STA 347 Lecture Notes

William He

Contents

1 Probability Measures, Independence and Conditional Probability

1.1 Probability Measures

Definition 1.1. *A sample space Ω is any non-empty set.*

Remark 1.2. *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 1.3. *We can have*

1. $\Omega = [0, 1]$
2. $\Omega = \{H, T\}$*, a coin flip*
3. $\Omega = \{$*all possible location of sand grain in the dessert of Arakis*$\}$

Definition 1.4. *Given a sample space Ω, an event is a subset $E \subseteq \Omega$.*

Remark 1.5. *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 $P(E) \in [0, 1]$. As we change the event E, the corresponding $P(E)$ will change, but every event E only has one value of $P(E)$ associated with it.*

Based on the above observation, modern probability theory formulates the notion of probability as a function $P(\cdot)$. 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., $E \subseteq \Omega$, 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) $E \subseteq Q$ 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 P.

Example 1.6.
- *Given $\Omega = [0, 1]$, $E = \mathbb{Q} \cap [0, 1]$. $P(E)$ would be the probability of whether the number selected from the sample space is rational. ($P(E) = 0$)*
- *$P(E^c)$ would be the probability of whether the number selected from the sample space is irrational. ($P(E) = 1$)*

Definition 1.7. *Given $\Omega, \mathcal{P}(\Omega)$ as the power set of Ω. $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is called a σ-algebra/σ-field on Ω, if and only if*

1. $\Omega \in \mathcal{F}$
2. *(Closed in Complements) If $E \in F$ then $E^c \in F$*
3. *(Closed in Countable Union) If $E_1, E_2, \ldots \in F$, 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 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 $P(\cdot)$ 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 $\mathcal{P}(\Omega)$ that we want it to satisfy.

Example 1.8. *$\mathcal{F} = \{\emptyset, \Omega\}$ for any Ω is a $\sigma-$algebra.*

Definition 1.9 (General Measure)**.** *Given Ω, a σ-algebra $\mathcal{F}$ on Ω. $\mu : \mathcal{F} \to [0, \infty)$ is a measure defined on $\mathcal{F}$ if*

1. *(non-negative) $\forall E \in \mathcal{F}, \mu(E) \geq 0$*
2. *(empty set has zero measure) $\mu(\emptyset) = 0$*
3. *(countably additive) $\forall E_1, E_2, \ldots \in \mathcal{F}$ s.t. $E_i \cap E_j = \emptyset, \forall i \neq j$, then*

$$\mu(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i)$$

We call $(\Omega, \mathcal{F}, \mu)$ a measurable space.

Definition 1.10 (Probability Measure)**.** *Given Ω, a σ-algebra $\mathcal{F}$ on Ω. $P : \mathcal{F} \to [0, 1]$ is a probability measure defined on $\mathcal{F}$ if*

1. *$\forall E \in \mathcal{F}, P(E) \in [0, 1]$*
2. *$P(\Omega) = 1$*
3. *$\forall E_1, E_2, \ldots \in \mathcal{F}$ s.t. $E_i \cap E_j = \emptyset, \forall i \neq j$, then $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$*

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

Remark 1.11. *Note that by the third condition, take $E_1 = \Omega, E_2 = \emptyset$, we have*

$$P(\bigcup_{i=1}^{\infty} E_i) = P(\Omega) = P(\Omega) + P(\emptyset) = \sum_{i=1}^{\infty} P(E_i)$$

This gives $P(\emptyset) = 0$ automatically.

Note also that we cannot change the second criteria to $P(\emptyset) = 0$, as we cannot derive $P(\Omega) = 1$ in this case.

Example 1.12. *Consider coin flipping twice, we have*

$$\Omega = \{HH, HT, TH, TT\}$$

We can define $\mathcal{F} = \mathcal{P}(\Omega)$.

We can also define (assuming the coin is fair) the probability measure satisfies $P(HH) = P(HT) = P(TH) = P(TT) = \frac{1}{4}$, where the rest follows the definition of a probability measure. In this case, we have $(\Omega, \mathcal{F}, P)$ is a probability space.

Remark 1.13. *For the general measure in Definition 1.9, it is a mathematically rigorous description and generalisation 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 $P(\Omega)$ upon which to define our P? The reason is that if we want P to satisfy the above axioms on the power set $P(\Omega)$, such 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 $(\Omega, \mathcal{F}, P)$ for accuracy.

Example 1.14 (Lebesgue/uniform measure on interval)**.** *Let $\Omega = [L, U]$ where $L < U \in \mathbb{R}$.*
Note that in this case, we are not able to set $\mathcal{F} = \mathcal{P}(\Omega)$
However, we can define $\mathcal{F} = \mathcal{B}(\Omega)$ (Borel $\sigma-$algebra).
We can now define $P([a, b]) = P((a, b]) = P([a, b)) = P((a, b)) = \frac{b-a}{U-L}$ for all $L \leq a \leq b \leq U$.
In this case, $(\Omega, \mathcal{F}, P)$ is a probability space.

Proof. We will assume this is true in class. □

Remark 1.15. *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 via the set $\{(-\infty, a] : a \in \mathbb{R}\}$.*

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 1.16 (Counting/Uniform measure on finite space)**.** *Suppose $\Omega = \{x_1, \ldots, x_n\}$ be finite.*
We can define $\mathcal{F} = \mathcal{P}(\Omega)$.
We can also define the probability measure as $P(A) = \frac{|A|}{|\Omega|}, \forall A \in \mathcal{F}$.
In this case, $(\Omega, \mathcal{F}, P)$ is also a probability space.

Proposition 1.17.

1. $\forall E \in \mathcal{F}, P(E^c) = 1 - P(E)$
2. $P(\emptyset) = 0$
3. $\forall E, F \in \mathcal{F}$ *s.t.* $E \subseteq F, P(E) \leq P(F)$
4. $\forall E, F \in \mathcal{F}, P(E \cup F) = P(E) + P(F) - P(E \cap F)$
5. $\forall E, F \in \mathcal{F}, P(E \cap F) = P(E) + P(F) - P(E \cup F)$
6. $\forall E, F \in \mathcal{F}, P(F \cap E^c) = P(F) - P(F \cap E)$

Proof. 1. Given any $E \in \mathcal{F}$, by the definition of a σ-algebra we have $E^c \in \mathcal{F}$, showing $P(E^c)$ is well-defined. Moreover, we know that $E \cap E^c = \emptyset$ and $E \cup E^c = \Omega$. Hence, we have

$$\begin{aligned} 1 &= P(\Omega) \qquad \text{[By Definition 1.10 Condition 2]} \\ &= P(E \cup E^c) \\ &= P(E) + P(E^c) \qquad \text{[By Definition 1.10 Condition 3]} \end{aligned}$$

This gives $P(E^c) = 1 - P(E)$ for arbitrary $E \in \mathcal{F}$.

2. By Part 1, take $E = \Omega$, we get $E^c = \emptyset$, showing $P(\emptyset)$ is well-defined. Hence,

$$\begin{aligned} P(\emptyset) &= P(E^c) \\ &= 1 - P(E) \\ &= 1 - P(\Omega) \\ &= 1 - 1 \\ &= 0 \end{aligned}$$

3. Given any $E, F \in \mathcal{F}$ such that $E \subseteq F$, we can write $F = (F \cap E) \cup (F \cap E^c) = E \cup (F \cap E^c)$. Note that $E \in \mathcal{F}$. By the definition of a σ-algebra we have $E^c \in \mathcal{F}$. Moreover, this also means $F \cup E^c \in \mathcal{F}$, showing $P(F \cup E^c)$ is well-defined.

   Since we also know that $E \cap (F \cap E^c) = \emptyset$, we have

$$\begin{aligned} P(F) &= P(E \cup (F \cap E^c)) \\ &= P(E) + P(F \cap E^c) \qquad \text{[By Definition 1.10 Condition 3]} \\ &\geq P(E) \qquad \text{[By Definition 1.10 Condition 1]} \end{aligned}$$

4. Given any $E, F \in \mathcal{F}$, note that we can write

$$\begin{aligned} E \cup F &= ((E \cup F) \cap E) \cup ((E \cup F) \cap E^c) \\ &= ((E \cap E) \cup (F \cap E)) \cup ((E \cap E^c) \cup (F \cap E^c)) \\ &= (E \cup (F \cap E)) \cup (F \cap E^c) \\ &= E \cup (F \cap E^c) \end{aligned}$$

   By definition of a σ-algebra, we have

$$F \in \mathcal{F} \implies F^c \in \mathcal{F} \implies F^c \cup E \in \mathcal{F} \implies (F^c \cup E)^c = F \cap E^c \in \mathcal{F}$$

   showing $P(F \cap E^c)$ is well-defined.

   Moreover, we also know that $E \cap (F \cap E^c) = \emptyset$. Hence, by the definition of a probability measure we have

$$P(E \cup F) = P(E) + P(F \cap E^c)$$

   Now for $P(F \cap E^c)$, note that we can write

$$F = (F \cap E) \cup (F \cap E^c)$$

   where $(F \cap E) \cap (F \cap E^c) = \emptyset$.

   Moreover, by the definition of a σ-algebra we also have

$$E, F \in \mathcal{F} \implies E^c, F^c \in \mathcal{F} \implies F^c \cup E^c \in \mathcal{F} \implies (F^c \cup E^c)^c = F \cap E \in \mathcal{F}$$

   showing $P(F \cap E)$ is well-defined.

   Hence, by the definition of a probability measure, we have

$$P(F) = P(F \cap E) + P(F \cap E^c) \implies P(F \cap E^c) = P(F) - P(F \cap E)$$

   Therefore, we obtain

$$P(E \cup F) = P(E) + P(F \cap E^c) = P(E) + P(F) - P(F \cap E) = P(E) + P(F) - P(E \cap F)$$

   as required.

5. Straight forward from Part 4.

6. Given any $E, F \in \mathcal{F}$, it is clear that we have

$$F = (F \cap E) \cup (F \cap E^c)$$

where $(F \cap E) \cap (F \cap E^c) = \emptyset$.

Moreover, we have

$$E, F \in \mathcal{F} \implies E^c, F^c \in \mathcal{F} \implies F^c \cup E^c, F^c \cup E \in \mathcal{F} \implies (F^c \cup E^c)^c = F \cap E, (F^c \cup E)^c = F \cap E^c \in \mathcal{F}$$

showing $P(F \cap E), P(F \cap E^c)$ are well-defined.

Hence, by definition of a probability measure, we have

$$P(F) = P(F \cap E) + P(F \cap E^c) \implies P(F \cap E^c) = P(F) - P(F \cap E)$$

□

Lemma 1.18 (Sub-additivity)**.** *Let* $E_1, E_2, \ldots \in \mathcal{F}$

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

Proof. Let $F_1 = E_1, F_i = E_i \cap (\bigcup_{k=1}^{i-1} F_k)^c$ for all $i \geq 2$.
Note that we have the following

1. F_i's are disjoint
2. $\bigcup_{i=1}^{\infty} F_i = \bigcup_{i=1}^{\infty} E_i$
3. $\forall i, F_i \subseteq E_i$

We have

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} E_i\right) &= P\left(\bigcup_{i=1}^{\infty} F_i\right) && \text{[By condition 2]} \\ &= \sum_{i=1}^{\infty} P(F_i) && \text{[By definition of a probability measure]} \\ &\leq \sum_{i=1}^{\infty} P(E_i) && \text{[By Proposition 1.17 (2)]} \end{aligned}$$

□

Lemma 1.19. *Let* P *be a uniform measure on* $[0, 1]$*. Then* $P(E) = 0$ *for any countable set* $E \subseteq [0, 1]$*.*

Proof. We can write $E = \{x_i\}_{i=1}^{\infty}$. Choose $\epsilon > 0$. Define $E_i(\epsilon) = [x_i - \frac{\epsilon}{2^{-i}}, x_i + \frac{\epsilon}{2^{-i}}] \cap [0, 1]$
We notice that

$$E \subseteq \bigcup_{i=1}^{\infty} E_i(\epsilon)$$

We also have

$$\begin{aligned} P(E) &\leq P\left(\bigcup_{i=1}^{\infty} E_i(\epsilon)\right) && \text{[By Proposition 1.17 (2)]} \\ &\leq \sum_{i=1}^{\infty} P(E_i(\epsilon)) && \text{[By Lemma 1.18]} \\ &\leq \sum_{i=1}^{\infty} \frac{2\epsilon}{2^{-i}} && \text{[By definition of a uniform measure]} \\ &= 2\epsilon \end{aligned}$$

Since ϵ is arbitrary, this shows that $P(E) = 0$. □

Proposition 1.20. *Let $\Omega = \mathbb{N}$, there is no uniform probability measure on Ω.*

Proof. Suppose for a contradiction that there is a uniform probability measure on Ω where by definition, there exists some $c \in [0, 1]$ such that $P(n) = c$ for all $n \in \mathbb{N}$. Denote $A_n = \{n\}$ for all $n \in \mathbb{N}$, we know that $A_n \cap A_m = \emptyset$ for all $n \neq m$. Moreover

$$\bigcup_{i=1}^{\infty} A_i = \mathbb{N} = \Omega$$

Case 1: If $c = 0$, we have

$$\begin{aligned} P(\Omega) &= P\left(\bigcup_{i=1}^{\infty} A_i\right) \\ &= \sum_{i=0}^{\infty} P(A_i) && \text{[By Definition of a Probability Measure]} \\ &= \sum_{i=0}^{\infty} 0 \\ &= 0 \\ &\neq 1 \end{aligned}$$

Contradiction to the definition of a probability measure.
Case 2: If $c \in (0, 1]$, we have

$$\begin{aligned} P(\Omega) &= P\left(\bigcup_{i=1}^{\infty} A_i\right) \\ &= \sum_{i=0}^{\infty} P(A_i) && \text{[By Definition of a Probability Measure]} \\ &= \sum_{i=0}^{\infty} c \\ &\neq 1 \end{aligned}$$

Contradiction to the definition of a probability measure.
Hence, there is no uniform probability measure on Ω. □

Lemma 1.21 (Inclusion-Exclusion Formula)**.** *Take $E_1, E_2, \ldots, E_n \in \mathcal{F}$, we have*

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{a_1=1}^{n} P(E_{a_1}) - \sum_{a_1<a_2} P(E_{a_1}\cap E_{a_2}) + \sum_{a_1<a_2<a_3} P(E_{a_1}\cap E_{a_2}\cap E_{a_3}) + \ldots + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} E_i\right)$$

Proof. Note that we can write the expression as follows

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} (-1)^{i+1} \sum_{1\le a_1<\ldots<a_i\le n} P\left(\bigcap_{j=1}^{i} E_{a_j}\right)$$

We show by inducting on n.

Base cases:

$n = 1$: Given $E_1 \in \mathcal{F}$, we have indeed

$$\begin{aligned} P\left(\bigcup_{i=1}^{1} E_i\right) &= P(E_1) \\ &= (-1)^{1+1} P\left(\bigcap_{i=1}^{1} E_i\right) \\ &= (-1)^{1+1} \sum_{1\le a_1\le 1} P\left(\bigcap_{i=1}^{1} E_i\right) \\ &= \sum_{i=1}^{1} (-1)^{i+1} \sum_{1\le a_1<\ldots<a_i\le 1} P\left(\bigcap_{j=1}^{i} E_{a_j}\right) \end{aligned}$$

which is true

$n = 2$: Given $E_1, E_2 \in \mathcal{F}$, we have

$$\begin{aligned} P\left(\bigcup_{i=1}^{2} E_i\right) &= P(E_1 \cup E_2) \\ &= P(E_1) + P(E_2) - P(E_1 \cap E_2) \qquad \text{[By Proposition 1.17]} \\ &= \sum_{a_1=1}^{2} P(E_{a_1}) - \sum_{1\le a_1<a_2\le 2} P(E_{a_1} \cap E_{a_2}) \\ &= (-1)^{1+1} \sum_{1\le a_1\le 2} P(E_{a_1}) + (-1)^{2+1} \sum_{1\le a_1<a_2\le 2} P(E_{a_1} \cap E_{a_2}) \\ &= \sum_{i=1}^{2} (-1)^{i+1} \sum_{1\le a_1<\ldots<a_i\le 2} P\left(\bigcap_{j=1}^{i} E_{a_j}\right) \end{aligned}$$

which is true

Inductive step:

Suppose this is true for k, that is given $E_1, \ldots, E_k \in \mathcal{F}$

$$P\left(\bigcup_{i=1}^{k} E_i\right) = \sum_{i=1}^{k} (-1)^{i+1} \sum_{1\le a_1<\ldots<a_i\le 1} P\left(\bigcap_{j=1}^{i} E_{a_j}\right)$$

We now show this is also true for $k+1$. Given $E_1, \ldots, E_{k+1} \in \mathcal{F}$, we have

$$
\begin{aligned}
P\left(\bigcup_{i=1}^{k+1} E_i\right) &= P\left(\bigcup_{i=1}^{k} E_i \cup E_{k+1}\right) \\
&= P\left(\bigcup_{i=1}^{k} E_i\right) + P(E_{k+1}) - P\left(\bigcup_{i=1}^{k} E_i \cap E_{k+1}\right) \\
&= \sum_{i=1}^{k}(-1)^{i+1} \sum_{1\le a_1<\ldots<a_i\le k} P\left(\bigcap_{j=1}^{i} E_{a_j}\right) + P(E_{k+1}) - P\left(\bigcup_{i=1}^{k} E_i \cap E_{k+1}\right) \qquad \text{[By Inductive Hypothesix]} \\
&= \sum_{i=1}^{k}(-1)^{i+1} \sum_{1\le a_1<\ldots<a_i\le k} P\left(\bigcap_{j=1}^{i} E_{a_j}\right) + P(E_{k+1}) - P\left(\left(\bigcup_{i=1}^{k} E_i\right) \cap E_{k+1}\right) \\
&= \sum_{i=1}^{k}(-1)^{i+1} \sum_{1\le a_1<\ldots<a_i\le k} P\left(\bigcap_{j=1}^{i} E_{a_j}\right) + P(E_{k+1}) - P\left(\bigcup_{i=1}^{k} (E_i \cap E_{k+1})\right) \\
&= \sum_{i=1}^{k}(-1)^{i+1} \sum_{1\le a_1<\ldots<a_i\le k} P\left(\bigcap_{j=1}^{i} E_{a_j}\right) + P(E_{k+1}) - \sum_{i=1}^{k}(-1)^{i+1} \sum_{1\le a_1<\ldots<a_i\le k} P\left(\bigcap_{j=1}^{i} E_{a_j} \cap E_{k+1}\right) \\
&= \sum_{i=1}^{k}(-1)^{i+1} \sum_{1\le a_1<\ldots<a_i\le k} P\left(\bigcap_{j=1}^{i} E_{a_j}\right) + P(E_{k+1}) + \sum_{i=1}^{k}(-1)^{(i+1)+1} \sum_{1\le a_1<\ldots<a_i\le k} P\left(\bigcap_{j=1}^{i} E_{a_j} \cap E_{k+1}\right) \\
&= \left((-1)^{1+1} \sum_{1\le a_1\le k} P\left(\bigcap_{j=1}^{1} E_{a_j}\right) + P(E_{k+1})\right) + (-1)^{(k+1)+1} \sum_{1\le a_1<\ldots<a_k\le k} P\left(\bigcap_{j=1}^{i} E_{a_j} \cap E_{k+1}\right) \\
&\quad + \left(\sum_{i=2}^{k}(-1)^{i+1} \sum_{1\le a_1<\ldots<a_i\le k} P\left(\bigcap_{j=1}^{i} E_{a_j}\right) + \sum_{i=1}^{k-1}(-1)^{(i+1)+1} \sum_{1\le a_1<\ldots<a_i\le k} P\left(\bigcap_{j=1}^{i} E_{a_j} \cap E_{k+1}\right)\right) \\
&= (-1)^{1+1} \sum_{1\le a_1\le k+1} P\left(\bigcap_{j=1}^{1} E_{a_j}\right) + (-1)^{(k+1)+1} \sum_{1\le a_1<\ldots<a_{k+1}\le k+1} P\left(\bigcap_{j=1}^{k+1} E_{a_j}\right) \\
&\quad + \sum_{i=2}^{k}(-1)^{i+1} \sum_{1\le a_1<\ldots<a_i\le k+1} P\left(\bigcap_{j=1}^{i} E_{a_j}\right) \\
&= \sum_{i=1}^{k}(-1)^{i+1} \sum_{1\le a_1<\ldots<a_i\le k+1} P\left(\bigcap_{j=1}^{i} E_{a_j}\right) \\
&= \sum_{a_1=1}^{n} P(E_{a_1}) - \sum_{a_1<a_2} P(E_{a_1} \cap E_{a_2}) + \sum_{a_1<a_2<a_3} P(E_{a_1} \cap E_{a_2} \cap E_{a_3}) + \ldots + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} E_i\right)
\end{aligned}
$$

□

Lemma 1.22 (Bonferroni's Inequality)**.** *Take* $E_1, E_2, \ldots, E_n \in \mathcal{F}$.

1. $P\left(\bigcup_{i=1}^{n} E_i\right) \le \sum_{i=1}^{n} P(E_i))$
2. $P\left(\bigcup_{i=1}^{n} E_i\right) \ge \sum_{i=1}^{n} P(E_i) - \sum_{i<j} P(E_i \cap E_j)$
3. $P\left(\bigcup_{i=1}^{n} E_i\right) \le \sum_{i=1}^{n} P(E_i) - \sum_{i<j} P(E_i \cap E_j) + \sum_{i<j<k} P(E_i \cap E_j \cap E_k)$

$\vdots$

n. Goes back to Lemma 1.21

Proof. We first rewrite the equations/inequalities. Given $E_1, \ldots, E_n \in \mathcal{F}$ the following holds

- $P\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^j (-1)^{i+1} \sum_{1 \leq a_1 < \ldots < a_i \leq n} P\left(\bigcap_{l=1}^i E_{a_l}\right)$, for all $1 \leq j \leq n$ odd
- $P\left(\bigcup_{i=1}^n E_i\right) \geq \sum_{i=1}^j (-1)^{i+1} \sum_{1 \leq a_1 < \ldots < a_i \leq n} P\left(\bigcap_{l=1}^i E_{a_l}\right)$, for all $1 \leq j \leq$ even

We prove by induction on n.

Base case:

$n = 1$: We have only one equation $P(\bigcup_{i=1}^1 E_i) = P(E_1) = \sum_{a_1=1}^1 P(E_{a_1})$, and this is satisfied.

$n = 2$: We have the inequality $P(\bigcup_{i=1}^2 E_i) \leq P(E_1) + P(E_2) = \sum_{a_1=1}^2 P(E_{a_1})$ is satisfied by the sub-additivity. The equation $P(\bigcup_{i=1}^2 E_i) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = \sum_{a_1=1}^n P(E_{a_1}) - \sum_{1 \leq a_1 < a_2 \leq 2} P(E_{a_1} \cap E_{a_2})$ is true by Proposition 1.17.

Inductive Step: Suppose this is true for $1, \ldots, k$ set. That is, for all $1 \leq j \leq k$, given any j sets $E_1, \ldots, E_j \in \mathcal{F}$, all the j equations/inequalities are satisfied. Now, we show this is also true for $k+1$. That is, given $E_1, \ldots, E_{k+1} \in \mathcal{F}$, all $k+1$ equations/inequalities are satisfied.

We first notice that

$$\begin{aligned} P\left(\bigcup_{i=1}^{k+1} E_i\right) &= P\left(E_1 \cup \bigcup_{i=2}^{k+1} E_i\right) \\ &= P(E_1) + P\left(\bigcup_{i=2}^{k+1} E_i\right) - P\left(E_1 \cap \bigcup_{i=2}^{k+1} E_i\right) \qquad \text{[By Proposition 1.17]} \\ &= P(E_1) + P\left(\bigcup_{i=2}^{k+1} E_i\right) - P\left(\bigcup_{i=2}^{k+1} (E_1 \cap E_i)\right) \end{aligned}$$

Case 1: The $k+1$th equation automatically holds by Lemma 1.21.

Case 2: Suppose we are looking at the jth equation where $1 \leq j < k+1$ and j is odd. We get

$$\begin{aligned} P\left(\bigcup_{i=1}^{k+1} E_i\right) &= P(E_1) + P\left(\bigcup_{i=2}^{k+1} E_i\right) - P\left(\bigcup_{i=2}^{k+1} (E_1 \cap E_i)\right) \\ &\leq P(E_1) + \sum_{i=1}^{j} (-1)^{i+1} \sum_{2 \leq a_1 < \ldots < a_i \leq k+1} P\left(\bigcap_{l=1}^{i} E_{a_l}\right) - \sum_{i=1}^{j-1} (-1)^{i+1} \sum_{2 \leq a_1 < \ldots < a_i \leq k+1} P\left(\bigcap_{l=1}^{i} (E_1 \cap E_{a_l})\right) \\ &= P(E_1) + \sum_{i=1}^{j} (-1)^{i+1} \sum_{2 \leq a_1 < \ldots < a_i \leq k+1} P\left(\bigcap_{l=1}^{i} E_{a_l}\right) - \sum_{i=2}^{j} (-1)^{i+1} \sum_{1 = a_1 < \ldots < a_i \leq k+1} P\left(\bigcap_{l=1}^{i} E_{a_l}\right) \\ &= \sum_{i=1}^{j} (-1)^{i+1} \sum_{2 \leq a_1 < \ldots < a_i \leq k+1} P\left(\bigcap_{l=1}^{i} E_{a_l}\right) - \sum_{i=1}^{j} (-1)^{i+1} \sum_{1 = a_1 < \ldots < a_i \leq k+1} P\left(\bigcap_{l=1}^{i} E_{a_l}\right) \\ &= \sum_{i=1}^{j} (-1)^{i+1} \sum_{1 \leq a_1 < \ldots < a_i \leq k+1} P\left(\bigcap_{l=1}^{i} E_{a_l}\right) \end{aligned}$$

This shows all the jth equation where j is odd and $1 \leq j < k+1$ are satisfied.

Case 3: Suppose we are looking at the jth equation where $1 \leq j < k+1$ is even. We get

$$\begin{aligned}
P\left(\bigcup_{i=1}^{k+1} E_i\right) &= P(E_1) + P\left(\bigcup_{i=2}^{k+1} E_i\right) - P\left(\bigcup_{i=2}^{k+1}(E_1 \cap E_i)\right) \\
&\geq P(E_1) + \sum_{i=1}^{j}(-1)^{i+1} \sum_{2\leq a_1<\ldots<a_i\leq k+1} P\left(\bigcap_{l=1}^{i} E_{a_l}\right) - \sum_{i=1}^{j-1}(-1)^{i+1} \sum_{2\leq a_1<\ldots<a_i\leq k+1} P\left(\bigcap_{l=1}^{i}(E_1 \cap E_{a_l})\right) \\
&= P(E_1) + \sum_{i=1}^{j}(-1)^{i+1} \sum_{2\leq a_1<\ldots<a_i\leq k+1} P\left(\bigcap_{l=1}^{i} E_{a_l}\right) - \sum_{i=2}^{j}(-1)^{i+1} \sum_{1=a_1<\ldots<a_i\leq k+1} P\left(\bigcap_{l=1}^{i} E_{a_l}\right) \\
&= \sum_{i=1}^{j}(-1)^{i+1} \sum_{2\leq a_1<\ldots<a_i\leq k+1} P\left(\bigcap_{l=1}^{i} E_{a_l}\right) - \sum_{i=1}^{j}(-1)^{i+1} \sum_{1=a_1<\ldots<a_i\leq k+1} P\left(\bigcap_{l=1}^{i} E_{a_l}\right) \\
&= \sum_{i=1}^{j}(-1)^{i+1} \sum_{1\leq a_1<\ldots<a_i\leq k+1} P\left(\bigcap_{l=1}^{i} E_{a_l}\right)
\end{aligned}$$

This shows all the jth equation where j is even and $1 \leq j < k+1$ are satisfied.
Hence, by induction, we showed that given $E_1, \ldots, E_n$, all the n equations/inequalities are satisfied.
□

Definition 1.23 (Monotone Sequence of Events)**.** *We say a sequnece of events $\{A_n\}$ is non-increasing if $A_{n+1} \subseteq A_n, \forall n$.*
It is non-decreasing if $A_n \subseteq A_{n+1}, \forall n$

Definition 1.24. *Let $\{A_n\}$ be a monotone sequence of events.*
The limit of $\{A_n\}$ is defined as

$$\begin{cases} \lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n = A & \text{if } \{A_n\} \text{ is non-decreasing} \\ \lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n = B & \text{if } \{A_n\} \text{ is non-increasing} \end{cases}$$

We say $\{A_n\}$ converge from above to A, or converge from below to B.

Proposition 1.25 (Continuity of Probability measure)**.** *If $\{A_n\}$ is monotone, then*

$$\lim_{n\to\infty} P(A_n) = P\left(\lim_{n\to\infty} A_n\right)$$

Proof. WLOG, assume $\{A_n\}$ is non-decreasing, and $\lim_{n\to\infty} A_n = A$, also define $A_0 = \emptyset$.
Define $B_1 = A_1, B_i = A_i \backslash A_{i-1}$ for all $i \geq 2$.
Note that again we have

$$\bigcup_{m=1}^{n} B_m = \bigcup_{m=1}^{n} (A_m \backslash A_{m-1}) = \bigcup_{m=1}^{n} A_m \cap \bigcup_{m=1}^{n} A_{m-1}^c = A_n$$

We also have

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

For $\{A_n\}$ being non-increasing, we can take the complement. □

Example 1.26. *Let $A_n = [0, 1 - \frac{1}{n}]$ and consider the uniform measure.*

Example 1.27. *Let*

$$A_n = \begin{cases} \Omega & n \text{ is odd} \\ \emptyset & n \text{ is even} \end{cases}$$

any probability measure does not converge.

1.2 Independence and Conditional Probability

Definition 1.28. *Given $E_1, E_2, \ldots \subseteq \Omega$. We call $E_1, E_2, \ldots$ are independent with respect to some P if for every $I \subseteq [n] = \{1, 2, \ldots, n\}$*

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

Proposition 1.29. *If $E_1, \ldots, E_n$ are independent then $E_1^c, E_2, \ldots, E_n$ are independent.*

Proof. Fix $I \subseteq [n]$, if $1 \notin I$, then we clearly have

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

by assumption.

Now, suppose $1 \in I$, then define $I' = I \backslash \{1\}$

$$\begin{aligned} P\left(E_1^c \bigcap_{i \in I'} E_i\right) &= P\left(\bigcap_{i \in I'} E_i \backslash E_1\right) \\ &= P\left(\bigcap_{i \in I'} E_i\right) - P\left(E_1 \cap \bigcap_{i \in I'} E_i\right) \\ &= \prod_{i \in I'} P(E_i) - P(E_1) \prod_{i \in I'} P(E_i) \\ &= (1 - P(E_1)) \prod_{i \in I'} P(E_i) \\ &= P(E_1^c) \prod_{i \in I'} P(E_i) \end{aligned}$$

□

Definition 1.30. *An infinite collection of events $\{E_\alpha : \alpha \in I\}$ where I can be uncountable are independent if for any finite subset $J \subseteq I$, the events $\{E_i : i \in J\}$ are independent.*

Definition 1.31. *Given a probability space $(\Omega, \mathcal{F}, P)$ and some $B \in F$ with $P(B) > 0$. Then the conditional probability of $A \in \mathcal{F}$ given B*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Lemma 1.32. *Two events A, B such that $P(A), P(B) > 0$ are independent if and only if*

$$P(A|B) = P(A) \qquad or \qquad P(B|A) = P(B)$$

Proof. ($\Rightarrow$): Suppose A, B are independent. Denote $A = E_1, B = E_2$, take $I = \{1, 2\} \subseteq [2]$, we have

$$P(A \cap B) = P(\bigcap_{i \in I} E_i) = \prod_{i \in I} P(E_i) = P(A)P(B)$$

As $P(B) > 0$, this gives

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

($\Leftarrow$): Denote $A = E_1, B = E_2$, take $I \in [n]$, we need to show

$$P(\bigcap_{i \in I} E_i) = \prod_{i \in I} P(E_i)$$

If $I = \{1\}$, we have

$$P(\bigcap_{i \in I} E_i) = P(A) = \prod_{i \in I} P(E_i)$$

Similarly, if $I = \{2\}$, we have

$$P(\bigcap_{i \in I} E_i) = P(A) = \prod_{i \in I} P(E_i)$$

If $I = \emptyset$, we have

$$P(\bigcap_{i\in I} E_i) = P(\Omega) = 1 = \prod_{i\in I} P(E_i)$$

If $I = \{1, 2\}$, we have two cases
Case 1: Suppose $P(A|B) = P(A)$. As $P(B) > 0$, we have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = P(A) \implies P(A \cap B) = P(\bigcap_{i\in I} E_i) = P(A)P(B) = \prod_{i\in I} P(E_i)$$

Case 2: Suppose $P(B|A) = P(B)$. As $P(A) > 0$, we have

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = P(B) \implies P(A \cap B) = P(\bigcap_{i\in I} E_i) = P(A)P(B) = \prod_{i\in I} P(E_i)$$

□

Definition 1.33. *Given the setup of Definition 1.31, define*

$$P_B(A) = P(A|B), \forall A \in \mathcal{F}$$

Proposition 1.34. *$P_B(\cdot)$ is a probability measure.*

Proof. First, we show $P_B(A) \in [0, 1]$ for all $A \in \mathcal{F}$. Indeed, we have

$$P_B(A) = P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Since both are probabilities $P(A \cap B), P(B) \geq 0$, we have $\frac{P(A\cap B)}{P(B)} \geq 0$.
Since $A \cap B \subseteq B$, we have $P(A \cap B) \subseteq P(B)$ by Proposition 1.17.
Hence, $\frac{P(A\cap B)}{P(B)} \leq 1$. This shows that $P_B(A) \in [0, 1]$.
Now, we also have

$$P_B(\Omega) = P(\Omega|B) = \frac{P(B \cap \Omega)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

Finally, take disjoint $E_1, \ldots \in \mathcal{F}$, we have

$$\begin{aligned}
P_B(\bigcup_{i=1}^{\infty} E_i) &= P(\bigcup_{i=1}^{\infty} |B) \\
&= \frac{P(\bigcup_{i=1}^{\infty} E_i \cap B)}{P(B)} \\
&= \frac{P(\bigcup_{i=1}^{\infty} (E_i \cap B))}{P(B)} \\
&= \frac{\sum_{i=1}^{\infty} P(E_i \cap B)}{P(B)} \qquad \text{[By definition of a probability measure]} \\
&= \sum_{i=1}^{\infty} P(E_i|B) \\
&= \sum_{i=1}^{\infty} P_B(E_i)
\end{aligned}$$

□

Proposition 1.35 (Law of Total Probability)**.** *Given $E_1, E_2, \ldots \mathcal{F}$ with $P(E_i) > 0, \forall i$. Moreover, suppose $E_i \cap E_j = \emptyset, \forall i \neq j$, and $\bigcup_{i=1}^{\infty} E_i = \Omega$. Then for $A \in \mathcal{F}$*

$$P(A) = \sum_{i=1}^{\infty} P(A|E_i)P(E_i)$$

Proof. We have

$$\begin{aligned}
P(A) &= P(A \cap \Omega) \\
&= P(A \cap \bigcup_{i=1}^{\infty} E_i) \\
&= P(\bigcup_{i=1}^{\infty}(A \cap E_i) \\
&= \sum_{i=1}^{\infty} P(A \cap E_i) \qquad \text{[By definition of a probability measure]} \\
&= \sum_{i=1}^{\infty} P(A|E_i)P(E_i) \qquad [P(E_i) > 0, \forall i]
\end{aligned}$$

□

Theorem 1.36 (Baye's Theorem)**.** *Given $A, B \in \mathcal{F}$ with $P(A), P(B) > 0$, then*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Proof. We have

$$\begin{aligned}
P(A|B) &= \frac{P(A \cap B)}{P(B)} \qquad [P(B) > 0] \\
&= \frac{P(B|A)P(A)}{P(B)} \qquad [P(A) > 0]
\end{aligned}$$

□

Proposition 1.37. *Given $A, E_1, \ldots, E_n \in \mathcal{F}$. Moreover, $E_i \cap E_j = \emptyset, \forall i \neq j$, and $\bigcup_{i=1}^{n} E_i = \Omega$.*

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

Proof. We have

$$\begin{aligned}
P(E_i|A) &= \frac{P(E_i \cap A)}{P(A)} \qquad [P(A) > 0] \\
&= \frac{P(A|E_i)P(E_i)}{P(A \cap \Omega)} \qquad [P(E_i) > 0] \\
&= \frac{P(A|E_i)P(E_i)}{P(A \cap \bigcup_{i=1}^{n} E_i)} \\
&= \frac{P(A|E_i)P(E_i)}{P(\bigcup_{i=1}^{n}(A \cap E_i))} \\
&= \frac{P(A|E_i)P(E_i)}{\sum_{j=1}^{n} P(A \cap E_j)} \qquad \text{[By definition of a probability measure]} \\
&= \frac{P(A|E_i)P(E_i)}{\sum_{j=1}^{n} P(A|E_j)P(E_j)} \qquad [P(E_j) > 0]
\end{aligned}$$

□

1.3 Exercises

Question 1.38. *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 $n \to \infty$?*

Question 1.39. *Suppose $E_1, E_2, \dots \subseteq \Omega$. Show that $P(E_i) = 1$ for all $i \in \mathbb{N}$ if and only if $P(\bigcap_{i=1}^{\infty} E_i) = 1$*

Question 1.40. *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 $k \in \{1, \dots, n\}$.*

Question 1.41. *Find an example of an Ω, and sets A, B, C such that $P(A \cap B \cap C) = P(A)P(B)P(C)$ but A, B, C are not independent.*

Hint: Ω does not need to have more than 4 elements

Question 1.42. *1. Prove that for $\{a_i\}_{i=1}^n$ with $a_i \in (0, 1)$,*

$$\prod_{i=1}^{n}(1 - a_i) = \sum_{I \subseteq [n]} (-1)^{|I|} \prod_{i \in I} a_i$$

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

Question 1.43. *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 $P(A_i) = \frac{1}{2}$ for all i.*

2 Random Variables and Distributions

2.1 Random Variables

Definition 2.1. *Given a sample space Ω, a random variable (R.V.) is a function $X : \Omega \to \mathbb{R}$.*

Example 2.2. *The sample space of a coin flip is $\{H, T\}$, we can define the random variable X as*

$$\begin{cases} X(H) = 1 \\ X(T) = 0 \end{cases}$$

Remark 2.3. *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*

$$\begin{cases} X(w) = 1 & \text{if } w = H \\ X(w) = 0 & \text{if } w = T \end{cases} \tag{1}$$

However, Definition 2.1 is not exactly sufficient. This is because when talking about R.V., we are almost always interested in quantities such as $P(X \in A)$ for some $A \subseteq \mathbb{R}$. This means that we need the event $\{X \in A\} = \{w \in \Omega : X(\omega) \in A\}$ to be a well-defined set (or measurable under P), i.e., it

needs to be in a σ-algebra $\mathcal{F}$, so that $P(X \in A)$ is well-defined. Most of A that we are interested in are all contained in the Borel σ-algebra of $\mathbb{R}$, that is, $A \in \mathcal{B}(\mathbb{R})$, which are all the sets generated by intervals in R. Mathematically, we write $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$ and we call it the pre-image of A under the mapping X.

NPOS starts:

Formally, we require $X^{-1}(A) \in \mathcal{F}$ 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 the following:

Definition 2.4 (True Definition of A Random Variable)**.** *A R.V. is a function $X : \Omega \to \mathbb{R}$ on a probability space $(\Omega, \mathcal{F}, P)$ that is $F\backslash\mathcal{B}$ measurable.*

More succinctly, we say that a R.V. is a measurable map $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

In fact, if we require all the events of the form $X^{-1}(A)$ 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, $\sigma(X) = \{X^{-1}(A) : A \in \mathcal{B}(\mathbb{R})\}$

The reason that $\sigma(X)$ is the smallest is that it contains only the information that can be learned about the random experiment through X, and nothing more. We can look at the following example

Example 2.5. *The sample space of two coin flips is $\{HH, HT, TH, TT\}$.*

The biggest σ-algebra from Ω is $P(\Omega)$. Now we can define

$$X(\omega) = \begin{cases} 1 & \text{if } \omega \in \{HH, HT\} \\ 0 & \text{if } \omega \in \{TH, TT\} \end{cases}$$

This gives $\sigma(X) = \{\emptyset, \Omega, \{HH, HT\}, \{TH, TT\}\}$, which is a strict subset of $P(\Omega)$. This is because knowing the value of X at best only tells us whether the first coin flip results in H or T, but we do not know anything about the second flip if we only have access to the value of X

Everything we have discussed here is not tested and is only intended to give you slightly deeper understanding of how random variable works. For further details, they are out of the scope of the course as it ventures deep 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.

NPOS ends

Proposition 2.6. *1. If $X = c$ for some $c \in \mathbb{R}, X$ is a random variable.*

2. $X = \mathbb{1}_A$ is the indicator of $A \in \mathcal{F}$ is a random variable.

3. If X, Y are both random variables. For some $c \in \mathbb{R}$, then $X + c, cX, X^2, X + Y, XY$ are all random variables.

4. If $Z_1, Z_2, \ldots$ are random variables. In addition, $\forall \omega \in \Omega, \lim_{n\to\infty} Z_n(\omega) \in \mathbb{R}$ then there is a random variable $Z = \lim_{n\to\infty} Z_n$.

Proof. (1) For every $\omega \in \Omega$, we get $X(\omega) = c \in \mathbb{R}$, which is indeed a random variable.

(2) For every $\omega \in \Omega$, we get

Case 1: If $w \in A$, we have $X(w) = 1$

Case 2: If $w \notin A$, we have $X(w) = 0$.

Therefore, for every $\omega \in \Omega, X(\omega)$, is a function that maps ω to a real value in $\{0, 1\}$. Thus, X is a random variable.

(3) $X + c$: Given $c \in \mathbb{R}$ and a random variable X, for every $\omega \in \Omega$, we get

$$(X + c)(\omega) = X(\omega) + c \in \mathbb{R} \qquad \text{[By real number is closed under addition and } X(\omega), c \in \mathbb{R}]$$

Hence, by definition $X + c$ is a random variable.

cX : Given $c \in \mathbb{R}$ and a random variable X, for every $\omega \in \Omega$, we get

$$(cX)(\omega) = c \cdot X(\omega) \in \mathbb{R} \qquad \text{[By real number is closed under multiplication and } X(\omega), c \in \mathbb{R}]$$

Hence, by definition cX is a random variable.

X^2 : Given a random variable X, for every $\omega \in \Omega$, we get

$$X^2(\omega) = X(\omega) \cdot X(\omega) \in \mathbb{R} \qquad \text{[By real number is closed under multiplication and } X(\omega) \in \mathbb{R}]$$

Hence, by definition X^2 is a random variable.

$X + Y$: Given a random variable X, Y, for every $\omega \in \Omega$, we get

$$(X+Y)(\omega) = X(\omega) + Y(\omega) \in \mathbb{R} \qquad \text{[By real number is closed under addition and } X(\omega), Y(\omega) \in \mathbb{R}]$$

Hence, by definition $X + Y$ is a random variable.

XY : Given a random variable X, Y, for every $\omega \in \Omega$, we get

$$(XY)(\omega) = X(\omega) \cdot Y(\omega) \in \mathbb{R} \qquad \text{[By real number is closed under multiplication and } X(\omega), Y(\omega) \in \mathbb{R}]$$

Hence, by definition XY is a random variable.

(4) Define $Z : \Omega \to \mathbb{R}$ by $Z(\omega) = \lim_{n\to\infty} Z_n(\omega)$ for all $\omega \in \Omega$. We show Z is a well-defined function. Given any $\omega \in \Omega, \lim_{n\to\infty} Z_n(\omega) \in \mathbb{R}$ exists hence $Z(\omega)$ maps to at least one real number. Since the limit is always unique we showed that $Z(\omega)$ maps to exactly one real number for all ω.

Hence, by definition Z is a random variable. □

Proposition 2.7. *If X is a random variable and $f : \mathbb{R} \to \mathbb{R}$ is continuous function. Then $f(X)$ is a random variable.*

Proof. Given $X : \Omega \to \mathbb{R}$ is a random variable and $f : \mathbb{R} \to \mathbb{R}$ a continuous function. We show $f(X) : \Omega \to \mathbb{R}$ is also a random variable.

Take any $\omega \in \Omega$, we have

$$f(X)(\omega) = f(X(\omega)) \in \mathbb{R} \qquad \text{[As } X(\omega) \in \mathbb{R}, f \text{ maps to real numbers]}$$

This by definition shows that $f(X)$ is a random variable. □

Remark 2.8. *Note that for the above proposition, we do not necessarily need f to be continuous for $f(X)$ to be a random variable. The least we require is that f is a measurable function; however, measure theory and measurability is out of scope.*

2.2 Distributions

Lemma 2.9. *A random variable X defined on a probability space $(\Omega, \mathcal{F}, P)$ induces a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. This μ has a property that*

$$\mu(A) = P(X \in A) = P(X^{-1}(A))$$

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

Proof. Fisrt, for any $A \in \mathcal{B}(\mathbb{R}), \mu(A) \in [0, 1]$ is trivial as P is a probability measure.

Next, since we have $X(\omega) \in \mathbb{R}$ for all $\omega \in \Omega$, we have $\mu(\mathbb{R}) = P(X \in \mathbb{R}) = P(\Omega) = 1$.

Finally, take disjoint $A_1, A_2, \ldots \in \mathcal{B}(\mathbb{R})$. Define $E_i = X^{-1}(A_i) = \{\omega \in \Omega : X(\omega) \in A_i\} \in \mathcal{F}$

Note that $E_1 = \{\omega \in \Omega, X(\omega) \in A_1\}, E_2 = \{\omega \in \Omega, X(\omega) \in A_2\}, \ldots$ are disjoint as X is a function, it is impossible for $X(\omega) \in A_i$ and $X(\omega) \in A_j$ for $i \neq j$. Hence

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

□

Remark 2.10. *μ can be called law/distribution of X.*

Definition 2.11. *A random variable X and probability measure P generate the cumulative distribution function (CDF) F defined by*

$$F(x) = \mu((-\infty, x]) = P(X \leq x)$$

Theorem 2.12. *A distribution function F for a random variable X uniquely defines the measure μ*

Proof. NPOS. □

Example 2.13 (Uniform Distribution)**.** *Suppose P is uniform on $[0,1]$ and $X(\omega) = \omega$. For $x \in [0,1]$*

$$F(x) = \mu([0,x))/\mu((-\infty, x)) = P(X(\omega) \in [0,x))/P(X(\omega) \in (-\infty, x)) = P(\omega \in [0,x))/P(\omega \in (-\infty, x)) = x$$

Theorem 2.14. *A CDF has the following property*

1. *For all $x \leq y \in \mathbb{R}, F(x) \leq F(y)$*
2. *$\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$*
3. *F is right continuous, $\lim_{x \to t^+} F(x) = F(t)$*

Proof. (1) If $x \leq y$, we have $\{\omega \in \Omega : X(w) \leq x\} = X^{-1}((-\infty, x]) \subseteq \{\omega \in \Omega : X(w) \leq y\} = X^{-1}((-\infty, y])$.

Hence, $F(x) = P(X \leq x) = P(X^{-1}((-\infty, x])) \leq P(X^{-1}((-\infty, y])) = P(X \leq y) = F(y)$

(2) Define two sequences $\{x_n\}_{n=1}^{\infty}$ and $\{y_n\}_{n=1}^{\infty}$ such that $x_n \to \infty, y_n \to -\infty$.

We have

$$\lim_{n \to \infty} \{\omega \in \Omega : X(\omega) \leq x_n\} = \{\omega \in \Omega : X(\omega) < \infty\}$$

and

$$\lim_{n \to \infty} \{\omega \in \Omega : X(\omega) \leq y_n\} = \{\omega \in \Omega : X(\omega) < -\infty\}$$

By Proposition 1.25, $\lim_{n \to \infty} F(x_n) = \lim_{n \to \infty} P(\omega \in \Omega : X(\omega) \leq x_n) = P(\{\omega \in \Omega : X(\omega) \leq \infty) = 1$.

Similarly, $\lim_{n \to \infty} F(y_n) = \lim_{n \to \infty} P(\omega \in \Omega : X(\omega) \leq y_n) = P(\{\omega \in \Omega : X(\omega) < \infty) = 0$

Since the sequence is arbitrary, we get $\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$

(3) Given $x \in \mathbb{R}$ and a sequence $x_n \to x^+$ converges from above. We get $\{\omega \in \Omega : X(\omega) \leq x_n\} \to \{w \in \Omega : X(\omega) \leq x\}$. Again, by Propositon 1.25 we get

$$\lim_{n \to \infty} P(\omega \in \Omega : X(\omega) \leq x_n) = P(\lim_{n \to \infty} \omega \in \Omega : X(\omega) \leq x_n) = P(\omega \in \Omega : X(\omega) \leq x) = F(x)$$

Since the sequence is arbitrary, we get F is right continuous. □

Definition 2.15. *The inverse CDF of random variable X is*

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

Note that if F is continuous, then F^{-1} reduces to the classical inverse of a continuous and strictly monotone function (by flipping x and y)

Remark 2.16. *A bit of a review from analysis: the supremum of a set $A \subset \mathbb{R}$ is the smallest real number s such that:*

1. *$a \leq s$ for all $a \in A$*
2. *For any $\epsilon > 0$, there exists $a \in A$ such that $a > s - \epsilon$*

Theorem 2.17 (Inverse CDF method)**.** *If F satisfies the three properties of Theorem 2.14, then 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 $x \in \mathbb{R}$, $t \in [0,1]$.

First, suppose $F^{-1}(t) \leq x$, then $\sup\{y : F(y) < t\} \leq x$, so for all $y > x$, $F(y) \geq t$ as F is non-decreasing. Since F is right-continuous, $F(x) \geq t$.

Conversely, if $F(x) \geq t$, then all $y > x$ satisfy $F(y) \geq t$ (again because F is non-decreasing), so $\sup\{y : F(y) < t\} \leq x$, hence $F^{-1}(t) \leq x$.

The above means that $\{t : F^{-1}(t) \leq x\} = \{t : t \leq F(x)\}$. Thus,

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

where last equality is given by the definition of a uniform distribution. So Y is a random variable with CDF F. □

Example 2.18. *If $U \sim Unif(0,1)$, then $-\lambda^{-1}\log(1-U) \sim \exp(\lambda)$ for $\lambda > 0$.*

Proof. Note that the CDF of $\exp(\lambda)$ R.V. is

$$F_X(x) = 1 - e^{-\lambda x}, \quad x > 0$$

Since F_X is continuous, we have $F_X^{-1}(y) = -\lambda^{-1}\log(1-y)$. Thus, by the Theorem 2.17, the random variable $Y = -\lambda^{-1}\log(1-U) = F_X^{-1}(U)$ has CDF $F_X(x)$, i.e., $Y \sim \text{Exp}(\lambda)$. □

Example 2.19 (Empirical CDF)**.** *Let $X_1, X_2, \ldots, X_n$ be random variables (observed data). The empirical CDF is defined as*

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

$F_n(x)$ is the CDF of some random variable.

Definition 2.20. *A random variable X is called discrete if its range is countable, i.e., $X(\omega) \in \{x_i\}_{i=1}^n$ or $X(\omega) \in \{x_i\}_{i=1}^\infty$ for any $\omega \in \Omega$.*

Definition 2.21. *X is a discrete random variable on $(\Omega, \mathcal{F}, P)$. Let $D \subseteq \mathbb{R}$ be the possible values that X can take.*

The probability mass function (PMF) of X is

$$p_X(x) = \begin{cases} P(X = x) > 0 & \text{if } x \in D \\ 0 & \text{if } x \notin D \end{cases}$$

We call D is the support of X.

Lemma 2.22. *A discrete random variable with PMF $P_X(x)$ and support D. It has CDF*

$$F_X(x) = \sum_{y \in D : y \leq x} P_X(y)$$

Proof. We have

$$\begin{aligned}
F_X(x) &= P(X \leq x) \qquad \text{[By definition]} \\
&= P(\{\omega \in \Omega : X(\omega) \leq x\}) \qquad \text{[By definition]} \\
&= P\left(\bigcup_{y \leq x} \{\omega \in \Omega : X(\omega) = y\}\right) \\
&= \sum_{y \leq x} P\left(\{\omega \in \Omega : X(\omega) = y\}\right) \qquad \text{[By definition of a probability measure]} \\
&= \sum_{y \in D : y \leq x} P\{\omega \in \Omega : X(\omega) = y\} + \sum_{y \notin D : y \leq x} P\{\omega \in \Omega : X(\omega) = y\} \\
&= \sum_{y \in D : y \leq x} P\{X = y\} + \sum_{y \notin D : y \leq x} P\{\emptyset\} \\
&= \sum_{y \in D : y \leq x} p_X(y)
\end{aligned}$$

□

Definition 2.23. *Two random variables X, Y are equal if for all $\omega \in \Omega, X(\omega) = Y(\omega)$.*

Two random variables X, Y are equal in distribution if for all $A \in \mathcal{B}(\mathbb{R}), \mu_X(A) = P(X \in A) = P(Y \in A) = \mu_Y(A)$.

Proposition 2.24. *$p_X(x)$ uniquely defines the distribution of a random variable.*

Note: *This does not guarantee two random variables are equal.*

Proof. If X and Y are discrete random variables with $p_X(x) = p_Y(x)$ for all x, then for every Borel set A,

$$\mu_X(A) = P(X \in A) = \sum_{x \in A} p_X(x) = \sum_{x \in A} p_Y(x) = P(Y \in A) = \mu_Y(A)$$

so $\mu_X = \mu_Y$. Therefore the p.m.f. uniquely characterises the distribution of a discrete random variable. □

Example 2.25 (Bernoulli)**.** *A Bernoulli random variable X with probability of success p has p.m.f.*

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

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

Example 2.26 (Binomial)**.** *A Binomial random variable X with probability of success p and number of trials n has p.m.f.*

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, \ldots, n$$

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

Example 2.27 (Geometric)**.** *A geometric random variable X with probability of success p has p.m.f.*

$$p_X(x) = (1-p)^{x-1} p, \quad x \in \mathbb{N}^+$$

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

Example 2.28 (Poisson)**.** *A Poisson random variable X with rate parameter $\lambda > 0$ has p.m.f.*

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x \in \mathbb{N}_0$$

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

Definition 2.29. *A random variable X is continuous if its CDF is continuous everywhere.*

Definition 2.30. *A random variable X is absolutely continuous if $\exists f : \mathbb{R} \to [0, \infty)$ s.t. $\forall A \in \mathcal{B}(\mathbb{R})$*

$$P(X \in A) = \int_A f(y)dy$$

It is equivalent to say

$$\forall x \in \mathbb{R}, F(x) = \int_{-\infty}^{x} f(y)dy$$

f is called the probability density function (p.d.f.) of X. The support of X, denoted by D is defined as $D = \{x \in \mathbb{R} : f(x) > 0\}$

Remark 2.31. *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 2.32. *If X is continuous random variable with CDF F and p.d.f. f then $\exists A \in \mathcal{B}(\mathbb{R})$ with $P(X \in A) = 1$ such that*

$$\forall x \in A, F'(x) = f(x)$$

Proof. NPOS □

Remark 2.33. *Note that the above lemma is different from the introductory level probability claim that $F'(X) = f(x)$ based on the fundamental theorem of calculus. This is because the p.d.f. can be discontinuous, and F' may not exist at the discontinuity points of f. The only thing we require is that the set where F is not differentiable has zero probability mass under the distribution of X. Although the above lemma is a direct result of the fundamental theorem of calculus under the Lebesgue integration regime. However, this goes into the realm of measure theory and we thus do not cover the proof.*

Theorem 2.34. *If $f : \mathbb{R} \to [0, \infty)$ such that*

$$\int_{\mathbb{R}} f(y)dy = 1$$

then f is the p.d.f of some R.V. X.

Proof. Since $f \geq 0$, then $\forall x \in \mathbb{R}$. We have

$$0 \leq \int_{-\infty}^{x} f(y)dy \leq \int_{\mathbb{R}} f(y)dy = 1$$

Now, define $F(x) = \int_{-\infty}^{x} f(y)dy$. We now claim F satisfies the three defining properties of CDF. Take $x_1, x_2 \in \mathbb{R}$ such that $x_1 \leq x_2$. We have

$$F(x) = \int_{-\infty}^{x_1} f(y)dy \leq \int_{-\infty}^{x_2} f(y)dy$$

We also have

$$\lim_{x\to-\infty} F(x) = \lim_{x\to-\infty} \int_{-\infty}^{x} f(y)dy = 0$$

and

$$\lim_{x\to\infty} F(x) = \lim_{x\to\infty} \int_{-\infty}^{x} f(y)dy = \int_{\mathbb{R}} f(y)dy = 0$$

□

Lemma 2.35. *If X has a p.d.f. then*

$$P(X = x) = 0, \forall x \in \mathbb{R}$$

Proof.

$$\begin{aligned} P(X = x) &= \lim_{\delta\to 0} P(x - \delta < X \leq x + \delta) \\ &= \lim_{\delta\to 0} \int_{x-\delta}^{x+\delta} f(y)dy \\ &= 0 \end{aligned}$$

□

Furthermore, X does not need to have a p.d.f. for the above to be true. X is continuous $\iff$ F is continuous.

Proposition 2.36. *A p.d.f. $f(x)$ uniquely defines the distribution of some R.V. X.*

Proof. If X, Y are absolutely continuous random variable with the same p.d.f. $f(x)$. Then, for any $A \in \mathcal{B}(\mathbb{R})$

$$\mu_Y(A) = P(Y \in A) = \int_A f(y)dy = P(X \in A) = \mu_X(A)$$

so $\mu_Y = \mu_X$. □

Example 2.37 (Uniform)**.** *A uniform random variable on the interval (a, b) with $a < b \in \mathbb{R}$ has p.d.f.*

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

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

Example 2.38 (Beta)**.** *A Beta random variable with parameters $a, b > 0$ has p.d.f.*

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

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

Definition 2.39. $\Gamma(\cdot)$ *is a real-number extension of factorial. Specifically, we have for all $z \in \mathbb{R}^+$*

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$$

When $z \in \mathbb{N}^+$, the expression can be simplified into

$$\Gamma(z) = (z-1)!$$

Example 2.40 (Exponential)**.** *An exponential random variable with rate parameter $\lambda > 0$ has p.d.f.*

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

We denote X as $X \sim \exp(\lambda)$.

Example 2.41 (Gamma)**.** *A Gamma random variable with shape parameter $\alpha > 0$ and rate parameter β has p.d.f.*

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

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

Example 2.42 (Normal/Gaussian)**.** *A normal or a Gaussian random variable with mean μ and standard deviation σ has p.d.f.*

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

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

Example 2.43. *The p.d.f. of a Gaussian random variable is indeed a proper p.d.f.*

Proof. Obviously $f(x) > 0, \forall x \in \mathbb{R}$.

We will only prove $\int_{\mathbb{R}} f(x)dx = 1$ when $\mu = 0, \sigma^2 = 1$.

We have $\phi(x) = \frac{1}{\sqrt{2}} \exp\left(-\frac{x^2}{2}\right)$

Let $I = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx$, we have

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx\right)\left(\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} e^{-\frac{y^2}{2}} dxdy \\ &= \int_{\mathbb{R}^2} e^{-\frac{x^2+y^2}{2}} dxdy \end{aligned}$$

Let $x = r\cos(\theta), y = r\sin(\theta), x^2 + y^2 = r^2$ and $dxdy = rdrd\theta$. We have

$$\begin{aligned} I^2 &= \int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2}} rdrd\theta \\ &= \int_0^{2\pi}\int_0^{\infty} e^{-u} dud\theta \\ &= \int_0^{2\pi} 1 d\theta \\ &= 2\pi \end{aligned}$$

This gives $I = \sqrt{2\pi}$. Hence

$$\int_{\mathbb{R}} f(x)dx = \int_{\mathbb{R}} -\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1$$

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then change the variables

$$z = \frac{x-\mu}{\sigma}, \qquad dx = \sigma\, dz$$

Then, we have

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 1.\end{aligned}$$

□

Theorem 2.44 (Change of Variables)**.** *Suppose X has continuous p.d.f. f_X with support (a, b) where $a, b \in \mathbb{R} \cup \{\infty\}$. We also have a function g that is strictly monotone and differentiable. Then $Y = g(X)$ has density.*

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

Y has support $(g(a), g(b))$ if g is increasing.
Y has support $(g(b), g(a))$ if g is decreasing.
If $g(\pm\infty)$ we define $g(\pm\infty) = \lim_{x\to\pm\infty} g(x)$

Proof. WLOG assuming g is increasing, we have

$$\begin{aligned}F_Y(y) &= P(Y \le y) \\ &= P(g(X) \le y) \\ &= P(X \le g^{-1}(y)) \\ &= F_X(g^{-1}(y))\end{aligned}$$

Hence,

$$\begin{aligned}f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= \frac{d}{dy} F_X(g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y) \\ &= f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \qquad [g \text{ is increasing } \implies g^{-1} \text{ is increasing}]\end{aligned}$$

For the support of Y, since $f_X(x) > 0 \iff x \in (a, b)$ and g is strictly increasing, we have $f_Y(y) > 0 \iff g(a) < y < g(b)$. □

Example 2.45 (log-Gaussian/Normal)**.** *Let $X \sim N(\mu, \sigma^2), Y = \exp(x)$*

Proof. Note that $y = e^x$ is strictly increasing and differentiable. Moreover X has continuous p.d.f with support $(-\infty.\infty)$.

By change of variables $Y = g(X)$ has support $(e^{-\infty}), e^{\infty}) = (0, \infty)$.

Moreover, the density function is

$$\begin{aligned}f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \\ &= \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log y - \mu)^2)}{2\sigma^2}\right) \qquad \forall y > 0\end{aligned}$$

□

2.3 Exercises

Question 2.46. *Show that the p.m.f. of all the listed discrete random variables in the current chapter are proper p.m.f. That is, show that these p.m.f. all sum up to 1.*

Question 2.47. *Show that the p.d.f. of all the listed continuous random variables in the current chapter are proper p.d.f. That is, show that these p.d.f. all integrate to 1.*

Question 2.48. *Consider flipping a fair coin. Consider the following random variables:*

$$X = \begin{cases} 1 & \text{if Head} \\ 0 & \text{if Tail} \end{cases} \qquad Y = \begin{cases} 0 & \text{if Head} \\ 1 & \text{if Tail} \end{cases}$$

1. *Do X and Y have the same distribution?*
2. *Does $X = Y$*

Question 2.49. *Let $U \sim Unif(0,1)$. Let $R = \sqrt{-2\log(U)}$, then R follows a Rayleigh distribution, i.e., $f_R(r) = r\exp(-\frac{r^2}{2}), r \geq 0$.*

Question 2.50. *Suppose that X and Y have p.d.f. f and g. Show that it is not possible for all $x \in \mathbb{R}, f(x) > g(x)$.*

Question 2.51. *Show that it is possible for a continuous random variables X to have infinitely many different p.d.f. Is it possible if X is discrete?*

Question 2.52. *Let X be a random variable such that $P(X > 0) > 0$. Prove that there exists a $\delta > 0$ such that $P(X \geq \delta) > 0$.*

Question 2.53. *Prove that for the CDF F of some random variable, F can only have countably many discontinuities.*

3 Random Vectors, Joint distributions and Conditional Distributions

3.1 Random Vectors and Joint distributions

Definition 3.1. *A random vector $X = (X_1, X_2, \ldots, X_n)$ where $X : \Omega \to \mathbb{R}^n$.*

Remark 3.2. *In the above, we have omitted the attention on measurability.*

Definition 3.3. *A joint distribution function of $X = (X_1, X_2, \ldots, X_n)$ is*

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n) = P(X_1 \leq x_1 \cap \ldots \cap X_n \leq x_n)$$

Remark 3.4. *We typically use the notation $P(A, B) = P(A \cap B)$*

Definition 3.5. *The R.V.s $X_1, X_2, \ldots, X_n$ on $(\Omega, \mathcal{F}, P)$ are independent $\iff \forall A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R})$*

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

Theorem 3.6. *R.V.s $X_1, \ldots, X_n$ on $(\Omega, \mathcal{F}, P)$ are independent $\iff \forall x_1, \ldots, x_n \in \mathbb{R}$*
$F(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$

Proof. ($\Rightarrow$) : NPOS.

($\Leftarrow$) : Given $X_1, \ldots, X_n$ are independent. Take any $(x_1, \ldots, x_n) \in \mathbb{R}^n$, we have $(-\infty, x_i] \in \mathcal{B}(\mathbb{R})$ for all x_i. Hence,

$$F(x_1, \ldots, x_n) = P\left(\bigcap_{i=1}^{n} \{X_i \in (-\infty, x_i]\}\right) = \prod_{i=1}^{n} P\left(X_i \in (-\infty, x_i]\right) = \prod_{i=1}^{n} F_{X_i}(x_i)$$

□

Definition 3.7. *An infinite collection of R.V.s $\{X_\alpha : \alpha \in I\}$ are independent if for all finite subset $J \subseteq I, \{X_i : i \in J\}$ are independent.*

Theorem 3.8. *A joint distribution function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies*

1. *If $(x_1, \ldots, x_n), (x'_1, \ldots, x'_n)$ satisfies $x_i \leq x'_i$ for all i, then $F(x_1, \ldots, x_n) \leq F(x'_1, \ldots, x'_n)$.*
2. *$\lim_{x_i \to -\infty} F(x_1, \ldots, x_n) = 0$ for any $i = \{1, \ldots, n\}$ and $\lim_{x_1, \ldots, x_n \to \infty} F(x_1, \ldots, x_n) = 1$*
3. *$\lim_{h \to 0^+} F(x_1 + h, \ldots, x_n) = \lim_{h \to 0^+} F(x_1, x_2 + h, \ldots, x_n) = \ldots = \lim_{h \to 0^+} F(x_1, x_2, \ldots, x_n + h) = F(x_1, \ldots, x_n)$*

Proof. (1) If $(x_1, \ldots, x_n), (x'_1, \ldots, x'_n)$ satisfies $x_i \leq x'_i$ for all i, we have $\{(\omega_1, \ldots \omega_n) \in \Omega^n : X_1(\omega_1) \leq x_1, \ldots, X_n(\omega_n) \leq x_n\} \subseteq \{(\omega_1, \ldots \omega_n) \in \Omega^n : X_1(\omega_1) \leq x'_1, \ldots, X_n(\omega_n) \leq x'_n\}$.

Hence, $F(x_1, \ldots, x_n) = P(X_1 \leq x_1 \cap \ldots \cap X_n \leq x_n) = P(\{(\omega_1, \ldots \omega_n) \in \Omega^n : X_1(\omega_1) \leq x_1, \ldots, X_n(\omega_n) \leq x_n\}) \leq P(\{(\omega_1, \ldots \omega_n) \in \Omega^n : X_1(\omega_1) \leq x'_1, \ldots, X_n(\omega_n) \leq x'_n\}) = P(X_1 \leq x'_1 \cap \ldots \cap X_n \leq x'_n) = F(x'_1, \ldots, x'_n)$

(2) Given arbitrary index i, take any $(x_1, \ldots, x_n) \in \mathbb{R}^n$, we get

$$\begin{aligned}
\lim_{x_i \to -\infty} F(x_1, \ldots, x_n) &= \lim_{x_i \to -\infty} P(X_1 \leq x_1 \cap X_n \leq x_n) \\
&= \lim_{x_i \to -\infty} P(\{(\omega \in \Omega : X_1(\omega) \leq x_1 \cap X_n(\omega) \leq x_n\}) \\
&\leq \lim_{x_i \to -\infty} P(\{(\omega \in \Omega : X_i(\omega) \leq x_i\}) \\
&= \lim_{x_i \to -\infty} P(X_i \leq x_i) \\
&= \lim_{x_i \to -\infty} F_{X_i}(x_i) \\
&= 0
\end{aligned}$$

This gives $\lim_{x_i \to -\infty} F(x_1, \ldots, x_n) = 0$.

Given monotone sequences $\{x_{i,j}\}_{j=1}^{\infty}$ such that $\lim_{j \to \infty} x_{i,j} = \infty$ for all $i = 1, 2, \ldots, n$.

We have for all $j = 1, 2, \ldots$

$$\{\omega \in \Omega : X_1(\omega) \leq x_{1,j}, \ldots, X_n(\omega) \leq x_{n,j}\} \subseteq \{\omega \in \Omega : X_1(\omega) \leq x_{1,j+1}, \ldots, X_n(\omega) \leq x_{n,j+1}\}$$

Moreover, we also have the sequence of monotone sets converges from below as $j \to \infty$

$$\{\omega \in \Omega : X_1(\omega) \leq x_{1,j}, \ldots, X_n(\omega) \leq x_{n,j}\} \to \{\omega \in \Omega : X_1(\omega) < \infty, \ldots, X_n(\omega) < \infty\}$$

Hence, by continuity of probability measure we have

$$\begin{aligned}
\lim_{j \to \infty} F(x_{1,j}, \ldots, x_{n,j}) &= \lim_{x_{1,j}, \ldots, x_{n,j} \to \infty} P(\{\omega \in \Omega : X_1(\omega) \leq x_{1,j}, \ldots, X_n(\omega) \leq x_{n,j}\}) \\
&= P(\{\omega \in \Omega : X_1(\omega) < \infty, \ldots, X_n(\omega) < \infty\}) \\
&= 1
\end{aligned}$$

Since the sequence is arbitrary, we just showed that $\lim_{x_1,\ldots,x_n} \to \infty F(x_1,\ldots,x_n) = 1$.

(3) Given $(x_1,\ldots,x_n) \in \mathbb{R}^n$, and fix an index $i \in [n]$. Take any decreasing sequence $(y_n)_{j=1}^{\infty} \to x_i$. We have

We have for all $j = 1, 2, \ldots$

$$\{\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_i(\omega) \leq y_j, X_n(\omega) \leq x_n\} \subseteq \{\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_i(\omega) \leq y_{j+1}, X_n(\omega) \leq x_n\}$$

Moreover, we also have the sequence of monotone sets converges from below as $j \to \infty$

$$\{\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_i(\omega) \leq y_j, X_n(\omega) \leq x_n\} \to \{\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_i(\omega) \leq x_j, X_n(\omega) \leq x_n\}$$

Hence, by continuity of probability measure we have

$$\begin{aligned}\lim_{j\to\infty} F(x_1,\ldots,y_j,\ldots x_n) &= \lim_{j\to\infty} P(\{\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_i(\omega) \leq y_j, \ldots, X_n(\omega) \leq x_n\}) \\ &= P(\{\omega \in \Omega : X_1(\omega) \leq x_i, \ldots, X_n(\omega) \leq x_n\}) \\ &= F(x_1,\ldots,x_n)\end{aligned}$$

Since, the sequence is arbitrary and i is arbitrary, we showed that

$$\lim_{h\to 0^+} F(x_1,\ldots,x_i+h,\ldots,x_n) = F(x_1,\ldots,x_n)$$

□

Definition 3.9. *$(X_1,\ldots,X_n)$ has a joint p.m.f. $p : \mathbb{R}^n \to [0,1]$ if there exists a countable space $D \subseteq \mathbb{R}^2$ such that*

$$p(x_1,\ldots,x_n) = \begin{cases} P(X_1 = x_1,\ldots,X_n = x_n) > 0 & (x_1,\ldots,x_n) \in D \\ 0 & \textit{otherwise} \end{cases}$$

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

Theorem 3.10. *If n R.V.s $X_1,\ldots,X_n$ are discrete then there exists a function $p : \mathbb{R}^n \to [0,1]$ such that p is the joint p.m.f of $(X_1,\ldots,X_n)$*

Proof. Let S_i be the support of X_i for each i.

Since $X_1,\ldots,X_n$ are discrete variables, $S_1,\ldots,S_n$ are countable.

Similarly $S = S_1 \times \ldots \times S_n$ are countable.

It is clear that $(x_1,\ldots,x_n) \notin S$ then $P(X_1 = x_1,\ldots,X_n = x_n) = 0$.

Now, let $D = \{(x_1,\ldots,x_n) \in S : P(X_1 = x_1,\ldots,X_n = x_n) > 0\}$. Since $D \subseteq S$, D is countable.

Now, we define

$$p(x_1,\ldots,x_n) = \begin{cases} P(X_1 = x_1,\ldots,X_n = x_n) & (x_1,\ldots,x_n) \in D \\ 0 & \text{otherwise} \end{cases}$$

Here p is the joint p.m.f. of $(X_1,\ldots,X_n)$. □

Definition 3.11. *Random Vector $(X_1,\ldots,X_n)$ has a joint p.d.f. $f : \mathbb{R}^2 \to [0,\infty)$ if $\forall A \in \mathcal{B}(\mathbb{R}^n)$*

$$P((x_1,\ldots,x_n) \in A) = \int_A f(x_1,\ldots,x_n)dx_1\ldots x_n$$

This is equivalent to $\forall (x_1,\ldots,x_n) \in \mathbb{R}^n$

$$F(x_1,\ldots,x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f(y_1,\ldots,y_n)dy_1\ldots dy_n$$

Definition 3.12. *If $(X_1, \ldots, X_n)$ has joint p.d.f. f, then its support is $D = \{(x_1, \ldots, x_n) \in \mathbb{R}^2 : f(x_1, \ldots, x_n) > 0\}$*

Lemma 3.13. *If $(X_1, \ldots, X_n)$ has joint distribution function F and joint p.d.f. f, there exists $A \subseteq \mathbb{R}^n$ with $P((X_1, \ldots, X_n) \in A) = 1$ such that $\forall (x_1, \ldots, x_n) \in A$*

$$\frac{\partial^n}{\partial x_1 \ldots \partial x_n} F(x_1, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F(x_n, \ldots, x_1) = \ldots = f(x_1, \ldots, x_n)$$

Proof. NPOS. □

Example 3.14 (Bivariate Gaussian)**.** *Multivariate Gaussian is useful, but we will only cover up to bivariate.*

A bivariate Gaussian random vector $\boldsymbol{X} = (X, Y)$, with mean vector $\mu = (\mu_X, \mu_Y)$ and a correlation coefficient $\rho \in (-1, 1)$ and variances σ_X^2, σ_Y^2 has a joint p.d.f.

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\{-\frac{1}{2(1-\rho^2)}[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2p\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}]\}$$

Definition 3.15 (Marginal Distribution)**.** *Suppose that $(X_1, \ldots, X_n)$ has a joint p.m.f. p and support D. Suppose that X_i has support D_i, the marginal p.m.f. of X is*

$$p_{X_1}(x) = \sum_{x_2 \in D_2, \ldots, x_n \in D_n} p(x_1, \ldots, x_n) \qquad \forall x \in D_1$$

$$\vdots$$

$$p_{X_n}(x) = \sum_{x_1 \in D_1, \ldots, x_{n-1} \in D_{n-1}} p(x_1, \ldots, x_n) \qquad \forall x \in D_n$$

Definition 3.16. *Suppose $(X_1, \ldots, X_n)$ has joint p.d.f. f and support D where X_i has support D_i. The marginal p.d.f. of X is*

$$f_{X_1}(x) = \int_{D_2} \ldots \int_{D_n} f(x, x_2, \ldots, x_n) dx_n \ldots dx_2 \qquad \forall x \in D_1$$

Theorem 3.17. *Two continuous R.V.s does not necessarily have a joint p.d.f.*

Proof. Let X_1 be a random variable with p.d.f. $f_1(x_1)$ and support D_1 and set $X_2 = X_1$.

Assume (X_1, X_2) has a joint p.d.f $f(x_1, x_2)$. Define $L = \{(x_1, x_2) : x_1 = x_2\}$. Since $X_1 = X_2, P((X_1, X_2)) \in L = 1$

On the other hand, by definition of joint p.d.f.

$$P((X_1, X_2) \in L) = \int_L f(x_1, x_2) dx_1 dx_2 = 0$$

Contradiction, therefore the joint p.d.f. $f(x_1, x_2)$ does not exists. □

Example 3.18. *For a bivariate Gaussian random vector $\boldsymbol{X} = (X_1, X_2)$ with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and a covariance matrix Σ, the marginal distribution of X_i's are all Gaussian.*

3.2 Conditional Distribution

Definition 3.19. *Consider a random vector $\boldsymbol{X} = (X_1, X_2)$ 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 p.m.f. of $X_1|X_2$ is*

$$p_{X_1|X_2}(x_1|x_2) = \frac{p(x_1, x_2)}{p_2(x_2)}, \forall (x_1, x_2) \in D$$

Similarly,

$$p_{X_2|X_1}(x_2|x_1) = \frac{p(x_1,x_2)}{p_1(x_1)}, \forall (x_1,x_2) \in D$$

Proposition 3.20. *If discrete R.V.s X_1, X_2 are independent. And they have a joint p.m.f. p and support D then*

$$p_{X_1|X_2}(x_1|x_2) = p_1(x_1), \forall (x_1,x_2) \in D$$
$$p_{X_2|X_1}(x_2|x_1) = p_2(x_2), \forall (x_1,x_2) \in D$$

Proof. Suppose X_1, X_2 are independent, given any $(x_1, x_2) \in D$, we have

$$\begin{aligned}
p_{X_1|X_2}(x_1|x_2) &= \frac{p(x_1,x_2)}{p_2(x_2)} \\
&= \frac{P(X_1 \in \{x_1\} \cap X_2 \in \{x_2\})}{P(X_2 \in \{x_2\})} \\
&= \frac{P(X_1 \in \{x_1\})P(X_2 \in \{x_2\})}{P(X_2 \in \{x_2\})} && \text{[By definition of independence]} \\
&= P(X_1 \in \{x_1\}) \\
&= p_1(x_1)
\end{aligned}$$

Similarly

$$\begin{aligned}
p_{X_2|X_1}(x_2|x_1) &= \frac{p(x_1,x_2)}{p_1(x_1)} \\
&= \frac{P(X_1 \in \{x_1\} \cap X_2 \in \{x_2\})}{P(X_1 \in \{x_1\})} \\
&= \frac{P(X_1 \in \{x_1\})P(X_2 \in \{x_2\})}{P(X_1 \in \{x_1\})} && \text{[By definition of independence]} \\
&= P(X_2 \in \{x_2\}) \\
&= p_2(x_2)
\end{aligned}$$

□

Definition 3.21. *Consider a random vector $\boldsymbol{X} = (X_1, X_2)$ with a joint p.d.f. f and support D, with $X_1 \sim f_1, X_2 \sim f_2$. Furthermore, for each random variable X_i, assume its p.d.f. is f_i. The conditional p.d.f. of $X_1|X_2$ is*

$$f_{X_1|X_2}(x_1|x_2) = \frac{f(x_1,x_2)}{f_2(x_2)}, \forall (x_1,x_2) \in D$$

Similarly,

$$f_{X_2|X_1}(x_2|x_1) = \frac{f(x_1,x_2)}{f_1(x_1)}, \forall (x_1,x_2) \in D$$

Proposition 3.22. *Consider a random vector $\boldsymbol{X} = (X_1, X_2)$ with joint p.d.f. and support D then X_1 and X_2 are independent if and only if*

$$f_{X_1|X_2}(x_1|x_2) = f_1(x_1), \forall (x_1,x_2) \in D$$
$$f_{X_2|X_1}(x_2|x_1) = f_2(x_2), \forall (x_1,x_2) \in D$$

Proof. ($\Leftarrow$): Suppose

$$f_{X_1|X_2}(x_1|x_2) = f_1(x_1), \forall (x_1,x_2) \in D$$
$$f_{X_2|X_1}(x_2|x_1) = f_2(x_2), \forall (x_1,x_2) \in D$$

It is enough to show $F(x_1, x_2) = \prod_{i=1}^{2} F_{X_i}(x_i)$ for all $(x_1, x_2) \in D_1 \times D_2$.
We have

$$\begin{aligned} F(x_1, x_2) &= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{X_1|X_2}(y_1|y_2) f_2(y_2) dy_2 dy_1 \\ &= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_1(y_1) f_2(y_2) dy_2 dy_1 \\ &= \int_{-\infty}^{x_1} f_1(y_1) \int_{-\infty}^{x_2} f_2(y_2) dy_2 dy_1 \\ &= \int_{-\infty}^{x_1} f_1(y_1) F_{X_2}(x_2) dy_1 \\ &= \prod_{i=1}^{2} F_{X_i}(x_i) \end{aligned}$$

($\Rightarrow$): Suppose X_1, X_2 are independent, given any $(x_1, x_2) \in D$, we have

$$\begin{aligned} f_{X_1|X_2}(x_1|x_2) &= \frac{f(x_1, x_2)}{f_2(x_2)} \\ &= \frac{P(X_1 \in \{x_1\} \cap X_2 \in \{x_2\})}{P(X_2 \in \{x_2\})} \\ &= \frac{P(X_1 \in \{x_1\}) P(X_2 \in \{x_2\})}{P(X_2 \in \{x_2\})} \qquad \text{[By definition of independence]} \\ &= P(X_1 \in \{x_1\}) \\ &= f_1(x_1) \end{aligned}$$

Similarly

$$\begin{aligned} p_{X_2|X_1}(x_2|x_1) &= \frac{p(x_1, x_2)}{p_1(x_1)} \\ &= \frac{P(X_1 \in \{x_1\} \cap X_2 \in \{x_2\})}{P(X_1 \in \{x_1\})} \\ &= \frac{P(X_1 \in \{x_1\}) P(X_2 \in \{x_2\})}{P(X_1 \in \{x_1\})} \qquad \text{[By definition of independence]} \\ &= P(X_2 \in \{x_2\}) \\ &= f_2(x_2) \end{aligned}$$

□

Lemma 3.23. *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, $X|Y$ and $Y|X$ are also random variables.*

Proof. Given (X, Y) is a random vector with a joint p.m.f. To show $X|Y$ is also a random variable, given any $y \in \mathbb{R}$ such that $p_Y(y) > 0$, we need to show that $p_{X|Y}(x|y)$ is a proper p.d.f. for our fixed y.

First, given any $x \in \mathbb{R}$, since the usual p.d.f. gives $p(x, y) \geq 0$ we have

$$p_{X|Y}(x|y) = \frac{p(x, y)}{p_Y(y)} \geq 0$$

Finally, we must also show that this integrates to 1, indeed, we have

$$\sum_{x\in\mathbb{R}} p_{X|Y}(x,y)dx = \sum_{x_1\in D_1} \frac{p(x,y)}{p_Y(y)}dx = \frac{1}{p_Y(y)} \sum_{x\in D_1} p(x,y)dx = \frac{1}{p_Y(y)} p_Y(y) = 1$$

where the second last equality is derived by the marginal distribution.

$Y|X$ is also a proper p.d.f. also follows in the same way.

Given (X,Y) is a random vector with a joint p.d.f. To show $X|Y$ is also a random variable, given any $y \in \mathbb{R}$ such that $f_Y(y) > 0$, we need to show that $f_{X|Y}(x|y)$ is a proper p.d.f. for our fixed y.

First, given any $x \in \mathbb{R}$, since the usual p.d.f. gives $f(x,y) \geq 0$ we have

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} \geq 0$$

Finally, we must also show that this integrates to 1, indeed, we have

$$\int_{-\infty}^{\infty} f_{X|Y}(x,y)dx = \int_{-\infty}^{\infty} \frac{f(x,y)}{f_Y(y)}dx = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f(x,y)dx = \frac{1}{f_Y(y)} f_Y(y) = 1$$

where the second last equality is derived by the marginal distribution.

$Y|X$ is also a proper p.d.f. also follows in the same way. □

Example 3.24. *Suppose (X,Y) is bivariate Gaussian, then $X|Y, Y|X$ are both Gaussian.*

Proof. Note that the joint Gaussian random vector (X,Y) with $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and a covariance matrix

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

has a p.d.f.

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - 2p\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right]\right)$$

We will show $X|Y = y$ is Gaussian for all $y \in D_2$, $Y|X$ will follow similarly. Given $y \in D_2$, we

show that $f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for some μ, σ. We get

$$
\begin{aligned}
& f_{X|Y}(x|y) \\
=& \frac{f_{X,Y}(x,y)}{f_Y(y)} \\
=& \frac{\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right]\right)}{\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(y-\mu_2)^2}{2\sigma_2^2}\right)} \\
=& \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right] + \frac{(y-\mu_2)^2}{2\sigma_2^2}\right) \\
=& \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right] + \frac{(1-\rho^2)(y-\mu_2)^2}{2\sigma_2^2(1-\rho^2)}\right) \\
=& \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{\rho^2(y-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right]\right) \\
=& \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{x-\mu_1}{\sigma_1} - \rho\frac{y-\mu_2}{\sigma_2}\right]^2\right) \\
=& \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}\frac{\left[x-\mu_1-\rho\sigma_1\frac{y-\mu_2}{\sigma_2}\right]^2}{\sigma_1^2}\right) \\
=& \frac{1}{\sqrt{2\pi\sigma_1^2(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}\frac{\left[x-\left(\mu_1+\frac{\rho\sigma_1}{\sigma_2}(y-\mu_2)\right)\right]^2}{\sigma_1^2}\right) \\
=& \frac{1}{\sqrt{2\pi(\sigma_1^2(1-\rho^2)}} \exp\left\{-\frac{\left[x-\left(\mu_1+\frac{\rho\sigma_1}{\sigma_2}(y-\mu_2)\right)\right]^2}{2\sigma_1^2(1-\rho^2)}\right\}
\end{aligned}
$$

This shows that $X|Y$ is a normal distribution with $\mu = \mu_1 + \frac{\rho\sigma_1}{\sigma_2}(y-\mu_2), \sigma = 2\sigma_1^2(1-\rho^2)$.

□

Theorem 3.25 (Multivariate change of Variable)**.** *Let $X = (X_1, X_2)$ with a joint p.d.f. f_X with $g : \mathbb{R}^2 \to \mathbb{R}^2$ be one-to-one, continuously differentiable with continuously differentiable inverse h.*

That is, for any $\boldsymbol{x} = (x_1, x_2), \boldsymbol{y} = (y_1, y_2) \in \mathbb{R}^2$, such that $\boldsymbol{y} = g(\boldsymbol{x}) = (g_1(\boldsymbol{x}), g_2(\boldsymbol{x}))$ we have $\boldsymbol{x} = h(\boldsymbol{y}) = (h_1(\boldsymbol{y}), h_2(\boldsymbol{y}))$.

Let $Y = (Y_1, Y_2) = g(X)$, then Y has joint p.d.f. $f_Y(\boldsymbol{y}) = f_X(h(\boldsymbol{y}))|\det J|$ where $J_{i,j} = \frac{\partial h_i(\boldsymbol{y})}{\partial y_j}$

Proof. Take any $B \in \mathbb{R}^2$

$$
\begin{aligned}
P(Y \in B) &= P(X \in h(B)) \\
&= \int_{h(B)} f_X(x) dx_1 dx_2 \\
&= \int_B f_X(h(y))|\det J(y)| dy_1 dy_2 \qquad \text{[Regular change of variables theorem]}
\end{aligned}
$$

Hence, we indeed get a joint p.d.f. $f_Y(y) = f_X(h(y))|\det J|$. □

Remark 3.26. *Note that g being one-to-one is in terms of the entire function of $g : \mathbb{R}^2 \to \mathbb{R}^2$, that is, knowing the value of $g(x)$, we know what x is. However, the individual component function g_1 or g_2*

does not need to be one-to-one, i.e., it is not necessary that we must know what x is by simply knowing the value of only one of g_1 or g_2.

Example 3.27. *Let $X_1, X_2 \sim Exp(\lambda), \lambda > 0$ be independent. Then $X_1 + X_2 \sim Gamma(2, \lambda)$*

Proof. Since $X_1, X_2 \sim Exp(\lambda), f_{X_i}(x) = \lambda e^{-\lambda x} \mathbb{1}\{x \geq 0\}$. This gives

$$\begin{aligned} f(x_1, x_2) &= f_{X_1}(x_1) f_{X_2}(x_2) \\ &= \lambda^2 e^{-\lambda(x_1+x_2)} \mathbb{1}\{x_1 \geq 0, x_2 \geq 0\} \end{aligned}$$

Define bijective map:

$$\begin{aligned} Y_1 = g_1(X_1, X_2) &= X_1 + X_2 \\ Y_2 = g_2(X_1, X_2) &= X_1 \end{aligned}$$

Define

$$\begin{aligned} X_1 = h_1(Y_1, Y_2) &= Y_2 \\ X_2 = h_2(Y_1, Y_2) &= Y_1 - Y_2 \end{aligned}$$

Since $x_1, x_2 \geq 0$, we have $y_1 \geq 0, 0 \leq y_2 \leq y_1$.
This gives

$$J = \begin{pmatrix} \frac{\partial h_1(y_1,y_2)}{\partial y_1} & \frac{\partial h_1(y_1,y_2)}{\partial y_2} \\ \frac{\partial h_2(y_1,y_2)}{\partial y_1} & \frac{\partial h_2(y_1,y_2)}{\partial y_2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}$$

We have $\det(J) = -1$. Since g, h are all continuously differentiable, and g is one-to-one, by multivariate chain rule we have

$$\begin{aligned} f_{Y_1Y_2}(y_1, y_2) &= f_{X_1X_2}(y_2, y_1 - y_2) |\det(J)| \\ &= \lambda^2 e^{-\lambda y_1} \mathbb{1}\{y_1 \geq 0, 0 \leq y_2 \leq y_1\} \end{aligned}$$

This gives

$$f_{Y_1}(y_1) = \int_0^{y_1} \lambda^2 e^{-\lambda y_1} dy_2 = \lambda^2 y_1 e^{-\lambda y_1} \qquad \forall y_1 \geq 0$$

Since

$$f_{\Gamma(2,\lambda)}(x) = \lambda^2 x e^{-\lambda} x \qquad \forall x \geq 0$$

Hence, $X_1 + X_2 \sim \Gamma(2, \lambda)$ □

3.3 Mixture Distribution

We have now discussed the joint distributions of the random vector (X_1, X_2) when X_1 and X_2 are either both discrete or continuous. But what happens when one of X_1 and X_2 is discrete and the other is continuous? Obviously, there is no such thing as joint p.m.f. or p.d.f. in such cases. Therefore, we require something else entirely and such scenarios are best described by what we call a mixture distribution.

The key here is that even though we cannot define a joint p.m.f. nor a joint p.d.f., we can however define the joint distribution of (X_1, X_2) by conditioning. Without loss of generality, assume that X_1 is a continuous random variable while X_2 is discrete. The following identity always holds for any $A \subseteq \mathbb{R}$ and $x_2 \in D_2$ where D_2 is the support of X_2:

$$P(X_1 \in A, X_2 = x_2) = P(X_1 \in A \mid X_2 = x_2) P(X_2 = x_2)$$

Notice that $P(X_2 = x_2)$ is nothing more than the p.m.f. of X_2. For $P(X_1 \in A \mid X_2 = x_2)$, it is simply the conditional probability distribution of X_1 given the event that $X_2 = x_2$. Note that even though X_1 is continuous, $X_1 \mid X_2 = x_2$ may not necessarily be continuous. In fact, we have not even concluded that whether $X_1(\omega) \mid X_2(\omega) = x_2$ is indeed a random variable! In more advanced probability course, it can be shown rigorously that $X_1 \mid X_2 = x_2$ is a proper random variable. However, to properly demonstrate this, we require measure theory which is out of the scope of this course. For simplicity, we therefore will only restrict our attention to the case where $X_1 \mid X_2 = x_2$ is always a continuous random variable.

Definition 3.28. *Let X_2 be discrete, with support D_2 and p.m.f. $p(x_2)$. Suppose for every $x_2 \in D_2, X_1|X_2 = x_2$ has a continuous density*

$$f_{X_1|X_2}(x_1|x_2)$$

The marginal distribution of X_1 is called a mixture of continuous distributions with p.d.f.

$$f_{X_1}(x_1) = \sum_{x_2 \in D_2} f_{X_1|X_2}(x_1|x_2)P(x_2)$$

Let X_2 be continuous, with support D_2 and p.d.f. $f_{X_2}(x_2)$. Suppose for every $x_2 \in D_2, X_1|X_2 = x_2$ has a discrete density

$$p_{X_1|X_2}(x_1|x_2)$$

The marginal distribution of X_1 is called a continuous mixture of discrete distribution with p.m.f.

$$p_{X_1}(x_1) = \int_{D_2} p_{X_1|X_2}(x_1|x_2)f_{X_2}(x_2)dx_2$$

Example 3.29. *Suppose $X \sim Bernoulli(\frac{1}{2})$ and also Y be such that*

$$Y \sim \begin{cases} \mathcal{N}(-1,1) & \text{if } x = 0 \\ \mathcal{N}(1,1) & \text{if } x = 1 \end{cases}$$

This is different from

$$Y = \frac{1}{2}Z_1 + \frac{1}{2}Z_2$$

where $Z_1 \sim \mathcal{N}(2,1), Z_2 \sim \mathcal{N}(1,1)$

In this case, we have

$$f_Y(y) = \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{(y+1)^2}{2}\right) + \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{(y-1)^2}{2}\right)$$

Theorem 3.30 (Baye's Theorem)**.** *Suppose (X, Y) is a random vector.*

1. *If (X, Y) has joint p.d.f. f, and X and Y have p.d.f. f_X and f_Y. Then*

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

2. *If (X, Y) has joint p.m.f. p, and X and Y have p.m.f. p_X and p_Y. Then*

$$p_{Y|X}(y \mid x) = \frac{p(x,y)}{\sum_y p_{X|Y}(x \mid y)p_Y(y)} = \frac{p_{X|Y}(x \mid y)p_Y(y)}{\sum_y p_{X|Y}(x \mid y)p_Y(y)}.$$

3. *If (X, Y) is specified by a mixture distribution where X is a mixture of continuous distributions based on a discrete Y. Then*

$$p_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y)p_Y(y)}{\sum_y f_{X|Y}(x \mid y)p_Y(y)}.$$

4. *If (X, Y) is specified by a mixture distribution where X is a continuous mixture of discrete distributions based on a continuous Y. Then*

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

Proof. We will only prove 1. As others are similar by just using definitions.

Since

$$f(x, y) = f_{X|Y}(x, y)f_Y(y)$$

We just need to show that

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

Clearly,

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

□

Example 3.31. *Suppose $X|Y = y \sim \mathcal{N}(y, \sigma^2)$ where $Y \sim \mathcal{N}(\mu, \tau^2)$.*

Then, $Y|X = x \sim \mathcal{N}\left(\frac{\sigma^2\mu+\tau^2 x}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)$

Proof. By Bayes' Theorem, the conditional p.d.f. of $Y|X = x$

$$\begin{aligned}f_{Y|X}(y|x) &= \frac{f_{X|Y}(x|y)f_Y(y)}{\int f_{X|Y}(x|y)f_Y(y)dy} \\ &= \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)} \\ &\propto f_{X|Y}(x|y)f_Y(y) \\ &\propto \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right)\exp\left(-\frac{(y-\mu)^2}{2\tau^2}\right) \\ &= \exp\left(-\frac{(x-y)^2}{2\sigma^2} - \frac{(y-\mu)^2}{2\tau^2}\right)\end{aligned}$$

Therefore, for the coefficient of the y^2 term, we have

$$b^2 = \frac{1}{\sigma^{-2} + \tau^{-2}} = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}$$

For the coefficient of the y term, we have

$$x\sigma^{-2} + \mu\tau^{-2} = ab^{-2}$$

which gives

$$a = \frac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2}$$

Therefore,

$$Y \mid X = x \sim \mathcal{N}\left(\frac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)$$

For the marginal of X, we have

$$\begin{aligned}
f_X(x) &= \int f_{X,Y}(x,y)dy \\
&= \int f_{X|Y}(x|y)f_Y(y)dy \\
&= \frac{f_{X|Y}(x|y)f_Y(y)}{f_{Y|X}(y|x)} \qquad \text{[Baye's Theorem]} \\
&\propto \exp\left(-\frac{(x-y)^2}{2\sigma^2} - \frac{(y-\mu)^2}{2\tau^2} + \frac{(y-a)^2}{2b^2}\right) \\
&= \exp\left(-\frac{(x-\mu)^2}{2(\sigma^2+\tau^2)}\right) \qquad \text{[Check]}
\end{aligned}$$

Hence, $X \sim \mathcal{N}(\mu, \sigma^2 + \tau^2)$. □

3.4 Exercises

Question 3.32. *Consider the joint Gaussian random vector $\mathbf{X} = (X_1, X_2)$. Prove that X_i are independent if and only if all the off-diagonal entry of the covariance matrix Σ is zero, i.e., they are uncorrelated.*

Question 3.33. *Suppose $X|Y = y \sim Bin(n, y)$ and $Y \sim Beta(a, b)$ where $n \in \mathbb{N}$ and $a, b \in \mathbb{R}^+$. Find the conditional p.d.f. of $Y|X$.*

Question 3.34. *Suppose $X|Y = y \sim Gamma(\alpha, y)$ and $Y \sim Exp(\lambda)$ where $\alpha, \lambda \in \mathbb{R}^+$. Find the conditional p.d.f. of $Y|X$.*

Question 3.35. *Suppose $X|\Lambda = \lambda \sim Pois(\lambda)$ and $\Lambda \sim Exp(\theta)$ where $\theta > 0$. Find the conditional p.d.f. of $\Lambda|X$*

Question 3.36. *Let $X_1 \sim Unif(0, 1)$. For all $n \geq 2$, set $X_n|X_{n-1} \sim Unif(X_{n-1}, 1)$. Find the p.d.f. of X_n.*

Question 3.37. *Suppose $P(Z = 0) = P(Z = 1) = 1/2$, $Y \sim \mathcal{N}(0, 1)$, and that Y and Z are independent. Let $X = YZ$. What is the CDF of X?*

Question 3.38. *Suppose $P(Z = 1) = P(Z = -1) = 1/2$, $Y \sim \mathcal{N}(0, 1)$, and that Y and Z are independent. Let $X = YZ$.*

(a) Prove that $X \sim \mathcal{N}(0, 1)$.

(b) Prove that $P(|X| = |Y|) = 1$.

(c) Prove that X and Y are not independent.

Question 3.39. *If X and Y are independent on a probability space $(\Omega, \mathcal{F}, P)$, then $f(X)$ and $g(Y)$ are independent on $(\Omega, \mathcal{F}, \mathcal{P})$ 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 $A \in \mathcal{B}(\mathbb{R})$).*

Question 3.40. *Let $(\Omega, \mathcal{F}, P)$ 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.*

Question 3.41. *Let U_1, U_2 be two independent $Unif(0,1)$ random variables. Let*

$$Z_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2),$$
$$Z_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2).$$

Find the joint distribution of (Z_1, Z_2). Are Z_1 and Z_2 independent?

4 Convergence of Random Variables

Definition 4.1 (Convergence of Sequences and Functions)**.** *A real sequence $\{x_n\}_{n=1}^{\infty}$ converges to x if*

$$\forall \epsilon > 0, \exists N \in \mathbb{N},\ s.t.\ n > N \implies |x_n - x| < \epsilon$$

A sequence of functions $\{f_n\}_{n=1}^{\infty}$ with $f_n : X \to \mathbb{R}$ converges pointwise to another function $f : X \to \mathbb{R}$ if

$$\forall \epsilon > 0, \forall x \in X, \exists N_x \in \mathbb{N},\ s.t.\ n > N_x \implies |f_n(x) - f(x)| < \epsilon$$

A sequence of functions $\{f_n\}_{n=1}^{\infty}$ with $f_n : X \to \mathbb{R}$ converges uniformly to another function $f : X \to \mathbb{R}$ if

$$\forall \epsilon > 0, \forall n \in \mathbb{N}, \forall x \in X\ s.t.\ n > N \implies |f_n(x) - f(x)| < \epsilon$$

In this case, it is also equivalent to write $\sup_{x \in X} |f_n(x) - f(x)| \to 0$.

Definition 4.2 (Limiting process)**.** *Let $A \subseteq \mathbb{R}$.*
We define $u = \sup A \iff \forall \epsilon > 0, \exists x \in A, u - \epsilon < x \wedge \forall x \in A, x \leq u$.
We also define $l = \inf A \iff \forall \epsilon > 0, \exists x \in A, l + \epsilon > x \wedge \forall x \in A, x \geq l$

Definition 4.3. *Let $\{x_n\}_{n=1}^{\infty}$ be a real sequence*

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty} (\sup_{m \geq n} x_m) = \inf_{n \geq 0} \sup_{m \geq n} x_m$$

$$\liminf_{n\to\infty} x_n = \lim_{n\to\infty} (\inf_{m \geq n} x_m) = \sup_{n \geq 0} \inf_{m \geq n} x_m$$

We also have

$$\liminf_{n\to\infty}(-x_n) = -\limsup_{n\to\infty}(x_n)$$

Below are some useful fundamental theorems from analysis. We will not be covering their proofs as you can find them in any analysis textbooks.

Theorem 4.4. *Suppose $\{x_n\}_{n=1}^{\infty}$ be a real sequence, we have*

$$\liminf_{n\to\infty} x_n \leq \limsup_{n\to\infty} x_n$$

$\lim_{n\to\infty} x_n$ exists in $\overline{R} = [-\infty, \infty]$ if and only if

$$\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n$$

In this case, $\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n = \lim_{n\to\infty} x_n$.

Theorem 4.5 (Bolzano-Weierstrass Theorem)**.** *Every bounded sequence of real numbers has a convergent subsequence.*

Theorem 4.6. *If $\{x_n\}_{n=1}^{\infty}$ is monotone (does not have to be strict). Then $\lim_{n\to\infty} x_n$ always exists in the $\overline{\mathbb{R}} = [-\infty, \infty]$.*

Theorem 4.7. *Let $\{x_n\}_{n=1}^{\infty}$ to be a real sequence, define $S_n = \sum_{k=1}^{n} x_k$. If $\lim_{n\to\infty} S_n$ exists and is finite then $x_n \to 0$. Moreover, $\lim_{m\to\infty} \sum_{k=m}^{\infty} x_k = 0$.*

Theorem 4.8. *If $f : X \to \mathbb{R}$ is continuous at some point $x \in X$, then $\forall\{x_n\}$ such that $x_n \to x$ we have $f(x_n) \to f(x)$.*

Theorem 4.9. *If f_n convergence uniformly to f on A and each f_n is continuous on A then f is also continuous on A.*

4.1 Limit Events

Definition 4.10. *Let $A_1, \ldots \subseteq \Omega$. Define the limit events as*

$$\limsup_{n\to\infty} A_n = \{A_n \text{ i.o. (infinitely often)}\} = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

Similarly,

$$\liminf_{n\to\infty} A_n = \{A_n \text{ a.a. (almost always)}\} = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$$

Corollary 4.11. $P(A_n \text{ i.o.}) = 1 - P(A_n^c \text{ a.a.})$

Proof. We have

$$\{A_n \text{ i.o.}\} = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k = \bigcap_{n=1}^{\infty} \left(\bigcap_{k=n}^{\infty} A_k^c \right)^c = \left(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c \right)^c = \{A_n^c \text{ a.a.}\}^c$$

□

Remark 4.12. *We require the introduction and discussion of limit events to talk about convergence of random variables. To better understand the limit events, we use $\{A_n \text{ a.a.}\}$ as an example.*

For the sequence of events $\{A_n\}$ to occur almost always (a.a.), it means that there exists $n \in \mathbb{N}^+$ such that for all $k \geq n$, A_k occurs. Now let $\varepsilon > 0$ and $A_n = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| < \varepsilon\}$ for a sequence of R.V.s $\{X_n\}_{n=1}^{\infty}$ and X. If $\{A_n \text{ a.a.}\}$ happens for all $\varepsilon > 0$, then we have the event $\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}$. Under the same setup, if there is a $\varepsilon > 0$ where $\{A_n^c \text{ i.o.}\}$ happens, then we have the event that $\{\omega \in \Omega : X_n(\omega) \nrightarrow X(\omega)\}$. This is exactly why we need the notion and precise definition of limit events so that we can properly talk about the convergence of random variables.

To translate the above to set operations, notice the quantifier "for any" and "there exists" have one-to-one correspondence to intersection and union. Specifically, the statement "there exists $n \in \mathbb{N}^+$" translates to "$\bigcup_{n=1}^{\infty}$" since it only needs to happen for one $n \in \mathbb{N}^+$, while "for all $k \geq n$, A_k occurs" translates to "$\bigcap_{k=n}^{\infty} A_k$" since we need all of A_k to occur when $k \geq n$. Combining things together, we have $\{A_n \text{ a.a.}\} = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$.

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

Proposition 4.13. $P(A_n \text{ a.a.}) \leq \liminf_{n\to\infty} P(A_n) \leq \limsup_{n\to\infty} P(A_n) \leq P(A_n \text{ i.o.})$

Proof. Note that

$$\bigcap_{k=n}^{\infty} A_k \subseteq \bigcap_{k=n+1}^{\infty} A_k \qquad \forall n$$

We have

$$\begin{aligned}
P(A_n \text{ a.a.}) &= P\left(\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k\right) \\
&= P\left(\lim_{n\to\infty}\bigcap_{k=n}^{\infty} A_k\right) \\
&= \lim_{n\to\infty} P\left(\bigcap_{k=n}^{\infty} A_k\right) \\
&= \liminf_{n\to\infty} P\left(\bigcap_{k=n}^{\infty} A_k\right) \qquad \text{[Bounded monotone sequence]} \\
&\leq \liminf_{n\to\infty} P(A_n) \\
&\leq \limsup_{n\to\infty} P(A_n) \\
&\leq \limsup_{n\to\infty} P\left(\bigcup_{k=n}^{\infty} A_k\right) \\
&= \lim_{n\to\infty} P\left(\bigcup_{k=n}^{\infty} A_k\right) \text{[Bounded monotone sequence]} \\
&= P\left(\lim_{n\to\infty}\bigcup_{k=n}^{\infty} A_k\right) \\
&= P\left(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k\right) \\
&= P(A_n \text{ i.o.})
\end{aligned}$$

□

Theorem 4.14 (Borel Cantelli Lemma)**.** *1. If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n \text{ i.o.}) = 0$*

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

Proof. (1) We have

$$\bigcup_{k=n+1}^{\infty} A_k \subseteq \bigcup_{k=n}^{\infty} A_k \qquad \forall n$$

Hence,

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

(2) We have

$$\bigcap_{k=n}^{\infty} A_k^c \subseteq \bigcap_{k=n+1}^{\infty} A_k^c \qquad \forall n$$

Hence

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

Hence, $P(A_n \text{ i.o.})) \geq 1 \implies P(A_n \text{ i.o.}) = 1$. □

Example 4.15. *For Part 1 of the Theorem 4.14, the converse is not true.*
Take $A_n = [0, \frac{1}{n}]$. *Then*

$$\{A_n \text{ i.o.}\} = \bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} [0, \frac{1}{k}] = \{0\}$$

However,

$$\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

Example 4.16. *Independence is always needed for Part 2 of the Theorem 4.14.*
Consider a fair coin toss, define $c_1, c_2, \ldots$ *such that* $\forall i \in \mathbb{R}^+, c_i = 1$ *if we toss head.*
Let $A_1, A_2, \ldots$ *so that* $A_i = \{c_i = 1\}$.

We have

$$\sum_{i=1}^{\infty} P(A_i) = \infty$$

But $P(A_n \text{ i.o.}) = \frac{1}{2}$.

4.2 Convergence

Definition 4.17. *A sequence of R.V.s* X_n *converges almost surely to* X *if*

$$P\left(\lim_{n\to\infty} X_n = X\right) = P(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$$

this is denoted as $X_n \to X$ *a.s.*

Proposition 4.18. *If* $\forall \epsilon > 0, P(|X_n - X| > \epsilon \text{ i.o.}) = 0$ *then* $X_n \to X$ *a.s.*

Proof. Recall that $\lim_{n\to\infty} X_n(\omega) = X(\omega) \iff \forall \epsilon > 0, |X_n(\omega) - X(\omega)| < \epsilon$ a.a.

$$\begin{aligned} P(\lim_{n\to\infty} X_n = X) &= P(\forall \epsilon > 0, |X_n(\omega) - X(\omega)| < \epsilon \text{ a.a.}) \\ &= 1 - P(\exists \epsilon > 0, |X_n(\omega) - X(\omega)| \geq \epsilon \text{ i.o.}) \end{aligned}$$

If we have $\exists \epsilon > 0, |X_n(\omega) - X(\omega)| \geq \epsilon$ i.o., then there exists $\epsilon \in \mathbb{Q}^+$ such that $|X_n(\omega) - X(\omega)| \geq \epsilon$ i.o.

Hence,

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

□

Corollary 4.19. *If* $\forall \epsilon > 0, \sum_{n=1}^{\infty} P(|X_n(\omega) - X(\omega)| \geq \epsilon) < \infty$ *then* $X_n \to X$ *a.s.*

Proof. If $\forall \epsilon > 0, \sum_{n=1}^{\infty} P(|X_n(\omega) - X(\omega) \geq \epsilon) < \infty$. Then by the Borel Cantelli Lemma, we have $\forall \epsilon > 0, P(|X_n(\omega) - X(\omega)| \geq \epsilon \text{ i.o.}) = 0$. And by 4.18, $X_n \to X$ a.s. □

Definition 4.20. *A sequence of R.V.s* X_n *converges in probability to* X *if for all* $\epsilon > 0$

$$\lim_{n\to\infty} P(|X_n - X| \leq \epsilon) = 1$$

This is denoted by $X_n \xrightarrow{P} X$

Example 4.21. *Let* $X_n \sim Exp(n)$*, then* $X_n \xrightarrow{P} 0$

Proof. Fix $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - 0| > \epsilon) = \lim_{n\to\infty} P(X_n > \epsilon) = \lim_{n\to\infty} e^{-n\epsilon} = 0$$

□

Proposition 4.22. *If $X_n \to X$ a.s. then $X_n \xrightarrow{P} X$*

Proof. Fix $\epsilon > 0$, denote $E_n = \{\omega \in \Omega : \exists m \geq n, |X_m(\omega) - X(\omega)| \geq \epsilon\}$.
Notice that $E_{n+1} \subseteq E_n$, if $\omega \in \bigcap_{n=1}^{\infty} E_n$ then we get $X_n(\omega) \not\to X(\omega)$. Thus, we get

$$\begin{aligned}
\lim_{n\to\infty} P(|X_n(\omega) - X(\omega)| \geq \epsilon) &\leq \lim_{n\to\infty} P(E_n) \\
&= P(\bigcap_{n=1}^{\infty} E_n) \qquad \text{[Continuity of probability]} \\
&\leq P(X_n \not\to X) \\
&= 0
\end{aligned}$$

□

Remark 4.23. *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 4.24. *Let X_n to be independent with $P(X_n = 1) = \frac{1}{n}, P(X_n = 0) = 1 - \frac{1}{n}$.*
This shows $X_n \xrightarrow{P} 0$.
But by Borel Cantelli, $P(X_n = 1$ i.o.$) = P(X_n \not\to X) = 1$. So $P(X_n \to 0) = 0$

Theorem 4.25. *If $X_n \xrightarrow{P} X$, there exists a subsequence $X_{n_k} \to X$ a.s.*

Proof. Since $X_n \xrightarrow{P} X, \lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$.
By definition $\forall k \in \mathbb{N}, \exists n_k$ s.t. $\forall n \geq n_k, P(|X_n - X|) > 2^{-k}) \leq 2^{-k}$.
Further, choose $n_{k+1} \geq n_k$. We can define

$$A_k = \{\omega \in \Omega : |X_{n_k}(\omega) - X(\omega)| > 2^{-k}\}$$

We get

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

By Borel Cantelli, $P(A_k$ i.o.$) = 0$.
Finally, observe $|X_{n_k}(\omega) - X(\omega)| > 2^{-k}$ only for finitely many times then $X_{n_k}(\omega) \to X(\omega)$.
Hence,

$$1 = P(A_k^c \text{ a.a.}) \leq P(X_{n_k} \to X) \leq 1$$

□

Theorem 4.26 (Continuous Mapping Theorem)**.** *If f is a continuous function*

1. *$X_n \to X$ a.s. $\implies f(X_n) \to f(X)$ a.s.*
2. *$X_n \xrightarrow{P} X \implies f(X_n) \xrightarrow{P} f(x)$*

Proof. Note that f is continuous $\iff$ for all $\omega, X_n(\omega) \to X(\omega) \implies f(X_n(\omega)) \to f(X(\omega))$.

(1) We have $P(\{\omega : X_n(\omega) \to X(\omega)\}) \leq P(\{\omega : f(X_n(\omega)) \to f(X(\omega)\})$. Since $P(\{\omega : X_n(\omega) \to X(\omega)\}) = 1$. We have $P(\{\omega : f(X_n(\omega)) \to f(X(\omega)\}) = 1$. Hence, $f(X_n(\omega)) \to f(X(\omega))$ a.s.

(2) We have $\lim_{n\to\infty} P(|X_n - X| \leq \epsilon) = 1 \implies \lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$. We want to show $\lim_{n\to\infty} P(|f(X_n) - f(X)| < \epsilon = 1$, for all ϵ.

Take any $\epsilon > 0$, as f is continuous, there exists a $\delta > 0$ s.t. $|X_n(\omega) - X(\omega)| \leq \delta \implies |f(X_n(\omega)) - f(X(\omega))| \leq \epsilon$, we obtain

$$1 = \lim_{n\to\infty} P(|X_n - X| < \delta) \leq \lim_{n\to\infty} P(|f(x_n) - f(x)| < \epsilon)$$

hence, $f(X_n) \xrightarrow{P} f(X)$ □

Example 4.27. *If $X_n \xrightarrow{P} 1, \log(X_n) \xrightarrow{P} 0$.*

Proof. Since $X_n \xrightarrow{P} 1$, we have for all ϵ

$$\lim_{n\to\infty} P(X_n - 1) \leq \epsilon) = 1$$

By Continuous Mapping Theorem, $f : \mathbb{R}^+ \to \mathbb{R}$ defined by $f(x) = \log(x)$ is a continuous function at $x = 1$. Therefore, we have $X_n \xrightarrow{P} 1 \implies f(X_n) \xrightarrow{P} \log(1) = 0$. □

4.3 Convergence of Random Vectors

Definition 4.28. *A sequence of random vectors $(X_n^{(1)}, \ldots, X_n^{(d)})$ converges almost surely to $(X^{(1)}, \ldots, X^{(d)})$ if*

$$P(\lim_{n\to\infty} X_n^{(1)} = X^{(1)} \wedge \ldots \wedge \lim_{n\to\infty} X_n^{(d)} = X^{(d)}) = 1$$

Definition 4.29. *A sequence of random vectors $(X_n^{(1)}, \ldots, X_n^{(d)})$ converges in probability to $(X^{(1)}, \ldots, X^{(d)})$ if $\forall \epsilon > 0$*

$$\lim_{n\to\infty} P(|(X_n^{(1)}, \ldots, X_n^{(d)}) - (X^{(1)}, \ldots, X^{(d)})| \geq \epsilon) = 0$$

Proposition 4.30. *$(X_n^{(1)}, \ldots, X_n^{(d)}) \to (X^{(1)}, \ldots, X^{(d)})$ a.s. if and only if $X_n^{(1)} \to X^{(1)}$ a.s. and ... and $X_n^{(d)} \to X^{(d)}$ a.s.*

Proof. ($\Rightarrow$) Suppose $(X_n^{(1)}, \ldots, X_n^{(d)}) \to (X^{(1)}, \ldots, X^{(d)})$ a.s. Then there exists a set $A \subseteq \Omega$ with $\mathbb{P}(A) = 1$ such that for all $\omega \in A$,

$$(X_n^{(1)}(\omega), \ldots, X_n^{(d)}(\omega)) \to (X^{(1)}(\omega), \ldots, X^{(d)}(\omega))$$

in $\mathbb{R}^d$. Then by definition of convergence in $\mathbb{R}^d$, it must be that:

$$X_n^{(1)}(\omega) \to X^{(1)}(\omega) \quad \text{and} \quad \ldots \quad \text{and} \quad X_n^{(d)}(\omega) \to X^{(d)}(\omega)$$

for all $\omega \in A$. Hence, $X_n^{(1)} \to X^{(1)}$ a.s. and ... and $X_n^{(d)} \to X^{(d)}$ a.s.

($\Leftarrow$) Suppose $X_n^{(1)} \to X^{(1)}$ a.s. and ... and $X_n^{(d)} \to X^{(d)}$ a.s. Let:

$$A_i = \{\omega \in \Omega : X_n^{(i)}(\omega) \to X^{(i)}(\omega)\} \qquad \forall i \in [d]$$

Then $P(A_i) = 1$ for all $i \in [d]$. Therefore, $P(\bigcap_{i=1}^d A_i) = 1$. Thus, $(X_n^{(1)}, \ldots, X_n^{(d)}) \to (X^{(1)}, \ldots, X^{(d)})$ a.s. □

Proposition 4.31. $(X_n^{(1)},\dots,X_n^{(d)}) \xrightarrow{P} (X^{(1)},\dots,X^{(d)}) \iff X_n^{(1)} \xrightarrow{P} X^{(1)}$ *and ... and* $X_n^{(d)} \xrightarrow{P} X^{(d)}$.

Proof. We have

$$
\begin{aligned}
&(X_n^{(1)},\dots,X_n^{(d)}) \xrightarrow{P} (X^{(1)},\dots,X^{(d)}) \\
\iff &\lim_{n\to\infty} P(\|(X_n^{(1)},\dots,X_n^{(d)}) - (X^{(1)},\dots,X^{(d)})\| \geq \epsilon) = 0 \qquad \forall \epsilon > 0 \\
\iff &\lim_{n\to\infty} P(\|(X_n^{(1)},\dots,X_n^{(d)}) - (X^{(1)},\dots,X^{(d)})\| < \epsilon) = 1 \qquad \forall \epsilon > 0 \\
\iff &\lim_{n\to\infty} P(\omega\in\Omega : \|(X_n^{(1)}(\omega),\dots,X_n^{(d)}(\omega)) - (X^{(1)}(\omega),\dots,X^{(d)}(\omega))\| < \epsilon) = 1 \qquad \forall \epsilon > 0 \\
\iff &\lim_{n\to\infty} P(\omega\in\Omega : (X_n^{(1)}(\omega),\dots,X_n^{(d)}(\omega)) \to (X^{(1)}(\omega),\dots,X^{(d)}(\omega))) = 1 \\
\iff &\lim_{n\to\infty} P(\omega\in\Omega : X_n^{(1)}(\omega) \to X^{(1)}(\omega),\dots,X_n^{(d)}(\omega) \to X^{(d)}(\omega)) = 1 \qquad \text{[Converge } \iff \text{ Converge component-wise]} \\
\iff &\lim_{n\to\infty} P(\omega\in\Omega : \|X_n^{(1)}(\omega) - X^{(1)}(\omega)\| < \epsilon,\dots,\|X_n^{(d)}(\omega) \to X^{(d)}(\omega)\| < \epsilon) = 1 \qquad \forall \epsilon > 0 \\
\iff &\lim_{n\to\infty} P(\omega\in\Omega : |X_n^{(1)}(\omega) - X^{(1)}(\omega)| < \epsilon,\dots,|X_n^{(d)}(\omega) \to X^{(d)}(\omega)| < \epsilon) = 1 \qquad \forall \epsilon > 0 \qquad \text{[Equivalence of norms]}
\end{aligned}
$$

Finally, we show this is equivalent to

$$\lim_{n\to\infty} P(\omega\in\Omega : |X_n^{(i)}(\omega) - X^{(i)}(\omega)| < \epsilon) = 1 \qquad \forall \epsilon > 0, \forall i \in [d]$$

($\Rightarrow$) : This is obvious, suppose

$$\lim_{n\to\infty} P(\omega\in\Omega : |X_n^{(1)}(\omega) - X^{(1)}(\omega)| < \epsilon,\dots,|X_n^{(d)}(\omega) \to X^{(d)}(\omega)| < \epsilon) = 1 \qquad \forall \epsilon > 0$$

For all $i \in [d]$, we have

$$P(\omega\in\Omega : |X_n^{(1)}(\omega) - X^{(1)}(\omega)| < \epsilon,\dots,|X_n^{(d)}(\omega) \to X^{(d)}(\omega)| < \epsilon) \leq P(\omega\in\Omega : |X_n^{(i)}(\omega) - X^{(i)}(\omega)| < \epsilon)$$

Take the limit we get

$$1 = \lim_{n\to\infty} P(\omega\in\Omega : |X_n^{(1)}(\omega) - X^{(1)}(\omega)| < \epsilon,\dots,|X_n^{(d)}(\omega) \to X^{(d)}(\omega)| < \epsilon) \leq \lim_{n\to\infty} P(\omega\in\Omega : |X_n^{(i)}(\omega) - X^{(i)}(\omega)| < \epsilon) \leq 1$$

($\Leftarrow$) : Suppose

$$\lim_{n\to\infty} P(\omega\in\Omega : |X_n^{(i)}(\omega) - X^{(i)}(\omega)| < \epsilon) = 1 \qquad \forall \epsilon > 0, \forall i \in [d]$$

We have

$$
\begin{aligned}
&\lim_{n\to\infty} P(\omega\in\Omega : |X_n^{(1)}(\omega) - X^{(1)}(\omega)| < \epsilon,\dots,|X_n^{(d)}(\omega) \to X^{(d)}(\omega)| < \epsilon) \\
=&\lim_{n\to\infty} P(\bigcap_{i=1}^{d} \{\omega\in\Omega : |X_n^{(i)}(\omega) - X^{(i)}(\omega)| < \epsilon\}) \\
=&1 - \lim_{n\to\infty} P(\bigcup_{i=1}^{d} \{\omega\in\Omega : |X_n^{(i)}(\omega) - X^{(i)}(\omega)| \geq \epsilon\}) \\
=&1 - \sum_{i=1}^{d} (-1)^{i+1} \sum_{1\leq a_1<\dots<a_i\leq d} P\left(\bigcap_{l=1}^{i} \{\omega\in\Omega : |X_n^{(i)}(\omega) - X^{(i)}(\omega)| \geq \epsilon\}\right) \\
=&1 - \sum_{i=1}^{d} (-1)^{i+1} \sum_{1\leq a_1<\dots<a_i\leq d} \sum_{1\leq a_1<\dots<a_i\leq d} 0 \\
=&1
\end{aligned}
$$

□

Theorem 4.32 (Multivariate Continuous Mapping Theorem)**.** *If $f : \mathbb{R}^2 \to \mathbb{R}$ is continuous*

1. $(X_n^{(1)}, \ldots, X_n^{(d)}) \to (X^{(1)}, \ldots, X^{(d)})$ *a.s.* $\implies f(X_n^{(1)}, \ldots, X_n^{(d)}) \to f(X^{(1)}, \ldots, X^{(d)})$ *a.s.*

2. $(X_n^{(1)}, \ldots, X_n^{(d)}) \xrightarrow{P} (X^{(1)}, \ldots, X^{(d)}) \implies f(X_n^{(1)}, \ldots, X_n^{(d)}) \xrightarrow{P} f(X^{(1)}, \ldots, X^{(d)})$

Proof. Note that f is continuous $\iff$ for all $\omega, (X_n^{(1)}(\omega), \ldots, X_n^{(d)}(\omega)) \to (X^{(1)}(\omega), \ldots, X^{(d)}(\omega)) \implies f(X_n^{(1)}(\omega), \ldots, X_n^{(d)}(\omega)) \to f(X^{(1)}(\omega), \ldots, X^{(d)}(\omega))$.

(1) We have $P(\{\omega : (X_n^{(1)}(\omega), \ldots, X_n^{(d)}(\omega)) \to (X^{(1)}(\omega), \ldots, X^{(d)}(\omega))\}) \leq P(\{\omega : f(X_n^{(1)}(\omega), \ldots, X_n^{(d)}(\omega)) \to f(X^{(1)}(\omega), \ldots, X^{(d)}(\omega))\})$.

Since $P(\{\omega : (X_n^{(1)}(\omega), \ldots, X_n^{(d)}(\omega)) \to (X^{(1)}(\omega), \ldots, X^{(d)}(\omega))\}) = 1$. We have $P(\{\omega : f(X_n^{(1)}(\omega), \ldots, X_n^{(d)}(\omega)) \to f(X^{(1)}(\omega), \ldots, X^{(d)}(\omega))\}) = 1$. Hence, $f(X_n^{(1)}, \ldots, X_n^{(d)}) \to f(X^{(1)}, \ldots, X^{(d)})$ a.s.

(2) We have for all $\epsilon > 0$

$$\lim_{n\to\infty} P(|(X_n^{(1)}, \ldots, X_n^{(d)}) - (X^{(1)}, \ldots, X^{(d)})| \geq \epsilon) = 0 \implies \lim_{n\to\infty} P(|(X_n^{(1)}, \ldots, X_n^{(d)}) - (X^{(1)}, \ldots, X^{(d)})| < \epsilon) = 1$$

We want to show for all $\epsilon > 0$

$$\lim_{n\to\infty} P(|f(X_n^{(1)}, \ldots, X_n^{(d)}) - f(X^{(1)}, \ldots, X^{(d)})| < \epsilon) = 1$$

Take any $\epsilon > 0$, as f is continuous, there exists a $\delta > 0$ s.t. for all $|(X_n^{(1)}, \ldots, X_n^{(d)}) - (X^{(1)}, \ldots, X^{(d)})| \leq \delta \implies |f(X_n^{(1)}, \ldots, X_n^{(d)}) - f(X^{(1)}, \ldots, X^{(d)})| \leq \epsilon$, we obtain

$$1 = \lim_{n\to\infty} P(|(X_n^{(1)}, \ldots, X_n^{(d)}) - (X^{(1)}, \ldots, X^{(d)})| < \delta) \leq \lim_{n\to\infty} P(|f(X_n^{(1)}, \ldots, X_n^{(d)}) - f(X^{(1)}, \ldots, X^{(d)})| < \epsilon)$$

hence, $f(X_n^{(1)}, \ldots, X_n^{(d)}) \xrightarrow{P} f(X^{(1)}, \ldots, X^{(d)})$ □

4.4 Exercises

Question 4.33. *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 4.2.3 are strict.*

Question 4.34. *Prove that for any collections $\{A_n\}$ and $\{B_n\}$,*

$$\limsup(A_n \cap B_n) \subseteq \limsup A_n \cap \limsup B_n$$

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

Question 4.35. *Prove that if $X_n \to X$ a.s., for all $\varepsilon > 0$*

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

Question 4.36. *Find a sequence of non-independent X_n and X such that $X_n \xrightarrow{P} X$ but $X_n \not\to X$ a.s.*

Question 4.37. *Show that $X_n \to X$ a.s. or $X_n \xrightarrow{P} X$ if and only if $(X_n - X) \to 0$ a.s. or $(X_n - X) \xrightarrow{P} 0$ respectively.*

Question 4.38. *Show that if $X_n - a_n \xrightarrow{P} 0$ and $a_n \to a$, then $X_n \xrightarrow{P} a$.*

Question 4.39. *If $X_n \xrightarrow{P} X$ and $X_n \leq X_{n+1}$ for all n, then $X_n \to X$ a.s.*

Question 4.40. *Let $\delta, \varepsilon > 0$, and let $X_1, X_2, \ldots$ be a sequence of independent non-negative random variables such that $P(X_i \geq \delta) \geq \varepsilon$ for all i. Prove that $\sum_{i=1}^{\infty} X_i = \infty$ a.s.*

Question 4.41. *Give an example of $X_1, X_2, \ldots$ such that $X_n/n \xrightarrow{P} 0$ and $X_n/n^2 \to 0$ a.s., but $\mathbb{P}(X_n/n \to 0) < 1$.*

5 Expectations, Conditional Expectations

5.1 Expectations

NPOS starts:

After discussion on random variables, the next important topic is then expectation. As seen in introductory probability courses (STA 257), expectation of a random variable X or its function $g(X)$ is nothing more than just integration: if X is discrete, then $E[g(X)] = \sum g(x)p(x)$; if X is continuous, then $E[g(X)] = \int g(x)f(x)dx$. However, as we shall see below, these definitions can break in certain situations.

Let $X \sim \text{Uniform}(0,1)$, and define the function

$$g(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \cap (0,1) \\ 0 & \text{otherwise} \end{cases}$$

We ask the question: what is $E[g(X)]$? Based on what we know from STA 257, since X is a continuous random variable, we have

$$E[g(X)] = \int_0^1 g(x)f(x)\,dx = \int_0^1 g(x)dx$$

However, g is a nowhere continuous function on $(0,1)$, the integral above, therefore, does not exist! Both approaches seem entirely reasonable but they gave completely different conclusions. What is happening here?

The key issue is that the notion of integration for random variables, that is, a measurable function that maps from the sample space to the real line, cannot be generally defined for all random variables through the standard Riemann integration that we are familiar with from first and second year calculus.

Riemann integration, as we all know, is built upon the idea of dividing up the domain of a function into many tiny rectangular pieces, and summing up the area of all these pieces. A crucial requirement for Riemann integral to exist is the limit of the lower Riemann sum is the same as the limit of the upper Riemann sum. That is, if f is a bounded function on $[x_0, x_n]$, then for any subdivision $\{x_0, x_1, \ldots, x_n\}$ satisfying $\lim_{n\to\infty} \max_{i\in[n]} |x_i - x_{i-1}| = 0$,

$$\underbrace{\lim_{n\to\infty} \sum_{i=1}^{n} \inf_{x\in(x_{i-1},x_i]} f(x)(x_i - x_{i-1})}_{\text{Lower sum}} = \underbrace{\lim_{n\to\infty} \sum_{i=1}^{n} \sup_{x\in(x_{i-1},x_i]} f(x)(x_i - x_{i-1})}_{\text{Upper sum}}$$

However, the above is not satisfied by the function g in our previous example since the LHS of the above is 0 while the RHS is 1.

To address this issue, the integration of a random variable must rely on an entirely different idea of integration called Lebesgue integration. Unlike Riemann integration, Lebesgue integration partitions the range (the y-axis) of the function into intervals and sums horizontal slices. Each slice collects together all the points in the domain where the function takes on values within a given range, and multiplies that range value by the measure (e.g., probability) of those points.

Mathematically, let $f : \mathcal{X} \to \mathbb{R}$ be a non-negative, bounded, measurable function and let $\{y_0, \ldots, y_n\}$ be a subdivision of the range $(y_0, y_n]$ such that $\lim_{n\to\infty} \max_{i\in[n]} |y_i - y_{i-1}| = 0$. The Lebesgue integral of f with respect to a measure μ is defined by

$$\int_{\mathcal{X}} f d\mu = \lim_{n\to\infty} \sum_{i=1}^{n} y_i \mu(\{x \in \mathcal{X} : f(x) \in (y_{i-1}, y_i]\})$$

In terms of a non-negative, bounded random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$, the rigorous definition of its expectation is then

$$E[X] = \int_\Omega X dP = \lim_{n\to\infty} \sum_{i=1}^{n} x_i P(\{\omega \in \Omega : x_{i-1} < X(\omega) \leq x_i\})$$

where $\{x_0, x_1, \ldots, x_n\}$ is a subdivision of the range of $X \in (x_0, x_n]$ with $\lim_{n\to\infty} \max_{i\in[n]} |x_i - x_{i-1}| = 0$. For general real-valued, non-bounded random variable X, the above definition can be generalised and we arrive at the same consistent definition. Note that if both the Lebesgue integral and Riemann integral of a function exists, then they are the same.

A massive advantage of Lebesgue integration is that we can generalise the notion of integration to any measurable functions whose domains can be any abstract space since we are no longer subdividing the domain. This is a big motivation for the Lebesgue formulation of the expectation of random variables.

Based on the Lebesgue integration formulation of the expectation of a random variable, we can see for the random variable $g(X)$ in our previous example, its expectation indeed exists and is equal to zero. In fact, what we should have done here is to view $Y = g(X)$ as a discrete random variable where $Y = 1$ with probability zero and $Y = 0$ with probability one. Computing the expectation of Y this way is the correct approach, but the deeper underlying logic is not because Y is a discrete random variable, rather that it is based on how expectation is actually defined for any random variable through Lebesgue integration.

Everything we have discussed previously is to provide you with a more rigorous and deeper understanding on how expectation of a random variable is defined. Further discussion on this requires extensive measure theory and is thus out of the scope of this course. In addition, many of the rigorous construction and derivation of the properties of expectations require us to build up from measure theory, we thus will assume many of these construction and properties to be granted and not cover their proof. It goes without saying that Lebesgue integration will not be tested in this course.

Below we review the basics of expectation and their properties, which are simply special cases that follow from the Lebesgue integration formulation.

NPOS ends.

Proposition 5.1 (Expectation for Discrete Random Variables)**.** *If X is a discrete random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with support $\mathcal{X} = \{x_1, x_2, \ldots\}$ and p.m.f. $p_X(x_i) = P(X = x_i)$, the expectation of X then reduces from the Lebesgue integral with respect to P to a counting integral over the support:*

$$E[X] = \int_\Omega X dP = \sum_{x_i \in \mathcal{X}} x_i p_X(x_i)$$

Example 5.2. *Take an arbitrary $A \subseteq \Omega$, let $Y(\omega) = \mathbb{1}_{\omega\in A}$ then $E(Y) = 0p_X(0) + 1p_X(1) = P(A)$.*

Example 5.3. *P is uniform measure on $\Omega = [0, 1]$, and*

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

Proposition 5.4 (Expectation for Continuous Random Variables)**.** *If X is a continuous random variable on $(\Omega, \mathcal{F}, P)$ with p.d.f. $f_X(x)$, the expectation of X then reduces from the Lebesgue integral with respect to P to a Riemann integral over the real line:*

$$E[X] = \int_\Omega X dP = \int_{\mathbb{R}} x f_X(x) dx$$

Definition 5.5. *A random variable X is called integrable if $E[|X|] < \infty$.*

Remark 5.6. *Expectation of a random variable need not be finite. For example, we can consider a random variable X with $P(X = x) = 6/(\pi^2 n^2)$ if $x \in \mathbb{N}^+$ and 0 otherwise. The expectation of X in this case is ∞.*

We have

$$E[X] = \sum_{x \in \mathbb{N}^+} x p_X(x) = \sum_{x \in \mathbb{N}^+} \frac{6}{\pi^2 x} = \infty$$

Example 5.7. *A Cauchy$(0,1)$ random variable X has the following density:*

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}. \quad (5.9)$$

X is not integrable. We have

$$E(|X|) = \int_{\mathbb{R}} |x| f(x) = \int_{\mathbb{R}} \frac{|x|}{\pi(1+x^2)} = \frac{2}{\pi} \int_0^\infty \frac{x}{(1+x^2)} dx = \frac{2}{\pi} \left[\frac{\log(1+x^2)}{2} \right]_0^\infty = \infty$$

Proposition 5.8. *If X and Y are random variables, the following properties hold.*

1. *If $X \geq 0$ a.s. $\iff P(X \geq 0) = 1$, then $E[X] \geq 0$*
2. *For all $a, b \in \mathbb{R}$, $E[aX + b] = aE[X] + b$*
3. *$E[X + Y] = E[X] + E[Y]$*

Proof. NPOS □

Lemma 5.9. 1. *If $X \leq Y$ a.s., that is $P(X \leq Y) = 1$, then $E[X] \leq E[Y]$*

2. *If $X = Y$ a.s., that is $P(X = Y) = 1$, then $E[X] = E[Y]$*
3. *$|E[X]| \leq E[|X|]$*

Proof. 1. $X \leq Y$ a.s. $\implies Y - X \geq 0$ a.s. $\implies E[Y - X] \geq 0 \implies E[Y] \geq E[X]$;

2. $X = Y$ a.s. $\implies X \leq Y$ a.s. and $Y \leq X$ a.s. so $E[X] \leq E[Y]$ and $E[Y] \leq E[X]$. Thus, $E[X] = E[Y]$;
3. Trivially, $X \leq |X|$ and $-X \leq |X|$. So $E[X] \leq E[|X|]$ and $-E[X] = E[-X] \leq E[|X|]$. This shows $|E[X]| \leq E[|X|]$.

□

Definition 5.10. *Let X be a R.V., the variance of X is defined as*

$$Var(X) = E[(X - E[X])^2]$$

We also have the standard deviation of X is $\sqrt{Var(X)}$

Lemma 5.11. *If X, Y are R.V.s with $Var(X), Var(Y)$, we have*

1. *$Var(X) \geq 0$*
2. *$\forall c \in \mathbb{R}, Var(cX) = c^2 Var(X)$*
3. *If X and Y are independent, $Var(X + Y) = Var(X) + Var(Y)$*

Theorem 5.12. *Let X be a R.V. with induced probability measure $\mu_X : \mathcal{B}(\mathbb{R}) \to [0,1]$. Then for any measurable $g : \mathbb{R} \to \mathbb{R}$*

$$E[g(X)] = \int_{\mathbb{R}} g(X) d\mu_X$$

provided that $\int_{\mathbb{R}} |g(X)| d\mu_X < \infty$.

Moreover, if X is discrete with p.m.f. p_X and support $\mathcal{X}$, we have

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

If X is continuous with p.d.f. f_X then

$$E[g(X)] = \int_{\mathbb{R}} g(x) f_X(x) dx$$

Proof. NPOS □

Remark 5.13. *Note that in general, $E[g(X)] \neq g(E[X])$. Expectation is a linear operation since, in essence, expectation is integration. Thus, expectation is not closed under any non-linear operation g.*

Equality only satisfies when g is linear, or X is constant.

Example 5.14. *Let X be a random variable, and $g(x) = (x - E[X])^2$ is a measurable function, then $E[g(X)] = Var(X)$ provided that $Var(X) < \infty$*

Indeed, if X is discrete with p.m.f. p_X

$$E[g(X)] = \sum_{x_i \in \mathcal{X}} g(x_i) p_X(x_i) = \sum_{x_i \in \mathcal{X}} (x_i - E[X])^2 p_X(x_i) = E[(X - E[X])^2] = Var(X)$$

The proof is similar for a continuous random variable X.

Theorem 5.15. *If X, Y are independent with $E[|X|], E[|Y|] < \infty$. Then*

$$E[XY] = E[X]E[Y]$$

Proof. NPOS. □

Definition 5.16. *Given X, Y as R.V.s, we define the covariance of random variables X and Y is*

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

The correlation between random variables X and Y is

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

If $Cov(X, Y) = Corr(X, Y) = 0$, then X, Y are said to be uncorrelated.

Lemma 5.17. *1.* $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$*;*

2. For any $a, b \in \mathbb{R}$, $\mathrm{Cov}(aX + bY, Z) = a\,\mathrm{Cov}(X, Z) + b\,\mathrm{Cov}(Y, Z)$.

Proof. 1. We get

$$\begin{aligned} Cov(X, Y) &= E[(X - E[X])(Y - E[Y]] \\ &= E[XY - E[X]Y - XE[Y] + E[X]E[Y]] \\ &= E[XY] - E[E[X]Y] - E[XE[Y]] + E[E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[E[Y]] \\ &= E[XY] - 2E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

2. We get

$$\begin{aligned}
Cov(aX+bY,Z) &= E[(aX+bY-E[aX+bY])(Z-E[Z])] \\
&= E[aXZ - aXE[Z] + bYZ - bYE[Z] - E[aX+bY]Z + E[aX+bY]E[Z]] \\
&= aE[XZ] - aE[Z]E[X] + bE[YZ] - bE[Z]E[Y] - E[aX+bY]E[Z] + E[aX+bY]E[E[Z]] \\
&= a(E[XZ] - E[Z]E[X]) + b(E[YZ] - E[Z]E[Y]) \\
&= aCov(X,Z) + bCov(Y,Z)
\end{aligned}$$

□

Remark 5.18. *Independence of two random variables implies no correlation. However, no correlation does NOT imply independence.*

5.2 Conditional Expectation

Definition 5.19. *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*

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

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

$$E[X \mid Y = y] = \int x f_{X|Y}(x \mid y) dx$$

if $X \mid Y = y$ is continuous. The above assumes that the sum/integral is defined.

Remark 5.20. *In general, we may not be only interested in the conditional expectation $E[X \mid Y = y]$ 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 $E[X \mid Y]$. 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 $E[X \mid Y]$ without conditioning on the specific value that Y takes? To this end, notice the definition of conditional expectation in Definition 5.19, we can see that if we change the value of y, the resulting conditional expectation will also change. Specifically, we can write

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

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

Lemma 5.21. *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*

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

if $X \mid Y$ is discrete, and

$$E[X \mid Y] = \int x f_{X|Y}(x \mid Y) dx$$

if $X \mid Y$ is continuous.

Proof. If $X|Y$ is discrete, for any $\omega \in \Omega$ since Y is a random variable $Y(\omega) = y \in \mathbb{R}$ we have

$$E[X|Y](\omega) = E[X|Y = Y(\omega)] = E[X|Y = y] = \sum_i x_i p_{X|Y}(x_i|y) \in \mathbb{R}$$

This gives

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

is a random variable.

If $X|Y$ is continuous, for any $\omega \in \Omega$ since Y is a random variable $Y(\omega) = y \in \mathbb{R}$ we have

$$E[X|Y](\omega) = E[X|Y = Y(\omega)] = E[X|Y = y] = \int x f_{X|Y}(x|y)dx \in \mathbb{R}$$

This gives

$$E[X|Y] = \int x f_{X|Y}(x|Y)dx$$

is a random variable. □

Theorem 5.22 (Tower property / Law of total expectation)**.** *Given random variables X and Y with $E[X] < \infty$, then*

$$E[X] = E[E[X \mid Y]]$$

Proof. We have

$$\begin{aligned}
E[E[X|Y]] &= \int E[X|Y = y] f_Y(y)dy \\
&= \int \int x f_{X|Y}(x|y) f_Y(y)dxdy \\
&= \int \int x \frac{f(x,y)}{f_Y(y)} f_Y(y)dxdy \\
&= \int \int x f(x,y)dxdy \\
&= \int \int x f(x,y)dydx \qquad \text{[Fubini's Theorem always work for joint p.d.f.]} \\
&= \int x \int f(x,y)dydx \\
&= \int x f_X(x)dx \\
&= E(X)
\end{aligned}$$

Similar can be proved if joint distribution is discrete. □

Theorem 5.23 (Law of total variance)**.** *Given random variables X and Y with $Var(X) < \infty$, then*

$$Var(X) = E[Var(X \mid Y)] + Var(E[X \mid Y])$$

where

$$Var(X \mid Y) = E[(X - E[X \mid Y])^2 \mid Y]$$

Proof.

$$\begin{aligned}
Var(X) &= E[(X - E[X])^2] \\
&= E[((X - E[X|Y]) + (E[X|Y] + E[X]))^2] \\
&= E[(X - E[X|Y])^2 + 2(X - E[X|Y])(X + E[X|Y]) + (E[X|Y] - E[X])^2] \\
&= E[(X - E[X|Y])^2] + 2E[(X - E[X|Y])(E[X|Y] - E[X])] + E[(E[X|Y] - E[X])^2] \\
&= E[E[(X - E[X|Y])^2|Y]] + 2E[(X - E[X|Y])(E[X|Y] - E[X])] + E[(E[X|Y] - E[X])^2] \qquad \text{[Tower Property]} \\
&= E[Var(X|Y)] + 2E[(X - E[X|Y])(E[X|Y] - E[X])] + E[(E[X|Y] - E[X])^2] \\
&= E[Var(X|Y)] + 2E[(X - E[X|Y])(E[X|Y] - E[X])] + E[(E[X|Y] - E[E[X|Y]])^2] \qquad \text{[Tower Property]} \\
&= E[Var(X|Y)] + 2E[(X - E[X|Y])(E[X|Y] - E[X])] + Var(E[X|Y]) \\
&= E[Var(X|Y)] + 2E[(X - E[X|Y])(E[X|Y] - E[X])] + Var(E[X|Y]) \\
&= E[Var(X|Y)] + 2E[E[(X - E[X|Y])(E[X|Y] - E[X])|Y]] + Var(E[X|Y]) \qquad \text{[Tower Property]} \\
&= E[Var(X|Y)] + 2E[(E[X|Y] - E[X])E[(X - E[X|Y])|Y]] + Var(E[X|Y]) \qquad \text{[}(E[X|Y] - E[X]) \text{ is constant]} \\
&= E[Var(X|Y)] + 2E[(E[X|Y] - E[X])(E[X|Y] - E[X|Y])] + Var(E[X|Y]) \\
&= E[Var(X|Y)] + Var(E[X|Y])
\end{aligned}$$

We first prove

$$Var(X|Y) = E[(X - E[X|Y])^2|Y] = E[X^2|Y] - (E[X|Y])^2$$

Indeed, we have

$$E[(X - E[X|Y])^2|Y] = \int x - \int x f_{X|Y}(x|y)dx$$

We have

$$\begin{aligned}
E[Var(X|Y)] + Var(E[X|Y]) &= E[E[X^2|Y] - (E[X|Y])^2] + E[(E[X|Y] - E[E[X|Y]])^2] \\
&= E[X^2] - E[(E[X|Y])^2] + E[E[X|Y]^2 - 2E[X|Y]E[E[X|Y]] + E[E[X|Y]]^2] \\
&= E[X^2] - E[(E[X|Y])^2] + E[E[X|Y]^2] - 2E[X]^2 + E[X]^2 \\
&= E[X^2] - E[X]^2 \\
&= Var(X)
\end{aligned}$$

□

Proposition 5.24 (Convolution Formula)**.** *If X and Y are independent with densities f_X and f_Y, for all $z \in \mathbb{R}$,*

$$P(X + Y \leq z) = \int F_X(z - y) f_Y(y)dy = \int F_Y(z - x) f_X(x)dx$$

Proof.

$$\begin{aligned}
P(X+Y\le z) &= P(X\le z-Y)\\
&= E[\mathbb{1}_{X\le z-Y}]\\
&= E[E[\mathbb{1}_{X\le z-Y}\mid Y]] \qquad \text{[Law of total expectation]}\\
&= \int E[\mathbb{1}_{X\le z-y}\mid Y=y]f_Y(y)dy\\
&= \int E[\mathbb{1}_{X\le z-y}]f_Y(y)dy \qquad \text{[By Independence]}\\
&= \int P(X\le z-y)f_Y(y)dy\\
&= \int F_X(z-y)f_Y(y)dy
\end{aligned}$$

Note that if p.d.f. is differentaible, we can differentiate, and we get

$$\frac{\partial}{\partial z}P(X+Y\le z) = F_X(z-y)f_Y(y)$$

□

5.3 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 $E[X_n] \to E[X]$? 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 5.25. *Let $U \sim \text{Unif}(0,1)$ and $X_1, X_2, \ldots$ be a sequence of random variables such that $X_n(\omega) = n\mathbb{1}_{U(\omega)\in(0,1/n)}$. We have $X_n \to X = 0$ a.s. But*

$$\lim_{n\to\infty} E[X_n] = 1 \ne E[X] = 0.$$

Definition 5.26. *A random variable X is called bounded if there exists $M < \infty$ such that $|X| \le M$ a.s.*

Definition 5.27. *Suppose $X \ge 0$ a.s., then*

$$\sup\{E[Y] : Y \text{ bounded},\ 0 \le Y \le X \text{ a.s.}\} = E[X].$$

Lemma 5.28. *Let $X \ge 0$ a.s., and recall the notation $a \wedge b = \min\{a, b\}$. Then,*

$$\lim_{n\to\infty} E(X \wedge n) = E[X].$$

Proof. Define $X_n = X \wedge n$. Clearly $X_n \le X$, so $E[X_n] \le E[X]$. Also, $E[X_n] \le E[X_{n+1}]$ for all n, so the limit exists, and thus $\lim_{n\to\infty} E[X_n] \le E[X]$.

Consider a bounded Y such that $0 \le Y \le X$ a.s.. Then for large n, $X_n \ge Y$ a.s., so $E[X_n] \ge E[Y]$. Hence,

$$\lim_{n\to\infty} E[X_n] \ge \sup\{E[Y] : Y \text{ bounded},\ 0 \le Y \le X \text{ a.s.}\} = E[X].$$

Therefore, $\lim_{n\to\infty} E[X_n] = E[X]$. □

Theorem 5.29 (Bounded Convergence Theorem)**.** *Suppose that $|X_n| \leq M$ a.s. and $X_n \xrightarrow{P} X$. Then, $E[X] = \lim_{n\to\infty} E[X_n]$.*

Note: *This is not equivalent to*

$$E[\lim_{n\to\infty} X_n] = \lim_{n\to\infty} E[X_n]$$

If $X_n \to X$ a.s., we can use the latter one.

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

$$\begin{aligned}
|E[X_n] - E[X]| &= |E[X_n - X]| \\
&\leq E[|X_n - X|] \\
&= E[|X_n - X|\mathbb{1}_{G_n}] + E[|X_n - X|\mathbb{1}_{G_n^c}] \\
&\leq 2MP(G_n) + \epsilon[1 - P(G_n)] \\
&= \epsilon + P(G_n)(2M - \epsilon) \\
&\to \epsilon \qquad [\text{As } P(G_n) \to 0]
\end{aligned}$$

By convergence in probability and arbitrary ϵ, $\lim_{n\to\infty} |E[X_n] - E[X]| = 0$. □

Theorem 5.30 (Fatou's Lemma)**.** *If $X_n \geq 0$ a.s. for all n, then $\liminf_{n\to\infty} E[X_n] \geq E[\liminf_{n\to\infty} X_n]$.*

Proof. For each n, define $Y_n = \inf_{m\geq n} X_m$. Clearly, $X_n \geq Y_n$ a.s., so $E[X_n] \geq E[Y_n]$ for all n, and

$$\liminf_{n\to\infty} E[X_n] \geq \liminf_{n\to\infty} E[Y_n].$$

Moreover, by definition of Y_n, we have $Y_n \uparrow Y = \liminf_{n\to\infty} X_n$ a.s. Since $Y_n \leq Y_{n+1}$ a.s., we have $E[Y_n] \leq E[Y_{n+1}]$. Thus, $\lim_{n\to\infty} E[Y_n]$ exists and equals $\liminf_{n\to\infty} E[Y_n]$.

Fix an arbitrary M, and observe that $Y_n \wedge M \uparrow Y \wedge M$ a.s. So by the bounded convergence theorem,

$$\liminf_{n\to\infty} E[Y_n] = \lim_{n\to\infty} E[Y_n] \geq \lim_{n\to\infty} E[Y_n \wedge M] = E[Y \wedge M].$$

Taking the limit as $M \to \infty$ we get

$$\begin{aligned}
\liminf_{n\to\infty} E[Y_n] &\geq \lim_{M\to\infty} \lim_{n\to\infty} E[Y_n \wedge M] \\
&= \lim_{M\to\infty} E[Y \wedge M] \\
&= E[Y] \qquad [Y \geq 0 \text{ a.s.}] \\
&= E[\liminf_{n\to\infty} X_n]
\end{aligned}$$

□

Theorem 5.31 (Monotone Convergence Theorem)**.** *If $X_n \geq 0$ a.s. and $X_n \uparrow X$ a.s., then $E[X_n] \uparrow E[X]$.*

Proof. Since $E[X_n] \leq E[X]$, we have $\lim_{n\to\infty} E[X_n] \leq E[X]$. But by Fatou's Lemma,

$$\lim_{n\to\infty} E[X_n] = \liminf_{n\to\infty} E[X_n] \geq E\left[\liminf_{n\to\infty} X_n\right] = E[X].$$

□

Example 5.32. *If $X_1, X_2, \ldots$ are a sequence of non-negative random variables, then*

$$E\left[\sum_{n=1}^{\infty} X_n\right] = \sum_{n=1}^{\infty} E[X_n].$$

Proof. Let $Y_m = \sum_{n=1}^{m} X_n$. Since X_n are non-negative, $Y_m \leq Y_{m+1}$ a.s. By MCT,

$$E\left[\sum_{n=1}^{\infty} X_n\right] = E\left[\lim_{m\to\infty} Y_m\right] = \lim_{m\to\infty} E[Y_m] = \lim_{m\to\infty} \sum_{n=1}^{m} E[X_n] = \sum_{n=1}^{\infty} E[X_n].$$

□

Theorem 5.33 (Dominated Convergence Theorem)**.** *If $X_n \to X$ a.s. and $|X_n| \leq Y$ a.s. for some integrable Y, then*

$$E[X] = \lim_{n\to\infty} E[X_n].$$

Proof. Since $|X_n| \leq Y$ a.s. for all n, we have $|X| \leq Y$ a.s., and thus X is integrable. Since $X_n + Y \geq 0$ a.s. for all n, by Fatou's Lemma,

$$\liminf_{n\to\infty} E[X_n + Y] \geq E\left[\liminf_{n\to\infty}(X_n + Y)\right] = E[X + Y].$$

Since all expectations here are finite, it follows that

$$\liminf_{n\to\infty} E[X_n] \geq E[X].$$

It also holds that $Y - X_n \geq 0$, so

$$\liminf_{n\to\infty} E[Y - X_n] \geq E\left[\liminf_{n\to\infty}(Y - X_n)\right] = E[Y - X].$$

Therefore,

$$\liminf_{n\to\infty} E[-X_n] \geq -E[X] \iff \limsup_{n\to\infty} E[X_n] \leq E[X].$$

Thus, we have

$$\limsup_{n\to\infty} E[X_n] \leq E[X] \leq \liminf_{n\to\infty} E[X_n],$$

which implies

$$E[X] = \lim_{n\to\infty} E[X_n].$$

□

5.4 Moment Generating Functions

Definition 5.34. *The moment-generating function (M.G.F.) of a random variable X is defined by*

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

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

Theorem 5.35 (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*

$$M_X^{(n)}(0) = E[X^n], \quad \textit{and} \quad E[|X^n|] < \infty \textit{ for all } n.$$

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

$$e^{|\lambda z|} \leq e^{\lambda z} + e^{-\lambda z},$$

we have

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

showing $e^{|\lambda X|}$ is integrable.

By Taylor expansion, we have

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

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

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

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

$$|E[X^n]| \leq E[|X^n|] \leq E[|X|^n],$$

the series $\sum_{n=0}^{\infty} \frac{\lambda^n E[X^n]}{n!}$ 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

$$|M_X(\lambda) - E[S'_k]| = \left|\sum_{n=k+1}^{\infty} \frac{\lambda^n E[X^n]}{n!}\right| \leq \sum_{n=k+1}^{\infty} \frac{\lambda^n E[|X|^n]}{n!} \to 0 \quad \text{as } k \to \infty.$$

Thus,

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

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

$$M_X^{(m)}(0) = E[X^m].$$

□

Theorem 5.36. *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. NPOS □

Lemma 5.37. *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*

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

Proof. We have

$$\begin{aligned}
M_S(\lambda) &= E[e^{\lambda S}] \\
&= E\left[\prod_{i=1}^{n} e^{\lambda X_i}\right] \\
&= \prod_{i=1}^{n} E[e^{\lambda X_i}] \qquad [e^{\lambda X_1}, \ldots, e^{\lambda X_n} \text{ are independent, prove see Question 5.38}] \\
&= \prod_{i=1}^{n} M_{X_i}(\lambda) \\
&< \infty
\end{aligned}$$

□

5.5 Exercises

Question 5.38. *Give an example of Gaussian random variables X and Y such that they are uncorrelated, i.e.,* $\mathrm{Cov}(X,Y)=0$*, but not independent.*

Question 5.39. *Compute the expected value and variance for all random variables listed in Chapter 2.*

Question 5.40. *Show that the sum of i.i.d. Poisson random variables is a Poisson random variable.*

Question 5.41. *Show that the sum of i.i.d. Exponential random variables is a Gamma random variable.*

Question 5.42. *Question 5.42 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 $\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2)$.*

Question 5.43. *Let $X_1,\dots,X_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Let $\chi^2=\sum_{i=1}^n X_i^2$. Show that $\chi^2 \sim \mathrm{Gamma}(n/2,1/2)$. χ^2 is also called the Chi-squared distribution with n degrees of freedom.*

Question 5.44. *Show that if $X\geq 0$ a.s. and $E[X]=0$, then $X=0$ a.s.*

Question 5.45. *Show that $X=a$ a.s. for some $a\in\mathbb{R}$ if and only if* $\mathrm{Var}(X)=0$.

Question 5.46 (You can't always under-perform)**.** *If X is a random variable with $E[X]<\infty$, then $P(X\geq E[X])>0$.*

Question 5.47. *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?*

Question 5.48. *Let $X_1,X_2,\dots$ be i.i.d. with $E[X_i]=\mu$ and* $\mathrm{Var}(X_i)=\sigma^2$*, and let N be an integer-valued random variable with $E[N]=m$ and* $\mathrm{Var}(N)=v$ *and independent from all X_i. Show that*

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

Question 5.49. *Show that if X only takes values in $\mathbb{N}$, $E[X]=\sum_{k=1}^\infty P(X\geq k)$.*
Hint: use MCT

Question 5.50. *Show that if $X\geq 0$ and $p>0$, $E[X^p]=\int_0^\infty px^{p-1}P(X\geq x)\,dx$.*

Question 5.51. *If X is an integrable random variable such that $X\geq 0$ a.s., show that*

$$\lim_{t\to\infty} tP(X>t)=0.$$

Question 5.52 (Optional)**.** *Let $X\geq 0$ a.s. so that $M_X(\lambda)<\infty$ and $E[Xe^{\lambda X}]<\infty$ for all $\lambda\in\mathbb{R}$.*

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

6 Properties of Expectation

6.1 Concentration of Measure

Theorem 6.1 (Markov's Inequality)**.** *If $X\geq 0$ a.s., then for all $a>0$,*

$$P(X\geq a)\leq \frac{E[X]}{a}$$

Proof. Define $Z(\omega) = a\mathbb{1}\{X(\omega) \geq a\}$. Then, $Z \leq X$ a.s., and

$$E[Z] = aP(X \geq a) \leq E[X]$$

□

Corollary 6.2 (Chebyshev's Inequality)**.** *For all $t \geq 0$,*

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

Proof. Take $Y = (X - E[X])^2$, by Markov

$$P(|X - E[X]| \geq t) = P(Y \geq t^2) \leq \frac{E[Y]}{t^2} = \frac{Var(Y)}{t^2}$$

□

Definition 6.3. *A function $f : \mathbb{R} \to \mathbb{R}$ is convex if for all $\lambda \in (0, 1)$ and $x, y \in \mathbb{R}$,*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

If the inequality is reversed, then f is called concave.

Lemma 6.4 (NPOS)**.** *If $f : \mathbb{R} \to \mathbb{R}$ is convex, then for all $a \in \mathbb{R}$, there exists $m \in \mathbb{R}$ such that for all $x \in \mathbb{R}$,*

$$f(x) \geq f(a) + m(x - a).$$

Theorem 6.5 (Jensen's Inequality)**.** *If f is a convex function and X is a random variable such that X and $f(X)$ are integrable,*

$$f(E[X]) \leq E[f(X)].$$

The inequality is flipped if f is concave.

In addition, $f(E[X]) = E[f(X)]$ if and only if $X = E[X]$ a.s. or $f(x) = mx + b$ for some $m, b \in \mathbb{R}$.

Proof. Since f is convex, by Lemma 6.4, set $a = E[X]$, then for all $x \in \mathbb{R}$, there exists $m \in \mathbb{R}$ such that

$$f(x) \geq f(E[X]) + m(x - E[X]).$$

This means that

$$f(X) \geq f(E[X]) + m(X - E[X]) \quad \text{a.s.}$$

Hence, taking the expectation, we get

$$E[f(X)] \geq f(E[X]) + m(E[X] - E[X]) = f(E[X]).$$

For the equality claim, suppose $f(E[X]) = E[f(X)]$. Since

$$f(X) \geq f(E[X]) + m(X - E[X]) \quad \text{a.s.},$$

then

$$Y = f(X) - f(E[X]) - m(X - E[X]) \geq 0 \quad \text{a.s.}$$

However, since $f(E[X]) = E[f(X)]$, it means $E[Y] = 0$. By Question 5.44, we thus have $Y = 0$ a.s. Therefore,

$$f(X) - f(E[X]) - m(X - E[X]) = 0 \quad \text{a.s.}$$

Certainly, the above holds if $X = E[X]$ a.s. If $X \neq E[X]$, then

$$\frac{f(X) - f(E[X])}{X - E[X]} = m \quad \text{a.s.}$$

The above can only be true if $f(x) = mx + b$ for some $b \in \mathbb{R}$. The other direction is trivial. □

Example 6.6. *1.* $(E[X])^2 \leq E[X^2]$

2. $E[\log X] \leq \log(E[X])$

Lemma 6.7. *If* $0 < p < q$*, then*

$$(E[|X|^p])^{1/p} \leq (E[|X|^q])^{1/q}$$

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

$$(E[|X|^p])^{\frac{q}{p}} \leq E\left[(|X|^p)^{\frac{q}{p}}\right] = E[|X|^q] \implies (E[|X|^p])^{1/p} \leq (E[|X|^q])^{1/q}$$

□

Corollary 6.8 (Chernoff's Inequality)**.** *For all* $t \geq 0$*,*

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

Proof. For $\lambda > 0$,

$$\begin{aligned} P(X \geq E[X] + t) &= P(X - E[X] \geq t) \\ &= P(e^{\lambda(X-E[X])} \geq e^{\lambda t}) \\ &\leq \frac{E[e^{\lambda(X-E[X])}]}{e^{\lambda t}} \\ &\leq M_{X-E[X]}(\lambda) e^{-\lambda t}. \quad \text{(Markov)} \end{aligned}$$

Since the above is true for all $\lambda > 0$, we have

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

□

Proposition 6.9 (Hölder's Inequality)**.** *If* $p, q > 1$ *such that* $\frac{1}{p} + \frac{1}{q} = 1$*, then*

$$E[|XY|] \leq (E[|X|^p])^{1/p} (E[|Y|^q])^{1/q}.$$

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

Otherwise, let

$$X^* = \frac{|X|}{(E[|X|^p])^{1/p}}, \quad Y^* = \frac{|Y|}{(E[|Y|^q])^{1/q}}.$$

Observe that $E[(X^*)^p] = E[(Y^*)^q] = 1$.

We now show that $\frac{1}{p}x^p + \frac{1}{q}y^q \geq xy$ for all $x, y \geq 0$. To see this, for all $y > 0$, let

$$h_y(x) = \frac{1}{p}x^p + \frac{1}{q}y^q - xy.$$

Then,

$$h_y'(x) = x^{p-1} - y, \quad h_y''(x) = (p-1)x^{p-2} \geq 0,$$

so the minimizer is at $x^* = y^{1/(p-1)}$. Plugging in, we get

$$\begin{aligned} h_y(x^*) &= \frac{1}{p}y^{p/(p-1)} + \frac{1}{q}y^q - y \cdot y^{1/(p-1)} \\ &= y^q \left(\frac{1}{p} + \frac{1}{q}\right) - y^q = 0. \end{aligned}$$

Thus,

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

so

$$E[|X^*Y^*|] = \frac{E[|XY|]}{(E[|X|^p])^{1/p}(E[|Y|^q])^{1/q}} \leq \frac{1}{p} + \frac{1}{q} = 1.$$

□

Corollary 6.10 (Cauchy-Schwarz Inequality)**.**

$$E[|XY|] \leq \sqrt{E[X^2]E[Y^2]}.$$

Proof. Apply Hölder's Inequality with $p = q = 2$. □

Example 6.11. *Let $X \sim \text{Exp}(1)$, then*

$$E[X^{1/4}] \leq 1.$$

Proof.

$$\begin{aligned} E[X^{1/4}] &= E[X^{1/4} \cdot 1] \\ &\leq \left(E[X^{p/4}]\right)^{1/p} (E[1^q])^{1/q}. \end{aligned}$$

Let $p = 4$, $q = \frac{4}{3}$, then

$$E[X^{1/4}] \leq (E[X])^{1/4} \cdot 1 = 1.$$

We can check by direct computation that $E[X^{1/4}] \approx 0.906$, which is very close to the bound of one. □

6.2 Exercises

Question 6.12. *Using a Taylor expansion, show that for a Rademacher random variable S (e.g., taking values 1 and -1 with probability $1/2$ each)*

$$E[e^{\lambda S}] \leq e^{\lambda^2/2}, \quad \forall \lambda \in \mathbb{R}.$$

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

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

Question 6.13. *Suppose $E[X_n] = 0$ and $E[X_n^2] = 1$ for all n. Prove that $P(X_n \geq n \text{ i.o.}) = 0$.*

Question 6.14. *Find a random variable X and $a > 0$ such that $P(X > a) \geq \frac{E[X]}{a}$, and identify what breaks in the proof of Markov's inequality for this example.*

Question 6.15. *Show that the functions $f(x) = \frac{x}{1+x}$ and $f(x) = \log(x)$ are concave for $x > 0$.*

Question 6.16. *For X such that $E[X] < \infty$ and $a \in \mathbb{R}$, prove that $E[\max\{X, a\}] \geq \max\{E[X], a\}$.*

Question 6.17. *For any X, Y, use Cauchy-Schwarz to show that $|\operatorname{Corr}(X, Y)| \leq 1$.*

Question 6.18. *Let $p \geq 0$, show that for random variables X, Y,*

$$E[|X + Y|^p] \leq 2^p(E[|X|^p] + E[|Y|^p]).$$

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.

Question 6.19. *Prove that if X is such that $E[X] = \mu < \infty$ and $\operatorname{Var}(X) = \sigma^2 < \infty$, for all $a > 0$*

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

Question 6.20. *Let $X_1, X_2, \ldots$ satisfy $E[X_n] = m < \infty$ and $\operatorname{Var}(X_n) = 1/\sqrt{n}$. Prove that $X_n \xrightarrow{P} m$.*

7 The Law of Large Numbers

7.1 Basic LLN

Theorem 7.1 (Basic Weak LLN)**.** *Let $X_1, X_2, \ldots$ be independent, and for all i, suppose $E[X_i] = \mu$ and $\mathrm{Var}(X_i) \le \sigma^2 < \infty$. Define $S_n = \sum_{i=1}^n X_i$. Then,*

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

Proof. By linearity, $E\left[\frac{S_n}{n}\right] = \mu$ and $Var\left(\frac{S_n}{n}\right) \le \frac{\sigma^2}{n}$. Then, for any $\epsilon > 0$, Chebyshev's inequality gives that

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \le \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0.$$

□

Theorem 7.2 (Basic Strong LLN)**.** *Let $X_1, X_2, \ldots$ be independent, and for all i, suppose $E[X_i] = \mu$ and $E[(X_i - \mu)^4] \le a < \infty$. Then,*

$$\frac{1}{n} S_n \xrightarrow{a.s.} \mu.$$

Proof. First, observe that

$$\begin{aligned} E[(X_i-\mu)^2] &= E[(X_i-\mu)^2 \mathbb{1}_{(X_i-\mu)^2>1}] + E[(X_i-\mu)^2 \mathbb{1}_{(X_i-\mu)^2\le 1}] \\ &\le E[(X_i-\mu)^4] + 1 \\ &\le a + 1 \end{aligned}$$

WLOG, suppose $\mu = 0$. Then we have

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

Next, for any $\epsilon > 0$, apply Markov's inequality:

$$P\left(\left|\frac{1}{n} S_n\right| > \epsilon\right) = P(S_n^4 > n^4\epsilon^4) \le \frac{E[S_n^4]}{n^4\epsilon^4} \le \frac{K}{n^2\epsilon^4}.$$

Since $\sum_{n=1}^\infty \frac{K}{n^2\varepsilon^4} < \infty$, by Borel–Cantelli, $P\left(\left|\frac{1}{n}S_n\right| > \epsilon \text{ i.o.}\right) = 0 \implies P\left(\lim_{n\to\infty}\left|\frac{1}{n}S_n\right| = 0\right) = 1$. □

7.2 Advanced Weak LLN

Definition 7.3. *A sequence of random variables $(X_i)_{i\in\mathcal{I}}$ with $E[X_i^2] < \infty$ are uncorrelated if for all $i \neq j$,*

$$E[X_iX_j] = E[X_i]E[X_j].$$

Theorem 7.4 (Uncorrelated Weak LLN)**.** *If $X_1, X_2, \ldots$ are uncorrelated with $E[X_i] = \mu$ and $Var(X_i) \leq \sigma^2 < \infty$, then*

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

Proof.

$$\begin{aligned} E\left[\left(\frac{1}{n}S_n - \mu\right)^2\right] &= E\left[\left(\frac{1}{n}S_n - E\left[\frac{1}{n}S_n\right]\right)^2\right] \\ &= Var\left(\frac{1}{n}S_n\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &\leq \frac{n\sigma^2}{n^2} \to 0. \end{aligned}$$

□

Theorem 7.5 (Advanced Weak LLN)**.** *If $X_1, X_2, \ldots$ are i.i.d. with $\lim_{x\to\infty} xP(|X_1| > x) = 0$. Define*

$$\mu_n = E[X_1\mathbb{1}_{\{|X_1|\leq n\}}]$$

then

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

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

$$\begin{aligned} P\left(\left|\frac{S_n}{n} - \mu_n\right| > \epsilon\right) &= P\left(\left|\frac{S_n}{n} - \frac{\overline{S_n}}{n} + \frac{\overline{S_n}}{n} - \mu_n\right| > \epsilon\right) \\ &\leq P\left(\left|\frac{S_n}{n} - \frac{\overline{S_n}}{n}\right| > \frac{\epsilon}{2}\right) + P\left(\left|\frac{\overline{S_n}}{n} - \mu_n\right| > \frac{\epsilon}{2}\right) \\ &\leq P(S_n \neq \overline{S}_n) + P\left(\left|\frac{\overline{S}_n}{n} - \mu_n\right| > \frac{\epsilon}{2}\right). \end{aligned}$$

For the first term,

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

For the second term, first observe that

$$E[\overline{S}_n] = E\left[\sum_{k=1}^{n} \overline{X}_k^{(n)}\right] = \sum_{k=1}^{n} E[X_k \mathbb{1}_{\{|X_k|\leq n\}}] = n\mu_n.$$

So, by Chebyshev's inequality,

$$\begin{aligned} P\left(\left|\frac{\overline{S}_n}{n} - \mu_n\right| > \frac{\epsilon}{2}\right) &\leq \frac{4}{n^2\epsilon^2} E\left[(\overline{S}_n - n\mu_n)^2\right] \\ &= \frac{4}{n^2\epsilon^2} Var(\overline{S}_n) \\ &= \frac{4}{n^2\epsilon^2} \sum_{k=1}^{n} Var(\overline{X}_k^{(n)}) \\ &= \frac{4}{n\epsilon^2} Var(\overline{X}_1^{(n)}) \\ &\leq \frac{4}{n\epsilon^2} E\left[\left(X_1 \mathbb{1}_{\{|X_1|\leq n\}}\right)^2\right] \end{aligned}$$

Finally, recalling that $E[X^p] = \int_0^\infty p x^{p-1} P(X \geq x) dx$, we compute:

$$\begin{aligned} E\left[\left(X_1 \mathbb{1}_{\{|X_1|\leq n\}}\right)^2\right] &= \int_0^\infty 2x P(|\overline{X}_k^{(n)}| \geq x) dx \\ &= \int_0^n 2x P(|X_k| \geq x) dx \\ &= 2\int_0^n x P(|X_1| \geq x) dx \end{aligned}$$

Since $0 \leq xP(|X_1| \geq x) \leq x$ for all x and $\sup_x xP(|X_1| \geq x) < \infty$. Thus, by BCT

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

□

Corollary 7.6. *If $X_1, X_2, \ldots$ are i.i.d. with $E[|X_1|] < \infty$ and $E[X_1] = \mu < \infty$, then*

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

Proof. Let $Y_n = |X_1| \mathbb{1}_{\{|X_1|>n\}}$. For each $\omega \in \Omega$, $|X_1(\omega)| < \infty$, so $Y_n \to 0$ a.s. Since $|Y_n| \leq |X_1|$ which is integrable, by the DCT we have $E[Y_n] \to 0$. Further, observe that $Y_n \geq n\mathbb{1}_{\{|X_1|>n\}}$, so

$$E[Y_n] \geq nP(|X_1| > n).$$

Thus, $\lim_{x\to\infty} xP(|X_1| > x) = 0$.

Next, consider $Z_n = X_1 \mathbb{1}_{\{|X_1|\leq n\}}$. Again, since $|X_1(\omega)| < \infty$, $Z_n \to X_1$ a.s. Thus, we can again apply DCT to obtain $\mu_n = E[Z_n] \to E[X_1] = \mu$. The result then follows by the Weak LLN. □

7.3 Advanced Strong LLN

Theorem 7.7 (Strong LLN)**.** *If $X_1, X_2, \ldots$ are i.i.d. with $E[|X_1|] < \infty$ and $E[X_1] = \mu < \infty$, then*

$$\frac{1}{n} S_n \xrightarrow{a.s.} \mu.$$

Proof. NPOS □

7.4 Applications

Theorem 7.8 (Glivenko-Cantelli)**.** *Let $X_1, X_2, \ldots$ be i.i.d. with CDF F and define the empirical distribution function*

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

Then,

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0 \quad a.s.$$

Proof. NPOS □

Theorem 7.9 (Basic Monte-Carlo Integration on an Interval)**.** *Let $g : [0,1] \to \mathbb{R}$ be an integrable function. Let $U_i \overset{i.i.d.}{\sim} Unif(0,1)$ for $i \in [n]$. Then*

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

Proof. Notice that

$$E[g(U)] = \int_0^1 g(u) f_U(u)\,du = \int_0^1 g(u)\,du,$$

and since g is integrable, then

$$\int_0^1 |g(u)|\,du < \infty,$$

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

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

□

Theorem 7.10 (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. random variables with p.d.f. f_X and support $\mathbb{R}$. Then*

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

Proof. Since g is integrable, then

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

and

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

is defined. Notice that

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

Thus, by the strong LLN,

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

□

Example 7.11 (Importance Sampling)**.** *Suppose X has p.d.f. f_X and Y has p.d.f. f_Y. In addition, assume that if $f_Y(x) > 0$, then $f_X(x) > 0$. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with the same distribution as X and*

$$w_i = \frac{f_Y(X_i)}{f_X(X_i)}.$$

Then for any function g such that $g(Y)$ is integrable,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_i) w_i = E[g(Y)], \quad \text{a.s.}$$

Proof. Since $g(Y)$ is integrable, then

$$E[g(Y)] = \int g(y) f_Y(y)\, dy < \infty.$$

Since $f_Y(y) = f_X(y) \cdot \frac{f_Y(y)}{f_X(y)}$, we can write:

$$E[g(Y)] = \int g(y) \frac{f_Y(y)}{f_X(y)} f_X(y)\, dy = E\left[g(X) \cdot \frac{f_Y(X)}{f_X(X)} \right].$$

Let $h(X) = g(X) \cdot \frac{f_Y(X)}{f_X(X)}$. Then:

$$E[h(X)] = E[g(Y)].$$

By the Strong Law of Large Numbers,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_i) w_i = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(X_i) = E[g(Y)] \quad \text{a.s.}$$

□

Example 7.12 (Self-Normalized Importance Sampling)**.** *Suppose X has p.d.f. f_X and Y has p.d.f. $\tilde{f}_Y$ known up to a normalizing constant, that is,*

$$f_Y(y) = \frac{\tilde{f}_Y(y)}{\int \tilde{f}_Y(y)\, dy} := \frac{\tilde{f}_Y(y)}{Z},$$

where Z is unknown. Assume that if $f_Y(x) > 0$, then $f_X(x) > 0$. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with the same distribution as X, and

$$w_i = \frac{\tilde{f}_Y(X_i)}{f_X(X_i)}.$$

Then for any function g such that $g(Y)$ is integrable,

$$\lim_{n\to\infty} \sum_{i=1}^{n} g(X_i) \hat{w}_i = E[g(Y)], \quad \text{a.s.},$$

where

$$\hat{w}_i = \frac{w_i}{\sum_{j=1}^{n} w_j}.$$

Proof. Define:

$$N_n = \sum_{i=1}^n g(X_i)w_i = \sum_{i=1}^n g(X_i)\frac{\tilde{f}_Y(X_i)}{f_X(X_i)}, \quad D_n = \sum_{i=1}^n w_i = \sum_{i=1}^n \frac{\tilde{f}_Y(X_i)}{f_X(X_i)}.$$

We have

$$E[g(X)w(X)] = \int g(x)\frac{\tilde{f}_Y(x)}{f_X(x)}f_X(x)\,dx = \int g(x)\tilde{f}_Y(x)\,dx,$$

and

$$E[w(X)] = \int \frac{\tilde{f}_Y(x)}{f_X(x)}f_X(x)\,dx = \int \tilde{f}_Y(x)\,dx = Z.$$

Therefore,

$$\frac{E[g(X)w(X)]}{E[w(X)]} = \frac{1}{Z}\int g(x)\tilde{f}_Y(x)\,dx = E[g(Y)].$$

By the Strong Law of Large Numbers,

$$\frac{N_n}{n} \to E[g(X)w(X)], \quad \frac{D_n}{n} \to E[w(X)] = Z, \quad \text{a.s.}$$

Consequently, by the continuous mapping theorem,

$$\sum_{i=1}^n g(X_i)\hat{w}_i = \frac{N_n}{D_n} = \frac{N_n/n}{D_n/n} \to \frac{E[g(X)w(X)]}{E[w(X)]} = E[g(Y)], \quad \text{a.s. as } n \to \infty.$$

□

7.5 Exercises

Question 7.13. *Suppose $X_1, X_2, \ldots$ are uncorrelated with $E[X_i] = \mu_i$ and $\lim_{i\to\infty} \frac{\text{Var}(X_i)}{i} = 0$. If $\nu_n = \frac{1}{n}E[S_n]$, show that*

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

Question 7.14. *Let $X_1, X_2, \ldots$ be i.i.d. such that $P(X_1 = (-1)^k k) = \frac{C}{k^2\log(k)}$ for all integers $k \geq 2$, where C is a constant so that the probabilities sum to 1. Show that $E[|X_1|] = \infty$ but there is a finite μ such that $\frac{S_n}{n} \xrightarrow{P} \mu$.*

Hint #1: $\mu_n = E[X_1 \mathbb{1}_{\{|X_1|\leq n\}}] = \sum_{k=2}^n (-1)^k k\frac{C}{k^2\log(k)}$ is an alternating sequence of real numbers that converge to zero, so from calculus class $\mu_n \to \mu$ 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, $\sum_{k=x+1}^\infty f(k) \leq \int_x^\infty f(y)dy \leq \sum_{k=x}^\infty f(k)$.

Question 7.15. *Let $X_1, X_2, \ldots$ be i.i.d. such that $P(X_1 > x) = \frac{e}{x\log(x)}$ for $x \geq e$. Show that $E[|X_1|] = \infty$, but $\frac{S_n}{n} - \mu_n \xrightarrow{P} 0$.*

Hint: Recall Question 5.50.

Question 7.16. *For any sequence of random variables X_n and $\varepsilon > 0$, show there exist constants $c_n \to \infty$ such that $P(|X_n| > \varepsilon c_n) < 2^{-n}$.*

Question 7.17. *For any sequence of random variables X_n, show there exist constants $c_n \to \infty$ such that*

$$\frac{X_n}{c_n} \to 0 \quad \textit{a.s.}$$

Question 7.18. *Suppose $X_1, X_2, \ldots$ are i.i.d. Show that $E[|X_1|] < \infty$ if and only if*

$$\frac{X_n}{n} \to 0 \quad a.s.$$

Question 7.19. *Suppose $X_1, X_2, \ldots$ are i.i.d. such that for all n, $P(X_n > x) = e^{-x}$. Show that*

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

8 Central Limit Theorems

8.1 Weak Convergence

Definition 8.1 (Weak convergence)**.** *A sequence $\{\mu_n\}$ of probability measures with distribution functions F_n converges weakly to a probability measure μ with distribution function F if, for every continuity point x of F,*

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

We write $\mu_n \Rightarrow \mu$, or $F_n \Rightarrow F$, or $X_n \Rightarrow X$ when $X_n \sim F_n$ and $X \sim F$.

Theorem 8.2 (Scheffé's Theorem)**.** *Let X_n have density f_n and X have density f. If $f_n(x) \to f(x)$ for all $x \in \mathbb{R}$, then $X_n \Rightarrow X$.*

Proof. Define $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$, so $x^+ + x^- = |x|$, $x^+ - x^- = x$, and $(-x)^+ = x^-$. Since f is a density, its distribution function F is continuous everywhere. Fix $y \in \mathbb{R}$:

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

Because $[f(x) - f_n(x)]^+ \le f(x)$ and f is integrable, the dominated convergence theorem gives

$$\lim_{n\to\infty} |F_n(y) - F(y)| \le \lim_{n\to\infty} 2\int_{\mathbb{R}} [f(x) - f_n(x)]^+ dx = 2\int_{\mathbb{R}} \lim_{n\to\infty} [f(x) - f_n(x)]^+\, dx = 0.$$

Hence $F_n(y) \to F(y)$ for every $y \in \mathbb{R}$, so $X_n \Rightarrow X$. □

Theorem 8.3 (Portmanteau Lemma)**.** *$F_n \Rightarrow F \iff$ for every bounded, continuous $h : \mathbb{R} \to \mathbb{R}$, whenever $X_n \sim F_n$ and $X \sim F$,*

$$E[h(X_n)] \longrightarrow E[h(X)].$$

Proof. NPOS. □

Corollary 8.4 (Continuous Mapping Theorem)**.** *If $X_n \Rightarrow X$, then $f(X_n) \Rightarrow f(X)$ for any continuous f.*

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

$$E[g(h(X_n))] = E[g \circ h(X_n)] \to E[g \circ h(X)] = E[g(h(X_n))]$$

□

Corollary 8.5. *If $X_n \xrightarrow{P} X$ then $X_n \Rightarrow X$.*

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 $X_{n_{j(k)}} \to X$ a.s. Thus, for bounded and continuous h, the continuous mapping theorem gives $h(X_{X_{n_{j(k)}}}) \to h(X)$ a.s. And bounded convergence theorem give that $E[h(X_{n_{j(k)}})] \to E[h(X)]$. Since $E[h(X_n)]$ is a sequence of real numbers, and we just showed for all subsequence has a convergent subsubsequence, this implies $E[h(X_n)] \to E[h(X)]$. □

Example 8.6. *The reverse does not hold. Let $X \sim \mathcal{N}(0,1)$ and $X_n = -X$ for $n \in \mathbb{N}$. Then, trivially $F_n = F$, so $X_n \Rightarrow X$, but*

$$P(|X_n - X| > \epsilon) = P(|X| > \epsilon/2) = c > 0.$$

Lemma 8.7. *If $X_n \Rightarrow c$ where c is constant, then $X_n \xrightarrow{P} c$.*

Proof. Let F be the CDF of the random variable $X \equiv c$, so that $F(y) = \mathbf{1}_{\{c \le y\}}$. Observe that F is continuous everywhere except $y = c$. Now, fix $\epsilon > 0$.

$$\begin{aligned}
\lim_{n\to\infty} P(|X_n - c| > \epsilon) &= \lim_{n\to\infty} \big(P(X_n < c - \epsilon) + P(X_n > c + \epsilon)\big) \\
&\le \lim_{n\to\infty} \big(P(X_n \le c - \epsilon/2) + P(X_n > c + \epsilon)\big) \\
&= \lim_{n\to\infty} F_n(c - \epsilon/2) + 1 - \lim_{n\to\infty} F_n(c + \epsilon) \\
&= F(c - \epsilon/2) + 1 - F(c + \epsilon) \\
&= 0.
\end{aligned}$$

□

Note: we need to set it to $c - \epsilon/2$ in the above instead of c because c is a discontinuity point of F, and there may be no convergence at this point.

Theorem 8.8 (Slutsky's Theorem)**.** *If $X_n \Rightarrow X$ and $Y_n \Rightarrow c$ for a constant c, then:*

1. *$X_n + Y_n \Rightarrow X + c$,*
2. *$X_n Y_n \Rightarrow Xc$,*
3. *$X_n / Y_n \Rightarrow X/c$ if $c \neq 0$.*

Proof. 1. Fix $\epsilon > 0$ and let z be a continuity point of the CDF of $X + c$. Observe that $X_n \le z - c - \epsilon$ and $Y_n \le c + \epsilon$ implies $X_n + Y_n \le z$. Thus,

$$\begin{aligned}
P(X_n + Y_n \le z) &\ge P(X_n \le z - c - \epsilon \ \cap \ Y_n \le c + \epsilon) \\
&= P(X_n \le z - c - \epsilon) + P(Y_n \le c + \epsilon) - P(X_n \le z - c - \epsilon \cup Y_n \le c + \epsilon) \\
&\ge P(X_n \le z - c - \epsilon) + P(Y_n \le c + \epsilon) - 1 \\
&= P(X_n \le z - c - \epsilon) - P(Y_n > c + \epsilon)
\end{aligned}$$

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

$$\lim_{n\to\infty} P(Y_n > c + \epsilon) = P(Y > c + \epsilon) = 0.$$

For arbitrarily small ϵ we can take $z - c - \epsilon$ to be a continuity point of the CDF of X (we can do this by Question 2.53), without loss of generality, so

$$\lim_{n\to\infty} P(X_n \le z - c - \epsilon) = P(X \le z - c - \epsilon).$$

That is,

$$\liminf_{n\to\infty} P(X_n + Y_n \le z) \ge P(X + c \le z - \epsilon).$$

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

$$\liminf_{n\to\infty} P(X_n + Y_n \le z) \ge P(X + c \le z).$$

Similarly, $X_n + Y_n \le z$ and $Y_n \ge c - \epsilon$ implies $X_n \le z - c + \epsilon$, so

$$\begin{aligned} P(X_n \le z - c + \epsilon) &\ge P(X_n + Y_n \le z \cap Y_n \ge c - \epsilon) \\ &= P(X_n + Y_n \le z) + P(Y_n \ge c - \epsilon) - P(X_n + Y_n \le z \cup Y_n \ge c - \epsilon) \\ &\ge P(X_n + Y_n \le z) + P(Y_n \ge c - \epsilon) - 1 \\ &= P(X_n + Y_n \le z) - P(Y_n < c - \epsilon) \end{aligned}$$

Rearranging this gives

$$P(X_n + Y_n \le z) \le P(X_n \le z - c + \epsilon) + P(Y_n < c - \epsilon)$$

Since $Y_n \to Y$, we get

$$\lim_{n\to\infty} P(Y_n < c - \epsilon) = P(Y < c - \epsilon) = 0$$

For arbitrary small ϵ we can take $z - c + \epsilon$ to be a continuity point of the CDF of X, WLOG, so

$$\lim_{n\to\infty} P(X_n \le z - c + \epsilon) = P(X \le z - c + \epsilon)$$

Therefore, we get

$$\limsup_{n\to\infty} P(X_n + Y_n \le z) \le P(X \le z - c + \epsilon) = P(X + c \le z - \epsilon)$$

Since z is a continuity point of $X + c$, take $\epsilon \to 0$ we get

$$\limsup_{n\to\infty} P(X_n + Y_n \le z) \le P(X + c \le z)$$

2. Since $X_n \Rightarrow X, Y_n \Rightarrow c$, by Continuous Mapping Theorem where $f(x) = \log(x)$ is continuous, we have $\log(X_n) \Rightarrow \log(X), \log(Y_n) \Rightarrow \log(c)$.

   By Part 1, $\log(X_n) + \log(Y_n) = \log(X) + \log(c)$. Again, by Continuous Mapping Theorem where $g(x) = e^x$ we get

   $$e^{\log(X_n)+\log(Y_n)} \Rightarrow e^{\log(X)+\log(c)} \implies X_n Y_n \Rightarrow cX$$

3. Since $X_n \Rightarrow X, Y_n \Rightarrow c$, by Continuous Mapping Theorem where $f(x) = -x$ is continuous, we have $-Y_n \Rightarrow -c$. By Continuous Mapping Theorem where $g(x) = \log(x)$ is continuous, we have $\log(X_n) \Rightarrow \log(X), \log(-Y_n) \Rightarrow \log(-c)$.

   By Part 1, $\log(X_n) + \log(-Y_n) = \log(X) + \log(-c)$. Again, by Continuous Mapping Theorem where $h(x) = e^x$ we get

   $$e^{\log(X_n)+\log(-Y_n)} \Rightarrow e^{\log(X)+\log(-c)} \implies \frac{X_n}{Y_n} \Rightarrow \frac{X}{c}$$

□

Proposition 8.9 (Delta Method)**.** *Suppose $a_n(X_n-\theta) \Rightarrow X$ 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*

$$a_n\big(g(X_n) - g(\theta)\big) \Rightarrow g'(\theta)X.$$

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

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

Define

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

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

$$g(\theta+h) = g(\theta) + g'(\theta)h + h\, r(h).$$

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

$$a_n(g(X_n) - g(\theta)) = g'(\theta)a_n(X_n-\theta) + a_n(X_n-\theta)r(X_n-\theta).$$

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

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

Choose $x > 0$ such that x and $-x$ are the continuous points of F_X and that $P(X \le -x) \le \delta/4$ and $P(X > x) \le \delta/4$. Since $a_n(X_n-\theta) \Rightarrow X$, there is a sufficiently large N such that, for all $n \ge N$, $a_n\epsilon > x$ and

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

and

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

which gives us

$$P(|X_n-\theta| > \epsilon) < \delta$$

for all $N \ge n$. Since δ is arbitrary, we showed that $X_n \xrightarrow{P} \theta$.

For any $\epsilon > 0$, since $r(h) \to 0$ as $h \to 0$, there exists $\delta > 0$ such that for all $|h| < \delta, |r(h)| < \epsilon$. Hence, we obtain

$$\lim_{n\to\infty} P(|r(X_n-\theta)| \ge \epsilon) \le \lim_{n\to\infty} P(|X_n-\theta| \ge \delta) = 0$$

Therefore, $r(X_n-\theta) \xrightarrow{P} 0$, and by Slutsky's theorem,

$$a_n(g(X_n) - g(\theta)) = g'(\theta)a_n(X_n-\theta) + a_n(X_n-\theta)r(X_n-\theta) \Rightarrow g'(\theta)X.$$

□

8.2 Characteristic Functions

Remark 8.10. *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 8.11 (Euler's Formula)**.** *For any $x \in \mathbb{R}$,*

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

Proof. See first-year calculus book. □

Definition 8.12. *For a random variable X taking values in $\mathbb{C}$, we define its expectation by*

$$E[X] = E[Re(X)] + iE[Im(X)].$$

Definition 8.13. *A random variable X has a unique characteristic function defined by*

$$\varphi(t) = E[e^{itX}] = E[\cos(tX)] + iE[\sin(tX)].$$

Proposition 8.14. *A characteristic function φ has the following properties:*

1. $\varphi(0) = 1$.
2. $\varphi(-t) = \overline{\varphi(t)}$.
3. $|\varphi(t)| \leq 1$.
4. $|\varphi(t+h) - \varphi(t)| \leq E\big[|e^{-ihX} - 1|\big]$.
5. $E[e^{it(aX+b)}] = e^{itb}\varphi(at)$.

Proof. 1. We have

$$\phi(0) = E[\cos(0)] + iE[\sin(0)] = E[1] = 1$$

2. We have

$$\phi(-t) = E[\cos(-tX)] + iE[\sin(-tX)] = E[\cos(tX)] + iE[-\sin(tX)] = E[\cos(tX)] - iE[\sin(tX)] = \overline{\phi(t)}$$

3. We have

$$|\phi(t)| = |E[e^{itX}]| \leq E[|e^{itX}|] = E[|\cos(tX) + i\sin(tX)|] = E\left[\sqrt{\cos^2(tX) + \sin^2(tX)}\right] = E[1] = 1$$

4. We have

$$\begin{aligned}
|\phi(t+h) - \phi(t)| &= |E[e^{i(t+h)X}] - E[e^{itX}]| \\
&= |E[e^{itX}(e^{ihX} - 1)]| \\
&\leq E[|e^{itX}(e^{ihX} - 1)|] \\
&= E[|e^{itX}||e^{ihX} - 1|] \\
&= E[|\cos(tX) + i\sin(tX)||\cos(hX) + i\sin(hX) - 1|] \\
&= E[|\cos(hX) + i\sin(hX) - 1|] \\
&= E\left[\sqrt{(\cos(hX) - 1)^2 + \sin^2(hX)}\right] \\
&= E[|\cos(hX) - i\sin(hX) - 1|] \\
&= E[|\cos(-hX) + i\sin(-hX) - 1|] \\
&= E[|e^{-ihX} - 1|]
\end{aligned}$$

5. We have

$$E[e^{it(aX+b)}] = E[e^{itaX}e^{itb}] = e^{itb}E[e^{itaX}] = e^{itb}\phi(at)$$

□

Proposition 8.15. *If X and Y are independent with characteristic functions φ_X and φ_Y, then $Z = X + Y$ has characteristic function*

$$\varphi_Z(t) = \varphi_X(t)\varphi_Y(t).$$

Proof.

$$\begin{aligned}
\varphi_Z(t) &= E[e^{it(X+Y)}] \\
&= E[e^{itX}e^{itY}] \\
&= E[e^{itX}]E[e^{itY}] \qquad \text{[By Independence]} \\
&= \varphi_X(t)\varphi_Y(t)
\end{aligned}$$

□

Theorem 8.16. *Common distributions have the following characteristic functions:*

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

Proof. Take $X \sim Ber(P)$, we have

$$\begin{aligned}
\phi(t) &= E[\cos(tX)] + iE[\sin(tX)] \\
&= \sum_{x\in\{0,1\}} \cos(tX)P(X = x) + i \sum_{x\in\{0,1\}} \sin(tX)P(X = x) \\
&= 1 - p + p\cos(t) + i\sin(t)p \\
&= 1 - p + p(\cos(t) + i\sin(t)) \\
&= 1 - p + pe^{it}
\end{aligned}$$

Take $X \sim Poisson(\lambda)$, we have

$$\begin{aligned}
\phi(t) &= E[\cos(tX)] + iE[\sin(tX)] \\
&= \sum_{x=0}^{\infty} \cos(tx)e^{-\lambda}\frac{\lambda^x}{x!} + i\sum_{x=0}^{\infty} \sin(tx)e^{-\lambda}\frac{\lambda^x}{x!} \\
&= e^{-\lambda}\left(\sum_{x=0}^{\infty} \frac{\lambda^x(\cos(tx) + i\sin(tx))}{x!}\right) \\
&= e^{-\lambda}\left(\sum_{x=0}^{\infty} \frac{\lambda^x e^{itx}}{x!}\right) \\
&= e^{-\lambda}\left(\sum_{x=0}^{\infty} \frac{(\lambda e^{it})^x}{x!}\right) \\
&= e^{-\lambda}e^{\lambda e^{it}} \\
&= \exp\{\lambda(e^{it} - 1\}
\end{aligned}$$

Take $X \sim Exp(\lambda)$, we have

$$\begin{aligned}
\phi(t) &= E[e^{itX}] \\
&= \int_0^\infty e^{itx}\lambda e^{-\lambda x}dx \\
&= \lambda\int_0^\infty e^{(it-\lambda)x}dx \\
&= \lambda\left[\frac{e^{(it-\lambda)x}}{it-\lambda}\right]_0^\infty \\
&= \lambda\left[\frac{e^{itx}}{e^{\lambda x}(it-\lambda)}\right]_0^\infty \\
&= \lambda\left(-\frac{1}{it-\lambda}\right) \\
&= \frac{\lambda}{\lambda-it}
\end{aligned}$$

Take $X \sim \mathcal{N}(\mu, \sigma^2)$, we have

$$\begin{aligned}
\phi(t) &= E[e^{itX}] \\
&= \int_{-\infty}^\infty e^{itx}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx \\
&= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{itx-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx \\
&= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}((x-\mu)^2-2itx\sigma^2)\right\}dx \\
&= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}((x-\mu-it\sigma^2)^2-2it\mu\sigma^2+t^2\sigma^4)\right\}dx \\
&= \exp\left\{i\mu t-\frac{t^2\sigma^2}{2}\right\}\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu-it\sigma^2)^2}{2\sigma^2}\right\}dx \\
&= \exp\left\{i\mu t-\frac{t^2\sigma^2}{2}\right\}
\end{aligned}$$

□

Theorem 8.17 (Leibniz's Rule for Expectation)**.** *Let X be a random variable and consider a function $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ such that $E[f(t, X)] < \infty$ 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 $E[Y] < \infty$ and $|f'(t, X)| \leq Y$ for $t \in (a, b)$. Then, for all $t \in (a, b)$,*

$$\frac{\partial}{\partial t}E[f(t, X)] = E[f'(t, X)].$$

Proof. First, observe that

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

and

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

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

$$\begin{aligned}\frac{\partial}{\partial t}E[f(t,X)] &= \lim_{h\to 0} E\left[\frac{f(t+h,X)-f(t,X)}{h}\right] \\ &= E\left[\lim_{h\to 0}\frac{f(t+h,X)-f(t,X)}{h}\right] \\ &= E[f'(t,X)].\end{aligned}$$

□

Proposition 8.18. *If X is a random variable with $E[|X|^k] < \infty$, then for $0 \le j \le k$,*

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

Proof. We prove by induction.

When $j = 0$, we get

$$\phi_X^{(0)}(t) = E[e^{itX}] = E[(iX)^0 e^{itX}]$$

Suppose this holds for some j. Then, for $j+1 \le k$

$$\begin{aligned}\varphi_X^{(j+1)}(t) &= \frac{\partial}{\partial t}\varphi_X^{(j)}(t) \\ &= \frac{\partial}{\partial t}E[(iX)^j e^{itX}] \\ &= E\left[(iX)^j \frac{\partial}{\partial t}e^{itX}\right] \qquad \text{[By Leibniz's Rule for Expectation]} \\ &= E[(iX)^{j+1}e^{itX}],\end{aligned}$$

where we can use Leibniz's Rule as

$$E[(iX)^j e^{itX}] = i^j E[X^j(\cos(tX) + i\sin(tX))] \le i^j E[X^j] + i^{j+1}E[X^j] < \infty$$

Moreover,

$$\begin{aligned}\left|(iX)^j \frac{\partial}{\partial t}\right| &= \left|(iX)^{j+1}e^{itX}\right| \\ &= \left|i^{j+1}X^{j+1}(\cos(tX) + i\sin(tX))\right| \\ &= |i^{j+1}||X^{j+1}||\cos(tX) + i\sin(tX)| \\ &= |X^{j+1}|\end{aligned}$$

where $E[|X^{j+1}|] < \infty$ □

Theorem 8.19 (Continuity Theorem)**.** *Let $\mu, \mu_1, \mu_2, \ldots$ be probability measures with characteristic functions $\varphi, \varphi_1, \varphi_2, \ldots$. Then, $\mu_n \Rightarrow \mu$ if and only if $\varphi_n(t) \to \varphi(t)$ for all $t \in \mathbb{R}$.*

Proof. NPOS □

Lemma 8.20. *If X_n, Y_n are independent for all n, and X, Y are independent with $X_n \Rightarrow X$ and $Y_n \Rightarrow Y$, then*

$$X_n + Y_n \Rightarrow X + Y.$$

Proof. Using characteristic functions,

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

Then apply the continuity theorem. □

Proposition 8.21 (Poisson Convergence Theorem)**.** *Suppose X_n is a sequence of random variables such that $X_n \sim Bin(n, p_n)$ with $np_n \to \lambda > 0$ as $n \to \infty$, then*

$$X_n \Rightarrow Poisson(\lambda).$$

Proof. The characteristic function of X_n is

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

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

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

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

8.3 Central Limit Theorem

Lemma 8.22. *For any random variable with $E[|X|^k] < \infty$ for $0 \le k \le m$,*

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

Proof. By Taylor Series approximation up to mth order centred at 0, we have

$$\phi_X(t) = \sum_{k=0}^{m} \frac{\phi_X^{(k)}(0)}{k!} t^k + o(t^m) = \sum_{k=0}^{m} \frac{E[(iX)^k]}{k!} t^k + o(|t|^m) = \sum_{k=0}^{m} \frac{(it)^k}{k!} E[X^k] + o(|t|^m)$$

□

Theorem 8.23. *If $X_1, X_2, \ldots$ are i.i.d. with $E[X_n] = 0$ and $E[X_n^2] = \sigma^2 < \infty$, then*

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

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

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

Thus,

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

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

8.4 Exercises

Question 8.24. *A Cauchy random variable has density defined by*

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

for all $x \in R$. Show that for all $t \neq 0$, the MGF

$$M_X(t) = E[e^{tX}] = \infty.$$

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.

Question 8.25. *Suppose $X_n \sim Gaussian(0, \sigma_n^2)$ and $X_n \Rightarrow X$ for some random variable X. Show that $\sigma_n \to \sigma$ for some $\sigma \in [0, \infty)$.*

Question 8.26. *Prove Scheffé's Theorem for discrete pmfs instead of densities.*

Question 8.27. *Find an example of random variables X_n with densities f_n such that*

$$X_n \Rightarrow Unif(0,1)$$

but

$$\{x : f_n(x) \to 1\} = \varnothing.$$

Question 8.28 (Optional)**.** *Prove that if $P_n \stackrel{d}{\Rightarrow} P$ and $P_n \stackrel{d}{\Rightarrow} P'$, then $P = P'$.* Hint: use the Portmanteau lemma.

Question 8.29. *Prove that if $F_n \stackrel{d}{\Rightarrow} F$ and x is such that there is at most one $a \in \mathbb{R}$ with $F(a) = x$, then*

$$F_n^{-1}(x) \to F^{-1}(x).$$

Hint: *For any $\epsilon > 0$, you can choose a y such that F is continuous at y (why?) and*

$$F^{-1}(x) - \epsilon < y < F^{-1}(x).$$

Question 8.30. *Find an example such that $X_n \Rightarrow X$ and $Y_n \Rightarrow Y$ but $X_n + Y_n \not\Rightarrow X + Y$.*

Question 8.31. *Let $X_1, X_2, \ldots$ be i.i.d. with characteristic function ϕ. Show that if $\phi'(0) = ia$, then*

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