MATH 2431: Honors Probability

HU-HTAKM

March 27, 2024

This lecture note is made based on the MATH 2431 lecture notes made by Prof. Bao, Zhigang in Spring 2023-24. I also include some of the stuff in the textbook "Probability and Random Processes" Third Edition written by G. Grimmett and D. Stirzaker to have better understanding in some specific topics. We follow the chapters based on the textbook.
Some proofs are written by me because they are not included in both the lecture notes or the textbook. Therefore, it is likely that they are wrong.
This course has the co-requisite of multivariable calculus. However, we highly recommend you know everything about multivariable calculus beforehand because those knowledge will be applied very early.

Notations	Meaning
$\mathbb{Q}$	Set of rational numbers
$\mathbb{R}$	Set of real numbers
$\emptyset$	Empty set
Ω	Sample space / Entire set
ω	Outcome
$\mathcal{F}, \mathcal{G}$	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 $\mathbb{R}$
$\mathbb{E}$	Expectation
ψ	Conditional expectation
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \cdots$	Matrix
$\mathbf{1}_A$	Indicator function

Table 1: Notations

Definition 0.1. This is definition.

Remark 0.1.1. This is remark.

Lemma 0.2. This is lemma.

Proposition 0.3. This is proposition.

Theorem 0.4. This is theorem.

Claim 0.4.1. This is claim.

Corollary 0.5. This is corollary.

Example 0.1. This is example.

Contents

Chapter 1

Events and their probabilities

1.1 Fundamental terminologies

In 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.
We start with some basic terminology.

Definition 1.1. Sample space Ω is the set of all outcomes of an experiment. Outcomes are denoted by ω.

Example 1.1. Coin flipping $\Omega = \{H, T\}$

Example 1.2. Die rolling $\Omega = \{1, 2, 3, 4, 5, 6\}$

Example 1.3. Life time of bulb $\Omega = [0, \infty)$

Example 1.4. Two coins flipping $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$

Many statements take the form of "the probability of event A is p", which events usually include some of the elements of sample space.

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

Remark 1.2.1. It is not necessary for all subset of Ω to be an event. However, we do not discuss this issue for the moment.

Example 1.5. Dice rolling $\Omega = \{1, 2, \cdots, 6\}$ Event: Even $(\{2, 4, 6\})$

Remark 1.2.2. If only the outcome $\omega = 2$ is given, there are many events that can obtain that outcome. E.g. $\{2\}$, $\{2, 4\}$, $\cdots$

Definition 1.3. Complement of a subset A is a subset $A^{\complement}$ which contains all elements in sample space Ω that is not in A.

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

Definition 1.4. Field 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 1.5. **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 $\Omega = A \cup A^{\complement} \cup \cdots \in \mathcal{F}$.

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

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

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

Remark 1.5.1. These two formulae may be useful.

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

1.2 Probability measure

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

Definition 1.6. **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 ($A_i \cap A_j = \emptyset$ for all $i \neq j$), then $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$. (Countable additivity)

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. However, there are in fact other measures that satisfy the definition of probability measure. E.g. Risk-neutral measure. We will discuss it later.
The following measures are not probability measures.

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

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

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

Definition 1.7. **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.

Example 1.11. Coin flip: $\Omega = \{H, T\}$, $\mathcal{F} = \{\emptyset, H, T, \Omega\}$. Let $\mathbb{P}(H) = p$ where $p \in [0, 1]$.

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

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

Example 1.12. Die roll: $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = \{0, 1\}^{\Omega}$. 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 1.8. Basic properties of $\mathbb{P}$:

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

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

Proof.

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

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

□

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 1.8.1. We have two types of set limit:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.3 Conditional probability

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

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

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

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

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

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

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

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

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

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

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

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

There is some cases when there are multiple events that allows certain event to occur.

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

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

Proof.

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

□

1.4 Independence

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

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

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

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

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

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

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

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

Proof.

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

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

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

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

Proof.

1. Using the properties of probability,

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

2.

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

□

Remark 1.18.1. If $A \perp\!\!\!\perp B$ and $A \cap B = \emptyset$, then $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$.

1.5 Product space

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

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

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

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

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

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

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

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

Definition 1.21. Product space of two probability spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ is the probability space $(\Omega_1 \times \Omega_2, \mathcal{G}, \mathbb{P}_{12})$ comprising a collection of ordered pairs $\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$, a σ-algebra $\mathcal{G} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ where $\mathcal{F}_1 \times \mathcal{F}_2 = \{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}$, and a probability measure $\mathbb{P}_{12} : \mathcal{F}_1 \times \mathcal{F}_2 \to [0, 1]$ given by:

$$\mathbb{P}_{12}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2) \qquad \text{for } A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$$

Chapter 2

Random variables and their distribution

2.1 Introduction of random variables

Sometimes, we are not interested in an experiment, but rather in 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 2.1. Random variable is a function $X : \Omega \to \mathbb{R}$ with the property that for any $x \in \mathbb{R}$,

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

Remark 2.1.1. More generally, 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**.

Remark 2.1.2. 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 2.1.1. Suppose $X^{-1}(B) \in \mathcal{F}$ for all open sets B. Then $X^{-1}(B') \in \mathcal{F}$ for all closed sets B'.

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

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

□

Remark 2.1.3. X needs to be $\mathcal{F}$-measurable because $\mathbb{P}(X \in A) = \mathbb{P}(\{\omega : X(\omega) \in A\}) = \mathbb{P}(X^{-1}(A))$. $X^{-1}(A)$ has to be in $\mathcal{F}$.

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

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

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

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

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

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

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

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

Proof.

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

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

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

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

□

Example 2.3. If we choose $\mathcal{F} = \{\emptyset, \Omega\}$, X is not a random variable.

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

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

2.2 CDF of random variables

Every random variable has its own distribution function.

Definition 2.4. (Cumulative) distribution function (CDF) of a random variable X is a function $F_X : \mathbb{R} \to [0,1]$ given by:

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

Example 2.4. From Example 2.1,

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

Lemma 2.5. 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 $F_X(x) \leq F_X(y)$.
3. F_X is right-continuous ($F_X(x + h) \to F_X(x)$ as $h \downarrow 0$)

Proof.

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

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

   Alternative proof:

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

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

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

   Alternative Proof:

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

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

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

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

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

   Alternative Proof:

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

□

Remark 2.5.1. F is not left-continuous because:

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

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

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

Proof.

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

□

Example 2.5. (**Constant variables**) Let $X : \Omega \to \mathbb{R}$ be defined by $X(\omega) = c$ for all $\omega \in \Omega$. For all $B \in \mathcal{B}$,

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

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

Example 2.6. (**Bernoulli variables**) Consider flipping coin once. Let $X : \Omega \to \mathbb{R}$ be defined by $X(H) = 1$ and $X(T) = 0$.

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

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

Example 2.7. Let A be an event in $\mathcal{F}$ and **indicator functions** $\mathbf{1}_A : \Omega \to \mathbb{R}$ such that for all $B \in \mathcal{B}(\mathbb{R})$:

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

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

2.3 PMF / PDF of random variables

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

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

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

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

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

Proof.

1. $$F_X(x) = \mathbb{P}(X \leq x) = \sum_{i=-\infty}^{x} \mathbb{P}(X = i) = \sum_{i \leq x} f_X(i)$$

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

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

□

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

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

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

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

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

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

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

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

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

2.4 JCDF of random variables

How do we deal with cases when there are more than 1 random variables?

Definition 2.10. 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 2.10.1. Both definition of random vectors is 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. We can also get X_2 is a random variable by choosing $A_1 = \mathbb{R}$ instead.
Therefore, we can get second definition from first definition.
By second definition, X_1, X_2 are random variable. 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 get first definition from second definition.
Therefore, two definitions are equivalent. □

Remark 2.10.1. 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 2.11. 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 2.11.1. We can replace all Borel sets by the form $[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$.

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

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

1. $\lim_{x,y\to-\infty} F_{X,Y}(x, y) = 0$ 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 2.13. Let X, Y be random variable. We can get a **marginal distribution** (marginal CDF) by having:

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

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

Definition 2.14. Random variable 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.
Joint (probability) mass function (JPMF) $f : \mathbb{R}^2 \to [0,1]$ is given by

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

Remark 2.14.1.

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

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

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

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

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

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

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

Remark 2.15.2. If X, Y are both continuous random variables, it is not always true that X, Y are jointly continuous.

Example 2.8. Let X be uniformly distributed on $[0,1]$ ($f_X(x) = \mathbf{1}_{[0,1]}$). This means $f_X(x)$ is 1 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.

Example 2.9. Assume that a special three-sided coin is provided. Each toss results in heads (H), tails (T) or edge (E) with equal probability. 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 $H_n + T_n + E_n = n$.

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

Chapter 3

Discrete random variables

3.1 Introduction of discrete random variables

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

Definition 3.1. 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 3.2. PMF $f_X : \mathbb{R} \to [0, 1]$ of a 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 3.3. 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 3.4. Discrete variables X and Y are **independent** if the events $\{X = x\}$ and $\{Y = y\}$ are independent for all x, y.
Equivalently, X and Y are independent if

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

Claim 3.4.1. 3 definitions are equivalent.

Proof.
We can get definition 2 from definition 1.

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

We can get definition 3 from definition 2.

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

We can get definition 1 from definition 3.

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

Therefore, 3 definitions are equivalent. □

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

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

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

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

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

Remark 3.4.3. We can generate σ-field using random variables.
σ-field generated by random variable $X = \sigma(X) = \{X^{-1}(E) : E \in \mathcal{B}(\mathbb{R})\} \subseteq \mathcal{F}$

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

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

Example 3.1. $\sigma(X) \perp\!\!\!\perp \sigma(Y) \iff X \perp\!\!\!\perp Y$

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

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

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

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

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

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

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

Example 3.2. (**Bernoulli random variable**) $X \sim \text{Bern}(p)$
Let $A \in \mathcal{F}$ be a specific event. A Bernoulli trial is success if A occurs. Let $X : \Omega \to \mathbb{R}$ be such that

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

Example 3.3. (**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 $(X_1, X_2, \cdots, X_n) = (\sigma_1, \sigma_2, \cdots, \sigma_n)$.
A_σ are mutually exclusive. Hence $\mathbb{P}(A) = \sum_\sigma \mathbb{P}(A_\sigma)$. There are totally $\binom{n}{k}$ different σ's in the sum.
By independence, we have

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

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

Example 3.4. (**Trinomial distribution**) Suppose we perform n trials, each of which result in three outcomes A, B and C, where A occurs with probability p, B with probability q, and C with probability $1 - p - q$.
Probability of r A's, w B's, and $n - r - w$ C's is

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

Example 3.5. (**Geometric distribution**) $W \sim \text{Geom}(p)$
Suppose we keep performing independent Bernoulli trials until the first success shows up. Let p be the probability of success and W be the **waiting time** which elapses before first success.

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

Example 3.6. (**Negative binomial distribution**) $W_r \sim \text{NBin}(r, p)$
Similar with examples of geometric distribution, let W_r be the waiting time for the r-th success. For $k \geq r$,

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

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

Example 3.7. (**Poisson distribution**) $X \sim \text{Poisson}(\lambda)$
Poisson variable is a random variable with Poisson PMF

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

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

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

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

Example 3.8. (**Poisson flips**) A coin is tossed once and heads turns up with probability p.
Let X and Y be the numbers of heads and tails respectively. X and Y are not independent since

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

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

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

3.2 Expectation of discrete random variables

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

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

Definition 3.7. Suppose we have discrete random variable X taking values from $\{x_1, x_2, \cdots\}$ with probability mass function $f_X(x)$. **Mean value**, **expectation**, or **expected value** of X is defined to be:

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

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

Example 3.9. 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 $Q(y)$?

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

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

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

whenever this sum is absolutely convergent.

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

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

□

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

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

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

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

□

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

Definition 3.10. If k is a positive integer, the **k-th moment** m_k of X is defined to be $m_k = \mathbb{E}(X^k)$.
The **k-th central moment** α_k is $\alpha_k = \mathbb{E}((X - \mathbb{E}X)^k) = \mathbb{E}((X - m_1)^k)$.
Mean μ of X is the 1st moment $m_1 = \mathbb{E}(X)$.
Variance of X is the 2nd central moment $\alpha_2 = \mathrm{Var}(X) = \mathbb{E}((X - m_1)^2) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}(X^2) - \mu^2$.
Standard deviation σ of X is defined as $\sqrt{\mathrm{Var}(X)}$.

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

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

We have the expectation and the variance of following distribution.

Example 3.10.

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

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

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

Proof.

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

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

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

□

Remark 3.11.1. More generally, we have

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

Example 3.11. 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 gets 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 number of additional cards we need to get in order to get a new type after having i distinct types.

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

$$\mathbb{E}Y_i = \frac{N}{N-i}$$

$$\mathbb{E}Y = \sum_{i=0}^{N-1}\mathbb{E}Y_i = N\left(\frac{1}{N} + \frac{1}{N-1} + \cdots + 1\right)$$

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

Proof.

$$\mathbb{E}(XY) = \sum_{x,y} xyf_{X,Y}(x,y) = \sum_{x,y} xyf_X(x)f_Y(y) = \sum_x xf_X(x) \sum_y yf_Y(y) = \mathbb{E}X\mathbb{E}Y$$

□

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

Proof.

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

□

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

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

Remark 3.14.1. The fact that X and Y are uncorrelated does not mean X and Y are independent.

Example 3.12. Let X be such that $f_X(0) = f_X(1) = f_X(-1) = \frac{1}{3}$ and Y be such that $Y = 0$ if $X \neq 0$ and $Y = 1$ if $X = 0$.

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

However,

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

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

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

Theorem 3.15. For random variables X and Y,

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

Proof.

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

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

2. When X and Y are uncorrelated,

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

□

Definition 3.16. Covariance of X and Y is:

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

Remark 3.16.1.

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

Remark 3.16.2. In general,

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

Example 3.13. If X_i are independent and $\mathrm{Var}(X_i) = 1$ for all i, then:

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

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

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

3.3 Conditional distribution of discrete random variables

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

Definition 3.17. Suppose $X, Y : \Omega \to \mathbb{R}$ are two 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

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

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

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

Remark 3.17.1. By definition,

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

.

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

Remark 3.17.3. If $X \perp\!\!\!\perp Y$, then $f_{Y|X}(y|x) = f_Y(y)$.

Conditional distributions still have properties of original distribution.

Lemma 3.18. 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 2.8.

□

Definition 3.19. Conditional expectation ψ of Y given $X = x$ is defined by:

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

Conditional expectation ψ of Y given X is defined by:

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

Example 3.14. Assume we roll a fair dice.

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

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

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

If we know that $X = x$, for example: $X = 1$ and $X = 0$

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

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

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

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

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

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

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

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

Remark 3.20.1. Under this definition,

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

Theorem 3.21. (Law of total expectation) Let $\psi(X) = \mathbb{E}(Y|X)$. Conditional expectation satisfies:

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

Proof.
By Lemma 3.8,

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

□

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

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

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

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

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

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

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

Theorem 3.22. Conditional expectation $\psi(X) = \mathbb{E}(Y|X)$ satisfies:

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

for any function g for which both expectations exist.

Proof.
By Lemma 3.8,

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

□

3.4 Convolution of discrete random variables

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

Theorem 3.23. Given two jointly discrete random variables X and Y.

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

Proof.
We have the disjoint union:

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

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

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

□

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

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

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

Chapter 4

Continuous random variables

4.1 Introduction of continuous random variables

We recall some definitions of continuous random variables.

Definition 4.1. 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 4.1.1. f_X is not prescribed uniquely since two integrable function which take identical values except at some specific point have the same integral.
But 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 4.1.2. 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 4.2. If random variable X has a density function f_X, then

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

Proof.

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

□

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

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

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

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

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

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

4.2 Expectation of continuous random variables

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

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

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

whenever this integral exists.

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

We have a special properties in the continuous random variable.

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

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

Proof.

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

□

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

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

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

Proof.

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

□

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

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

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

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

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

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

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

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

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

□

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

Definition 4.9. Given a positive integer k and a random variable X. **k-th moment** is defined to be

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

k-th central moment is defined to be

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

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

We have some important continuous distributions.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

□

These are some properties that are used frequently.

Lemma 4.10. The normal distribution has the following properties:

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

Proof.

1. Let $z = bx + a$.

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

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

2. Let $x = bz + a$.

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

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

3. Let $y = bz + a$.

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

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

□

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

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

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

Example 4.5. (**Cauchy distribution**) $X \sim$ Cauchy
Random variable X has a Cauchy distribution if:

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

It has the expectation

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

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

4.3 Joint distribution function of continuous random variables

Again, we recall the definition of joint distribution function.

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

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

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

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

We also recall the definition of marginal distribution function.

Definition 4.13. Marginal distribution function (Marginal PDF) of X given Y is

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

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

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

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

We can obtain the following important lemma.

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

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

Proof.

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

□

Example 4.6. Assume that a plane is ruled by horizontal lines separated by D and a needle of length $L \le D$ 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 $\Theta \sim \mathrm{U}\left[0, \frac{\pi}{2}\right]$.

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

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

Suppose that we throw the needle for n times.

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

It is useful to combine two normal distributions.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.4 Conditional distribution of continuous random variables

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

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

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

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

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

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

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

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

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

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

Remark 4.16.2. More generally, for two continuous random variables X and Y and $f_X(x) > 0$,

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

Example 4.10. Let X and Y have a JPDF:

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

Compute $f_X(x)$ and $f_{Y|X}(y|x)$. If $0 \le x \le 1$,

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

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

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

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

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

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

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

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

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

We 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 $(X|Y=y) \sim \mathrm{N}(\rho y, 1-\rho^2)$. If $\rho \to 1$, $X \to Y$. If $\rho \to -1$, $X \to -Y$.
In general, there exists a $Z \sim \mathrm{N}(0,1)$ 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 $\mathbf{A}$, if

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

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

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

Definition 4.17. Given an event $X = x$. **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 an random variable X. Conditional expectation of Y is defined by:

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

Again we also have the same properties of conditional distribution.

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

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

Proof.

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

□

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

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

Proof.

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

□

4.5 Functions of continuous random variables

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

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

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

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

Example 4.14. Let $X \sim \mathrm{N}(0,1)$. Let $Y = g(X) = X^2$. What is $f_Y(y)$?

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

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

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

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

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

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

$$F_Y(y) = \mathbb{P}(g(X) \le Y) = \mathbb{P}(X \le g^{-1}(Y)) = F_X(g^{-1}(Y))$$

$$f_Y(y) = F_Y'(y) = f_X(g^{-1}(y))\frac{\partial}{\partial y}g^{-1}(y) = f_X(g^{-1}(y))\left|\frac{\partial}{\partial y}g^{-1}(y)\right|$$

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

$$F_Y(y) = \mathbb{P}(g(X) \le y) = \mathbb{P}(X \ge g^{-1}(y)) = 1 - F_X(g^{-1}(y))$$

$$f_Y(y) = F_Y'(y) = -f_X(y^{-1}(y))\frac{\partial}{\partial y}g^{-1}(y) = f_X(g^{-1}(y))\left|\frac{\partial}{\partial y}g^{-1}(y)\right|$$

□

We can consider the multivariable case.

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

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

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

Then

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

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

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

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

Example 4.17. More specifically, if $X_1 \sim \mathrm{N}(0, 1)$ and $X_1 \perp\!\!\!\perp X_2$,

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

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

Example 4.18. If $X_1 \sim \mathrm{U}[0, 1]$ and $X_2 \sim \mathrm{U}[0, 1]$ and $X_1 \perp\!\!\!\perp X_2$, for all $x_1, x_2 \in \mathbb{R}$,

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

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

Theorem 4.21. If X and Y have JPDF $f_{X,Y}$, then $X + Y$ has a PDF

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

Proof.

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

□

Definition 4.22. Given $X \perp\!\!\!\perp Y$. **Convolution** f_{X+Y} $(f_X * f_Y)$ of PDFs of X and Y is the PDF of $X+Y$:

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

Example 4.19. If $X \sim \mathrm{U}[0,1]$ and $Y \sim \mathrm{U}[0,1]$. In case of $X \perp\!\!\!\perp Y$,

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

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

Example 4.20. If $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2, \cdots, n$ and they are independent, then $\sum_{i=1}^n X_i \sim \mathrm{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$.

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

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

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

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

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

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

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

□

Summary

Definition

Definition 1. **Sample space** Ω is the set of all outcomes ω of an experiment.

Definition 2. **Event** A is a subset of sample space. Outcomes are **elementary events**.

Definition 3. **Complement** of subset A is a subset $A^{\complement}$ which contains all elements in sample space Ω that is not in A.

Definition 4. **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 5. **Measurable space** $(\Omega, \mathcal{F})$ is a pair comprising a sample space Ω and a σ-field $\mathcal{F}$.

Definition 6. **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 7. **Probability space** $(\Omega, \mathcal{F}, \mathbb{P})$ is a triple comprising a sample space Ω, a σ-field $\mathcal{F}$ of certain subsets of Ω, and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$.

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

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

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

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

Definition 9. Event is **null** is $\mathbb{P}(A) = 0$.

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

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

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

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

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

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

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

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

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

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

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

Definition 14. Product space $(\Omega_1 \times \Omega_2, \mathcal{G}, \mathbb{P}(12))$ of two probability spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ is a probability space comprising a collection of ordered pairs $\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2 : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$, a σ-algebra $\mathcal{G} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ where $\mathcal{F}_1 \times \mathcal{F}_2 = \{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}$, and a probability measure $\mathbb{P}_{12} : \mathcal{F}_1 \times \mathcal{F}_2 \to [0,1]$ given by

$$\mathbb{P}_{12}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2) \qquad \text{for } A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

If it is absolutely convergent.

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

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

k-th cnetral moment α_k is defined to be

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

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

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

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

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

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

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

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

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

Definition 27. Given an event $X = x$. **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 an random variable X. Conditional expectation of random variable Y is defined by:

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

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

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

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

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

Named Property

Property 1. (Inclusion-exclusion 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 2. (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 3. (Law of total expectation) Let $\psi(X) = \mathbb{E}(Y|X)$. Conditional expectation satisfies:

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

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

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

Distributions

For discrete random variables,

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

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

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

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

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

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

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

$$f_W(k) = \mathbb{P}(W > k-1) - \mathbb{P}(W > k) = p(1-p)^{k-1} \quad F_W(k) = 1 - \mathbb{P}(W > k) = 1 - (1-p)^k \quad \mathbb{E}X = p^{-1} \quad \text{Var}(X) = (1-p)p^{-2}$$

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

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

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

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

For continuous random variables,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Chapter 5

Generating function

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 Bernoulli 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 $\frac{1}{2}(n + b - a)$ is 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-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 $S_1, S_2, \cdots, S_n \neq 0$ iff $S_1 S_2 \cdots S_n \neq 0$.
Event $S_1 S_2 \cdots S_n \neq 0$ occurs iff 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,

$$\mathbb{P}(S_1 S_2 \cdots S_n \neq 0, S_n = b) = \frac{b}{n} N_n(0,b) p^{\frac{1}{2}(n+b)} q^{\frac{1}{2}(n-b)} = \frac{b}{n}\mathbb{P}(S_n = b).$$

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 $N_n^r(0,b) = N_n(0, 2r-b)$.

$$\mathbb{P}(M_n \geq r, S_n = b) = N_n^r(0,b) p^{\frac{1}{2}(n+b)} q^{\frac{1}{2}(n-b)} = \left(\frac{q}{p}\right)^{r-b} N_n(0, 2r-b) p^{\frac{1}{2}(n+2r-b)} q^{\frac{1}{2}(n-2r+b)} = \left(\frac{q}{p}\right)^{r-b} \mathbb{P}(S_n = 2r-b)$$

□