Probability Theory

Daniela Rodriguez

Universidad Torcuato Di Tella and CONICET

December 23, 2025

Contents

Chapter 1

Probability spaces

1.1 Introduction

Probability is the mathematics of uncertainty. One of the main distinctions we need to make is between the concept of a random experiment and a deterministic experiment. In probability, the main concept is a random experiment, whose outcome cannot be determined in advance. The toss of a coin and an election poll are simple examples

Consider the following experiment: a box has 4 red balls, 6 blue balls and 10 green balls. We pick a ball at random. What is the probability of that ball being red? We are taught in school that this should be the number of red balls over the total number of balls, so $\frac{4}{20} = 0.2 = 20\%$ and this is indeed true under certain assumptions. To understand the assumptions we are implicitly making when doing this computation, we ask the following questions:

> What is a probability as a mathematical object?
>
> Would the answer be the same if some balls are of difference sizes?

Note that when we ask about a probability, we need to determine the event whose probability we are interested in – while the probability of a specific event (e.g. 'the ball is red') is a number in $[0, 1]$, the probability on its own, is a map that attaches to each event a number. There are three possible outcomes for this experiment: red, blue and green.

The set of all elementary events is called a **Sample space** and is denoted by Ω. Elements of Ω, that is elementary events, are denoted by $\omega \in \Omega$.

As a mathematical object, Ω is any non-empty set – in this case, $\Omega = \{\text{red, blue, green}\}$

Example 1.

- *We throw a coin. There are two results: heads or tails. Thus* $\Omega = \{H,\ T\}$, $|\Omega| = 2$.
- *We throw a die. There are six possible results. Thus* $\Omega = \{1, 2, 3, 4, 5, 6\}$, $|\Omega| = 6$.

Often we are not interested in a concrete result of an experiment but we just want to know if this result belongs to a subset of Ω. Such subsets are called **events** and we denote them with capital letters: A, B, C, D etc.

Example 2.

- *We throw two dice. Let A be an event that the sum of spots is equal to 5. Then* $\Omega = \{(i, j) \mid 1 \leq i \leq 6, 1 \leq j \leq 6\}$, $A = \{(1, 4), (2, 3), (3, 2), (4, 1)\}$.
- *We throw a coin until we get heads. Let A be an event where at most 3 trials were done.* $\Omega = \{(H), (T,H), (T,T,H), (T,T,T,H), \dots\}$. $A = \{(H), (T,H), (T,T,H)\}$.

Since we'll be working with sets, we'll need to perform operations on them. So, let's review the notation.

Remark 1.

1. Ω - *the sample space*
2. $\emptyset$ - *the impossible event*
3. $A \cap B$ - *events A and B both occurred*
4. $A \cap B = \emptyset$ - *events A and B are mutually exclusive*
5. $A \cup B$ - *either A or B occurred*
6. $A^c = \Omega \setminus A$ - *A did not happen*
7. $A \setminus B = A \cap B^c$ - *A happened and B did not happen*
8. $A \subseteq B$ - *event $A \neq \emptyset$ leads to event B*

For example, we can ask for the probability that 'the ball is either blue or green' (which would have been equivalent to 'ball is not red'). In words, an event is a statement that you can tell whether it is true or not, after seeing the outcome of the experiment. In this case, all possible events are

'The ball is none of the three colours or any other colour' – mathematically, this will be denoted by the empty set $\emptyset$, since it contains none of the possible outcomes.
'The ball is red' – denoted by $\{red\}$
'The ball is blue' – denoted by $\{blue\}$
'The ball is green' – denoted by $\{green\}$
'The ball is either red or blue' – denoted by $\{red, blue\}$
'The ball is either red or green' – denoted by $\{red, green\}$
'The ball is either blue or green' – denoted by $\{blue, green\}$
'The ball is any of red, blue or green' – denoted by $\{red, blue, green\} = \Omega$.

From this exhaustive list, it is clear that all events are subset of Ω and in fact, in this case at least, all subsets of Ω are events.

Assume we have a fixed sample space Ω. We want to distinguish a family of events, that we want to consider. We call the collection of all events the event space, usually denoted by $\mathcal{F}$.

A first choice is to take: 2^Ω = all subsets of Ω. This is a good choice when Ω is at most countable set. When $|\Omega| > \aleph_0$, 2^Ω is too big and there are problems with defining probability on 2^Ω. The set of all possible subsets of a given set is often denoted by $\mathcal{P}(\Omega)$ or 2^Ω and it is called the power set. We will discuss these problems later. That is why we need to choose a smaller family. On the other hand, $\mathcal{F}$ should be closed with respect to taking unions, intersections and complements.

Definition 1. *A family $\mathcal{F}$ of subsets of Ω is called a σ-**algebra** if:*

1. $\emptyset \in \mathcal{F}$
2. $A \in \mathcal{F} \implies A^c \in \mathcal{F}$
3. $A_1, A_2, \ldots, A_n, \cdots \in \mathcal{F} \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$

*A pair $(\Omega, \mathcal{F})$ is called a **measurable space**.*

Remark 2. *For any Ω the pair $(\Omega, 2^\Omega)$ is a measurable space. Also for any nonempty subset $A \subset \Omega$ the smallest σ-algebra that contains A is $\sigma(A) = \{\emptyset, A, A^c, \Omega\}$.*

Fact 1. *Let $(\Omega, \mathcal{F})$ be a measurable space. Then*

1. $A, B \in \mathcal{F} \implies A \setminus B \in \mathcal{F}$
2. $A_1, A_2, \ldots, A_n, \cdots \in \mathcal{F} \implies \bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$

Proof.

1. $A \setminus B = A \cap B^c = (A^c \cup B)^c$
2. $\bigcap_{k=1}^{\infty} A_k = (\bigcup_{k=1}^{\infty} A_k^c)^c$, but $A_k^c \in \mathcal{F}$.

Definition 2. ***Borel σ-algebra of*** $\mathbb{R}^d$, $\mathcal{B}(\mathbb{R}^d)$, *is the smallest σ-algebra that contains all open subsets of $\mathbb{R}^d$. Elements of $\mathcal{B}(\mathbb{R}^d)$ are called* ***Borel subsets****.*

Example 3. *Let $d = 1$, $\mathcal{B}(\mathbb{R})$ is the smallest σ-algebra that contains all open sets in $\mathbb{R}$.*

- *Intervals (a, b), where $a, b \in \mathbb{R}$ are in $\mathcal{B}(\mathbb{R})$,*
- $(a, b] \in \mathcal{B}(\mathbb{R})$ *as* $(a, b] = \bigcap_{n=1}^{\infty} \left(a, b + \frac{1}{n}\right)$,
- $[a, b) \in \mathcal{B}(\mathbb{R})$ *as* $[a, b) = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, b\right)$,
- $[a, b] \in \mathcal{B}(\mathbb{R})$ *as* $[a, b] = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, b + \frac{1}{n}\right)$,
- $\{a\} \in \mathcal{B}(\mathbb{R})$ *as* $\{a\} = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, a + \frac{1}{n}\right)$,
- $(-\infty, a) \in \mathcal{B}(\mathbb{R})$ *(or* $(-\infty, a) = \bigcup_{n=1}^{\infty}(-n, a)$*)* $\implies [a, \infty) \in \mathcal{B}(\mathbb{R})$,
- $(a, \infty) \in \mathcal{B}(\mathbb{R})$ *(or* $(a, \infty) = \bigcup_{n=1}^{\infty}(a, n)$*)* $\implies (-\infty, a] \in \mathcal{B}(\mathbb{R})$,
- $\mathbb{Q} \in \mathcal{B}(\mathbb{R})$ *as a countable union of points,*
- $\mathbb{R} \setminus \mathbb{Q} \in \mathcal{B}(\mathbb{R})$ *as the complement of $\mathbb{Q}$.*

Remark 3. *Not every subset of $\mathbb{R}^d$ is a Borel subset. An example of non-Borel set is a Vitali set $V \subset [0, 1]$.*

1.2 Probability

We next define the notion of probability. First, however, in order to get some intuition we consider the **frequency of an event**. Assume that we can repeat an experiment n times. Each repetitions happens under the same conditions. We define a relative frequency of an event $A \in \mathcal{F}$ in the series of n experiments by:

$$f_n(A) = \frac{\#\text{ experiments in which } A \text{ happened}}{n}$$

When n is large we expect that $f_n(A)$ is close to the chance A occurs in a single trial. We easily check that f_n takes values in $[0, 1]$ and

1. $f_n(\Omega) = 1$
2. If $A_1, \ldots, A_j$ are pairwise disjoint, then

$$f_n\left(\bigcup_{k=1}^{j} A_k\right) = \sum_{k=1}^{j} f_n(A_k).$$

This is because # experiments in which $\bigcup_{k=1}^{j} A_k$ happened is
$\sum_{k=1}^{j}$(# experiments in which A_k happened).

These are fundamental properties that a probability map should have. Are these sufficient for infinite probability spaces? Let us consider the following:

Example 4. *Let $\Omega = \mathbb{N}^* = \{1, 2, ...\}$ be the positive natural numbers and $\mathcal{F} = \mathcal{P}(\Omega)$. Suppose that $P(\{n\}) = \frac{1}{2^n}$, for every $n \geq 1$. What would we expect the event $\{2n | n \geq 1\}$ ('the outcome is an even number') to be?*

Intuitively, what we would do is to sum up the probabilities corresponding to the outcome being even, i.e.,

$$P(\{2n|n \geq 1\}) = \sum_{n=1}^{\infty} P(\{2n\}) = \sum_{n=1}^{\infty} \frac{1}{2^{2n}} = \sum_{n=1}^{\infty} \frac{1}{4^n}$$

(Note that the event $\{n\}$ corresponds to 'the outcome is n'). The computation then follows as:

$$\sum_{n=1}^{\infty} \frac{1}{4^n} = \frac{1}{1 - \frac{1}{4}} - 1 = \frac{1}{\frac{3}{4}} - 1 = \frac{1}{3}$$

The computation above cannot be justified, unless we extend the property of finite additivity to also hold for countable unions of disjoint events. Indeed, this is what we do!

Definition 3. *Given a sample space Ω and an event space $\mathcal{F}$, a function $P : \mathcal{F} \to \mathbb{R}$ is called a* probability measure *if it satisfies*

- *$P(B) \in [0, 1]$ for every $B \in \mathcal{F}$;*
- *$P(\Omega) = 1$;*
- *(Countable additivity) For every $A_n \in \mathcal{F}$, $n > 1$ disjoint events (i.e. for all $m, n >$ such that $m \neq n$, $A_m \cap A_n = \emptyset$),*

$$P\left[\bigcup_{n=1}^{\infty} A_n\right] = \sum_{n=1}^{\infty} P(A_n).$$

We now give the definition of an abstract probability space.

Definition 4. *A* probability space *is defined as the triplet $(\Omega, \mathcal{F}, P)$, where*

- *Ω (the* sample space*) is the set of all possible outcomes of the experiment (we always assume that it is not empty);*
- *$\mathcal{F}$ is an event space of subsets of Ω.*
- *P is a probability measure on $\mathcal{F}$.*

Theorem 1. *Assume $(\Omega, \mathcal{F}, P)$ is a probability space, and $A, B, A_1, \ldots, A_n, \cdots \in \mathcal{F}$, then*

1) $P(\emptyset) = 0$

2) $P(A^c) = 1 - P(A)$

3) If $A \subseteq B$, then $P(B \setminus A) = P(B) - P(A)$ and $P(B) \geq P(A)$

4) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

5) $P\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} P(A_k)$

Proof.

1. $\Omega \cup \emptyset = \Omega$, $\Omega \cap \emptyset = \emptyset \implies 1 = P(\Omega) = P(\Omega) + P(\emptyset) \implies P(\emptyset) = 0$.
2. $\Omega = A \cup A^c$ and $A \cap A^c = \emptyset \implies 1 = P(\Omega) = P(A) + P(A^c) \implies P(A^c) = 1 - P(A)$.
3. $B = (B \setminus A) \cup A$ and $(B \setminus A) \cap A = \emptyset \implies P(B) = P(B \setminus A) + P(A)$.
4. $A \cup B = (A \setminus (A \cap B)) \cup (A \cap B) \cup (B \setminus (A \cap B))$.
$$P(A \cup B) = P(A \setminus (A \cap B)) + P(A \cap B) + P(B \setminus (A \cap B))$$
$$= (P(A) - P(A \cap B)) + P(A \cap B) + (P(B) - P(A \cap B)).$$
5. Let $B_1 = A_1$, $B_2 = A_2 \setminus B_1$, $B_3 = A_3 \setminus (B_1 \cup B_2)$, $\ldots$, $B_n = A_n \setminus \bigcup_{k=1}^{n-1} B_k$. The B_k's are mutually exclusive. $\bigcup_{k=1}^{\infty} A_k = \bigcup_{k=1}^{\infty} B_k$, and $B_n \subset A_n$, for all n.
Thus $P\left(\bigcup_{k=1}^{\infty} A_k\right) = P\left(\bigcup_{k=1}^{\infty} B_k\right) = \sum_{k=1}^{\infty} P(B_k) \leq \sum_{k=1}^{\infty} P(A_k)$.

We can use countable additivity to compute the probability of a union of disjoint events. How can we compute the probability of any union of events? The following proposition gives us a way to do this.

Theorem 2 (Inclusion-Exclusion Formula)**.** *Assume $A_1, A_2, \ldots, A_n \in \mathcal{F}$, then:*

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} P(A_{i_1} \cap \cdots \cap A_{i_k})$$

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

The sum $\sum_{1 \leq i_1 < \cdots < i_k \leq n}$ is to be interpreted as the sum going through all k-tuples $(i_1, \ldots, i_k)$ of numbers $\{1, \ldots, n\}$ with no repetition (inequalities are strict).

This Formula uses concise notation and is not straightforward to interpret. To understand it better, let us consider some specific cases.

For $n = 2$:

$$P\left[\bigcup_{k=1}^{2} A_k\right] = \sum_{k=1}^{2} (-1)^{k-1} \sum_{1 \leq i_1 < \cdots < i_k \leq 2} P(A_{i_1} \cap \cdots \cap A_{i_k})$$
$$= \sum_{1 \leq i \leq 2} P(A_i) - \sum_{1 \leq i_1 < i_2 \leq 2} P(A_{i_1} \cap A_{i_2})$$
$$= P(A_1) + P(A_2) - P(A_1 \cap A_2).$$

For $n = 3$:

$$P\left[\bigcup_{k=1}^{3} A_k\right] = \sum_{k=1}^{3} (-1)^{k-1} \sum_{1 \leq i_1 < \cdots < i_k \leq 3} P(A_{i_1} \cap \cdots \cap A_{i_k})$$
$$= \sum_{1 \leq i \leq 3} P(A_i) - \sum_{1 \leq i_1 < i_2 \leq 3} P(A_{i_1} \cap A_{i_2}) + \sum_{1 \leq i_1 < i_2 < i_3 \leq 3} P(A_{i_1} \cap A_{i_2} \cap A_{i_3})$$
$$= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3)$$
$$- P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3).$$

Proof. (by induction): By Theorem 1 we know that Theorem 2 is true for $n = 2$. Assume it is true for $n \geq 2$. We need to show that it is true for $n+1$. $P(A_1 \cup A_2 \cup \cdots \cup A_n \cup A_{n+1}) = P((A_1 \cup A_2 \cup \cdots \cup A_n) \cup A_{n+1})$
$= P(A_1 \cup A_2 \cup \cdots \cup A_n) + P(A_{n+1}) - P((A_1 \cup A_2 \cup \cdots \cup A_n) \cap A_{n+1})$
$= P(A_1 \cup A_2 \cup \cdots \cup A_n) + P(A_{n+1}) - P((A_1 \cap A_{n+1}) \cup \cdots \cup (A_n \cap A_{n+1}))$.
Now apply the induction hypothesis to $P(A_1 \cup \cdots \cup A_n)$ and $P((A_1 \cap A_{n+1}) \cup \cdots \cup (A_n \cap A_{n+1}))$.
The full expansion becomes:

$$P(A_1 \cup A_2 \cup \cdots \cup A_n \cup A_{n+1}) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} P(A_{i_1} \cap \cdots \cap A_{i_k}) + P(A_{n+1})$$

$$- \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} P(A_{i_1} \cap \cdots \cap A_{i_k} \cap A_{n+1})$$

Combining these three formulas we get the result.

Example 5. ***Equiprobability.***: *Ω - a finite set, $\mathcal{F} = 2^\Omega$, all elementary events have the same probability. Then for all $\omega \in \Omega$ and for all $A \subseteq \Omega$ $(A \in \mathcal{F})$*

$$P(\{\omega\}) = \frac{1}{|\Omega|} \tag{1.1}$$

$$P(A) = \frac{|A|}{|\Omega|} \tag{1.2}$$

Since $\forall \omega_1, \omega_2 \in \Omega$, $P(\{\omega_1\}) = P(\{\omega_2\})$, let $p \in [0,1]$ such that $p := P(\{\omega\}) \forall \omega \in \Omega$.
Since P is a probability measure

$$1 = P(\Omega) = \sum_{\omega \in \Omega} P(\{\omega\}) = \sum_{\omega \in \Omega} p = p \sum_{\omega \in \Omega} 1 = p|\Omega|.$$

$$p = \frac{1}{|\Omega|},$$

We see that for $A \subseteq \Omega$ as

$$P(A) = \sum_{\omega \in A} P(\{\omega\}) = \sum_{\omega \in A} p = p \sum_{\omega \in A} 1 = p|A| = \frac{|A|}{|\Omega|}.$$

Then for any $A \in \mathcal{F}$

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

Example 6. *$\Omega = \{\omega_1, \ldots, \omega_n, \ldots\}$ - a countable set. Let $p_1, \ldots, p_n, \ldots$ - sequence of non-negative numbers s.t. $\sum_{k=1}^{\infty} p_k = 1$. We can choose $\mathcal{F} = 2^\Omega$ and $P(\{\omega_i\}) = p_i$. This choice defines the probability space $(\Omega, \mathcal{F}, P)$, and for any $A \in \mathcal{F}$ we have*

$$P(A) = \sum_{k=1}^{\infty} \mathbf{1}_A(\omega_k) p_k$$

where

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

Example 7. ***Uniform probability****: We choose $\Omega \subset \mathbb{R}^d$, that is Ω is a Borel subset of $\mathbb{R}^d$ and we assume that $0 < |\Omega| < \infty$, where $|\Omega| = \int_{\mathbb{R}^d} \mathbf{1}_\Omega$ is the Lebesgue measure[1]. Let $\mathcal{F} = \mathcal{B}(\Omega)$ the smallest $\sigma-$algebra that contains all open subsets of Ω and for $A \in \mathcal{F} = \mathcal{B}(\Omega)$,*

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

Then $(\Omega, \mathcal{F}, P)$ is a probability space. We use this probability space to describe experiments where a point(s) are randomly chosen from Ω.

[1]This integral is actually the Lebesgue integral. For a Riemann integrable function, the Lebesgue integral is equal to the Riemann integral. We will always consider subsets of $\mathbb{R}^d$ whose characteristic functions are Riemann integrable.

1.3 Combinatorial

As we saw earlier, if the sample space is finite and we can assume equal probability, the problems of calculating probabilities are reduced to being able to compute the sizes of the sets involved. In other words, we need to be good at counting the number of elements in various sets. The science of counting is called combinatorics. Next, we will consider some simple combinatorial rules and their application in probability theory when a uniform distribution is appropriate.

Counting Permutations

Suppose four friends go to a restaurant, and each checks his or her coat. At the end of the meal, the four coats are randomly returned to the four people. What is the probability that each of the four people gets his or her own coat? Here the total number of different ways the coats can be returned is equal to $4 \times 3 \times 2 \times 1$, or 4! (i.e., four factorial). This is because the first coat can be returned to any of the four friends, the second coat to any of the three remaining friends, and so on. Only one of these assignments is correct. Hence, the probability that each of the four people gets his or her own coat is equal to $\frac{1}{4!}$, or $\frac{1}{24}$.

Here we are counting permutations, or sequences of elements from a set where no element appears more than once. We can use the multiplication principle to count permutations more generally. For example, suppose that we have the set S with n different objects and we want to count the number of permutations of length $k \leq n$ obtained from S, i.e., we want to count the number of elements of the set

$$\{s_1, \ldots, s_k : s_i \in S, s_i \neq s_j \text{ when } i \neq j\}$$

Then we have n choices for the first element s_1, $n-1$ choices for the second element, and finally $n-k+1$ choices for the last element. So there are

$$n(n-1)\ldots(n-k+1)$$

permutations of length k from a set of n elements. This can also be written as $\frac{n!}{(n-k)!}$. Notice that when $k = n$ there are $n! = n(n-1)\ldots 2 \cdot 1$.

Counting Subsets

Suppose 10 fair coins are flipped. What is the probability that exactly seven of them are heads? Here each possible sequence of 10 heads or tails (e.g., H H H T T T H T T T, T H T T T T H H H T, etc.) is equally likely, and by the multiplication principle the total number of possible outcomes is equal to 2 multiplied by itself 10 times, or $2^{10} = 1024$. Hence, the probability of any particular sequence occurring is $\frac{1}{1024}$. But of these sequences, how many have exactly seven heads?

To answer this, notice that we may specify such a sequence by giving the positions of the seven heads, which involves choosing a subset of size 7 from the set of possible indices $\{1, \ldots, 10\}$. There

are $\frac{10!}{3!} = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4$ different permutations of length 7 from $\{1, \ldots, 10\}$, and each such permutation specifies a sequence of seven heads and three tails. But we can permute the indices specifying where the heads go in 7! different ways without changing the sequence of heads and tails. So the total number of outcomes with exactly seven heads is equal to $\frac{10!}{3!7!} = 120$. The probability that exactly seven of the 10 coins are heads is therefore equal to $\frac{120}{1024}$, or just under 12%.

In general, if we have a set S of n elements, then the number of different subsets of size k that we can construct by choosing elements from S is

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

which is called the binomial coefficient. This follows by the same argument, namely, there are $\frac{n!}{(n-k)!}$ permutations of length k obtained from the set; each such permutation, and the $k!$ permutations obtained by permuting it, specify a unique subset of S.

Counting Sequences of Subsets and Partitions

When we want to divide a larger set into several smaller, non-overlapping subsets, we can use a powerful counting method based on the multiplication principle. For example, how many different ways can a deck of 52 cards be divided up into four hands of 13 cards each, with the hands labelled North, East, South, and West, respectively?

1. **North Hand (N):** Choose 13 cards from 52: $\binom{52}{13}$
2. **East Hand (E):** 39 cards remain. Choose 13 from 39: $\binom{39}{13}$
3. **South Hand (S):** 26 cards remain. Choose 13 from 26: $\binom{26}{13}$
4. **West Hand (W):** 13 cards remain. Choose 13 from 13: $\binom{13}{13} = 1$

The total number of ways is the product of these combinations:

$$\text{Total Ways} = \binom{52}{13} \cdot \binom{39}{13} \cdot \binom{26}{13} \cdot \binom{13}{13}$$

Expanding the factorial terms shows a cancellation pattern (telescoping product):

$$\text{Total Ways} = \left(\frac{52!}{13!39!}\right) \cdot \left(\frac{39!}{13!26!}\right) \cdot \left(\frac{26!}{13!13!}\right) \cdot 1 = \frac{52!}{13!13!13!13!}$$

This equals

$$\binom{52}{13, 13, 13, 13} = \frac{52!}{13!13!13!13!} \approx 5.364 \times 10^{28}$$

which is a very large number.

In general, suppose we have a set S of n elements and we want to count the number of elements of

$$\{S_1, S_2, \ldots, S_l : S_i \subset S, |S_i| = k_i, S_i \cap S_j = \emptyset \text{ when } i \neq j\}$$

namely, we want to count the number of sequences of l subsets of a set where no two subsets have any elements in common and the i-th subset has k_i elements. By the multiplication principle, this equals

$$\binom{n}{k_1}\binom{n-k_1}{k_2}\cdots\binom{n-k_1-\cdots-k_{l-1}}{k_l} = \frac{n!}{k_1!\cdots k_l!(n-k_1-\cdots-k_l)!}$$

because we can choose the elements of S_1 in $\binom{n}{k_1}$ ways, choose the elements of S_2 in $\binom{n-k_1}{k_2}$ ways, etc.

When we have that $S = S_1 \cup S_2 \cup \cdots \cup S_l$, in addition to the individual sets being mutually disjoint, then we are counting the number of ordered partitions of a set of n elements with k_1 elements in the first set, k_2 elements in the second set, etc. In this case, the previous expression equals

$$\binom{n}{k_1, k_2, \ldots, k_l} = \frac{n!}{k_1! k_2! \cdots k_l!}$$

which is called the multinomial coefficient.

1.4 Borel σ-algebra

Fact 2. *Any open set $A \subset \mathbb{R}$ is a countable union of open intervals.*

Proof. Let A be a non-empty open set in $\mathbb{R}$. For each point $x \in A$, there exists an open interval (a, b) such that $x \in (a, b) \subset A$.

Consider the set of all open intervals with rational endpoints that are subsets of A:

$$S = \{(p, q) \mid p, q \in \mathbb{Q} \text{ and } (p, q) \subset A\}$$

The set S is countable because the set of all pairs of rational numbers $\mathbb{Q} \times \mathbb{Q}$ is countable.

We now show that $A = \bigcup_{(p,q)\in S}(p, q)$.

($\subseteq$) Let $x \in A$. Since A is an open set, there exists an open interval (a, b) such that $x \in (a, b) \subset A$. Because the rational numbers are dense in $\mathbb{R}$, we can choose rational numbers p and q such that $a < p < x < q < b$. This implies that $x \in (p, q)$ and $(p, q) \subset (a, b) \subset A$. Therefore, the interval (p, q) is in the set S, and thus x is in the union $\bigcup_{(p,q)\in S}(p, q)$.

($\supseteq$) This direction is straightforward. Every interval (p, q) in the union is a subset of A by the definition of the set S. Therefore, their union must also be a subset of A.

Since both inclusions hold, we have $A = \bigcup_{(p,q)\in S}(p, q)$. As the set S is countable, A is a countable union of open intervals.

Definition 5. *We say that σ-algebra $\mathcal{F}$ is **generated by a set** $\mathcal{S} \subset \mathcal{F}$ if $\mathcal{F}$ is the smallest σ-algebra that contains $\mathcal{S}$. We write $\mathcal{F} = \sigma(\mathcal{S})$.*

Recall that we have defined the Borel σ-algebra, $\mathcal{B}(\mathbb{R})$, as the one generated by open intervals on the real line. However, a fundamental result is that this definition is equivalent to considering the σ-algebra generated by semi-infinite intervals of the form $(-\infty, a]$. Below, we will formally demonstrate this equivalence.

Theorem 3. *The Borel σ-algebra $\mathcal{B}(\mathbb{R})$ is generated by intervals of the form $(-\infty, a]$, $a \in \mathbb{R}$*

Proof. Let $\mathcal{D}$ denote the collection of intervals of the form $(-\infty, a]$, $a \in \mathbb{R}$. In the example 3 we prove that $\mathcal{D} \subset \mathcal{B}(\mathbb{R})$, therefore $\sigma(\mathcal{D}) \subset \mathcal{B}(\mathbb{R})$.

Now in order to prove the other inclusion, let us denote by $\mathcal{O}$ the collection of all open intervals. Of course $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{O})$.

One easily sees that any open interval $(a, b) \in \sigma(\mathcal{D})$:

$$(a, b) = \bigcup_{n=1}^{\infty}(-\infty, b + \frac{1}{n}] \setminus (-\infty, a] = \bigcup_{n=1}^{\infty}((-\infty, +\frac{1}{n}] \cap (-\infty, a]^c).$$

Thus $\mathcal{O} \subset \sigma(\mathcal{D})$ which means $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{O}) \subset \sigma(\mathcal{D})$.

Finally, we have that $\sigma(\mathcal{D}) = \mathcal{B}(\mathbb{R})$.

Remark 4. *Analogous statements are true for $\mathbb{R}^d$, $d > 1$, that is*

1. *any open set in $\mathbb{R}^d$ is a countable union of $(a_1, b_1) \times \cdots \times (a_d, b_d)$.*
2. *sets $\{(-\infty, a_1] \times \cdots \times (-\infty, a_d] \mid (a_1, \ldots, a_d) \in \mathbb{R}^d\}$ generate $\mathcal{B}(\mathbb{R}^d)$.*

Chapter 2

Conditional Probability and Independence

2.1 Conditional Probability

Suppose that before rolling a fair die, you bet one pound that the outcome is 3. Your friend sees the result before you and tells you that the die shows an even number. Would you continue your bet or withdraw from it? What about if you are told that the outcome is odd? How does this partial information about the outcome changes the probability?

We model the probability space corresponding to rolling a fair die by taking $\Omega = \{1, 2, ..., 6\}$, $\mathcal{F}$ to be the power set and P as the uniform probability on it.

Then, the event our friend tells us happened is $B = \{2, 4, 6\}$, and its probability is $P(B) = \frac{3}{6} = \frac{1}{2} > 0$. The favourable event for us is $A = \{3\}$ and its probability is $P(A) = \frac{1}{6}$.

Knowing that the outcome is even can be interpreted as changing the sample space from Ω to B. Intuitively, we would assume that the probability on the new sample space remains uniform, but the probability of each outcome changes from $\frac{1}{6}$ to $\frac{1}{3}$ since there are now only 3 outcomes.

Given that our preferred outcome 3 is not in the new sample space, we would expect the probability of getting 3 to be 0 and thus it would make sense to withdraw from the bet.

If, in the other hand, we were told that the outcome is odd, then we could reformalise the probability space as one with sample space $B^c = \{1, 3, 5\}$ and we would expect the probability of winning the bet to be $\frac{1}{3}$, as it is one of 3 possible outcomes.

Intuitively, we'd expect the updated probability, given event B has occurred, to be calculated considering the ways to win, divided by the total possible outcomes after getting the new information. But is this a mathematically sound way to define probability?"

Definition 6. *Assume $(\Omega, \mathcal{F}, P)$ is a probability space and $A, B \in \mathcal{F}$ are events such that $P(B) > 0$. The **conditional probability** of an event A under the condition that B occurred is given by:*

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

Proposition 1. *For fixed $B \in \mathcal{F}$ that satisfies $P(B) > 0$, the triple $(\Omega, \mathcal{F}, P(\cdot|B))$ is a probability space*

Proof.

1. For any $E \in \mathcal{F}$, $P(E|B) = \frac{P(E \cap B)}{P(B)} \geq 0$.
2. $P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$.

3. $(A_n)_{n=1}^{\infty}$ - countable sequence of mutually exclusive events (pairwise disjoint)

$$P\left(\bigcup_{n=1}^{\infty} A_n \middle| B\right) = \frac{P\left(\left(\bigcup_{n=1}^{\infty} A_n\right) \cap B\right)}{P(B)} = \frac{P\left(\bigcup_{n=1}^{\infty}(A_n \cap B)\right)}{P(B)}$$

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

Fact 3.

1. *$A, B \in \mathcal{F}$ and $A \cap B = \emptyset \implies P(A|B) = 0$.*
2. *$A, B \in \mathcal{F}$ and $B \subset A \implies P(A|B) = 1$.*

Example 8. *An experiment consists of tossing a fair coin 7 times.*

(a) Describe the probability space associated to it.

(b) Let C be the event corresponding to getting a prime number of heads. What is $P(C)$?

(c) Let B be the event "H occurs at least 6 times". What is $P(C|B)$?

Solution.

(a) $\Omega = \{(a_1, ..., a_7) : a_i \in \{H, T\}\}$, $\mathcal{F}$ the power set of Ω and P the uniform probability, i.e. P is such that $\forall A \in \mathcal{F}$, $P(A) = \frac{|A|}{|\Omega|}$.

Recall that $|\Omega| = 2^7$.

(b) For $i = 1, ..., 7$ let A_i be the event "we get exactly i heads".

The elements of A_i can be uniquely characterised via the position $Heads$ appearing in the sequence. Hence by the fundamental counting principle $|A_i| = \binom{7}{i}$.

Thus $P(A_i) = \frac{1}{2^7}\binom{7}{i}$. Now, notice that $A_i \cap A_j = \emptyset$ for $i \neq j$ (no outcome has both i and j heads) and $C = A_2 \cup A_3 \cup A_5 \cup A_7$.

Then, by finite additivity

$$P(C) = P(A_2) + P(A_3) + P(A_5) + P(A_7) = \frac{1}{2^7}\binom{7}{2} + \frac{1}{2^7}\binom{7}{3} + \frac{1}{2^7}\binom{7}{5} + \frac{1}{2^7}\binom{7}{7} = \frac{78}{128}.$$

(c) B is the event "H appears at least 6 times", so $B = A_6 \cup A_7$. Notice that,

$$P(B) = P(A_6) + P(A_7) = \frac{1}{2^7}\binom{7}{6} + \frac{1}{2^7}\binom{7}{7} = \frac{7!}{6!1!} \cdot \frac{1}{2^7} + \frac{7!}{7!0!} \cdot \frac{1}{2^7} = \frac{7+1}{2^7} = \frac{8}{128} = \frac{1}{16} > 0.$$

Now, we can compute $P(C|B)$. By definition, $P(C|B) = \frac{P(C \cap B)}{P(B)}$.

Since $C \cap B = (A_2 \cup A_3 \cup A_5 \cup A_7) \cap (A_6 \cup A_7) = A_7$, we have

$$P(C|B) = \frac{P(A_7)}{P(B)} = \frac{\frac{1}{2^7}\binom{7}{7}}{\frac{8}{128}} = \frac{\frac{1}{128}}{\frac{8}{128}} = \frac{1}{8}.$$

Theorem 4 (Multiplication Rule)**.** *Assume $(\Omega, \mathcal{F}, P)$ is a probability space and $A_1, \ldots, A_n \in \mathcal{F}$ are events satisfying $P(A_1 \cap A_2 \cap \cdots \cap A_n) > 0$. Then*

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_n|A_1 \cap \cdots \cap A_{n-1})P(A_{n-1}|A_1 \cap \cdots \cap A_{n-2}) \ldots P(A_2|A_1)P(A_1)$$

Proof.

$$P(A_n|A_1 \cap \cdots \cap A_{n-1}) \cdot P(A_{n-1}|A_1 \cap \cdots \cap A_{n-2}) \cdot \cdots \cdot P(A_2|A_1) \cdot P(A_1)$$
$$= \frac{P(A_n \cap (\cdots \cap A_{n-1}))}{P(A_1 \cap \cdots \cap A_{n-1})} \cdot \frac{P(A_{n-1} \cap (\cdots \cap A_{n-2}))}{P(A_1 \cap \cdots \cap A_{n-2})} \cdot \cdots \cdot \frac{P(A_1 \cap A_2)}{P(A_1)} \cdot P(A_1)$$
$$= P(A_1 \cap A_2 \cap \cdots \cap A_n)$$

Example 9. *Imagine you have a box with n balls inside. $n-1$ of them are white, and just one is black. Your goal is to keep pulling out balls, one at a time, until you find the black ball. We want to compute the probability that you'll finally get the black ball on your k-th try.*

Let A_i be an event that the i-th ball is white. Then

$$P(\textit{Black on } k^{th} \textit{ try and White on previous } k-1 \textit{ tries}) = P(A_k^c \cap A_{k-1} \cap \cdots \cap A_1)$$

Using the Multiplication Rule, this breaks down to:

$$P(A_k^c|A_{k-1} \cap \cdots \cap A_1)P(A_{k-1}|A_{k-2} \cap \cdots \cap A_1) \ldots P(A_2|A_1)P(A_1)$$

- **Case a)** *Without Replacement (You don't put the balls back)*

  The probability of pulling a white ball on the i-th try, given that the previous $i-1$ balls were white:

  $P(A_i|A_{i-1} \cap \cdots \cap A_1) = \frac{n-i}{n-i+1}$.

  Because if you've already pulled out $i-1$ white balls, you have $n-(i-1)$ balls left, and $n-1-(i-1)$ of them are white.

  The probability of pulling the black ball on the k-th try, given that the previous $k-1$ balls were white: $P(A_k^c|A_{k-1} \cap \cdots \cap A_1) = \frac{1}{n-k+1}$.

  So, when you multiply all these probabilities together, something neat happens:

  $$\frac{1}{n-k+1} \cdot \frac{n-k+1}{n-k+2} \cdot \cdots \cdot \frac{n-2}{n-1} \cdot \frac{n-1}{n} = \frac{1}{n}$$

  Almost everything cancels out!

  This makes intuitive sense because, in the grand scheme of things, each ball has an equal chance of being the one you pick at any given position when you're not putting them back.

- **Case b)** *With Replacement (You always put the ball back)*

  This means the total number of balls, and the number of white and black balls, always stays the same!

  The probability of pulling a white ball on any given try (because you put it back):

  $P(A_i|A_{i-1} \cap \cdots \cap A_1) = \frac{n-1}{n}$.

  The probability of pulling the black ball on any given try: $P(A_k^c|A_{k-1} \cap \cdots \cap A_1) = \frac{1}{n}$.

  So, to get a white ball $k-1$ times in a row, and then the black ball on the k-th try, you multiply their individual probabilities:

  $$\frac{1}{n} \cdot \frac{n-1}{n} \cdot \cdots \cdot \frac{n-1}{n} \text{ (where } \frac{n-1}{n} \text{ appears } k-1 \text{ times)} = \frac{1}{n}\left(\frac{n-1}{n}\right)^{k-1}$$

  The probability decreases as k gets larger. Why? Because every time you pull a ball, you have a really good chance of pulling a white one (since there are so many of them). So, getting to the k-th try without finding the black ball becomes less likely as k increases.

The law of total probabilities allows us to compute the probability of an event, by conditioning on all possible instances of a 'different event', or, more formally, on every set in a partition of the sample space.

Definition 7. *We say that family of events $(B_k)_{k=1}^n$ is a **partition of** Ω if $\bigcup_{k=1}^n B_k = \Omega$ and B_k's are pairwise disjoint. We define a countable partition in the analogous way.*

Theorem 5 (Law of the total probability)**.** *Assume $(\Omega, \mathcal{F}, P)$ is a probability space, $A \in \mathcal{F}$ and $(B_k)_k$ is a partition (finite or countable) of Ω such that $P(B_k) > 0$ for every k. Then:*

$$P(A) = \sum_k P(A|B_k)P(B_k)$$

Proof. Events $(A \cap B_k)_k$ are pairwise disjoint and $A = A \cap \Omega = A \cap \bigcup_k B_k = \bigcup_k (A \cap B_k)$.

$$P(A) = P\left(\bigcup_k (A \cap B_k)\right) = \sum_k P(A \cap B_k) = \sum_k P(A|B_k)P(B_k)$$

Bayes' theorem allows us to compute the conditional probability of one event, given another in terms of the reverse conditional probabilities. It is particularly useful in Statistics, leading to a whole area called Bayesian Statistics: while in probability, we are interested in computing probabilities given a 'model' (i.e. sufficient information that determine the probabilities), in statistics, we are interested in choosing a model, given the events that we observe. Bayes' theorem allows us to connect the two.

Theorem 6 (Bayes' Theorem)**.** *With the same conditions as in the above theorem, if $P(A) > 0$ then for every k*

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{P(A)} = \frac{P(A|B_k)P(B_k)}{\sum_n P(A|B_n)P(B_n)}$$

Proof.

$$P(B_k|A) = \frac{P(A \cap B_k)}{P(A)} = \frac{P(A|B_k) \cdot P(B_k)}{\sum_n P(A|B_n)P(B_n)}$$

Example 10. *A disease has incidence of $\frac{1}{100}$ over the population. The available diagnostic test is such that:*

- *if you have the disease, the test is positive with probability $\frac{72}{100}$*
- *if you don't have the disease, the test is positive with probability $\frac{5}{1000}$*

A person gets a positive result. What is the probability they actually have the disease?

The two events of interest are

$D = \{$the person has the disease$\}$ and $P = \{$the person tests positive$\}$. We are interested in $P(D|P)$.

The information we are given is $P(D) = \frac{1}{100}$, $P(P|D) = \frac{72}{100}$ and $P(P|D^c) = \frac{5}{1000}$. By Bayes' Theorem:

$$P(D|P) = \frac{P(P|D)P(D)}{P(P)} = \frac{P(P|D)P(D)}{P(P|D)P(D) + P(P|D^c)P(D^c)}$$

We know $P(D^c) = 1 - P(D) = 1 - \frac{1}{100} = \frac{99}{100}$.

Substituting the values:

$$P(D|P) = \frac{\frac{72}{100} \cdot \frac{1}{100}}{\frac{72}{100} \cdot \frac{1}{100} + \frac{5}{1000} \cdot \frac{99}{100}} = \frac{720}{1215} \approx 0.59$$

2.2 Independence of events

One way to think of independence is that knowledge about occurrence of one of the events will neither increase nor decrease the chance that the other occurs. Assume A and B are events and $P(B) > 0$. Events A and B should be independent if the information that B occurred does not influence the probability that A occurs. Thus we should have $P(A|B) = P(A)$ which means $P(A \cap B) = P(A)P(B)$.

Definition 8. *Assume $(\Omega, \mathcal{F}, P)$ is a probability space. Events $A, B \in \mathcal{F}$ are **independent** if $P(A \cap B) = P(A)P(B)$.*

Remark 5. *The notions of "independent" and "disjoint" events are very different. In fact, these notions are normally incompatible. In fact, two disjoint events are independent if and only if the probability of one of them is 0.*

Additionally, you can prove as an exercise that if $P(A) = 0$, then A and B are independent for any $B \in \mathcal{F}$. The same holds if $P(A) = 1$.

Definition 9. *Events $A_1, A_2, \ldots, A_n$ are **mutually independent** (or independent) if for all $2 \leq k \leq n$ and for any sequence $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ we have:*

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdot \ldots P(A_{i_k})$$

Definition 10. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $A_1, A_2, \ldots, A_n$ be events. We say that the events $A_1, \ldots, A_n$ are **pairwise independent** if A_j and A_k are independent for every choice of j and k distinct.*

Remark 6. *Note that in case $n = 2$, pairwise independence is obviously the same as mutual independence. However, in case $n = 3$, pairwise independence means:*

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1)P(A_2) \\ P(A_1 \cap A_3) &= P(A_1)P(A_3) \\ P(A_2 \cap A_3) &= P(A_2)P(A_3) \end{aligned}$$

whereas mutual independence means:

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1)P(A_2) \\ P(A_1 \cap A_3) &= P(A_1)P(A_3) \\ P(A_2 \cap A_3) &= P(A_2)P(A_3) \\ P(A_1 \cap A_2 \cap A_3) &= P(A_1)P(A_2)P(A_3) \end{aligned}$$

This illustrates that mutual independence is stronger than pairwise independence.

It is hard to write down the list for larger values of n. For instance, if $n = 5$, pairwise independence involves $\binom{5}{2} = 10$ conditions to be checked, and mutual independence involves $2^5 - 5 - 1 = 26$ conditions to be checked.

Example 11. *Two dice are rolled. Let*

$$\begin{aligned} A_1 &= \{\text{the first die is even}\} \\ A_2 &= \{\text{the second die is odd}\} \\ A_3 &= \{\text{sum of the dice is 7}\} \end{aligned}$$

These events are pairwise independent, since

$$\begin{aligned} P(A_1 \cap A_2) &= \frac{1}{4} = P(A_1)P(A_2) \\ P(A_1 \cap A_3) &= \frac{1}{12} = P(A_1)P(A_3) \\ P(A_2 \cap A_3) &= \frac{1}{12} = P(A_2)P(A_3). \end{aligned}$$

That means, for each pair of events in this family, knowledge about occurrence of one of them will not affect the odds that any of the other two occurs. However, A_1, A_2, A_3 are not mutually independent, because

$$P(A_1 \cap A_2 \cap A_3) = \frac{1}{12} \neq \frac{1}{24} = P(A_1)P(A_2)P(A_3).$$

Proposition 2. *If two events A and B are independent, then the events A and B^c are also independent.*

Proof. Note that $P(A \cap B^c) = P(A) - P(A \cap B)$. Furthermore, since A and B are independent events, $P(A \cap B) = P(A)P(B)$. It now follows that

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A)P(B) \\ &= P(A)[1 - P(B)] \\ &= P(A)P(B^c). \end{aligned}$$

Therefore, the events A and B^c are independent.

The proof of the analogous result for the events A^c and B is similar, and the proof for the events A^c and B^c.

For any (possibly infinite) number of events we define independence by:

Definition 11. *Assume $\{A_i\}_{i \in I}$ is a family of events. We say that these events are **independent** if for any n and pairwise different $i_1, i_2, \ldots, i_n \in I$ the events $A_{i_1}, \ldots, A_{i_n}$ are independent.*

Definition 12. *Assume $(\Omega, \mathcal{F}, P)$ is a probability space and $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n$ are σ-algebras contained in $\mathcal{F}$. We say that σ-algebras $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are **independent** if for any $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2, \ldots, A_n \in \mathcal{F}_n$ we have:*

$$P(A_1 \cap \cdots \cap A_n) = P(A_1) \ldots P(A_n)$$

2.3 Sequences of Events

The following theorem, which will be useful to us later, is known as the continuity of probability over sets.

Theorem 7. *Assume $(\Omega, \mathcal{F}, P)$ is a probability space and $(A_n)_{n=1}^{\infty}$ is a sequence of events*

1. *If A_n is increasing, i.e. $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$, then $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n \to \infty} P(A_n)$.*
2. *If A_n is decreasing, i.e. $A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots$, then $P\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n \to \infty} P(A_n)$.*

Proof.

1. Let $B_1 = A_1$, $B_2 = A_2 \setminus A_1$, $B_3 = A_3 \setminus A_2$, ...
   The B_k's are mutually exclusive.
   $\bigcup_{k=1}^{\infty} B_k = \bigcup_{k=1}^{\infty} A_k$.
   $A_n = \bigcup_{k=1}^{n} B_k$
   This implies
   $P\left(\bigcup_{k=1}^{\infty} A_k\right) = P\left(\bigcup_{k=1}^{\infty} B_k\right) = \sum_{k=1}^{\infty} P(B_k) = \lim_{n \to \infty} \sum_{k=1}^{n} P(B_k) = \lim_{n \to \infty} P(A_n)$.
2. A_n - decreasing $\implies$ A_n^c is increasing.

$P\left(\bigcap_{k=1}^{\infty} A_k\right) = P\left(\left(\bigcup_{k=1}^{\infty} A_k^c\right)^c\right) = 1 - P\left(\bigcup_{k=1}^{\infty} A_k^c\right) = 1 - \lim_{n\to\infty} P(A_n^c) = 1 - \lim_{n\to\infty}(1 - P(A_n)) = \lim_{n\to\infty} P(A_n)$.

Definition 13. *Assume $(A_n)_{n=1}^{\infty}$ is a sequence of events, i.e. $A_k \in \mathcal{F}$ for every k. We say that: Events in $(A_n)_{n=1}^{\infty}$ occur **infinitely often** if A_n occurs for infinite number of indices $n \in \mathbb{N}$.*

$$\{A_n \text{ } i.o.\} = \{\omega \mid \omega \in A_n \text{ for infinitely many indices } n \in \mathbb{N}\}$$

*Events in $(A_n)_{n=1}^{\infty}$ occur **finitely often** if A_n occurs for at most finite number of indices $n \in \mathbb{N}$*

$$\{A_n \text{ } f.o.\} = \{\omega \mid \omega \in A_n \text{ for finitely many indices } n \in \mathbb{N}\}$$

Theorem 8.

$$\{A_n \text{ } i.o.\} = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n \quad \text{and} \quad \{A_n \text{ } f.o.\} = \{A_n \text{ } i.o.\}^c = \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n^c.$$

Proof. We first show that $\{A_n \text{ i.o.}\} \subset \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n$. Suppose $\omega \in A_n$ for infinite number of indices. Then for any $m \geq 1$ we have that $\omega \in \bigcup_{n=m}^{\infty} A_n$. Hence $\omega \in \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n$.

Suppose next that $\omega \notin \{A_n \text{ i.o.}\}$. Then $\omega \in A_n$ for only finitely many indices n. Let k be the largest of these indices. Then $\omega \notin \bigcup_{n=m}^{\infty} A_n$ for all $m > k$. Thus $\omega \notin \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n$.

Remark 7. *Note that from the definition of σ-algebra, $\{A_n \text{ } i.o.\} \in \mathcal{F}$ and $\{A_n \text{ } f.o.\} \in \mathcal{F}$.*

Theorem 9 (Borel-Cantelli Lemmas)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $(A_n)_{n=1}^{\infty}$ be a sequence of events*

a) If $\sum_{n=1}^{\infty} P(A_n) < \infty$ then $P(\{A_n \text{ } i.o.\}) = 0$.

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

Proof.

a) $\{A_n \text{ i.o.}\} = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n = \bigcap_{m=1}^{\infty} B_m$, where $B_1 \supset B_2 \supset B_3 \ldots$

$$P(\{A_n \text{ i.o.}\}) = P\left(\bigcap_{m=1}^{\infty} B_m\right) = \lim_{m\to\infty} P(B_m) = \lim_{m\to\infty} P\left(\bigcup_{n=m}^{\infty} A_n\right) \leq \lim_{m\to\infty} \sum_{n=m}^{\infty} P(A_n) = 0$$

where the last equality follows from the convergence of $\sum_{n=1}^{\infty} P(A_n) < \infty$.

b) We will show $P(\{A_n \text{ f.o.}\}) = 0$. $\{A_n \text{ f.o.}\} = \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n^c = \bigcup_{m=1}^{\infty} C_m$, where $C_1 \subset C_2 \subset C_3 \subset \ldots$.

$$P(\{A_n \text{ f.o.}\}) = P\left(\bigcup_{m=1}^{\infty} C_m\right) = \lim_{m\to\infty} P(C_m) = \lim_{m\to\infty} P\left(\bigcap_{n=m}^{\infty} A_n^c\right)$$

For any fixed m, A_n^c are independent events.

$$P\left(\bigcap_{n=m}^{\infty} A_n^c\right) = \lim_{k\to\infty} P\left(\bigcap_{n=m}^{m+k} A_n^c\right) = \lim_{k\to\infty} \prod_{n=m}^{m+k} P(A_n^c) = \lim_{k\to\infty} \prod_{n=m}^{m+k} (1 - P(A_n))$$

Using the inequality $1 - x \leq e^{-x}$ for $x \in \mathbb{R}$:

$$\lim_{k\to\infty} \prod_{n=m}^{m+k} (1 - P(A_n)) \leq \lim_{k\to\infty} \prod_{n=m}^{m+k} e^{-P(A_n)} = \lim_{k\to\infty} e^{-\sum_{n=m}^{m+k} P(A_n)} = e^{-\sum_{n=m}^{\infty} P(A_n)}$$

Since $\sum_{n=1}^{\infty} P(A_n) = \infty$, we have $\sum_{n=m}^{\infty} P(A_n) = \infty$ for any m. Therefore,

$$P(\{A_n \text{ f.o.}\}) = \lim_{m\to\infty} 0 = 0.$$

For part (a), we can think that if the probabilities of each event occurring are so small that, even when you sum them all (infinitely), the result is a finite number, then it is almost certain that each individual event will occur a finite number of times.

That is to say, the events are so rare and become rare so quickly that there is not enough probability for them to occur infinitely often. In contrast, for part (b), if the probabilities sum to infinity, and if the events are independent of each other (the outcome of one does not affect the next), and their probabilities, when you sum them, give an infinite value, then it is certain that they will occur infinitely often. Although each event may have a small probability, if the events are independent and their probabilities do not decrease quickly enough, the "chance" for an event to happen is so persistent that it guarantees you will see it happen again and again, forever. Independence is key here because it ensures that each "attempt" is a fresh opportunity, unaffected by what happened before.

2.4 π-λ Theorem

Assume we have a measurable space $(\Omega, \mathcal{F})$ with two probability measures P_1 and P_2. We want to verify if $P_1 = P_2$, that is

$$P_1(A) = P_2(A) \text{ for all } A \in \mathcal{F}.$$

Can we check for equality on a smaller number of elements than all of $\mathcal{F}$? Perhaps on the generators of $\mathcal{F}$? This is true for a specific type of generator. Next, we will define the necessary concepts and introduce the main theorem, whose proof is beyond the scope of this course.

Definition 14. *A collection $\mathcal{L}$ of subsets of a set Ω is called a* ***λ-system*** *if*

1. $\emptyset \in \mathcal{L}$.
2. *If* $A \in \mathcal{L}$ *then* $A^c \in \mathcal{L}$.
3. $\mathcal{L}$ *is closed under countable disjoint unions, i.e., if* $A_1, A_2, \cdots \in \mathcal{L}$ *and* $A_i \cap A_j = \emptyset, \forall i \neq j$, *then* $\bigcup_{n=1}^{\infty} A_n \in \mathcal{L}$.

The fundamental difference lies in property 3. A σ-algebra requires closure under countable infinite unions to hold for any sequence of sets, whether they are disjoint or not. In contrast, a λ-system only requires this closure to be true for unions of sets that are pairwise disjoint.

This distinction means that every σ-algebra is a λ-system, but the reverse is not true. The main utility of λ-systems is found in the monotone class theorem, a powerful tool in measure theory that allows you to prove a property holds for an entire σ-algebra by simply verifying it for a smaller set of generators.

Example 12. $(\Omega, \mathcal{F}, P_1), (\Omega, \mathcal{F}, P_2)$ *- two probability spaces. Then* $\mathcal{L} = \{A \in \mathcal{F} \mid P_1(A) = P_2(A)\}$ *is a λ-system contained in $\mathcal{F}$:*

1. $P_1(\emptyset) = P_2(\emptyset) = 0$, *so* $\emptyset \in \mathcal{L}$.
2. *If* $A \in \mathcal{L}$ *then* $P_1(A) = P_2(A)$. *Then* $P_1(A^c) = 1 - P_1(A) = 1 - P_2(A) = P_2(A^c)$, *so* $A^c \in \mathcal{L}$.

3. *If* $P_1(A_1) = P_2(A_1), \ldots, P_1(A_n) = P_2(A_n), \ldots$ *and* $A_i \cap A_j = \emptyset$ *then of course* $P_1(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P_1(A_n) = \sum_{n=1}^{\infty} P_2(A_n) = P_2(\bigcup_{n=1}^{\infty} A_n)$, *so* $\bigcup_{n=1}^{\infty} A_n \in \mathcal{L}$.

Definition 15. *A collection* $\mathcal{D}$ *of subsets of a set* Ω *is called a* ***π-system*** *if it is closed under finite intersections, that is if* $A, B \in \mathcal{D}$ *then* $A \cap B \in \mathcal{D}$.

Example 13. *The set of all open intervals,* $\mathcal{O} = \{(a, b) \mid a, b \in \mathbb{R}\}$ *is a* π*-system. Similarly* $\mathcal{D} = \{(-\infty, a] \mid a \in \mathbb{R}\}$ *is a* π*-system of subsets of* $\mathbb{R}$.

Theorem 10 (Dynkin's π-λ Theorem)**.** *Let* $\mathcal{D}$ *be a* π*-system of subsets of* Ω *and* $\mathcal{L}$ *a* λ*-system of subsets of* Ω. *Assume* $\mathcal{D} \subset \mathcal{L}$. *Then* $\sigma(\mathcal{D}) \subset \mathcal{L}$, *that is* $\mathcal{L}$ *contains the* σ*-algebra generated by* $\mathcal{D}$.

Remark 8. *In order to show that two probability measures on* $(\Omega, \mathcal{F})$ *satisfy* $P_1 = P_2$, *we need to only show that* $\mathcal{L} = \{A \in \mathcal{F} \mid P_1(A) = P_2(A)\} = \mathcal{F}$. *But we know that* $\mathcal{L}$ *is a* λ*-system. Hence it is enough to show that* $\mathcal{L}$ *contains a* π*-system that generates* $\mathcal{F}$. *By Dynkin's* $\pi - \lambda$ *theorem this would imply that* $\mathcal{F} \subset \mathcal{L}$.

The idea is that it can be very difficult to prove a property holds for every set in a σ-algebra, which can be an enormous collection of sets. However, if we can show that the property holds for a smaller, more manageable collection of sets (like a π-system), and also show that the collection of sets for which the property holds is a λ-system, then Dynkin's $\pi - \lambda$ theorem guarantees that the property must hold for the entire σ-algebra generated by that smaller collection.

This simplifies proofs enormously. Instead of proving something for all of $\mathcal{B}(\mathbb{R})$(the Borel σ-algebra), you can just prove it for all open intervals (a, b), and then use this powerful theorem to extend the result to the entire σ-algebra.

Example 14. $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_1)$ *and* $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_2)$. *Let* $\mathcal{D} = \{(-\infty, a] \mid a \in \mathbb{R}\}$. $\mathcal{D}$ *is a* π*-system and* $\sigma(\mathcal{D}) = \mathcal{B}(\mathbb{R})$. *Thus* $P_1(A) = P_2(A)$ *for all* $A \in \mathcal{F}$ *if* $P_1((-\infty, a]) = P_2((-\infty, a])$ *for all* $a \in \mathbb{R}$.

More details of this Section could be founded in the book of Alan Karr (1993), Probability, Springer.

2.5 Constructing a Product Probability Space

Assume that we have n experiments, the i-th experiment is described by $(\Omega_i, \mathcal{F}_i, P_i)$. We want to build a probability space $(\Omega, \mathcal{F}, P)$ that describes independent execution of these experiments.

The Sample Space (Ω)

The total set of outcomes, Ω, is the Cartesian product of the individual sample spaces:

$$\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$$

This is intuitive: an outcome in the new space is an n-tuple $(\omega_1, \omega_2, \ldots, \omega_n)$, where each ω_i is an outcome from experiment i.

The Event Space ($\mathcal{F}$)

The event space $\mathcal{F}$ needs to be a σ-algebra that contains all relevant information from each individual experiment.

- First, we "embed" the events of each individual experiment into the larger product space. The collection $\mathcal{F}_i'$ represents all events where we care about the outcome of experiment i, but not the others. An event $A_i \in \mathcal{F}_i$ corresponds to the "slice" $\Omega_1 \times \cdots \times A_i \times \cdots \times \Omega_n$ in the product space.

- The **product σ-algebra**, denoted $\mathcal{F} = \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n$, is defined as the smallest σ-algebra containing all of these embedded event collections. That is, $\mathcal{F} = \sigma(\mathcal{F}'_1, \mathcal{F}'_2, \ldots, \mathcal{F}'_n)$.
- This σ-algebra is generated by the set of all "rectangular" events: $\mathcal{G} = \{A_1 \times \cdots \times A_n \mid A_i \in \mathcal{F}_i\}$. These rectangular sets are the basic building blocks of our combined events.

The Probability Measure (P)

The definition of the probability measure P is the most crucial step, as it is where we impose the condition of **independence**.

- We want the embedded event collections $\mathcal{F}'_1, \ldots, \mathcal{F}'_n$ to be independent σ-algebras.
- The probability of an event from an individual experiment, like A_i, should remain unchanged in the new space: $P(\Omega_1 \times \cdots \times A_i \times \cdots \times \Omega_n) = P_i(A_i)$.
- To satisfy these conditions, we must define the probability of the generating rectangular sets as the product of the individual probabilities.

The following equation formalizes this and is a direct consequence of the independence requirement:

$$P(A_1 \times \cdots \times A_n) = P((A_1 \times \Omega_2 \times \cdots \times \Omega_n) \cap (\Omega_1 \times A_2 \times \Omega_3 \times \cdots \times \Omega_n) \cap \cdots \cap (\Omega_1 \times \cdots \times \Omega_{n-1} \times A_n))$$
$$= \prod_{i=1}^{n} P(\Omega_1 \times \cdots \times A_i \times \cdots \times \Omega_n) = \prod_{i=1}^{n} P_i(A_i).$$

The Product Measure Theorem

Fact 4. *Let $(\Omega_1, \mathcal{F}_1, P_1), \ldots, (\Omega_n, \mathcal{F}_n, P_n)$ be n probability spaces. Then there exists a unique probability measure P on the product space $(\Omega_1 \times \cdots \times \Omega_n, \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n)$ such that*

$$P(A_1 \times \cdots \times A_n) = P_1(A_1) \ldots P_n(A_n)$$

*for any events $A_i \in \mathcal{F}_i$. This measure is called the **product measure** and we denote it by $P = P_1 \otimes P_2 \otimes \cdots \otimes P_n$.*

Proof. The uniqueness of the measure follows from the fact that the collection of rectangular sets $\{A_1 \times A_2 \times \cdots \times A_n \mid A_i \in \mathcal{F}_i\}$ is a π-system that generates the product σ-algebra $\mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n$. The existence proof is more complicated and we skip it.

Remark 9. *In summary, if $(\Omega_i, \mathcal{F}_i, P_i)$ describes the i-th experiment, then the product space $(\Omega_1 \times \cdots \times \Omega_n, \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n, P_1 \otimes \cdots \otimes P_n)$ describes a series of n independent experiments. Analogous construction can be done for a countable product of probability spaces.*

2.6 Apendix 1: The Monty Hall Problem

Game Setup and Rules: You are on the TV game show *Let's Make a Deal*. The host, Monty Hall, invites you to play the following game. In front of you are three doors, all closed. You are told that two of them have a goat behind them, but behind one is a new car. The car is equally likely to be behind any of the three doors, meaning

$$P(\text{Car behind any door}) = 1/3.$$

The Host's Reveal and the Decision: You are asked to choose an initial door (let's call it Door 1). After you choose, Monty Hall opens one of the two remaining doors. **He will always open a door to reveal a goat.** (He never reveals the car). You are then permitted to switch your initial choice to the other closed door. Should you switch your choice?

We analyze the probability of winning based on the initial choice:

- Let C be the event that the player initially chooses the door with the car.
- Let C^c denote the event that the player initially chooses a door with a goat.

The initial probabilities are:

$$P(C) = 1/3 \quad \text{and} \quad P(C^c) = 2/3$$

The decision to switch or not switch is evaluated using the Law of Total Probability, conditioning on the initial choice (C or C^c).

Option 1: Switch Doors

If you switch, you win only if your **initial choice was incorrect** (i.e., you chose a goat).

$$P(\text{Win}|\text{Switch}) = P(\text{Win}|C)P(C) + P(\text{Win}|C^c)P(C^c)$$

- $P(\text{Win}|C) = 0$: If you initially chose the car, switching guarantees a loss (you switch to the remaining goat).
- $P(\text{Win}|C^c) = 1$: If you initially chose a goat, Monty opens the *other* goat door, meaning the remaining closed door **must** be the car. Switching guarantees a win.

$$P(\text{Win}|\text{Switch}) = (0 \cdot 1/3) + (1 \cdot 2/3) = \frac{2}{3}$$

Option 2: Don't Switch Doors

If you do not switch, you win only if your **initial choice was correct** (i.e., you chose the car).

$$P(\text{Win}|\text{Don't Switch}) = P(\text{Win}|C)P(C) + P(\text{Win}|C^c)P(C^c)$$

- $P(\text{Win}|C) = 1$: If you initially chose the car, staying guarantees a win.
- $P(\text{Win}|C^c) = 0$: If you initially chose a goat, staying guarantees a loss.

$$P(\text{Win}|\text{Don't Switch}) = (1 \cdot 1/3) + (0 \cdot 2/3) = \frac{1}{3}$$

Monty Hall reveals new information. He can only open a door to show you a goat. The initial probability of your chosen door having the car (**1/3**) does not change. Instead, the entire **2/3** probability of the car being behind one of the two unchosen doors is consolidated onto the single remaining closed door. Therefore, it is always strategically advantageous to switch your choice, as your probability of winning increases from 1/3 to 2/3.

2.7 Apendix 2: Polya's Urn Scheme

Consider the following experiment called the **Polya Scheme**. From an urn containing B white balls, $B \geq 1$, and R red balls, $R \geq 1$, we successively and randomly draw n balls, $n \geq 2$. Each ball drawn is returned to the urn along with c additional balls of the same color, where $c \geq 1$.

Let R_n be the event that a red ball is drawn on the n-th draw.

Probability of Drawing a Red Ball on the First Draw

The probability of obtaining a red ball on the first draw is simply the initial proportion:

$$P(R_1) = \frac{R}{B+R}$$

Probability of Drawing a Red Ball on the Second Draw

We use the Law of Total Probability, conditioning on the outcome of the first draw (R_1 or B_1):

$$P(R_2) = P(R_2 \mid R_1)P(R_1) + P(R_2 \mid B_1)P(B_1)$$

Substituting the probabilities based on the contents of the urn after the first draw:

$$P(R_2) = \left(\frac{R+c}{R+B+c}\right)\left(\frac{R}{R+B}\right) + \left(\frac{R}{R+B+c}\right)\left(\frac{B}{R+B}\right)$$

$$P(R_2) = \frac{R(R+c)+RB}{(R+B+c)(R+B)} = \frac{R(R+c+B)}{(R+B+c)(R+B)} = \frac{R}{R+B}$$

Probability of Drawing a Red Ball on the n-th Draw

This result is solved using mathematical induction. The hypothesis is that the probability of drawing a red ball on any draw $k < n$ is equal to the initial proportion.

Inductive Hypothesis: Assume that for any urn initially containing M red balls and K white balls, the probability of drawing a red ball on the $(n-1)$-th draw is $\frac{M}{M+K}$.

We apply the Law of Total Probability, conditioning on the outcome of the first draw:

$$P(R_n) = P(R_n \mid R_1)P(R_1) + P(R_n \mid B_1)P(B_1)$$

We apply the inductive hypothesis:

- $P(R_n \mid R_1)$: If R_1 occurred, the urn now contains $R+c$ red balls and B white balls. The probability of drawing a red ball on the next draw is $\frac{R+c}{R+c+B}$.
- $P(R_n \mid B_1)$: If B_1 occurred, the urn now contains R red balls and $B+c$ white balls. The probability of drawing a red ball on the next draw is $\frac{R}{R+B+c}$.

Substituting these into the formula (and noting the calculation is identical to $P(R_2)$):

$$P(R_n) = \left(\frac{R+c}{R+B+c}\right)\left(\frac{R}{R+B}\right) + \left(\frac{R}{R+B+c}\right)\left(\frac{B}{R+B}\right) = \frac{R}{R+B}$$

This remarkable result shows that the probability of drawing a red ball remains $\frac{R}{R+B}$ for all draws n.

We provide the R code used to simulate the evolution of the proportion of red balls in Polya's Urn, demonstrating a non-stationary and exchangeable process.

```
simulate_polya_urn <- function(R, B, c, n) {
 # Vectors to store the number of balls and the proportion of red balls
 proportions <- rep(0, n)
 reds <- rep(0, n)
 whites <- rep(0, n)

 # Initialize the urn
 reds[1] <- R
 whites[1] <- B
 proportions[1] <- reds[1] / (reds[1] + whites[1])

 for (i in 2:n) {
  # Draw a ball at random (1 = red, 0 = white)
  # Probability is based on the proportion in the previous step (i-1)
```

```
  is_red <- rbinom(1, 1, prob = proportions[i-1])

  # Update the urn based on the drawn color
  if (is_red == 1) {
   reds[i] <- reds[i-1] + c # c red balls are added
   whites[i] <- whites[i-1]
  } else { # is_red == 0
   whites[i] <- whites[i-1] + c # c white balls are added
   reds[i] <- reds[i-1]
  }

  # Store the proportion of red balls in the current iteration
  proportions[i] <- reds[i] / (reds[i] + whites[i])
 }

 return(cbind(reds, whites, proportions))
}

set.seed(999)
result <- simulate_polya_urn(R = 5, B = 10, c = 3, n = 1000)

# The result can be plotted to visualize the convergence to a random variable
Initial_Proportion <- 5 / (5 + 10)
```

Chapter 3

Random Variables

3.1 Introduction

We are not always concerned with the full details of a random experiment, but instead with a specific numerical outcome. For example, a gambler might only care about their final winnings or losses, not the sequence of coin flips that led to the result. We can think of these numerical outcomes as functions which map the experiment's results to a real number.

A fair coin is tossed twice. The sample space is $\Omega = \{HH, HT, TH, TT\}$. For an outcome $\omega \in \Omega$, let $X(\omega)$ be the number of heads. The values of this function are:

$$X(HH) = 2, \quad X(HT) = X(TH) = 1, \quad X(TT) = 0.$$

Now, imagine a gambler betting all their money on this experiment. They gamble cumulatively, so their fortune is doubled each time a head appears and becomes zero upon the appearance of a tail. Their subsequent fortune, W, is a function defined as:

$$W(HH) = 4, \quad W(HT) = W(TH) = W(TT) = 0.$$

The function X or W are what we call random variables. To formally define the concept of a random variable, we first need the following definition.

Definition 16. *Assume $(\Omega, \mathcal{F}, P)$ is a probability space. A function $X : \Omega \to \mathbb{R}^d$ is called* ***measurable*** *if*

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\} \in \mathcal{F}, \quad \text{for any } B \in \mathcal{B}(\mathbb{R}^d).$$

That is, for every set $B \in \mathcal{B}$ (a Borel set), the preimage $X^{-1}(B)$ must be an element of the σ-algebra $\mathcal{F}$.

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

If you can't trace a numerical outcome back to a valid event in the original experiment, you can't assign a probability to it. The measurable condition guarantees that you always can. It's the mathematical requirement that lets you connect numerical results to the underlying probabilities of the experiment itself.

Definition 17. *Any measurable function $X : \Omega \to \mathbb{R}^d$ is called a* ***d-dimensional random variable****. For $d = 1$ we will call a 1-dimensional random variable a* ***random variable****.*

The core idea is that for a function to be considered a random variable, it must be "measurable." This means that for any well-behaved set of outcomes you can define (a Borel set), the set of all original experiment results that lead to those outcomes must belong to the probability space's sigma-algebra.

Definition 18. *A function $f : \mathbb{R}^d \to \mathbb{R}^k$ is called a* ***Borel function*** *if $f^{-1}(A) \in \mathcal{B}(\mathbb{R}^d)$ for any $A \in \mathcal{B}(\mathbb{R}^k)$.*

A function is Borel-measurable if the preimage of every Borel set is also a Borel set. Checking this for every set in the Borel sigma-algebra would be impossible, as this collection of sets is vast.

The following lemma provides a much more efficient approach. It states that you don't need to check all Borel sets; you only need to check the sets that generate the sigma-algebra. This is a powerful shortcut because we can often choose a small, simple set of generators, like the collection of all open sets or all semi-infinite intervals.

Lemma 1. *Let $\mathcal{S} \subset \mathcal{B}(\mathbb{R}^k)$ be a generating set of $\mathcal{B}(\mathbb{R}^k)$. A function $f : \mathbb{R}^d \to \mathbb{R}^k$ is a Borel function if*

$$\forall A \in \mathcal{S}, \quad f^{-1}(A) \in \mathcal{B}(\mathbb{R}^d)$$

Proof. Let $\mathcal{G} = \{A \subset \mathbb{R}^k \mid f^{-1}(A) \in \mathcal{B}(\mathbb{R}^d)\}$. One easily shows that $\mathcal{G}$ is a σ-algebra. This follows from the fact that

1. $\emptyset \in \mathcal{G}$ because $f^{-1}(\emptyset) = \emptyset \in \mathcal{B}(\mathbb{R}^d)$.
2. If $A \in \mathcal{G}$ then $f^{-1}(A) \in \mathcal{B}(\mathbb{R}^d)$. Hence $f^{-1}(A^c) = (f^{-1}(A))^c$ also belongs to $\mathcal{B}(\mathbb{R}^d)$. Thus $A^c \in \mathcal{G}$.
3. If $(A_k)_k$ are in $\mathcal{G}$ then for every k we have $f^{-1}(A_k) \in \mathcal{B}(\mathbb{R}^d)$. But $f^{-1}(\bigcup_k A_k) = \bigcup_k f^{-1}(A_k) \in \mathcal{B}(\mathbb{R}^d)$. Hence $\bigcup_k A_k \in \mathcal{G}$.

By our assumption $\mathcal{G}$ contains $\mathcal{S}$. Thus it contains $\sigma(\mathcal{S}) = \mathcal{B}(\mathbb{R}^k)$.

This means that to prove a function is Borel-measurable, it is sufficient to check that the preimages of the generating sets are Borel sets.

Corollary 1. *All continuous functions are Borel functions.*

Since continuous functions are defined by the property that the preimage of every open set is open, and the collection of all open sets is a generating set for the Borel sigma-algebra, it follows directly from the lemma that all continuous functions are Borel-measurable.

Fact 5. *If $(\Omega, \mathcal{F}, P)$ is a probability space, X is a d-dimensional random variable, and $f : \mathbb{R}^d \to \mathbb{R}^k$ is a Borel function, then $f(X) : \Omega \to \mathbb{R}^k$ is a k-dimensional random variable.*

3.2 The distribution of a random variable

When working with a random variable X, we're often only interested in the values it takes on the real number line, $\mathbb{R}$, rather than the original, possibly more complex, probability space $(\Omega, \mathcal{F}, P)$. We can simplify our approach by creating a new probability measure, P_X, that lives directly on $\mathbb{R}$.

This new measure is called the **distribution** of X. While P_X is derived from the original measure P and the function X, it's important to note that this process isn't reversible. In other words, you can't reconstruct the original probability space or the random variable X from P_X alone. The space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ is often more manageable than the original one, as it lets us focus on the properties of the random variable without the complexity of the underlying space Ω.

Definition 19. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and X a d-dimentional random variable. The distribution of the random variable X is the probability measure on $\mathbb{R}^d$ denoted by P_X and given by*

$$P_X(B) = P(X^{-1}(B)) = P(X \in B) = P(\{\omega \in \Omega \mid X(\omega) \in B\})$$

for any $B \in \mathcal{B}(\mathbb{R}^d)$.

In the previous example, the distribution $P_X(B)$ is determined by its values

$$P_X(\{0\}) = P_X(\{2\}) = \frac{1}{2}, \quad P_X(\{1\}) = \frac{1}{4},$$

and $P_X(\mathbb{R} \setminus \{0, 1, 2\}) = 0$.

Proposition 3. *$(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P_X)$ is a new probability space.*

Proof. Recall definition of probability space. We check the three conditions:

1. $P_X(B) = P(X \in B) \in [0, 1]$ since P is a probability measure.
2. $P_X(\mathbb{R}^d) = P(\{\omega \in \Omega : X(\omega) \in \mathbb{R}^d\}) = P(\Omega) = 1$.
3. If $B_1, B_2, B_3, \cdots \in \mathcal{B}(\mathbb{R}^d)$ are disjoint, then

$$\begin{aligned} P_X\left(\bigcup_{n=1}^{\infty} B_n\right) &= P\left(\left\{\omega : X(\omega) \in \bigcup_{n=1}^{\infty} B_n\right\}\right) \\ &= P\left(\bigcup_{n=1}^{\infty} \{\omega : X(\omega) \in B_n\}\right) \\ &= \sum_{n=1}^{\infty} P(\{\omega : X(\omega) \in B_n\}) \\ &= \sum_{n=1}^{\infty} P_X(B_n). \end{aligned}$$

In the above steps, we used: the definition of P_X; that the pre-image of a union is the union of the pre-images; that the pre-image of disjoint sets are disjoint, combined with the countable additivity of P; and the definition of P_X again. This proves the proposition.

Definition 20. *Assume $(\Omega, \mathcal{F}, P)$ is a probability space and $X = (X_1, X_2, \ldots, X_d)$ is a d-dimensional random variable. The **distribution function** of X is $F_X : \mathbb{R}^d \to [0, 1]$ given by*

$$F_X(x_1, x_2, \ldots, x_d) = P((X_1 \le x_1) \cap (X_2 \le x_2) \cap \cdots \cap (X_d \le x_d)) = P_X((-\infty, x_1] \times (-\infty, x_2] \times \cdots \times (-\infty, x_d])$$

Example 15. *A fair coin is tossed twice. The sample space is $\Omega = \{HH, HT, TH, TT\}$, and each outcome has a probability of $1/4$. For an outcome $\omega \in \Omega$, let $X(\omega)$ be the number of heads. The values of this function are:*

$$X(HH) = 2, \quad X(HT) = 1, \quad X(TH) = 1, \quad X(TT) = 0.$$

The cumulative distribution function (CDF) for this random variable is:

$$F_X(t) = P(X \le t) = \begin{cases} 0 & t < 0 \\ 1/4 & 0 \le t < 1 \\ 3/4 & 1 \le t < 2 \\ 1 & t \ge 2 \end{cases}$$

Example 16. *We choose a point at random from a disk of radius R. The probability space is $(\Omega, \mathcal{F}, P) = (D(0, R), \mathcal{B}(D(0, R)), \frac{|\cdot|}{\pi R^2})$, where $|\cdot|$ denotes the area. Let $X(\omega) = \rho(\omega, 0)$ be the distance of the chosen point from the center of the disk. Then its cumulative distribution function (CDF) is given by*

$$F_X(t) = P(X \le t) = \begin{cases} 0 & t < 0 \\ \frac{t^2}{R^2} & 0 \le t < R \\ 1 & t \ge R \end{cases}$$

Next, we will show some properties of the distribution function.

Theorem 11. *Assume $X : \Omega \to \mathbb{R}$ is a random variable. Then*

1. *F_X is non-decreasing.*
2. *$\lim_{t\to\infty} F_X(t) = 1$, $\lim_{t\to-\infty} F_X(t) = 0$.*
3. *F_X is right continuous.*
4. *for any $t \in \mathbb{R}$ the limit*
$$F_X(t^-) := \lim_{s\uparrow t} F_X(s)$$
exists and $F_X(t^-) = P(X < t) = P_X((-\infty, t))$.
5. *F_X is not continuous at a point t_0 if $P(X = t_0) > 0$. More precisely*
$$P(X = t_0) = F_X(t_0) - F_X(t_0^-)$$
6. *For any $a < b$ we have*
$$\begin{aligned} P(a \le X \le b) &= F_X(b) - F_X(a^-) \\ P(a < X \le b) &= F_X(b) - F_X(a) \\ P(a \le X < b) &= F_X(b^-) - F_X(a^-) \\ P(a < X < b) &= F_X(b^-) - F_X(a) \end{aligned}$$

Proof.

1. If $t_1 \le t_2$ then $(-\infty, t_1] \subset (-\infty, t_2]$ and $F_X(t_1) = P_X((-\infty, t_1]) \le P_X((-\infty, t_2]) = F_X(t_2)$.
2. Let $(t_n)_{n=1}^\infty$ be a sequence increasing to $+\infty$, then $\mathbb{R} = \bigcup_n (-\infty, t_n] = \bigcup_n A_n$. Moreover, $A_1 \subset A_2 \subset A_3 \subset \cdots \subset A_n$, that is A_n is an increasing sequence of events. Hence
$$1 = P_X(\mathbb{R}) = P_X\left(\bigcup_n A_n\right) = \lim_{n\to\infty} P_X((-\infty, t_n]) = \lim_{n\to\infty} F_X(t_n).$$
Let now $(t_n)_{n=1}^\infty$ be a sequence decreasing to $-\infty$. Then $\bigcap_{n=1}^\infty (-\infty, t_n]$ is a decreasing sequence of events and
$$\lim_{n\to\infty} F_X(t_n) = \lim_{n\to\infty} P_X((-\infty, t_n]) = P_X\left(\bigcap_{n=1}^\infty (-\infty, t_n]\right) = P_X(\emptyset) = 0.$$
3. Let's take $t \in \mathbb{R}$ and $(t_n)_{n=1}^\infty$ a decreasing sequence s.t. $\lim_{n\to\infty} t_n = t$, $t_n \ne t$ for any n.
$$\begin{aligned} \lim_{n\to\infty} F_X(t_n) - F_X(t) &= \lim_{n\to\infty} (P_X((-\infty, t_n]) - P_X((-\infty, t])) \\ &= \lim_{n\to\infty} P_X((-\infty, t_n] \setminus (-\infty, t]) = \lim_{n\to\infty} P_X((t, t_n]). \end{aligned}$$
Since $(t, t_n]$ forms a decreasing sequence of sets whose intersection is $\emptyset$,
$$= P_X\left(\bigcap_{n=1}^\infty (t, t_n]\right) = P_X(\emptyset) = 0.$$

4. Let's take $t \in \mathbb{R}$ and $(t_n)_{n=1}^{\infty}$ an increasing sequence st. $\lim_{n\to\infty} t_n = t$, $t_n \neq t$ for any n.

$$F_X(t^-) = \lim_{n\to\infty} F_X(t_n) = \lim_{n\to\infty} P_X((-\infty, t_n]) = P_X\left(\bigcup_{n=1}^{\infty}(-\infty, t_n]\right) = P_X((-\infty, t)) = P(X < t).$$

5. $F_X(t_0) - F_X(t_0^-) = P_X((-\infty, t_0]) - P_X((-\infty, t_0)) = P_X((-\infty, t_0] \setminus (-\infty, t_0)) = P_X(\{t_0\}) = P(X = t_0)$, for any $t_0 \in \mathbb{R}$.

6.

$$\begin{aligned}
P(a \le X \le b) &= P_X([a, b]) = P_X((-\infty, b] \setminus (-\infty, a)) \\
&= P_X((-\infty, b]) - P_X((-\infty, a)) = F_X(b) - F_X(a^-) \\
P(a < X \le b) &= P_X((a, b]) = P_X((-\infty, b]) - P_X((-\infty, a]) = F_X(b) - F_X(a) \\
P(a \le X < b) &= P_X([a, b)) = P_X((-\infty, b)) - P_X((-\infty, a)) = F_X(b^-) - F_X(a^-) \\
P(a < X < b) &= P_X((a, b)) = P_X((-\infty, b)) - P_X((-\infty, a]) = F_X(b^-) - F_X(a)
\end{aligned}$$

The first three properties characterize the distribution functions.

Remark 10. *If F satisfies 1), 2), 3) then F is the distribution function of some random variable X.*

Theorem 12. *Let X, Y be two random variables. Then*

$$\forall t \in \mathbb{R}, F_X(t) = F_Y(t) \iff \forall B \in \mathcal{B}(\mathbb{R}^d), P_X(B) = P_Y(B).$$

Proof. Let $\mathcal{L} = \{A \in \mathcal{B}(\mathbb{R}^d) \mid P_X(A) = P_Y(A)\}$. $\mathcal{L}$ is a λ-system, $\mathcal{L} \subset \mathcal{B}(\mathbb{R}^d)$. Since $F_X = F_Y$ we have $P_X((-\infty, a_1] \times \cdots \times (-\infty, a_d]) = P_Y((-\infty, a_1] \times \cdots \times (-\infty, a_d])$ for all $(a_1, \ldots, a_d) \in \mathbb{R}^d$. But $\mathcal{D} = \{(-\infty, a_1] \times \cdots \times (-\infty, a_d] \mid (a_1, \ldots, a_d) \in \mathbb{R}^d\}$ is a π-system and $\sigma(\mathcal{D}) = \mathcal{B}(\mathbb{R}^d)$. Thus the result follows from the Dynkin π-λ theorem (because $\mathcal{D} \subset \mathcal{L}$, $\sigma(\mathcal{D}) \subset \mathcal{L}$, $\mathcal{B}(\mathbb{R}^d) \subset \mathcal{L}$ and $\mathcal{L} \subset \mathcal{B}(\mathbb{R}^d) \implies \mathcal{L} = \mathcal{B}(\mathbb{R}^d)$).

Definition 21. *Let $X = (X_1, X_2, \ldots, X_d)$ be a d-dimensional random vector and let $1 \le i_1 < i_2 < \cdots < i_k \le d$ with $k < d$. The random vector $(X_{i_1}, X_{i_2}, \ldots, X_{i_k})$ is a random vector and its distribution is called a **marginal distribution** of X.*

Remark 11. *Marginal distributions do not uniquely determine the joint distribution of $(X_1, X_2, \ldots, X_d)$.*

Example 17. *Consider two different random vectors, $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$, each defined on the sample space $\Omega = \{0, 1\} \times \{0, 1\}$.*

Case 1: Random Vector X Let the joint distribution of X be defined as:

$$\begin{aligned}
P(X_1 = 0, X_2 = 0) &= 1/4 \\
P(X_1 = 0, X_2 = 1) &= 1/4 \\
P(X_1 = 1, X_2 = 0) &= 1/4 \\
P(X_1 = 1, X_2 = 1) &= 1/4
\end{aligned}$$

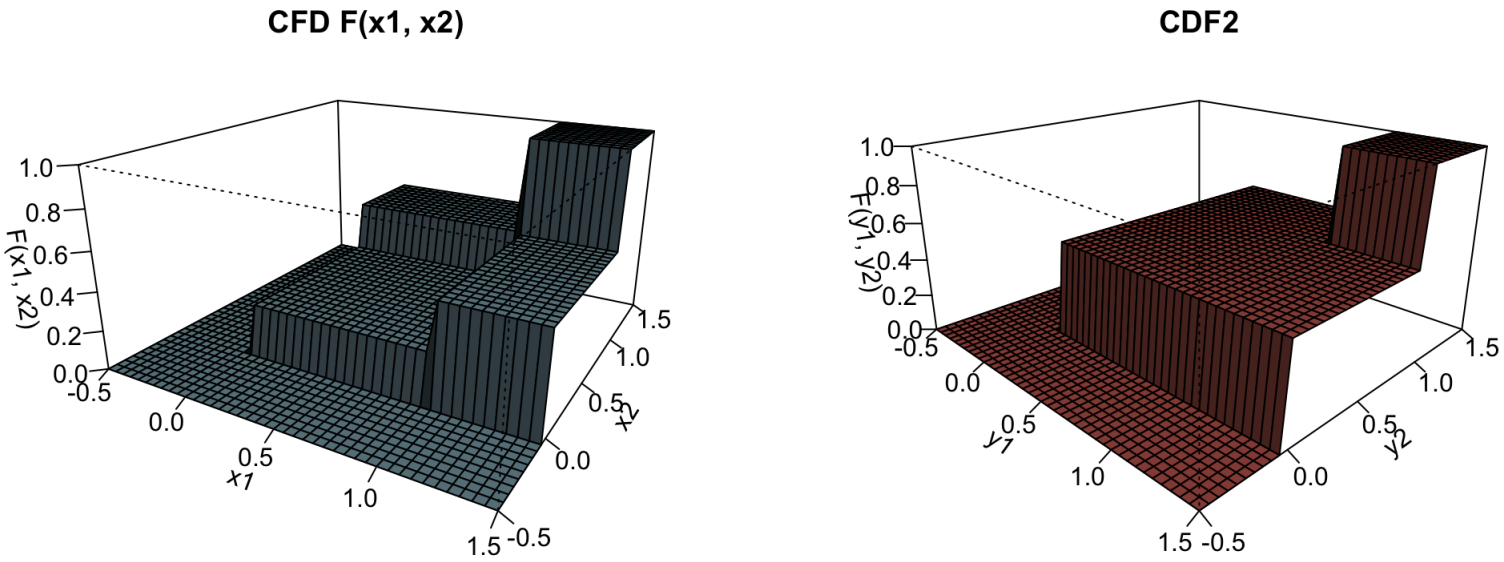
The marginal distributions are thus $P(X_1 = 0) = P(X_1 = 1) = 1/2$ and $P(X_2 = 0) = P(X_2 = 1) = 1/2$.

Case 2: Random Vector Y Now, let's define a second random vector Y with a different joint distribution:

$$P(Y_1 = 0, Y_2 = 0) = 1/2$$
$$P(Y_1 = 1, Y_2 = 1) = 1/2$$
$$P(Y_1 = 0, Y_2 = 1) = 0$$
$$P(Y_1 = 1, Y_2 = 0) = 0$$

The marginal distributions for Y_1 and Y_2 are also $P(Y_1 = 0) = P(Y_1 = 1) = 1/2$ and $P(Y_2 = 0) = P(Y_2 = 1) = 1/2$.

The marginal distributions of X and Y are identical, but their joint distributions are distinct. This shows that marginal distributions do not uniquely determine the joint distribution.



a) case 1 b) case 2

3.3 $\sigma-$ algebrea generated by X

The sigma-algebra generated by a random variable represents all the information or knowledge contained in the random variable X.

A random variable X assigns a number to each outcome in the sample space. The sigma-algebra $\sigma(X)$ collects all the possible events whose occurrence or non-occurrence you can determine just by knowing the value of X.

Fact 6. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $X : \Omega \to \mathbb{R}^d$ a d-dim r.v. The set*

$$\sigma(X) = \{X^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R}^d)\} \subset \mathcal{F}$$

*is a σ-algebra that we will call **σ-algebra generated by** X.*

Proof.

1. $\emptyset \in \sigma(X)$ as $X^{-1}(\emptyset) = \emptyset \in \mathcal{B}(\mathbb{R}^d)$.
2. Assume $A \in \sigma(X)$ then there is $B \in \mathcal{B}(\mathbb{R}^d)$ such that $A = X^{-1}(B)$. But $B^c \in \mathcal{B}(\mathbb{R}^d)$ and $X^{-1}(B^c) = (X^{-1}(B))^c = A^c$. Thus $A^c \in \sigma(X)$.

3. $A_1, A_2, \ldots$ sequence of events from $\sigma(X)$. Then there is $B_1, B_2, \ldots$ in $\mathcal{B}(\mathbb{R}^d)$ s.t. $A_1 = X^{-1}(B_1), \ldots, A_n = X^{-1}(B_n), \ldots$

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} X^{-1}(B_n) = X^{-1}\left(\bigcup_{n=1}^{\infty} B_n\right) \in \sigma(X).$$

For example, imagine X is a random variable representing the number of heads in two coin flips. The possible values are 0, 1, or 2.

The event "there was at least one head" is in $\sigma(X)$ because you can determine if it happened by observing the value of X (i.e., $X \geq$).

The event "the first flip was heads" is not in $\sigma(X)$ because knowing that $X = 1$ (one head) doesn't tell you whether it was the first or second flip.

The following definition introduces the notion of independence for random variables.

Definition 22. *Assume $(\Omega, \mathcal{F}, P)$ is a probability space and $\{X_i\}_{i\in I}$ is a collection of random variables $X_i : \Omega \to \mathbb{R}^{d_i}$. We say that random variables $\{X_i\}_{i\in I}$ are **independent** if the σ-algebras $\{\sigma(X_i)\}_{i\in I}$ are independent.*

This definition is equivalent to showing that for any finite subcollection of Borel sets, the joint probability is equal to the product of the individual probabilities. This characterization then gives a practical and intuitive way to check for independence, which is often used in applications.

More precisely, we have the following remark

Remark 12. *$\{X_i : \Omega \to \mathbb{R}^{d_i}\}_{i\in I}$ are independent if $\forall n$ and pairwise different $i_1, i_2, \ldots, i_n \in I$ and $B_1 \in \mathcal{B}(\mathbb{R}^{d_{i_1}}), \ldots, B_n \in \mathcal{B}(\mathbb{R}^{d_{i_n}})$ we have*

$$P(X_{i_1}^{-1}(B_1) \cap X_{i_2}^{-1}(B_2) \cap \cdots \cap X_{i_n}^{-1}(B_n)) = P(X_{i_1} \in B_1 \cap \cdots \cap X_{i_n} \in B_n) = P(X_{i_1} \in B_1) \ldots P(X_{i_n} \in B_n)$$

Example 18. *Let $A_1, \ldots, A_n \in \mathcal{F}$. Consider the indicator function*

$$\mathbf{1}_{A_i}(\omega) = \begin{cases} 1 & \omega \in A_i \\ 0 & \omega \notin A_i \end{cases}.$$

The events $A_1, A_2, \ldots, A_n$ are independent if and only if the random variables $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_n}$ are independent.

This follows from the fact that the σ-algebra generated by the indicator function $\mathbf{1}_{A_i}$ is identical to the σ-algebra generated by the event A_i. To see this, consider the preimages of a Borel set $B \subset \mathbb{R}$ under $\mathbf{1}_{A_i}$:

$$\mathbf{1}_{A_i}^{-1}(B) = \begin{cases} A_i & \text{if } 0 \notin B \text{ and } 1 \in B \\ A_i^c & \text{if } 0 \in B \text{ and } 1 \notin B \\ \emptyset & \text{if } 0 \notin B \text{ and } 1 \notin B \\ \Omega & \text{if } 0 \in B \text{ and } 1 \in B \end{cases}$$

Thus, the σ-algebra generated by $\mathbf{1}_{A_i}$ is $\sigma(\mathbf{1}_{A_i}) = \{\emptyset, \Omega, A_i, A_i^c\}$.

Since the σ-algebra generated by the event A_i is $\sigma(A_i) = \{\emptyset, \Omega, A_i, A_i^c\}$, we have $\sigma(\mathbf{1}_{A_i}) = \sigma(A_i)$.

Therefore, the condition that the random variables $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_n}$ are independent (i.e., their generated σ-algebras are independent) is exactly the same condition as the events $A_1, \ldots, A_n$ being independent.

Theorem 13. *Assume $X_1, X_2, \ldots, X_n$ are random variables. The following conditions are equivalent:*

1. *$X_1, \ldots, X_n$ are independent.*
2. *For any $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$ the events $\{X_1 \in B_1\}, \ldots, \{X_n \in B_n\}$ are independent.*
3. *$P_{(X_1,\ldots,X_n)} = P_{X_1} \otimes \cdots \otimes P_{X_n}$.*
4. *$\forall (x_1, \ldots, x_n) \in \mathbb{R}^n$ we have $F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = F_{X_1}(x_1) \ldots F_{X_n}(x_n)$.*

Proof. We will prove the equivalences in the order $1 \implies 2 \iff 4 \iff 3 \implies 2$. Since $2 \implies 1$ is also true by definition, the cycle of implications will complete the proof.

1 $\implies$ 2: This follows directly from the definition of independence of random variables. If the σ-algebras $\{\sigma(X_i)\}_{i=1}^n$ are independent, then any collection of events $B_i \in \sigma(X_i)$ for $i = 1, \ldots, n$ are independent. The events $\{X_i \in B_i\}$ for $B_i \in \mathcal{B}(\mathbb{R})$ are precisely of this form.

2 $\implies$ 4: Let us assume condition (2) holds. For any $(x_1, \ldots, x_n) \in \mathbb{R}^n$, we can choose the Borel sets $B_k = (-\infty, x_k]$. By condition (2), the events $\{X_1 \in (-\infty, x_1]\}, \ldots, \{X_n \in (-\infty, x_n]\}$ are independent. Therefore,

$$F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = P((X_1 \leq x_1) \cap (X_2 \leq x_2) \cap \cdots \cap (X_n \leq x_n)) = P(\{X_1 \in B_1\} \cap \cdots \cap \{X_n \in B_n\})$$

$$= P(X_1 \in B_1) \ldots P(X_n \in B_n) = P_{X_1}((-\infty, x_1]) \ldots P_{X_n}((-\infty, x_n]) = F_{X_1}(x_1) \ldots F_{X_n}(x_n).$$

Thus, condition (4) holds.

4 $\implies$ 3: By Dynkin $\pi - \lambda$ theorem, the product measure $P_{X_1} \otimes \cdots \otimes P_{X_n}$ and the measure $P_{(X_1,\ldots,X_n)}$ are equal iff they agree on a gererating set of $\mathcal{B}(\mathbb{R}^n)$ that is also a $\pi - system$. The set $D = \{(-\infty, x_1] \times \cdots \times (-\infty, x_n] : x_i \in R\}$ has this property and

$$F_{P'}(x_1, \ldots, x_n) = (P_{X_1} \otimes \cdots \otimes P_{X_n})((-\infty, x_1] \times \cdots \times (-\infty, x_n])$$

$$= P_{X_1}((-\infty, x_1]) \cdot \cdots \cdot P_{X_n}((-\infty, x_n]) = F_{X_1}(x_1) \cdot \cdots \cdot F_{X_n}(x_n).$$

The by Dynkin $\pi - \lambda$ theorem $P_{X_1} \otimes \cdots \otimes P_{X_n} = P_{(X_1,\ldots,X_n)}$.

3 $\implies$ 2: Assume condition (3) holds. For any Borel sets $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$, we can consider the event $\{X_1 \in B_1 \cap \cdots \cap X_n \in B_n\}$, which is the preimage of the product set $B_1 \times \cdots \times B_n$ under the random vector $(X_1, \ldots, X_n)$. By the definition of the joint distribution,

$$P(X_1 \in B_1 \cap \cdots \cap X_n \in B_n) = P_{(X_1,\ldots,X_n)}(B_1 \times \cdots \times B_n).$$

Since $P_{(X_1,\ldots,X_n)} = P_{X_1} \otimes \cdots \otimes P_{X_n}$, the property of product measures gives

$$(P_{X_1} \otimes \cdots \otimes P_{X_n})(B_1 \times \cdots \times B_n) = P_{X_1}(B_1) \ldots P_{X_n}(B_n) = P(X_1 \in B_1) \cdot \cdots \cdot P(X_n \in B_n).$$

Thus, the events $\{X_1 \in B_1\}, \ldots, \{X_n \in B_n\}$ are independent.

Theorem 14. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X, Y : \Omega \to \mathbb{R}$ be two independent random variables and $f, g : \mathbb{R} \to \mathbb{R}$ be two Borel functions. Then the random variables $f(X)$ and $g(Y)$ are independent.*

Proof. We need to show that the σ-algebras $\sigma(f(X))$ and $\sigma(g(Y))$ are independent. The σ-algebra generated by $f(X)$ is given by

$$\sigma(f(X)) = \{(f \circ X)^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R})\} = \{X^{-1}(f^{-1}(B)) \mid B \in \mathcal{B}(\mathbb{R})\}$$

Similarly, the σ-algebra generated by $g(Y)$ is

$$\sigma(g(Y)) = \{(g \circ Y)^{-1}(B) \mid B \in \mathcal{B}(\mathbb{R})\} = \{Y^{-1}(g^{-1}(B)) \mid B \in \mathcal{B}(\mathbb{R})\}$$

Let $B_1, B_2 \in \mathcal{B}(\mathbb{R})$. Since f and g are Borel functions, the preimages $A_1 = f^{-1}(B_1)$ and $A_2 = g^{-1}(B_2)$ are also in the Borel σ-algebra on $\mathbb{R}$, i.e., $A_1, A_2 \in \mathcal{B}(\mathbb{R})$.

We want to show that for any such sets, we have:

$$P(f(X) \in B_1 \text{ and } g(Y) \in B_2) = P(f(X) \in B_1)P(g(Y) \in B_2)$$

The events are:

$$\{f(X) \in B_1\} = \{\omega \in \Omega \mid f(X(\omega)) \in B_1\} = \{\omega \in \Omega \mid X(\omega) \in f^{-1}(B_1)\} = \{X \in A_1\}$$
$$\{g(Y) \in B_2\} = \{\omega \in \Omega \mid g(Y(\omega)) \in B_2\} = \{\omega \in \Omega \mid Y(\omega) \in g^{-1}(B_2)\} = \{Y \in A_2\}$$

Since X and Y are independent and $A_1, A_2 \in \mathcal{B}(\mathbb{R})$, we know that the events $\{X \in A_1\}$ and $\{Y \in A_2\}$ are independent. Therefore,

$$P(f(X) \in B_1 \cap g(Y) \in B_2) = P(X \in A_1 \cap Y \in A_2)$$

$$= P(X \in A_1)P(Y \in A_2) = P(f(X) \in B_1)P(g(Y) \in B_2)$$

This shows that the σ-algebras $\sigma(f(X))$ and $\sigma(g(Y))$ are independent, which completes the proof.

3.4 Discret random variables

Definition 23. *Let $X : \Omega \to \mathbb{R}^d$ be a d-dimensional random variable and let $S_X = \{x \in \mathbb{R}^d \mid P(X = x) > 0\}$. We say that X has a discrete distribution if $P(X \in S_X) = 1$ ($P_X(S_X) = 1$). A random variable with a discrete distribution is called a **discrete random variable**.*

Remark 13. *The set S_X of points with positive probability is at most countable. This can be seen by writing it as a countable union of finite sets:*

$$S_X = \bigcup_{n=1}^{\infty} \left\{ x \in \mathbb{R}^d \mid P(X = x) > \frac{1}{n} \right\}.$$

Each set in the union is finite because if it contained n or more elements, the total probability would exceed 1.

The **Probability Mass Function (PMF)** of a random variable X, denoted as p_X, is the essential tool used to describe the behavior of a **discrete random variable** (X). The PMF assigns a probability to every possible value the random variable X can take. The probability mass of a specific value x, denoted $p_X(x)$, is simply the probability of the event that the random variable X equals that value x:

$$p_X(x) = P(X = x)$$

In essence, the PMF is the map that tells us how likely each specific outcome of the discrete random variable is to occur.

The PMF must satisfy two fundamental properties, derived from the axioms of probability:

1. The sum of the probabilities of all possible values of X must equal 1. This is because the events $\{X = x\}$ are mutually disjoint and form a partition of the sample space.

$$\sum_{x \in S_X} p_X(x) = 1$$

2. For any $B \in \mathcal{B}(\mathbb{R}^d)$, the probability of B can be calculated as a sum over the points in S_X that are in B:

$$P_X(B) = P_X(B \cap S_X) = \sum_{x \in B \cap S_X} P_X(\{x\}) = \sum_{x \in B \cap S_X} P(X = x).$$

We establishe an important convention to avoid ambiguity: Upper case characters (e.g., X) denote the random variable. Lower case characters (e.g., x) denote the real numerical values the random variable can take.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ be a d-dimensional discrete random vector with the Joint PMF:

$$p_{\mathbf{X}}(x_1, x_2, \ldots, x_d) = P(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d)$$

To find the marginal PMF of a single variable, X_i, we must perform a process calle marginalization. This is achieved by summing the joint PMF over all possible combinations of the other $d - 1$ variables in the vector.

For example, to find the marginal PMF of any variable X_i is given by summing over all variables X_j where $j \neq i$:

$$\mathbf{p_{X_i}(x_i)} = \sum_{\mathbf{x_j, j \neq i}} \mathbf{p_X(x_1, x_2, \ldots, x_d)}$$

The same principle is used to find the marginal PMF of a subset of k variables, say $\mathbf{Y} = (X_1, \ldots, X_k)$, where $k < d$. We sum the joint PMF over all possible values of the remaining $d - k$ variables (X_{k+1} through X_d):

$$\mathbf{p_Y(x_1, \ldots, x_k)} = \sum_{\mathbf{x_{k+1}}} \sum_{\mathbf{x_{k+2}}} \cdots \sum_{\mathbf{x_d}} \mathbf{p_X(x_1, x_2, \ldots, x_d)}$$

The process of marginalization ensures that the resulting marginal PMF $p_{X_i}(x_i)$ represents the total probability of X_i taking the value x_i, considering all potential scenarios for the variables we summed over. This is a direct application of the Law of Total Probability.

Theorem 15. *Discrete random variables $X_1, \ldots, X_n$ are independent if and only if for any $x_1 \in S_{X_1}, \ldots, x_n \in S_{X_n}$ we have*

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1) \cdot \cdots \cdot P(X_n = x_n)$$

Proof. ($\Longrightarrow$) Trivial as (12) follows from the definition of independent random variables.

($\Longleftarrow$) To simplify notation we show only for $n = 2$ and $d = 1$ (for $n > 2$ the proof is analogous to $n = 2$). For any $B_1, B_2 \in \mathcal{B}(\mathbb{R})$.

$$P(X_1 \in B_1 \cap X_2 \in B_2) = P((X_1 \in S_{X_1} \cap B_1) \cap (X_2 \in S_{X_2} \cap B_2)) =$$

$$= \sum_{x_1 \in S_{X_1} \cap B_1} \sum_{x_2 \in S_{X_2} \cap B_2} P(X_1 = x_1 \cap X_2 = x_2) =$$

$$= \sum_{x_1 \in S_{X_1} \cap B_1} \sum_{x_2 \in S_{X_2} \cap B_2} P(X_1 = x_1) P(X_2 = x_2) = P(X_1 \in B_1) P(X_2 \in B_2)$$

Example 19. *In a certain population, a worker over 30 years old is chosen. Let:*

- $X =$ *Amount of years of education received*
- $Y =$ *Salary earned (in thousands of pesos)*

It is known that the joint probability mass function (PMF) of the random vector (X, Y) *is given by* $p_{XY}(x, y)$*, shown in the following table:*

$Y \setminus X$	7	12	18	24
40	0.14	0.23	0.02	0.01
100	0.06	0.16	0.25	0.03
150	0.00	0.01	0.03	0.06

The marginal PMF $p_X(x)$ *is found by summing the probabilities across the rows (for a fixed* x*), and* $p_Y(y)$ *is found by summing across the columns (for a fixed* y*).*

- *Marginal PMF for* Y *(*p_Y*):*

$$p_Y(40) = 0.14 + 0.23 + 0.02 + 0.01 = 0.40$$
$$p_Y(100) = 0.06 + 0.16 + 0.25 + 0.03 = 0.50$$
$$p_Y(150) = 0.00 + 0.01 + 0.03 + 0.06 = 0.10$$

- *Marginal PMF for* X *(*p_X*):*

$$p_X(7) = 0.14 + 0.06 + 0.00 = 0.20$$
$$p_X(12) = 0.23 + 0.16 + 0.01 = 0.40$$
$$p_X(18) = 0.02 + 0.25 + 0.03 = 0.30$$
$$p_X(24) = 0.01 + 0.03 + 0.06 = 0.10$$

The marginal PMFs are summarized in the following table (which extends the original joint PMF):

$Y \setminus X$	7	12	18	24	$p_Y(y)$
40	0.14	0.23	0.02	0.01	0.40
100	0.06	0.16	0.25	0.03	0.50
150	0.00	0.01	0.03	0.06	0.10
$p_X(x)$	0.20	0.40	0.30	0.10	1.00

For X *and* Y *to be independent, the condition* $p_{XY}(x, y) = p_X(x) p_Y(y)$ *must hold for all pairs* (x, y)*.*
Let's check the case $(X = 7, Y = 150)$*:*

- *Original Joint PMF:* $p_{XY}(7, 150) = 0.00$
- *Product of Marginal PMFs:* $p_X(7) p_Y(150) = (0.20) \times (0.10) = 0.02$

Since $p_{XY}(7, 150) = 0.00 \neq 0.02 = p_X(7) p_Y(150)$*, the variables* X *and* Y *are* ***NOT independent****.*

3.4.1 Important discrete distributions

One-point distribution

δ_a, $a \in \mathbb{R}^d$. A random variable X has a δ_a distribution if $P(X = a) = 1$, which means its support is $S_X = \{a\}$.

Key Property: This is the degenerate distribution, representing a situation with zero randomness. The probability mass is concentrated entirely at a single point a.

Two-point distribution

A random variable X has a two-point distribution centered at $\{a, b\}$ if $P(X = a) = p$ and $P(X = b) = 1 - p$, for some $p \in (0, 1)$. If $a = 0$ and $b = 1$, we call X a Bernoulli random variable.

Key Property: The Bernoulli distribution is the simplest non-degenerate distribution; it models a single event with only two outcomes: Success (1) or Failure (0).

Binomial distribution

A random variable X has a $B(n, p)$ distribution if its support is $S_X = \{0, 1, \ldots, n\}$ and for each $k \in S_X$, its probability mass function is given by

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

Interpretation: Consider a series of n independent experiments defined on the product probability space $(\Omega_1 \times \cdots \times \Omega_n, \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n, P_1 \otimes \cdots \otimes P_n)$. For each experiment, we consider a Bernoulli random variable $X_i : \Omega_i \to \mathbb{R}$ with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$. We can interpret $X_i = 1$ as a success and $X_i = 0$ as a failure. For the series of n experiments, we can define the random variable $X : \Omega_1 \times \cdots \times \Omega_n \to \mathbb{R}$ as the sum of the outcomes:

$$X((\omega_1, \ldots, \omega_n)) = X_1(\omega_1) + \cdots + X_n(\omega_n).$$

Then X represents the total number of successes in the series of n independent experiments.

The Binomial distribution answers: How many successes will occur in a fixed number of attempts?

Example 20. *Flipping a Coin Imagine you flip a fair coin* 10 *times. You want to know the probability of getting exactly* 7 *heads.*

- *Trial Type: Bernoulli (Heads or Tails).*
- *Fixed Number of Trials (n):* 10 *flips.*
- *Success Probability (p):* $P(Heads) = 0.5$.
- *Random Variable (X): The total number of Heads observed.*
- *Distribution:* $X \sim B(n = 10, p = 0.5)$.

The total number of trials must be 10. *The question concerns the count of successful outcomes (Heads) within that fixed set.*

Geometric distribution

Geom(p): A random variable X has a *Geom*(p) distribution if its probability mass function is $P(X = k) = (1 - p)^k p$ for $k \in S_X = \{0, 1, 2, \ldots\}$.
Interpretation: Consider an infinite sequence of independent Bernoulli trials, where X_i is as defined in the binomial interpretation. Let X denote the number of failures before the first success. Then X has a geometric distribution because the probability of k failures followed by one success is given by

$$P(X = k) = P(X_1 = 0 \cap \cdots \cap X_k = 0 \cap X_{k+1} = 1) =$$

$$= P(X_1 = 0) \ldots P(X_k = 0) P(X_{k+1} = 1) = (1 - p)^k p.$$

The Geometric distribution answers: How many failures will occur until the first success? (The number of attempts is not fixed).

Example 21. *Finding a Rare Card Imagine you are looking for a specific rare trading card that has a 5% chance of appearing in any single pack ($p = 0.05$). You keep buying packs until you find it.*

- *Trial Type: Bernoulli (Find Card or Don't Find Card).*
- *Success Probability (p): $P(\text{Find Card}) = 0.05$.*
- *Random Variable (X): The number of packs you failed to buy before finally getting the rare card.*
- *Distribution: $X \sim Geom(p = 0.05)$.*

The number of trials is not fixed. The trials stop as soon as the first success occurs.

Table 3.1: Comparison of Binomial and Geometric Structure

Feature	Binomial ($B(n,p)$)	Geometric ($Geom(p)$)
Number of Trials	Fixed (n)	Variable (until first success)
Random Variable (X)	Counts the number of successes	Counts the number of failures
Stopping Rule	Stop when n trials are completed.	Stop when the first success occurs.

Poisson distribution

$Pois(\lambda)$, $\lambda > 0$. A random variable X has a $Pois(\lambda)$ distribution if its support is $S_X = \{0, 1, 2, \dots\}$ and its probability mass function is

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Theorem 16. *Assume $(p_n)_{n=1}^{\infty}$ is a sequence of probabilities with $p_n \in (0,1)$ such that $\lim_{n\to\infty} np_n = \lambda > 0$. Then for any fixed integer $k \geq 0$,*

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

In other words, if X_n is a Binomial random variable with distribution $B(n, p_n)$ and X is a Poisson random variable with distribution $Pois(\lambda)$, and the condition $np_n \approx \lambda$ holds for large n, then the distribution of X_n can be approximated by the distribution of X.

Proof. Let $P(X_n = k)$ be the Binomial probability mass function (PMF):

$$P(X_n = k) = \binom{n}{k} p_n^k (1-p_n)^{n-k}$$

We use the given condition $\lim_{n\to\infty} np_n = \lambda$. Let $\lambda_n = np_n$, so that $p_n = \frac{\lambda_n}{n}$. Substituting this into the PMF and expanding the binomial coefficient $\binom{n}{k} = \frac{n!}{k!(n-k)!}$:

$$P(X_n = k) = \frac{n!}{k!(n-k)!} \left(\frac{\lambda_n}{n}\right)^k \left(1 - \frac{\lambda_n}{n}\right)^{n-k}$$

We rearrange the terms into three distinct factors for easier limit calculation:

$$P(X_n = k) = \left[\frac{1}{k!}\right] \cdot \left[\frac{n!}{(n-k)!n^k}\right] \cdot \left[\lambda_n^k \left(1 - \frac{\lambda_n}{n}\right)^{n-k}\right]$$

Now, we evaluate the limit of each factor as $n \to \infty$ (since k is a fixed integer):

$$\lim_{n\to\infty} \left[\frac{1}{k!}\right] = \frac{1}{k!}$$

$$\frac{n!}{(n-k)!n^k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k}$$

This can be written as a product of k fractions:

$$\frac{n}{n} \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-k+1}{n} = 1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right)$$

As $n \to \infty$, every term $(1 - j/n)$ approaches 1. Thus:

$$\lim_{n\to\infty} \left[\frac{n!}{(n-k)!n^k}\right] = 1 \cdot 1 \cdots 1 = 1$$

We split the power $(n-k)$ into n and $-k$ terms:

$$\lambda_n^k \left(1 - \frac{\lambda_n}{n}\right)^{n-k} = \lambda_n^k \left(1 - \frac{\lambda_n}{n}\right)^{n} \left(1 - \frac{\lambda_n}{n}\right)^{-k}$$

We find the limit of each sub-part:

1. $\lim_{n\to\infty} \lambda_n^k = \lambda^k$.
2. Using the definition of the limit for the exponential function, $\lim_{x\to\infty}(1 + \frac{c}{x})^x = e^c$:

$$\lim_{n\to\infty} \left(1 - \frac{\lambda_n}{n}\right)^{n} = e^{-\lambda}.$$

3. Since $\lim_{n\to\infty} \frac{\lambda_n}{n} = 0$:

$$\lim_{n\to\infty} \left(1 - \frac{\lambda_n}{n}\right)^{-k} = (1-0)^{-k} = 1.$$

Therefore, the limit of the third factor is:

$$\lim_{n\to\infty} \left[\lambda_n^k \left(1 - \frac{\lambda_n}{n}\right)^{n-k}\right] = \lambda^k \cdot e^{-\lambda} \cdot 1 = \lambda^k e^{-\lambda}$$

Multiplying the limits of the three factors (Constant, Combinatorial, and Exponential):

$$\lim_{n\to\infty} P(X_n = k) = \left[\frac{1}{k!}\right] \cdot [1] \cdot \left[\lambda^k e^{-\lambda}\right] = \frac{\lambda^k}{k!} e^{-\lambda}$$

This is the PMF for a Poisson distribution with parameter λ.

3.4.2 Functions of Random Variables

We know that if $Y = g(X)$ is a function of a random variable X, then Y is also a random variable. If X is discrete with PMF p_X, then Y is also discrete, and its PMF p_Y can be calculated using the PMF of X. In particular, to obtain $p_Y(y)$ for any y, we add the probabilities of all values of x such that $g(x) = y$:

$$p_Y(y) = \sum_{\{x|g(x)=y\}} p_X(x). \tag{3.1}$$

Example 22. *Let $Y = |X|$ and let us apply the preceding formula for the PMF p_Y to the case where*

$$p_X(x) = \begin{cases} 1/9 & \text{if } x \text{ is an integer in the range } [-4, 4], \\ 0 & \text{otherwise.} \end{cases}$$

The possible values of Y are $y = 0, 1, 2, 3, 4$. To compute $p_Y(y)$ for some given value y from this range, we must add $p_X(x)$ over all values x such that $|x| = y$. In particular, there is only one value of X that corresponds to $y = 0$, namely $x = 0$. Thus,

$$p_Y(0) = p_X(0) = \frac{1}{9}.$$

Also, there are two values of X that correspond to each $y = 1, 2, 3, 4$ (i.e., $x = y$ and $x = -y$), so for example,

$$p_Y(1) = p_X(-1) + p_X(1) = \frac{1}{9} + \frac{1}{9} = \frac{2}{9}.$$

Thus, the PMF of Y is

$$p_Y(y) = \begin{cases} 2/9 & \text{if } y = 1, 2, 3, 4, \\ 1/9 & \text{if } y = 0, \\ 0 & \text{otherwise.} \end{cases}$$

For another related example, let $Z = X^2$. To obtain the PMF of Z, we can view it either as the square of the random variable X or as the square of the random variable Y. By applying the formula $p_Z(z) = \sum_{\{x|x^2=z\}} p_X(x)$ or the formula $p_Z(z) = \sum_{\{y|y^2=z\}} p_Y(y)$, we obtain

$$p_Z(z) = \begin{cases} 2/9 & \text{if } z = 1, 4, 9, 16, \\ 1/9 & \text{if } z = 0, \\ 0 & \text{otherwise.} \end{cases}$$

3.5 Continous Distribution

Definition 24. *A d-dim random variable X has a continuous distribution if there exists a Borel function $g : \mathbb{R}^d \to [0, \infty)$ such that*

$$P(X \in B) = P_X(B) = \int_B g(x)dx, \quad \forall B \in \mathcal{B}(\mathbb{R}^d)$$

The function g is called the density of P_X (or the density of X).

Remark 14.

1. *If a random variable X has a continuous distribution with density g, then for any single point $x \in \mathbb{R}^d$, the probability is zero, i.e., $P(X = x) = 0$. Consequently, the set of points with positive probability, S_X, is empty.*

2. *The density of a continuous distribution is not unique. If g is a density for a distribution P_X, then any other Borel function $\tilde{g} : \mathbb{R}^d \to [0, \infty)$ is also a density for P_X if and only if $g = \tilde{g}$ almost everywhere.*

3. *A Borel function $g : \mathbb{R}^d \to [0, \infty)$ is a valid density for some probability distribution if and only if its integral over $\mathbb{R}^d$ equals 1, i.e., $\int_{\mathbb{R}^d} g(x)dx = 1$. The probability measure for this distribution is then defined by $P(B) = \int_B g(x)dx$ for any Borel set B.*

4. *A random variable X has a continuous distribution if and only if its cumulative distribution function (CDF) F_X can be expressed as a multidimensional integral of a non-negative Borel function g:*

$$F_X(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_d} g(y_1, \ldots, y_d) dy_1 dy_2 \ldots dy_d.$$

*The **($\implies$)** direction follows directly from the definition of a continuous distribution. For the **($\impliedby$)** direction, if such a function g exists, we can show that $\int_{\mathbb{R}^d} g(x)dx = 1$. This function g then defines a continuous distribution, say μ. Since the CDF of μ is identical to F_X, it follows that the distributions themselves are equal, i.e., $P_X = \mu$.*

5. *If the cumulative distribution function F_X of a random variable X is differentiable, then the density function g can be found by taking the mixed partial derivative:*

$$g(x_1, \ldots, x_d) = \frac{\partial^d F_X(x_1, \ldots, x_d)}{\partial x_1 \ldots \partial x_d},$$

This equality holds almost everywhere. If this function g is non-negative and $\int_{\mathbb{R}^d} g(x)dx = 1$, then X has a continuous distribution with density g.

Example 23. *Consider a point chosen uniformly at random from a disk of radius R. Let X be the random variable representing the distance of this point from the center of the disk. The probability of the point being at a distance less than or equal to t is the ratio of the area of a disk of radius t to the area of the entire disk of radius R. The cumulative distribution function (CDF) of X is therefore:*

$$F_X(t) = P(X \leq t) = \begin{cases} 0 & t < 0 \\ \frac{\pi t^2}{\pi R^2} = \frac{t^2}{R^2} & 0 \leq t < R \\ 1 & t \geq R \end{cases}$$

The probability density function (PDF), $g(t)$, is found by taking the derivative of the CDF:

$$g(t) = F'_X(t) = \begin{cases} 0 & t < 0 \text{ or } t > R \\ \frac{2t}{R^2} & 0 \leq t < R \end{cases}$$

To verify that $g(t)$ is a valid density function, we integrate it over its domain:

$$\int_{-\infty}^{\infty} g(t)dt = \int_0^R \frac{2t}{R^2} dt = \frac{1}{R^2}[t^2]_0^R = \frac{R^2 - 0}{R^2} = 1.$$

Since the integral equals 1, we have confirmed that $g(t)$ is the probability density function for the random variable X.

Theorem 17. *If the random vector $X = (X_1, \ldots, X_d)$ has a continuous distribution with joint density $g : \mathbb{R}^d \to [0, \infty)$, then its marginal distributions are also continuous. The marginal density of any component X_i is given by*

$$g_{X_i}(x_i) = \int_{\mathbb{R}^{d-1}} g(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)\, dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_d.$$

More generally, to find the joint density of a sub-vector $(X_{i_1}, \ldots, X_{i_k})$, we integrate the joint density

g over all variables x_j for $j \notin \{i_1, \ldots, i_k\}$.

Proof. Let B be any Borel set in $\mathcal{B}(\mathbb{R})$. By the definition of the joint density, the probability of the event $\{X_i \in B\}$ is given by the integral of the joint density over the set $\mathbb{R}^{i-1} \times B \times \mathbb{R}^{d-i}$:

$$P(X_i \in B) = P\left((X_1, \ldots, X_d) \in \mathbb{R}^{i-1} \times B \times \mathbb{R}^{d-i}\right) = \int_{\mathbb{R}^{i-1} \times B \times \mathbb{R}^{d-i}} g(x_1, \ldots, x_d)\, dx_1 \ldots dx_d.$$

By Fubini's Theorem, we can rearrange the integrals to separate the variable x_i from the others:

$$P(X_i \in B) = \int_B \left(\int_{\mathbb{R}^{d-1}} g(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)\, dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_d \right) dx_i.$$

The expression within the parentheses depends only on x_i. We can therefore define the marginal density for X_i as:

$$g_{X_i}(x_i) = \int_{\mathbb{R}^{d-1}} g(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)\, dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_d.$$

This shows that $P(X_i \in B) = \int_B g_{X_i}(x_i)\, dx_i$, which, by definition, means X_i has a continuous distribution with density g_{X_i}.

3.5.1 Important continuous distributions

Uniform Distribution on D

$U(D)$. Let $D \in \mathcal{B}(\mathbb{R}^d)$ be a Borel set with finite volume $|D| < \infty$. A random variable X has a uniform distribution on D if its probability density function is

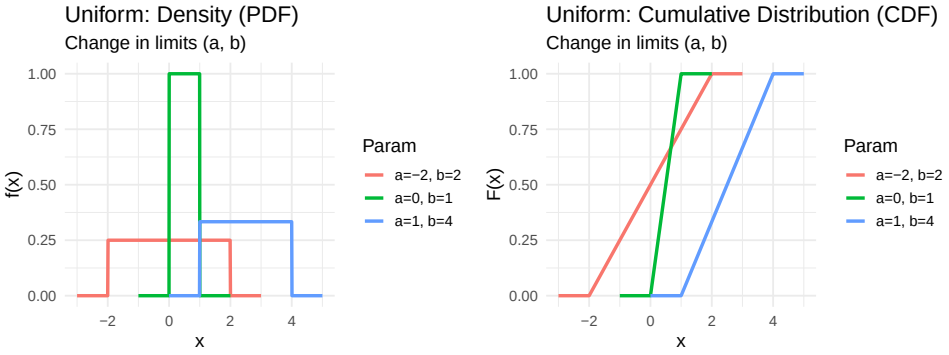
$$g(x) = \begin{cases} \frac{1}{|D|} & x \in D \\ 0 & x \notin D \end{cases}$$

For any Borel set $B \in \mathcal{B}(\mathbb{R}^d)$, the probability that X falls within B is given by

$$P(X \in B) = \int_B g(x)dx = \frac{|B \cap D|}{|D|}.$$

In the one-dimensional case where $D = [a, b]$, the density function simplifies to $g(x) = \frac{1}{b-a}\mathbf{1}_{[a,b]}(x)$. The corresponding cumulative distribution function is

$$F_X(x) = \int_{-\infty}^{x} g(s)ds = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x < b \\ 1 & x \ge b \end{cases}$$



Exponential Distribution

$Exp(\lambda)$, with parameter $\lambda > 0$. A random variable X has an exponential distribution if its density function is

$$g(x) = \lambda e^{-\lambda x} \mathbf{1}_{[0,\infty)}(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases}$$

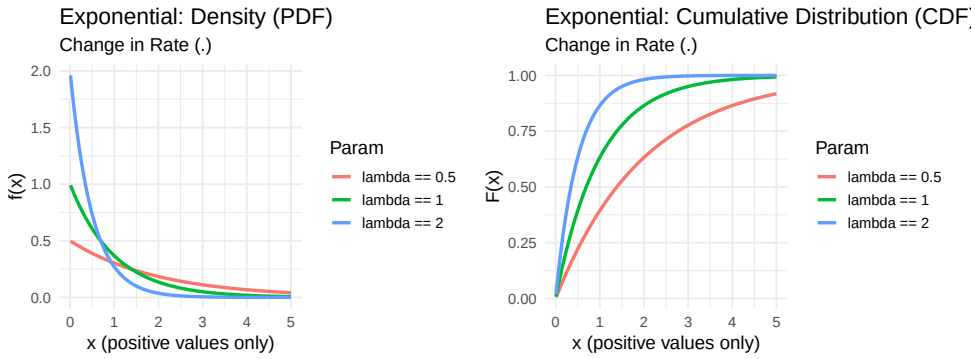
The cumulative distribution function is

$$F_X(x) = \int_{-\infty}^{x} g(s)ds = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}.$$

Interpretation: The Memoryless Property. The exponential distribution is commonly used to model the waiting time for a random event. Its key characteristic is the "lack of memory," which means the probability of the event occurring in the future is independent of how long we have already waited. This property is formally expressed as

$$P(X > t + s | X > s) = P(X > t)$$

for all $t, s \geq 0$. Let $f(t) = P(X > t)$. The memoryless property implies $f(t + s) = f(t)f(s)$, which leads to the functional equation whose only continuous solution is $f(t) = e^{-\lambda t}$ for some $\lambda > 0$, where $\lambda = -f'(0)$.



Gamma Distribution

$\Gamma(\alpha, \lambda)$, with shape parameter $\alpha > 0$ and rate parameter $\lambda > 0$. A random variable X has a Gamma distribution if its support is $x \in [0, \infty)$ and its probability density function is

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

The constant $\Gamma(\alpha)$ in the denominator is the Gamma function, defined by the integral:

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

- Properties of the Gamma Function:
  - For any integer $n \geq 1$: $\Gamma(n) = (n-1)!$.
  - For $\alpha > 0$: $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$.
  - $\Gamma(1/2) = \sqrt{\pi}$.
- Relationship to Other Distributions:

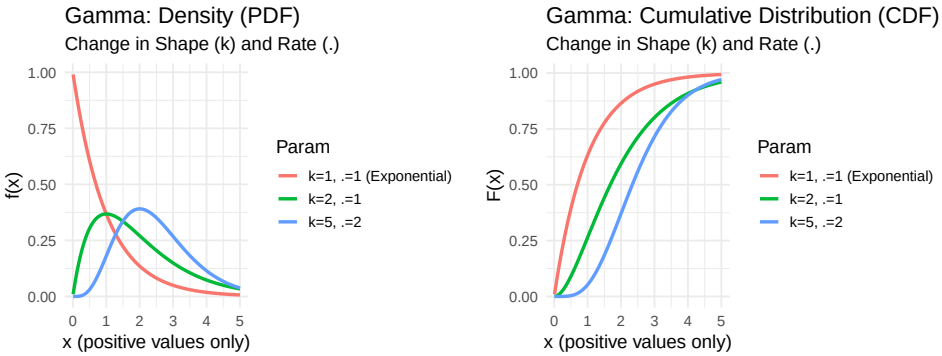
1. Generalization of the Exponential: The Gamma distribution $\Gamma(\alpha, \lambda)$ is a direct generalization of the Exponential distribution $Exp(\lambda)$. Specifically,

$$\Gamma(1, \lambda) \equiv Exp(\lambda).$$

2. Sum of Exponentials: If $X_1, X_2, \ldots, X_n$ are n independent and identically distributed (i.i.d.) Exponential random variables, $X_i \sim Exp(\lambda)$, then their sum $Y = \sum_{i=1}^{n} X_i$ follows a Gamma distribution: $Y \sim \Gamma(n, \lambda)$.
3. Chi-squared (χ^2) Distribution: The Chi-squared distribution with ν degrees of freedom is a special case of the Gamma distribution:

$$\chi^2(\nu) \equiv \Gamma\left(\frac{\nu}{2}, \frac{1}{2}\right).$$

The Gamma distribution, particularly when α is an integer (n), is used to model the waiting time until the n-th event occurs in a Poisson process with rate λ. This reflects the summation property mentioned above.



Multivariate Normal Distribution

$N(m, A)$, where $m \in \mathbb{R}^d$ is the mean vector and A is a symmetric positive definite matrix. A random vector X has a normal distribution if its density function is

$$g(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(A)}} \exp\left(-\frac{1}{2}(x-m)^T A^{-1}(x-m)\right).$$

To confirm that $g(x)$ is a valid density, we must show that $\int_{\mathbb{R}^d} g(x)dx = 1$. This can be demonstrated by a change of variables, which transforms the integral into a product of standard Gaussian integrals. Let $x - m = Oy$, where O is an orthogonal matrix that diagonalizes A^{-1}. The Jacobian of this transformation is $|\det(O)| = 1$. The integral becomes

$$\int_{\mathbb{R}^d} g(x)dx = \frac{1}{(2\pi)^{d/2}\sqrt{\det(A)}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2} y^T (O^T A^{-1} O) y\right) dy.$$

The matrix $O^T A^{-1} O$ is a diagonal matrix with eigenvalues $a_1, \ldots, a_d$.

$$\int_{\mathbb{R}^d} g(x)dx = \frac{1}{(2\pi)^{d/2}\sqrt{\det(A)}} \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}\sum_{k=1}^{d} a_k y_k^2\right) dy = \frac{1}{(2\pi)^{d/2}\sqrt{\det(A)}} \prod_{k=1}^{d} \int_{-\infty}^{\infty} e^{-\frac{1}{2}a_k y_k^2} dy_k.$$

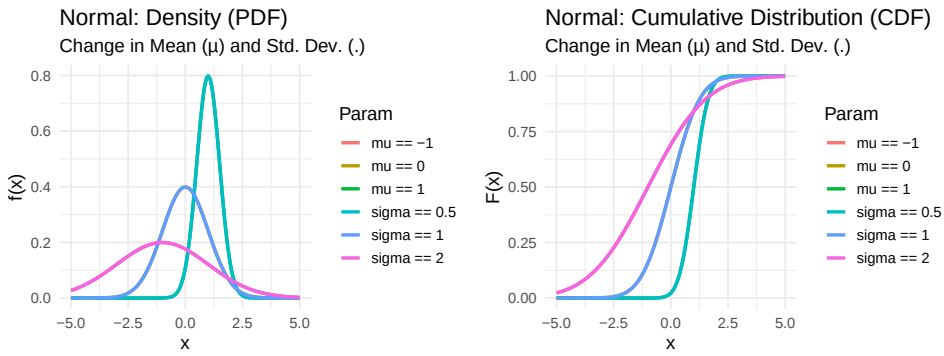
Using the well-known integral identity $\int_{-\infty}^{\infty} e^{-cy^2} dy = \sqrt{\frac{\pi}{c}}$, we get

$$= \frac{1}{(2\pi)^{d/2}\sqrt{\det(A)}} \prod_{k=1}^{d} \sqrt{\frac{2\pi}{a_k}} = \frac{\sqrt{\det(A^{-1})}}{(2\pi)^{d/2}} \frac{(2\pi)^{d/2}}{\sqrt{\prod_{k=1}^{d} a_k}} = \frac{\sqrt{\det(A^{-1})}}{\sqrt{\det(O^T A^{-1} O)}} = \frac{\sqrt{\det(A^{-1})}}{\sqrt{\det(A^{-1})}} = 1.$$

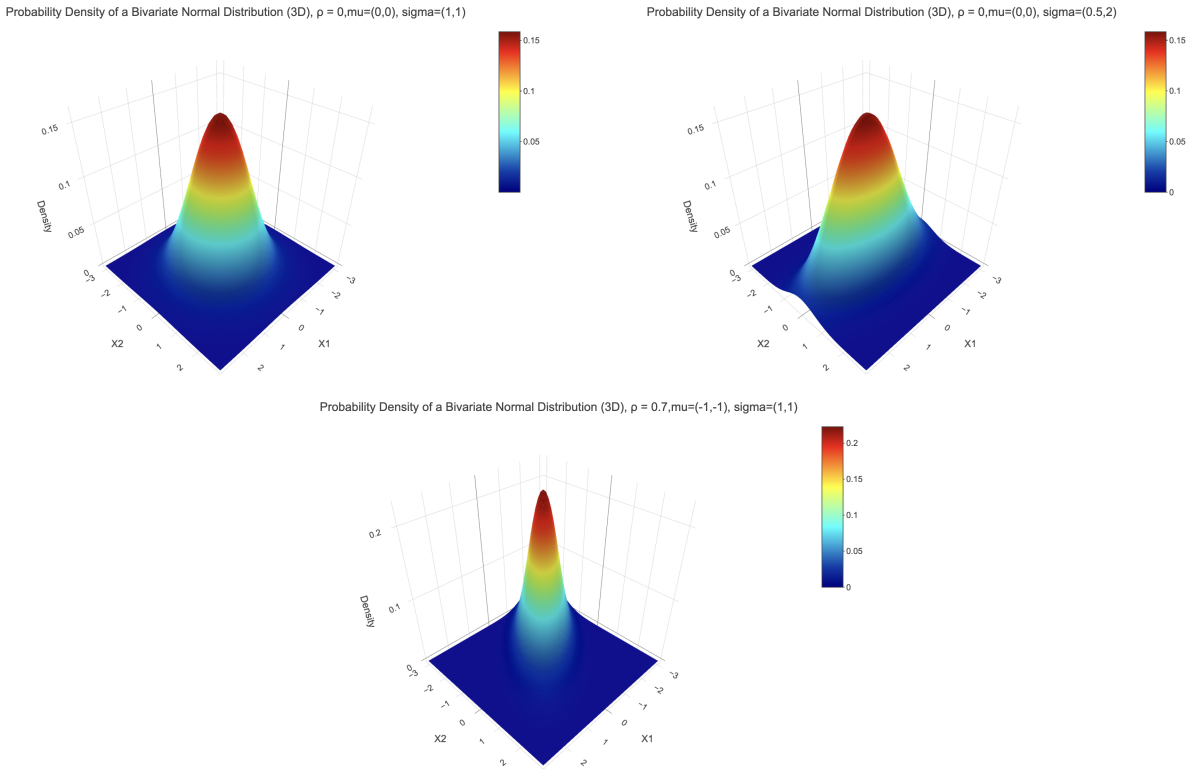
The One-Dimensional Case. When $d = 1$, the m is a scalar $m \in \mathbb{R}$ and the A is a scalar $A = \sigma^2 > 0$. This distribution is denoted by $N(m, \sigma^2)$, and its density is

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

The special case where $m = 0$ and $\sigma = 1$ is called the **standard normal distribution**, denoted by $N(0, 1)$.



The following plots are de densities of 3 Bivariate normal distribution using different combinations of parameters. **Scenario 1: Zero Correlation** ($\rho = 0$) represents the case where X_1 and X_2 are independent. The resulting surface is a perfectly circular, symmetric bell shape. **Scenario 2: Zero Correlation** ($\rho = 0$) butusing two diferent variances. **Scenario 3: Strong Positive Correlation** ($\rho = 0.7$) visualizes a distribution where the variables X_1 and X_2 move in opposite directions. The surface is oriented diagonally from top-right to bottom-left.



Example 24. *The true lifetime (X) of a light bulb follows an Exponential distribution with parameter λ. A replacement policy is established: the light bulb is changed if it fails* ***before*** *2000 hours of use, or if it reaches 2000 hours of use while still operational. We consider the observed usage time (Y) of the light bulb. The variable Y is a* ***mixed random variable*** *because it possesses both a continuous portion and a discrete probability mass point.*

If the light bulb fails before 2000 hours ($X < 2000$), we observe the actual failure time, $Y = X$. The probability density function (PDF) for Y in this interval is the same as that of X, truncated to the interval $(0, 2000)$:

$$f_Y(y) = f_X(y) = \lambda e^{-\lambda y}, \quad \text{for } 0 < y < 2000$$

If the light bulb is still working at 2000 hours ($X \geq 2000$), it is replaced, and the registered usage time is exactly $Y = 2000$. The probability of this event is the probability mass:

$$P(Y = 2000) = P(X \geq 2000) = e^{-\lambda \cdot 2000}$$

The distribution of the observed usage time Y is a mixed distribution, defined by the continuous PDF over $(0, 2000)$ and the discrete mass point at $y = 2000$.

Therefore, this variable is ***neither purely continuous nor purely discrete.***

Theorem 18. *Let $X_1, X_2, \ldots, X_n$ be continuous random variables with respective probability density functions (PDFs) $g_1, g_2, \ldots, g_n$. These random variables are independent if and only if their joint PDF, $g(x_1, \ldots, x_n)$, is equal to the product of their individual PDFs:*

$$g(x_1, \ldots, x_n) = g_1(x_1) \cdot g_2(x_2) \cdot \cdots \cdot g_n(x_n).$$

Proof. ($\Longrightarrow$) Assume that $X_1, \ldots, X_n$ are independent. By definition, their joint cumulative distribution function (CDF) is the product of their individual CDFs:

$$F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = F_{X_1}(x_1) \ldots F_{X_n}(x_n).$$

By expressing each individual CDF as an integral of its PDF, we get:

$$= \left(\int_{-\infty}^{x_1} g_1(y_1)\, dy_1 \right) \ldots \left(\int_{-\infty}^{x_n} g_n(y_n)\, dy_n \right).$$

This product of integrals can be written as a single multiple integral over the region $(-\infty, x_1] \times \cdots \times (-\infty, x_n]$:

$$= \int_{(-\infty,x_1]\times\cdots\times(-\infty,x_n]} g_1(y_1) \ldots g_n(y_n)\, dy_1 \ldots dy_n.$$

By the definition of a joint PDF, the joint density $g(x_1, \ldots, x_n)$ is the function that is integrated to get the joint CDF. Thus, we conclude that:

$$g(x_1, \ldots, x_n) = g_1(x_1) \ldots g_n(x_n).$$

($\Longleftarrow$) Now, assume that the joint PDF is the product of the individual PDFs: $g(x_1, \ldots, x_n) = g_1(x_1) \ldots g_n(x_n)$. The joint CDF is defined as the integral of the joint PDF:

$$F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = \int_{(-\infty,x_1]\times\cdots\times(-\infty,x_n]} g(y_1, \ldots, y_n)\, dy_1 \ldots dy_n.$$

Substituting our assumption for the joint PDF, we have:

$$= \int_{(-\infty,x_1]\times\cdots\times(-\infty,x_n]} g_1(y_1) \ldots g_n(y_n)\, dy_1 \ldots dy_n.$$

This multiple integral can be separated into a product of single integrals:

$$= \left(\int_{-\infty}^{x_1} g_1(y_1)\, dy_1 \right) \dots \left(\int_{-\infty}^{x_n} g_n(y_n)\, dy_n \right).$$

Each of these individual integrals is the definition of the respective CDF. Therefore, we have shown that the joint CDF is the product of the marginal CDFs:

$$= F_{X_1}(x_1) \dots F_{X_n}(x_n).$$

By definition, this means that the random variables $X_1, \dots, X_n$ are independent.

Example 25. *Sum of Two Uniform Random Variables Let X_1 and X_2 be independent random variables, both uniformly distributed on the interval $[0,1]$, denoted $X_1, X_2 \sim U([0,1])$. The joint probability density function (PDF), $f(x_1, x_2)$, is:*

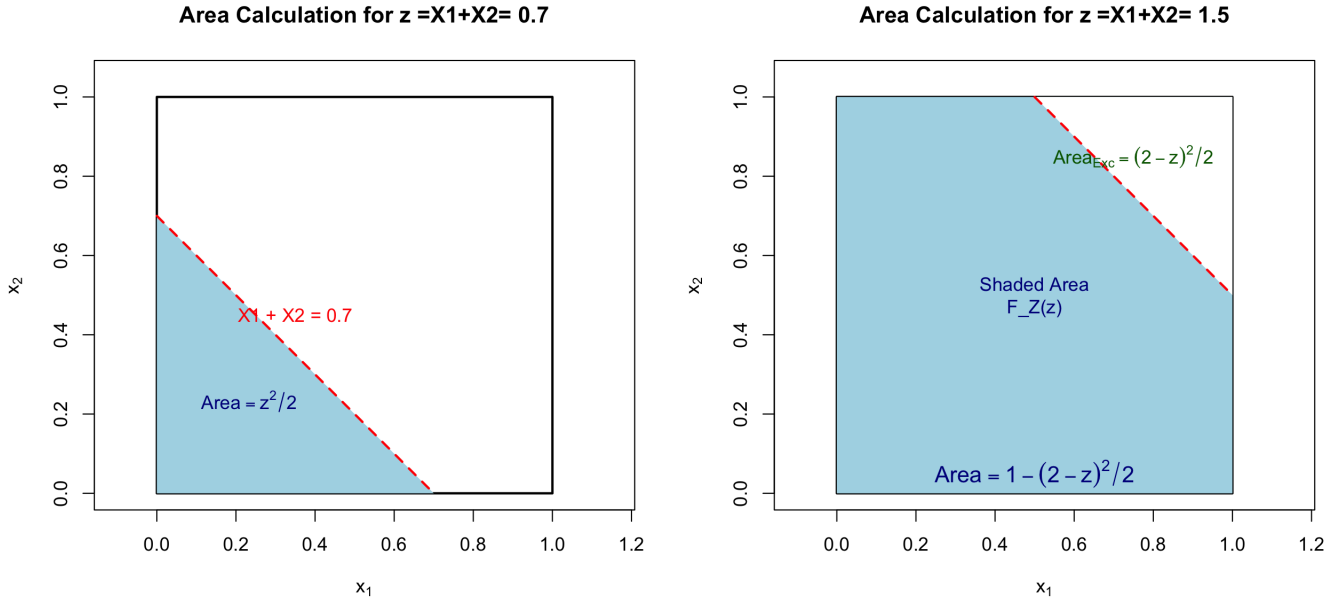
$$f(x_1, x_2) = \mathbf{1}_{[0,1]}(x_1)\mathbf{1}_{[0,1]}(x_2) = \begin{cases} 1 & \text{if } 0 \le x_1 \le 1 \text{ and } 0 \le x_2 \le 1 \\ 0 & \text{otherwise} \end{cases}$$

We seek the probability density function (PDF) of the sum, $Z = X_1 + X_2$, by first calculating the Cumulative Distribution Function (CDF), $F_Z(z) = P(Z \le z)$.

$$F_Z(z) = P(X_1 + X_2 \le z) = \iint_{x_1+x_2 \le z} f(x_1, x_2)\, dx_1\, dx_2$$

*Since the joint density is 1 inside the unit square $[0,1] \times [0,1]$, the probability $F_Z(z)$ is simply the **area** of the region defined by $x_1 + x_2 \le z$ **inside the unit square**. The total area of the square is 1.*

We analyze the cases based on the value of z:



1. ***Case*** *$z < 0$: The region $x_1 + x_2 \le z$ does not intersect the unit square.*

$$F_Z(z) = 0$$

2. ***Case*** *$0 \le z < 1$: The region $x_1 + x_2 \le z$ cuts the unit square and forms a right triangle with vertices $(0,0)$, $(z,0)$, and $(0,z)$. The area of this triangle is $\frac{1}{2} \cdot \text{base} \cdot \text{height} = \frac{1}{2} z \cdot z$.*

$$F_Z(z) = \frac{z^2}{2}$$

3. ***Case*** $1 \leq z < 2$*: The region* $x_1 + x_2 \leq z$ *covers the entire unit square, except for a small triangle in the top right corner defined by* $x_1 + x_2 > z$*. The base and height of this excluded triangle are both equal to* $1 - (z-1) = 2 - z$*. The excluded area is* $\frac{1}{2}(2-z)^2$*. The included area* $F_Z(z)$ *is:*

$$F_Z(z) = (\textit{Total Area}) - (\textit{Excluded Area}) = 1 - \frac{(2-z)^2}{2}$$

4. ***Case*** $z \geq 2$*: The region* $x_1 + x_2 \leq z$ *completely covers the unit square.*

$$F_Z(z) = 1$$

To obtain the PDF, $f_Z(z)$*, we differentiate the CDF* $F_Z(z)$ *with respect to* z*:*

- *If* $0 \leq z < 1$*:* $f_Z(z) = \frac{d}{dz}\left(\frac{z^2}{2}\right) = z$
- *If* $1 \leq z < 2$*:* $f_Z(z) = \frac{d}{dz}\left(1 - \frac{(2-z)^2}{2}\right) = -\frac{1}{2} \cdot 2(2-z) \cdot (-1) = 2 - z$

Therefore, the probability density function is:

$$f_Z(z) = \begin{cases} 0 & z < 0 \\ z & 0 \leq z < 1 \\ 2 - z & 1 \leq z < 2 \\ 0 & z \geq 2 \end{cases}$$

This result is known as the ***triangular distribution****.*

The elegant geometric solution derived above, which relies on calculating the area of the region $X_1 + X_2 \leq z$ *inside the unit square, is mathematically equivalent to the* ***Convolution Theorem*** *for the sum of independent random variables.*

For two independent continuous random variables X_1 *and* X_2 *with PDFs* $f_{X_1}(x)$ *and* $f_{X_2}(x)$*, the PDF of their sum* $Z = X_1 + X_2$ *is given by the convolution integral:*

$$f_Z(z) = (f_{X_1} * f_{X_2})(z) = \int_{-\infty}^{\infty} f_{X_1}(u) f_{X_2}(z-u)\, du$$

In this specific case, both PDFs are indicator functions over $[0,1]$*:* $f_{X_1}(x) = f_{X_2}(x) = \mathbf{1}_{[0,1]}(x)$*. Substituting this into the convolution integral yields:*

$$f_Z(z) = \int_{-\infty}^{\infty} \mathbf{1}_{[0,1]}(u) \mathbf{1}_{[0,1]}(z-u)\, du$$

The term $\mathbf{1}_{[0,1]}(u)$ *restricts the integration limits to* $0 \leq u \leq 1$*. The term* $\mathbf{1}_{[0,1]}(z-u)$ *imposes the constraint* $0 \leq z - u \leq 1$*, which can be rewritten as:*

$$z - 1 \leq u \leq z$$

Combining the restrictions on the integration variable u*:*

$$\max(0, z-1) \leq u \leq \min(1, z)$$

Since the integrand is 1 when these conditions are met, the integral simplifies to the length of the integration interval:

$$f_Z(z) = \min(1, z) - \max(0, z-1)$$

Evaluating this length for the relevant ranges of z*:*

- *If* $0 \leq z < 1$*:* $f_Z(z) = \min(1, z) - \max(0, z-1) = z - 0 = z$

- *If* $1 \leq z < 2$*:* $f_Z(z) = \min(1, z) - \max(0, z-1) = 1 - (z-1) = 2 - z$

This confirms that the result obtained geometrically through the CDF method ($f_Z(z) = z$ for $0 \leq z < 1$ and $f_Z(z) = 2 - z$ for $1 \leq z < 2$) is identical to the result derived from the formal convolution integral.

Theorem 19. *Let X_1 and X_2 be independent continuous random variables with probability density functions (PDFs) $g_1(x)$ and $g_2(x)$, respectively. Then the random variable $X_1 + X_2$ has a PDF given by the convolution of g_1 and g_2, denoted as $g_1 * g_2(x)$:*

$$(g_1 * g_2)(x) = \int_{\mathbb{R}} g_1(x-y) g_2(y)\, dy.$$

Proof. The proof relies on finding the cumulative distribution function (CDF) of $X_1 + X_2$ and then differentiating it to find the PDF. Let $Z = X_1 + X_2$. The CDF of Z is given by:

$$F_Z(z) = P(Z \leq z) = P(X_1 + X_2 \leq z).$$

This probability can be expressed as a joint integral over the region where $x_1 + x_2 \leq z$:

$$F_Z(z) = P((X_1, X_2) \in \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 + x_2 \leq z\}).$$

Since X_1 and X_2 are independent, their joint PDF is the product of their marginal PDFs, $g(x_1, x_2) = g_1(x_1) g_2(x_2)$. Therefore, we can write the joint integral as:

$$F_Z(z) = \iint_{\{(x_1,x_2) | x_1 + x_2 \leq z\}} g_1(x_1) g_2(x_2)\, dx_1\, dx_2.$$

We can evaluate this integral using Fubini's theorem by integrating with respect to x_1 first, for a fixed x_2. The condition $x_1 + x_2 \leq z$ is equivalent to $x_1 \leq z - x_2$.

$$F_Z(z) = \int_{\mathbb{R}} g_2(x_2) \left(\int_{-\infty}^{z-x_2} g_1(x_1)\, dx_1 \right) dx_2.$$

Let's consider the inner integral. By the Fundamental Theorem of Calculus, the inner integral is the CDF of X_1 evaluated at $z - x_2$.

$$F_Z(z) = \int_{\mathbb{R}} g_2(x_2) F_{X_1}(z - x_2)\, dx_2.$$

To find the PDF of Z, we differentiate its CDF with respect to z:

$$g_Z(z) = \frac{d}{dz} F_Z(z) = \frac{d}{dz} \int_{\mathbb{R}} g_2(x_2) F_{X_1}(z - x_2)\, dx_2.$$

Using the Leibniz integral rule, we can move the differentiation inside the integral:

$$g_Z(z) = \int_{\mathbb{R}} g_2(x_2) \frac{d}{dz} F_{X_1}(z - x_2)\, dx_2.$$

By the chain rule, $\frac{d}{dz}F_{X_1}(z-x_2) = F'_{X_1}(z-x_2)\cdot(1) = g_1(z-x_2)$. Substituting this back into the integral, we get:

$$g_Z(z) = \int_{\mathbb{R}} g_2(x_2)g_1(z-x_2)\,dx_2.$$

This is the convolution of g_1 and g_2, which completes the proof.

Example 26. *Sum of Two Normal Random Variables Let X_1 and X_2 be independent normal random variables, $X_1 \sim N(m_1,\sigma_1^2)$ and $X_2 \sim N(m_2,\sigma_2^2)$. The PDFs are $g_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(x-m_i)^2}{2\sigma_i^2}\right)$ for $i=1,2$. The convolution of these two densities is known to be another normal density. A detailed calculation shows that:*

$$(g_1 * g_2)(x) = \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\left(-\frac{(x-(m_1+m_2))^2}{2(\sigma_1^2+\sigma_2^2)}\right).$$

This means that the sum of two independent normal random variables is also a normal random variable, with its mean and variance being the sums of the individual means and variances:

$$X_1 + X_2 \sim N(m_1+m_2, \sigma_1^2+\sigma_2^2).$$

This property generalizes to any number of independent normal random variables. If $X_i \sim N(m_i,\sigma_i^2)$ for $i=1,\ldots,n$, and they are all independent, then their sum is also a normal random variable:

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

3.6 Transformation of Random Variables

Theorem 20 (Change of Variables Formula for Densities)**.** *Let X be a continuous random variable with probability density function (PDF) $g_X(\mathbf{x})$. Let $\mathbf{Y} = \phi(\mathbf{X})$, where $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is a transformation that is injective and continuously differentiable (C^1) on the support of $\mathbf{X}$. Then the PDF of $\mathbf{Y}$, denoted $g_Y(\mathbf{y})$, is given by:*

$$g_Y(\mathbf{y}) = g_X(\phi^{-1}(\mathbf{y}))\left|\det\left(D\phi^{-1}(\mathbf{y})\right)\right|,$$

where $D\phi^{-1}(\mathbf{y})$ is the Jacobian matrix of the inverse transformation.

Proof. The proof is based on the change of variables formula for integrals. We consider both the one-dimensional and multi-dimensional cases.

Case 1: One-Dimensional Random Variable ($d=1$)

Let X be a continuous random variable with PDF $g_X(x)$, and let $Y = \phi(X)$, where $\phi : \mathbb{R} \to \mathbb{R}$ is injective and C^1. For any measurable set $B \in \mathcal{B}(\mathbb{R})$, the probability of Y being in B is:

$$P(Y \in B) = P(\phi(X) \in B) = P(X \in \phi^{-1}(B)).$$

By the definition of the PDF, this probability is the integral of $g_X(x)$ over the set $\phi^{-1}(B)$:

$$P(Y \in B) = \int_{\phi^{-1}(B)} g_X(x)\,dx.$$

We perform a change of variables in the integral. Let $y = \phi(x)$, so $x = \phi^{-1}(y)$. The differential element transforms as $dx = \left|\frac{d\phi^{-1}(y)}{dy}\right| dy$. Applying this substitution, the integral over the set $\phi^{-1}(B)$

in the x-domain transforms into an integral over the set B in the y-domain:

$$P(Y \in B) = \int_B g_X(\phi^{-1}(y)) \left| \frac{d\phi^{-1}(y)}{dy} \right| dy.$$

By comparing this with the definition of the density of Y, $P(Y \in B) = \int_B g_Y(y)\, dy$, we identify the PDF of Y as:

$$g_Y(y) = g_X(\phi^{-1}(y)) \left| \frac{d\phi^{-1}(y)}{dy} \right|.$$

Case 2: Multi-Dimensional Random Variable ($d > 1$)

Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a continuous random vector with joint PDF $g_X(\mathbf{x})$, and let $\mathbf{Y} = \phi(\mathbf{X})$, where $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is injective and C^1. For any measurable set $B \in \mathcal{B}(\mathbb{R}^d)$, the probability of $\mathbf{Y}$ being in B is:

$$P(\mathbf{Y} \in B) = P(\phi(\mathbf{X}) \in B) = P(\mathbf{X} \in \phi^{-1}(B)).$$

This probability is the integral of the joint PDF $g_X(\mathbf{x})$ over the set $\phi^{-1}(B)$:

$$P(\mathbf{Y} \in B) = \int_{\phi^{-1}(B)} g_X(x_1, \ldots, x_d)\, dx_1 \ldots dx_d.$$

We apply the multivariate change of variables theorem for integrals. Let $\mathbf{y} = (y_1, \ldots, y_d) = \phi(\mathbf{x})$, so the inverse transformation is $\mathbf{x} = \phi^{-1}(\mathbf{y})$. The differential volume element $dx_1 \ldots dx_d$ transforms as:

$$dx_1 \ldots dx_d = \left| \det \left(D\phi^{-1}(\mathbf{y}) \right) \right| dy_1 \ldots dy_d,$$

where $D\phi^{-1}(\mathbf{y})$ is the Jacobian matrix of the inverse transformation:

$$D\phi^{-1}(\mathbf{y}) = \begin{pmatrix} \frac{\partial \phi_1^{-1}}{\partial y_1} & \cdots & \frac{\partial \phi_1^{-1}}{\partial y_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial \phi_d^{-1}}{\partial y_1} & \cdots & \frac{\partial \phi_d^{-1}}{\partial y_d} \end{pmatrix} (\mathbf{y}).$$

Substituting this into the integral and transforming the domain from $\phi^{-1}(B)$ to B, we get:

$$P(\mathbf{Y} \in B) = \int_B g_X(\phi^{-1}(\mathbf{y})) \left| \det \left(D\phi^{-1}(\mathbf{y}) \right) \right| dy_1 \ldots dy_d.$$

Comparing this to the definition of the joint PDF of $\mathbf{Y}$, $P(\mathbf{Y} \in B) = \int_B g_Y(\mathbf{y})\, d\mathbf{y}$, we can identify the density of $\mathbf{Y}$ as:

$$g_Y(\mathbf{y}) = g_X(\phi^{-1}(\mathbf{y})) \left| \det \left(D\phi^{-1}(\mathbf{y}) \right) \right|.$$

Example 27. *Let $X_1 \sim Exp(\lambda)$ and $X_2 \sim Exp(\lambda)$ be two independent and identically distributed (i.i.d.) random variables. We want to compute the density of $Y_1 = X_1 + X_2$*

Since X_1 and X_2 are independent, their joint PDF is the product of their marginal PDFs:

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) = (\lambda e^{-\lambda x_1})(\lambda e^{-\lambda x_2}) = \lambda^2 e^{-\lambda(x_1+x_2)} \tag{3.2}$$

The support region is $\mathcal{X} = \{(x_1, x_2) : x_1 > 0, x_2 > 0\}$.

We define the transformation $\mathbf{Y} = \phi(\mathbf{X})$ to isolate the sum:

$$Y_1 = X_1 + X_2 \quad (\textit{The sum of interest})$$
$$Y_2 = X_2 \quad (\textit{Auxiliary variable})$$

Inverse Transformation $\mathbf{X} = \phi^{-1}(\mathbf{Y})$

We solve for X_1 and X_2 in terms of Y_1 and Y_2:

$$X_1 = Y_1 - Y_2$$
$$X_2 = Y_2$$

Calculating the Jacobian $|D\phi^{-1}|$: We compute the partial derivatives of the inverse transformation:

$$\frac{\partial \phi_1^{-1}}{\partial y_1} = 1, \quad \frac{\partial \phi_1^{-1}}{\partial y_2} = -1$$
$$\frac{\partial \phi_2^{-1}}{\partial y_1} = 0, \quad \frac{\partial \phi_2^{-1}}{\partial y_2} = 1$$

The Jacobian determinant is:

$$|D\phi^{-1}| = \det \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = (1)(1) - (-1)(0) = 1$$

Joint Density f_{Y_1,Y_2}: Applying the formula $f_{\mathbf{Y}}(y_1, y_2) = f_{\mathbf{X}}(x_1, x_2) \cdot |D\phi^{-1}|$. Substitute $x_1 = y_1 - y_2$ and $x_2 = y_2$ into the initial density:

$$f_{Y_1,Y_2}(y_1, y_2) = \lambda^2 e^{-\lambda((y_1-y_2)+y_2)} \cdot 1 = \lambda^2 e^{-\lambda y_1}$$

The original support conditions are $x_1 > 0$ and $x_2 > 0$. Substituting the inverse relations:

1. $x_2 > 0 \implies Y_2 > 0$
2. $x_1 > 0 \implies Y_1 - Y_2 > 0 \implies Y_2 < Y_1$

Since $X_1, X_2 > 0$, the sum Y_1 must also be positive, $Y_1 > 0$.

The support region of (Y_1, Y_2) is $\{(y_1, y_2) : 0 < y_2 < y_1, \quad y_1 > 0\}$

Now, we find the marginal density of $Y_1 = X_1 + X_2$ by integrating $f_{Y_1,Y_2}(y_1, y_2)$ with respect to y_2 over its range, from 0 to y_1:

$$f_{Y_1}(y_1) = \int_0^{y_1} f_{Y_1,Y_2}(y_1, y_2) dy_2 = \int_0^{y_1} \lambda^2 e^{-\lambda y_1} dy_2$$

Since $\lambda^2 e^{-\lambda y_1}$ is constant with respect to y_2:

$$f_{Y_1}(y_1) = \lambda^2 e^{-\lambda y_1} \int_0^{y_1} 1 dy_2 = \lambda^2 e^{-\lambda y_1} [y_2]_0^{y_1}$$

$$\mathbf{f_{Y_1}(y_1) = y_1 \lambda^2 e^{-\lambda y_1}}, \quad \textit{for } \mathbf{y_1 > 0} \tag{3.3}$$

This is the PDF of a Gamma distribution with parameters $\alpha = 2$ and scale parameter $\theta = 1/\lambda$, denoted Gamma$(2, \lambda)$.

3.7 Distribution of the Minimum and Maximum of I.I.D. Variables

In the context of transforming random variables, there are particular cases of great importance in statistics where the calculation of the resulting distribution can be significantly simplified. The general multivariate change-of-variables theorem is powerful, but it is often unnecessary when the transformation takes the form of order statistics, such as the maximum ($Y_{\max}$) or the minimum ($Y_{\min}$).

Next, we will explore how the distribution of the maximum and the minimum of independent and identically distributed (i.i.d.) variables can be derived in a direct and elegant way using only the common Cumulative Distribution Function (CDF), without resorting to the general multivariate change-of-variables theorem.

Theorem 21. *Let $X_1, X_2, \ldots, X_n$ be a sequence of n independent and identically distributed (i.i.d.) random variables. Assume they share a common Cumulative Distribution Function (CDF) $F_X(x)$ and a common Probability Density Function (PDF) $f_X(x)$.*

Let $Y_{\min}$ be the minimum of the variables: $Y_{\min} = \min(X_1, X_2, \ldots, X_n)$. Let $Y_{\max}$ be the maximum of the variables: $Y_{\max} = \max(X_1, X_2, \ldots, X_n)$.

The CDF and PDF for $Y_{\max}$ and $Y_{\min}$ are given by:

$$F_{Y_{\max}}(y) = [F_X(y)]^n$$
$$f_{Y_{\max}}(y) = n[F_X(y)]^{n-1} f_X(y)$$

$$F_{Y_{\min}}(y) = 1 - [1 - F_X(y)]^n$$
$$f_{Y_{\min}}(y) = n[1 - F_X(y)]^{n-1} f_X(y)$$

Proof.

The CDF of $Y_{\max}$ is the probability that the maximum value is less than or equal to y. This occurs if and only if *all* X_i are less than or equal to y.

$$\begin{aligned} F_{Y_{\max}}(y) &= P(Y_{\max} \leq y) \\ &= P(X_1 \leq y, X_2 \leq y, \ldots, X_n \leq y) \end{aligned}$$

Since $X_1, \ldots, X_n$ are independent:

$$\begin{aligned} F_{Y_{\max}}(y) &= P(X_1 \leq y)P(X_2 \leq y) \cdots P(X_n \leq y) \\ &= F_X(y)F_X(y) \cdots F_X(y) \\ &= [F_X(y)]^n \end{aligned}$$

The PDF is obtained by differentiating the CDF with respect to y, applying the chain rule:

$$f_{Y_{\max}}(y) = \frac{d}{dy} F_{Y_{\max}}(y) = \frac{d}{dy}[F_X(y)]^n = n[F_X(y)]^{n-1} f_X(y)$$

The CDF of $Y_{\min}$ is easiest to find by considering the complementary event: $Y_{\min} > y$. The minimum value is greater than y if and only if *all* X_i are greater than y.

$$P(Y_{\min} > y) = P(X_1 > y, X_2 > y, \ldots, X_n > y)$$

Since $X_1, \ldots, X_n$ are independent:

$$\begin{aligned} P(Y_{\min} > y) &= P(X_1 > y)P(X_2 > y) \cdots P(X_n > y) \\ &= [1 - F_X(y)][1 - F_X(y)] \cdots [1 - F_X(y)] \\ &= [1 - F_X(y)]^n \end{aligned}$$

Therefore, the CDF is:

$$F_{Y_{\min}}(y) = 1 - P(Y_{\min} > y) = 1 - [1 - F_X(y)]^n$$

The PDF is obtained by differentiating the CDF with respect to y:

$$\begin{aligned} f_{Y_{\min}}(y) &= \frac{d}{dy}\left(1 - [1 - F_X(y)]^n\right) \\ &= -n[1 - F_X(y)]^{n-1} \cdot \frac{d}{dy}(1 - F_X(y)) \\ &= -n[1 - F_X(y)]^{n-1} \cdot (-f_X(y)) \\ &= n[1 - F_X(y)]^{n-1} f_X(y) \end{aligned}$$

Chapter 4

Expected Value

The concept of the expected value, or mean, is a fundamental pillar of probability theory and statistics. It represents the "long-run" average value of a random variable and provides a central measure of its distribution.

Before formally defining the Expected Value $E[X]$, let's consider what single number best represents the "long-term average" result of a random experiment. In random phenomena, not all outcomes are equally likely. Therefore, simply calculating the arithmetic mean of the possible results is misleading. To find a true average, we must weight each possible outcome by its probability of occurrence*

This weighted sum gives us the value we expect to see, on average, if the experiment is repeated a large number of times.

Consider a Discrete Random Variable X (e.g., a gain in a game) which can take only three specific values, each with a given probability:

Value of X (x_i)	**Probability $P(X = x_i)$**	**Contribution ($x_i \cdot P(X = x_i)$)**
1	0.20	$1 \cdot 0.20 = 0.20$
3	0.50	$3 \cdot 0.50 = 1.50$
5	0.30	$5 \cdot 0.30 = 1.50$

To find the expected average, we calculate the sum of the contributions:

$$\text{Representative Value} = (1 \cdot 0.20) + (3 \cdot 0.50) + (5 \cdot 0.30) = 3.20$$

This value, **3.20**, is what we formally call the Expected Value, $E[X]$. This calculation has a powerful physical interpretation: The Expected Value $E[X]$ is the Center of Mass (or center of gravity) of the probability distribution.

Imagine placing the possible values (1, 3, 5) on a horizontal beam. The probabilities (0.20, 0.50, 0.30) represent the weights (masses) placed at those points. The calculation of 3.20 tells us the exact point where a fulcrum must be placed to perfectly balance the entire beam.

In our example, even though $X = 3$ has the highest probability (50% of the mass), the heavier weight placed on the larger value ($X = 5$) pulls the point of equilibrium to the right, landing at $E[X] = 3.20$.

The most common definitions of the expected value for a continuous random variable involve an integral of the product of the variable's value and its probability density function. However, an alternative and equally valid definition can be formulated using the cumulative distribution function. This approach is particularly useful in theoretical contexts and offers a different perspective on the expectation. We begin with a general definition based on the cumulative distribution function and then demonstrate its relationship to the more commonly known integral definition involving the probability density function for continuous random variables.

Definition 25. *Let (Ω, Σ, P) be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable with the cumulative distribution function $F_X(t) = P(X \leq t)$. The **mean value** or **expected value** of X, denoted by $E(X)$, is given by*

$$E(X) = -\int_{-\infty}^{0} F_X(t)\,dt + \int_{0}^{\infty} (1 - F_X(t))\,dt. \tag{4.1}$$

*This expression is sometimes referred to as the **tail-integral formula** for the expectation.*

Remark 15. *The existence of $E(X)$ depends on the finiteness of the integrals in the sum (4.1).*

1. *If both integrals, $\int_{-\infty}^{0} F_X(t)\,dt$ and $\int_{0}^{\infty}(1-F_X(t))\,dt$, are finite, then $E(X)$ exists and is a finite number.*
2. *If at least one of these integrals diverges (i.e., is infinite), then the expected value $E(X)$ is not well-defined.*

For continuous random variables, the tail-integral formula for the expected value can be shown to be equivalent to the more familiar definition involving the probability density function.

Theorem 22. *Let X be a continuous random variable with a probability density function $g(t)$. If $\int_{-\infty}^{+\infty} |t| g(t)\,dt < \infty$, then the expected value $E(X)$ exists and is given by*

$$E(X) = \int_{-\infty}^{+\infty} t \cdot g(t)\,dt.$$

Proof. We will prove this by showing that the two terms of the tail-integral formula (4.1) correspond to the two parts of the integral over the density function, specifically for $t < 0$ and $t \geq 0$.

For the first term, we have

$$\begin{aligned}
-\int_{-\infty}^{0} F_X(t)\,dt &= -\int_{-\infty}^{0} \left(\int_{-\infty}^{t} g(u)\,du \right) dt = -\int_{-\infty}^{0} \int_{-\infty}^{t} g(u)\,du\,dt \\
&= -\int_{-\infty}^{0} \int_{u}^{0} g(u)\,dt\,du \quad \text{(changing the order of integration)} \\
&= -\int_{-\infty}^{0} g(u) \left(\int_{u}^{0} dt \right) du = -\int_{-\infty}^{0} g(u)[t]_{u}^{0}\,du \\
&= -\int_{-\infty}^{0} g(u)(0-u)\,du = \int_{-\infty}^{0} u \cdot g(u)\,du.
\end{aligned}$$

The change of the order of integration is justified by Fubini's theorem, since we assume the absolute integral of the density function multiplied by the variable is finite.

For the second term, we proceed similarly:

$$\begin{aligned}\int_0^\infty (1 - F_X(t))\,dt &= \int_0^\infty \left(1 - \int_{-\infty}^t g(u)\,du\right) dt = \int_0^\infty \left(\int_t^\infty g(u)\,du\right) dt \\ &= \int_0^\infty \int_t^\infty g(u)\,du\,dt \\ &= \int_0^\infty \int_0^u g(u)\,dt\,du \quad \text{(changing the order of integration)} \\ &= \int_0^\infty g(u)\left(\int_0^u dt\right) du = \int_0^\infty g(u)[t]_0^u\,du \\ &= \int_0^\infty g(u)(u-0)\,du = \int_0^\infty u \cdot g(u)\,du.\end{aligned}$$

Combining the two results, we get

$$E(X) = \int_{-\infty}^0 t \cdot g(t)\,dt + \int_0^\infty t \cdot g(t)\,dt = \int_{-\infty}^{+\infty} t \cdot g(t)\,dt.$$

This completes the proof.

Example 28. *Consider a variable X that is uniformly distributed over the interval $[0, 4]$. Its Probability Density Function (PDF) is $f(x) = \frac{1}{4}$ for $0 \le x \le 4$, and 0 otherwise.*

$$E[X] = \int_0^4 x \cdot \frac{1}{4} dx = \frac{1}{4}\left[\frac{x^2}{2}\right]_0^4 = \frac{1}{4}\left(\frac{4^2}{2} - \frac{0^2}{2}\right) = \frac{8}{4} = 2$$

Since the distribution is perfectly uniform and symmetric between 0 and 4, the Center of Mass (Expected Value) is precisely the midpoint, **2**.

Theorem 23. *Let X be a discrete random variable with $S_X = \{x : P(X = x) > 0\} = \{x_1, x_2, \dots\}$ and a probability mass function $p(x)$. If $\sum_{x \in S_X} |x|p(x) < \infty$, then the expected value $E(X)$ exists and is given by*

$$E(X) = \sum_{x \in S_X} xp(x)$$

Proof. With lock of generality we assume that $X > 0$ and let the support of X be $S_X = \{x_1, x_1, \dots\}$ with $\dots 0 < x_1 < x_2 < \dots$. Then

$$E(X) = -\int_{-\infty}^0 F_X(t)\,dt + \int_0^{-\infty} (1 - F_X(t))\,dt$$

We can split the integral into a sum over the intervals between consecutive values. Where we define $x_0 = 0$. In each interval $[x_{i-1}, x_i)$, the CDF $F_X(t)$ is constant and equal to $F_X(x_{i-1})$.

$$
\begin{aligned}
E(X) &= -\int_{-\infty}^{0} F_X(t)\,dt + \int_{0}^{-\infty} (1 - F_X(t))\,dt \\
&= 0 + \sum_{i=0}^{+\infty} \int_{x_i}^{x_{i+1}} (1 - F_X(t)) = \sum_{i=0}^{+\infty} (x_{i+1} - x_i)(1 - F_X(x_i)) \\
&= \sum_{i=0}^{+\infty} x_{i+1}(F_X(x_{i+1}) - F_X(x_i)) = \sum_{i=1}^{+\infty} x_i(F_X(x_i) - F_X(x_{i-1})) \\
&= \sum_{i=1}^{+\infty} x_i p(x_i)
\end{aligned}
$$

Fact 7. *Let $X : \Omega \to R$ be a random variable. Then for all $B \in \mathcal{B}(\mathbb{R})$ we have*

$$P(X \in B) = E(\mathbf{1}_{X^{-1}(B)})$$

In particular $P(X \leq a) = E(\mathbf{1}_{X^{-1}(-\infty,a]})$, $P(X = a) = E(\mathbf{1}_{X^{-1}\{a\}})$ and $P(X \geq a) = E(\mathbf{1}_{X^{-1}[a,+\infty)})$

Proof. By definition, the indicator random variable $\mathbf{1}_{X^{-1}(B)}$ is given by:

$$\mathbf{1}_{X^{-1}(B)}(\omega) = \begin{cases} 1 & \text{if } \omega \in X^{-1}(B) \\ 0 & \text{if } \omega \notin X^{-1}(B) \end{cases}$$

The condition $\omega \in X^{-1}(B)$ is equivalent to $X(\omega) \in B$. Therefore, the indicator variable can be written as:

$$\mathbf{1}_{X^{-1}(B)}(\omega) = \begin{cases} 1 & \text{if } X(\omega) \in B \\ 0 & \text{if } X(\omega) \notin B \end{cases}$$

By the definition of the expected value of a discrete random variable, we have:

$$E\left[\mathbf{1}_{X^{-1}(B)}\right] = 1 \cdot P\left(\mathbf{1}_{X^{-1}(B)} = 1\right) + 0 \cdot P\left(\mathbf{1}_{X^{-1}(B)} = 0\right).$$

The event $\left\{\omega \mid \mathbf{1}_{X^{-1}(B)}(\omega) = 1\right\}$ is the set $\{\omega \mid X(\omega) \in B\}$. Thus,

$$P\left(\mathbf{1}_{X^{-1}(B)} = 1\right) = P(X \in B).$$

Substituting this back, we obtain the desired result:

$$E\left[\mathbf{1}_{X^{-1}(B)}\right] = 1 \cdot P(X \in B) = P(X \in B).$$

Fact 8. *Let X and Y be random variables such that $E[X]$ and $E[Y]$ exist. Then for any constants $a, b \in \mathbb{R}$, the following properties hold:*

1. *$E[b] = b$, where b is a constant random variable (i.e., $P(X = b) = 1$).*
2. *$E[aX] = aE[X]$.*
3. *$E[X + Y] = E[X] + E[Y]$ (Linearity of Expectation).*
4. *If $X \geq 0$, then $E[X] \geq 0$. Consequently, if $X \geq Y$, then $E[X] \geq E[Y]$.*
5. *If $X > 0$, then $E[X] > 0$.*
6. *$E[|X|] \geq |E[X]|$ (Triangle Inequality for Expectation).*

Proof. We will prove each property in turn, considering both discrete and continuous cases where appropriate.

1. For a discrete random variable X with $P(X = b) = 1$, the expected value is simply the sum of all possible outcomes multiplied by their probabilities. In this case, there is only one outcome:
$$E[b] = b \cdot P(X = b) = b \cdot 1 = b.$$

2. For a constant $a \neq 0$, let $Z = aX$. The CDF of Z is $F_Z(z) = P(Z \leq z) = P(aX \leq z)$.
If $a > 0$, then $F_Z(z) = P(X \leq z/a) = F_X(z/a)$.
If $a < 0$, then $F_Z(z) = P(X \geq z/a) = 1 - F_X(()z/a)^-)$.
We do calculation only for $a > 0$
The expected value is given by the integral formula:
$$E[aX] = E[Z] = \int_0^\infty (1 - F_X(t/a))\, dt - \int_{-\infty}^0 F_X(t/a)\, dt.$$
Let $x = t/a$, so $t = ax$ and $dt = a\, dx$.
$$E[aX] = \int_0^\infty (1 - F_X(x))a\, dx - \int_{-\infty}^0 F_X(x)a\, dx$$
$$= a\left(\int_0^\infty (1 - F_X(x))\, dx - \int_{-\infty}^0 F_X(x)\, dx\right) = aE[X].$$

3. We prove this for a continuous random variable. Let $Z = X + Y$ with joint density $g(x, y)$.
$$E[X + Y] = E[Z] = \int_{-\infty}^\infty z\, g_Z(z)\, dz = \int_{-\infty}^\infty z\left(\int_{-\infty}^\infty g(t, z - t)\, dt\right) dz.$$
We rearrange the integral using a change of variables (t, y) where $y = z - t$, and then swap the order of integration (Fubini's Theorem):
$$E[X + Y] = \int_{-\infty}^\infty \int_{-\infty}^\infty (t + y)\, g(t, y)\, dt\, dy = \int_{-\infty}^\infty \int_{-\infty}^\infty t\, g(t, y)\, dt\, dy + \int_{-\infty}^\infty \int_{-\infty}^\infty y\, g(t, y)\, dt\, dy.$$
The first term is $E[X]$ and the second is $E[Y]$, so $E[X + Y] = E[X] + E[Y]$. The proof for discrete variables is similar, using sums instead of integrals.

4. This follows from the definition of expectation. If $X \geq 0$,
$$E(X) = \int_0^\infty (1 - F_X(t))\, dt \geq 0$$

5. If $X > 0$, $F_X(0) = 0$ and since F_X is right continuous, there exist $\delta > 0$ such that $F_X(t) < \frac{1}{2}$ when $t \in [0, \delta)$. Thus
$$E(X) = \int_0^\infty (1 - F_X(t))\, dt \geq \frac{\delta}{2} > 0$$

6. For any random variable X, we have the inequality $-|X| \leq X \leq |X|$. Using property 4, we can take the expectation of this inequality:

$$E[-|X|] \leq E[X] \leq E[|X|].$$

By property 2, $E[-|X|] = -E[|X|]$, so we have:

$$-E[|X|] \leq E[X] \leq E[|X|].$$

This is the definition of the inequality $|E[X]| \leq E[|X|]$.

When a random variable X is transformed by a function f, a new random variable $Y = f(X)$ is created. A natural question that arises is how to compute the expected value of this new variable, $E(Y)$. The naive approach would be to first determine the probability distribution of Y, $F_Y(y)$, and then use the standard definition of expectation (4.1). However, finding the distribution of Y can often be a complicated task.

The Law of the Unconscious Statistician provides a powerful shortcut. It states that the expected value of $Y = f(X)$ can be computed directly using the original distribution of X, without any reference to the distribution of Y. This theorem simplifies calculations significantly and is a cornerstone of advanced probability and statistical theory.

Theorem 24 (Law of the Unconscious Statistician)**.** *Let $X : \Omega \to \mathbb{R}$ be a random variable and let $f : \mathbb{R} \to \mathbb{R}$ be a Borel measurable function. Let $Y = f(X)$. If $E(|f(X)|)$ is finite, then the expected value of Y can be computed directly using the probability distribution of X:*

a) If X is a continuous random variable with probability density function $g_X(x)$, then

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

b) If X is a discrete random variable with a countable set of possible values S_X and probability mass function $P(X = x)$, then

$$E(f(X)) = \sum_{x \in S_X} f(x) P(X = x).$$

Proof. We provide a proof for both the continuous and discrete cases.

Case (a): Continuous Random Variable Let's start with the definition of expectation for the random variable $Y = f(X)$ using its cumulative distribution function $F_Y(y)$.

$$E(Y) = -\int_{-\infty}^{0} F_Y(y)\, dy + \int_{0}^{\infty} (1 - F_Y(y))\, dy.$$

For a continuous random variable X with density $g_X(x)$, these probabilities can be expressed as integrals over the density function. Let $A_y^- = \{x \in \mathbb{R} \mid f(x) \leq y\}$ and $A_y^+ = \{x \in \mathbb{R} \mid f(x) > y\}$.

$$E(Y) = -\int_{-\infty}^{0} P(f(X) \leq y)\, dy + \int_{0}^{\infty} P(f(X) > y)\, dy$$

$$= -\int_{-\infty}^{0} \left(\int_{A_y^-} g_X(x)\, dx \right) dy + \int_{0}^{\infty} \left(\int_{A_y^+} g_X(x)\, dx \right) dy.$$

Assuming the integrals exist and the conditions for Fubini's theorem hold (which is guaranteed by the assumption that $E(|f(X)|)$ is finite), we can change the order of integration. This is a crucial step. The regions of integration must be carefully re-evaluated. The first double integral is over the region $\{(x, y) \in \mathbb{R} \times (-\infty, 0] \mid f(x) \leq y\}$. When we change the order, the inner integral is with respect to y, and its bounds are determined by x. If $f(x) \leq 0$, the y values range from $f(x)$ up to 0. If $f(x) > 0$, there are no such y.

$$\int_{-\infty}^{0} \int_{A_y^-} g_X(x)\,dx\,dy = \int_{-\infty}^{\infty} \left(\int_{\max(f(x),-\infty)}^{0} \mathbb{I}(f(x) \leq 0)\,dy \right) g_X(x)\,dx$$

$$= \int_{\{x:f(x)\leq 0\}} \left(\int_{f(x)}^{0} dy \right) g_X(x)\,dx = \int_{\{x:f(x)\leq 0\}} -f(x)g_X(x)\,dx.$$

Thus,

$$-\int_{-\infty}^{0} P(f(X) \leq y)\,dy = -\int_{\{x:f(x)\leq 0\}} -f(x)g_X(x)\,dx = \int_{\{x:f(x)\leq 0\}} f(x)g_X(x)\,dx.$$

Similarly, for the second term, the region is $\{(x, y) \in \mathbb{R} \times [0, \infty) \mid f(x) > y\}$.

$$\int_{0}^{\infty} \int_{A_y^+} g_X(x)\,dx\,dy = \int_{-\infty}^{\infty} \left(\int_{0}^{\infty} \mathbb{I}(f(x) > y)\,dy \right) g_X(x)\,dx$$

$$= \int_{\{x:f(x)>0\}} \left(\int_{0}^{f(x)} dy \right) g_X(x)\,dx = \int_{\{x:f(x)>0\}} f(x)g_X(x)\,dx.$$

Combining the two results gives:

$$E(f(X)) = \int_{\{x:f(x)\leq 0\}} f(x)g_X(x)\,dx + \int_{\{x:f(x)>0\}} f(x)g_X(x)\,dx = \int_{-\infty}^{\infty} f(x)g_X(x)\,dx.$$

This completes the proof for the continuous case. An alternative, more general proof can be given using measure theory and the change of variables formula for integrals.

Case (b): Discrete Random Variable Let S_X be the countable set of values that X can take, and $S_Y = \{y \mid P(Y = y) > 0\} = \{f(x) \mid x \in S_X, P(X = x) > 0\}$. The definition of expectation for a discrete random variable Y is:

$$E(Y) = \sum_{y \in S_Y} yP(Y = y).$$

We know that the probability of Y taking a specific value y is the sum of the probabilities of all the x values for which $f(x) = y$:

$$E(Y) = \sum_{y \in S_Y} y \left(\sum_{x \in S_X, f(x)=y} P(X = x) \right).$$

Now, we replace the value y with its equivalent representation $f(x)$ for each term in the inner sum, since for those x values, $f(x) = y$:

$$E(Y) = \sum_{y \in S_Y} \sum_{x \in S_X, f(x)=y} f(x)P(X = x).$$

This double summation can be re-ordered. The inner sum groups all x values that map to the same y. The outer sum then sums over all possible y. This is equivalent to a single summation over all possible x values, where each x is mapped to its corresponding $f(x)$:

$$E(Y) = \sum_{x \in S_X} f(x)P(X = x).$$

This concludes the proof for the discrete case.

Theorem 25. *Let X and Y be two independent random variables. If the expectations $E(|X|)$ and $E(|Y|)$ are finite, then the expectation of their product $E(X \cdot Y)$ exists and is equal to the product of their individual expectations, i.e., $E(X \cdot Y) = E(X) \cdot E(Y)$.*

Proof. We will prove this theorem for both continuous and discrete random variables.

Case of Continuous Random Variables Let X and Y be continuous random variables with probability density functions (PDFs) $g_X(x)$ and $g_Y(y)$, respectively. Since X and Y are independent, their joint PDF is the product of their marginal PDFs: $g_{X,Y}(x, y) = g_X(x)g_Y(y)$.

Let $Z = XY$, for continuos random variables $g_Z(z) = \int_{-\infty}^{\infty} g_X(x)g_Y(\frac{z}{x})\frac{1}{|x|}dx$

$$E(X.Y) = E(Z) = \int_{-\infty}^{\infty} z g_z(z)\, dz = \int_{-\infty}^{\infty} z \int_{-\infty}^{\infty} g_X(x)g_Y(\frac{z}{x})\frac{1}{|x|}dxdz$$

changing the variable $y = \frac{z}{x}$,

$$E(Z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy g_X(x)g_Y(y)dxdy = \left(\int_{-\infty}^{\infty} x g_X(x)\, dx\right)\left(\int_{-\infty}^{\infty} y g_Y(y)\, dy\right) = E(X) \cdot E(Y).$$

Case of Discrete Random Variables is similar

Remark 16. *It is very difficult to prove the Theorem to the general definition. The converse of this theorem is not generally true. That is, if $E(XY) = E(X)E(Y)$, it does not necessarily imply that X and Y are independent. A counterexample can be constructed where the variables are dependent but their product's expectation still factorizes.*

The concept of expectation can be naturally extended from a single random variable to a vector of random variables, often called a random vector.

Definition 26. *Given a d-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)$, its expected value, denoted by $E(\mathbf{X})$, is a vector in $\mathbb{R}^d$ whose components are the expected values of the individual random variables:*

$$E(\mathbf{X}) = (E(X_1), E(X_2), \ldots, E(X_d)).$$

This definition allows us to extend many properties of expectation from one dimension to multiple dimensions.

Fact 9. *The expected value of a random vector can be computed using its joint probability distribution.*

1. *For a continuous random vector $\mathbf{X}$ with joint probability density function $g_{\mathbf{X}}(x_1, \ldots, x_d)$:*

$$E(\mathbf{X}) = \int_{\mathbb{R}^d} \mathbf{x} g_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} = \left(\int_{\mathbb{R}^d} x_1 g_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}, \ldots, \int_{\mathbb{R}^d} x_d g_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}\right).$$

2. *For a discrete random vector $\mathbf{X}$ with joint probability mass function $P(\mathbf{X} = \mathbf{x})$:*

$$E(\mathbf{X}) = \sum_{\mathbf{x} \in S_{\mathbf{X}}} \mathbf{x} P(\mathbf{X} = \mathbf{x}) = \left(\sum_{\mathbf{x} \in S_{\mathbf{X}}} x_1 P(\mathbf{X} = \mathbf{x}), \ldots, \sum_{\mathbf{x} \in S_{\mathbf{X}}} x_d P(\mathbf{X} = \mathbf{x})\right).$$

Note that the integral and sum above are over the entire support of the random vector.

Fact 10 (Linearity of Expectation for Random Vectors)**.** *For two d-dimensional random vectors $\mathbf{X}$ and $\mathbf{Y}$ and scalars $a, b \in \mathbb{R}$, the linearity of expectation still holds:*

$$E(a\mathbf{X} + b\mathbf{Y}) = aE(\mathbf{X}) + bE(\mathbf{Y}).$$

Fact 11 (Multidimensional Law of the Unconscious Statistician)**.** *For any Borel measurable function $f : \mathbb{R}^d \to \mathbb{R}$, if $E(|f(\mathbf{X})|)$ is finite, its expected value can be calculated using the joint distribution of $\mathbf{X}$:*

1. *For a continuous random vector $\mathbf{X}$:*

$$E(f(\mathbf{X})) = \int_{\mathbb{R}^d} f(x_1, \ldots, x_d) g_{\mathbf{X}}(x_1, \ldots, x_d)\, dx_1 \ldots dx_d.$$

2. *For a discrete random vector $\mathbf{X}$:*

$$E(f(\mathbf{X})) = \sum_{x_1 \in S_{X_1}} \cdots \sum_{x_d \in S_{X_d}} f(x_1, \ldots, x_d) P(X_1 = x_1, \ldots, X_d = x_d).$$

4.1 Variance

The Variance $\mathrm{Var}(X)$ is the core metric used to understand the spread or dispersion of a random variable. It quantifies the uncertainty by measuring how far results tend to fall from the Mean, or Expected Value (μ).

We will analyze three discrete random variables, X_1, X_2, and X_3, which all share the same Expected Value.

Case 1: X_1

The extreme values of -1 and 1 have significant probabilities (1/4), balancing the central probability (1/2).

Table 4.1: Distribution X_1

Value (x)	**Probability** $P(X = x)$
-1	1/4
0	1/2
1	1/4

Case 2: X_2

The value range is identical to X_1 (from -1 to 1), but the probability is heavily concentrated at the mean (0.98 at 0), making the extreme outcomes very rare. This is a case of **low variance due to concentrated probability**.

Table 4.2: Distribution X_2

Value (x)	**Probability** $P(X = x)$
-1	0.01
0	0.98
1	0.01

Table 4.3: Distribution X_3

Value (x)	**Probability** $P(X = x)$
-0.1	1/4
0	1/2
0.1	1/4

Case 3: X_3

This variable has the same probability distribution shape as X_1 $(1/4, 1/2, 1/4)$, but the range of values is much smaller (± 0.1). This is a case of **low variance due to a small range of values**.

In all cases, the Expectation is zero because of symmetry around zero:

$$E[X] = \sum x \cdot P(X = x) = (-1)(p_{-1}) + (0)(p_0) + (1)(p_1) = 0$$

However, the example show diferent level of dispersion arround the mean.

Definition 27 (Variance and Standard Deviation)**.** *The* ***variance*** *of a one-dimensional random variable X is defined as the expected squared deviation from its mean:*

$$Var(X) = E\left[(X - E[X])^2\right].$$

An alternative and often more convenient formula for calculation is:

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

The positive square root of the variance, $\sqrt{Var(X)}$, is called the ***standard deviation****.*

In the previous variables we can compute the variance:

$$\text{Var}(X_1) = E[X_1^2] = (-1)^2 \cdot \frac{1}{4} + (0)^2 \cdot \frac{1}{2} + (1)^2 \cdot \frac{1}{4} = 1 \cdot \frac{1}{4} + 0 + 1 \cdot \frac{1}{4} = \mathbf{0.5}$$

$$\text{Var}(X_2) = E[X_2^2] = (-1)^2 \cdot 0.01 + (0)^2 \cdot 0.98 + (1)^2 \cdot 0.01 = 1 \cdot 0.01 + 0 + 1 \cdot 0.01 = 0.02$$

Despite having large extreme values (up to ± 1), concentrating 98% of the probability on the mean (0) results in a very low variance: $\text{Var}(\mathbf{X_2}) = \mathbf{0.02}$.

$$\text{Var}(X_3) = E[X_3^2] = (-0.1)^2 \cdot \frac{1}{4} + (0)^2 \cdot \frac{1}{2} + (0.1)^2 \cdot \frac{1}{4} = 0.01 \cdot 0.25 + 0 + 0.01 \cdot 0.25 = 0.005$$

Having the exact same probability structure as X_1, the reduced distance of the values from the mean (0.1 vs 1) drastically lowers the variance, making it the smallest: $\text{Var}(\mathbf{X_3}) = \mathbf{0.005}$.

Table 4.4: Variance Results Comparison

Variable	**Mean** ($E[X]$)	**Variance** (**Var**(X))
X_1	0	0.500
X_2	0	0.020
X_3	0	0.005

Variance successfully distinguishes between distributions that, despite sharing the same mean, represent vastly different levels of dispersion . X_3 shows that reducing the absolute difference between the possible outcomes and the mean significantly lowers the variance. X_2 shows that even if extreme values are large, concentrating the probability mass near the mean will result in a low variance.

Definition 28 (Covariance Matrix)**.** *The **covariance matrix** of a d-dimensional random variable* $\mathbf{X} = (X_1, \ldots, X_d)$ *is a square matrix* $\mathbf{C} \in \mathbb{R}^{d \times d}$ *whose elements are the covariances between each pair of its components. The element* c_{ij} *at the i-th row and j-th column is the covariance of* X_i *and* X_j*:*

$$c_{ij} = Cov(X_i, X_j) = E\left[(X_i - E[X_i])(X_j - E[X_j])\right].$$

An equivalent formula for calculating covariance is:

$$Cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j].$$

The diagonal elements of the covariance matrix, c_{ii}*, correspond to the variance of each component:* $c_{ii} = Cov(X_i, X_i) = Var(X_i)$*.*

Fact 12 (Independence and Covariance)**.** *If two random variables* X *and* Y *are independent, then their covariance is zero. That is,* $Cov(X, Y) = 0$*.*

Proof. By the definition of covariance, we have $\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$. We can expand this expression:

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

The last step uses the linearity of expectation and the fact that the expectation of a constant is the constant itself. Since X and Y are independent, we know that $E[XY] = E[X]E[Y]$. Substituting this into the expression, we get:

$$\text{Cov}(X, Y) = E[X]E[Y] - E[X]E[Y] = 0.$$

Remark 17. *The converse of Fact 12 is **not** true in general. That is, having* $Cov(X, Y) = 0$ *does not necessarily imply that* X *and* Y *are independent. For example,*

$Y \backslash X$	-1	0	1
-1	0.2	0	0.2
0	0	0.2	0
1	0.2	0	0.2

$Cov(X, Y) = 0$*, but they are NOT independent.*

Remark 18. *For random variables with a joint normal (Gaussian) distribution, the converse is true. If two jointly normally distributed random variables have a covariance of zero, they are independent.*

Theorem 26. *For any random variables* X, Y, X_1, X_2 *on a probability space* $(\Omega, \mathcal{F}, P)$ *and any constants* $a, b \in \mathbb{R}$*, the following properties hold:*

1. ***Symmetry:*** $Cov(X, Y) = Cov(Y, X)$*.*
2. ***Bilinearity:*** $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$
3. $Cov(aX + b, Y) = a \cdot Cov(X, Y)$*.*
4. ***Relationship to Variance:*** $Cov(X, X) = Var(X)$*.*

5. ***Positive Semidefiniteness:*** *$Cov(X, X) = Var(X) \geq 0$, with $Var(X) = 0$ if and only if X is a constant random variable (i.e., $P(X = E(X)) = 1$).*

Proof.

1. **Symmetry:** By the definition of covariance,

$$\text{Cov}(X, Y) = E((X - E(X))(Y - E(Y))) = E((Y - E(Y))(X - E(X))) = \text{Cov}(Y, X).$$

   The equality holds due to the commutativity of real number multiplication.

2. **Bilinearity:** We use the linearity of expectation:

$$\begin{aligned}
\text{Cov}(X_1 + X_2, Y) &= E(((X_1 + X_2) - E(X_1 + X_2))(Y - E(Y))) \\
&= E((X_1 + X_2 - E(X_1) - E(X_2))(Y - E(Y))) \\
&= E(((X_1 - E(X_1)) + (X_2 - E(X_2)))(Y - E(Y))) \\
&= E((X_1 - E(X_1))(Y - E(Y))) + E((X_2 - E(X_2))(Y - E(Y))) \\
&= \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y).
\end{aligned}$$

3. We use the linearity of expectation:

$$\begin{aligned}
\text{Cov}(aX + b, Y) &= E(((aX + b) - E(aX + b))(Y - E(Y))) \\
&= E((aX + b - (aE(X) + b))(Y - E(Y))) \\
&= E((aX - aE(X))(Y - E(Y))) \\
&= E(a(X - E(X))(Y - E(Y))) \\
&= aE((X - E(X))(Y - E(Y))) = a \cdot \text{Cov}(X, Y).
\end{aligned}$$

4. **Relationship to Variance:** By the definition of covariance,

$$\text{Cov}(X, X) = E((X - E(X))(X - E(X))) = E((X - E(X))^2) = \text{Var}(X).$$

5. **Positive Semidefiniteness:** First, $\text{Var}(X) = E((X - E(X))^2) \geq 0$ because the random variable $(X - E(X))^2$ is non-negative for all $\omega \in \Omega$, and the expectation of a non-negative random variable is non-negative.

   The condition $\text{Var}(X) = 0$ holds if and only if $E((X - E(X))^2) = 0$. For any non-negative random variable Z, its expectation is zero if and only if $Z = 0$ almost surely. In our case, $Z = (X - E(X))^2$, so $E((X - E(X))^2) = 0 \iff P((X - E(X))^2 = 0) = 1$. This is equivalent to $P(X - E(X) = 0) = 1$, which means $P(X = E(X)) = 1$. This implies that X is a constant random variable with value $E(X)$.

Remark 19. *The bilinearity of covariance, combined with the symmetry property, means that $Cov(X, Y)$ is linear in both of its arguments. This allows us to expand expressions involving sums of random variables inside the covariance operator.*

Remark 20. *The properties of covariance are analogous to those of an inner product on a vector space, where $Cov(X, Y)$ is like an inner product and $Var(X) = Cov(X, X)$ is like the square of the norm.*

Theorem 27 (Variance of a Sum)**.** *Given a set of pairwise independent random variables $X_1, \ldots, X_d$, the variance of their sum is equal to the sum of their individual variances:*

$$Var\left(\sum_{k=1}^{d} X_k\right) = \sum_{k=1}^{d} Var(X_k).$$

Proof. We start with the definition of variance using covariance:

$$\mathrm{Var}\left(\sum_{i=1}^{d} X_i\right) = \mathrm{Cov}\left(\sum_{i=1}^{d} X_i, \sum_{j=1}^{d} X_j\right).$$

By the bilinearity of the covariance operator, we can expand the expression:

$$= \sum_{i=1}^{d}\sum_{j=1}^{d} \mathrm{Cov}(X_i, X_j).$$

We can split the double sum into two parts: the diagonal terms where $i = j$ and the off-diagonal terms where $i \neq j$.

$$= \sum_{i=1}^{d} \mathrm{Cov}(X_i, X_i) + \sum_{i\neq j} \mathrm{Cov}(X_i, X_j).$$

By definition, $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$. Since the random variables are pairwise independent for $i \neq j$, Fact 12 states that $\mathrm{Cov}(X_i, X_j) = 0$.

$$= \sum_{i=1}^{d} \mathrm{Var}(X_i) + \sum_{i\neq j} 0 = \sum_{i=1}^{d} \mathrm{Var}(X_i).$$

Fact 13. *For any random variable X and constants $a, b \in \mathbb{R}$, we have $Var(aX + b) = a^2\, Var(X)$.*

Proof. We begin with the definition of variance:

$$\mathrm{Var}(aX + b) = E\left[((aX + b) - E[aX + b])^2\right].$$

Using the linearity of expectation, we know that $E[aX + b] = aE[X] + b$. We substitute this into the equation:

$$= E\left[(aX + b - (aE[X] + b))^2\right] = E\left[(aX - aE[X])^2\right].$$

Factoring out the constant a from the expression inside the square:

$$= E\left[(a(X - E[X]))^2\right] = E\left[a^2(X - E[X])^2\right].$$

Finally, we use the property that constants can be factored out of the expectation operator:

$$= a^2 E\left[(X - E[X])^2\right] = a^2\mathrm{Var}(X).$$

Fact 14 (The Covariance Matrix is Positive Semidefinite)**.** *The covariance matrix of a d-dimensional random variable $\mathbf{X} = (X_1, \ldots, X_d)$ is positive semidefinite.*

Proof. By definition, a matrix C is positive semidefinite if and only if for all vectors $\mathbf{v} \in \mathbb{R}^d$, we have $\mathbf{v}^T C \mathbf{v} \geq 0$. Let C be the covariance matrix of $\mathbf{X}$, with entries $C_{jk} = \text{Cov}(X_j, X_k)$. Let $\mathbf{v} = [v_1, \ldots, v_d]^T \in \mathbb{R}^d$. Consider the random variable $Y = \sum_{j=1}^d v_j X_j$. By the positive semidefiniteness property of variance (part 5 of Theorem 26), we know that $\text{Var}(Y) \geq 0$. Now, we can express $\text{Var}(Y)$ in terms of the covariance matrix:

$$\begin{aligned} 0 \leq \text{Var}(Y) &= \text{Var}\left(\sum_{j=1}^d v_j X_j\right) \\ &= \text{Cov}\left(\sum_{j=1}^d v_j X_j, \sum_{k=1}^d v_k X_k\right) \\ &= \sum_{j=1}^d \sum_{k=1}^d v_j v_k \text{Cov}(X_j, X_k) \\ &= \sum_{j=1}^d \sum_{k=1}^d v_j C_{jk} v_k = \mathbf{v}^T C \mathbf{v}. \end{aligned}$$

Since $\mathbf{v}^T C \mathbf{v} \geq 0$ for any arbitrary vector $\mathbf{v} \in \mathbb{R}^d$, the covariance matrix C is positive semidefinite.

Definition 29 (Correlation Coefficient)**.** *The **correlation coefficient** of two random variables X and Y is a standardized measure of their linear relationship, defined as*

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

This coefficient is defined only if the variances of X and Y are non-zero and finite.

Theorem 28. *The correlation coefficient of any two random variables X and Y satisfies $-1 \leq r_{X,Y} \leq 1$.*

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

$$\begin{aligned} 0 &\leq \text{Cov}(X - tY, X - tY) \\ &= \text{Cov}(X,X) - 2t\text{Cov}(X,Y) + t^2\text{Cov}(Y,Y). \end{aligned}$$

This is a quadratic in t. For the quadratic to be always non-negative, its discriminant must be non-positive. Alternatively, choose $t = \frac{\text{Cov}(X,Y)}{\text{Cov}(Y,Y)}$. Substitute this value of t:

$$\begin{aligned} 0 &\leq \text{Cov}(X,X) - 2\frac{\text{Cov}(X,Y)}{\text{Cov}(Y,Y)}\text{Cov}(X,Y) + \left(\frac{\text{Cov}(X,Y)}{\text{Cov}(Y,Y)}\right)^2 \text{Cov}(Y,Y) \\ &= \text{Cov}(X,X) - 2\frac{(\text{Cov}(X,Y))^2}{\text{Cov}(Y,Y)} + \frac{(\text{Cov}(X,Y))^2}{\text{Cov}(Y,Y)} \\ &= \text{Cov}(X,X) - \frac{(\text{Cov}(X,Y))^2}{\text{Cov}(Y,Y)}. \end{aligned}$$

Thus, $\text{Cov}(X,X)\text{Cov}(Y,Y) \geq (\text{Cov}(X,Y))^2 \implies (r_{X,Y})^2 \leq 1$. This implies $r \in [-1,1]$.

Notice that the equality holds only when $\text{Cov}(X-tY, X-tY) = 0$, which means $X - tY = c \in \mathbb{R}$, which means $X = tY + c$. And $r = 0 \iff \text{Cov}(X, Y) = 0$. Thus for independent X, Y we have $r_{X,Y} = 0$.

Remark 21. *The value of the correlation coefficient provides a clear interpretation of the linear relationship between two random variables:*

- $r_{X,Y} = 1$*: This indicates a* ***perfect positive linear relationship****. If the value of one variable increases, the value of the other variable increases proportionally. The relationship can be expressed as* $Y = aX + b$ *for some constants* $a > 0$ *and* $b \in \mathbb{R}$*.*
- $r_{X,Y} = -1$*: This indicates a* ***perfect negative linear relationship****. As the value of one variable increases, the value of the other decreases proportionally. The relationship can be expressed as* $Y = aX + b$ *for some constants* $a < 0$ *and* $b \in \mathbb{R}$*.*
- $r_{X,Y} = 0$*: This indicates* ***no linear relationship****. It implies that the variables are uncorrelated, meaning there is no tendency for them to vary together in a straight-line pattern. However, a correlation of zero does **not** imply independence. The variables can still be strongly related in a non-linear way (e.g.,* $Y = X^2$*, where the correlation is often zero, but the variables are clearly dependent).*
- $0 < |r_{X,Y}| < 1$*: This indicates a* ***partial linear relationship****. The closer the absolute value of the correlation is to 1, the stronger the linear association between the variables.*

Theorem 29. *Let* $\mathbf{X} \sim N(\mathbf{m}, \mathbf{A})$ *be a d-dimensional random variable with a multivariate normal distribution, where* $\mathbf{m} \in \mathbb{R}^d$ *is the mean vector and* $\mathbf{A} \in \mathbb{R}^{d \times d}$ *is a symmetric, positive definite matrix. Then, the mean vector of* $\mathbf{X}$ *is* $E[\mathbf{X}] = \mathbf{m}$*, and its covariance matrix is* $\mathbf{C} = \mathbf{A}$*. The marginal distribution of each component* X_i *is a univariate normal distribution,* $X_i \sim N(m_i, \sigma_i^2)$*, where* m_i *is the i-th component of* $\mathbf{m}$ *and* $\sigma_i^2 = Var(X_i)$*.*

Proof. We left as an exercise for the reader.

Theorem 30 (Uncorrelated Implies Independent for Normal RVs)**.** *Assume* $\mathbf{X} = (X_1, \ldots, X_d)$ *is a random variable with a multivariate normal distribution* $N(\mathbf{m}, \mathbf{C})$*. The components* $X_1, \ldots, X_d$ *are independent if and only if they are uncorrelated. That is,* $X_1, \ldots, X_d$ *are independent if and only if* $Cov(X_i, X_j) = 0$ *for all* $i \neq j$*.*

Proof. ($\implies$) The forward direction has been proven previously; for any random variables, independence implies zero covariance.

($\impliedby$) We need to show that if $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$, then the joint probability density function (PDF) of $\mathbf{X}$ can be factored into the product of the marginal PDFs of its components.

The covariance matrix $\mathbf{C}$ is given by $C_{ij} = \text{Cov}(X_i, X_j)$. If the components are uncorrelated, then $\mathbf{C}$ is a diagonal matrix, where $C_{ij} = 0$ for $i \neq j$ and $C_{ii} = \text{Var}(X_i) = \sigma_i^2$.

The joint PDF of a multivariate normal distribution is given by:

$$f_{\mathbf{X}}(\mathbf{t}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\mathbf{C})}} e^{-\frac{1}{2}(\mathbf{t}-\mathbf{m})^T \mathbf{C}^{-1}(\mathbf{t}-\mathbf{m})}.$$

Since $\mathbf{C}$ is a diagonal matrix, its inverse $\mathbf{C}^{-1}$ is also a diagonal matrix, with entries $(C^{-1})_{ii} = 1/\sigma_i^2$. The determinant of $\mathbf{C}$ is simply the product of its diagonal entries: $\det(\mathbf{C}) = \prod_{i=1}^{d} \sigma_i^2$. The

quadratic form in the exponent simplifies to a sum:

$$(\mathbf{t}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{t}-\mathbf{m}) = \sum_{i=1}^{d}\frac{(t_i-m_i)^2}{\sigma_i^2}.$$

Substituting these into the joint PDF formula, we get:

$$\begin{aligned} f_{\mathbf{X}}(t_1,\ldots,t_d) &= \frac{1}{(2\pi)^{d/2}\sqrt{\prod_{i=1}^{d}\sigma_i^2}} e^{-\frac{1}{2}\sum_{i=1}^{d}\frac{(t_i-m_i)^2}{\sigma_i^2}} \\ &= \frac{1}{\prod_{i=1}^{d}\sqrt{2\pi\sigma_i^2}}\prod_{i=1}^{d} e^{-\frac{(t_i-m_i)^2}{2\sigma_i^2}} \\ &= \prod_{i=1}^{d}\left(\frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(t_i-m_i)^2}{2\sigma_i^2}}\right). \end{aligned}$$

The last expression is the product of the individual marginal PDFs, since each factor is the PDF of a univariate normal distribution $N(m_i,\sigma_i^2)$. Thus, the joint PDF factors, which proves that the random variables $X_1,\ldots,X_d$ are independent.

4.2 Expectation and Variance of Common Distributions

Theorem 31 (One-Point Distribution)**.** *Let X be a random variable such that $P(X=c)=1$ for some constant $c\in\mathbb{R}$.*

- $E[X]=c$
- $Var(X)=0$

Proof. By definition, the expectation is $E[X]=\sum_x xP(X=x)=c\cdot P(X=c)=c\cdot 1=c$. The variance is $\mathrm{Var}(X)=E[(X-E[X])^2]=E[(X-c)^2]=(c-c)^2P(X=c)=0\cdot 1=0$.

Theorem 32 (Two-Point Distribution (Bernoulli))**.** *Let X be a random variable with $P(X=1)=p$ and $P(X=0)=1-p$.*

- $E[X]=p$
- $Var(X)=p(1-p)$

Proof. $E[X]=\sum_x xP(X=x)=1\cdot P(X=1)+0\cdot P(X=0)=1\cdot p+0\cdot(1-p)=p$.

$\mathrm{Var}(X)=E[X^2]-(E[X])^2$. $E[X^2]=\sum_x x^2P(X=x)=1^2\cdot p+0^2\cdot(1-p)=p$. $\mathrm{Var}(X)=p-p^2=p(1-p)$.

Theorem 33 (Binomial Distribution)**.** *Let $X\sim B(n,p)$, with PMF $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$.*

- $E[X]=np$

- $Var(X) = np(1-p)$

Proof.

Expectation We can write $X = \sum_{i=1}^{n} Y_i$, where $Y_i \sim B(1,p)$ are independent Bernoulli trials. By linearity of expectation, $E[X] = E\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} E[Y_i]$. From the two-point distribution, we know $E[Y_i] = p$. Therefore, $E[X] = \sum_{i=1}^{n} p = np$.

Variance Since the Y_i are independent, the variance of their sum is the sum of their variances: $\text{Var}(X) = \text{Var}\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} \text{Var}(Y_i)$. From the two-point distribution, we know $\text{Var}(Y_i) = p(1-p)$. Therefore, $\text{Var}(X) = \sum_{i=1}^{n} p(1-p) = np(1-p)$.

Theorem 34 (Geometric Distribution)**.** *Let $X \sim Geom(p)$, with PMF $P(X = k) = (1-p)^{k-1}p$ for $k = 1, 2, \ldots$.*

- $E[X] = \frac{1}{p}$
- $Var(X) = \frac{1-p}{p^2}$

Proof.

Expectation $E[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = p\sum_{k=1}^{\infty} k(1-p)^{k-1}$. Recall the geometric series formula $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$ for $|x| < 1$.

Differentiating with respect to x: $\sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2}$. Let $x = 1-p$. Since $0 < p < 1$, we have $|1-p| < 1$.

$E[X] = p \cdot \frac{1}{(1-(1-p))^2} = p \cdot \frac{1}{p^2} = \frac{1}{p}$.

Variance We use the formula $\text{Var}(X) = E[X^2] - (E[X])^2$.

$E[X^2] = \sum_{k=1}^{\infty} k^2(1-p)^{k-1}p$.

We use the identity $\sum_{k=1}^{\infty} k^2x^{k-1} = \frac{1+x}{(1-x)^3}$. Let $x = 1-p$:

$E[X^2] = p \cdot \frac{1+(1-p)}{(1-(1-p))^3} = p \cdot \frac{2-p}{p^3} = \frac{2-p}{p^2}$.

$\text{Var}(X) = E[X^2] - (E[X])^2 = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{2-p-1}{p^2} = \frac{1-p}{p^2}$.

Theorem 35 (Poisson Distribution)**.** *Let $X \sim Pois(\lambda)$, with PMF $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, 2, \ldots$.*

- $E[X] = \lambda$
- $Var(X) = \lambda$

Proof.

Expectation $E[X] = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=1}^{\infty} k\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!}$.

Let $j = k-1$. $E[X] = e^{-\lambda}\sum_{j=0}^{\infty} \frac{\lambda^{j+1}}{j!} = e^{-\lambda}\lambda\sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = e^{-\lambda}\lambda e^{\lambda} = \lambda$.

Variance We use $\text{Var}(X) = E[X(X-1)] + E[X] - (E[X])^2$.
$E[X(X-1)] = \sum_{k=0}^{\infty} k(k-1)\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=2}^{\infty} k(k-1)\frac{\lambda^k}{k!}$.
$E[X(X-1)] = e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^k}{(k-2)!}$. Let $j = k-2$.
$E[X(X-1)] = e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^{j+2}}{j!} = e^{-\lambda}\lambda^2\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} = e^{-\lambda}\lambda^2 e^{\lambda} = \lambda^2$. $\text{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Theorem 36 (Uniform Distribution)**.** *Let $X \sim U(a,b)$, with PDF $f(x) = \frac{1}{b-a}$ for $x \in [a,b]$.*

- $E[X] = \frac{a+b}{2}$
- $Var(X) = \frac{(b-a)^2}{12}$

Proof.

Expectation $E[X] = \int_{-\infty}^{\infty} x f(x)\,dx = \int_a^b x\frac{1}{b-a}\,dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{1}{b-a}\left(\frac{b^2-a^2}{2}\right) = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}$.

Variance $\text{Var}(X) = E[X^2] - (E[X])^2$.
$E[X^2] = \int_a^b x^2\frac{1}{b-a}\,dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} = \frac{b^2+ab+a^2}{3}$.
$\text{Var}(X) = \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{b^2+ab+a^2}{3} - \frac{a^2+2ab+b^2}{4}$
$\text{Var}(X) = \frac{4(b^2+ab+a^2)-3(a^2+2ab+b^2)}{12} = \frac{4b^2+4ab+4a^2-3a^2-6ab-3b^2}{12} = \frac{a^2-2ab+b^2}{12} = \frac{(b-a)^2}{12}$.

Theorem 37 (Exponential Distribution)**.** *Let $X \sim Exp(\lambda)$, with PDF $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.*

- $E[X] = \frac{1}{\lambda}$
- $Var(X) = \frac{1}{\lambda^2}$

Proof. We use integration by parts, $\int u\,dv = uv - \int v\,du$.

Expectation $E[X] = \int_0^{\infty} x\lambda e^{-\lambda x}\,dx$.
Let $u = x$, $dv = \lambda e^{-\lambda x}\,dx$. Then $du = dx$, $v = -e^{-\lambda x}$.
$E[X] = \left[-xe^{-\lambda x}\right]_0^{\infty} - \int_0^{\infty} -e^{-\lambda x}\,dx = (0-0) + \int_0^{\infty} e^{-\lambda x}\,dx = \left[-\frac{1}{\lambda}e^{-\lambda x}\right]_0^{\infty} = 0 - \left(-\frac{1}{\lambda}\right) = \frac{1}{\lambda}$.

Variance We compute $E[X^2]$ using integration by parts.
$E[X^2] = \int_0^{\infty} x^2\lambda e^{-\lambda x}\,dx$.
Let $u = x^2$, $dv = \lambda e^{-\lambda x}\,dx$. Then $du = 2x\,dx$, $v = -e^{-\lambda x}$.
$E[X^2] = \left[-x^2e^{-\lambda x}\right]_0^{\infty} - \int_0^{\infty} -e^{-\lambda x}(2x)\,dx = 0 + \frac{2}{\lambda}\int_0^{\infty} x\lambda e^{-\lambda x}\,dx$.
The integral is just $E[X]$, so $E[X^2] = \frac{2}{\lambda}E[X] = \frac{2}{\lambda}\cdot\frac{1}{\lambda} = \frac{2}{\lambda^2}$.
$\text{Var}(X) = E[X^2] - (E[X])^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{2-1}{\lambda^2} = \frac{1}{\lambda^2}$.

Theorem 38 (Normal Distribution)**.** *Let $X \sim N(\mu,\sigma^2)$, with PDF $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.*

- $E[X] = \mu$

- $Var(X) = \sigma^2$

Proof.

Let $Z \sim N(0,1)$. We know $E[Z] = 0$ and $\text{Var}(Z) = 1$, for higher dimension random variable with normal distribution.

A normal variable $X \sim N(m, \sigma^2)$ can be expressed as a linear transformation of Z:

$$X = \sigma Z + m.$$

Using the linearity of expectation $E[X] = E[\sigma Z + m] = \sigma E[Z] + m = \sigma(0) + m = m$.

Using the properties of variance, $\text{Var}(aY + b) = a^2\text{Var}(Y)$:

$$\text{Var}(X) = \text{Var}(\sigma Z + m) = \sigma^2\text{Var}(Z) = \sigma^2(1) = \sigma^2.$$

4.3 Inequality

Theorem 39 (Markov's Inequality)**.** *Let X be a non-negative random variable such that its expected value $E[X]$ exists and is finite. Then for any constant $t > 0$, the following inequality holds:*

$$P(X \geq t) \leq \frac{E[X]}{t}.$$

Proof. Given any $t > 0$, let random variables Y, Z be the following,

$$Y(\omega) = X(\omega)\mathbf{1}_{X(\omega)<t}, \quad Z(\omega) = X(\omega)\mathbf{1}_{X(\omega)\geq t}.$$

Clearly, $X = Y + Z$, and since X is non-negative, Y, Z are non-negative. $E(X) = E(Y) + E(Z)$. Since $Y \geq 0$, $E(Y) \geq 0$.

$$E(X) \geq E(Z) = E(X \cdot \mathbf{1}_{X\geq t}) \geq E(t \cdot \mathbf{1}_{X\geq t})$$

$$= tE(\mathbf{1}_{X\geq t}) = tP(X \geq t).$$

Therefore, $P(X \geq t) \leq \frac{E(X)}{t}$.

Theorem 40 (Variants of Markov's Inequality)**.** *Let X be a random variable.*

1. ***Power Form of Markov's Inequality:*** *If X is non-negative and $E[X^p]$ exists for some $p > 0$, then for any $t > 0$,*

$$P(X \geq t) = P(X^p \geq t^p) \leq \frac{E[X^p]}{t^p}.$$

2. ***Exponential Form of Markov's Inequality:*** *If the moment-generating function $M_X(\lambda) = E[e^{\lambda X}]$ exists for some $\lambda > 0$, then for any $t \in \mathbb{R}$,*

$$P(X \geq t) = P(e^{\lambda X} \geq e^{\lambda t}) \leq \frac{E[e^{\lambda X}]}{e^{\lambda t}}.$$

3. ***Chebyshev's Inequality:*** *If the variance of X exists, then for any $t > 0$,*

$$P(|X - E[X]| \geq t) \leq \frac{Var(X)}{t^2}.$$

Proof. All three inequalities are direct applications of the standard Markov's Inequality, $P(Y \geq a) \leq E[Y]/a$, by a suitable choice of the non-negative random variable Y.

1. Let $Y = X^p$.
2. Let $Y = e^{\lambda X}$.
3. Let $Y = (X - E[X])^2$.

Remark 22 (The 3-σ Rule)**.** *Let $\sigma = \sqrt{Var(X)}$ be the standard deviation. By Chebyshev's inequality, we have:*

$$P(|X - E[X]| \geq 3\sigma) \leq \frac{Var(X)}{(3\sigma)^2} = \frac{\sigma^2}{9\sigma^2} = \frac{1}{9} \approx 0.111.$$

This result indicates that the probability of a random variable's value deviating by more than three standard deviations from its mean is relatively small, regardless of its distribution. This provides a useful rule of thumb often used in statistics to identify outliers or to estimate the concentration of a distribution.

Definition 30 (Convex and Concave Functions)**.** *Let $f : I \to \mathbb{R}$ be a function defined on an interval $I \subseteq \mathbb{R}$.*

- *f is **convex** if for any two points $x_1, x_2 \in I$ and any $\lambda \in [0, 1]$, the following inequality holds:*

  $$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

  *Geometrically, this means the line segment connecting any two points on the graph of f lies **above or on** the graph.*

  $$f(x) \geq f(x_0) + f'(x_0)(x - x_0) \quad \forall x, x_0 \in \mathbb{R}$$

- *f is **concave** if for any two points $x_1, x_2 \in I$ and any $\lambda \in [0, 1]$, the following inequality holds:*

  $$f(\lambda x_1 + (1 - \lambda)x_2) \geq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

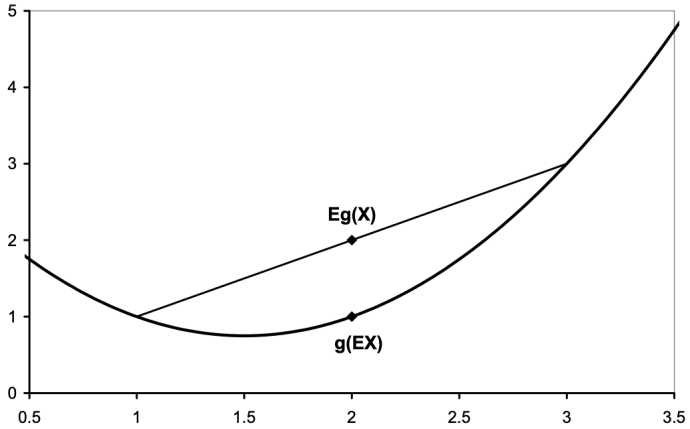
  *Geometrically, this means the line segment connecting any two points on the graph of f lies **below or on** the graph.*

For a differentiable function, an equivalent definition is that a function is convex if it lies above its tangent lines, and concave if it lies below its tangent lines.

$$f(x) \leq f(x_0) + f'(x_0)(x - x_0) \quad \forall x, x_0 \in \mathbb{R}$$

Theorem 41 (Jensen's Inequality)**.** *Let X be a random variable and $g : \mathbb{R} \to \mathbb{R}$ a function. Assume that both $E[X]$ and $E[g(X)]$ exist and are finite. Then:*

- *If g is **convex**, then $E[g(X)] \geq g(E[X])$.*
- *If g is **concave**, then $E[g(X)] \leq g(E[X])$.*



Proof. The proof for the convex case uses the tangent line property of convex functions. For any given point $x_0 \in \mathbb{R}$ and for any $x \in \mathbb{R}$:

$$g(x) \geq g(x_0) + g'(x_0)(x - x_0).$$

This inequality holds for the random variable X as well. Thus, for every outcome ω in the sample space:

$$g(X(\omega)) \geq g(x_0) + g'(x_0)(X(\omega) - x_0).$$

Taking the expectation of both sides of this inequality, we get:

$$E[g(X)] \geq E[g(x_0) + g'(x_0)(X - x_0)].$$

By the linearity of expectation, we can simplify the right-hand side:

$$E[g(X)] \geq g(x_0) + g'(x_0)(E[X] - x_0).$$

Now, let's choose a specific value for x_0 to simplify the expression. We select $x_0 = E[X]$. Substituting this into the inequality gives:

$$E[g(X)] \geq g(E[X]) + g'(x_0)(E[X] - E[X]) = g(E[X]) + g'(x_0) \cdot 0 = g(E[X]).$$

This completes the proof for the convex case.

The proof for the concave case is similar. If a function g is concave, then the function $-g$ is convex. By the convex case of Jensen's inequality, we have:

$$E[-g(X)] \geq -g(E[X]).$$

Multiplying both sides by -1 reverses the inequality sign, giving the desired result:

$$E[g(X)] \leq g(E[X]).$$

4.4 Moments of a Random Variable

Definition 31 (Moment of Order k)**.** *The moment of order k of a random variable X, denoted $E[X^k]$ or μ'_k, is defined as the expected value of X^k, provided the expectation exists.*

The first four moments provide key characteristics of the distribution:

- **First Moment ($k = 1$):** $E[X] = \mu$. Represents the central tendency or **position**.

- **Second Moment ($k = 2$):** $E[X^2] = \sigma^2 + \mu^2$. Relates to the **dispersion** (variance) of the variable.
- **Third Moment ($k = 3$):** $E[X^3]$. Used to define the **asymmetry** or skewness of the distribution.
- **Fourth Moment ($k = 4$):** $E[X^4]$. Used to define the **kurtosis** (peakedness or flatness) of the distribution.

Definition 32 (Moment Generating Function, $M_X(t)$)**.** *The Moment Generating Function of a random variable X is a function $M_X : \mathbb{R} \to \mathbb{R}$, defined as the expected value:*

$$M_X(t) = E[e^{tX}]$$

This expectation must exist for all t in some open interval $(-h, h)$ for some $h > 0$. This technical condition ensures that $M_X(t)$ is differentiable at $t = 0$.

Specifically, $M_X(t)$ is calculated as:

1. ***Discrete Case:*** *If X has probability mass function $p(x)$,*

$$M_X(t) = \sum_x e^{tx} p(x)$$

2. ***Continuous Case:*** *If X has probability density function $f(x)$,*

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

Assuming the MGF $M_X(t)$ exists in a neighborhood of $t = 0$, we can express it as a Taylor series expansion of e^{tX}:

$$e^{tX} = \sum_{k=0}^{\infty} \frac{(tX)^k}{k!} = \sum_{k=0}^{\infty} \frac{X^k t^k}{k!}$$

Taking the expectation (assuming the interchange of summation and expectation is valid):

$$M_X(t) = E[e^{tX}] = E\left[\sum_{k=0}^{\infty} \frac{X^k t^k}{k!}\right] = \sum_{k=0}^{\infty} E[X^k] \frac{t^k}{k!}$$

The coefficient of $\frac{t^k}{k!}$ in the Taylor expansion of $M_X(t)$ around $t = 0$ is precisely the k-th moment, $E[X^k]$.

Example 29 (Binomial Distribution, $X \sim \text{Binomial}(n, p)$)**.** *The probability mass function (PMF) is $p(k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, \ldots, n$.*

$$M_X(t) = E[e^{tX}] = \sum_{k=0}^{n} e^{tk} P(X = k) = \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k}$$

By grouping terms:

$$M_X(t) = \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k}$$

Using the Binomial Theorem $\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = (a+b)^n$ with $a = pe^t$ and $b = 1 - p$:

$$\mathbf{M_X(t) = (pe^t + 1 - p)^n}$$

Example 30 (Poisson Distribution, $X \sim \text{Poisson}(\lambda)$)**.** *The PMF is* $p(k) = \frac{e^{-\lambda}\lambda^k}{k!}$ *for* $k = 0, 1, 2, \ldots$.

$$M_X(t) = E[e^{tX}] = \sum_{k=0}^{\infty} e^{tk} P(X = k) = \sum_{k=0}^{\infty} e^{tk} \frac{e^{-\lambda}\lambda^k}{k!}$$

Factoring out the constant $e^{-\lambda}$ *and grouping terms:*

$$M_X(t) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!}$$

Recognizing the Maclaurin series for e^a, $\sum_{k=0}^{\infty} \frac{a^k}{k!} = e^a$, *with* $a = \lambda e^t$*:*

$$M_X(t) = e^{-\lambda} e^{\lambda e^t} = \exp\{\lambda(\mathbf{e^t - 1})\}$$

Example 31 (Exponential Distribution, $X \sim \text{Exponential}(\lambda)$)**.** *The probability density function (PDF) is* $f(x) = \lambda e^{-\lambda x}$ *for* $x > 0$.

$$M_X(t) = E[e^{tX}] = \int_0^{\infty} e^{tx} f(x) dx = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx$$

$$M_X(t) = \lambda \int_0^{\infty} e^{-(\lambda - t)x} dx$$

For the integral to converge, we require $\lambda - t > 0$*, or* $t < \lambda$*. Evaluating the integral:*

$$M_X(t) = \lambda \left[\frac{e^{-(\lambda - t)x}}{-(\lambda - t)} \right]_0^{\infty} = \frac{\lambda}{\lambda - t} \left[-e^{-(\lambda - t)x} \right]_0^{\infty}$$

$$M_X(t) = \frac{\lambda}{\lambda - t} \left[0 - (-e^0) \right] = \frac{\lambda}{\lambda - \mathbf{t}}, \quad \textit{for } t < \lambda$$

Example 32 (Normal Distribution, $X \sim \text{Normal}(\mu, \sigma^2)$)**.** *The PDF is* $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right\}$.

$$M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} dx$$

The exponent of the integral is:

$$tx - \frac{1}{2\sigma^2}(x - \mu)^2 = -\frac{1}{2\sigma^2} \left[(x - \mu)^2 - 2\sigma^2 tx \right]$$

By completing the square for the term in brackets, we find:

$$(x - \mu)^2 - 2\sigma^2 tx = \left(x - (\mu + \sigma^2 t) \right)^2 - (\mu + \sigma^2 t)^2 + \mu^2$$

Substituting this back into the exponent and separating the constant terms from the integral:

$$M_X(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{1}{2\sigma^2} \left[(\mu + \sigma^2 t)^2 - \mu^2 \right] \right\} \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2\sigma^2} \left(x - (\mu + \sigma^2 t) \right)^2 \right\} dx$$

The integral is $\int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2\sigma^2} (x - \mu^*)^2 \right\} dx$*, where* $\mu^* = \mu + \sigma^2 t$*. This integral is equal to* $\sigma\sqrt{2\pi}$*. Substituting this integral value:*

$$M_X(t) = \exp\left\{ \frac{1}{2\sigma^2} \left[\mu^2 + 2\mu\sigma^2 t + \sigma^4 t^2 - \mu^2 \right] \right\}$$

$$M_X(t) = \exp\left\{ \frac{1}{2\sigma^2} \left[2\mu\sigma^2 t + \sigma^4 t^2 \right] \right\} = \exp\left\{ \mu\mathbf{t} + \frac{\boldsymbol{\sigma^2 t^2}}{\mathbf{2}} \right\}$$

Theorem 42 (Moment Generation)**.** *Let X be a random variable with MGF $M_X(t)$. If $M_X(t)$ is finite in an open interval containing $t = 0$, then the n-th moment of X exists and is given by:*

$$E[X^n] = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}$$

Proof. We aim to show that $M_X^{(n)}(t) = E[X^n e^{tX}]$. By definition, $M_X(t) = E[e^{tX}]$. We can interchange the differentiation and expectation operators since the MGF is assumed to be differentiable in a neighborhood of $t = 0$:

$$M_X^{(n)}(t) = \frac{d^n}{dt^n} E[e^{tX}] = E\left[\frac{d^n}{dt^n}(e^{tX})\right] = E[X^n e^{tX}]$$

Evaluating at $t = 0$:

$$E[X^n] = M_X^{(n)}(0) = E[X^n e^0]$$

This proof holds for both continuous and discrete cases upon justifying the interchange of the derivative and the integral/summation.

Theorem 43 (Uniqueness Theorem)**.** *If the moment generating function of a random variable exists, it is unique. Furthermore, the MGF uniquely determines the probability density or mass function of the random variable, up to a set of probability zero.*

Proof. (Special Case: X assumes a finite set of non-negative integers): Let the range of X be a finite set of non-negative integers $R_X = \{0, 1, \ldots, n\}$, and let $p(j) = P(X = j)$. The MGF is:

$$M_X(t) = \sum_{j=0}^{n} e^{tj} p(j)$$

By setting $z = e^t$, $M_X(t)$ is equivalent to the Probability Generating Function (PGF), $H(z)$:

$$H(z) = \sum_{j=0}^{n} z^j p(j)$$

Since $H(z)$ is a polynomial in z of degree n, its coefficients $p(j)$ are uniquely determined by $H(z)$. Specifically, $p(j)$ is the coefficient of z^j:

$$p(j) = \frac{1}{j!} H^{(j)}(0)$$

Since the MGF uniquely determines $H(z)$, and $H(z)$ uniquely determines the probabilities $p(j)$ (the PMF), the MGF uniquely determines the distribution of X in this specific finite discrete case.

Theorem 44 (MGF of a Linear Transformation)**.** *Let X be a random variable with MGF $M_X(t)$, and let $Y = aX + b$, where a and b are constants. Then the MGF of Y is:*

$$M_Y(t) = e^{bt} M_X(at)$$

Proof. By the definition of the MGF:

$$M_Y(t) = E[e^{tY}] = E[e^{t(aX+b)}]$$

Applying the exponent rule $e^{A+B} = e^A e^B$:

$$M_Y(t) = E[e^{atX} e^{bt}]$$

Since e^{bt} is a constant with respect to the expectation over X, it can be pulled outside the expectation operator:

$$M_Y(t) = e^{bt} E[e^{(at)X}]$$

We recognize $E[e^{(at)X}]$ as the definition of the MGF of X, $M_X(u)$, evaluated at $u = at$:

$$M_Y(t) = e^{bt} M_X(at)$$

Theorem 45 (MGF of the Sum of Independent Variables)**.** *Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables. The MGF of their sum $S_n = X_1 + \cdots + X_n$ is the product of their individual MGFs:*

$$M_{S_n}(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t)$$

Proof. By the definition of the MGF for the sum S_n:

$$M_{S_n}(t) = E[e^{tS_n}] = E[e^{t(X_1+X_2+\cdots+X_n)}]$$

$$M_{S_n}(t) = E[e^{tX_1} e^{tX_2} \cdots e^{tX_n}]$$

Since $X_1, X_2, \ldots, X_n$ are independent, the functions $e^{tX_1}, e^{tX_2}, \ldots, e^{tX_n}$ are also independent. Therefore, the expectation of their product is equal to the product of their individual expectations:

$$M_{S_n}(t) = E[e^{tX_1}] E[e^{tX_2}] \cdots E[e^{tX_n}]$$

Recognizing each term as the MGF of the corresponding variable:

$$M_{S_n}(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t)$$

Example 33. *Let X be a random variable with moments $\mu'_0 = 1$ and $\mu'_k = E[X^k]$ given by:*

$$\mu'_k = \frac{1}{2} + \frac{1}{4} \cdot 2^k, \quad \text{for } k \geq 1$$

Calculate the distribution of X. The MGF is derived from the Taylor series expansion:

$$M_X(t) = \sum_{k=0}^{\infty} \mu'_k \frac{t^k}{k!} = 1 + \sum_{k=1}^{\infty} \left(\frac{1}{2} + \frac{1}{4} 2^k \right) \frac{t^k}{k!}$$

After algebraic rearrangement (as detailed in the body of the document):

$$M_X(t) = \frac{1}{4} e^{0t} + \frac{1}{2} e^{1t} + \frac{1}{4} e^{2t}$$

This MGF corresponds to a discrete distribution with probability mass function $p(x)$:

$$P(X = x) = \begin{cases} \frac{1}{4} & \text{if } x = 0 \\ \frac{1}{2} & \text{if } x = 1 \\ \frac{1}{4} & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

By the Uniqueness Theorem, this must be the distribution of X.

Chapter 5

Convergences and Limit theorems

A few lectures ago we discussed the notion of finitely and infinitely often. Recall that

Definition 33. *1. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(A_n)_{n=1}^{\infty}$ a sequence of events. We define*

- $\{A_n \text{ i.o.}\} := \{\omega \in \Omega \mid \omega \in A_n \text{ for infinitely many indices } n \in \mathbb{N}\}$
- $\{A_n \text{ f.o.}\} := \{\omega \in \Omega \mid \omega \in A_n \text{ for finitely many indices } n \in \mathbb{N}\}$

We also showed that

Fact 15. *1. Subsets $\{A_n \text{ i.o.}\}$ and $\{A_n \text{ f.o.}\}$ are events as they can be expressed in the form*

1. $\{A_n \text{ i.o.}\} = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n$
2. $\{A_n \text{ f.o.}\} = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n^c$

Moreover,

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

We showed the Borel Cantelli Lemma

Theorem 46. *2. Let $(A_n)_{n=1}^{\infty}$ be a sequence of events in the same space $(\Omega, \mathcal{F}, P)$,*

1. *If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(\{A_n \text{ i.o.}\}) = 0$, which means $P(\{A_n \text{ f.o.}\}) = 1$.*
2. *If $\sum_{n=1}^{\infty} P(A_n) = \infty$ and $(A_n)_{n=1}^{\infty}$ are independent events, then $P(\{A_n \text{ i.o.}\}) = 1$.*

5.1 Types of convergence for sequences of random variables

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(X_n)_{n=1}^{\infty}$ a sequence of random variables. Recall that $X_n : \Omega \to \mathbb{R}$ are functions, and we can consider their pointwise and uniform convergence.

Definition 34 (Pointwise and Uniform Convergence)**.** *Let $(X_n)_{n=1}^{\infty}$ be a sequence of functions and $X : \Omega \to \mathbb{R}$.*

- *We say that $X_n \to X$ **pointwise** on Ω if for every $\omega \in \Omega$, the sequence of real numbers $X_n(\omega)$ converges to $X(\omega)$. That is,*

$$\forall \omega \in \Omega, \forall \epsilon > 0, \exists N_\omega \in \mathbb{N} \text{ such that } \forall n \geq N_\omega, |X_n(\omega) - X(\omega)| < \epsilon.$$

- *We say that $X_n \rightrightarrows X$ **uniformly** on Ω if the rate of convergence is independent of ω. That is,*

$$\forall \epsilon > 0, \exists N \in \mathbb{N} \text{ such that } \forall n \geq N, \forall \omega \in \Omega, |X_n(\omega) - X(\omega)| < \epsilon.$$

Pointwise and uniform convergence are deterministic concepts from real analysis. In probability theory, we introduce a new type of convergence called almost sure convergence, which allows for convergence to fail on a set of outcomes that has probability zero.

Definition 35 (Almost Sure Convergence)**.** *A sequence of random variables $(X_n)_{n=1}^{\infty}$ is said to converge **almost surely** to a random variable X, denoted by $X_n \xrightarrow{a.s.} X$, if the set of outcomes where convergence occurs has probability one. More precisely, if the set*

$$A_X = \{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$$

satisfies $P(A_X) = 1$. The set A_X is also known as the set of convergence.

In order for this definition to be meaningful in a rigorous probabilistic sense, we must first show that the set A_X is a measurable set (i.e., an event) belonging to the σ-algebra $\mathcal{F}$.

Note that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ is equivalent to the following formal definition of a limit:

$$\forall \epsilon > 0, \exists N \in \mathbb{N} \text{ such that } \forall n \geq N, |X_n(\omega) - X(\omega)| < \epsilon.$$

To show that the set A_X is an event, we need to express it as a countable combination of sets that are already known to be events. We can simplify the condition by considering a sequence of decreasing ϵ values, such as $\epsilon = 1/k$ for $k \in \mathbb{N}$. The limit condition holds for all $\epsilon > 0$ if and only if it holds for this specific sequence of ϵ values.

Thus, the set of convergence A_X can be written as:

$$A_X = \bigcap_{k=1}^{\infty} \left\{ \omega \in \Omega \mid \exists N \in \mathbb{N} \text{ s.t. } \forall n \geq N, |X_n(\omega) - X(\omega)| < \frac{1}{k} \right\}.$$

We can express this set using standard set notation:

$$A_X = \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{\omega \in \Omega \mid |X_n(\omega) - X(\omega)| < \frac{1}{k}\}.$$

Since X_n and X are random variables (and thus measurable functions), the function $Y_n(\omega) = |X_n(\omega) - X(\omega)|$ is also a random variable.

By the definition of a random variable, for any constant c, the set $\{\omega \in \Omega \mid Y_n(\omega) < c\}$ is an event, i.e., it belongs to the σ-algebra $\mathcal{F}$. Therefore, for any $k \in \mathbb{N}$, the set $\{\omega : |X_n(\omega) - X(\omega)| < 1/k\}$ is an event. The intersection and union of a countable number of events are also events.

The expression for A_X involves a countable intersection and a countable union of events. Specifically, it is a countable intersection of the sets $B_k = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \{\omega : |X_n(\omega) - X(\omega)| < 1/k\}$, where each B_k is a countable union of countable intersections of events.

Remark 23. *If a sequence of random variables X_n converges to X pointwise on Ω, then it also converges almost surely. The converse is not necessarily true.*

Example 34. *Let $(\Omega, \mathcal{F}, P)$ be the probability space where $\Omega = [0,1]$, $\mathcal{F}$ is the Borel σ-algebra $\mathcal{B}([0,1])$, and P is the Lebesgue measure (geometric probability). Let $X(\omega) = 0$ for all $\omega \in \Omega$. Consider a countable set of points $\{a_i\}_{i=1}^{\infty} \subset [0,1]$. Define the sequence of random variables $(X_n)_{n=1}^{\infty}$ as:*

$$X_n(t) = \begin{cases} n & \text{if } t \in \{a_i\}_{i=1}^{n} \\ 0 & \text{otherwise} \end{cases}$$

1. ***Almost Sure Convergence:*** *The set of convergence is $A_X = \{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = 0\}$. For any $\omega \notin \{a_i\}_{i=1}^{\infty}$, we have $X_n(\omega) = 0$ for all n, so $X_n(\omega) \to 0$. Since the set $\{a_i\}_{i=1}^{\infty}$ is countable, its Lebesgue measure is zero. Therefore, the probability of its complement is one: $P([0,1] \setminus \{a_i\}_{i=1}^{\infty}) = 1$. The set of convergence A_X contains $[0,1] \setminus \{a_i\}_{i=1}^{\infty}$, so $P(A_X) \geq P([0,1] \setminus \{a_i\}_{i=1}^{\infty}) = 1$. Thus, $X_n \xrightarrow{a.s.} X$.*

2. ***Pointwise Convergence:*** *The sequence does not converge pointwise on Ω. At any point $t = a_k$ in the countable set, the sequence $X_n(a_k)$ eventually takes the value n and diverges to infinity, so $\lim_{n\to\infty} X_n(a_k) = \infty$.*

This example shows that almost sure convergence is a weaker condition than pointwise convergence.

Theorem 47. *Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables and X a random variable. The sequence converges almost surely to X if and only if for every $\epsilon > 0$, the probability of the event $\{|X_n - X| \geq \epsilon\}$ occurring only a finite number of times is one. This can be stated formally as:*

$$X_n \xrightarrow{a.s.} X \iff \forall \epsilon > 0, P(\{|X_n - X| \geq \epsilon \text{ } i.o.\}) = 0.$$

Proof.

($\implies$) Assume that $X_n \xrightarrow{\text{a.s.}} X$. By definition, this means $P(\lim_{n\to\infty} X_n = X) = 1$. The set of convergence is $A_X = \{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$, and we have $P(A_X) = 1$. The complement of this set, A_X^c, is the set of non-convergence, and $P(A_X^c) = 0$.

$$A_X^c = \bigcup_{k=1}^{\infty} \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} \{\omega \in \Omega \mid |X_n(\omega) - X(\omega)| \geq \frac{1}{k}\} = \bigcup_{k=1}^{\infty} \{\omega \mid |X_n(\omega) - X(\omega)| \geq 1/k \text{ i.o.}\}.$$

Since $P(A_X^c) = 0$, this implies that $P(\{|X_n - X| \geq 1/k \text{ i.o.}\}) = 0$ for all $k \in \mathbb{N}$. Now, consider an arbitrary $\epsilon > 0$. We can always find a natural number m such that $1/m \leq \epsilon$. This implies that

$$\{\omega \mid |X_n - X| \geq \epsilon \text{ i.o.}\} \subseteq \{\omega \mid |X_n - X| \geq 1/m \text{ i.o.}\}.$$

Then, $P(\{|X_n - X| \geq \epsilon \text{ i.o.}\}) \leq P(\{|X_n - X| \geq 1/k \text{ i.o.}\}) = 0$. Thus, $P(\{|X_n - X| \geq \epsilon \text{ i.o.}\}) = 0$ for any $\epsilon > 0$.

($\impliedby$) Assume that for every $\epsilon > 0$, we have $P(\{|X_n - X| \geq \epsilon \text{ i.o.}\}) = 0$. This means that for every $k \in \mathbb{N}$, $P(\{|X_n - X| \geq 1/k \text{ i.o.}\}) = 0$. We want to show that $P(A_X) = 1$, or equivalently, $P(A_X^c) = 0$. We know that A_X^c is the countable union of sets with probability zero:

$$A_X^c = \bigcup_{k=1}^{\infty} \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} \{\omega \in \Omega \mid |X_n(\omega) - X(\omega)| \geq \frac{1}{k}\} = \bigcup_{k=1}^{\infty} \{\omega \mid |X_n(\omega) - X(\omega)| \geq 1/k \text{ i.o.}\}.$$

$$P(A_X^c) = P\left(\bigcup_{k=1}^{\infty} \{\omega \mid |X_n - X| \geq 1/k \text{ i.o.}\}\right) \leq \sum_{k=1}^{\infty} P(\{|X_n - X| \geq 1/k \text{ i.o.}\}).$$

Since each term in the sum is zero, the sum is also zero.

$$P(A_X^c) \leq \sum_{k=1}^{\infty} 0 = 0.$$

This implies $P(A_X^c) = 0$, so $P(A_X) = 1$. Therefore, $X_n \xrightarrow{\text{a.s.}} X$.

Lemma 2. $X_n \xrightarrow{a.s.} X \iff \forall \epsilon > 0, \lim_{N\to\infty} P\left(\bigcup_{n=N}^{\infty} \{|X_n - X| \geq \epsilon\}\right) = 0.$

Proof. Let $\epsilon > 0$ be given. Define the events $A_N(\epsilon) = \bigcup_{n=N}^{\infty} \{|X_n - X| \geq \epsilon\}$. This is a decreasing sequence of events, i.e., $A_1(\epsilon) \supseteq A_2(\epsilon) \supseteq A_3(\epsilon) \supseteq \dots$. By the continuity of the probability measure

for a decreasing sequence of events, we have:

$$\lim_{N\to\infty} P(A_N(\epsilon)) = P\left(\bigcap_{N=1}^{\infty} A_N(\epsilon)\right).$$

The intersection of these events is precisely the event that the condition $|X_n - X| \geq \epsilon$ occurs infinitely often (i.o.):

$$\bigcap_{N=1}^{\infty} A_N(\epsilon) = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} \{|X_n - X| \geq \epsilon\} = \{|X_n - X| \geq \epsilon \text{ i.o.}\}.$$

Therefore, the statement $\lim_{N\to\infty} P\left(\bigcup_{n=N}^{\infty}\{|X_n - X| \geq \epsilon\}\right) = 0$ is equivalent to $P(\{|X_n - X| \geq \epsilon \text{ i.o.}\}) = 0$. From a previous theorem (Theorem 47), we know that $X_n \xrightarrow{\text{a.s.}} X$ if and only if $P(\{|X_n - X| \geq \epsilon \text{ i.o.}\}) = 0$ for all $\epsilon > 0$. This concludes the proof.

Definition 36. *Given a probability space $(\Omega, \mathcal{F}, P)$ and a sequence of random variables $(X_n)_{n=1}^{\infty}$.*

- *X_n converges to X **in mean square** ($X_n \xrightarrow{m.s.} X$) if*

$$\lim_{n\to\infty} E[(X_n - X)^2] = 0.$$

- *X_n converges to X **in probability** ($X_n \xrightarrow{p.} X$) if for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P(\{|X_n - X| \geq \epsilon\}) = 0.$$

- *X_n converges to X **almost surely** ($X_n \xrightarrow{a.s.} X$) if for every $\epsilon > 0$,*

$$\lim_{N\to\infty} P\left(\bigcup_{n=N}^{\infty}\{|X_n - X| \geq \epsilon\}\right) = 0.$$

Lemma 3. *Given a probability space $(\Omega, \mathcal{F}, P)$, a sequence of random variables $(X_n)_{n=1}^{\infty}$, and a random variable X:*

1. *If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{p.} X$.*
2. *If $X_n \xrightarrow{m.s.} X$, then $X_n \xrightarrow{p.} X$.*

Proof.

1. Assume $X_n \xrightarrow{\text{a.s.}} X$. By a previous theorem, this is equivalent to the condition that for any $\epsilon > 0$, $P(\{|X_n - X| \geq \epsilon \text{ i.o.}\}) = 0$. The event $\{|X_n - X| \geq \epsilon \text{ i.o.}\}$ can be expressed as a countable intersection of events:

$$\{|X_n - X| \geq \epsilon \text{ i.o.}\} = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} \{|X_n - X| \geq \epsilon\}.$$

Let $A_N^{\epsilon} = \bigcup_{n=N}^{\infty}\{|X_n - X| \geq \epsilon\}$. This forms a decreasing sequence of events, i.e., $A_1^{\epsilon} \supseteq A_2^{\epsilon} \supseteq \ldots$. By the continuity of the probability measure for decreasing sequences of events, we have:

$$\lim_{N\to\infty} P(A_N^{\epsilon}) = P\left(\bigcap_{N=1}^{\infty} A_N^{\epsilon}\right) = P(\{|X_n - X| \geq \epsilon \text{ i.o.}\}).$$

Since we assumed almost sure convergence, the right-hand side is 0. Therefore,

$$\lim_{N\to\infty} P(A_N^\epsilon) = 0.$$

We want to show that $X_n \xrightarrow{\text{p.}} X$, which means $\lim_{n\to\infty} P(\{|X_n - X| \geq \epsilon\}) = 0$. By the definition of A_n^ϵ, the event $\{|X_n - X| \geq \epsilon\}$ is a subset of A_n^ϵ.

$$\{|X_n - X| \geq \epsilon\} \subseteq \bigcup_{k=n}^{\infty} \{|X_k - X| \geq \epsilon\} = A_n^\epsilon.$$

By the monotonicity of probability, $0 \leq P(\{|X_n - X| \geq \epsilon\}) \leq P(A_n^\epsilon) = 0$.

2. Assume $X_n \xrightarrow{\text{m.s.}} X$. We need to show that $X_n \xrightarrow{\text{p.}} X$. By Markov's inequality, for any $\epsilon > 0$,

$$P(\{|X_n - X| \geq \epsilon\}) = P(\{(X_n - X)^2 \geq \epsilon^2\}) \leq \frac{E[(X_n - X)^2]}{\epsilon^2}.$$

Taking the limit of both sides of the inequality gives:

$$\lim_{n\to\infty} P(\{|X_n - X| \geq \epsilon\}) \leq \frac{\lim_{n\to\infty} E[(X_n - X)^2]}{\epsilon^2} = \frac{0}{\epsilon^2} = 0.$$

Example 35. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $X(\omega) \equiv 0$. Let $(X_n)_{n=1}^\infty$ be a sequence of **independent** Poisson random variables, with $X_n \sim Pois(1/n)$.*

Recall that for $Y \sim Pois(\lambda)$, we have $E[Y] = \lambda$ and $Var(Y) = \lambda$.

Therefore, for our sequence, $E[X_n] = 1/n$ and $Var(X_n) = 1/n$.

The second moment is $E[X_n^2] = Var(X_n) + (E[X_n])^2 = \frac{1}{n} + \left(\frac{1}{n}\right)^2 = \frac{1}{n} + \frac{1}{n^2}$.

***Convergence in Mean Square:** $X_n \xrightarrow{m.s.} 0$. We check the definition directly:*

$$\lim_{n\to\infty} E[(X_n - 0)^2] = \lim_{n\to\infty} E[X_n^2] = \lim_{n\to\infty} \left(\frac{1}{n} + \frac{1}{n^2}\right) = 0.$$

***Convergence in Probability:** $X_n \xrightarrow{p.} 0$. By Markov's inequality, for any $\epsilon > 0$,*

$$P(|X_n - 0| \geq \epsilon) = P(|X_n| \geq \epsilon) \leq \frac{E[|X_n|^2]}{\epsilon^2} = \frac{E[X_n^2]}{\epsilon^2} = \frac{1/n + 1/n^2}{\epsilon^2}.$$

Taking the limit as $n \to \infty$, we have:

$$\lim_{n\to\infty} P(|X_n| \geq \epsilon) \leq \lim_{n\to\infty} \frac{1/n + 1/n^2}{\epsilon^2} = 0.$$

***Almost Sure Convergence:** X_n does **not** converge almost surely to 0. We check the condition $P(\{|X_n - 0| \geq \epsilon \text{ i.o.}\}) = 1$.*

Let ϵ be any value in $(0, 1)$. Since X_n takes non-negative integer values, the event $|X_n| \geq \epsilon$ is the same as $X_n \geq \varepsilon$. We consider the sum of probabilities of these events:

$$\sum_{n=1}^{\infty} P(|X_n| \geq \epsilon) = \sum_{n=1}^{\infty} P(X_n \geq \varepsilon) = \sum_{n=1}^{\infty} (1 - P(X_n = 0)).$$

Using the Poisson PMF, $P(X_n = 0) = \frac{(1/n)^0 e^{-1/n}}{0!} = e^{-1/n}$. The sum becomes:

$$\sum_{n=1}^{\infty} (1 - e^{-1/n}).$$

To determine convergence, we compare with the harmonic series $\sum 1/n$, which is known to diverge. Then, the series $\sum(1-e^{-1/n})$ diverges. Because the events $\{|X_n| \geq \epsilon\}$ are independent, by the second Borel-Cantelli lemma, a divergent sum of probabilities implies the event occurs infinitely often with probability one.

$$P(\{|X_n| \geq \epsilon \text{ } i.o.\}) = 1.$$

Since this probability is 1 (not 0), X_n does not converge almost surely to 0.

5.2 Properties of Convergence

Theorem 48. *Given random variables $(X_n)_{n=1}^{\infty}$, $(Y_n)_{n=1}^{\infty}$, X, Y on a probability space $(\Omega, \mathcal{F}, P)$ with $X_n \xrightarrow{a.s.} X$ and $Y_n \xrightarrow{a.s.} Y$, then*

1. $X_n + Y_n \xrightarrow{a.s.} X + Y$
2. $X_n Y_n \xrightarrow{a.s.} XY$

Proof. By the definition of almost sure convergence, we have

$$P(\{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1 \quad \text{and} \quad P(\{\omega \in \Omega \mid \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}) = 1.$$

Let's denote these events as A_X and A_Y respectively. We are given $P(A_X) = 1$ and $P(A_Y) = 1$.

$$1 = P(A_X \cup A_Y) = P(A_X) + P(A_Y) - P(A_X \cap A_Y) = 1 + 1 - P(A_X \cap A_Y).$$

Then we have that $P(A_X \cap A_Y) = 1$. Therefore, for every $\omega \in A_X \cap A_Y$, we have

$$\lim_{n\to\infty} (X_n(\omega) + Y_n(\omega)) = X(\omega) + Y(\omega)$$

and

$$\lim_{n\to\infty} (X_n(\omega) Y_n(\omega)) = X(\omega) Y(\omega).$$

This means that $A_X \cap A_Y \subseteq \{\omega \in \Omega \mid \lim_{n\to\infty}(X_n(\omega) + Y_n(\omega)) = X(\omega) + Y(\omega)\}$ and $A_X \cap A_Y \subseteq \{\omega \in \Omega \mid \lim_{n\to\infty}(X_n(\omega)Y_n(\omega)) = X(\omega)Y(\omega)\}$.

Since $P(A_X \cap A_Y) = 1$, the probabilities of these larger events must also be 1. Thus, $X_n + Y_n \xrightarrow{\text{a.s.}} X + Y$ and $X_n Y_n \xrightarrow{\text{a.s.}} XY$.

Theorem 49. *Given random variables $(X_n)_{n=1}^{\infty}, (Y_n)_{n=1}^{\infty}$, X, Y on a probability space $(\Omega, \mathcal{F}, P)$ with $X_n \xrightarrow{p.} X$ and $Y_n \xrightarrow{p.} Y$, then*

1. $X_n + Y_n \xrightarrow{p.} X + Y$
2. $X_n Y_n \xrightarrow{p.} XY$

Proof. We will prove the first statement. The proof for the second statement is similar and is left as an exercise.

Given any $\varepsilon > 0$, we need to show that $\lim_{n\to\infty} P(|X_n + Y_n - X - Y| > \varepsilon) = 0$.

By the triangle inequality, we have

$$\varepsilon < |(X_n + Y_n) - (X + Y)| = |(X_n - X) + (Y_n - Y)| \leq |X_n - X| + |Y_n - Y|.$$

Furthermore, if $|X_n - X| + |Y_n - Y| > \varepsilon$, it must be that either $|X_n - X| > \varepsilon/2$ or $|Y_n - Y| > \varepsilon/2$ (or both). This allows us to establish the following set inclusion:

$$P(|X_n + Y_n - X - Y| > \varepsilon) \leq P\left(|X_n - X| > \frac{\varepsilon}{2}\right) + P\left(|Y_n - Y| > \frac{\varepsilon}{2}\right).$$

By the definition of convergence in probability, we have $\lim_{n\to\infty} P(|X_n - X| > \varepsilon/2) = 0$ and $\lim_{n\to\infty} P(|Y_n - Y| > \varepsilon/2) = 0$. Taking the limit as $n \to \infty$ on both sides of the inequality, we find that

$$\lim_{n\to\infty} P(|X_n + Y_n - X - Y| > \varepsilon) \leq 0 + 0 = 0.$$

Since probabilities are non-negative, the limit must be equal to 0. Hence, $X_n + Y_n \xrightarrow{\text{p.}} X + Y$.

5.3 Law of large numbers

Lemma 4. *Let X be a random variable and assume that for some integer $n \geq 2$, we have $E|X|^n < \infty$. Then for any $1 \leq r \leq n$, we have $(E|X|^r)^{1/r} \leq (E|X|^n)^{1/n}$. A simpler, but often useful, result for $1 \leq r < n$ is $E|X|^r \leq E|X|^n + 1$.*

Proof. Define the function $g(y) = y^{n/r}$. Since $1 \leq r \leq n$, the exponent $p = \frac{n}{r} \geq 1$, which implies that $g(y)$ is a fconvex function for $y \geq 0$. Applying Jensen's Inequality, $g(E[Y]) \leq E[g(Y)]$, where $Y = |X|^r$:

Substitute Y into the inequality: $g(E[|X|^r]) \leq E[g(|X|^r)]$. Apply the function $g(y) = y^{n/r}$: $(E|X|^r)^{n/r} \leq E\left[(|X|^r)^{n/r}\right]$. Simplify the expression: $(E|X|^r)^{n/r} \leq E|X|^n$. Taking the n-th root (raising both sides to $1/n$): $(E|X|^r)^{1/r} \leq (E|X|^n)^{1/n}$. Now, we can split the expectation into two parts based on the value of $|X|$.

$$E|X|^r = E[|X|^r \mathbb{I}_{\{|X|\leq 1\}}] + E[|X|^r \mathbb{I}_{\{|X|>1\}}].$$

For the first term, since $|X| \leq 1$, we have $|X|^r \leq 1$. Thus, $E[|X|^r \mathbb{I}_{\{|X|\leq 1\}}] \leq E[1 \cdot \mathbb{I}_{\{|X|\leq 1\}}] = P(|X| \leq 1) \leq 1$. For the second term, since $|X| > 1$, and $r \leq n$, we have $|X|^r \leq |X|^n$. Thus, $E[|X|^r \mathbb{I}_{\{|X|>1\}}] \leq E[|X|^n \mathbb{I}_{\{|X|>1\}}] \leq E[|X|^n] < \infty$. Combining the two parts, we have $E|X|^r \leq 1 + E|X|^n$.

Theorem 50 (Strong Law of Large Numbers under Moment Conditions)**.** *Let $(X_i)_{i=1}^{\infty}$ be a sequence of independent random variables on a probability space $(\Omega, \mathcal{F}, P)$. Assume that for some constant c,*

1. *$E(X_i) = 0$ for all i,*
2. *$E(X_i^4) < c$ for all i.*

Then the sequence of partial sums $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges almost surely to 0, i.e., $\bar{X}_n \xrightarrow{\text{a.s.}} 0$.

Proof. The proof relies on the Borel-Cantelli Lemma. To show $\bar{X}_n \xrightarrow{\text{a.s.}} 0$, we must show that for any $\varepsilon > 0$, the sum of probabilities $\sum_{n=1}^{\infty} P(|\bar{X}_n| > \varepsilon)$ is finite.

By the Markov inequality, this can be achieved by bounding a higher moment of $\bar{X}_n$. We will use the fourth moment. We first note that from the given conditions, $E(X_i) = 0$. By Lemma 4, $E(X_i^2) \leq E(X_i^4) + 1 < c + 1$.

Now, let's calculate the fourth moment of $\bar{X}_n$:

$$E(\bar{X}_n^4) = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^4\right] = \frac{1}{n^4}E\left[\left(\sum_{i=1}^{n} X_i\right)^4\right] = \frac{1}{n^4}\sum_{i,j,k,l=1}^{n} E(X_iX_jX_kX_l).$$

Due to the independence of the random variables and the fact that $E(X_i) = 0$, the expectation $E(X_iX_jX_kX_l)$ is non-zero only in two cases:

1. All indices are the same, i.e., $i = j = k = l$. There are n such terms, and each has the form $E(X_i^4)$.
2. The indices form two distinct pairs, e.g., $i = j \neq k = l$. There are $3\binom{n}{2}$ such terms. The number of ways to choose two distinct indices i and j is $\binom{n}{2}$. The number of ways to pair them up (e.g. $X_i^2X_j^2$) is 3 (e.g., $i = j, k = l$ or $i = k, j = l$ or $i = l, j = k$). Each term has the form $E(X_i^2X_j^2) = E(X_i^2)E(X_j^2)$ due to independence.

Combining these cases, we have:

$$E\left[\left(\sum_{i=1}^{n} X_i\right)^4\right] = \sum_{i=1}^{n} E(X_i^4) + 3\sum_{i\neq j} E(X_i^2)E(X_j^2).$$

Using our bounds on the moments:

$$\sum_{i=1}^{n} E(X_i^4) < \sum_{i=1}^{n} c = nc.$$

$$3\sum_{i\neq j} E(X_i^2)E(X_j^2) < 3\sum_{i\neq j}(c+1)(c+1) = 3n(n-1)(c+1)^2.$$

(Note: The number of terms in $\sum_{i\neq j}$ is $n(n-1)$).

Thus, we can bound the fourth moment of $\bar{X}_n$:

$$\begin{aligned} E(\bar{X}_n^4) &= \frac{1}{n^4}E\left[\left(\sum_{i=1}^{n} X_i\right)^4\right] < \frac{nc + 3n(n-1)(c+1)^2}{n^4} = \frac{c}{n^3} + \frac{3(n-1)(c+1)^2}{n^3} \\ &= \frac{c}{n^3} + 3(c+1)^2\left(\frac{1}{n^2} - \frac{1}{n^3}\right). \end{aligned}$$

Therefore, the sum $\sum_{n=1}^{\infty} E(\bar{X}_n^4)$ converges since the terms are asymptotically proportional to $1/n^2$.

Finally, by Markov's inequality, for any $\varepsilon > 0$:

$$\sum_{n=1}^{\infty} P(|\bar{X}_n| > \varepsilon) = \sum_{n=1}^{\infty} P(\bar{X}_n^4 > \varepsilon^4) \leq \sum_{n=1}^{\infty} \frac{E(\bar{X}_n^4)}{\varepsilon^4}.$$

Since $\sum_{n=1}^{\infty} E(\bar{X}_n^4)$ is finite, the right side is a finite sum. By the first Borel-Cantelli Lemma, if $\sum_{n=1}^{\infty} P(|\bar{X}_n| > \varepsilon) < \infty$, then the probability of the event $\{|\bar{X}_n| > \varepsilon \text{ infinitely often}\}$ is zero. This is the definition of $\bar{X}_n$ almost sure convergence to 0.

Remark 24. *Note that the previous theorem assumes $E(X_i^4) < \infty$. However, a weaker result can be obtained if the condition is less restrictive, for instance, if we only have $E(X_i^2) < \infty$. More precisely, if the random variables X_i are independent with $E(X_i) = 0$ and $E(X_i^2) = \sigma^2 < \infty$, we can show that $\bar{X}_n$ converges in mean square and in probability.*

1. Mean Square Convergence *First, we show that $\bar{X}_n$ converges to 0 in mean square. The*

expected value of $\bar{X}_n^2$ is:

$$\begin{aligned} E(\bar{X}_n^2) &= Var(\bar{X}_n) \\ &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} E(X_i^2) \end{aligned}$$

Since $E(X_i^2) = \sigma^2$ for all i, we have:

$$E(\bar{X}_n^2) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

As $n \to \infty$, it follows that $E(\bar{X}_n^2) \to 0$.

2. Convergence in Probability (Weak Law of Large Numbers) *Next, we show that $\bar{X}_n$ converges to 0 in probability. For any $\varepsilon > 0$, we use Markov's inequality:*

$$P(|\bar{X}_n - 0| > \varepsilon) = P(\bar{X}_n^2 > \varepsilon^2) \leq \frac{E(\bar{X}_n^2)}{\varepsilon^2}$$

Using the result from the previous step, we substitute $E(\bar{X}_n^2)$:

$$P(|\bar{X}_n| > \varepsilon) \leq \frac{\sigma^2/n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

As $n \to \infty$, the right-hand side of the inequality goes to 0, which implies that $P(|\bar{X}_n| > \varepsilon) \to 0$.

Application 1 (Relative frequency.)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space associated with a random experiment, and let $A \in \mathcal{F}$ be an event. How can we estimate the probability $P(A)$?*

The intuition behind the answer is this. The probability $P(A)$ can be estimated by repeating the experiment a large number of times.

We perform the experiment n times independently.

We count the number of times the event A occurs, which we denote by $N_n(A)$.

The relative frequency $p_n = \frac{N_n(A)}{n}$ should approximate the true probability $P(E)$.

Then the relative frequency converges almost surely to the true probability:

$$\frac{N_n(A)}{n} \xrightarrow{a.s.} P(A) = p.$$

Consider an infinite sequence of independent repetitions of the experiment. This can be described by the product probability space:

$$(\Omega^\infty, \mathcal{F}^\infty, P^\infty) = \left(\prod_{k=1}^{\infty}\Omega, \prod_{k=1}^{\infty}\mathcal{F}, \prod_{k=1}^{\infty}P\right).$$

We define a sequence of indicator random variables $X_n : \Omega^\infty \to \mathbb{R}$ for each repetition:

$$X_n(\omega_1, \omega_2, \ldots) = \mathbf{I}_A(\omega_n) = \begin{cases} 1 & \text{if } \omega_n \in A \\ 0 & \text{if } \omega_n \notin A \end{cases}$$

These random variables are clearly independent because the σ-algebras $\sigma(X_n)$ are independent by construction, as $\sigma(X_n) \subset \mathcal{F}_n$ and the $\mathcal{F}_n$ are independent for distinct n.

The expected value of each variable is

$$E(X_k) = 1 \cdot P(X_k = 1) + 0 \cdot P(X_k = 0) = P(A) = p.$$

Now, let's define a new sequence of random variables $Y_k = X_k - p$. This sequence has a zero mean and can only take values $(1-p)$ or $(-p)$. Thus, Y_k^4 can only take values $(1-p)^4$ or $(-p)^4 = p^4$ andthe fourth moment is a constant that depends only on p which is clearly finite, so $E(Y_k^4) < \infty$.

Thus, the sequence Y_k satisfies the assumptions of Theorem 50

$$\frac{1}{n}\sum_{k=1}^{n} Y_k = \frac{1}{n}\sum_{k=1}^{n}(X_k - p) = \left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) - p = \frac{N_n(A)}{n} - p \xrightarrow{a.s.} 0$$

By Theorem 4, the sequence $\frac{1}{n}\sum_{k=1}^{n} Y_k$ converges almost surely to 0, i.e., $\bar{X}_n \xrightarrow{a.s.} P(A)$.

Application 2 (Empirical distribution function.)**.** *Similarly, we can estimate the value of the distribution function of some random variable X on $(\Omega, \mathcal{F}, P)$. For this purpose let's consider infinite repetition of an experiment described by the probability space $(\Omega, \mathcal{F}, P)$. For any fixed $t \in \mathbb{R}$ denote A_t the set $(\infty, t])$ and let*

$$F_{n,t} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{I}_A(X_k).$$

Clearly, $E(\mathbf{I}_A(X_k)) = P(X_k \leq t) = F(t)$. Thus $F_{n,t}$ converges almost surely to $F(t)$,

$$F_{n,t} \xrightarrow{a.s.} F(t)$$

Application 3 (Weierstrass Approximation Theorem.)**.** *Using the Law of Large Numbers to prove the Weierstrass Approximation Theorem is a classic and beautiful example of how probability theory can be used to solve problems in analysis.*

Theorem 51 (Weierstrass Approximation Theorem)**.** *Let f be a continuous function on the closed interval $[0, 1]$. For any $\epsilon > 0$, there exists a polynomial $P(x)$ such that:*

$$|f(x) - P(x)| < \epsilon \quad \text{for all } x \in [0, 1].$$

The proof works by constructing a specific polynomial, the Bernstein polynomial $B_n(x)$, that is guaranteed to converge to $f(x)$ We first define the approximating polynomials.

Definition 37. *For a continuous function f on $[0, 1]$ and $n \geq 1$, the n^{th} Bernstein polynomial of f is the function*

$$B_n f(x) = \sum_{k=0}^{n} f\left(\tfrac{k}{n}\right)\binom{n}{k} x^k (1-x)^{n-k} \text{ for each } x \in [0, 1].$$

For each n, $B_n f(x)$ is a polynomial of degree at most n, whose computation requires only the values $f(0), f(1/n), \ldots, f(1)$.

Note that $B_n f(x) = E_x[f(S_n/n)]$, where $S_n \sim Binomial(n, x)$ represents the number of successes in n independent trials, each with a probability of success $x \in [0, 1]$. This observation is crucial to proving the following approximation theorem.

Since f is bounded and uniformly continuous, there is M such that $|f(x)| \leq M$ for all $x \in [0, 1]$, and given $\epsilon > 0$ there is $\delta > 0$ such that $|f(x) - f(y)| < \epsilon/2$ whenever $|x - y| < \delta$. Then, for each

$x \in [0,1]$,

$$
\begin{aligned}
|B_n f(x) - f(x)| &= \left| E\left[f\left(\frac{S_n}{n}\right) - f(x) \right] \right| \leq E\left| f\left(\frac{S_n}{n}\right) - f(x) \right| \\
&= E[|f(S_n/n) - f(x)| \cdot I\{|S_n/n - x| \leq \delta\}] + E[|f(S_n/n) - f(x)| \cdot I\{|S_n/n - x > \delta\}] \\
&\leq \frac{\epsilon}{2} \cdot P\{|S_n/n - x| \leq \delta\} + 2M \cdot P\{|S_n/n - x| > \delta\} \\
&\leq \frac{\epsilon}{2} + 2MP\{|S_n/n - x| > \delta\} \\
&\leq \frac{\epsilon}{2} + 2M\frac{Var(S_n/n)}{\delta^2} \quad \textit{[by Chebyshev's inequality]} \\
&= \frac{\epsilon}{2} + 2M\frac{x(1-x)/n}{\delta^2} \\
&\leq \frac{\epsilon}{2} + \frac{2M}{n\delta^2}\frac{1}{4} \quad \textit{[since } x(1-x) \leq 1/4 \textit{ for all } p\textit{]} \\
&= \frac{\epsilon}{2} + \frac{M}{2n\delta^2}
\end{aligned}
$$

uniformly in x. Thus, $\sup_x |B_n f(x) - f(x)| \leq \epsilon$ for n sufficiently large.

5.4 Convergence in distribution

The final form of convergence we consider is not convergence of random variables themselves, but rather, their distribution functions.

Definition 38 (Convergence in Distribution)**.** *The sequence of random variables $(X_n)_{n=1}^{\infty}$ is said to **converge in distribution** to a random variable X if the limit of the cumulative distribution functions (CDFs) of X_n is equal to the CDF of X at all points where the latter is continuous. This is denoted by $X_n \xrightarrow{d} X$.*

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$$

for all $t \in \mathbb{R}$ at which F_X is continuous.

This definition of convergence in distribution can be cumbersome due to the condition regarding the continuity points of the limit distribution function F_X.The following criterion for convergence in distribution is superior to Definition 38 in that one need not deal with continuity points of the limit distribution function.

Definition 39. *Denote by $C_b(\mathbb{R})$ the set of all bounded, continuous functions $f : \mathbb{R} \to \mathbb{R}$.*

Theorem 52. *A sequence of random variables $(X_n)_{n=1}^{\infty}$ converges in distribution to a random variable X if and only if $E[f(X_n)] \to E[f(X)]$ for every function $f \in C_b(\mathbb{R})$.*

This theorem, often referred to as a key part of the Portmanteau Theorem, provides a powerful and convenient way to check for convergence in distribution. It transforms the problem from checking a limit at every continuity point of a possibly complex distribution function to checking the convergence of expectations for a large, but manageable, class of test functions.

Proof.

Sufficiency $\implies$ **:** Suppose that $X_n \xrightarrow{d} X$. Given $f \in C_b(\mathbb{R})$, let $M = \sup_x |f(x)|$, which is finite. Then, given $\epsilon > 0$, choose $K > 0$ such that $\pm K$ are continuity points of F_X and such that $P\{|X| > K\} < \frac{\epsilon}{M}$. This is possible since F_X has at most a countable number of discontinuities and since $\lim_{x\to\infty} P\{|X| > x\} = 0$.

By convergence in distribution, $P\{|X_n| > K\} < \frac{2\epsilon}{M}$ for all large values of n. With K and ϵ

remaining fixed, there is a step function $g = \sum_{i=1}^{k} a_i 1_{(x_{i-1},x_i]}$, with $-K = x_0 < \cdots < x_k = K$, such that each x_i is a continuity point of F_X and $\sup_{x\in[-K,K]} |f(x) - g(x)| < \epsilon$. Then, for n sufficiently large,

$$\begin{aligned}
|E[f(X_n)] - E[f(X)]| &\le |E[f(X_n); \{|X_n| \le K\}] - E[f(X); \{|X| \le K\}]| \\
&\quad + E[|f(X_n)|; \{|X_n| > K\}] + E[|f(X)|; \{|X| > K\}] \\
&\le |E[f(X_n); \{|X_n| \le K\}] - E[f(X); \{|X| \le K\}]| + 3\epsilon \\
&\le 3\epsilon + |E[f(X_n); \{|X_n| \le K\}] - E[g(X_n)]| \\
&\quad + |E[f(X); \{|X| \le K\}] - E[g(X)]| + |E[g(X_n)] - E[g(X)]| \\
&\le 5\epsilon + |E[g(X_n)] - E[g(X)]|.
\end{aligned}$$

But, because $X_n \xrightarrow{d} X$ and the x_i are continuity points of F_X,

$$E[g(X_n)] = \sum_{i=1}^{k} a_i [F_{X_n}(x_i) - F_{X_n}(x_{i-1})] \to \sum_{i=1}^{k} a_i [F_X(x_i) - F_X(x_{i-1})] = E[g(X)],$$

and, hence, $E[f(X_n)] \to E[f(X)]$.

Necessity $\Longleftarrow$: Suppose that $E[f(X_n)] \to E[f(X)]$ for every $f \in C_b(\mathbb{R})$.

Given a continuity point t of F_X and $\epsilon > 0$, there exists $f \in C_b(\mathbb{R})$ such that

$$1_{(-\infty,t]} \le f \le 1_{(-\infty,t+\epsilon]}.$$

For example, for some m one may take

$$f(x) = \begin{cases} 1 & x \le t \\ 1 - m(x-t) & x \in [t, t+\epsilon] \\ 0 & x \ge t+\epsilon \end{cases}$$

Then, since

$$F_{X_n}(t) = E[1_{(-\infty,t]}(X_n)] \le E[f(X_n)]$$

and

$$E[f(X)] \le E[1_{(-\infty,t+\epsilon]}(X)] = F_X(t+\epsilon),$$

letting $n \to \infty$ gives

$$\limsup_{n\to\infty} F_{X_n}(t) \le \lim_{n\to\infty} E[f(X_n)] = E[f(X)] \le E[1_{(-\infty,t+\epsilon]}(X)] = F_X(t+\epsilon).$$

Letting $\epsilon \downarrow 0$ and invoking the right-continuity of F_X yields

$$\limsup_{n\to\infty} F_{X_n}(t) \le F_X(t).$$

Next, given $\epsilon > 0$, let $f \in C_b(\mathbb{R})$be such that $1_{(-\infty,t-\epsilon]} \le f \le 1_{(-\infty,t]}$. Then,

$$F_X(t-\epsilon) = E[1_{(-\infty,t-\epsilon]}(X)] \le E[f(X)] = \lim_{n\to\infty} E[f(X_n)] \le \liminf_{n\to\infty} F_{X_n}(t),$$

and consequently, because t is a continuity point of F_X,

$$F_X(t) = F_X(t-) \leq \liminf_{n\to\infty} F_{X_n}(t).$$

Combining the two inequalities, we have:

$$F_X(t) \leq \liminf_{n\to\infty} F_{X_n}(t) \leq \limsup_{n\to\infty} F_{X_n}(t) \leq F_X(t),$$

which completes the proof that $F_{X_n}(t) \to F_X(t)$.

Example 36. *A sequence of random variables, $(X_n)_{n=1}^{\infty}$, with $S_{X_n} = \{1, 2, 3, ..., n\}$, $P(X_n = k) = \frac{1}{n}$, and $Y_n = \frac{X_n}{n}$. Show that $Y_n \xrightarrow{D} Y$, where $Y \sim U[0, 1]$.*

To show that Y_n converges in distribution to Y, we need to prove that the cumulative distribution function (CDF) of Y_n, denoted as $F_{Y_n}(t)$, converges to the CDF of Y, $F_Y(t)$, at all continuity points of $F_Y(t)$. The CDF of a uniform random variable on $[0, 1]$ is:

$$F_Y(t) = \begin{cases} 0 & \text{if } t < 0 \\ t & \text{if } 0 \leq t < 1 \\ 1 & \text{if } t \geq 1 \end{cases}$$

Since $F_Y(t)$ is continuous everywhere, we must show that $\lim_{n\to\infty} F_{Y_n}(t) = F_Y(t)$ for all $t \in \mathbb{R}$.

The CDF of Y_n is given by $F_{Y_n}(t) = P(Y_n \leq t) = P\left(\frac{X_n}{n} \leq t\right) = P(X_n \leq nt)$. We consider three cases based on the value of t.

Case 1: $t < 0$ For any $t < 0$, the value nt is negative. Since the support of X_n is $\{1, 2, ..., n\}$, which contains only positive integers, the event $\{X_n \leq nt\}$ is impossible.

$$F_{Y_n}(t) = P(X_n \leq nt) = 0.$$

Thus, $\lim_{n\to\infty} F_{Y_n}(t) = 0$, which matches $F_Y(t)$ for $t < 0$.

Case 2: $t \geq 1$ For any $t \geq 1$, we have $nt \geq n$. Since the maximum possible value of X_n is n, the event $\{X_n \leq nt\}$ is certain.

$$F_{Y_n}(t) = P(X_n \leq nt) = 1.$$

Thus, $\lim_{n\to\infty} F_{Y_n}(t) = 1$, which matches $F_Y(t)$ for $t \geq 1$.

Case 3: $0 \leq t < 1$ For $0 \leq t < 1$, the value of nt is between 0 and n. The event $\{X_n \leq nt\}$ corresponds to X_n taking on integer values from 1 up to $\lfloor nt \rfloor$. Since each of these outcomes has a probability of $\frac{1}{n}$, the CDF is:

$$F_{Y_n}(t) = P(X_n \leq nt) = \sum_{k=1}^{\lfloor nt \rfloor} P(X_n = k) = \sum_{k=1}^{\lfloor nt \rfloor} \frac{1}{n} = \frac{\lfloor nt \rfloor}{n}.$$

To find the limit as $n \to \infty$, we use the property of the floor function: $x - 1 < \lfloor x \rfloor \leq x$. Applying this to nt, we get:

$$\frac{nt - 1}{n} < \frac{\lfloor nt \rfloor}{n} \leq \frac{nt}{n}$$

$$t - \frac{1}{n} < \frac{\lfloor nt \rfloor}{n} \leq t.$$

As $n \to \infty$, the lower bound $t - \frac{1}{n}$ converges to t, and the upper bound is already t. By the Squeeze Theorem, the limit of the middle term must also be t.

$$\lim_{n\to\infty} \frac{\lfloor nt \rfloor}{n} = t.$$

This matches $F_Y(t)$ for $0 \leq t < 1$.

Theorem 53. *If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$.*

Proof. Let t be a continuity point of the cumulative distribution function F_X of X. Our goal is to show that $F_{X_n}(t) \to F_X(t)$ as $n \to \infty$. To do this, we will prove that

$$F_X(t) \leq \liminf_{n\to\infty} F_{X_n}(t) \leq \limsup_{n\to\infty} F_{X_n}(t) \leq F_X(t).$$

We will use the following two facts:

1. Convergence in probability implies that for any $\epsilon > 0$, we have $P\{|X_n - X| > \epsilon\} \to 0$ as $n \to \infty$. This is the definition of convergence in probability.

2. The cumulative distribution function F_X is right-continuous.

Part 1: Proving $\limsup_{n\to\infty} F_{X_n}(t) \leq F_X(t)$

For any $\epsilon > 0$, we can write $F_{X_n}(t)$ using the law of total probability based on the event $\{|X_n - X| \leq \epsilon\}$:

$$\begin{aligned} F_{X_n}(t) &= P\{X_n \leq t\} \\ &= P\{X_n \leq t, |X_n - X| \leq \epsilon\} + P\{X_n \leq t, |X_n - X| > \epsilon\} \end{aligned}$$

The second term, $P\{X_n \leq t, |X_n - X| > \epsilon\}$, is bounded by $P\{|X_n - X| > \epsilon\}$.

$$\begin{aligned} F_{X_n}(t) &= P\{X_n \leq t\} \\ &= P\{X_n \leq t, |X_n - X| \leq \epsilon\} + P\{|X_n - X| > \epsilon\} \end{aligned}$$

For the first term, the condition $|X_n - X| \leq \epsilon$ is equivalent to $-\epsilon \leq X_n - X \leq \epsilon$, which implies $X - \epsilon \leq X_n \leq X + \epsilon$. If we have $X_n \leq t$ then $X \leq X_n + \epsilon \leq t + \epsilon$. Therefore,

$$P\{X_n \leq t, |X_n - X| \leq \epsilon\} \leq P\{X \leq t + \epsilon\} = F_X(t + \epsilon).$$

Combining these inequalities, we get:

$$F_{X_n}(t) \leq F_X(t + \epsilon) + P\{|X_n - X| > \epsilon\}$$

Taking the limit superior as $n \to \infty$, we use the fact that $X_n \xrightarrow{p} X$, which implies $P\{|X_n - X| > \epsilon\} \to 0$:

$$\limsup_{n\to\infty} F_{X_n}(t) \leq F_X(t + \epsilon)$$

This inequality holds for any $\epsilon > 0$. Since F_X is right-continuous, we can let $\epsilon \downarrow 0$:

$$\limsup_{n\to\infty} F_{X_n}(t) \leq \lim_{\epsilon\downarrow 0} F_X(t + \epsilon) = F_X(t)$$

Part 2: Proving $\liminf_{n\to\infty} F_{X_n}(t) \geq F_X(t)$

For any $\epsilon > 0$, we use a similar approach, but we start with $F_X(t - \epsilon)$:

$$\begin{aligned} F_X(t - \epsilon) &= P\{X \leq t - \epsilon\} \\ &= P\{X \leq t - \epsilon, |X_n - X| \leq \epsilon\} + P\{X \leq t - \epsilon, |X_n - X| > \epsilon\} \\ &= P\{X \leq t - \epsilon, |X_n - X| \leq \epsilon\} + P\{|X_n - X| > \epsilon\} \end{aligned}$$

For the first term, the condition $X \leq t - \epsilon$ and $|X_n - X| \leq \epsilon$ implies $X_n \leq X + \epsilon \leq (t - \epsilon) + \epsilon = t$. Thus,

$$P\{X \leq t - \epsilon, |X_n - X| \leq \epsilon\} \leq P\{X_n \leq t\} = F_{X_n}(t).$$

Combining these gives:

$$F_X(t - \epsilon) \leq F_{X_n}(t) + P\{|X_n - X| > \epsilon\}$$

Rearranging the inequality, we get:

$$F_{X_n}(t) \geq F_X(t - \epsilon) - P\{|X_n - X| > \epsilon\}$$

Taking the limit inferior as $n \to \infty$, we again use the fact that $P\{|X_n - X| > \epsilon\} \to 0$:

$$\liminf_{n\to\infty} F_{X_n}(t) \geq F_X(t - \epsilon)$$

Since t is a continuity point of F_X, we can let $\epsilon \downarrow 0$ and the right side converges to $F_X(t)$:

$$\liminf_{n\to\infty} F_{X_n}(t) \geq \lim_{\epsilon\downarrow 0} F_X(t - \epsilon) = F_X(t)$$

From Part 1 and Part 2, we have shown:

$$F_X(t) \leq \liminf_{n\to\infty} F_{X_n}(t) \leq \limsup_{n\to\infty} F_{X_n}(t) \leq F_X(t)$$

This implies that $\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$ for every continuity point t of F_X. This is the definition of convergence in distribution.

5.4.1 The Central Limit Theorem

The Central Limit Theorem (CLT) is a cornerstone of statistical and probability theory. In essence, it asserts that the ditribution of the sum of a sufficiently large collection of independent and identically distributed (i.i.d.) random variables, regardless of their individual underlying distribution, converges to the distribution to a Normal variable.

Theorem 54 (The Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables, where each variable possesses a finite expected value μ and a finite, positive variance σ^2. Then, the distribution of the standardized sum*

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

converges to the standard normal distribution as $n \to \infty$. That is, for any real number a,

$$\lim_{n\to\infty} P\left(Z_n \leq a\right) = \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$$

where $\Phi(a)$ is the cumulative distribution function (CDF) of the standard normal variable $Z \sim N(0, 1)$.

The rigorous demonstration of the CLT relies critically on the property that the convergence of the sequence of Moment Generating Functions (MGFs) implies the convergence of the corresponding cumulative distribution functions (CDFs). This relationship is formalized by the following lemma, often referred to as the Continuity Theorem for MGFs.

Lemma 5. *Let $Z_1, Z_2, \ldots$ be a sequence of random variables with distribution functions F_{Z_n} and Moment Generating Functions $M_{Z_n}(t)$, for $n \geq 1$. Let Z be a random variable with CDF F_Z and*

MGF $M_Z(t)$. If $M_{Z_n}(t)$ converges to $M_Z(t)$ for all t in some open interval containing 0, then $F_{Z_n}(t)$ converges to $F_Z(t)$ for all continuity points t of F_Z.

Since the MGF of a standard normal random variable is $M_Z(t) = e^{t^2/2}$, the objective of the proof is to establish that the MGF of the standardized sum Z_n converges to $e^{t^2/2}$ as $n \to \infty$.

Proof. Let X be a representative random variable from the sequence, with $M_X(t)$ being its MGF. We first simplify the analysis by assuming $\mu = 0$ and $\sigma^2 = 1$. The standardized sum reduces to:

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

The MGF of Z_n is obtained using the properties of MGFs for independent sums and linear scaling:

$$M_{Z_n}(t) = E\left[\exp\left(tZ_n\right)\right] = E\left[\exp\left(\frac{t}{\sqrt{n}} \sum_{i=1}^{n} X_i\right)\right] = \left[M_X\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

We analyze the limit of the logarithm of $M_{Z_n}(t)$. Let $L(t) = \ln M_X(t)$. We aim to show that $\lim_{n\to\infty} nL(t/\sqrt{n}) = t^2/2$.

The derivatives of $L(t)$ evaluated at $t = 0$ are critical, as they relate to the moments of X:

$$\begin{aligned} L(0) &= \ln M_X(0) = 0 \\ L'(0) &= \mu = 0 \quad \text{(Since we assumed } \mu = 0\text{)} \\ L''(0) &= \sigma^2 = 1 \quad \text{(Since we assumed } \sigma^2 = 1 \text{ and } \mu = 0\text{)} \end{aligned}$$

We rewrite the limit in an indeterminate form suitable for L'Hôpital's Rule:

$$\lim_{n\to\infty} nL\left(\frac{t}{\sqrt{n}}\right) = \lim_{n\to\infty} \frac{L(t/\sqrt{n})}{1/n} \quad \left(\frac{0}{0} \text{ form}\right)$$

Applying L'Hôpital's Rule (differentiating with respect to n):

$$\lim_{n\to\infty} \frac{\frac{d}{dn}L(t/\sqrt{n})}{\frac{d}{dn}(1/n)} = \lim_{n\to\infty} \frac{L'(t/\sqrt{n}) \cdot t \cdot (-\frac{1}{2}n^{-3/2})}{(-n^{-2})} = \lim_{n\to\infty} \frac{\frac{t}{2}L'(t/\sqrt{n})}{n^{-1/2}} \quad \left(\frac{0}{0} \text{ form}\right)$$

Applying L'Hôpital's Rule a second time:

$$= \lim_{n\to\infty} \frac{\frac{t}{2} \cdot L''(t/\sqrt{n}) \cdot t \cdot (-\frac{1}{2}n^{-3/2})}{-\frac{1}{2}n^{-3/2}}$$

Canceling the common factor $(-\frac{1}{2}n^{-3/2})$:

$$= \lim_{n\to\infty} \frac{t^2}{2} L''\left(\frac{t}{\sqrt{n}}\right)$$

Since $L''(t)$ is continuous at $t = 0$:

$$= \frac{t^2}{2} L''(0) = \frac{t^2}{2}(1) = \frac{t^2}{2}$$

Thus, $\lim_{n\to\infty} \ln M_{Z_n}(t) = t^2/2$, which yields $\lim_{n\to\infty} M_{Z_n}(t) = e^{t^2/2}$. By the Continuity Theorem, the CDF of Z_n converges to the standard normal CDF.

For variables with general parameters μ and σ^2, the result follows by considering the standardized variables $X_i^* = (X_i - \mu)/\sigma$. These variables are i.i.d. with $E[X_i^*] = 0$ and $\text{Var}(X_i^*) = 1$. The standardized sum Z_n is equivalent to the standardized sum of the X_i^*:

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i^*$$

The proof then applies directly to the X_i^* sequence, establishing the theorem for the general case.

Example 37. *The number of students who enroll in a course, denoted by X, is a Poisson random variable with mean $\lambda = 100$. The professor will teach the course in two separate sections if the number enrolling is 120 or more. Calculate the approximate probability that the professor will have to teach two sections, $P\{X \geq 120\}$, using the Central Limit Theorem.*

A key property of the Poisson distribution is its infinite divisibility. Specifically, a Poisson random variable X with mean λ can be rigorously represented as the sum of n independent and identically distributed Poisson random variables, each with mean λ/n.

$$X \sim Poisson(\lambda = 100) \quad \Longrightarrow \quad X = \sum_{i=1}^{100} X_i, \quad \text{where } X_i \sim Poisson(1)$$

Since X is the sum of a large number ($n = 100$) of i.i.d. random variables, the Central Limit Theorem applies, and the distribution of X can be accurately approximated by a Normal distribution. [Image of Poisson Distribution approximating Normal Distribution]

For a Poisson distribution, the mean and the variance are both equal to λ:

- *Mean:* $E[X] = E[\sum_{i=1}^{100} X_i] = 100E[X_1] = 100.1 = \lambda$.
- *Variance:* $Var(X) = Var[\sum_{i=1}^{100} X_i] = 100\,Var[X_1] = \lambda = 100$.

The CLT establishes that the distribution of $Z = \frac{X - E[X]}{\sqrt{Var(X)}}$ can be approximated with $N(0, 1)$. We calculate $P\{X \geq 120\}$ by standardizing the discrete value $x = 120$ directly:

$$P\{X \geq 120\} \approx P\left\{Z \geq \frac{120 - E[X]}{\sqrt{Var(X)}}\right\}$$

$$P\{X \geq 120\} \approx P\left\{Z \geq \frac{120 - 100}{10}\right\} = P\left\{Z \geq \frac{20}{10}\right\} = P\{Z \geq 2.0\}$$

$$P\{Z \geq 2.0\} = 1 - P\{Z < 2.0\} = 1 - \Phi(2.0)$$

Using the standard normal CDF $\Phi(a)$:

$$\Phi(2.0) \approx 0.9772$$

The approximate probability is:

$$P\{X \geq 120\} \approx 1 - 0.9772 = \mathbf{0.0228}$$

5.4.2 Slutsky's Theorem

Slutsky's Theorem is a powerful result that extends the applicability of the CLT by allowing for the combination of random variables that converge in distribution (like the standardized mean from the CLT) with those that converge merely in probability .

Theorem 55 (Slutsky's Theorem)**.** *Suppose that X_n is a sequence of random variables such that X_n converges in distribution to X, and Y_n is a sequence of random variables such that Y_n converges in probability to a constant c . Then:*

1. $X_n + Y_n \xrightarrow{d} X + c$
2. $X_n Y_n \xrightarrow{d} cX$

Proof. We will provide a sketch of the proof for the case, $X_n + Y_n \xrightarrow{d} X + c$.

The definition of convergence in distribution, $X_n \xrightarrow{d} X$, implies that $P(X_n \leq a) \to P(X \leq a)$ at continuity points a of the CDF of X. The definition of convergence in probability, $Y_n \xrightarrow{p} c$, implies that for any $\epsilon > 0$, $\lim_{n\to\infty} P(|Y_n - c| > \epsilon) = 0$.

Let $F_n(a) = P(X_n + Y_n \leq a)$ be the CDF of $X_n + Y_n$. We want to show $\lim_{n\to\infty} F_n(a) = P(X \leq a - c)$. For any $\epsilon > 0$:

$$P(X_n + Y_n \leq a) = P(X_n + Y_n \leq a, |Y_n - c| \leq \epsilon) + P(X_n + Y_n \leq a, |Y_n - c| > \epsilon)$$

Since $|Y_n - c| \leq \epsilon$ is equivalent to $c - \epsilon \leq Y_n \leq c + \epsilon$, we have:

$$P(X_n + Y_n \leq a) \leq P(X_n + c + \epsilon \leq a) + P(|Y_n - c| > \epsilon) = P(X_n \leq a - c - \epsilon) + P(|Y_n - c| > \epsilon)$$

As $n \to \infty$, $P(|Y_n - c| > \epsilon) \to 0$ (due to $Y_n \xrightarrow{p} c$), and $P(X_n \leq a - c - \epsilon) \to P(X \leq a - c - \epsilon)$ (due to $X_n \xrightarrow{d} X$). Thus, $\limsup_{n\to\infty} P(X_n + Y_n \leq a) \leq P(X \leq a - c - \epsilon)$.

Similarly, we can establish a lower bound:

$$P(X_n + Y_n \leq a) \geq P(X_n + Y_n \leq a, |Y_n - c| \leq \epsilon)$$

$$P(X_n + Y_n \leq a) \geq P(X_n + c - \epsilon \leq a) - P(|Y_n - c| > \epsilon) = P(X_n \leq a - c + \epsilon) - P(|Y_n - c| > \epsilon)$$

As $n \to \infty$:

$$\liminf_{n\to\infty} P(X_n + Y_n \leq a) \geq P(X \leq a - c + \epsilon)$$

Since this holds for any $\epsilon > 0$, by letting $\epsilon \to 0$ and assuming continuity of the CDF of X, we conclude that:

$$\lim_{n\to\infty} P(X_n + Y_n \leq a) = P(X \leq a - c)$$

which means $X_n + Y_n \xrightarrow{d} X + c$.

5.4.3 The Delta Method

The Delta Method is a general technique used to determine the asymptotic distribution of a function of an asymptotically normal estimator. It utilizes the Taylor series expansion to linearize the function, making it indispensable for finding the limiting variance of complex statistics.

Theorem 56 (The Delta Method)**.** *Suppose that $\sqrt{n}(T_n - \theta)$ converges in distribution to a Normal random variable $N(0, \sigma^2)$, where T_n is an estimator for θ. That is:*

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2)$$

Let g be a function such that $g'(\theta)$ exists and is non-zero. Then, the distribution of the transformed

statistic converges as follows:

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} N(0, [g'(\theta)]^2\sigma^2)$$

In simpler terms, if the estimator T_n is asymptotically Normal, then the transformed estimator $g(T_n)$ is also asymptotically Normal, and its variance is scaled by the square of the derivative of the transformation function, $g'(\theta)$.

Proof. The proof relies on the first-order Taylor expansion of the function $g(x)$ around the true parameter θ:

$$g(T_n) = g(\theta) + g'(\theta)(T_n - \theta) + R_n$$

where R_n is the remainder term. The crucial insight is that for consistent estimators T_n, the remainder term R_n is negligible in the asymptotic limit. Specifically, we can write the Taylor expansion such that the remainder is $o_p(n^{-1/2})$:

$$g(T_n) - g(\theta) = g'(\theta)(T_n - \theta) + o_p(T_n - \theta)$$

Since $T_n \xrightarrow{p} \theta$, we can rewrite the equation by multiplying by $\sqrt{n}$:

$$\sqrt{n}(g(T_n) - g(\theta)) = g'(\theta)\left(\sqrt{n}(T_n - \theta)\right) + \sqrt{n} \cdot o_p(T_n - \theta)$$

By definition of o_p, the term $\sqrt{n} \cdot o_p(T_n - \theta)$ converges to 0 in probability.

We are given the initial convergence:

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2)$$

Let $Z_n = \sqrt{n}(T_n - \theta)$. By the Continuous Mapping Theorem, multiplying by the constant $g'(\theta)$ preserves the distributional convergence, scaling the variance:

$$g'(\theta)Z_n \xrightarrow{d} g'(\theta)N(0, \sigma^2) = N(0, [g'(\theta)]^2\sigma^2)$$

Since the remainder term converges to zero in probability, $R_n^* = \sqrt{n} \cdot o_p(T_n - \theta) \xrightarrow{p} 0$.

Applying Slutsky's Theorem (part 1, where $X_n = g'(\theta)Z_n$ and $Y_n = R_n^*$):

$$\sqrt{n}(g(T_n) - g(\theta)) = g'(\theta)Z_n + R_n^* \xrightarrow{d} N(0, [g'(\theta)]^2\sigma^2) + 0$$

Thus,

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} N(0, [g'(\theta)]^2\sigma^2)$$