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k$ always exists. This form of almost one-to-one similarity with lim sup for real sequence is thus carried over and used for set operations.

Proposition 3.1.3.

$$\mathbb{P}(A_n \text{ a.a.}) \leq \liminf_{n\longrightarrow\infty} \mathbb{P}(A_n) \leq \limsup_{n\longrightarrow\infty} \mathbb{P}(A_n) \leq \mathbb{P}(A_n \text{ i.o.}). \tag{3.3}$$

Proof. Observe that $\bigcap_{k=n}^{\infty} A_k \subseteq \bigcap_{k=n+1}^{\infty} A_k$ for all n. So,

$$\begin{aligned}\mathbb{P}(A_n \text{ a.a.}) &= P\left(\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k\right)\\ &= \lim_{n\to\infty} P\left(\bigcap_{k=n}^{\infty} A_k\right)\\ &= \liminf_{n\longrightarrow\infty} P\left(\bigcap_{k=n}^{\infty} A_k\right)\\ &\leq \liminf_{n\longrightarrow\infty} P\left(A_n\right).\end{aligned} \tag{3.4}$$

The second inequality is by definition. The third inequality is an exercise. □

Theorem 3.1.4 (Borel-Cantelli Lemma).

i) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$;

ii) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and $\{A_n\}$ are independent, then $\mathbb{P}(A_n \text{ i.o.}) = 1$.

Proof. For i), observe that $\bigcup_{k=n+1}^{\infty} A_k \subseteq \bigcup_{k=n}^{\infty} A_k$ for all n. Thus,

$$\begin{aligned} \mathbb{P}(A_n \text{ i.o.}) &= \mathbb{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k\right) \\ &= \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{k=n}^{\infty} A_k\right) \\ &\leq \lim_{n\to\infty} \sum_{k=n}^{\infty} \mathbb{P}(A_k) \\ &= 0. \end{aligned} \tag{3.5}$$

For ii), observe that $\bigcap_{k=n}^{\infty} A_k^c \subseteq \bigcap_{k=n+1}^{\infty} A_k^c$ for all n. Thus,

$$\begin{aligned} 1 - \mathbb{P}(A_n \text{ i.o.}) &= \mathbb{P}((A_n \text{ i.o.})^c) \\ &= \mathbb{P}\left(\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k^c\right) \\ &= \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{k=n}^{\infty} A_k^c\right) \\ &= \lim_{n\to\infty} \prod_{k=n}^{\infty} [1 - \mathbb{P}(A_k)] \\ &\leq \lim_{n\to\infty} \prod_{k=n}^{\infty} e^{-\mathbb{P}(A_k)} \\ &= \lim_{n\to\infty} e^{-\sum_{k=n}^{\infty} \mathbb{P}(A_k)} \\ &= \lim_{n\to\infty} 0 \\ &= 0. \end{aligned} \tag{3.6}$$

□

Example 3.1.5 (converse does not hold for i)). Uniform measure and $A_n = [0, 1/n]$. Then,

$$\begin{aligned} A_n \text{ i.o.} &= \bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} [0, 1/k] \\ &= \{0\}, \end{aligned} \tag{3.7}$$

but $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} 1/n = \infty$.

Example 3.1.6 (independence is needed for ii)). Define $c_1, c_2, \dots$ such that $c_i = 1$ if a fair coin toss lands with head for all $i \in \mathbb{N}^+$. Let $A_1, A_2, \dots$ be such that $A_i = \{c_1 = 1\}$. Then, $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} 1/2 = \infty$ and $\mathbb{P}(A_n \text{ i.o.}) = \mathbb{P}(A_n) = 1/2$.

3.2 Types of Convergence

Definition 3.2.1. A sequence of random variables X_n *converges almost surely* to X if

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1, \tag{3.8}$$

and is denoted by $X_n \longrightarrow X$ a.s.

Proposition 3.2.2. If for all $\varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon \text{ i.o.}) = 0$, then $X_n \longrightarrow X$ a.s.

Proof. Consider that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ if and only if for all $\varepsilon > 0$, $|X_n(\omega) - X(\omega)| < \varepsilon$ for all but finitely many n. Thus,

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = \mathbb{P}(\forall\varepsilon > 0, |X_n(\omega) - X(\omega)| < \varepsilon \text{ a.a.}) = 1 - \mathbb{P}(\exists\varepsilon > 0, |X_n(\omega) - X(\omega)| \geq \varepsilon \text{ i.o.}). \tag{3.9}$$

Next, if $\exists\varepsilon > 0, |X_n(\omega) - X(\omega)| \geq \varepsilon$ i.o., then it implies $\exists\varepsilon \in \mathbb{Q}_+, |X_n(\omega) - X(\omega)| \geq \varepsilon$ i.o., thus,

$$\begin{aligned}\mathbb{P}(\exists\varepsilon > 0, |X_n(\omega) - X(\omega)| \geq \varepsilon \text{ i.o.}) &\leq \mathbb{P}(\exists\varepsilon \in \mathbb{Q}_+, |X_n(\omega) - X(\omega)| \geq \varepsilon \text{ i.o.}) \\ &\leq \sum_{\varepsilon\in\mathbb{Q}_+} \mathbb{P}(|X_n(\omega) - X(\omega)| \geq \varepsilon \text{ i.o.}) \\ &= 0.\end{aligned} \tag{3.10}$$

□

Corollary 3.2.3. If for all $\varepsilon > 0$, $\sum_{n=1}^{\infty} \mathbb{P}(|X_n(\omega) - X(\omega)| \geq \varepsilon) < \infty$, $X_n \longrightarrow X$ a.s.

Proof. Borel-Cantelli combined with the assumption implies the hypothesis of Proposition 3.2.2. □

Definition 3.2.4. A sequence of random variables X_n *converges in probability* to X if for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \leq \varepsilon) = 1, \tag{3.11}$$

and is denoted by $X_n \xrightarrow{P} X$.

Proposition 3.2.5. If $X_n \longrightarrow X$ a.s. then $X_n \xrightarrow{P} X$.

Proof. Fix $\varepsilon > 0$ and let $E_n = \{\omega : \exists m \geq n, |X_m(\omega) - X(\omega)| \geq \varepsilon\}$. Observe that $E_{n+1} \subseteq E_n$, and if $\omega \in \bigcap_{n=1}^{\infty} E_n$ then $X_n(\omega) \not\to X(\omega)$. Thus, using continuity of probability,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(|X_n(\omega) - X(\omega)| \geq \varepsilon) &\leq \lim_{n\to\infty} \mathbb{P}(E_n) \\ &= \mathbb{P}\left(\bigcap_{n=1}^{\infty} E_n\right) \\ &\leq \mathbb{P}(X_n \not\to X) \\ &= 0.\end{aligned} \tag{3.12}$$

□

Remark. Convergence almost surely is a much stronger mode of convergence than convergence in probability. The intuition behind it is that for X_n to converge to X almost surely, we require a *fixed* event $N \subseteq \Omega$ with $\mathbb{P}(N) = 0$ such that $X_n(\omega) \to X(\omega)$ for any $\omega \notin N$. However, for X_n to converge to X in probability, there is no requirement for such a set N, we only require the set N_n on which X_n and X differ by more than ϵ satisfies $\mathbb{P}(N_n) \to 0$ as $n \to \infty$ for any ϵ. Note that N_n may change with n and need not be fixed.

Example 3.2.6. X_n independent with $\mathbb{P}(X_n = 1) = 1/n$, $\mathbb{P}(X_n = 0) = 1 - 1/n$. For all $\varepsilon > 0$,

$$\mathbb{P}(X_n > \varepsilon) = 1/n \longrightarrow 0. \tag{3.13}$$

So, $X_n \xrightarrow{P} 0$. But, $P(X_n = 1 \text{ i.o.}) = 1$, so $P(X_n \longrightarrow 0) = 0$.

Theorem 3.2.7. If $X_n \xrightarrow{P} X$, there exists a subsequence such that $X_{n_k} \longrightarrow X$ a.s.

Proof. By definition, for each $k \in \mathbb{N}$, there exists n_k such that for $n \geq n_k$

$$\mathbb{P}(|X_n - X| > 2^{-k}) \leq 2^{-k}. \tag{3.14}$$

Further, choose these such that $n_{k+1} \geq n_k$, and define the sets

$$A_k = \{\omega : |X_{n_k}(\omega) - X(\omega)| > 2^{-k}\}. \tag{3.15}$$

Clearly,

$$\sum_{k=1}^{\infty} \mathbb{P}(A_k) \leq \sum_{k=1}^{\infty} 2^{-k} < \infty, \tag{3.16}$$

so by Borel Cantelli $\mathbb{P}(A_k \text{ i.o.}) = 0$. Finally, observe that $|X_{n_k}(\omega) - X(\omega)| > 2^{-k}$ only finitely many times implies $X_{n_k}(\omega) \longrightarrow X(\omega)$, so

$$1 = \mathbb{P}[(A_k \text{ i.o.})^c] \leq \mathbb{P}(X_{n_k} \longrightarrow X). \tag{3.17}$$

□

Theorem 3.2.8 (Continuous Mapping Theorem). If f is a continuous function then

i) $X_n \longrightarrow X$ a.s. implies that $f(X_n) \longrightarrow f(X)$ a.s.;

ii) $X_n \xrightarrow{P} X$ implies that $f(X_n) \xrightarrow{P} f(X)$;

Proof.

i) f continuous means that $X_n(\omega) \to X(\omega)$ implies $f(X_n(\omega)) \to f(X(\omega))$, so

$$1 = \mathbb{P}(X_n \to X) \leq \mathbb{P}(f(X_n) \to f(X)); \tag{3.18}$$

ii) Exercise.
For all $\varepsilon > 0$, there exists $\delta > 0$ such that $|X_n(\omega) - X(\omega)| \leq \delta$ implies $|f(X_n(\omega)) - f(X(\omega))| \leq \varepsilon$, but

$$1 = \lim_{n\to\infty} \mathbb{P}(|X_n - X| \leq \delta) \leq \lim_{n\to\infty} \mathbb{P}(|f(X_n) - f(X)| \leq \varepsilon); \tag{3.19}$$

□

3.3 Exercises

Exercise 3.1. Prove all left over exercises in this chapter.

Exercise 3.2. Consider $\Omega = \{a, b, c\}$ with the measure $\mathbb{P}(a) = \mathbb{P}(b) = \mathbb{P}(c) = 1/3$. Find examples of $A_n \subseteq \Omega$ such that the inequalities in Proposition 3.1.3 are strict.

Exercise 3.3. Prove that for any collections $\{A_n\}$ and $\{B_n\}$,

$$\limsup(A_n \cap B_n) \subseteq \limsup A_n \cap \limsup B_n, \tag{3.20}$$

and find example where the inclusion is strict and where it is equality.

Exercise 3.4. Find an example of $\mathbb{P}$ and A_n such that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ but $\mathbb{P}(A_n \text{ i.o.}) = 1$.

Exercise 3.5. Prove that if $X_n \longrightarrow X$ a.s., for all $\varepsilon > 0$

$$\mathbb{P}(|X_n - X| \geq \varepsilon \text{ i.o.}) = 0. \tag{3.21}$$

Exercise 3.6. Find a sequence X_n and X such that $X_n \xrightarrow{P} X$ but $X_n \not\to X$ a.s.

Exercise 3.7. Show that $X_n \longrightarrow X$ a.s. or $X_n \xrightarrow{P} X$ if and only if $(X_n - X) \longrightarrow 0$ a.s. or $(X_n - X) \xrightarrow{P} 0$ respectively.

Exercise 3.8. Show that if $X_n - a_n \xrightarrow{P} 0$ and $a_n \to a$, then $X_n \xrightarrow{P} a$.

Exercise 3.9. If $X_n \xrightarrow{P} X$ and $X_n \leq X_{n+1}$ for all n, then $X_n \to X$ a.s.

Exercise 3.10. Let $\delta, \epsilon > 0$, and let $X_1, X_2, \ldots$ be a sequence of independent non-negative random variables such that $\mathbb{P}(X_i \geq \delta) \geq \epsilon$ for all i. Prove that $\sum_{i=1}^{\infty} X_i = \infty$ a.s.

LECTURE 4

EXPECTATION

This chapter reviews the basic definitions and certain properties of expectation. We then discuss the scenarios under which we can exchange the operations of limit and expectation. However, rigorous treatment on the subject requires building things from the ground up and involves quite a bit of abstract integration theory from measure theory. Thus, many results presented in this chapter will be taken as granted and not proven.

4.1 Basic Definition and Properties

Definition 4.1.1 (Expectation for Discrete Random Variables). If X is a discrete random variable with support being a countable set $\mathcal{X} = \{x_1, x_2, \dots\}$ with p.m.f. $p_X(x_i) = \mathbb{P}(X = x_i)$, the *expectation* of X is then

$$\mathbb{E}[X] = \sum_{x_i \in \mathcal{X}} x_i p_X(x_i). \tag{4.1}$$

Example 4.1.2. For an arbitrary set $A \subset \Omega$, let $Y(\omega) = \mathbb{I}\{\omega \in A\}$. Then, $\mathbb{E}[Y] = \mathbb{P}(A)$.

Example 4.1.3. $\mathbb{P}$ is uniform measure on $\Omega = [0, 1]$, and

$$X(\omega) = \begin{cases} 5, & \omega > 1/3 \\ 3, & \omega \le 1/3 \end{cases}. \tag{4.2}$$

Definition 4.1.4 (Expectation for Continuous Random Variables). If X is a continuous random variable with p.d.f. $f_X(x)$, the *expectation* of X is then

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x) \mathrm{d}x. \tag{4.3}$$

Definition 4.1.5. A random variable X is called *integrable* if $\mathbb{E}[|X|] < \infty$.

Remark. Expectation of a random variable need not be finite. For example, we can consider a random variable X with $\mathbb{P}(X = x) = 6/(\pi^2 n^2)$ if $x \in \mathbb{N}^+$ and 0 otherwise. The expectation of X in this case is ∞.

Example 4.1.6. A Cauchy$(0, 1)$ random variable X has the following density

$$f(x) = \frac{1}{\pi(1 + x^2)}, \ x \in \mathbb{R}. \tag{4.4}$$

X is not integrable.

Proposition 4.1.7. If X and Y are random variables, the following properties hold.

i) If $X \geq 0$ a.s., $\mathbb{E}[X] \geq 0$;
ii) For all $a \in \mathbb{R}$, $\mathbb{E}[aX] = a\mathbb{E}[X]$;
iii) $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Proof. We will take these as granted. □

Lemma 4.1.8. If properties i) to iii) hold, the following properties also hold.

iv) If $X \leq Y$ a.s. then $\mathbb{E}X \leq \mathbb{E}Y$;
v) If $X = Y$ a.s. then $\mathbb{E}X = \mathbb{E}Y$;
vi) $|\mathbb{E}X| \leq \mathbb{E}\,|X|$.

Proof. iv) $X \leq Y$ a.s. $\implies Y - X \geq 0$ a.s. $\implies \mathbb{E}(Y-X) \geq 0 \implies \mathbb{E}Y \geq \mathbb{E}X$;
v) Exercise.
vi) Exercise.

□

Definition 4.1.9. Let X be a random variable, the *variance* is defined as

$$\operatorname{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)^2]. \tag{4.5}$$

The *standard deviation* of X is $\sqrt{\operatorname{Var}(X)}$.

Theorem 4.1.10. Let g be a measurable function, X be a random variable. Then

$$\mathbb{E}[g(X)] = \sum_{x_i \in \mathcal{X}} g(x_i) p_X(x_i), \tag{4.6}$$

if X is discrete, while

$$\mathbb{E}[g(X)] = \int_R g(x) f_X(x) dx, \tag{4.7}$$

if X is continuous.

Proof. We will take this as granted. □

Remark. In general, $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$. Expectation is a *linear* operation since, in essence, expectation is integration. Thus, expectation is not closed under any non-linear operation g.

Example 4.1.11. Let X be a random variable, and $g(x) = (x - \mathbb{E}X)^2$, then $\mathbb{E}[g(X)] = \operatorname{Var}(X)$.

Theorem 4.1.12. If X and Y are independent with $\mathbb{E}X, \mathbb{E}Y < \infty$, then $\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$.

Proof. We will take this as granted. □

Definition 4.1.13. The *variance* of a random variable X is defined as $\operatorname{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)^2]$. The *covariance* of random variables X and Y is $\operatorname{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$. The *correlation* between random variables X and Y is $\operatorname{Corr}(X,Y) = \operatorname{Cov}(X,Y)/\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}$. If $\operatorname{Cov}(X,Y) = 0$, then X, Y are said to be uncorrelated.

Lemma 4.1.14. $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Proof. Exercise. □

Remark. Independence of two random variables implies no correlation. However, no correlation does NOT imply independence.

4.2 Conditional Expectation

Definition 4.2.1. Given random variables X and Y, and the conditional p.m.f. or p.d.f. of $X \mid Y = y$ is $p_{X|Y}(x \mid y)$ or $f_{X|Y}(x \mid y)$. The *conditional expectation* of $X \mid Y = y$ is

$$\mathbb{E}[X \mid Y = y] = \sum_i x_i p_{X|Y}(x_i \mid y) \tag{4.8}$$

if $X \mid Y = y$ is discrete, and

$$\mathbb{E}[X \mid Y = y] = \int x f_{X|Y}(x \mid y) \mathrm{d}x \tag{4.9}$$

if $X \mid Y = y$ is continuous.

Remark. In general, we may not be only interested in the conditional expectation $\mathbb{E}[X \mid Y = y]$ for a fixed y. Usually, we are more interested in a much more vague conditioning statement, such as $Y \in A$ for some $A \subseteq \mathbb{R}$ instead of $Y = y$. In the most general case, we are interested in $\mathbb{E}[X \mid Y]$. This quantity tells us what the expectation of X is given the information of Y, without the need to specify what that information exactly is. However, how do we make sense of $\mathbb{E}[X \mid Y]$ without conditioning on the specific value that Y takes? To this end, notice the definition of conditional expectation in 4.2.1, we can see that if we change the value of y, the resulting conditional expectation will also change. Specifically, we can write

$$\mathbb{E}[X \mid Y = y] = \mathbb{E}[X \mid \{\omega \in \Omega : Y(\omega) = y\}]. \tag{4.10}$$

The ultimate reason that y changes is the change of ω. This means that the value of $\mathbb{E}[X \mid Y] = \mathbb{E}[X \mid Y(\omega)]$ changes with $\omega \in \Omega$. Thus, the general quantity $\mathbb{E}[X \mid Y]$ that we are the most interested is in fact a function that maps from Ω to $\mathbb{R}$. Therefore, $\mathbb{E}[X \mid Y]$ is nothing more than a *random variable*!

Lemma 4.2.2. Given random variables X and Y, and the conditional p.m.f. or p.d.f. of $X \mid Y = y$ is $p_{X|Y}(x \mid y)$ or $f_{X|Y}(x \mid y)$. The *conditional expectation* of $X \mid Y$ is a random variable and

$$\mathbb{E}[X \mid Y] = \sum_i x_i p_{X|Y}(x_i \mid Y) \tag{4.11}$$

if $X \mid Y$ is discrete, and

$$\mathbb{E}[X \mid Y] = \int x f_{X|Y}(x \mid Y) \mathrm{d}x \tag{4.12}$$

if $X \mid Y$ is continuous.

Theorem 4.2.3 (Tower property/Law of total expectation). Given random variables X and Y with $\mathbb{E}[X] < \infty$, then

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]. \tag{4.13}$$

Proof.

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X \mid Y]] &= \int \mathbb{E}[X \mid Y=y] f_Y(y)\mathrm{d}y \\ &= \int\int x f_{X|Y}(x \mid y)\mathrm{d}x f_Y(y)\mathrm{d}y \\ &= \int\int x \frac{f(x,y)}{f_Y(y)} f_Y(y)\mathrm{d}x\mathrm{d}y \\ &= \int\int x f(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int x \int f(x,y)\mathrm{d}y\mathrm{d}x \\ &= \int x f_X(x)\mathrm{d}x = \mathbb{E}[X].\end{aligned} \tag{4.14}$$

□

Theorem 4.2.4 (Law of total variance). Given random variables X and Y with $\mathrm{Var}(X) < \infty$, then

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]). \tag{4.15}$$

where

$$\mathrm{Var}(X \mid Y) = \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2 \mid Y]. \tag{4.16}$$

Proof. Exercise. □

Proposition 4.2.5 (Convolution Formula). If X and Y are independent with densities f_X and f_Y, for all $z \in \mathbb{R}$,

$$\mathbb{P}(X + Y \leq z) = \int F_X(z-y) f_Y(y) dy. \tag{4.17}$$

Proof.

$$\begin{aligned}\mathbb{P}(X + Y \leq z) &= \mathbb{E}[\mathbb{I}(X \leq z - Y)] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{I}(X \leq z - Y) \mid Y]] \\ &= \int \mathbb{E}[\mathbb{I}(X \leq z - y) \mid Y = y] f_Y(y)\mathrm{d}y \\ &= \int \mathbb{E}[\mathbb{I}(X \leq z - y)] f_Y(y)\mathrm{d}y \\ &= \int \mathbb{P}(X \leq z - y) f_Y(y)\mathrm{d}y \\ &= \int F_X(z-y) f_Y(y)\mathrm{d}y\end{aligned} \tag{4.18}$$

□

4.3 Limit Theorems

We now discuss several important theorems that provide conditions on when we can exchange the operations of limit and integration. In essence, if we are given certain mode of convergence of a sequence of random variables, e.g. $X_n \to X$ a.s., do we have $\mathbb{E}X_n \to \mathbb{E}X$? As a cautionary note, the convergence of random variables usually DOES NOT translate to convergence in expectation. However, under certain conditions, convergence in expectation does hold.

Example 4.3.1. Let $U \sim \text{Unif}(0,1)$ and $X_1, X_2, \dots$ be a sequence of random variables s.t. $X_n(\omega) = n\mathbb{I}(U(\omega) \in (0, 1/n))$. We have $X_n \to X = 0$ a.s. But $\lim_{n\to\infty} \mathbb{E}[X_n] = 1 \neq \mathbb{E}[X] = 0$.

Definition 4.3.2. A random variable X is called *bounded* if there is a $M < \infty$ s.t. $|X| \leq M$ a.s.

Proposition 4.3.3. Suppose $X \geq 0$ a.s., then

$$\sup\{\mathbb{E}Y : Y \text{ bounded}, 0 \leq Y \leq X \text{ a.s.}\} = \mathbb{E}X. \tag{4.19}$$

Proof. We will take this as granted. □

Lemma 4.3.4. Let $X \geq 0$ a.s., and recall notation $a \wedge b = \min\{a, b\}$. Then,

$$\lim_{n\to\infty} \mathbb{E}(X \wedge n) = \mathbb{E}X. \tag{4.20}$$

Proof. Define $X_n = X \wedge n$. Clearly $X_n \leq X$, so $\mathbb{E}X_n \leq \mathbb{E}X$. Also $\mathbb{E}X_n \leq \mathbb{E}X_{n+1}$ for all n, so the limit exists, and thus $\lim_{n\to\infty} \mathbb{E}X_n \leq \mathbb{E}X$. Consider a bounded Y such that $0 \leq Y \leq X$ a.s. Then, for large n, $\mathbb{E}X_n \geq \mathbb{E}Y$, so $\lim_{n\to\infty} \mathbb{E}X_n \geq \sup\{\mathbb{E}Y : Y \text{ bounded}, 0 \leq Y \leq X \text{ a.s.}\} = \mathbb{E}X$. □

Theorem 4.3.5 (Bounded Convergence Theorem). Suppose that $|X_n| \leq M$ a.s. and $X_n \xrightarrow{P} X$. Then, $\mathbb{E}X = \lim_{n\to\infty} \mathbb{E}X_n$.

Proof. Fix $\varepsilon > 0$ and define $G_n = \{|X_n - X| > \varepsilon\}$. Then,

$$\begin{aligned}
|\mathbb{E}X_n - \mathbb{E}X| &= |\mathbb{E}(X_n - X)| \\
&\leq \mathbb{E}\,|X_n - X| \\
&= \mathbb{E}\left[|X_n - X|\,\mathbb{I}_{G_n}\right] + \mathbb{E}\left[|X_n - X|\,\mathbb{I}_{G_n^c}\right] \\
&\leq 2M\mathbb{P}(G_n) + \varepsilon[1 - \mathbb{P}(G_n)] \\
&= \varepsilon + \mathbb{P}(G_n)[2M - \varepsilon].
\end{aligned} \tag{4.21}$$

By convergence in probability and arbitrary ε, $\lim_{n\to\infty} |\mathbb{E}X_n - \mathbb{E}X| = 0$. Real analysis fact that this implies the result. □

Theorem 4.3.6 (Fatou's Lemma). If $X_n \geq 0$ a.s. for all n, $\liminf_{n\longrightarrow\infty} \mathbb{E}X_n \geq \mathbb{E}(\liminf_{n\longrightarrow\infty} X_n)$.

Proof. For each n, define $Y_n = \inf_{m\geq n} X_m$. Clearly, $X_n \geq Y_n$ a.s., thus, $\liminf_{n\longrightarrow\infty} \mathbb{E}X_n \geq \liminf_{n\longrightarrow\infty} \mathbb{E}Y_n$. Moreover, $Y_n \uparrow \liminf_{n\longrightarrow\infty} X_n = Y$ a.s., so fix an arbitrary M, and observe $(Y_n \wedge M) \longrightarrow (Y \wedge M)$ a.s., so by BCT we have $\liminf_{n\longrightarrow\infty} \mathbb{E}Y_n \geq \lim_{n\to\infty} \mathbb{E}(Y_n \wedge M) = \mathbb{E}(Y \wedge M)$. Taking the limit as $M \longrightarrow \infty$ and applying Lemma 4.3.4 gives the result. □

Theorem 4.3.7 (Monotone Convergence Theorem). If $X_n \geq 0$ a.s. and $X_n \uparrow X$ a.s., $\mathbb{E}X_n \uparrow \mathbb{E}X$.

Proof. Since $\mathbb{E}X_n \leq \mathbb{E}X$, $\lim_{n\to\infty} \mathbb{E}X_n \leq \mathbb{E}X$. But, by Fatou's,

$$\lim_{n\to\infty} \mathbb{E}X_n = \liminf_{n\longrightarrow\infty} \mathbb{E}X_n \geq \mathbb{E}(\liminf_{n\longrightarrow\infty} X_n) = \mathbb{E}X. \tag{4.22}$$

□

Example 4.3.8. If $X_1, X_2, \dots$ are a sequence of non-negative random variables, then $\mathbb{E}[\sum_{n=1}^{\infty} X_n] = \sum_{n=1}^{\infty} \mathbb{E}[X_n]$.

Theorem 4.3.9 (Dominated Convergence Theorem). If $X_n \longrightarrow X$ a.s. and $|X_n| \leq Y$ a.s. for some integrable Y, $\mathbb{E}X = \lim_{n\to\infty} \mathbb{E}X_n$.

Proof. Since $|X_n| \leq Y$ a.s. for all n, $|X| \leq Y$ a.s. and thus X is integrable. Since $X_n + Y \geq 0$ a.s. for all n, by Fatou's $\liminf_{n\longrightarrow\infty} \mathbb{E}(X_n + Y) \geq \mathbb{E}(\liminf_{n\longrightarrow\infty} X_n + Y) = \mathbb{E}(X + Y)$. Thus, $\liminf_{n\longrightarrow\infty} \mathbb{E}X_n \geq \mathbb{E}X$. It also holds that $Y - X_n \geq 0$, so $\liminf_{n\longrightarrow\infty} \mathbb{E}(Y - X_n) \geq \mathbb{E}(\liminf_{n\longrightarrow\infty} Y - X_n) = \mathbb{E}(Y - X)$. Rearranging (and using that the expectations are all finite) gives $\mathbb{E}X \geq \limsup_{n\longrightarrow\infty} \mathbb{E}X_n$, so the result holds. □

4.4 Moment-Generating Functions

Definition 4.4.1. The *moment-generating function* (M.G.F.) of a random variable X is defined by

$$M_X(\lambda) = \mathbb{E}[e^{\lambda X}], \tag{4.23}$$

for any $\lambda \in \mathbb{R}$ such that $\mathbb{E}[e^{\lambda X}] < \infty$.

Theorem 4.4.2 (Moment generating property of M.G.F). For a random variable X, suppose its M.G.F. $M_X(\lambda) < \infty$ for all $\lambda \in (-\delta, \delta)$ for some $\delta > 0$, then $M_X^{(n)}(0) = \mathbb{E}[X^n]$, and $\mathbb{E}[|X^n|] < \infty$ for all n.

Proof. Let $\lambda \in (0, \delta)$. Using the inequality

$$e^{\lambda|x|} \leq e^{\lambda x} + e^{-\lambda x}, \tag{4.24}$$

we have

$$\mathbb{E}[e^{\lambda|X|}] \leq M_X(\lambda) + M_X(-\lambda) < \infty. \tag{4.25}$$

By Talor expansion, we have

$$e^{\lambda|X|} = \sum_{n=0}^{\infty} \frac{\lambda^n |X|^n}{n!}. \tag{4.26}$$

Let $S_k = \sum_{n=0}^{k} \frac{\lambda^n|X|^n}{n!}$, we thus have $S_k \uparrow e^{\lambda|X|}$ a.s., and since $|S_k| = S_k \leq e^{\lambda|X|}$, by DCT or MCT, we have

$$\mathbb{E}[e^{\lambda|X|}] = \lim_{k\to\infty} \mathbb{E}[S_k] = \lim_{k\to\infty} \mathbb{E}\left[\sum_{n=0}^{k} \frac{\lambda^n|X|^n}{n!}\right] = \sum_{n=0}^{\infty} \frac{\lambda^n \mathbb{E}[|X|^n]}{n!} < \infty. \tag{4.27}$$

This means that $\mathbb{E}[|X^n|] < \infty$ for all n. Furthermore, since

$$\mathbb{E}[X^n] \leq |\mathbb{E}[X^n]| \leq \mathbb{E}[|X^n|], \tag{4.28}$$

the series $\sum_{n=0}^{\infty} \frac{\lambda^n \mathbb{E}[X^n]}{n!}$ is absolutely convergent for any $\lambda \in (-\delta, \delta)$. Let $S_k' = \sum_{n=0}^{k} \frac{\lambda^n X^n}{n!}$ and we have

$$|M_X(\lambda) - \mathbb{E}[S_k']| \leq \sum_{n=k+1}^{\infty} \frac{\lambda^n \mathbb{E}[|X|^n]}{n!} \to 0, \text{ as } k \to \infty. \tag{4.29}$$

Thus,

$$M_X(\lambda) = \sum_{n=0}^{\infty} \frac{\lambda^n \mathbb{E}[X^n]}{n!}. \tag{4.30}$$

Take the m-th derivative of $M_X(\lambda)$ using the previous expansion for $\lambda \in (-\delta, \delta)$ and set $\lambda = 0$, we have

$$M_X^{(m)}(0) = \mathbb{E}[X^m]. \tag{4.31}$$

□

Theorem 4.4.3. For random variables X and Y, suppose $M_X(\lambda) = M_Y(\lambda) < \infty$ for all $\lambda \in (-\delta, \delta)$ and some $\delta > 0$, then X and Y have the same distribution.

Proof. Out of scope. □

Lemma 4.4.4. Let $X_1, \ldots, X_n$ be independent. Denote $S = \sum_{i=1}^n X_i$. If $M_{X_i}(\lambda) < \infty$ for all $\lambda \in (-\delta, \delta)$ for some $\delta > 0$ and all $i \in \{1, \ldots, n\}$, then

$$M_S(\lambda) = \prod_{i=1}^n M_{X_i}(\lambda). \tag{4.32}$$

Proof. Exercise. □

4.5 Exercises

Exercise 4.1. Prove all left over exercises in this chapter.

Exercise 4.2. Give an example of Gaussian random variables X and Y such that they are uncorrelated, i.e., $\operatorname{Cov}(X, Y) = 0$, but not independent.

Exercise 4.3. Compute the expected value and variance for all random variables listed in Chapter 2.

Exercise 4.4. Compute the M.G.F. for all random variables listed in Chapter 2.

Exercise 4.5. Show that the sum of i.i.d.Poisson random variables is a Poisson random variable.

Exercise 4.6. Show that the sum of i.i.d.Exponential random variables is a Gamma random variable.

Exercise 4.7. Show that the sum of independent Gaussian random variables is still Gaussian. Specifically, if $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, then $\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2)$.

Exercise 4.8. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. Let $\chi^2 = \sum_{i=1}^n X_i^2$. Show that $\chi^2 \sim \operatorname{Gamma}(n/2, 1/2)$. χ^2 is also called the Chi-squared distribution with n degrees of freedom.

Exercise 4.9. Show that the following properties hold:

i) $\operatorname{Var}(X) \geq 0$,
ii) For any $c \in \mathbb{R}$, $\operatorname{Var}(cX) = c^2 \operatorname{Var}(X)$,
iii) If X and Y are independent, $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$.
iv) For any $a, b \in \mathbb{R}$, $\operatorname{Cov}(aX + bY, Z) = a \operatorname{Cov}(X, Z) + b \operatorname{Cov}(Y, Z)$.

Exercise 4.10. Show that if $X \geq 0$ a.s and $\mathbb{E}[X] = 0$, then $X = 0$ a.s.

Exercise 4.11. Show that $X = a$ a.s for some $a \in \mathbb{R}$ if and only if $\operatorname{Var}(X) = 0$.

Exercise 4.12 (You can't always under-perform). If X is a random variable with $\mathbb{E}[X] < \infty$, then $\mathbb{P}(X \geq \mathbb{E}[X]) > 0$.

Exercise 4.13. Show that if $X + Y$ is integrable, it does not imply X and Y are integrable. What if adding the condition that X and Y are independent?

Exercise 4.14. Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{E}X_i = \mu$ and $\operatorname{Var}(X_i) = \sigma^2$, and let N be integer valued random variable with $\mathbb{E}N = m$ and $\operatorname{Var}(N) = v$ and independent from all X_i. Show that

$$\operatorname{Var}\left(\sum_{i=1}^N X_i\right) = \sigma^2 m + \mu^2 v. \tag{4.33}$$

Exercise 4.15. Follow Exercise 2.5, find $\lim_{n\to\infty} \mathbb{E}\left[\prod_{i=1}^{n} X_n\right]$.

Exercise 4.16. Show that if X only takes values in $\mathbb{N}$, $\mathbb{E}X = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k)$. (Hint: use MCT)

Exercise 4.17. Show that if $X \geq 0$ and $p > 0$, $\mathbb{E}X^p = \int_0^\infty px^{p-1}\mathbb{P}(X \geq x)dx$.

Exercise 4.18. If X is an integrable random variable s.t. $X \geq 0$ a.s., show that

$$\lim_{t\to\infty} t\mathbb{P}(X > t) = 0. \tag{4.34}$$

(Hint: use DCT)

Exercise 4.19. Let $X \geq 0$ a.s. so that $M_X(\lambda) < \infty$ and $\mathbb{E}[Xe^{\lambda X}] < \infty$ for all $\lambda \in \mathbb{R}$.

1. Show that $M_X'(\lambda) = \mathbb{E}[Xe^{\lambda X}]$.
2. Suppose now we remove the condition that $X \geq 0$ a.s., but we have $M_X(\lambda) < \infty$ and $\mathbb{E}[|X|e^{\lambda X}] < \infty$ for all $\lambda \in \mathbb{R}$. Show once again that $M_X'(\lambda) = \mathbb{E}[Xe^{\lambda X}]$.

LECTURE 5

PROPERTIES OF EXPECTATION

5.1 Concentration of Measure

Theorem 5.1.1 (Markov's Inequality). If $X \geq 0$ a.s., then for all $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}X}{a}. \tag{5.1}$$

Proof. Define $Z(\omega) = a\mathbb{I}\{X(\omega) \geq a\}$. Then, $Z \leq X$ a.s., and $\mathbb{E}Z = aP(X \geq a)$. □

Corollary 5.1.2 (Chebyshev's Inequality). For all $t \geq 0$,

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}. \tag{5.2}$$

Proof. Apply Markov's to $Y = (X - \mathbb{E}X)^2$. □

Corollary 5.1.3 (Chernoff's Inequality). For all $t \geq 0$,

$$\mathbb{P}(X \geq \mathbb{E}X + t) \leq \inf_{\lambda > 0} M_{X-\mathbb{E}X}(\lambda)e^{-\lambda t}. \tag{5.3}$$

Proof. For $\lambda \geq 0$,

$$\begin{aligned}\mathbb{P}(X \geq \mathbb{E}X + t) &= \mathbb{P}(X - \mathbb{E}X \geq t) \\ &= \mathbb{P}(e^{\lambda(X-\mathbb{E}X)} \geq e^{\lambda t}) \\ &\leq M_{X-\mathbb{E}X}(\lambda)e^{-\lambda t}. \text{ (Markov).}\end{aligned} \tag{5.4}$$

Since the above is true for all $\lambda \geq 0$, then

$$\mathbb{P}(X \geq \mathbb{E}X + t) \leq \inf_{\lambda > 0} M_{X-\mathbb{E}X}(\lambda)e^{-\lambda t}. \tag{5.5}$$

□

Definition 5.1.4. A function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is *convex* if for all $\lambda \in (0, 1)$ and $x, y \in \mathbb{R}$,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y). \tag{5.6}$$

If '$\leq$' is changed to '$\geq$', then f is called *concave.*

Lemma 5.1.5. If Y satisfies $\mathbb{E}Y = 0$ and $a \leq Y \leq b$ a.s., for all $\lambda > 0$

$$M_Y(\lambda) \leq e^{\lambda^2(b-a)^2/8}. \tag{5.7}$$

Proof. Since $Y \in [a, b]$, we can write it as a convex combination via $Y = \alpha a + (1 - \alpha)b$ for some $\alpha \in [0, 1]$. In particular, this holds for $\alpha = (b - Y)/(b - a)$. Since $e^{\lambda Y}$ is convex in Y,

$$e^{\lambda Y} = e^{\lambda[\alpha a+(1-\alpha)b]} \leq \alpha e^{\lambda a} + (1 - \alpha)e^{\lambda b} = \frac{b - Y}{b - a}e^{\lambda a} + \frac{Y - a}{b - a}e^{\lambda b}. \tag{5.8}$$

Taking expectation of both sides gives

$$\mathbb{E}e^{\lambda Y} \leq \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b}. \tag{5.9}$$

Then, define $p = b/(b-a)$ and $u = (b-a)\lambda$. Also, consider the function

$$\varphi(u) = \log(pe^{\lambda a} + (1-p)e^{\lambda b}) = \lambda a + \log(p + (1-p)e^{\lambda(b-a)}) = (p-1)u + \log(p + (1-p)e^u). \tag{5.10}$$

Then, from Taylor expanding we see that there exists a $\xi \in \mathbb{R}$ such that

$$\varphi(u) = \varphi(0) + \varphi'(0)u + \frac{1}{2}\varphi''(\xi)u^2. \tag{5.11}$$

Observe that $\varphi(0) = 0$ and

$$\varphi'(x) = (p-1) + \frac{(1-p)e^x}{p + (1-p)e^x} = (p-1) + 1 - \frac{p}{p+(1-p)e^x}, \tag{5.12}$$

so $\varphi'(0) = 0$ as well. Finally,

$$\varphi''(x) = \frac{p(1-p)e^x}{[p+(1-p)e^x]^2}, \tag{5.13}$$

which you can check satisfies $\varphi''(x) \leq 1/4$ for all $x \in \mathbb{R}$. That is,

$$\mathbb{E}e^{\lambda Y} \leq e^{\varphi(u)} \leq e^{u^2/8} \leq e^{\lambda^2(b-a)^2/8}. \tag{5.14}$$

□

Theorem 5.1.6 (Hoeffding's Inequality). Suppose $X_1, X_2, \dots$ are independent with $a_i \leq X_i \leq b_i$ a.s. for all i. Then, letting $S_n = \sum_{i=1}^n X_i$, for all $t > 0$

$$\mathbb{P}\left(|S_n - \mathbb{E}S_n| \geq t\right) \leq 2\exp\left\{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}. \tag{5.15}$$

Proof. First, we apply Chernoff's inequality to obtain

$$\mathbb{P}(S_n - \mathbb{E}S_n \geq t) \leq \inf_{\lambda > 0} e^{-\lambda t} M_{S_n - \mathbb{E}S_n}(\lambda). \tag{5.16}$$

By independence and the above lemma,

$$M_{S_n - \mathbb{E}S_n}(\lambda) = \prod_{i=1}^n M_{X_i - \mathbb{E}X_i}(\lambda) \leq \prod_{i=1}^n e^{\lambda^2(b_i - a_i)^2/8} = \exp\left\{\frac{\lambda^2}{8}\sum_{i=1}^n (b_i - a_i)^2\right\}. \tag{5.17}$$

Thus,

$$\mathbb{P}(S_n - \mathbb{E}S_n \geq t) \leq \inf_{\lambda > 0} \exp\left\{-\lambda t + \frac{\lambda^2}{8}\sum_{i=1}^n (b_i - a_i)^2\right\}. \tag{5.18}$$

Taking $\lambda = \frac{4t}{\sum_{i=1}^n (b_i - a_i)^2}$ gives one direction of the result. To get the other direction consider $-X_1, \dots, -X_n$. □

5.2 Various Inequalities

Proposition 5.2.1 (Jensen's Inequality). If f is a convex function and X is a random variable such that $f(X)$ is integrable,

$$f(\mathbb{E}X) \leq \mathbb{E}f(X). \tag{5.19}$$

The inequality is flipped if f is concave.

Proof. Since f is convex, there exists a function g such that $g(x) = ax + b$ such that $g(x) \leq f(x)$ for all x and $g(\mathbb{E}X) = f(\mathbb{E}X)$. Then,

$$\mathbb{E}f(X) \geq \mathbb{E}g(X) = \mathbb{E}[aX + b] = a\mathbb{E}X + b = g(\mathbb{E}X) = f(\mathbb{E}X). \tag{5.20}$$

If concave then $-f$ is convex. □

Example 5.2.2. $(\mathbb{E}X)^2 \leq \mathbb{E}X^2$, $\mathbb{E}\log(X) \leq \log(\mathbb{E}X)$.

Lemma 5.2.3. If $0 < p < q$,

$$[\mathbb{E}\,|X|^p]^{1/p} \leq [\mathbb{E}\,|X|^q]^{1/q}. \tag{5.21}$$

Proof. Since $q/p > 1$, $f(x) = x^{q/p}$ is convex. Thus,

$$[\mathbb{E}\,|X|^p]^{q/p} \leq \mathbb{E}(|X|^p)^{q/p} = \mathbb{E}\,|X|^q\,. \tag{5.22}$$

□

Definition 5.2.4 (L^p space). Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For any $p \geq 1$, we say that a random variable $X \in L^p$, if $\mathbb{E}[|X|^p] < \infty$, and we can define a norm

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p} \tag{5.23}$$

Proposition 5.2.5 (Holder's Inequality). If $p, q > 1$ such that $1/p + 1/q = 1$,

$$\mathbb{E}\,|XY| \leq [\mathbb{E}\,|X|^p]^{1/p}[\mathbb{E}\,|Y|^q]^{1/q}. \tag{5.24}$$

Proof. If $[\mathbb{E}\,|X|^p]^{1/p} = 0$, $|X|^p = 0$ a.s. so $X = 0$ a.s., which implies $|XY| = 0$ a.s. The same holds if $[\mathbb{E}\,|Y|^q]^{1/q} = 0$.

Otherwise, let

$$X^* = \frac{|X|}{[\mathbb{E}\,|X|^p]^{1/p}} \text{ and } Y^* = \frac{|Y|}{[\mathbb{E}\,|Y|^q]^{1/q}}. \tag{5.25}$$

Observe that $\mathbb{E}(X^*)^p = \mathbb{E}(Y^*)^q = 1$.

We now show $\frac{1}{p}x^p + \frac{1}{q}y^q \geq xy$ for all $x, y \geq 0$. To see this, let $h_y(x) = \frac{1}{p}x^p + \frac{1}{q}y^q - xy$. Then, $h_y'(x) = x^{p-1} - y$ and $h_y''(x) = (p-1)x^{p-2} \geq 0$, so the minimizer is $x^* = y^{1/(p-1)}$. Plugging this in, $h_y(x^*) = \frac{1}{p}y^{p/(p-1)} + \frac{1}{q}y^q - y^{1/(p-1)}y = y^q\left(\frac{1}{p} + \frac{1}{q}\right) - y^q = 0$.

Thus, $X^*Y^* \leq \frac{1}{p}(X^*)^p + \frac{1}{q}(Y^*)^q$, so

$$\frac{\mathbb{E}\,|XY|}{[\mathbb{E}\,|X|^p]^{1/p}[\mathbb{E}\,|Y|^p]^{1/p}} \leq \frac{1}{p} + \frac{1}{q} = 1. \tag{5.26}$$

□

Corollary 5.2.6 (Cauchy-Schwarz Inequality).

$$\mathbb{E}\,|XY| \leq \sqrt{\mathbb{E}X^2\mathbb{E}Y^2}. \tag{5.27}$$

Proposition 5.2.7 (Minkowski's Inequality). If $p > 1$,

$$[\mathbb{E}\,|X+Y|^p]^{1/p} \leq [\mathbb{E}\,|X|^p]^{1/p} + [\mathbb{E}\,|Y|^p]^{1/p}. \tag{5.28}$$

Proof. Let $q = \frac{p}{p-1}$ so that $1/p + 1/q = 1$. Then, by Holder's,

$$\mathbb{E}\left[|X|\,|X+Y|^{p-1}\right] \leq [\mathbb{E}\,|X|^p]^{1/p}\left[\mathbb{E}\,|X+Y|^{q(p-1)}\right]^{1/q} = [\mathbb{E}\,|X|^p]^{1/p}\,[\mathbb{E}\,|X+Y|^p]^{(p-1)/p}. \tag{5.29}$$

Similarly,

$$\mathbb{E}\left[|Y|\,|X+Y|^{p-1}\right] \leq [\mathbb{E}\,|Y|^p]^{1/p}\,[\mathbb{E}\,|X+Y|^p]^{(p-1)/p}. \tag{5.30}$$

Thus,

$$\mathbb{E}\,|X+Y|^p \leq \mathbb{E}\left[(|X|+|Y|)\,|X+Y|^{p-1}\right] \leq \left([\mathbb{E}\,|X|^p]^{1/p} + [\mathbb{E}\,|Y|^p]^{1/p}\right)[\mathbb{E}\,|X+Y|^p]^{(p-1)/p}. \tag{5.31}$$

Rearrange to get the result. □

5.3 L^p Convergence

Definition 5.3.1 (Convergence in L^p). A sequence of random variables $X_1, X_2, \ldots$ converges in $L^p (p \geq 1)$ to a random variable X if

$$\lim_{n\to\infty} \mathbb{E}|X_n - X|^p = 0. \tag{5.32}$$

We denote this by

$$X_n \xrightarrow{L^p} X. \tag{5.33}$$

In particular, if $p = 2$, we say X_n converges in mean square to X.

Proposition 5.3.2. If $X_n \xrightarrow{L^p} X$, then $X_n \xrightarrow{P} X$. The converse does not hold.

Proof. By Markov,

$$\mathbb{P}(|X_n - X| \geq \varepsilon) \leq \frac{\mathbb{E}|X_n - X|^p}{\varepsilon^p} \to 0. \tag{5.34}$$

To show that the converse does not hold, let $U \sim \text{Unif}(0,1)$ and $X_n = 2^n\mathbb{I}\{0 \leq U \leq 1/n\}$. Clearly,

$$X_n \xrightarrow{P} 0. \tag{5.35}$$

However, $\mathbb{E}|X_n|^p = \frac{2^{np}}{n} \to \infty$. □

Remark. Convergence in L^p and almost sure convergence does not imply one another.

Lemma 5.3.3. If $X_n \xrightarrow{L^p} X$, then

$$\lim_{n\to\infty} \mathbb{E}|X_n| = \mathbb{E}|X|. \tag{5.36}$$

Proof. Since $p \geq 1$, by the reverse triangle inequality and Jensen's inequality,

$$\begin{aligned} |\mathbb{E}|X_n| - \mathbb{E}|X||^p &\leq (\mathbb{E}|X_n - X|)^p \\ &\leq \mathbb{E}|X_n - X|^p \to 0. \end{aligned} \tag{5.37}$$

Thus,

$$\mathbb{E}|X_n| \to \mathbb{E}|X|. \tag{5.38}$$

□

Proposition 5.3.4. If $1 \leq p < q$, and that $X_n \xrightarrow{L^q} X$, then $X_n \xrightarrow{L^p} X$.

Proof. By Holder's inequality, for any X,

$$\|X\|_p = \|X \times 1\|_p \leq \|1\|_{\frac{1}{1/p-1/q}} \|X\|_q = \|X\|_q \,. \tag{5.39}$$

Thus, the required is easily shown. □

5.4 Exercises

Exercise 5.1. Using a Taylor expansion, show that for a *Rademacher* random variable S (e.g., taking values 1 and -1 with probability $1/2$ each)

$$\mathbb{E}e^{\lambda S} \leq e^{\lambda^2/2}, \ \forall \lambda \in \mathbb{R}. \tag{5.40}$$

Then, letting $Z = \sum_{i=1}^{n} S_i$ for i.i.d. Rademachers S_i, show that for $t \geq 0$,

$$\mathbb{P}(Z \geq t) \leq e^{-t^2/(2n)}. \tag{5.41}$$

Exercise 5.2. Suppose $\mathbb{E}X_n = 0$ and $\mathbb{E}(X_n^2) = 1$ for all n. Prove that $\mathbb{P}(X_n \geq n \text{ i.o.}) = 0$.

Exercise 5.3. Find a random variable X and $a > 0$ such that $\mathbb{P}(X > a) \geq \mathbb{E}X/a$, and identify what breaks in the proof of Markov's inequality for this example.

Exercise 5.4. Show that the functions $f(x) = x/(1+x)$ and $f(x) = \log(x)$ are concave for $x > 0$.

Exercise 5.5. For X such that $\mathbb{E}X < \infty$ and $a \in \mathbb{R}$, prove that $\mathbb{E}[\max\{X, a\}] \geq \max\{\mathbb{E}X, a\}$.

Exercise 5.6. For any X, Y, use Cauchy-Schwarz to show that $|\text{Corr}(X, Y)| \leq 1$.

Exercise 5.7. Let $p \geq 0$, show that for random variables X, Y,

$$\mathbb{E}|X + Y|^p \leq 2^p(\mathbb{E}|X|^p + \mathbb{E}|Y|^p). \tag{5.42}$$

Moreover, show that if $p \geq 1$, 2^p in the above can be replaced by 2^{p-1}. If $0 \leq p \leq 1$, 2^p can be replaced by 1.

Exercise 5.8.

1. Show that for $a, b \geq 0$ and $1 \leq p \leq 2$,

$$(a^p + b^p)^2 \geq (a^2 + b^2)^p. \tag{5.43}$$

2. Suppose X, Y are independent, and that Y is symmetric, i.e., Y and $-Y$ have the same distribution. Let $1 \leq p \leq 2$, show that $\mathbb{E}|X + Y|^p \leq \mathbb{E}|X|^p + \mathbb{E}|Y|^p$.

1. Let $f(x) = 2\log(x^p + 1) - p\log(x^2 + 1)$ with $x \geq 0$. Then

$$f'(x) = \frac{2p(x^{p-1} - x)}{(x^p + 1)(x^2 + 1)}. \tag{5.44}$$

Therefore, $f'(x) = 0$ iff $x = 0$ or $x = 1$. Moreover, $f'(x) \geq 0$ for $x \in [0, 1]$ and $f'(x) \leq 0$ for $x \in [1, \infty)$. Thus, $f(x)$ reaches global maximum at $x = 1$. Further, $f(0) = 0$ and $f(1) = (2 - p)\log(2) \geq 0$, so $f(x) \geq 0$ for $x \in [0, 1]$. In addition, $f(1) > 0$ and $f'(x) < 0$ on $[1, \infty)$ if $p \neq 2$, and $\lim_{x\to\infty} f(x) = 0$. Therefore, when $p \neq 2$, $f(x) > 0$ on $[1, \infty)$, otherwise, contradiction. If $p = 2$, then $f(x) \equiv 0$ for all $x \geq 0$. Hence, for all $x \geq 0$, $f(x) \geq 0$. Therefore, for $x \geq 0$ and $p \in [1, 2]$,

$$2\log(x^p + 1) \geq p\log(x^2 + 1). \tag{5.45}$$

Now let $x = a/b$ with $a \geq 0, b > 0$ and rearrange, we have

$$(a^p + b^p)^2 \geq (a^2 + b^2)^p. \tag{5.46}$$

If $b = 0$, the result still holds trivially for any $a \geq 0$.

2. By results from 1, we have for any $a, b \geq 0$,

$$a^p + b^p \geq (a^2 + b^2)^{p/2}. \tag{5.47}$$

Let $a = \left|\frac{u+v}{2}\right|$ and $b = \left|\frac{u-v}{2}\right|$, we have

$$\begin{aligned}\left|\frac{u+v}{2}\right|^p + \left|\frac{u-v}{2}\right|^p &\geq \left(\left|\frac{u+v}{2}\right|^2 + \left|\frac{u+v}{2}\right|^2\right)^{p/2} \\ &= \left(\frac{u^2}{2} + \frac{v^2}{2}\right)^{p/2} \\ &\geq \frac{1}{2}(|u|^p + |v|^p) \quad (y = x^{p/2} \text{ concave}).\end{aligned} \tag{5.48}$$

Now let $X = u + v$ and $Y = u - v$, we have

$$|X|^p + |Y|^p \geq 2^{p-1}\left(\left|\frac{X+Y}{2}\right|^p + \left|\frac{X-Y}{2}\right|^p\right), \tag{5.49}$$

or equivalently,

$$2(|X|^p + |Y|^p) \geq |X+Y|^p + |X-Y|^p. \tag{5.50}$$

Now taking expectation for the above. Since X, Y are independent and $Y \overset{d}{=} -Y$, we have $\mathbb{E}|X+Y|^p = \mathbb{E}|X-Y|^p$. Thus,

$$\mathbb{E}|X+Y|^p \leq \mathbb{E}|X|^p + \mathbb{E}|Y|^p \tag{5.51}$$

as required.

Exercise 5.9. Prove that if X is such that $\mathbb{E}X = \mu < \infty$ and $\text{Var}(X) = \sigma^2 < \infty$, for all $a > 0$

$$\mathbb{P}(X - \mu \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2}. \tag{5.52}$$

Exercise 5.10. Let $X_1, X_2, \ldots$ satisfy $\mathbb{E}X_n = m < \infty$ and $\text{Var}(X_n) = 1/\sqrt{n}$. Prove that $X_n \xrightarrow{P} m$.

Exercise 5.11. Give an example of $X_1, X_2, \ldots$ such that $X_n/n \xrightarrow{P} 0$ and $X_n/n^2 \longrightarrow 0$ a.s., but $\mathbb{P}(X_n/n \longrightarrow 0) < 1$.

LECTURE 6

LAWS OF LARGE NUMBERS

6.1 Familiar Results

Theorem 6.1.1 (Basic Weak LLN). Let $X_1, X_2, \dots$ be independent, and for all i, suppose $\mathbb{E}X_i = \mu$ and $\mathrm{Var}(X_i) \le \sigma^2 < \infty$. Define $S_n = \sum_{i=1}^n X_i$. Then,

$$\frac{1}{n} S_n \xrightarrow{P} \mu. \tag{6.1}$$

Proof. By linearity, $\mathbb{E}S_n/n = \mu$ and $\mathrm{Var}(S_n/n) \le \sigma^2/n$. Then, for any $\varepsilon > 0$, Chebyshev's gives that

$$\lim_{n\to\infty} \mathbb{P}(|S_n/n - \mu| > \varepsilon) \le \lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} = 0. \tag{6.2}$$

□

Theorem 6.1.2 (Basic Strong LLN). Let $X_1, X_2, \dots$ be independent, and for all i, suppose $\mathbb{E}X_i = \mu$ and $\mathbb{E}(X_i - \mu)^4 \le a < \infty$. Then,

$$\frac{1}{n} S_n \longrightarrow \mu \text{ a.s.} \tag{6.3}$$

Proof. First, observe that $\mathbb{E}(X_i - \mu)^2 \le \mathbb{E}(X_i - \mu)^4 + 1$, by considering the case when the variance is smaller and greater than 1. Without loss of generality, we can suppose $\mu = 0$. Then, we have that

$$\begin{aligned}
\mathbb{E}S_n^4 &= \mathbb{E}\left(\sum_{i=1}^n X_i\right)^4 \\
&= \mathbb{E}\Big(\sum_{i=1}^n X_i^4 + k_1 \sum_{i=1}^n \sum_{j\neq i} X_i^3 X_j + k_2 \sum_{i=1}^n \sum_{j\neq i} X_i^2 X_j^2 + k_3 \sum_{i=1}^n \sum_{j\neq i} \sum_{k\neq j,i} X_i^2 X_j X_k \\
&\qquad + k_4 \sum_{i=1}^n \sum_{j\neq i} \sum_{k\neq j,i} \sum_{\ell\neq j,i,k} X_i X_j X_k X_\ell\Big) \\
&= \mathbb{E}\sum_{i=1}^n X_i^4 + k_2 \mathbb{E}\sum_{i=1}^n \sum_{j\neq i} X_i^2 X_j^2 \\
&\le na + k_2 n(n-1)(a+1)^2 \\
&\le Kn^2.
\end{aligned}$$

Next, for any $\varepsilon > 0$, we can apply Markov's to get

$$\mathbb{P}\left(\left|\frac{1}{n}S_n\right| > \varepsilon\right) = \mathbb{P}(S_n^4 > n^4\varepsilon^4) \le \frac{\mathbb{E}S_n^4}{n^4\varepsilon^4} \le \frac{K}{n^2\varepsilon^4}. \tag{6.4}$$

Since $\sum_{n=1}^\infty \frac{K}{n^2\varepsilon^4} < \infty$, by Borel-Cantelli the result holds. □

6.2 Advanced Weak LLN

Definition 6.2.1. A sequence of random variables $(X_i)_{i\in\mathcal{I}}$ with $\mathbb{E}X_i^2 < \infty$ are uncorrelated if for all $i \neq j$, $\mathbb{E}X_iX_j = \mathbb{E}X_i\mathbb{E}X_j$.

Theorem 6.2.2 (Uncorrelated Weak LLN). If $X_1, X_2, \dots$ are uncorrelated with $\mathbb{E}X_i = \mu$ and $\mathrm{Var}(X_i) \leq \sigma^2 < \infty$,

$$\mathbb{E}\left(\frac{1}{n}S_n - \mu\right)^2 \to 0. \tag{6.5}$$

Proof.

$$\mathbb{E}\left(\frac{1}{n}S_n - \mu\right)^2 = \mathrm{Var}\left(\frac{1}{n}S_n\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) \leq \frac{n\sigma^2}{n^2} \to 0. \tag{6.6}$$

□

Theorem 6.2.3 (Weak LLN). If $X_1, X_2, \dots$ are i.i.d. with $\lim_{x\to\infty} x\mathbb{P}(|X_1| > x) = 0$, for $\mu_n = \mathbb{E}\left(X_1\mathbb{I}\left\{|X_1| \leq n\right\}\right)$,

$$\frac{1}{n}S_n - \mu_n \xrightarrow{P} 0. \tag{6.7}$$

Proof. Fix $\varepsilon > 0$. Let $\bar{X}_k^{(n)} = X_k\mathbb{I}\{|X_k| \leq n\}$ and $\bar{S}_n = \sum_{k=1}^n \bar{X}_k^{(n)}$. Then,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu_n\right| > \varepsilon\right) \leq \mathbb{P}(S_n \neq \bar{S}_n) + \mathbb{P}\left(\left|\frac{\bar{S}_n}{n} - \mu_n\right| > \frac{\varepsilon}{2}\right). \tag{6.8}$$

For the first term,

$$\begin{aligned}\mathbb{P}(S_n \neq \bar{S}_n) &\leq \mathbb{P}\left(\bigcup_{k=1}^n \{\bar{X}_k^{(n)} \neq X_k\}\right)\\ &\leq \sum_{k=1}^n \mathbb{P}(\bar{X}_k^{(n)} \neq X_k)\\ &= \sum_{k=1}^n \mathbb{P}(|X_k| > n)\\ &= n\mathbb{P}(|X_1| > n)\\ &\longrightarrow 0.\end{aligned} \tag{6.9}$$

For the second term, first observe that

$$\mathbb{E}\bar{S}_n = \mathbb{E}\sum_{k=1}^n \bar{X}_k^{(n)} = \sum_{k=1}^n \mathbb{E}[X_k\mathbb{I}\{|X_k| \leq n\}] = n\mu_n. \tag{6.10}$$

So, by Chebyshev's,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\bar{S}_n}{n}-\mu_n\right|>\frac{\varepsilon}{2}\right)&\leq\frac{4}{n^2\varepsilon^2}\mathbb{E}\left(\bar{S}_n-n\mu_n\right)^2\\&=\frac{4}{n^2\varepsilon^2}\operatorname{Var}\left(\bar{S}_n\right)\\&=\frac{4}{n^2\varepsilon^2}\sum_{k=1}^{n}\operatorname{Var}\left(\bar{X}_k^{(n)}\right)\\&=\frac{4}{n\varepsilon^2}\operatorname{Var}\left(\bar{X}_1^{(n)}\right)\\&\leq\frac{4}{n\varepsilon^2}\mathbb{E}(X_1\mathbb{I}\{|X_1|\leq n\})^2.\end{aligned}\tag{6.11}$$

Finally, recalling that $\mathbb{E}X^p=\int_0^\infty px^{p-1}\mathbb{P}(X\geq x)dx$,

$$\begin{aligned}\mathbb{E}(X_1\mathbb{I}\{|X_1|\leq n\})^2&=\int_0^\infty 2x\mathbb{P}\left(\left|\bar{X}_k^{(n)}\right|\geq x\right)dx\\&=\int_0^n 2x\mathbb{P}(|X_k|\geq x)dx\\&=2\int_0^n x\mathbb{P}(|X_1|\geq x)dx.\end{aligned}\tag{6.12}$$

Since $0\leq x\mathbb{P}(|X_1|\geq x)\leq x$ for all x and goes to 0, $\sup_x x\mathbb{P}(|X_1|\geq x)<\infty$. Thus,

$$\lim_{n\to\infty}\frac{1}{n}\int_0^n x\mathbb{P}(|X_1|\geq x)dx=\lim_{n\to\infty}\int_0^1 ny\mathbb{P}(|X_1|>ny)dy=\int_0^1\lim_{n\to\infty}ny\mathbb{P}(|X_1|>ny)dy=0.\tag{6.13}$$

□

Corollary 6.2.4. If $X_1,X_2,\ldots$ are i.i.d. with $\mathbb{E}|X_1|<\infty$ and $\mathbb{E}X_1=\mu<\infty$,

$$\frac{1}{n}S_n\xrightarrow{P}\mu.\tag{6.14}$$

Proof. Let $Y_n=|X_1|\mathbb{I}\{|X_1|>n\}$. For each $\omega\in\Omega$, $|X_1(\omega)|<\infty$, so $Y_n\longrightarrow 0$ a.s. Since $|Y_n|\leq|X_1|$ which is integrable, by the DCT we have $\mathbb{E}Y_n\longrightarrow 0$. Further, observe that $Y_n\geq n\mathbb{I}\{|X_1|>n\}$, so $\mathbb{E}Y_n\geq n\mathbb{P}(|X_1|>n)$. Thus, $\lim_{x\to\infty}x\mathbb{P}(|X_1|>x)=0$.

Next, consider $Z_n=X_1\mathbb{I}\{|X_1|\leq n\}$. Again, since $|X_1(\omega)|<\infty$, $Z_n\longrightarrow X_1$ a.s. Thus, we can again apply DCT to obtain $\mu_n=\mathbb{E}Z_n\longrightarrow\mathbb{E}X_1=\mu$. The result then follows by the Weak LLN. □

6.3 Advanced Strong LLN

Theorem 6.3.1 (Strong LLN). If $X_1,X_2,\ldots$ are i.i.d. with $\mathbb{E}|X_1|<\infty$ and $\mathbb{E}X_1=\mu<\infty$,

$$\frac{1}{n}S_n\longrightarrow\mu\text{ a.s.}\tag{6.15}$$

Proof. Not required to be able to prove for this class. The main techniques are well summarized in the proof of the Weak LLN.

□

6.4 Applications

Proposition 6.4.1 (Polynomial Approximation). If $f : [0,1] \to \mathbb{R}$ is continuous, the Bernstein approximation

$$f_n(x) = \sum_{m=0}^{n} \binom{n}{m} x^m (1-x)^{n-m} f(m/n) \tag{6.16}$$

satisfies

$$\sup_{x \in [0,1]} |f_n(x) - f(x)| \longrightarrow 0. \tag{6.17}$$

Proof. Fix $x \in [0,1]$ and let $X_1, X_2, \dots$ be independent with $\mathbb{P}(X_i = 1) = x$ and $\mathbb{P}(X_i = 0) = 1 - x$. Then, $\mathbb{E}X_i = x$ and $\operatorname{Var}(X_i) = x(1-x)$, and further $\mathbb{P}(S_n = m) = \binom{n}{m} x^m (1-x)^{n-m}$ (the Binomial pmf). That is, $\mathbb{E}f(S_n/n) = \sum_{m=0}^{n} \mathbb{P}(S_n = m) f(m/n) = \sum_{m=0}^{n} \binom{n}{m} x^m (1-x)^{n-m} f(m/n) = f_n(x)$. Now, fix $\varepsilon > 0$, and observe that since $[0,1]$ bounded and compact, f is uniformly continuous. That is, there exists $\delta > 0$ such that $|x - y| < \delta$ implies $|f(x) - f(y)| < \varepsilon$. Then,

$$\begin{aligned} & |f_n(x) - f(x)| \\ = \ & |\mathbb{E}f(S_n/n) - f(x)| \\ \leq \ & \mathbb{E}\, |f(S_n/n) - f(x)| \\ = \ & \mathbb{E}\left[|f(S_n/n) - f(x)| \, \mathbb{I}\{|S_n/n - x| < \delta\}\right] + \mathbb{E}\left[|f(S_n/n) - f(x)| \, \mathbb{I}\{|S_n/n - x| \geq \delta\}\right] \\ \leq \ & \varepsilon + 2M\mathbb{P}(|S_n/n - x| \geq \delta), \end{aligned} \tag{6.18}$$

where $M = \sup_{x \in [0,1]} |f(x)| < \infty$ since f is continuous on a closed interval. Now, by Chebyshev's inequality,

$$\mathbb{P}(|S_n/n - x| \geq \delta) \leq \frac{\mathbb{E}[(S_n/n - x)^2]}{\delta^2} = \frac{\operatorname{Var}(S_n/n)}{\delta^2} = \frac{\sum_{i=1}^{n} \operatorname{Var}(X_i)}{n^2 \delta^2} = \frac{x(1-x)}{n\delta^2} \leq \frac{1}{4n\delta^2}.$$

Since this was true for all $x \in [0,1]$, $\lim_{n\to\infty} \sup_{x \in [0,1]} |f_n(x) - f(x)| \leq \varepsilon$, but ε was arbitrary so the proposition follows. □

Theorem 6.4.2 (Glivenko-Cantelli). Let $X_1, X_2, \dots$ be i.i.d. with CDF F and define the empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i \leq x\}.$$

Then,

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \longrightarrow 0 \qquad \text{a.s.} \tag{6.19}$$

Proof. Out of scope. □

Theorem 6.4.3 (Basic Monte-Carlo Integration on an Interval). Let $g : [0,1] \to \mathbb{R}$ be an integrable function. Let $U_i \overset{\text{i.i.d.}}{\sim} \operatorname{Unif}(0,1)$ for $i \in [n]$. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(U_i) = \int_0^1 g(x) dx \text{ a.s.} \tag{6.20}$$

Proof. Notice that

$$\mathbb{E}[g(U)] = \int_0^1 g(u) f_U(u) du = \int_0^1 g(u) du, \tag{6.21}$$

and since g is integrable, then

$$\int_0^1 |g(u)| du < \infty, \tag{6.22}$$

and $\mathbb{E}[g(U)]$ is finite. Thus, by the strong LLN,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(U_i) = \mathbb{E}[g(U)] = \int_0^1 g(x) dx \text{ a.s.} \tag{6.23}$$

□

Theorem 6.4.4 (General Monte-Carlo Integration on $\mathbb{R}$). Let $g : \mathbb{R} \to \mathbb{R}$ be an integrable function. Let $X_i,\ i \in [n]$ be continuous i.i.d. R.V.s with p.d.f. f_X and support $\mathbb{R}$. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{g(X_i)}{f_X(X_i)} = \int_{\mathbb{R}} g(x) dx \text{ a.s.} \tag{6.24}$$

Proof. Since g is integrable, then

$$\int_{\mathbb{R}} |g(x)| dx < \infty, \tag{6.25}$$

and

$$\int_{\mathbb{R}} g(x) dx \tag{6.26}$$

is defined. Notice that

$$\int_{\mathbb{R}} g(x) dx = \int_{\mathbb{R}} \frac{g(x)}{f_X(x)} f_X(x) dx = \mathbb{E}\left[\frac{g(X)}{f_X(X)}\right] \tag{6.27}$$

Thus, by the strong LLN,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{g(X_i)}{f_X(X_i)} = \mathbb{E}\left[\frac{g(X)}{f_X(X)}\right] = \int_{\mathbb{R}} g(x) dx \text{ a.s.} \tag{6.28}$$

□

6.5 Exercises

Exercise 6.1. Suppose $X_1, X_2, \dots$ are uncorrelated with $\mathbb{E}X_i = \mu_i$ and $\lim_{i\to\infty} \frac{\text{Var}(X_i)}{i} = 0$. If $\nu_n = \frac{1}{n}\mathbb{E}S_n$, show that

$$\lim_{n\to\infty} \mathbb{E}\Big[\Big(\frac{S_n}{n} - \nu_n\Big)^2\Big] = 0.$$

Exercise 6.2. Let $X_1, X_2, \dots$ be i.i.d. such that $\mathbb{P}(X_1 = (-1)^k k) = \frac{C}{k^2\log(k)}$ for all integers $k \geq 2$, where C is a constant so that the probabilities sum to 1. Show that $\mathbb{E}\,|X_1| = \infty$ but there is a finite μ such that $\frac{S_n}{n} \xrightarrow{P} \mu$.

Hint #1: $\mu_n = \mathbb{E}[X_1\mathbb{I}\{|X_1 \leq n|\}] = \sum_{k=2}^n (-1)^k k \frac{C}{k^2\log(k)}$ is an alternating sequence of real numbers that congerge to zero, so from calculus class $\mu_n \longrightarrow \mu$ for some real number μ. Thus, it suffices to show $\frac{S_n}{n} - \mu_n \xrightarrow{P} 0$.
Hint #2: For any positive and decreasing function f, $\sum_{k=x+1}^\infty f(k) \leq \int_x^\infty f(y)dy \leq \sum_{k=x}^\infty f(k)$.

Exercise 6.3. Let $X_1, X_2, \dots$ be i.i.d. such that $\mathbb{P}(X_1 > x) = \frac{e}{x\log(x)}$ for $x \geq e$. Show that $\mathbb{E}\,|X_1| = \infty$, but $\frac{S_n}{n} - \mu_n \xrightarrow{P} 0$.
Hint: Recall Exercise 4.17.

Exercise 6.4. For any sequence of random variables X_n and $\varepsilon > 0$, show there exist constants $c_n \to \infty$ such that $\mathbb{P}(|X_n| > \varepsilon c_n) < 2^{-n}$.

Exercise 6.5. For any sequence of random variables X_n, show there exist constants $c_n \to \infty$ such that

$$\frac{X_n}{c_n} \longrightarrow 0 \text{ a.s.}$$

Exercise 6.6. Suppose $X_1, X_2, \dots$ are i.i.d. Show that $\mathbb{E}\,|X_1| < \infty$ if and only if

$$\frac{X_n}{n} \longrightarrow 0 \text{ a.s.}$$

Exercise 6.7. Suppose $X_1, X_2, \dots$ are i.i.d. such that for all n, $\mathbb{P}(X_n > x) = e^{-x}$. Show that

$$\limsup_{n\to\infty} \frac{X_n}{\log(n)} = 1 \text{ a.s.} \tag{6.29}$$

LECTURE 7

CENTRAL LIMIT THEOREMS

7.1 Weak Convergence

Definition 7.1.1. A sequence of probability measures μ_n with corresponding distribution functions F_n *converge weakly* to a probability measure μ with distribution function F if for all continuity points x of F

$$\lim_{n\to\infty} F_n(x) = F(x), \tag{7.1}$$

and is denoted by $\mu_n \Longrightarrow \mu$, or $F_n \Longrightarrow F$, or $X_n \Longrightarrow X$ if $X_n \sim F_n$ and $X \sim F$.

Theorem 7.1.2 (Scheffé's Theorem). Let X_n each have density f_n and X have density f. Then, if $f_n(x) \longrightarrow f(x)$ for all $x \in \mathbb{R}$, $X_n \Longrightarrow X$.

Proof. First, denote $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$. We have $x^+ + x^- = |x|$, $x^+ - x^- = x$, and $(-x)^+ = x^-$. Observe that the existence of a density implies F is continuous everywhere.

$$\begin{aligned}
|F_n(y) - F(y)| &= \left|\int_{-\infty}^{y} [f_n(x) - f(x)]\, dx\right| \\
&\leq \int_{-\infty}^{y} |f_n(x) - f(x)|\, dx \\
&\leq \int_{\mathbb{R}} |f_n(x) - f(x)|\, dx \\
&= \int_{\mathbb{R}} [f_n(x) - f(x)]^+\, dx + \int_{\mathbb{R}} [f_n(x) - f(x)]^-\, dx \\
&= \int_{\mathbb{R}} f_n(x) - f(x) + [f_n(x) - f(x)]^-\, dx + \int_{\mathbb{R}} [f_n(x) - f(x)]^-\, dx \\
&= 2\int_{\mathbb{R}} [f_n(x) - f(x)]^-\, dx \\
&= 2\int_{\mathbb{R}} [f(x) - f_n(x)]^+\, dx.
\end{aligned} \tag{7.2}$$

Then, observe that $[f(x) - f_n(x)]^+ \leq [f(x)]^+ = f(x)$, so by DCT,

$$\lim_{n\to\infty} |F_n(y) - F(y)| \leq 2\int \lim_{n\to\infty} [f(x) - f_n(x)]^+\, dx = 0. \tag{7.3}$$

□

Theorem 7.1.3 (Portmanteau Lemma). $F_n \Longrightarrow F$ if and only if for all bounded and continuous $h : \mathbb{R} \to \mathbb{R}$, when $X_n \sim F_n$ and $X \sim F$ then

$$\mathbb{E}h(X_n) \longrightarrow \mathbb{E}h(X).$$

Proof. Out of scope. □

Corollary 7.1.4. If $X_n \Longrightarrow X$, then $f(X_n) \Longrightarrow f(X)$ for any continuous f.

Proof. For any continuous and bounded g, $g \circ f$ is also continuous and bounded. □

Corollary 7.1.5. If $X_n \xrightarrow{P} X$ then $X_n \Longrightarrow X$.

Proof. Consider any subsequence X_{n_j}, and observe that $X_{n_j} \xrightarrow{P} X$ as well. Also, recall that there exists a further subsequence $X_{n_{j(k)}} \longrightarrow X$ a.s. Thus, for bounded and continuous h, the continuous mapping theorem and bounded convergence theorem give that $\mathbb{E}h(X_{n_{j(k)}}) \longrightarrow \mathbb{E}h(X)$. Since $\mathbb{E}h(X_n)$ is a sequence of real numbers, this implies $\mathbb{E}h(X_n) \longrightarrow \mathbb{E}h(X)$. □

Example 7.1.6. The reverse does not hold. Let $X \sim \text{Gaussian}(0,1)$ and $X_n = -X$ for $n \in \mathbb{N}$. Then, trivially $F_n = F$, so $X_n \Longrightarrow X$, but

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X| > \varepsilon/2) = c > 0. \tag{7.4}$$

Lemma 7.1.7. If $X_n \Longrightarrow c$ where c is constant, $X_n \xrightarrow{P} c$.

Proof. First, let F be the CDF of the random variable $X \equiv c$, so that $F(y) = \mathbb{I}\{c \leq y\}$. Observe that F is continuous everywhere except $y = c$. Now, fix $\varepsilon > 0$.

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(|X_n - c| > \varepsilon) &= \lim_{n\to\infty} [\mathbb{P}(X_n < c - \varepsilon) + \mathbb{P}(X_n > c + \varepsilon)] \\ &\leq \lim_{n\to\infty} [\mathbb{P}(X_n \leq c - \varepsilon/2) + \mathbb{P}(X_n > c + \varepsilon)] \\ &= \lim_{n\to\infty} F_n(c - \varepsilon/2) + 1 - \lim_{n\to\infty} F_n(c + \varepsilon) \\ &= F(c - \varepsilon/2) + 1 - F(c + \varepsilon) \\ &= 0.\end{aligned} \tag{7.5}$$

Note: we need to set it to $c - \varepsilon/2$ in the above instead of c because c is a discontinuity point of F, and there may not be convergence at this point. □

Theorem 7.1.8 (Slutsky's Theorem). If $X_n \Longrightarrow X$ and $Y_n \Longrightarrow c$ for a constant c,

- $X_n + Y_n \Longrightarrow X + c$,
- $X_n Y_n \Longrightarrow Xc$,
- $X_n/Y_n \Longrightarrow X/c$ if $c \neq 0$.

Proof.

- Fix $\varepsilon > 0$ and z a continuity point of the CDF of $X + c$. Observe that $X_n \leq z - c - \varepsilon$ and $Y_n \leq c + \varepsilon$ implies $X_n + Y_n \leq z$. Thus,

$$\begin{aligned}&\mathbb{P}(X_n + Y_n \leq z) \\ &\geq \mathbb{P}(X_n \leq z - c - \varepsilon \cap Y_n \leq c + \varepsilon) \\ &= \mathbb{P}(X_n \leq z - c - \varepsilon) + \mathbb{P}(Y_n \leq c + \varepsilon) - \mathbb{P}(X_n \leq z - c - \varepsilon \cup Y_n \leq c + \varepsilon) \\ &\geq \mathbb{P}(X_n \leq z - c - \varepsilon) + \mathbb{P}(Y_n \leq c + \varepsilon) - 1 \\ &= \mathbb{P}(X_n \leq z - c - \varepsilon) - \mathbb{P}(Y_n > c + \varepsilon).\end{aligned} \tag{7.6}$$

Now, observe that the CDF of Y_n is continuous everywhere except at c. So,

$$\lim_{n\to\infty} \mathbb{P}(Y_n > c + \varepsilon) = \mathbb{P}(Y > c + \varepsilon) = 0. \tag{7.7}$$

For arbitrarily small ε we can take $z-c-\varepsilon$ to be a continuity point of the CDF of X without loss of generality, so

$$\lim_{n\to\infty} \mathbb{P}(X_n \leq z-c-\varepsilon) = \mathbb{P}(X \leq z-c-\varepsilon). \tag{7.8}$$

That is,

$$\liminf_{n\to\infty} \mathbb{P}(X_n + Y_n \leq z) \geq \mathbb{P}(X + c \leq z-\varepsilon). \tag{7.9}$$

Since z is a continuity point of the CDF of $X+c$, taking $\varepsilon \to 0$ gives

$$\liminf_{n\to\infty} \mathbb{P}(X_n + Y_n \leq z) \geq \mathbb{P}(X + c \leq z). \tag{7.10}$$

Similarly, $X_n + Y_n \leq z$ and $Y_n \geq c-\varepsilon$ implies $X_n \leq z-c+\varepsilon$, so

$$\mathbb{P}(X_n \leq z-c+\varepsilon) \geq \mathbb{P}(X_n + Y_n \leq z) - \mathbb{P}(Y_n < c-\varepsilon). \tag{7.11}$$

Rearranging and taking limits in the same way gives

$$\limsup_{n\to\infty} \mathbb{P}(X_n + Y_n \leq z) \leq \mathbb{P}(X + c \leq z). \tag{7.12}$$

The remaining two are left as exercises. □

Proposition 7.1.9 (Delta Method). Suppose $a_n(X_n - \theta) \Longrightarrow X$ with a_n not dependent on θ and $a_n \to \infty$ as $n \to \infty$. If g is a function that is differentiable at θ with non-zero derivative. Then

$$a_n(g(X_n) - g(\theta)) \Longrightarrow g'(\theta)X. \tag{7.13}$$

Proof. Since g is differentiable at θ, it means that

$$g'(\theta) = \lim_{h\to 0} \frac{g(\theta+h) - g(\theta)}{h}. \tag{7.14}$$

Define

$$r(h) = \frac{g(\theta+h) - g(\theta)}{h} - g'(\theta),\ \ h \neq 0, \tag{7.15}$$

and $r(0) = 0$. Clearly, $r(h) \to 0$ as $h \to 0$, and

$$g(\theta+h) = g(\theta) + g'(\theta)h + hr(h). \tag{7.16}$$

Let $h = X_n - \theta$ and multiply by a_n, we have

$$a_n(g(X_n) - g(\theta)) = g'(\theta)a_n(X_n - \theta) + a_n(X_n - \theta)r(X_n - \theta). \tag{7.17}$$

Note that $a_n(X_n - \theta) \Longrightarrow X$ implies $X_n \xrightarrow{P} \theta$. To see this, fix $\epsilon, \delta > 0$,

$$\begin{aligned} \mathbb{P}(|X_n - \theta| > \epsilon) &= \mathbb{P}(a_n|X_n - \theta| > a_n\epsilon) \\ &= \mathbb{P}(a_n(X_n - \theta) < -a_n\epsilon) + \mathbb{P}(a_n(X_n - \theta) > a_n\epsilon). \end{aligned} \tag{7.18}$$

Choose $x > 0$ s.t. x and $-x$ are the continuous points of F_X and that $\mathbb{P}(X \leq -x) \leq \delta/4$ and $\mathbb{P}(X > x) \leq \delta/4$. Since $a_n(X_n - \theta) \Longrightarrow X$, there is a sufficiently large N s.t. for all $n \geq N$, $a_n\epsilon > x$ and

$$\begin{aligned} \mathbb{P}(a_n(X_n - \theta) < -a_n\epsilon) &< \mathbb{P}(a_n(X_n - \theta) \leq -x) \\ &< \mathbb{P}(X \leq -x) + \delta/4 \leq \delta/2, \end{aligned} \tag{7.19}$$

and

$$\begin{aligned}\mathbb{P}(a_n(X_n-\theta) > a_n\epsilon) &< \mathbb{P}(a_n(X_n-\theta) > x) \\ &< \mathbb{P}(X > x) + \delta/4 \leq \delta/2,\end{aligned} \tag{7.20}$$

which gives us

$$\mathbb{P}(|X_n-\theta|> \epsilon) < \delta \tag{7.21}$$

for all $N \geq n$. Therefore, $r(X_n-\theta) \xrightarrow{P} 0$ (*why?*), and by Slutsky's theorem,

$$a_n(g(X_n)-g(\theta)) = g'(\theta)a_n(X_n-\theta) + a_n(X_n-\theta)r(X_n-\theta) \Longrightarrow g'(\theta)X. \tag{7.22}$$

□

7.2 Characteristic Functions

Remark. This section requires the use of complex numbers. Recall i is defined as the solution to $x^2=-1$, and a complex number $x \in \mathbb{C}$ is defined by real numbers $a, b \in \mathbb{R}$ such that $x = a+ib$. The *modulus* of x is $|x| = \sqrt{a^2+b^2}$ and the *complex conjugate* is $\bar{x} = a - bi$. The *real* and *imaginary* parts of x are defined respectively by $\mathrm{Re}(x) = a$ and $\mathrm{Im}(x) = b$.

Lemma 7.2.1 (Euler's Formula). For any $x \in \mathbb{R}$,

$$e^{ix} = \cos(x) + i\sin(x) \tag{7.23}$$

Proof. See first year calculus book. □

Definition 7.2.2. For a random variable taking X taking values in $\mathbb{C}$, we define its expectation by

$$\mathbb{E}X = \mathbb{E}\mathrm{Re}(X) + i\mathbb{E}\mathrm{Im}(X). \tag{7.24}$$

Definition 7.2.3. A random variable X has a unique *characteristic function* defined by

$$\varphi(t) = \mathbb{E}e^{itX} = \mathbb{E}\cos(tX) + i\mathbb{E}\sin(tX). \tag{7.25}$$

Proposition 7.2.4. A characteristic function φ has the following properties:

i) $\varphi(0) = 1$.
ii) $\varphi(-t) = \overline{\varphi(t)}$.
iii) $|\varphi(t)| \leq 1$.
iv) $|\varphi(t+h) - \varphi(t)| \leq \mathbb{E}\left|e^{-ihX} - 1\right|$.
v) $\mathbb{E}e^{it(aX+b)} = e^{itb}\varphi(at)$.

Proposition 7.2.5. If X and Y are independent with characteristic functions φ_X and φ_Y, then $Z = X+Y$ has characteristic function

$$\varphi_Z(t) = \varphi_X(t)\varphi_Y(t). \tag{7.26}$$

Proof.

$$\varphi_Z(t) = \mathbb{E}[e^{it(X+Y)}] = \mathbb{E}[e^{itX}e^{itY}] = \mathbb{E}[e^{itX}]\mathbb{E}[e^{itY}] = \varphi_X(t)\varphi_Y(t). \tag{7.27}$$

□

Theorem 7.2.6. Common distributions have the following characteristic functions.

- $X \sim \text{Ber}(p)$: $\varphi(t) = 1 - p + pe^{it}$.
- $X \sim \text{Poisson}(\lambda)$: $\varphi(t) = \exp\{\lambda(e^{it} - 1)\}$.
- $X \sim \text{Exp}(\lambda)$: $\varphi(t) = \frac{\lambda}{\lambda - it}$.
- $X \sim \text{Gaussian}(\mu, \sigma^2)$: $\varphi(t) = e^{i\mu t - \sigma^2 t^2/2}$.

Proof. Exercise.
(*Treat i as a constant – we won't worry about the technical details of complex analysis here.*) □

Theorem 7.2.7. Let X be a random variable and consider a function $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}f(t, X) < \infty$ and $\frac{\partial}{\partial t} f(t, X) = f'(t, X)$ exists for all $t \in (a, b)$. Further, suppose there is a random variable Y with $\mathbb{E}Y < \infty$ and $|f'(t, X)| \leq Y$ a.s. for $t \in (a, b)$. Then, for all $t \in (a, b)$,

$$\frac{\partial}{\partial t}\mathbb{E}f(t, X) = \mathbb{E}f'(t, X). \tag{7.28}$$

Proof. First, observe that

$$f'(t, X) = \lim_{h \to 0} \frac{f(t + h, X) - f(t, X)}{h}, \tag{7.29}$$

so

$$\left| \frac{f(t + h, X) - f(t, X)}{h} \right| \leq Y \tag{7.30}$$

for small h. Then, using the dominated convergence theorem,

$$\begin{aligned} \frac{\partial}{\partial t}\mathbb{E}f(t, X) &= \lim_{h \to 0} \frac{\mathbb{E}f(t + h, X) - \mathbb{E}f(t, X)}{h} \\ &= \lim_{h \to 0} \mathbb{E}\frac{f(t + h, X) - f(t, X)}{h} \\ &= \mathbb{E} \lim_{h \to 0} \frac{f(t + h, X) - f(t, X)}{h} \\ &= \mathbb{E}f'(t, X). \end{aligned} \tag{7.31}$$

□

Proposition 7.2.8. If X is a random variable with $\mathbb{E}\,|X|^k < \infty$, then for $0 \leq j \leq k$,

$$\varphi_X^{(j)}(t) = \mathbb{E}[(iX)^j e^{itX}]. \tag{7.32}$$

Proof. This is proved by induction. When $j = 0$, this is just the definition of φ. Suppose it holds for some j. Then,

$$\begin{aligned} \varphi_X^{(j+1)}(t) &= \frac{\partial}{\partial t}\varphi_X^{(j)}(t) \\ &= \frac{\partial}{\partial t}\mathbb{E}[(iX)^j e^{itX}] \\ &= \mathbb{E}\left[(iX)^j \frac{\partial}{\partial t} e^{itX}\right] \\ &= \mathbb{E}[(iX)^{j+1} e^{itX}], \end{aligned} \tag{7.33}$$

where we used the previous proposition to swap the order of expectation and differentiation. □

Theorem 7.2.9 (Continuity Theorem). Let $\mu, \mu_1, \mu_2, \dots$ be probability measures with characteristic functions $\varphi, \varphi_1, \varphi_2, \dots$. Then, $\mu_n \Longrightarrow \mu$ if and only if $\varphi_n(t) \longrightarrow \varphi(t)$ for all $t \in \mathbb{R}$.

Proof. Beyond the scope of this course. □

Lemma 7.2.10. If X_n, Y_n are independent for all n and X, Y are independent with $X_n \Longrightarrow X$ and $Y_n \Longrightarrow Y$, then

$$X_n + Y_n \Longrightarrow X + Y. \tag{7.34}$$

Proof. Using characteristic functions,

$$\lim_{n\to\infty} \varphi_{X_n+Y_n}(t) = \lim_{n\to\infty} \varphi_{X_n}(t)\varphi_{Y_n}(t) = \varphi_X(t)\varphi_Y(t) = \varphi_{X+Y}(t). \tag{7.35}$$

Then apply the continuity theorem. □

Proposition 7.2.11 (Poisson Convergence Theorem). Suppose X_n is a sequence of random variables s.t. $X_n \sim \text{Bin}(n, p_n)$ with $np_n \to \lambda > 0$ as $n \to \infty$, then $X_n \Longrightarrow \text{Poisson}(\lambda)$.

Proof. The characteristic function of X_n is

$$\begin{aligned} \varphi_n(t) &= (1 - p_n + p_n e^{it})^n \\ &= \left(1 - \frac{np_n - np_n e^{it}}{n}\right)^n \\ &\triangleq \left(1 - \frac{a_n(t)}{n}\right)^n, \end{aligned} \tag{7.36}$$

where $a_n(t) = np_n - np_n e^{it}$. Note that $\lim_{n\to\infty} a_n(t) = \lambda(1 - e^{it})$ for any t, thus

$$\lim_{n\to\infty} \varphi_n(t) = \exp[\lambda(e^{it} - 1)] \triangleq \varphi(t) \tag{7.37}$$

for all $t \in \mathbb{R}$. But $\varphi(t)$ is the characteristic function of Poisson(λ). By the continuity theorem, $X_n \Longrightarrow \text{Poisson}(\lambda)$. □

7.3 Central Limit Theorem

Lemma 7.3.1. For any random variable with $\mathbb{E}\,|X|^k < \infty$ for $0 \le k \le m$,

$$\varphi_X(t) = \sum_{k=0}^{m} \frac{(it)^k}{k!} \mathbb{E} X^k + o(|t|^m). \tag{7.38}$$

Proof. Exercise. *Hint: use Taylor series error approximation.* □

Theorem 7.3.2. If $X_1, X_2, \dots$ are i.i.d. with $\mathbb{E} X_n = 0$ and $\mathbb{E} X_n^2 = \sigma^2 < \infty$,

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \Longrightarrow \text{Gaussian}(0, \sigma^2). \tag{7.39}$$

Proof. Using the previous lemma with $m = 2$,

$$\varphi_{X_i}(t) = 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2). \tag{7.40}$$

Thus,

$$\lim_{n\to\infty} \varphi_{Y_n}(t) = \lim_{n\to\infty} [\varphi_{X_i}(t/\sqrt{n})]^n = \lim_{n\to\infty} \left[1 - \frac{1}{2n}\sigma^2 t^2 + o(t^2/n)\right]^n = e^{-\sigma^2 t^2/2}. \tag{7.41}$$

This is the characteristic function of Gaussian$(0, \sigma^2)$, so the result follows from the continuity theorem. □

7.4 Exercises

Exercise 7.1. A Cauchy random variable has density defined by $f(x) = \frac{1}{\pi(1+x^2)}$ for all $x \in \mathbb{R}$. Show that for all $t \neq 0$, the MGF $M_X(t) = \mathbb{E}e^{tX} = \infty$.
Note: this is just a calculus question to motivate why we use characteristic functions. I won't ask you something like this on an exam.

Exercise 7.2 (Durrett 3.3.9). Suppose $X_n \sim \text{Gaussian}(0, \sigma_n^2)$ and $X_n \Longrightarrow X$ for some random variable X. Show that $\sigma_n \longrightarrow \sigma$ for some $\sigma \in [0, \infty)$.

Exercise 7.3. Prove Scheffé's Theorem for discrete pmfs instead of densities.

Exercise 7.4. Find an example of random variables X_n with densities f_n such that $X_n \Longrightarrow \text{Unif}(0,1)$ but $\{x : f_n(x) \longrightarrow 1\} = \emptyset$.

Exercise 7.5. Prove that if $\mathbb{P}_n \Longrightarrow \mathbb{P}$ and $\mathbb{P}_n \Longrightarrow \mathbb{P}'$, then $\mathbb{P} = \mathbb{P}'$. *Hint: use the Portmanteau lemma.*

Exercise 7.6. Prove that if $F_n \Longrightarrow F$ and x is such that there is at most one $a \in \mathbb{R}$ with $F(a) = x$, then $F_n^{-1}(x) \longrightarrow F^{-1}(x)$. *Hint: For any $\varepsilon > 0$, you can choose a y such that F is continuous at y (why?) and $F^{-1}(x) - \varepsilon < y < F^{-1}(x)$.*

Exercise 7.7. Prove the remainder of Theorem 7.1.8.

Exercise 7.8. Find an example such that $X_n \Longrightarrow X$ and $Y_n \Longrightarrow Y$ but $X_n + Y_n \not\Longrightarrow X + Y$.

Exercise 7.9. Prove Theorem 7.2.6.

Exercise 7.10. Prove Lemma 7.3.1.

Exercise 7.11. Let $X_1, X_2, \ldots$ be i.i.d. with characteristic function ψ. Show that if $\psi'(0) = ia$, then

$$\frac{S_n}{n} \xrightarrow{P} a.$$

LECTURE 8

STOCHASTIC PROCESS

8.1 Poisson Process

Definition 8.1.1 (Counting Process). Let $t \in [0, \infty)$, $\{N(t)\}_{t \geq 0}$ is called a *counting process* if

- $N(t) \in \mathbb{N}$,
- $s < t \Longrightarrow N(s) \leq N(t)$,
- For $s < t$, the increment $N(s, t] \triangleq N(t) - N(s)$ is the number of events within $(s, t]$.

Definition 8.1.2 (Simple Counting Process). A counting process $\{N(t)\}_{t \geq 0}$ is called *simple* if for all $t \in [0, \infty)$,

$$\mathbb{P}\left(\lim_{\Delta t \downarrow 0} N(t, t + \Delta t] \leq 1\right) = 1. \tag{8.1}$$

Remark. For simplicity, we will just refer to simple counting process as counting process.

Definition 8.1.3 (Poisson process). Let $\{N(t)\}_{t \geq 0}$ be a counting process. Suppose $N(t)$ satisfies the following:

- $N(0) = 0$ a.s.
- (*independent increments*) $\forall 0 \leq t_1 < t_2 < \cdots < t_m$, $N(t_1, t_2], \ldots, N(t_{m-1}, t_m]$ are independent.
- $N(s, t] \sim \text{Poisson}(\lambda(t - s))$ for all $0 \leq s < t$.

Then $\{N(t)\}_{t \geq 0}$ is called a Poisson process with intensity λ.

Proposition 8.1.4. A Poisson process has stationary increments, i.e., the distribution of $N(t, t + h]$ only depends on h for all $t \geq 0$.

Proof. By definition. □

Definition 8.1.5. A function $f : \mathbb{R} \to \mathbb{R}$ is called $o(h)$ for $h \to 0$, if

$$\lim_{h \to 0} \frac{f(h)}{h} = 0. \tag{8.2}$$

Theorem 8.1.6 (Alternative Definition of Poisson Process). The counting process $\{N(t)\}_{t \geq 0}$ is a Poisson process with intensity $\lambda > 0$ if

- $N(0) = 0$ a.s.
- The process has independent and stationary increments.
- $\mathbb{P}(N(h) = 1) = \lambda h + o(h),\ \forall t, h > 0.$
- $\mathbb{P}(N(h) \geq 2) = o(h),\ \forall t, h > 0.$

Proof. The only thing we need to show based on the above alternative definition is $N(s,t] \sim \text{Poisson}(\lambda(t-s))$ for all $0 \leq s < t$. Since $N(t)$ is stationary, it suffices to show that $N(t) \sim \text{Poisson}(\lambda t)$. For $n \in \mathbb{N}$, write $I_{n,k} = ((k-1)t/n, kt/n]$. Let $X_{k,n} = \mathbb{I}(N(I_{n,k}) > 0)$. Since $\{N(t)\}_{t\geq 0}$ has independent and stationary increments, $X_{k,n}$'s are i.i.d. Bernoulli random variables with probability of success $p_n = \mathbb{E}[X_{k,n}] = \frac{\lambda t}{n} + o\left(\frac{t}{n}\right)$. Thus, $\sum_{k=1}^n X_{k,n} \sim \text{Bin}(n, p_n)$. Now notice that since $\{N(t)\}_{t\geq 0}$ is simple, $\sum_{k=1}^n X_{k,n} \to N(t)$ a.s., and $np_n \to \lambda t$ as $n \to \infty$. Thus, by the Poisson convergence theorem, we have $N(t) \sim \text{Poisson}(\lambda t)$ as required. □

Theorem 8.1.7. Let $X_1, X_2, \ldots$ be i.i.d.Exp(λ). Define $S_n = \sum_{i=1}^n X_i$, and

$$N(t) = \max\{n : S_n \leq t\},\ t \geq 0, \tag{8.3}$$

where $N(t) = 0$ if $t < S_1 = X_1$. $\{N(t)\}_{t\geq 0}$ is a Poisson process with intensity λ. X_i's are called the waiting time.

Proof. i) $N(0) = 0$ trivial. ii) We first show $N(t, t+h] \sim \text{Poisson}(\lambda h)$ for any $t, h \geq 0$. For any $k \in \mathbb{N}^+$, $S_k \sim \text{Gamma}(k, \lambda)$. By law of total probability, we have

$$\begin{aligned}
\mathbb{P}(N(t) = k) &= \mathbb{P}\left(S_k \leq t < S_{k+1}\right) \\
&= \int_0^t \mathbb{P}\left(X_{k+1} > t - s \mid S_k = s\right) f\left(S_k = s\right) \mathrm{d}s \\
&= \int_0^t \mathbb{P}\left(X_{k+1} > t - s\right) f\left(S_k = s\right) \mathrm{d}s \\
&= \int_0^t e^{-\lambda(t-s)} \frac{\lambda^k}{(k-1)!} s^{k-1} e^{-\lambda s} \mathrm{d}s \\
&= e^{-\lambda t} \frac{\lambda^k}{(k-1)!} \int_0^t s^{k-1} \mathrm{d}s \\
&= e^{-\lambda t} \frac{(\lambda t)^k}{k!}.
\end{aligned} \tag{8.4}$$

If $k = 0$, then

$$\mathbb{P}(N(t) = 0) = \mathbb{P}(X_1 \geq t) = e^{-\lambda t}. \tag{8.5}$$

Thus, $N(t) \sim \text{Poisson}(\lambda t)$ for all $t \geq 0$.

Now fix $t > 0$, we have $S_{N(t)} \leq t < S_{N(t)+1}$. The waiting time from t to the first event after t occurs at $S_{N(t)+1} - t$, while the next event is at $X_{N(t)+2}$, and so on. Define

$$X_1^{(t)} = S_{N(t)+1} - t,\ X_i^{(t)} = X_{N(t)+i},\ i \geq 2. \tag{8.6}$$

$\{X_i^{(t)}\}_{i\geq 1}$ is thus the waiting time after time t. Now notice that $N(t, t+h] \geq k$ is equivalent to $S_{N(t)+k} \leq t + h$, but $S_{N(t)+k} \leq t + h$ is the same as $S_k^{(t)} \triangleq \sum_{i=1}^k X_i^{(t)} \leq h$. Therefore,

$$N(t, t+h] = \max\{n : S_n^{(t)} \leq h\}. \tag{8.7}$$

Thus, for any $t, h \geq 0$, $N(t, t+h]$ is constructed in the exact same way as $N(h)$ using $\{X_i^{(t)}\}_{i\geq 1}$. Let $n \geq 0$, $j \geq 1$, and $s_i \geq 0$ for $i = 1, \dots, j$, since X_i's are independent,

$$\begin{aligned}\mathbb{P}\left(N(t) = n, X_1^{(t)} > s_1, \dots, X_j^{(t)} > s_j\right) &= \mathbb{P}\left(S_n \leq t < S_{n+1}, S_{n+1} - t > s_1, \dots, X_{n+j} > y_j\right) \\ &= \mathbb{P}(S_n \leq t < S_{n+1}, S_{n+1} - t > s_1) \prod_{i=2}^{j} \mathbb{P}(X_{n+i} > s_i) \\ &= \mathbb{P}(S_n \leq t, X_{n+1} > t + s_1 - S_n) \prod_{i=2}^{j} \mathbb{P}(X_{n+i} > s_i) \\ &= \mathbb{P}(S_n \leq t, X_{n+1} > t + s_1 - S_n) \prod_{i=2}^{j} e^{-\lambda s_i}.\end{aligned} \tag{8.8}$$

For $\mathbb{P}(S_n \leq t, X_{n+1} > t + s_1 - S_n)$, we have

$$\begin{aligned}\mathbb{P}(S_n \leq t, X_{n+1} > t + s_1 - S_n) &= \mathbb{E}[\mathbb{I}(S_n \leq t, X_{n+1} > t + s_1 - S_n)] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{I}(S_n \leq t, X_{n+1} > t + s_1 - S_n) \mid S_n]] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{I}(S_n \leq t)\mathbb{I}(X_{n+1} > t + s_1 - S_n) \mid S_n]] \\ &= \mathbb{E}[\mathbb{I}(S_n \leq t)\mathbb{E}[\mathbb{I}(X_{n+1} > t + s_1 - S_n) \mid S_n]] \\ &= \mathbb{E}[\mathbb{I}(S_n \leq t)\mathbb{P}(X_{n+1} > t + s_1 - S_n \mid S_n)] \\ &= e^{-\lambda s_1}\mathbb{E}[\mathbb{I}(S_n \leq t)\mathbb{P}(X_{n+1} > t - S_n \mid S_n)] \\ &= e^{-\lambda s_1}\mathbb{E}[\mathbb{E}[\mathbb{I}(S_n \leq t, X_{n+1} > t - S_n) \mid S_n]] \\ &= e^{-\lambda s_1}\mathbb{P}(S_n \leq t, X_{n+1} > t - S_n) \\ &= e^{-\lambda s_1}\mathbb{P}(S_n \leq t < S_{n+1}) = e^{-\lambda s_1}\mathbb{P}(N(t) = n).\end{aligned} \tag{8.9}$$

Thus,

$$\begin{aligned}\mathbb{P}\left(N(t) = n, X_1^{(t)} > s_1, \dots, X_j^{(t)} > s_j\right) &= \mathbb{P}(N(t) = n) \prod_{i=1}^{j} e^{-\lambda s_i} \\ &= \mathbb{P}(N(t) = n) \prod_{i=1}^{j} \mathbb{P}(X_i > s_i) \\ &= \mathbb{P}(N(t) = n)\mathbb{P}(X_1 > s_1, \dots, X_j > s_j)\end{aligned} \tag{8.10}$$

By law of total probability, we then have

$$\mathbb{P}\left(X_1^{(t)} > s_1, \dots, X_j^{(t)} > s_j\right) = \mathbb{P}(X_1 > s_1, \dots, X_j > s_j). \tag{8.11}$$

The LHS of the above completely determines the distribution of $\{X_i^{(t)}\}_{i=1}^{j}$ for any $j \geq 1$, which is independent of t, and so is $S_n^{(t)}$. Hence, $N(t, t+h] \overset{d}{=} N(h) \sim \text{Poisson}(\lambda h)$.

iii) For the independent increments, consider the event $\cap_{i=1}^{u}\{N(s_i) = m_i\}$ where $j = m_u + 1$. Define

$$H = \left\{(x_1, \dots, x_j) \in \mathbb{R}^j : \sum_{l=1}^{m_i} x_l \leq s_i < \sum_{l=1}^{m_i+1} x_l,\ 1 \leq i \leq u\right\}. \tag{8.12}$$

Then $\cap_{i=1}^{u}\{N(s_i) = m_i\} = \{(X_1, \ldots, X_j) \in H\}$. Similarly, we have $\cap_{i=1}^{u}\{N(t, t+s_i] = m_i\} = \{(X_1^{(t)}, \ldots, X_j^{(t)}) \in H\}$ for any $t \geq 0$. By Equation 8.10,

$$\begin{aligned}\mathbb{P}\left(N(t) = n, \bigcap_{i=1}^{u}\{N(t, t+s_i] = m_i\}\right) &= \mathbb{P}\left(N(t) = n, (X_1^{(t)}, \ldots, X_j^{(t)}) \in H\right) \\ &= \mathbb{P}(N(t) = n)\mathbb{P}\left((X_1, \ldots, X_j) \in H\right) \\ &= \mathbb{P}(N(t) = n)\mathbb{P}\left(\bigcap_{i=1}^{u}\{N(s_i) = m_i\}\right),\end{aligned} \tag{8.13}$$

and thus

$$\mathbb{P}\left(\bigcap_{i=1}^{u}\{N(t, t+s_i] = m_i\}\right) = \mathbb{P}\left(\bigcap_{i=1}^{u}\{N(s_i) = m_i\}\right). \tag{8.14}$$

Now for any $k \geq 0$, and $0 = t_0 < t_1 < \cdots < t_k$ and $(n_1, \ldots, n_k) \in \mathbb{N}^k$, we prove

$$\mathbb{P}\left(\bigcap_{i=1}^{k}\{N(t_{i-1}, t_i] = n_i\}\right) = \prod_{i=1}^{k}\mathbb{P}(N(t_{i-1}, t_i] = n_i). \tag{8.15}$$

by induction.

For $k = 2$, setting $u = 1$ in Equation 8.13 and the fact that $N(t_1, t_2]$ has the same distribution as $N(t_2 - t_1)$ gives the result.

Suppose for some $k > 2$,

$$\mathbb{P}\left(\bigcap_{i=1}^{k}\{N(t_{i-1}, t_i] = n_i\}\right) = \prod_{i=1}^{k}\mathbb{P}(N(t_{i-1}, t_i] = n_i). \tag{8.16}$$

For $k+1$, we have

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i=1}^{k+1}\{N(t_{i-1}, t_i] = n_i\}\right) &= \mathbb{P}\left(N(t_1) = n_1, \bigcap_{i=1}^{k}\{N(t_i, t_{i+1}] = n_{i+1}\}\right) \\ &= \mathbb{P}(N(t_1) = n_1)\mathbb{P}\left(\bigcap_{i=1}^{k}\{N(t_i - t_1, t_{i+1} - t_1] = n_{i+1}\}\right) \\ &= \mathbb{P}(N(t_1) = n_1)\prod_{i=1}^{k}\mathbb{P}(N(t_i - t_1, t_{i+1} - t_1] = n_{i+1}) \\ &= \mathbb{P}(N(t_1) = n_1)\prod_{i=1}^{k}\mathbb{P}(N(t_i, t_{i+1}] = n_{i+1}) \\ &= \mathbb{P}(N(t_1) = n_1)\prod_{i=2}^{k+1}\mathbb{P}(N(t_{i-1}, t_i] = n_i) \\ &= \prod_{i=1}^{k+1}\mathbb{P}(N(t_{i-1}, t_i] = n_i).\end{aligned} \tag{8.17}$$

□

Remark (Stochastic Process). The Poisson process we have discussed so far is an example of a stochastic process, that is, a collection of random variables indexed by a parameter representing time. In the case of Poisson process, time is continuous. In other cases, it can be discrete, i.e., time progresses in jumps.

Moreover, based on the waiting time characterization of the Poisson process, we can see that it also has the renewal property. This means that we can cut off the process at any time $t > 0$ and forget about what happened before t and the process starting at time t will be exactly the same as if it started at 0. This is also called memoryless property of the Poisson process, which is a product from the construction of Poisson process using i.i.d.exponential random variables.

8.2 Properties of Poisson Process

Theorem 8.2.1 (Superposition of Poisson Process). Consider two independent Poisson processes $\{N_1(t)\}_{t\geq 0}$ with intensity λ_1 and $\{N_2(t)\}_{t\geq 0}$ with intensity λ_2. The combined process $\{N(t)\}_{t\geq 0}$ where $N(t) = N_1(t) + N_2(t)$ is another Poisson process with intensity $\lambda_1 + \lambda_2$.

Proof. i) $N(0) = 0$ trivial.

ii) For independent increments, denote $N_i^1 = N_1(t_{i-1}, t_i]$, $N_i^2 = N_2(t_{i-1}, t_i]$. Notice that for $0 = t_0 < t_1 < \cdots < t_k$, $N_i = N(t_{i-1}, t_i] = N_i^1 + N_i^2$ for $1 \leq i \leq k$. But $N_i^1 + N_i^2$ are independent for all $1 \leq i \leq k$ since $N_1(t)$ and $N_2(t)$ are Poisson processes. Thus, N_i's are independent for all $1 \leq i \leq k$.

iii) For simplicity, denote $N = N(t, t+h]$, $N_1 = N_1(t, t+h]$, and $N_2 = N_2(t, t+h]$. Since N_1 and N_2 are independent, $N_1 \sim \text{Poisson}(\lambda_1 t)$ and $N_2 \sim \text{Poisson}(\lambda_2 t)$, $N = N_1 + N_2 \sim \text{Poisson}((\lambda_1 + \lambda_2)t)$. Hence, $N(t)$ is a Poisson process with intensity $\lambda_1 + \lambda_2$. □

Theorem 8.2.2 (Thinning of Poisson Process). Let $\{N(t)\}_{t\geq 0}$ be a Poisson process with intensity $\lambda_1 + \lambda_2$. Suppose $N(t)$ is the superposition of two independent Poisson processes $\{N_1(t)\}_{t\geq 0}$ and $\{N_2(t)\}_{t\geq 0}$ with intensities λ_1 and λ_2 respectively. Then given $N(t) > 0$,

$$\mathbb{P}(\text{An event in } N(t) \text{ comes from } N_i(t) \mid N(t) > 0) = \frac{\lambda_i}{\lambda_1 + \lambda_2}. \tag{8.18}$$

Proof. Let A ={An event in $N(t)$ comes from $N_1(t)$}, $A_n = A \cap \{N(t) = n\}$, for $n \in \mathbb{N}^+$. First, consider

$$\begin{aligned}
\mathbb{P}(A_n \mid N(t) > 0) &= \frac{\mathbb{P}(N_1(t) + N_2(t) = n, \text{An event is from } N_1(t))}{\mathbb{P}(N(t) > 0)} \\
&= \left(1 - e^{-(\lambda_1+\lambda_2)t}\right)^{-1} \sum_{s=0}^{n} \mathbb{P}(\text{An event is from } N_1(t), N_1(t) = s, N_2(t) = n - s) \\
&= \left(1 - e^{-(\lambda_1+\lambda_2)t}\right)^{-1} \sum_{s=0}^{n} \mathbb{P}(\text{An event is from } N_1(t) \mid N_1(t) = s, N_2(t) = n - s) \times \\
&\qquad \mathbb{P}(N_1(t) = s, N_2(t) = n - s).
\end{aligned} \tag{8.19}$$

Notice that $\mathbb{P}(\text{An event is from } N_1(t) \mid N_1(t) = s, N_2(t) = n - s) = s/n$, thus,

$$\begin{aligned}
\mathbb{P}(A_n \mid N(t) > 0) &= \left(1 - e^{-(\lambda_1+\lambda_2)t}\right)^{-1} \sum_{s=0}^{n} \frac{s}{n} \mathbb{P}(N_1(t) = s)\mathbb{P}(N_2(t) = n - s) \\
&= \left(1 - e^{-(\lambda_1+\lambda_2)t}\right)^{-1} e^{-(\lambda_1+\lambda_2)t} \frac{\lambda_1 t}{n!} \sum_{s=1}^{n} \binom{n-1}{s-1} (\lambda_1 t)^{s-1} (\lambda_2 t)^{n-s} \\
&= \left(1 - e^{-(\lambda_1+\lambda_2)t}\right)^{-1} e^{-(\lambda_1+\lambda_2)t} \frac{\lambda_1 t}{n!} \sum_{j=0}^{n-1} \binom{n-1}{j} (\lambda_1 t)^{j} (\lambda_2 t)^{n-1-j} \\
&= \left(1 - e^{-(\lambda_1+\lambda_2)t}\right)^{-1} e^{-(\lambda_1+\lambda_2)t} \frac{\lambda_1 t}{n!} (\lambda_1 t + \lambda_2 t)^{n-1},\ n > 0.
\end{aligned} \tag{8.20}$$

Now,

$$\begin{aligned}
\mathbb{P}(A \mid N(t) > 0) &= \sum_{n=1}^{\infty} \mathbb{P}(A_n \mid N(t) > 0) \\
&= \left(1 - e^{-(\lambda_1+\lambda_2)t}\right)^{-1} e^{-(\lambda_1+\lambda_2)t} \lambda_1 t \sum_{n=1}^{\infty} \frac{(\lambda_1 t + \lambda_2 t)^{n-1}}{n!} \\
&= \left(1 - e^{-(\lambda_1+\lambda_2)t}\right)^{-1} e^{-(\lambda_1+\lambda_2)t} \frac{\lambda_1}{\lambda_1 + \lambda_2} \sum_{n=1}^{\infty} \frac{(\lambda_1 t + \lambda_2 t)^{n}}{n!} \\
&= \left(1 - e^{-(\lambda_1+\lambda_2)t}\right)^{-1} e^{-(\lambda_1+\lambda_2)t} \frac{\lambda_1}{\lambda_1 + \lambda_2} \left(e^{(\lambda_1+\lambda_2)t} - 1\right) \\
&= \frac{\lambda_1}{\lambda_1 + \lambda_2}.
\end{aligned} \tag{8.21}$$

□

Proposition 8.2.3. Let $\{N(t)\}_{t \geq 0}$ be a Poisson process with intensity λ. Given $N(t) = 1$, and let the location of the event in $(0, t]$ be S, then $S \sim \text{Unif}(0, t)$.

Proof.

$$\begin{aligned}
\mathbb{P}(S \leq s \mid N(t) = 1) &= \frac{\mathbb{P}(S \leq s, N(t) = 1)}{\mathbb{P}(N(t) = 1)} \\
&= \frac{\mathbb{P}(N(s) = 1, N(s, t] = 0)}{\mathbb{P}(N(t) = 1)} \\
&= \frac{\mathbb{P}(N(s) = 1)\mathbb{P}(N(s, t] = 0)}{\mathbb{P}(N(t) = 1)} \\
&= \frac{e^{-\lambda s} s e^{-\lambda(t-s)}}{e^{-\lambda t} t} = s/t.
\end{aligned} \tag{8.22}$$

□

Proposition 8.2.4. Let $\{N(t)\}_{t \geq 0}$ be a Poisson process with intensity λ. Given $N(t) = n$, and let the location of the n events in $(0, t]$ be $S_1 < \cdots < S_n$, then the joint p.d.f. of $(S_1, \ldots, S_n)$ is

$$f(s_1, \ldots, s_n) = \frac{n!}{t^n},\ 0 < s_1 < \cdots < s_n < t. \tag{8.23}$$

Proof. Exercise. □

Remark. The previous proposition means that condition on $N(t)$, the locations of the events are distributed as the order statistics of $N(t)$ number of uniform random variables in $(0, t]$.

Lemma 8.2.5 (Binomial Process). Let $\{N(t)\}_{t\geq 0}$ be a Poisson process with intensity λ. Given $N(t) = n$, then $N(s) \sim \text{Bin}(n, s/t)$ for $s \in (0, t]$, and $\{N(s)\}_{0\leq s\leq t}$ is called a Binomial process.

Proof. For any $0 \leq k \leq n$,

$$\begin{aligned}
\mathbb{P}(N(s) = k \mid N(t) = n) &= \frac{\mathbb{P}(N(s) = k, N(t) = n)}{\mathbb{P}(N(t) = n)} \\
&= \frac{\mathbb{P}(N(s) = k, N(s, t] = n - k)}{\mathbb{P}(N(t) = n)} \\
&= \frac{\mathbb{P}(N(s) = k)\mathbb{P}(N(s, t] = n - k)}{\mathbb{P}(N(t) = n)} \\
&= \frac{e^{-\lambda s}(\lambda s)^k/k!\, e^{-\lambda(t-s)}(\lambda(t - s))^{n-k}/(n - k)!}{e^{-\lambda t}(\lambda t)^n/n!} \\
&= \binom{n}{k}\frac{(\lambda s)^k(\lambda(t - s))^{n-k}}{(\lambda t)^n} \\
&= \binom{n}{k}(s/t)^k(1 - s/t)^{n-k}.
\end{aligned} \tag{8.24}$$

□

8.3 Exercise

Exercise 8.1. Prove left-over exercises in the Chapter.

Exercise 8.2. Show that the minimum of exponential random variables is again exponential.

Exercise 8.3. Let $\{N(t)\}_{t\geq 0}$ be a Poisson process with intensity λ. Denote S_n the time of the n-th event in the process. Define $X_n = S_n - S_{n-1}$ with $X_0 = 0$. Show that $X_1, X_2, \dots$ are i.i.d.Exp(λ).

Exercise 8.4. A bus platform is now empty of passengers, and the next bus will depart in t minutes. Passengers arrive to the platform according to a Poisson process at rate λ. What is the expected waiting time of an arriving passenger?

Exercise 8.5. Let $\{N_1(t)\}_{t\geq 0}$ and $\{N_2(t)\}_{t\geq 0}$ be two independent Poisson processes with intensity λ_1 and λ_2 respectively. Show that for any $n \in \mathbb{N}^+$ and $t > 0$,

$$\mathbb{P}(\text{the last event in } (0,t] \text{ is from } N_2 \mid N_1(t) + N_2(t) = n) = \frac{\lambda_2}{\lambda_1 + \lambda_2}. \tag{8.25}$$

Exercise 8.6. Let $\{N(t)\}_{t\geq 0}$ be a Poisson process with intensity λ. Set $A_t = t - S_{N(t)}$ the time back to the most recent event, and $B_t = S_{N(t)+1} - t$ the time forward to the next event. Show that A_t and B_t are independent and that $B_t \sim \text{Exp}(\lambda)$, and A_t has distribution $\min\{\text{Exp}(\lambda), t\}$. That is

$$\mathbb{P}(A_t \leq x) = \begin{cases} 0, \ x < 0 \\ 1 - e^{-\lambda x}, \ 0 \leq x < t, \\ 1, \ x \geq t. \end{cases} \tag{8.26}$$

Exercise 8.7. Suppose X and Y are Poisson random variables with parameters λ_1 and λ_2 respectively. Show that

$$|\mathbb{P}(X = n) - \mathbb{P}(Y = n)| \leq |\lambda_1 - \lambda_2|. \tag{8.27}$$

Exercise 8.8 (Spatial Poisson Process). Point processes can also be generalized to spatial scenario where the indexing parameter for events is changed from time t to a spatial location s. The main difference is that for spatial point processes, events in the process no longer has ordering as in the case of Poisson process w.r.t. time. Specifically, Let $\mathbf{X}$ be a spatial point process in $\mathbb{R}^d$, then $\mathbf{X}$ consists of a collection of random point locations in $\mathbb{R}^d$. Moreover, $\mathbf{X}$ is called a spatial Poisson process with intensity $\lambda > 0$ if

- If $A \subset \mathbb{R}^d$ with $|A| < \infty$, where $|A|$ is the volume of A, let $N(A)$ be the number of points from $\mathbf{X}$ that are within A, then $N(A) \sim \text{Poisson}(\lambda|A|)$.
- For any $A, B \subset \mathbb{R}^d$ s.t. $A \cap B = \emptyset$, $N(A)$ and $N(B)$ are independent.

Suppose that $\mathbf{X}$ is a spatial Poisson process in $\mathbb{R}^d$ with intensity $\lambda > 0$, let $s \in \mathbb{R}^d$ be a randomly chosen fixed location. Show that

- if $d_1(s, \mathbf{X})$ is the (Euclidean) distance from the nearest neighbor in $\mathbf{X}$ to s, then

$$\mathbb{P}(d_1(s, \mathbf{X}) > x) = \exp\left(-B_d x^d\right), \tag{8.28}$$

  where $B_d = \frac{\lambda \pi^{d/2}}{\Gamma(d/2+1)}$.

- if $d_k(s, \mathbf{x})$ is the distance from the k-th nearest neighbor in $\mathbf{X}$ to s, derive

$$\mathbb{P}(d_k(s, \mathbf{X}) > x). \tag{8.29}$$

- Prove the superposition and thinning properties of spatial Poisson process.

Exercise 8.9. Let $X_1, X_2, \ldots$ be i.i.d.Exp(λ) random variables. Suppose $Y_1, Y_2, \ldots$ are i.i.d.Bernoulli(p) random variables, $(X_1, X_2, \ldots)$ and $(Y_1, Y_2, \ldots)$ are independent. Let $S_n = \sum_{i=1}^n Y_i X_i$ and

$$N(t) = \max\{n : S_n \leq t\}, \tag{8.30}$$

with $N(t) = 0$ for $t < S_1 = Y_1 X_1$. Show that $\{N(t)\}_{t \geq 0}$ is a Poisson process with intensity $p\lambda$.