MATH 2431: Honors Probability

HU-HTAKM
Website: https://htakm.github.io/htakm_test/

Last major change: May 31, 2024
Last small update (fixed typo): September 14, 2025

This lecture note is based on the HKUST MATH 2431 lecture notes by Prof. Bao, Zhigang (Spring 2023-24). To clarify certain topics, I have also included material from the textbook "Probability and Random Processes" (Third Edition) by G. Grimmett and D. Stirzaker. The chapters follow the textbook's structure.
Some proofs are written by myself, as they are not found in either the lecture notes or the textbook. These may contain errors. If you notice any, you likely possess a strong understanding of the topic or a keen eye for detail. ;)
This course requires a co-requisite in multivariable calculus (MATH 2011 and MATH 2023 for HKUST students). However, it is strongly recommended to be familiar with multivariable calculus beforehand, as it is used early in the course. Knowledge of mathematical analysis is also very helpful.
If any topics are unclear or not well explained, it is likely due to my non-mathematics background. ;)

Notations	Meaning
$\mathbb{N}_+$	Set of positive integers
$\mathbb{N}$	Set of natural numbers
$\mathbb{Z}$	Set of integers
$\mathbb{Q}$	Set of rational numbers
$\mathbb{R}$	Set of real numbers
$\emptyset$	Empty set
Ω	Sample space / Entire set
ω	Outcome
$\mathcal{F}, \mathcal{G}, \mathcal{H}$	σ-field / σ-algebra
$A, B, C, \cdots$	Events
$A^{\complement}$	Complement of events
$\mathbb{P}$	Probability measure
X	Random variable
$\mathcal{B}(\mathbb{R})$	Borel σ-field of $\mathbb{R}$
f_X	PMF/PDF of X
F_X	CDF of X
$\mathbf{1}_A$	Indicator function
$\mathbb{E}$	Expectation
ψ	Conditional expectation
$\mathbf{u}, \mathbf{v}, \mathbf{w}, \cdots$	Vector
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \cdots$	Matrix
$\mathbf{X}$	Random vector
G_X	Probability generating function of X
M_X	Moment generating function of X
ϕ	CF / PDF of $X \sim \mathrm{N}(0,1)$
Φ	CDF of $X \sim \mathrm{N}(0,1)$

(a) Notations

Abbreviations	Meaning
CDF	Cumulative distribution function
JCDF	Joint cumulative distribution function
PMF	Probability mass function
JPMF	Joint probability mass function
PDF	Probability density function
JPDF	Joint probability density function
PGF	Probability generating function
MGF	Moment generating function
CF	Characteristic function
JCF	Joint characteristic function
i.i.d.	independent and identically distributed
WLLN	Weak Law of Large Numbers
SLLN	Strong Law of Large Numbers
CLT	Central Limit Theorem
BCI	Borel-Cantelli Lemma I
BCII	Borel-Cantelli Lemma II
i.o.	infinitely often
f.o.	finitely often
a.s.	almost surely

(b) Abbreviations

Definition 0.1. This is definition.

Remark 0.1.1. This is remark.

Lemma 0.2. This is lemma.

Proposition 0.3. This is proposition.

Theorem 0.4. This is theorem.

Claim 0.4.1. This is claim.

Corollary 0.5. This is corollary.

Example 0.1. This is example.

Contents

Chapter 1

Events and their probabilities

1.1 Fundamental terminologies

In everyday life, we often perceive the future as largely unpredictable. This belief is reflected in our understanding of random phenomena, to which we assign both quantitative and qualitative meanings.
We start with some basic terminology.

Definition 1.1. A **sample space** is the set of all outcomes of an experiment and is denoted by Ω. **Outcomes** are denoted by ω.

Example 1.1. For a coin flip, the sample space is $\Omega = \{H, T\}$.

Example 1.2. For a die roll, the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Example 1.3. For the lifetime of a bulb, the sample space is $\Omega = [0, \infty)$.

Example 1.4. For two coins flipping, the sample space is $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$.

Many statements in probability take the form "the probability of event A is p," where the events typically include certain elements of the sample space.

Definition 1.2. An **event** is a subset of the sample space. Outcomes are **elementary events**.

Remark 1.2.1. Not every subset of Ω must be considered an event. However, we will not address this issue at present.

Example 1.5. For a dice roll, the sample space is $\Omega = \{1, 2, \cdots, 6\}$. An example of an event is rolling an even number: $A = \{2, 4, 6\}$.

Remark 1.2.2. If only the outcome $\omega = 2$ is given, then there are many events that could result in this outcome. For example, $\{2\}$, $\{2, 4\}$, etc.

Definition 1.3. The **complement** of a subset A is the set $A^{\complement}$, which contains all elements in the sample space Ω that are not in A.

We can define a collection of subsets of the sample space.

Definition 1.4. A **field** $\mathcal{F}$ is any collection of subsets of Ω which satisfies the following conditions:

1. If $A \in \mathcal{F}$, then $A^{\complement} \in \mathcal{F}$.
2. If $A, B \in \mathcal{F}$, then $A \cup B \in \mathcal{F}$ and $A \cap B = (A^{\complement} \cup B^{\complement})^{\complement} \in \mathcal{F}$. (Closed under *finite* unions or intersections)
3. $\emptyset \in \mathcal{F}$ and $\Omega = A \cup A^{\complement} \in \mathcal{F}$.

We are particularly interested in σ-fields, which are closed under countably infinite unions.

Definition 1.5. A **σ-field** (or **σ-algebra**) $\mathcal{F}$ is any collection of subsets of Ω which satisfies the following conditions:

1. If $A \in \mathcal{F}$, then $A^{\complement} \in \mathcal{F}$.
2. If $A_1, A_2, \cdots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$. (Closed under *countably infinite* unions)
3. $\emptyset \in \mathcal{F}$ and $\Omega = A \cup A^{\complement} \cup \cdots \in \mathcal{F}$.

Remark 1.5.1. From this point onwards, $\mathcal{F}$ will denote the σ-field.

Example 1.6. Smallest σ-field: $\mathcal{F} = \{\emptyset, \Omega\}$.

Example 1.7. If A is any subset of Ω, then $\mathcal{F} = \{\emptyset, A, A^{\complement}, \Omega\}$ is a σ-field.

Example 1.8. Largest σ-field: Power set of Ω: $2^{\Omega} = \{0,1\}^{\Omega} := \{\text{All subsets of } \Omega\}$.
When Ω is infinite, the power set is too large a collection for probabilities to be assigned reasonably.

Remark 1.5.2. The following two formulae are particularly useful:

$$(a,b) = \bigcup_{n=1}^{\infty} \left[a + \frac{1}{n}, b - \frac{1}{n}\right] \qquad [a,b] = \bigcap_{n=1}^{\infty} \left[a - \frac{1}{n}, b + \frac{1}{n}\right]$$

1.2 Probability measure

We wish to discuss the likelihood of the occurrence of events.
Now that we have defined some fundamental terminologies, we can define probability.

Definition 1.6. A **measurable space** $(\Omega, \mathcal{F})$ is a pair comprising a sample space Ω and a σ-field $\mathcal{F}$.
A **measure** μ on a measurable space $(\Omega, \mathcal{F})$ is a function $\mu : \mathcal{F} \to [0, \infty]$ satisfying:

1. $\mu(\emptyset) = 0$.
2. If $A_i \in \mathcal{F}$ for all i and they are disjoint ($A_i \cap A_j = \emptyset$ for all $i \neq j$), then $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$. (Countable additivity)

A **probability measure** $\mathbb{P}$ is a measure with $\mathbb{P}(\Omega) = 1$.

You might wonder, "Isn't this just probability?" The probability we are familiar with is indeed a probability measure, which we will define shortly. However, there exist other measures that satisfy the definition of a probability measure, such as the risk-neutral measure.
The following examples are not probability measures:

Example 1.9. Lebesgue measure: $\mu((a,b)) = b - a$, $\Omega = \mathbb{R}$.

Example 1.10. Counting measure: $\mu(A) = \#\{A\}$, $\Omega = \mathbb{R}$.

We can combine a measurable space and a measure to form a measure space.

Definition 1.7. A **measure space** is the triple $(\Omega, \mathcal{F}, \mu)$, comprising:

1. A sample space Ω.
2. A σ-field $\mathcal{F}$ of certain subsets of Ω.
3. A measure μ on $(\Omega, \mathcal{F})$.

A **probability space** $(\Omega, \mathcal{F}, \mathbb{P})$ is a measure space with a probability measure $\mathbb{P}$ as the measure.

Example 1.11. Consider a coin flip. The sample space is $\Omega = \{H, T\}$, and the σ-field is $\mathcal{F} = \{\emptyset, H, T, \Omega\}$. Let $\mathbb{P}(H) = p$, where $p \in [0, 1]$. Define $A = \{\omega \in \Omega : \omega = H\}$. Then:

$$\mathbb{P}(A) = \begin{cases} 0, & A = \emptyset \\ p, & A = \{H\} \\ 1 - p, & A = \{T\} \\ 1, & A = \Omega \end{cases}$$

If $p = \frac{1}{2}$, then the coin is fair.

Example 1.12. Consider a die roll. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, and the σ-field is $\mathcal{F} = \{0, 1\}^{\Omega}$. Let $p_i = \mathbb{P}(\{i\})$, where $i \in \Omega$. For all $A \in \mathcal{F}$:

$$\mathbb{P}(A) = \sum_{i \in A} p_i$$

If $p_i = \frac{1}{6}$ for all i, then the die is fair, and $\mathbb{P}(A) = \frac{|A|}{6}$.

The following properties are fundamental and form the basis of probability theory:

Lemma 1.8. Basic properties of $\mathbb{P}$:

1. $\mathbb{P}(A^{\complement}) = 1 - \mathbb{P}(A)$.
2. If $A \subseteq B$, then $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$.
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. If A and B are disjoint, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
4. (**Inclusion-exclusion formula**) For any set of events $\{A_1, \cdots, A_n\}$:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \cdots + (-1)^{n+1} \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n)$$

Proof.

1. $A \cup A^{\complement} = \Omega$ and $A \cap A^{\complement} = \emptyset \Longrightarrow \mathbb{P}(A \cup A^{\complement}) = \mathbb{P}(A) + \mathbb{P}(A^{\complement}) = 1$
2. $A \subseteq B \Longrightarrow B = A \cup (B \setminus A) \Longrightarrow \mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$
3. $A \cup B = A \cup (B \setminus A) \Longrightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(A) + \mathbb{P}(B \setminus (A \cap B)) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
4. By induction. When $n = 1$, it is obviously true. Assume it is true for some positive integers m. When $n = m + 1$,

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^{m+1} A_i\right) &= \mathbb{P}\left(\bigcup_{i=1}^{m} A_i\right) + \mathbb{P}(A_{m+1}) - \mathbb{P}\left(\bigcup_{i=1}^{m} A_i \cap A_{m+1}\right) \qquad \text{(Item 3)} \\ &= \sum_{i=1}^{m+1} \mathbb{P}(A_i) - \sum_{1 \leq i < j \leq m} \mathbb{P}(A_i \cap A_j) + \cdots (-1)^{m+1} \mathbb{P}\left(\bigcap_{i=1}^{m} A_i\right) - \mathbb{P}\left(\bigcup_{i=1}^{m} A_i \cap A_{m+1}\right) \\ &= \sum_{i=1}^{m+1} \mathbb{P}(A_i) - \sum_{1 \leq i < j \leq m+1} \mathbb{P}(A_i \cap A_j) + \cdots + (-1)^{m+2} \mathbb{P}\left(\bigcap_{i=1}^{m+1} A_i\right) \end{aligned}$$

Therefore, by induction, the Inclusion-exclusion formula is true for any set of events $\{A_1, \cdots, A_n\}$ for any $n \in \mathbb{N}_+$.

□

We recall the continuity of function $f : \mathbb{R} \to \mathbb{R}$. f is continuous at some point x if for all x_n, $x_n \to x$ when $n \to \infty$. We have:

$$\lim_{n \to \infty} f(x_n) = f\left(\lim_{n \to \infty} x_n\right) = f(x)$$

Similarly, we say a set function μ is continuous if for all A_n with $A = \lim_{n \to \infty} A_n$, we have:

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

Remark 1.8.1. Given a sequence of sets A_n. We have two types of set limit:

$$\limsup_{n\to\infty} A_n = \lim_{n\uparrow\infty} \sup_{m\geq n} A_m = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \{\omega \in \Omega : \omega \in A_n \text{ for infinitely many } n\}$$

$$\liminf_{n\to\infty} A_n = \lim_{n\uparrow\infty} \inf_{m\geq n} A_m = \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m = \{\omega \in \Omega : \omega \in A_n \text{ for all but finitely many } n\}$$

Apparently, $\liminf_{n\to\infty} A_n \subseteq \limsup_{n\to\infty} A_n$

Definition 1.9. We say a sequence of events A_n **converges** and $\lim_{n\to\infty} A_n$ exists if:

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

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $A_1, A_2, \cdots \in \mathcal{F}$ such that $A = \lim_{n\to\infty} A_n$ exists, then:

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

From the definition, we can get the following important lemma.

Lemma 1.10. If $A_1, A_2, \cdots$ are an increasing sequence of events ($A_1 \subseteq A_2 \subseteq \cdots$), then:

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n)$$

Similarly, if $A_1, A_2, \cdots$ are a decreasing sequence of events ($A_1 \supseteq A_2 \supseteq \cdots$), then:

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n)$$

Proof.
For $A_1 \subseteq A_2 \subseteq \cdots$, let $B_n = A_n \setminus A_{n-1}$

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{m=1}^{\infty} \mathbb{P}(B_m) = \lim_{n\to\infty} \sum_{m=1}^{n} \mathbb{P}(B_m) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{m=1}^{n} B_m\right) = \lim_{n\to\infty} \mathbb{P}(A_n)$$

For $A_1 \supseteq A_2 \supseteq \cdots$, we get $A^{\complement} = \bigcup_{i=1}^{\infty} A_i^{\complement}$ and $A_1^{\complement} \subseteq A_2^{\complement} \subseteq \cdots$. Therefore,

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = 1 - \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n^{\complement}\right) = 1 - \lim_{n\to\infty} \mathbb{P}(A_n^{\complement}) = \lim_{n\to\infty} \mathbb{P}(A_n)$$

□

We can give some terminology to some special probabilities.

Definition 1.11. An event A is **null** if $\mathbb{P}(A) = 0$.

Remark 1.11.1. Null events need not to be impossible. For example, the probability of choosing a point in a plane is 0.

Definition 1.12. An event A occurs **almost surely** if $\mathbb{P}(A) = 1$.

1.3 Conditional probability

Sometimes, we are interested in the probability of a certain event given that another event has occurred.

Definition 1.13. If $\mathbb{P}(B) > 0$, then the **conditional probability** that A occurs given that B occurs is:

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

Remark 1.13.1. For any event A, $\mathbb{P}(A)$ can be regarded as $\mathbb{P}(A|\Omega)$.

Remark 1.13.2. When $\mathbb{P}(E) = \mathbb{P}(E|F)$, E and F are said to be **independent**.

Remark 1.13.3. Given an event B. $\mathbb{P}(\cdot|B)$ is also a probability measure on $\mathcal{F}$.

Example 1.13. Two fair dice are thrown. Given that the first shows 3, what is the probability that the sum of number shown exceeds 6?

$$\mathbb{P}(\text{Sum} > 3|\text{First die shows 3}) = \frac{\frac{3}{36}}{\frac{1}{6}} = \frac{1}{6}$$

It is obvious that a certain event occurs when another event either occurs or not occurs.

Lemma 1.14. For any events A and B such that $0 < \mathbb{P}(B) < 1$:

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^{\complement})\mathbb{P}(B^{\complement})$$

Proof.
$A = (A \cap B) \cup (A \cap B^{\complement}) \implies \mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^{\complement}) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^{\complement})\mathbb{P}(B^{\complement})$ □

There are some cases when multiple events allow a certain event to occur.

Lemma 1.15. (Law of total probability) Let $\{B_1, B_2, \cdots, B_n\}$ be a partition of Ω ($B_i \cap B_j = \emptyset$ for all $i \neq j$ and $\bigcup_{i=1}^{n} = \Omega$). Suppose that $\mathbb{P}(B_i) > 0$ for all i. Then:

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

Proof.

$$\mathbb{P}(A) = \mathbb{P}(A \cap \Omega) = \mathbb{P}\left(A \cap \left(\bigcup_{i=1}^{n} B_i\right)\right) = \mathbb{P}\left(\bigcup_{i=1}^{n}(A \cap B_i)\right) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

□

1.4 Independence

In general, the probability of a certain event is affected by the occurrence of other events. There are some exceptions.

Definition 1.16. Two events A and B are **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. It is denoted by $A \perp\!\!\!\perp B$.
More generally, a family of events $\{A_i : i \in I\}$ is **(mutually) independent** if for all subsets J of I:

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

Remark 1.16.1. If the family of events $\{A_i : i \in I\}$ has the property that $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for all $i \neq j$, then it is **pairwise independent**.

Example 1.14. Roll for dice twice: $\Omega = \{1, 2, \cdots, 6\} \times \{1, 2, \cdots, 6\}$ and $\mathcal{F} = 2^\Omega$
Let A be event that the sum is 7. Event $A = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$.
Let B be event that the first roll is 4. Event $B = \{(4,1), (4,2), (4,3), (4,4), (4,5), (4,6)\}$
Let C be event that the second roll is 3. Event $C = \{(1,3), (2,3), (3,3), (4,3), (5,3), (6,3)\}$

$$\mathbb{P}(A \cap B) = \mathbb{P}((4,3)) = \frac{1}{36} = \frac{1}{6}\left(\frac{1}{6}\right) = \mathbb{P}(A)\mathbb{P}(B)$$
$$\mathbb{P}(B \cap C) = \mathbb{P}((4,3)) = \frac{1}{36} = \frac{1}{6}\left(\frac{1}{6}\right) = \mathbb{P}(B)\mathbb{P}(C)$$
$$\mathbb{P}(A \cap C) = \mathbb{P}((4,3)) = \frac{1}{36} = \frac{1}{6}\left(\frac{1}{6}\right) = \mathbb{P}(A)\mathbb{P}(C)$$
$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}((4,3)) = \frac{1}{36} \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$$

Therefore, events A, B and C are pairwise independent, but not mutually independent.

Proposition 1.17. If events A and B are independent, then so are $A \perp\!\!\!\perp B^\complement$ and $A^\complement \perp\!\!\!\perp B^\complement$.

Proof.

$$\mathbb{P}(A \cap B^\complement) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A)(1 - \mathbb{P}(B)) = \mathbb{P}(A)\mathbb{P}(B^\complement)$$

Therefore, $A \perp\!\!\!\perp B^\complement$ and also $A^\complement \perp\!\!\!\perp B^\complement$. □

Proposition 1.18. If events A, B, C are independent, then:

1. $A \perp\!\!\!\perp (B \cup C)$
2. $A \perp\!\!\!\perp (B \cap C)$

Proof.

1. Using the properties of probability,

$$\begin{aligned}\mathbb{P}(A \cap (B \cup C)) &= \mathbb{P}((A \cap B) \cup (A \cap C)) \\ &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap C) - \mathbb{P}(A \cap B \cap C) \\ &= \mathbb{P}(A)\mathbb{P}(B) + \mathbb{P}(A)\mathbb{P}(C) - \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) \\ &= \mathbb{P}(A)\mathbb{P}(B \cup C)\end{aligned}$$

2.

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

□

Remark 1.18.1. If events A and B are independent and $A \cap B = \emptyset$, then either $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$.

1.5 Product space

There are many σ-fields you can generate using a collection of subset of Ω. However, many of those may be too big to be useful. Therefore, we have the following definition.

Definition 1.19. Let A be a collection of subsets of Ω. The **σ-field generated by** A is:

$$\sigma(A) = \bigcap_{A \subseteq \mathcal{G}} \mathcal{G}$$

where $\mathcal{G}$ is also a σ-field.

Remark 1.19.1. $\sigma(A)$ is the smallest σ-field containing A.

Example 1.15. Let $\Omega = \{1, 2, \cdots, 6\}$ and $A = \{\{1\}\} \subseteq 2^{\Omega}$. $\sigma(A) = \{\emptyset, \{1\}, \{2, 3, \cdots, 6\}, \Omega\}$

Corollary 1.20. Suppose $(\mathcal{F}_i)_{i \in I}$ is a system of σ-fields in Ω. Then:

$$\bigcap_{i \in I} \mathcal{F}_i = \{A \in \Omega : A \in \mathcal{F}_i \text{ for all } i \in I\}$$

Now that we know which σ-field we should generate, we can finally combine two probability spaces together to form a new probability space.

Definition 1.21. Product space of two probability spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ is the probability space $(\Omega_1 \times \Omega_2, \mathcal{G}, \mathbb{P}_{12})$ comprising:

1. a collection of ordered pairs $\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$
2. a σ-algebra $\mathcal{G} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ where $\mathcal{F}_1 \times \mathcal{F}_2 = \{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}$
3. a probability measure $\mathbb{P}_{12} : \mathcal{F}_1 \times \mathcal{F}_2 \to [0, 1]$ given by:

$$\mathbb{P}_{12}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2)$$

for $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$.

Chapter 2

Random variables and their distribution

2.1 Introduction of random variables

Sometimes, we are not interested in the experiment itself but rather in the consequences of its random outcomes. These consequences can be represented as functions mapping a sample space to the real number field. Such functions are called "random variables."

Definition 2.1. A **random variable** is a function $X : \Omega \to \mathbb{R}$ with the property that for any $x \in \mathbb{R}$,

$$X^{-1}((-\infty, x]) = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}.$$

Remark 2.1.1. More generally, a random variable is a function X such that for all intervals $A \subseteq \mathbb{R}$,

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

Such a function is said to be **$\mathcal{F}$-measurable**.

Remark 2.1.2. The intervals can be replaced by any of the following classes:

1. (a, b) for all $a < b$,
2. $(a, b]$ for all $a < b$,
3. $[a, b)$ for all $a < b$,
4. $[a, b]$ for all $a < b$,
5. $(-\infty, x]$ for all $x \in \mathbb{R}$.

This is due to the following reasons:

1. X^{-1} can be interchanged with any set functions.
2. $\mathcal{F}$ is a σ-field.

Claim 2.1.1. Suppose $X^{-1}(B) \in \mathcal{F}$ for all open sets B. Then $X^{-1}(B') \in \mathcal{F}$ for all closed sets B'.

Proof.
For any $a, b \in \mathbb{R}$,

$$X^{-1}([a, b]) = X^{-1}\left(\bigcap_{n=1}^{\infty}\left(a - \frac{1}{n}, b + \frac{1}{n}\right)\right) = \bigcap_{n=1}^{\infty} X^{-1}\left(\left(a - \frac{1}{n}, b + \frac{1}{n}\right)\right) \in \mathcal{F}.$$

□

Remark 2.1.3. The $\mathcal{F}$-measurability of X is necessary because $\mathbb{P}(X \in A) = \mathbb{P}(\{\omega : X(\omega) \in A\}) = \mathbb{P}(X^{-1}(A))$. Thus, $X^{-1}(A)$ must belong to $\mathcal{F}$.

Example 2.1. A fair coin is tossed twice. $\Omega = \{HH, HT, TH, TT\}$. For all $\omega \in \Omega$, let $X(\omega)$ be the number of heads.

$$X(\omega) = \begin{cases} 0, & \omega \in \{TT\} \\ 1, & \omega \in \{HT, TH\} \\ 2, & \omega \in \{HH\} \end{cases} \qquad X^{-1}((-\infty, x]) = \begin{cases} \emptyset, & x < 0 \\ \{TT\}, & x \in [0, 1) \\ \{HT, TH, TT\}, & x \in [1, 2) \\ \Omega, & x \in [2, \infty) \end{cases}$$

If we choose $\mathcal{F} = \{\emptyset, \Omega\}$, then X is not a random variable. If we choose $\mathcal{F} = 2^{\Omega}$, then X is a random variable.

Before we continue, it is best if we know about Borel set first.

Definition 2.2. Borel set is a set which can be obtained by taking countable union, intersection or complement repeatedly. (Countably many steps)

Definition 2.3. Borel σ-field of $\mathbb{R}$ is a σ-field $\mathcal{B}(\mathbb{R})$ that is generated by all open sets. It is a collection of Borel sets.

Example 2.2. $\{(a, b), [a, b], \{a\}, \mathbb{Q}, \mathbb{R} \setminus \mathbb{Q}\} \subset \mathcal{B}(\mathbb{R})$. Note that closed sets can be generated by open sets.

Remark 2.3.1. In modern way of understanding, $(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} (\mathbb{R}, \mathcal{B}, \mathbb{P} \circ X^{-1})$

Claim 2.3.1. $\mathbb{P} \circ X^{-1}$ is a probability measure on $(\mathbb{R}, \mathcal{B})$.

Proof.

1. For all $B \in \mathcal{B}$, $\mathbb{P} \circ X^{-1}(B) = \mathbb{P}(\{\omega : X(\omega) \in B\}) \in [0, 1]$

$$\mathbb{P} \circ X^{-1}(\emptyset) = \mathbb{P}(\{\omega : X(\omega) \in \emptyset\}) = \mathbb{P}(\emptyset) = 0$$
$$\mathbb{P} \circ X^{-1}(\mathbb{R}) = \mathbb{P}(\{\omega : X(\omega) \in \mathbb{R}\}) = \mathbb{P}(\Omega) = 1$$

2. For any disjoint $B_1, B_2, \cdots \in \mathcal{B}$,

$$\mathbb{P} \circ X^{-1}\left(\bigcup_{i=1}^{\infty} B_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} X^{-1}(B_i)\right) = \sum_{i=1}^{\infty} \mathbb{P}(X^{-1}(B_i)) = \sum_{i=1}^{\infty} \mathbb{P} \circ X^{-1}(B_i)$$

□

Remark 2.3.2. We can derive the probability of all $A \in \mathcal{B}$.

$$\begin{aligned} \mathbb{P}([a, b]) &= \mathbb{P}((-\infty, b]) - \mathbb{P}((-\infty, a)) \\ &= \mathbb{P}((-\infty, b]) - \mathbb{P}\left(\bigcup_{n=1}^{\infty} \left(-\infty, a - \frac{1}{n}\right]\right) \\ &= \mathbb{P}((-\infty, b]) - \lim_{n \to \infty} \mathbb{P}\left(\left(-\infty, a - \frac{1}{n}\right]\right) \end{aligned}$$

2.2 CDF of random variables

Every random variable has an associated distribution function.

Definition 2.4. The **(cumulative) distribution function** (CDF) of a random variable X is a function $F_X : \mathbb{R} \to [0,1]$ defined as:

$$F_X(x) = \mathbb{P}(X \leq x) := \mathbb{P} \circ X^{-1}((-\infty, x]).$$

Example 2.3. From Example 2.1,

$$\mathbb{P}(\omega) = \frac{1}{4}, \qquad F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0, & x < 0, \\ \frac{1}{4}, & 0 \leq x < 1, \\ \frac{3}{4}, & 1 \leq x < 2, \\ 1, & x \geq 2. \end{cases}$$

Lemma 2.5. The CDF F_X of a random variable X satisfies the following properties:

1. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.
2. If $x < y$, then $F_X(x) \leq F_X(y)$.
3. F_X is right-continuous ($F_X(x + h) \to F_X(x)$ as $h \downarrow 0$).

Proof.

1. Let $B_n = \{\omega \in \Omega : X(\omega) \leq -n\} = \{X \leq -n\}$. Since $B_1 \supseteq B_2 \supseteq \cdots$, by Lemma 1.10,

$$\lim_{x \to -\infty} F_X(x) = \mathbb{P}\left(\lim_{i \to \infty} B_i\right) = \mathbb{P}(\emptyset) = 0.$$

   Alternative proof:

$$\lim_{x \to -\infty} F_X(x) = \lim_{x \to -\infty} \mathbb{P} \circ X^{-1}((-\infty, x]) = \lim_{n \to \infty} \mathbb{P} \circ X^{-1}((-\infty, -n]) = \mathbb{P} \circ X^{-1}(\emptyset) = 0$$

   Let $C_n = \{\omega \in \Omega : X(\omega) \leq n\} = \{X \leq n\}$. Since $C_1 \subseteq C_2 \subseteq \cdots$, by Lemma 1.10,

$$\lim_{x \to \infty} F_X(x) = \mathbb{P}\left(\lim_{i \to \infty} C_i\right) = \mathbb{P}(\Omega) = 1.$$

   Alternative Proof:

$$\lim_{x \to \infty} F_X(x) = \lim_{x \to \infty} \mathbb{P} \circ X^{-1}((-\infty, x]) = \mathbb{P} \circ X^{-1}(\mathbb{R}) = 1.$$

2. Let $A(x) = \{X \leq x\}, A(x, y) = \{x < X \leq y\}$. Then $A(y) = A(x) \cup A(x, y)$ is a disjoint union.

$$F_X(y) = \mathbb{P}(A(y)) = \mathbb{P}(A(x)) + \mathbb{P}(A(x, y)) = F_X(x) + \mathbb{P}(x < X \leq y) \geq F_X(x)$$

3. Let $B_n = \{\omega \in \Omega : X(\omega) \leq x + \frac{1}{n}\}$. Since $B_1 \supseteq B_2 \supseteq \cdots$, by Lemma 1.10,

$$\lim_{h \downarrow 0} F_X(x + h) = \mathbb{P}\left(\bigcap_{i=1}^{\infty} B_i\right) = \mathbb{P}\left(\lim_{n \to \infty} B_n\right) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = F_X(x)$$

   Alternative Proof:

$$\lim_{h \downarrow 0} F_X(x + h) = \lim_{h \downarrow 0} \mathbb{P} \circ X^{-1}((-\infty, x + h]) = \lim_{n \to \infty} \mathbb{P} \circ X^{-1}\left(\left(-\infty, x + \frac{1}{n}\right]\right) = \mathbb{P} \circ X^{-1}((-\infty, x]) = F_X(x)$$

□

Remark 2.5.1. F is not left-continuous because:

$$\lim_{h \downarrow 0} F_X(x - h) = \lim_{n \to \infty} \mathbb{P} \circ X^{-1}\left(\left(-\infty, x - \frac{1}{n}\right)\right) = \mathbb{P} \circ X^{-1}((-\infty, x)) = F_X(x) - \mathbb{P} \circ X^{-1}(\{x\})$$

Lemma 2.6. Let F_X be the CDF of a random variable X. Then

1. $\mathbb{P}(X > x) = 1 - F_X(x)$.
2. $\mathbb{P}(x < X \leq y) = F_X(y) - F_X(x)$.

Proof.

1. $\mathbb{P}(X > x) = \mathbb{P}(\Omega \setminus \{X \leq x\}) = \mathbb{P}(\Omega) - \mathbb{P}(X \leq x) = 1 - F_X(x)$.
2. $\mathbb{P}(x < X \leq y) = \mathbb{P}(\{X \leq y\} \setminus \{X \leq x\}) = \mathbb{P}(X \leq y) - \mathbb{P}(X \leq x) = F_X(y) - F_X(x)$.

□

Example 2.4. (**Constant variables**) Let $X : \Omega \to \mathbb{R}$ be defined by $X(\omega) = c$ for all $\omega \in \Omega$. For all $B \in \mathcal{B}$,

$$F_X(x) = \mathbb{P} \circ X^{-1}(B) = \begin{cases} 0, & B \cap \{c\} = \emptyset \\ 1, & B \cap \{c\} = \{c\} \end{cases}$$

X is constant almost surely if there exists $c \in \mathbb{R}$ such that $\mathbb{P}(X = c) = 1$.

Example 2.5. (**Bernoulli variables**) Consider flipping coin once. Let $X : \Omega \to \mathbb{R}$ be defined by $X(H) = 1$ and $X(T) = 0$.

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1 - p, & 0 \leq x < 1 \\ 1, & x \geq 1 \end{cases}$$

X have **Bernoulli distribution**, denoted by $\text{Bern}(p)$.

Example 2.6. Let A be an event in $\mathcal{F}$ and **indicator functions** $\mathbf{1}_A : \Omega \to \mathbb{R}$ such that for all $B \in \mathcal{B}(\mathbb{R})$:

$$\mathbf{1}_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \in A^{\complement} \end{cases} \qquad \mathbf{1}_A^{-1}(B) = \begin{cases} \emptyset, & B \cap \{0,1\} = \emptyset \\ A^{\complement}, & B \cap \{0,1\} = \{0\} \\ A, & B \cap \{0,1\} = \{1\} \\ \Omega, & B \cap \{0,1\} = \{0,1\} \end{cases} \qquad \mathbb{P} \circ \mathbf{1}_A^{-1}(B) = \begin{cases} 0, & B \cap \{0,1\} = \emptyset \\ \mathbb{P}(A^{\complement}), & B \cap \{0,1\} = \{0\} \\ \mathbb{P}(A), & B \cap \{0,1\} = \{1\} \\ 1, & B \cap \{0,1\} = \{0,1\} \end{cases}$$

Then $\mathbf{1}_A$ is a Bernoulli random variable taking values 1 and 0 with probabilities $\mathbb{P}(A)$ and $\mathbb{P}(A^{\complement})$ respectively.

2.3 PMF / PDF of random variables

We can classify some random variables into either discrete or continuous. This two will be further discussed in the next two chapters.

Definition 2.7. Random variable X is **discrete** if it takes value in some countable subsets $\{x_1, x_2, \cdots\}$ only of $\mathbb{R}$.
Discrete random variable X has **probability mass function** (PMF) $f_X : \mathbb{R} \to [0,1]$ given by:

$$f_X(x) = \mathbb{P}(X = x) = \mathbb{P} \circ X^{-1}(\{x\})$$

Lemma 2.8. Relationship between PMF f_X and CDF F_X of a random variable X:

1. $F_X(x) = \sum_{i \leq x} f_X(i)$
2. $f_X(x) = F_X(x) - \lim_{y \uparrow x} F_X(y)$

Proof.

1. $$F_X(x) = \mathbb{P}(X \leq x) = \sum_{i : x_i \leq x} \mathbb{P}(X = x_i) = \sum_{y \leq x} f_X(y)$$

2. Let $B_n = \{x - \frac{1}{n} < X \leq x\}$. Since $B_1 \supseteq B_2 \supseteq \cdots$, by Lemma 1.10,

$$F_X(x) - \lim_{y \uparrow x} F_X(y) = \mathbb{P}\left(\bigcap_{i=1}^{\infty} B_i\right) = \mathbb{P}\left(\lim_{n \to \infty} B_n\right) = \mathbb{P}\left(\left\{\lim_{n \to \infty} \left(x - \frac{1}{n}\right) < X \leq x\right\}\right) = \mathbb{P}(X = x)$$

□

This is problematic when random variable X is continuous because using PMF will get the result of $f_X(x) = 0$ for all x. Therefore, we would need another definition for continuous random variable.

Definition 2.9. Random variable X is called **continuous** if its distribution function can be expressed as:

$$F_X(x) = \int_{-\infty}^{x} f(u)\,du \qquad\qquad x \in \mathbb{R}$$

for some integrable **probability density function** (PDF) $f_X : \mathbb{R} \to [0, \infty)$ of X.

Remark 2.9.1. For small $\delta > 0$:

$$\mathbb{P}(x < X \leq x + \delta) = F_X(x + \delta) - F_X(x) = \int_{x}^{x+\delta} f_X(u)\,du \approx f_X(x)\delta$$

Remark 2.9.2. On discrete random variable, the distribution is **atomic** because the distribution function has jump discontinuities at values $x_1, x_2, \cdots$ and is constant in between.

Remark 2.9.3. On continuous random variable, the CDF of a continuous variable is **absolutely continuous**.
Not every continuous function can be written as $\int_{-\infty}^{x} f_X(u)\,du$. E.g. Canton function

Remark 2.9.4. It is possible that a random variable is neither continuous nor discrete.

2.4 JCDF of random variables

How do we deal with cases when there are more than one random variables?

Definition 2.10. Let $X_1, X_2 : \Omega \to \mathbb{R}$ be random variables. We define **random vector** $\mathbf{X} = (X_1, X_2) : \Omega^2 \to \mathbb{R}^2$ with properties

$$\mathbf{X}^{-1}(D) = \{\omega \in \Omega : \mathbf{X}(\omega) = (X_1(\omega), X_2(\omega)) \in D\} \in \mathcal{F}$$

for all $D \in \mathcal{B}(\mathbb{R}^2)$.
We can also say $\mathbf{X} = (X_1, X_2)$ is a random vector if both $X_1, X_2 : \Omega \to \mathbb{R}$ are random variables. That means:

$$X_a^{-1}(B) \in \mathcal{F}$$

for all $B \in \mathcal{B}(\mathbb{R}), a = 1, 2$.

Claim 2.10.1. Both definitions of random vectors are equivalent.

Proof.
By first definition, $\mathbf{X}^{-1}(A_1 \times A_2) \in \mathcal{F}$. If we choose $A_2 = \mathbb{R}$,

$$\begin{aligned}\mathbf{X}^{-1}(A_1 \times \mathbb{R}) &= \{\omega \in \Omega : (X_1(\omega), X_2(\omega)) \in A_1 \times \mathbb{R}\} \\ &= \{\omega \in \Omega : X_1(\omega) \in A_1\} \cap \{\omega \in \Omega : X_2(\omega) \in \mathbb{R}\} \\ &= X_1^{-1}(A_1)\end{aligned}$$

This means X_1 is a random variable. Using similar method, we can also find that X_2 is a random variable.
Therefore, we can obtain the second definition from the first definition.
By second definition, X_1 and X_2 are random variables. Therefore,

$$\begin{aligned}\mathbf{X}^{-1}(A_1 \times A_2) &= \{\omega \in \Omega : (X_1(\omega), X_2(\omega)) \in A_1 \times A_2\} \\ &= \{\omega \in \Omega : X_1(\omega) \in A_1\} \cap \{\omega \in \Omega : X_2(\omega) \in A_2\} \\ &= X_1^{-1}(A_1) \cap X_2^{-1}(A_2) \in \mathcal{F}\end{aligned}$$

Therefore, we can obtain the first definition from the second definition.
Therefore, two definitions are equivalent. □

Remark 2.10.1. We can write $\mathbb{P} \circ \mathbf{X}^{-1}(D) = \mathbb{P}(\mathbf{X} \in D) = \mathbb{P}(\{\omega \in \Omega : \mathbf{X}(\omega) = (X_1(\omega), X_2(\omega)) \in D\})$.

Of course, there is a distribution function corresponding to the random vector.

Definition 2.11. Joint distribution function (JCDF) $F_{\mathbf{X}} : \mathbb{R}^2 \to [0,1]$ is defined as

$$F_{\mathbf{X}}(x_1, x_2) = F_{X_1,X_2}(x_1, x_2) = \mathbb{P} \circ \mathbf{X}^{-1}((-\infty, x_1] \times (-\infty, x_2]) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2)$$

Remark 2.11.1. We can replace all Borel sets by the form $[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$.

Joint distribution function has quite similar properties with normal distribution function.

Lemma 2.12. JCDF $F_{X,Y}$ of random vector (X, Y) has the following properties:

1. $\lim_{(x,y)\to(-\infty,-\infty)} F_{X,Y}(x, y) = 0$ and $\lim_{(x,y)\to(\infty,\infty)} F_{X,Y}(x, y) = 1$.
2. If $x_1 \leq y_1$ and $x_2 \leq y_2$, then $F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$.
3. $F_{X,Y}$ is continuous from above, in that $F_{X,Y}(x + u, y + v) \to F_{X,Y}(x, y)$ as $u, v \downarrow 0$.

We can find the probability distribution of one random variable by disregarding another variable. We get the following distribution.

Definition 2.13. Let X, Y be random variables. We can get a **marginal distribution** (marginal CDF) by having:

$$F_X(x) = \mathbb{P} \circ X^{-1}((-\infty, x]) = \mathbb{P}\left(X^{-1}((-\infty, x]) \cap Y^{-1}((-\infty, \infty))\right) = \lim_{y\uparrow\infty} \mathbb{P}\left(X^{-1}((-\infty, x]) \cap Y^{-1}((-\infty, y])\right) = \lim_{y\uparrow\infty} F_{X,Y}(x, y)$$

Joint distribution function also has its probability mass function and probability density function too.

Definition 2.14. Two random variables X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$ are **jointly discrete** if the vector (X, Y) takes values in some countable subset of $\mathbb{R}^2$ only. The corresponding **joint (probability) mass function** (JPMF) $f : \mathbb{R}^2 \to [0, 1]$ is given by

$$f_{X,Y}(x, y) = \mathbb{P}((X, Y) = (x, y)) = \mathbb{P} \circ (X, Y)^{-1}(\{x, y\}) \qquad F_{X,Y}(x, y) = \sum_{u\leq x} \sum_{v\leq y} f(u, v) \qquad x, y \in \mathbb{R}$$

Remark 2.14.1.

$$f_{X,Y}(x, y) = F_{X,Y}(x, y) - F_{X,Y}(x^-, y) - F_{X,Y}(x, y^-) + F_{X,Y}(x^-, y^-)$$

Remark 2.14.2. More generally, for all $B \in \mathcal{B}(\mathbb{R}^2)$,

$$\mathbb{P} \circ (X, Y)^{-1}(B) = \sum_{(u,v)\in B} f_{X,Y}(u, v)$$

Definition 2.15. Two random variables X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$ are **jointly continuous** if the **joint probability density function** (JPDF) $f : \mathbb{R}^2 \to [0, \infty)$ of (X, Y) can be expressed as:

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x\, \partial y} F_{X,Y}(x, y) \qquad F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v)\, du\, dv \qquad x, y \in \mathbb{R}$$

Remark 2.15.1. More generally, for all $B \in \mathcal{B}(\mathbb{R}^2)$,

$$\mathbb{P} \circ (X, Y)^{-1}(B) = \mathbb{P}((X, Y) \in B) = \iint_B f_{X,Y}(u, v)\, du\, dv$$

Example 2.7. Assume that a special three-sided coin is provided. Each toss results in head (H), tail (T) or edge (E) with equal probabilities. What is the probability of having h heads, t tails and e edges after n tosses?
Let H_n, T_n, E_n be the numbers of such outcomes in n tosses of the coin. The vector (H_n, T_n, E_n) satisfy $H_n + T_n + E_n = n$.

$$\mathbb{P}((H_n, T_n, E_n) = (h, t, e)) = \frac{n!}{h!t!e!} \left(\frac{1}{3}\right)^n$$

Remark 2.15.2. It is not generally true for two continuous random variables X and Y to be jointly continuous.

Example 2.8. Let X be uniformly distributed on $[0,1]$ ($f_X(x) = \mathbf{1}_{[0,1]}$). This means $f_X(x) = 1$ when $x \in [0,1]$ and 0 otherwise. Let $Y = X$ ($Y(\omega) = X(\omega)$ for all $\omega \in \Omega$). That means $(X,Y) = (X,X)$. Let $B = \{(x,y) : x = y \text{ and } x \in [0,1]\} \in \mathcal{B}(\mathbb{R}^2)$. Since $y = x$ is just a line,

$$\mathbb{P} \circ (X,Y)^{-1}(B) = 1$$

$$\iint_B f_{X,Y}(u,v)\,du\,dv = 0 \neq \mathbb{P} \circ (X,Y)^{-1}(B)$$

Therefore, X and Y are not jointly continuous.

Chapter 3

Discrete random variables

3.1 Introduction of discrete random variables

Let us revisit some key definitions related to discrete random variables from the previous chapter.

Definition 3.1. A random variable X is said to be **discrete** if it takes values in a countable subset $\{x_1, x_2, \cdots\}$ of $\mathbb{R}$.
The **(cumulative) distribution function** (CDF) of a discrete random variable X is the function $F_X : \mathbb{R} \to [0,1]$ defined as:

$$F_X(x) = \mathbb{P}(X \leq x).$$

The **probability mass function** (PMF) of a discrete random variable X is the function $f_X : \mathbb{R} \to [0,1]$ defined as:

$$f_X(x) = \mathbb{P}(X = x).$$

The CDF and PMF are related by the following equations:

$$F_X(x) = \sum_{i:x_i \leq x} f_X(x_i), \qquad f_X(x) = F_X(x) - \lim_{y \uparrow x} F_X(y).$$

Lemma 3.2. The PMF $f_X : \mathbb{R} \to [0,1]$ of a discrete random variable X satisfies the following properties:

1. The set of x values for which $f_X(x) \neq 0$ is countable.
2. $\sum_i f_X(x_i) = 1$, where $x_1, x_2, \cdots$ are the values of x such that $f_X(x) \neq 0$.

Next, we recall the definitions of joint distribution and joint mass functions.

Definition 3.3. For jointly discrete random variables X and Y, the **joint probability mass function** (JPMF) $f_{X,Y} : \mathbb{R}^2 \to [0,1]$ is defined as:

$$f_{X,Y}(x,y) = \mathbb{P}((X,Y) = (x,y)) = \mathbb{P} \circ (X,Y)^{-1}(\{x,y\}), \qquad F_{X,Y}(x,y) = \sum_{u \leq x} \sum_{v \leq y} f(u,v), \qquad x, y \in \mathbb{R}.$$

Recall that two events A and B are independent if the occurrence of A does not affect the probability of B occurring.

Definition 3.4. Discrete random variables X and Y are **independent** if the events $\{X = x\}$ and $\{Y = y\}$ are independent for all x, y. Equivalently, X and Y are independent if:

1. $\mathbb{P}((X,Y) \in A \times B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$ for all $A, B \in \mathcal{B}(\mathbb{R})$.
2. $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for all $x, y \in \mathbb{R}$.
3. $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$.

Claim 3.4.1. Three definitions are equivalent.

Proof.
We can get definition 2 from definition 1.

$$F_{X,Y}(x,y) = \mathbb{P}(X \le x, Y \le y) = \mathbb{P}(X \le x)\mathbb{P}(Y \le y) = F_X(x)F_Y(y)$$

We can get definition 3 from definition 2.

$$\begin{aligned} f_{X,Y}(x,y) &= F_{X,Y}(x,y) - F_{X,Y}(x^-,y) - F_{X,Y}(x,y^-) + F_{X,Y}(x^-,y^-) \\ &= F_X(x)F_Y(y) - F_X(x^-)F_Y(y) - F_X(x)F_Y(y^-) + F_X(x^-)F_Y(y^-) \\ &= (F_X(x) - F_X(x^-))(F_Y(y) - F_Y(y^-)) = f_X(x)f_Y(y) \end{aligned}$$

We can get definition 1 from definition 3.

$$\mathbb{P} \circ (X,Y)^{-1}(E \times F) = \sum_{(x,y)\in E\times F} f_{X,Y}(x,y) = \sum_{x\in E}\sum_{y\in F} f_X(x)f_Y(y) = (\mathbb{P}\circ X^{-1}(E))(\mathbb{P}\circ Y^{-1}(F))$$

Therefore, three definitions are equivalent. □

Remark 3.4.1. More generally, let $X_1, X_2, \cdots, X_n : \Omega \to \mathbb{R}$ be discrete random variables. They are **independent** if

1. For all $A_i \in \mathcal{B}(\mathbb{R})$,
$$\mathbb{P} \circ (X_1, X_2, \cdots, X_n)^{-1}(A_1 \times A_2 \times \cdots \times A_n) = \prod_{i=1}^{n} \mathbb{P} \circ X_i^{-1}(A_i)$$
2. For all $x_i \in \mathbb{R}$,
$$F_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$$
3. For all $x_i \in \mathbb{R}$,
$$f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

Recall that we say $A_1, A_2, \cdots, A_n$ are independent if for any $I \subseteq \{1, 2, \cdots, n\}$:

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

Remark 3.4.2. From the definition, we can see that $X \perp\!\!\!\perp Y$ means that $X^{-1}(E) \perp\!\!\!\perp Y^{-1}(F)$ for all $E, F \in \mathcal{B}(\mathbb{R})$.

Remark 3.4.3. We can generate σ-field using random variables by defining σ-field generated by random variable X

$$\sigma(X) = \{X^{-1}(E) : E \in \mathcal{B}(\mathbb{R})\} \subseteq \mathcal{F}$$

From the remarks, we can extend the definition of independence from random variables to σ-fields.

Definition 3.5. Let $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ be two σ-fields. We say $\mathcal{G}$ and $\mathcal{H}$ are **independent** if $A \perp\!\!\!\perp B$ for all $A \in \mathcal{G}, B \in \mathcal{H}$.

Remark 3.5.1. $\sigma(X) \perp\!\!\!\perp \sigma(Y) \iff X \perp\!\!\!\perp Y$

Theorem 3.6. Given two random variables X and Y. If $X \perp\!\!\!\perp Y$ and we have two functions $g, h : \mathbb{R} \to \mathbb{R}$ such that $g(X)$ and $h(Y)$ are still random variables, then $g(X) \perp\!\!\!\perp h(Y)$.

Proof.
For all $A, B \in \mathcal{B}$,

$$\begin{aligned} \mathbb{P}((g(X), h(Y)) \in A \times B) &= \mathbb{P}(g(X) \in A, h(Y) \in B) \\ &= \mathbb{P}(X \in \{x : g(x) \in A\}, Y \in \{y : h(y) \in B\}) \\ &= \mathbb{P}(X \in \{x : g(x) \in A\})\mathbb{P}(Y \in \{y : h(y) \in B\}) \\ &= \mathbb{P}(g(X) \in A)\mathbb{P}(h(Y) \in B) \end{aligned}$$

Therefore, $g(X) \perp\!\!\!\perp h(Y)$. □

Remark 3.6.1. We assume a product space $(\Omega, \mathcal{F}, \mathbb{P})$ of two probability space $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$.
$(\Omega = \Omega_1 \times \Omega_2,\ \mathcal{F} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2),\ \mathbb{P}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2))$.
Any pair of events of the form $E_1 \times \Omega_2$ and $\Omega_1 \times E_2$ are independent.

$$\mathbb{P}((E_1 \times \Omega_2) \cap (\Omega_1 \times E_2)) = \mathbb{P}(E_1 \times E_2) = \mathbb{P}_1(E_1)\mathbb{P}_2(E_2) = \mathbb{P}(E_1 \times \Omega_2)\mathbb{P}(\Omega_1 \times E_2)$$

We have some important examples of random variables that have wide number of applications.

Example 3.1. (**Bernoulli random variable**) $X \sim \text{Bern}(p)$
Let $A \in \mathcal{F}$ be a specific event. A Bernoulli trial is considered a success if A occurs. Let $X : \Omega \to \mathbb{R}$ be such that

$$X(\omega) = \mathbf{1}_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \in A^{\complement} \end{cases} \qquad \mathbb{P}(A) = \mathbb{P}(X = 1) = p \qquad \mathbb{P}(A^{\complement}) = \mathbb{P}(X = 0) = 1 - p$$

Example 3.2. (**Binomial distribution**) $Y \sim \text{Bin}(n, p)$
Suppose we perform n independent Bernoulli trials $X_1, X_2, \cdots, X_n$. Let $Y = X_1 + X_2 + \cdots + X_n$ be total number of successes.

$$f_Y(k) = \mathbb{P}(Y = k) = \mathbb{P}\left(\sum_{i=1}^{k} X_i = k\right) = \mathbb{P}(\{\#\{i : X_i = 1\} = k\})$$

We denote $A = \{\#\{i : X_i = 1\} = k\} = \bigcup_\sigma A_\sigma$ where $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_n)$ can be any sequence satisfying $\#\{i : \sigma_i = 1\} = k$ and $A_\sigma :=$ events that $(X_1, X_2, \cdots, X_n) = (\sigma_1, \sigma_2, \cdots, \sigma_n)$. Events A_σ are mutually exclusive. Hence $\mathbb{P}(A) = \sum_\sigma \mathbb{P}(A_\sigma)$.
There are totally $\binom{n}{k}$ different σ's in the sum. By independence, we have

$$\mathbb{P}(A_\sigma) = \mathbb{P}(X_1 = \sigma_1, X_2 = \sigma_2, \cdots, X_n = \sigma_n) = \mathbb{P}(X_1 = \sigma_1)\mathbb{P}(X_2 = \sigma_2)\cdots\mathbb{P}(X_n = \sigma_n) = p^k(1-p)^{n-k}$$

Hence, $f_Y(k) = \mathbb{P}(A) = \binom{n}{k}p^k(1-p)^{n-k}$.

Example 3.3. (**Trinomial distribution**) Suppose we perform n trials, each of which result in three outcomes A, B and C, where A occurs with probability p, B with probability q, and C with probability $1 - p - q$. Probability of r A's, w B's, and $n - r - w$ C's is

$$\mathbb{P}(\#A = r, \#B = w, \#C = n - r - w) = \frac{n!}{r!w!(n-r-w)!}p^r q^w (1-p-q)^{n-r-w}$$

Example 3.4. (**Geometric distribution**) $W \sim \text{Geom}(p)$
Suppose we keep performing independent Bernoulli trials until the first success shows up. Let p be the probability of success and W be the **waiting time** which elapses before first success.

$$\mathbb{P}(W > k) = (1-p)^k \qquad \mathbb{P}(W = k) = \mathbb{P}(W > k-1) - \mathbb{P}(W > k) = p(1-p)^{k-1}$$

Example 3.5. (**Negative binomial distribution**) $W_r \sim \text{NBin}(r, p)$
Similar with examples of geometric distribution, let W_r be the waiting time for the r-th success. For $k \geq r$,

$$f_{W_r}(k) = \mathbb{P}(W_r = k) = \binom{k-1}{r-1}p^r(1-p)^{k-r}$$

Remark 3.6.2. W_r is the sum of r independent geometric variables.

Example 3.6. (**Poisson distribution**) $X \sim \text{Poisson}(\lambda)$
Poisson variable is a discrete random variable with Poisson PMF:

$$f_X(k) = \frac{\lambda^k}{k!}e^{-\lambda} \qquad k = 0, 1, 2, \cdots$$

for some parameter $\lambda > 0$.
This is used for approximation of binomial random variable $\text{Bin}(n, p)$ when n is large, p is small and np is moderate.
Let $X \sim \text{Bin}(n, p)$ and $\lambda = np$.

$$\mathbb{P}(X = k) = \binom{n}{k}p^k(1-p)^{n-k} = \frac{n!}{(n-k)!k!}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!}\left(\frac{n!}{n^k(n-k)!}\right)\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^k} \approx \frac{\lambda^k}{k!}(1)\left(\frac{e^{-\lambda}}{1}\right) = \frac{\lambda^k}{k!}e^{-\lambda}$$

We have an interesting example concerning independence with Poisson distribution involved.

Example 3.7. (**Poisson flips**) A coin is tossed once and head turns up with probability p.
Let random variables X and Y be the numbers of heads and tails respectively. X and Y are not independent since

$$\mathbb{P}(X=1,Y=1)=0 \qquad\qquad \mathbb{P}(X=1)\mathbb{P}(Y=1)=p(1-p)\neq 0$$

Suppose now that the coin is tosses N times, where N has the Poisson distribution with parameter λ.
In this case, random variables X and Y are independent since

$$\begin{aligned}
\mathbb{P}(X=x,Y=y) &= \mathbb{P}(X=x,Y=y|N=x+y)\mathbb{P}(N=x+y)\\
&= \binom{x+y}{x}p^x(1-p)^y\frac{\lambda^{x+y}}{(x+y)!}e^{-\lambda}\\
&= \frac{(\lambda p)^x(\lambda(1-p))^y}{x!y!}e^{-\lambda}\\
\mathbb{P}(X=x)\mathbb{P}(Y=y) &= \sum_{i\geq x}\mathbb{P}(X=x|N=i)\mathbb{P}(N=i)\sum_{j\geq y}\mathbb{P}(Y=y|N=j)\mathbb{P}(N=j)\\
&= \sum_{i\geq x}\binom{i}{x}p^x(1-p)^{i-x}\frac{\lambda^i}{i!}e^{-\lambda}\sum_{j\geq y}\binom{j}{y}p^{j-y}(1-p)^y\frac{\lambda^j}{j!}e^{-\lambda}\\
&= \frac{(\lambda p)^x}{x!}e^{-\lambda}\left(\sum_{i\geq x}\frac{(\lambda(1-p))^{i-x}}{(i-x)!}\right)\frac{(\lambda(1-p))^y}{y!}e^{-\lambda}\left(\sum_{j\geq y}\frac{(\lambda p)^{j-y}}{(j-y)!}\right)\\
&= \frac{(\lambda p)^x}{x!}e^{-\lambda+\lambda(1-p)}\frac{(\lambda(1-p))^y}{y!}e^{-\lambda+\lambda p}\\
&= \frac{(\lambda p)^x(\lambda(1-p))^y}{x!y!}e^{-\lambda}=\mathbb{P}(X=x,Y=y)
\end{aligned}$$

3.2 Expectation of discrete random variables

In real-world scenarios, we often want to determine the expected outcome based on calculated probabilities.
The expected result is typically a theoretical approximation of the empirical average.
Assume we have random variables $X_1, X_2, \cdots, X_N$ that take values in $\{x_1, x_2, \cdots, x_n\}$ with a probability mass function $f_X(x)$.
The empirical average is given by:

$$\mu=\frac{1}{N}\sum_{i=1}^{N}X_i\approx\frac{1}{N}\sum_{i=1}^{n}x_iNf(x_i)=\sum_{i=1}^{N}x_if(x_i).$$

Definition 3.7. Suppose we have a discrete random variable X taking values from $\{x_1, x_2, \cdots\}$ with PMF $f_X(x)$. The **mean value**, **expectation**, or **expected value** of X is defined as:

$$\mathbb{E}X=\mathbb{E}(X):=\sum_i x_if_X(x_i)=\sum_{x:f_X(x)>0}xf_X(x),$$

whenever this sum is absolutely convergent. Otherwise, we say $\mathbb{E}X$ does not exist.

Example 3.8. Suppose a product is sold seasonally. Let b represent the net profit per sold unit, ℓ the net loss per unsold unit, and X the number of products ordered by customers. If y units are stocked, the expected profit $Q(y)$ is given by:

$$Q(y)=\begin{cases}bX-(y-X)\ell, & X\leq y,\\ yb, & X>y.\end{cases}$$

Lemma 3.8. If discrete random variable X has a PMF f_X and $g : \mathbb{R} \to \mathbb{R}$ such that $g(X)$ is still a discrete random variable, then

$$\mathbb{E}(g(X)) = \sum_x g(x) f_X(x)$$

whenever this sum is absolutely convergent.

Proof.
Denote by $Y := g(X)$.

$$\begin{aligned}\sum_x g(x)f_X(x) = \sum_y \sum_{x:g(x)=y} g(x)f_X(x) = \sum_y y\left(\sum_{x:g(x)=y} f_X(x)\right) &= \sum_y y\left(\sum_{x:g(x)=y} \{\omega \in \Omega : X(\omega) = x\}\right) \\ &= \sum_y y\mathbb{P}(\{\omega \in \Omega : g(X(\omega)) = y\}) \\ &= \sum_y y\mathbb{P}(\{\omega \in \Omega : Y(\omega) = y\}) \\ &= \sum_y y f_Y(y) = \mathbb{E}Y = \mathbb{E}g(X)\end{aligned}$$

□

Lemma 3.9. Let (X, Y) be a discrete random vector with JPMF $f_{X,Y}(x, y)$. Let $g : \mathbb{R}^2 \to \mathbb{R}$ such that $g(X, Y)$ is a discrete random variable. Then

$$\mathbb{E}g(X, Y) = \sum_{x,y} g(x, y) f_{X,Y}(x, y)$$

Proof.
Denote by $Z := g(X, Y)$.

$$\begin{aligned}\sum_{x,y} g(x,y)f_{X,Y}(x,y) = \sum_z \sum_{x,y:g(x,y)=z} g(x,y)f_{X,Y}(x,y) = \sum_z z\left(\sum_{x,y:g(x,y)=z} f_{X,Y}(x,y)\right) &= \sum_z z\left(\sum_{x,y:g(x,y)=z} \mathbb{P}((X,Y)=(x,y))\right) \\ &= \sum_z z\mathbb{P}(\{\omega \in \Omega : g(X,Y)(\omega) = z\}) \\ &= \sum_z z\mathbb{P}(\{\omega \in \Omega : Z(\omega) = z\}) = \sum_z z f_Z(z) = \mathbb{E}Z = \mathbb{E}g(X,Y)\end{aligned}$$

□

The lemmas have provided a method to calculate the moments of a discrete distribution. Most of the time, we only care about the expectation and variance.

Definition 3.10. Let $k \in \mathbb{N}_+$. We have a special term for each of the following expectations:

1. The **k-th moment** m_k of X is defined to be $m_k = \mathbb{E}(X^k)$.
2. The **k-th central moment** α_k is $\alpha_k = \mathbb{E}((X - \mathbb{E}X)^k) = \mathbb{E}((X - m_1)^k)$.
3. **Mean** of X is the 1st moment $m_1 = \mathbb{E}(X)$ and is denoted by μ.
4. **Variance** of X is the 2nd central moment $\alpha_2 = \mathrm{Var}(X) = \mathbb{E}((X - m_1)^2) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}(X^2) - \mu^2$.
5. **Standard deviation** of X is defined as $\sqrt{\mathrm{Var}(X)}$ and is denoted by σ.

Remark 3.10.1. Not all random variables have k-th moments for all $k \in \mathbb{N}_+$.

Remark 3.10.2. We cannot use collection of moments to uniquely determine a distribution that has k-th moments for all $k \in \mathbb{N}$.

We have the expectation and the variance of following distribution.

Example 3.9.

Bernoulli :	$\mathbb{E}X = p$	$\mathrm{Var}(X) = p(1-p)$
Binomial :	$\mathbb{E}X = np$	$\mathrm{Var}(X) = np(1-p)$
Geometric :	$\mathbb{E}X = p^{-1}$	$\mathrm{Var}(X) = (1-p)p^{-2}$
Poisson :	$\mathbb{E}X = \lambda$	$\mathrm{Var}(X) = \lambda$

Theorem 3.11. Expectation operator $\mathbb{E}$ has the following properties:

1. If $X \geq 0$, then $\mathbb{E}X \geq 0$.
2. If $a, b \in \mathbb{R}$, then $\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$.
3. The random variable $\mathbf{1}$, taking the value 1 always, has expectation $\mathbb{E}(1) = 1$.

Proof.

1. Since $f_X(x) \geq 0$ for all x, $\mathbb{E}X = \sum_x x f_X(x) \geq 0$ if $X \geq 0$.
2. Let $g(X, Y) = aX + bY$. Then,
$$\begin{aligned}\mathbb{E}(aX + bY) = \sum_{x,y}(ax + by)f_{X,Y}(x, y) &= a\sum_x x\left(\sum_y f_{X,Y}(x, y)\right) + b\sum_y y\left(\sum_x f_{X,Y}(x, y)\right) \\ &= a\sum_x x f_X(x) + b\sum_y y f_Y(y) = a\mathbb{E}X + b\mathbb{E}Y\end{aligned}$$
3. $\mathbb{E}(1) = 1(1) = 1$.

□

Remark 3.11.1. More generally, we have
$$\mathbb{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mathbb{E}X_i$$

Example 3.10. Assume we have N different types of card and each time one gets a card to be any one of the N types. Each types is equally likely to be gotten. What is the expected number of types of card we can get if we get n cards?
Let $X = X_1 + X_2 + \cdots + X_N$ where $X_i = 1$ if at least one type i card is among the n cards and otherwise 0.

$$\mathbb{E}X_i = \mathbb{P}(X_i = 1) = 1 - \left(\frac{N-1}{N}\right)^n$$
$$\mathbb{E}X = \sum_{i=1}^N \mathbb{E}X_i = N\left(1 - \left(\frac{N-1}{N}\right)^n\right)$$

What is the expected number of cards one needs to collect in order to get all N types?
Let $Y = Y_0 + Y_1 + \cdots + Y_{N-1}$ where Y_i is the number of additional cards we need to get in order to get a new type after having i distinct types.

$$\mathbb{P}(Y_i = k) = \left(\frac{i}{N}\right)^{k-1}\frac{N-i}{N} \qquad (Y_i \sim \mathrm{Geom}\left(\tfrac{N-i}{N}\right))$$
$$\mathbb{E}Y_i = \frac{N}{N-i}$$
$$\mathbb{E}Y = \sum_{i=0}^{N-1}\mathbb{E}Y_i = N\left(\frac{1}{N} + \frac{1}{N-1} + \cdots + 1\right)$$

Lemma 3.12. If two discrete random variables X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y$.

Proof.

$$\mathbb{E}(XY) = \sum_{x,y} xy f_{X,Y}(x, y) = \sum_{x,y} xy f_X(x) f_Y(y) = \sum_x x f_X(x)\sum_y y f_Y(y) = \mathbb{E}X\mathbb{E}Y$$

□

Lemma 3.13. Given two discrete random variables X and Y. Let $g, h : \mathbb{R} \to \mathbb{R}$ such that $g(X), h(Y)$ are still discrete random variables. If $X \perp\!\!\!\perp Y$ and $\mathbb{E}(g(X)h(Y)), \mathbb{E}g(X)$ and $\mathbb{E}h(Y)$ exist, then $\mathbb{E}(g(X)h(Y)) = \mathbb{E}g(X)\mathbb{E}h(Y)$.

Proof.

$$\mathbb{E}(g(X)h(Y)) = \sum_{x,y} g(x)h(y)f_{X,Y}(x,y) = \sum_{x,y} g(x)h(y)f_X(x)f_Y(y) = \sum_x g(x)f_X(x)\sum_y h(y)f_Y(y) = \mathbb{E}g(X)\mathbb{E}h(Y)$$

□

We can now say that two independent random variables are uncorrelated when they are independent.

Definition 3.14. Random variables X and Y are **uncorrelated** if $\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y$.

Remark 3.14.1. If X and Y are independent, then they are uncorrelated. The converse is generally not true.

Example 3.11. Let X be such that $f_X(0) = f_X(1) = f_X(-1) = \frac{1}{3}$ and Y be such that $Y = 0$ if $X \neq 0$ and $Y = 1$ if $X = 0$.

$$\mathbb{E}(XY) = 0 \qquad \mathbb{E}X = 0 = \mathbb{E}(XY)$$

However,

$$\mathbb{P}(X = 0, Y = 0) = 0 \qquad \mathbb{P}(X = 0) \neq 0 \qquad \mathbb{P}(Y = 0) \neq 0 \qquad \mathbb{P}(X = 0)\mathbb{P}(Y = 0) \neq 0$$

Therefore, X and Y are uncorrelated, but they are not independent.

We can now use the properties of expectations to deduce the properties of variance.

Theorem 3.15. For random variables X and Y,

1. $\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X)$ for $a \in \mathbb{R}$.
2. $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$ if X and Y are uncorrelated.

Proof.

1. Using linearity of $\mathbb{E}$,

$$\operatorname{Var}(aX + b) = \mathbb{E}((aX + b - \mathbb{E}(aX + b))^2) = \mathbb{E}(a^2(X - \mathbb{E}X)^2) = a^2\mathbb{E}((X - \mathbb{E}X)^2) = a^2 \operatorname{Var}(X)$$

2. When X and Y are uncorrelated,

$$\begin{aligned}\operatorname{Var}(X + Y) &= \mathbb{E}((X + Y - \mathbb{E}(X + Y))^2) \\ &= \mathbb{E}((X - \mathbb{E}X)^2 + 2(XY - \mathbb{E}X\mathbb{E}Y) + (Y - \mathbb{E}Y)^2) \\ &= \operatorname{Var}(X) + 2(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)) + \operatorname{Var}(Y) \\ &= \operatorname{Var}(X) + \operatorname{Var}(Y)\end{aligned}$$

□

Definition 3.16. Covariance of two random variables X and Y is:

$$\operatorname{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y$$

Remark 3.16.1.

$$\operatorname{Var}(X) = \operatorname{cov}(X, X)$$

Remark 3.16.2. In general, for any random variables $X_1, X_2, \cdots, X_n$,

$$\operatorname{Var}(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^n \operatorname{Var}(X_i) + 2\sum_{i<j}(\mathbb{E}(X_iX_j) - \mathbb{E}X_i\mathbb{E}X_j) = \sum_{i=1}^n \operatorname{Var}(X_i) + 2\sum_{i<j}\operatorname{cov}(X_i, X_j)$$

Remark 3.16.3. If X_i are (pairwise) independent or uncorrelated, we can get that $\operatorname{cov}(X_i, X_j) = 0$ for all $i \neq j$.

Example 3.12. If X_i are independent and $\mathrm{Var}(X_i) = 1$ for all i, then:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = n$$

If $X_i = X$ for all i and $\mathrm{Var}(X) = 1$, then:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathrm{Var}(nX) = n^2$$

3.3 Conditional distribution of discrete random variables

In the first chapter, we have discussed the conditional probability $\mathbb{P}(B|A)$. We can use this to define a distribution function.

Definition 3.17. Suppose $X, Y : \Omega \to \mathbb{R}$ are two discrete random variables. **Conditional distribution** of Y given $X = x$ for any x such that $\mathbb{P}(X = x) > 0$ is defined by

$$\mathbb{P}(Y \in \cdot | X = x)$$

Conditional distribution function (Conditional CDF) of Y given $X = x$ for any x such that $\mathbb{P}(X = x) > 0$ is defined by

$$F_{Y|X}(y|x) = \mathbb{P}(Y \leq y | X = x)$$

Conditional mass function (Conditional PMF) of Y given $X = x$ or any x such that $\mathbb{P}(X = x) > 0$ is defined by

$$f_{Y|X}(y|x) = \mathbb{P}(Y = y | X = x)$$

Remark 3.17.1. By definition,

$$f_{Y|X}(y|x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)} = \frac{\mathbb{P}(Y = y, X = x)}{\sum_v \mathbb{P}((X, Y) = (x, v))}$$

.

Remark 3.17.2. For any $x \in \mathbb{R}$, the conditional PMF $f_{Y|X}(y|x)$ is a probability mass function in y.

Remark 3.17.3. If X and Y are independent, then $f_{Y|X}(y|x) = f_Y(y)$.

Conditional distributions still have properties of original distribution.

Lemma 3.18. Given two discrete random variables X and Y. Conditional distributions have following properties:

1. $F_{Y|X}(y|x) = \sum_{v \leq y} f_{Y|X}(v|x)$
2. $f_{Y|X}(y|x) = F_{Y|X}(y|x) - F_{Y|X}(y^-|x)$

Proof.

1. $$\sum_{v \leq y} f_{Y|X}(v|x) = \sum_{v \leq y} \mathbb{P}(Y = v | X = x) = \mathbb{P}(Y \leq y | X = x) = F_{Y|X}(y|x)$$
2. This is just Lemma 2.8.

□

Definition 3.19. Given two discrete random variables X and Y. **Conditional expectation** ψ of Y given $X = x$ for any x is defined by:

$$\psi(x) = \mathbb{E}(Y|X = x) = \sum_y y f_{Y|X}(y|x)$$

Conditional expectation ψ of Y given X is defined by:

$$\psi(X) = \mathbb{E}(Y|X)$$

Example 3.13. Assume we roll a fair dice.

$$\Omega = \{1, 2, \cdots, 6\} \qquad Y(\omega) = \omega \qquad X(\omega) = \begin{cases} 1, & \omega \in \{2, 4, 6\} \\ 0, & \omega \in \{1, 3, 5\} \end{cases}$$

We try to guess Y. If we do not have any information about X,

$$\mathbb{E}Y = \operatorname*{argmin}_e(\mathbb{E}((Y - e)^2)) = 3.5$$

If we know that $X = x$, in which we have two cases: $X = 1$ and $X = 0$

$$f_{Y|X}(y|1) = \frac{\mathbb{P}(X = 1, Y = y)}{\mathbb{P}(X = 1)} = \begin{cases} \frac{1}{3}, & y = 2, 4, 6 \\ 0, & y = 1, 3, 5 \end{cases} \qquad f_{Y|X}(y|0) = \frac{\mathbb{P}(X = 0, Y = y)}{\mathbb{P}(X = 0)} = \begin{cases} 0, & y = 2, 4, 6 \\ \frac{1}{3}, & y = 1, 3, 5 \end{cases}$$

$$\mathbb{E}(Y|X = 1) = \sum_y y f_{Y|X}(y|1) = \frac{2 + 4 + 6}{3} = 4 \qquad \mathbb{E}(Y|X = 0) = \frac{1 + 3 + 5}{3} = 3$$

Finally, if we want to guess Y based on the future information of X,

$$\psi(X) = \mathbb{E}(Y|X) = 4(\mathbf{1}_{X=1}) + 3(\mathbf{1}_{X=0})$$

Example 3.14. If $Y = X$, then $\psi(X) = \mathbb{E}(Y|X) = \mathbb{E}(X|X) = X$.

Example 3.15. If $Y \perp\!\!\!\perp X$, then $\psi(X) = \mathbb{E}Y$.

In fact, we can extend the definition of conditional expectation into σ-field.

Definition 3.20. Given a random variable Y and a σ-field $\mathcal{H} \subseteq \mathcal{F}$.
$\mathbb{E}(Y|\mathcal{H})$ is any random variable Z satisfying the following two properties:

1. Z is $\mathcal{H}$-measurable. ($Z^{-1}(B) \in \mathcal{H}$ for all $B \in \mathcal{B}(\mathbb{R})$)
2. $\mathbb{E}(Y\mathbf{1}_A) = \mathbb{E}(Z\mathbf{1}_A)$ for all $A \in \mathcal{H}$.

Remark 3.20.1. Under this definition,

$$\mathbb{E}(Y|X) = \mathbb{E}(Y|\sigma(X))$$

Theorem 3.21. (Law of total expectation) Given two discrete random variables X and Y. Conditional expectation $\psi(X) = \mathbb{E}(Y|X)$ satisfies:

$$\mathbb{E}(\psi(X)) = \mathbb{E}(Y)$$

Proof.
By Lemma 3.8,

$$\mathbb{E}(\psi(X)) = \sum_x \psi(x) f_X(x) = \sum_{x,y} y f_{Y|X}(y|x) f_X(x) = \sum_{x,y} y f_{X,Y}(x, y) = \sum_y y f_Y(y) = \mathbb{E}(Y)$$

□

Example 3.16. A miner is trapped in a mine with doors, each will lead to a tunnel. Tunnel 1 will help the miner reach safety after 3 hours respectively. However, tunnel 2 and 3 will send the miner back after 5 and 7 hours respectively.
What is the expected amount of time the miner need to reach safety? (Assume that the miner is memoryless)
Let X be the amount of time to reach safety, Y be the door number he chooses for the first time.

$$\mathbb{E}X = \mathbb{E}(\mathbb{E}(X|Y)) = \sum_{k=1}^{3} \mathbb{E}(X|Y=k)\mathbb{P}(Y=k) = 3\left(\frac{1}{3}\right) + (\mathbb{E}X+5)\left(\frac{1}{3}\right) + (\mathbb{E}X+7)\left(\frac{1}{3}\right)$$
$$\mathbb{E}X = 15$$

What is the expected amount of time the miner needed to reach safety after he chose the second door and sent back?
Let $\widetilde{X}$ be the time for the miner to reach safety after the first round.

$$\mathbb{E}(X|Y=2) = \sum_x x f_{X|Y}(x|2) = \sum_x x\frac{\mathbb{P}(X=x,Y=2)}{\mathbb{P}(Y=2)} = \sum_x x\frac{\mathbb{P}(\widetilde{X}=x-5,Y=2)}{\mathbb{P}(Y=2)} = \sum_{\widetilde{x}} (\widetilde{x}+5)\mathbb{P}(\widetilde{X}=\widetilde{x}) = \mathbb{E}X+5$$

Example 3.17. We consider a sum of random number of random variables.
Let N be the number of customers and X_i be the amount of money spent by the i-th customers.
Assume that N and X_i's are all independent and $\mathbb{E}X_i = \mathbb{E}X$, what is the expected total amount of money spent by all N customers?

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{N} X_i\right) &= \mathbb{E}\left(\mathbb{E}\left(\sum_{i=1}^{N} X_i \middle| N\right)\right) \\
&= \sum_{n=0}^{\infty} \mathbb{E}\left(\sum_{i=1}^{N} X_i \middle| N=n\right)\mathbb{P}(N=n) \\
&= \sum_{n=0}^{\infty}\sum_y y\left(\frac{\mathbb{P}\left(\sum_{i=1}^{N} X_i = y, N=n\right)}{\mathbb{P}(N=n)}\right)\mathbb{P}(N=n) \\
&= \sum_{n=0}^{\infty}\sum_y y\mathbb{P}\left(\sum_{i=1}^{n} X_i = y\right)\mathbb{P}(N=n) \\
&= \sum_{n=0}^{\infty} \mathbb{E}\left(\sum_{i=1}^{n} X_i\right)\mathbb{P}(N=n) \\
&= \sum_{n=0}^{\infty} n\mathbb{E}X\mathbb{P}(N=n) = \mathbb{E}N\mathbb{E}X
\end{aligned}$$

The following theorem is the generalization of Law of total expectation.

Theorem 3.22. Given two discrete random variables X and Y. Conditional expectation $\psi(X) = \mathbb{E}(Y|X)$ satisfies:

$$\mathbb{E}(\psi(X)g(X)) = \mathbb{E}(Yg(X))$$

for any function g for which both expectations exist.

Proof.
By Lemma 3.8,

$$\mathbb{E}(\psi(X)g(X)) = \sum_x \psi(x)g(x)f_X(x) = \sum_{x,y} y f_{Y|X}(y|x)g(x)f_X(x) = \sum_{x,y} y f_{X,Y}(x,y)g(x) = \mathbb{E}(Yg(X))$$

□

3.4 Convolution of discrete random variables

Finally, a lot of times, we consider the sum of the two variables. For example, the number of heads in n tosses of a coin. However, there are situations that are more complicated, especially when the summands are dependent. We tries to find a formula for describing the mass function of the sum $Z = X + Y$.

Theorem 3.23. Given two jointly discrete random variables X and Y. The probability of sum of two random variables is given by:

$$\mathbb{P}(X + Y = z) = \sum_x f_{X,Y}(x, z - x) = \sum_y f_{X,Y}(z - y, y)$$

Proof.
We have the disjoint union:

$$\{X + Y = z\} = \bigcup_x (\{X = x\} \cap \{Y = z - x\})$$

At most countably many of its contributions have non-zero probability. Therefore,

$$\mathbb{P}(X + Y = z) = \sum_x \mathbb{P}(X = x, Y = z - x) = \sum_x f(x, z - x)$$

□

Definition 3.24. Convolution f_{X+Y} $(f_X * f_Y)$ of PMFs of two independent discrete random variables X and Y is the PMF of $X + Y$:

$$f_{X+Y}(z) = \mathbb{P}(X + Y = z) = \sum_x f_X(x) f_Y(z - x) = \sum_y f_X(z - y) f_Y(y)$$

There is an important example that has a wide range of applications in real life. However, we will not discuss this here. You can find the example in Appendix A.

Chapter 4

Continuous random variables

4.1 Introduction to Continuous Random Variables

We begin by recalling the definition of continuous random variables.

Definition 4.1. A random variable X is **continuous** if its **cumulative distribution function** (CDF) $F_X(x)$ can be expressed as:

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(u)\,du$$

for some integrable probability density function (PDF) $f_X : \mathbb{R} \to [0, \infty)$.

Remark 4.1.1. The PDF f_X is not uniquely defined, as two integrable functions that differ only on a set of measure zero yield the same integral. However, if F_X is **differentiable** at u, we define $f_X(u) = F_X'(u)$.

Note that we use the same notation f for both mass functions and density functions, as they serve analogous purposes.

Remark 4.1.2. The value $f_X(x)$ is not a probability. However, $f_X(x)\,dx = \mathbb{P}(x < X \leq x + dx)$ can be interpreted as an infinitesimal probability element.

Lemma 4.2. If a continuous random variable X has a density function f_X, then:

1. $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.
2. $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$.
3. $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\,dx$.

Proof.

1. $$\int_{-\infty}^{\infty} f_X(x)\,dx = \lim_{x\to\infty} F_X(x) = 1.$$

2. $$\mathbb{P}(X = x) = \lim_{h\to 0} \int_{x-h}^{x} f_X(x)\,dx = F_X(x) - \lim_{h\to\infty} F(x-h) = F_X(x) - F_X(x) = 0.$$

3. $$\mathbb{P}(a \leq X \leq b) = F(b) - F(a) = \int_{-\infty}^{b} f_X(x)\,dx - \int_{-\infty}^{a} f_X(x)\,dx = \int_a^b f_X(x)\,dx.$$

□

Remark 4.2.1. More generally, for an interval B, we have:

$$\mathbb{P}(X \in B) = \int_B f_X(x)\,dx.$$

We also recall the definition of independence. This definition also works for continuous random variables.

Definition 4.3. Two continuous random variables X and Y are called **independent** if for all $x, y \in \mathbb{R}$,

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

Theorem 4.4. Let two continuous random variables X and Y be independent. Suppose $g(X)$ and $h(Y)$ are still continuous random variables, then $g(X)$ and $h(Y)$ are independent.

4.2 Expectation of continuous random variables

In a continuous random variable X, the probability in every single point x is 0. Therefore, in order to make sense of the expectation of continuous random variable, we naturally give the following definition.

Definition 4.5. Expectation of a continuous random variable X with density function f is given by:

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x)\,dx$$

whenever this integral exists.

Remark 4.5.1. We usually can define $\mathbb{E}X$ only if $\mathbb{E}|X|$ exists.

We have a special properties in the continuous random variable.

Lemma 4.6. (Tail sum formula) If continuous random variable X has a PDF f_X with $f_X(x) = 0$ when $x < 0$, and a CDF F_X, then

$$\mathbb{E}X = \int_0^{\infty} (1 - F_X(x))\,dx$$

Proof.

$$\int_0^{\infty} (1 - F_X(x))\,dx = \int_0^{\infty} \mathbb{P}(X > x)\,dx = \int_0^{\infty}\int_x^{\infty} f_X(y)\,dy\,dx = \int_0^{\infty}\int_0^{y} f_X(y)\,dx\,dy = \int_0^{\infty} y f_X(y)\,dy = \mathbb{E}X$$

□

The following lemma is a formula I developed just for proving the next theorem.

Lemma 4.7. If continuous random variable X has a PDF f_X with $f_X(x) = 0$ when $x > 0$, and a CDF F_X, then

$$\mathbb{E}X = \int_{-\infty}^{0} -F_X(x)\,dx$$

Proof.

$$\int_{-\infty}^{0} -F_X(x)\,dx = \int_{-\infty}^{0}\int_{-\infty}^{x} -f_X(y)\,dy\,dx = \int_{-\infty}^{0}\int_{y}^{0} -f_X(y)\,dx\,dy = \int_{-\infty}^{0} y f_X(y)\,dy = \mathbb{E}X$$

□

Similar to discrete random variable, we can ask what is $\mathbb{E}g(X)$ for a function g.

Theorem 4.8. If X and $g(X)$ are continuous random variables, then

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$$

Proof.
We first consider that $g(x) \geq 0$ for all x. Let $Y = g(X)$ and $B = \{x : g(x) > y\}$. By Lemma 4.6,

$$\mathbb{E}(g(X)) = \int_0^\infty \mathbb{P}(g(X) > y)\, dy = \int_0^\infty \int_B f_X(x)\, dx\, dy = \int_0^\infty \int_0^{g(x)} f_X(x)\, dy\, dx = \int_0^\infty g(x) f_X(x)\, dx$$

We then consider that $g(x) \leq 0$ for all x. Let $Z = g(X)$ and $C = \{x : g(x) < z\}$. By Lemma 4.7,

$$\mathbb{E}(g(X)) = \int_{-\infty}^0 -F_Z(z)\, dz = \int_{-\infty}^0 \int_C -f_X(x)\, dx\, dz = \int_{-\infty}^0 \int_{g(x)}^0 -f_X(x)\, dz\, dx = \int_{-\infty}^0 g(x) f_X(x)\, dx$$

Now we combined both formulas into one. If $g(X)$ is a random variable,

$$\mathbb{E}(g(X)) = \int_0^\infty g(x) f_X(x)\, dx + \int_{-\infty}^0 g(x) f_X(x)\, dx = \int_{-\infty}^\infty g(x) f_X(x)\, dx$$

□

Similar to discrete random variables, this theorem also provided a method to calculate the moments of a continuous distribution.

Definition 4.9. Given $k \in \mathbb{N}_+$ and a continuous random variable X. The **k-th moment** is defined to be

$$\mathbb{E}X^k = \int_{-\infty}^\infty x^k f_X(x)\, dx$$

The **k-th central moment** is defined to be

$$\mathbb{E}((X - \mathbb{E}X)^k) = \int_{-\infty}^\infty (x - \mathbb{E}X)^k f_X(x)\, dx$$

Variance is defined as $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$.

We have some important continuous distributions.

Example 4.1. (**Uniform distribution**) $X \sim \mathrm{U}[a, b]$
Random variable X is **uniform** on $[a, b]$ if CDF and PDF of X is

$$F_X(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a < x \leq b \\ 1, & x > b \end{cases} \qquad f_X(x) = \begin{cases} \frac{1}{b-a}, & a < x \leq b \\ 0, & \text{Otherwise} \end{cases}$$

Example 4.2. (**Inverse transform sampling**) If we have an invertible CDF $G(x)$. How can we generate a random variable Y with the given distribution function?
We only need to generate an uniform random variable $U \sim \mathrm{U}[0, 1]$. We claim that $Y = G^{-1}(U)$ has the distribution function $G(x)$.

$$F_Y(x) = \mathbb{P}(Y \leq x) = \mathbb{P}(G^{-1}(U) \leq x) = \mathbb{P}(U \leq G(x)) = F_U(G(x)) = G(x)$$

Example 4.3. (**Exponential distribution**) $X \sim \mathrm{Exp}(\lambda)$
Random variable X is **exponential** with parameter $\lambda > 0$ if CDF and PDF of X is

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Example 4.4. (**Normal distribution / Gaussian distribution**) $X \sim \mathrm{N}(\mu, \sigma^2)$
Random variable X is **normal** if it has two parameters μ and σ^2, and its PDF and CDF is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$$

This distribution is the most important distribution.
The random variable X is **standard normal** if $\mu = 0$ and $\sigma^2 = 1$. ($X \sim \mathrm{N}(0,1)$)

$$f_X(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad F_X(x) = \Phi(x) = \int_{-\infty}^{x} \phi(u)\,du$$

Claim 4.9.1. $\phi(x)$ is a probability distribution function.

Proof.
Let $I = \int_{-\infty}^{\infty} \phi(x)\,dx$.

$$I^2 = \int_{-\infty}^{\infty} \phi(x)\,dx \int_{-\infty}^{\infty} \phi(y)\,dy = \frac{1}{2\pi} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,dx\,dy$$

Let $x = r\cos\theta$ and $y = r\sin\theta$ where $r \in [0,\infty)$ and $\theta \in [0, 2\pi]$

$$I^2 = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2}} r\,dr\,d\theta = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2}}\,d\left(\frac{r^2}{2}\right)\,d\theta = \frac{1}{2\pi}\int_0^{2\pi} d\theta = 1$$

□

These are some properties that are used frequently.

Lemma 4.10. The normal distribution has the following properties:

1. Let $X \sim \mathrm{N}(0,1)$. If a random variable $Y = bX + a$ for some $a, b \in \mathbb{R}$ and $b \neq 0$, then $Y \sim \mathrm{N}(a, b^2)$.
2. Let $X \sim \mathrm{N}(a, b^2)$ for some $a, b \in \mathbb{R}$ and $b \neq 0$. If a random variable $Y = \frac{X-a}{b}$, then $Y \sim \mathrm{N}(0,1)$.
3. If $Y \sim \mathrm{N}(a, b^2)$, then $\mathbb{E}Y = a$ and $\mathrm{Var}(Y) = b^2$.

Proof.

1. Let $z = bx + a$.

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}\left(X \le \frac{y-a}{b}\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{y-a}{b}} e^{-\frac{x^2}{2}}\,dx = \frac{1}{\sqrt{2\pi b^2}}\int_{-\infty}^{y} e^{-\frac{(z-a)^2}{2b^2}}\,dz$$

   Therefore, $Y \sim \mathrm{N}(a, b^2)$.

2. Let $x = bz + a$.

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X \le by + a) = \frac{1}{\sqrt{2\pi b^2}}\int_{-\infty}^{by+a} e^{-\frac{(x-a)^2}{2b^2}}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-\frac{z^2}{2}}\,dz$$

   Therefore, $Y \sim \mathrm{N}(0,1)$.

3. Let $y = bz + a$.

$$\mathbb{E}Y = \frac{1}{\sqrt{2\pi b^2}}\int_{-\infty}^{\infty} y e^{-\frac{(y-a)^2}{2b^2}}\,dy = \frac{1}{\sqrt{2\pi}}\left(\int_{-\infty}^{\infty} bze^{-\frac{z^2}{2}}\,dz + \int_{-\infty}^{\infty} ae^{-\frac{z^2}{2}}\,dz\right) = \frac{a}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\,dz = a(1) = a$$

$$\mathrm{Var}(Y) = \frac{1}{\sqrt{2\pi b^2}}\int_{-\infty}^{\infty} (y-a)^2 e^{-\frac{(y-a)^2}{2b^2}}\,dy = \frac{b^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}}\,dz = \frac{-b^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z\,d\left(e^{-\frac{z^2}{2}}\right) = \frac{b^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\,dz = b^2$$

□

Lemma 4.11. If $X \sim \mathrm{N}(a, b^2)$, then:

$$\mathbb{P}(s \le X \le t) = \mathbb{P}\left(\frac{s-a}{|b|} \le \frac{X-a}{|b|} \le \frac{t-a}{|b|}\right) = \Phi\left(\frac{t-a}{|b|}\right) - \Phi\left(\frac{s-a}{|b|}\right)$$

Proof.
Just apply Lemma 4.2 and you would get the equation. □

Example 4.5. (**Cauchy distribution**) $X \sim \text{Cauchy}$
Random variable X has a Cauchy distribution if it has a PDF:

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

It has the expectation

$$\mathbb{E}|X| = \int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)}\,dx = 2\int_0^{\infty} \frac{x}{\pi(1+x^2)}\,dx = \infty$$

There are also plenty of other continuous distributions. For example, Gamma distribution, Beta distribution, Weibull distribution, etc. However, they are too complicated and we will not discuss them here.

4.3 Joint distribution function of continuous random variables

Again, we recall the definition of joint distribution function.

Definition 4.12. Joint distribution function (JCDF) of two continuous random variables X and Y is the function $F : \mathbb{R}^2 \to [0,1]$ such that:

$$F_{X,Y}(x,y) = \mathbb{P}(X \le x, Y \le y)$$

Two continuous random variables X and Y are **jointly continuous** if the have a **joint density function** (JPDF) $f : \mathbb{R}^2 \to [0,\infty)$ such that:

$$F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\,du\,dv \qquad f_{X,Y}(x,y) = \frac{\partial^2}{\partial x\,\partial y}F_{X,Y}(x,y) \qquad \mathbb{P}((X,Y)\in D) = \iint_D f_{X,Y}(x,y)\,dx\,dy$$

We also recall the definition of marginal distribution function.

Definition 4.13. Given two continuous random variables X and Y. Marginal distribution function (Marginal PDF) of X given Y is

$$F_X(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{\infty}\int_{-\infty}^{x} f_{X,Y}(u,v)\,du\,dv = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(u,v)\,dv\,du$$
$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,u)\,dv$$

Similarly, we have the following extension of Theorem 4.8. However, we are not going to prove it here.

Theorem 4.14. If X and Y are jointly continuous random variables and $g(X,Y)$ is continuous random variable, then

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy$$

We can obtain the following important lemma.

Lemma 4.15. If X and Y are jointly continuous random variables, then for any $a, b \in \mathbb{R}$,

$$\mathbb{E}(aX+bY) = a\mathbb{E}X + b\mathbb{E}Y$$

Proof.

$$\begin{aligned}\mathbb{E}(aX+bY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(ax+by)f_{X,Y}(x,y)\,dx\,dy\\ &= \int_{-\infty}^{\infty} ax f_X(x)\,dx + \int_{-\infty}^{\infty} by f_Y(y)\,dy\\ &= a\mathbb{E}X + b\mathbb{E}Y\end{aligned}$$

□

Example 4.6. Assume that a plane is ruled by horizontal lines separated by D and a needle of length $L \leq D$ is cast randomly on the plane. What is the probability that the needle intersects some lines?
Let X be the distance from center of the needle to the nearest line and Θ be the acute angle between the needle and vertical line. We have $\mathbb{P}(\text{Intersection}) = \mathbb{P}\left(\frac{L}{2}\cos\Theta \geq X\right)$.
Assume that $X \perp\!\!\!\perp \Theta$. We have $X \sim \mathrm{U}\left[0, \frac{D}{2}\right]$ and $\Theta \sim \mathrm{U}\left[0, \frac{\pi}{2}\right]$.

$$f_{X,\Theta}(x,\theta) = \begin{cases} \frac{4}{D\pi}, & 0 \leq x \leq \frac{D}{2}, 0 \leq \theta \leq \frac{\pi}{2} \\ 0, & \text{Otherwise} \end{cases}$$

$$\mathbb{P}\left(\frac{L}{2}\cos\Theta \geq X\right) = \iint_{\frac{L}{2}\cos\theta \geq x} \frac{4}{D\pi}\mathbf{1}_{0\leq x\leq \frac{D}{2}}\mathbf{1}_{0\leq\theta\leq\frac{\pi}{2}}\,dx\,d\theta = \int_0^{\frac{\pi}{2}}\int_0^{\frac{L}{2}\cos\theta}\frac{4}{D\pi}\,dx\,d\theta = \frac{2L}{D\pi}$$

Suppose that we throw the needle for n times.

$$\frac{\#\{\text{Intersection}\}}{n} \approx \mathbb{P}(\text{Intersection}) = \frac{2L}{D\pi}$$

Combining two normal distributions into a joint distribution can be really useful.

Example 4.7. (**Standard bivariate normal distribution**) Two continuous random variables X and Y are **standard bivariate normal** if they have JPDF:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}\right)$$

where ρ is a constant satisfying $-1 < \rho < 1$.

Remark 4.15.1. If $X \sim \mathrm{N}(0,1)$ and $Y \sim \mathrm{N}(0,1)$,

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{\infty}\exp\left(-\frac{(x-\rho y)^2+(1-\rho^2)y}{2(1-\rho^2)}\right)dx \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi(1-\rho^2)}}e^{-\frac{(x-\rho y)^2}{2(1-\rho^2)}}\,dx \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}} \end{aligned}$$

Remark 4.15.2. ρ is the **correlation coefficient** between X and Y and is given by

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

Remark 4.15.3. If $X \sim \mathrm{N}(0,1)$ and $Y \sim \mathrm{N}(0,1)$,

$$\begin{aligned} \mathrm{cov}(X,Y) &= \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y = \mathbb{E}(XY) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{y}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}\frac{x}{\sqrt{2\pi(1-\rho^2)}}e^{-\frac{(x-\rho y)^2}{2(1-\rho^2)}}\,dx\,dy \\ &= \int_{-\infty}^{\infty}\frac{y}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}\rho y\,dy = \rho\int_{-\infty}^{\infty}y^2\phi(y)\,dy = \rho \end{aligned}$$

Example 4.8. (**Bivariate normal distribution**) Two continuous random variables X and Y are **bivariate normal** with means μ_X and μ_Y, variance σ_X^2 and σ_Y^2, and correlation coefficient ρ if JPDF is given by

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right)$$

There are some remarks that may be important to know about.

Remark 4.15.4. X and Y are bivariate normal and uncorrelated $\iff$ X and Y are independent normal.

Remark 4.15.5. X and Y are jointly continuous and they are both normal does not mean they are bivariate normal.

Example 4.9. Consider a JPDF of random variables X and Y

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi} e^{-\frac{1}{2}(x^2+y^2)}, & xy > 0 \\ 0, & xy \leq 0 \end{cases}$$

As you can see, this is not a bivariant normal distribution.
However, if you look at their marginal PDF,

$$f_X(x) = \int_0^\infty \frac{1}{\pi} e^{-\frac{1}{2}(x^2+y^2)}\, dy = \frac{1}{2\pi} \int_{-\infty}^\infty e^{-\frac{1}{2}(x^2+y^2)}\, dy = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad x > 0$$

$$f_X(x) = \int_{-\infty}^0 \frac{1}{\pi} e^{-\frac{1}{2}(x^2+y^2)}\, dy = \frac{1}{2\pi} \int_{-\infty}^\infty e^{-\frac{1}{2}(x^2+y^2)}\, dy = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad x < 0$$

This is the same to $f_Y(x)$.
Therefore, X and Y are jointly continuous and they are both normal does not mean they are bivariant normal.

Remark 4.15.6. Two random variables X and Y are jointly continuous and uncorrelated Gaussian does not mean they are independent Gaussian.

4.4 Conditional distribution of continuous random variables

Recall the definition of conditional distribution function of discrete random variable Y given $X = x$.

$$F_{Y|X}(y|x) = \mathbb{P}(Y \leq y | X = x) = \frac{\mathbb{P}(Y \leq y, X = x)}{\mathbb{P}(X = x)}$$

However, for the continuous random variables, $\mathbb{P}(X = x) = 0$ for all x. We take a limiting point of view.
Suppose the probability distribution function $f_X(x) > 0$,

$$\begin{aligned} F_{Y|X}(y|x) = \mathbb{P}(Y \leq y | x \leq X \leq x + dx) &= \frac{\mathbb{P}(Y \leq y, x \leq X \leq x + dx)}{\mathbb{P}(x \leq X \leq x + dx)} \\ &= \frac{\int_{-\infty}^y \int_x^{x+dx} f_{X,Y}(u,v)\, du\, dv}{\int_x^{x+dx} f_X(u)\, du} \\ &\approx \frac{\int_{-\infty}^y f_{X,Y}(x,v)\, dx\, dv}{f_X(x)\, dx} \\ &= \int_{-\infty}^y \frac{f_{X,Y}(x,v)}{f_X(x)}\, dv \end{aligned}$$

Definition 4.16. Suppose $X, Y : \Omega \to \mathbb{R}$ are two continuous random variables with PDF $f_X(x) > 0$ for some $x \in \mathbb{R}$. **Conditional distribution function** (Conditional CDF) of Y given $X = x$ is defined by

$$F_{Y|X}(y|x) = \mathbb{P}(Y \leq y | X = x) = \int_{-\infty}^y \frac{f_{X,Y}(x,v)}{f_X(x)}\, dv$$

Conditional density function (Conditional PDF) of Y given $X = x$ is defined by

$$f_{Y|X}(y|x) = \frac{\partial}{\partial y} F_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

Remark 4.16.1. Since $f_X(x)$ can also be computed from $f(x,y)$, we can simply compute

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{\int_{-\infty}^\infty f_{X,Y}(x,y)\, dy}$$

Remark 4.16.2. More generally, for two continuous random variables X and Y with PDF $f_X(x) > 0$ for some $x \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}(Y \in A | X = x) &= \int_A \frac{f_{X,Y}(x,v)}{f_X(x)}\,dv \\ &= \int_A f_{Y|X}(y|x)\,dy\end{aligned}$$

Example 4.10. Assume that two jointly continuous random variables X and Y have a JPDF:

$$f_{X,Y}(x,y) = \begin{cases}\frac{1}{x}, & 0 \le y \le x \le 1 \\ 0, & \text{Otherwise}\end{cases} = \frac{1}{x}\mathbf{1}_{0\le y\le x\le 1}$$

We want to compute $f_X(x)$ and $f_{Y|X}(y|x)$. For $x \in [0,1]$,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_{-\infty}^{\infty} \frac{1}{x}\mathbf{1}_{0\le y\le x\le 1}\,dy = \int_0^x \frac{1}{x}\,dy = 1$$

Therefore, $X \sim \mathrm{U}[0,1]$.
For $0 \le y \le x$ and $0 \le x \le 1$,

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{1}{x}$$

Therefore, $(Y|X=x) \sim \mathrm{U}[0,x]$.

Example 4.11. We want to find $\mathbb{P}(X^2+Y^2 \le 1)$ with two jointly continuous random variables X and Y having JPDF in Example 4.10. Let $Y \in A_x = \{y : |y| \le \sqrt{1-x^2}\}$.

$$\begin{aligned}\mathbb{P}(X^2+Y^2 \le 1 | X = x) = \mathbb{P}(|Y| \le \sqrt{1-x^2} | X = x) &= \int_{A_x} f_{Y|X}(y|x)\,dy \\ &= \int_{A_x \cap [0,1]} \frac{1}{x}\,dy \\ &= \int_0^{\min\{x, \sqrt{1-x^2}\}} \frac{1}{x}\,dy \\ &= \min\{1, \sqrt{x^{-2}-1}\}\end{aligned}$$

$$\begin{aligned}\mathbb{P}(X^2+Y^2 \le 1) &= \iiint_{x^2+y^2\le 1} f_{X,Y}(x,y)\,dy\,dx \\ &= \iiint_{x^2+y^2\le 1} f_{Y|X}(y|x)\,dy f_X(x)\,dx \\ &= \int_0^1 \min\{1, \sqrt{x^{-2}-1}\}\,dx \\ &= \int_0^{\frac{1}{\sqrt{2}}} dx + \int_{\frac{1}{\sqrt{2}}}^1 \sqrt{x^{-2}-1}\,dx \\ &= \frac{1}{\sqrt{2}} + \int_{\frac{\pi}{4}}^{\frac{\pi}{2}} \left(\frac{1}{\sin\theta} - \sin\theta\right) d\theta \qquad (x = \sin\theta) \\ &= \ln\left(\tan\frac{\theta}{2}\right)\Big|_{\frac{\pi}{4}}^{\frac{\pi}{2}} = \ln(1) - \ln(\sqrt{2}-1) = \ln(1+\sqrt{2})\end{aligned}$$

Example 4.12. Assume that random variables $X \sim \mathrm{N}(0,1)$ and $Y \sim \mathrm{N}(0,1)$ are standard bivariate normal. For $-1 < \rho < 1$,

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right)$$

We want to find $f_{X|Y}(x|y)$.

$$\begin{aligned}
f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\
&= \sqrt{2\pi} e^{\frac{1}{2}y^2} f_{X,Y}(x,y) && (C_{1,y} = \sqrt{2\pi} e^{\frac{1}{2}y^2}) \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{\frac{1}{2}y^2 - \frac{y^2}{2(1-\rho^2)}} \exp\left(-\frac{x^2 - 2\rho xy}{2(1-\rho^2)}\right) && (C_{2,y} = \tfrac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{\left(\frac{1}{2} - \frac{1}{2(1-\rho^2}\right)y^2}) \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} e^{\left(\frac{1}{2} - \frac{1}{2(1-\rho^2)} - \frac{\rho^2}{2(1-\rho^2)}\right)y^2} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right) && (C_{3,y} = \tfrac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}) \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right)
\end{aligned}$$

Therefore, we have $(X|Y=y) \sim \mathrm{N}(\rho y, 1-\rho^2)$. As $\rho \to 1$, we have $X \to Y$. As $\rho \to -1$, we have $X \to -Y$.
In general, there exists a random variable $Z \sim \mathrm{N}(0,1)$ such that

$$X = \rho Y + \sqrt{1-\rho^2} Z \qquad (X|Y=y) = \rho y + \sqrt{1-\rho^2} Z \qquad \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \rho & \sqrt{1-\rho^2} \\ 1 & 0 \end{pmatrix} \begin{pmatrix} Y \\ Z \end{pmatrix}$$

We can see that bivariate normal distribution is a linear transform of two independent normal distribution.
More generally, for any orthogonal matrix $\mathbf{A}$, we have two random variables W and U such that if they can be obtained by:

$$\begin{pmatrix} W \\ U \end{pmatrix} = \begin{pmatrix} \rho & \sqrt{1-\rho^2} \\ 1 & 0 \end{pmatrix} \mathbf{A} \begin{pmatrix} Y \\ Z \end{pmatrix}$$

then W and U will also be bivariate normal with ρ.

With conditional density function defined, we can now define conditional expectation.

Definition 4.17. Given two continuous random variables X and Y and an event $X = x$ for some $x \in \mathbb{R}$. **Conditional expectation** of Y is defined by:

$$\psi(x) = \mathbb{E}(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\, dy$$

Given a continuous random variable X. Conditional expectation of Y is defined by:

$$\psi(X) = \mathbb{E}(Y|X)$$

Again we also have the same properties of conditional distribution.

Lemma 4.18. (Law of total expectation) Conditional expectation $\psi(X) = \mathbb{E}(Y|X)$ for continuous random variables X and Y satisfies:

$$\mathbb{E}Y = \mathbb{E}(\psi(X))$$

Proof.

$$\begin{aligned}
\mathbb{E}(\psi(X)) &= \int_{-\infty}^{\infty} \psi(x) f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X}(y|x) f_X(x)\, dy\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x,y)\, dy\, dx \\
&= \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} y f_Y(y)\, dy = \mathbb{E}Y
\end{aligned}$$

□

Lemma 4.19. Conditional expectation $\psi(X) = \mathbb{E}(Y|X)$ for continuous random variables X and Y satisfies:

$$\mathbb{E}(Yg(X)) = \mathbb{E}(\psi(X)g(X))$$

Proof.

$$\begin{aligned}\mathbb{E}(\psi(X)g(X)) &= \int_{-\infty}^{\infty} \psi(x)g(x)f_X(x)\,dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf_{Y|X}(y|x)f_X(x)g(x)\,dy\,dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf_{X,Y}(x,y)g(x)\,dy\,dx \\ &= \mathbb{E}(Yg(X))\end{aligned}$$

□

4.5 Functions of continuous random variables

Given a continuous random variable X and a function g such that $g(X)$ is still a random variable, we have $\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x)f_X(x)\,dx$. Therefore, we only need $f_x(x)$ to compute $\mathbb{E}g(X)$. However, very often, we want to know the distribution of $g(X)$.

Example 4.13. Assume that X is continuous random variable with PDF $f_X(x)$. Let $Y = g(X)$ be a continuous random variable. How do we find the PDF $f_Y(y)$? We work with $F_Y(y)$ first. Let $g^{-1}(A) = \{x \in \mathbb{R} : g(x) \in A\}$.

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(g(X) \in (-\infty, y]) = \mathbb{P}(X \in g^{-1}((-\infty, y])) = \int_{g^{-1}((-\infty,y])} f_X(x)\,dx$$

$$f_Y(y) = \frac{\partial}{\partial y}\int_{g^{-1}((-\infty,y])} f_X(x)\,dx$$

Example 4.14. Let $X \sim \mathrm{N}(0,1)$. Let $Y = g(X) = X^2$. We want to find the PDF $f_Y(y)$.

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(-\sqrt{y} \le X \le \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y}) - 1$$

$$f_Y(y) = F'(y) = 2\phi(\sqrt{y})\left(\frac{1}{2\sqrt{y}}\right) = \frac{1}{\sqrt{y}}\phi(\sqrt{y}) = \begin{cases}\frac{1}{\sqrt{2\pi y}}\exp\left(\frac{-y}{2}\right), & y > 0 \\ 0, & y < 0\end{cases}$$

We have $X^2 \sim \chi^2(1)$. (This is a distribution)

Theorem 4.20. In case that $g(x)$ is strictly monotonic (strictly increasing or strictly decreasing) and differentiable, let $Y = g(X)$. We have

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))\left|\frac{\partial}{\partial y}g^{-1}(y)\right|, & \text{if } y = g(x) \text{ for some } x \\ 0, & \text{Otherwise}\end{cases}$$

Proof.
If $g(x)$ is a strictly increasing function,

$$\begin{aligned}F_Y(y) &= \mathbb{P}(g(X) \le y) = \mathbb{P}(X \le g^{-1}(y)) = F_X(g^{-1}(y)) \\ f_Y(y) &= F_Y'(y) = f_X(g^{-1}(y))\frac{\partial}{\partial y}g^{-1}(y) = f_X(g^{-1}(y))\left|\frac{\partial}{\partial y}g^{-1}(y)\right|\end{aligned}$$

If $g(x)$ is a strictly decreasing function,

$$\begin{aligned}F_Y(y) &= \mathbb{P}(g(X) \le y) = \mathbb{P}(X \ge g^{-1}(y)) = 1 - F_X(g^{-1}(y)) \\ f_Y(y) &= F_Y'(y) = -f_X(y^{-1}(y))\frac{\partial}{\partial y}g^{-1}(y) = f_X(g^{-1}(y))\left|\frac{\partial}{\partial y}g^{-1}(y)\right|\end{aligned}$$

□

We can consider the multivariable case.

Example 4.15. Suppose two random variables X and Y are jointly continuous with JPDF $f_{X,Y}$. Given that $U = g(X,Y)$ and $V = h(X,Y)$. What is $f_{U,V}(u,v)$? For simplifying the process, we need to first make some following assumptions.

1. X, Y can be uniquely solved from U, V. (There exists only 1 pair of functions a, b such that $X = a(U,V)$ and $Y = b(U,V)$)
2. The function g and h are differentiable and the Jacobian determinant

$$J(x,y) = \begin{vmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \\ \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \end{vmatrix} \neq 0$$

Then

$$f_{U,V}(u,v) = \frac{1}{|J(x,y)|} f_{X,Y}(x,y) = \begin{cases} \frac{1}{|J(a(u,v),b(u,v))|} f_{X,Y}(a(u,v),b(u,v)), & (u,v) = (g(x,y),h(x,y)) \text{ for some } x,y \\ 0, & \text{Otherwise} \end{cases}$$

Example 4.16. Given two jointly continuous random variables X_1, X_2 and their JPDF f_{X_1,X_2}.
Let $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$.

$$X_1 = \frac{Y_1 + Y_2}{2} = a(Y_1, Y_2) \qquad X_2 = \frac{Y_1 - Y_2}{2} = b(Y_1, Y_2) \qquad J(x_1,x_2) = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} = -2$$

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{|J(x_1,x_2)|} f_{X_1,X_2}(x_1,x_2) = \frac{1}{2} f_{X_1,X_2}\left(\frac{y_1+y_2}{2}, \frac{y_1-y_2}{2}\right)$$

More specifically, if $X_1 \sim \mathrm{N}(0,1)$, $X_2 \sim \mathrm{N}(0,1)$ and $X_1 \perp\!\!\!\perp X_2$,

$$\begin{aligned} f_{X_1,X_2}(x_1,x_2) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_1^2+x_2^2)} \\ f_{Y_1,Y_2}(y_1,y_2) &= \frac{1}{2} f_{X_1,X_2}\left(\frac{y_1+y_2}{2}, \frac{y_1-y_2}{2}\right) \\ &= \frac{1}{4\pi} e^{-\frac{1}{2}\left(\left(\frac{1}{2}(y_1+y_2)\right)^2 + \left(\frac{1}{2}(y_1-y_2)\right)^2\right)} \\ &= \frac{1}{4\pi} e^{-\frac{1}{4}(y_1^2+y_2^2)} \end{aligned}$$

Therefore, $Y_1 \perp\!\!\!\perp Y_2$ and we have $Y_1 \sim \mathrm{N}(0,2)$ and $Y_2 \sim \mathrm{N}(0,2)$.

Example 4.17. Given two random variables $X_1 \sim \mathrm{U}[0,1]$ and $X_2 \sim \mathrm{U}[0,1]$. If $X_1 \perp\!\!\!\perp X_2$, for all $x_1, x_2 \in \mathbb{R}$,

$$\begin{aligned} f_{X_1,X_2}(x_1,x_2) &= \begin{cases} 1, & x_1, x_2 \in [0,1] \\ 0, & \text{Otherwise} \end{cases} \\ f_{Y_1,Y_2}(y_1,y_2) &= \frac{1}{2} f_{X_1,X_2}\left(\frac{y_1+y_2}{2}, \frac{y_1-y_2}{2}\right) \\ &= \frac{1}{2} \mathbf{1}_{0 \leq y_1+y_2 \leq 2, 0 \leq y_1-y_2 \leq 2} \end{aligned}$$

Similar to discrete random variables, we can find the distribution of $X + Y$ when X and Y are jointly continuous.

Theorem 4.21. If two jointly continuous random variables X and Y have JPDF $f_{X,Y}$, then $X + Y$ has a PDF

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x)\, dx = \int_{-\infty}^{\infty} f_{X,Y}(z-y, y)\, dy$$

Proof.

$$\begin{aligned}
F_{X+Y}(z) &= \mathbb{P}(X+Y \le z) \\
&= \iint_{x+y\le z} f_{X,Y}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{z-y} f_{X,Y}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{z} f_{X,Y}(v-y,y)\,dv\,dy && (v = x+y) \\
&= \int_{-\infty}^{z}\int_{-\infty}^{\infty} f_{X,Y}(v-y,y)\,dy\,dv \\
f_{X+Y}(z) = F'_{X+Y}(z) &= \int_{-\infty}^{\infty} f_{X,Y}(z-y,y)\,dy = \int_{-\infty}^{\infty} f_{X,Y}(x,z-x)\,dx
\end{aligned}$$

□

Definition 4.22. Given two independent continuous random variables X and Y. **Convolution** f_{X+Y} $(f_X * f_Y)$ of PDFs of X and Y is the PDF of $X+Y$:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)\,dy = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)\,dx$$

Example 4.18. If $X \sim \mathrm{U}[0,1]$ and $Y \sim \mathrm{U}[0,1]$. In case of $X \perp\!\!\!\perp Y$,

$$\begin{aligned}
f_X(t) = f_Y(t) &= \begin{cases} 1, & 0 \le t \le 1 \\ 0, & \text{Otherwise} \end{cases} \\
f_{X+Y}(z) &= \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)\,dy \\
&= \int_0^1 f_X(z-y)\,dy \\
&= \int_0^1 \mathbf{1}_{0\le z-y\le 1}\,dy \\
&= \int_{\max\{0,z-1\}}^{\min\{1,z\}} dy && (z-1 \le y \le z) \\
&= \min\{1,z\} - \max\{0,z-1\} = \begin{cases} z, & 0 \le z \le 1 \\ 2-z, & 1 \le z \le 2 \\ 0, & \text{Otherwise} \end{cases}
\end{aligned}$$

The following example states that sum of independent normal random variables is still normal.

Example 4.19. If $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2, \cdots, n$ and they are independent, then $\sum_{i=1}^n X_i \sim \mathrm{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$.

Claim 4.22.1. It suffices to prove for the case $n = 2$.

Proof.
We first consider a special case when $X \sim \mathrm{N}(0, \sigma^2)$, $Y \sim \mathrm{N}(0, 1)$ and $X \perp\!\!\!\perp Y$.

$$\begin{aligned}
f_{X+Y}(z) &= \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-y)^2}{2\sigma^2}\right)\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)\right) dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma} \exp\left(-\frac{z^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2}(-2yz + y^2(1+\sigma^2))\right) dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma} \exp\left(-\frac{z^2}{2\sigma^2} + \frac{z^2}{2\sigma^2(1+\sigma^2)}\right) \exp\left(-\frac{1+\sigma^2}{2\sigma^2}\left(\frac{z^2}{(1+\sigma^2)^2} - \frac{2yz}{1+\sigma^2} + y^2\right)\right) dy \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{1+\sigma^2}} \exp\left(-\frac{z^2}{2\sigma^2} + \frac{z^2}{2\sigma^2(1+\sigma^2)}\right)\left(\frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{1+\sigma^2}}}\right) \exp\left(-\frac{\left(y - \frac{z}{1+\sigma^2}\right)^2}{2\left(\frac{\sigma}{\sqrt{1+\sigma^2}}\right)^2}\right) dy \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1+\sigma^2}} \exp\left(-\frac{z^2}{2(1+\sigma^2)}\right)
\end{aligned}$$

Therefore, $X + Y \sim \mathrm{N}(0, 1+\sigma^2)$. In general case when $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ and $X_1 \perp\!\!\!\perp X_2$.

$$X_1 + X_2 = \sigma_2\left(\frac{X_1 - \mu_1}{\sigma_2} + \frac{X_2 - \mu_2}{\sigma_2}\right) + \mu_1 + \mu_2$$

We get $\frac{X_1-\mu_1}{\sigma_2} \sim \mathrm{N}\left(0, \frac{\sigma_1^2}{\sigma_2^2}\right)$. Now we can apply this to special case and we get $\frac{X_1-\mu_1}{\sigma_2} + \frac{X_2-\mu_2}{\sigma_2} \sim \mathrm{N}\left(0, 1 + \frac{\sigma_1^2}{\sigma_2^2}\right)$.
Therefore, $X_1 + X_2 \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. By induction, if $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2, \cdots, n$ and they are independent, then

$$\sum_{i=1}^{n} X_i \sim \mathrm{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

□

Summary

Definition

Definition 1. **Sample space** Ω is the set of all possible outcomes ω of an experiment. It represents the universe of all potential results.

Definition 2. **Event** A is a subset of the sample space. Individual outcomes within A are referred to as **elementary events**.

Definition 3. The **complement** of a subset A is the set $A^{\complement}$, which includes all elements in the sample space Ω that are not part of A.

Definition 4. A **σ-field** (or **σ-algebra**) $\mathcal{F}$ is a collection of subsets of Ω that satisfies the following conditions:

1. If $A \in \mathcal{F}$, then $A^{\complement} \in \mathcal{F}$.
2. If $A_i \in \mathcal{F}$ for all i, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.
3. The empty set $\emptyset \in \mathcal{F}$.

Definition 5. A **measurable space** $(\Omega, \mathcal{F})$ consists of a sample space Ω and a σ-field $\mathcal{F}$.

Definition 6. A **probability measure** $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a function defined on a measurable space $(\Omega, \mathcal{F})$ that satisfies:

1. $\mathbb{P}(\emptyset) = 0$.
2. $\mathbb{P}(\Omega) = 1$.
3. If $A_i \in \mathcal{F}$ for all i and the sets A_i are disjoint, then:

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

Definition 7. A **probability space** $(\Omega, \mathcal{F}, \mathbb{P})$ consists of:

1. A sample space Ω.
2. A σ-field $\mathcal{F}$ of subsets of Ω.
3. A probability measure $\mathbb{P}$ defined on $(\Omega, \mathcal{F})$.

Definition 8. We say a sequence of events A_n **converges** and $\lim_{n\to\infty} A_n$ exists if

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

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $A_i \in \mathcal{F}$ for all i such that $A = \lim_{n\to\infty} A_n$ exists. Then

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

Definition 9. Event A is **null** if $\mathbb{P}(A) = 0$. This means A has no chance of occurring.

Definition 10. Event A is **almost surely** if $\mathbb{P}(A) = 1$. This indicates A occurs with certainty.

Definition 11. Given $\mathbb{P}(B) > 0$. **Conditional probability** that A occurs given that B occurs is:

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

Definition 12. Events A and B are independent ($A \perp\!\!\!\perp B$) if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
Given A_k for all $k \in I$. If for all $i \neq j$,

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$$

then they are **pairwise independent**.
If additionally, for all subsets $J \subseteq I$,

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

then they are **(mutually) independent**.

Definition 13. Let A be a collection of subsets of Ω. The **σ-field generated by** A is:

$$\sigma(A) = \bigcap_{A \subseteq \mathcal{G}} \mathcal{G}$$

where $\mathcal{G}$ are also σ-field. $\sigma(A)$ is the smallest σ-field containing A.

Definition 14. Product space of two probability spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ is the probability space $(\Omega_1 \times \Omega_2, \mathcal{G}, \mathbb{P}_{12})$ comprising:

1. a collection of ordered pairs $\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$
2. a σ-algebra $\mathcal{G} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ where $\mathcal{F}_1 \times \mathcal{F}_2 = \{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}$
3. a probability measure $\mathbb{P}_{12} : \mathcal{F}_1 \times \mathcal{F}_2 \to [0, 1]$ given by:

$$\mathbb{P}_{12}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2)$$

for $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$.

Definition 15. Random variable is a function $X : \Omega \to \mathbb{R}$ with the property that:

$$X^{-1}((-\infty, x]) = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

for any $X \in \mathbb{R}$. We say the function is **$\mathcal{F}$-measurable**.

Definition 16. Borel set is a set which can be obtained by taking countable union, intersection or complement repeatedly.

Definition 17. Borel σ-field $\mathcal{B}(\mathbb{R})$ of $\mathbb{R}$ is a σ-field that is generated by all open sets. It is a collection of Borel sets.

Definition 18. (Cumulative) distribution function (CDF) of a random variable X is a function $F_X : \mathbb{R} \to [0, 1]$ given by

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P} \circ X^{-1}((-\infty, x])$$

In **discrete** case, **probabilty mass function** (PMF) of discrete random variable X is the function $f : \mathbb{R} \to [0, 1]$ given by:

$$f_X(x) = \mathbb{P}(X = x) = \mathbb{P} \circ X^{-1}(\{x\}) \qquad F_X(x) = \sum_{i : x_i \leq x} f(x_i) \qquad f_X(x) = F_X(x) - \lim_{y \uparrow x} F_X(y)$$

In **continuous** case, **probability density function** (PDF) of continuous random variable X is the function $f : \mathbb{R} \to [0, \infty)$ given by:

$$F_X(x) = \int_{-\infty}^{x} f(u)\, du \qquad f_X(x) = \frac{\partial}{\partial x} F_X(x)$$

Definition 19. Let $X_i : \Omega \to \mathbb{R}$ for all $1 \leq i \leq n$ be random variables. **Random vector** $\vec{X} = (X_1, X_2, \cdots, X_n) : \Omega \to \mathbb{R}^n$ with properties:

$$\vec{X}^{-1}(D) = \{\omega \in \Omega : \vec{X}(\omega) = (X_1(\omega), X_2(\omega), \cdots, X_n(\omega)) \in D\} \in \mathcal{F}$$

for all $D \in \mathcal{B}(\mathbb{R}^n)$.
We can also say $\vec{X}$ is a random vector if

$$X_i^{-1}(B) \in \mathcal{F}$$

for all $B \in \mathcal{B}(\mathbb{R})$ and i.

Definition 20. Given a random vector (X, Y). **Joint distribution function** (JCDF) $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$ is defined as:

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \mathbb{P} \circ (X, Y)^{-1}((-\infty, x] \times (-\infty, y])$$

In discrete case, **joint probability mass function** (JPMF) of **jointly discrete** random variable X and Y is the function $f_{X,Y} : \mathbb{R}^2 \to [0, 1]$ given by:

$$f_{X,Y}(x, y) = \mathbb{P}((X, Y) = (x, y)) = \mathbb{P} \circ (X, Y)^{-1}(\{x, y\}) \qquad F_{X,Y}(x, y) = \sum_{u \leq x} \sum_{v \leq y} f(u, v)$$

In continuous case, **joint probability density function** (JPDF) of **jointly continuous** random variable X and Y is the function $f_{X,Y} : \mathbb{R}^2 \to [0, \infty)$ given by:

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \, \partial y} F_{X,Y}(x, y) \qquad F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v) \, du \, dv$$

Definition 21. Let X and Y be random variables. **Marginal distribution function** (Marginal CDF) is given by:

$$F_X(x) = \mathbb{P}(X^{-1}((-\infty, x]) \cap Y^{-1}((-\infty, \infty))) = \lim_{y \to \infty} F_{X,Y}(x, y)$$

In discrete case, **marginal mass function** (Marginal PMF) is given by:

$$f_X(x) = \sum_y f_{X,Y}(x, y)$$

In continuous case, **marginal density function** (Marginal PDF) is given by:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy$$

Definition 22. Given a random variable X. **Mean value**, **expectation**, or **expected value** of X is given by:

$$\mathbb{E}X = \begin{cases} \sum_{x : f_X(x) > 0} x f_X(x), & X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_X(x) \, dx, & X \text{ is continuous} \end{cases}$$

If it is absolutely convergent.

Definition 23. Given $k \in \mathbb{N}_+$ and a random variable X. **k-th moment** m_k is defined to be:

$$\mathbb{E}(X^k) = \begin{cases} \sum_x x^k f_X(x), & X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^k f_X(x) \, dx, & X \text{ is continuous} \end{cases}$$

k-th cnetral moment α_k is defined to be

$$\mathbb{E}((X - \mathbb{E}X)^k) = \begin{cases} \sum_x (x - \mathbb{E}X)^k f_X(x), & X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x - \mathbb{E}X)^k f_X(x) \, dx, & X \text{ is continuous} \end{cases}$$

Mean μ is the 1st moment $\mu = m_1 = \mathbb{E}X$.
Variance is the 2nd central moment $\alpha_2 = \mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}X)^2) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$.
Standard deviation σ is defined as $\sigma = \sqrt{\mathrm{Var}(X)}$.

Definition 24. Two random variables X and Y are **uncorrelated** if $\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y$.

Definition 25. **Covariance** of two random variables X and Y is:

$$\operatorname{cov}(X,Y) = \mathbb{E}((X-\mathbb{E}X)(Y-\mathbb{E}Y)) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y$$

Definition 26. Given two random variables X and Y. **Conditional distribution function** (Conditional CDF) of Y given $X = x$ for any x is defined by:

$$F_{Y|X}(y|x) = \mathbb{P}(Y \le y | X = x) = \begin{cases} \frac{\mathbb{P}(Y\le y, X=x)}{\mathbb{P}(X=x)}, & X \text{ is discrete} \\ \int_{-\infty}^{y} \frac{f_{X,Y}(x,v)}{f_X(x)}\, dv, & X \text{ is continuous} \end{cases}$$

In discrete case, **conditional mass function** (Conditional PMF) of Y given $X = x$ is defined by:

$$f_{Y|X}(y|x) = \begin{cases} \frac{\mathbb{P}(Y=y, X=x)}{\mathbb{P}(X=x)}, & X \text{ is discrete} \\ \frac{\partial}{\partial y} F_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, & X \text{ is continuous} \end{cases}$$

Definition 27. Given two random variables X and Y, and an event $X = x$ for some X. **Conditional expectation** of random variable Y is defined by:

$$\psi(x) = \mathbb{E}(Y|X=x) = \begin{cases} \sum_y y f_{Y|X}(y|x), & X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\, dy, & X \text{ and } Y \text{ are continuous} \end{cases}$$

Given a random variable X. Conditional expectation of random variable Y is defined by:

$$\psi(X) = \mathbb{E}(Y|X) = \begin{cases} \sum_x \psi(x), & X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \psi(x)\, dx, & X \text{ are continuous} \end{cases}$$

Definition 28. Given $X \perp\!\!\!\perp Y$. In discrete case, **convolution** f_{X+Y} ($f_X * f_Y$) of PMFs of random variables X and Y is the PMF of $X + Y$:

$$f_{X+Y}(z) = \mathbb{P}(X+Y-z) = \sum_x f_X(x) f_Y(z=x) = \sum_y f_X(z-y) f_Y(y)$$

In continuous case, **convolution** of PDFs of random variables X and Y is the PDF of $X + Y$:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx$$

Named Property

Property 1. (Inclusion-Exclusion Principle) For any finite collection of events $A_1, A_2, \ldots, A_n$:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \cdots + (-1)^{n+1} \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n).$$

Property 2. (Law of Total Probability) Let $\{B_1, B_2, \ldots, B_n\}$ be a partition of Ω such that $B_i \cap B_j = \emptyset$ for all $i \neq j$ and $\bigcup_{i=1}^{n} B_i = \Omega$. If $\mathbb{P}(B_i) > 0$ for all i, then:

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

Property 3. (Law of Total Expectation) Let $\psi(X) = \mathbb{E}(Y|X)$. Conditional expectation satisfies:

$$\mathbb{E}(\psi(X)) = \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$$

Property 4. (Tail sum formula) If X has a PDF f_X with $f_X(x) = 0$ when $x < 0$, and a CDF F_X, then:

$$\mathbb{E}X = \int_0^{\infty} (1 - F_X(x))\, dx$$

Distributions

For discrete random variables,

Example 1. (Bernoulli distribution) $X \sim \text{Bern}(p)$
Suppose we perform 1 Bernoulli trial. Let p be probability of success and X be number of successes.

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1-p, & 0 \le x < 1 \\ 1, & x \ge 1 \end{cases} \qquad f_X(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \\ 0, & \text{Otherwise} \end{cases} \qquad \mathbb{E}X = p \qquad \text{Var}(X) = p(1-p)$$

Example 2. (Binomial distribution) $Y \sim \text{Bin}(n, p)$
Suppose we perform n independent Bernoulli trials. Let p be the probability of success and $Y = X_1 + X_2 + \cdots + X_n$ be total number of successes.

$$f_Y(k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad F_Y(k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \qquad \mathbb{E}X = np \qquad \text{Var}(X) = np(1-p)$$

Example 3. (Trinomial distribution)
Suppose we perform n trials with three outcomes A, B and C, where the probability of occurrence is p, q and $1-p-q$ respectively. Let X be number of occurrence of A and Y be number of occurrence of B.
Probability of x A's, y B's and $n-x-y$ C's is:

$$f_{X,Y}(x, y) = \frac{n!}{x!y!(n-x-y)!} p^x q^y (1-p-q)^{n-x-y}$$

Example 4. (Geometric distribution) $W \sim \text{Geom}(p)$ $X \sim \text{Geom}(p)$
Suppose we keep performing independent Bernoulli trials until the first success shows up. Let p be probability of success. Let W be the waiting time which elapses before first success. For $k \ge 1$,

$$f_W(k) = p(1-p)^{k-1} \qquad F_W(k) = 1 - (1-p)^k \qquad \mathbb{E}W = \frac{1}{p} \qquad \text{Var}(W) = \frac{1-p}{p^2}$$

Let X be number of failures before first success. For $k \ge 0$,

$$f_X(k) = p(1-p)^k \qquad F_X(k) = 1 - (1-p)^{k+1} \qquad \mathbb{E}X = \frac{1-p}{p} \qquad \text{Var}(X) = \frac{1-p}{p^2}$$

Example 5. (Negative Binomial distribution) $W_r \sim \text{NBin}(r, p)$ $X \sim \text{NBin}(r, p)$
Suppose we keep performing independent Bernoulli trials until the first success shows up. Let p be the probability of success. Let W_r be the waiting time which elapses before r-th success. For any $k \ge r$,

$$f_{W_r}(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r} \qquad \mathbb{E}W_r = \frac{r}{p} \qquad \text{Var}(W_r) = \frac{r(1-p)}{p^2}$$

Let X be number of failures before the r-th success. For any $k \ge 0$,

$$f_X(k) = \binom{k+r-1}{r-1} p^r (1-p)^k \qquad \mathbb{E}X = \frac{r(1-p)}{p} \qquad \text{Var}(X) = \frac{r(1-p)}{p^2}$$

Example 6. (Poisson distribution) $X \sim \text{Poisson}(\lambda)$
Suppose we perform n independent Bernoulli trials. Let p be the probability of success, $\lambda = np$ and $X \sim \text{Bin}(n, p)$. When n is large, p is small, and np is moderate:

$$f_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda} \qquad F_X(k) = \sum_{i=0}^{k} \frac{\lambda^i}{i!} e^{-\lambda} \qquad \mathbb{E}X = \lambda \qquad \text{Var}(X) = \lambda$$

For continuous random variables,

Example 7. (Uniform distribution) $X \sim \mathrm{U}[a,b]$
Random variable X is uniform on $[a,b]$ is PDF and CDF is:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{Otherwise} \end{cases} \qquad F_X(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases}$$

Example 8. (Exponential distribution) $X \sim \mathrm{Exp}(\lambda)$
Random variable X is exponential with parameter $\lambda > 0$ if PDF and CDF is:

$$f_X(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \ge 0 \end{cases} \qquad F_X(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \ge 0 \end{cases}$$

Example 9. (Normal distribution / Gaussian distribution) $X \sim \mathrm{N}(\mu, \sigma^2)$
Random variable X is normal if it has two parameter μ and σ^2, and its PDF and CDF is:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad F_X(x) = \int_{-\infty}^{x} f_X(u)\,du \qquad \mathbb{E}X = \mu \qquad \mathrm{Var}(X) = \sigma^2$$

Random variable X is standard normal if $\mu = 0$ and $\sigma^2 = 1$. ($X \sim \mathrm{N}(0,1)$)

$$f_X(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \qquad F_X(x) = \Phi(x) = \int_{-\infty}^{x} \phi(u)\,du \qquad \mathbb{E}X = 0 \qquad \mathrm{Var}(X) = 1$$

Example 10. (Cauchy distribution) $X \sim \mathrm{Cauchy}$
Random variable X has a Cauchy distribution if:

$$f_X(x) = \frac{1}{\pi(1+x^2)} \qquad \mathbb{E}|X| = \int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)}\,dx = \infty$$

Example 11. (Bivariate normal distribution) Two random variables X and Y are bivariate normal with μ_X and μ_Y, variance σ_X^2 and σ_Y^2, and correlation coefficient ρ if:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right)$$

Two random variables X and Y are standard bivariate normal if $\mu_X = \mu_Y = 0$ and $\sigma_X^2 = \sigma_Y^2 = 1$.

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right)$$

Chapter 5

Generating function

5.1 Introduction of generating functions

A sequence of numbers $a = \{a_i : i = 0, 1, 2, \cdots\}$ can hold a significant amount of information. For example, the values of a PMF describe the distribution of discrete random variables.
A concise way to represent this information is by encapsulating the numbers in a generating function.

Definition 5.1. For any sequence $\{a_n : n = 0, 1, 2, \cdots\}$, the generating function is defined as

$$G_a(s) = \sum_{i=0}^{\infty} a_i s^i = \lim_{N\uparrow\infty} \sum_{i=0}^{N} a_i s^i$$

for $s \in \mathbb{R}$ if the limit exists.

Remark 5.1.1. It can be observed that

$$a_i = \frac{G_a^{(i)}(0)}{i!}$$

Example 5.1. In some cases, it is not possible to interchange a countable sum with derivatives.
Let $b_n(x) = \frac{\sin nx}{n}$ such that $a_1(x) = b_1(x)$ and $a_n(x) = b_n(x) - b_{n-1}(x)$.

$$\sum_{n=0}^{\infty} a_n(x) = \lim_{N\uparrow\infty} \sum_{i=0}^{\infty} a_n(x) = \lim_{N\uparrow\infty} \frac{\sin Nx}{N} = 0 \qquad \text{(Squeeze Theorem)}$$

$$\lim_{N\uparrow\infty} \frac{\partial}{\partial x} \sum_{i=0}^{\infty} a_i(x) = 0$$

$$\lim_{N\uparrow\infty} \sum_{i=0}^{N} \frac{\partial}{\partial x} a_n(x) = \lim_{N\uparrow\infty} \cos Nx \quad \text{does not exist}$$

Convolutions are frequently encountered in probability theory, and generating functions serve as a valuable tool for analyzing them.

Definition 5.2. Let $a = \{a_i : i \geq 0\}$ and $b = \{b_i : i \geq 0\}$ be two sequences of real numbers. The **convolution** $c = a * b = \{c_i : i \geq 0\}$ of $\{a_i\}$ and $\{b_i\}$ is defined as

$$c_n = \sum_{i=0}^{n} a_i b_{n-i}$$

Example 5.2. If $a_n = f_X(n)$ and $b_n = f_Y(n)$, then $c_n = f_{X+Y}(n)$.

Claim 5.2.1. If sequences a and b have generating functions $G_a(s)$ and $G_b(s)$ respectively, then

$$G_c(s) = G_a(s) G_b(s)$$

Proof.

$$G_c(s) = \sum_{n=0}^{\infty} c_n s^n = \sum_{n=0}^{\infty} \sum_{i=0}^{n} a_i b_{n-i} s^i s^{n-i} = \sum_{i=0}^{\infty} a_i s^i \sum_{n=i}^{\infty} b_{n-i} s^{n-i} = \sum_{i=0}^{\infty} a_i s^i \sum_{j=0}^{\infty} b_j s^j = G_a(s) G_b(s)$$

□

Example 5.3. Suppose that $X \perp\!\!\!\perp Y$. Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$. What is the distribution of $Z = X + Y$?
Recall that $f_Z = f_X * f_Y$. We let $a_n = f_X(n)$ and $b_n = f_Y(n)$.

$$G_{f_X}(s) = \sum_{i=0}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} s^i = e^{\lambda(s-1)}$$
$$G_{f_Y}(s) = e^{\mu(s-1)}$$
$$G_{f_Z}(s) = e^{(\lambda+\mu)(s-1)}$$

Suppose that X is a discrete random variable taking values in the non-negative integers. We can see how the generating function works in probability.

Definition 5.3. Probability generating function (PGF) of a non-negative random variable X is

$$G_X(s) = \mathbb{E}s^X = \sum_{i=0}^{\infty} s^i f_X(i)$$

We can see that the definition is a power series. We may want to know whether the series is convergent.

Definition 5.4. Radius of convergence R of power series is the half size of an interval such that the power series $f(s)$ is convergent. If $s \in (-R, R)$, then $f(s)$ is convergent. If $s \in [-R, R]^{\complement}$, then $f(s)$ is divergent.
We can obtain the radius of convergence by applying the root test:

$$R = \frac{1}{\limsup_{n\to\infty} \sqrt[n]{|a_n|}}$$

Remark 5.4.1. We need to perform additional tests to find whether the power series converges at $s = -R$ and $s = R$.

Remark 5.4.2. Sometimes, it is hard to compute R using the root test. One convenient way to compute R is using the ratio test. If the limit exists,

$$R = \lim_{n\to\infty} \left| \frac{a_n}{a_{n+1}} \right|$$

Here are some properties of power series involving the radius of convergence. We will not prove them since the proof is not important.

Theorem 5.5. If R is the radius of convergence of $G_a(s) = \sum_{i=0}^{\infty} a_i s^i$, then

1. $G_a(s)$ converges absolutely for all $|s| < R$ and diverges for all $|s| > R$.
2. $G_a(s)$ can be differentiated or integrated for any fixed number of times term by term if $|s| < R$.

$$\frac{\partial^i}{\partial s^i} \sum_{n=0}^{\infty} a_n s^n = \sum_{n=0}^{\infty} \frac{\partial^i}{\partial s^i} a_n s^n$$

3. If $R > 0$ and $G_a(s) = G_b(s)$ for all $|s| \le R'$ for some $0 < R' \le R$, then $a_n = b_n$ for all n.

Remark 5.5.1. For any sequence $\{a_n : n \ge 0\}$, if the radius of convergence of $G_a(s)$ is positive, then $\{a_n : n \ge 0\}$ is uniquely determined by $G_a(s)$ via

$$a_n = \frac{1}{n!} G_a^{(n)}(0)$$

Remark 5.5.2. If $a_n = f_X(n)$ for some random variables X, then $R \ge 1$ for $G_X(s) = G_a(s)$ since

$$\sum_{n=0}^{\infty} f_X(n) s^n$$

converges when $s \in [-1, 1]$.

Example 5.4. Let $X \sim \text{Poisson}(\lambda)$ and $a_n = f_X(n) = \frac{\lambda^n e^{-\lambda}}{n!}$. By the ratio test,

$$\frac{a_n}{a_{n+1}} = \frac{n+1}{\lambda} \to \infty$$

Therefore, $R = \infty$.

Example 5.5. Let X has a PMF $a_n = f_X(n) = \frac{c}{n^2}$. By the ratio test,

$$\frac{a_n}{a_{n+1}} = \frac{(n+1)^2}{n} \to 1$$

Therefore, $R = 1$.

In fact, when $s = 1$, we can find the expectation of a distribution.

Example 5.6. By having $s = 1$,

$$\frac{\partial}{\partial s} G_X(s)\bigg|_{s=1} = \frac{\partial}{\partial s} \sum_{i=0}^{\infty} f_X(i) s^i \bigg|_{s=1} = \sum_{i=0}^{\infty} i f_X(i) s^i \bigg| = \sum_{i=0}^{\infty} i f_X(i) = \mathbb{E}X$$

There is an important theorem regarding $s = 1$. Again, we are not going to prove it.

Theorem 5.6. (Abel's Theorem) Suppose that $a_n \geq 0$ for all n. If a has a generating function $G_a(s)$ and radius of convergence $R = 1$, then if $\sum_{n=0}^{\infty}$ converges in $\mathbb{R} \cup \{\infty\}$, we have

$$\lim_{s \uparrow 1} G_a(s) = \sum_{n=0}^{\infty} a_n \lim_{s \uparrow 1} s^n = \sum_{n=0}^{\infty} a_n$$

Example 5.7. We have some PGF of random variable X.

$X \sim \text{Bern}(p)$	$G_X(s) = ps^1 + (1-p)s^0 = 1 - p + ps$
$X \sim \text{Bin}(n, p)$	$G_X(s) = (1 - p + ps)^n$
$X \sim \text{Geom}(p)$	$G_X(s) = \sum_{n=1}^{\infty} (1-p)^{n-1} p s^n = \frac{ps}{1 - s(1-p)}$
$X \sim \text{Poisson}(\lambda)$	$G_X(s) = e^{\lambda(s-1)}$

We already know that by computing the derivatives of G at $s = 0$, we can get the probability sequence. The following theorem shows that we can get the moment sequence by computing the derivatives of G at $s = 1$.

Theorem 5.7. If random variable X has a PGF $G_X(s)$, then

1. $\mathbb{E}X = \lim_{s \uparrow 1} G'(s) = G'(1)$
2. $\mathbb{E}(X(X-1)\cdots(X-k+1)) = G^{(k)}(1)$
3. $\text{Var}(X) = G''(1) + G'(1) - (G'(1))^2$

Proof.

1. This is proved in Example 5.6.
2. Let $s < 1$.

$$G^{(k)}(s) = \frac{\partial^k}{\partial s^k} \sum_n f_X(n) s^n = \sum_n n(n-1)\cdots(n-k+1) s^{n-k} f_X(n) = \mathbb{E}(s^{X-k} X(X-1)\cdots(X-k+1))$$

   By applying Abel's Theorem, we obtain

$$G^{(k)}(1) = \mathbb{E}(X(X-1)\cdots(X-k+1))$$

3.

$$\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}(X(X-1)) + \mathbb{E}X - (\mathbb{E}X)^2 = G''(1) + G'(1) - (G'(1))^2$$

□

From Example 5.3, we can generalize it to study the sum of many other independent discrete random variables.

Theorem 5.8. If $X \perp\!\!\!\perp Y$, then $G_{X+Y}(s) = G_X(s)G_Y(s)$.

Proof.

$$G_{X+Y}(s) = \sum_{z=0}^{\infty}\sum_{x=0}^{z} f_X(x)f_Y(z-x)s^z = \sum_{x=0}^{\infty} f_X(x)s^x \sum_{z=x}^{\infty} f_Y(z-x)s^{z-x} = \sum_{x=0}^{\infty} f_X(x)\sum_{y=0}^{\infty} f_Y(z)s^y = G_X(s)G_Y(s)$$

□

Interestingly, we can also use generating functions to deal with the sum of a random number of independent random variables.

Theorem 5.9. Let $X_1, X_2, \cdots$ be a sequence of independent identically distributed (i.i.d.) random variables with common PGF $G_X(s)$ and N be a random variable independent of X_i for all i with PGF $G_N(s)$. If $T = X_1 + X_2 + \cdots + X_N$, then

$$G_T(s) = G_N(G_X(s))$$

Proof.

$$G_T(s) = \mathbb{E}s^T = \mathbb{E}(\mathbb{E}(s^T|N)) = \sum_n \mathbb{E}(s^T|N=n)\mathbb{P}(N=n) = \sum_n \mathbb{E}(s^{X_1+X_2+\cdots+X_n}|N=n)\mathbb{P}(N=n) = \sum_n (G_X(s))^n\mathbb{P}(N=n)$$
$$= G_N(G_X(s))$$

□

Example 5.8. The sum of a Poisson number of independent Bernoulli random variables is still Poisson.
Let $G_N(t) = e^{\lambda(t-1)}$ and $G_X(s) = 1 - p + ps$.

$$G_T(s) = G_N(G_X(s)) = e^{\lambda(1-p+ps-1)} = e^{\lambda p(s-1)}$$

Therefore, $T \sim \text{Poisson}(\lambda p)$.

When JPMF exists, there obviously will be a joint PGF.

Definition 5.10. Let random variables X_1, X_2 be both non-negative integer-valued, jointly discrete with JPMF f_{X_1,X_2}.
Joint probability generating function (JPGF) is defined by

$$G_{X_1,X_2}(s_1,s_2) = \mathbb{E}s_1^{X_1}s_2^{X_2} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} s_1^i s_2^j f_{X_1,X_2}(i,j)$$

Remark 5.10.1. We can find that

$$f_{X_1,X_2}(i,j) = \left(\frac{\partial^i}{\partial s_1^i}\frac{\partial^j}{\partial s_2^j}\frac{G_{X_1,X_2}(s_1,s_2)}{i!j!}\right)\Bigg|_{(s_1,s_2)=(0,0)}$$

Theorem 5.11. Random variables X, Y are independent if and only if $G_{X,Y}(s,t) = G_X(s)G_Y(t)$.

Proof.
If $X \perp\!\!\!\perp Y$,

$$G_{X,Y}(s,t) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} s^i t^j f_{X,Y}(i,j) = \sum_{i=0}^{\infty} s^i f_X(i)\sum_{j=0}^{\infty} t^j f_Y(j) = G_X(s)G_Y(t)$$

If $G_{X,Y}(s,t) = G_X(s)G_Y(t)$, we consider the coefficient of terms $s^i t^j$ for all $i \geq 0$ and $j \geq 0$. We can see that

$$f_{X,Y}(i,j) = f_X(i)f_Y(j)$$

Therefore, $X \perp\!\!\!\perp Y$. □

Remark 5.11.1. We know that if $X_1 \perp\!\!\!\perp X_2$, then $G_{X_1+X_2}(s) = \mathbb{E}s^{X_1+X_2} = \mathbb{E}s^{X_1}s^{X_2} = G_{X_1}(s)G_{X_2}(s)$. Converse may not be true.

5.2 Applications of generating functions

The following example involves a simple random walk, which is discussed in Appendix A. Generating functions are particularly valuable when studying random walks. So far, we have only considered random variables X taking finite values. In this application, we encounter variables that can take the value $+\infty$. For such variables X, $G_X(s)$ converges as long as $|s| < 1$ and

$$\lim_{s\uparrow 1} G_X(s) = \sum_k \mathbb{P}(X = k) = 1 - \mathbb{P}(X = \infty)$$

Definition 5.12. A random variable X is **defective** if $\mathbb{P}(X = \infty) > 0$.

Remark 5.12.1. It is not surprising that the expectation is infinite when the random variable is defective.

With this generalization, we can start discussing random walks.

Example 5.9. (Recurrence and transience of random walk) Let S_n be the position of the particle after n moves, and X_i be independent and identically distributed random variables mentioned in Appendix A. For $n \geq 0$,

$$S_n = \sum_{i=1}^{n} X_i \qquad S_0 = 0 \qquad \mathbb{P}(X_i = 1) = p \qquad \mathbb{P}(X_i = -1) = q = 1 - p$$

Let T_0 be the number of moves until the particle makes its first return to the origin.

$$T_0 = \min\{i \geq 1 : S_i = 0\}$$

Is T_0 a defective random variable? How do we calculate $\mathbb{P}(T_0 = \infty)$?
Let $p_0(n)$ be the probability of the particle returning to the origin at n moves, and P_0 be the generating function of p_0.
Let $f_0(n)$ be the probability of the particle first returning to the origin at n moves, and F_0 be the generating function of f_0.

$$p_0(n) = \mathbb{P}(S_n = 0) = \begin{cases} \binom{n}{\frac{n}{2}} p^{\frac{n}{2}} q^{\frac{n}{2}}, & n \text{ is even} \\ 0, & n \text{ is odd} \end{cases} \qquad P_0(s) = \lim_{N\uparrow\infty} \sum_{n=0}^{N} p_0(n) s^n$$

$$f_0(n) = \mathbb{P}(S_1 \neq 0, S_2 \neq 0, \cdots, S_{n-1} \neq 0, S_n = 0) = \mathbb{P}(T_0 = n) \qquad F_0(s) = \lim_{N\uparrow\infty} \sum_{n=1}^{N} f_0(n) s^n$$

Theorem 5.13. From the definitions in Example 5.9, we have

1. $P_0(s) = 1 + P_0(s)F_0(s)$
2. $P_0(s) = (1 - 4pqs^2)^{-\frac{1}{2}}$
3. $F_0(s) = 1 - (1 - 4pqs^2)^{\frac{1}{2}}$

Proof.

1. Let $A_n = \{S_n = 0\}$ and $B_k = \{S_1 \neq 0, S_2 \neq 0, \cdots, S_{k-1} \neq 0, S_k = 0\}$. $p_0(n) = \mathbb{P}(A_n)$ and $f_0(k) = \mathbb{P}(B_k)$.
   By using the Law of total probability,

$$\begin{aligned}
\mathbb{P}(A_n) &= \sum_{i=1}^{n} \mathbb{P}(A_n|B_i)\mathbb{P}(B_i) \\
p_0(n) &= \sum_{i=1}^{n} \mathbb{P}(S_n = 0|S_1 \neq 0, S_2 \neq 0, \cdots, S_{i-1} \neq 0, S_i = 0) f_0(i) \\
&= \sum_{i=1}^{n} \mathbb{P}(S_n = 0|S_i = 0) f_0(i) && \text{(Markov property in Lemma A.1)} \\
&= \sum_{i=1}^{n} \mathbb{P}(S_{n-i} = 0) f_0(i) && \text{(Temporarily homogeneous property in Lemma A.1)} \\
&= \sum_{i=1}^{n} p_0(n-k) f_0(i) \\
p_0(0) &= 1
\end{aligned}$$

$$\begin{aligned}
P_0(s) = \sum_{k=0}^{\infty} p_0(k)s^k = 1 + \sum_{k=1}^{\infty} p_0(k)s^k &= 1 + \sum_{k=1}^{\infty}\sum_{i=1}^{k} p_0(k-i)f_0(i)s^k \\
&= 1 + \sum_{i=1}^{\infty}\sum_{k=i}^{\infty} p_0(k-i)s^{k-i}f_0(i)s^i \\
&= 1 + P_0(s)F_0(s)
\end{aligned}$$

2. *If you want to understand the proof, search "Central binomial coefficient" in Wikipedia*
   We know that $S_n = 0$ if n is even. Therefore,

$$\begin{aligned}
P_0(s) = \lim_{N\uparrow\infty} \sum_{n=0}^{N} p_0(n)s^n &= \lim_{N\uparrow\infty} \sum_{i=0}^{N} \binom{2i}{i} p^i q^i s^{2i} \\
&= \lim_{N\uparrow\infty} \sum_{i=1}^{N} (-1)^i 4^i \binom{\frac{-1}{2}}{i} p^i q^i s^{2i} && \left(\binom{\frac{-1}{2}}{i} \text{ is a generalized binomial coefficient}\right) \\
&= \frac{1}{\sqrt{1-4pqs^2}}
\end{aligned}$$

3. By applying (1) and (2), we can get

$$F_0(s) = \frac{P_0(s) - 1}{P_0(s)} = 1 - \sqrt{1-4pqs^2}$$

□

From this theorem, we can get the following corollary.

Corollary 5.14. The probability that the particle ever returns to the origin is

$$\sum_{n=1}^{\infty} f_0(n) = F_0(1) = 1 - |p-q|$$

Probability that the particle will not return to origin ever is

$$\mathbb{P}(T_0 = \infty) = |p-q|$$

Proof.
By using Theorem 5.13, since $p + q = 1$,

$$F_0(1) = 1 - (1-4pq)^{\frac{1}{2}} = 1 - (p^2 - 2pq + q^2)^{\frac{1}{2}} = 1 - |p-q|$$

□

Remark 5.14.1. A random walk is **recurrent** if it has at least one recurrent point. ($\mathbb{P}(X < \infty) = 1$)
A random walk is **transient** if it has no recurrent points. ($\mathbb{P}(X = \infty) > 0$)
Notice that when $p = q = \frac{1}{2}$, $\mathbb{P}(T_0 = \infty)$ and therefore the random walk is recurrent.
If $p \neq q$, then $\mathbb{P}(T_0 = \infty) \neq 0$ and so the random walk is transient.

Example 5.10. We use the Example 5.9 again. How do we calculate $\mathbb{E}T_0$ if $p = q = \frac{1}{2}$?

$$F_0(s) = 1 - \sqrt{1 - s^2} \qquad F_0'(s) = \frac{s}{\sqrt{1 - s^2}} \qquad \mathbb{E}T_0 = \lim_{s \uparrow 1} F_0'(s) = \infty$$

This means that although we find that the particle almost certainly returns to the origin, the expectation for the number of steps needed to return to the origin is still infinite.

We move on to our next important application, which is the Branching Process. Many scientists have been interested in reproduction in a population. Accurate models for evolution are extremely difficult to handle, but some non-trivial models are tractable. We will investigate one of the models.

Example 5.11. (Galton-Watson process) This process investigates a population that evolves in generations.
Let Z_n be the number of individuals of the n-th generation and $X_i^{(m)}$ be the number of offspring of the i-th individual of the m-th generation. We have:

$$Z_{n+1} = \begin{cases} X_1^{(n)} + X_2^{(n)} + \cdots + X_{Z_n}^{(n)}, & Z_n \geq 1 \\ 0, & Z_n = 0 \end{cases}$$

We make some following assumptions:

1. Family sizes of the individuals of the branching process form a collection of independent random variables. ($X_i^{(k)}$'s are independent)
2. All family sizes have the same probability mass function f and generating function G. ($X_i^{(k)}$'s are identically distributed)

Assume that $Z_0 = 1$. Note that $Z_1 = X_1^{(0)}$

Theorem 5.15. Let $G_n(s) = \mathbb{E}s^{Z_n}$ and $G(s) = G_1(s) = \mathbb{E}s^{Z_1} = \mathbb{E}s^{X_i^{(m)}}$ for all i and m. Then

$$G_n(s) = G(G(\cdots(G(s))\cdots)) = G(G_{n-1}(s)) = G_{n-1}(G(s))$$

is the n-fold iteration of G.
This further implies

$$G_{m+n}(s) = G_m(G_n(s)) = G_n(G_m(s))$$

Proof.
When $n = 2$,

$$G_2(s) = \mathbb{E}s^{Z_2} = \mathbb{E}s^{X_1^{(1)} + X_2^{(1)} + \cdots + X_{Z_1}^{(1)}} = G_{Z_1}\left(G_{X_1^{(1)}}(s)\right) = G(G(s))$$

.
When $n = m + 1$ for some m,

$$G_{m+1}(s) = \mathbb{E}s^{Z_{m+1}} = \mathbb{E}s^{X_1^{(m)} + X_2^{(m)} + \cdots + X_{Z_m}^{(m)}} = G_{Z_m}\left(G_{X_1^{(m)}}(s)\right) = G_m(G(s))$$

□

In principle, the above theorem tells us the distribution of Z_n. However, it may not be easy to compute $G_n(s)$.
The moments of Z_n can be computed easier.

Lemma 5.16. Let $\mathbb{E}Z_1 = \mathbb{E}X_i^{(m)} = \mu$ and $\mathrm{Var}(Z_1) = \sigma^2$. Then

$$\mathbb{E}Z_n = \mu^n \qquad \mathrm{Var}(Z_n) = \begin{cases} n\sigma^2, & \mu = 1 \\ \frac{\sigma^2(\mu^n - 1)\mu^{n-1}}{\mu - 1}, & \mu \neq 1 \end{cases}$$

Proof.
Using Theorem 5.15, we can get

$$\begin{aligned}
\mathbb{E}Z_2 &= G_2'(1) = G'(G(1))G'(1) = G'(1)\mu = \mu^2 \\
\mathbb{E}Z_n &= G_n'(1) = G'(G_{n-1}(1))G_{n-1}'(1) = G'(1)\mu^{n-1} = \mu^n \\
G_1''(1) &= \sigma^2 + (G'(1))^2 - G'(1) = \sigma^2 + \mu^2 - \mu \\
G_2''(1) &= G''(G(1))(G'(1))^2 + G'(G(1))G''(1) = G''(1)(\mu^2 + \mu) \\
G_n''(1) &= G''(G_{n-1}(1))(G_{n-1}'(1))^2 + G'(G_{n-1}(1))G_{n-1}''(1) \\
&= (\sigma^2 + \mu^2 - \mu)\mu^{2n-2} + \mu G_{n-1}''(1) \\
&= \mu^{2n-2}(\sigma^2 + \mu^2 - \mu) + \mu^{2n-3}(\sigma^2 + \mu^2 - \mu) + \cdots + \mu^{n-1}(\sigma^2 + \mu^2 - \mu) \\
&= \frac{\mu^{n-1}(\sigma^2 + \mu^2 - \mu)(\mu^n - 1)}{\mu - 1}
\end{aligned}$$

If $\mu = 1$,

$$\mathrm{Var}(Z_n) = G_n''(1) + G_n'(1) - (G_n'(1))^2 = \sigma^2 + G_{n-1}''(1) + 1 - 1 = n\sigma^2$$

If $\mu \neq 1$,

$$\mathrm{Var}(Z_n) = G_n''(1) + G_n'(1) - (G_n'(1))^2 = \frac{\mu^{n-1}(\sigma^2 + \mu^2 - \mu)(\mu^n - 1)}{\mu - 1} + \mu^n - \mu^{2n} = \frac{\mu^{n-1}\sigma^2(\mu^n - 1)}{\mu - 1}$$

□

Example 5.12. Does this process eventually lead to extinction?
Note that

$$\{\text{ultimate extinction}\} = \bigcup_n \{Z_n = 0\} = \lim_{n\uparrow\infty} \{Z_n = 0\}$$

$$\mathbb{P}(\text{ultimate extinction}) = \mathbb{P}\left(\lim_{n\uparrow\infty} \{Z_n = 0\}\right) = \lim_{n\uparrow\infty} \mathbb{P}(Z_n = 0) = \lim_{n\uparrow\infty} G_n(0)$$

Let $\eta_n = G_n(0)$ and $\eta = \lim_{n\uparrow\infty} \eta_n$.

Theorem 5.17. We have that η is the smallest non-negative root of the equation

$$s = G(s)$$

Furthermore,

1. $\eta = 1$ if $\mu < 1$
2. $\eta < 1$ if $\mu > 1$
3. $\eta = 1$ if $\mu = 1$ and $\sigma^2 > 0$
4. $\eta = 0$ if $\mu = 1$ and $\sigma^2 = 0$

Proof.

$$\eta_n = G_n(0) = G(G_{n-1}(0)) = G(\eta_{n-1})$$

We know that η_n is bounded. Therefore, $\eta_n \uparrow \eta$ for some $\eta \in [0, 1]$.

$$\eta = \lim_{n\uparrow\infty} \eta_n = \lim_{n\uparrow\infty} G(\mu_{n-1}) = G\left(\lim_{n\uparrow\infty} \eta_{n-1}\right) = G(\eta)$$

Suppose that there exists another non-negative root ψ.

$$\begin{aligned}
\eta_1 &= G(0) \leq G(\psi) = \psi \\
\eta_2 &= G(\eta_1) \leq G(\psi) = \psi
\end{aligned}$$

By induction, $\eta_n \leq \psi$ for all n and therefore $\eta \leq \psi$. Therefore, η is the smallest non-negative root of the equation $s = G(s)$.

$$G''(s) = \sum_{i=2}^{\infty} i(i-1)s^{i-2}\mathbb{P}(Z_1 = i) \geq 0$$

Therefore, G is non-decreasing and also either convex or a straight line.
When $\mu \neq 1$, we can find that two curves $y = G(s)$ and $y = s$ intersect at $s = 1$ and $s = k \in \mathbb{R}$.
We know that $\eta \leq 1$ since η is the smallest root. In order to intersect at $s = \eta$, $G'(\eta) \leq 1$.
If $\mu = G'(1) < 1$, then $\eta = 1$. If $\mu = G'(1) > 1$, then $\eta = k$ such that $G'(k) \leq 1$.
In the case when $\mu = G'(1) = 1$, we need to further analyze whether $y = G(s)$ intersects $y = s$ at 1 point or infinite points.

$$\sigma^2 = G''(1) + G'(1) - (G'(1))^2 = G''(1)$$

If $\sigma^2 = G''(1) > 0$, then $\eta = 1$. If $\sigma^2 = G''(1) = 0$, then $\eta = 0$. □

5.3 Expectation revisited

Recall that the expectations are given respectively by

$$\mathbb{E}X = \begin{cases} \sum x f_X(x), & X \text{ is discrete} \\ \int x f_X(x)\, dx, & X \text{ is continuous} \end{cases}$$

We want a notation which incorporates both these cases. Suppose that X has a CDF F. We can rewrite the equations as

$$\mathbb{E}X = \begin{cases} \sum x\, dF_X(x), & dF_X(x) = F_X(x) - \lim_{y \uparrow x} F_X(y) = f_X(x) \\ \int x\, dF_X(x), & dF_X(x) = \frac{\partial F}{\partial x}\, dx = f_X(x)\, dx \end{cases}$$

Instead of using the regular Riemann integral, which cannot deal with discrete case, we can use the Riemann-Stieltjes integral, which is a generalization of the Riemann integral.

$$\int_a^b g(x)\, dx = \lim_{\max_i |x_{i+1} - x_i|} \sum_i g(x_i^*)(x_{i+1} - x_i) \qquad \int_a^b g(x)\, dF(x) = \lim_{\max_i |x_{i+1} - x_i|} \sum_i g(x_i^*)(F(x_{i+1}) - F(x_i))$$

if the limit does not depend on the choice of $x_i^* \in [x_i, x_{i+1})$.

Definition 5.18. Expectation of a random variable X is given by:

$$\mathbb{E}X = \int x\, dF_X$$

Lemma 5.19. If $g : \mathbb{R} \to \mathbb{R}$ such that $g(X)$ is also a random variable, then

$$\mathbb{E}(g(X)) = \int g(x)\, dF_X$$

Example 5.13. If g is regular (differentiable at every point and every values in the domain maps to a value in range), then

$$\sum_i g(x_i^*)(F(x_{i+1} - F(x_i)) \approx \sum_i g(x_i^*) f(x_i^*)(x_{i+1} - x_i) \approx \int g(x) f(x)\, dx$$

Example 5.14. In irregular case, assume that the function g is the Dirichlet function. That is

$$\mathbf{1}_{\mathbb{Q}}(x) = \begin{cases} 1, & x \in \mathbb{Q} \\ 0, & x \notin \mathbb{Q} \end{cases} \qquad \sum_i g(x_i^*)(F(x_{i+1}) - F(x_i)) = \sum_i g(x_i^*)(x_{i+1} - x_i)$$

Since the limit depends on the choice of x_i^*, Riemann-Stieltjes integral of $\mathbf{1}_{\mathbb{Q}}(x)$ with respect to $F(x) = x$ is not well defined. Therefore, $\mathbb{E}\mathbf{1}_{\mathbb{Q}}(X)$ cannot be defined as a Riemann-Stieltjes integral. However, on the other hand,

$$\mathbb{E}\mathbf{1}_{\mathbb{Q}}(X) = \mathbb{P}(\mathbf{1}_{\mathbb{Q}}(x) = 1) = \mathbb{P} \circ X^{-1}(\mathbb{Q} \cap [0, 1]) = 0$$

With this notation, we can also change how we define PGF.

Definition 5.20. Probability generating function of a random variable X is given by:

$$\mathbb{E}s^X = \int s^x\, dF_X$$

5.4 Moment generating function and Characteristic function

Now that we have unified the notations, we can now properly apply the probability generating function.
For a more general variables X, it is best if we substitute $s = e^t$. We get the following definition.

Definition 5.21. Moment generating function (MGF) of a random variable X is the function $M : \mathbb{R} \to [0, \infty)$ given by:

$$M_X(t) = \mathbb{E}(e^{tX}) = \int e^{tx}\, dF_X$$

Remark 5.21.1. The definition of MGF only requires replacing s by e^t in PGF. MGF is easier for computing moments, but less convenient for computing distribution.

Remark 5.21.2. MGFs are related to Laplace transforms.

We can easier get the following lemma.

Lemma 5.22. Given a MGF $M_X(t)$ of a random variable X.

1. For any $k \geq 0$,
$$\mathbb{E}X^k = M^{(k)}(0)$$
2. The function M can be expanded via Taylor's Theorem within its radius of convergence.
$$M(t) = \sum_{i=0}^{\infty} \frac{\mathbb{E}X^k}{k!} t^k$$
3. If X and Y are independent, then
$$M_{X+Y}(t) = M_X(t) M_Y(t)$$

Proof.

1. $$M^{(k)}(0) = \frac{\partial^k}{\partial t^k} \int e^{tx}\, dF_X(x)\Big|_{t=0} = \int x^k e^{tx}\, dF_X(x)\Big|_{t=0} = \int x^k\, dF_X(x) = \mathbb{E}X^k$$
2. Just using (1) and Taylor's Theorem and you get the answer.
3. This is just Theorem 5.8.

□

Remark 5.22.1. $M_X(0) = 1$ for all random variables X.

Example 5.15. Let $X \sim \text{Exp}(1)$. For all $x > 0$, if $t < 1$,

$$f_X(x) = e^{-x} \qquad M_X(t) = \int_0^\infty e^{tx}\, dF_X(x) = \int_0^\infty e^{(t-1)x}\, dx = \frac{1}{1-t}$$

Example 5.16. Let $X \sim$ Cauchy.

$$f_X(x) = \frac{1}{\pi(1+x^2)} \qquad M_X(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{e^{tx}}{1+x^2}\, dx$$

$M_X(t)$ exists only at $t = 0$. We get $M_X(0) = 1$.

Moment generating functions provide a useful technique but the integrals used to define may not be finite. There is another class of functions which finiteness is guaranteed.

Definition 5.23. Characteristic function (CF) of a random variable X is the function $\phi_X : \mathbb{R} \to \mathbb{C}$ given by:

$$\phi_X(t) = \mathbb{E}(e^{itX}) = \int e^{itx}\, dF_X(x) = \mathbb{E}\cos(tX) + i\mathbb{E}\sin(tX) \qquad i = \sqrt{-1}$$

Remark 5.23.1. $\phi_X(t)$ is essentially a Fourier Transform.

Lemma 5.24. CF ϕ_X of a random variable X has the following properties:

1. $\phi_X(0) = 1$. $|\phi_X(t)| \leq 1$ for all t
2. $\phi_X(t)$ is uniformly continuous

Proof.

1. For all t,

$$\phi_X(0) = \int dF_X(x) = 1$$
$$|\phi_X(t)| = \left|\int (\cos(tx) + i\sin(tx))\, dF_X(x)\right| \leq \int |\cos(tx) + i\sin(tx)|\, dF_X(x) = \int dF_X(x) = 1$$

2.
$$\sup_t |\phi_X(t+c) - \phi_X(t)| = \sup_t \left|\int (e^{i(t+c)x} - e^{itx})\, dF_X(x)\right| \leq \sup_t \left(\int \left|e^{itx}\right| \left|e^{icx-1}\right|\, dF_X(x)\right)$$

When $c \downarrow 0$, the supremum $\to 0$. Therefore, $\phi_X(t)$ is uniformly continuous.

□

Theorem 5.25. There are some properties of ϕ_X of a random variable X regarding derivatives and moments.

1. If $\phi_X^{(k)}(0)$ exists, then
$$\begin{cases} \mathbb{E}\,|X|^k < \infty, & k \text{ is even} \\ \mathbb{E}\,|X|^{k-1} < \infty, & k \text{ is odd} \end{cases}$$
2. If $\mathbb{E}\,|X|^k < \infty$, then $\phi_X^{(k)}(0)$ exists. We have
$$\phi_X(t) = \sum_{j=0}^{k} \frac{\phi_X^{(j)}(0)}{j!} t^j + o(t^k) = \sum_{j=0}^{k} \frac{\mathbb{E}X^j}{j!} (it)^j + o(t^k)$$

Proof.
We use the Taylor's Theorem.
$$\phi_X(t) = \sum_{j=0}^{k} \frac{\phi_X^{(j)}(0)}{j!} t^j + o(t^k) = \sum_{j=0}^{k} \frac{\mathbb{E}X^j}{j!} (it)^j + o(t^k)$$

1.
$$\phi_X^{(k)}(0) = i^k \mathbb{E}X^k$$
If k is even, we have $\phi_X^{(k)}(0) = (-1)^{\frac{k}{2}} \mathbb{E}X^k = (-1)^{\frac{k}{2}} \mathbb{E}\,|X|^k$ exists. Therefore, $\mathbb{E}\,|X|^k < \infty$.
If k is odd, we know that $\phi_X^{(k-1)}(0)$ exists if $\phi_X^{(k)}(0)$ exists.
Therefore, with $\phi_X^{(k-1)}(0) = (-1)^{\frac{k-1}{2}} \mathbb{E}X^{k-1} = (-1)^{\frac{k-1}{2}} \mathbb{E}\,|X|^{k-1}$, $\mathbb{E}\,|X|^{k-1} < \infty$.

2. Again using the formula in (1). We have
$$\frac{\phi_X^{(k)}(0)}{i^k} = \mathbb{E}X^k \leq \mathbb{E}\,|X|^k < \infty$$
Therefore, $\phi_X^{(k)}(0)$ exists. The formula can be obtained from the Taylor's theorem formula.

□

Theorem 5.26. If $X \perp\!\!\!\perp Y$, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$

Proof.

$$\phi_{X+Y}(t) = \mathbb{E}(e^{it(X+Y)}) = \mathbb{E}(e^{itX})\mathbb{E}(e^{itY}) = \phi_X(t)\phi_Y(t)$$

□

Again and again, we have a joint characteristic function.

Definition 5.27. Joint characteristic function (JCF) $\phi_{X,Y}$ of two random variables X, Y is given by

$$\phi_{X,Y}(s,t) = \mathbb{E}(e^{i(sX+tY)})$$

We have another way to prove that two random variables are independent.

Theorem 5.28. Two random variables X, Y are independent if and only if for all s and t,

$$\phi_{X,Y}(s,t) = \phi_X(s)\phi_Y(t)$$

Proof.
If $X \perp\!\!\!\perp Y$,

$$\phi_{X,Y}(s,t) = \mathbb{E}(e^{i(sX+tY)}) = \mathbb{E}(e^{isX})\mathbb{E}(e^{itY}) = \phi_X(s)\phi_Y(t)$$

Currently, it is not suffice to prove the inverse. We will need to use a theorem later. (Example 5.22) □

Example 5.17. Let $X \sim \text{Bern}(p)$. We have

$$\phi_X(t) = \mathbb{E}(e^{itX}) = q + pe^{it}$$

Example 5.18. Let $X \sim \text{Bin}(n,p)$. We have

$$\phi_X(t) = (q + pe^{it})^n$$

Example 5.19. Let $X \sim \text{Exp}(1)$. We have

$$\phi_X(t) = \int e^{(it-1)x}\,dx = \frac{1}{1-it}$$

Example 5.20. Let $X \sim \text{Cauchy}$. We have

$$\phi_X(t) = e^{-|t|}$$

Example 5.21. Let $X \sim \text{N}(\mu, \sigma^2)$. Using the fact that for any $u \in \mathbb{C}$, not just in $\mathbb{R}$,

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-u)^2}{2\sigma^2}\right) dx = 1$$

We have

$$\begin{aligned}
\phi_X(t) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{itx} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - (2\mu + 2\sigma^2 it)x + \mu^2}{2\sigma^2}\right) dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(\mu+\sigma^2 it)^2 - \mu^2}{2\sigma^2}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{(x-(\mu+\sigma^2 it))^2}{2\sigma^2}\right) dx \\
&= \exp\left(\frac{\mu^2 + 2\sigma^2 i\mu t - \sigma^4 t^2 - \mu^2}{2\sigma^2}\right) \\
&= \exp\left(i\mu t - \frac{1}{2}\sigma^2 t^2\right)
\end{aligned}$$

Remark 5.28.1. We have a function called **cumulant generating function** defined by $\log \phi_X(t)$. Normal distribution is the only distribution we have learnt whose cumulant generating function has finite terms, which is:

$$\log \phi_X(t) = i\mu t - \frac{1}{2}\sigma^2 t^2$$

5.5 Inversion and continuity theorems

There are two major ways that characteristic functions are useful. One of them is that we can use characteristic function of a random variable to generate a probability density function of that random variable.

Theorem 5.29. (Fourier Inverse Transform for continuous case) If a random variable X is continuous with a PDF f_X and a CF ϕ_X, then

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_X(t)\,dt$$

at all point x which f_X is differentiable.
If X has a CDF F_X, then

$$F_X(b) - F_X(a) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_a^b e^{-itx}\phi_X(t)\,dx\,dt$$

Proof.
We give you a non-rigorous proof. Let

$$I(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{itx}\phi_X(t)\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\int_{-\infty}^{\infty} e^{ity}f_X(y)\,dy\,dt$$
$$I_\varepsilon(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\int_{-\infty}^{\infty} e^{ity}f_X(y)\,dy\,e^{-\frac{1}{2}\varepsilon^2t^2}\,dt$$

We want to show that $I_\varepsilon(x) \to I(x)$ when $\varepsilon \downarrow 0$.

$$\begin{aligned}
I_\varepsilon(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\varepsilon^2t^2+i(y-x)t}f_X(y)\,dt\,dy \\
&= \frac{1}{\sqrt{2\pi\varepsilon^2}}\left(\frac{1}{\sqrt{2\pi\frac{1}{\varepsilon^2}}}\right)\int_{-\infty}^{\infty}\exp\left(-\frac{(y-x)^2}{2\varepsilon^2}\right)f_X(y)\int_{-\infty}^{\infty}\exp\left(-\frac{-\left(t-i\frac{y-x}{\varepsilon}\right)^2}{2\left(\frac{1}{\varepsilon^2}\right)}\right)dt\,dy \\
&= \frac{1}{\sqrt{2\pi\varepsilon^2}}\int_{-\infty}^{\infty}\exp\left(-\frac{(y-x)^2}{2\epsilon}\right)f_X(y)\,dy
\end{aligned}$$

Let $Z \sim \mathrm{N}(0,1)$ and $Z_\varepsilon = \varepsilon Z$. $I_\varepsilon(x)$ is the PDF of $\varepsilon Z + X$. Therefore, we can say that $f_{\varepsilon Z+X}(x) \to f_X(x)$ when $\varepsilon \downarrow 0$.
Note that this proof is not rigorous. □

Theorem 5.30. (Inversion Theorem) If a random variable X have a CDF F_X and a CF ϕ_X, we define $\overline{F}_X : \mathbb{R} \to [0,1]$ by

$$\overline{F}_X(x) = \frac{1}{2}\left(F_X(x) + F_X(x^-)\right)$$

Then for all $a \le b$,

$$\overline{F}_X(b) - \overline{F}_X(a) = \int_{-\infty}^{\infty}\frac{e^{-iat} - e^{-ibt}}{2\pi it}\phi_X(t)\,dt$$

Remark 5.30.1. We can say $\overline{F}_X$ represents the average of limit going from two directions.

Example 5.22. With the Inversion Theorem, we can now prove Theorem 5.28.
Given two random variables X, Y. We want to first extend the Fourier Inverse Transform into multivariable case.
If $\phi_{X,Y}(s,t) = \phi_X(s)\phi_Y(t)$, then for any $a \le b$ and $c \le d$,

$$\begin{aligned}
\overline{F}_{X,Y}(b,d) - \overline{F}_{X,Y}(b,c) - \overline{F}_{X,Y}(a,d) + \overline{F}_{X,Y}(a,c) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{(e^{-ias} - e^{-ibs})(e^{-ict} - e^{-idt})}{-4\pi^2t^2}\phi_X(s)\phi_Y(t)\,ds\,dt \\
&= (\overline{F}_X(b) - \overline{F}_X(a))\int_{-\infty}^{\infty}\frac{e^{-ict} - e^{-idt}}{2\pi it}\phi_Y(t)\,dt \\
&= (\overline{F}_X(b) - \overline{F}_X(a))(\overline{F}_Y(d) - \overline{F}_Y(c)) \\
&= \overline{F}_X(b)\overline{F}_Y(d) - \overline{F}_X(b)\overline{F}_Y(c) - \overline{F}_X(a)\overline{F}_Y(d) + \overline{F}_X(a)\overline{F}_Y(c)
\end{aligned}$$

From the definition of independent random variables, we prove that $X \perp\!\!\!\perp Y$ if $\phi_{X,Y}(s,t) = \phi_X(s)\phi_Y(t)$.

Another way is to evaluate the convergence of a sequence of cumulative distribution function.

Definition 5.31. (Convergence of distribution function sequence [Weak convergence]) A sequence of CDF $F_1, F_2, \cdots$ **converges** to a CDF F, written as $F_n \to F$, if at each point x where F is continuous,

$$F_n(x) \to F(x)$$

Example 5.23. Assume we have two sequences of CDF.

$$F_n(x) = \begin{cases} 0, & x < \frac{1}{n} \\ 1, & x \geq \frac{1}{n} \end{cases} \qquad G_n(x) = \begin{cases} 0, & x < -\frac{1}{n} \\ 1, & x \geq -\frac{1}{n} \end{cases}$$

If we have $n \to \infty$, we get

$$F(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases} \qquad G(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

This is problematic because $F(x)$ in this case is not a distribution function because it is not right-continuous.
Therefore, it is needed to define the convergence so that both sequences $\{F_n\}$ and $\{G_n\}$ have the same limit.

We can modify a bit on the definition to say each distribution function in the sequence represents a different random variable.

Definition 5.32. (Convergence in distribution for random variables) Let $X, X_1, X_2, \cdots$ be a family of random variables with PDF $F, F_1, F_2, \cdots$, we say $X_n \to X$, written as $X_n \xrightarrow{D} X$ or $X_n \Rightarrow X$, if $F_n \to F$.

Remark 5.32.1. For this convergence definition, we do not care about the closeness of X_n and X as functions of ω.

Remark 5.32.2. Sometimes, we also write $X_n \Rightarrow F$ or $X_n \xrightarrow{D} F$.

With the definition, sequence of characteristic functions can be used to determine whether the sequence of cumulative distribution function converges.

Theorem 5.33. (Lévy continuity theorem) Suppose that $F_1, F_2, \cdots$ is a sequence of CDF with CF $\phi_1, \phi_2, \cdots$, then

1. If $F_n \to F$ for some CDF F with CF ϕ, then $\phi_n \to \phi$ pointwise.
2. If $\phi_n \to \phi$ pointwise for some CF ϕ, and ϕ is continuous at O $(t = 0)$, then ϕ is the CF of some CDF F and $F_n \to F$.

We have a more general definition of convergence.

Definition 5.34. (**Vague convergence**) Given a sequence of CDF $F_1, F_2, \cdots$. Suppose that $F_n(x) \to G(x)$ at all continuity point of G but G may not be a CDF. Then we say $F_n \to G$ **vaguely**, written as $F_n \xrightarrow{v} G$.

Example 5.24. If

$$F_n(x) = \begin{cases} 0, & x < \frac{1}{n} \\ \frac{1}{2}, & \frac{1}{n} \leq x < n \\ 1, & x \geq n \end{cases} \qquad G(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2}, & x \geq 0 \end{cases}$$

We can see that $F_n \xrightarrow{v} G$ if $n \to \infty$ and G is not a CDF.

Remark 5.34.1. In Theorem 5.33 (2), the statement that ϕ is continuous at O can be replaced by any of the following statements:

1. $\phi(t)$ is a continuous function of t
2. $\phi(t)$ is a CF of some CDF
3. The sequence $\{F_n\}_{n=1}^{\infty}$ is tight, i.e. for all $\epsilon > 0$, there exists $M_\epsilon > 0$ such that

$$\sup_n (F_n(-M_\epsilon) + 1 - F_n(M_\epsilon)) \leq \epsilon$$

Example 5.25. Let $X_n \sim \mathrm{N}(0, n^2)$ and let ϕ_n be the CF of X_n. Then

$$\phi_n(t) = \exp\left(-\frac{1}{2}n^2t^2\right) \to \phi(t) = \begin{cases} 0, & t \neq 0 \\ 1, & t = 0 \end{cases}$$

5.6 Two limit theorems

In this section, we introduce two fundamental theorems in probability theory: the Law of Large Numbers and the Central Limit Theorem.

Theorem 5.35. (**Weak Law of Large Numbers** [WLLN]) Let $X_1, X_2, \cdots$ be i.i.d. random variables. Assume that $\mathbb{E}|X_1| < \infty$ and $\mathbb{E}X_1 = \mu$. Let $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{1}{n}S_n \xrightarrow{D} \mu$$

Proof.
We recall the Taylor expansion of $\phi_\xi(s)$ at 0. If $\mathbb{E}|\xi|^k < \infty$ and s is small, then

$$\phi_\zeta(s) = \sum_{j=0}^{k} \frac{\mathbb{E}\xi^j}{j!}(is)^j + o(s^k)$$

For any $t \in \mathbb{R}$, let $\phi_{X_1}(s) = \mathbb{E}(e^{isX_1})$.

$$\begin{aligned}\phi_n(t) = \mathbb{E}\left(\exp\left(\frac{it}{n}S_n\right)\right) = \mathbb{E}\left(\prod_{i=1}^n \exp\left(\frac{itX_i}{n}\right)\right) = \left(\mathbb{E}\left(\exp\left(\frac{itX_1}{n}\right)\right)\right)^n = \left(\phi_{X_1}\left(\frac{t}{n}\right)\right)^n = \left(1 + \frac{it}{n}\mathbb{E}X_1 + o\left(\frac{t}{n}\right)\right)^n \\ = \left(1 + \frac{i\mu t}{n} + o\left(\frac{t}{n}\right)\right)^n \\ \to e^{i\mu t}\end{aligned}$$

By Lévy continuity theorem, we get that $\frac{1}{n}S_n \xrightarrow{D} \mu$. □

Theorem 5.36. (**Central Limit Theorem** [CLT]) Let $X_1, X_2, \cdots$ be i.i.d. random variables with $\mathbb{E}|X_1|^2 < \infty$ and $\mathbb{E}X_1 = \mu$, $\mathrm{Var}(X_1) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{1}{\sigma}\sqrt{n}\left(\frac{1}{n}S_n - \mu\right) = \frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{D} \mathrm{N}(0,1)$$

Proof.
Let $Y_i = \frac{X_i - \mu}{\sigma}$. We have $\mathbb{E}Y_i = 0$ and $\mathrm{Var}(Y_i) = 1$.

$$\begin{aligned}\frac{S_n - n\mu}{\sqrt{n}\sigma} &= \sum_{i=1}^n \frac{1}{\sqrt{n}}\frac{X_i - \mu}{\sigma} = \sum_{i=1}^n \frac{Y_i}{\sqrt{n}} \\ \phi_n(t) &= \mathbb{E}\left(\exp\left(it\sum_{\ell-1}^n \frac{Y_\ell}{\sqrt{n}}\right)\right) \\ &= \left(\mathbb{E}\left(\exp\left(\frac{itY_1}{\sqrt{n}}\right)\right)\right)^n \\ &= \left(\phi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right)^n \\ &= \left(1 + \frac{it}{\sqrt{n}}\mathbb{E}Y_1 + \frac{1}{2}\left(\frac{it}{\sqrt{n}}\right)^2\mathbb{E}(Y_i^2) + o\left(\frac{t^2}{n}\right)\right)^n && \text{(Taylor expansion)} \\ &= \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \\ &\to e^{-\frac{1}{2}t^2}\end{aligned}$$

By Lévy continuity theorem, $\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{D} \mathrm{N}(0,1)$. □

Central Limit Theorem can be generalized in several directions, one of which concerns independent random variables instead of i.i.d. random variables.

Theorem 5.37. Let $X_1, X_2, \cdots$ be independent random variables satisfying $\mathbb{E}X_i = 0$, $\mathrm{Var}(X_i) = \sigma_i^2$, $\mathbb{E}\left|X_i\right|^3 < \infty$ and such that

$$\frac{1}{(\sigma(n))^3}\sum_{i=1}^{n}\mathbb{E}\left|X_i^3\right| \to 0 \text{ as } n \to \infty \tag{*}$$

where $(\sigma(n))^2 = \mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \sigma_i^2$. Then

$$\frac{1}{\sigma(n)}\sum_{i=1}^{n} X_i \xrightarrow{D} \mathrm{N}(0,1)$$

Remark 5.37.1. The condition (*) means that none of the random variables X_i can be significant in the sum S_n.

$$\frac{1}{(\sigma(n))^3}\sum_{i=1}^{n}|X_i|^3 \lesssim \frac{1}{\sigma(n)}\max_{i=1,2,\cdots,n}|X_i|\left(\frac{1}{(\sigma(n))^2}\right)\sum_{i=1}^{n}(X_i)^2 \approx \frac{1}{\sigma(n)}\max_{i=1,2,\cdots,n}|X_i| \to 0$$

Chapter 6

Markov chains (Skipped, read the book for reference)

Chapter 7

Convergence of Random Variables

In Chapter 5, we discussed convergence in distribution. However, this is not the only significant mode of convergence for random variables. In this chapter, we will explore other modes of convergence.

7.1 Modes of Convergence

We will discuss various modes of convergence for a sequence of random variables.
Let us first recall the convergence modes for real functions. Let $f, f_1, f_2, \cdots : [0,1] \to \mathbb{R}$.

1. Pointwise Convergence
   We say $f_n \to f$ pointwise if, for all $x \in [0,1]$,
   $$f_n(x) \to f(x) \text{ as } n \to \infty$$
2. Convergence in Norm $\|\cdot\|$
   We say $f_n \to f$ in norm $\|\cdot\|$ if
   $$\|f_n - f\| \to 0 \text{ as } n \to \infty$$
3. Convergence in Lebesgue (Uniform) Measure
   We say $f_n \to f$ in uniform measure μ if, for all $\epsilon > 0$,
   $$\mu\left(\{x \in [0,1] : |f_n(x) - f(x)| > \epsilon\}\right) \to 0 \text{ as } n \to \infty$$

These definitions can be extended to define convergence modes for random variables.

Definition 7.1. (**Almost Sure Convergence**) We say $X_n \to X$ **almost surely**, denoted as $X_n \xrightarrow{\text{a.s.}} X$, if

$$\mathbb{P}(\{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}) = 1 \qquad \text{or} \qquad \mathbb{P}(\{\omega \in \Omega : X_n(\omega) \not\to X(\omega) \text{ as } n \to \infty\}) = 0$$

Remark 7.1.1. $X_n \xrightarrow{\text{a.s.}} X$ is an adaptation of pointwise convergence for functions.

Remark 7.1.2. Almost sure convergence is often referred to as:

1. $X_n \to X$ almost everywhere ($X_n \xrightarrow{\text{a.e.}} X$)
2. $X_n \to X$ with probability 1 ($X_n \to X$ w.p. 1)

Definition 7.2. (Convergence in r-th Mean) Let $r \geq 1$. We say $X_n \to X$ **in r-th mean**, denoted as $X_n \xrightarrow{r} X$, if

$$\mathbb{E}\,|X_n - X|^r \to 0 \text{ as } n \to \infty$$

Example 7.1. If $r = 1$, we say $X_n \to X$ in mean or expectation.
If $r = 2$, we say $X_n \to X$ in mean square.

Definition 7.3. (Convergence in Probability) We say $X_n \to X$ **in probability**, denoted as $X_n \xrightarrow{\mathbb{P}} X$, if, for all $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| > \varepsilon) \to 0 \text{ as } n \to \infty$$

Definition 7.4. (Convergence in Distribution) We say $X_n \to X$ **in distribution**, denoted as $X_n \xrightarrow{D} X$, if, at continuity points of $\mathbb{P}(X \leq x)$,

$$F_n(x) = \mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x) = F(x) \text{ as } n \to \infty$$

Before exploring the relationships between different convergence modes, we first introduce some key formulas.

Lemma 7.5. (Markov's Inequality) If X is any random variable with a finite mean, then for all $a > 0$,

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}\,|X|}{a}$$

Proof.

$$\mathbb{P}(|X| \geq a) = \mathbb{E}(\mathbf{1}_{|X| \geq a}) \leq \mathbb{E}\left(\frac{|X|}{a}\mathbf{1}_{|X|>a}\right) \leq \frac{\mathbb{E}\,|X|}{a}$$

□

Remark 7.5.1. For any non-negative function φ that is increasing on $[0, \infty)$,

$$\mathbb{P}(|X| \geq a) = \mathbb{P}(\varphi(|X|) \geq \varphi(a)) \leq \frac{\mathbb{E}(\varphi(|X|))}{\varphi(a)}$$

The following inequality requires Hölder's inequality (see Appendix C) for its proof. Therefore, we will not prove it here.

Lemma 7.6. (Lyapunov's Inequality) Let Z be any random variable. For all $r \geq s > 0$,

$$(\mathbb{E}\,|Z|^s)^{\frac{1}{s}} \leq (\mathbb{E}\,|Z|^r)^{\frac{1}{r}}$$

We also need to understand how to achieve almost sure convergence.

Lemma 7.7. Let

$$A_n(\varepsilon) = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\} \qquad B_m(\varepsilon) = \bigcup_{n=m}^{\infty} A_n(\varepsilon)$$

We have

1. $X_n \xrightarrow{\text{a.s.}} X$ if and only if $\lim_{m\uparrow\infty} \mathbb{P}(B_m(\varepsilon)) = 0$ for all $\varepsilon > 0$.
2. $X_n \xrightarrow{\text{a.s.}} X$ if $\sum_{n=1}^{\infty} \mathbb{P}(A_n(\varepsilon)) < \infty$ for all $\varepsilon > 0$.

Proof.

1. We denote $C = \{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}$.
   If $\omega \in C$, it means that for all $\varepsilon > 0$, there exists $n_0 > 0$ such that $|X_n(\omega) - X(\omega)| \leq \varepsilon$ for all $n \geq n_0$.
   This also implies that for all $\varepsilon > 0$, $|X_n(\omega) - X(\omega)| > \varepsilon$ for finitely many n.
   If $\omega \in C^{\complement}$, it means that for all $\varepsilon > 0$, $|X_n(\omega) - X(\omega)| > \varepsilon$ for infinitely many n. ($\omega \in \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n(\varepsilon)$)
   Therefore,
   $$C^{\complement} = \bigcup_{\varepsilon>0} \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n(\varepsilon)$$
   If $\mathbb{P}(C^{\complement}) = 0$, then for all $\varepsilon > 0$,
   $$\mathbb{P}\left(\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n(\varepsilon)\right) = 0$$
   We can also find that
   $$\mathbb{P}\left(\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n(\varepsilon)\right) = 0 \qquad \Longrightarrow \qquad \mathbb{P}(C^{\complement}) = \mathbb{P}\left(\bigcup_{\varepsilon>0} \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n(\varepsilon)\right) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n\left(\frac{1}{k}\right)\right) = 0$$
   Therefore, $X_n \xrightarrow{\text{a.s.}} X$ if and only if $\lim_{m\uparrow\infty} \mathbb{P}(B_m(\varepsilon)) = 0$ for all $\varepsilon > 0$.
2. From (1), for all $\varepsilon > 0$,
   $$\sum_{n=1}^{\infty} \mathbb{P}(A_n(\varepsilon)) < \infty \implies \lim_{m\to\infty} \sum_{n=m}^{\infty} \mathbb{P}(A_n(\varepsilon)) = 0 \implies \lim_{m\to\infty} \mathbb{P}(B_m(\varepsilon)) = 0 \implies (X_n \xrightarrow{\text{a.s.}} X)$$

□

Lemma 7.8. There exist sequences that

1. converge almost surely but not in mean.
2. converge in mean but not almost surely.

Proof.

1. We consider
$$X_n = \begin{cases} n^3, & \text{Probability } = n^{-2} \\ 0, & \text{Probability } = 1 - n^{-2} \end{cases}$$
By applying Lemma 7.7, for some $\varepsilon > 0$.
$$\mathbb{P}(|X_n(\omega) - X(\omega)| > \varepsilon) = \frac{1}{n^2} \qquad \sum_{n=1}^{\infty} \mathbb{P}(|X_n(\omega) - X(\omega)| > \varepsilon) < \infty$$
Therefore, the sequence converges almost surely. However,
$$\mathbb{E}\,|X_n - X| = n^3 \left(\frac{1}{n^2}\right) = n \to \infty$$
Therefore, the sequence does not converge in mean.

2. We consider
$$X_n = \begin{cases} 1, & \text{Probability } = n^{-1} \\ 0, & \text{Probability } = 1 - n^{-1} \end{cases}$$
In mean, as $n \to \infty$ we have
$$\mathbb{E}\,|X_n - X| = 1 \left(\frac{1}{n}\right) = \frac{1}{n} \to 0$$
However, by applying Lemma 7.7, if $\varepsilon \in (0, 1)$, for all n
$$\begin{aligned} \mathbb{P}(B_m(\varepsilon)) &= 1 - \lim_{r\to\infty} \mathbb{P}(X_n = 0 \text{ for all } n \text{ such that } m \le n \le r) \\ &= 1 - \lim_{r\to\infty} \prod_{i=m}^{r} \frac{i-1}{i} \\ &= 1 - \lim_{r\to\infty} \frac{m-1}{r} \to 1 \neq 0 \end{aligned}$$
Therefore, the sequence does not converge almost surely.

□

We can now deduce the following implications. Roughly speaking, convergence in distribution is the weakest among all convergence modes, since it only cares about the distribution of X_n.

Theorem 7.9. The following implications hold:

1. (a) $(X_n \xrightarrow{\text{a.s.}} X) \implies (X_n \xrightarrow{\mathbb{P}} X)$
   (b) $(X_n \xrightarrow{r} X) \implies (X_n \xrightarrow{\mathbb{P}} X)$
   (c) $(X_n \xrightarrow{\mathbb{P}} X) \implies (X_n \xrightarrow{D} X)$
2. If $r \ge s \ge 1$, then $(X_n \xrightarrow{r} X) \implies (X_n \xrightarrow{s} X)$
3. No other implications holds in general.

Proof.

1. (a) From Lemma 7.7, for all $\varepsilon > 0$,
$$\mathbb{P}(A_m(\varepsilon)) \le \mathbb{P}\left(\bigcup_{n=m}^{\infty} A_n(\varepsilon)\right) = \mathbb{P}(B_m(\varepsilon)) \to 0$$
Therefore, $(X_n \xrightarrow{\text{a.s.}} X) \implies (X_n \xrightarrow{\mathbb{P}} X)$

   (b) From Markov's inequality, since $r \ge 1$,
$$0 \le \mathbb{P}(|X - X_n| > \varepsilon) = \mathbb{P}(|X - X_n|^r > \varepsilon^r) \le \frac{\mathbb{E}\,|X_n - X|^r}{\varepsilon^r}$$
Therefore, if $X_n \xrightarrow{r} X$, then $\mathbb{E}\,|X_n - X|^r \to 0$. We have $\mathbb{P}(|X - X_n| > \varepsilon) \to 0$ and thus $X_n \xrightarrow{\mathbb{P}} X$.

   (c)
$$\begin{aligned}
&\mathbb{P}(X_n \le x) = \mathbb{P}(X_n \le x, X \le x + \varepsilon) + \mathbb{P}(X_n \le x, X > x + \varepsilon) \le \mathbb{P}(X \le x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon)\\
&\mathbb{P}(X \le y) \le \mathbb{P}(X_n \le y + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon)\\
&\mathbb{P}(X_n \le x) \ge \mathbb{P}(X \le x - \varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \qquad (y = x - \varepsilon)
\end{aligned}$$
Since $X_n \xrightarrow{\mathbb{P}} X$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ for all $\varepsilon > 0$. Therefore,
$$\mathbb{P}(X \le x - \varepsilon) \le \liminf_{n\to\infty} \mathbb{P}(X_n \le x) \le \limsup_{n\to\infty} \mathbb{P}(X_n \le x) \le \mathbb{P}(X \le x + \varepsilon)$$
By having $\varepsilon \downarrow 0$,
$$\mathbb{P}(X \le x) \le \liminf_{n\to\infty} \mathbb{P}(X_n \le x) \le \limsup_{n\to\infty} \mathbb{P}(X_n \le x) \le \mathbb{P}(X \le x)$$
Therefore, $\lim_{n\to\infty} \mathbb{P}(X_n \le x) = \mathbb{P}(X \le x)$ and thus $X_n \xrightarrow{D} X$.

2. Since $X_n \xrightarrow{r} X$, $\mathbb{E}\,|X_n - X| \to 0$ as $n \to \infty$. By Lyapunov's inequality, if $r \ge s$,
$$\mathbb{E}\,|X_n - X|^s \le (\mathbb{E}\,|X_n - X|^r)^{\frac{s}{r}} \to 0$$

3. Let $\Omega = \{H, T\}$ and $\mathbb{P}(H) = \mathbb{P}(T) = \frac{1}{2}$. Let
$$X_{2m}(\omega) = \begin{cases} 1, & \omega = H \\ 0, & \omega = T \end{cases} \qquad X_{2m+1}(\omega) = \begin{cases} 0, & \omega = H \\ 1, & \omega = T \end{cases}$$
Since $F(x)$ and $F_n(x)$ for all n are all the same, $X_n \xrightarrow{D} X$. However, for $\varepsilon \in [0, 1]$, $\mathbb{P}(|X_n - X| > \varepsilon) \not\to 0$.
Therefore, $(X_n \xrightarrow{D} X) \not\implies (X_n \xrightarrow{\mathbb{P}} X)$.
Let $r = 1$ and
$$X_n = \begin{cases} n, & \text{probability } = \frac{1}{n} \\ 0, & \text{probability } = 1 - \frac{1}{n} \end{cases} \qquad X = 0$$
We get that $\mathbb{P}(|X_n - X| > \varepsilon) = \frac{1}{n} \to 0$. However, $\mathbb{E}\,|X_n - X| = n\left(\frac{1}{n}\right) = 1 \not\to 0$. Therefore, $(X_n \xrightarrow{\mathbb{P}} X) \not\implies (X_n \xrightarrow{r} X)$.
Let $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}([0, 1])$ and $\mathbb{P}$ be uniform.
Let I_i be such that $I_{\frac{1}{2}m(m-1)+1}, I_{\frac{1}{2}m(m-1)+2}, \cdots, I_{\frac{1}{2}m(m-1)+m}$ is a partition of $[0, 1]$ for all m.
We have $I_1 = [0, 1], I_2 \cup I_3 = [0, 1], \cdots$. Let
$$X_n(\omega) = \mathbf{1}_{I_n(\omega)} = \begin{cases} 1, & \omega \in I_n \\ 0, & \omega \in I_n^{\complement} \end{cases} \qquad X(\omega) = 0 \text{ for all } \omega \in \Omega$$
For all $\varepsilon \in [0, 1]$, $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(I_n) = \frac{1}{n} \to 0$ for some n if $n \to \infty$.
However, for any given $\omega \in \Omega$, although 1 becomes less often due to decreasing probability, it never dies out.
Therefore, $X_n(\omega) \not\to 0 = X(\omega)$ and $\mathbb{P}(\{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}) = 0$, and thus, $(X_n \xrightarrow{\mathbb{P}} X) \not\implies (X_n \xrightarrow{\text{a.s.}} X)$.
If $r \ge s \ge 1$, let
$$X_n = \begin{cases} n, & \text{probability } = n^{-\left(\frac{r+s}{2}\right)} \\ 0, & \text{probability } = 1 - n^{-\left(\frac{r+s}{2}\right)} \end{cases} \qquad X = 0$$
$$\mathbb{E}\,|X_n - X|^s = n^s\left(n^{-\left(\frac{r+s}{2}\right)}\right) = n^{\frac{s-r}{2}} \to 0 \qquad \mathbb{E}\,|X_n - X|^r = n^r\left(n^{-\left(\frac{r+s}{2}\right)}\right) = n^{\frac{r-s}{2}} \to \infty$$
Therefore, if $r \ge s \ge 1$, $(X_n \xrightarrow{s} X) \not\implies (X_n \xrightarrow{r} X)$.
We have proven that $(X_n \xrightarrow{\text{a.s.}} X) \not\implies (X_n \xrightarrow{r} X)$ and $(X_n \xrightarrow{r} X) \not\implies (X_n \xrightarrow{\text{a.s.}} X)$ in Lemma 7.8.

□

By applying Theorem 7.9, we can easily obtain this lemma.

Lemma 7.10. The following implications hold:

1. $(X_n \xrightarrow{1} X) \implies (X_n \xrightarrow{\mathbb{P}} X)$

Some of the implications do not hold in general but they hold if we apply some restrictions.

Theorem 7.11. (Partial Converse Statements) The following implications hold:

1. If $X_n \xrightarrow{D} c$, where c is a constant, then $X_n \xrightarrow{\mathbb{P}} c$.
2. If $X_n \xrightarrow{\mathbb{P}} X$ and $\mathbb{P}(|X_n| \leq k) = 1$ for all n with some fixed constant $k > 0$, then $X_n \xrightarrow{r} X$ for all $r \geq 1$.

Proof.

1. Since $X_n \xrightarrow{D} X$, $\mathbb{P}(X_n \leq x) \to \mathbb{P}(c \leq x)$ as $n \to \infty$. For all $\varepsilon > 0$,

$$\mathbb{P}(|X_n - c| \geq \varepsilon) = \mathbb{P}(X_n \leq c - \varepsilon) + \mathbb{P}(X_n \geq c + \varepsilon) = \mathbb{P}(X_n \leq c - \varepsilon) + 1 - \mathbb{P}(X_n < c + \varepsilon)$$

We can get that $\mathbb{P}(X_n \leq c - \varepsilon) \to \mathbb{P}(c \leq c - \varepsilon) = 0$. For $\mathbb{P}(X_n < c + \varepsilon)$,

$$\mathbb{P}\left(X_n \leq c + \frac{\varepsilon}{2}\right) \leq \mathbb{P}(X_n < c + \varepsilon) \leq \mathbb{P}(X_n \leq c + 2\varepsilon)$$

$$\mathbb{P}\left(X_n \leq c + \frac{\varepsilon}{2}\right) \to \mathbb{P}\left(c \leq c + \frac{\varepsilon}{2}\right) = 1 \qquad \mathbb{P}(X_n \leq c + 2\varepsilon) \to \mathbb{P}(c \leq c + 2\varepsilon) = 1$$

Therefore, $\mathbb{P}(X_n < c + \varepsilon) \to 1$. We have

$$\mathbb{P}(|X_n - c| \geq \varepsilon) \to 0 + 1 - 1 = 0$$

Therefore, $X_n \xrightarrow{\mathbb{P}} c$.

2. Since $X_n \xrightarrow{\mathbb{P}} X$, $X_n \xrightarrow{D} X$. We have $\mathbb{P}(|X_n| \leq k) \to \mathbb{P}(|X| \leq k) = 1$.
Therefore, for all $\varepsilon > 0$, if $|X_n - X| \leq \varepsilon$, $|X_n - X| \leq |X_n| + |X| \leq 2k$.

$$\begin{aligned}
\mathbb{E}\,|X_n - X|^r &= \mathbb{E}\left(|X_n - X|^r \mathbf{1}_{|X_n - X| \leq \varepsilon}\right) + \mathbb{E}\left(|X_n - X|^r \mathbf{1}_{|X_n - X| > \varepsilon}\right) \\
&\leq \varepsilon^r \mathbb{E}\left(\mathbf{1}_{|X_n - X| \leq \varepsilon}\right) + (2k)^r \mathbb{E}\left(\mathbf{1}_{|X_n - X| > \varepsilon}\right) \\
&\leq \varepsilon^r + ((2k)^r - \varepsilon^r)\mathbb{P}(|X_n - X| > \varepsilon)
\end{aligned}$$

Since $X_n \xrightarrow{\mathbb{P}} X$, as $n \to \infty$, $\mathbb{E}\,|X_n - X|^r \to \varepsilon^r$. If we send $\varepsilon \downarrow 0$, $\mathbb{E}\,|X_n - X|^r \to 0$ and therefore $X_n \xrightarrow{r} X$.

□

Note that any sequence $\{X_n\}$ which satisfies $X_n \xrightarrow{\mathbb{P}} X$ necessarily contains a subsequence $\{X_{n_i} : 1 \leq i < \infty\}$ which converges almost surely.

Theorem 7.12. If $X_n \xrightarrow{\mathbb{P}} X$, then there exists a non-random increasing sequence of integers $n_1, n_2, \cdots$ such that as $i \to \infty$,

$$X_{n_i} \xrightarrow{\text{a.s.}} X$$

Proof.
Since $X_n \xrightarrow{\mathbb{P}} X$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$ for all $\varepsilon > 0$.
We can pick an increasing sequence $n_1, n_2, \cdots$ of positive integers such that

$$\mathbb{P}(|X_{n_i} - X| > i^{-1}) \leq i^{-2}$$

For any $\varepsilon > 0$,

$$\sum_{i > \varepsilon^{-1}} \mathbb{P}(|X_{n_i} - X| > \varepsilon) \leq \sum_{i > \varepsilon^{-1}} \mathbb{P}(|X_{n_i} - X| > i^{-1}) \leq \sum_i i^{-2} < \infty$$

By Lemma 7.7, we get the $X_{n_i} \xrightarrow{\text{a.s.}} X$ as $i \to \infty$ □

7.2 Other Versions of the Weak Law of Large Numbers

Let us revisit and introduce additional versions of the Weak Law of Large Numbers (WLLN) and their applications.

Theorem 7.13. (L^2-WLLN) Let $X_1, X_2, \cdots, X_n$ be uncorrelated random variables with $\mathbb{E}X_i = \mu$ and $\mathrm{Var}(X_i) \le c < \infty$ for all i. Let $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{S_n}{n} \xrightarrow{2} \mu$$

Proof.

$$\mathbb{E}\left(\frac{S_n}{n} - \mu\right)^2 = \frac{\mathbb{E}(S_n - \mathbb{E}S_n)^2}{n^2} = \frac{1}{n^2}\mathrm{Var}(S_n) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) \le \frac{c}{n} \to 0$$

Therefore, $\frac{S_n}{n} \xrightarrow{2} \mu$. □

Remark 7.13.1. From this theorem, we can immediately conclude that

$$\left(\frac{S_n}{n} \xrightarrow{2} \mu\right) \implies \left(\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu\right)$$

Remark 7.13.2. Note that in the i.i.d. case, the existence of variance is not required.

There are numerous applications of the Weak Law of Large Numbers.

Example 7.2. (Bernstein Approximation) Let f be a continuous function on $[0,1]$, and define

$$f_n(x) = \sum_{m=0}^n \binom{n}{m} x^m (1-x)^{n-m} f\left(\frac{m}{n}\right). \qquad \text{(Bernstein Polynomial)}$$

We aim to show that as $n \to \infty$,

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

Remark 7.13.3. Let $x \in [0,1]$. To better approach this question, we can let $X_{1,x}, X_{2,x}, \cdots, X_{n,x} \sim \mathrm{Bern}(x)$ be i.i.d. random variables. Let $S_{n,x} = \sum_{i=1}^n X_{i,x} \sim \mathrm{Bin}(n,x)$.

$$\mathbb{P}(S_{n,x} = m) = \binom{n}{m} x^m (1-x)^{n-m},$$

$$f_n(x) = \sum_{m=0}^n \mathbb{P}(S_{n,x} = m) f\left(\frac{m}{n}\right) = \mathbb{E}\left(f\left(\frac{S_{n,x}}{n}\right)\right).$$

By the WLLN, $\frac{S_{n,x}}{n} \xrightarrow{\mathbb{P}} x$.

Remark 7.13.4. (Continuous Mapping Theorem) Let f be a uniformly continuous function. For all $\varepsilon > 0$, there exists δ_ε such that

$$\text{if} \quad \left|\frac{S_{n,x}}{n} - x\right| \le \delta_\varepsilon, \quad \text{then} \quad \left|f\left(\frac{S_{n,x}}{n}\right) - f(x)\right| \le \varepsilon.$$

By the contrapositive,

$$\mathbb{P}\left(\omega \in \Omega : \left|f\left(\frac{S_{n,x}(\omega)}{n}\right) - f(x)\right| > \varepsilon\right) \le \mathbb{P}\left(\omega \in \Omega : \left|\frac{S_{n,x}(\omega)}{n} - x\right| > \delta_\varepsilon\right) \to 0.$$

From this, we conclude that $f\left(\frac{S_{n,x}}{n}\right) \xrightarrow{\mathbb{P}} f(x)$.

Note: For non-uniformly continuous functions, the analysis is more complex. Further research is recommended.

Example 7.3. By establishing that $f\left(\frac{S_{n,x}}{n}\right) \xrightarrow{\mathbb{P}} f(x)$, and noting that there exists a constant M such that $\|f\|_\infty \le M$ (due to f being continuous on $[0,1]$), we have

$$\begin{aligned}\left|\mathbb{E}\left(f\left(\frac{S_{n,x}}{n}\right)\right) - f(x)\right| &\le \mathbb{E}\left|f\left(\frac{S_{n,x}}{n}\right) - f(x)\right| = \mathbb{E}\left(\left|f\left(\frac{S_{n,x}}{n}\right) - f(x)\right| \mathbf{1}_{\left|\frac{S_{n,x}}{n}-x\right| \le \delta_\varepsilon}\right) + \mathbb{E}\left(\left|f\left(\frac{S_{n,x}}{n}\right) - f(x)\right| \mathbf{1}_{\left|\frac{S_{n,x}}{n}-x\right| > \delta_\varepsilon}\right) \\ &\le \varepsilon + 2M\mathbb{P}\left(\left|\frac{S_{n,x}}{n} - x\right| > \delta_\varepsilon\right)\end{aligned}$$

$$\begin{aligned}\sup_{x\in[0,1]}\left|\mathbb{E}\left(f\left(\frac{S_{n,x}}{n}\right)\right) - f(x)\right| &= \varepsilon + 2M \sup_{x\in[0,1]}\left(\mathbb{P}\left(\left|\frac{S_{n,x}-nx}{n}\right| > \delta_\varepsilon\right)\right) \\ &\le \varepsilon + 2M \sup_{x\in[0,1]}\left(\frac{\mathbb{E}\left|S_{n,x}-nx\right|^2}{n^2\delta_\varepsilon^2}\right) \qquad \text{(Markov's Inequality and Lyapunov's Inequality)} \\ &\le \varepsilon + 2M \sup_{x\in[0,1]}\left(\frac{\mathrm{Var}(S_{n,x})}{n^2\delta_\varepsilon^2}\right) = \varepsilon + 2M \sup_{x\in[0,1]}\left(\frac{x(1-x)}{n\delta_\varepsilon^2}\right) \qquad (\mathbb{E}S_{n,x} = nx) \\ &\le \varepsilon + \frac{M}{2n\delta_\varepsilon^2}.\end{aligned}$$

$$\limsup_{n\to\infty}\sup_{x\in[0,1]}\left|\mathbb{E}\left(f\left(\frac{S_{n,x}}{n}\right)\right) - f(x)\right| \le \varepsilon \to 0.$$

Therefore, we conclude that $\sup_{x\in[0,1]}|f_n(x) - f(x)| \to 0$ as $n \to \infty$.

Example 7.4. (Borel's Geometric Concentration) Let μ_n be the uniform probability measure on the n-dimensional cube $[-1,1]^n$. Let $\mathcal{H}$ be a hyperplane that is orthogonal to a principal diagonal of $[-1,1]$ ($\mathcal{H} = (1,\cdots,1)^\perp$).
Let $\mathcal{H}_r = \{x \in [-1,1]^n : \mathrm{dist}(x:\mathcal{H}) \le r\}$. Then for any given $\varepsilon > 0$, $\mu_n(\mathcal{H}_{\varepsilon\sqrt{n}}) \to 1$ as $n \to \infty$.
We can prove this by letting $X_1, X_2, \cdots \sim \mathrm{U}[-1,1]$ be i.i.d. random variables and $\mathbb{E}X_i = 0$. Let $X = (X_1, X_2, \cdots, X_n)$.
For all $B \in [-1,1]^n$, $\mu_n(B) = \mathbb{P}(X \in B) = \mathbb{P} \circ X^{-1}(B)$.

$$\begin{aligned}\mu_n(\mathcal{H}_{\varepsilon\sqrt{n}}) &= \mathbb{P}(\mathrm{dist}(X,\mathcal{H}) \le \varepsilon\sqrt{n}) \\ &= \mathbb{P}\left(\frac{|\langle X,(1,\cdots,1)\rangle|}{\|(1,\cdots,1)\|_2} \le \varepsilon\sqrt{n}\right) \\ &= \mathbb{P}\left(\left|\frac{\sum_{i=1}^n X_i}{n}\right| \le \varepsilon\right) \\ &= \mathbb{P}\left(\left|\frac{S_n}{n} - \mathbb{E}X_1\right| \le \varepsilon\right) \\ &\to 1 \qquad \text{(WLLN)}\end{aligned}$$

We do not necessarily need to stick to a given sequence of random variables $X_1, X_2, \cdots$ in the Law of Large Numbers.

Theorem 7.14. (WLLN for Triangular Array) Let $\{X_{n,j}\}_{1\le j\le n<\infty}$ be a triangular array. Let $S_n = \sum_{i=1}^n X_{n,i}$, $\mu_n = \mathbb{E}S_n$ and $\sigma_n^2 = \mathrm{Var}(S_n)$. Suppose that for some sequence b_n,

$$\frac{\sigma_n^2}{b_n^2} = \mathbb{E}\left(\frac{S_n - \mu_n}{b_n}\right)^2 \to 0$$

Then we have

$$\frac{S_n - \mu_n}{b_n} \xrightarrow{\mathbb{P}} 0$$

Proof.

$$\mathbb{E}\left(\frac{S_n - \mu_n}{b_n}\right)^2 = \frac{\mathrm{Var}(S_n)}{b_n^2} \to 0$$

Therefore, $\frac{S_n-\mu_n}{b_n} \xrightarrow{2} 0$ and thus $\frac{S_n-\mu_n}{b_n} \xrightarrow{\mathbb{P}} 0$. □

Remark 7.14.1. We should choose b_n that is no larger than $\mathbb{E}S_n$ if possible.

Example 7.5. (Coupon Collector's Problem) Let $X_1, X_2, \cdots$ be i.i.d. uniform random variables on $\{1, 2, \cdots, n\}$.
Let $\tau_k^n = \inf\{m : |\{X_1, X_2, \cdots, X_m\}| = k\}$ be the waiting time for picking k distinct types.
What is the asymptotic behavior of τ_n^n?
It is easy to see that $\tau_1^n = 1$. By convention, $\tau_0^n = 0$.
For $1 \le k \le n$, let $X_{n,k} = \tau_k^n - \tau_{k-1}^n$ be the additional waiting time for picking k distinct types when we have $k-1$ types.
Notice that

$$\tau_n^n = \sum_{k=1}^{n} X_{n,k}$$

We know that

$$\mathbb{P}(X_{n,k} = \ell) = \left(\frac{k-1}{n}\right)^{\ell-1}\left(1 - \frac{k-1}{n}\right) \qquad \Longrightarrow \qquad X_{n,k} \sim \text{Geom}\left(1 - \frac{k-1}{n}\right)$$

We claim that $X_{n,k}$ are independent for all k. For a constant c,

$$\mathbb{E}\tau_n^n = \sum_{k=1}^{n} \mathbb{E}X_{n,k} = \sum_{k=1}^{n}\left(1 - \frac{k-1}{n}\right)^{-1} = \sum_{m=1}^{n} \frac{n}{m} \sim n\log n$$

$$\text{Var}(\tau_n^n) = \sum_{k=1}^{n} \text{Var}(X_{n,k}) = \sum_{k=1}^{n}\left(\left(1 - \frac{k-1}{n}\right)^{-2} - \left(1 - \frac{k-1}{n}\right)^{-1}\right) \le \sum_{k=1}^{n}\left(1 - \frac{k-1}{n}\right)^{-2} = \sum_{m=1}^{n}\frac{n^2}{m^2} \le cn^2$$

By WLLN, if we choose $b_n = n\log n$, then we have

$$\frac{\text{Var}(\tau_n^n)}{b_n^2} \to 0 \implies \frac{\tau_n^n - \sum_{m=1}^{n}\frac{n}{m}}{n\log n} \xrightarrow{\mathbb{P}} 0$$

Therefore, $\frac{\tau_n^n}{n\log n} \xrightarrow{\mathbb{P}} 1$

Example 7.6. (An Occupancy Problem) r balls are put at random into n bins. All n^r configurations are equally likely.
Let A_i be the event that the i-th bin is empty, and let N_n be the number of empty bins $= \sum_{i=1}^{n} \mathbf{1}_{A_i}$.
How can we prove that if $\frac{r}{n} \to c$ as $n \to \infty$,

$$\frac{N_n}{n} \xrightarrow{\mathbb{P}} e^{-c}?$$

We can see that

$$\begin{aligned}
\frac{\mathbb{E}N_n}{n} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\mathbf{1}_{A_i} = \mathbb{P}(A_i) = \left(1 - \frac{1}{n}\right)^r \to e^{-c}, \\
\text{Var}(N_n) &= \mathbb{E}(N_n^2) - (\mathbb{E}N_n)^2 \\
&= \mathbb{E}\left(\sum_{i=1}^{n}\mathbf{1}_{A_i}\right)^2 - \left(\mathbb{E}\left(\sum_{i=1}^{n}\mathbf{1}_{A_i}\right)\right)^2 \\
&= \sum_{i=1}^{n}(\mathbb{P}(A_1) - (\mathbb{P}(A_1))^2) + \sum_{i\neq j}(\mathbb{P}(A_i \cap A_j) - (\mathbb{P}(A_1))^2) \\
&= n\left(\left(1 - \frac{1}{n}\right)^r - \left(1 - \frac{1}{n}\right)^{2r}\right) + n(n-1)\left(\left(1 - \frac{2}{n}\right)^r - \left(1 - \frac{1}{n}\right)^{2r}\right) \\
&= o(n^2)
\end{aligned}$$

By using the WLLN, let $b_n = n$,

$$\frac{\text{Var}(N_n)}{b_n^2} \to 0 \implies \frac{N_n - \mathbb{E}N_n}{n} \xrightarrow{\mathbb{P}} 0$$

Therefore, $\frac{N_n}{n} \xrightarrow{\mathbb{P}} e^{-c}$.

7.3 Borel-Cantelli Lemmas

Let $A_1, A_2, \dots$ be a sequence of events in $(\Omega, \mathcal{F})$. We are particularly interested in

$$\limsup_{n\to\infty} A_n = \{A_n \text{ i.o.}\} = \bigcap_m \bigcup_{n=m}^{\infty} A_n$$

Theorem 7.15. (Borel-Cantelli Lemmas) For any sequence of events $A_n \in \mathcal{F}$,

1. (BCI) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then
$$\mathbb{P}(A_n \text{ i.o.}) = 0$$
2. (BCII) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and A_n are independent, then
$$\mathbb{P}(A_n \text{ i.o.}) = 1$$

Proof.

1. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$,
$$\mathbb{P}(A_n \text{ i.o.}) = \lim_{m\to\infty} \mathbb{P}\left(\bigcup_{n=m}^{\infty} A_n\right) \le \lim_{m\to\infty} \sum_{n=m}^{\infty} \mathbb{P}(A_n) = 0$$
2. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and A_n are independent, we have
$$\mathbb{P}\left(\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} A_n^{\complement}\right) = \lim_{m\uparrow\infty} \mathbb{P}\left(\bigcap_{n=m}^{\infty} A_n^{\complement}\right) = \lim_{m\uparrow\infty}\lim_{r\uparrow\infty} \mathbb{P}\left(\bigcap_{n=m}^{r} A_n^{\complement}\right) = \lim_{m\uparrow\infty}\lim_{r\uparrow\infty} \prod_{n=m}^{r} \mathbb{P}(A_n^{\complement}) = \lim_{m\uparrow\infty} \prod_{n=m}^{\infty} (1-\mathbb{P}(A_n))$$
$$\le \lim_{m\uparrow\infty} \prod_{n=m}^{\infty} e^{-\mathbb{P}(A_n)} = \lim_{m\uparrow\infty} \exp\left(-\sum_{n=m}^{\infty} \mathbb{P}(A_n)\right) = 0 \qquad (1-x \le e^{-x} \text{ if } x \ge 0)$$
$$\mathbb{P}(A_n \text{ i.o.}) = \mathbb{P}\left(\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty} A_n\right) = 1 - \mathbb{P}\left(\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} A_n^{\complement}\right) = 1$$

□

Remark 7.15.1. BCII can be considered a partial converse of BCI.

Remark 7.15.2. i.o. stands for "infinitely often," while f.o. stands for "finitely often."

We will now explore applications of the Borel-Cantelli Lemmas.

Example 7.7. (Infinite Monkey Problem) Assume there is a keyboard with N keys, each representing a distinct letter. Given a string of letters S of length m, a monkey randomly hits any key at each round.
How can we prove that, almost surely, the monkey will type the given string S infinitely many times?
Let E_k be the event that the m-string S is typed starting from the k-th hit. Note that E_k are not independent.
To produce an independent sequence, consider E_{mk+1}, where each string is m letters apart from the next.
For any i, $\mathbb{P}(E_i) = \left(\frac{1}{N}\right)^m$. By BCII,
$$\sum_{k=0}^{\infty} \mathbb{P}(E_{mk+1}) = \infty \implies \mathbb{P}(E_{mk+1} \text{ i.o.}) = 1$$
Therefore, $\mathbb{P}(E_k \text{ i.o.}) = 1$.

Recall that if $X_n \xrightarrow{\mathbb{P}} X$, there exists a non-random increasing sequence of integers $n_1, n_2, \dots$ such that $X_{n_i} \xrightarrow{\text{a.s.}} X$ as $i \to \infty$. We can use the Borel-Cantelli Lemmas to prove a similar theorem.

Theorem 7.16. $X_n \xrightarrow{\mathbb{P}} X$ if and only if for all subsequences $X_{n(m)}$, there exists a further subsequence

$$X_{n(m_k)} \xrightarrow{\text{a.s.}} X$$

Proof.

($\Longrightarrow$) Let ε_k be a sequence of positive numbers such that $\varepsilon_k \downarrow 0$ as $k \uparrow \infty$. For any k, there exists an $n(m_k) > n(m_{k-1})$ such that

$$\mathbb{P}(\left|X_{n(m_k)} - X\right| > \varepsilon_k) \le 2^{-k} \qquad (X_n \xrightarrow{\mathbb{P}} X)$$

Since $\sum_{k=1}^{\infty} \mathbb{P}(\left|X_{n(m_k)} - X\right| > \varepsilon_k) < \infty$, by BCI,

$$\mathbb{P}(\left|X_{n(m_k)} - X\right| > \varepsilon_k \text{ i.o.}) = 0 \qquad \mathbb{P}(\left|X_{n(m_k)} - X\right| > \varepsilon_k \text{ f.o.}) = 1$$

For all $\varepsilon > 0$, $\varepsilon_k \le \varepsilon$ for all $k \ge k_0$. If $\varepsilon_k \le \varepsilon$,

$$\{\left|X_{n(m_k)} - X\right| > \varepsilon_k\} \supseteq \{\left|X_{n(m_k)} - X\right| > \varepsilon\}$$

If $\omega \in \{\left|X_{n(m_k)} - X\right| > \varepsilon_k\}$ for finitely many k, then $\omega \in \{\left|X_{n(m_k)} - X\right| > \varepsilon\}$ for finitely many k. Therefore, for all $\varepsilon > 0$

$$\mathbb{P}(\left|X_{n(m_k)} - X\right| > \varepsilon \text{ i.o.}) = 0$$

($\Longleftarrow$) For all $\varepsilon > 0$, let $a_n = \mathbb{P}(|X_n - X| > \varepsilon)$.
For all $n(m)$, there exists $n(m_k)$ such that $X_{n(m_k)} \xrightarrow{\text{a.s.}} X$. We have

$$(X_{n(m_k)} \xrightarrow{\text{a.s.}} X) \implies (X_{n(m_k)} \xrightarrow{\mathbb{P}} X) \implies a_{n(m_k)} \to 0$$

Therefore, for any a_n and $a_{n(m)}$, there exists further $a_{n(m_k)} \to 0$.
We have $a_n \to 0 \implies (X_n \xrightarrow{\mathbb{P}} X)$.

□

We have a theorem that has conditions quite similar to the Law of Large Numbers. However, notice that $\mathbb{E}\,|X_1| = \infty$ here.

Theorem 7.17. If $X_1, X_2, \cdots$ are i.i.d. random variables with $\mathbb{E}\,|X_i| = \infty$. Then

$$\mathbb{P}(|X_n| \ge n \text{ i.o.}) = 1$$

Let $S_n = \sum_{i=1}^{n} X_i$. Then

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \text{ exists in } (-\infty, \infty)\right) = 0$$

Proof.

$$\mathbb{E}\,|X_1| = \int_0^{\infty} \mathbb{P}(|X_1| > t)\, dt \le \sum_{n=0}^{\infty} \mathbb{P}(|X_1| > n)$$

Since $\{|X_n| > n\}$ is a collection of independent events, by BCII, $\mathbb{P}(|X_n| > n \text{ i.o.}) = 1$.
For the second statement, let $C = \{\omega \in \Omega : \lim_{n\to\infty} \frac{S_n(\omega)}{n} \text{ exists in } \mathbb{R}\}$.
Assume that $\omega \in C$, then

$$\frac{S_n(\omega)}{n} - \frac{S_{n+1}(\omega)}{n+1} = \frac{S_n(\omega)}{n(n+1)} - \frac{X_{n+1}(\omega)}{n+1}$$

Since $\frac{S_n}{n}$ converges, $\frac{S_n(\omega)}{n} - \frac{S_{n+1}(\omega)}{n+1} \to 0$ and $\frac{S_n(\omega)}{n(n+1)} \to 0$. We get that $\frac{X_{n+1}(\omega)}{n+1} \to 0$.
However, that means $|X_{n+1}| < n+1$ for an arbitrary large n. Therefore, $\omega \notin \{|X_n| \ge n \text{ i.o.}\}$.
From that, we get that $\mathbb{P}(C) = 0$ since $\mathbb{P}(|X_n| \ge n \text{ i.o.}) = 1$. □

The next result extends BCII.

Theorem 7.18. If $A_1, A_2, \cdots$ are pairwise independent and $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then as $n \to \infty$,

$$\frac{\sum_{m=1}^{n} \mathbf{1}_{A_m}}{\sum_{m=1}^{n} \mathbb{P}(A_m)} \xrightarrow{\text{a.s.}} 1$$

Proof.
Let $X_n = \mathbf{1}_{A_n}$, $S_n = \sum_{i=1}^{n} X_i$ and $\mathbb{E}S_n = \sum_{m=1}^{n} \mathbb{P}(A_m)$.
Notice that pairwise independence is already enough for $\text{cov}(X_i, X_j) = 0$ for all $i \neq j$.
Using Markov's inequality, for any $\varepsilon > 0$, we get as $n \to \infty$

$$\mathbb{P}\left(\left|\frac{S_n - \mathbb{E}S_n}{\mathbb{E}S_n}\right| > \varepsilon\right) \leq \frac{\mathbb{E}(S_n - \mathbb{E}S_n)^2}{\varepsilon^2(\mathbb{E}S_n)^2} = \frac{\text{Var}(S_n)}{\varepsilon^2(\mathbb{E}S_n)^2} = \sum_{m=1}^{n} \frac{\text{Var}(\mathbf{1}_{A_m})}{\varepsilon^2(\mathbb{E}S_n)^2} = \sum_{m=1}^{n} \frac{\mathbb{E}\mathbf{1}_{A_m}}{\varepsilon^2(\mathbb{E}S_n)^2} = \frac{1}{\varepsilon^2 \mathbb{E}S_n} \to 0$$

Therefore, we get that $\frac{S_n - \mathbb{E}S_n}{\mathbb{E}S_n} \xrightarrow{\mathbb{P}} 0$.
Now, we can choose a desirable subsequence to prove almost surely convergence. Let $n_k = \inf\{n : \mathbb{E}S_n \geq k^2\}$.
We can get that $\mathbb{E}S_{n_k} \geq k^2$ and $\mathbb{E}S_{n_k} = \mathbb{E}S_{n_k - 1} + \mathbb{E}\mathbf{1}_{A_{n_k}} < k^2 + 1$. Again by Markov's inequality,

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\left|\frac{S_{n_k} - \mathbb{E}S_{n_k}}{\mathbb{E}S_{n_k}}\right| > \varepsilon\right) \leq \sum_{k=1}^{\infty} \frac{1}{\varepsilon^2 \mathbb{E}S_{n_k}} \leq \sum_{k=1}^{\infty} \frac{1}{\varepsilon^2(k^2+1)} < \infty$$

By BCI, we have that as $k \to \infty$,

$$\frac{S_{n_k}}{\mathbb{E}S_{n_k}} \xrightarrow{\text{a.s.}} 1 \qquad\qquad \mathbb{P}\left(\frac{S_{n_k}}{\mathbb{E}S_{n_k}} \to 1 \text{ as } k \to \infty\right) = 1$$

Let $C = \{\omega \in \Omega : \frac{S_{n_k}(\omega)}{\mathbb{E}S_{n_k}} \to 1 \text{ as } k \to \infty\}$. For $\omega \in C$, for all $n_k \leq n < n_{k+1}$, we have $S_{n_k}(\omega) \leq S_n(\omega) \leq S_{n_{k+1}}(\omega)$.

$$\frac{S_{n_k}(\omega)}{\mathbb{E}S_{n_k+1}} \leq \frac{S_n(\omega)}{\mathbb{E}S_n} \leq \frac{S_{n_k+1}(\omega)}{\mathbb{E}S_{n_k}}$$

Since $\frac{S_{n_k}(\omega)}{\mathbb{E}S_{n_k+1}} = \frac{S_{n_k}(\omega)}{\mathbb{E}S_{n_k}} \left(\frac{\mathbb{E}S_{n_k}}{\mathbb{E}S_{n_k+1}}\right) \to 1$ and $\frac{S_{n_k+1}(\omega)}{\mathbb{E}S_{n_k+1}} = \frac{S_{n_k+1}(\omega)}{\mathbb{E}S_{n_k+1}} \left(\frac{\mathbb{E}S_{n_k+1}}{\mathbb{E}S_{n_k}}\right) \to 1$, we get that for any $\omega \in C$,

$$\frac{S_n(\omega)}{\mathbb{E}S_n} \to 1$$

Therefore, we have

$$\mathbb{P}\left(\frac{S_n}{\mathbb{E}S_n} \to 1\right) \geq \mathbb{P}\left(\frac{S_{n_k}}{\mathbb{E}S_{n_k}} \to 1 \text{ as } k \to \infty\right) = 1$$

As a result, we get that

$$\frac{S_n}{\mathbb{E}S_n} \xrightarrow{\text{a.s.}} 1$$

□

If the events $A_1, A_2, \cdots$ in the Borel-Cantelli Lemmas are independent, then $\mathbb{P}(A)$ is either 0 or 1 depending on whether $\sum \mathbb{P}(A_n)$ converges. The following is a simple version.

Theorem 7.19. (Borel Zero-one Law) Let $A_1, A_2, \cdots \in \mathcal{F}$ and $\mathcal{A} = \sigma(A_1, A_2, \cdots)$. Suppose that

1. $A \in \mathcal{A}$
2. A is independent with any finite collection of $A_1, A_2, \cdots$

Then $\mathbb{P}(A) = 0$ or 1.

Proof (Non-rigorous).
Suppose that $A_1, A_2, \cdots$ are independent. Let $A = \limsup_n A_n$.
We know that $A = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n$. Therefore, $A \in \mathcal{A} = \sigma(A_1, A_2, \cdots)$.
For all k, we can also have $A = \bigcap_{m=k+1}^{\infty} \bigcup_{n=m}^{\infty} A_n$. Therefore, A is independent with any $A_i \in \sigma(A_1, A_2, \cdots, A_k)$.
Setting $k \to \infty$, we have that A is independent of all elements in $\mathcal{A}$, which also include itself.
Therefore, $\mathbb{P}(A) = \mathbb{P}(A \cap A) = (\mathbb{P}(A))^2 \implies \mathbb{P}(A) = 0$ or 1. □

Let $X_1, X_2, \cdots$ be a collection of random variables. For any subcollection $\{X_i : i \in I\}$, write $\sigma(X_i : i \in I)$ for the smallest σ-field with reference to which each of X_i is measurable.

Definition 7.20. Let $\mathcal{H}_n = \sigma(X_{n+1}, X_{n+2}, \cdots)$. We have $\mathcal{H}_n \supseteq \mathcal{H}_{n+1} \supseteq \cdots$. **Tail σ-field** is defined as

$$\mathcal{H}_\infty = \bigcap_n \mathcal{H}_n$$

Remark 7.20.1. If $E \in \mathcal{H}_\infty$, then E is called **tail event**.

Example 7.8. $\{\limsup_{n\to\infty} X_n = \infty\}$ is a tail event.

Example 7.9. $\{\sum_n X_n \text{ converges}\}$ is a tail event.

Example 7.10. $\{\sum_n X_n \text{ converges to } 1\}$ is not a tail event.

We get another version of zero-one law.

Theorem 7.21. (Kolmogorov's zero-one law) If $H \in \mathcal{H}_\infty$, then $\mathbb{P}(H) = 0$ or 1.

We continue to explore more into tail events.

Definition 7.22. We define **tail function** to be $Y : \Omega \to \mathbb{R} \cup \{-\infty, \infty\}$, which is a generalized random variables that is a function of $X_1, X_2, \cdots$. It is independent of any finite collection of X_i's and is $\mathcal{H}_\infty$-measurable.

Example 7.11. Let $Y(\omega) = \limsup_{n\to\infty} X_n(\omega)$ for all $\omega \in \Omega$. $F_Y(y) = \mathbb{P}(Y \le y) = 0$ or 1 for all $y \in \mathbb{R} \cup \{-\infty, \infty\}$.
$\{Y \le y\}$ is a tail event.

Theorem 7.23. If Y is a tail function of independent sequence of random variables $X_1, X_2, \cdots$, then there exists $-\infty \le k \le \infty$,

$$\mathbb{P}(Y = k) = 1$$

Again let $X_1, X_2, \cdots$ be i.i.d. random variables and let $S_n = \sum_{i=1}^n X_i$.
Recall that if $\mathbb{E}|X_1| < \infty$,

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = \mathbb{E}X_1\right) = 1$$

If $\mathbb{E}|X_1| = \infty$,

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \text{ exists}\right) = 0$$

Using tail function, the random variables are not necessarily identically distributed.

Theorem 7.24. Let $X_1, X_2, \cdots$ be independent random variables. Let $S_n = \sum_{i=1}^n X_i$. Then

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \text{ exists}\right) = 0 \text{ or } 1$$

Proof.
Let $Z_1 = \limsup_{n\to\infty} \frac{S_n}{n}$ and $Z_2 = \liminf_{n\to\infty} \frac{S_n}{n}$. We claim that both Z_1 and Z_2 are tail functions of X_i's. For any k,

$$Z_1(\omega) = \limsup_{n\to\infty} \left(\frac{1}{n}\sum_{i=1}^k X_i(\omega) + \frac{1}{n}\sum_{i=k+1}^n X_i(\omega)\right) \qquad Z_2(\omega) = \liminf_{n\to\infty} \left(\frac{1}{n}\sum_{i=1}^k X_i(\omega) + \frac{1}{n}\sum_{i=k+1}^n X_i(\omega)\right)$$

Therefore, both Z_1 and Z_2 do not depend on any finite collection of X_i. We say that $\{Z_1 = Z_2\}$ is a tail event.
Therefore, by Kolmogorov's zero-one law.

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \text{ exists}\right) = \mathbb{P}(Z_1 = Z_2) = 0 \text{ or } 1$$

□

Example 7.12. (Random power series) Let $X_1, X_2, \cdots$ be i.i.d. exponential random variables with parameter $\lambda = 1$. We consider a random power series

$$p(z;\omega) = \sum_{n=0}^{\infty} X_n(\omega) z^n$$

The formula for radius of convergence is

$$R(\omega) = \frac{1}{\limsup_{n\to\infty} |X_n(\omega)|^{\frac{1}{n}}}$$

We can get that $R(\omega)$ is a tail function of X_i's. Therefore, there exists C such that $\mathbb{P}(R = C) = 1$ ($R = C$ almost surely)
We want to find the value of C.
We claim that $C = 1$.

$$\mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} = 1\right) = 1$$

It suffices to show that for all $\varepsilon > 0$,

$$\mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} \le 1+\varepsilon\right) = 1 \qquad \mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} \ge 1-\varepsilon\right) = 1$$

We first prove the first one.

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|X_n|^{\frac{1}{n}} > 1+\varepsilon\right) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > (1+\varepsilon)^n) = \sum_{n=1}^{\infty} e^{-(1+\varepsilon)^n} < \infty$$

Therefore, by BCI,

$$\mathbb{P}(|X_n|^{\frac{1}{n}} > 1+\varepsilon \text{ i.o.}) = 0 \implies \mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} \le 1+\varepsilon\right) = 1$$

Similarly,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|X_n|^{\frac{1}{n}} > 1-\varepsilon\right) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > (1-\varepsilon)^n) = \sum_{n=1}^{\infty} e^{-(1-\varepsilon)^n} = \infty$$

Therefore, by BCII,

$$\mathbb{P}(|X_n|^{\frac{1}{n}} > 1-\varepsilon \text{ i.o.}) = 1 \implies \mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} \ge 1-\varepsilon\right) = 1$$

By sending $\varepsilon \downarrow 0$, we get

$$\mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} = 1\right) = 1$$

Therefore, $C = 1$.

7.4 Strong Law of Large Numbers

Let us revisit the Weak Law of Large Numbers (WLLN). Consider $X_1, X_2, \ldots$ as a sequence of i.i.d. random variables with $\mathbb{E}(X_1) = \mu$. Define $S_n = \sum_{i=1}^n X_i$. Then, as $n \to \infty$,

$$\frac{S_n}{n} \xrightarrow{D} \mu \qquad \frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu$$

The Strong Law of Large Numbers (SLLN) is a more robust version of WLLN. Below, we prove one version of SLLN.

Theorem 7.25. (**Strong Law of Large Numbers** [SLLN]) Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}X_1 = \mu$ and $\mathbb{E}|X_1| < \infty$. Define $S_n = \sum_{i=1}^n X_i$. Then,

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mu$$

Note that the proof of SLLN is intricate and will not be covered here. Instead, we present a simpler version of SLLN.

Theorem 7.26. (SLLN with $\mathbb{E}X_i^4 < \infty$) Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}X_1 = 0$ and $\mathbb{E}(X_1^4) < \infty$. Define $S_n = \sum_{i=1}^n X_i$. Then,

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} 0$$

Proof.

$$\mathbb{E}S_n^4 = \mathbb{E}\left(\sum_{i=1}^n X_i\right)^4 = \sum_{i,j,k,\ell=1}^n \mathbb{E}X_iX_jX_kX_\ell$$

The expectation is non-zero only if there are two pairs of random variables with identical values.

$$\mathbb{E}S_n^4 = 3\sum_{i\neq j}\mathbb{E}X_i^2\mathbb{E}X_j^2 + \sum_i \mathbb{E}X_i^4 = O(n^2)$$

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right| \geq \varepsilon\right) \leq \frac{\mathbb{E}S_n^4}{(n\varepsilon)^4} = O\left(\frac{1}{n^2}\right)$$

Consequently, for all $\varepsilon > 0$,

$$\sum_{n=1}^\infty \mathbb{P}\left(\left|\frac{S_n}{n}\right| > \varepsilon\right) < \infty$$

Thus, $\frac{S_n}{n} \xrightarrow{\text{a.s.}} 0$. □

Theorem 7.27. (SLLN with $\mathbb{E}X_1^2 < \infty$) Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}X_1^2 < \infty$ and $\mathbb{E}X_i = \mu$. Define $S_n = \sum_{i=1}^n X_i$. Then,

$$\frac{S_n}{n} \xrightarrow{2} \mu \qquad\qquad \frac{S_n}{n} \xrightarrow{\text{a.s.}} \mu$$

Proof.
We first demonstrate convergence in mean square. Since $\mathbb{E}X_1^2 < \infty$, as $n \to \infty$,

$$\mathbb{E}\left(\frac{S_n}{n} - \mu\right)^2 = \frac{\mathbb{E}(S_n - n\mu)^2}{n^2} = \frac{\text{Var}(S_n)}{n^2} = \frac{\text{Var}(X_1)}{n} \to 0$$

For almost sure convergence, we know that convergence in probability implies the existence of almost sure convergence for some subsequence of $\frac{S_n}{n}$ to μ. Let $n_i = i^2$. Using Markov's inequality, for all $\varepsilon > 0$,

$$\sum_i \mathbb{P}\left(\frac{\left|S_{i^2} - i^2\mu\right|}{i^2} > \varepsilon\right) \leq \sum_i \frac{\mathbb{E}\left|S_{i^2} - i^2\mu\right|^2}{i^4\varepsilon^2} = \sum_i \frac{\text{Var}(S_{i^2})}{i^4\varepsilon^2} = \sum_i \frac{\text{Var}(X_1)}{i^2\varepsilon^2} < \infty$$

Therefore, $\frac{S_{i^2}}{i^2} \xrightarrow{\text{a.s.}} \mu$. However, we need to address the gaps.
Assume X_i are non-negative. Then $S_{i^2} \leq S_n \leq S_{(i+1)^2}$ if $i^2 \leq n \leq (i+1)^2$.
We can deduce that

$$\frac{S_{i^2}}{(i+1)^2} \leq \frac{S_n}{n} \leq \frac{S_{(i+1)^2}}{i^2}$$

Since $\frac{S_{i^2}}{i^2} \xrightarrow{\text{a.s.}} \mu$, and $\frac{i^2}{(i+1)^2} \to 1$ as $i \to \infty$, we conclude that for non-negative X_i, as $n \to \infty$,

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mu$$

For general X_i, we can write $X_n = X_n^+ - X_n^-$ where

$$X_n^+(\omega) = \max\{X_n(\omega), 0\} \qquad\qquad X_n^-(\omega) = -\min\{X_n(\omega), 0\}$$

Both $X_n^+(\omega)$ and $X_n^-(\omega)$ are non-negative.
Furthermore, $X_n^+ \leq |X_n|$ and $X_n^- \leq |X_n|$. Thus, $\mathbb{E}(X_n^+)^2 < \infty$ and $\mathbb{E}(X_n^-)^2 < \infty$. By the earlier conclusion for non-negative random variables, we find that as $n \to \infty$,

$$\frac{S_n}{n} = \frac{1}{n}\left(\sum_{i=1}^n X_i^+ - \sum_{i=1}^n X_i^-\right) \xrightarrow{\text{a.s.}} \mathbb{E}X_1^+ - \mathbb{E}X_1^- = \mathbb{E}X_1$$

Therefore, $\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mu$. □

Why do we need SLLN? There are a lot of applications that specifically need SLLN.

Example 7.13. (Renewal Theory) Assume that we have a light bulb. We change it immediately when it burnt out.
Let X_i be the lifetime of i-th bulb and $T_n = X_1 + X_2 + \cdots + X_n$ be the time to replace the n-th bulb.
Let $N_t = \sup\{n : T_n \leq t\}$ be number of bulbs that have burnt out by time t. T_{N_t} is the exact time that N_t's bulb burnt out.
Since we are dealing with practical bulb, assume that $X_1, X_2, \cdots$ are i.i.d. random variables with $0 < X_i < \infty$ and $\mathbb{E}X_1 < \infty$.

Theorem 7.28. Let $\mathbb{E}X_1 = \mu$. As $t \to \infty$,

$$\frac{t}{N_t} \xrightarrow{\text{a.s.}} \mu$$

Proof.
Since $T_{N_t} \leq t < T_{N_t+1}$,

$$\frac{T_{N_t}}{N_t} \leq \frac{t}{N_t} < \frac{T_{N_t+1}}{N_t+1}\left(\frac{N_t+1}{N_t}\right)$$

By SLLN, we know that $\frac{T_n}{n} \xrightarrow{\text{a.s.}} \mu$. Since $\frac{T_n}{n}$ and $\frac{T_{N_t}}{N_t}$ are the same sequence, we get that

$$\frac{T_{N_t}}{N_t} \xrightarrow{\text{a.s.}} \mu \qquad\qquad \frac{T_{N_t+1}}{N_t+1} \xrightarrow{\text{a.s.}} \mu$$

For all $\omega \in \Omega$, $t < T_{N_t+1} = X_1(\omega) + X_2(\omega) + \cdots + X_{N_t(\omega)+1}(\omega)$.
As $t \to \infty$, it forces $N_t(\omega) \to \infty$. Therefore, $\frac{N_t+1}{N_t} \xrightarrow{\text{a.s.}} 1$. Combining all of this, we get $\frac{t}{N_t} \xrightarrow{\text{a.s.}} \mu$. □

Claim 7.28.1. If $X_n \xrightarrow{\mathbb{P}} X_\infty$, then $N_m \xrightarrow{\text{a.s.}} \infty$ as $m \to \infty$.

Remark 7.28.1. For this claim, it is not necessary that $X_{N_m} \xrightarrow{\text{a.s}} X_\infty$ or $X_{N_m} \xrightarrow{\mathbb{P}} X_\infty$.

Example 7.14. Recall the example that we use in Theorem 7.9 to prove $(X_n \xrightarrow{\mathbb{P}} X) \not\Longrightarrow (X_n \xrightarrow{\text{a.s.}} X)$. Let $\Omega = [0, 1]$. Let

$$Y_{m,k} = \mathbf{1}_{I_{m,k}} = \begin{cases} 1, & \omega \in \left[\frac{k-1}{m}, \frac{k}{m}\right] \\ 0, & \text{Otherwise} \end{cases}$$

Let X_n be the enumeration of $Y_{m,k}$. i.e. $X_1 = Y_{1,1}$, $X_2 = Y_{2,1}$, $X_3 = Y_{2,2}$, $\cdots$.
From the proof of the theorem, we got that $X_n \xrightarrow{\mathbb{P}} X_\infty = 0$ but $X_n \not\xrightarrow{\text{a.s.}} X_\infty$.
For each $\omega \in \Omega$, and each $m \geq 1$, there exists k such that $\omega \in \left[\frac{k-1}{m}, \frac{k}{m}\right]$. We denote these as $k_m(\omega)$.
Let $N_m(\omega) = \sum_{i=1}^{m-1} i + k_m(\omega)$. We get that $X_{N_m(\omega)}(\omega) = Y_{m,k_m(\omega)}(\omega) = 1$.
However, $X_\infty = 0$. That means, $X_{N_m} \not\xrightarrow{\mathbb{P}} X_\infty$ and $X_{N_m} \not\xrightarrow{\text{a.s.}} X_\infty$.

We move to our next examples, which is the Glivenko-Cantelli Theorem. It is also called the Fundamental Theorem of Statistics.

Theorem 7.29. (Glivenko-Cantelli Theorem) Assume that $X \sim F(x)$ where $F(x)$ is unknown. Let $X_1, X_2, \cdots$ be i.i.d. random samples of X. We define the empirical distribution function, which is also a distribution function of a histogram.

$$F_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{X_i \leq x} \qquad\qquad F_N(x;\omega) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{X_i(\omega) \leq x}$$

We have that

$$\sup_x |F_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0$$

Proof.
We only proof for the case when $F(x)$ is continuous.
For each m, there exists $-\infty = x_0 < x_1 < \cdots < x_m = \infty$ such that $F(x_i) - F(x_{i-1}) = \frac{1}{m}$.
For all $x \in [x_{i-1}, x_i)$,

$$F_N(x) - F(x) \leq F_N(x_i) - F(x_{i-1}) = F_N(x_i) - F(x_i) + \frac{1}{m}$$
$$F_N(x) - F(x) \geq F_N(x_{i-1}) - F(x_i) = F_N(x_{i-1}) - F(x_{i-1}) - \frac{1}{m}$$

From this, we get

$$-\sup_i |F_N(x_i) - F(x_i)| - \frac{1}{m} \leq F_N(x) - F(x) \leq \sup_i |F_N(x_i) - F(x_i)| + \frac{1}{m} \implies \sup_x |F_N(x) - F(x)| \leq \sup_i |F_N(x_i) - F(x_i)| + \frac{1}{m}$$

By SLLN, when we fix x, we get

$$F_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{X_i \le x} \xrightarrow{\text{a.s.}} \mathbb{E}\mathbf{1}_{X_1 \le x} = \mathbb{P}(X_1 \le x) = F(x) \qquad \mathbb{P}(\{\omega \in \Omega : F_N(x;\omega) \to F(x) \text{ as } N \to \infty\}) = 1$$

Let $C_x = \{\omega \in \Omega : F_N(x;\omega) \to F(x) \text{ as } N \to \infty\}$. Notice that if $\omega \in \bigcap_{i=1}^{\infty} C_{x_i}$, $\sup_i |F_N(x_i;\omega) - F(x_i)| \to 0$.

$$\limsup_N \sup_x |F_N(x) - F(x)| \le \frac{1}{m}$$

If $\omega \in \bigcap_{m=1}^{\infty}\bigcap_{i=1}^{m} C_{x_i}$,

$$\limsup_N \sup_x |F_N(x) - F(x)| = 0$$

Therefore, since $\bigcap_{m=1}^{\infty}\bigcap_{i=1}^{m} C_{x_i} \subseteq \{\omega \in \Omega : \sup_x |F_N(x;\omega) - F(x)| \to 0 \text{ as } N \to \infty\}$ and $\mathbb{P}(C_{x_i}) = 1$ by SLLN,

$$\mathbb{P}(\{\omega \in \Omega : \sup_x |F_N(x;\omega) - F(x)| \to 0 \text{ as } N \to \infty\}) = 1$$

□

We will end here. Of course, there are still a lot of examples that we haven't explored (including some mentioned during the lectures that I'm too lazy to include here). We also skipped a lot of proofs in some of the theorems. It is up to you to explore further, either in other courses or in the future world of mathematics.

Appendix A

Random walk

Example A.1. (**Simple random walk**) Consider a particle moving along the real line. At each step, it moves either one unit to the right with probability p or one unit to the left with probability $q = 1 - p$.
Let S_n represent the particle's position after n steps, with $S_0 = a$. Then:

$$S_n = a + \sum_{i=1}^{n} X_i$$

where $X_1, X_2, \ldots$ are independent random variables taking the value 1 with probability p and -1 with probability q.
The random walk is called **symmetric** if $p = q = \frac{1}{2}$.

Lemma A.1. A simple random walk has the following properties:

1. It is **spatially homogeneous**: $\mathbb{P}(S_n = j|S_0 = a) = \mathbb{P}(S_n = j + b|S_0 = a + b)$.
2. It is **temporally homogeneous**: $\mathbb{P}(S_n = j|S_0 = a) = \mathbb{P}(S_{m+n} = j|S_m = a)$.
3. It satisfies the **Markov property**: $\mathbb{P}(S_{m+n} = j|S_0, S_1, \ldots, S_m) = \mathbb{P}(S_{m+n} = j|S_m)$, $n \geq 0$.

Proof.

1. $\mathbb{P}(S_n = j|S_0 = a) = \mathbb{P}(\sum_{i=1}^{n} X_i = j - a) = \mathbb{P}(S_n = j + b|S_0 = a + b)$.
2. $$\mathbb{P}(S_n = j|S_0 = a) = \mathbb{P}\left(\sum_{i=1}^{n} X_i = j - a\right) = \mathbb{P}\left(\sum_{i=m+1}^{m+n} X_i = j - a\right) = \mathbb{P}(S_{m+n} = j|S_m = a).$$
3. If S_m is known, the distribution of S_{m+n} depends only on $X_{m+1}, X_{m+2}, \ldots, X_{m+n}$, and is independent of $S_0, S_1, \ldots, S_{m-1}$.

□

Example A.2. (Probability via sample path counting) Define a **sample path** $\vec{s} = (s_0, s_1, \ldots, s_n)$ as the outcome or realization of the random walk, where $s_0 = a$ and $s_{i+1} - s_i = \pm 1$.

$$\mathbb{P}((S_0, S_1, \ldots, S_n) = (s_0, s_1, \ldots, s_n)) = p^r q^\ell, \qquad r = \#\{i : s_{i+1} - s_i = 1\}, \qquad \ell = \#\{i : s_{i+1} - s_i = -1\}.$$

Example A.3. Let $M_n^r(a, b)$ denote the number of paths $(s_0, s_1, \ldots, s_n)$ with $s_0 = a$, $s_n = b$, and exactly r steps to the right.

$$\mathbb{P}(S_n = b) = \sum_r M_n^r(a, b) p^r q^{n-r}.$$

Using the equations $r + \ell = n$ and $r - \ell = b - a$, we find $r = \frac{1}{2}(n + b - a)$ and $\ell = (n - b + a)$. If $\frac{1}{2}(n + b - a) \in \{0, 1, \ldots, n\}$,

$$\mathbb{P}(S_n = b) = \binom{n}{\frac{1}{2}(n + b - a)} p^{\frac{1}{2}(n+b-a)} q^{\frac{1}{2}(n-b+a)}.$$

Otherwise, $\mathbb{P}(S_n = b) = 0$.

Theorem A.2. (Reflection principle) Let $N_n(a,b)$ represent the number of possible paths from $(0,a)$ to (n,b), and let $N_n^0(a,b)$ denote the number of such paths that pass through some point $(k,0)$ on the x-axis. If $a,b>0$, then:

$$N_n^0(a,b) = N_n(-a,b).$$

Proof.
Each path from $(0,-a)$ to (n,b) intersects the x-axis at some earliest point $(k,0)$.
Reflect the segment of the path with $0 \le x \le k$ across the x-axis to obtain a path joining $(0,a)$ to (n,b) that intersects the x-axis.
This operation establishes a one-to-one correspondence between these collections of paths. □

Lemma A.3.

$$N_n(a,b) = \binom{n}{\frac{1}{2}(n+b-a)}.$$

Proof.
Consider a path from $(0,a)$ to (n,b), and let α and β represent the number of positive and negative steps, respectively.
Then $\alpha+\beta=n$ and $\alpha-\beta=b-a$, which implies $\alpha=\frac{1}{2}(n+b-a)$.
The number of such paths corresponds to the number of ways to choose α positive steps from n available steps. Thus,

$$N_n(a,b) = \binom{n}{\alpha} = \binom{n}{\frac{1}{2}(n+b-a)}.$$

□

Example A.4. We aim to determine the probability that the walk does not revisit its starting point during the first n steps.
Without loss of generality, assume $S_0=0$, so that $S_1,S_2,\dots,S_n \neq 0$ if and only if $S_1S_2\cdots S_n \neq 0$.
The event $S_1S_2\cdots S_n \neq 0$ occurs if and only if the path of the walk does not intersect the x-axis during the interval $[1,n]$.
If $b>0$, the first step must be $(1,1)$. By Lemma A.3 and the Reflection Principle, the number of such paths is:

$$\begin{aligned} N_{n-1}(1,b) - N_{n-1}^0(1,b) &= N_{n-1}(1,b) - N_{n-1}(-1,b) \\ &= \binom{n-1}{\frac{1}{2}(n+b-2)} - \binom{n-1}{\frac{1}{2}(n+b)} \\ &= \left(\frac{n+b}{2n} - \frac{n-b}{2n}\right)\binom{n}{\frac{1}{2}(n+b)} \\ &= \frac{b}{n}\binom{n}{\frac{1}{2}(n+b)}. \end{aligned}$$

There are $\frac{1}{2}(n+b)$ rightward steps and $\frac{1}{2}(n-b)$ leftward steps. Therefore,

$$\mathbb{P}(S_1S_2\cdots S_n \neq 0, S_n=b) = \frac{b}{n}N_n(0,b)p^{\frac{1}{2}(n+b)}q^{\frac{1}{2}(n-b)} = \frac{b}{n}\mathbb{P}(S_n=b).$$

Example A.5. Let $M_n = \max\{S_i : 0 \le i \le n\}$ denote the maximum value attained by the random walk up to time n. Suppose $S_0=0$, so $M_n \ge 0$. Clearly, $M_n \ge S_n$.

Theorem A.4. Suppose $S_0=0$. Then, for $r \ge 1$,

$$\mathbb{P}(M_n \ge r, S_n=b) = \begin{cases} \mathbb{P}(S_n=b), & \text{if } b \ge r \\ \left(\frac{q}{p}\right)^{r-b}\mathbb{P}(S_n=2r-b), & \text{if } b<r. \end{cases}$$

Consequently, for $r \ge 1$,

$$\mathbb{P}(M_n \ge r) = \mathbb{P}(S_n \ge r) + \sum_{b=-\infty}^{r-1}\left(\frac{q}{p}\right)^{r-b}\mathbb{P}(S_n=2r-b) = \mathbb{P}(S_n=r) + \sum_{c=r+1}^{\infty}\left(1+\left(\frac{q}{p}\right)^{c-r}\right)\mathbb{P}(S_n=c).$$

For the symmetric case where $p=q=\frac{1}{2}$,

$$\mathbb{P}(M_n \ge r) = 2\mathbb{P}(S_n \ge r+1) + \mathbb{P}(S_n=r).$$

Proof.
Assume $r \geq 1$ and $b < r$. Let $N_n^r(0,b)$ denote the number of paths from $(0,0)$ to (n,b) that include at least one point with height r (i.e., some point (i,r) with $0 < i < n$).
For a path π, let (i_π, r) be the earliest such point.
Reflect the segment of the path with $i_\pi \leq x \leq n$ across the line $y = r$ to obtain a path π' joining $(0,0)$ to $(n, 2r-b)$.
Thus, $N_n^r(0,b) = N_n(0, 2r-b)$.

$$\mathbb{P}(M_n \geq r, S_n = b) = N_n^r(0,b) p^{\frac{1}{2}(n+b)} q^{\frac{1}{2}(n-b)} = \left(\frac{q}{p}\right)^{r-b} N_n(0, 2r-b) p^{\frac{1}{2}(n+2r-b)} q^{\frac{1}{2}(n-2r+b)} = \left(\frac{q}{p}\right)^{r-b} \mathbb{P}(S_n = 2r-b).$$

□

Appendix B

Terminologies in other fields of mathematics

Definition B.1. The **supremum** of a subset S is the smallest upper bound x such that $x \geq a$ for all $a \in S$. It is denoted as:

$$x = \sup S$$

Definition B.2. The **infimum** of a subset S is the greatest lower bound x such that $x \leq b$ for all $b \in S$. It is denoted as:

$$x = \inf S$$

Definition B.3. The **limit superior** and **limit inferior** of a sequence $x_1, x_2, \ldots$ are defined as:

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty} \sup_{m\geq n} x_m, \qquad \liminf_{n\to\infty} x_n = \lim_{n\to\infty} \inf_{m\geq n} x_m.$$

Definition B.4. An infinite series $\sum_{n=0}^{\infty} a_n$ is **absolutely convergent** if there exists a real number L such that:

$$\sum_{n=0}^{\infty} |a_n| = L.$$

Rearranging or grouping terms in an absolutely convergent series does not change its sum.
A series is **conditionally convergent** if it converges but does not satisfy the condition of absolute convergence.

Definition B.5. (**Monotonicity**) A **monotonic** function is one that is either entirely non-increasing or entirely non-decreasing. A **strictly monotonic** function is one that is either entirely strictly increasing or strictly decreasing.

Definition B.6. The **arguments of the maxima** are the input values at which a function achieves its maximum output. It is defined as:

$$\operatorname*{argmax}_{x\in S} f(x) = \{x \in S : f(x) \geq f(s) \text{ for all } s \in S\}.$$

Definition B.7. The **arguments of the minima** are the input values at which a function achieves its minimum output. It is defined as:

$$\operatorname*{argmin}_{x\in S} f(x) = \{x \in S : f(x) \leq f(s) \text{ for all } s \in S\}.$$

Definition B.8. (**Linearity**) A **linear** function f satisfies the following two properties:

1. $f(x + y) = f(x) + f(y)$.
2. $f(ax) = af(x)$ for all a.

Definition B.9. A **regular** function f satisfies the following conditions:

1. It is single-valued (each input in the domain maps to exactly one output).
2. It is analytic (f can be expressed as a convergent power series).

Definition B.10. Let V be the space of all real functions on $[0,1]$. A **norm** $\|\cdot\| : V \to \mathbb{R}$ of a function f satisfies:

1. $\|f\| \geq 0$ for all $f \in V$.
2. If $\|f\| = 0$, then $f = 0$.
3. $\|af\| = |a| \, \|f\|$ for all $f \in V$ and $a \in \mathbb{R}$.
4. (Triangle inequality) $\|f + g\| \leq \|f\| + \|g\|$ for all $f, g \in V$.

The L_p norm for $p \geq 1$ is defined as:

$$\|f\|_p = \left(\int_0^1 |f(x)|^p \, dx \right)^{\frac{1}{p}}.$$

The **infinity norm** of a function $f \in V$ is defined as:

$$\|f\|_\infty = \max_{0 \leq x \leq 1} |f(x)| \, .$$

Definition B.11. Two functions f and g are **asymptotically equivalent** ($f \sim g$) if and only if:

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 1.$$

Appendix C

Some useful inequalities

Theorem C.1. (Triangle inequality) Let X and Y be random variables. Then:

$$|X+Y| \leq |X| + |Y|.$$

Theorem C.2. (Reverse triangle inequality) Let X and Y be random variables. Then:

$$|X-Y| \geq ||X| - |Y||.$$

Theorem C.3. (Cauchy-Schwarz inequality) Let X and Y be random variables. Then:

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

Theorem C.4. (Covariance inequality) Let X and Y be random variables. Then:

$$|\mathrm{cov}(X,Y)|^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y).$$

Theorem C.5. (Markov's inequality) Let X be a random variable with a finite mean. For all $k > 0$ and any non-negative function γ that is increasing on $[0, \infty)$:

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

Theorem C.6. (Chebyshev's inequality) Let X be a random variable with $\mathbb{E}X = \mu$ and $\mathrm{Var}(X) = \sigma^2$. For all $k > 0$:

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

Theorem C.7. (Hölder's inequality) Let X and Y be random variables. For any $p > 1$, let $q = \frac{p}{p-1}$. Then:

$$\mathbb{E}\,|XY| \leq (\mathbb{E}\,|X|^p)^{\frac{1}{p}} (\mathbb{E}\,|Y|^q)^{\frac{1}{q}}.$$

Theorem C.8. (Lyapunov's inequality) Let X be a random variable. For all $0 < s \leq r$:

$$(\mathbb{E}\,|X|^s)^{\frac{1}{s}} \leq (\mathbb{E}\,|X|^r)^{\frac{1}{r}}.$$

Theorem C.9. (Minkowski inequality) Let X and Y be random variables. For any $r \geq 1$:

$$(\mathbb{E}\,|X+Y|^r)^{\frac{1}{r}} \leq (\mathbb{E}\,|X|^r)^{\frac{1}{r}} + (\mathbb{E}\,|Y|^r)^{\frac{1}{r}}.$$

Theorem C.10. (Jensen's inequality) Let X be a random variable and γ a convex function. Then:

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

For better memorization:
Triangle inequality $\implies$ Reverse triangle inequality
Markov's inequality $\implies$ Chebyshev's inequality
Hölder's inequality $\implies$ Cauchy-Schwarz inequality $\implies$ Covariance inequality

Appendix D

Some other distributions

Example D.1. (Gamma distribution) $X \sim \Gamma(\alpha, \beta)$
A random variable X follows a gamma distribution with parameters α and β if:

$$f(x) = \frac{x^{\alpha-1}e^{-\beta x}\beta^{\alpha}}{\Gamma(\alpha)}, \qquad \mathbb{E}X = \frac{\alpha}{\beta}, \qquad \mathrm{Var}(X) = \frac{\alpha}{\beta^2}, \qquad M_X(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha}, \qquad G_X(t) = \left(1 - \frac{it}{\beta}\right)^{-\alpha},$$

where $\Gamma(\alpha)$ is the gamma function. If α is a positive integer, $\Gamma(\alpha) = (\alpha - 1)!$.

Example D.2. (Chi-squared distribution) $Y \sim \chi^2(k)$
Suppose $X_1, X_2, \ldots, X_n$ are independent standard normal random variables. Let $Y = \sum_{i=1}^{n} X_i^2$. Then Y follows a χ^2-distribution with parameter k if:

$$f(x) = \begin{cases} \frac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}, & x \geq 0, \\ 0, & x < 0, \end{cases} \qquad \mathbb{E}Y = k, \qquad \mathrm{Var}(Y) = 2k, \qquad M_Y(t) = (1-2t)^{-\frac{k}{2}}, \qquad G_Y(t) = (1-2it)^{-\frac{k}{2}}.$$