Probability

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

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

This is a rewritten version of my lecture notes named "MATH 2431: Honor Probability". I realised that the ordering of topics is a bit of a mess and there are some missing topics that will probably be necessary for future courses.
My plan is to make this notes a mixture of MATH 2421 (Probability) and MATH 2431 (Honor Probability). Again, if you can find any typos, you are already pretty good at the topics or you have good eyes. ;)
Please note that all proofs that may 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
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \cdots$	Matrix
G_X	Probability generating function of X
M_X	Moment generating function of X
ϕ	CF / PDF of $X \sim \mathrm{N}(0,1)$
Φ	CDF of $X \sim \mathrm{N}(0,1)$

(a) Notations

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

(b) Abbreviations

Definition 0.1. This is definition.

Remark 0.1.1. This is remark.

Lemma 0.2. This is lemma.

Proposition 0.3. This is proposition.

Theorem 0.4. This is theorem.

Claim 0.4.1. This is claim.

Corollary 0.5. This is corollary.

Example 0.1. This is example.

Contents

Chapter 1

Combinatorial Analysis

1.1 Probabilities

In our life, we mostly believe that the future is largely unpredictable. We express this belief in chance behaviour and assign quantitative and qualitative meanings to its usages.
Therefore, we create the concept of "probability", which tries to create a numerical descriptions of how likely an event would occur.

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

If we want to determine the probability, we can use random experiments.

Definition 1.2. **Experiment** is a process that has a random outcome.

Example 1.1. Example of an experiment:

1. Randomly picking a number from 1 to 10
2. Randomly toss a coin

The most basic way to find a 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 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 Principal of Counting, the total number of possible outcomes is:

$$30 \times 29 = 870$$

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

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

Remark 1.4.2. If we want to find the total number of permutations, we can see it as finding the number of outcomes for putting one object into the 1st position, 2nd position, and so on.

Example 1.3. 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. 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, can be obtained by:

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

We also have cases when two ends of the ordered arrangement are connected together. When we want to find the number of arrangements for putting people into seats of circular table, rotating clockwise or anti-clockwise should be considered as same arrangement.

Theorem 1.6. Given a set with n distinct elements. The number of arrangements of elements into the circle is:

$$(n-1)!$$

1.2 Combinations

Now, assume that we don't care about the ordering of choosing objects. We only want to choose k objects from n objects.
Then we have the following definition.

Definition 1.7. If $k \leq n$, k**-combination** of a set with n distinct elements is an unordered arrangement of k elements of the set.

Theorem 1.8. Let n and k be integers with $0 \leq k \leq n$. The number of k-combination of a set with n distinct elements, denoted by C_k^n or $\binom{n}{k}$, can be obtained 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 arrangement of choosing k objects from n objects is:

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

We don't care about the order. The number of unordered arrangement of choosing k objects from n objects would be:

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

□

Remark 1.8.1. By convention, when n is non-negative integers and $k < 0$ or $k > n$, we define:

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

We can immediately derive the following corollary.

Corollary 1.9. Let n be integers. For all integers k with $0 \leq k \leq n$, we have:

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

Proof.
Expanding both sides, we get:

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

□

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

Theorem 1.10. (**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 have the famous Binomial Theorem.

Theorem 1.11. (**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 coefficient**.

Corollary 1.12. Let n be a non-negative integer. We have:

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

Proof.
Using the Binomial Theorem and substitute $x = 1$ 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 1.13. Let n be a positive integer. We have:

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

Proof.
Using the Binomial Theorem and substitute $x = -1$ 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 1.14. Let n be a positive integer. We have:

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

Proof.
Using the Binomial Theorem and substitute $x = 2$ 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 1.15. (**Vandermonde's Identity**) Let $m, n, r \in \mathbb{Z}$ with $0 \le r \le m$ and $0 \le r \le n$. We have:

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

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

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

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

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

□

Corollary 1.16. Let n be a non-negative integer. We have:

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

Proof.
Using the Vandermonde's Identity and substitute $m = n$ 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 1.17. Let n and r be integers such that $0 \le r \le n$. We have:

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

Proof.
We can use the 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 in the same type?

Theorem 1.18. Given a set with n elements. If we choose n_i objects to 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 1.18.1. When you choose k objects from n objects, you may consider it as classifying k objects as chosen and remaining $n - k$ objects as not chosen.

We have a more generalized version of Binomial Theorem.

Theorem 1.19. (**Multinomial Theorem**) Let n be a non-negative integers. 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.

The formula is way too complicated! Luckily, we have a formula to know how many terms the above equation has.

Theorem 1.20. There are $\binom{n+r-1}{r-1}$ distinct non-negative integer-valued vectors $(x_1, x_2, \cdots, x_r)$ that satisfies:

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

where $x_i \geq 0$ for all i.

We can also use the above to find number of ways to put n objects into r objects. If in addition every box must have a box, we use the theorem below.

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

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

where $x_i \geq 0$ for all i.

Example 1.4. Suppose that a cookie shop has 4 different kinds of cookies. How many different ways can 6 cookies be chosen?
Number of ways to choose 6 cookies is the number of 6-combinations with repetition from set with 4 elements, which is:

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

Chapter 2

Events and their probabilities

We have now understood how we can find the number of possibilities. However, we still haven't started rigorously formalize probability. Therefore, we need to define some basic terminologies around 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", which the events usually include some of the elements of sample space.

Definition 2.1. These are the basic object of probabilities.

1. **Experiment** is an activity that produces distinct and well-defined possibilities called **outcomes**, denoted by ω.
2. **Sample space** is the set of all outcomes of an experiment, denoted by Ω.
3. **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 space:

1. Die rolling: $\Omega = \{1, 2, 3, 4, 5, 6\}$
2. Life time of bulb: $\Omega = [0, \infty)$
3. Two coins flipping: $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$

Remark 2.1.1. It is not necessary for all subset of Ω to be an event. However, we do not discuss this issue for the moment.

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

Remark 2.1.2. If only the outcome $\omega = 2$ is given, then there exists many events that can obtain this outcome. E.g. $\{2\}$, $\{2, 4\}$, $\cdots$

2.2 Event Operations

We can perform operations to events, similar to sets.

Definition 2.2. Given two events A and B.

1. **Union** of A and B is an event $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$.
2. **Intersection** of A and B is an event $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$.
3. **Complement** of A is an event containing all elements in sample space Ω that is not in A. It is denoted by A^{C}.
4. **Complement** of B in A is an event $A \setminus B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\}$.
5. **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 inclusion of all outcomes of events in another events.

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 and $A_1, A_2, \cdots, A_k$ be a sequence of 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: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \qquad A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
4. De Morgan's Law: $(\cup_{i=1}^k A_i)^\complement = \cap_{i=1}^k A_i^\complement \qquad (\cap_{i=1}^k A_i)^\complement = \cup_{i=1}^k A_i^\complement$

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

Lemma 2.6. For any events A and B, we have:

$$A = (A \cap B) \cup (A \cap B^\complement)$$

Proof.
By distributive law,

$$(A \cap B) \cup (A \cap B^\complement) = A \cap (B \cup B^\complement) = A \cap \Omega = A$$

□

We may start defining probability. Let's start with defining a collection of subsets of the sample space.

Definition 2.7. Field $\mathcal{F}$ is any collection of subsets of Ω which satisfies the following conditions:

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

We are more interested on σ-field that is closed under countably infinite unions.

Definition 2.8. σ-field (or **σ-algebra**) $\mathcal{F}$ is any collection of subsets of Ω which satisfies the following conditions:

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

Remark 2.8.1. From this point onwards, $\mathcal{F}$ represents the σ-field.

Example 2.3. Smallest σ-field: $\mathcal{F} = \{\emptyset, \Omega\}$

Example 2.4. If A is any subset of Ω, then $\mathcal{F} = \{\emptyset, A, A^\complement, \Omega\}$ is a σ-field.

Example 2.5. 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 2.8.2. These two formulae will be very useful.

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

2.3 Probability measure and Kolmogorov axioms

We wish to be able to discuss the likelihoods of the occurrences of events.
Now that we define some fundamental terminologies, we can finally define probability.

Definition 2.9. Measurable space $(\Omega, \mathcal{F})$ is a pair comprising a sample space Ω and a σ-field $\mathcal{F}$.
Measure μ on a measurable space $(\Omega, \mathcal{F})$ is a function $\mu : \mathcal{F} \to [0, \infty]$ satisfying:

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

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 in fact other measures that satisfy the definition of probability measure. E.g. Risk-neutral measure.
The following measures are not probability measures.

Example 2.6. Lebesgue measure: $\mu((a,b)) = b - a$, $\Omega = \mathbb{R}$

Example 2.7. Counting measure: $\mu(A) = \#\{A\}$, $\Omega = \mathbb{R}$

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

Definition 2.10. 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 $(\Omega, \mathcal{F})$

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

Using probability measure, we can use the axiom of probability to formalize probability.

Definition 2.11. (**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.

Example 2.8. We consider a coin flip. We can find that sample space $\Omega = \{H, T\}$ and σ-field $\mathcal{F} = \{\emptyset, H, T, \Omega\}$. Let $\mathbb{P}(H) = p$ where $p \in [0, 1]$. We define $A = \{\omega \in \Omega : \omega = H\}$. Then we can get:

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

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

Example 2.9. We consider a die roll. We can find that sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ and σ-field $\mathcal{F} = \{0, 1\}^{\Omega}$. 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. $\mathbb{P}(A) = \frac{|A|}{6}$.

The following properties are important and build a foundation of probability.

Lemma 2.12. Basic properties of $\mathbb{P}$:

1. $\mathbb{P}(A^{\complement}) = 1 - \mathbb{P}(A)$.
2. If $A \subseteq B$, then $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$.
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. If A and B are disjoint, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
4. **Inclusion-exclusion formula**

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

Proof.

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

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

□

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

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

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

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

Remark 2.12.1. Given a sequence of sets A_n. We have two types of set limit:

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

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

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

Definition 2.13. We say a sequence of events A_n **converges** and $\lim_{n\to\infty} A_n$ exists if:

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

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

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

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

Lemma 2.14. If $A_1, A_2, \cdots$ are an increasing sequence of events ($A_1 \subseteq A_2 \subseteq \cdots$), then:

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

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

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

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

$$\mathbb{P}\left(\bigcup_{n\to\infty} A_n\right) = \mathbb{P}\left(\bigcup_{n\to\infty} A_n\right) = \sum_{i=1}^{\infty} \mathbb{P}(B_n) = \lim_{N\to\infty} \sum_{i=1}^{N} \mathbb{P}(B_n) = \lim_{N\to\infty} \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_N\right) = \lim_{N\to\infty} \mathbb{P}(A_N)$$

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

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

□

We can give some terminology to some special probabilities.

Definition 2.15. Event A is **null** if $\mathbb{P}(A) = 0$.

Remark 2.15.1. Null events need not to be impossible. For example, the probability of choosing a point in a plane is 0.

Definition 2.16. Event A occurs **almost surely** if $\mathbb{P}(A) = 1$.

2.4 Conditional probability

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

Definition 2.17. 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 2.17.1. For any event A, $\mathbb{P}(A)$ can be regarded as $\mathbb{P}(A|\Omega)$.

Remark 2.17.2. When $\mathbb{P}(E) = \mathbb{P}(E|F)$, E and F are said to be **independent**.

Example 2.10. Two fair dice are thrown. Given that the first shows 3, what is the probability that the sum of number shown exceeds 6?

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

Lemma 2.18. 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 from definition of 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 $\mathbb{P}(\Omega|B) = 0$. 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 the 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 probability based on previous events. This is useful if you are dealing with a sequence of events in time.

Lemma 2.19. (**General Multiplication Rule**) Let $A_1, A_2, \cdots, A_n$ be a sequence of events. We have:

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

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

Lemma 2.20. For any events A and B such that $0 < \mathbb{P}(B) < 1$,

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

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

Example 2.11. In medical cases, we usually identify the efficiency and effectiveness of a type of medical test. We have a name for each types of results:

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

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

Lemma 2.21. (**Law of total probability**) Let $\{B_1, B_2, \cdots, B_n\}$ 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.

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

□

At this point, we can finally prove a theorem that is used in a lot of field outside of mathematics. Imagine that you know the probability of getting each type of disease and the probability of having a specific symptom if you have the disease. If a patient have the symptom, what is the chance that he gets the type of disease you are considering?

Theorem 2.22. (**Bayes' Theorem**) Suppose that a sequence of events $A_1, A_2, \cdots, A_n$ is a partition of sample space. Assume further that $\mathbb{P}(A_i) > 0$ for all i. Let B be any event, then for any i:

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

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

Definition 2.23. Two events A and B are **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. It is denoted by $A \perp\!\!\!\perp B$.

Remark 2.23.1. If events A and B are independent and $A \cap B = \emptyset$, then either $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$.

It is relatively simple to prove the following.

Lemma 2.24. For any two events A and B, 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 2.25. If events A and B are independent, then so are $A \perp\!\!\!\perp B^{\complement}$ and $A^{\complement} \perp\!\!\!\perp B^{\complement}$.

Proof.

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

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

Proposition 2.26. If events A, B, C are independent, then:

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

Proof.

1. Using the properties of probability,

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

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

□

Sometimes, we may deal with more than 2 events. We have a more specific way to describe their relationship.

Definition 2.27. Given a family of events $\{A_i : i \in I$ for some $I \in \mathbb{N}_+$.

1. If $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for any $i \neq j$, it is **pairwise independent**.
2. If additionally that 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 it is **mutually independent**.

Remark 2.27.1. Usually, when we say multiple events are independent, we are saying they are mutually independent.

Example 2.12. Roll for dice twice: $\Omega = \{1, 2, \cdots, 6\} \times \{1, 2, \cdots, 6\}$ and $\mathcal{F} = 2^{\Omega}$
Let A be event that the sum is 7. Event $A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$.
Let B be event that the first roll is 4. Event $B = \{(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6)\}$
Let C be event that the second roll is 3. Event $C = \{(1, 3), (2, 3), (3, 3), (4, 3), (5, 3), (6, 3)\}$

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

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 subset of Ω. However, many of those may be too big to be useful. Therefore, we have the following definition.

Definition 2.28. 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 2.28.1. $\sigma(A)$ is the smallest σ-field containing A.

Example 2.13. Let $\Omega = \{1, 2, \cdots, 6\}$ and $A = \{\{1\}\} \subseteq 2^{\Omega}$. $\sigma(A) = \{\emptyset, \{1\}, \{2, 3, \cdots, 6\}, \Omega\}$

Corollary 2.29. Suppose $(\mathcal{F}_i)_{i \in I}$ is a system of σ-fields in Ω. Then:

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

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

Definition 2.30. 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 2.14. Assume that we want to consider the probabilities of getting a head in coin flipping and getting a 5 in die tossing at the same time. We already know that:

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

The probability space we are considering is the produce 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

3.1 Introduction of random variables

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

Definition 3.1. 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, 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 following classes:

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

It is due to following reasons:

1. X^{-1} can be interchanged with any set functions.
2. $\mathcal{F}$ 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 $a, b \in \mathbb{R}$,

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

$\square$

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 and $c, d \in \mathbb{R}$.

1. If $Y = cX + d$, then Y is a random variable.
2. If $Z = X^2$, then Z is a random variable.

Proof.
Let $y \in \mathbb{R}$ and $z \in \mathbb{R}_{\geq 0}$

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

□

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

Definition 3.3. Borel set is a set which can be obtained by taking countable union, intersection or complement repeatedly. (Countably many steps)

Definition 3.4. 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 way of 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 $B \in \mathcal{B}$, $\mathbb{P} \circ X^{-1}(B) = \mathbb{P}(\{\omega : X(\omega) \in B\}) \in [0, 1]$
$$\mathbb{P} \circ X^{-1}(\emptyset) = \mathbb{P}(\{\omega : X(\omega) \in \emptyset\}) = \mathbb{P}(\emptyset) = 0$$
$$\mathbb{P} \circ X^{-1}(\mathbb{R}) = \mathbb{P}(\{\omega : X(\omega) \in \mathbb{R}\}) = \mathbb{P}(\Omega) = 1$$
2. For any disjoint $B_1, B_2, \dots \in \mathcal{B}$,
$$\mathbb{P} \circ X^{-1}\left(\bigcup_{i=1}^{\infty} B_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} X^{-1}(B_i)\right) = \sum_{i=1}^{\infty} \mathbb{P}(X^{-1}(B_i)) = \sum_{i=1}^{\infty} \mathbb{P} \circ X^{-1}(B_i)$$

□

Remark 3.4.2. We can derive the probability of all $A \in \mathcal{B}$.

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

3.2 CDF of random variables

Every random variable has its own distribution function.

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

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

Example 3.3. From Example 3.1,

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

Lemma 3.6. CDF F_X of a random variable X has the following properties:

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

Proof.

1. Let $B_n = \{\omega \in \Omega : X(\omega) \le -n\} = \{X \le -n\}$. Since $B_1 \supseteq B_2 \supseteq \cdots$, by Lemma 2.14,

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

   Alternative proof:

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

   Let $C_n = \{\omega \in \Omega : X(\omega) \le n\} = \{X \le n\}$. Since $C_1 \subseteq C_2 \subseteq \cdots$, by Lemma 2.14,

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

   Alternative Proof:

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

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

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

3. Let $B_n = \{\omega \in \Omega : X(\omega) \le x + \frac{1}{n}\}$. Since $B_1 \supseteq B_2 \supseteq \cdots$, by Lemma 2.14,

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

   Alternative Proof:

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

□

Remark 3.6.1. F is not left-continuous because:

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

Lemma 3.7. Let F_X be the CDF of a random variable X. Then

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

Proof.

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

□

In some cases, we want to find a number where a specific percentage of outcomes go 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:

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

3.3 PMF / PDF of random variables

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

Definition 3.9. Random variable X is **discrete** if it takes value in some countable subsets $\{x_1, x_2, \cdots\}$ only of $\mathbb{R}$.
Discrete random variable X has **probability mass function** (PMF) $f_X : \mathbb{R} \to [0,1]$ given by:

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

Lemma 3.10. Relationship between PMF f_X and CDF F_X of a random variable X:

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

Proof.

1. $$F_X(x) = \mathbb{P}(X \leq x) = \sum_{i=-\infty}^{x} \mathbb{P}(X = i) = \sum_{i\leq x} f_X(i)$$
2. Let $B_n = \{x - \frac{1}{n} < X \leq x\}$. Since $B_1 \supseteq B_2 \supseteq \cdots$, by Lemma 2.14,

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

□

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

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

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

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

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

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

Remark 3.11.2. On discrete random variable, the distribution is **atomic** because the distribution function has jump discontinuities at values $x_1, x_2, \cdots$ and is constant in between.

Remark 3.11.3. On continuous random variable, the CDF of a continuous variable is **absolutely continuous**.
Not every continuous function can be written as $\int_{-\infty}^{x} f_X(u)\,du$. E.g. Canton function

Remark 3.11.4. It is possible that a random variable is neither continuous nor discrete.

3.4 JCDF of random variables

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

Definition 3.12. Let $X_1, X_2 : \Omega \to \mathbb{R}$ be random variables. We define **random vector** $\vec{X} = (X_1, X_2) : \Omega \to \mathbb{R}^2$ with properties

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

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

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

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

Claim 3.12.1. Both definitions of random vectors are equivalent.

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

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

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

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

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

Remark 3.12.1. We can write $\mathbb{P} \circ \vec{X}^{-1}(D) = \mathbb{P}(\vec{X} \in D) = \mathbb{P}(\{\omega \in \Omega : \vec{X}(\omega) = (X_1(\omega), X_2(\omega)) \in D\})$.

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

Definition 3.13. Joint distribution function (JCDF) $F_{\vec{X}} : \mathbb{R}^2 \to [0, 1]$ is defined as

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

Remark 3.13.1. We can replace all Borel sets by the form $[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$.

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

Lemma 3.14. JCDF $F_{X,Y}$ of random vector (X, Y) has the following properties:

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

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

Definition 3.15. Let X, Y be random variables. We can get a **marginal distribution** (marginal CDF) by having:

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

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

Definition 3.16. Two random variables X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$ are **jointly discrete** if the vector (X, Y) takes values in some countable subset of $\mathbb{R}^2$ only. The corresponding **joint (probability) mass function** (JPMF) $f : \mathbb{R}^2 \to [0, 1]$ is given by

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

Remark 3.16.1.

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

Remark 3.16.2. More generally, for all $B \in \mathcal{B}(\mathbb{R}^2)$,

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

Definition 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:

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

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

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

Example 3.4. Assume that a special three-sided coin is provided. Each toss results in head (H), tail (T) or edge (E) with equal probabilities. What is the probability of having h heads, t tails and e edges after n tosses?
Let H_n, T_n, E_n be the numbers of such outcomes in n tosses of the coin. The vector (H_n, T_n, E_n) satisfy $H_n + T_n + E_n = n$.

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

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

Example 3.5. Let X be uniformly distributed on $[0, 1]$ ($f_X(x) = \mathbf{1}_{[0,1]}$). This means $f_X(x) = 1$ when $x \in [0, 1]$ and 0 otherwise.
Let $Y = X$ ($Y(\omega) = X(\omega)$ for all $\omega \in \Omega$). That means $(X, Y) = (X, X)$.
Let $B = \{(x, y) : x = y \text{ and } x \in [0, 1]\} \in \mathcal{B}(\mathbb{R}^2)$. Since $y = x$ is just a line,

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

Therefore, X and Y are not jointly continuous.

Chapter 4

Discrete random variables

4.1 Introduction of discrete random variables

Let's recall some of the definitions on discrete random variable in previous chapter.

Definition 4.1. Random variable X is **discrete** if it takes value in some countable subsets $\{x_1, x_2, \cdots\}$ only of $\mathbb{R}$.
(Cumulative) distribution function (CDF) of discrete random variable X is the function $F_X : \mathbb{R} \to [0, 1]$ given by:

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

Probability mass function (PMF) of discrete random variable X is the function $f_X : \mathbb{R} \to [0, 1]$ given by:

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

CDF and PMF are related by

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

Lemma 4.2. PMF $f_X : \mathbb{R} \to [0, 1]$ of a discrete random variable X satisfies:

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

We also recall the definition of joint distribution function and joint mass function.

Definition 4.3. For jointly discrete random variables X and Y, **joint probability mass function** (JPMF) $f_{X,Y} : \mathbb{R}^2 \to [0, 1]$ is given by

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

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

Definition 4.4. Discrete random variables X and Y are **independent** if the events $\{X = x\}$ and $\{Y = y\}$ are independent for all x, y. Equivalently, X and Y are independent if

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

Claim 4.4.1. Three definitions are equivalent.

Proof.
We can get definition 2 from definition 1.

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

We can get definition 3 from definition 2.

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

We can get definition 1 from definition 3.

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

Therefore, three definitions are equivalent. □

Remark 4.4.1. More generally, let $X_1, X_2, \cdots, X_n : \Omega \to \mathbb{R}$ be discrete random variables. They are **independent** if

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

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

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

Remark 4.4.2. From the definition, we can see that $X \perp\!\!\!\perp Y$ means that $X^{-1}(E) \perp\!\!\!\perp Y^{-1}(F)$ for all $E, F \in \mathcal{B}(\mathbb{R})$.

Remark 4.4.3. We can generate σ-field using random variables by defining σ-field generated by random variable X

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

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

Definition 4.5. Let $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ be two σ-fields. We say $\mathcal{G}$ and $\mathcal{H}$ are **independent** if $A \perp\!\!\!\perp B$ for all $A \in \mathcal{G}, B \in \mathcal{H}$.

Remark 4.5.1. $\sigma(X) \perp\!\!\!\perp \sigma(Y) \iff X \perp\!\!\!\perp Y$

Theorem 4.6. Given two random variables X and Y. If $X \perp\!\!\!\perp Y$ and we have two functions $g, h : \mathbb{R} \to \mathbb{R}$ such that $g(X)$ and $h(Y)$ are still random variables, then $g(X) \perp\!\!\!\perp h(Y)$.

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

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

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

Remark 4.6.1. We assume a product space $(\Omega, \mathcal{F}, \mathbb{P})$ of two probability space $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$.
($\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$, $\mathbb{P}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2)$).
Any pair of events of the form $E_1 \times \Omega_2$ and $\Omega_1 \times E_2$ are independent.

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

4.2 Expectation of discrete random variables

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

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

Definition 4.7. Suppose we have a discrete random variable X taking values from $\{x_1, x_2, \cdots\}$ with PMF $f_X(x)$. **Mean value**, **expectation**, or **expected value** of X is defined to be:

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

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

Example 4.1. Suppose a product is sold seasonally. Let b be net profit for each sold unit, ℓ be net loss for each left unit, and X be number of products ordered by customer. If y units are stocked, what is the expected profit $Q(y)$?

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

Theorem 4.8. (**Tail Sum Formula**) For non-negative integer-valued discrete random variable X, we have:

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

Proof.

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

□

Lemma 4.9. If discrete random variable X has a PMF f_X and $g : \mathbb{R} \to \mathbb{R}$ such that $g(X)$ is still a discrete random variable, then

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

whenever this sum is absolutely convergent.

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

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

□

Lemma 4.10. Let (X,Y) be a discrete random vector with JPMF $f_{X,Y}(x,y)$. Let $g:\mathbb{R}^2\to\mathbb{R}$ such that $g(X,Y)$ is a discrete random variable. Then

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

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

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

□

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

Definition 4.11. Let $k\in\mathbb{N}_+$. We have a special term for each of the following expectations:

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

Remark 4.11.1. Not all random variables have k-th moments for all $k\in\mathbb{N}_+$.

Remark 4.11.2. We cannot use collection of moments to uniquely determine a distribution that has k-th moments for all $k\in\mathbb{N}$.

Theorem 4.12. Expectation operator $\mathbb{E}$ has the following properties:

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

Proof.

1. Since $f_X(x)\geq 0$ for all x, $\mathbb{E}X=\sum_x xf_X(x)\geq 0$ if $X\geq 0$.
2. Let $g(X,Y)=aX+bY$. Then,

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

3. $\mathbb{E}(1)=1(1)=1$.

□

Remark 4.12.1. More generally, we have

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

Example 4.2. Assume we have N different types of card and each time one gets a card to be any one of the N types. Each types is equally likely to be gotten. What is the expected number of types of card we can get if we get n cards?
Let $X = X_1 + X_2 + \cdots + X_N$ where $X_i = 1$ if at least one type i card is among the n cards and otherwise 0.

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

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

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

Lemma 4.13. If two discrete random variables X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y$.

Proof.

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

□

Lemma 4.14. Given two discrete random variables X and Y. Let $g, h : \mathbb{R} \to \mathbb{R}$ such that $g(X), h(Y)$ are still discrete random variables. If $X \perp\!\!\!\perp Y$ and $\mathbb{E}(g(X)h(Y)), \mathbb{E}g(X)$ and $\mathbb{E}h(Y)$ exist, then $\mathbb{E}(g(X)h(Y)) = \mathbb{E}g(X)\mathbb{E}h(Y)$.

Proof.

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

□

4.3 Conditional distribution of discrete random variables

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

Definition 4.15. Suppose $X, Y : \Omega \to \mathbb{R}$ are two discrete random variables. **Conditional distribution** of Y given $X = x$ for any x such that $\mathbb{P}(X = x) > 0$ is defined by

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

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

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

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

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

Remark 4.15.1. By definition,

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

.

Remark 4.15.2. For any $x \in \mathbb{R}$, the conditional PMF $f_{Y|X}(y|x)$ is a probability mass function in y.

Remark 4.15.3. If X and Y are independent, then $f_{Y|X}(y|x) = f_Y(y)$.

Conditional distributions still have properties of original distribution.

Lemma 4.16. Given two discrete random variables X and Y. Conditional distributions have following properties:

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

Proof.

1. $$\sum_{v \le y} f_{Y|X}(v|x) = \sum_{v \le y} \mathbb{P}(Y = v|X = x) = \mathbb{P}(Y \le y|X = x) = F_{Y|X}(y|x)$$
2. This is just Lemma 3.10.

□

Definition 4.17. Given two discrete random variables X and Y. **Conditional expectation** ψ of Y given $X = x$ for any x is defined by:

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

Conditional expectation ψ of Y given X is defined by:

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

Example 4.3. Assume we roll a fair dice.

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

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

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

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

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

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

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

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

Example 4.4. If $Y = X$, then $\psi(X) = \mathbb{E}(Y|X) = \mathbb{E}(X|X) = X$.

Example 4.5. If $Y \perp\!\!\!\perp X$, then $\psi(X) = \mathbb{E}Y$.

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

Definition 4.18. Given a random variable Y and a σ-field $\mathcal{H} \subseteq \mathcal{F}$.
$\mathbb{E}(Y|\mathcal{H})$ is any random variable Z satisfying the following two properties:

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

Remark 4.18.1. Under this definition,

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

Theorem 4.19. (Law of total expectation) Given two discrete random variables X and Y. Conditional expectation $\psi(X) = \mathbb{E}(Y|X)$ satisfies:

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

Proof.
By Lemma 4.9,

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

□

Example 4.6. A miner is trapped in a mine with doors, each will lead to a tunnel.
Tunnel 1 will help the miner reach safety after 3 hours respectively.
However, tunnel 2 and 3 will send the miner back after 5 and 7 hours respectively.
What is the expected amount of time the miner need to reach safety? (Assume that the miner is memoryless)
Let X be the amount of time to reach safety, Y be the door number he chooses for the first time.

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

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

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

Example 4.7. We consider a sum of random number of random variables.
Let N be the number of customers and X_i be the amount of money spent by the i-th customers.
Assume that N and X_i's are all independent and $\mathbb{E}X_i = \mathbb{E}X$, what is the expected total amount of money spent by all N customers?

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

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

Theorem 4.20. Given two discrete random variables X and Y. Conditional expectation $\psi(X) = \mathbb{E}(Y|X)$ satisfies:

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

for any function g for which both expectations exist.

Proof.
By Lemma 4.9,

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

□

4.4 Convolution of discrete random variables

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

Theorem 4.21. Given two jointly discrete random variables X and Y. The probability of sum of two random variables is given by:

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

Proof.
We have the disjoint union:

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

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

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

□

Definition 4.22. Convolution f_{X+Y} $(f_X * f_Y)$ of PMFs of two independent discrete random variables X and Y is the PMF of $X+Y$:

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

4.5 Examples of discrete random variables

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

Definition 4.23. Parametric distribution of a discrete random variable is a distribution where the PMF depends on one or more parameters.

The following examples are some of the most useful distributions.

Example 4.8. (**Constant variables**) Let $X : \Omega \to \mathbb{R}$ be defined by $X(\omega) = c$ for all $\omega \in \Omega$. For all $B \in \mathcal{B}$,

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

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

Example 4.9. (**Bernoulli distribution**) $X \sim \text{Bern}(p)$
Let $A \in \mathcal{F}$ be a specific event. A Bernoulli trial is considered a success if A occurs. Let $X : \Omega \to \mathbb{R}$ be such that

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

The expectation and variance of Bernoulli random variable is:

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

Example 4.10. Let A be an event in $\mathcal{F}$ and **indicator functions** $\mathbf{1}_A : \Omega \to \mathbb{R}$ such that for all $B \in \mathcal{B}(\mathbb{R})$:

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

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

Example 4.11. (**Binomial distribution**) $Y \sim \text{Bin}(n, p)$
Suppose we perform n independent Bernoulli trials $X_1, X_2, \cdots, X_n$. Let $Y = X_1 + X_2 + \cdots + X_n$ be total number of successes.

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

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

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

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

Theorem 4.24. If $X \sim \text{Bin}(n, p)$, then $\mathbb{E}X = np$ and $\text{Var}(X) = np(1-p)$.

Proof.

$$\mathbb{E}X = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} = np$$

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}(X(X-1)) + \mathbb{E}X - (\mathbb{E}X)^2 = np - n^2p^2 + \sum_{k=2}^{\infty} k(k-1) \binom{n}{k} p^k (1-p)^{n-k} \\ &= np - n^2p^2 + \sum_{k=1}^{n} \frac{n!}{(k-2)!(n-k)!} p^k (1-p)^{n-k} \\ &= np - n^2p^2 + n(n-1)p^2 \sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2} (1-p)^{n-k} \\ &= np - n^2p^2 + n(n-1)p^2 = np(1-p) \end{aligned}$$

□

Example 4.12. (**Trinomial distribution**) Suppose we perform n trials, each of which result in three outcomes A, B and C, where A occurs with probability p, B with probability q, and C with probability $1-p-q$. Probability of r A's, w B's, and $n-r-w$ C's is

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

Example 4.13. (**Geometric distribution**) $W \sim \text{Geom}(p)$
Suppose we keep performing independent Bernoulli trials until the first success shows up. Let p be the probability of success and W be the **waiting time** which elapses before first success.

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

Theorem 4.25. If $X \sim \text{Geom}(p)$, then $\mathbb{E}X = \frac{1}{p}$ and $\text{Var}(X) = \frac{1-p}{p^2}$.

Proof.

$$\begin{aligned}
\mathbb{E}X &= \sum_{k=1}^{\infty} kp(1-p)^{k-1} = p\sum_{k=1}^{\infty} k(1-p)^{k-1} = \frac{p}{p^2} = \frac{1}{p} \\
\text{Var}(X) &= \mathbb{E}(X(x-1)) - \mathbb{E}X + (\mathbb{E}X)^2 = \frac{1}{p} - \frac{1}{p^2} + \sum_{k=2}^{\infty} k(k-1)p(1-p)^{k-1} \\
&= \frac{1}{p} - \frac{1}{p^2} + p(1-p)\sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2} \\
&= \frac{2p(1-p)}{p^3} = \frac{2(1-p)}{p^2}
\end{aligned}$$

□

Example 4.14. (Alternative Geometric distribution) Suppose we keep performing independent Bernoulli trials until the first success shows up. Let p be the probability of success and 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 4.25.1. Conventionally, when we consider the geometric distribution, we usually refer to the one related to waiting time instead of number of failures.

Example 4.15. (**Negative binomial distribution**) $W_r \sim \text{NBin}(r, p)$
Similar with examples of geometric distribution, let W_r be the waiting time for the r-th success. For $k \geq r$,

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

Remark 4.25.2. W_r is the sum of r independent geometric variables.

Theorem 4.26. If $X \sim \text{NBin}(r, p)$, then $\mathbb{E}X = \frac{r}{p}$ and $\text{Var}(X) = \frac{r(1-p)}{p^2}$.

Proof.
Assume that $X_i \sim \text{Geom}(p)$ for all i. Since X is the sum of r independent geometric random variable, we get:

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

□

Example 4.16. (**Poisson distribution**) $X \sim \text{Poisson}(\lambda)$
Poisson variable is a discrete random variable with Poisson PMF:

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

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

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

Theorem 4.27. If $X \sim \text{Poisson}(\lambda)$, then $\mathbb{E}X = \lambda$ and $\text{Var}(X) = \lambda$.

Proof.

$$\begin{aligned}
\mathbb{E}X &= \sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!}e^{-\lambda} = \lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda} = \lambda \\
\text{Var}(X) &= \mathbb{E}(X(X-1)) + \mathbb{E}X - (\mathbb{E}X)^2 = \lambda - \lambda^2 + \sum_{k=2}^{\infty}\frac{\lambda^k}{(k-2)!}e^{-\lambda} \\
&= \lambda - \lambda^2 + \lambda^2\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}e^{-\lambda} \\
&= \lambda - \lambda^2 + \lambda^2 = \lambda
\end{aligned}$$

$\square$

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

Example 4.17. (**Poisson flips**) A coin is tossed once and head turns up with probability p.
Let random variables X and Y be the numbers of heads and tails respectively. X and Y are not independent since

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

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

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x, Y = y|N = x+y)\mathbb{P}(N = x+y) = \binom{x+y}{x}p^x(1-p)^y\frac{\lambda^{x+y}}{(x+y)!}e^{-\lambda} = \frac{(\lambda p)^x(\lambda(1-p))^y}{x!y!}e^{-\lambda}$$

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

Theorem 4.28. If the number of occurrence of an event in unit time or space followings Poisson distribution with rate λ, then the number of occurrence in t units times or spaces follows Poisson(λt).

Proof.
Number of occurrence of an event in t unit time or space is equivalent to the sum of t numbers of occurrence of an event in unit time or space.
Let $X_1 \sim \text{Poisson}(\mu)$, $X_2 \sim \text{Poisson}(\lambda)$ are independent. For any $k \geq 2$,

$$\begin{aligned}
\mathbb{P}(X_1 + X_2 = k) &= \sum_{i=0}^{k} \left(\frac{\mu^i}{i!}e^{-\mu}\right)\left(\frac{\lambda^{k-i}}{(k-i)!}e^{-\lambda}\right) \\
&= \frac{1}{k!}e^{-(\mu+\lambda)} \sum_{i=0}^{k} \frac{k!}{i!(k-i)!}\mu^i \lambda^{k-i} \\
&= \frac{1}{k!}e^{-(\mu+\lambda)} \sum_{i=0}^{k} \binom{k}{i}\mu^i \lambda^{k-i} \\
&= \frac{(\mu-\lambda)^k}{k!}e^{-(\mu+\lambda)}
\end{aligned}$$

Therefore, $(X_1 + X_2) \sim \text{Poisson}(\lambda + \mu)$.
When $t = 2$, the number of occurrence in 2 units times or spaces follows Poisson(2λ).
By induction, the number of occurrence in t units times or spaces follows Poisson(λt). □

Example 4.18. (**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 $x = 0, 1, \cdots, \min(m, n)$.
The expectation and variance of hypergeometric random variable is:

$$\mathbb{E}X = \frac{mn}{N} \qquad \text{Var}(X) = \frac{mn}{N}\left(\frac{(m-1)(n-1)}{N-1} + 1 - \frac{mn}{N}\right)$$

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

Chapter 5

Continuous random variables

5.1 Introduction of continuous random variables

We recall some definitions of continuous random variables.

Definition 5.1. Random variable X is **continuous** if its distribution function (CDF) $F_X(x)$ 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. PDF f_X is not prescribed uniquely since two integrable function which take identical values except at some specific point have the same integral. However, if F_X is **differentiable** at u, we set $f_X(u) = F'_X(u)$.

Note that we have used the same letter f for mass functions and density functions since both are performing similar task.

Remark 5.1.2. 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 element of probability.

Lemma 5.2. If continuous random variable X has a density function f_X, then

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

Proof.

1. $$\int_{-\infty}^{\infty} f_X(x)\, dx = \lim_{x \to \infty} F_X(x) = 1$$
2. $$\mathbb{P}(X = x) = \lim_{h \to 0} \int_{x-h}^{x} f_X(x)\, dx = F_X(x) - \lim_{h \to \infty} F(x - h) = F_X(x) - F_X(x) = 0$$
3. $$\mathbb{P}(a \leq X \leq b) = F(b) - F(a) = \int_{-\infty}^{b} f_X(x)\, dx - \int_{-\infty}^{a} f_X(x)\, dx = \int_a^b f_X(x)\, dx$$

□

Remark 5.2.1. More generally, for an interval B, we have

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

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

Definition 5.3. Two continuous random variables X and Y are called **independent** if for all $x, y \in \mathbb{R}$,

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

Theorem 5.4. Let two continuous random variables X and Y be independent. Suppose $g(X)$ and $h(Y)$ are still continuous random variables, then $g(X)$ and $h(Y)$ are independent.

5.2 Expectation of continuous random variables

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

Definition 5.5. **Expectation** of a continuous random variable X with density function f is given by:

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

whenever this integral exists.

Remark 5.5.1. We usually can define $\mathbb{E}X$ only if $\mathbb{E}|X|$ exists.

We have a special properties in the continuous random variable.

Lemma 5.6. (Tail sum formula) If continuous random variable X has a PDF f_X with $f_X(x) = 0$ when $x < 0$, and a CDF F_X, then

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

Proof.

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

□

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

Lemma 5.7. If continuous random variable X has a PDF f_X with $f_X(x) = 0$ when $x > 0$, and a CDF F_X, then

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

Proof.

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

□

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

Theorem 5.8. If X and $g(X)$ are continuous random variables, then

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

Proof.
We first consider that $g(x) \geq 0$ for all x. Let $Y = g(X)$ and $B = \{x : g(x) > y\}$. By Lemma 5.6,

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

We then consider that $g(x) \leq 0$ for all x. Let $Z = g(X)$ and $C = \{x : g(x) < z\}$. By Lemma 5.7,

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

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

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

$\square$

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

Definition 5.9. Given $k \in \mathbb{N}_+$ and a continuous random variable X.

1. The **k-th moment** of X is defined to be:
$$\mathbb{E}X^k = \int_{-\infty}^\infty x^k f_X(x)\, dx$$
2. The **k-th central moment** of X is defined to be:
$$\mathbb{E}((X - \mathbb{E}X)^k) = \int_{-\infty}^\infty (x - \mathbb{E}X)^k f_X(x)\, dx$$
3. **Variance** of X is defined as $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$.
4. **Standard deviation** of X is defined as $\sigma = \sqrt{\mathrm{Var}(X)}$.

5.3 Joint distribution function of continuous random variables

Again, we recall the definition of joint distribution function.

Definition 5.10. 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 the have a **joint density function** (JPDF) $f : \mathbb{R}^2 \to [0, \infty)$ such that:

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

We also recall the definition of marginal distribution function.

Definition 5.11. Given two continuous random variables X and Y. Marginal distribution function (Marginal PDF) of X given Y is

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

Similarly, we have the following extension of Theorem 5.8. However, we are not going to prove it here.

Theorem 5.12. If X and Y are jointly continuous random variables and $g(X, Y)$ is continuous random variable, then

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

We can obtain the following important lemma.

Lemma 5.13. If X and Y are jointly continuous random variables, then for any $a, b \in \mathbb{R}$,

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

Proof.

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

□

5.4 Conditional distribution of continuous random variables

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

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

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

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

Definition 5.14. Suppose $X, Y : \Omega \to \mathbb{R}$ are two continuous random variables with PDF $f_X(x) > 0$ for some $x \in \mathbb{R}$. **Conditional distribution function** (Conditional CDF) of Y given $X = x$ is defined by

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

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

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

Remark 5.14.1. Since $f_X(x)$ can also be computed from $f(x, y)$, we can simply compute

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

Remark 5.14.2. More generally, for two continuous random variables X and Y with PDF $f_X(x) > 0$ for some $x \in \mathbb{R}$,

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

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

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

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

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

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

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

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

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

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

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

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

Definition 5.15. Given two continuous random variables X and Y and an event $X = x$ for some $x \in \mathbb{R}$. **Conditional expectation** of Y is defined by:

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

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

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

Again we also have the same properties of conditional distribution.

Lemma 5.16. (Law of total expectation) Conditional expectation $\psi(X) = \mathbb{E}(Y|X)$ for continuous random variables X and Y satisfies:

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

Proof.

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

□

Lemma 5.17. Conditional expectation $\psi(X) = \mathbb{E}(Y|X)$ for continuous random variables X and Y satisfies:

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

Proof.

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

□

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

Theorem 5.18. If two jointly continuous random variables X and Y have JPDF $f_{X,Y}$, then $X+Y$ has a PDF

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

Proof.

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

□

Definition 5.19. Given two independent continuous random variables X and Y. **Convolution** f_{X+Y} $(f_X * f_Y)$ of PDFs of X and Y is the PDF of $X+Y$:

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

5.5 Examples of continuous random variables

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

Example 5.3. (**Uniform distribution**) $X \sim \mathrm{U}[a,b]$
Random variable X is **uniform** on $[a,b]$ if CDF and PDF of X is

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

Example 5.4. If $X \sim \mathrm{U}[0,1]$ and $Y \sim \mathrm{U}[0,1]$. In case of $X \perp\!\!\!\perp Y$,

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

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 center of the needle to the nearest line and Θ be the acute angle between the needle and vertical line. We have $\mathbb{P}(\text{Intersection}) = \mathbb{P}\left(\frac{L}{2}\cos\Theta \ge X\right)$.
Assume that $X \perp\!\!\!\perp \Theta$. We have $X \sim \mathrm{U}\left[0, \frac{D}{2}\right]$ and $\Theta \sim \mathrm{U}\left[0, \frac{\pi}{2}\right]$.

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

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

Suppose that we throw the needle for n times.

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

Example 5.6. (**Inverse transform sampling**) If we have an invertible CDF $G(x)$. How can we generate a random variable Y with the given distribution function?
We only need to generate an uniform random variable $U \sim \mathrm{U}[0,1]$. We claim that $Y = G^{-1}(U)$ has the distribution function $G(x)$.

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

Example 5.7. (**Exponential distribution**) $X \sim \mathrm{Exp}(\lambda)$
Random variable X is **exponential** with parameter $\lambda > 0$ if CDF and PDF of X is

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

Theorem 5.20. Exponential distribution has memoryless property. It means that for all $s > 0$ and $t > 0$,

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

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

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

□

Theorem 5.21. If $X \sim \mathrm{Exp}(\lambda)$, then $\mathbb{E}X = \frac{1}{\lambda}$ and $\mathrm{Var}(X) = \frac{1}{\lambda^2}$.

Proof.

$$\begin{aligned}
\mathbb{E}X &= \int_0^\infty x\lambda e^{-\lambda x}\,dx = -xe^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x}\,dx = -\frac{1}{\lambda}e^{-\lambda x}\Big|_0^\infty = \frac{1}{\lambda} \\
\mathrm{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}X)^2 = -\frac{1}{\lambda^2} + \int_0^\infty x^2\lambda e^{-\lambda x}\,dx \\
&= -\frac{1}{\lambda} - x^2e^{-\lambda x}\Big|_0^\infty + \int_0^\infty 2xe^{-\lambda x}\,dx \\
&= -\frac{1}{\lambda} + \frac{2}{\lambda}\mathbb{E}X \\
&= -\frac{1}{\lambda} + \frac{2}{\lambda^2} = \frac{1}{\lambda^2}
\end{aligned}$$

□

Example 5.8. (**Normal distribution / Gaussian distribution**) $X \sim \mathrm{N}(\mu, \sigma^2)$
Random variable X is **normal** if it has two parameters μ and σ^2, and its PDF and CDF is

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

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

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

Claim 5.21.1. $\phi(x)$ is a probability distribution function.

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

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

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

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

□

These are some properties that are used frequently.

Lemma 5.22. The normal distribution has the following properties:

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

Proof.

1. Let $z = bx + a$.
$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}\left(X \leq \frac{y-a}{b}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{y-a}{b}} e^{-\frac{x^2}{2}}\,dx = \frac{1}{\sqrt{2\pi b^2}} \int_{-\infty}^{y} e^{-\frac{(z-a)^2}{2b^2}}\,dz$$
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, $Y \sim \mathrm{N}(0,1)$.
3. Let $y = bz + a$.
$$\mathbb{E}Y = \frac{1}{\sqrt{2\pi b^2}} \int_{-\infty}^{\infty} y e^{-\frac{(y-a)^2}{2b^2}}\,dy = \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{\infty} bz e^{-\frac{z^2}{2}}\,dz + \int_{-\infty}^{\infty} a e^{-\frac{z^2}{2}}\,dz\right) = \frac{a}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\,dz = a(1) = a$$
$$\mathrm{Var}(Y) = \frac{1}{\sqrt{2\pi b^2}} \int_{-\infty}^{\infty} (y-a)^2 e^{-\frac{(y-a)^2}{2b^2}}\,dy = \frac{b^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}}\,dz = \frac{-b^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z d\left(e^{-\frac{z^2}{2}}\right) = \frac{b^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\,dz = b^2$$

□

Lemma 5.23. 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.
Just apply Lemma 5.2 and you would get the equation. □

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

Theorem 5.24. If $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2, \cdots, n$ and they are independent, then $\sum_{i=1}^n X_i \sim \mathrm{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$.

Proof.
We first consider a special case when $X \sim \mathrm{N}(0, \sigma^2)$, $Y \sim \mathrm{N}(0,1)$ and $X \perp\!\!\!\perp Y$.
$$\begin{aligned}
f_{X+Y}(z) &= \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\,dy \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-y)^2}{2\sigma^2}\right) \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)\right) dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma} \exp\left(-\frac{z^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2}(-2yz + y^2(1+\sigma^2))\right) dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma} \exp\left(-\frac{z^2}{2\sigma^2} + \frac{z^2}{2\sigma^2(1+\sigma^2)}\right) \exp\left(-\frac{1+\sigma^2}{2\sigma^2}\left(\frac{z^2}{(1+\sigma^2)^2} - \frac{2yz}{1+\sigma^2} + y^2\right)\right) dy \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{1+\sigma^2}} \exp\left(-\frac{z^2}{2\sigma^2} + \frac{z^2}{2\sigma^2(1+\sigma^2)}\right) \left(\frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{1+\sigma^2}}}\right) \exp\left(-\frac{\left(y - \frac{z}{1+\sigma^2}\right)^2}{2\left(\frac{\sigma}{\sqrt{1+\sigma^2}}\right)^2}\right) dy \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1+\sigma^2}} \exp\left(-\frac{z^2}{2(1+\sigma^2)}\right)
\end{aligned}$$
Therefore, $X + Y \sim \mathrm{N}(0, 1+\sigma^2)$.
In general case when $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ and $X_1 \perp\!\!\!\perp X_2$.
$$X_1 + X_2 = \sigma_2\left(\frac{X_1 - \mu_1}{\sigma_2} + \frac{X_2 - \mu_2}{\sigma_2}\right) + \mu_1 + \mu_2$$

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

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

$\square$

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

Example 5.9. (**Standard bivariate normal distribution**) Two continuous random variables X and Y are **standard bivariate normal** if they have JPDF:

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

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

Remark 5.24.1. 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 5.24.2. ρ is called the **population correlation coefficient** between X and Y. It will be discussed in Chapter 6.

Remark 5.24.3. If $X \sim \mathrm{N}(0,1)$ and $Y \sim \mathrm{N}(0,1)$,

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

In most of the cases, normal random variables X and Y do not have a mean of 0 and a variance 1. If we also include mean and variance into the distribution, we would have 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, variance σ_X^2 and σ_Y^2, and correlation coefficient ρ if JPDF is given by

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

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

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

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

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

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

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

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

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

Remark 5.24.4. X and Y are bivariate normal and uncorrelated $\iff$ X and Y are independent normal.

Remark 5.24.5. X and Y are jointly continuous and they are both normal does not mean they are bivariate normal.

Example 5.12. Consider a JPDF of random variables X and Y

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

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

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

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

Remark 5.24.6. Two random variables X and Y are jointly continuous and uncorrelated Gaussian does not mean they are independent Gaussian.

Example 5.13. (**Cauchy distribution**) $X \sim \text{Cauchy}(\theta)$
Random variable X has a Cauchy distribution if it has a PDF:

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

It has the expectation

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

Remark 5.24.7. If $X \sim \mathrm{N}(0,1)$ and $Y \sim \mathrm{N}(0,1)$, then $\frac{X}{Y} \sim \text{Cauchy}(0)$.

Example 5.14. (**Gamma distribution**)$X \sim \Gamma(\alpha, \lambda)$
Random variable X has a gamma distribution with parameters α and λ if it has a PDF:

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

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, $\Gamma(\alpha) = (\alpha-1)!$. The expectation and variance are:

$$\mathbb{E}X = \frac{\alpha}{\lambda} \qquad \text{Var}(X) = \frac{\alpha}{\lambda^2}$$

Remark 5.24.8. When $\alpha = 1$, it is an exponential distribution.

If we substitute $\lambda = \frac{1}{2}$ and $\alpha = \frac{k}{2}$ where k is a parameter, we get the Chi-squared distribution.

Example 5.15. (**Chi-squared distribution**) $Y \sim \chi^2(k)$
Assume that $X_1, X_2, \cdots, X_n$ 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 k if it has a PDF:

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

The expectation and variance are:

$$\mathbb{E}Y = k \qquad \text{Var}(Y) = 2k$$

Remark 5.24.9. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Example 5.16. (**Beta distribution**) $X \sim \text{Beta}(a, b)$
Random variable X has a beta distribution with parameters a and b if it has a PDF:

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

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

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

The expectation and variance are:

$$\mathbb{E}X = \frac{a}{a+b} \qquad \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

5.6 Functions of continuous random variables

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

Example 5.17. Assume that X is continuous random variable with PDF $f_X(x)$. Let $Y = g(X)$ be a continuous random variable. How do we find the PDF $f_Y(y)$? We work with $F_Y(y)$ first. Let $g^{-1}(A) = \{x \in \mathbb{R} : g(x) \in A\}$.

$$
\begin{aligned}
F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \in (-\infty, y]) = \mathbb{P}(X \in g^{-1}((-\infty, y])) = \int_{g^{-1}((-\infty,y])} f_X(x)\,dx \\
f_Y(y) &= \frac{\partial}{\partial y}\int_{g^{-1}((-\infty,y])} f_X(x)\,dx
\end{aligned}
$$

Example 5.18. Let $X \sim \mathrm{N}(0,1)$. Let $Y = g(X) = X^2$. We want to find the PDF $f_Y(y)$.

$$
\begin{aligned}
F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y}) - 1 \\
f_Y(y) &= F'(y) = 2\phi(\sqrt{y})\left(\frac{1}{2\sqrt{y}}\right) = \frac{1}{\sqrt{y}}\phi(\sqrt{y}) = \begin{cases} \frac{1}{\sqrt{2\pi y}}\exp\left(\frac{-y}{2}\right), & y > 0 \\ 0, & y < 0 \end{cases}
\end{aligned}
$$

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

Theorem 5.25. In case that $g(x)$ is strictly monotonic (strictly increasing or strictly decreasing) and differentiable, let $Y = g(X)$. We have

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

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

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

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

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

$\square$

We can consider the multivariable case.

Example 5.19. Suppose two random variables X and Y are jointly continuous with JPDF $f_{X,Y}$. Given that $U = g(X,Y)$ and $V = h(X,Y)$. What is $f_{U,V}(u,v)$? For simplifying the process, we need to first make some following assumptions.

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

2. The function g and h are differentiable and the Jacobian determinant

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

Then

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

Example 5.20. Given two jointly continuous random variables X_1, X_2 and their JPDF f_{X_1,X_2}. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$.

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

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

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

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

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

Example 5.21. 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 $x_1, x_2 \in \mathbb{R}$,

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

Summary

Definition

Definition 1. Given a set with n distinct elements.

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

Definition 2. If $k \leq n$, 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 object of probabilities.

1. **Experiment** is an activity that produces distinct and well-defined possibilities called **outcomes**, denoted by ω.
2. **Sample space** is the set of all outcomes of an experiment, denoted by Ω.
3. **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. **Union** of A and B is an event $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$.
2. **Intersection** of A and B is an event $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$.
3. **Complement** of A is an event containing all elements in sample space Ω that is not in A. It is denoted by A^{C}.
4. **Complement** of B in A is an event $A \setminus B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\}$.
5. **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 of the outcomes in A are also in B, then we say A is **contained** in B, written as $A \subset B$ or $B \supset A$.

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. **σ-field** (**σ-algebra**) $\mathcal{F}$ is any collection of subsets of Ω which satisfied 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. **Measurable space** $(\Omega, \mathcal{F})$ is a pair comprising a sample space Ω and a σ-field $\mathcal{F}$.

Definition 10. **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 $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Definition 11. **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. Event A is **null** if $\mathbb{P}(A) = 0$.

Definition 14. Event A is **almost surely** if $\mathbb{P}(A) = 1$.

Definition 15. Given $\mathbb{P}(B) > 0$. **Conditional probability** that A occurs given that B occurs is:
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

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

Definition 18. **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. **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. **Borel set** is a set which can be obtained by taking countable union, intersection or complement repeatedly.

Definition 21. **Borel σ-field** $\mathcal{B}(\mathbb{R})$ of $\mathbb{R}$ is a σ-field that is generated by all open sets. It is a collection of Borel sets.

Definition 22. **(Cumulative) distribution function** (CDF) of a random variable X is a function $F_X : \mathbb{R} \to [0,1]$ given by

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

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

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

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

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

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

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

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

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

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

Definition 25. Given a random vector (X, Y). **Joint distribution function** (JCDF) $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$ is defined as:

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

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

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

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

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

Definition 26. Let X and Y be random variables. **Marginal distribution function** (Marginal CDF) is given by:

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

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

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

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

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

Definition 27. Given a random variable X. **Mean value**, **expectation**, or **expected value** of X is given by:

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

If it is absolutely convergent.

Definition 28. Given $k \in \mathbb{N}_+$ and a random variable X. ***k*-th moment** m_k is defined to be:

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

***k*-th cnetral moment** α_k is defined to be

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

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

Definition 29. Given two random variables X and Y. **Conditional distribution function** (Conditional CDF) of Y given $X = x$ for any x is defined by:

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

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

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

Definition 30. Given two random variables X and Y, and an event $X = x$ for some X. **Conditional expectation** of random variable Y is defined by:

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

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

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

Definition 31. Given $X \perp\!\!\!\perp Y$. In discrete case, **convolution** f_{X+Y} $(f_X * f_Y)$ of PMFs of random variables X and Y is the PMF of $X + Y$:

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

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

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

Definition 32. Parametric distribution of a discrete random variable is a distribution where the PMF 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 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 coefficient**.

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 integers. 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 $\bigcup_{i=1}^{n} = \Omega$). 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 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 $\psi(X) = \mathbb{E}(Y|X)$. Conditional expectation satisfies:

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

Property 11. (Tail sum formula) If 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 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 1 Bernoulli trial. Let p be probability of success and X be number of successes.

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

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

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

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

$$f_{X,Y}(x, y) = \binom{n}{r, w, n-r-w} p^x q^y (1-p-q)^{n-x-y}$$

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

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

Above is the conventional geometric distribution.
Let X be number of failures before first success. For $k \ge 0$,

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

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

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

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

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

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

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

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 $0 \le k \le \min(m, n)$,

$$f_X(k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \qquad \mathbb{E}X = \frac{mn}{N} \qquad \text{Var}(X) = \frac{mn}{N}\left(\frac{(m-1)(n-1)}{N-1} + 1 - \frac{mn}{N}\right)$$

For continuous random variables,

Example 8. (Uniform distribution) $X \sim \text{U}[a, b]$
Random variable X is uniform on $[a, b]$ is PDF and CDF is:

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

Example 9. (Exponential distribution) $X \sim \text{Exp}(\lambda)$
Random variable X is exponential with parameter $\lambda > 0$ if PDF and CDF is:

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

Example 10. (Normal distribution / Gaussian distribution) $X \sim \text{N}(\mu, \sigma^2)$
Random variable X is normal if it has two parameters μ and σ^2, and its PDF and CDF is:

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

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 μ_X and μ_Y, variance 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. (Cauchy distribution) $X \sim \text{Cauchy}(\theta)$
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 13. (Gamma distribution) $X \sim \Gamma(\alpha, \lambda)$
Random variable X has a gamma distribution with parameters α and λ if:

$$f_X(x) = \begin{cases} 0, & x < 0 \\ \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}, & x \ge 0 \end{cases} \qquad \Gamma(\alpha) = \int_{0}^{\infty} e^{-y} y^{\alpha-1}\, dy \qquad \mathbb{E}X = \frac{\alpha}{\lambda} \qquad \text{Var}(X) = \frac{\alpha}{\lambda^2}$$

where $\Gamma(\alpha)$ is the gamma function with $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$. If α is a positive integer, then $\Gamma(\alpha) = (\alpha-1)!$.

Example 14. (Chi-squared distribution) $Y \sim \chi^2(k)$
Assume that $X_1, X_2, \cdots, X_n$ are independent standard normal random variables. Let $Y = \sum_{i=1}^{n} X_i^2$. Random variable Y has a χ^2-distribution with parameter k if:

$$f_Y(x) = \begin{cases} 0, & x < 0 \\ \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})}, & x \geq 0 \end{cases} \qquad \mathbb{E}Y = k \qquad \mathrm{Var}(Y) = 2k$$

Example 15. (Beta distribution) $X \sim \mathrm{Beta}(a, b)$
Random variable X has a beta distribution with parameters a and b if:

$$f_X(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, & 0 < x < 1 \\ 0, & \text{Otherwise} \end{cases} \quad B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad \mathbb{E}X = \frac{a}{a+b} \quad \mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b)}$$

where $B(a, b)$ is the beta function.

Chapter 6

Properties of Expectation and Generating function

6.1 Expectation of sum of random variables

In the previous chapters, you can find that a lot of distribution involves the sum of other distributions. It is necessary that we observe the properties involving expectations. Let's start with the most simple one.

Theorem 6.1. Given a random variable X. For any $a \leq b$, if $a \leq X \leq b$, then $a \leq \mathbb{E}X \leq b$.

Proof.
In discrete case,

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

In continuous case,

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

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

This establish that the lower bound and upper bound of random variable X are also the lower and upper bound of its expectation. We should also remind the linearity of expectation.

Theorem 6.2. (**Linearity of expectation**) Let $X_1, X_2, \cdots, X_n$ be random variables. We have:

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

for all $a_i \in \mathbb{R}$ for all i.

Example 6.1. Suppose that we have a set of N balls, which m of them are red and the rest are blue. We choose n of these balls, without replacement. What is the expected number of red balls in our sample?
We let $X \sim \text{Hypergeometric}(N, m, n)$. How do we find $\mathbb{E}X$? Let $X = X_1 + X_2 + \cdots + X_m$, where for all i,

$$X_i = \begin{cases} 1, & \text{if the } i\text{-th red ball is selected} \\ 0, & \text{Otherwise} \end{cases}$$

From this, for all i, 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, $\mathbb{E}X = \frac{mn}{N}$.

We sometimes want to find the tendency in the linear relationship between two random variables. We say it is called the covariance of two random variables.

Definition 6.3. Covariance of two random variables X and Y is:

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

Remark 6.3.1. Magnitude of covariance is the geometric mean of the variances of two random variables. The sign represents the linear relationship between two random variables. If the sign is positive, this means two random variables show similar behaviour. If the sign is negative, this means two random variables show opposite behaviour.

Remark 6.3.2.

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

Remark 6.3.3. In general, for any random variables $X_1, X_2, \cdots, X_n$,

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

Remark 6.3.4. If X_i are (pairwise) independent or uncorrelated, we can get that $\operatorname{cov}(X_i, X_j) = 0$ for all $i \neq j$.

Example 6.2. 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 correlation coefficient.

Definition 6.4. 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 find 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 6.4.1. Population correlation coefficient ρ of random variables X and Y satisfies $-1 < \rho < 1$.

Remark 6.4.2. If ρ is near 1 or near -1, then it shows a strong linear relationship between X and Y

Remark 6.4.3. The constant ρ used in bivariate normal distribution is the population correlation coefficient.

Lemma 6.5. 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 6.5.1. If X and Y are independent, then they are uncorrelated. The converse is generally not true.

Example 6.3. Let X be such that $f_X(0) = f_X(1) = f_X(-1) = \frac{1}{3}$ and Y be such that $Y = 0$ if $X \neq 0$ and $Y = 1$ if $X = 0$.

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

However,

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

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

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

Theorem 6.6. For random variables X and Y,

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

Proof.

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

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

2. When X and Y are uncorrelated,

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

□

In a lot of cases, we do not know the actual distribution of the population. We can only predict the distribution based on the samples we can get.

Definition 6.7. Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed random variables with CDF F, mean μ and variance σ^2. **Sample mean**, denoted by $\overline{X}$, is defined by:

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

Theorem 6.8. Given a sample mean $\overline{X}$ defined above. We have $\mathbb{E}\overline{X} = \mu$ and $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$.

Proof.

$$\begin{aligned}\mathbb{E}\overline{X} &= \mathbb{E}\left(\sum_{\frac{1}{n}\sum_{k=1}^{n} X_k}\right) = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}X_k = \frac{1}{n}\sum_{k=1}^{n}\mu = \mu \\ \mathrm{Var}(\overline{X}) &= \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{Var}(X_k) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}\end{aligned}$$

□

Definition 6.9. Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed random variables with CDF F, mean μ and variance σ^2. Let $\overline{X}$ be the sample mean. **Sample variance**, denoted by S^2, is defined by:

$$S^2 = \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{n - 1}$$

Remark 6.9.1. Notice that the denominator is $n-1$.

Theorem 6.10. Given a sample variance S^2 defined above. We have $\mathbb{E}S^2 = \sigma^2$.

Proof.

$$\begin{aligned}
\mathbb{E}S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i - \overline{X})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\mathbb{E}(X_i-\mu)^2 + \mathbb{E}(\overline{X}-\mu)^2 - 2\mathbb{E}((X_i-\mu)(\overline{X}-\mu))) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{Var}(X_i) + \mathrm{Var}(\overline{X}) - 2\operatorname{cov}(X_i,\overline{X})) \\
&= \frac{n\sigma^2}{n-1} + \frac{\sigma^2}{n-1} - \frac{2}{n-1}\sum_{i=1}^{n}\operatorname{cov}\left(X_i, \frac{1}{n}\sum_{j=1}^{n}X_j\right) \\
&= \frac{n\sigma^2}{n-1} + \frac{\sigma^2}{n-1} - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{cov}(X_i,X_j) \\
&= \frac{n\sigma^2}{n-1} + \frac{\sigma^2}{n-1} - \frac{2\sigma^2}{n-1} = \sigma^2
\end{aligned}$$

□

6.2 Introduction of generating functions

A sequence of number $a = \{a_i : i = 0, 1, 2, \cdots\}$ may contain a lot of information. For example, values of PMF tells us the distribution of a discrete random variables.
A concise way of storing this information is to wrap up the numbers together in a generating function.

Definition 6.11. For any sequence $\{a_n : n = 0, 1, 2, \cdots\}$, we defined the generating function by

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

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

Remark 6.11.1. We can observe that

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

Example 6.4. Sometimes, we cannot interchange countable sum with derivatives.
Let $b_n(x) = \frac{\sin nx}{n}$ such that $a_1(x) = b_1(x)$ and $a_n(x) = b_n(x) - b_{n-1}(x)$.

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

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

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

Convolutions are common in probability theory, and generating functions can provide a tool for studying them.

Definition 6.12. Let $a = \{a_i : i \geq 0\}$ and $b = \{b_i : i \geq 0\}$ be two sequence of real numbers. **Convolution** $c = a * b = \{c_i : i \geq 0\}$ of $\{a_i\}$ and $\{b_i\}$ is defined by

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

Example 6.5. If $a_n = f_X(n)$ and $b_n = f_Y(n)$, then $c_n = f_{X+Y}(n)$.

Claim 6.12.1. If sequences a and b have generating functions $G_a(s)$ and $G_b(s)$ respectively, then

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

Proof.

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

$\square$

Example 6.6. 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 $b_n = f_Y(n)$.

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

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

Definition 6.13. **Probability generating function** (PGF) of a non-negative random variable X is

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

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

Definition 6.14. **Radius of convergence** R of power series is the half size of an interval such that the power series $f(s)$ is convergent. If $s \in (-R, R)$, then $f(s)$ is convergent. If $s \in [-R, R]^{\complement}$, then $f(s)$ is divergent.
We can obtain the radius of convergence by applying root test:

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

Remark 6.14.1. We need to perform additional tests to find whether the power series converges at $s = -R$ and $s = R$.

Remark 6.14.2. Sometimes, it is hard to compute R using root test. One convenient way to compute R is using the ratio test. If the limit exists,

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

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

Theorem 6.15. If R is the radius of convergence of $G_a(s) = \sum_{i=0}^{\infty} a_i s^i$, then

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

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

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

Remark 6.15.1. For any sequence $\{a_n : n \geq 0\}$, if radius of convergence of $G_a(s)$ is positive, then $\{a_n : n \geq 0\}$ is uniquely determined by $G_a(s)$ via

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

Remark 6.15.2. If $a_n = f_X(n)$ for some random variables 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 6.7. Let $X \sim \text{Poisson}(\lambda)$ and $a_n = f_X(n) = \frac{\lambda^n e^{-\lambda}}{n!}$. By ratio test,

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

Therefore, $R = \infty$.

Example 6.8. Let X has a PMF $a_n = f_X(n) = \frac{c}{n^2}$. By ratio test,

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

Therefore, $R = 1$.

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

Example 6.9. By having $s = 1$,

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

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

Theorem 6.16. (**Abel's Theorem**) Suppose that $a_n \geq 0$ for all n. If a has a generating function $G_a(s)$ and radius of convergence $R = 1$, then if $\sum_{n=0}^{\infty}$ converges in $\mathbb{R} \cup \{\infty\}$, we have

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

Example 6.10. We have some PGF of random variable X.

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

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

Theorem 6.17. If random variable X has a PGF $G_X(s)$, then

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

Proof.

1. This is proved in Example 6.9.
2. Let $s<1$.

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

   By applying Abel's Theorem, we obtain

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

3.

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

□

From Example 6.6, we can generalize it to study the sum of many other independent discrete random variables.

Theorem 6.18. If $X \perp\!\!\!\perp Y$, then $G_{X+Y}(s) = G_X(s)G_Y(s)$.

Proof.

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

□

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

Theorem 6.19. Let $X_1, X_2, \cdots$ be a sequence of independent identically distributed (i.i.d.) random variables with common PGF $G_X(s)$ and N be a random variable independent of X_i for all i with PGF $G_N(s)$. If $T = X_1 + X_2 + \cdots + X_N$, then

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

Proof.

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

□

Example 6.11. The sum of a Poisson number of independent Bernoulli random variables is still Poisson.
Let $G_N(t) = e^{\lambda(t-1)}$ and $G_X(s) = 1 - p + ps$.

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

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

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

Definition 6.20. Let random variables X_1, X_2 be both non-negative integer-valued, jointly discrete with JPMF f_{X_1,X_2}. **Joint probability generating function** (JPGF) is defined by

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

Remark 6.20.1. 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 6.21. Random variables X, Y are independent if and only if $G_{X,Y}(s, t) = G_X(s)G_Y(t)$.

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

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

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

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

Therefore, $X \perp\!\!\!\perp Y$. $\square$

Remark 6.21.1. We know that if $X_1 \perp\!\!\!\perp X_2$, then $G_{X_1+X_2}(s) = \mathbb{E}s^{X_1+X_2} = \mathbb{E}s^{X_1}s^{X_2} = G_{X_1}(s)G_{X_2}(s)$. Converse may not be true.

6.3 Applications of generating functions

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

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

Definition 6.22. A random variable X is **defective** if $\mathbb{P}(X = \infty) > 0$.

Remark 6.22.1. It is no surprise that expectation is infinite when random variable is defective.

With this generalization, we can start discussing random walk.

Example 6.12. (**Recurrence and transience of random walk**) Let S_n the position of the particles after n moves and X_i be independent and identically distributed random variables mentioned in Appendix A. For $n \geq 0$,

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

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

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

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

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

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

Theorem 6.23. From the definitions in Example 6.12, we have

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

Proof.

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

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

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

2. *If you want to understand the proof, search "Central binomial coefficient" in Wikipedia*
   We know that $S_n = 0$ if n is even. Therefore,

$$P_0(s) = \lim_{N \uparrow \infty} \sum_{n=0}^{N} p_0(n) s^n = \lim_{N \uparrow \infty} \sum_{i=0}^{N} \binom{2i}{i} p^i q^i s^{2i} = \lim_{N \uparrow \infty} \sum_{i=1}^{N} (-1)^i 4^i \binom{\frac{-1}{2}}{i} p^i q^i s^{2i} = \frac{1}{\sqrt{1 - 4pqs^2}}$$

($\binom{\frac{-1}{2}}{i}$ is a generalized binomial coefficient)

3. By applying (1) and (2), we can get

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

□

From this theorem, we can get the following corollary.

Corollary 6.24. The probability that the particle ever returns to the origin is

$$\sum_{n=1}^{\infty} f_0(n) = F_0(1) = 1 - |p-q|$$

Probability that the particle will not return to origin ever is

$$\mathbb{P}(T_0 = \infty) = |p-q|$$

Proof.
By using Theorem 6.23, since $p+q=1$,

$$F_0(1) = 1 - (1-4pq)^{\frac{1}{2}} = 1 - (p^2 - 2pq + q^2)^{\frac{1}{2}} = 1 - |p-q|$$

□

Remark 6.24.1. Random walk is **recurrent** if it has at least one recurrent point. ($\mathbb{P}(X < \infty) = 1$)
Random walk is **transient** if it has no recurrent points. ($\mathbb{P}(X = \infty) > 0$)
Notice that when $p = q = \frac{1}{2}$, $\mathbb{P}(T_0 = \infty)$ and therefore random walk is recurrent.
If $p \neq q$, then $\mathbb{P}(T_0 = \infty) \neq 0$ and so the random walk is transient.

Example 6.13. We use the Example 6.12 again. How do we calculate $\mathbb{E}T_0$ if $p = q = \frac{1}{2}$?

$$F_0(s) = 1 - \sqrt{1-s^2} \qquad F_0'(s) = \frac{s}{\sqrt{1-s^2}} \qquad \mathbb{E}T_0 = \lim_{s \uparrow 1} F_0'(s) = \infty$$

This means that although we find that the particle almost certainly return to origin, the expectation for number of steps needed to return to origin is still infinite.

We move on to our next important application, which is the Branching Process.
Many scientists have been interested in reproduction in a population. Accurate models for evolution are extremely difficult to handle, but some non-trivial models are tractable. We will investigate one of the models.

Example 6.14. (**Galton-Watson process**) This process investigates a population that evolves in generations.
Let Z_n be number of individuals of the n-th generation and $X_i^{(m)}$ be number of offspring of the i-th individual of the m-th generation. We have:

$$Z_{n+1} = \begin{cases} X_1^{(n)} + X_2^{(n)} + \cdots + X_{Z_n}^{(n)}, & Z_n \geq 1 \\ 0, & Z_n = 0 \end{cases}$$

We make some following assumptions:

1. Family sizes of the individuals of the branching process form a collection of independent random variables. ($X_i^{(k)}$'s are independent)

2. All family sizes have the same probability mass function f and generating function G. ($X_i^{(k)}$'s are identically distributed)

Assume that $Z_0 = 1$. Note that $Z_1 = X_1^{(0)}$

Theorem 6.25. Let $G_n(s) = \mathbb{E}s^{Z_n}$ and $G(s) = G_1(s) = \mathbb{E}s^{Z_1} = \mathbb{E}s^{X_i^{(m)}}$ for all i and m. Then

$$G_n(s) = G(G(\cdots(G(s))\cdots)) = G(G_{n-1}(s)) = G_{n-1}(G(s))$$

is the n-fold iteration of G.
This further implies

$$G_{m+n}(s) = G_m(G_n(s)) = G_n(G_m(s))$$

Proof.
When $n = 2$,

$$G_2(s) = \mathbb{E}s^{Z_2} = \mathbb{E}s^{X_1^{(1)}+X_2^{(1)}+\cdots+X_{Z_1}^{(1)}} = G_{Z_1}\left(G_{X_1^{(1)}}(s)\right) = G(G(s))$$

.
When $n = m + 1$ for some m,

$$G_{m+1}(s) = \mathbb{E}s^{Z_{m+1}} = \mathbb{E}s^{X_1^{(m)}+X_2^{(m)}+\cdots+X_{Z_m}^{(m)}} = G_{Z_m}\left(G_{X_1^{(m)}}(s)\right) = G_m(G(s))$$

□

In principle, the above theorem tells us the distribution of Z_n. However, it may not be easy to compute $G_n(s)$.
The moments of Z_n can be computed easier.

Lemma 6.26. Let $\mathbb{E}Z_1 = \mathbb{E}X_i^{(m)} = \mu$ and $\mathrm{Var}(Z_1) = \sigma^2$. Then

$$\mathbb{E}Z_n = \mu^n \qquad \mathrm{Var}(Z_n) = \begin{cases} n\sigma^2, & \mu = 1 \\ \frac{\sigma^2(\mu^n-1)\mu^{n-1}}{\mu-1}, & \mu \neq 1 \end{cases}$$

Proof.
Using Theorem 6.25, we can get

$$\begin{aligned}
\mathbb{E}Z_2 &= G_2'(1) = G'(G(1))G'(1) = G'(1)\mu = \mu^2 \\
\mathbb{E}Z_n &= G_n'(1) = G'(G_{n-1}(1))G_{n-1}'(1) = G'(1)\mu^{n-1} = \mu^n \\
G_1''(1) &= \sigma^2 + (G'(1))^2 - G'(1) = \sigma^2 + \mu^2 - \mu \\
G_2''(1) &= G''(G(1))(G'(1))^2 + G'(G(1))G''(1) = G''(1)(\mu^2 + \mu) \\
G_n''(1) &= G''(G_{n-1}(1))(G_{n-1}'(1))^2 + G'(G_{n-1}(1))G_{n-1}''(1) \\
&= (\sigma^2 + \mu^2 - \mu)\mu^{2n-2} + \mu G_{n-1}''(1) \\
&= \mu^{2n-2}(\sigma^2 + \mu^2 - \mu) + \mu^{2n-3}(\sigma^2 + \mu^2 - \mu) + \cdots + \mu^{n-1}(\sigma^2 + \mu^2 - \mu) \\
&= \frac{\mu^{n-1}(\sigma^2 + \mu^2 - \mu)(\mu^n - 1)}{\mu - 1}
\end{aligned}$$

If $\mu = 1$,

$$\mathrm{Var}(Z_n) = G_n''(1) + G_n'(1) - (G_n'(1))^2 = \sigma^2 + G_{n-1}''(1) + 1 - 1 = n\sigma^2$$

If $\mu \neq 1$,

$$\mathrm{Var}(Z_n) = G_n''(1) + G_n'(1) - (G_n'(1))^2 = \frac{\mu^{n-1}(\sigma^2 + \mu^2 - \mu)(\mu^n - 1)}{\mu - 1} + \mu^n - \mu^{2n} = \frac{\mu^{n-1}\sigma^2(\mu^n - 1)}{\mu - 1}$$

□

Example 6.15. Does this process eventually lead to extinct?
Note that

$$\{\text{ultimate extinction}\} = \bigcup_n \{Z_n = 0\} = \lim_{n\uparrow\infty} \{Z_n = 0\}$$

$$\mathbb{P}(\text{ultimate extinction}) = \mathbb{P}\left(\lim_{n\uparrow\infty} \{Z_n = 0\}\right) = \lim_{n\uparrow\infty} \mathbb{P}(Z_n = 0) = \lim_{n\uparrow\infty} G_n(0)$$

Let $\eta_n = G_n(0)$ and $\eta = \lim_{n\uparrow\infty} \eta_n$.

Theorem 6.27. We have that η is the smallest non-negative root of the equation

$$s = G(s)$$

Furthermore,

1. $\eta = 1$ if $\mu < 1$
2. $\eta < 1$ if $\mu > 1$
3. $\eta = 1$ if $\mu = 1$ and $\sigma^2 > 0$
4. $\eta = 0$ if $\mu = 1$ and $\sigma^2 = 0$

Proof.

$$\eta_n = G_n(0) = G(G_{n-1}(0)) = G(\eta_{n-1})$$

We know that η_n is bounded. Therefore, $\eta_n \uparrow \eta$ for some $\eta \in [0, 1]$.

$$\eta = \lim_{n\uparrow\infty} \eta_n = \lim_{n\uparrow\infty} G(\mu_{n-1}) = G\left(\lim_{n\uparrow\infty} \eta_{n-1}\right) = G(\eta)$$

Suppose that there exists another non-negative root ψ.

$$\eta_1 = G(0) \le G(\psi) = \psi$$
$$\eta_2 = G(\eta_1) \le G(\psi) = \psi$$

By induction, $\eta_n \le \psi$ for all n and therefore $\eta \le \psi$. Therefore, η is the smallest non-negative root of the equation $s = G(s)$.

$$G''(s) = \sum_{i=2}^{\infty} i(i-1)s^{i-2}\mathbb{P}(Z_1 = i) \ge 0$$

Therefore, G is non-decreasing and also either convex or a straight line.
When $\mu \ne 1$, we can find that two curves $y = G(s)$ and $y = s$ intersects at $s = 1$ and $s = k \in \mathbb{R}$.
We know that $\eta \le 1$ since η is the smallest root. In order to intersect at $s = \eta$, $G'(\eta) \le 1$.
If $\mu = G'(1) < 1$, then $\eta = 1$.
If $\mu = G'(1) > 1$, then $\eta = k$ such that $G'(k) \le 1$.
In the case when $\mu = G'(1) = 1$, we need to further analyse whether $y = G(s)$ intersects $y = s$ at 1 point or infinite points.

$$\sigma^2 = G''(1) + G'(1) - (G'(1))^2 = G''(1)$$

If $\sigma^2 = G''(1) > 0$, then $\eta = 1$.
If $\sigma^2 = G''(1) = 0$, then $\eta = 0$. □

6.4 Expectation revisited

Recall that the expectations are given respectively by

$$\mathbb{E}X = \begin{cases} \sum x f_X(x), & X \text{ is discrete} \\ \int x f_X(x)\, dx, & X \text{ is continuous} \end{cases}$$

We want a notation which incorporates both these cases. Suppose that X has a CDF F. We can rewrite the equations as

$$\mathbb{E}X = \begin{cases} \sum x\, dF_X(x), & dF_X(x) = F_X(x) - \lim_{y\uparrow x} F_X(y) = f_X(x) \\ \int x\, dF_X(x), & dF_X(x) = \frac{\partial F}{\partial x}\, dx = f_X(x)\, dx \end{cases}$$

Instead of using the regular Riemann integral, which cannot deal with discrete case, we can use the Riemann-Stieltjes integral, which is a generalization of the Riemann integral.

$$\int_a^b g(x)\, dx = \lim_{\max_i |x_{i+1}-x_i|} \sum_i g(x_i^*)(x_{i+1} - x_i)$$
$$\int_a^b g(x)\, dF(x) = \lim_{\max_i |x_{i+1}-x_i|} \sum_i g(x_i^*)(F(x_{i+1}) - F(x_i))$$

if the limit does not depend on the choice of $x_i^* \in [x_i, x_{i+1})$.

Definition 6.28. Expectation of a random variable X is given by:

$$\mathbb{E}X = \int x\, dF_X$$

Lemma 6.29. 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 6.29.1. The notation of $\int g(x)\, dF_X(x)$ does not mean Riemann-Stieltjes integral.

Example 6.16. If g is regular (differentiable at every point and every values in the domain maps to a value in range), then

$$\sum_i g(x_i^*)(F(x_{i+1} - F(x_i)) \approx \sum_i g(x_i^*) f(x_i^*)(x_{i+1} - x_i) \approx \int g(x) f(x)\, dx$$

Example 6.17. In irregular case, assume that the function g is the Dirichlet function. That is

$$\mathbf{1}_{\mathbb{Q}}(x) = \begin{cases} 1, & x \in \mathbb{Q} \\ 0, & x \notin \mathbb{Q} \end{cases} \qquad \sum_i g(x_i^*)(F(x_{i+1}) - F(x_i)) = \sum_i g(x_i^*)(x_{i+1} - x_i)$$

Since the limit depends on the choice of x_i^*, Riemann-Stieltjes integral of $\mathbf{1}_{\mathbb{Q}}(x)$ with respect to $F(x) = x$ is not well defined.
Therefore, $\mathbb{E}\mathbf{1}_{\mathbb{Q}}(X)$ cannot be defined as a Riemann-Stieltjes integral.
However, on the other hand,

$$\mathbb{E}\mathbf{1}_{\mathbb{Q}}(X) = \mathbb{P}(\mathbf{1}_{\mathbb{Q}}(x) = 1) = \mathbb{P} \circ X^{-1}(\mathbb{Q} \cap [0, 1]) = 0$$

With this notation, we can also change how we define PGF.

Definition 6.30. Probability generating function of a random variable X is given by:

$$\mathbb{E}s^X = \int s^x\, dF_X$$

6.5 Moment generating function and Characteristic function

Now that we have unified the notations, we can now properly apply the probability generating function.
For a more general variables X, it is best if we substitute $s = e^t$. We get the following definition.

Definition 6.31. Moment generating function (MGF) of a random variable X is the function $M : \mathbb{R} \to [0, \infty)$ given by:

$$M_X(t) = \mathbb{E}(e^{tX}) = \int e^{tx}\, dF_X$$

Remark 6.31.1. The definition of MGF only requires replacing s by e^t in PGF. MGF is easier for computing moments, but less convenient for computing distribution.

Remark 6.31.2. MGFs are related to Laplace transforms.

We can easier get the following lemma.

Lemma 6.32. Given a MGF $M_X(t)$ of a random variable X.

1. For any $k \geq 0$,
$$\mathbb{E}X^k = M^{(k)}(0)$$
2. The function M can be expanded via Taylor's Theorem within its radius of convergence.
$$M(t) = \sum_{i=0}^{\infty} \frac{\mathbb{E}X^k}{k!} t^k$$
3. If X and Y are independent, then
$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

Proof.

1. $$M^{(k)}(0) = \frac{\partial^k}{\partial t^k} \int e^{tx}\, dF_X(x)\bigg|_{t=0} = \int x^k e^{tx}\, dF_X(x)\bigg|_{t=0} = \int x^k\, dF_X(x) = \mathbb{E}X^k$$
2. Just using (1) and Taylor's Theorem and you get the answer.
3. This is just Theorem 6.18.

□

Remark 6.32.1. $M_X(0) = 1$ for all random variables X.

Example 6.18. Let $X \sim \mathrm{Exp}(1)$. For all $x > 0$, if $t < 1$,

$$f_X(x) = e^{-x} \qquad M_X(t) = \int_0^\infty e^{tx}\, dF_X(x) = \int_0^\infty e^{(t-1)x}\, dx = \frac{1}{1-t}$$

Example 6.19. Let $X \sim$ Cauchy.

$$f_X(x) = \frac{1}{\pi(1+x^2)} \qquad M_X(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{e^{tx}}{1+x^2}\, dx$$

$M_X(t)$ exists only at $t = 0$. We get $M_X(0) = 1$.

Moment generating functions provide a useful technique but the integrals used to define may not be finite. There is another class of functions which finiteness is guaranteed.

Definition 6.33. Characteristic function (CF) of a random variable X is the function $\phi_X : \mathbb{R} \to \mathbb{C}$ given by:

$$\phi_X(t) = \mathbb{E}(e^{itX}) = \int e^{itx}\, dF_X(x) = \mathbb{E}\cos(tX) + i\mathbb{E}\sin(tX) \qquad i = \sqrt{-1}$$

Remark 6.33.1. $\phi_X(t)$ is essentially a Fourier Transform.

Lemma 6.34. CF ϕ_X of a random variable X has the following properties:

1. $\phi_X(0) = 1$. $|\phi_X(t)| \leq 1$ for all t
2. $\phi_X(t)$ is uniformly continuous

Proof.

1. For all t,
$$\phi_X(0) = \int dF_X(x) = 1$$
$$|\phi_X(t)| = \left|\int (\cos(tx) + i\sin(tx))\, dF_X(x)\right| \leq \int |\cos(tx) + i\sin(tx)|\, dF_X(x) = \int dF_X(x) = 1$$

2.
$$\sup_t |\phi_X(t+c) - \phi_X(t)| = \sup_t \left|\int (e^{i(t+c)x} - e^{itx})\, dF_X(x)\right| \leq \sup_t \left(\int \left|e^{itx}\right| \left|e^{icx-1}\right| dF_X(x)\right)$$
When $c \downarrow 0$, the supremum $\to 0$. Therefore, $\phi_X(t)$ is uniformly continuous.

□

Theorem 6.35. There are some properties of ϕ_X of a random variable X regarding derivatives and moments.

1. If $\phi_X^{(k)}(0)$ exists, then
$$\begin{cases} \mathbb{E}\,|X|^k < \infty, & k \text{ is even} \\ \mathbb{E}\,|X|^{k-1} < \infty, & k \text{ is odd} \end{cases}$$
2. If $\mathbb{E}\,|X|^k < \infty$, then $\phi_X^{(k)}(0)$ exists. We have
$$\phi_X(t) = \sum_{j=0}^{k} \frac{\phi_X^{(j)}(0)}{j!} t^j + o(t^k) = \sum_{j=0}^{k} \frac{\mathbb{E}X^j}{j!}(it)^j + o(t^k)$$

Proof.
We use the Taylor's Theorem.
$$\phi_X(t) = \sum_{j=0}^{k} \frac{\phi_X^{(j)}(0)}{j!} t^j + o(t^k) = \sum_{j=0}^{k} \frac{\mathbb{E}X^j}{j!}(it)^j + o(t^k)$$

1.
$$\phi_X^{(k)}(0) = i^k \mathbb{E}X^k$$
If k is even, we have $\phi_X^{(k)}(0) = (-1)^{\frac{k}{2}}\mathbb{E}X^k = (-1)^{\frac{k}{2}}\mathbb{E}\,|X|^k$ exists. Therefore, $\mathbb{E}\,|X|^k < \infty$.
If k is odd, we know that $\phi_X^{(k-1)}(0)$ exists if $\phi_X^{(k)}(0)$ exists.
Therefore, with $\phi_X^{(k-1)}(0) = (-1)^{\frac{k-1}{2}}\mathbb{E}X^{k-1} = (-1)^{\frac{k-1}{2}}\mathbb{E}\,|X|^{k-1}$, $\mathbb{E}\,|X|^{k-1} < \infty$.
2. Again using the formula in (1). We have
$$\frac{\phi_X^{(k)}(0)}{i^k} = \mathbb{E}X^k \leq \mathbb{E}\,|X|^k < \infty$$
Therefore, $\phi_X^{(k)}(0)$ exists. The formula can be obtained from the Taylor's theorem formula.

□

Theorem 6.36. If $X \perp\!\!\!\perp Y$, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$

Proof.
$$\phi_{X+Y}(t) = \mathbb{E}(e^{it(X+Y)}) = \mathbb{E}(e^{itX})\mathbb{E}(e^{itY}) = \phi_X(t)\phi_Y(t)$$

□

Again and again, we have a joint characteristic function.

Definition 6.37. Joint characteristic function (JCF) $\phi_{X,Y}$ of two random variables X, Y is given by

$$\phi_{X,Y}(s,t) = \mathbb{E}(e^{i(sX+tY)})$$

We have another way to prove that two random variables are independent.

Theorem 6.38. Two random variables X, Y are independent if and only if for all s and t,

$$\phi_{X,Y}(s,t) = \phi_X(s)\phi_Y(t)$$

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

$$\phi_{X,Y}(s,t) = \mathbb{E}(e^{i(sX+tY)}) = \mathbb{E}(e^{isX})\mathbb{E}(e^{itY}) = \phi_X(s)\phi_Y(t)$$

Currently, it is not suffice to prove the inverse. We will need to use a theorem later. (Example 6.25) □

Example 6.20. Let $X \sim \text{Bern}(p)$. We have

$$\phi_X(t) = \mathbb{E}(e^{itX}) = q + pe^{it}$$

Example 6.21. Let $X \sim \text{Bin}(n,p)$. We have

$$\phi_X(t) = (q + pe^{it})^n$$

Example 6.22. Let $X \sim \text{Exp}(1)$. We have

$$\phi_X(t) = \int e^{(it-1)x}\,dx = \frac{1}{1-it}$$

Example 6.23. Let $X \sim \text{Cauchy}$. We have

$$\phi_X(t) = e^{-|t|}$$

Example 6.24. Let $X \sim \text{N}(\mu, \sigma^2)$. Using the fact that for any $u \in \mathbb{C}$, not just in $\mathbb{R}$,

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-u)^2}{2\sigma^2}\right) dx = 1$$

We have

$$\begin{aligned}
\phi_X(t) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{itx} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - (2\mu + 2\sigma^2 it)x + \mu^2}{2\sigma^2}\right) dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(\mu + \sigma^2 it)^2 - \mu^2}{2\sigma^2}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{(x - (\mu + \sigma^2 it))^2}{2\sigma^2}\right) dx \\
&= \exp\left(\frac{\mu^2 + 2\sigma^2 i\mu t - \sigma^4 t^2 - \mu^2}{2\sigma^2}\right) \\
&= \exp\left(i\mu t - \frac{1}{2}\sigma^2 t^2\right)
\end{aligned}$$

Remark 6.38.1. We have a function called **cumulant generating function** defined by $\log \phi_X(t)$. Normal distribution is the only distribution we have learnt whose cumulant generating function has finite terms, which is:

$$\log \phi_X(t) = i\mu t - \frac{1}{2}\sigma^2 t^2$$

6.6 Inversion and continuity theorems

There are two major ways that characteristic functions are useful. One of them is that we can use characteristic function of a random variable to generate a probability density function of that random variable.

Theorem 6.39. (Fourier Inverse Transform for continuous case) If a random variable X is continuous with a PDF f_X and a CF ϕ_X, then

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_X(t)\,dt$$

at all point x which f_X is differentiable.
If X has a CDF F_X, then

$$F_X(b) - F_X(a) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_a^b e^{-itx}\phi_X(t)\,dx\,dt$$

Proof.
We give you a non-rigorous proof. Let

$$\begin{aligned} I(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{itx}\phi_X(t)\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\int_{-\infty}^{\infty} e^{ity} f_X(y)\,dy\,dt \\ I_\varepsilon(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\int_{-\infty}^{\infty} e^{ity} f_X(y)\,dy\,e^{-\frac{1}{2}\varepsilon^2 t^2}\,dt \end{aligned}$$

We want to show that $I_\varepsilon(x) \to I(x)$ when $\varepsilon \downarrow 0$.

$$\begin{aligned} I_\varepsilon(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\varepsilon^2 t^2 + i(y-x)t} f_X(y)\,dt\,dy \\ &= \frac{1}{\sqrt{2\pi\varepsilon^2}}\left(\frac{1}{\sqrt{2\pi\frac{1}{\varepsilon^2}}}\right)\int_{-\infty}^{\infty}\exp\left(-\frac{(y-x)^2}{2\varepsilon^2}\right) f_X(y)\int_{-\infty}^{\infty}\exp\left(-\frac{-\left(t - i\frac{y-x}{\varepsilon}\right)^2}{2\left(\frac{1}{\varepsilon^2}\right)}\right)dt\,dy \\ &= \frac{1}{\sqrt{2\pi\varepsilon^2}}\int_{-\infty}^{\infty}\exp\left(-\frac{(y-x)^2}{2\epsilon}\right) f_X(y)\,dy \end{aligned}$$

Let $Z \sim \mathrm{N}(0,1)$ and $Z_\varepsilon = \varepsilon Z$. $I_\varepsilon(x)$ is the PDF of $\varepsilon Z + X$.
Therefore, we can say that $f_{\varepsilon Z+X}(x) \to f_X(x)$ when $\varepsilon \downarrow 0$.
Note that this proof is not rigorous. □

Theorem 6.40. (Inversion Theorem) If a random variable X have a CDF F_X and a CF ϕ_X, we define $\overline{F}_X : \mathbb{R} \to [0,1]$ by

$$\overline{F}_X(x) = \frac{1}{2}\left(F_X(x) + F_X(x^-)\right)$$

Then for all $a \leq b$,

$$\overline{F}_X(b) - \overline{F}_X(a) = \int_{-\infty}^{\infty}\frac{e^{-iat} - e^{-ibt}}{2\pi it}\phi_X(t)\,dt$$

Remark 6.40.1. We can say $\overline{F}_X$ represents the average of limit going from two directions.

Example 6.25. With the Inversion Theorem, we can now prove Theorem 6.38.
Given two random variables X, Y. We want to first extend the Fourier Inverse Transform into multivariable case.
If $\phi_{X,Y}(s,t) = \phi_X(s)\phi_Y(t)$, then for any $a \leq b$ and $c \leq d$,

$$\begin{aligned} \overline{F}_{X,Y}(b,d) - \overline{F}_{X,Y}(b,c) - \overline{F}_{X,Y}(a,d) + \overline{F}_{X,Y}(a,c) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{(e^{-ias} - e^{-ibs})(e^{-ict} - e^{-idt})}{-4\pi^2 t^2}\phi_X(s)\phi_Y(t)\,ds\,dt \\ &= (\overline{F}_X(b) - \overline{F}_X(a))\int_{-\infty}^{\infty}\frac{e^{-ict} - e^{-idt}}{2\pi it}\phi_Y(t)\,dt \\ &= (\overline{F}_X(b) - \overline{F}_X(a))(\overline{F}_Y(d) - \overline{F}_Y(c)) \\ &= \overline{F}_X(b)\overline{F}_Y(d) - \overline{F}_X(b)\overline{F}_Y(c) - \overline{F}_X(a)\overline{F}_Y(d) + \overline{F}_X(a)\overline{F}_Y(c) \end{aligned}$$

From the definition of independent random variables, we prove that $X \perp\!\!\!\perp Y$ if $\phi_{X,Y}(s,t) = \phi_X(s)\phi_Y(t)$.

Another way is to evaluate the convergence of a sequence of cumulative distribution function.

Definition 6.41. (Convergence of distribution function sequence [Weak convergence]) A sequence of CDF $F_1, F_2, \cdots$ **converges** to a CDF F, written as $F_n \to F$, if at each point x where F is continuous,

$$F_n(x) \to F(x)$$

Example 6.26. Assume we have two sequences of CDF.

$$F_n(x) = \begin{cases} 0, & x < \frac{1}{n} \\ 1, & x \geq \frac{1}{n} \end{cases} \qquad G_n(x) = \begin{cases} 0, & x < -\frac{1}{n} \\ 1, & x \geq -\frac{1}{n} \end{cases}$$

If we have $n \to \infty$, we get

$$F(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases} \qquad G(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

This is problematic because $F(x)$ in this case is not a distribution function because it is not right-continuous. Therefore, it is needed to define the convergence so that both sequences $\{F_n\}$ and $\{G_n\}$ have the same limit.

We can modify a bit on the definition to say each distribution function in the sequence represents a different random variable.

Definition 6.42. (Convergence in distribution for random variables) Let $X, X_1, X_2, \cdots$ be a family of random variables with PDF $F, F_1, F_2, \cdots$, we say $X_n \to X$, written as $X_n \xrightarrow{D} X$ or $X_n \Rightarrow X$, if $F_n \to F$.

Remark 6.42.1. For this convergence definition, we do not care about the closeness of X_n and X as functions of ω.

Remark 6.42.2. Sometimes, we also write $X_n \Rightarrow F$ or $X_n \xrightarrow{D} F$.

With the definition, sequence of characteristic functions can be used to determine whether the sequence of cumulative distribution function converges.

Theorem 6.43. (**Lévy continuity theorem**) Suppose that $F_1, F_2, \cdots$ is a sequence of CDF with CF $\phi_1, \phi_2, \cdots$, then

1. If $F_n \to F$ for some CDF F with CF ϕ, then $\phi_n \to \phi$ pointwise.
2. If $\phi_n \to \phi$ pointwise for some CF ϕ, and ϕ is continuous at O $(t = 0)$, then ϕ is the CF of some CDF F and $F_n \to F$.

We have a more general definition of convergence.

Definition 6.44. (**Vague convergence**) Given a sequence of CDF $F_1, F_2, \cdots$. Suppose that $F_n(x) \to G(x)$ at all continuity point of G but G may not be a CDF. Then we say $F_n \to G$ **vaguely**, written as $F_n \xrightarrow{v} G$.

Example 6.27. If

$$F_n(x) = \begin{cases} 0, & x < \frac{1}{n} \\ \frac{1}{2}, & \frac{1}{n} \leq x < n \\ 1, & x \geq n \end{cases} \qquad G(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2}, & x \geq 0 \end{cases}$$

We can see that $F_n \xrightarrow{v} G$ if $n \to \infty$ and G is not a CDF.

Remark 6.44.1. In Theorem 6.43 (2), the statement that ϕ is continuous at O can be replaced by any of the following statements:

1. $\phi(t)$ is a continuous function of t
2. $\phi(t)$ is a CF of some CDF
3. The sequence $\{F_n\}_{n=1}^{\infty}$ is tight, i.e. for all $\epsilon > 0$, there exists $M_\epsilon > 0$ such that

$$\sup_n (F_n(-M_\epsilon) + 1 - F_n(M_\epsilon)) \leq \epsilon$$

Example 6.28. Let $X_n \sim \mathrm{N}(0, n^2)$ and let ϕ_n be the CF of X_n. Then

$$\phi_n(t) = \exp\left(-\frac{1}{2}n^2t^2\right) \to \phi(t) = \begin{cases} 0, & t \neq 0 \\ 1, & t = 0 \end{cases}$$

6.7 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 6.45. (**Weak Law of Large Numbers** [WLLN]) Let $X_1, X_2, \cdots$ be i.i.d. random variables. Assume that $\mathbb{E}\,|X_1| < \infty$ and $\mathbb{E}X_1 = \mu$. Let $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{1}{n}S_n \xrightarrow{D} \mu$$

Proof.
We recall the Taylor expansion of $\phi_\xi(s)$ at 0. If $\mathbb{E}\,|\xi|^k < \infty$ and s is small, then

$$\phi_\zeta(s) = \sum_{j=0}^k \frac{\mathbb{E}\xi^j}{j!}(is)^j + o(s^k)$$

For any $t \in \mathbb{R}$, let $\phi_{X_1}(s) = \mathbb{E}(e^{isX_1})$.

$$\begin{aligned}\phi_n(t) = \mathbb{E}\left(\exp\left(\frac{it}{n}S_n\right)\right) = \mathbb{E}\left(\prod_{i=1}^n \exp\left(\frac{itX_i}{n}\right)\right) = \left(\mathbb{E}\left(\exp\left(\frac{itX_1}{n}\right)\right)\right)^n = \left(\phi_{X_1}\left(\frac{t}{n}\right)\right)^n &= \left(1 + \frac{it}{n}\mathbb{E}X_1 + o\left(\frac{t}{n}\right)\right)^n \\ &= \left(1 + \frac{i\mu t}{n} + o\left(\frac{t}{n}\right)\right)^n \\ &\to e^{i\mu t}\end{aligned}$$

By Lévy continuity theorem, we get that $\frac{1}{n}S_n \xrightarrow{D} \mu$. □

Theorem 6.46. (**Central Limit Theorem** [CLT]) Let $X_1, X_2, \cdots$ be i.i.d. random variables with $\mathbb{E}\,|X_1|^2 < \infty$ and $\mathbb{E}X_1 = \mu$, $\mathrm{Var}(X_1) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{1}{\sigma}\sqrt{n}\left(\frac{1}{n}S_n - \mu\right) = \frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{D} \mathrm{N}(0,1)$$

Proof.
Let $Y_i = \frac{X_i - \mu}{\sigma}$. We have $\mathbb{E}Y_i = 0$ and $\mathrm{Var}(Y_i) = 1$.

$$\begin{aligned}\frac{S_n - n\mu}{\sqrt{n}\sigma} &= \sum_{i=1}^n \frac{1}{\sqrt{n}}\frac{X_i - \mu}{\sigma} = \sum_{i=1}^n \frac{Y_i}{\sqrt{n}} \\ \phi_n(t) &= \mathbb{E}\left(\exp\left(it\sum_{\ell-1}^n \frac{Y_\ell}{\sqrt{n}}\right)\right) \\ &= \left(\mathbb{E}\left(\exp\left(\frac{itY_1}{\sqrt{n}}\right)\right)\right)^n \\ &= \left(\phi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right)^n \\ &= \left(1 + \frac{it}{\sqrt{n}}\mathbb{E}Y_1 + \frac{1}{2}\left(\frac{it}{\sqrt{n}}\right)^2\mathbb{E}(Y_i^2) + o\left(\frac{t^2}{n}\right)\right)^n && \text{(Taylor expansion)} \\ &= \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \\ &\to e^{-\frac{1}{2}t^2}\end{aligned}$$

By Lévy continuity theorem, $\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{D} \mathrm{N}(0,1)$. □

Central Limit Theorem can be generalized in several directions, one of which concerns about independent random variables instead of i.i.d. random variables.

Theorem 6.47. Let $X_1, X_2, \cdots$ be independent random variables satisfying $\mathbb{E}X_i = 0$, $\mathrm{Var}(X_i) = \sigma_i^2$, $\mathbb{E}|X_i|^3 < \infty$ and such that

$$\frac{1}{(\sigma(n))^3}\sum_{i=1}^{n}\mathbb{E}\left|X_i^3\right| \to 0 \text{ as } n \to \infty \tag{*}$$

where $(\sigma(n))^2 = \mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n}\sigma_i^2$. Then

$$\frac{1}{\sigma(n)}\sum_{i=1}^{n} X_i \xrightarrow{D} \mathrm{N}(0,1)$$

Remark 6.47.1. The condition (*) means that none of the random variables X_i can be significant in the sum S_n.

$$\frac{1}{(\sigma(n))^3}\sum_{i=1}^{n}|X_i|^3 \lesssim \frac{1}{\sigma(n)}\max_{i=1,2,\cdots,n}|X_i|\left(\frac{1}{(\sigma(n))^2}\right)\sum_{i=1}^{n}(X_i)^2 \approx \frac{1}{\sigma(n)}\max_{i=1,2,\cdots,n}|X_i| \to 0$$

This theorem is a special case of Central Limit Theorem. It is more about the sum of Bernoulli random variables converges to a normal distribution.

Theorem 6.48. (**De Moivre-Laplace Limit Theorem**) Suppose that $X \sim \mathrm{Bin}(n,p)$. Then for any $a < b$, as $n \to \infty$,

$$\mathbb{P}\left(a < \frac{X - np}{\sqrt{np(1-p)}} \le b\right) \to \Phi(b) - \Phi(a)$$

Proof.
Before we start the proof, we need to know about the Stirling's formula:

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Our target is to transform the PMF of Binomial random variable into the PDF of standard normal distribution. For $0 \le k \le n$,

$$\begin{aligned}
\binom{n}{k}p^k(1-p)^{n-k} &= \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \\
&\sim \sqrt{\frac{n}{2\pi k(n-k)}}\frac{n^n}{k^k(n-k)^{n-k}}p^k(1-p)^{n-k} && \text{(Using Stirling's formula)} \\
&= \sqrt{\frac{n}{2\pi k(n-k)}}\left(\frac{np}{k}\right)^k\left(\frac{n(1-p)}{n-k}\right)^{n-k} \\
&\sim \sqrt{\frac{1}{2\pi np(1-p)}}\exp\left(-k\ln\left(\frac{k}{np}\right) + (k-n)\ln\left(\frac{n-k}{n(1-p)}\right)\right) && \left(\tfrac{k}{n} \to p\right)
\end{aligned}$$

We know that $\mathbb{E}X = np$ and $\mathrm{Var}(X) = np(1-p)$.
For any integer k we choose between 0 and n, there exists an arbitrary finite point c such that $k = np + c\sqrt{np(1-p)}$.
To simplify, let $q = 1-p$. Using the Taylor series of $\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + o(x^3)$, we get:

$$\begin{aligned}
\binom{n}{k}p^k q^{n-k} &\sim \sqrt{\frac{1}{2\pi npq}}\exp\left((-np - c\sqrt{npq})\ln\left(\frac{np + c\sqrt{npq}}{np}\right) + (np + c\sqrt{npq} - n)\ln\left(\frac{n - np - c\sqrt{npq}}{nq}\right)\right) \\
&= \frac{1}{\sqrt{2\pi npq}}\exp\left((-np - c\sqrt{npq})\ln\left(1 + c\sqrt{\frac{q}{np}}\right) + (c\sqrt{npq} - nq)\ln\left(1 - c\sqrt{\frac{p}{nq}}\right)\right) \\
&= \frac{1}{\sqrt{2\pi npq}}\exp\left((-np - c\sqrt{npq})\left(c\sqrt{\frac{q}{np}} - \frac{c^2q}{2np} + o(n^{-1})\right) + (c\sqrt{npq} - nq)\left(-c\sqrt{\frac{p}{nq}} - \frac{c^2p}{2nq} + o(n^{-1})\right)\right) \\
&= \frac{1}{\sqrt{2\pi npq}}\exp\left((-c\sqrt{npq} - c^2q + \frac{1}{2}c^2q + o(1)) + (-c^2p + c\sqrt{npq} - \frac{1}{2}c^2p + o(1))\right) \\
&\sim \frac{1}{\sqrt{2\pi npq}}\exp\left(\frac{1}{2}c^2\right) \\
&= \frac{1}{\sqrt{2\pi npq}}e^{-\frac{(k-np)^2}{2npq}}
\end{aligned}$$

Therefore, as $n \to \infty$, $\frac{X-np}{\sqrt{np(1-p)}} \xrightarrow{D} N(0,1)$ and the theorem is proven. □

Chapter 7

Convergence of random variables

We have mentioned the convergence in distribution in Chapter 5. However, this is not the only important type of convergence mode of random variables. In this chapter, we will introduce some other convergence modes.

7.1 Modes of convergence

Many modes of convergence of a sequence of random variables will be discussed.
Let us recall the convergence mode of real function. Let $f, f_1, f_2, \cdots : [0,1] \to \mathbb{R}$.

1. Pointwise convergence
   We say $f_n \to f$ pointwise if for all $x \in [0,1]$,
   $$f_n(x) \to f(x) \text{ as } n \to \infty$$
2. Convergence in norm $\|\cdot\|$
   We say $f_n \to f$ in norm $\|\cdot\|$ if
   $$\|f_n - f\| \to 0 \text{ as } n \to \infty$$
3. Convergence in Lebesgue (uniform) measure
   We say $f_n \to f$ in uniform measure μ if for all $\epsilon > 0$,
   $$\mu\left(\{x \in [0,1] : |f_n(x) - f(x)| > \epsilon\}\right) \to 0 \text{ as } n \to \infty$$

We can use these definitions to define convergence modes of random variables.

Definition 7.1. (**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 7.1.1. $X_n \xrightarrow{\text{a.s.}} X$ almost surely is an adaptation to the pointwise convergence for function.

Remark 7.1.2. Very often, we also call almost surely convergence:

1. $X_n \to X$ almost everywhere ($X_n \xrightarrow{\text{a.e.}} X$)
2. $X_n \to X$ with probability 1 ($X_n \to X$ w.p. 1)

Definition 7.2. (Convergence in r-th mean) Let $r \geq 1$. We say $X_n \to X$ **in r-th mean**, written as $X_n \xrightarrow{r} X$, if

$$\mathbb{E}\,|X_n - X|^r \to 0 \text{ as } n \to \infty$$

Example 7.1. If $r = 1$, we say $X_n \to X$ in mean or expectation.
If $r = 2$, we say $X_n \to X$ in mean square.

Definition 7.3. (Convergence in probability) We say $X_n \to X$ **in probability**, 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 7.4. (Convergence in distribution) We say that $X_n \to X$ **in distribution**, written as $X_n \xrightarrow{D} X$, if at continuity point of $\mathbb{P}(X \le x)$,

$$F_n(x) = \mathbb{P}(X_n \le x) \to \mathbb{P}(X \le x) = F(x) \text{ as } n \to \infty$$

Before we tackle the relationships between different convergence mode, we first need to introduce some formulas.

Lemma 7.5. (Markov's inequality) If X is any random variables with finite mean, then for all $a > 0$,

$$\mathbb{P}(|X| \ge a) \le \frac{\mathbb{E}|X|}{a}$$

Proof.

$$\mathbb{P}(|X| \ge a) = \mathbb{E}(\mathbf{1}_{|X|\ge a}) \le \mathbb{E}\left(\frac{|X|}{a}\mathbf{1}_{|X|>a}\right) \le \frac{\mathbb{E}|X|}{a}$$

□

Remark 7.5.1. For any non-negative function φ that is increasing on $[0, \infty)$,

$$\mathbb{P}(|X| \ge a) = \mathbb{P}(\varphi(|X|) \ge \varphi(a)) \le \frac{\mathbb{E}(\varphi(|X|))}{\varphi(a)}$$

Following inequality needs Hölder inequality (In Appendix C) in order to be proven. Therefore, we will not prove it here.

Lemma 7.6. (Lyapunov's inequality) Let Z be any random variables. For all $r \ge s > 0$,

$$(\mathbb{E}|Z|^s)^{\frac{1}{s}} \le (\mathbb{E}|Z|^r)^{\frac{1}{r}}$$

We also need to know how we can obtain almost sure convergence.

Lemma 7.7. Let

$$A_n(\varepsilon) = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\} \qquad B_m(\varepsilon) = \bigcup_{n=m}^{\infty} A_n(\varepsilon)$$

We have

1. $X_n \xrightarrow{\text{a.s.}} X$ if and only if $\lim_{m\uparrow\infty} \mathbb{P}(B_m(\varepsilon)) = 0$ for all $\varepsilon > 0$
2. $X_n \xrightarrow{\text{a.s.}} X$ if $\sum_{n=1}^{\infty} \mathbb{P}(A_n(\varepsilon)) < \infty$ for all $\varepsilon > 0$

Proof.

1. We denote $C = \{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}$.
If $\omega \in C$, that means for all $\varepsilon > 0$, there exists $n_0 > 0$ such that $|X_n(\omega) - X(\omega)| \le \varepsilon$ for all $n \ge n_0$.
This also means that for all $\varepsilon > 0$, $|X_n(\omega) - X(\omega)| > \varepsilon$ for finitely many n.
If $\omega \in C^{\complement}$, that means that for all $\varepsilon > 0$, $|X_n(\omega) - X(\omega)| > \varepsilon$ for infinitely many n. ($\omega \in \bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty} A_n(\varepsilon)$)
Therefore,
$$C^{\complement} = \bigcup_{\varepsilon>0}\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty} A_n(\varepsilon)$$
If $\mathbb{P}(C^{\complement}) = 0$, then for all $\varepsilon > 0$,
$$\mathbb{P}\left(\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty} A_n(\varepsilon)\right) = 0$$
We can also find that
$$\mathbb{P}\left(\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty} A_n(\varepsilon)\right) = 0 \qquad \Longrightarrow \qquad \mathbb{P}(C^{\complement}) = \mathbb{P}\left(\bigcup_{\varepsilon>0}\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty} A_n(\varepsilon)\right) = \mathbb{P}\left(\bigcup_{k=1}^{\infty}\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty} A_n\left(\frac{1}{k}\right)\right) = 0$$
Therefore, $X_n \xrightarrow{\text{a.s.}} X$ if and only if $\lim_{m\uparrow\infty} \mathbb{P}(B_m(\varepsilon)) = 0$ for all $\varepsilon > 0$
2. From (1), for all $\varepsilon > 0$,
$$\sum_{n=1}^{\infty} \mathbb{P}(A_n(\varepsilon)) < \infty \implies \lim_{m\to\infty}\sum_{n=m}^{\infty} \mathbb{P}(A_n(\varepsilon)) = 0 \implies \lim_{m\to\infty} \mathbb{P}(B_m(\varepsilon)) = 0 \implies (X_n \xrightarrow{\text{a.s.}} X)$$

□

Lemma 7.8. There exist sequences that

1. converge almost surely but not in mean
2. converge in mean but not almost surely

Proof.

1. We consider

$$X_n = \begin{cases} n^3, & \text{Probability } = n^{-2} \\ 0, & \text{Probability } = 1 - n^{-2} \end{cases}$$

By applying Lemma 7.7, for some $\varepsilon > 0$.

$$\mathbb{P}(|X_n(\omega) - X(\omega)| > \varepsilon) = \frac{1}{n^2} \qquad \sum_{n=1}^{\infty} \mathbb{P}(|X_n(\omega) - X(\omega)| > \varepsilon) < \infty$$

Therefore, the sequence converges almost surely. However,

$$\mathbb{E}\,|X_n - X| = n^3 \left(\frac{1}{n^2}\right) = n \to \infty$$

Therefore, the sequence does not converge in mean.

2. We consider

$$X_n = \begin{cases} 1, & \text{Probability } = n^{-1} \\ 0, & \text{Probability } = 1 - n^{-1} \end{cases}$$

In mean, as $n \to \infty$ we have

$$\mathbb{E}\,|X_n - X| = 1\left(\frac{1}{n}\right) = \frac{1}{n} \to 0$$

However, by applying Lemma 7.7, if $\varepsilon \in (0, 1)$, for all n

$$\begin{aligned} \mathbb{P}(B_m(\varepsilon)) &= 1 - \lim_{r\to\infty} \mathbb{P}(X_n = 0 \text{ for all } n \text{ such that } m \le n \le r) \\ &= 1 - \lim_{r\to\infty} \prod_{i=m}^{r} \frac{i-1}{i} \\ &= 1 - \lim_{r\to\infty} \frac{m-1}{r} \to 1 \neq 0 \end{aligned}$$

Therefore, the sequence does not converge almost surely.

□

We can now deduce the following implications. Roughly speaking, convergence in distribution is the weakest among all convergence modes, since it only cares about the distribution of X_n.

Theorem 7.9. The following implications hold:

1. (a) $(X_n \xrightarrow{\text{a.s.}} X) \implies (X_n \xrightarrow{\mathbb{P}} X)$
   (b) $(X_n \xrightarrow{r} X) \implies (X_n \xrightarrow{\mathbb{P}} X)$
   (c) $(X_n \xrightarrow{\mathbb{P}} X) \implies (X_n \xrightarrow{D} X)$
2. If $r \ge s \ge 1$, then $(X_n \xrightarrow{r} X) \implies (X_n \xrightarrow{s} X)$
3. No other implications holds in general.

Proof.

1. (a) From Lemma 7.7, for all $\varepsilon > 0$,
$$\mathbb{P}(A_m(\varepsilon)) \leq \mathbb{P}\left(\bigcup_{n=m}^{\infty} A_n(\varepsilon)\right) = \mathbb{P}(B_m(\varepsilon)) \to 0$$
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)
$$\begin{aligned}
&\mathbb{P}(X_n \leq x) = \mathbb{P}(X_n \leq x, X \leq x+\varepsilon) + \mathbb{P}(X_n \leq x, X > x+\varepsilon) \leq \mathbb{P}(X \leq x+\varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon) \\
&\mathbb{P}(X \leq y) \leq \mathbb{P}(X_n \leq y+\varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon) \\
&\mathbb{P}(X_n \leq x) \geq \mathbb{P}(X \leq x-\varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \qquad (y = x - \varepsilon)
\end{aligned}$$
Since $X_n \xrightarrow{\mathbb{P}} X$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ for all $\varepsilon > 0$. Therefore,
$$\mathbb{P}(X \leq x-\varepsilon) \leq \liminf_{n\to\infty} \mathbb{P}(X_n \leq x) \leq \limsup_{n\to\infty} \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x+\varepsilon)$$
By having $\varepsilon \downarrow 0$,
$$\mathbb{P}(X \leq x) \leq \liminf_{n\to\infty} \mathbb{P}(X_n \leq x) \leq \limsup_{n\to\infty} \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x)$$
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 $r \geq s$,
$$\mathbb{E}\,|X_n - X|^s \leq (\mathbb{E}\,|X_n - X|^r)^{\frac{s}{r}} \to 0$$

3. Let $\Omega = \{H, T\}$ and $\mathbb{P}(H) = \mathbb{P}(T) = \frac{1}{2}$. Let
$$X_{2m}(\omega) = \begin{cases} 1, & \omega = H \\ 0, & \omega = T \end{cases} \qquad X_{2m+1}(\omega) = \begin{cases} 0, & \omega = H \\ 1, & \omega = T \end{cases}$$
Since $F(x)$ and $F_n(x)$ for all n are all the same, $X_n \xrightarrow{D} X$. However, for $\varepsilon \in [0,1]$, $\mathbb{P}(|X_n - X| > \varepsilon) \not\to 0$.
Therefore, $(X_n \xrightarrow{D} X) \not\Longrightarrow (X_n \xrightarrow{\mathbb{P}} X)$.
Let $r = 1$ and
$$X_n = \begin{cases} n, & \text{probability } = \frac{1}{n} \\ 0, & \text{probability } = 1 - \frac{1}{n} \end{cases} \qquad X = 0$$
We get that $\mathbb{P}(|X_n - X| > \varepsilon) = \frac{1}{n} \to 0$. However, $\mathbb{E}\,|X_n - X| = n\left(\frac{1}{n}\right) = 1 \not\to 0$. Therefore, $(X_n \xrightarrow{\mathbb{P}} X) \not\Longrightarrow (X_n \xrightarrow{r} X)$.
Let $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}([0,1])$ and $\mathbb{P}$ be uniform.
Let I_i be such that $I_{\frac{1}{2}m(m-1)+1}, I_{\frac{1}{2}m(m-1)+2}, \cdots, I_{\frac{1}{2}m(m-1)+m}$ is a partition of $[0,1]$ for all m.
We have $I_1 = [0,1], I_2 \cup I_3 = [0,1], \cdots$. Let
$$X_n(\omega) = \mathbf{1}_{I_n(\omega)} = \begin{cases} 1, & \omega \in I_n \\ 0, & \omega \in I_n^{\complement} \end{cases} \qquad X(\omega) = 0 \text{ for all } \omega \in \Omega$$
For all $\varepsilon \in [0,1]$, $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(I_n) = \frac{1}{n} \to 0$ for some n if $n \to \infty$.
However, for any given $\omega \in \Omega$, although 1 becomes less often due to decreasing probability, it never dies out.
Therefore, $X_n(\omega) \not\to 0 = X(\omega)$ and $\mathbb{P}(\{\omega \in \Omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}) = 0$, and thus, $(X_n \xrightarrow{\mathbb{P}} X) \not\Longrightarrow (X_n \xrightarrow{\text{a.s.}} X)$.
If $r \geq s \geq 1$, let
$$X_n = \begin{cases} n, & \text{probability } = n^{-\left(\frac{r+s}{2}\right)} \\ 0, & \text{probability } = 1 - n^{-\left(\frac{r+s}{2}\right)} \end{cases} \qquad X = 0$$
$$\mathbb{E}\,|X_n - X|^s = n^s\left(n^{-\left(\frac{r+s}{2}\right)}\right) = n^{\frac{s-r}{2}} \to 0 \qquad \mathbb{E}\,|X_n - X|^r = n^r\left(n^{-\left(\frac{r+s}{2}\right)}\right) = n^{\frac{r-s}{2}} \to \infty$$
Therefore, if $r \geq s \geq 1$, $(X_n \xrightarrow{s} X) \not\Longrightarrow (X_n \xrightarrow{r} X)$.
We have proven that $(X_n \xrightarrow{\text{a.s.}} X) \not\Longrightarrow (X_n \xrightarrow{r} X)$ and $(X_n \xrightarrow{r} X) \not\Longrightarrow (X_n \xrightarrow{\text{a.s.}} X)$ in Lemma 7.8.

□

We can easily obtain this lemma.

Lemma 7.10. The following implications hold:

1. $(X_n \xrightarrow{1} X) \implies (X_n \xrightarrow{\mathbb{P}} X)$

Proof.
Just use the Theorem 7.9 with $r = 1$ and you get the answer. □

Some of the implications does not hold in general but they hold if we apply some restrictions.

Theorem 7.11. (Partial converse statements) The following implications hold:

1. If $X_n \xrightarrow{D} c$, where c is a constant, then $X_n \xrightarrow{\mathbb{P}} c$.
2. If $X_n \xrightarrow{\mathbb{P}} X$ and $\mathbb{P}(|X_n| \leq k) = 1$ for all n with some fixed constant $k > 0$, then $X_n \xrightarrow{r} X$ for all $r \geq 1$.

Proof.

1. Since $X_n \xrightarrow{D} X$, $\mathbb{P}(X_n \leq x) \to \mathbb{P}(c \leq x)$ as $n \to \infty$. For all $\varepsilon > 0$,
$$\mathbb{P}(|X_n - c| \geq \varepsilon) = \mathbb{P}(X_n \leq c - \varepsilon) + \mathbb{P}(X_n \geq c + \varepsilon) = \mathbb{P}(X_n \leq c - \varepsilon) + 1 - \mathbb{P}(X_n < c + \varepsilon)$$
We can get that $\mathbb{P}(X_n \leq c - \varepsilon) \to \mathbb{P}(c \leq c - \varepsilon) = 0$. For $\mathbb{P}(X_n < c + \varepsilon)$,
$$\mathbb{P}\left(X_n \leq c + \frac{\varepsilon}{2}\right) \leq \mathbb{P}(X_n < c + \varepsilon) \leq \mathbb{P}(X_n \leq c + 2\varepsilon)$$
$$\mathbb{P}\left(X_n \leq c + \frac{\varepsilon}{2}\right) \to \mathbb{P}\left(c \leq c + \frac{\varepsilon}{2}\right) = 1 \qquad \mathbb{P}(X_n \leq c + 2\varepsilon) \to \mathbb{P}(c \leq c + 2\varepsilon) = 1$$
Therefore, $\mathbb{P}(X_n < c + \varepsilon) \to 1$. We have
$$\mathbb{P}(|X_n - c| \geq \varepsilon) \to 0 + 1 - 1 = 0$$
Therefore, $X_n \xrightarrow{\mathbb{P}} c$.
2. Since $X_n \xrightarrow{\mathbb{P}} X$, $X_n \xrightarrow{D} X$. We have $\mathbb{P}(|X_n| \leq k) \to \mathbb{P}(|X| \leq k) = 1$.
Therefore, for all $\varepsilon > 0$, if $|X_n - X| \leq \varepsilon$, $|X_n - X| \leq |X_n| + |X| \leq 2k$.
$$\begin{aligned}\mathbb{E}\,|X_n - X|^r &= \mathbb{E}\left(|X_n - X|^r \mathbf{1}_{|X_n - X| \leq \varepsilon}\right) + \mathbb{E}\left(|X_n - X|^r \mathbf{1}_{|X_n - X| > \varepsilon}\right)\\ &\leq \varepsilon^r \mathbb{E}\left(\mathbf{1}_{|X_n - X| \leq \varepsilon}\right) + (2k)^r \mathbb{E}\left(\mathbf{1}_{|X_n - X| > \varepsilon}\right)\\ &\leq \varepsilon^r + ((2k)^r - \varepsilon^r)\mathbb{P}(|X_n - X| > \varepsilon)\end{aligned}$$
Since $X_n \xrightarrow{\mathbb{P}} X$, as $n \to \infty$, $\mathbb{E}\,|X_n - X|^r \to \varepsilon^r$. If we send $\varepsilon \downarrow 0$, $\mathbb{E}\,|X_n - X|^r \to 0$ and therefore $X_n \xrightarrow{r} X$.

□

Note that any sequence $\{X_n\}$ which satisfies $X_n \xrightarrow{\mathbb{P}} X$ necessarily contains a subsequence $\{X_{n_i} : 1 \leq i < \infty\}$ which converges almost surely.

Theorem 7.12. If $X_n \xrightarrow{\mathbb{P}} X$, then there exists a non-random increasing sequence of integers $n_1, n_2, \cdots$ such that as $i \to \infty$,
$$X_{n_i} \xrightarrow{\text{a.s.}} X$$

Proof.
Since $X_n \xrightarrow{\mathbb{P}} X$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$ for all $\varepsilon > 0$.
We can pick an increasing sequence $n_1, n_2, \cdots$ of positive integers such that
$$\mathbb{P}(|X_{n_i} - X| > i^{-1}) \leq i^{-2}$$
For any $\varepsilon > 0$,
$$\sum_{i > \varepsilon^{-1}} \mathbb{P}(|X_{n_i} - X| > \varepsilon) \leq \sum_{i > \varepsilon^{-1}} \mathbb{P}(|X_{n_i} - X| > i^{-1}) \leq \sum_i i^{-2} < \infty$$
By Lemma 7.7, we get the $X_{n_i} \xrightarrow{\text{a.s.}} X$ as $i \to \infty$ □

7.2 Other versions of Weak Law of Large Numbers

We can revisit and introduce some other versions of weak law of large numbers and their applications.

Theorem 7.13. (L^2-WLLN) Let $X_1, X_2, \cdots, X_n$ be uncorrelated random variables with $\mathbb{E}X_i = \mu$, $\mathrm{Var}(X_i) \leq c < \infty$ for all i. Let $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{S_n}{n} \xrightarrow{2} \mu$$

Proof.

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

Therefore, $\frac{S_n}{n} \xrightarrow{2} \mu$. □

Remark 7.13.1. From this theorem, we can immediately find that

$$\left(\frac{S_n}{n} \xrightarrow{2} \mu\right) \implies \left(\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu\right)$$

Remark 7.13.2. Note that in the i.i.d. case, we do not require the existence of variance.

There are wide range of applications for just Weak Law of Large Numbers.

Example 7.2. (Bernstein approximation) Let f be continuous on $[0, 1]$ and let

$$f_n(x) = \sum_{m=0}^n \binom{n}{m} x^n (1-x)^{n-m} f\left(\frac{m}{n}\right) \qquad \text{(Bernstein polynomial)}$$

We want to show that as $n \to \infty$.

$$\sup_{x\in[0,1]} |f_n(x) - f(x)| \to 0$$

Remark 7.13.3. Let $x \in [0, 1]$. To better approach this question, we can let $X_{1,x}, X_{2,x}, \cdots, X_{n,x} \sim \mathrm{Bern}(x)$ be i.i.d. random variables. Let $S_{n,x} = \sum_{i=1}^n X_{i,x} \sim \mathrm{Bin}(n, x)$.

$$\mathbb{P}(S_{n,x} = m) = \binom{n}{m} x^m (1-x)^{n-m}$$

$$f_n(x) = \sum_{m=0}^n \mathbb{P}(S_{n,x} = m) f\left(\frac{m}{n}\right) = \mathbb{E}\left(f\left(\frac{S_{n,x}}{n}\right)\right)$$

We know that by WLLN, $\frac{S_{n,x}}{n} \xrightarrow{\mathbb{P}} x$

Remark 7.13.4. (Continuous mapping theorem) Let f be an uniformly continuous function. For all $\varepsilon > 0$, there exists δ_ε such that

$$\text{if} \quad \left|\frac{S_{n,x}}{n} - x\right| \leq \delta_\varepsilon \quad \text{then} \quad \left|f\left(\frac{S_{n,x}}{n}\right) - f(x)\right| \leq \varepsilon$$

We can obtain by converse,

$$\mathbb{P}\left(\omega \in \Omega : \left|f\left(\frac{S_{n,x}(\omega)}{n}\right) - f(x)\right| > \varepsilon\right) \leq \mathbb{P}\left(\omega \in \Omega : \left|\frac{S_{n,x}(\omega)}{n} - x\right| > \delta_\varepsilon\right) \to 0$$

From this, we can find that $f\left(\frac{S_{n,x}}{n}\right) \xrightarrow{\mathbb{P}} f(x)$.
Note that for non-uniformly continuous function, it is a bit more complicated. It is best if you do some searching on that.

Example 7.3. By obtaining that $f\left(\frac{S_{n,x}}{n}\right) \xrightarrow{\mathbb{P}} f(x)$, since there exists a number M such that $\|f\|_\infty \le M$ (due to f being continuous at $[0,1]$),

$$\begin{aligned}\left|\mathbb{E}\left(f\left(\frac{S_{n,x}}{n}\right)\right) - f(x)\right| \le \mathbb{E}\left|f\left(\frac{S_{n,x}}{n}\right) - f(x)\right| &= \mathbb{E}\left(\left|f\left(\frac{S_{n,x}}{n}\right) - f(x)\right| \mathbf{1}_{\left|\frac{S_{n,x}}{n} - x\right| \le \delta_\varepsilon}\right) + \mathbb{E}\left(\left|f\left(\frac{S_{n,x}}{n}\right) - f(x)\right| \mathbf{1}_{\left|\frac{S_{n,x}}{n} - x\right| > \delta_\varepsilon}\right) \\ &\le \varepsilon + 2M\mathbb{P}\left(\left|\frac{S_{n,x}}{n} - x\right| > \delta_\varepsilon\right)\end{aligned}$$

$$\begin{aligned}\sup_{x\in[0,1]}\left|\mathbb{E}\left(f\left(\frac{S_{n,x}}{n}\right)\right) - f(x)\right| &= \varepsilon + 2M \sup_{x\in[0,1]}\left(\mathbb{P}\left(\left|\frac{S_{n,x} - nx}{n}\right| > \delta_\varepsilon\right)\right) \\ &\le \varepsilon + 2M \sup_{x\in[0,1]}\left(\frac{\mathbb{E}\left|S_{n,x} - nx\right|^2}{n^2\delta_\varepsilon^2}\right) && \text{(Markov's inequality and Lyapunov's inequality)} \\ &\le \varepsilon + 2M \sup_{x\in[0,1]}\left(\frac{\mathrm{Var}(S_{n,x})}{n^2\delta_\varepsilon^2}\right) = \varepsilon + 2M \sup_{x\in[0,1]}\left(\frac{x(1-x)}{n\delta_\varepsilon^2}\right) && (\mathbb{E}S_{n,x} = nx) \\ &\le \varepsilon + \frac{M}{2n\delta_\varepsilon^2} \\ \limsup_{n\to\infty}\sup_{x\in[0,1]}\left|\mathbb{E}\left(f\left(\frac{S_{n,x}}{n}\right)\right) - f(x)\right| &\le \varepsilon \to 0\end{aligned}$$

Therefore, we can find that $\sup_{x\in[0,1]}|f_n(x) - f(x)| \to 0$ as $n \to \infty$.

Example 7.4. (Borel's geometric concentration) Let μ_n be the uniform probability measure on the n-dimensional cube $[-1,1]^n$.
Let $\mathcal{H}$ be a hyperplane that is orthogonal to a principal diagonal of $[-1,1]$ ($\mathcal{H} = (1,\cdots,1)^\perp$).
Let $\mathcal{H}_r = \{x \in [-1,1]^n : \mathrm{dist}(x:\mathcal{H}) \le r\}$.
Then for any given $\varepsilon > 0$, $\mu_n(\mathcal{H}_{\varepsilon\sqrt{n}}) \to 1$ as $n \to \infty$.
We can prove this by letting $X_1, X_2, \cdots \sim \mathrm{U}[-1,1]$ be i.i.d. random variables and $\mathbb{E}X_i = 0$. Let $X = (X_1, X_2, \cdots, X_n)$.
For all $B \in [-1,1]^n$, $\mu_n(B) = \mathbb{P}(X \in B) = \mathbb{P} \circ X^{-1}(B)$.

$$\begin{aligned}\mu_n(\mathcal{H}_{\varepsilon\sqrt{n}}) &= \mathbb{P}(\mathrm{dist}(X,\mathcal{H}) \le \varepsilon\sqrt{n}) \\ &= \mathbb{P}\left(\frac{|\langle X, (1,\cdots,1)\rangle|}{\|(1,\cdots,1)\|_2} \le \varepsilon\sqrt{n}\right) \\ &= \mathbb{P}\left(\left|\frac{\sum_{i=1}^n X_i}{n}\right| \le \varepsilon\right) \\ &= \mathbb{P}\left(\left|\frac{S_n}{n} - \mathbb{E}X_1\right| \le \varepsilon\right) \\ &\to 1 && \text{(WLLN)}\end{aligned}$$

We do not necessarily need to stick to a given sequence of random variables $X_1, X_2, \cdots$ in Law of Large Numbers.

Theorem 7.14. (WLLN for triangular array) Let $\{X_{n,j}\}_{1\le j\le n<\infty}$ be a triangular array. Let $S_n = \sum_{i=1}^n X_{n,i}$, $\mu_n = \mathbb{E}S_n$ and $\sigma_n^2 = \mathrm{Var}(S_n)$. Suppose that for some sequences b_n,

$$\frac{\sigma_n^2}{b_n^2} = \mathbb{E}\left(\frac{S_n - \mu_n}{b_n}\right)^2 \to 0$$

Then we have

$$\frac{S_n - \mu_n}{b_n} \xrightarrow{\mathbb{P}} 0$$

Proof.

$$\mathbb{E}\left(\frac{S_n - \mu_n}{b_n}\right)^2 = \frac{\mathrm{Var}(S_n)}{b_n^2} \to 0$$

Therefore, $\frac{S_n - \mu_n}{b_n} \xrightarrow{2} 0$ and thus $\frac{S_n - \mu_n}{b_n} \xrightarrow{\mathbb{P}} 0$. □

Remark 7.14.1. We should choose b_n that no larger than $\mathbb{E}S_n$ if possible.

Example 7.5. (Coupon collector's problem) Let $X_1, X_2, \cdots$ be i.i.d. uniform random variables on $\{1, 2, \cdots, n\}$.
Let $\tau_k^n = \inf\{m : |\{X_1, X_2, \cdots, X_m\}| = k\}$ be the waiting time for picking k distinct types.
What is the asymptotic behavior of τ_n^n?
It is easy to see that $\tau_1^n = 1$. By convention, $\tau_0^n = 0$.
For $1 \le k \le n$, let $X_{n,k} = \tau_k^n - \tau_{k-1}^n$ be the additional waiting time for picking k distinct types when we have $k-1$ types.
Notice that

$$\tau_n^n = \sum_{k=1}^{n} X_{n,k}$$

We know that

$$\mathbb{P}(X_{n,k} = \ell) = \left(\frac{k-1}{n}\right)^{\ell-1}\left(1 - \frac{k-1}{n}\right) \qquad \Longrightarrow \qquad X_{n,k} \sim \text{Geom}\left(1 - \frac{k-1}{n}\right)$$

We claim that $X_{n,k}$ are independent for all k. For a constant c,

$$\begin{aligned}
\mathbb{E}\tau_n^n &= \sum_{k=1}^{n} \mathbb{E}X_{n,k} = \sum_{k=1}^{n}\left(1 - \frac{k-1}{n}\right)^{-1} = \sum_{m=1}^{n} \frac{n}{m} \sim n \log n \\
\text{Var}(\tau_n^n) &= \sum_{k=1}^{n} \text{Var}(X_{n,k}) = \sum_{k=1}^{n}\left(\left(1 - \frac{k-1}{n}\right)^{-2} - \left(1 - \frac{k-1}{n}\right)^{-1}\right) \le \sum_{k=1}^{n}\left(1 - \frac{k-1}{n}\right)^{-2} = \sum_{m=1}^{n} \frac{n^2}{m^2} \le cn^2
\end{aligned}$$

By WLLN, if we choose $b_n = n \log n$, then we have

$$\frac{\text{Var}(\tau_n^n)}{b_n^2} \to 0 \implies \frac{\tau_n^n - \sum_{m=1}^{n} \frac{n}{m}}{n \log n} \xrightarrow{\mathbb{P}} 0$$

Therefore, $\frac{\tau_n^n}{n \log n} \xrightarrow{\mathbb{P}} 1$

Example 7.6. (An occupancy problem) r balls are put at random into n bins. All n^r are equally likely.
Let A_i be event that the i-th bin is empty and N_n be number of empty bins $= \sum_{i=1}^{n} \mathbf{1}_{A_i}$.
How to prove that if $\frac{r}{n} \to c$ as $n \to \infty$,

$$\frac{N_n}{n} \xrightarrow{\mathbb{P}} e^{-c}$$

We can see that

$$\begin{aligned}
\frac{\mathbb{E}N_n}{n} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\mathbf{1}_{A_i} = \mathbb{P}(A_i) = \left(1 - \frac{1}{n}\right)^r \to e^{-c} \\
\text{Var}(N_n) &= \mathbb{E}(N_n^2) - (\mathbb{E}N_n)^2 \\
&= \mathbb{E}\left(\sum_{i=1}^{n} \mathbf{1}_{A_i}\right)^2 - \left(\mathbb{E}\left(\sum_{i=1}^{n} \mathbf{1}_{A_i}\right)\right)^2 \\
&= \sum_{i=1}^{n}(\mathbb{P}(A_1) - (\mathbb{P}(A_1))^2) + \sum_{i \ne j}(\mathbb{P}(A_i \cap A_j) - (\mathbb{P}(A_1))^2) \\
&= n\left(\left(1 - \frac{1}{n}\right)^r - \left(1 - \frac{1}{n}\right)^{2r}\right) + n(n-1)\left(\left(1 - \frac{2}{n}\right)^r - \left(1 - \frac{1}{n}\right)^{2r}\right) \\
&= o(n^2)
\end{aligned}$$

By using WLLN, let $b_n = n$,

$$\frac{\text{Var}(N_n)}{b_n^2} \to 0 \implies \frac{N_n - \mathbb{E}N_n}{n} \xrightarrow{\mathbb{P}} 0$$

Therefore, $\frac{N_n}{n} \xrightarrow{\mathbb{P}} e^{-c}$.

7.3 Borel-Cantelli Lemmas

Let $A_1, A_2, \cdots$ be a sequence of events in $(\Omega, \mathcal{F})$. We are more interested in

$$\limsup_{n\to\infty} A_n = \{A_n \text{ i.o}\} = \bigcap_m \bigcup_{n=m}^{\infty} A_n$$

Theorem 7.15. (Borel-Cantelli Lemmas) For any sequence of events $A_n \in \mathcal{F}$,

1. (BCI) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then
$$\mathbb{P}(A_n \text{ i.o.}) = 0$$
2. (BCII) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and A_n's 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's are independent, we have
$$\begin{aligned}\mathbb{P}\left(\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} A_n^{\complement}\right) &= \lim_{m\uparrow\infty} \mathbb{P}\left(\bigcap_{n=m}^{\infty} A_n^{\complement}\right) = \lim_{m\uparrow\infty}\lim_{r\uparrow\infty} \mathbb{P}\left(\bigcap_{n=m}^{r} A_n^{\complement}\right) = \lim_{m\uparrow\infty}\lim_{r\uparrow\infty} \prod_{n=m}^{r} \mathbb{P}(A_n^{\complement}) = \lim_{m\uparrow\infty} \prod_{n=m}^{\infty} (1 - \mathbb{P}(A_n)) \\ &\le \lim_{m\uparrow\infty} \prod_{n=m}^{\infty} e^{-\mathbb{P}(A_n)} = \lim_{m\uparrow\infty} \exp\left(-\sum_{n=m}^{\infty} \mathbb{P}(A_n)\right) = 0 \qquad (1 - x \le e^{-x} \text{ if } x \ge 0) \\ \mathbb{P}(A_n \text{ i.o.}) &= \mathbb{P}\left(\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty} A_n\right) = 1 - \mathbb{P}\left(\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty} A_n^{\complement}\right) = 1\end{aligned}$$

□

Remark 7.15.1. We can say that BCII is a partial converse statement of BCI.

Remark 7.15.2. i.o. is an abbreviation of "infinitely often". Similarly, f.o. is an abbreviation if "finitely often".

We will now explore how we can apply Borel-Cantelli Lemmas in multiple applications.

Example 7.7. (Infinite Monkey Problem) Assume there is a keyboard with N keys, each with distinct letters. Given a string of letters S of length m. We have a monkey which randomly hits any key at any round.
How do we prove that almost surely, the monkey will type up the given string S for infinitely many times?
Let E_k be the event that the m-string S is typed starting from the k-th hit. Note that E_k's are not independent.
In order to produce an independent sequence, we can consider E_{mk+1}. where each string is m letters apart from next one.
For any i, $\mathbb{P}(E_i) = \left(\frac{1}{N}\right)^m$. By BCII,

$$\sum_{k=0}^{\infty} \mathbb{P}(E_{mk+1}) = \infty \implies \mathbb{P}(E_{mk+1} \text{ i.o.}) = 1$$

Therefore, $\mathbb{P}(E_k \text{ i.o.}) = 1$

Recall that if $X_n \xrightarrow{\mathbb{P}} X$, there exists a non-random increasing sequence of integers $n_1, n_2, \cdots$ such that $X_{n_i} \xrightarrow{\text{a.s.}} X$ as $i \to \infty$. We can use Borel-Cantelli Lemmas to prove a theorem that is pretty similar.

Theorem 7.16. $X_n \xrightarrow{\mathbb{P}} X$ if and only if for all subsequence $X_{n(m)}$, there is a further subsequence

$$X_{n(m_k)} \xrightarrow{\text{a.s.}} X$$

Proof.

($\Longrightarrow$) Let ε_k be a sequence of positive numbers such that $\varepsilon_k \downarrow 0$ if $k \uparrow \infty$. For any k, there exists an $n(m_k) > n(m_{k-1})$ such that

$$\mathbb{P}(\left|X_{n(m_k)} - X\right| > \varepsilon_k) \leq 2^{-k} \qquad (X_n \xrightarrow{\mathbb{P}} X)$$

Since $\sum_{k=1}^{\infty} \mathbb{P}(\left|X_{n(m_k)} - X\right| > \varepsilon_k) < \infty$, By BCI,

$$\mathbb{P}(\left|X_{n(m_k)} - X\right| > \varepsilon_k \text{ i.o.}) = 0 \qquad \mathbb{P}(\left|X_{n(m_k)} - X\right| > \varepsilon_k \text{ f.o.}) = 1$$

For all $\varepsilon > 0$, $\varepsilon_k \leq \varepsilon$ for all $k \geq k_0$. If $\varepsilon_k \leq \varepsilon$,

$$\{\left|X_{n(m_k)} - X\right| > \varepsilon_k\} \supseteq \{\left|X_{n(m_k)} - X\right| > \varepsilon\}$$

If $\omega \in \{\left|X_{n(m_k)} - X\right| > \varepsilon_k\}$ for finitely many k, then $\omega \in \{\left|X_{n(m_k)} - X\right| > \varepsilon\}$ for finitely many k. Therefore, for all $\varepsilon > 0$

$$\mathbb{P}(\left|X_{n(m_k)} - X\right| > \varepsilon \text{ i.o.}) = 0$$

($\Longleftarrow$) For all $\varepsilon > 0$, let $a_n = \mathbb{P}(|X_n - X| > \varepsilon)$.
For all $n(m)$, there exists $n(m_k)$ such that $X_{n(m_k)} \xrightarrow{\text{a.s.}} X$. We have

$$(X_{n(m_k)} \xrightarrow{\text{a.s.}} X) \implies (X_{n(m_k)} \xrightarrow{\mathbb{P}} X) \implies a_{n(m_k)} \to 0$$

Therefore, for any a_n and $a_{n(m)}$, there exists further $a_{n(m_k)} \to 0$.
We have $a_n \to 0 \implies (X_n \xrightarrow{\mathbb{P}} X)$.

□

We have a theorem that have conditions quite similar to Law of Large Numbers. However, notice that $\mathbb{E}\,|X_1| = \infty$ here.

Theorem 7.17. If $X_1, X_2, \cdots$ are i.i.d. random variables with $\mathbb{E}\,|X_i| = \infty$. Then

$$\mathbb{P}(|X_n| \geq n \text{ i.o.}) = 1$$

Let $S_n = \sum_{i=1}^{n} X_i$. Then

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \text{ exists in } (-\infty, \infty)\right) = 0$$

Proof.

$$\mathbb{E}\,|X_1| = \int_0^{\infty} \mathbb{P}(|X_1| > t)\,dt \leq \sum_{n=0}^{\infty} \mathbb{P}(|X_1| > n)$$

Since $\{|X_n| > n\}$ is a collection of independent events, by BCII, $\mathbb{P}(|X_n| > n \text{ i.o.}) = 1$.
For the second statement, let $C = \{\omega \in \Omega : \lim_{n\to\infty} \frac{S_n(\omega)}{n} \text{ exists in } \mathbb{R}\}$.
Assume that $\omega \in C$, then

$$\frac{S_n(\omega)}{n} - \frac{S_{n+1}(\omega)}{n+1} = \frac{S_n(\omega)}{n(n+1)} - \frac{X_{n+1}(\omega)}{n+1}$$

Since $\frac{S_n}{n}$ converges, $\frac{S_n(\omega)}{n} - \frac{S_{n+1}(\omega)}{n+1} \to 0$ and $\frac{S_n(\omega)}{n(n+1)} \to 0$. We get that $\frac{X_{n+1}(\omega)}{n+1} \to 0$.
However, that means $|X_{n+1}| < n+1$ for an arbitrary large n. Therefore, $\omega \notin \{|X_n| \geq n \text{ i.o.}\}$.
From that, we get that $\mathbb{P}(C) = 0$ since $\mathbb{P}(|X_n| \geq n \text{ i.o.}) = 1$. □

The next result extends BCII.

Theorem 7.18. If $A_1, A_2, \cdots$ are pairwise independent and $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then as $n \to \infty$,

$$\frac{\sum_{m=1}^{n} \mathbf{1}_{A_m}}{\sum_{m=1}^{n} \mathbb{P}(A_m)} \xrightarrow{\text{a.s.}} 1$$

Proof.
Let $X_n = \mathbf{1}_{A_n}$, $S_n = \sum_{i=1}^{n} X_i$ and $\mathbb{E}S_n = \sum_{m=1}^{n} \mathbb{P}(A_m)$.
Notice that pairwise independence is already enough for $\text{cov}(X_i, X_j) = 0$ for all $i \neq j$.
Using Markov's inequality, for any $\varepsilon > 0$, we get as $n \to \infty$

$$\mathbb{P}\left(\left|\frac{S_n - \mathbb{E}S_n}{\mathbb{E}S_n}\right| > \varepsilon\right) \leq \frac{\mathbb{E}(S_n - \mathbb{E}S_n)^2}{\varepsilon^2(\mathbb{E}S_n)^2} = \frac{\text{Var}(S_n)}{\varepsilon^2(\mathbb{E}S_n)^2} = \sum_{m=1}^{n} \frac{\text{Var}(\mathbf{1}_{A_m})}{\varepsilon^2(\mathbb{E}S_n)^2} = \sum_{m=1}^{n} \frac{\mathbb{E}\mathbf{1}_{A_m}}{\varepsilon^2(\mathbb{E}S_n)^2} = \frac{1}{\varepsilon^2\mathbb{E}S_n} \to 0$$

Therefore, we get that $\frac{S_n - \mathbb{E}S_n}{\mathbb{E}S_n} \xrightarrow{\mathbb{P}} 0$.
Now, we can choose a desirable subsequence to prove almost surely convergence. Let $n_k = \inf\{n : \mathbb{E}S_n \geq k^2\}$.
We can get that $\mathbb{E}S_{n_k} \geq k^2$ and $\mathbb{E}S_{n_k} = \mathbb{E}S_{n_k - 1} + \mathbb{E}\mathbf{1}_{A_{n_k}} < k^2 + 1$. Again by Markov's inequality,

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\left|\frac{S_{n_k} - \mathbb{E}S_{n_k}}{\mathbb{E}S_{n_k}}\right| > \varepsilon\right) \leq \sum_{k=1}^{\infty} \frac{1}{\varepsilon^2\mathbb{E}S_{n_k}} \leq \sum_{k=1}^{\infty} \frac{1}{\varepsilon^2(k^2+1)} < \infty$$

By BCI, we have that as $k \to \infty$,

$$\frac{S_{n_k}}{\mathbb{E}S_{n_k}} \xrightarrow{\text{a.s.}} 1 \qquad\qquad \mathbb{P}\left(\frac{S_{n_k}}{\mathbb{E}S_{n_k}} \to 1 \text{ as } k \to \infty\right) = 1$$

Let $C = \{\omega \in \Omega : \frac{S_{n_k}(\omega)}{\mathbb{E}S_{n_k}} \to 1 \text{ as } k \to \infty\}$. For $\omega \in C$, for all $n_k \leq n < n_{k+1}$, we have $S_{n_k}(\omega) \leq S_n(\omega) \leq S_{n_{k+1}}(\omega)$.

$$\frac{S_{n_k}(\omega)}{\mathbb{E}S_{n_k+1}} \leq \frac{S_n(\omega)}{\mathbb{E}S_n} \leq \frac{S_{n_k+1}(\omega)}{\mathbb{E}S_{n_k}}$$

Since $\frac{S_{n_k}(\omega)}{\mathbb{E}S_{n_k+1}} = \frac{S_{n_k}(\omega)}{\mathbb{E}S_{n_k}}\left(\frac{\mathbb{E}S_{n_k}}{\mathbb{E}S_{n_k+1}}\right) \to 1$ and $\frac{S_{n_k+1}(\omega)}{\mathbb{E}S_{n_k+1}} = \frac{S_{n_k+1}(\omega)}{\mathbb{E}S_{n_k+1}}\left(\frac{\mathbb{E}S_{n_k+1}}{\mathbb{E}S_{n_k}}\right) \to 1$, we get that for any $\omega \in C$,

$$\frac{S_n(\omega)}{\mathbb{E}S_n} \to 1$$

Therefore, we have

$$\mathbb{P}\left(\frac{S_n}{\mathbb{E}S_n} \to 1\right) \geq \mathbb{P}\left(\frac{S_{n_k}}{\mathbb{E}S_{n_k}} \to 1 \text{ as } k \to \infty\right) = 1$$

As a result, we get that

$$\frac{S_n}{\mathbb{E}S_n} \xrightarrow{\text{a.s.}} 1$$

□

If the events $A_1, A_2, \cdots$ in the Borel-Cantelli Lemmas are independent, then $\mathbb{P}(A)$ is either 0 or 1 depending on whether $\sum \mathbb{P}(A_n)$ converges. The following is a simple version.

Theorem 7.19. (Borel Zero-one Law) Let $A_1, A_2, \cdots \in \mathcal{F}$ and $\mathcal{A} = \sigma(A_1, A_2, \cdots)$. Suppose that

1. $A \in \mathcal{A}$
2. A is independent with any finite collection of $A_1, A_2, \cdots$

Then $\mathbb{P}(A) = 0$ or 1.

Proof (Non-rigorous).
Suppose that $A_1, A_2, \cdots$ are independent. Let $A = \limsup_n A_n$.
We know that $A = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n$. Therefore, $A \in \mathcal{A} = \sigma(A_1, A_2, \cdots)$.
For all k, we can also have $A = \bigcap_{m=k+1}^{\infty} \bigcup_{n=m}^{\infty} A_n$. Therefore, A is independent with any $A_i \in \sigma(A_1, A_2, \cdots, A_k)$.
Setting $k \to \infty$, we have that A is independent of all elements in $\mathcal{A}$, which also include itself.
Therefore, $\mathbb{P}(A) = \mathbb{P}(A \cap A) = (\mathbb{P}(A))^2 \implies \mathbb{P}(A) = 0$ or 1. □

Let $X_1, X_2, \cdots$ be a collection of random variables. For any subcollection $\{X_i : i \in I\}$, write $\sigma(X_i : i \in I)$ for the smallest σ-field with reference to which each of X_i is measurable.

Definition 7.20. Let $\mathcal{H}_n = \sigma(X_{n+1}, X_{n+2}, \cdots)$. We have $\mathcal{H}_n \supseteq \mathcal{H}_{n+1} \supseteq \cdots$. **Tail σ-field** is defined as

$$\mathcal{H}_\infty = \bigcap_n \mathcal{H}_n$$

Remark 7.20.1. If $E \in \mathcal{H}_\infty$, then E is called **tail event**.

Example 7.8. $\{\limsup_{n\to\infty} X_n = \infty\}$ is a tail event.

Example 7.9. $\{\sum_n X_n \text{ converges}\}$ is a tail event.

Example 7.10. $\{\sum_n X_n \text{ converges to } 1\}$ is not a tail event.

We get another version of zero-one law.

Theorem 7.21. (Kolmogorov's zero-one law) If $H \in \mathcal{H}_\infty$, then $\mathbb{P}(H) = 0$ or 1.

We continue to explore more into tail events.

Definition 7.22. We define **tail function** to be $Y : \Omega \to \mathbb{R} \cup \{-\infty, \infty\}$, which is a generalized random variables that is a function of $X_1, X_2, \cdots$. It is independent of any finite collection of X_i's and is $\mathcal{H}_\infty$-measurable.

Example 7.11. Let $Y(\omega) = \limsup_{n\to\infty} X_n(\omega)$ for all $\omega \in \Omega$. $F_Y(y) = \mathbb{P}(Y \le y) = 0$ or 1 for all $y \in \mathbb{R} \cup \{-\infty, \infty\}$. $\{Y \le y\}$ is a tail event.

Theorem 7.23. If Y is a tail function of independent sequence of random variables $X_1, X_2, \cdots$, then there exists $-\infty \le k \le \infty$,

$$\mathbb{P}(Y = k) = 1$$

Again let $X_1, X_2, \cdots$ be i.i.d. random variables and let $S_n = \sum_{i=1}^n X_i$.
Recall that if $\mathbb{E}|X_1| < \infty$,

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = \mathbb{E}X_1\right) = 1$$

If $\mathbb{E}|X_1| = \infty$,

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \text{ exists}\right) = 0$$

Using tail function, the random variables are not necessarily identically distributed.

Theorem 7.24. Let $X_1, X_2, \cdots$ be independent random variables. Let $S_n = \sum_{i=1}^n X_i$. Then

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \text{ exists}\right) = 0 \text{ or } 1$$

Proof.
Let $Z_1 = \limsup_{n\to\infty} \frac{S_n}{n}$ and $Z_2 = \liminf_{n\to\infty} \frac{S_n}{n}$. We claim that both Z_1 and Z_2 are tail functions of X_i's. For any k,

$$Z_1(\omega) = \limsup_{n\to\infty} \left(\frac{1}{n}\sum_{i=1}^k X_i(\omega) + \frac{1}{n}\sum_{i=k+1}^n X_i(\omega)\right) \qquad Z_2(\omega) = \liminf_{n\to\infty} \left(\frac{1}{n}\sum_{i=1}^k X_i(\omega) + \frac{1}{n}\sum_{i=k+1}^n X_i(\omega)\right)$$

Therefore, both Z_1 and Z_2 do not depend on any finite collection of X_i. We say that $\{Z_1 = Z_2\}$ is a tail event.
Therefore, by Kolmogorov's zero-one law.

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \text{ exists}\right) = \mathbb{P}(Z_1 = Z_2) = 0 \text{ or } 1$$

$\square$

Example 7.12. (Random power series) Let $X_1, X_2, \cdots$ be i.i.d. exponential random variables with parameter $\lambda = 1$. We consider a random power series

$$p(z;\omega) = \sum_{n=0}^{\infty} X_n(\omega) z^n$$

The formula for radius of convergence is

$$R(\omega) = \frac{1}{\limsup_{n\to\infty} |X_n(\omega)|^{\frac{1}{n}}}$$

We can get that $R(\omega)$ is a tail function of X_i's. Therefore, there exists C such that $\mathbb{P}(R = C) = 1$ ($R = C$ almost surely)
We want to find the value of C.
We claim that $C = 1$.

$$\mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} = 1\right) = 1$$

It suffices to show that for all $\varepsilon > 0$,

$$\mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} \leq 1+\varepsilon\right) = 1 \qquad \mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} \geq 1-\varepsilon\right) = 1$$

We first prove the first one.

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|X_n|^{\frac{1}{n}} > 1+\varepsilon\right) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > (1+\varepsilon)^n) = \sum_{n=1}^{\infty} e^{-(1+\varepsilon)^n} < \infty$$

Therefore, by BCI,

$$\mathbb{P}(|X_n|^{\frac{1}{n}} > 1+\varepsilon \text{ i.o.}) = 0 \implies \mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} \leq 1+\varepsilon\right) = 1$$

Similarly,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|X_n|^{\frac{1}{n}} > 1-\varepsilon\right) = \sum_{n=1}^{\infty} \mathbb{P}(|X_n| > (1-\varepsilon)^n) = \sum_{n=1}^{\infty} e^{-(1-\varepsilon)^n} = \infty$$

Therefore, by BCII,

$$\mathbb{P}(|X_n|^{\frac{1}{n}} > 1-\varepsilon \text{ i.o.}) = 1 \implies \mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} \geq 1-\varepsilon\right) = 1$$

By sending $\varepsilon \downarrow 0$, we get

$$\mathbb{P}\left(\limsup_{n\to\infty} |X_n|^{\frac{1}{n}} = 1\right) = 1$$

Therefore, $C = 1$.

7.4 Strong Law of Large Numbers

We recall the Weak Law of Large Numbers. Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables with $\mathbb{E}(X_1) = \mu$. Let $S_n = \sum_{i=1}^n X_i$. Then as $n \to \infty$,

$$\frac{S_n}{n} \xrightarrow{D} \mu \qquad \frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu$$

By name, WLLN indeed has a stronger version. It is called Strong Law of Large Numbers. We prove one of the versions of SLLN.

Theorem 7.25. (**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 $\mathbb{E}|X_1| < \infty$. Let $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mu$$

Note that the proof for SLLN is very complicated, and we will not prove it here. Instead, we will prove a different version of SLLN.

Theorem 7.26. (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 $\mathbb{E}(X_1^4) < \infty$. Let $S_n = \sum_{i=1}^n X_i$, then

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} 0$$

Proof.

$$\mathbb{E}S_n^4 = \mathbb{E}\left(\sum_{i=1}^n X_i\right)^4 = \sum_{i,j,k,\ell=1}^n \mathbb{E}X_iX_jX_kX_\ell$$

The expectation is non-zero if there are 2 pairs of the random variables with same value.

$$\mathbb{E}S_n^4 = 3\sum_{i\neq j}\mathbb{E}X_i^2\mathbb{E}X_j^2 + \sum_i \mathbb{E}X_i^4 = O(n^2)$$

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right| \geq \varepsilon\right) \leq \frac{\mathbb{E}S_n^4}{(n\varepsilon)^4} = O\left(\frac{1}{n^2}\right)$$

Therefore, we get that for all $\varepsilon > 0$

$$\sum_{n=1}^\infty \mathbb{P}\left(\left|\frac{S_n}{n}\right| > \varepsilon\right) < \infty$$

Therefore, $\frac{S_n}{n} \xrightarrow{\text{a.s.}} 0$. □

Theorem 7.27. (SLLN with $\mathbb{E}X_1^2 < \infty$) Let $X_1, X_2, \cdots$ be i.i.d. random variables with $\mathbb{E}X_1^2 < \infty$ and $\mathbb{E}X_i = \mu$. Let $S_n = \sum_{i=1}^n X_i$. Then

$$\frac{S_n}{n} \xrightarrow{2} \mu \qquad\qquad \frac{S_n}{n} \xrightarrow{\text{a.s.}} \mu$$

Proof.
We first show convergence in mean square. Since $\mathbb{E}X_1^2 < \infty$, as $n \to \infty$,

$$\mathbb{E}\left(\frac{S_n}{n} - \mu\right)^2 = \frac{\mathbb{E}(S_n - n\mu)^2}{n^2} = \frac{\text{Var}(S_n)}{n^2} = \frac{\text{Var}(X_1)}{n} \to 0$$

For the almost sure convergence, we know that convergence in probability implies the existence of almost sure convergence of some subsequence of $\frac{S_n}{n}$ to μ. We write $n_i = i^2$. By using Markov's inequality, for all $\varepsilon > 0$,

$$\sum_i \mathbb{P}\left(\frac{\left|S_{i^2} - i^2\mu\right|}{i^2} > \varepsilon\right) \leq \sum_i \frac{\mathbb{E}\left|S_{i^2} - i^2\mu\right|^2}{i^4\varepsilon^2} = \sum_i \frac{\text{Var}(S_{i^2})}{i^4\varepsilon^2} = \sum_i \frac{\text{Var}(X_1)}{i^2\varepsilon^2} < \infty$$

Therefore, we get that $\frac{S_{i^2}}{i^2} \xrightarrow{\text{a.s.}} \mu$. However, we need to fill the gaps.
We suppose the X_i are non-negative. We have $S_{i^2} \leq S_n \leq S_{(i+1)^2}$ if $i^2 \leq n \leq (i+1)^2$.
We can get that

$$\frac{S_{i^2}}{(i+1)^2} \leq \frac{S_n}{n} \leq \frac{S_{(i+1)^2}}{i^2}$$

Since we get that $\frac{S_{i^2}}{i^2} \xrightarrow{\text{a.s.}} \mu$, by having $\frac{i^2}{(i+1)^2} \to 1$ as $i \to \infty$, we get that whenever X_i are non-negative, as $n \to \infty$

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mu$$

For general X_i, we can write $X_n = X_n^+ - X_n^-$ where

$$X_n^+(\omega) = \max\{X_n(\omega), 0\} \qquad\qquad X_n^-(\omega) = -\min\{X_n(\omega), 0\}$$

Therefore, both $X_n^+(\omega)$ and $X_n^-(\omega)$ are non-negative.
Furthermore, $X_n^+ \leq |X_n|$ and $X_n^- \leq |X_n|$. Therefore, $\mathbb{E}(X_n^+)^2 < \infty$ and $\mathbb{E}(X_n^-)^2 < \infty$. By previous conclusion for non-negative random variables, we get as $n \to \infty$,

$$\frac{S_n}{n} = \frac{1}{n}\left(\sum_{i=1}^n X_i^+ - \sum_{i=1}^n X_i^-\right) \xrightarrow{\text{a.s.}} \mathbb{E}X_1^+ - \mathbb{E}X_1^- = \mathbb{E}X_1$$

Therefore, $\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mu$. □

Why do we need SLLN? There are a lot of applications that specifically need SLLN.

Example 7.13. (Renewal Theory) Assume that we have a light bulb. We change it immediately when it burnt out.
Let X_i be the lifetime of i-th bulb and $T_n = X_1 + X_2 + \cdots + X_n$ be the time to replace the n-th bulb.
Let $N_t = \sup\{n : T_n \le t\}$ be number of bulbs that have burnt out by time t. T_{N_t} is the exact time that N_t's bulb burnt out.
Since we are dealing with practical bulb, assume that $X_1, X_2, \cdots$ are i.i.d. random variables with $0 < X_i < \infty$ and $\mathbb{E}X_1 < \infty$.

Theorem 7.28. Let $\mathbb{E}X_1 = \mu$. As $t \to \infty$,

$$\frac{t}{N_t} \xrightarrow{\text{a.s.}} \mu$$

Proof.
Since $T_{N_t} \le t < T_{N_t+1}$,

$$\frac{T_{N_t}}{N_t} \le \frac{t}{N_t} < \frac{T_{N_t+1}}{N_t+1}\left(\frac{N_t+1}{N_t}\right)$$

By SLLN, we know that $\frac{T_n}{n} \xrightarrow{\text{a.s.}} \mu$. Since $\frac{T_n}{n}$ and $\frac{T_{N_t}}{N_t}$ are the same sequence, we get that

$$\frac{T_{N_t}}{N_t} \xrightarrow{\text{a.s.}} \mu \qquad\qquad \frac{T_{N_t+1}}{N_t+1} \xrightarrow{\text{a.s.}} \mu$$

For all $\omega \in \Omega$, $t < T_{N_t+1} = X_1(\omega) + X_2(\omega) + \cdots + X_{N_t(\omega)+1}(\omega)$.
As $t \to \infty$, it forces $N_t(\omega) \to \infty$. Therefore, $\frac{N_t+1}{N_t} \xrightarrow{\text{a.s.}} 1$. Combining all of this, we get $\frac{t}{N_t} \xrightarrow{\text{a.s.}} \mu$. $\square$

Claim 7.28.1. If $X_n \xrightarrow{\mathbb{P}} X_\infty$, then $N_m \xrightarrow{\text{a.s.}} \infty$ as $m \to \infty$.

Remark 7.28.1. For this claim, it is not necessary that $X_{N_m} \xrightarrow{\text{a.s}} X_\infty$ or $X_{N_m} \xrightarrow{\mathbb{P}} X_\infty$.

Example 7.14. Recall the example that we use in Theorem 7.9 to prove $(X_n \xrightarrow{\mathbb{P}} X) \not\Longrightarrow (X_n \xrightarrow{\text{a.s.}} X)$. Let $\Omega = [0,1]$. Let

$$Y_{m,k} = \mathbf{1}_{I_{m,k}} = \begin{cases} 1, & \omega \in \left[\frac{k-1}{m}, \frac{k}{m}\right] \\ 0, & \text{Otherwise} \end{cases}$$

Let X_n be the enumeration of $Y_{m,k}$. i.e. $X_1 = Y_{1,1}$, $X_2 = Y_{2,1}$, $X_3 = Y_{2,2}$, $\cdots$.
From the proof of the theorem, we got that $X_n \xrightarrow{\mathbb{P}} X_\infty = 0$ but $X_n \not\xrightarrow{\text{a.s.}} X_\infty$.
For each $\omega \in \Omega$, and each $m \ge 1$, there exists k such that $\omega \in \left[\frac{k-1}{m}, \frac{k}{m}\right]$. We denote these as $k_m(\omega)$.
Let $N_m(\omega) = \sum_{i=1}^{m-1} i + k_m(\omega)$. We get that $X_{N_m(\omega)}(\omega) = Y_{m,k_m(\omega)}(\omega) = 1$.
However, $X_\infty = 0$. That means, $X_{N_m} \not\xrightarrow{\mathbb{P}} X_\infty$ and $X_{N_m} \not\xrightarrow{\text{a.s.}} X_\infty$.

We move to our next examples, which is the Glivenko-Cantelli Theorem. It is also called the Fundamental Theorem of Statistics.

Theorem 7.29. (Glivenko-Cantelli Theorem) Assume that $X \sim F(x)$ where $F(x)$ is unknown. Let $X_1, X_2, \cdots$ be i.i.d. random samples of X. We define the empirical distribution function, which is also a distribution function of a histogram.

$$F_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{X_i \le x} \qquad\qquad F_N(x;\omega) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{X_i(\omega) \le x}$$

We have that

$$\sup_x |F_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0$$

Proof.
We only proof for the case when $F(x)$ is continuous.
For each m, there exists $-\infty = x_0 < x_1 < \cdots < x_m = \infty$ such that $F(x_i) - F(x_{i-1}) = \frac{1}{m}$.
For all $x \in [x_{i-1}, x_i)$,

$$F_N(x) - F(x) \le F_N(x_i) - F(x_{i-1}) = F_N(x_i) - F(x_i) + \frac{1}{m}$$
$$F_N(x) - F(x) \ge F_N(x_{i-1}) - F(x_i) = F_N(x_{i-1}) - F(x_{i-1}) - \frac{1}{m}$$

From this, we get

$$-\sup_i |F_N(x_i) - F(x_i)| - \frac{1}{m} \le F_N(x) - F(x) \le \sup_i |F_N(x_i) - F(x_i)| + \frac{1}{m} \implies \sup_x |F_N(x) - F(x)| \le \sup_i |F_N(x_i) - F(x_i)| + \frac{1}{m}$$

By SLLN, when we fix x, we get

$$F_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{X_i \le x} \xrightarrow{\text{a.s.}} \mathbb{E}\mathbf{1}(X_1 \le x) = \mathbb{P}(X_1 \le x) = F(x) \qquad \mathbb{P}(\{\omega \in \Omega : F_N(x;\omega) \to F(x) \text{ as } N \to \infty\}) = 1$$

Let $C_x = \{\omega \in \Omega : F_N(x;\omega) \to F(x) \text{ as } N \to \infty\}$. Notice that if $\omega \in \bigcap_{i=1}^{\infty} C_{x_i}$, $\sup_i |F_N(x_i;\omega) - F(x_i)| \to 0$.

$$\limsup_N \sup_x |F_N(x) - F(x)| \le \frac{1}{m}$$

If $\omega \in \bigcap_{m=1}^{\infty}\bigcap_{i=1}^{m} C_{x_i}$,

$$\limsup_N \sup_x |F_N(x) - F(x)| = 0$$

Therefore, since $\bigcap_{m=1}^{\infty}\bigcap_{i=1}^{m} C_{x_i} \subseteq \{\omega \in \Omega : \sup_x |F_N(x;\omega) - F(x)| \to 0 \text{ as } N \to \infty\}$ and $\mathbb{P}(C_{x_i}) = 1$ by SLLN,

$$\mathbb{P}(\{\omega \in \Omega : \sup_x |F_N(x;\omega) - F(x)| \to 0 \text{ as } N \to \infty\}) = 1$$

$\square$

We will end here. Of course, there are still a lot of examples that we haven't explored (including some mentioned during the lectures that I'm too lazy to include here). We also skipped a lot of proofs in some of the theorems. It is up to you to explore further, either in other courses or in the future world of mathematics.

Appendix A

Random walk

Example A.1. (**Simple random walk**) Consider a particle moving on the real line. Every step it moves to the right by 1 with probability p, and to the left by 1 with probability $q = 1 - p$.
Let S_n be the position of the particles after n moves and let $S_0 = a$. Then:

$$S_n = a + \sum_{i=1}^{n} X_i$$

where $X_1, X_2, \cdots$ is a sequence of independently random variables taking 1 with probability p and -1 with probability q.
Random walk is **symmetric** if $p = q = \frac{1}{2}$.

Lemma A.1. Simple random walk has the following properties:

1. It is **spatially homogeneous**: $\mathbb{P}(S_n = j|S_0 = a) = \mathbb{P}(S_n = j + b|S_0 = a + b)$.
2. It is **temporarily homogeneous**: $\mathbb{P}(S_n = j|S_0 = a) = \mathbb{P}(S_{m+n} = j|S_m = a)$.
3. It has **Markov property**: $\mathbb{P}(S_{m+n} = j|S_0, S_1, \cdots, S_m) = \mathbb{P}(S_{m+n} = j|S_m)$, $n \geq 0$.

Proof.

1. $\mathbb{P}(S_n = j|S_0 = a) = \mathbb{P}(\sum_{i=1}^{n} X_i = j - a) = \mathbb{P}(S_n = j + b|S_0 = a + b$
2. $$\mathbb{P}(S_n = j|S_0 = a) = \mathbb{P}\left(\sum_{i=1}^{n} X_i = j - a\right) = \mathbb{P}\left(\sum_{i=m+1}^{m+n} X_i = j - a\right) = \mathbb{P}(S_{m+n} = j|S_m = a)$$
3. If we know S_m, then distribution of S_{m+n} depends only on $X_{m+1}, X_{m+2}, \cdots, X_{m+n}$ and $S_0, S_1, \cdots, S_{m-1}$ does not influence the dependency.

□

Example A.2. (Probability via sample path counting) Let **sample path** $\vec{s} = (s_0, s_1, \cdots, s_n)$ (outcome/realization of the random walk), with $s_0 = a$ and $s_{i+1} - s_i = \pm 1$.

$$\mathbb{P}((S_0, S_1, \cdots, S_n) = (s_0, s_1, \cdots, s_n)) = p^r q^\ell \qquad r = \#\{i : s_{i+1} - s_i = 1\} \qquad \ell = \#\{i : s_{i+1} - s_i = -1\}$$

Example A.3. Let $M_n^r(a, b)$ be number of paths $(s_0, s_1, \cdots, s_n)$ with $s_0 = a$, $s_n = b$ and having r rightward steps.

$$\mathbb{P}(S_n = b) = \sum_r M_n^r(a, b) p^r q^{n-r}$$

By equations $r + \ell = n$ and $r - \ell = b - a$, $r = \frac{1}{2}(n + b - a)$ and $\ell = (n - b + a)$.
If $\frac{1}{2}(n + b - a) \in \{0, 1, \cdots, n\}$,

$$\mathbb{P}(S_n = b) = \binom{n}{\frac{1}{2}(n + b - a)} p^{\frac{1}{2}(n+b-a)} q^{\frac{1}{2}(n-b+a)}$$

Otherwise, $\mathbb{P}(S_n = b) = 0$.

Theorem A.2. (Reflection principle) Let $N_n(a,b)$ be number of possible paths from $(0,a)$ to (n,b) and let $N_n^0(a,b)$ be number of such paths which contains some point $(k,0)$ on the x-axis. If $a,b>0$, then:

$$N_n^0(a,b)=N_n(-a,b)$$

Proof.
Each path from $(0,-a)$ to (n,b) intersects the x-axis at some earliest point $(k,0)$.
Reflect the segment of the path with $0\le x\le k$ in the x-axis to obtain a path joining $(0,a)$ to (n,b) which intersects the x-axis.
This operation gives a one-to-one correspondence between the collections of such paths. □

Lemma A.3.

$$N_n(a,b)=\binom{n}{\frac{1}{2}(n+b-a)}$$

Proof.
Choose a path from $(0,a)$ to (n,b) and let α and β be numbers of positive and negative steps in this path respectively.
Then $\alpha+\beta=n$ and $\alpha-\beta=b-a$, which we have $\alpha=\frac{1}{2}(n+b-a)$.
Number of such paths is the number of ways of picking α positive steps from n available. Therefore,

$$N_n(a,b)=\binom{n}{\alpha}=\binom{n}{\frac{1}{2}(n+b-a)}$$

□

Example A.4. We want to find the probability that the walk does not revisit its starting point in the first n steps.
Without loss of generality, we assume $S_0=0$ so that $S_1,S_2,\cdots,S_n\neq 0$ if and only if $S_1S_2\cdots S_n\neq 0$.
Event $S_1S_2\cdots S_n\neq 0$ occurs if and only if the path of the walk does not visit the x-axis in the time interval $[1,n]$.
If $b>0$, first step must be $(1,1)$, so, by Lemma A.3 and Reflection principle, number of such path is:

$$\begin{aligned}
N_{n-1}(1,b)-N_{n-1}^0(1,b)&=N_{n-1}(1,b)-N_{n-1}(-1,b)\\
&=\binom{n-1}{\frac{1}{2}(n+b-2)}-\binom{n-1}{\frac{1}{2}(n+b)}\\
&=\left(\frac{n+b}{2n}-\frac{n-b}{2n}\right)\binom{n}{\frac{1}{2}(n+b)}\\
&=\frac{b}{n}\binom{n}{\frac{1}{2}(n+b)}
\end{aligned}$$

There are $\frac{1}{2}(n+b)$ rightward steps and $\frac{1}{2}(n-b)$ leftward steps. Therefore,

$$\mathbb{P}(S_1S_2\cdots S_n\neq 0, S_n=b)=\frac{b}{n}N_n(0,b)p^{\frac{1}{2}(n+b)}q^{\frac{1}{2}(n-b)}=\frac{b}{n}\mathbb{P}(S_n=b).$$

Example A.5. Let $M_n=\max\{S_i:0\le i\le n\}$ be the maximum value attained by random walk up to time n. Suppose that $S_0=0$ so that $M_n\ge 0$. We have $M_n\ge S_n$.

Theorem A.4. Suppose that $S_0=0$. Then, for $r\ge 1$,

$$\mathbb{P}(M_n\ge r,S_n=b)=\begin{cases}\mathbb{P}(S_n=b), & \text{if } b\ge r\\ \left(\frac{q}{p}\right)^{r-b}\mathbb{P}(S_n=2r-b), & \text{if } b<r\end{cases}$$

It follows that, for $r\ge 1$,

$$\mathbb{P}(M_n\ge r)=\mathbb{P}(S_n\ge r)+\sum_{b=-\infty}^{r-1}\left(\frac{q}{p}\right)^{r-b}\mathbb{P}(S_n=2r-b)=\mathbb{P}(S_n=r)+\sum_{c=r+1}^{\infty}\left(1+\left(\frac{q}{p}\right)^{c-r}\right)\mathbb{P}(S_n=c)$$

For symmetric case when $p=q=\frac{1}{2}$,

$$\mathbb{P}(M_n\ge r)=2\mathbb{P}(S_n\ge r+1)+\mathbb{P}(S_n=r)$$

Proof.
Assume that $r \geq 1$ and $b < r$. Let $N_n^r(0,b)$ be number of paths from $(0,0)$ to (n,b) which include some points having height r (Some point (i,r) with $0 < i < n$).
For a path π, let (i_π, r) be the earliest point.
We reflect the segment of path with $i_\pi \leq x \leq n$ in the line $y = r$ to obtain path π' joining $(0,0)$ to $(n, 2r-b)$.
We have $N_n^r(0,b) = N_n(0, 2r-b)$.

$$\mathbb{P}(M_n \geq r, S_n = b) = N_n^r(0,b) p^{\frac{1}{2}(n+b)} q^{\frac{1}{2}(n-b)} = \left(\frac{q}{p}\right)^{r-b} N_n(0, 2r-b) p^{\frac{1}{2}(n+2r-b)} q^{\frac{1}{2}(n-2r+b)} = \left(\frac{q}{p}\right)^{r-b} \mathbb{P}(S_n = 2r-b)$$

□

Appendix B

Terminologies in other fields of mathematics

Definition B.1. **Supremum** of subset S is the lowest upper bound x such that for all $a \in S$, $x \geq a$. We write it as

$$x = \sup S$$

Definition B.2. **Infimum** of subset S is the highest lower bound x such that for all $b \in S$, $x \leq b$. We write it as

$$x = \inf S$$

Definition B.3. **Limit superior** and **limit inferior** of a sequence $x_1, x_2, \cdots$ are defined by

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty} \sup_{m\geq n} x_m \qquad \liminf_{n\to\infty} x_n = \lim_{n\to\infty} \inf_{m\geq n} x_m$$

Definition B.4. Infinite series $\sum_{n=0}^{\infty} a_n$ is **absolutely convergent** if for some real numbers L,

$$\sum_{n=0}^{\infty} |a_n| = L$$

Any groupings and rearrangings of absolutely convergent infinite series do not change the result of the infinite series.
An infinite series is **conditionally convergent** if it converges but does not satisfy the condition.

Definition B.5. (**Monotonicity**) **Monotonic** function is a function that is either entirely non-increasing or entirely non-decreasing.
Strictly monotonic function is a function that is either entirely strictly increasing or entirely strictly decreasing.

Definition B.6. **Arguments of the maxima** are the input points at which a function 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. **Arguments of the minima** are the input points at which a function 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**) **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. **Regular** function is a function f that is

1. single-valued (any values in the domain will map to exactly one value)
2. analytic (f can be written as a convergent power series)

Definition B.10. Let V be a space of all real functions on $[0,1]$. $\|\cdot\| : V \to \mathbb{R}$ 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 to be

$$\|f\|_\infty = \max_{0 \leq x \leq 1} |f(x)|$$

Definition B.11. Functions f and g are **asymptotic equivalent** ($f \sim g$) if and only if

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 1$$

Appendix C

Some useful inequalities

Theorem C.1. (Triangle inequality) Let X and Y be random variables. Then

$$|X+Y| \leq |X| + |Y|$$

Theorem C.2. (Reverse triangle inequality) Let X and Y be random variables. Then

$$|X-Y| \geq ||X| - |Y||$$

Theorem C.3. (Cauchy-Schwarz inequality) Let X and Y be random variables. Then

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

Theorem C.4. (Covariance inequality) Let X and Y be random variables. Then

$$|\text{cov}(X,Y)|^2 \leq \text{Var}(X)\,\text{Var}(Y)$$

Theorem C.5. (Markov's inequality) Let X be a random variable with 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 $\text{Var}(X) = \sigma^2$. 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 variables and γ be a convex function. Then

$$\gamma(\mathbb{E}X) \leq \mathbb{E}(\gamma(X))$$

For better memorization,
Triangle inequality $\Longrightarrow$ Reverse triangle inequality
Markov's inequality $\Longrightarrow$ Chebyshev's inequality
Hölder's inequality $\Longrightarrow$ Cauchy-Schwarz inequality $\Longrightarrow$ Covariance inequality