Probability

Version 2.1

HU-HTAKM

Website: https://htakm.github.io/htakm_test/

Last major change: September 29, 2025

Last small update: October 1, 2025

This is a rewritten version of my lecture notes titled ”MATH 2431: Honor Probability.” I realized that the ordering of topics was somewhat disorganized, and there were some missing topics that might be necessary for future courses. My plan is to make these notes a combination of MATH 2421 (Probability) and MATH 2431 (Honor Probability). Again, if you can find any typos, you are either already proficient in the topics or have sharp eyes. ;)
Please note that all proofs that could be combinatorial proofs are omitted.

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
$\overline{X}$	Sample mean of X
S_{n-1}^2	Sample variance of X
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
JMGF	Joint 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 a definition.

Remark 0.1.1. This is a remark.

Lemma 0.2. This is a lemma.

Proposition 0.3. This is a proposition.

Theorem 0.4. This is a theorem.

Claim 0.4.1. This is a claim.

Corollary 0.5. This is a corollary.

Example 0.1. This is an example.

Contents

Chapter 1

Combinatorial Analysis

1.1 Probabilities

In our lives, we often believe that the future is largely unpredictable. We express this belief through chance behavior and assign both quantitative and qualitative meanings to its usage. Therefore, we introduce the concept of "probability," which aims to provide a numerical description of how likely an event is to occur.

Definition 1.1. **Probability** is a numerical measurement of how likely an event is to occur.

To determine probabilities, we can use random experiments.

Definition 1.2. An **experiment** is a process that produces a random outcome.

Example 1.1. Examples of experiments:

1. Randomly picking a number from 1 to 10.

2. Randomly tossing a coin.

The most basic way to calculate probability is by counting.

Theorem 1.3. (**Fundamental Principle of Counting**) Suppose that m_i represents the number of outcomes of the i-th event. The total number of outcomes of n independent events is the product of the number of outcomes for each individual event:

$$\prod_{i=1}^{n} m_i$$

Example 1.2. Assume that we choose a president and a vice-president from 30 people. By the Fundamental Principle of Counting, the total number of possible outcomes is:

$$30 \times 29 = 870$$

Example 1.3. Assume that we toss a coin 5 times. By the Fundamental Principle of Counting, the total number of possible outcomes is:

$$2^5 = 32$$

Example 1.4. Assume that we roll a die 3 times. By the Fundamental Principle of Counting, the total number of possible outcomes is:

$$6^3 = 216$$

Sometimes, we want to focus on how many ways we can arrange a set of objects.

Definition 1.4. Given a set with n distinct elements:

1. A **permutation** of the set is an ordered arrangement of all elements of the set.
2. If $k \leq n$, a k**-permutation** of the set is an ordered arrangement of k elements of the set.

Remark 1.4.1. The n-permutation is simply the regular permutation.

Remark 1.4.2. To find the total number of permutations, we can think of it as determining the number of outcomes for placing one object in the 1st position, 2nd position, and so on.

Example 1.5. Given a set $\{1, 2, 3\}$:

1. The ordered arrangement $(3, 1, 2)$ is a permutation of the set.
2. The ordered arrangement $(3, 2)$ is a 2-permutation of the set.

We have a formula to find the number of permutations.

Theorem 1.5. Given a set with n distinct elements, the number of permutations of the set is:

$$n!$$

where $n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$ and $0! = 1$.

Theorem 1.6. Let n and k be integers with $0 \leq k \leq n$. The number of k-permutations of a set with n distinct elements, denoted by P_k^n, is given by:

$$P_k^n = \frac{n!}{(n-k)!}$$

Example 1.6. Given a set $\{1, 2, 3, 4\}$:

1. The number of permutations is:
$$P_4^4 = \frac{4!}{(4-4)!} = \frac{24}{1} = 24$$
2. The number of 2-permutations is:
$$P_2^4 = \frac{4!}{(4-2)!} = \frac{24}{2} = 12$$

We also consider cases where the two ends of the ordered arrangement are connected together. For example, when arranging people around a circular table, rotating clockwise or counterclockwise results in the same arrangement.

Theorem 1.7. Given a set with n distinct elements, the number of arrangements of elements in a circle is:

$$(n-1)!$$

Example 1.7. Given a set $\{1, 2, 3, 4\}$, the number of arrangements in a circle is:

$$(4-1)! = 3! = 6$$

1.2 Combinations

Now, assume that we do not care about the ordering of chosen objects. We only want to choose k objects from n objects. Then, we have the following definition.

Definition 1.8. If $k \leq n$, a **k-combination** of a set with n distinct elements is an unordered arrangement of k elements of the set.

Theorem 1.9. Let n and k be integers with $0 \leq k \leq n$. The number of k-combinations of a set with n distinct elements, denoted by C_k^n or $\binom{n}{k}$, is given by:

$$C_k^n = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Proof.
We know that the number of permutations of k objects is $k!$, and the number of ordered arrangements of choosing k objects from n objects is:

$$P_k^n = \frac{n!}{(n-k)!}$$

Since we do not care about the order, the number of unordered arrangements of choosing k objects from n objects is:

$$\binom{n}{k} = \frac{P_k^n}{k!} = \frac{n!}{k!(n-k)!}$$

□

Remark 1.9.1. By convention, when n is a non-negative integer and $k < 0$ or $k > n$, we define:

$$\binom{n}{k} = 0$$

Example 1.8. Given a set $\{1, 2, 3, 4\}$:

1. The number of 2-combinations is:
$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{24}{4} = 6$$
2. The number of 3-combinations is:
$$\binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{24}{6} = 4$$

We can immediately derive the following corollary.

Corollary 1.10. Let n be an integer. For all integers k with $0 \leq k \leq n$, we have:

$$\binom{n}{k} = \binom{n}{n-k}$$

Proof.

$$\binom{n}{n-k} = \frac{n!}{(n-k)!(n-n+k)!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

□

From here, we will provide some important combinatorial identities that could be very useful.

Theorem 1.11. (**Pascal's Identity**) Let n and k be integers with $0 < k < n$. Then:

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

Proof.
From the right-hand side, we have:

$$\binom{n-1}{k-1} + \binom{n-1}{k} = \frac{(n-1)!}{(k-1)!(n-k)!} + \frac{(n-1)!}{k!(n-k-1)!} = \frac{(k+n-k)(n-1)!}{k!(n-k)!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

□

We now introduce the famous Binomial Theorem.

Theorem 1.12. (**Binomial Theorem**) Let n be a non-negative integer. Then:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

where for all k, $\binom{n}{k}$ is called the **binomial coefficient**.

Corollary 1.13. Let n be a non-negative integer. Then:

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n$$

Proof.
Using the Binomial Theorem and substituting $x = 1$ and $y = 1$, we have:

$$2^n = (1+1)^n = \sum_{k=0}^{n} \binom{n}{k} (1)^k (1)^{n-k} = \sum_{k=0}^{n} \binom{n}{k}$$

□

Corollary 1.14. Let n be a positive integer. Then:

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} = 0$$

Proof.
Using the Binomial Theorem and substituting $x = -1$ and $y = 1$, we have:

$$0 = (-1+1)^n = \sum_{k=0}^{n} \binom{n}{k} (-1)^k (1)^{n-k} = \sum_{k=0}^{n} (-1)^k \binom{n}{k}$$

□

Corollary 1.15. Let n be a positive integer. Then:

$$\sum_{k=0}^{n} 2^k \binom{n}{k} = 3^n$$

Proof.
Using the Binomial Theorem and substituting $x = 2$ and $y = 1$, we have:

$$3^n = (2+1)^n = \sum_{k=0}^{n} \binom{n}{k} (2)^k (1)^{n-k} = \sum_{k=0}^{n} 2^k \binom{n}{k}$$

□

Using the Binomial Theorem, we can prove the following identity.

Theorem 1.16. (**Vandermonde's Identity**) Let m, n, r be positive integers with $0 \leq r \leq m$ and $0 \leq r \leq n$. Then:

$$\binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{r-k}\binom{n}{k}$$

Proof.
Assume that we want to find $(x+y)^{m+n}$. Using the Binomial Theorem, we have:

$$\begin{aligned}\sum_{r=0}^{m+n} \binom{m+n}{r} x^r y^{m+n-r} = (x+y)^{m+n} &= \left(\sum_{i=0}^{m} \binom{m}{i} x^i y^{m-i}\right)\left(\sum_{j=0}^{n} \binom{n}{j} x^j y^{n-j}\right) \\ &= \sum_{i=0}^{m}\sum_{j=0}^{n} \binom{m}{i}\binom{n}{j} x^{i+j} y^{m+n-i-j} \\ &= \sum_{r=0}^{m+n}\sum_{k=0}^{r} \binom{m}{r-k}\binom{n}{k} x^r y^{m+n-r} \qquad \text{(Setting } r = i+j \text{ and } k = j\text{)}\end{aligned}$$

If we look at each binomial coefficient, we find that:

$$\binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{r-k}\binom{n}{k}$$

□

Corollary 1.17. Let n be a non-negative integer. Then:

$$\binom{2n}{n} = \sum_{k=0}^{n} \binom{n}{k}^2$$

Proof.
Using Vandermonde's Identity and substituting $m = n$ and $r = n$, we have:

$$\binom{2n}{n} = \sum_{k=0}^{n} \binom{n}{n-k}\binom{n}{k} = \sum_{k=0}^{n} \binom{n}{k}^2$$

□

Theorem 1.18. Let n and r be integers such that $0 \leq r \leq n$. Then:

$$\binom{n+1}{r+1} = \sum_{i=r}^{n} \binom{i}{r}$$

Proof.
We can use Pascal's Identity on the right-hand side:

$$\sum_{i=r}^{n} \binom{i}{r} = \sum_{i=r+1}^{n} \binom{i}{r} + 1 = \sum_{i=r+1}^{n} \binom{i}{r} + \binom{r+1}{r+1} = \binom{n}{r} + \binom{n}{r+1} = \binom{n+1}{r+1}$$

□

1.3 Multinomial

What about when some objects are of the same type?

Theorem 1.19. Given a set with n elements, if we divide n_i objects into the i-th group for all i and $n_1 + n_2 + \cdots + n_k = n$, then the number of different combinations, denoted by $\binom{n}{n_1, n_2, \cdots, n_k}$, is:

$$\binom{n}{n_1, n_2, \cdots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}$$

Remark 1.19.1. When you choose k objects from n objects, you may consider it as classifying k objects as chosen and the remaining $n - k$ objects as not chosen.

We have a more generalized version of the Binomial Theorem.

Theorem 1.20. (**Multinomial Theorem**) Let n be a non-negative integer. Then:

$$(x_1 + x_2 + \cdots + x_k)^n = \sum_{(n_1, n_2, \cdots, n_k): n_1 + n_2 + \cdots + n_k = n} \binom{n}{n_1, n_2, \cdots, n_k} x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k}$$

where $(n_1, n_2, \cdots, n_k)$ are all non-negative integer-valued vectors.

The formula is quite complex! Fortunately, we have a formula to determine how many terms the above equation contains.

Theorem 1.21. There are $\binom{n+r-1}{r-1}$ distinct non-negative integer-valued vectors $(x_1, x_2, \cdots, x_r)$ that satisfy:

$$x_1 + x_2 + \cdots + x_r = n$$

where $x_i \geq 0$ for all i.

Example 1.9. Suppose that a cookie shop has 4 different kinds of cookies. How many different ways can 6 cookies be chosen? The number of ways to choose 6 cookies is the number of 6-combinations with repetition from a set with 4 elements, which is:

$$\binom{4+6-1}{6} = \binom{9}{6} = 84$$

We can also use the above to find the number of ways to distribute n objects into r boxes. If, in addition, every box must contain at least one object, we have the following.

Theorem 1.22. There are $\binom{n-1}{r-1}$ distinct positive integer-valued vectors $(x_1, x_2, \cdots, x_r)$ that satisfy:

$$x_1 + x_2 + \cdots + x_r = n$$

where $x_i > 0$ for all i.

Example 1.10. Suppose that a cookie shop has 4 different kinds of cookies. How many different ways can 6 cookies be chosen if at least one cookie of each kind must be chosen? The number of ways to choose 6 cookies is the number of 6-combinations with repetition from a set with 4 elements, where each element must be chosen at least once, which is:

$$\binom{6-1}{4-1} = \binom{5}{3} = 10$$

Chapter 2

Events and Their Probabilities

We have now understood how to find the number of possibilities. However, we still have not rigorously formalized probability. Therefore, we need to define some basic terminology related to probability.

2.1 Fundamentals

We start with some basic terminology. Many statements in probability take the form of ”the probability of event A is p,” where events usually include some of the elements of the sample space.

Definition 2.1. These are the basic objects of probability:

1. An **experiment** is an activity that produces distinct and well-defined possibilities called **outcomes**, denoted by ω.
2. The **sample space** is the set of all outcomes of an experiment, denoted by Ω.
3. An **event** is a subset of the sample space and is usually represented by $A, B, C, \cdots$.
4. Outcomes are called **elementary events**.

Example 2.1. Examples of sample spaces:

1. Rolling a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
2. Lifetime of a bulb: $\Omega = [0, \infty)$
3. Flipping two coins: $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$

Remark 2.1.1. If the outcome ω is in event A, we say that event A has occurred.

Remark 2.1.2. It is not necessary for all subsets of Ω to be events. However, we do not discuss this issue for the moment.

Example 2.2. Events for rolling a die: Odd ($A = \{1, 3, 5\}$), Even ($A = \{2, 4, 6\}$), $\cdots$

Example 2.3. Events for flipping two coins: At least one head ($A = \{(H, H), (H, T), (T, H)\}$), Two tails ($A = \{(T, T)\}$), $\cdots$

Remark 2.1.3. If only the outcome $\omega = 2$ is given, then there exist many events that can include this outcome, e.g., $\{2\}$, $\{2, 4\}$, $\cdots$

Example 2.4. If we roll a die and the outcome is 2, then the event that occurred could be $\{2\}$, $\{1, 2\}$, $\{2, 4, 6\}$, $\{1, 2, 3, 4, 5, 6\}$, etc.

2.2 Event Operations

We can perform operations on events, similar to sets.

Definition 2.2. Given two events A and B:

1. The **union** of A and B is an event:
$$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$$
2. The **intersection** of A and B is an event:
$$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$$
3. The **complement** of A is an event containing all elements in the sample space Ω that are not in A, or equivalently:
$$A^{\mathsf{C}} = \{\omega \in \Omega : \omega \notin A\}$$
4. The **relative complement** of B in A is an event:
$$A \setminus B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\}$$
5. The **symmetric difference** of A and B is an event:
$$A \Delta B = \{\omega \in \Omega : \omega \in A \cup B \text{ and } \omega \notin A \cap B\}$$

We also need to define the inclusion of all outcomes of one event in another event.

Definition 2.3. For any two events A and B, if all of the outcomes in A are also in B, then we say A is **contained** in B, written as $A \subset B$ or $B \supset A$.

Remark 2.3.1. If $A \subset B$, the occurrence of A necessarily implies the occurrence of B.

We can describe the events in a sample space.

Definition 2.4. Given a sequence of events $A_1, A_2, \cdots, A_k$:

1. For any i and j, if $A_i \cap A_j = \emptyset$, then A_i and A_j are called **disjoint**.
2. If $A_i \cap A_j = \emptyset$ for all i and j, the sequence of events is called **mutually exclusive**.
3. If $A_1 \cup A_2 \cup \cdots \cup A_k = \Omega$, the sequence of events is called **exhaustive**.
4. If the sequence is both mutually exclusive and exhaustive, it is called a **partition**.

We have some fundamental laws for event operations.

Theorem 2.5. Let A, B, C be any three events:

1. Commutative Law:
$$A \cup B = B \cup A \qquad A \cap B = B \cap A$$
2. Associative Law:
$$A \cup (B \cup C) = (A \cup B) \cup C \qquad A \cap (B \cap C) = (A \cap B) \cap C$$
3. Distributive Law:
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \qquad A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

Theorem 2.6. (**De Morgan's Law**) For any k, the sequence of events $A_1, A_2, \cdots, A_k$ satisfies:

$$\left(\bigcup_{i=1}^{k} A_i\right)^{\mathsf{C}} = \bigcap_{i=1}^{k} A_i^{\mathsf{C}} \qquad \left(\bigcap_{i=1}^{k} A_i\right)^{\mathsf{C}} = \bigcup_{i=1}^{k} A_i^{\mathsf{C}}$$

We can also split any event into a union of two events.

Lemma 2.7. For any events A and B, we have:

$$A = (A \cap B) \cup (A \cap B^{\mathsf{C}})$$

Proof.
By the distributive law:

$$(A \cap B) \cup (A \cap B^{\mathsf{C}}) = A \cap (B \cup B^{\mathsf{C}}) = A \cap \Omega = A$$

□

We may now start defining probability. Let us begin by defining a collection of subsets of the sample space.

Definition 2.8. A **field** $\mathcal{F}$ is any collection of subsets of the sample space Ω that satisfies the following conditions:

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

We are more interested in σ-fields that are closed under countably infinite unions.

Definition 2.9. A **σ-field** (or **σ-algebra**) $\mathcal{F}$ is any collection of subsets of the sample space Ω that satisfies the following conditions:

1. If $A \in \mathcal{F}$, then $A^{\mathsf{C}} \in \mathcal{F}$.
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 $\Omega = A \cup A^{\mathsf{C}} \cup \cdots \in \mathcal{F}$.

Remark 2.9.1. All σ-fields are fields. The converse is not necessarily true.

Remark 2.9.2. From this point onwards, $\mathcal{F}$ represents the σ-field.

Example 2.5. Smallest σ-field: $\mathcal{F} = \{\emptyset, \Omega\}$

Example 2.6. If A is any subset of Ω, then $\mathcal{F} = \{\emptyset, A, A^{\mathsf{C}}, \Omega\}$ is a σ-field.

Example 2.7. 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 2.9.3. For any $a < b$:

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

2.3 Probability Measure and Kolmogorov Axioms

Now that we have defined some fundamental terminology, we can finally define probability.

Definition 2.10. 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. (Countable additivity) If $A_i \in \mathcal{F}$ for all i and they are disjoint, then

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

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

You may ask, "Isn't it just probability?" The probability that we know is indeed a probability measure, which we will define soon. However, there are other measures that satisfy the definition of a probability measure, e.g., the risk-neutral measure.

Example 2.8. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $\mathcal{F} = \{0, 1\}^{\Omega}$. Define $\mathbb{P}(A) = \frac{|A|}{6}$ for all $A \in \mathcal{F}$. Then $\mathbb{P}$ is a probability measure.

The following measures are not probability measures:

Example 2.9. Lebesgue measure: $\mu((a, b)) = b - a$, $\Omega = \mathbb{R}$

Example 2.10. Counting measure: $\mu(A) = \#\{A\}$, $\Omega = \mathbb{R}$

We can combine a measurable space and a measure into a measure space.

Definition 2.11. 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.

Kolmogorov's axioms of probability use axioms to formalize probability.

Definition 2.12. (**Kolmogorov Axioms of Probability**) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, with sample space Ω, σ-field $\mathcal{F}$, and probability measure $\mathbb{P}$.

1. The probability of an event is a non-negative real number. For all $E \in \mathcal{F}$:

$$\mathbb{P}(E) \in \mathbb{R} \qquad\qquad \mathbb{P}(E) \geq 0$$

2. The probability that at least one of the elementary events in the entire sample space will occur is 1:

$$\mathbb{P}(\Omega) = 1$$

3. Any countable sequence of disjoint events $\{E_i\}_i \subset \mathcal{F}$ satisfies:

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

By this definition, we call $\mathbb{P}(A)$ the **probability** of the event A.

Example 2.11. 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 2.12. 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 we have:

$$\mathbb{P}(A) = \frac{|A|}{6}$$

The following properties are important and form the foundation of probability.

Lemma 2.13. For any $A, B \in \mathcal{F}$, $\mathbb{P}$ satisfies the following properties:

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 $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

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. $A \cup B = A \cup (B \setminus A) \Longrightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(A) + \mathbb{P}(B \setminus (A \cap B)) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

□

Theorem 2.14. (**Inclusion-Exclusion Formula**) For any set of events $\{A_1, \cdots, A_n\} \in \mathcal{F}$:

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

Proof.
By induction. When $n = 1$, it is obviously true. Assume it is true for some positive integer m. When $n = m + 1$:

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

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

We recall the continuity of a function $f : \mathbb{R} \to \mathbb{R}$. The function f is continuous at some point x if, for all sequences x_n such that $x_n \to x$ as $n \to \infty$, 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 sequences of sets $\{A_i\}_i$ 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 2.14.1. Given a sequence of sets $\{A_i\}_i \subset \mathcal{F}$, we have two types of set limits:

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

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

Definition 2.15. We say a sequence of events $\{A_i\}_i \subset \mathcal{F}$ **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 $\{A_i\}_i \subset \mathcal{F}$ 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 this definition, we derive the following important lemma.

Lemma 2.16. If $\{A_i\}_i \subset \mathcal{F}$ is 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 $\{A_i\}_i \subset \mathcal{F}$ is 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 $B_n = A_n \setminus A_{n-1}$. Then:

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

For $A_1 \supseteq A_2 \supseteq \cdots$, we have $A^{\complement} = \bigcup_{i=1}^{\infty} A_i^{\complement}$ 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 assign terminology to some special probabilities.

Definition 2.17. An event A is **null** if $\mathbb{P}(A) = 0$.

Remark 2.17.1. Null events need not be impossible. For example, $\mathbb{P}(\text{Choosing a specific point in a plane}) = 0$.

Definition 2.18. An event A occurs **almost surely** if $\mathbb{P}(A) = 1$.

2.4 Conditional Probability

Let $\mathcal{F}$ be a σ-field of the sample space Ω. Sometimes, we are interested in the probability of a certain event given that another event has occurred.

Definition 2.19. For any $A, B \in \mathcal{F}$, 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 2.19.1. For any $A \in \mathcal{F}$, $\mathbb{P}(A)$ can be regarded as $\mathbb{P}(A|\Omega)$.

Remark 2.19.2. For any $E, F \in \mathcal{F}$, when $\mathbb{P}(E) = \mathbb{P}(E|F)$, E and F are said to be **independent**.

Example 2.13. Two fair dice are thrown. Given that the first shows 3, what is the probability that the sum of the numbers shown exceeds 6?

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

Lemma 2.20. For any $B \in \mathcal{F}$, if $\mathbb{P}(B) > 0$, $\mathbb{P}(\cdot|B)$ is a probability measure on $\mathcal{F}$.

Proof.
We prove this from the definition of a probability measure.

1. We prove $\mathbb{P}(\emptyset|B) = 0$. Since $\mathbb{P}(B) > 0$:

$$\mathbb{P}(\emptyset|B) = \frac{\mathbb{P}(\emptyset \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(B)} = 0.$$

2. We prove $\mathbb{P}(\Omega|B) = 1$. Since $\mathbb{P}(B) > 0$:

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

3. We prove countable additivity. Since $\mathbb{P}(B) > 0$, for any disjoint sequence of events $A_i \in \mathcal{F}$ for all i:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \middle| B\right) = \frac{1}{\mathbb{P}(B)}\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \cap B\right) = \frac{1}{\mathbb{P}(B)}\mathbb{P}\left(\bigcup_{i=1}^{\infty}(A_i \cap B)\right) = \frac{1}{\mathbb{P}(B)}\sum_{i=1}^{\infty}\mathbb{P}(A_i \cap B) = \sum_{i=1}^{\infty}\mathbb{P}(A_i|B).$$

Therefore, for any $B \in \mathcal{F}$, if $\mathbb{P}(B) > 0$, then $\mathbb{P}(\cdot|B)$ is a probability measure. □

We may create a series of probabilities based on previous events. This is useful when dealing with a sequence of events over time.

Lemma 2.21. (**General Multiplication Rule**) Let $\{A_1, \cdots, A_n\} \subset \mathcal{F}$ be a sequence of events. We have:

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2)\cdots\mathbb{P}(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Proof.

$$\begin{aligned}
\mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2)\cdots\mathbb{P}(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}) &= \mathbb{P}(A_1 \cap A_2)\mathbb{P}(A_3|A_1 \cap A_2)\cdots\mathbb{P}(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}) \\
&= \mathbb{P}(A_1 \cap A_2 \cap A_3)\cdots\mathbb{P}(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}) \\
&= \mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right).
\end{aligned}$$

□

It is evident that a certain event occurs when another event either occurs or does not occur.

Lemma 2.22. For any $A, B \in \mathcal{F}$ such that $0 < \mathbb{P}(B) < 1$:

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^{\mathsf{C}})\mathbb{P}(B^{\mathsf{C}}).$$

Proof.

$$A = (A \cap B) \cup (A \cap B^{\mathsf{C}}) \implies \mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^{\mathsf{C}}) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^{\mathsf{C}})\mathbb{P}(B^{\mathsf{C}}).$$

□

Example 2.14. In medical cases, we often evaluate the efficiency and effectiveness of a medical test. Each type of result has a specific name:

1. True positive (TP): Sick individuals correctly identified as sick (Found positive and correct).
2. False positive (FP): Healthy individuals incorrectly identified as sick (Found positive but incorrect).
3. True negative (TN): Healthy individuals correctly identified as healthy (Found negative and correct).
4. False negative (FN): Sick individuals incorrectly identified as healthy (Found negative but incorrect).

There are cases where multiple events contribute to the occurrence of a certain event.

Lemma 2.23. (**Law of Total Probability**) Let $\{B_1, \cdots, B_n\} \subset \mathcal{F}$ be a partition of Ω. 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.

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

□

At this point, we can prove a theorem widely used in fields outside mathematics. Imagine knowing the probability of each type of disease and the probability of having a specific symptom given the disease. If a patient has the symptom, what is the probability they have the disease you are considering?

Theorem 2.24. (**Bayes' Theorem**) Suppose that a sequence of events $\{A_1, \cdots, A_n\} \subset \mathcal{F}$ forms a partition of the sample space. Assume further that $\mathbb{P}(A_i) > 0$ for all i. For any $B \in \mathcal{F}$:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{k=1}^{n} \mathbb{P}(B|A_k)\mathbb{P}(A_k)}, \qquad \text{for } i = 1, \cdots, n.$$

Proof.

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{k=1}^{n} \mathbb{P}(B|A_k)\mathbb{P}(A_k)}.$$

□

2.5 Independence

In general, the probability of a certain event is influenced by the occurrence of other events. However, there are exceptions.

Definition 2.25. Two events A and B are **independent**, denoted by $A \perp\!\!\!\perp B$, if:

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

Remark 2.25.1. If either $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$, then $A \perp\!\!\!\perp B$.

Remark 2.25.2. If events A and B are independent and $A \cap B = \emptyset$, then either $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$.

The following is relatively simple to prove.

Lemma 2.26. For any $A, B \in \mathcal{F}$, if $A \perp\!\!\!\perp B$, then:

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

Proof.

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

□

Proposition 2.27. If events A and B are independent, then $A \perp\!\!\!\perp B^{\mathsf{C}}$ and $A^{\mathsf{C}} \perp\!\!\!\perp B^{\mathsf{C}}$.

Proof.

$$\begin{aligned}\mathbb{P}(A \cap B^{\mathsf{C}}) &= \mathbb{P}(A) - \mathbb{P}(A \cap B)\\ &= \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B)\\ &= \mathbb{P}(A)(1 - \mathbb{P}(B))\\ &= \mathbb{P}(A)\mathbb{P}(B^{\mathsf{C}}).\end{aligned}$$

Therefore, $A \perp\!\!\!\perp B^{\mathsf{C}}$ and also $A^{\mathsf{C}} \perp\!\!\!\perp B^{\mathsf{C}}$. □

Proposition 2.28. If events A, B, C are independent, then $A \perp\!\!\!\perp (B \cup C)$ and $A \perp\!\!\!\perp (B \cap C)$.

Proof.
Using the properties of probability:

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

□

Sometimes, we may deal with more than two events. We have a more specific way to describe their relationships.

Definition 2.29. Given a family of events $\{A_i : i \in I\} \subset \mathcal{F}$ for some $I \subset \mathbb{N}_+$:

1. If $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for any $i \neq j$, the events are **pairwise independent**.
2. If, additionally, for all subsets J of I:

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

then the events are **mutually independent**.

Remark 2.29.1. Usually, when we say multiple events are independent, we mean they are mutually independent.

Example 2.15. Consider tossing a fair coin three times. Define the following events:

$$\begin{aligned}
A &= \{\text{The first toss is a head}\} = \{HHH, HHT, HTH, HTT\},\\
B &= \{\text{The second toss is a head}\} = \{HHH, HTH, THH, THT\},\\
C &= \{\text{The third toss is a head}\} = \{HHH, HHT, THH, TTH\}.
\end{aligned}$$

We can observe the intersections between each pair of events and the overall intersection:

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \mathbb{P}(\{HHH, HHT\}) = \frac{2}{8} = \frac{1}{2}\left(\frac{1}{2}\right) = \mathbb{P}(A)\mathbb{P}(B),\\
\mathbb{P}(B \cap C) &= \mathbb{P}(\{HHH, THH\}) = \frac{2}{8} = \frac{1}{2}\left(\frac{1}{2}\right) = \mathbb{P}(B)\mathbb{P}(C),\\
\mathbb{P}(A \cap C) &= \mathbb{P}(\{HHH, HHT\}) = \frac{2}{8} = \frac{1}{2}\left(\frac{1}{2}\right) = \mathbb{P}(A)\mathbb{P}(C),\\
\mathbb{P}(A \cap B \cap C) &= \mathbb{P}(\{HHH\}) = \frac{1}{8} = \frac{1}{2}\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C).
\end{aligned}$$

Therefore, events A, B, and C are mutually independent.

Remark 2.29.2. Not all pairwise independent events are mutually independent.

Example 2.16. Consider rolling a die twice: $\Omega = \{1, 2, \cdots, 6\} \times \{1, 2, \cdots, 6\}$ and $\mathcal{F} = 2^\Omega$. Define the following events:

$$\begin{aligned}
A &= \{\text{The sum is 7}\} = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\},\\
B &= \{\text{The first roll is 4}\} = \{(4,1), (4,2), (4,3), (4,4), (4,5), (4,6)\},\\
C &= \{\text{The second roll is 3}\} = \{(1,3), (2,3), (3,3), (4,3), (5,3), (6,3)\}.
\end{aligned}$$

We can observe the intersections between each pair of events and the overall intersection:

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

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

2.6 Product Space

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

Definition 2.30. 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 2.30.1. $\sigma(A)$ is the smallest σ-field containing A.

Example 2.17. Let $\Omega = \{1, 2, \cdots, 6\}$ and $A = \{\{1\}\} \subseteq 2^{\Omega}$. Then $\sigma(A) = \{\emptyset, \{1\}, \{2, 3, \cdots, 6\}, \Omega\}$.

Corollary 2.31. 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 to generate, we can combine two probability spaces to form a new probability space.

Definition 2.32. The **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 $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$.

Example 2.18. Assume that we want to consider the probabilities of getting a head in a coin flip and getting a 5 in a die toss simultaneously. We already know:

$$\begin{aligned} &\Omega_1 = \{H, T\}, && \mathcal{F}_1 = \{\emptyset, \{H\}, \{T\}, \Omega_1\}, && \mathbb{P}_1 = \mathbb{P}(\cdot|\Omega_1), && \text{(Coin flipping)} \\ &\Omega_2 = \{1, 2, 3, 4, 5, 6\}, && \mathcal{F}_2 = 2^{\Omega_2}, && \mathbb{P}_2 = \mathbb{P}(\cdot|\Omega_2). && \text{(Die tossing)} \end{aligned}$$

The probability space we are considering is the product space $(\Omega_1 \times \Omega_2, \mathcal{G}, \mathbb{P}_{12})$, where:

$$\begin{aligned} \Omega_1 \times \Omega_2 &= \{(H, 1), (H, 2), (H, 3), (H, 4), (H, 5), (H, 6), (T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6)\}, \\ \mathcal{G} &= 2^{\Omega}, \\ \mathbb{P}_{12} &= \mathbb{P}(\cdot|\Omega_1 \times \Omega_2) = \mathbb{P}(\cdot|\Omega_1)\mathbb{P}(\cdot|\Omega_2). \end{aligned}$$

Chapter 3

Random Variables and Their Distribution

In this chapter, let $\mathcal{F}$ be a σ-field of a sample space Ω, and let $\mathbb{P}$ be a probability measure.

3.1 Introduction to Random Variables

Sometimes, we are not interested in an experiment itself but rather in the consequence of its random outcome. We can consider this consequence as a function that maps a sample space into the real number field. We call these functions ”random variables.”

Definition 3.1. A **random variable** (r.v.) 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 3.1.1. More generally, a random variable is a function X with the property that for all intervals $A \subseteq \mathbb{R}$:

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

We say the function is **$\mathcal{F}$-measurable**. Any function that is $\mathcal{F}$-measurable is a random variable.

Remark 3.1.2. All intervals can be replaced by any of the following classes:

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

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

Claim 3.1.1. 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 $a, b \in \mathbb{R}$:

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

□

Example 3.1. 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.

We can create new random variables from X.

Lemma 3.2. Given a random variable X:

1. If $Y = cX + d$ for some $c, d \in \mathbb{R}$, then Y is a random variable.
2. If $Z = X^2$, then Z is a random variable.

Proof.

1. Let $y \in \mathbb{R}$. If $c = 0$, then:
$$Y^{-1}((-\infty, y]) = \{\omega \in \Omega : d \leq y\} \in \mathcal{F}.$$
Otherwise:
$$Y^{-1}((-\infty, y]) = \{\omega \in \Omega : Y(\omega) \leq y\} = \begin{cases} \left\{\omega \in \Omega : X(\omega) \leq \frac{y-d}{c}\right\} \in \mathcal{F}, & c > 0, \\ \left\{\omega \in \Omega : X(\omega) \geq \frac{y-d}{c}\right\} \in \mathcal{F}, & c < 0. \end{cases}$$
Therefore, for any $c, d \in \mathbb{R}$, $Y = cX + d$ is a random variable.
2. Let z be a positive number. We have:
$$Z^{-1}([0, z]) = \{\omega \in \Omega : 0 \leq Z(\omega) \leq z\} = \{\omega \in \Omega : -\sqrt{z} \leq X(\omega) \leq \sqrt{z}\} \in \mathcal{F}.$$
Therefore, $Z = X^2$ is a random variable.

□

Before we continue, it is best to understand Borel sets first.

Definition 3.3. A **Borel set** is a set that can be obtained by taking countable unions, intersections, or complements repeatedly (countably many steps).

Definition 3.4. The **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 3.2. $\{(a, b), [a, b], \{a\}, \mathbb{Q}, \mathbb{R} \setminus \mathbb{Q}\} \subset \mathcal{B}(\mathbb{R})$. Note that closed sets can be generated by open sets.

Remark 3.4.1. In modern understanding: $(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{X} (\mathbb{R}, \mathcal{B}, \mathbb{P} \circ X^{-1})$.

Claim 3.4.1. $\mathbb{P} \circ X^{-1}$ is a probability measure on $(\mathbb{R}, \mathcal{B})$.

Proof.

1. For all $B \in \mathcal{B}$, $\mathbb{P} \circ X^{-1}(B) = \mathbb{P}(\{\omega : X(\omega) \in B\}) \in [0, 1]$:
$$\mathbb{P} \circ X^{-1}(\emptyset) = \mathbb{P}(\{\omega : X(\omega) \in \emptyset\}) = \mathbb{P}(\emptyset) = 0, \qquad \mathbb{P} \circ X^{-1}(\mathbb{R}) = \mathbb{P}(\{\omega : X(\omega) \in \mathbb{R}\}) = \mathbb{P}(\Omega) = 1.$$
2. For any disjoint $\{B_i\}_i \subset \mathcal{B}$:
$$\mathbb{P} \circ X^{-1}\left(\bigcup_{i=1}^{\infty} B_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} X^{-1}(B_i)\right) = \sum_{i=1}^{\infty} \mathbb{P}(X^{-1}(B_i)) = \sum_{i=1}^{\infty} \mathbb{P} \circ X^{-1}(B_i).$$

□

Remark 3.4.2. We can derive the probability of all $[a, b] \in \mathcal{B}$:

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

3.2 CDF of Random Variables

Every random variable has its own distribution function.

Definition 3.5. The **(Cumulative) Distribution Function** (CDF) of a random variable X is a function $F_X : \mathbb{R} \to [0,1]$ given by:

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

Example 3.3. From Example 3.1:

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

Remark 3.5.1. $F_X(x)$ is non-decreasing because if $x < y$, then $\{X \leq x\} \subseteq \{X \leq y\}$.

Remark 3.5.2. $F_X(x)$ is bounded because $0 \leq F_X(x) \leq 1$ for all $x \in \mathbb{R}$.

Lemma 3.6. The CDF F_X of a random variable X has the following properties for any $x, y \in \mathbb{R}$:

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 $h \to 0$).

Proof.

1. Let $B_n = \{\omega \in \Omega : X(\omega) \leq -n\}$. Since $B_1 \supseteq B_2 \supseteq \cdots$, by Lemma 2.16:

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

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 $C_n = \{\omega \in \Omega : X(\omega) \leq n\}$. Since $C_1 \subseteq C_2 \subseteq \cdots$, by Lemma 2.16:

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

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 $A(x) = \{X \leq x\}$ and $A(x,y) = \{x < X \leq y\}$. 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 $B_n = \left\{\omega \in \Omega : X(\omega) \leq x + \frac{1}{n}\right\}$. Since $B_1 \supseteq B_2 \supseteq \cdots$, by Lemma 2.16:

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

Alternative proof:

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

□

Remark 3.6.1. F is not left-continuous because:

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

Remark 3.6.2. Two random variables X and Y are **identically distributed** if they have the same CDF, i.e., $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$.

Lemma 3.7. Let F_X be the CDF of a random variable X. Then for any $x, y \in \mathbb{R}$ such that $x < y$:

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

□

In some cases, we want to find a number where a specific percentage of outcomes fall below it. This is very useful if you know that the random variable follows a certain distribution.

Definition 3.8. The ***q*-th quantile** of a random variable X is defined as a number z_q such that:

$$\mathbb{P}(X \leq z_q) = q.$$

Remark 3.8.1. If F_X is continuous and strictly increasing, then $z_q = F_X^{-1}(q)$.

Example 3.4. The median is the 0.5-th quantile. The quartiles are the 0.25-th and 0.75-th quantiles.

3.3 PMF / PDF of Random Variables

We can classify some (not all) random variables as either discrete or continuous. These two types will be further discussed in the next two chapters.

Definition 3.9. A random variable X is **discrete** if it takes values in some countable subset $\{x_1, x_2, \cdots\} \subset \mathbb{R}$.

Definition 3.10. A discrete random variable X has a **probability mass function** (PMF) $f_X : \mathbb{R} \to [0,1]$ given by:

$$f_X(x) = \mathbb{P}(X = x) = \mathbb{P} \circ X^{-1}(\{x\}), \qquad \text{for } x \in \mathbb{R}.$$

Remark 3.10.1. Some textbooks use $p_X(x)$ to denote the PMF to prevent confusion with the PDF.

Remark 3.10.2. For discrete random variables, the distribution is **atomic** because the distribution function has jump discontinuities at values $x_1, x_2, \cdots$ and is constant in between.

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

Definition 3.11. A random variable X is called **continuous** if its distribution function can be expressed as:

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

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

Remark 3.11.1. For small $\delta > 0$:

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

Remark 3.11.2. For continuous random variables, the CDF is **absolutely continuous**.

Remark 3.11.3. Not every continuous function can be written as $\int_{-\infty}^{x} f_X(u)\, du$. For example, the Cantor function.

Remark 3.11.4. It is possible for a random variable to be neither continuous nor discrete.

3.4 JCDF of Random Variables

Often, we want to study the simultaneous behavior of multiple events. In such cases, we involve two or more random variables that share the same sample space.

Definition 3.12. Let $X_1, X_2, \dots, X_n$ be n random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. We define a **random vector $\mathbf{X}$** with the property:

$$\mathbf{X}^{-1}(B) = \{\omega \in \Omega : \mathbf{X}(\omega) = (X_1(\omega), \dots, X_n(\omega)) \in B\} \in \mathcal{F}, \qquad \text{for } B \in \mathcal{B}(\mathbb{R}^n).$$

We can also say $\mathbf{X}$ is a random vector if X_i are random variables for $i = 1, \dots, n$. This means:

$$X_i^{-1}(B) \in \mathcal{F}, \qquad \text{for } B \in \mathcal{B}(\mathbb{R}),$$

for $i = 1, \dots, n$.

Claim 3.12.1. Both definitions of random vectors are equivalent.

Proof.
By the first definition, $\mathbf{X}^{-1}(A_1 \times \cdots \times A_n) \in \mathcal{F}$. If we choose $A_2 = A_3 = \cdots = A_n = \mathbb{R}$:

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

This means X_1 is a random variable. Similarly, we can also find that X_i is a random variable for $i = 2, \dots, n$. Therefore, we can derive the second definition from the first definition.
By the second definition, X_i are random variables for $i = 1, \dots, n$. Therefore:

$$\begin{aligned}\mathbf{X}^{-1}(A_1 \times \cdots \times A_n) &= \{\omega \in \Omega : (X_1(\omega), \dots, X_n(\omega)) \in A_1 \times \cdots \times A_n\} \\ &= \{\omega \in \Omega : X_1(\omega) \in A_1\} \cap \cdots \cap \{\omega \in \Omega : X_n(\omega) \in A_n\} \\ &= X_1^{-1}(A_1) \cap \cdots \cap X_n^{-1}(A_n) \in \mathcal{F}.\end{aligned}$$

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

Remark 3.12.1. Alternatively, for any $B \in \mathcal{B}(\mathbb{R}^n)$:

$$\mathbb{P} \circ \mathbf{X}^{-1}(B) = \mathbb{P}(\mathbf{X} \in B) = \mathbb{P}(\{\omega \in \Omega : \mathbf{X}(\omega) = (X_1(\omega), \dots, X_n(\omega)) \in B\}).$$

Remark 3.12.2. We can replace all Borel sets with the form $[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$.

In general, we can define a distribution function corresponding to the random vector.

Definition 3.13. The **Joint (Cumulative) Distribution Function** (JCDF) $F_{\mathbf{X}} : \mathbb{R}^n \to [0, 1]$ is defined as:

$$F_{\mathbf{X}}(x_1, \dots, x_n) = F_{X_1,\dots,X_n}(x_1, \dots, x_n) = \mathbb{P} \circ \mathbf{X}^{-1}((-\infty, x_1] \times \cdots \times (-\infty, x_n]) = \mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n),$$

for $x_1, \dots, x_n \in \mathbb{R}$.

Remark 3.13.1. The random variables being of the same type is not necessary to define the JCDF.

The joint distribution function has properties similar to those of a normal distribution function.

Lemma 3.14. The JCDF $F_{X,Y}$ of two random variables X and 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 $x_1 \le y_1$ and $x_2 \le y_2$, then $F_{X,Y}(x_1, y_1) \le F_{X,Y}(x_2, y_2)$.
3. $F_{X,Y}$ is continuous from above, meaning $F_{X,Y}(x + u, y + v) \to F_{X,Y}(x, y)$ as $u \to 0^+$ and $v \to 0^+$.

Most of the time, we consider the case when the random variables are either all discrete or all continuous. For simplicity, we consider only two random variables from the same probability space.

Definition 3.15. 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.

Definition 3.16. Given two jointly discrete random variables X and Y, the **Joint (Probability) Mass Function** (JPMF) $f_{X,Y} : \mathbb{R}^2 \to [0, 1]$ is given by:

$$f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) = \mathbb{P} \circ (X, Y)^{-1}(\{(x, y)\}), \qquad \text{for } x, y \in \mathbb{R}.$$

Remark 3.16.1. If two random variables X and Y are jointly discrete, we can find the JCDF by:

$$F_{X,Y}(x, y) = \sum_{u \leq x} \sum_{v \leq y} f(u, v), \qquad \text{for } x, y \in \mathbb{R}.$$

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

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

Example 3.5. Assume that a special three-sided coin is provided. Each toss results in a 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) satisfies $H_n + T_n + E_n = n$.

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

Similar to the idea of defining the probability density function, we can define the joint probability density function for two jointly continuous random variables.

Definition 3.17. 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:

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

Remark 3.17.1. More generally, for all $B \in \mathcal{B}(\mathbb{R}^2)$:

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

Remark 3.17.2. It is not generally true for two continuous random variables X and Y to be jointly continuous.

Example 3.6. 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 $Y = X$. 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:

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

Therefore, X and Y are not jointly continuous.

Remark 3.17.3.

$$f_{X,Y}(x, y) = F_{X,Y}(x, y) - F_{X,Y}(x^-, y) - F_{X,Y}(x, y^-) + F_{X,Y}(x^-, y^-), \qquad \text{for } x, y \in \mathbb{R}.$$

3.5 Marginal Distribution of Random Variables

We can find the probability distribution of one random variable by disregarding another variable. This results in the following distribution:

Definition 3.18. Let X, Y be jointly discrete random variables. We can obtain a **marginal distribution** (marginal CDF) as follows:

$$F_X(x) = \mathbb{P} \circ X^{-1}((-\infty, x]) = \mathbb{P}\left(X^{-1}((-\infty, x]) \cap Y^{-1}((-\infty, \infty))\right), \qquad \text{for } x \in \mathbb{R}.$$

Remark 3.18.1. For all $x \in \mathbb{R}$, the marginal distribution can be obtained by:

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

Definition 3.19. Let X, Y be jointly discrete random variables with a JPMF $f_{X,Y}$. The **marginal PMF** of X is given by:

$$f_X(x) = \mathbb{P}(X = x) = \sum_{y \in \mathbb{R}} \mathbb{P}(X = x, Y = y) = \sum_{y \in \mathbb{R}} f_{X,Y}(x, y), \qquad \text{for } x \in \mathbb{R}.$$

Example 3.7. From Example 3.1, let X be the number of heads and Y be the number of tails when tossing a coin twice. The JPMF is given by:

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{4}, & (x, y) = (0, 2), (1, 1), (2, 0), \\ 0, & \text{otherwise.} \end{cases}$$

The marginal PMF of X is given by:

$$f_X(x) = \sum_{y \in \mathbb{R}} f_{X,Y}(x, y) = \begin{cases} \frac{1}{4}, & x = 0, \\ \frac{1}{2}, & x = 1, \\ \frac{1}{4}, & x = 2, \\ 0, & \text{otherwise.} \end{cases}$$

Definition 3.20. Let X, Y be jointly continuous random variables with a JPDF $f_{X,Y}$. The **marginal PDF** of X is given by:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy, \qquad \text{for } x \in \mathbb{R}.$$

Example 3.8. Let X and Y be jointly continuous random variables with a JPDF:

$$f_{X,Y}(x, y) = \begin{cases} 2, & 0 < x < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

The marginal PDF of X is given by:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy = \int_x^1 2\, dy = 2(1 - x), \qquad \text{for } 0 < x < 1.$$

Chapter 4

Discrete Random Variables

4.1 Introduction to Discrete Random Variables

Let us recall some of the definitions of discrete random variables from the previous chapter.

Definition 4.1. A random variable X is **discrete** if it takes values in some countable subset $\{x_1, x_2, \dots\} \subset \mathbb{R}$.

Definition 4.2. The **Probability Mass Function** (PMF) of a discrete random variable X is the function $f_X : \mathbb{R} \to [0,1]$ given by:

$$f_X(x) = \mathbb{P}(X = x), \qquad \text{for } x \in \mathbb{R}.$$

Definition 4.3. The **(Cumulative) Distribution Function** (CDF) of a discrete random variable X is the function $F_X : \mathbb{R} \to [0,1]$ given by:

$$F_X(x) = \mathbb{P}(X \le x) = \sum_{i : x_i \le x} f_X(x_i), \qquad \text{for } x \in \mathbb{R}.$$

Lemma 4.4. The relationship between the PMF f_X and the CDF F_X of a random variable X for any $x, y \in \mathbb{R}$ is as follows:

1. $F_X(x) = \sum_{y \le x} f_X(y)$.
2. $f_X(x) = F_X(x) - \lim_{y \to x^-} F_X(y)$.

Proof.

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

2. Let $B_n = \left\{x - \frac{1}{n} < X \le x\right\}$. Since $B_1 \supseteq B_2 \supseteq \cdots$, by Lemma 2.16:

$$\begin{aligned} F_X(x) - \lim_{y \to x^-} F_X(y) &= \mathbb{P}\left(\bigcap_{n=1}^{\infty} B_n\right) \\ &= \mathbb{P}\left(\lim_{n \to \infty} B_n\right) \\ &= \mathbb{P}\left(\left\{\lim_{n \to \infty}\left(x - \frac{1}{n}\right) < X \le x\right\}\right) \\ &= \mathbb{P}(X = x). \end{aligned}$$

□

Lemma 4.5. The PMF $f_X : \mathbb{R} \to [0,1]$ of a discrete random variable X satisfies the following properties:

1. $\{x \in \mathbb{R} : f_X(x) \ne 0\}$ is countable.
2. $\sum_i f_X(x_i) = 1$, where $x_1, x_2, \dots$ are the values of x such that $f_X(x) \ne 0$.

Let us also recall the definitions of the joint distribution function and the joint mass function.

Definition 4.6. For jointly discrete random variables X and Y, the **Joint Probability Mass Function** (JPMF) $f_{X,Y} : \mathbb{R}^2 \to [0,1]$ is given by:

$$f_{X,Y}(x,y) = \mathbb{P}((X,Y) = (x,y)) = \mathbb{P} \circ (X,Y)^{-1}(\{x,y\}), \qquad \text{for } x, y \in \mathbb{R}.$$

Definition 4.7. For jointly discrete random variables X and Y with a JPMF $f_{X,Y}$, the **Joint Cumulative Distribution Function** (JCDF) $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ is given by:

$$F_{X,Y}(x,y) = \sum_{u \le x} \sum_{v \le y} f(u,v), \qquad \text{for } x, y \in \mathbb{R}.$$

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

Definition 4.8. 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 $A, B \in \mathcal{B}(\mathbb{R})$, or
2. $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for all $x, y \in \mathbb{R}$, or
3. $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$.

Claim 4.8.1. These three definitions are equivalent.

Proof.

2 → 1

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

3 → 2

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

1 → 3

$$\begin{aligned} \mathbb{P} \circ (X,Y)^{-1}(A \times B) &= \sum_{(x,y) \in A \times B} f_{X,Y}(x,y) \\ &= \sum_{x \in A} \sum_{y \in B} f_X(x)f_Y(y) \\ &= (\mathbb{P} \circ X^{-1}(A))(\mathbb{P} \circ Y^{-1}(B)). \end{aligned}$$

Therefore, these three definitions are equivalent. □

Remark 4.8.1. More generally, let $X_1, X_2, \dots, X_n$ be discrete random variables. They are **independent** if:

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

Definition 4.9. Given two discrete random variables X and Y, the **marginal probability mass function** (marginal PMF) of X is given by:

$$f_X(x) = \sum_{y \in \mathbb{R}} f_{X,Y}(x, y), \qquad \text{for } x \in \mathbb{R}.$$

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

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

Remark 4.9.1. 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 4.9.2. We can generate a σ-field using random variables by defining the σ-field generated by a random variable X as:

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

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

Definition 4.10. Let $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ be two σ-fields. We say $\mathcal{G}$ and $\mathcal{H}$ are **independent** if, for all $A \in \mathcal{G}$ and $B \in \mathcal{H}$:

$$A \perp\!\!\!\perp B.$$

Remark 4.10.1. For any discrete random variables X and Y:

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

Theorem 4.11. 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 4.11.1. We assume a product space $(\Omega, \mathcal{F}, \mathbb{P})$ of two probability spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$. Any pair of events of the form $E_1 \times \Omega_2$ and $\Omega_1 \times E_2$ are independent:

$$\begin{aligned}
\mathbb{P}((E_1 \times \Omega_2) \cap (\Omega_1 \times E_2)) &= \mathbb{P}(E_1 \times E_2) \\
&= \mathbb{P}_1(E_1)\mathbb{P}_2(E_2) \\
&= \mathbb{P}(E_1 \times \Omega_2)\mathbb{P}(\Omega_1 \times E_2).
\end{aligned}$$

4.2 Conditional Distribution of Discrete Random Variables

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

Definition 4.12. Suppose X, Y are two discrete random variables. The **Conditional Distribution** of Y given $X = x$ for any $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) > 0$ is defined as:

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

The **Conditional Mass Function** (Conditional PMF) of Y given $X = x$ for any $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) > 0$ is defined as:

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

The **Conditional Distribution Function** (Conditional CDF) of Y given $X = x$ for any $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) > 0$ is defined as:

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

Remark 4.12.1. By definition:

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

Remark 4.12.2. For any $x \in \mathbb{R}$, the conditional PMF $f_{Y|X}(y|x)$ is a probability mass function in y.

Remark 4.12.3. If X and Y are independent, then:

$$f_{Y|X}(y|x) = f_Y(y), \qquad \text{for } x, y \in \mathbb{R}.$$

Conditional distributions retain the properties of the original distribution.

Lemma 4.13. Given two discrete random variables X and Y, conditional distributions have the following properties for $x, y \in \mathbb{R}$:

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. Using Lemma 4.4:

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{1}{f_X(x)}\mathbb{P}(X = x, Y = y) \\ &= \frac{1}{f_X(x)}\left(\mathbb{P}(X = x, Y \leq y) - \lim_{z \to y^-} \mathbb{P}(X = x, Y \leq z)\right) \\ &= \mathbb{P}(Y \leq y|X = x) - \lim_{z \to y^-} \mathbb{P}(Y \leq z|X = x) \\ &= F_{Y|X}(y|x) - F_{Y|X}(y^-|x). \end{aligned}$$

□

4.3 Convolution of Discrete Random Variables

Often, we consider the sum of two variables, such as the number of heads in n tosses of a coin. However, more complex situations arise, especially when the summands are dependent. We aim to find a formula for describing the mass function of the sum $Z = X + Y$.

Theorem 4.14. Given two jointly discrete random variables X and Y, the probability of the sum of the two random variables is given by:

$$\mathbb{P}(X+Y=z) = \sum_x f_{X,Y}(x, z-x) = \sum_y f_{X,Y}(z-y, y), \qquad \text{for } z \in \mathbb{R}.$$

Proof.
We have the disjoint union:

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

At most countably many of its distributions have non-zero probability. Therefore:

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

□

Example 4.1. From Example 3.1, let X be the number of heads and Y be the number of tails when tossing a coin twice. The JPMF is given by:

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{4}, & (x,y) = (0,2),(1,1),(2,0), \\ 0, & \text{otherwise.} \end{cases}$$

Let $Z = X+Y$ be the total number of heads and tails. The PMF of Z is given by:

$$f_Z(z) = \sum_x f_{X,Y}(x, z-x) = \begin{cases} \frac{1}{4}, & z=2, \\ \frac{1}{2}, & z=1, \\ \frac{1}{4}, & z=0, \\ 0, & \text{otherwise.} \end{cases}$$

Definition 4.15. The **Convolution** f_{X+Y} $(f_X * f_Y)$ of the PMFs of two independent discrete random variables X and Y is the PMF of $X+Y$:

$$f_{X+Y}(z) = \mathbb{P}(X+Y=z) = \sum_x f_X(x) f_Y(z-x) = \sum_y f_X(z-y) f_Y(y), \qquad \text{for } z \in \mathbb{R}.$$

Remark 4.15.1. The convolution operation is commutative and associative. That is, for any independent discrete random variables X, Y, Z:

$$f_X * f_Y = f_Y * f_X, \qquad (f_X * f_Y) * f_Z = f_X * (f_Y * f_Z).$$

4.4 Examples of Discrete Random Variables

Here are some important examples of random variables that have a wide range of applications.

Definition 4.16. The **Parametric Distribution** of a discrete random variable is a distribution where the PMF depends on one or more parameters.

The following examples illustrate some of the most useful distributions.

Example 4.2. (**Constant Variables**) For a constant c, let X 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 4.3. (**Bernoulli Distribution**) $X \sim \mathrm{Bern}(p)$
Let $A \in \mathcal{F}$ be a specific event. A Bernoulli trial is considered a success if A occurs. Let X be such that:

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

Example 4.4. Let $A \in \mathcal{F}$ and **Indicator Functions** $\mathbf{1}_A : \Omega \to \mathbb{R}$ such that, for all $B \in \mathcal{B}(\mathbb{R})$:

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

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

Example 4.5. (**Binomial Distribution**) $Y \sim \mathrm{Bin}(n,p)$
Suppose we perform n independent Bernoulli trials $X_1, X_2, \ldots, X_n$. Let $Y = X_1 + X_2 + \cdots + X_n$ be the 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 $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ can be any sequence satisfying $\#\{i : \sigma_i = 1\} = k$, and $A_\sigma :=$ events that $(X_1, X_2, \ldots, X_n) = (\sigma_1, \sigma_2, \ldots, \sigma_n)$. The events A_σ are mutually exclusive. Hence,

$$\mathbb{P}(A) = \sum_\sigma \mathbb{P}(A_\sigma).$$

There are a total of $\binom{n}{k}$ different σ's in the sum. By independence, we have:

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

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

Example 4.6. (**Trinomial Distribution**) Suppose we perform n trials, each of which results in three outcomes A, B, and C, where A occurs with probability p, B with probability q, and C with probability $1 - p - q$. The probability of r A's, w B's, and $n - r - w$ C's is:

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

Remark 4.16.1. The multinomial distribution is a generalization of the binomial distribution. It describes the probabilities of counts for k different outcomes in n independent trials, where each outcome has a fixed probability.

Example 4.7. (**Geometric Distribution**) $W \sim \text{Geom}(p)$
Suppose we keep performing independent Bernoulli trials until the first success occurs. Let p be the probability of success, and let W be the **waiting time** that elapses before the 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 4.8. (Alternative Geometric Distribution) Suppose we keep performing independent Bernoulli trials until the first success occurs. Let p be the probability of success, and let W' be the number of failures before the first success.

$$\mathbb{P}(W' = k) = p(1-p)^k, \qquad \mathbb{E}W' = \frac{1-p}{p}, \qquad \text{Var}(W') = \frac{1-p}{p^2}.$$

Remark 4.16.2. Conventionally, when we refer to the geometric distribution, we usually mean the one related to waiting time rather than the number of failures.

Example 4.9. (**Negative Binomial Distribution**) $W_r \sim \text{NBin}(r, p)$
Similar to the examples of the 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 4.16.3. W_r is the sum of r independent geometric variables.

Example 4.10. (**Hypergeometric Distribution**) $X \sim \text{Hypergeometric}(N, m, n)$
Suppose that we have a set of N balls. There are m red balls and $N - m$ blue balls. We choose n of these balls without replacement and define X to be the number of red balls in our sample. Then:

$$\mathbb{P}(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}},$$

for $k = 0, 1, \ldots, \min(m, n)$.

Example 4.11. (**Poisson Distribution**) $X \sim \text{Poisson}(\lambda)$
A **Poisson variable** is a discrete random variable with the Poisson PMF:

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

for some parameter $\lambda > 0$.

Remark 4.16.4. This is used to approximate a 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 $\lambda = np$. For $k = 0, 1, \ldots, n$:

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

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

Theorem 4.17. If $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ are independent, then:

$$X + Y \sim \text{Poisson}(\lambda + \mu).$$

Proof.
For any $k \geq 0$:

$$\begin{aligned}
\mathbb{P}(X+Y=k) &= \sum_{i=0}^{k} \mathbb{P}(X=i, Y=k-i) = \sum_{i=0}^{k} \mathbb{P}(X=i)\mathbb{P}(Y=k-i) \\
&= \sum_{i=0}^{k} \left(\frac{\lambda^i}{i!}e^{-\lambda}\right)\left(\frac{\mu^{k-i}}{(k-i)!}e^{-\mu}\right) = \frac{1}{k!}e^{-(\lambda+\mu)} \sum_{i=0}^{k} \frac{k!}{i!(k-i)!}\lambda^i \mu^{k-i} \\
&= \frac{1}{k!}e^{-(\lambda+\mu)} \sum_{i=0}^{k} \binom{k}{i} \lambda^i \mu^{k-i} = \frac{(\lambda+\mu)^k}{k!}e^{-(\lambda+\mu)}.
\end{aligned}$$

Therefore, $X + Y \sim \text{Poisson}(\lambda + \mu)$. □

Lemma 4.18. If the number of occurrences of an event in unit time or space follows the Poisson distribution with rate λ, then the number of occurrences in t units of time or space follows Poisson(λt).

Proof.
Let $X_1, X_2, \ldots, X_t$ be independent Poisson variables with parameter λ. Then, by Theorem 4.17 and induction:

$$X_1 + X_2 + \cdots + X_t \sim \text{Poisson}(\lambda t).$$

Therefore, the number of occurrences in t units of time or space follows Poisson(λt). □

Remark 4.18.1. The Poisson distribution is often used to model the number of events occurring within a fixed interval of time or space, such as the number of phone calls received by a call center in an hour or the number of decay events from a radioactive source in a given time period.

We have an interesting example concerning independence involving the Poisson distribution.

Example 4.12. (**Poisson Flips**) A coin is tossed once, and heads turn up with probability p.
Let X and Y be the numbers of heads and tails, respectively. X and Y are not independent since:

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

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

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

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

Chapter 5

Continuous Random Variables

5.1 Introduction to Continuous Random Variables

In this chapter, we discuss continuous random variables. We begin with the definition of a continuous random variable.

Definition 5.1. A random variable X is **continuous** if its **distribution function** (CDF) F_X can be written 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 5.1.1. The PDF f_X is not uniquely prescribed since two integrable functions that take identical values except at some specific points have the same integral. However, if F_X is **differentiable** at u, we set $f_X(u) = F'_X(u)$.

Remark 5.1.2. More generally, for an interval B, we have:

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

Note that we use the same letter f for mass functions and density functions since both perform similar tasks.

Remark 5.1.3. For $x \in \mathbb{R}$, the numerical value $f_X(x)$ is not a probability. However, we can consider:

$$f_X(x)\, dx = \mathbb{P}(x < X \leq x + dx),$$

as an element of probability.

Lemma 5.2. 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.
$$\mathbb{P}(a \leq X \leq b) = F(b) - F(a) = \int_{-\infty}^{b} f_X(x)\, dx - \int_{-\infty}^{a} f_X(x)\, dx = \int_a^b f_X(x)\, dx.$$

□

Similar to the discrete case, there is a joint distribution function for two random variables.

Definition 5.3. The **joint distribution function** (JCDF) of two continuous random variables X and Y is the function $F : \mathbb{R}^2 \to [0, 1]$ such that:

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

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

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

Remark 5.3.1. More generally, for $B \in \mathcal{B}(\mathbb{R}^2)$:

$$\mathbb{P}((X, Y) \in B) = \iint_B f_{X,Y}(x, y)\, dx\, dy.$$

We also recall the definition of the marginal distribution function.

Definition 5.4. Given two continuous random variables X and Y, the **marginal probability density function** (Marginal PDF) of X is:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, u)\, du.$$

The definition of independence also applies to continuous random variables.

Definition 5.5. 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 5.6. Two continuous random variables X and Y are independent if and only if, for all $x, y \in \mathbb{R}$:

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

Proof.
If $f_{X,Y}(x, y) = f_X(x) f_Y(y)$, then:

$$F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v)\, du\, dv = \int_{-\infty}^{y} \int_{-\infty}^{x} f_X(u) f_Y(v)\, du\, dv = F_X(x) \int_{-\infty}^{y} f_Y(v)\, dv = F_X(x) F_Y(y).$$

If X and Y are independent, then:

$$\int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v)\, du\, dv = \int_{-\infty}^{y} \int_{-\infty}^{x} f_X(u) f_Y(v)\, du\, dv.$$

By the Fundamental Theorem of Calculus, we have:

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

□

Theorem 5.7. Consider two continuous and independent random variables X and Y. For any two functions g and h, if $g(X)$ and $h(Y)$ are still continuous random variables, then $g(X)$ and $h(Y)$ are independent.

5.2 Conditional Distribution of Continuous Random Variables

Recall the definition of the conditional distribution function of a 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 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 5.8. Suppose $X, Y : \Omega \to \mathbb{R}$ are two continuous random variables with PDF $f_X(x) > 0$ for some $x \in \mathbb{R}$. The **conditional distribution function** (Conditional CDF) of Y given $X = x$ for some $\mathbf{x} \in \mathbb{R}$ is defined by:

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

The **conditional probability density function** (Conditional PDF) of Y given $X = x$ for some $\mathbf{x} \in \mathbb{R}$ 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 5.8.1. For any $x, y \in \mathbb{R}$, since $f_X(x)$ can also be computed from $f_{X,Y}(x,y)$, 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 5.8.2. More generally, for two continuous random variables X and Y with PDF $f_X(x) > 0$ for some $x \in \mathbb{R}$:

$$\mathbb{P}(Y \in A | X = x) = \int_A \frac{f_{X,Y}(x,v)}{f_X(x)}\, dv = \int_A f_{Y|X}(y|x)\, dy, \qquad \text{for } A \in \mathcal{B}(\mathbb{R}).$$

Example 5.1. Assume that two jointly continuous random variables X and Y have a JPDF:

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

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

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

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

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

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

Example 5.2. We aim to find $\mathbb{P}(X^2+Y^2 \le 1)$, where X and Y are two jointly continuous random variables with the JPDF given in Example 5.1. Let $Y \in A_x = \{y : |y| \le \sqrt{1-x^2}\}$.

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

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

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

Theorem 5.9. If two jointly continuous random variables X and Y have a JPDF $f_{X,Y}$, then $X+Y$ has a PDF:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x)\,dx = \int_{-\infty}^{\infty} f_{X,Y}(z-y, y)\,dy, \qquad \text{for } z \in \mathbb{R}.$$

Proof.

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

□

Definition 5.10. Given two independent continuous random variables X and Y, the **convolution** f_{X+Y} $(f_X * f_Y)$ of the PDFs of X and Y is the PDF of $X+Y$:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\,dy = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx, \qquad \text{for } z \in \mathbb{R}.$$

5.3 Examples of Continuous Random Variables

Similar to discrete random variables, we have some useful parametric distributions.

Definition 5.11. The **parametric distribution** of a continuous random variable is a distribution where the PDF depends on one or more parameters.

Example 5.3. (**Uniform Distribution**) $X \sim \mathrm{U}[a, b]$
A random variable X is **uniform** on $[a, b]$ for $a < b$ if the CDF and PDF of X are:

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

Example 5.4. If $X \sim \mathrm{U}[0, 1]$ and $Y \sim \mathrm{U}[0, 1]$, and $X \perp\!\!\!\perp Y$, then for all $t \in \mathbb{R}$:

$$f_X(t) = f_Y(t) = \begin{cases} 1, & 0 \le t \le 1 \\ 0, & \text{Otherwise} \end{cases}.$$

Therefore, for all $z \in \mathbb{R}$:

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

Example 5.5. Assume that a plane is ruled by horizontal lines separated by D, and a needle of length $L \le D$ is cast randomly on the plane. What is the probability that the needle intersects some lines?
Let X be the distance from the center of the needle to the nearest line, and Θ be the acute angle between the needle and the vertical line. Assume that $X \perp\!\!\!\perp \Theta$. We have $X \sim \mathrm{U}\left[0, \frac{D}{2}\right]$ and $\Theta \sim \mathrm{U}\left[0, \frac{\pi}{2}\right]$.

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

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

Suppose that we throw the needle n times.

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

By measuring the number of intersections, we can estimate the value of π.

Example 5.6. (**Inverse Transform Sampling**) Let $U \sim \mathrm{U}[0, 1]$. For a continuous random variable X with CDF F_X, we want to find a function $g : \mathbb{R} \to \mathbb{R}$ such that $g(U) \sim X$. Since F_X is non-decreasing, we can define the **generalized inverse** of F_X as:

$$F_X^{-1}(y) = \inf\{x : F_X(x) \ge y\}, \qquad \text{for } y \in [0, 1].$$

We set $g = F_X^{-1}$. For $x \in \mathbb{R}$:

$$\mathbb{P}(F_X^{-1}(U) \le x) = \mathbb{P}(U \le F_X(x)) = F_X(x).$$

Therefore, $F_X^{-1}(U) \sim X$.

Example 5.7. (**Exponential Distribution**) $X \sim \mathrm{Exp}(\lambda)$
A random variable X is **exponentially distributed** with parameter $\lambda > 0$ if the CDF and PDF of X are:

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

Theorem 5.12. The exponential distribution has the memoryless property. This means that for all $s > 0$ and $t > 0$:

$$\mathbb{P}(X > s + t | X > s) = \mathbb{P}(X > t).$$

Proof.
Assume that $X \sim \mathrm{Exp}(\lambda)$.

$$\mathbb{P}(X > s + t | X > s) = \frac{\mathbb{P}(\{X > s + t\} \cap \{X > s\})}{\mathbb{P}(X > s)} = \frac{\mathbb{P}(X > s + t)}{\mathbb{P}(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbb{P}(X > t).$$

□

Example 5.8. (**Normal Distribution / Gaussian Distribution**) $X \sim \mathrm{N}(\mu, \sigma^2)$
A random variable X is **normally distributed** if it has two parameters μ and σ^2, and its PDF and CDF are:

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

This distribution is one of the most important distributions.
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 5.12.1. $\phi(x)$ is a probability density 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 $\theta \in [0, 2\pi]$:

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

Since $\phi(x) > 0$, $I = 1$. Therefore, $\phi(x)$ is a probability density function. □

These are some properties that are frequently used.

Lemma 5.13. The normal distribution has the following properties:

1. Let $X \sim \mathrm{N}(0,1)$. If $Y = bX + a$ 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 $Y = \frac{X-a}{b}$, then $Y \sim \mathrm{N}(0,1)$.

Proof.

1. Let $z = bx + a$.

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

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

2. Let $x = bz + a$.

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

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

□

Lemma 5.14. If $X \sim \mathrm{N}(\mu, \sigma^2)$, then for all $s \leq t$:

$$\mathbb{P}(s \leq X \leq t) = \mathbb{P}\left(\frac{s-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{t-\mu}{\sigma}\right) = \Phi\left(\frac{t-\mu}{\sigma}\right) - \Phi\left(\frac{s-\mu}{\sigma}\right).$$

Proof.
Since $\sigma > 0$, $\frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1)$ by Lemma 5.13. Therefore:

$$\begin{aligned}\mathbb{P}(s \leq X \leq t) &= \int_s^t \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \int_{\frac{s-\mu}{\sigma}}^{\frac{t-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz = \Phi\left(\frac{t-\mu}{\sigma}\right) - \Phi\left(\frac{s-\mu}{\sigma}\right). \qquad (z = \tfrac{x-\mu}{\sigma})\end{aligned}$$

□

This is a very important theorem, as it states that the sum of normal distributions is still normal.

Theorem 5.15. If $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots, n$ and they are independent, then:

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

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

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

Therefore, $X + Y \sim \mathrm{N}(0, 1+\sigma^2)$. In the general case where $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$, 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 $\frac{X_1-\mu_1}{\sigma_2} \sim \mathrm{N}\left(0, \frac{\sigma_1^2}{\sigma_2^2}\right)$. Applying this to the special case, we find:

$$\frac{X_1-\mu_1}{\sigma_2} + \frac{X_2-\mu_2}{\sigma_2} \sim \mathrm{N}\left(0, 1 + \frac{\sigma_1^2}{\sigma_2^2}\right).$$

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 $i = 1, 2, \ldots, n$ and they are independent, then:

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

□

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

Example 5.9. (**Standard Bivariate Normal Distribution**) Two continuous random variables X and Y are **standard bivariate normal** if they have the 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 5.15.1. 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 5.15.2. ρ is called the **population correlation coefficient** between X and Y. It will be discussed in Chapter 6.

In most cases, normal random variables X and Y do not have a mean of 0 and a variance of 1. If we also include the mean and variance in the distribution, we obtain the following distribution.

Example 5.10. (**Bivariate Normal Distribution**) Two continuous random variables X and Y are **bivariate normal** with means μ_X and μ_Y, variances σ_X^2 and σ_Y^2, and correlation coefficient ρ if the JPDF is given by:

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

Example 5.11. 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 $f_{X|Y}(x|y)$.

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

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:

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

This means that we can generate bivariate normal random variables X and Y from two independent standard normal random variables Y and Z. More generally, for any orthogonal matrix $\mathbf{A}$ (i.e. $\mathbf{A}^T\mathbf{A} = \mathbf{I}$), if we let:

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

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

There are some remarks that may be important to know.

Remark 5.15.3. X and Y are bivariate normal and uncorrelated if and only if X and Y are independent normal. We will discuss what uncorrelatedness means.

Remark 5.15.4. X and Y being jointly continuous and both normal does not imply that they are bivariate normal.

Example 5.12. Consider the 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 bivariate normal distribution.
However, if you look at their marginal PDFs:

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

This is the same as $f_Y(x)$.
Therefore, X and Y being jointly continuous and both normal does not imply that they are bivariate normal.

Remark 5.15.5. Two random variables X and Y being jointly continuous and uncorrelated Gaussian does not imply that they are independent Gaussian.

More generally, we can create a multivariate normal distribution from more than two random variables.

Example 5.13. (**Multivariate Normal Distribution**) $\mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
A random vector $\mathbf{X}$ with dimension p is p-dimensional normal with $p \times 1$ mean vector $\boldsymbol{\mu}$ and $p \times p$ variance-covariance matrix $\boldsymbol{\Sigma}$ if:

$$f(\mathbf{x}) = (2\pi)^{-\frac{p}{2}}\,|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\,e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}.$$

Remark 5.15.6. The elements a_{ij} in the i-th row and j-th column of the variance-covariance matrix $\boldsymbol{\Sigma}$ are obtained by:

$$a_{ij} = \mathrm{cov}(X_i, X_j).$$

We will discuss how to calculate the covariance in the next chapter.

Example 5.14. (**Cauchy Distribution**) $X \sim \mathrm{Cauchy}(\theta)$
A random variable X has a Cauchy distribution if it has the PDF:

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

Remark 5.15.7. If $X \sim \mathrm{N}(0,1)$ and $Y \sim \mathrm{N}(0,1)$, then $\frac{X}{Y} \sim \mathrm{Cauchy}(0)$.

Example 5.15. (**Gamma Distribution**) $X \sim \text{Gamma}(\alpha, \lambda)$
A random variable X has a gamma distribution with parameters α and λ if it has the PDF:

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)}\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}, & x \geq 0 \\ 0, & x < 0 \end{cases} = \frac{1}{\Gamma(\alpha)}\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}\mathbf{1}_{x\geq 0},$$

where $\Gamma(\alpha)$ is called the **gamma function**, defined by:

$$\Gamma(\alpha) = \int_0^\infty e^{-y}y^{\alpha-1}\,dy.$$

Note that $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$. If α is a positive integer, $\Gamma(\alpha) = (\alpha-1)!$.

Lemma 5.16. When $\alpha = 1$, the gamma distribution becomes an exponential distribution.

Proof.
Let $X \sim \text{Gamma}(1, \lambda)$. The PDF is:

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

□

Example 5.16. (**Chi-Squared Distribution**) $Y \sim \chi^2(n)$
Assume that $X_1, X_2, \ldots, X_n$ are independent standard normal random variables. Let $Y = \sum_{i=1}^n X_i^2$. We say Y has a χ^2-distribution with parameter n if it has the PDF:

$$f_Y(x) = \begin{cases} \frac{1}{\Gamma(\frac{n}{2})}2^{-\frac{n}{2}}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, & x \geq 0 \\ 0, & x < 0 \end{cases} = \frac{1}{\Gamma(\frac{n}{2})}2^{-\frac{n}{2}}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}\mathbf{1}_{x\geq 0}.$$

Remark 5.16.1. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Lemma 5.17. A random variable $\chi^2(n)$ is equivalent to $\text{Gamma}(\frac{n}{2}, \frac{1}{2})$.

Proof.
Let $X \sim \text{Gamma}(\frac{n}{2}, \frac{1}{2})$. Substituting $\alpha = \frac{n}{2}$ and $\lambda = \frac{1}{2}$ into the PDF of X, we have:

$$f_X(x) = \begin{cases} \frac{1}{2\Gamma(\frac{n}{2})}e^{-\frac{x}{2}}\left(\frac{x}{2}\right)^{\frac{n}{2}-1}, & x \geq 0 \\ 0, & x < 0 \end{cases} = \begin{cases} \frac{1}{\Gamma(\frac{n}{2})}2^{-\frac{n}{2}}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, & x \geq 0 \\ 0, & x < 0 \end{cases}.$$

Therefore, $X \sim \chi^2(n)$. □

Lemma 5.18. Given that $V \sim \chi^2(n_1)$ and $W \sim \chi^2(n_2)$, if V and W are independent, then $V+W \sim \chi^2(n_1+n_2)$.

Proof.
Let $V = X_1^2 + X_2^2 + \cdots + X_{n_1}^2$ and $W = Z_1^2 + Z_2^2 + \cdots + Z_{n_2}^2$, where $X_i, Z_j \sim \text{N}(0,1)$ for all i, j.

$$V + W = X_1^2 + X_2^2 + \cdots + X_{n_1}^2 + Z_1^2 + Z_2^2 + \cdots + Z_{n_2}^2 \sim \chi^2(n_1+n_2),$$

□

We can derive further distributions from the chi-squared distribution.

Example 5.17. (**Student's t-Distribution**) $W \sim t(n)$
Given $Y \sim \chi^2(n)$ and $Z \sim \mathrm{N}(0,1)$, if Y and Z are independent, let:

$$W = \frac{Z}{\sqrt{\frac{Y}{n}}}.$$

The random variable W follows the t-distribution with n degrees of freedom and has the PDF:

$$f(w) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{w^2}{n}\right)^{-\frac{n+1}{2}}.$$

Remark 5.18.1. If $W \sim t(1)$, then from the PDF:

$$f(w) = \frac{\Gamma(1)}{\sqrt{\pi}\Gamma\left(\frac{1}{2}\right)}(1+w^2)^{-1} = \frac{1}{\pi(1+w^2)}.$$

This means that $W \sim \mathrm{Cauchy}(0)$.

Remark 5.18.2. Fixing $Y = y$ for some constant $y \neq 0$, we can easily find that $W \sim \mathrm{N}(0, \frac{n}{y})$.

Example 5.18. (**Beta Distribution**) $X \sim \mathrm{Beta}(a,b)$
A random variable X has a beta distribution with parameters a and b if it has the PDF:

$$f_X(x) = \begin{cases} \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}, & 0<x<1 \\ 0, & \text{Otherwise} \end{cases} = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}\mathbf{1}_{0<x<1},$$

where $B(a,b)$ is called the **beta function**, defined as:

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Example 5.19. (**F-Distribution**) $F \sim F(r_1, r_2)$
Assume that X and Y are independent random variables with $X \sim \chi^2(r_1)$ and $Y \sim \chi^2(r_2)$. Let:

$$F = \frac{\frac{X}{r_1}}{\frac{Y}{r_2}}.$$

Then F has an F-distribution with r_1 and r_2 degrees of freedom, and its PDF is:

$$f_F(w) = \frac{\Gamma\left(\frac{r_1+r_2}{2}\right)}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)}\left(\frac{r_1}{r_2}\right)^{\frac{r_1}{2}} w^{\frac{r_1}{2}-1}\left(1+\frac{r_1 w}{r_2}\right)^{-\frac{r_1+r_2}{2}},$$

where $0 < w < \infty$.

Lemma 5.19. If $U \sim F(r_1, r_2)$, then $\frac{1}{U} \sim F(r_2, r_1)$.

Proof.
By definition:

$$U = \frac{\frac{X}{r_1}}{\frac{Y}{r_2}},$$

where $X \sim \chi^2(r_1)$ and $Y \sim \chi^2(r_2)$. Therefore:

$$\frac{1}{U} = \frac{\frac{Y}{r_2}}{\frac{X}{r_1}} \sim F(r_2, r_1).$$

□

5.4 Functions of Continuous Random Variables

Given a continuous random variable X and a function g such that $g(X)$ is also 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 determine the distribution of $g(X)$.

Example 5.20. Assume that X is a 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 first work with $F_Y(y)$. 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 5.21. 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 5.20. In the case where $g(x)$ 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:

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

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

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

□

We can consider the multivariable case.

Example 5.22. 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 $f_{U,V}(u,v)$? To simplify the process, we need to make the following assumptions:

1. X and Y can be uniquely solved from U and V. (There exists only one pair of functions a and b such that $X = a(U,V)$ and $Y = b(U,V)$.)

2. The functions g and h are differentiable, and the Jacobian determinant:

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

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 5.23. Given two jointly continuous random variables X_1 and X_2 with 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 $X_1 \perp\!\!\!\perp X_2$:

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

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 5.24. We do the same thing as the previous example, but instead, we have two independent random variables $X_1 \sim \mathrm{U}[0, 1]$ and $X_2 \sim \mathrm{U}[0, 1]$. For all $x_1, x_2 \in \mathbb{R}$:

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

Chapter 6

Expectation

In this chapter, if we are discussing more than one random variable, we assume that either all are discrete or all are continuous. In a practical sense, it is possible to have some discrete and others continuous, but we will not tackle this situation. It is relatively easy to prove all the theorems and lemmas in this case once you know how to prove them when all are discrete or all are continuous.

6.1 Introduction to Expectation

In real life, we often want to know the expected final result given the probabilities we calculated. The result is usually a theoretical approximation of the empirical average. Assume we have random variables $X_1, X_2, \ldots, X_N$ which take values in $\{x_1, x_2, \ldots, x_n\}$ with probability mass function $f_X(x)$. We get an empirical average:

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

This gives us the formula for the expectation of a discrete random variable. However, for continuous random variables, the probability at every single point is 0. To make sense of this, we use the probability density function to obtain the expectation:

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx.$$

Definition 6.1. The **mean value**, **expectation**, or **expected value** of a discrete random variable X with PMF f_X is defined as:

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

whenever this sum is absolutely convergent.
The **expectation** of a continuous random variable X with PDF f_X is defined as:

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

whenever this integral exists.

Remark 6.1.1. Due to absolute convergence, we can usually define $\mathbb{E}X$ only if $\mathbb{E}\,|X|$ exists.

Example 6.1. Suppose a product is sold seasonally. Let b be the net profit for each sold unit, ℓ be the net loss for each unsold unit, and X be the number of products ordered by customers. If y units are stocked, what is the expected profit $Q(y)$?

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

Theorem 6.2. Given two random variables X and Y, the expectation operator $\mathbb{E}$ has the following properties:

1. For any $a \leq b$, if $a \leq X \leq b$, then $a \leq \mathbb{E}X \leq b$.
2. If $X \geq 0$, then $\mathbb{E}X \geq 0$.
3. If $a, b \in \mathbb{R}$, then $\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$.

Proof.

1. If X is discrete, then:

$$\mathbb{E}X = \sum_x x f_X(x) \geq \sum_x a f_X(x) = a, \qquad \mathbb{E}X = \sum_x x f_X(x) \leq \sum_x b f_X(x) = b.$$

If X is continuous, then:

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x) \geq \int_{-\infty}^{\infty} a f_X(x) = a, \qquad \mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x) \leq \int_{-\infty}^{\infty} b f_X(x) = b.$$

Therefore, we have $a \leq \mathbb{E}X \leq b$.

2. If X is discrete, since $f_X(x) \geq 0$ for all $x \geq 0$, $\mathbb{E}X = \sum_x x f_X(x) \geq 0$ if $X \geq 0$.
If X is continuous, since $f_X(x) \geq 0$ for all $x \geq 0$, $\mathbb{E}X = \int_0^\infty x f_X(x)\, dx \geq 0$.

3. When X and Y are discrete:

$$\begin{aligned}
\mathbb{E}(aX + bY) &= \sum_{x,y} (ax + by) f_{X,Y}(x, y) \\
&= a \sum_x x \left(\sum_y f_{X,Y}(x, y) \right) + b \sum_y y \left(\sum_x f_{X,Y}(x, y) \right) \\
&= a \sum_x x f_X(x) + b \sum_y y f_Y(y) = a\mathbb{E}X + b\mathbb{E}Y.
\end{aligned}$$

When X and Y are continuous:

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

□

Using induction, we can immediately obtain the following result.

Lemma 6.3. (**Linearity of Expectation**) More generally, for any sequence of random variables $\{X_1, \ldots, X_n\}$, we have:

$$\mathbb{E}\left(\sum_{i=1}^n a_i X_i \right) = \sum_{i=1}^n a_i \mathbb{E}X_i.$$

Theorem 6.4. (**Tail Sum Formula**) If a discrete random variable X has a PMF f_X that satisfies $f_X(x) = 0$ when $x < 0$, then:

$$\mathbb{E}X = \sum_{k=0}^{\infty} \mathbb{P}(X > k).$$

If a continuous random variable X has a PDF f_X that satisfies $f_X(x) = 0$ when $x < 0$, then:

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

Proof.
For a discrete random variable X with $f_X(x)$ for any $x < 0$:

$$\begin{aligned}\sum_{k=0}^{\infty} \mathbb{P}(X > k) &= \sum_{k=1}^{\infty} \mathbb{P}(X \geq k) \\ &= \sum_{k=1}^{\infty} \sum_{i=k}^{\infty} \mathbb{P}(X = i) \\ &= \sum_{k=1}^{\infty} k\mathbb{P}(X = k) = \mathbb{E}X.\end{aligned}$$

For a continuous random variable X with $f_X(x)$ for any $x < 0$:

$$\begin{aligned}\int_0^{\infty} \mathbb{P}(X > x)\, dx &= \int_0^{\infty} \int_x^{\infty} f_X(y)\, dy\, dx \\ &= \int_0^{\infty} \int_0^{y} f_X(y)\, dx\, dy \\ &= \int_0^{\infty} y f_X(y)\, dy = \mathbb{E}X.\end{aligned}$$

□

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

Lemma 6.5. If a 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.

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

□

Theorem 6.6. Given a function $g : \mathbb{R} \to \mathbb{R}$ and a random variable X:

1. If X is discrete with a PMF $f_X(x)$, and $g(X)$ is still a discrete random variable, then:

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

whenever this sum is absolutely convergent.

2. If X is continuous with a PDF $f_X(x)$, and $g(X)$ is still a continuous random variable, then:

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

whenever this integral exists.

Proof.

1. Let $Y = g(X)$. We have:

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

2. Consider that $g(x) \geq 0$ for all x. Let $B = \{x : g(x) > y\}$. By Lemma 6.4:

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

Now consider that $g(x) \leq 0$ for all x. Let $C = \{x : g(x) < z\}$. By Lemma 6.5:

$$\begin{aligned}
\mathbb{E}g(X) &= \int_{-\infty}^{0} -F_{g(X)}(z)\, dz \\
&= \int_{-\infty}^{0} \int_C -f_X(x)\, dx\, dz \\
&= \int_{-\infty}^{0} \int_{g(x)}^{0} -f_X(x)\, dz\, dx \\
&= \int_{-\infty}^{0} g(x) f_X(x)\, dx.
\end{aligned}$$

Combining both cases, if $g(X)$ is a random variable, then:

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

□

Theorem 6.7. Given a function $g : \mathbb{R}^2 \to \mathbb{R}$ and two random variables X and Y:

1. If X and Y are jointly discrete with JPMF $f_{X,Y}(x,y)$, and $g(X,Y)$ is a discrete random variable, then:

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

2. If X and Y are jointly continuous with JPDF $f_{X,Y}(x,y)$, and $g(X,Y)$ is a continuous random variable, then:

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

Proof.

1. Let $Z = g(X,Y)$. We have:

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

2. *We will not prove this case. Just note that it works similarly to the previous theorem.*

□

Remark 6.7.1. We may generalize this to a random vector.

We have some special terms for specific expectations.

Definition 6.8. Let k be a positive integer. We have special terms for the following expectations:

1. The **k-th moment** m_k of X is defined as:

$$m_k = \mathbb{E}(X^k).$$

2. The **k-th central moment** α_k is defined as:

$$\alpha_k = \mathbb{E}(X - \mathbb{E}X)^k = \mathbb{E}(X - m_1)^k.$$

Remark 6.8.1. Not all random variables have k-th moments for all positive integers k.

Remark 6.8.2. We cannot use a finite number of moments to uniquely determine a distribution with k-th moments for all positive integers k.

Example 6.2. Let $X \sim \mathrm{N}(\mu, \sigma^2)$. We have:

$$m_1 = \mathbb{E}X = \mu, \quad m_2 = \mathbb{E}(X^2) = \sigma^2 + \mu^2, \quad m_3 = \mathbb{E}(X^3) = 3\mu\sigma^2 + \mu^3, \quad m_4 = \mathbb{E}(X^4) = 3\sigma^4 + 6\mu^2\sigma^2 + \mu^4.$$

Example 6.3. Let $Y \sim \mathrm{Exp}(\lambda)$. We have:

$$m_1 = \mathbb{E}Y = \frac{1}{\lambda}, \qquad m_2 = \mathbb{E}(Y^2) = \frac{2}{\lambda^2}, \qquad m_3 = \mathbb{E}(Y^3) = \frac{6}{\lambda^3}, \qquad m_4 = \mathbb{E}(Y^4) = \frac{24}{\lambda^4}.$$

Definition 6.9. Given a random variable X:

1. The **mean** of X is the 1st moment, denoted by μ, defined as:
$$\mu = \mathbb{E}X.$$
2. The **variance** of X is the 2nd central moment, denoted by $\mathrm{Var}(X)$, defined as:
$$\mathrm{Var}(X) = \mathbb{E}(X-\mu)^2 = \mathbb{E}(X^2) - \mu^2.$$
3. The **standard deviation** of X, denoted by σ, is defined as:
$$\sigma = \sqrt{\mathrm{Var}(X)}.$$

Lemma 6.10. If two random variables X and Y are independent, then:
$$\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y.$$

Proof.
If X and Y are both discrete:
$$\begin{aligned}\mathbb{E}(XY) &= \sum_{x,y} xyf_{X,Y}(x,y)\\ &= \sum_{x,y} xyf_X(x)f_Y(y)\\ &= \sum_x xf_X(x)\sum_y yf_Y(y)\\ &= \mathbb{E}X\mathbb{E}Y.\end{aligned}$$

If X and Y are both continuous:
$$\begin{aligned}\mathbb{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x,y)\,dy\,dx\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)\,dy\,dx\\ &= \int_{-\infty}^{\infty} xf_X(x)\,dx\int_{-\infty}^{\infty} yf_Y(y)\,dy\\ &= \mathbb{E}X\mathbb{E}Y.\end{aligned}$$

□

Remark 6.10.1. The converse is not generally true.

We may generalize this to a function of X and Y. This is very important, as it implies that two resultant random variables, $g(X)$ and $h(Y)$, are "uncorrelated" as long as the two random variables from the domain, X and Y, are independent.

Theorem 6.11. Given two random variables X and Y and two functions $g, h : \mathbb{R} \to \mathbb{R}$ such that $g(X)$ and $h(Y)$ are still random variables. Let X and Y be independent. If $\mathbb{E}(g(X)h(Y))$, $\mathbb{E}g(X)$, and $\mathbb{E}h(Y)$ exist, then:

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}g(X)\mathbb{E}h(Y).$$

Proof.
If X and Y are both discrete:

$$\begin{aligned}
\mathbb{E}(g(X)h(Y)) &= \sum_{x,y} g(x)h(y)f_{X,Y}(x,y) \\
&= \sum_{x,y} g(x)h(y)f_X(x)f_Y(y) \\
&= \sum_{x} g(x)f_X(x) \sum_{y} h(y)f_Y(y) \\
&= \mathbb{E}g(X)\mathbb{E}h(Y).
\end{aligned}$$

If X and Y are both continuous:

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

□

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

Theorem 6.12. For random variables X and Y:

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

Proof.

1. Using the linearity of $\mathbb{E}$:

$$\begin{aligned}
\operatorname{Var}(aX + b) &= \mathbb{E}((aX + b - \mathbb{E}(aX + b))^2) \\
&= \mathbb{E}(a^2(X - \mathbb{E}X)^2) \\
&= a^2\mathbb{E}((X - \mathbb{E}X)^2) \\
&= a^2 \operatorname{Var}(X).
\end{aligned}$$

2. When X and Y are uncorrelated:

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

□

6.2 Conditional Expectation

Sometimes, it is not practical to find $\mathbb{E}X$ itself. What if we want to find the expectation of X given that another result occurred? Similar to conditional probability, we may also have conditional expectation.

Definition 6.13. Given two random variables X and Y:

1. If X and Y are discrete, then the **conditional expectation** ψ of Y given $X = x$ for any x is defined as:

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

2. If X and Y are continuous, then the **conditional expectation** ψ of Y given $X = x$ for any x is defined as:

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

The **conditional expectation** ψ of Y given X is defined as:

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

Example 6.4. Assume we roll a fair die.

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

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 $X = x$, 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 6.5. If $Y = X$, then $\psi(X) = \mathbb{E}(Y|X) = \mathbb{E}(X|X) = X$.

Example 6.6. If X and Y are independent, then $\psi(X) = \mathbb{E}(Y|X) = \mathbb{E}(Y)$.

Lemma 6.14. Given two random variables X and Y, the following properties hold:

1. $\mathbb{E}(aY + bZ|X) = a\mathbb{E}(Y|X) + b\mathbb{E}(Z|X)$.
2. If $Y \geq 0$, then $\mathbb{E}(Y|X) \geq 0$.
3. If X and Y are independent, then $\mathbb{E}(Y|X) = \mathbb{E}(Y)$.

Proof.

1. If X, Y, and Z are discrete, then for all x:

$$\begin{aligned}\mathbb{E}(aY + bZ|X = x) &= \sum_{y,z}(ay + bz)\mathbb{P}(Y = y, Z = z|X = x) \\ &= a\sum_{y,z} y\mathbb{P}(Y = y, Z = z|X = x) + b\sum_{y,z} z\mathbb{P}(Y = y, Z = z|X = x) \\ &= a\sum_{y} yf_{Y|X}(y|x) + b\sum_{z} zf_{Z|X}(z|x) \\ &= a\mathbb{E}(Y|X = x) + b\mathbb{E}(Z|X = x).\end{aligned}$$

If X, Y, and Z are continuous, then for all x:

$$\begin{aligned}\mathbb{E}(aY + bZ|X = x) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(ay + bz)\frac{f_{Y,Z,X}(y, z, x)}{f_X(x)}\,dy\,dz \\ &= a\int_{-\infty}^{\infty} y\int_{-\infty}^{\infty}\frac{f_{Y,Z,X}(y, z, x)}{f_X(x)}\,dz\,dy + b\int_{-\infty}^{\infty} z\int_{-\infty}^{\infty}\frac{f_{Y,Z,X}(y, z, x)}{f_X(x)}\,dy\,dz \\ &= a\int_{-\infty}^{\infty} y\frac{f_{Y,X}(y, x)}{f_X(x)}\,dy + b\int_{-\infty}^{\infty} z\frac{f_{Z,X}(z, x)}{f_X(x)}\,dz \\ &= a\mathbb{E}(Y|X = x) + b\mathbb{E}(Z|X = x).\end{aligned}$$

Therefore, $\mathbb{E}(aY + bZ|X) = a\mathbb{E}(Y|X) + b\mathbb{E}(Z|X)$.

2. If X and Y are discrete, then for all x:

$$\mathbb{E}(Y|X = x) = \sum_{y} yf_{Y|X}(y|x) \geq 0.$$

If X and Y are continuous, then for all x:

$$\mathbb{E}(Y|X = x) = \int_{-\infty}^{\infty} yf_{Y|X}(y|x)\,dy \geq 0.$$

Therefore, $\mathbb{E}(Y|X) \geq 0$ if $Y \geq 0$.

3. If X and Y are discrete, then for all x:

$$\mathbb{E}(Y|X = x) = \sum_{y} yf_{Y|X}(y|x) = \sum_{y} yf_Y(y) = \mathbb{E}Y.$$

If X and Y are continuous, then for all x:

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

Therefore, if X and Y are independent, then $\mathbb{E}(Y|X) = \mathbb{E}Y$.

□

In fact, we can extend the definition of conditional expectation to σ-fields.

Definition 6.15. 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 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 6.15.1. Under this definition:

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

Theorem 6.16. (**Law of Total Expectation**) Given two random variables X and Y, the conditional expectation $\psi(X) = \mathbb{E}(Y|X)$ satisfies:

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

Proof.
We can apply Theorem 6.6. If X and Y are discrete:

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

If X and Y are continuous:

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

The proof is similar if one of them is discrete and the other is continuous. □

Example 6.7. A miner is trapped in a mine with doors, each leading to a tunnel. Tunnel 1 will help the miner reach safety after 3 hours. However, tunnels 2 and 3 will send the miner back after 5 and 7 hours, respectively. What is the expected amount of time the miner needs to reach safety? (Assume that the miner is memoryless.) Let X be the amount of time to reach safety, and Y be the door number he chooses first.

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

$\mathbb{E}X = 15.$

Example 6.8. Continuing from the previous example, what is the expected amount of time the miner needs to reach safety if he chooses door 2 first? Let $\widetilde{X}$ be the amount of time to reach safety if the miner chooses door 2 first.

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

Example 6.9. A store has N customers in a day, where N is a random variable with $\mathbb{E}N < \infty$. Each customer spends an amount X_i, where X_i's are i.i.d. random variables with $\mathbb{E}X < \infty$. 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 a generalization of the Law of Total Expectation.

Theorem 6.17. Given two random variables X and Y, the 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.
We can apply Theorem 6.6. If X and Y are discrete,

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

If X and Y are continuous,

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

□

If there is conditional expectation, there is also conditional variance.

Definition 6.18. Given two random variables X and Y, the **conditional variance** is defined as:

$$\mathrm{Var}(Y|X) = \mathbb{E}((Y - \mathbb{E}(Y|X))^2|X)$$

We can obtain the variance of a random variable based on conditional variance.

Theorem 6.19. (**Law of Total Variance**) Given two random variables X and Y, we have:

$$\mathrm{Var}(Y) = \mathbb{E}(\mathrm{Var}(Y|X)) + \mathrm{Var}(\mathbb{E}(Y|X))$$

Proof.
Using Theorem 6.16 and Theorem 6.17,

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

□

Sometimes, we want to find the tendency in the linear relationship between two random variables. This is called the covariance of two random variables.

Definition 6.20. The **covariance** of two random variables X and Y is:

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

Remark 6.20.1. The magnitude of covariance is the geometric mean of the variances of two random variables.

Remark 6.20.2. The sign represents the linear relationship between the two random variables.

1. If the sign is positive, the two random variables show similar behavior.
2. If the sign is negative, the two random variables show opposite behavior.

Lemma 6.21. For any random variables X, Y, and Z, we have:

1. $\mathrm{Var}(X) = \mathrm{cov}(X, X)$.
2. $\mathrm{cov}(X, Y) = \mathrm{cov}(Y, X)$.
3. $\mathrm{cov}(X, Y + Z) = \mathrm{cov}(X, Y) + \mathrm{cov}(X, Z)$.
4. If X and Y are uncorrelated, then $\mathrm{cov}(X, Y) = 0$.

Proof.

1. $$\mathrm{cov}(X, X) = \mathbb{E}((X - \mathbb{E}X)(X - \mathbb{E}X)) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathrm{Var}(X)$$

2. $$\mathrm{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y = \mathbb{E}(YX) - \mathbb{E}Y\mathbb{E}X = \mathrm{cov}(Y, X)$$

3. $$\mathrm{cov}(X, Y+Z) = \mathbb{E}(X(Y+Z)) - \mathbb{E}X\mathbb{E}(Y+Z) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y + \mathbb{E}(XZ) - \mathbb{E}X\mathbb{E}Z = \mathrm{cov}(X, Y) + \mathrm{cov}(X, Z)$$

4. If X and Y are independent, then

$$\mathrm{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y = \mathbb{E}X\mathbb{E}Y - \mathbb{E}X\mathbb{E}Y = 0$$

□

Remark 6.21.1. In general, for any random variables $X_1, X_2, \cdots, X_n$,

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

Example 6.10. If X_i are independent and $\operatorname{Var}(X_i) = 1$ for all i, then:

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

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

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

We usually only care about the normalized covariance, which is called the correlation coefficient.

Definition 6.22. The **population correlation coefficient** between two random variables X and Y, denoted by ρ, is given by:

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

We can determine the relationship between X and Y based on their correlation coefficient.

1. If $\rho > 0$, then X and Y are **positively correlated**.
2. If $\rho < 0$, then X and Y are **negatively correlated**.
3. If $\rho = 0$, then X and Y are **uncorrelated**.

Remark 6.22.1. The population correlation coefficient ρ of random variables X and Y satisfies $-1 < \rho < 1$.

Remark 6.22.2. If ρ is near 1 or near -1, it indicates a strong linear relationship between X and Y.

Remark 6.22.3. The constant ρ used in the bivariate normal distribution is the population correlation coefficient.

Example 6.11. If $X \sim \mathrm{N}(0, 1)$ and $Y \sim \mathrm{N}(0, 1)$,

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

Lemma 6.23. Two random variables are uncorrelated if $\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y$.

Proof.
Based on the definition of the correlation coefficient, $\rho = 0$ if and only if $\operatorname{cov}(X, Y) = 0$. Therefore,

$$\operatorname{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y = 0 \iff \mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y$$

□

Remark 6.23.1. If X and Y are independent, then they are uncorrelated. The converse is generally not true.

Example 6.12. Let X be such that $f_X(0) = f_X(1) = f_X(-1) = \frac{1}{3}$ and $Y = \mathbf{1}_{X=0}$.

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

However,

$$\mathbb{P}(X = 0, Y = 0) = 0 \qquad \mathbb{P}(X = 0) = \frac{1}{3} \qquad \mathbb{P}(Y = 0) = \frac{2}{3} \qquad \mathbb{P}(X = 0)\mathbb{P}(Y = 0) = \frac{2}{9} \neq 0$$

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

When would the converse be true? It turns out it is true when X and Y are uncorrelated and bivariate normal.

Theorem 6.24. Random variables $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ are bivariate normal and uncorrelated if and only if X and Y are independent normal.

Proof.
Since X and Y are uncorrelated, $\mathrm{cov}(X, Y) = 0$ and thus the population correlation coefficient $\rho = 0$. Therefore, we have:

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

Therefore, X and Y are independent if X and Y are uncorrelated bivariate normal.
If X and Y are independent normal, then we have:

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

Therefore, X and Y are both uncorrelated and bivariate normal with $\rho = 0$. □

6.3 Expectation and Variance of Distributions

In this section, we will primarily focus on finding the expectation and variance of distributions we have discussed.

Theorem 6.25. Given a discrete variable X:

1. If $X \sim \text{Bern}(p)$, then:
$$\mathbb{E}X = p, \qquad \text{Var}(X) = p(1-p).$$
2. If $X \sim \text{Bin}(n,p)$, then:
$$\mathbb{E}X = np, \qquad \text{Var}(X) = np(1-p).$$

Proof.
If $X \sim \text{Bern}(p)$:

$$\mathbb{E}X = 0(1-p) + 1(p) = p, \qquad \text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = p - p^2 = p(1-p).$$

If $X \sim \text{Bin}(n,p)$, then by definition, $X = Y_1 + \cdots + Y_n$ where $Y_i \sim \text{Bern}(p)$ and are independent. Therefore:

$$\mathbb{E}X = \sum_{i=1}^{n} \mathbb{E}Y_i = np, \qquad \text{Var}(X) = \sum_{i=1}^{n} \text{Var}(Y_i) = np(1-p).$$

□

Theorem 6.26. If $X \sim \text{Geom}(p)$, then:

$$\mathbb{E}X = \frac{1}{p}, \qquad \text{Var}(X) = \frac{1-p}{p^2}.$$

Proof.

$$\begin{aligned}
\mathbb{E}X &= \sum_{k=1}^{\infty} kp(1-p)^{k-1} = p\sum_{k=1}^{\infty} k(1-p)^{k-1} = \frac{p}{p^2} = \frac{1}{p},\\
\text{Var}(X) &= \mathbb{E}X - (\mathbb{E}X)^2 + \mathbb{E}(X(X-1))\\
&= \frac{1}{p} - \frac{1}{p^2} + \sum_{k=2}^{\infty} k(k-1)p(1-p)^{k-1}\\
&= \frac{1}{p} - \frac{1}{p^2} + \frac{2p(1-p)}{p^3} = \frac{1-p}{p^2}.
\end{aligned}$$

□

Example 6.13. Suppose that there are N types of cards, and each time we obtain a card, it is equally likely to be any one of the N types. We want to find the expected number of types of cards we can get if we obtain n cards, and the expected number of cards needed to get all N types. 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. The expected number of types of cards we can get after obtaining n cards is:

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

To find the expected number of cards needed to get all N types, let Y be the number of cards needed. We can write $Y = Y_0 + Y_1 + \cdots + Y_{N-1}$, where Y_i is the number of additional cards needed to get a new type of card after we have already obtained i types. Therefore, the expected number of cards needed to get all N types is:

$$\begin{aligned}
\mathbb{P}(Y_i = k) &= \left(\frac{i}{N}\right)^{k-1} \frac{N-i}{N}, && \left(Y_i \sim \text{Geom}\left(\tfrac{N-i}{N}\right)\right)\\
\mathbb{E}Y &= \sum_{i=0}^{N-1} \mathbb{E}Y_i = \sum_{i=0}^{N-1} \frac{N}{N-i} = N\left(\frac{1}{N} + \frac{1}{N-1} + \cdots + 1\right).
\end{aligned}$$

Theorem 6.27. If $X \sim \text{NBin}(r, p)$, then:

$$\mathbb{E}X = \frac{r}{p}, \qquad \text{Var}(X) = \frac{r(1-p)}{p^2}.$$

Proof.
Assume that $X_i \sim \text{Geom}(p)$ for all i. Since X is the sum of r independent geometric random variables, we get:

$$\mathbb{E}X = \sum_{k=1}^{r} \mathbb{E}X_k = \frac{r}{p}, \qquad \text{Var}(X) = \sum_{k=1}^{r} \text{Var}(X_k) = \frac{r(1-p)}{p^2}.$$

□

Theorem 6.28. If $X \sim \text{Poisson}(\lambda)$, then:

$$\mathbb{E}X = \lambda, \qquad \text{Var}(X) = \lambda.$$

Proof.

$$\begin{aligned}
\mathbb{E}X &= \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} = \lambda, \\
\text{Var}(X) = \mathbb{E}(X(X-1)) + \mathbb{E}X - (\mathbb{E}X)^2 &= \lambda - \lambda^2 + \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} e^{-\lambda} \\
&= \lambda - \lambda^2 + \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} e^{-\lambda} \\
&= \lambda - \lambda^2 + \lambda^2 = \lambda.
\end{aligned}$$

□

Theorem 6.29. If $X \sim \text{Hypergeometric}(N, m, n)$, then the expectation and variance are:

$$\mathbb{E}X = \frac{mn}{N}, \qquad \text{Var}(X) = \frac{mn}{N}\left(\frac{(m-1)(n-1)}{N-1} + 1 - \frac{mn}{N}\right).$$

We are not going to prove the variance. To prove the expectation, we imagine the following scenario.

Example 6.14. There are N balls in a box, of which m are red and $N-m$ are blue. We randomly select n balls from the box without replacement. Let X be the number of red balls selected. We want to find $\mathbb{E}X$. Let $X = X_1 + X_2 + \cdots + X_m$, where

$$X_i = \begin{cases} 1, & \text{if the } i\text{-th red ball is selected,} \\ 0, & \text{otherwise.} \end{cases}, \qquad \text{for } i = 1, 2, \cdots, m.$$

Since each ball is equally likely to be selected, we have:

$$\mathbb{E}X_i = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n}{N}.$$

Therefore, $\mathbb{E}X = \frac{mn}{N}$.

Theorem 6.30. If $X \sim \mathrm{U}[a,b]$, then:

$$\mathbb{E}X = \frac{1}{2}(a+b), \qquad \mathrm{Var}(X) = \frac{1}{12}(b-a)^2.$$

Proof.

$$\begin{aligned}
\mathbb{E}X &= \int_a^b \frac{x}{b-a}\,dx \\
&= \frac{1}{2}(a+b), \\
\mathrm{Var}(X) &= -(\mathbb{E}X)^2 + \mathbb{E}(X^2) \\
&= -\frac{1}{4}(a+b)^2 + \int_a^b \frac{x^2}{b-a}\,dx \\
&= -\frac{1}{4}(a^2+2ab+b^2) + \frac{1}{3}(a^2+ab+b^2) \\
&= \frac{1}{12}(a^2-2ab+b^2) \\
&= \frac{1}{12}(b-a)^2.
\end{aligned}$$

□

Theorem 6.31. If $X \sim \mathrm{Exp}(\lambda)$, then:

$$\mathbb{E}X = \frac{1}{\lambda}, \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

Proof.

$$\begin{aligned}
\mathbb{E}X &= \int_0^\infty x\lambda e^{-\lambda x}\,dx \\
&= -xe^{-\lambda x}\big|_0^\infty + \int_0^\infty e^{-\lambda x}\,dx \\
&= -\frac{1}{\lambda}e^{-\lambda x}\Big|_0^\infty \\
&= \frac{1}{\lambda}, \\
\mathrm{Var}(X) &= -(\mathbb{E}X)^2 + \mathbb{E}(X^2) \\
&= -\frac{1}{\lambda^2} + \int_0^\infty x^2\lambda e^{-\lambda x}\,dx \\
&= -\frac{1}{\lambda^2} - x^2e^{-\lambda x}\big|_0^\infty + \int_0^\infty 2xe^{-\lambda x}\,dx \\
&= -\frac{1}{\lambda^2} + \frac{2}{\lambda}\mathbb{E}X \\
&= -\frac{1}{\lambda^2} + \frac{2}{\lambda^2} \\
&= \frac{1}{\lambda^2}.
\end{aligned}$$

□

Theorem 6.32. If $X \sim \mathrm{N}(\mu, \sigma^2)$, then:

$$\mathbb{E}X = \mu, \qquad \operatorname{Var}(X) = \sigma^2.$$

Proof.
Let $x = \sigma z + \mu$ for some z.

$$\begin{aligned}
\mathbb{E}X &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} y e^{-\frac{(y-\mu)^2}{2\sigma^2}}\, dy \\
&= \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{\infty} \sigma z e^{-\frac{z^2}{2}}\, dz + \int_{-\infty}^{\infty} \mu e^{-\frac{z^2}{2}}\, dz \right) \\
&= \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\, dz \\
&= \mu, \\
\operatorname{Var}(X) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (y-\mu)^2 e^{-\frac{(y-\mu)^2}{2\sigma^2}}\, dy \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}}\, dz \\
&= \frac{-\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z\, d\left(e^{-\frac{z^2}{2}}\right) \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\, dz \\
&= \sigma^2.
\end{aligned}$$

□

Theorem 6.33. All Cauchy random variables $X \sim \mathrm{Cauchy}(\theta)$ do not have defined expectation and variance.

Proof.
We check if $\mathbb{E}\,|X|$ is infinite.

$$\mathbb{E}\,|X| = \int_{-\infty}^{\infty} \frac{|x|}{\pi(1+(x-\theta)^2)}\, dx = 2\int_{0}^{\infty} \frac{x}{\pi(1+(x-\theta)^2)}\, dx = \infty.$$

Therefore, expectation and thus variance do not exist. □

Theorem 6.34. If $X \sim \chi^2(n)$, then $\mathbb{E}X = n$ and $\operatorname{Var}(X) = 2n$.

Proof.
Since $X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ where $Z_i \sim \mathrm{N}(0,1)$ are independent, we have:

$$\mathbb{E}X = \sum_{i=1}^{n} \mathbb{E}(Z_i^2) = n, \qquad \operatorname{Var}(X) = \sum_{i=1}^{n} \operatorname{Var}(Z_i^2) = 2n.$$

□

Solving the following expectations is out of our scope.

Theorem 6.35. Given a continuous random variable X:

1. If $X \sim \mathrm{Gamma}(a, \lambda)$, then $\mathbb{E}X = \frac{\alpha}{\lambda}$ and $\operatorname{Var}(X) = \frac{\alpha}{\lambda^2}$.
2. If $X \sim t(n)$, then:

$$\mathbb{E}X = \begin{cases} 0, & n > 1, \\ \text{undefined}, & \text{otherwise.} \end{cases} \qquad \operatorname{Var}(X) = \begin{cases} \frac{n}{n-2}, & n > 2, \\ \infty, & 2 < n \leq 4, \\ \text{undefined}, & \text{otherwise.} \end{cases}$$

3. If $X \sim \mathrm{Beta}(a, b)$, then $\mathbb{E}X = \frac{a}{a+b}$ and $\operatorname{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$.

6.4 Combining Expectation from Discrete and Continuous Random Variables

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 that incorporates both these cases. Suppose that X has a CDF F_X. We can rewrite the equations as:

$$\mathbb{E}X = \begin{cases} \sum x\,dF_X(x), & dF_X(x) = F_X(x) - \lim_{y\to x^-} F_X(y) = f_X(x), \\ \int x\,dF_X(x), & dF_X(x) = \frac{\partial F_X}{\partial x}\,dx = f_X(x)\,dx. \end{cases}$$

Instead of using the regular Riemann integral, which cannot handle the discrete case, we can use the Riemann-Stieltjes integral, which is a generalization of the Riemann integral:

$$\int_a^b g(x)\,dx = \lim_{\max_i |x_{i+1}-x_i|} \sum_i g(x_i^*)(x_{i+1}-x_i),$$
$$\int_a^b g(x)\,dF(x) = \lim_{\max_i |x_{i+1}-x_i|} \sum_i g(x_i^*)(F(x_{i+1})-F(x_i)),$$

if the limit does not depend on the choice of $x_i^* \in [x_i, x_{i+1})$.

Definition 6.36. The **expectation** of a random variable X is given by:

$$\mathbb{E}X = \int x\,dF_X.$$

Lemma 6.37. If $g : \mathbb{R} \to \mathbb{R}$ such that $g(X)$ is also a random variable, then:

$$\mathbb{E}(g(X)) = \int g(x)\,dF_X.$$

Remark 6.37.1. The notation $\int g(x)\,dF_X(x)$ does not imply the Riemann-Stieltjes integral.

Example 6.15. If g is regular (differentiable at every point, and every value in the domain maps to a value in the 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 6.16. In the 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^*, the 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:

$$\mathbb{E}\mathbf{1}_{\mathbb{Q}}(X) = \mathbb{P}(\mathbf{1}_{\mathbb{Q}}(x) = 1) = \mathbb{P} \circ X^{-1}(\mathbb{Q} \cap [0,1]) = 0.$$

Summary

Definition

Definition 1. Given a set with n distinct elements:

1. A **permutation** of the set is an ordered arrangement of all elements of the set.
2. If $k \leq n$, a k**-permutation** of the set is an ordered arrangement of k elements of the set.

Definition 2. If $k \leq n$, a k**-combination** of a set with n distinct elements is an unordered arrangement of k elements of the set.

Definition 3. These are the basic objects of probability:

1. An **experiment** is an activity that produces distinct and well-defined possibilities called **outcomes**, denoted by ω.
2. A **sample space** is the set of all outcomes of an experiment, denoted by Ω.
3. An **event** is a subset of the sample space and is usually represented by $A, B, C, \cdots$.
4. Outcomes are called **elementary events**.

Definition 4. Given two events A and B:

1. The **union** of A and B is an event $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$.
2. The **intersection** of A and B is an event $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$.
3. The **complement** of A is an event containing all elements in the sample space Ω that are not in A. It is denoted by A^{C}.
4. The **relative complement** of B in A is an event $A \setminus B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\}$.
5. The **symmetric difference** of A and B is an event $A \Delta B = \{\omega \in \Omega : \omega \in A \cup B \text{ and } \omega \notin A \cap B\}$.

Definition 5. For any two events A and B, if all outcomes in A are also in B, then we say A is **contained** in B, written as $A \subset B$ or $B \supset A$.

Definition 6. Given a sequence of events $A_1, A_2, \cdots, A_k$:

1. For any i and j, if $A_i \cap A_j = \emptyset$, then A_i and A_j are called **disjoint**.
2. If $A_i \cap A_j = \emptyset$ for all i and j, the sequence of events is called **mutually exclusive**.
3. If $A_1 \cup A_2 \cup \cdots \cup A_k = \Omega$, the sequence of events is called **exhaustive**.
4. If the sequence is both mutually exclusive and exhaustive, it is called a **partition**.

Definition 7. (**Kolmogorov Axioms of Probability**) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, with sample space Ω, σ-field $\mathcal{F}$, and probability measure $\mathbb{P}$:

1. The probability of an event is a non-negative real number. For all $E \in \mathcal{F}$:
$$\mathbb{P}(E) \in \mathbb{R}, \qquad \mathbb{P}(E) \geq 0.$$
2. The probability that at least one of the elementary events in the entire sample space will occur is 1:
$$\mathbb{P}(\Omega) = 1.$$
3. Any countable sequence of disjoint events $E_1, E_2, \cdots$ satisfies:
$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i).$$

By this definition, we call $\mathbb{P}(A)$ the **probability** of the event A.

Definition 8. A **σ-field** (**σ-algebra**) $\mathcal{F}$ is any 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. $\emptyset \in \mathcal{F}$.

Definition 9. A **measurable space** $(\Omega, \mathcal{F})$ is a pair comprising a sample space Ω and a σ-field $\mathcal{F}$.

Definition 10. A **probability measure** $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a measure on a measurable space $(\Omega, \mathcal{F})$ satisfying:

1. $\mathbb{P}(\emptyset) = 0$.
2. $\mathbb{P}(\Omega) = 1$.
3. If $A_i \in \mathcal{F}$ for all i and they are disjoint, then $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Definition 11. A **probability space** $(\Omega, \mathcal{F}, \mathbb{P})$ is a triple comprising:

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

Definition 12. 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 13. An event A is **null** if $\mathbb{P}(A) = 0$.

Definition 14. An event A occurs **almost surely** if $\mathbb{P}(A) = 1$.

Definition 15. Given $\mathbb{P}(B) > 0$, the **conditional probability** that A occurs given that B occurs is:
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Definition 16. 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 17. 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-fields. $\sigma(A)$ is the smallest σ-field containing A.

Definition 18. The **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 19. A **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 20. A **Borel set** is a set that can be obtained by taking countable unions, intersections, or complements repeatedly.

Definition 21. The **Borel σ-field** $\mathcal{B}(\mathbb{R})$ of $\mathbb{R}$ is a σ-field generated by all open sets. It is a collection of Borel sets.

Definition 22. The **(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 the **discrete** case, the **probability mass function** (PMF) of a discrete random variable X is the function $f : \mathbb{R} \to [0,1]$ given by:

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

In the **continuous** case, the **probability density function** (PDF) of a 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 23. The **q-th quantile** of a random variable X is defined as a number z_q such that:

$$\mathbb{P}(X \leq z_q) = q.$$

Definition 24. Let $X_i : \Omega \to \mathbb{R}$ for all $1 \leq i \leq n$ be random variables. A **random vector** $\mathbf{X} = (X_1, X_2, \cdots, X_n) : \Omega \to \mathbb{R}^n$ has the following properties:

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

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

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

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

Definition 25. Given a random vector (X, Y), the **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 the discrete case, the **joint probability mass function** (JPMF) of **jointly discrete** random variables 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 the continuous case, the **joint probability density function** (JPDF) of **jointly continuous** random variables 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 26. Let X and Y be random variables. The **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 the discrete case, the **marginal mass function** (Marginal PMF) is given by:

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

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

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

Definition 27. Given a random variable X, the **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 28. Given $k \in \mathbb{N}_+$ and a random variable X, the **k-th moment** m_k is defined as:

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

The **k-th central moment** α_k is defined as:

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

The **mean** μ is the 1st moment $\mu = m_1 = \mathbb{E}X$.
The **variance** is the 2nd central moment $\alpha_2 = \text{Var}(X) = \mathbb{E}((X - \mathbb{E}X)^2) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$.
The **standard deviation** σ is defined as $\sigma = \sqrt{\text{Var}(X)}$.

Definition 29. Given two random variables X and Y, the **conditional distribution function** (Conditional CDF) of Y given $X = x$ for any x is defined as:

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

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

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

Definition 30. Given two random variables X and Y, and an event $X = x$ for some X, the **conditional expectation** of the random variable Y is defined as:

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

Given a random variable X, the conditional expectation of the random variable Y is defined as:

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

Definition 31. Given $X \perp\!\!\!\perp Y$, in the discrete case, the **convolution** f_{X+Y} $(f_X * f_Y)$ of the 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 the continuous case, the **convolution** of the PDFs of random variables X and Y is the PDF of $X + Y$:

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

Definition 32. A **parametric distribution** of a random variable is a distribution where the PMF or PDF depends on one or more parameters.

Named Properties

Property 1. (**Fundamental Principle of Counting**) Suppose that m_i represents the number of outcomes of the i-th event. The total number of outcomes of n independent events is the product of the number of outcomes for each individual event:

$$\prod_{i=1}^{n} m_i.$$

Property 2. (**Pascal's Identity**) Let n and k be integers with $0 < k < n$. Then:

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Property 3. (**Binomial Theorem**) Let n be a non-negative integer. We have:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k},$$

where $\binom{n}{k}$ for all k are called the **binomial coefficients**.

Property 4. (**Vandermonde's Identity**) Let $m, n, r \in \mathbb{Z}$ with $0 \leq r \leq m$ and $0 \leq r \leq n$. We have:

$$\binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{r-k} \binom{n}{k}.$$

Property 5. (**Multinomial Theorem**) Let n be a non-negative integer. We have:

$$(x_1 + x_2 + \cdots + x_k)^n = \sum_{(n_1, n_2, \cdots, n_k): n_1 + n_2 + \cdots + n_k = n} \binom{n}{n_1, n_2, \cdots, n_k} x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k},$$

where $(n_1, n_2, \cdots, n_k)$ are all non-negative integer-valued vectors.

Property 6. (**Inclusion-Exclusion Formula**)

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

Property 7. (**General Multiplication Rule**) Let $A_1, A_2, \cdots, A_n$ be a sequence of events. We have:

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2) \cdots \mathbb{P}(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Property 8. (**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 $\bigcup_{i=1}^{n} B_i = \Omega$).
If $\mathbb{P}(B_i) > 0$ for all i, then:

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

Property 9. (**Bayes' Theorem**) Suppose that a sequence of events $A_1, A_2, \cdots, A_n$ is a partition of the sample space. Assume further that $\mathbb{P}(A_i) > 0$ for all i. Let B be any event. Then, for any i:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{k=1}^{n} \mathbb{P}(B|A_k)\mathbb{P}(A_k)}.$$

Property 10. (**Law of Total Expectation**) Let $\psi(X) = \mathbb{E}(Y|X)$. The conditional expectation satisfies:

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

Property 11. (**Tail Sum Formula**) If a discrete random variable X has a PMF f_X with $f_X(x) = 0$ when $x < 0$, then:

$$\mathbb{E}X = \sum_{k=0}^{\infty} \mathbb{P}(X > k).$$

If a continuous random variable X has a PDF f_X with $f_X(x) = 0$ when $x < 0$, and a CDF F_X, then:

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

Distributions

For discrete random variables:

Example 1. (**Bernoulli Distribution**) $X \sim \text{Bern}(p)$
Suppose we perform one Bernoulli trial. Let p be the probability of success, and X be the 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 the total number of successes.

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

Example 3. (**Trinomial Distribution**)
Suppose we perform n trials with three outcomes: A, B, and C, where the probabilities of occurrence are p, q, and $1-p-q$, respectively. Let X be the number of occurrences of A, and Y be the number of occurrences of B. The probability of x occurrences of A, y occurrences of B, and $n-x-y$ occurrences of C is:

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

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 occurs. Let p be the probability of success.
Let W be the waiting time that elapses before the first success. For $k \ge 1$:

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

The above is the conventional geometric distribution.
Let X be the number of failures before the 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 r-th success occurs. Let p be the probability of success.
Let W_r be the waiting time that elapses before the 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 the 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:

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

Example 7. (**Hypergeometric Distribution**) $X \sim \text{Hypergeometric}(N, m, n)$
Suppose that we have a set of N balls. There are m red balls and $N - m$ blue balls. We choose n of these balls without replacement. Let X be the number of red balls in our sample. For $0 \leq k \leq \min(m, n)$:

$$f_X(k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}, \qquad \mathbb{E}X = \frac{mn}{N}, \qquad \text{Var}(X) = \frac{mn}{N}\left(\frac{(m-1)(n-1)}{N-1} + 1 - \frac{mn}{N}\right).$$

For continuous random variables:

Example 8. (**Uniform Distribution**) $X \sim \text{U}[a, b]$
A random variable X is uniform on $[a, b]$. Its PDF and CDF are:

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

Example 9. (**Exponential Distribution**) $X \sim \text{Exp}(\lambda)$
A random variable X is exponential with parameter $\lambda > 0$. Its PDF and CDF are:

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

Example 10. (**Normal Distribution / Gaussian Distribution**) $X \sim \text{N}(\mu, \sigma^2)$
A random variable X is normal if it has two parameters μ and σ^2. Its PDF and CDF are:

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

A random variable X is standard normal if $\mu = 0$ and $\sigma^2 = 1$ ($X \sim \text{N}(0, 1)$):

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

Example 11. (**Bivariate Normal Distribution**)
Two random variables X and Y are bivariate normal with means μ_X and μ_Y, variances σ_X^2 and σ_Y^2, and population 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 $\sigma_X^2 = \sigma_Y^2 = 1$:

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

Example 12. (**Multivariate Normal Distribution**) $\mathbf{X} \sim \text{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
A random vector $\mathbf{X}$ with dimension p is p-dimensional normal with $p \times 1$ mean vector $\boldsymbol{\mu}$ and $p \times p$ variance-covariance matrix $\boldsymbol{\Sigma}$ if:

$$f(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}(\mathbf{x}-\boldsymbol{\mu})}.$$

Example 13. (**Cauchy Distribution**) $X \sim \text{Cauchy}(\theta)$
A random variable X has a Cauchy distribution with parameter θ if:

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

Example 14. (**Gamma Distribution**) $X \sim \text{Gamma}(\alpha, \lambda)$
A random variable X has a gamma distribution with parameters α and λ if:

$$f_X(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{\Gamma(\alpha)}\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}, & x \geq 0 \end{cases}, \qquad \Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1}\, dy, \qquad \mathbb{E}X = \frac{\alpha}{\lambda}, \qquad \text{Var}(X) = \frac{\alpha}{\lambda^2}.$$

Here, $\Gamma(\alpha)$ is the gamma function, defined as $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$. If α is a positive integer, then $\Gamma(\alpha) = (\alpha-1)!$.

Example 15. (**Chi-Squared Distribution**) $Y \sim \chi^2(n)$
Assume that $X_1, X_2, \cdots, X_n$ are independent standard normal random variables. Let $Y = \sum_{i=1}^n X_i^2$. The random variable Y has a χ^2-distribution with parameter n if:

$$f_Y(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{\Gamma\left(\frac{n}{2}\right)} 2^{-\frac{n}{2}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x \geq 0 \end{cases}, \qquad \mathbb{E}Y = n, \qquad \text{Var}(Y) = 2n.$$

Example 16. (**Student's t-Distribution**) $W \sim t(n)$
Given $Y \sim \chi^2(n)$ and $Z \sim \text{N}(0, 1)$, if Y and Z are independent, let:

$$W = \frac{Z}{\sqrt{\frac{Y}{n}}}.$$

The random variable W follows the t-distribution with n degrees of freedom, and:

$$f(w) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{w^2}{n}\right)^{-\frac{n+1}{2}}, \quad \mathbb{E}W = \begin{cases} \text{Undefined}, & n \leq 1, \\ 0, & n > 1 \end{cases}, \quad \text{Var}(W) = \begin{cases} \text{Undefined}, & n \leq 1, \\ \infty, & 1 < n \leq 2, \\ \frac{n}{n-2}, & n > 2 \end{cases}.$$

Here, $\Gamma(\alpha)$ is the gamma function.

Example 17. (**Beta Distribution**) $X \sim \text{Beta}(a, b)$
A random variable X has a beta distribution with parameters a and b if:

$$f_X(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, & 0 < x < 1, \\ 0, & \text{otherwise} \end{cases}, \qquad B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

$$\mathbb{E}X = \frac{a}{a+b}, \qquad \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Here, $B(a, b)$ is the beta function.

Example 18. (**F-Distribution**) $F \sim F(r_1, r_2)$
Assume that X and Y are independent random variables with $X \sim \chi^2(r_1)$ and $Y \sim \chi^2(r_2)$. Let:

$$F = \frac{\frac{X}{r_1}}{\frac{Y}{r_2}}.$$

Then F has an F-distribution with r_1 and r_2 degrees of freedom, and:

$$f_F(w) = \frac{\Gamma\left(\frac{r_1+r_2}{2}\right)}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)} \left(\frac{r_1}{r_2}\right)^{\frac{r_1}{2}} w^{\frac{r_1}{2}-1} \left(1 + \frac{r_1 w}{r_2}\right)^{-\frac{r_1+r_2}{2}},$$

where $0 < w < \infty$.

Chapter 7

Generating Function

7.1 Introduction to Generating Functions

A sequence of numbers $a = \{a_i : i = 0, 1, 2, \cdots\}$ may contain a lot of information. For example, the values of a PMF describe the distribution of a discrete random variable. A concise way of storing this information is to encapsulate the numbers in a generating function.

Definition 7.1. For any sequence $\{a_n : n = 0, 1, 2, \cdots\}$, we define the generating function as:

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

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

Remark 7.1.1. We can observe that:

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

Remark 7.1.2. Sometimes, we cannot interchange a countable sum with derivatives.

Example 7.1. Let $b_n(x) = \frac{\sin nx}{n}$ such that $a_1(x) = b_1(x)$ and $a_n(x) = b_n(x) - b_{n-1}(x)$.

$$\sum_{n=0}^{\infty} a_n(x) = \lim_{n\to\infty} \sum_{i=0}^{n} a_n(x) = \lim_{n\to\infty} \frac{\sin nx}{n} = 0, \qquad \text{(Squeeze Theorem)}$$

$$\lim_{n\to\infty} \frac{\partial}{\partial x} \sum_{i=0}^{n} a_i(x) = 0,$$

$$\lim_{n\to\infty} \sum_{i=0}^{n} \frac{\partial}{\partial x} a_n(x) = \lim_{n\to\infty} \cos nx \quad \text{does not exist.}$$

Convolutions are common in probability theory, and generating functions provide a useful tool for studying them.

Definition 7.2. 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 7.2. If $a_n = f_X(n)$ and $b_n = f_Y(n)$, then $c_n = f_{X+Y}(n)$.

Lemma 7.3. 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.

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

□

From the definition of a generating function, we see that it is a power series. We may want to determine whether the series is convergent.

Definition 7.4. The **radius of convergence** R of a 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 $s \in [-R, R]^{\mathsf{C}}$, then $f(s)$ is divergent. We can determine the radius of convergence by applying the root test:

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

Remark 7.4.1. Additional tests are required to determine whether the power series converges at $s = -R$ and $s = R$.

Remark 7.4.2. Sometimes, it is difficult to compute R using the root test. A convenient alternative is 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 proofs are not essential.

Theorem 7.5. 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 term by term for any fixed number of times 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 7.5.1. 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:

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

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 7.6. The **probability generating function** (PGF) of a non-negative random variable X is:

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

Using this, we can determine the distribution of a random variable with the following theorem.

Theorem 7.7. Given two random variables X and Y with corresponding PGFs, if the two PGFs are the same, then X and Y have the same distribution.

This is particularly useful for finding the distribution of a random variable.

Example 7.3. 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 $b_n = f_Y(n)$.

$$G_X(s) = \sum_{i=0}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} s^i = e^{\lambda(s-1)}, \qquad G_Y(s) = e^{\mu(s-1)}, \qquad G_Z(s) = e^{(\lambda+\mu)(s-1)}.$$

We may conclude that $Z \sim \text{Poisson}(\lambda + \mu)$.

Remark 7.7.1. If $a_n = f_X(n)$ for some random variable X, then $R \geq 1$ for $G_X(s) = G_a(s)$ since:

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

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

Example 7.4. Let $X \sim \text{Poisson}(\lambda)$ and $a_n = f_X(n) = \frac{\lambda^n e^{-\lambda}}{n!}$. By the ratio test, as $n \to \infty$:

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

Therefore, $R = \infty$.

Example 7.5. Let X have a PMF $a_n = f_X(n) = \frac{c}{n^2}$. By the ratio test, as $n \to \infty$:

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

Therefore, $R = 1$.

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

Theorem 7.8. (**Abel's Theorem**) Suppose that $a_n \geq 0$ for all n. If a has a generating function $G_a(s)$ and a radius of convergence $R = 1$, then if $\sum_{n=0}^{\infty} a_n$ converges in $\mathbb{R} \cup \{\infty\}$, we have:

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

Example 7.6. If a random variable $X \sim \text{Bern}(p)$ for some p, then:

$$G_X(s) = ps^1 + (1-p)s^0 = 1 - p + ps.$$

Example 7.7. If a random variable $X \sim \text{Bin}(p)$ for some n and p, then:

$$G_X(s) = (1 - p + ps)^n.$$

Example 7.8. If a random variable $X \sim \text{Geom}(p)$ for some p, then:

$$G_X(s) = \sum_{n=1}^{\infty} (1-p)^{n-1} ps^n = \frac{ps}{1 - s(1-p)}.$$

Example 7.9. If a random variable $X \sim \text{Poisson}(\lambda)$ for some λ, then:

$$G_X(s) = e^{\lambda(s-1)}.$$

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

Theorem 7.9. If a random variable X has a PGF $G_X(s)$, then:

1. $\mathbb{E}X = \lim_{s\to 1^-} G'(s) = G'(1)$,
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. By evaluating at $s = 1$:

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

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

□

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

Theorem 7.10. Let $X_1, X_2, \cdots$ be a sequence of independent and identically distributed (i.i.d.) random variables with a common PGF $G_X(s)$, and let 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.

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

□

Example 7.10. 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 the JPMF exists, there will obviously be a joint PGF.

Definition 7.11. Let random variables X_1 and X_2 be non-negative integer-valued, jointly discrete with JPMF f_{X_1,X_2}. The **joint probability generating function** (JPGF) is defined as:

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

Remark 7.11.1. We can find that:

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

Theorem 7.12. Random variables X and Y are independent if and only if:

$$G_{X,Y}(s,t) = G_X(s)G_Y(t).$$

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

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

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 $j \geq 0$:

$$\begin{aligned} s^i t^j f_{X,Y}(i,j) &= s^i f_X(i) t^j f_Y(j), \\ f_{X,Y}(i,j) &= f_X(i) f_Y(j). \end{aligned}$$

Therefore, $X \perp\!\!\!\perp Y$. □

Theorem 7.13. If random variables X and Y are independent, then:

$$G_{X+Y}(t) = G_X(t)G_Y(t).$$

Proof.

$$G_{X+Y}(t) = \mathbb{E}(t^{X+Y}) = \mathbb{E}(t^X)\mathbb{E}(t^Y) = G_X(t)G_Y(t).$$

□

Remark 7.13.1. The converse does not necessarily hold true.

7.2 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:

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

Definition 7.14. A random variable X is **defective** if $\mathbb{P}(X = \infty) > 0$.

Remark 7.14.1. It is no surprise that the expectation is infinite when the random variable is defective.

With this generalization, we can start discussing random walks.

Example 7.11. (**Recurrence and Transience of Random Walk**) Let Y_n be the position of a particle after n moves in a simple random walk on $\mathbb{Z}$, and let X_i be independent and identically distributed random variables mentioned in Appendix A. For $n \geq 0$:

$$Y_n = \sum_{i=1}^{n} X_i, \qquad Y_0 = 0, \qquad \mathbb{P}(X_i = 1) = p, \qquad \mathbb{P}(X_i = -1) = q = 1 - p.$$

Define the first return time to the origin as:

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

We want to determine whether T_0 is a defective random variable. To do this, we need to calculate $\mathbb{P}(T_0 = \infty)$. Let $p_0(n)$ be the probability that the particle is at the origin after n moves, and let P_0 be the generating function of p_0. Let $f_0(n)$ be the probability that the particle returns to the origin for the first time after n moves, and let F_0 be the generating function of f_0.

$$\begin{aligned}
p_0(n) &= \mathbb{P}(Y_n = 0) = \begin{cases} \binom{n}{\frac{n}{2}} p^{\frac{n}{2}} q^{\frac{n}{2}}, & n \text{ is even}, \\ 0, & n \text{ is odd} \end{cases}, \\
P_0(s) &= \lim_{N \to \infty} \sum_{n=0}^{N} p_0(n) s^n, \\
f_0(n) &= \mathbb{P}(Y_1 \neq 0, Y_2 \neq 0, \cdots, Y_{n-1} \neq 0, Y_n = 0) = \mathbb{P}(T_0 = n), \\
F_0(s) &= \lim_{N \to \infty} \sum_{n=1}^{N} f_0(n) s^n.
\end{aligned}$$

Theorem 7.15. From the definitions in Example 7.11, 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. By using the Law of Total Probability:

$$\begin{aligned}
p_0(n) &= \sum_{i=1}^{n} \mathbb{P}(Y_n = 0 | Y_1 \neq 0, Y_2 \neq 0, \cdots, Y_{i-1} \neq 0, Y_i = 0) f_0(i), \\
&= \sum_{i=1}^{n} \mathbb{P}(Y_n = 0 | Y_i = 0) f_0(i), && \text{(Markov property in Lemma A.1)} \\
&= \sum_{i=1}^{n} \mathbb{P}(Y_{n-i} = 0) f_0(i), && \text{(Temporarily homogeneous property in Lemma A.1)} \\
&= \sum_{i=1}^{n} p_0(n-k) f_0(i), \\
p_0(0) &= 1, \\
P_0(s) &= \sum_{k=0}^{\infty} p_0(k)s^k = 1 + \sum_{k=1}^{\infty} p_0(k)s^k, \\
&= 1 + \sum_{k=1}^{\infty}\sum_{i=1}^{k} p_0(k-i) f_0(i) s^k, \\
&= 1 + \sum_{i=1}^{\infty}\sum_{k=i}^{\infty} p_0(k-i) s^{k-i} f_0(i) s^i, \\
&= 1 + P_0(s)F_0(s).
\end{aligned}$$

2. *If you want to understand the proof, search "Central binomial coefficient" in Wikipedia.*
   We know that $Y_n = 0$ if n is even. Therefore:

$$\begin{aligned}
P_0(s) &= \lim_{N\to\infty} \sum_{n=0}^{N} p_0(n)s^n, \\
&= \lim_{N\to\infty} \sum_{i=0}^{N} \binom{2i}{i} p^i q^i s^{2i}, \\
&= \lim_{N\to\infty} \sum_{i=1}^{N} (-1)^i 4^i \binom{\frac{-1}{2}}{i} p^i q^i s^{2i}, && \left(\binom{\frac{-1}{2}}{i} \text{ is a generalized binomial coefficient}\right) \\
&= \frac{1}{\sqrt{1 - 4pqs^2}}.
\end{aligned}$$

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 derive the following corollary.

Corollary 7.16. The probability that the particle ever returns to the origin is:

$$\sum_{n=1}^{\infty} f_0(n) = F_0(1) = 1 - |p-q|\,.$$

The probability that the particle will never return to the origin is:

$$\mathbb{P}(T_0 = \infty) = |p-q|\,.$$

Proof.
By using Theorem 7.15, since $p + q = 1$:

$$\begin{aligned} F_0(1) &= 1 - (1-4pq)^{\frac{1}{2}}, \\ &= 1 - (p^2 - 2pq + q^2)^{\frac{1}{2}}, \\ &= 1 - |p-q|\,. \end{aligned}$$

□

Definition 7.17. A random walk is **recurrent** if it has at least one recurrent point, or equivalently:

$$\mathbb{P}(X < \infty) = 1.$$

A random walk is **transient** if it has no recurrent points, or equivalently:

$$\mathbb{P}(X = \infty) > 0.$$

Remark 7.17.1. When $p = q = \frac{1}{2}$, $\mathbb{P}(T_0 = \infty) = 0$, and therefore the random walk is recurrent.

Remark 7.17.2. If $p \neq q$, then $\mathbb{P}(T_0 = \infty) \neq 0$, and so the random walk is transient.

Example 7.12. Revisit Example 7.11. When $p = q = \frac{1}{2}$:

$$F_0(s) = 1 - \sqrt{1-s^2}, \qquad F_0'(s) = \frac{s}{\sqrt{1-s^2}}, \qquad \mathbb{E}T_0 = \lim_{s \to 1^-} F_0'(s) = \infty.$$

Therefore, even though the random walk is recurrent, the expected return time is infinite.

We now move on to our next important application, the Branching Process.
Many scientists have been interested in modeling reproduction in a population. Accurate models for evolution are extremely difficult to handle, but some non-trivial models are tractable. We will investigate one such model.

Example 7.13. (**Galton-Watson Process**) Consider a population in which each individual in generation n produces some number of offspring according to a fixed probability distribution, and these offspring form generation $n+1$. This process continues indefinitely. This is known as a **branching process** or **Galton-Watson process**.
Let Z_n be the size of generation n. Let $X_i^{(n)}$ be the number of offspring produced by the i-th individual in generation n. Then:

$$Z_{n+1} = \begin{cases} X_1^{(n)} + X_2^{(n)} + \cdots + X_{Z_n}^{(n)}, & Z_n \geq 1, \\ 0, & Z_n = 0. \end{cases}$$

We make the following assumptions:

1. All individuals reproduce independently. ($X_i^{(k)}$'s are independent.)

2. All individuals reproduce according to the same distribution. ($X_i^{(k)}$'s are identically distributed.)

We also assume that $Z_0 = 1$ (the population starts with one individual). We are interested in the distribution of Z_n for all n.

Remark 7.17.3. $Z_1 = X_1^{(0)}$

Theorem 7.18. 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 $n = 2$:

$$\begin{aligned} G_2(s) &= \mathbb{E}s^{Z_2}, \\ &= \mathbb{E}s^{X_1^{(1)}+X_2^{(1)}+\cdots+X_{Z_1}^{(1)}}, \\ &= G_{Z_1}\left(G_{X_1^{(1)}}(s)\right), \\ &= G(G(s)). \end{aligned}$$

When $n = m+1$ for some m:

$$\begin{aligned} G_{m+1}(s) &= \mathbb{E}s^{Z_{m+1}}, \\ &= \mathbb{E}s^{X_1^{(m)}+X_2^{(m)}+\cdots+X_{Z_m}^{(m)}}, \\ &= G_{Z_m}\left(G_{X_1^{(m)}}(s)\right), \\ &= G_m(G(s)). \end{aligned}$$

□

In principle, the above theorem describes the distribution of Z_n. However, computing $G_n(s)$ explicitly can be quite challenging. Fortunately, the moments of Z_n can be computed more easily.

Lemma 7.19. Let $\mathbb{E}Z_1 = \mathbb{E}X_i^{(m)} = \mu$ and $\operatorname{Var}(Z_1) = \sigma^2$. Then:

$$\mathbb{E}Z_n = \mu^n, \qquad \operatorname{Var}(Z_n) = \begin{cases} n\sigma^2, & \mu = 1, \\ \frac{\sigma^2(\mu^n-1)\mu^{n-1}}{\mu-1}, & \mu \neq 1. \end{cases}$$

Proof.
Using Theorem 7.18, we find:

$$\begin{aligned}
\mathbb{E}Z_2 &= G_2'(1), \\
&= G'(G(1))G'(1), \\
&= G'(1)\mu, \\
&= \mu^2, \\
\mathbb{E}Z_n &= G_n'(1), \\
&= G'(G_{n-1}(1))G_{n-1}'(1), \\
&= G'(1)\mu^{n-1}, \\
&= \mu^n. \\
G_1''(1) &= \sigma^2 + (G'(1))^2 - G'(1), \\
&= \sigma^2 + \mu^2 - \mu, \\
G_2''(1) &= G''(G(1))(G'(1))^2 + G'(G(1))G''(1), \\
&= G''(1)(\mu^2 + \mu), \\
G_n''(1) &= G''(G_{n-1}(1))(G_{n-1}'(1))^2 + G'(G_{n-1}(1))G_{n-1}''(1), \\
&= (\sigma^2 + \mu^2 - \mu)\mu^{2n-2} + \mu G_{n-1}''(1), \\
&= \mu^{2n-2}(\sigma^2 + \mu^2 - \mu) + \mu^{2n-3}(\sigma^2 + \mu^2 - \mu) + \cdots + \mu^{n-1}(\sigma^2 + \mu^2 - \mu), \\
&= \frac{\mu^{n-1}(\sigma^2 + \mu^2 - \mu)(\mu^n - 1)}{\mu - 1}.
\end{aligned}$$

If $\mu = 1$:

$$\begin{aligned}
\operatorname{Var}(Z_n) &= G_n''(1) + G_n'(1) - (G_n'(1))^2, \\
&= \sigma^2 + G_{n-1}''(1) + 1 - 1, \\
&= n\sigma^2.
\end{aligned}$$

If $\mu \neq 1$:

$$\begin{aligned}
\operatorname{Var}(Z_n) &= G_n''(1) + G_n'(1) - (G_n'(1))^2, \\
&= \frac{\mu^{n-1}(\sigma^2 + \mu^2 - \mu)(\mu^n - 1)}{\mu - 1} + \mu^n - \mu^{2n}, \\
&= \frac{\mu^{n-1}\sigma^2(\mu^n - 1)}{\mu - 1}.
\end{aligned}$$

□

Example 7.14. We are particularly interested in the probability of ultimate extinction, i.e., the event that the population eventually dies out. Note that if $Z_n = 0$ for some n, then $Z_m = 0$ for all $m > n$. Therefore:

$$\begin{aligned}
\{\text{ultimate extinction}\} &= \bigcup_n \{Z_n = 0\} = \lim_{n\to\infty} \{Z_n = 0\}, \\
\mathbb{P}(\text{ultimate extinction}) &= \mathbb{P}\left(\lim_{n\to\infty} \{Z_n = 0\}\right) = \lim_{n\to\infty} \mathbb{P}(Z_n = 0) = \lim_{n\to\infty} G_n(0).
\end{aligned}$$

Let $\eta_n = G_n(0)$ and $\eta = \lim_{n\to\infty} \eta_n$.

Theorem 7.20. The probability of ultimate extinction η is the smallest non-negative root of the equation:

$$s = G(s).$$

Furthermore:

1. If $\mu \leq 1$, then $\eta = 1$.
2. If $\mu > 1$, then $\eta < 1$ and $G'(\eta) \leq 1$.
3. If $\mu = 1$ and $\sigma^2 > 0$, then $\eta = 1$.
4. If $\mu = 1$ and $\sigma^2 = 0$, then $\eta = 0$.

Proof.

$$\eta_n = G_n(0) = G(G_{n-1}(0)) = G(\eta_{n-1}).$$

Since $G(s)$ is continuous and non-decreasing on $[0, 1]$, η_n is a non-decreasing sequence bounded above by 1. Therefore, η_n converges to some limit $\eta \leq 1$.

$$\eta = \lim_{n\to\infty} \eta_n = \lim_{n\to\infty} G(\mu_{n-1}) = G\left(\lim_{n\to\infty} \eta_{n-1}\right) = G(\eta).$$

Thus, η is a non-negative root of the equation $s = G(s)$.
Let ψ be any non-negative root of the equation $s = G(s)$. We will show that $\eta \leq \psi$. Since $G(s)$ is non-decreasing on $[0, 1]$ and $\psi \geq 0$:

$$\eta_1 = G(0) \leq G(\psi) = \psi, \qquad \eta_2 = G(\eta_1) \leq G(\psi) = \psi.$$

By induction, we find that $\eta_n \leq \psi$ for all n. Taking the limit as $n \to \infty$, we obtain $\eta \leq \psi$.
Next, we analyze the properties of $G(s)$ on $[0, 1]$:

$$G''(s) = \sum_{i=2}^{\infty} i(i-1)s^{i-2}\mathbb{P}(Z_1 = i) \geq 0.$$

Thus, $G(s)$ is convex on $[0, 1]$. Since $G(1) = 1$, the line $y = s$ intersects the curve $y = G(s)$ at $s = 1$.
If $\mu = G'(1) \leq 1$, then $G(s)$ lies above the line $y = s$ for all $s \in [0, 1)$, and so the only intersection is at $s = 1$. Thus, $\eta = 1$.
If $\mu = G'(1) > 1$, then $G(s)$ intersects the line $y = s$ at some $s = k \in (0, 1)$. Since $G(s)$ is convex, there can be only one such intersection. Thus, $\eta = k < 1$.
Furthermore, since $G(s)$ is convex and intersects the line $y = s$ at $s = \eta$ and $s = 1$, we have:

$$\sigma^2 = G''(1) + G'(1) - (G'(1))^2 = G''(1).$$

If $\mu = 1$ and $\sigma^2 > 0$, then $G''(1) > 0$. Therefore, $G(s)$ is strictly convex, and the line $y = s$ intersects the curve $y = G(s)$ only at $s = 1$. Thus, $\eta = 1$.
If $\mu = 1$ and $\sigma^2 = 0$, then $G''(1) = 0$. Therefore, $G(s)$ is linear, and since $G(1) = 1$ and $G'(1) = 1$, we have $G(s) = s$. Thus, $\eta = 0$. □

7.3 Moment Generating Function and Characteristic Function

Recall that both discrete and continuous distributions can be unified into one framework. We can redefine the PGF accordingly.

Definition 7.21. The **probability generating function** of a random variable X is given by:

$$\mathbb{E}s^X = \int s^x \, dF_X.$$

For more general random variables X, it is convenient to substitute $s = e^t$. This leads to the following definition.

Definition 7.22. The **moment generating function** (MGF) of a random variable X is the function $M : \mathbb{R} \to [0, \infty)$ defined as:

$$M_X(t) = \mathbb{E}(e^{tX}) = \int e^{tx} \, dF_X.$$

Remark 7.22.1. The definition of the MGF is obtained by replacing s with e^t in the PGF. While the MGF is more convenient for computing moments, it is less useful for deriving distributions.

Remark 7.22.2. MGFs are closely related to Laplace transforms.

Definition 7.23. The **joint moment generating function** (JMGF) of two random variables X and Y is defined as:

$$M_{X,Y}(s,t) = \mathbb{E}(e^{sX+tY}).$$

The following lemma can be derived easily.

Lemma 7.24. Given the MGF $M_X(t)$ of a random variable X:

1. For any $k \geq 0$:
$$\mathbb{E}X^k = M_X^{(k)}(0).$$
2. The function M can be expanded using Taylor's theorem within its radius of convergence:
$$M_X(t) = \sum_{i=0}^{\infty} \frac{\mathbb{E}X^i}{i!} t^i.$$
3. If X and Y are independent, then:
$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

Proof.

1. $$M^{(k)}(0) = \frac{\partial^k}{\partial t^k} \int e^{tx} \, dF_X(x) \bigg|_{t=0} = \int x^k e^{tx} \, dF_X(x) \bigg|_{t=0} = \int x^k \, dF_X(x) = \mathbb{E}X^k.$$
2. This follows directly from (1) and Taylor's theorem.
3. Substitute $s = e^t$ into Theorem 7.13.

□

Lemma 7.25. If $X_1, X_2, \cdots, X_n$ are independent, then:

$$M_{X_1,X_2,\cdots,X_n}(t_1, t_2, \cdots, t_n) = M_{X_1}(t_1)M_{X_2}(t_2)\cdots M_{X_n}(t_n).$$

Proof.
By independence:

$$M_{X_1,X_2,\cdots,X_n}(t_1, t_2, \cdots, t_n) = \mathbb{E}(e^{t_1X_1+t_2X_2+\cdots+t_nX_n}) = \mathbb{E}(e^{t_1X_1})\mathbb{E}(e^{t_2X_2})\cdots\mathbb{E}(e^{t_nX_n}) = M_{X_1}(t_1)M_{X_2}(t_2)\cdots M_{X_n}(t_n).$$

□

Remark 7.25.1. $M_X(0) = 1$ for all random variables X.

Example 7.15. Let $X \sim \text{Bern}(p)$. We have:

$$M_X(t) = \mathbb{E}(e^{tX}) = q + pe^t.$$

Example 7.16. Let $X \sim \text{Bin}(n, p)$. We have:

$$M_X(t) = (q + pe^t)^n.$$

Example 7.17. Let $X \sim \text{Geom}(p)$. We have:

$$f_X(x) = p(1-p)^{x-1}, \qquad M_X(t) = \sum_{k=1}^{\infty} e^{tk} p(1-p)^{k-1} = \frac{pe^t}{1 - e^t(1-p)},$$

$$f_X(x) = p(1-p)^x, \qquad M_X(t) = \sum_{k=0}^{\infty} e^{tk} p(1-p)^k = \frac{p}{1 - e^t(1-p)}.$$

Example 7.18. Let $X \sim \text{NBin}(r, p)$. We have:

$$f_X(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \qquad M_X(t) = \left(\frac{pe^t}{1 - e^t(1-p)}\right)^n,$$

$$f_X(x) = \binom{x+r-1}{r-1} p^r (1-p)^x, \qquad M_X(t) = \left(\frac{p}{1 - e^t(1-p)}\right)^n.$$

Example 7.19. Let $X \sim \text{Poisson}(\lambda)$. We have:

$$M_X(t) = \sum_{k=0}^{\infty} \frac{\lambda^k e^{tk-\lambda}}{k!} = e^{\lambda(e^t - 1)}.$$

Example 7.20. Let $X \sim \text{U}[a, b]$ for some $a < b$. We have:

$$M_X(t) = \begin{cases} \int_a^b \frac{e^{tx}}{b-a}\, dx = \frac{e^{tb} - e^{ta}}{t(b-a)}, & t \neq 0, \\ 1, & t = 0. \end{cases}$$

Example 7.21. Let $X \sim \text{Exp}(\lambda)$. For $t < \lambda$, we have:

$$M_X(t) = \int_0^{\infty} \lambda e^{x(t-\lambda)}\, dx = \frac{\lambda}{\lambda - t}.$$

Example 7.22. Let $X \sim \text{N}(\mu, \sigma^2)$. We have:

$$\begin{aligned} M_X(t) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{tx} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx, \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - 2(\mu - \sigma^2 t)x + \mu^2}{2\sigma^2}\right) dx, \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(\mu - \sigma^2 t)^2 - \mu^2}{2\sigma^2}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{(x - (\mu - \sigma^2 t))^2}{2\sigma^2}\right) dx, \\ &= \exp\left(\frac{-2\mu\sigma^2 t + \sigma^4 t^2}{2\sigma^2}\right), \\ &= \exp\left(\frac{1}{2}\sigma^2 t^2 - \mu t\right). \end{aligned}$$

Example 7.23. Let $X \sim \text{Cauchy}(0)$.

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

$M_X(t)$ exists only at $t = 0$. We get $M_X(0) = 1$.

Example 7.24. Let $X \sim \text{Gamma}(\alpha, \lambda)$. If $t < \lambda$, we have:

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-\alpha}.$$

Example 7.25. Let $X \sim \chi^2(k)$. If $t < \frac{1}{2}$, we have:

$$M_X(t) = (1-2t)^{-\frac{k}{2}}.$$

Example 7.26. Let $X \sim \text{Beta}(\alpha, \beta)$. We have:

$$M_X(t) = 1 + \sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!}.$$

Remark 7.25.2. Not all distributions have MGFs.

Example 7.27. The MGF of $X \sim t(n)$ is undefined.

Moment generating functions provide a useful technique, but the integrals used to define them may not always converge. There is another class of functions for which convergence is guaranteed.

Definition 7.26. The **characteristic function** (CF) of a random variable X is the function $\phi_X : \mathbb{R} \to \mathbb{C}$ defined as:

$$\phi_X(t) = \mathbb{E}e^{itX} = \int e^{itx}\,dF_X(x) = \mathbb{E}\text{cos}(tX) + i\mathbb{E}\text{sin}(tX), \qquad i = \sqrt{-1}.$$

Remark 7.26.1. $\phi_X(t)$ is essentially a Fourier transform.

Lemma 7.27. The CF ϕ_X of a random variable X has the following properties:

1. $\phi_X(0) = 1$. Moreover, $|\phi_X(t)| \leq 1$ for all t.
2. $\phi_X(t)$ is uniformly continuous.

Proof.

1. For all t:

$$\begin{aligned}\phi_X(0) &= \int dF_X(x) = 1,\\ |\phi_X(t)| &= \left|\int (\cos(tx) + i\sin(tx))\,dF_X(x)\right| \leq \int |\cos(tx) + i\sin(tx)|\,dF_X(x) = \int dF_X(x) = 1.\end{aligned}$$

2. $$\sup_t |\phi_X(t+c) - \phi_X(t)| = \sup_t \left|\int (e^{i(t+c)x} - e^{itx})\,dF_X(x)\right| \leq \sup_t\left(\int \left|e^{itx}\right|\left|e^{icx} - 1\right|\,dF_X(x)\right).$$

As $c \to 0$, the supremum approaches 0. Therefore, $\phi_X(t)$ is uniformly continuous.

□

Theorem 7.28. The characteristic function ϕ_X of a random variable X has the following properties regarding derivatives and moments:

1. If $\phi_X^{(k)}(0)$ exists, then:
$$\begin{cases} \mathbb{E}\left|X\right|^k < \infty, & k \text{ is even}, \\ \mathbb{E}\left|X\right|^{k-1} < \infty, & k \text{ is odd}. \end{cases}$$

2. If $\mathbb{E}\left|X\right|^k < \infty$, then $\phi_X^{(k)}(0)$ exists. Moreover:
$$\phi_X(t) = \sum_{j=0}^{k} \frac{\phi_X^{(j)}(0)}{j!} t^j + o(t^k) = \sum_{j=0}^{k} \frac{\mathbb{E}X^j}{j!} (it)^j + o(t^k).$$

Proof.
Using 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 $\phi_X^{(k)}(0) = (-1)^{\frac{k}{2}} \mathbb{E}X^k = (-1)^{\frac{k}{2}} \mathbb{E}\left|X\right|^k$, which exists. Therefore, $\mathbb{E}\left|X\right|^k < \infty$.
If k is odd, note that $\phi_X^{(k-1)}(0)$ exists if $\phi_X^{(k)}(0)$ exists.
Thus, with $\phi_X^{(k-1)}(0) = (-1)^{\frac{k-1}{2}} \mathbb{E}X^{k-1} = (-1)^{\frac{k-1}{2}} \mathbb{E}\left|X\right|^{k-1}$, we conclude $\mathbb{E}\left|X\right|^{k-1} < \infty$.

2. Using the formula in (1):
$$\frac{\phi_X^{(k)}(0)}{i^k} = \mathbb{E}X^k \leq \mathbb{E}\left|X\right|^k < \infty.$$
Therefore, $\phi_X^{(k)}(0)$ exists. The formula follows directly from Taylor's theorem.

□

Theorem 7.29. If $X \perp\!\!\!\perp Y$, then:
$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t).$$

Proof.
$$\phi_{X+Y}(t) = \mathbb{E}(e^{it(X+Y)}) = \mathbb{E}(e^{itX})\mathbb{E}(e^{itY}) = \phi_X(t)\phi_Y(t).$$

□

The concept of a joint characteristic function is defined as follows:

Definition 7.30. The **joint characteristic function** (JCF) $\phi_{X,Y}$ of two random variables X and Y is given by:

$$\phi_{X,Y}(s,t) = \mathbb{E}(e^{i(sX+tY)}).$$

This leads to another method for proving the independence of two random variables:

Theorem 7.31. Two random variables X and 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 $X \perp\!\!\!\perp Y$:

$$\phi_{X,Y}(s,t) = \mathbb{E}(e^{i(sX+tY)}) = \mathbb{E}(e^{isX})\mathbb{E}(e^{itY}) = \phi_X(s)\phi_Y(t).$$

To prove the converse, additional theorems are required (see Example 7.33). □

Example 7.28. Let $X \sim \mathrm{Bern}(p)$. We have:

$$\phi_X(t) = \mathbb{E}(e^{itX}) = q + pe^{it}.$$

Example 7.29. Let $X \sim \mathrm{Bin}(n,p)$. We have:

$$\phi_X(t) = (q + pe^{it})^n.$$

Example 7.30. Let $X \sim \mathrm{Exp}(1)$. We have:

$$\phi_X(t) = \int e^{(it-1)x}\,dx = \frac{1}{1-it}.$$

Example 7.31. Let $X \sim$ Cauchy. We have:

$$\phi_X(t) = e^{-|t|}.$$

Example 7.32. Let $X \sim \mathrm{N}(\mu,\sigma^2)$. 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:

$$\begin{aligned}
\phi_X(t) &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{itx}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx,\\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2-(2\mu+2\sigma^2 it)x+\mu^2}{2\sigma^2}\right)dx,\\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{(\mu+\sigma^2 it)^2-\mu^2}{2\sigma^2}\right)\int_{-\infty}^{\infty}\exp\left(-\frac{(x-(\mu+\sigma^2 it))^2}{2\sigma^2}\right)dx,\\
&= \exp\left(\frac{\mu^2+2\sigma^2 i\mu t-\sigma^4t^2-\mu^2}{2\sigma^2}\right),\\
&= \exp\left(i\mu t-\frac{1}{2}\sigma^2t^2\right).
\end{aligned}$$

Remark 7.31.1. The **cumulant generating function** is defined as $\log\phi_X(t)$. The normal distribution is the only distribution we have encountered whose cumulant generating function has a finite number of terms:

$$\log\phi_X(t) = i\mu t - \frac{1}{2}\sigma^2t^2.$$

7.4 Inversion and Continuity Theorems

Characteristic functions are useful in two major ways. One of them is that the characteristic function of a random variable can be used to generate the probability density function of that random variable.

Theorem 7.32. (**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 points x where 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 (Non-Rigorous).
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 $I_\varepsilon(x) \to I(x)$ as $\varepsilon \to 0$.

$$\begin{aligned} I_\varepsilon(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\varepsilon^2t^2+i(y-x)t}f_X(y)\,dt\,dy,\\ &= \frac{1}{\sqrt{2\pi\varepsilon^2}}\left(\frac{1}{\sqrt{2\pi\frac{1}{\varepsilon^2}}}\right)\int_{-\infty}^{\infty}\exp\left(-\frac{(y-x)^2}{2\varepsilon^2}\right)f_X(y)\int_{-\infty}^{\infty}\exp\left(-\frac{-\left(t-i\frac{y-x}{\varepsilon}\right)^2}{2\left(\frac{1}{\varepsilon^2}\right)}\right)dt\,dy,\\ &= \frac{1}{\sqrt{2\pi\varepsilon^2}}\int_{-\infty}^{\infty}\exp\left(-\frac{(y-x)^2}{2\varepsilon^2}\right)f_X(y)\,dy. \end{aligned}$$

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 conclude that $f_{\varepsilon Z+X}(x) \to f_X(x)$ as $\varepsilon \to 0$. □

Theorem 7.33. (**Inversion Theorem**) If a random variable X has 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 $a \leq b$:

$$\overline{F}_X(b) - \overline{F}_X(a) = \int_{-\infty}^{\infty}\frac{e^{-iat}-e^{-ibt}}{2\pi it}\phi_X(t)\,dt.$$

Remark 7.33.1. The function $\overline{F}_X$ represents the average of the limits from both directions.

Example 7.33. Using the Inversion Theorem, we can now prove Theorem 7.31.
Given two random variables X and Y, we first extend the Fourier Inverse Transform to the multivariable case.
If $\phi_{X,Y}(s,t) = \phi_X(s)\phi_Y(t)$, then for any $a \leq b$ and $c \leq d$:

$$\begin{aligned} \overline{F}_{X,Y}(b,d) - \overline{F}_{X,Y}(b,c) - \overline{F}_{X,Y}(a,d) + \overline{F}_{X,Y}(a,c) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{(e^{-ias}-e^{-ibs})(e^{-ict}-e^{-idt})}{-4\pi^2t^2}\phi_X(s)\phi_Y(t)\,ds\,dt,\\ &= (\overline{F}_X(b) - \overline{F}_X(a))\int_{-\infty}^{\infty}\frac{e^{-ict}-e^{-idt}}{2\pi it}\phi_Y(t)\,dt,\\ &= (\overline{F}_X(b) - \overline{F}_X(a))(\overline{F}_Y(d) - \overline{F}_Y(c)),\\ &= \overline{F}_X(b)\overline{F}_Y(d) - \overline{F}_X(b)\overline{F}_Y(c) - \overline{F}_X(a)\overline{F}_Y(d) + \overline{F}_X(a)\overline{F}_Y(c). \end{aligned}$$

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 application is to evaluate the convergence of a sequence of cumulative distribution functions.

Definition 7.34. (**Convergence of Distribution Function Sequence [Weak Convergence]**) A sequence of CDFs $F_1, F_2, \cdots$ **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 7.34. Assume we have two sequences of CDFs:

$$F_n(x) = \begin{cases} 0, & x < \frac{1}{n}, \\ 1, & x \geq \frac{1}{n}, \end{cases} \qquad G_n(x) = \begin{cases} 0, & x < \frac{-1}{n}, \\ 1, & x \geq \frac{-1}{n}. \end{cases}$$

As $n \to \infty$, 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 since it is not right-continuous. Therefore, it is necessary to define convergence so that both sequences $\{F_n\}$ and $\{G_n\}$ have the same limit.

We can modify the definition slightly to state that each distribution function in the sequence represents a different random variable.

Definition 7.35. (**Convergence in Distribution for Random Variables**) Let $X, X_1, X_2, \cdots$ be a family of random variables with CDFs $F, F_1, F_2, \cdots$. We say $X_n \to X$, written as $X_n \xrightarrow{D} X$ or $X_n \Rightarrow X$, if $F_n \to F$.

Remark 7.35.1. For this convergence definition, we do not consider the closeness of X_n and X as functions of ω.

Remark 7.35.2. Sometimes, we also write $X_n \Rightarrow F$ or $X_n \xrightarrow{D} F$.

Using this definition, a sequence of characteristic functions can be used to determine whether the sequence of cumulative distribution functions converges.

Theorem 7.36. (**Lévy Continuity Theorem**) Suppose that $F_1, F_2, \cdots$ is a sequence of CDFs with CFs $\phi_1, \phi_2, \cdots$. 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 0 ($t = 0$), then ϕ is the CF of some CDF F, and $F_n \to F$.

Remark 7.36.1. In Lévy Continuity Theorem (2), the statement that ϕ is continuous at 0 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 7.35. Let $X_n \sim \mathrm{N}(0, \frac{1}{n^2})$, and let ϕ_n be the CF of X_n. Then

$$\phi_n(t) = \exp\left(-\frac{t^2}{2n^2}\right) \to \phi(t) = 1.$$

Since $\phi(t)$ is continuous at 0, by Lévy Continuity Theorem (2), we conclude that $X_n \xrightarrow{D} 0$.

Example 7.36. Let $X_n \sim \mathrm{N}(0, n^2)$, and let ϕ_n be the CF of X_n. Then

$$\phi_n(t) = \exp\left(-\frac{1}{2}n^2t^2\right) \to \phi(t) = \begin{cases} 0, & t \neq 0 \\ 1, & t = 0 \end{cases}.$$

We have a more general definition of convergence.

Definition 7.37. (**Vague convergence**) Given a sequence of CDFs $F_1, F_2, \cdots$, suppose that $F_n(x) \to G(x)$ at all continuity points 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 7.37. 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 $F_n \xrightarrow{v} G$ as $n \to \infty$, and G is not a CDF.

7.5 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 7.38. (**Weak Law of Large Numbers**) [WLLN] Let $X_1, X_2, \cdots$ be i.i.d. random variables. Assume that $\mathbb{E}\left|X_1\right| < \infty$ and $\mathbb{E}X_1 = \mu$. We have:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{D} \mu.$$

Proof.
We recall the Taylor expansion of $\phi_\xi(s)$ at 0. If $\mathbb{E}\left|\xi\right|^k < \infty$ and s is small, then

$$\phi_\varsigma(s) = \sum_{j=0}^{k} \frac{\mathbb{E}\xi^j}{j!}(is)^j + o(s^k).$$

For any $t \in \mathbb{R}$, let $\phi_{X_1}(s) = \mathbb{E}e^{isX_1}$.

$$\begin{aligned}
\phi_n(t) &= \mathbb{E}\left(\exp\left(\frac{it}{n}\sum_{i=1}^{n} X_i\right)\right)\\
&= \mathbb{E}\left(\prod_{i=1}^{n}\exp\left(\frac{itX_i}{n}\right)\right)\\
&= \left(\mathbb{E}\left(\exp\left(\frac{itX_1}{n}\right)\right)\right)^n\\
&= \left(\phi_{X_1}\left(\frac{t}{n}\right)\right)^n\\
&= \left(1 + \frac{it}{n}\mathbb{E}X_1 + o\left(\frac{t}{n}\right)\right)^n\\
&= \left(1 + \frac{i\mu t}{n} + o\left(\frac{t}{n}\right)\right)^n \to e^{i\mu t}.
\end{aligned}$$

By Lévy continuity theorem, we get that $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{D} \mu$. □

Theorem 7.39. (**Central Limit Theorem**) [CLT] Let $X_1, X_2, \cdots$ be i.i.d. random variables with $\mathbb{E}\left|X_1\right|^2 < \infty$ and $\mathbb{E}X_1 = \mu$, $\mathrm{Var}(X_1) = \sigma^2$. Then

$$\frac{1}{\sigma}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{D} \mathrm{N}(0,1).$$

Proof.
Let $Y_i = \frac{X_i - \mu}{\sigma}$. We have $\mathbb{E}Y_i = 0$ and $\mathrm{Var}(Y_i) = 1$.

$$\begin{aligned}
\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} &= \sum_{i=1}^{n} \frac{1}{\sqrt{n}}\frac{X_i - \mu}{\sigma} = \sum_{i=1}^{n}\frac{Y_i}{\sqrt{n}}\\
\phi_n(t) &= \mathbb{E}\left(\exp\left(it\sum_{\ell=1}^{n}\frac{Y_\ell}{\sqrt{n}}\right)\right)\\
&= \left(\mathbb{E}\left(\exp\left(\frac{itY_1}{\sqrt{n}}\right)\right)\right)^n\\
&= \left(\phi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right)^n\\
&= \left(1 + \frac{it}{\sqrt{n}}\mathbb{E}Y_1 + \frac{1}{2}\left(\frac{it}{\sqrt{n}}\right)^2\mathbb{E}Y_1^2 + o\left(\frac{t^2}{n}\right)\right)^n \qquad \text{(Taylor expansion)}\\
&= \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \to e^{-\frac{1}{2}t^2}.
\end{aligned}$$

By Lévy continuity theorem, $\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{D} \mathrm{N}(0,1)$. □

The Central Limit Theorem can be generalized in several directions, one of which concerns independent random variables instead of i.i.d. random variables.

Theorem 7.40. Let $X_1, X_2, \cdots$ be independent random variables satisfying $\mathbb{E}X_i = 0$, $\text{Var}(X_i) = \sigma_i^2$, $\mathbb{E}\left|X_i\right|^3 < \infty$, 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 $(\sigma(n))^2 = \text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \sigma_i^2$. Then

$$\frac{1}{\sigma(n)}\sum_{i=1}^{n} X_i \xrightarrow{D} \text{N}(0,1).$$

Remark 7.40.1. The condition (*) means that none of the random variables X_i can dominate the sum.

$$\frac{1}{(\sigma(n))^3}\sum_{i=1}^{n}\left|X_i\right|^3 \lesssim \frac{1}{\sigma(n)}\max_{i=1,2,\cdots,n}\left|X_i\right|\left(\frac{1}{(\sigma(n))^2}\right)\sum_{i=1}^{n}(X_i)^2 \approx \frac{1}{\sigma(n)}\max_{i=1,2,\cdots,n}\left|X_i\right| \to 0.$$

This theorem is a special case of the Central Limit Theorem that focuses on the sum of Bernoulli random variables.

Theorem 7.41. (**De Moivre-Laplace Limit Theorem**) Suppose that $X \sim \text{Bin}(n,p)$. Then for any $a < b$, as $n \to \infty$,

$$\mathbb{P}\left(a < \frac{X - np}{\sqrt{np(1-p)}} \leq b\right) \to \Phi(b) - \Phi(a).$$

Proof.
Our goal is to transform the PMF of the Binomial random variable into the PDF of the standard normal distribution. Let $q = 1 - p$. For $0 \leq k \leq n$, by applying Stirling's formula,

$$\begin{aligned}
\binom{n}{k}p^k q^{n-k} &= \frac{n!}{k!(n-k)!}p^k q^{n-k} \\
&\sim \sqrt{\frac{n}{2\pi k(n-k)}}\frac{n^n}{k^k(n-k)^{n-k}}p^k q^{n-k} = \sqrt{\frac{n}{2\pi k(n-k)}}\left(\frac{np}{k}\right)^k\left(\frac{nq}{n-k}\right)^{n-k} \\
&\qquad\qquad\qquad\qquad \text{(Using Stirling's formula: } n! \sim \sqrt{2\pi n}\left(\tfrac{n}{e}\right)^n\text{)} \\
&\sim \frac{1}{\sqrt{2\pi npq}}\exp\left(-k\ln\left(\frac{k}{np}\right) + (k-n)\ln\left(\frac{n-k}{nq}\right)\right). \qquad \left(\tfrac{k}{n} \to p\right)
\end{aligned}$$

With $\mathbb{E}X = np$ and $\text{Var}(X) = np(1-p)$, for any integer k chosen between 0 and n, there exists an arbitrary finite point c such that $k = np + c\sqrt{np(1-p)}$. Using the Taylor series expansion $\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + o(x^3)$, we get:

$$\begin{aligned}
\binom{n}{k}p^k q^{n-k} &\sim \frac{1}{\sqrt{2\pi npq}}\exp\left((-np - c\sqrt{npq})\ln\left(\frac{np + c\sqrt{npq}}{np}\right) + (np + c\sqrt{npq} - n)\ln\left(\frac{n - np - c\sqrt{npq}}{nq}\right)\right) \\
&= \frac{1}{\sqrt{2\pi npq}}\exp\left((-np - c\sqrt{npq})\ln\left(1 + c\sqrt{\frac{q}{np}}\right) + (c\sqrt{npq} - nq)\ln\left(1 - c\sqrt{\frac{p}{nq}}\right)\right) \\
&= \frac{1}{\sqrt{2\pi npq}}\exp\left((-np - c\sqrt{npq})\left(c\sqrt{\frac{q}{np}} - \frac{c^2 q}{2np} + o(n^{-1})\right) + (c\sqrt{npq} - nq)\left(-c\sqrt{\frac{p}{nq}} - \frac{c^2 p}{2nq} + o(n^{-1})\right)\right) \\
&= \frac{1}{\sqrt{2\pi npq}}\exp\left((-c\sqrt{npq} - c^2 q + \frac{1}{2}c^2 q + o(1)) + (-c^2 p + c\sqrt{npq} - \frac{1}{2}c^2 p + o(1))\right) \\
&\sim \frac{1}{\sqrt{2\pi npq}}\exp\left(\frac{1}{2}c^2\right) = \frac{1}{\sqrt{2\pi npq}}\exp\left(-\frac{(k-np)^2}{2npq}\right).
\end{aligned}$$

Therefore, as $n \to \infty$, $\frac{X-np}{\sqrt{np(1-p)}} \xrightarrow{D} \text{N}(0,1)$, and the theorem is proven. □

Remark 7.41.1. Let $X_1, X_2, \cdots, X_n$ be a random sample of a population $X \sim \text{Bern}(p)$. We have:

$$\mathbb{P}\left(a < \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq b\right) \to \Phi(b) - \Phi(a).$$

7.6 Sampling

In many cases, we do not know the actual distribution of the population. We can only predict the distribution based on the samples we obtain. This section is closer to statistics than probability, so we will not delve deeply into it.

Definition 7.42. A set of random variables $\{X_1, X_2, \cdots, X_n\}$ is called a **random sample** of a random variable X with PMF or PDF $f_X(x)$ and CDF $F_X(x)$ if they are independent and identically distributed (i.i.d.).

1. The **sample mean** of X, denoted by $\overline{X}$, is defined as:
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$
2. The **sample variance** of X, denoted by S_{n-1}^2, is defined as:
$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Remark 7.42.1. Notice that the denominator is $n-1$.

Theorem 7.43. Given a sample mean $\overline{X}$ of a random variable X, we have $\mathbb{E}\overline{X} = \mu$ and $\operatorname{Var}(\overline{X}) = \frac{\sigma^2}{n}$.

Proof.

$$\begin{aligned}
\mathbb{E}\overline{X} = \mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}X_k = \frac{1}{n}\sum_{k=1}^{n}\mu = \mu, \\
\operatorname{Var}(\overline{X}) = \frac{1}{n^2}\sum_{k=1}^{n}\operatorname{Var}(X_k) &= \frac{n\sigma^2}{n^2} = \frac{1}{n}\sigma^2.
\end{aligned}$$

□

Theorem 7.44. Given a sample variance S_{n-1}^2 of a random variable X, we have $\mathbb{E}S_{n-1}^2 = \sigma^2$.

Proof.

$$\begin{aligned}
\mathbb{E}S_{n-1}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i - \overline{X})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\big(\mathbb{E}(X_i-\mu)^2 + \mathbb{E}(\overline{X}-\mu)^2 - 2\mathbb{E}((X_i-\mu)(\overline{X}-\mu))\big) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\big(\operatorname{Var}(X_i) + \operatorname{Var}(\overline{X}) - 2\operatorname{cov}(X_i, \overline{X})\big) \\
&= \frac{n\sigma^2}{n-1} + \frac{\sigma^2}{n-1} - \frac{2}{n-1}\sum_{i=1}^{n}\operatorname{cov}\left(X_i, \frac{1}{n}\sum_{j=1}^{n}X_j\right) \\
&= \frac{n\sigma^2}{n-1} + \frac{\sigma^2}{n-1} - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}(X_i, X_j) \\
&= \frac{n\sigma^2}{n-1} + \frac{\sigma^2}{n-1} - \frac{2\sigma^2}{n-1} = \sigma^2.
\end{aligned}$$

□

By using the CLT, we can estimate μ_X if we know the value of σ_X^2.

Theorem 7.45. Let $X_1, X_2, \cdots, X_n$ be a random sample from the population $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$. We have:

$$\overline{X} \sim \mathrm{N}\left(\mu_X, \frac{\sigma_X^2}{n}\right).$$

Proof.
By Theorem 5.15 and the properties of the Normal distribution,

$$X_1 + X_2 + \cdots + X_n \sim \mathrm{N}(n\mu_X, n\sigma_X^2).$$

By Lemma 5.13, we have:

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \sim \mathrm{N}\left(\mu_X, \frac{\sigma_X^2}{n}\right).$$

□

What if we want to estimate σ_X^2 using μ_X?

Theorem 7.46. Let $X_1, X_2, \cdots, X_n$ be a random sample from the population $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$. Then we have:

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu_X}{\sigma_X}\right)^2 \sim \chi^2(n).$$

Proof.
Using the properties of the Normal distribution, for any $i = 1, \cdots, n$, we have:

$$\frac{X_i - \mu_X}{\sigma_X} \sim \mathrm{N}(0, 1).$$

Therefore, by the definition of the χ^2-distribution,

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu_X}{\sigma_X}\right)^2 \sim \chi^2(n).$$

□

How do we find σ_X if μ_X is unknown? We can use the following theorem.

Theorem 7.47. Let $X_1, X_2, \cdots, X_n$ be a random sample from the population $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$. Then we have:

1. $\overline{X}$ and S_{n-1}^2 are independent.
2. $$\frac{(n-1)S_{n-1}^2}{\sigma_X^2} \sim \chi^2(n-1).$$

Proof.

1. It suffices to prove that $\overline{X}$ and $X_i - \overline{X}$ are independent for any $i = 1, 2, \cdots, n$.
   We know that $\overline{X} \sim \mathrm{N}(\mu_X, \frac{\sigma_X^2}{n})$ and $X_i - \overline{X} \sim \mathrm{N}(0, \frac{(n+1)\sigma_X^2}{n})$.
   $$\mathrm{cov}(\overline{X}, X_i - \overline{X}) = 0.$$
2. We may find that:
   $$\begin{aligned}\frac{(n-1)S_{n-1}^2}{\sigma_X^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma_X^2} = \sum_{i=1}^{n} \frac{((X_i - \mu_X) + (\mu_X - \overline{X}))^2}{\sigma_X^2} &= \sum_{i=1}^{n} \frac{(X_i - \mu_X)^2}{\sigma_X^2} - \frac{n(\overline{X} - \mu_X)^2}{\sigma_X^2} \\ &= \sum_{i=1}^{n} \left(\frac{X_i - \mu_X}{\sigma_X}\right)^2 - \left(\frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}\right)^2.\end{aligned}$$
   We know that $\frac{X_i - \mu_X}{\sigma_X} \sim \mathrm{N}(0, 1)$. Let:
   $$U = \frac{(n-1)S_{n-1}^2}{\sigma_X^2}, \qquad V = \left(\frac{X_1 - \mu_X}{\sigma_X}\right)^2.$$
   Using Theorem 7.46 and by definition, we have:
   $$U + V \sim \chi^2(n), \qquad V \sim \chi^2(1).$$
   We can use the MGF to prove the theorem. Note that for any $i = 1, 2, \cdots, n$, $\overline{X}$ and $X_i - \overline{X}$ are independent.
   $$M_U(t) = \frac{M_{U+V}(t)}{M_V(t)} = (1 - 2t)^{-\frac{n-1}{2}} \implies U = \frac{(n-1)S_{n-1}^2}{\sigma_X^2} \sim \chi^2(n-1).$$

□

Theorem 7.48. Let $X_1, X_2, \cdots, X_n$ be a random sample from the population $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$. We have:
$$\frac{\overline{X} - \mu_X}{\frac{S_{n-1}}{\sqrt{n}}} \sim t(n-1).$$

Proof.
By Theorems 7.45 and 7.47, we let:
$$U = \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \sim \mathrm{N}(0, 1), \qquad V = \frac{(n-1)S_{n-1}^2}{\sigma_X^2} \sim \chi^2(n-1).$$

By the definition of the t-distribution, we have:
$$\frac{\overline{X} - \mu_X}{\frac{S_{n-1}}{\sqrt{n}}} = \frac{U}{\sqrt{\frac{V}{n-1}}} \sim t(n-1).$$

□

Chapter 8

Convergence of random variables

We discussed convergence in distribution in Chapter 7. However, this is not the only important mode of convergence for random variables. In this chapter, we introduce other modes of convergence.

8.1 Modes of convergence

Several modes of convergence for a sequence of random variables will be discussed.
Let us 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 $\varepsilon > 0$,
   $$\mu\left(\{x \in [0,1] : |f_n(x) - f(x)| > \varepsilon\}\right) \to 0 \text{ as } n \to \infty.$$

These definitions can be used to define convergence modes for random variables.

Definition 8.1. (**Almost sure convergence**) We say $X_n \to X$ **almost surely**, written 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 8.1.1. $X_n \xrightarrow{\text{a.s.}} X$ is an adaptation of pointwise convergence for functions.

Remark 8.1.2. 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 8.2. (**Convergence in r-th mean**) Let $r \geq 1$. We say $X_n \to X$ **in r-th mean**, written as $X_n \xrightarrow{r} X$, if

$$\mathbb{E}\,|X_n - X|^r \to 0 \text{ as } n \to \infty.$$

Example 8.1. 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 8.3. (**Convergence in probability**) We say $X_n \to X$ **in probability**, written 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 8.4. (**Convergence in distribution**) We say that $X_n \to X$ **in distribution**, written as $X_n \xrightarrow{D} X$, if at every continuity point 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 formulas.

Lemma 8.5. (**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 8.5.1. 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 (in Appendix C) for its proof. Therefore, we will not prove it here.

Lemma 8.6. (**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 obtain almost sure convergence.

Lemma 8.7. 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 $\lim_{m\to\infty} \mathbb{P}(B_m(\varepsilon)) = 0$ 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 $\varepsilon > 0$,
   $$\mathbb{P}\left(\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n(\varepsilon)\right) = 0, \qquad \Longrightarrow \mathbb{P}(C^{\complement}) = \mathbb{P}\left(\bigcup_{\varepsilon>0} \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n(\varepsilon)\right) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n\left(\frac{1}{k}\right)\right) = 0.$$
   Therefore, $X_n \xrightarrow{\text{a.s.}} X$ if and only if $\lim_{m\to\infty} \mathbb{P}(B_m(\varepsilon)) = 0$ for all $\varepsilon > 0$.
2. From (1), for all $\varepsilon > 0$,
   $$\sum_{n=1}^{\infty} \mathbb{P}(A_n(\varepsilon)) < \infty \implies \lim_{m\to\infty} \sum_{n=m}^{\infty} \mathbb{P}(A_n(\varepsilon)) = 0 \implies \lim_{m\to\infty} \mathbb{P}(B_m(\varepsilon)) = 0 \implies (X_n \xrightarrow{\text{a.s.}} X).$$

□

Now we can explore the relationships between different convergence modes. Roughly speaking, convergence in distribution is the weakest among all convergence modes, as it only concerns the distribution of X_n.

Theorem 8.8. 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 $r \geq s \geq 1$, then $(X_n \xrightarrow{r} X) \implies (X_n \xrightarrow{s} X)$.

Proof.

1. (a) From Lemma 8.7, for all $\varepsilon > 0$,

$$\mathbb{P}(A_m(\varepsilon)) \leq \mathbb{P}\left(\bigcup_{n=m}^{\infty} A_n(\varepsilon)\right) = \mathbb{P}(B_m(\varepsilon)) \to 0.$$

Therefore, $(X_n \xrightarrow{\text{a.s.}} X) \implies (X_n \xrightarrow{\mathbb{P}} X)$.

(b) From Markov's inequality, since $r \geq 1$,

$$0 \leq \mathbb{P}(|X - X_n| > \varepsilon) = \mathbb{P}(|X - X_n|^r > \varepsilon^r) \leq \frac{\mathbb{E}\,|X_n - X|^r}{\varepsilon^r}.$$

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)

$$\mathbb{P}(X_n \leq x) = \mathbb{P}(X_n \leq x, X \leq x + \varepsilon) + \mathbb{P}(X_n \leq x, X > x + \varepsilon) \leq \mathbb{P}(X \leq x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon).$$

$$\mathbb{P}(X \leq y) \leq \mathbb{P}(X_n \leq y + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon), \quad \mathbb{P}(X_n \leq x) \geq \mathbb{P}(X \leq x - \varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon). \tag*{$(y = x - \varepsilon)$}$$

Since $X_n \xrightarrow{\mathbb{P}} X$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ for all $\varepsilon > 0$. Therefore,

$$\mathbb{P}(X \leq x - \varepsilon) \leq \liminf_{n\to\infty} \mathbb{P}(X_n \leq x) \leq \limsup_{n\to\infty} \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x + \varepsilon).$$

By letting $\varepsilon \to 0$,

$$\mathbb{P}(X \leq x) \leq \liminf_{n\to\infty} \mathbb{P}(X_n \leq x) \leq \limsup_{n\to\infty} \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x).$$

Therefore, $\lim_{n\to\infty} \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x)$, 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 $r \geq s$,

$$\mathbb{E}\,|X_n - X|^s \leq (\mathbb{E}\,|X_n - X|^r)^{\frac{s}{r}} \to 0.$$

□

Remark 8.8.1. The converses of the implications in Theorem 8.8 do not hold in general.

Example 8.2. Let $\Omega = \{H, T\}$ and $\mathbb{P}(H) = \mathbb{P}(T) = \frac{1}{2}$. Define

$$X_{2m}(\omega) = \begin{cases} 1, & \omega = H, \\ 0, & \omega = T, \end{cases} \qquad X_{2m+1}(\omega) = \begin{cases} 0, & \omega = H, \\ 1, & \omega = T. \end{cases}$$

Since $F(x) = F_n(x)$ for all n, we have $X_n \xrightarrow{D} X$. However, for $\varepsilon \in [0, 1]$,

$$\mathbb{P}(|X_n - X| > \varepsilon) \not\to 0 \qquad \text{as } n \to \infty.$$

Therefore,

$$(X_n \xrightarrow{D} X) \not\Longrightarrow (X_n \xrightarrow{\mathbb{P}} X).$$

Example 8.3. Let $r = 1$ and

$$X_n = \begin{cases} n, & \text{probability } = \frac{1}{n}, \\ 0, & \text{probability } = 1 - \frac{1}{n}, \end{cases} \qquad X = 0.$$

As $n \to \infty$, we have:

$$\mathbb{P}(|X_n - X| > \varepsilon) = \frac{1}{n} \to 0, \qquad \mathbb{E}\,|X_n - X| = n\left(\frac{1}{n}\right) = 1 \not\to 0.$$

Therefore,

$$(X_n \xrightarrow{\mathbb{P}} X) \not\Longrightarrow (X_n \xrightarrow{r} X).$$

Example 8.4. Let $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}([0, 1])$, and $\mathbb{P}$ be uniform. Let I_i be such that:

$$I_{\frac{1}{2}m(m-1)+1}, I_{\frac{1}{2}m(m-1)+2}, \cdots, I_{\frac{1}{2}m(m-1)+m}$$

are m disjoint intervals that partition $[0, 1]$ with length $\frac{1}{m}$ for all $m = 1, 2, \cdots$. We have $I_1 = [0, 1]$, $I_2 \cup I_3 = [0, 1]$, $\cdots$. Define

$$X_n(\omega) = \mathbf{1}_{I_n(\omega)} = \begin{cases} 1, & \omega \in I_n, \\ 0, & \omega \in I_n^{\complement}, \end{cases} \qquad X(\omega) = 0 \text{ for all } \omega \in \Omega.$$

For all $\varepsilon \in [0, 1]$,

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(I_n) = \frac{1}{n} \to 0 \qquad \text{as } n \to \infty.$$

Therefore, $X_n \xrightarrow{\mathbb{P}} X$. However, for any $\omega \in \Omega$, there exists a subsequence $X_k(\omega)$ such that $\omega \in I_k$ for all k. Thus,

$$X_k(\omega) = \mathbf{1}_{I_k(\omega)} = 1.$$

Therefore, $X_n(\omega) \not\to X(\omega)$ for all $\omega \in \Omega$. We have:

$$(X_n \xrightarrow{\mathbb{P}} X) \not\Longrightarrow (X_n \xrightarrow{\text{a.s.}} X).$$

Example 8.5. If $r \geq s \geq 1$, let

$$X_n = \begin{cases} n, & \text{probability } = n^{-\left(\frac{r+s}{2}\right)}, \\ 0, & \text{probability } = 1 - n^{-\left(\frac{r+s}{2}\right)}, \end{cases} \qquad X = 0.$$

As $n \to \infty$, we have:

$$\mathbb{E}\left|X_n - X\right|^s = n^s \left(n^{-\left(\frac{r+s}{2}\right)}\right) = n^{\frac{s-r}{2}} \to 0, \qquad \mathbb{E}\left|X_n - X\right|^r = n^r \left(n^{-\left(\frac{r+s}{2}\right)}\right) = n^{\frac{r-s}{2}} \to \infty.$$

Therefore, if $r \geq s \geq 1$,

$$(X_n \xrightarrow{s} X) \;\not\Longrightarrow\; (X_n \xrightarrow{r} X).$$

Example 8.6. Consider

$$X_n = \begin{cases} n^3, & \text{Probability } = n^{-2}, \\ 0, & \text{Probability } = 1 - n^{-2}. \end{cases}$$

By applying Lemma 8.7, 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, $X_n \xrightarrow{\text{a.s.}} X$. However, in mean, as $n \to \infty$, we have

$$\mathbb{E}\left|X_n - X\right| = n^3 \left(\frac{1}{n^2}\right) = n \to \infty.$$

Therefore, the sequence does not converge in mean. We have:

$$(X_n \xrightarrow{\text{a.s.}} X) \;\not\Longrightarrow\; (X_n \xrightarrow{r} X).$$

Example 8.7. Consider

$$X_n = \begin{cases} 1, & \text{Probability } = n^{-1}, \\ 0, & \text{Probability } = 1 - n^{-1}. \end{cases}$$

As $n \to \infty$, we have:

$$\mathbb{E}\left|X_n - X\right| = 1\left(\frac{1}{n}\right) = \frac{1}{n} \to 0.$$

Therefore, $X_n \xrightarrow{1} X$. However, for any $m > 1$, by applying Lemma 8.7, for some $\varepsilon \in (0,1)$,

$$\begin{aligned} \mathbb{P}(B_m(\varepsilon)) &= 1 - \lim_{r\to\infty} \mathbb{P}(X_n = 0 \text{ for all } n \text{ such that } m \leq n \leq r) \\ &= 1 - \lim_{r\to\infty} \prod_{i=m}^{r} \frac{i-1}{i} \\ &= 1 - \lim_{r\to\infty} \frac{m-1}{r} \to 1 \neq 0. \end{aligned}$$

Therefore, the sequence does not converge almost surely. We have:

$$(X_n \xrightarrow{1} X) \;\not\Longrightarrow\; (X_n \xrightarrow{\text{a.s.}} X).$$

Some partial converses of the implications in Theorem 8.8 do hold.

Theorem 8.9. 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 find 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

$$\mathbb{P}(|X_n - c| \geq \varepsilon) \to 0 + 1 - 1 = 0.$$

Hence, $X_n \xrightarrow{\mathbb{P}} c$.

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 $\varepsilon \to 0$, $\mathbb{E}\,|X_n - X|^r \to 0$ and therefore $X_n \xrightarrow{r} X$.

□

Note that any sequence $\{X_n\}$ that satisfies $X_n \xrightarrow{\mathbb{P}} X$ always has a subsequence that converges almost surely to X.

Theorem 8.10. If $X_n \xrightarrow{\mathbb{P}} X$, then there exists a subsequence $X_{n_1}, X_{n_2}, \cdots$ such that

$$X_{n_i} \xrightarrow{\text{a.s.}} X.$$

Proof.
Since $X_n \xrightarrow{\mathbb{P}} X$, for all $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| > \varepsilon) \to 0 \text{ as } n \to \infty.$$

We can find a subsequence $X_{n_1}, X_{n_2}, \cdots$ such that for all $i \geq 1$,

$$\mathbb{P}(|X_{n_i} - X| > i^{-1}) \leq i^{-2}.$$

For any fixed $\varepsilon > 0$, we have

$$\sum_{i > \varepsilon^{-1}} \mathbb{P}(|X_{n_i} - X| > \varepsilon) \leq \sum_{i > \varepsilon^{-1}} \mathbb{P}(|X_{n_i} - X| > i^{-1}) \leq \sum_i i^{-2} < \infty.$$

By Lemma 8.7, we get that $X_{n_i} \xrightarrow{\text{a.s.}} X$ as $i \to \infty$. □

So far, we have only discussed convergence of a single sequence of random variables. We now consider the convergence of two sequences of random variables.

Theorem 8.11. (**Slutsky's Theorem**) If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{\mathbb{P}} c$, for a constant c, then:

1. $X_n + Y_n \xrightarrow{D} X + c$.
2. $X_n Y_n \xrightarrow{D} Xc$.
3. $\frac{X_n}{Y_n} \xrightarrow{D} \frac{X}{c}$ if $c \neq 0$.

Proof.

1. Suppose that $c > 0$ and pick δ such that $0 < \delta < c$. We can find N such that $\mathbb{P}(|Y_n - c| > \delta) < \delta$ for $n \geq N$. For all x, we have:
$$\mathbb{P}(X_n + Y_n \leq x) \leq \mathbb{P}(X_n + Y_n \leq x, |Y_n - c| \leq \delta) + \mathbb{P}(|Y_n - c| > \delta) \leq \mathbb{P}(X_n \leq x - c + \delta) + \delta,$$
$$\mathbb{P}(X_n + Y_n > x) \leq \mathbb{P}(X_n + Y_n > x, |Y_n - c| \leq \delta) + \delta \leq \mathbb{P}(X_n > x - c - \delta) + \delta.$$
By sending $n \to \infty$ and $\delta \to 0$, we find that $\mathbb{P}(X_n + Y_n \leq x) \to \mathbb{P}(X + c \leq x)$ when $c > 0$.
A similar argument can be used to prove that $\mathbb{P}(X_n + Y_n \leq x) \to \mathbb{P}(X + c \leq x)$ when $c < 0$.
Suppose that $c = 0$. Choose an arbitrarily small number $\delta > 0$ and a number N such that $\mathbb{P}(|Y_n| > \delta) < \delta$ for $n \geq N$. For all x, we have:
$$\mathbb{P}(X_n + Y_n \leq x) \leq \mathbb{P}(X_n + Y_n \leq x, |Y_n| \leq \delta) + \mathbb{P}(|Y_n| > \delta) \leq \mathbb{P}(X_n \leq x + \delta) + \delta,$$
$$\mathbb{P}(X_n + Y_n > x) \leq \mathbb{P}(X_n + Y_n \leq x, |Y_n| \leq \delta) + \mathbb{P}(|Y_n| > \delta) \leq \mathbb{P}(X_n \leq x - \delta) + \delta.$$
By sending $n \to \infty$ and $\delta \to 0$, we find that $\mathbb{P}(X_n + Y_n \leq x) \to \mathbb{P}(X + c \leq x)$ when $c = 0$.
Therefore, $X_n + Y_n \xrightarrow{D} X + c$.

2. Suppose that $c > 0$ and pick δ such that $0 < \delta < c$. We can find N such that $\mathbb{P}(|Y_n - c| > \delta) < \delta$ for $n \geq N$. For all x, we have:
$$\mathbb{P}(X_n Y_n \leq x) \leq \mathbb{P}(X_n Y_n \leq x, |Y_n - c| \leq \delta) + \mathbb{P}(|Y_n - c| > \delta) \leq \mathbb{P}\left(X_n \leq \frac{x}{c - \delta}\right) + \delta,$$
$$\mathbb{P}(X_n Y_n > x) \leq \mathbb{P}(X_n Y_n > x, |Y_n - c| \leq \delta) + \delta \leq \mathbb{P}\left(X_n > \frac{x}{c + \delta}\right) + \delta.$$
By sending $n \to \infty$ and $\delta \to 0$, we find that $\mathbb{P}(X_n Y_n \leq x) \to \mathbb{P}(Xc \leq x)$ when $c > 0$.
A similar argument can be used to prove that $\mathbb{P}(X_n Y_n \leq x) \to \mathbb{P}(Xc \leq x)$ when $c < 0$.
Suppose that $c = 0$. Choose an arbitrarily small number $\delta > 0$ and a number N such that $\mathbb{P}(|Y_n| > \delta) < \delta$ for $n \geq N$. For all x, we have:
$$\mathbb{P}(X_n Y_n \leq x) \leq \mathbb{P}(X_n Y_n \leq x, |Y_n| \leq \delta) + \mathbb{P}(|Y_n| > \delta) \leq \mathbb{P}(-\delta X_n \leq x) + \delta,$$
$$\mathbb{P}(X_n Y_n > x) \leq \mathbb{P}(X_n Y_n > x, |Y_n| \leq \delta) + \mathbb{P}(|Y_n| > \delta) \leq \mathbb{P}(\delta X_n > x) + \delta.$$
By sending $n \to \infty$ and $\delta \to 0$, we find that $\mathbb{P}(X_n Y_n \leq x) \to \mathbb{P}(0 \leq x)$ when $c = 0$.
Therefore, $X_n Y_n \xrightarrow{D} Xc$.

3. It suffices to prove that $Y_n^{-1} \xrightarrow{\mathbb{P}} c^{-1}$ if $Y_n \xrightarrow{\mathbb{P}} c$, or equivalently by Theorem 8.9, $Y_n \xrightarrow{D} c$.
If $Y_n \xrightarrow{D} c$, then $\mathbb{P}(Y_n \leq x) \to \mathbb{P}(c \leq x)$ as $n \to \infty$ for all x. When $x \geq 0$, as $n \to \infty$,
$$\mathbb{P}\left(\frac{1}{Y_n} \leq x\right) = \mathbb{P}(Y_n < 0) + \mathbb{P}\left(Y_n \geq \frac{1}{x}\right) \to \mathbb{P}(c < 0) + \mathbb{P}\left(c \geq \frac{1}{x}\right) = \mathbb{P}\left(\frac{1}{c} \leq x\right).$$
When $x < 0$, as $n \to \infty$,
$$\mathbb{P}\left(\frac{1}{Y_n} \leq x\right) = \mathbb{P}\left(\frac{1}{x} \leq Y_n < 0\right) \to \mathbb{P}\left(\frac{1}{x} \leq c < 0\right) = \mathbb{P}\left(\frac{1}{c} \leq x\right).$$
Therefore, $\frac{1}{Y_n} \xrightarrow{D} \frac{1}{c}$ and thus $\frac{1}{Y_n} \xrightarrow{\mathbb{P}} \frac{1}{c}$. By using (2), $\frac{X_n}{Y_n} \xrightarrow{D} \frac{X}{c}$.

□

Theorem 8.12. (**Continuous Mapping Theorem**) Given a sequence of random variables $\{X_n\}$ and a random variable X, let f be a function with the set of discontinuity points D_f such that $\mathbb{P}(X \in D_f) = 0$. Then:

1. If $X_n \xrightarrow{D} X$, then $f(X_n) \xrightarrow{D} f(X)$.
2. If $X_n \xrightarrow{\mathbb{P}} X$, then $f(X_n) \xrightarrow{\mathbb{P}} f(X)$.
3. If $X_n \xrightarrow{\text{a.s.}} X$, then $f(X_n) \xrightarrow{\text{a.s.}} f(X)$.

Proof.

1. *Current knowledge in these notes does not suffice to prove (1). Search for "Weak convergence of measures" on Wikipedia to start.*

2. Fix an arbitrary $\varepsilon > 0$. For any $\delta > 0$, consider a set B_δ defined as:

$$B_\delta = \{x : x \notin D_f \text{ and there exists } y \text{ such that } |x - y| < \delta \text{ and } |f(x) - f(y)| > \varepsilon\}.$$

Suppose that $|f(X) - f(X_n)| > \varepsilon$. This means either $|X - X_n| \geq \delta$, $X \in D_f$, or $X \in B_\delta$. Therefore,

$$\mathbb{P}(|f(X_n) - f(X)| > \varepsilon) \leq \mathbb{P}(|X_n - X| \geq \delta) + \mathbb{P}(X \in D_f) + \mathbb{P}(X \in B_\delta).$$

Since $X_n \xrightarrow{\mathbb{P}} X$, $\mathbb{P}(|X_n - X| \geq \delta) \to 0$ as $n \to \infty$ for all $\delta > 0$.
Since $\mathbb{P}(X \in D_f) = 0$ by assumption and $\mathbb{P}(X \in B_\delta) \to 0$ as $\delta \to 0$, by first sending $n \to \infty$ and then $\delta \to 0$, we get:

$$\mathbb{P}(|f(X_n) - f(X)| > \varepsilon) \to 0.$$

Generalizing to all $\varepsilon > 0$, we get $f(X_n) \xrightarrow{\mathbb{P}} f(X)$.

3. By the definition of almost sure convergence, we want to show that:

$$\begin{aligned}\mathbb{P}(\{\omega \in \Omega : f(X_n(\omega)) \to f(X(\omega)) \text{ as } n \to \infty\}) &\geq \mathbb{P}(\{\omega \in \Omega : f(X_n) \to f(X) \text{ as } n \to \infty \text{ and } X \notin D_f\}) \\ &\geq \mathbb{P}(\{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty \text{ and } X \notin D_f\}).\end{aligned}$$

Since $\mathbb{P}(X \in D_f) = 0$ by assumption, and $X_n \xrightarrow{\text{a.s.}} X$,

$$\mathbb{P}(\{\omega \in \Omega : X_n \to X \text{ as } n \to \infty \text{ and } X \notin D_f\}) = \mathbb{P}(\{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}) = 1.$$

Therefore, $\mathbb{P}(\{\omega \in \Omega : f(X_n(\omega)) \to f(X(\omega)) \text{ as } n \to \infty\}) = 1$, and thus $f(X_n) \xrightarrow{\text{a.s.}} f(X)$.

□

8.2 Different Versions and Applications of the Weak Law of Large Numbers

Recall the Weak Law of Large Numbers (WLLN) for i.i.d. random variables with finite mean and variance, which states that the sample average converges in probability to the population mean.

Theorem 8.13. (**WLLN for i.i.d. Case**) Let $X_1, X_2, \cdots$ be i.i.d. random variables with $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mu.$$

There are many applications of the WLLN. We present one such application here.

Example 8.8. (**Bernstein Approximation**) Let $f : [0,1] \to \mathbb{R}$ be a continuous function. For each $n = 1, 2, \cdots$, define

$$f_n(x) = \sum_{m=0}^{n} \binom{n}{m} x^m (1-x)^{n-m} f\left(\frac{m}{n}\right). \qquad \text{(Bernstein Polynomial)}$$

Then,

$$\sup_{x\in[0,1]} |f_n(x) - f(x)| \to 0 \qquad \text{as } n \to \infty.$$

Proof.
For each fixed $x \in [0,1]$, let $Y_{n,x} \sim \mathrm{Bin}(n, x)$. Then,

$$\mathbb{P}(Y_{n,x} = m) = \binom{n}{m} x^m (1-x)^{n-m},$$
$$f_n(x) = \sum_{m=0}^{n} \mathbb{P}(Y_{n,x} = m) f\left(\frac{m}{n}\right) = \mathbb{E}\left(f\left(\frac{Y_{n,x}}{n}\right)\right).$$

By the WLLN and the Continuous Mapping Theorem, as $n \to \infty$,

$$\frac{Y_{n,x}}{n} \xrightarrow{\mathbb{P}} x \implies f\left(\frac{Y_{n,x}}{n}\right) \xrightarrow{\mathbb{P}} f(x).$$

Since f is continuous on the compact set $[0,1]$, it is uniformly continuous and bounded. For all $\varepsilon > 0$, there exists a $\delta_\varepsilon > 0$ such that for all $x, y \in [0,1]$,

$$|x - y| \le \delta_\varepsilon \implies |f(x) - f(y)| \le \varepsilon.$$

Let $M > 0$ be such that for all $x \in [0,1]$, $|f(x)| \le M$. Then,

$$\begin{aligned}
\left|\mathbb{E}\left(f\left(\frac{Y_{n,x}}{n}\right)\right) - f(x)\right| &\le \mathbb{E}\left|f\left(\frac{Y_{n,x}}{n}\right) - f(x)\right| \\
&= \mathbb{E}\left(\left|f\left(\frac{Y_{n,x}}{n}\right) - f(x)\right| \mathbf{1}_{\left|\frac{Y_{n,x}}{n} - x\right| \le \delta_\varepsilon}\right) + \mathbb{E}\left(\left|f\left(\frac{Y_{n,x}}{n}\right) - f(x)\right| \mathbf{1}_{\left|\frac{Y_{n,x}}{n} - x\right| > \delta_\varepsilon}\right) \\
&\le \varepsilon + 2M\mathbb{P}\left(\left|\frac{Y_{n,x} - nx}{n}\right| > \delta_\varepsilon\right). \\
\sup_{x\in[0,1]}\left|\mathbb{E}\left(f\left(\frac{Y_{n,x}}{n}\right)\right) - f(x)\right| &= \varepsilon + 2M \sup_{x\in[0,1]} \mathbb{P}\left(\left|\frac{Y_{n,x} - nx}{n}\right| > \delta_\varepsilon\right) \\
&\le \varepsilon + \frac{2M}{n\delta_\varepsilon} \sup_{x\in[0,1]} \mathbb{E}\left|Y_{n,x} - nx\right| && \text{(Markov's Inequality)} \\
&\le \varepsilon + \frac{2M}{n\delta_\varepsilon} \sup_{x\in[0,1]} \sqrt{\mathbb{E}(Y_{n,x} - nx)^2} && \text{(Lyapunov's Inequality)} \\
&\le \varepsilon + \frac{2M}{n\delta_\varepsilon} \sup_{x\in[0,1]} \sqrt{\mathrm{Var}(Y_{n,x})} = \varepsilon + \frac{2M}{n\delta_\varepsilon} \sup_{x\in[0,1]} \sqrt{nx(1-x)} && (\mathbb{E}Y_{n,x} = nx) \\
&\le \varepsilon + \frac{M}{\delta_\varepsilon\sqrt{n}}.
\end{aligned}$$

$$\limsup_{n\to\infty} \sup_{x\in[0,1]} \left|\mathbb{E}\left(f\left(\frac{Y_{n,x}}{n}\right)\right) - f(x)\right| \le \varepsilon \to 0.$$

Therefore, $\sup_{x\in[0,1]} |f_n(x) - f(x)| \to 0$ as $n \to \infty$. □

Example 8.9. (**Borel's Geometric Concentration**) Let μ_n be the uniform distribution on $[-1,1]^n$. Let $\mathcal{H} = \{\mathbf{x} \in [-1,1]^n : \langle \mathbf{x}, (1,\cdots,1)\rangle = 0\}$ be a hyperplane orthogonal to a principal diagonal in $\mathbb{R}^n$. For $r > 0$, let $\mathcal{H}_r = \{\mathbf{x} \in [-1,1]^n : \operatorname{dist}(\mathbf{x},\mathcal{H}) \leq r\}$ be the r-neighbourhood of $\mathcal{H}$. For all $\varepsilon > 0$,

$$\mu_n(\mathcal{H}_{\varepsilon\sqrt{n}}) \to 1 \qquad \text{as } n \to \infty.$$

Proof.
We can prove this by letting $X_1, X_2, \cdots \sim \mathrm{U}[-1,1]$ be i.i.d. random variables with $\mathbb{E}X_i = 0$. Let $\mathbf{X} = (X_1,\cdots,X_n)$. Then, $\mathbf{X} \sim \mu_n$. For all $B \subseteq [-1,1]^n$, $\mu_n(B) = \mathbb{P}(\mathbf{X} \in B)$. Therefore,

$$\begin{aligned}
\mu_n(\mathcal{H}_{\varepsilon\sqrt{n}}) &= \mathbb{P}(\mathbf{X} \in \mathcal{H}_{\varepsilon\sqrt{n}}) \\
&= \mathbb{P}(\operatorname{dist}(\mathbf{X},\mathcal{H}) \leq \varepsilon\sqrt{n}) \\
&= \mathbb{P}\left(\frac{|\langle X,(1,\cdots,1)\rangle|}{\|(1,\cdots,1)\|_2} \leq \varepsilon\sqrt{n}\right) \\
&= \mathbb{P}\left(\left|\frac{\sum_{i=1}^n X_i}{n}\right| \leq \varepsilon\right) \\
&= 1 - \mathbb{P}\left(\left|\frac{\sum_{i=1}^n X_i}{n}\right| > \varepsilon\right) \\
&\to 1. && \text{(WLLN)}
\end{aligned}$$

□

Another version of the WLLN is the L^2-WLLN, which states that the sample average converges in L^2 to the population mean under weaker conditions.

Theorem 8.14. (L^2**-WLLN**) Let $X_1, X_2, \cdots$ be uncorrelated random variables with $\mathbb{E}(X_i) = \mu$ and $\operatorname{Var}(X_i) \leq c < \infty$ for all i. Then:

$$\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{2} \mu.$$

Proof.

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu\right)^2 = \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^n X_i - \mathbb{E}\left(\sum_{i=1}^n X_i\right)\right)^2 = \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \operatorname{Var}(X_i) \leq \frac{c}{n} \to 0.$$

Therefore, $\frac{1}{n}\sum_{i=1}^n \xrightarrow{2} \mu$. □

Remark 8.14.1. By Theorem 8.8, convergence in r-th mean implies convergence in probability. Therefore, by the L^2-WLLN, we also have:

$$\left(\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{2} \mu\right) \implies \left(\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{\mathbb{P}} \mu\right).$$

Remark 8.14.2. In the L^2-WLLN, we do not require the random variables to be identically distributed. We only require them to be uncorrelated and have the same mean and bounded variance.

Remark 8.14.3. The L^2-WLLN can be generalized to the case where $\operatorname{Var}(X_i)$ are not uniformly bounded. If $\operatorname{Var}(X_i) < \infty$ for all i and $\frac{1}{n^2}\sum_{i=1}^n \operatorname{Var}(X_i) \to 0$, then we still have $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{2} \mu$.

A triangular array is a collection of random variables $\{X_{n,j}\}_{1\le j\le n<\infty}$ such that for each fixed n, $X_{n,1}, X_{n,2}, \cdots, X_{n,n}$ are independent random variables. The WLLN can be generalized to triangular arrays.

Theorem 8.15. (**WLLN for Triangular Array**) Let $\{X_{n,j}\}_{1\le j\le n<\infty}$ be a triangular array. Define $Y_n = \sum_{i=1}^n X_{n,i}$, $\mu_n = \mathbb{E}Y_n$, and $\sigma_n^2 = \operatorname{Var}(Y_n)$. Suppose that for some sequence b_n,

$$\frac{\sigma_n^2}{b_n^2} = \mathbb{E}\left(\frac{Y_n-\mu_n}{b_n}\right)^2 \to 0.$$

Then we have:

$$\frac{Y_n-\mu_n}{b_n} \xrightarrow{\mathbb{P}} 0.$$

Proof.

$$\mathbb{E}\left(\frac{Y_n-\mu_n}{b_n}\right)^2 = \frac{\operatorname{Var}(Y_n)}{b_n^2} \to 0.$$

Therefore, $\frac{Y_n-\mu_n}{b_n} \xrightarrow{2} 0$, and thus $\frac{Y_n-\mu_n}{b_n} \xrightarrow{\mathbb{P}} 0$. □

Remark 8.15.1. We should choose b_n to be no larger than $\mathbb{E}Y_n$ if possible.

Example 8.10. (**Coupon Collector's Problem**) There are n different types of coupons. Each time we pick a coupon, it is equally likely to be any one of the n types, independent of previous picks. Let τ_k^n be the number of picks needed to get k distinct types of coupons. We want to find the asymptotic behaviour of τ_n^n as $n\to\infty$. Let $X_{n,k}$ be the number of picks needed to get the k-th new type of coupon after we have already collected $k-1$ distinct types. Then, $X_{n,1}, X_{n,2}, \cdots, X_{n,n}$ are independent random variables.
Note that $X_{n,1}=1$ since we always get a new type of coupon on the first pick. Also, for $2\le k\le n$, we have:

$$\tau_n^n = \sum_{k=1}^n X_{n,k}.$$

For $2\le k\le n$ and $\ell = 1,2,\cdots$, we have:

$$\mathbb{P}(X_{n,k}=\ell) = \left(\frac{k-1}{n}\right)^{\ell-1}\left(1-\frac{k-1}{n}\right) \implies X_{n,k}\sim \operatorname{Geom}\left(1-\frac{k-1}{n}\right).$$

Therefore,

$$\mathbb{E}X_{n,k} = \left(1-\frac{k-1}{n}\right)^{-1}, \qquad \operatorname{Var}(X_{n,k}) = \left(1-\frac{k-1}{n}\right)^{-2} - \left(1-\frac{k-1}{n}\right)^{-1}.$$

We can see that:

$$\begin{aligned}
\mathbb{E}\tau_n^n &= \sum_{k=1}^n \mathbb{E}X_{n,k} = \sum_{k=1}^n \left(1-\frac{k-1}{n}\right)^{-1} = n\sum_{m=1}^n \frac{1}{m} \sim n\log n,\\
\operatorname{Var}(\tau_n^n) &= \sum_{k=1}^n \operatorname{Var}(X_{n,k}) = \sum_{k=1}^n \left(\left(1-\frac{k-1}{n}\right)^{-2} - \left(1-\frac{k-1}{n}\right)^{-1}\right)\\
&\le \sum_{k=1}^n \left(1-\frac{k-1}{n}\right)^{-2} = n^2\sum_{m=1}^n \frac{1}{m^2} \sim \frac{\pi^2}{6}n^2.
\end{aligned}$$

By using the WLLN for triangular array, let $b_n = n\log n$:

$$\frac{\operatorname{Var}(\tau_n^n)}{b_n^2} \le \frac{\frac{\pi^2}{6}n^2}{(n\log n)^2} = \frac{\pi^2}{6\log^2 n} \to 0.$$

Therefore, $\frac{\tau_n^n - \mathbb{E}\tau_n^n}{n\log n} \xrightarrow{\mathbb{P}} 0$, or equivalently,

$$\frac{\tau_n^n}{n\log n} \xrightarrow{\mathbb{P}} 1.$$

Example 8.11. (**Empty Bins Problem**) We throw r balls into n bins independently and uniformly at random. Let N_n be the number of empty bins after throwing the r balls. We want to find the asymptotic behaviour of N_n as $n \to \infty$ and $\frac{r}{n} \to c$ for some constant $c > 0$.
For $1 \leq i \leq n$, let A_i be the event that the i-th bin is empty after throwing the r balls. Then,

$$N_n = \sum_{i=1}^{n} \mathbf{1}_{A_i}.$$

Note that $\mathbf{1}_{A_i}$ are identically distributed but not independent. However, they are pairwise negatively correlated since for $i \neq j$,

$$\begin{aligned}
\frac{\mathbb{E}N_n}{n} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\mathbf{1}_{A_i}) \\
&= \mathbb{P}(A_1) \\
&= \left(1 - \frac{1}{n}\right)^r \\
&\sim e^{-c}, \\
\mathrm{Var}(N_n) &= \mathbb{E}(N_n - \mathbb{E}N_n)^2 \\
&= \mathbb{E}\left(\sum_{i=1}^{n}(\mathbf{1}_{A_i} - \mathbb{E}\mathbf{1}_{A_i})\right)^2 \\
&= \sum_{i=1}^{n} \mathbb{E}\left(\mathbf{1}_{A_i} - \mathbb{E}\mathbf{1}_{A_i}\right)^2 + 2\sum_{i \neq j} \mathrm{cov}(\mathbf{1}_{A_i}, \mathbf{1}_{A_j}) \\
&= \sum_{i=1}^{n} \mathrm{Var}(\mathbf{1}_{A_i}) + 2\sum_{i \neq j} \left(\mathbb{E}(\mathbf{1}_{A_i}\mathbf{1}_{A_j}) - \mathbb{E}\mathbf{1}_{A_i}\mathbb{E}\mathbf{1}_{A_j}\right) \\
&= n\mathbb{P}(A_1)(1 - \mathbb{P}(A_1)) + n(n-1)(\mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1)\mathbb{P}(A_2)) \\
&= n\left(1 - \frac{1}{n}\right)^r \left(1 - \left(1 - \frac{1}{n}\right)^r\right) + n(n-1)\left(1 - \frac{2}{n}\right)^r - n(n-1)\left(1 - \frac{1}{n}\right)^{2r} \\
&\sim n\left(e^{-c} - e^{-2c}\right) + n^2\left(e^{-2c} - e^{-2c}\right) = o(n^2).
\end{aligned}$$

By using the WLLN for triangular array, let $b_n = n$:

$$\frac{\mathrm{Var}(N_n)}{b_n^2} = \frac{\mathrm{Var}(N_n)}{n^2} \to 0.$$

Therefore, $\frac{N_n - \mathbb{E}N_n}{n} \xrightarrow{\mathbb{P}} 0$, or equivalently,

$$\frac{N_n}{n} \xrightarrow{\mathbb{P}} e^{-c}.$$

8.3 Borel-Cantelli Lemmas

Let $A_1, A_2, \cdots$ 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 8.16. (**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:
$$\mathbb{P}\left(\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} A_n^{\complement}\right) = \lim_{m\to\infty} \mathbb{P}\left(\bigcap_{n=m}^{\infty} A_n^{\complement}\right) = \lim_{m\to\infty}\lim_{r\to\infty} \mathbb{P}\left(\bigcap_{n=m}^{r} A_n^{\complement}\right) = \lim_{m\to\infty}\lim_{r\to\infty} \prod_{n=m}^{r} \mathbb{P}(A_n^{\complement}) = \lim_{m\to\infty} \prod_{n=m}^{\infty} (1 - \mathbb{P}(A_n)),$$
$$\le \lim_{m\to\infty} \prod_{n=m}^{\infty} e^{-\mathbb{P}(A_n)} = \lim_{m\to\infty} \exp\left(-\sum_{n=m}^{\infty} \mathbb{P}(A_n)\right) = 0, \qquad (1 - x \le e^{-x} \text{ if } x \ge 0)$$
$$\mathbb{P}(A_n \text{ i.o.}) = \mathbb{P}\left(\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty} A_n\right) = 1 - \mathbb{P}\left(\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} A_n^{\complement}\right) = 1.$$

□

Remark 8.16.1. BCII can be considered a partial converse of BCI.

Remark 8.16.2. ”i.o.” stands for ”infinitely often,” while ”f.o.” stands for ”finitely often.”

We will now explore how the Borel-Cantelli Lemmas can be applied in various scenarios.

Example 8.12. (**Infinite Monkey Problem**) Suppose there are N letters in the alphabet. A monkey is hitting keys on a typewriter at random, with each letter being equally likely to be any one of the N letters, independent of previous hits. Given a specific string S of length m, such as ”monkey,” we want to determine the probability that the monkey will type the string S infinitely many times.

Let E_k be the event that the monkey types the string S starting from the k-th letter. Then, $E_1, E_2, \cdots$ are not independent since if the monkey types S starting from the k-th letter, it will continue to type S for the next $m-1$ letters. However, $E_1, E_{m+1}, E_{2m+1}, \cdots$ are independent. Note that $\mathbb{P}(E_k) = N^{-m}$ for all k.

Since $\sum_{k=0}^{\infty} \mathbb{P}(E_{mk+1}) = \sum_{k=0}^{\infty} N^{-m} = \infty$, by BCII:

$$\mathbb{P}(E_{mk+1} \text{ i.o.}) = 1.$$

Therefore, the monkey will type the string S infinitely many times with probability 1.

Recall that if $X_n \xrightarrow{\mathbb{P}} X$, there exists a non-random increasing sequence of integers $n_1, n_2, \cdots$ such that $X_{n_i} \xrightarrow{\text{a.s.}} X$ as $i \to \infty$.
We now present a theorem that provides a necessary and sufficient condition for convergence in probability.

Theorem 8.17. $X_n \xrightarrow{\mathbb{P}} X$ 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.

($\Longrightarrow$) For all $\varepsilon > 0$, let $\varepsilon_k = \min\{\varepsilon, 1/k\}$. Since $X_n \xrightarrow{\mathbb{P}} X$, for all k, there exists $n(m_k)$ such that:

$$\mathbb{P}(\left|X_{n(m_k)} - X\right| > \varepsilon_k) < \frac{1}{2^k}.$$

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 $\varepsilon > 0$, there exists k such that $\varepsilon_k \leq \varepsilon$. We have:

$$\{\left|X_{n(m_k)} - X\right| > \varepsilon\} \subseteq \{\left|X_{n(m_k)} - X\right| > \varepsilon_k\}.$$

Therefore,

$$\mathbb{P}(\left|X_{n(m_k)} - X\right| > \varepsilon \text{ i.o.}) = 0 \implies \mathbb{P}(\left|X_{n(m_k)} - X\right| \leq \varepsilon \text{ f.o.}) = 1.$$

($\Longleftarrow$) Let $a_n = \mathbb{P}(|X_n - X| > \varepsilon)$ for some $\varepsilon > 0$. Assume that $a_n \not\to 0$.
There exists a subsequence $a_{n(m)}$ such that $a_{n(m)} \to a > 0$.
This implies that for some $\delta > 0$ and all m sufficiently large, $a_{n(m)} > \delta$ and $\mathbb{P}(\left|X_{n(m)} - X\right| > \varepsilon) > \delta$.
Therefore, there does not exist a further subsequence $X_{n(m_k)}$ such that $X_{n(m_k)} \xrightarrow{\text{a.s.}} X$. This contradicts our assumption. Hence, $a_n \to 0$. Note that:

$$a_n = \mathbb{P}(|X_n - X| > \varepsilon) \leq 1.$$

Since $a_n \to 0$, we have $X_n \xrightarrow{\mathbb{P}} X$.

□

We now present a theorem with conditions similar to the Law of Large Numbers. However, note that $\mathbb{E}\,|X_1| = \infty$ in this case.

Theorem 8.18. If $X_1, X_2, \cdots$ are i.i.d. random variables with $\mathbb{E}\,|X_i| = \infty$, then:

$$\mathbb{P}(|X_n| \geq n \text{ i.o.}) = 1, \qquad \mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \text{ exists in } (-\infty, \infty)\right) = 0.$$

Proof.

$$\mathbb{E}\,|X_1| = \int_0^\infty \mathbb{P}(|X_1| > t)\,dt \leq \sum_{n=0}^{\infty} \mathbb{P}(|X_1| > n).$$

Since $\mathbb{E}\,|X_1| = \infty$, we have $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| \geq n) = \infty$. By BCII, $\mathbb{P}(|X_n| \geq n \text{ i.o.}) = 1$.
Let $Y_n = \sum_{i=1}^{n} X_i$ and $C = \{\omega \in \Omega : \frac{Y_n(\omega)}{n} \text{ exists in } (-\infty, \infty)\}$. Assume that $\mathbb{P}(C) > 0$.
For all $\omega \in C$, we have:

$$\frac{Y_n(\omega)}{n} - \frac{Y_{n+1}(\omega)}{n+1} = \frac{Y_n(\omega)}{n(n+1)} - \frac{X_{n+1}(\omega)}{n+1}.$$

Since $\frac{Y_n}{n}$ converges, we have $\frac{Y_n(\omega)}{n(n+1)} \to 0$. Therefore, $\frac{X_{n+1}(\omega)}{n+1} \to 0$, or equivalently, $X_{n+1}(\omega) = o(n+1)$.
This implies that there exists $N(\omega)$ such that for all $n \geq N(\omega)$, $|X_n(\omega)| < n$. Therefore, $\omega \notin \{|X_n| \geq n \text{ i.o.}\}$.
We have shown that $C \subseteq \{|X_n| < n \text{ f.o.}\}$. Since $\mathbb{P}(|X_n| \geq n \text{ i.o.}) = 1$, we have $\mathbb{P}(C) = 0$. □

We now present a theorem that weakens the independence condition in BCII to pairwise independence.

Theorem 8.19. 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 $Y_n = \sum_{m=1}^{n} \mathbf{1}_{A_m}$. Note that $\mathbb{E}Y_n = \sum_{m=1}^{n} \mathbb{P}(A_m) \to \infty$ as $n \to \infty$.
Also, $\text{Var}(Y_n) = \sum_{m=1}^{n} \text{Var}(\mathbf{1}_{A_m}) = \sum_{m=1}^{n} \mathbb{P}(A_m)(1 - \mathbb{P}(A_m)) \leq \sum_{m=1}^{n} \mathbb{P}(A_m)$.
By Markov's inequality, for all $\varepsilon > 0$:

$$\mathbb{P}\left(\left|\frac{Y_n - \mathbb{E}Y_n}{\mathbb{E}Y_n}\right| > \varepsilon\right) \leq \frac{\text{Var}(Y_n)}{\varepsilon^2(\mathbb{E}Y_n)^2} \leq \frac{1}{\varepsilon^2 \mathbb{E}Y_n} \to 0.$$

Therefore, we get that $\frac{Y_n - \mathbb{E}Y_n}{\mathbb{E}Y_n} \xrightarrow{\mathbb{P}} 0$.
Let $n_k = \min\{n : \mathbb{E}Y_n \geq k^2\}$. Note that n_k is an increasing sequence of integers.
Since $\mathbb{E}Y_n \to \infty$ as $n \to \infty$, we have $n_k \to \infty$ as $k \to \infty$. Also, $\mathbb{E}Y_{n_k} \geq k^2$ and $\mathbb{E}Y_{n_k - 1} < k^2$.
By Markov's inequality, for all $\varepsilon > 0$:

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\left|\frac{Y_{n_k} - \mathbb{E}Y_{n_k}}{\mathbb{E}Y_{n_k}}\right| > \varepsilon\right) \leq \sum_{k=1}^{\infty} \frac{1}{\varepsilon^2 \mathbb{E}Y_{n_k}} \leq \sum_{k=1}^{\infty} \frac{1}{\varepsilon^2 k^2} < \infty.$$

By BCI, we have:

$$\mathbb{P}\left(\left|\frac{Y_{n_k} - \mathbb{E}Y_{n_k}}{\mathbb{E}Y_{n_k}}\right| > \varepsilon \text{ i.o.}\right) = 0, \qquad \mathbb{P}\left(\left|\frac{Y_{n_k} - \mathbb{E}Y_{n_k}}{\mathbb{E}Y_{n_k}}\right| \leq \varepsilon \text{ f.o.}\right) = 1.$$

Let $C = \left\{\omega \in \Omega : \frac{Y_{n_k}(\omega)}{\mathbb{E}Y_{n_k}} \to 1 \text{ as } k \to \infty\right\}$. Then, $\mathbb{P}(C) = 1$. For any $\omega \in C$ and $n_k \leq n < n_{k+1}$, we have:

$$\frac{Y_{n_k}(\omega)}{\mathbb{E}Y_{n_k}} \cdot \frac{\mathbb{E}Y_{n_k}}{\mathbb{E}Y_{n_k+1}} \leq \frac{Y_n(\omega)}{\mathbb{E}Y_n} \leq \frac{Y_{n_{k+1}}(\omega)}{\mathbb{E}Y_{n_{k+1}}} \cdot \frac{\mathbb{E}Y_{n_{k+1}}}{\mathbb{E}Y_{n_k+1}}.$$

Since $\frac{\mathbb{E}Y_{n_k}}{\mathbb{E}Y_{n_k+1}} \to 1$ and $\frac{\mathbb{E}Y_{n_{k+1}}}{\mathbb{E}Y_{n_k+1}} \to 1$ as $k \to \infty$, we have:

$$\frac{Y_n(\omega)}{\mathbb{E}Y_n} \to 1 \text{ as } n \to \infty.$$

Therefore, $C \subseteq \left\{\omega \in \Omega : \frac{Y_n(\omega)}{\mathbb{E}Y_n} \to 1 \text{ as } n \to \infty\right\}$. We have:

$$\mathbb{P}\left(\frac{Y_n}{\mathbb{E}Y_n} \to 1 \text{ as } n \to \infty\right) = 1 \Longleftrightarrow \frac{\sum_{m=1}^{n} \mathbf{1}_{A_m}}{\sum_{m=1}^{n} \mathbb{P}(A_m)} \xrightarrow{\text{a.s.}} 1.$$

□

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. We now present a more general version of this result.

Theorem 8.20. (**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 of 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 $A = \limsup_n A_n = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n$.
Let $B_m = \bigcup_{n=m}^{\infty} A_n$. Since $B_1 \supseteq B_2 \supseteq \cdots$, we have $A = \lim_{m \to \infty} B_m$.
For all k, $B_k \in \sigma(A_k, A_{k+1}, \cdots)$, which is independent of $\sigma(A_1, A_2, \cdots, A_{k-1})$.
Therefore, B_k is independent of any $A_i \in \sigma(A_1, A_2, \cdots, A_{k-1})$.
Setting $k \to \infty$, we have that $B_k \to A$ is independent of all elements in $\mathcal{A}$, which also includes itself.
Therefore, $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A)$, which implies that $\mathbb{P}(A) = 0$ or 1. □

We now explore more about sequences of random variables. Let $X_1, X_2, \cdots$ be independent random variables. For any subcollection $X_{i_1}, X_{i_2}, \cdots, X_{i_k}$, we can write $\sigma(X_{i_1}, X_{i_2}, \cdots, X_{i_k})$ as the smallest σ-field with respect to which each X_{i_j} is measurable.

Definition 8.21. Let $\mathcal{H}_n = \sigma(X_{n+1}, X_{n+2}, \cdots)$. We have $\mathcal{H}_n \supseteq \mathcal{H}_{n+1} \supseteq \cdots$. The **tail σ-field** is defined as:

$$\mathcal{H}_\infty = \bigcap_n \mathcal{H}_n.$$

An event $E \in \mathcal{F}$ is called a **tail event** if $E \in \mathcal{H}_\infty$.

Remark 8.21.1. $E \in \mathcal{H}_\infty$ means that E is independent of any finite collection of X_i's.

Example 8.13. $\{\limsup_{n\to\infty} X_n = \infty\}$ and $\{\liminf_{n\to\infty} X_n = -\infty\}$ are tail events.

Example 8.14. $\{\sum_n X_n \text{ converges}\}$ is a tail event.

Example 8.15. $\{\sum_n X_n \text{ converges to } 1\}$ is not a tail event.

We now present Kolmogorov's Zero-One Law, which is a special case of the Borel Zero-One Law.

Theorem 8.22. (**Kolmogorov's Zero-One Law**) If $H \in \mathcal{H}_\infty$, then $\mathbb{P}(H) = 0$ or 1.

Proof.
Since $H \in \mathcal{H}_\infty$, H is independent of any finite collection of X_i's. By the Borel Zero-One Law, $\mathbb{P}(H) = 0$ or 1. □

We now define a tail function.

Definition 8.23. We define a **tail function** to be $Y : \Omega \to \mathbb{R} \cup \{-\infty, \infty\}$, such that Y is a measurable function with respect to the tail σ-field $\mathcal{H}_\infty$. In other words, for all $y \in \mathbb{R} \cup \{-\infty, \infty\}$, $\{Y \leq y\} \in \mathcal{H}_\infty$.

Remark 8.23.1. Y is a function of the "tail" of the sequence $X_1, X_2, \cdots$.

Remark 8.23.2. Y is a tail function if and only if Y is independent of any finite collection of X_i's.

Example 8.16. Let $Y = \limsup_{n\to\infty} X_n$. Then, Y is a tail function.
For any $y \in \mathbb{R}$:

$$\{Y \leq y\} = \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{X_n \leq y\} \in \mathcal{H}_\infty.$$

Therefore, $\{Y \leq y\}$ is a tail event.

Theorem 8.24. If Y is a tail function of independent random variables $X_1, X_2, \cdots$, then there exists $-\infty \leq k \leq \infty$ such that:

$$\mathbb{P}(Y = k) = 1.$$

Proof.
For all $y \in \mathbb{R} \cup \{-\infty, \infty\}$, $\{Y \leq y\} \in \mathcal{H}_\infty$. By Kolmogorov's Zero-One Law, $\mathbb{P}(Y \leq y) = 0$ or 1.
Let $k = \inf\{y : \mathbb{P}(Y \leq y) = 1\}$. We have:

$$\mathbb{P}(Y < k) = \mathbb{P}\left(\bigcup_{m=1}^{\infty} \left\{Y \leq k - \frac{1}{m}\right\}\right) = \lim_{m\to\infty} \mathbb{P}\left(Y \leq k - \frac{1}{m}\right) = 0, \qquad \mathbb{P}(Y > k) = 1 - \mathbb{P}(Y \leq k) = 0.$$

Therefore, $\mathbb{P}(Y = k) = 1$. □

We now apply the above theorem to the sequence of independent and identically distributed (i.i.d.) random variables $X_1, X_2, \cdots$. Recall that if $X_1, X_2, \cdots$ are i.i.d. with $\mathbb{E}\,|X_1| < \infty$, then:

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}X_1\right) = 1.$$

If $\mathbb{E}\,|X_1| = \infty$, then:

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \text{ exists}\right) = 0.$$

We now present a theorem that does not require the random variables to be identically distributed.

Theorem 8.25. If $X_1, X_2, \cdots$ are independent random variables, then:

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \text{ exists}\right) = 0 \text{ or } 1.$$

Proof.
Let:

$$Z_1(\omega) = \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i(\omega), \qquad Z_2(\omega) = \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i(\omega).$$

Note that $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i(\omega)$ exists if and only if $Z_1(\omega) = Z_2(\omega)$. For all k, we have:

$$Z_1(\omega) = \limsup_{n\to\infty} \frac{1}{n}\left(\sum_{i=1}^{k} X_i(\omega) + \sum_{i=k+1}^{n} X_i(\omega)\right) = \limsup_{n\to\infty} \frac{1}{n}\sum_{i=k+1}^{n} X_i(\omega),$$

$$Z_2(\omega) = \liminf_{n\to\infty} \frac{1}{n}\left(\sum_{i=1}^{k} X_i(\omega) + \sum_{i=k+1}^{n} X_i(\omega)\right) = \liminf_{n\to\infty} \frac{1}{n}\sum_{i=k+1}^{n} X_i(\omega).$$

Therefore, Z_1 and Z_2 do not depend on any finite collection of X_i's.
This implies that Z_1 and Z_2 are tail functions of X_i's. By Theorem 8.24, there exist $-\infty \le c_1, c_2 \le \infty$ such that:

$$\mathbb{P}(Z_1 = c_1) = 1, \qquad \mathbb{P}(Z_2 = c_2) = 1.$$

Therefore:

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \text{ exists}\right) = \mathbb{P}(Z_1 = Z_2) = \begin{cases} 1, & c_1 = c_2, \\ 0, & c_1 \neq c_2. \end{cases}$$

□

Example 8.17. (**Random power series**) Let $X_1, X_2, \cdots$ be i.i.d. exponential random variables with parameter $\lambda = 1$. Consider the random power series:

$$\sum_{n=1}^{\infty} X_n(\omega) z^n.$$

For each $\omega \in \Omega$, the radius of convergence is given by:

$$R(\omega) = \frac{1}{\limsup_{n\to\infty} |X_n(\omega)|^{\frac{1}{n}}}.$$

Note that R is a tail function of $X_1, X_2, \cdots$. By Theorem 8.24, there exists C such that $\mathbb{P}(R = C) = 1$ ($R = C$ almost surely). We will show that $C = 1$. Note that:

$$\mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} = \frac{1}{C}\right) = 1.$$

For all $\varepsilon > 0$, we have:

$$\mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} \leq 1 + \varepsilon\right) = 1, \qquad \mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} \geq 1 - \varepsilon\right) = 1.$$

We first prove the first equation. Note that:

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|X_n|^{\frac{1}{n}} > 1 + \varepsilon\right) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > (1+\varepsilon)^n) = \sum_{n=1}^{\infty} e^{-(1+\varepsilon)^n} < \infty.$$

Therefore, by BCI:

$$\mathbb{P}(|X_n|^{\frac{1}{n}} > 1 + \varepsilon \text{ i.o.}) = 0 \implies \mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} \leq 1 + \varepsilon\right) = 1.$$

We now prove the second equation. Note that:

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|X_n|^{\frac{1}{n}} > 1 - \varepsilon\right) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > (1-\varepsilon)^n) = \sum_{n=1}^{\infty} e^{-(1-\varepsilon)^n} = \infty.$$

Therefore, by BCII:

$$\mathbb{P}(|X_n|^{\frac{1}{n}} > 1 - \varepsilon \text{ i.o.}) = 1 \implies \mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} \geq 1 - \varepsilon\right) = 1.$$

Since the above two equations hold for all $\varepsilon > 0$, we have:

$$\mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} = 1\right) = 1.$$

Therefore, $C = 1$.

8.4 Strong Law of Large Numbers

We now present the Law of Large Numbers. The Weak Law of Large Numbers (WLLN) states that the sample average converges in probability to the expected value. Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables with $\mathbb{E}(X_1) = \mu$. As $n \to \infty$:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{D} \mu, \qquad \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mu.$$

By its name, the Weak Law of Large Numbers (WLLN) indeed has a stronger version, called the Strong Law of Large Numbers. We prove one of the versions of SLLN.

Theorem 8.26. (**Strong Law of Large Numbers**) [SLLN] Let $X_1, X_2, \cdots$ be i.i.d. random variables with $\mathbb{E}X_1 = \mu$ and $\mathbb{E}|X_1| < \infty$. We have:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu.$$

Note that the proof for SLLN is more complicated than that for WLLN. We will not prove this version of SLLN here. Instead, we will prove two simpler versions of SLLN with stronger conditions. We first prove SLLN with the fourth moment condition.

Theorem 8.27. (**SLLN with $\mathbb{E}X_i^4 < \infty$**) Let $X_1, X_2, \cdots$ be i.i.d. random variables with $\mathbb{E}X_1 = 0$ and $\mathbb{E}X_1^4 < \infty$. We have:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} 0.$$

Proof.
Let $Y_n = \sum_{i=1}^{n} X_i$. We will use the fourth moment condition to show that $\frac{Y_n}{n} \xrightarrow{\text{a.s.}} 0$. Note that:

$$\mathbb{E}Y_n^4 = \sum_{i,j,k,\ell=1}^{n} \mathbb{E}(X_i X_j X_k X_\ell).$$

Since X_i are i.i.d. with $\mathbb{E}X_i = 0$, if any of i, j, k, ℓ appears only once, then $\mathbb{E}(X_i X_j X_k X_\ell) = 0$.
Therefore, the only non-zero terms in the above summation are those with pairs of indices. There are two cases:

1. $i = j, k = \ell$ or $i = k, j = \ell$ or $i = \ell, j = k$ with $i \neq k$. There are $3n(n-1)$ such terms, each equal to $\mathbb{E}X_1^2\mathbb{E}X_1^2$.
2. $i = j = k = \ell$. There are n such terms, each equal to $\mathbb{E}X_1^4$.

Therefore,

$$\mathbb{E}Y_n^4 = 3n(n-1)(\mathbb{E}X_1^2)^2 + n\mathbb{E}X_1^4 = O(n^2).$$

By Markov's inequality, for all $\varepsilon > 0$:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq \varepsilon\right) \leq \frac{1}{n^4\varepsilon^4}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right)^4 = O\left(\frac{1}{n^2}\right).$$

Therefore, for all $\varepsilon > 0$:

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq \varepsilon\right) < \infty.$$

By BCI, for all $\varepsilon > 0$:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq \varepsilon \text{ i.o.}\right) = 0.$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} 0.$$

□

Theorem 8.28. **(SLLN with $\mathbb{E}X_1^2 < \infty$)** Let $X_1, X_2, \cdots$ be i.i.d. random variables with $\mathbb{E}X_1^2 < \infty$ and $\mathbb{E}X_i = \mu$. We have:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{2} \mu, \qquad \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu.$$

Proof.
Let $Y_n = \sum_{i=1}^n X_i$. We first prove the convergence in mean square. Note that:

$$\mathbb{E}\left(\frac{Y_n}{n} - \mu\right)^2 = \frac{\mathbb{E}(Y_n - n\mu)^2}{n^2} = \frac{\operatorname{Var}(Y_n)}{n^2} = \frac{\operatorname{Var}(X_1)}{n}.$$

Therefore, as $n \to \infty$:

$$\frac{Y_n}{n} \xrightarrow{2} \mu.$$

We now prove the almost sure convergence. Since convergence in mean square implies convergence in probability, by Theorem 8.10, there exists a subsequence $n_1 < n_2 < \cdots$ such that $\frac{Y_{n_i}}{n_i} \xrightarrow{\text{a.s.}} \mu$.
For simplicity, we choose $n_i = i^2$. By Markov's inequality, for all $\varepsilon > 0$:

$$\sum_{i=1}^{\infty} \mathbb{P}\left(\left|\frac{Y_{i^2}}{i^2} - \mu\right| > \varepsilon\right) \le \sum_{i=1}^{\infty} \frac{\operatorname{Var}(X_1)}{i^2\varepsilon^2} < \infty.$$

By BCI, for all $\varepsilon > 0$:

$$\mathbb{P}\left(\left|\frac{Y_{i^2}}{i^2} - \mu\right| > \varepsilon \text{ i.o.}\right) = 0.$$

Therefore, $\frac{Y_{i^2}}{i^2} \xrightarrow{\text{a.s.}} \mu$. We need to show that $\frac{Y_n}{n} \xrightarrow{\text{a.s.}} \mu$. Note that for any n, there exists i such that $i^2 \le n \le (i+1)^2$, and thus when X_i are non-negative:

$$\frac{Y_{i^2}}{(i+1)^2} \le \frac{Y_n}{n} \le \frac{Y_{(i+1)^2}}{i^2}.$$

Since $\frac{Y_{i^2}}{i^2} \xrightarrow{\text{a.s.}} \mu$, we have:

$$\frac{Y_{i^2}}{(i+1)^2} = \frac{Y_{i^2}}{i^2} \cdot \frac{i^2}{(i+1)^2} \xrightarrow{\text{a.s.}} \mu, \qquad \frac{Y_{(i+1)^2}}{i^2} = \frac{Y_{(i+1)^2}}{(i+1)^2} \cdot \frac{(i+1)^2}{i^2} \xrightarrow{\text{a.s.}} \mu.$$

By the Squeeze Theorem, we conclude that whenever X_i are non-negative, as $n \to \infty$:

$$\frac{Y_n}{n} \xrightarrow{\text{a.s.}} \mu.$$

For general X_i, we can write $X_n = X_n^+ - X_n^-$ where:

$$X_n^+(\omega) = \max\{X_n(\omega), 0\}, \qquad X_n^-(\omega) = -\min\{X_n(\omega), 0\}.$$

Both X_n^+ and X_n^- are non-negative random variables.
Furthermore, $X_1^+, X_2^+, \cdots$ are i.i.d. with $\mathbb{E}X_1^+ < \infty$, and $X_1^-, X_2^-, \cdots$ are i.i.d. with $\mathbb{E}X_1^- < \infty$. Note that:

$$\mathbb{E}(X_1^+)^2 \le \mathbb{E}(X_1^+)^2 + 2\mathbb{E}X_1^+\mathbb{E}X_1^- + \mathbb{E}(X_1^-)^2 = \mathbb{E}(X_1^+ - X_1^-)^2 = \mathbb{E}X_1^2 < \infty.$$

Similarly, $\mathbb{E}(X_1^-)^2 < \infty$. By the above result for non-negative random variables, we have:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^+ \xrightarrow{\text{a.s.}} \mathbb{E}X_1^+, \quad \frac{1}{n}\sum_{i=1}^{n} X_i^- \xrightarrow{\text{a.s.}} \mathbb{E}X_1^-.$$

Therefore, as $n \to \infty$:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} X_i^+ - \frac{1}{n}\sum_{i=1}^{n} X_i^- \xrightarrow{\text{a.s.}} \mathbb{E}X_1^+ - \mathbb{E}X_1^- = \mathbb{E}X_1 = \mu.$$

□

Why do we need SLLN? There are many applications that specifically require SLLN.

Example 8.18. (**Renewal Theory**) Consider a bulb that has a random lifetime. When the bulb burns out, it is immediately replaced by a new identical bulb with another random lifetime. This process continues indefinitely. Let $X_1, X_2, \cdots$ be the lifetimes of the bulbs. Let $T_n = X_1 + X_2 + \cdots + X_n$ be the time that the n-th bulb burns out. Let $N_t = \max\{n : T_n \leq t\}$ be the number of bulbs that have burned out by time t. Note that

$$T_{N_t} \leq t < T_{N_t+1}.$$

We want to find the average lifetime of the bulbs, which is given by $\frac{t}{N_t}$.
We assume that $X_1, X_2, \cdots$ are i.i.d. random variables with $0 < X_i < \infty$ and $\mathbb{E}X_1 < \infty$. Let $\mathbb{E}X_1 = \mu$. As $t \to \infty$:

$$\frac{t}{N_t} \xrightarrow{\text{a.s.}} \mu.$$

Proof.
Since $X_i > 0$, T_n is an increasing sequence. Therefore, N_t is also an increasing sequence. Note that:

$$\frac{T_{N_t}}{N_t} \leq \frac{t}{N_t} < \frac{T_{N_t+1}}{N_t+1} \cdot \frac{N_t+1}{N_t}.$$

By SLLN, as $n \to \infty$:

$$\frac{T_n}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n} \xrightarrow{\text{a.s.}} \mu.$$

Therefore, as $t \to \infty$:

$$\frac{T_{N_t}}{N_t} \xrightarrow{\text{a.s.}} \mu, \qquad \frac{T_{N_t+1}}{N_t+1} \xrightarrow{\text{a.s.}} \mu.$$

Since $T_{N_t} \leq t < T_{N_t+1}$, as $t \to \infty$, $N_t \to \infty$ almost surely. Therefore, $\frac{N_t+1}{N_t} \xrightarrow{\text{a.s.}} 1$.
By the Squeeze Theorem, as $t \to \infty$:

$$\frac{t}{N_t} \xrightarrow{\text{a.s.}} \mu.$$

□

Remark 8.28.1. If $X_n \xrightarrow{\mathbb{P}} X_\infty$, then $N_m \xrightarrow{\text{a.s.}} \infty$ as $m \to \infty$. To prove this, it is not necessary that $X_{N_m} \xrightarrow{\text{a.s.}} X_\infty$ or $X_{N_m} \xrightarrow{\mathbb{P}} X_\infty$.

Example 8.19. Recall the Example 8.4. Let $\Omega = [0, 1]$. For each $m \geq 1$, we have a partition of Ω into m subintervals:

$$Y_{m,k} = \mathbf{1}_{I_{m,k}} = \begin{cases} 1, & \omega \in \left[\frac{k-1}{m}, \frac{k}{m}\right], \\ 0, & \text{Otherwise.} \end{cases}$$

Note that for each fixed m, $Y_{m,1}, Y_{m,2}, \cdots, Y_{m,m}$ are mutually exclusive and exhaustive.
We define a sequence of random variables $X_1, X_2, \cdots$ by arranging $Y_{m,k}$ in a triangular array, i.e.:

$$\begin{array}{lll} X_1 = Y_{1,1}, & & \\ X_2 = Y_{2,1}, & X_3 = Y_{2,2}, & \\ X_4 = Y_{3,1}, & X_5 = Y_{3,2}, & X_6 = Y_{3,3}, \\ \vdots & \vdots & \vdots \end{array}$$

We have that for each fixed m, $Y_{m,1}, Y_{m,2}, \cdots, Y_{m,m}$ are mutually independent.
Therefore, $X_1, X_2, \cdots$ are independent. Based on the example, let $X_\infty = 0$, we have that $X_n \xrightarrow{\mathbb{P}} 0$.
For each $\omega \in [0, 1]$, there exists a unique $k_m(\omega)$ such that $Y_{m,k_m(\omega)}(\omega) = 1$. Let $N_m(\omega) = \frac{(m-1)m}{2} + k_m(\omega)$.
Therefore, $N_m \xrightarrow{\text{a.s.}} \infty$ as $m \to \infty$. However, for all $\omega \in [0, 1]$ and $m \geq 1$, $X_{N_m(\omega)}(\omega) = 1$.
Therefore, $X_{N_m} \not\to X_\infty$ almost surely or in probability.

We now present another important application of SLLN, the Glivenko-Cantelli Theorem. This theorem states that the empirical distribution function converges uniformly to the true distribution function almost surely. It is also called the Fundamental Theorem of Statistics.

Theorem 8.29. (**Glivenko-Cantelli Theorem**) Let $X \sim F$, where F is an unknown distribution function. Let $X_1, X_2, \cdots, X_N$ be i.i.d. random variables with the same distribution as X.
Let $F_N(x)$ be the empirical distribution function based on $X_1, X_2, \cdots, X_N$, which is also a distribution function of a histogram with N samples.

$$F_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{X_i \le x}, \qquad\qquad F(x) = \mathbb{P}(X \le x).$$

We have:

$$\sup_x |F_N(x) - F(x)| \xrightarrow{\text{a.s.}} 0.$$

Proof.
We only prove the case when $F(x)$ is continuous.
For any $m \ge 1$, there exist $x_0 = -\infty < x_1 < x_2 < \cdots < x_m = +\infty$ such that for all $1 \le i \le m$:

$$F(x_i) - F(x_{i-1}) = \frac{1}{m}.$$

Note that such x_i exist because $F(x)$ is continuous and non-decreasing with $0 \le F(x) \le 1$.
For any $x \in \mathbb{R}$, there exists i such that $x_{i-1} \le x \le x_i$. Since $F_N(x)$ and $F(x)$ are non-decreasing, we have:

$$\begin{aligned} F_N(x) - F(x) &\le F_N(x_i) - F(x_{i-1}) = F_N(x_i) - F(x_i) + \frac{1}{m}, \\ F_N(x) - F(x) &\ge F_N(x_{i-1}) - F(x_i) = F_N(x_{i-1}) - F(x_{i-1}) - \frac{1}{m}. \end{aligned}$$

Therefore,

$$|F_N(x) - F(x)| \le \sup_i |F_N(x_i) - F(x_i)| + \frac{1}{m} \implies \sup_x |F_N(x) - F(x)| \le \sup_i |F_N(x_i) - F(x_i)| + \frac{1}{m}.$$

By SLLN, for each fixed x:

$$F_N(x;\omega) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{X_i(\omega) \le x} \xrightarrow{\text{a.s.}} F(x), \qquad\qquad \mathbb{P}(F_N(x) \to F(x) \text{ as } N \to \infty) = 1.$$

Let $C_x = \{\omega \in \Omega : F_N(x;\omega) \to F(x) \text{ as } N \to \infty\}$. For each fixed m, if $\omega \in \bigcap_{i=1}^m C_{x_i}$:

$$\begin{aligned} \sup_i |F_N(x_i;\omega) - F(x_i)| &\to 0 \text{ as } N \to \infty, \\ \limsup_N \sup_x |F_N(x;\omega) - F(x)| &\le \frac{1}{m}. \end{aligned}$$

Since the above holds for all $m \ge 1$:

$$\limsup_N \sup_x |F_N(x;\omega) - F(x)| = 0 \implies \sup_x |F_N(x;\omega) - F(x)| \to 0 \text{ as } N \to \infty.$$

Note that $\mathbb{P}(C_{x_i}) = 1$ for all $1 \le i \le m$. Therefore:

$$\mathbb{P}\left(\bigcap_{i=1}^{m} C_{x_i}\right) = 1 \implies \mathbb{P}\left(\sup_x |F_N(x) - F(x)| \to 0 \text{ as } N \to \infty\right) = 1.$$

□

Of course, there are still many examples that we haven't explored (including some mentioned during the lectures that I am too lazy to include here). We also skipped many 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**) A particle starts at position $a \in \mathbb{Z}$. At each time step, it moves one unit to the right with probability p and one unit to the left with probability $q = 1 - p$. The moves are independent of each other. Let Y_n be the position of the particle at time n. We have:

$$Y_n = a + \sum_{i=1}^{n} X_i,$$

where X_i are i.i.d. random variables with:

$$\mathbb{P}(X_i = 1) = p, \quad \mathbb{P}(X_i = -1) = q.$$

Such a process $\{Y_n, n \geq 0\}$ is called a **simple random walk** on $\mathbb{Z}$. If $p = q = \frac{1}{2}$, the random walk is called a **symmetric simple random walk**.

Lemma A.1. The simple random walk $\{Y_n, n \geq 0\}$ has the following properties:

1. It is **spatially homogeneous**: $\mathbb{P}(Y_n = j|Y_0 = a) = \mathbb{P}(Y_n = j + b|Y_0 = a + b)$.
2. It is **temporally homogeneous**: $\mathbb{P}(Y_n = j|Y_0 = a) = \mathbb{P}(Y_{m+n} = j|Y_m = a)$, $m, n \geq 0$.
3. It has the **Markov property**: $\mathbb{P}(Y_{m+n} = j|Y_m = a, Y_{m-1} = a_{m-1}, \cdots, Y_0 = a_0) = \mathbb{P}(Y_{m+n} = j|Y_m = a)$.

Proof.

1. $$\mathbb{P}(Y_n = j|Y_0 = a) = \mathbb{P}\left(\sum_{i=1}^{n} X_i = j - a\right) = \mathbb{P}\left(\sum_{i=1}^{n} X_i = j + b - (a + b)\right) = \mathbb{P}(Y_n = j + b|Y_0 = a + b).$$

2. $$\mathbb{P}(Y_n = j|Y_0 = a) = \mathbb{P}\left(\sum_{i=1}^{n} X_i = j - a\right) = \mathbb{P}\left(\sum_{i=m+1}^{m+n} X_i = j - a\right) = \mathbb{P}(Y_{m+n} = j|Y_m = a).$$

3. If we know Y_m, then the distribution of Y_{m+n} depends only on $X_{m+1}, X_{m+2}, \cdots, X_{m+n}$, and $Y_0, Y_1, \cdots, Y_{m-1}$ do not influence the dependency. Therefore,

$$\mathbb{P}(Y_{m+n} = j|Y_m = a, Y_{m-1} = a_{m-1}, \cdots, Y_0 = a_0) = \mathbb{P}(Y_{m+n} = j|Y_m = a).$$

□

Example A.2. (**Sample Path Counting**) A sample path of the random walk is a sequence $(s_0, s_1, \cdots, s_n)$ with $s_0 = a$ and $|s_{i+1} - s_i| = 1$ for all $0 \leq i \leq n-1$. The number of such paths is 2^n.
Each path corresponds to a unique outcome in the sample space.
Let $M_n^r(a, b)$ be the number of paths from $(0, a)$ to (n, b) with r rightward steps and $\ell = n - r$ leftward steps. We have:

$$M_n^r(a,b) = \begin{cases} \binom{n}{r}, & r - \ell = b - a, \\ 0, & \text{Otherwise.} \end{cases}$$

Since $r + \ell = n$ and $r - \ell = b - a$, we have $r = \frac{1}{2}(n + b - a)$ and $\ell = \frac{1}{2}(n - b + a)$.
Therefore, if $n + b - a$ is even and $|b - a| \leq n$, then:

$$\mathbb{P}(Y_n = b) = M_n^{\frac{1}{2}(n+b-a)}(a,b) p^{\frac{1}{2}(n+b-a)} q^{\frac{1}{2}(n-b+a)} = \binom{n}{\frac{1}{2}(n+b-a)} p^{\frac{1}{2}(n+b-a)} q^{\frac{1}{2}(n-b+a)}.$$

Otherwise, $\mathbb{P}(Y_n = b) = 0$.

Theorem A.2. (**Reflection Principle**) Let $N_n(a, b)$ be the number of paths from $(0, a)$ to (n, b), and let $N_n^0(a, b)$ be the number of such paths that intersect the x-axis. We have:

$$N_n^0(a, b) = N_n(-a, b).$$

Proof.
Let k be the first time that the path intersects the x-axis, i.e. $k = \min\{i : s_i = 0\}$.
We reflect the path from time k onwards about the x-axis. The new path starts from $(0, -a)$ and ends at (n, b).
Conversely, given a path from $(0, -a)$ to (n, b), we can find the first time that the path intersects the x-axis and reflect the path from that time onwards about the x-axis to obtain a path from $(0, a)$ to (n, b) that intersects the x-axis. Therefore, there is a one-to-one correspondence between paths from $(0, a)$ to (n, b) that intersect the x-axis and paths from $(0, -a)$ to (n, b). Hence, $N_n^0(a, b) = N_n(-a, b)$. □

Lemma A.3. The number of paths from $(0, a)$ to (n, b) is given by:

$$N_n(a,b) = \binom{n}{\frac{1}{2}(n+b-a)}.$$

Proof.
Let α be the number of positive steps and β be the number of negative steps. We have $\alpha + \beta = n$ and $\alpha - \beta = b - a$. Therefore, $\alpha = \frac{1}{2}(n + b - a)$ and $\beta = \frac{1}{2}(n - b + a)$.
The number of paths from $(0, a)$ to (n, b) is the number of ways to choose α positive steps from n steps. Therefore:

$$N_n(a,b) = \binom{n}{\alpha} = \binom{n}{\frac{1}{2}(n+b-a)}.$$

□

Example A.3. Let $Y_0, Y_1, \cdots, Y_n$ be a simple random walk. We want to know the probability that the path does not revisit the starting point in the time interval $[1, n]$.
Without loss of generality, we assume that $Y_0 = 0$. We want to find $\mathbb{P}(Y_1 Y_2 \cdots Y_n \neq 0, Y_n = b)$.
Note that if $n + b$ is odd or $|b| > n$, then $\mathbb{P}(Y_1 Y_2 \cdots Y_n \neq 0, Y_n = b) = 0$. Assume that $n + b$ is even and $|b| \leq n$.
Let $N_n(0, b)$ be the number of paths from $(0, 0)$ to (n, b), and let $N_n^0(0, b)$ be the number of such paths that intersect the x-axis. If $b > 0$, the first step must be $(1, 1)$. By the Reflection Principle and Lemma A.3:

$$\begin{aligned} N_{n-1}(1,b) - N_{n-1}^0(1,b) &= N_{n-1}(1,b) - N_{n-1}(-1,b) \\ &= \binom{n-1}{\frac{1}{2}(n-1+b-1)} - \binom{n-1}{\frac{1}{2}(n-1+b+1)} \\ &= \binom{n-1}{\frac{1}{2}(n+b)-1} - \binom{n-1}{\frac{1}{2}(n+b)} = \frac{b}{n}\binom{n}{\frac{1}{2}(n+b)}. \end{aligned}$$

Therefore, by Lemma A.3:

$$\mathbb{P}(Y_1 Y_2 \cdots Y_n \neq 0, Y_n = b) = \frac{N_{n-1}(1,b) - N_{n-1}^0(1,b)}{N_n(0,b)} \mathbb{P}(Y_n = b) = \frac{b}{n} \mathbb{P}(Y_n = b).$$

Example A.4. Let $Y_0, Y_1, \cdots, Y_n$ be a simple random walk. We want to find the distribution of the maximum position reached by the random walk in the time interval $[0, n]$:

$$M_n = \max\{Y_0, Y_1, \cdots, Y_n\}.$$

Suppose that $Y_0 = 0$ so that $M_n \geq 0$. If $b \geq r$, then $M_n \geq r$ whenever $Y_n = b$. Therefore, for $b \geq r$:

$$\mathbb{P}(M_n \geq r, Y_n = b) = \mathbb{P}(Y_n = b).$$

It follows that for $r \geq 0$:

$$\mathbb{P}(M_n \geq r) = \sum_{b=r}^{\infty} \mathbb{P}(Y_n = b).$$

Theorem A.4. Suppose that $Y_0 = 0$. For $r \geq 1$ and $b < r$:

$$\mathbb{P}(M_n \geq r, Y_n = b) = \left(\frac{q}{p}\right)^{r-b} \mathbb{P}(Y_n = 2r - b).$$

Proof.
Let $N_n^r(0, b)$ be the number of paths from $(0, 0)$ to (n, b) that intersect the line $y = r$. For a path π, let (i_π, r) be the earliest intersection point of π and the line $y = r$. We reflect the path π from time i_π onwards about the line $y = r$. The new path starts from $(0, 0)$ and ends at $(n, 2r - b)$.
Conversely, given a path from $(0, 0)$ to $(n, 2r - b)$, we can find the earliest intersection point of the path and the line $y = r$ and reflect the path from that time onwards about the line $y = r$ to obtain a path from $(0, 0)$ to (n, b) that intersects the line $y = r$. Therefore, there is a one-to-one correspondence between paths from $(0, 0)$ to (n, b) that intersect the line $y = r$ and paths from $(0, 0)$ to $(n, 2r - b)$. Hence, $N_n^r(0, b) = N_n(0, 2r - b)$.
By Example A.2 and Lemma A.3:

$$\begin{aligned}
\mathbb{P}(M_n \geq r, Y_n = b) &= \frac{N_n^r(0, b)}{N_n(0, b)} \mathbb{P}(Y_n = b) = \frac{N_n(0, 2r - b)}{N_n(0, b)} \mathbb{P}(Y_n = b) \\
&= \frac{\binom{n}{\frac{1}{2}(n+2r-b)}}{\binom{n}{\frac{1}{2}(n+b)}} \binom{n}{\frac{1}{2}(n + b)} p^{\frac{1}{2}(n+b)} q^{\frac{1}{2}(n-b)} = \binom{n}{\frac{1}{2}(n + 2r - b)} p^{\frac{1}{2}(n+b)} q^{\frac{1}{2}(n-b)} \\
&= \left(\frac{q}{p}\right)^{r-b} \binom{n}{\frac{1}{2}(n + 2r - b)} p^{\frac{1}{2}(n+2r-b)} q^{\frac{1}{2}(n-2r+b)} = \left(\frac{q}{p}\right)^{r-b} \mathbb{P}(Y_n = 2r - b).
\end{aligned}$$

□

Remark A.4.1. Combining the two cases, for $r \geq 1$:

$$\mathbb{P}(M_n \geq r) = \sum_{b=r}^{\infty} \mathbb{P}(Y_n = b) + \sum_{b=-\infty}^{r-1} \left(\frac{q}{p}\right)^{r-b} \mathbb{P}(Y_n = 2r - b).$$

If $p = q = \frac{1}{2}$, then for $r \geq 1$:

$$\mathbb{P}(M_n \geq r) = \sum_{b=r}^{\infty} \mathbb{P}(Y_n = b) + \sum_{b=-\infty}^{r-1} \mathbb{P}(Y_n = 2r - b) = \sum_{b=r}^{\infty} \mathbb{P}(Y_n = b) + \sum_{c=r+1}^{\infty} \mathbb{P}(Y_n = c) = 2 \sum_{b=r}^{\infty} \mathbb{P}(Y_n = b).$$

Appendix B

Terminologies in Other Fields of Mathematics

Definition B.1. The **supremum** of a subset S is the lowest upper bound x such that for all $a \in S$, $x \geq a$. It is written as:

$$x = \sup S.$$

Definition B.2. The **infimum** of a subset S is the highest lower bound x such that for all $b \in S$, $x \leq b$. It is written as:

$$x = \inf S.$$

Definition B.3. The **limit superior** and **limit inferior** of a sequence $x_1, x_2, \cdots$ 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, for some real number L:

$$\sum_{n=0}^{\infty} |a_n| = L.$$

Any grouping or rearranging of an absolutely convergent infinite series does not change the result of the series. An infinite series is **conditionally convergent** if it converges but does not satisfy the condition for absolute convergence.

Definition B.5. (**Monotonicity**) A **monotonic** function is a function that is either entirely non-increasing or entirely non-decreasing. A **strictly monotonic** function is a function that is either entirely strictly increasing or entirely strictly decreasing.

Definition B.6. The **arguments of the maxima** are the input points at which a function's output is maximized. 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 points at which a function's output is minimized. 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 is a function f that 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 is a function f that is:

1. Single-valued (any value in the domain maps to exactly one value).
2. Analytic (f can be written as a convergent power series).

Definition B.10. Let V be the space of all real functions on $[0, 1]$. A mapping $\|\cdot\| : V \to \mathbb{R}$ is a **norm** of a function f if:

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. Let f and g be real-valued functions.

1. We write $f(x) = O(g(x))$ if and only if:
$$\limsup_{x \to \infty} \frac{f(x)}{g(x)} < \infty.$$
This is called **big O notation**.
2. We write $f(x) = o(g(x))$ if and only if:
$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0.$$
This is called **small o notation**.
3. We write $f(x) = \Omega(g(x))$ if and only if:
$$\liminf_{x \to \infty} \frac{f(x)}{g(x)} > 0.$$
This is called **big Omega notation**.
4. We write $f(x) = \omega(g(x))$ if and only if:
$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = \infty.$$
This is called **small omega notation**.
5. Functions f and g are **asymptotically equivalent** ($f \sim g$) if and only if:
$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 1.$$
6. If $f(x) = O(g(x))$ and $g(x) = O(f(x))$, then we write:
$$f(x) \asymp g(x).$$

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:

$$|\operatorname{cov}(X,Y)|^2 \leq \operatorname{Var}(X)\operatorname{Var}(Y).$$

Theorem C.5. (**Markov's Inequality**) Let X be a random variable with a finite mean. Then, 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 $\operatorname{Var}(X) = \sigma^2$. Then, 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 be 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.

Change Log

1.0 Create the notes

1.1 Add the definition ”Parametric distribution” in Chapter 5.5 ”Examples of continuous random variables”

1.2 Add the Student’s t-distribution and the properties of chi-squared distribution
Add theorems that relate sample mean and sample variance with distributions
Create a new Chapter 7.8 ”Sampling”
Add a remark in De Moivre-Laplace Limit Theorem
Fix some typos

1.3 Modify the wording of some theorems
Add the definition of random sample and combine it with sample mean and sample variance
Add the definitions related to asymptotic notations
Change how Change Log is produced

1.4 Add a Lemma linking Gamma distribution and Chi-squared distribution
Add Slutsky’s Theorem
Add multivariate normal distribution
Modify the appearance of random vectors

2.0 Combining all expectation related topic into a separate chapter
Greatly modify the ordering of topics
Add MGF of some of the distribution taught
Add a theorem that allows estimating population variance from population mean
Finish the proof of a theorem that allows estimating population variance from sample variance
Add a theorem that correlates uncorrelated bivariate normal and independent normal
Add a lemma regarding the properties of covariance
Add definition of conditional variance and Law of total variance
Modify notations for regular convergence

2.1 Use the updated document template
Create a new Chapter 3.5 ”Marginal Distribution of Random Variables”
Add more examples and remarks
Reorder items for better logical flow
Rewrite definitions, theorems and lemmas
Fix typos with the help of AI