Contents

1 Combinatorics

1.1 Basic Counting Principles

An important motivation to study combinatorics is to count the **number of ways** in which an event may occur. Intuitively, we have two approaches to count.

The first approach is to categorise the event into **non-overlapping cases**. This means that we break an event into mutually exclusive sub-events, after which we can count the number of ways for each sub-event to occur. The agregate of these counts is the total number of ways for the original event to occur.

Those familiar with basic set theory may consider E to be the set containing all distinct ways for an event to occur. By breaking up the event, we essentially establish a **partition** of E, so that the sum of cardinalities of all the elements in that partition equals the cardinality of E.

This motivates us to write the following principle using set notations.

Theorem 1.1.1 ▸ Addition Principle (AP)

Let $k \in \mathbb{N}^+$ and let $A_1, A_2, \cdots, A_k$ be k finite sets which are pairwise disjoint, i.e. for all i, j such that $1 \leq i, j \leq k$, $A_i \cap A_j = \varnothing$ whenever $i \neq j$, then

$$\left|\bigcup_{i=1}^{k} A_i\right| = \sum_{i=1}^{k} |A_i|.$$

Proof. The case where $k = 1$ is trivial.
Suppose that when $k = n$, we have

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{i=1}^{n} |A_i|$$

for any n finite sets which are pairwise disjoint. Let A_{n+1} be an arbitrary finite set

which is disjoint with any of the A_i's from the n sets. So we have:

$$\begin{aligned}\left|\bigcup_{i=1}^{n+1} A_i\right| &= \left|\left(\bigcup_{i=1}^{n} A_i\right) \cup A_{n+1}\right| \\ &= \left|\bigcup_{i=1}^{n} A_i\right| + |A_{n+1}| - \left|\left(\bigcup_{i=1}^{n} A_i\right) \cap A_{n+1}\right| \\ &= \left(\sum_{i=1}^{n} |A_i|\right) + |A_{n+1}| - |\varnothing| \\ &= \sum_{i=1}^{n+1} |A_i|.\end{aligned}$$

Therefore, the original statement holds for all $k \in \mathbb{N}^+$. □

In more casual language, this means that if an event E_k has n_k distinct ways to occur, then there is $\sum_{i=1}^{k} n_k$ ways for at least one of the events $E_1, E_2, \cdots, E_k$ to occur, provided that E_i and E_j can never occur concurrently whenever $i \neq j$.

Given an event E, the other approach to count the number of ways for it to occur is to break E up internally into **non-overlapping stages**.

With set notations, we can write the i-th stage for E to occur as e_i, and so a way for E to occur can be represented by an ordered tuple $(e_1, e_2, \cdots, e_k)$, where k is the total number of stages to undergo for E to occur.

Let E_i denote the set of all distinct ways to undergo the i-th stage of E, then it is easy to see that E is just the **Cartesian product** of all the E_i's. Hence, we derive the following principle:

Theorem 1.1.2 ▸ Multiplication Principle (MP)

Let $k \in \mathbb{N}^+$ and let $A_1, A_2, \cdots, A_k$ be k pairwise disjoint finite sets, then

$$\left|\prod_{i=1}^{k} A_i\right| = \prod_{i=1}^{k} |A_i|.$$

Proof. The case where $k = 1$ is trivial.
Suppose that when $k = n$, we have

$$\left|\prod_{i=1}^{n} A_i\right| = \prod_{i=1}^{n} |A_i|$$

for any n finite sets which are pairwise disjoint. Let A_{n+1} be an arbitrary finite set which is disjoint with any of the A_i's from the n sets. Take $a_i, a_j \in A_{n+1}$. Note that for all $\boldsymbol{a} \in \prod_{i=1}^{n} A_i$, $(\boldsymbol{a}, a_i) \neq (\boldsymbol{a}, a_j)$ whenever $a_i \neq a_j$. This means that

$$\begin{aligned}\left|\prod_{i=1}^{n+1} A_i\right| &= \left|\prod_{i=1}^{n} A_i \times A_{n+1}\right| \\ &= \left|\prod_{i=1}^{n} A_i\right| |A_{n+1}| \\ &= \left(\prod_{i=1}^{n} |A_i|\right) |A_{n+1}| \\ &= \prod_{i=1}^{n+1} |A_i|\end{aligned}$$

Therefore, the original statement holds for all $k \in \mathbb{N}^+$. □

In more casual language, this means that if an event E requires k stages to be undergone before it occurs and the i-th stage has n_i ways to complete, then there is $\prod_{i=1}^{k} n_k$ ways for E to occur, provided that no two different stages complete concurrently.

Next, consider the following naïve question:

> If $|A| = a$ and $|B| = b$, what is $|A \cup B|$?

The solution is simple: we first count how many elements are in A and then count how many elements are in B, but note that there are some elements which might be in both sets and are thus counted twice, so we need to subtract these elements away.

Proposition 1.1.3 ▸ Cardinality of the Union of Two Finite Sets

Let A and B be two finite sets, then

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

In particular, if A and B are disjoint, then $|A \cup B| = |A| + |B|$.

Proposition 1.5.1 can be generalised for a finite sequence of finite sets.

Theorem 1.1.4 ▶ Principle of Inclusion and Exclusion (PIE)

Let $E_1, E_2, \cdots, E_n$ be a sequence of finite sets. In general, we have

$$\left|\bigcup_{i=1}^{n} E_i\right| = \sum_{j=1}^{n}\left[(-1)^{j+1}\left(\sum_{1\le k_1\le k_2\le\cdots\le k_j\le n}\left|\bigcap_{r=1}^{j} E_{k_r}\right|\right)\right].$$

Proof. Define a function $f_S : S \to \{0,1\}$ by

$$f_S(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{if } x \notin S \end{cases}.$$

Let $E = \bigcup_{i=1}^{n} E_i$. Consider

$$\prod_{i=1}^{n}\left(f_E(x) - f_{E_i}(x)\right). \tag{*}$$

For any x, if $x \in E$, then $x \in E_k$ for some $k \in \{x \in \mathbb{N} : x \le n\}$, which means that $f_E(x) - f_{E_k}(x) = 0$; if $x \notin E$, then $f_E(x) = f_{E_i}(x) = 0$ for all $i \in \{x \in \mathbb{N} : x \le n\}$. In either case, (*) is identically 0.

Note that for any sets A, B, $f_{A\cap B}(x) = f_A(x)f_B(x)$. In particular, $f_{E\cap E_i}(x) = f_{E_i}(x)$ for all $i = 1, 2, \cdots, n$. Expanding (*), we have

$$f_E(x) = \left[f_E(x)\right]^n = \sum_{j=1}^{n}\left\{(-1)^{j+1}\left[\sum_{1\le k_1\le k_2\le\cdots\le k_j\le n}\left(\prod_{r=1}^{j} f_{E_{k_r}}(x)\right)\right]\right\}.$$

Summing up $f_E(x)$ over all x leads to the identity in the original statement. □

We can generalise this further. Note that every set is uniquely determined by some formula ϕ with some parameter p. This means that the statement $x \in X$ is logically equivalent to "x satisfies a property $\phi(x, p)$". This leads to the general principle of inclusion and exclusion.

Theorem 1.1.5 ▶ General Principle of Inclusion and Exclusion (GPIE)

Let S be a finite set with $|S| = n$ and let $\{P_1, P_2, \cdots P_q\}$ be a set of properties. Suppose that $E(m)$ is the number of elements in S that satisfy exactly m of the q properties and that $\omega(P_{i_1}P_{i_2}\cdots P_{i_m})$ is the number of elements in S that satisfy exactly $P_{i_1}, P_{i_2}, \cdots, P_{i_m}$, such that

$$\omega(m) = \sum \omega(P_{i_1}P_{i_2}\cdots P_{i_m}),$$

then

$$E(m) = \sum_{k=m}^{q} \left[(-1)^{k-m} \binom{k}{m} \omega(k) \right].$$

Proof. Let $x \in S$ be an arbitrary element which satisfies exactly t properties. If $t < m$ or $q < t$, then x contributes a count of 0 to both sides. If $t = m$, then x contributes a count of 1 to $E(m)$ and to $\omega(m)$ and contributes a count of 0 to $\omega(k)$ for all $k > m$.

If $q \geq t > m$, then x contributes a count of 0 to the left-hand side. On the right-hand side, x is counted $\binom{t}{k}$ times for $k = m, m+1, \cdots, t$. So the total contribution of x to the right-hand side is

$$\begin{aligned}
\sum_{k=m}^{t} \left[(-1)^{k-m} \binom{t}{k} \binom{k}{m} \right] &= \sum_{k=m}^{t} \left[(-1)^{k-m} \binom{t}{m} \binom{t-m}{k-m} \right] \\
&= \sum_{k=0}^{t-m} \left[(-1)^{k} \binom{t}{m} \binom{t-m}{k} \right] \\
&= \binom{t}{m} \sum_{k=0}^{t-m} \left[(-1)^{k} \binom{t-m}{k} \right] \\
&= 0.
\end{aligned}$$

Therefore, every $x \in S$ contributes an equal count to both sides of the identity, so the identity holds. □

In particular, we see that

$$E(0) = \sum_{i=0}^{q} \left[(-1)^{i} \omega(i) \right].$$

The PIE can be used to study derangements, the concept of which arises from the following question:

> Given an increasing sequence of consecutive integers from 1 to n inclusive. How many permutations of the sequence are there such that the integer at index i in the permuted sequence is not i?

Here we give a rigorous definition along with a generalised notion of *fixed points*.

Definition 1.1.6 ▸ Derangement

Let $N_n = [n] - \{0\}$. A **derangement** of N_n, denoted by D_n, is defined as the permutation $a_1a_2\cdots a_n$ of N_n such that $a_i \neq i$ for all $i = 1, 2, \cdots, n$. We say that an r-permutation of N_n has a **fixed point** at i if $a_i = i$. The number of r-permutations of N_n with k fixed points is $D(n, r, k)$.

With some careful computation, we arrive at the following formula:

Theorem 1.1.7 ▸ A Formula for $D(n, r, k)$

For $n \geq r \geq k \geq -$ and $r \geq 1$,

$$D(n,r,k) = \frac{C_k^r}{(n-r)!}\sum_{i=0}^{r-k}(-1)^i\binom{r-k}{i}(n-k-i)!.$$

Proof. Let $a_1a_2\cdots a_r$ be a permutation and let P_i be a property satisfied by a permutation with $a_i = i$, then $D(n, r, k) = E(k)$. Suppose A is a permutation with at least t fixed points, then the rest $(r-t)$ digits form an $(r-t)$-permutation of $(n-t)$ integers. Therefore,

$$\omega(t) = \binom{r}{t}\frac{(n-t)!}{(n-r)!}.$$

By Theorem 1.5.3,

$$\begin{aligned}
D(n,r,k) &= \sum_{i=0}^{r-k}\left[(-1)^i\binom{k+i}{k}\omega(k+i)\right] \\
&= \sum_{i=0}^{r-k}\left[(-1)^i\binom{k+i}{k}\binom{r}{k+i}\frac{(n-k-i)!}{(n-r)!}\right] \\
&= \frac{1}{(n-r)!}\sum_{i=0}^{r-k}\left[(-1)^i\binom{r}{k}\binom{r-k}{i}(n-k-i)!\right] \\
&= \frac{C_k^r}{(n-r)!}\sum_{i=0}^{r-k}(-1)^i\binom{r-k}{i}(n-k-i)!.
\end{aligned}$$

□

Using Theorem 1.1.7, we see that the number of derangements is just

$$D_n = D(n, n, 0) = n! \sum_{i=0}^{n} \frac{(-1)^i}{i!}.$$

Relating this to probability, we see that given a sequence of n integers, the probability of us having a derangement is given by:

$$\lim_{n\to\infty} P(\text{Having a derangement}) = \lim_{n\to\infty} \frac{D_n}{n!} = \mathrm{e}^{-1}.$$

Another application of the PIE is to investigate a standardised method to count the number of prime numbers between 2 and n inclusive. We will first introduce an ancient algorithm discovered by Greek mathematician Eratosthenes.

Technique 1.1.8 ▸ Sieve of Eratosthenes

Let $n \in \mathbb{N}$ and $n \geq 2$.

- Define L to be the list of all integers from 2 to n inclusive, arranged in ascending order.
- For every $i \in L$, while $i \leq \sqrt{n}$:
  - If i is a prime, remove all multiples of i from L.
- All remaining integers in L are primes.

To extend the problem further, we can instead consider counting the number of positive integers up to n inclusive which are **co-prime** to n.

Theorem 1.1.9 ▸ Euler φ-Function

Let $n \in \mathbb{N}^+$ and $\varphi(n)$ be the number of positive integers which are smaller than n and co-prime to n. If $p_1, p_2, \cdots, p_k$ are distinct prime numbers such that

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

for positive m_i's, then

$$\varphi(n) = n \prod_{i=1}^{k} \left(1 - \frac{1}{p_i}\right).$$

Proof. Let S_t be the set of positive integers at most n which are divisible by at least t

of the k primes. For $t = 1, 2, \cdots, k$, consider

$$|S_t| = \sum_{i_1<i_2<\cdots<i_t} \left\lfloor \frac{n}{\prod_{j=1}^{t} p_{i_j}} \right\rfloor = \sum_{i_1<i_2<\cdots<i_t} \frac{n}{\prod_{j=1}^{t} p_{i_j}}$$

By Theorem 1.5.3,

$$\begin{aligned} \varphi(n) &= n - \left| \left(\bigcup_{t=1}^{k} S_t \right) \right| \\ &= n \prod_{i=1}^{k} \left(1 - \frac{1}{p_i}\right). \end{aligned}$$

□

1.2 Permutations

A fundamental problem in combinatorics is described as follows: given a set S, how many ways are there to arrange r elements in S, i.e. how many **distinct sequences** can be formed using the elements in S without repetition? The process of selecting elements from S and arranging them as a sequence is known as **permutation**.

Note that forming a sequence using r elements from a set S is an event consisting of r stages, as we need to select an element for each of the r terms of the sequence. Suppose S has n elements. For the first term of the sequence, we can choose any of the elements in S, so there is n ways to do it. For the second term, since we cannot repeat the elements, we are left with $n - 1$ choices.

Continue choosing elements in this way, we realise that if we choose the terms sequentially, when we reach the k-th term we will be left with $n - k + 1$ options as the previous $(k - 1)$ terms have taken away $(k - 1)$ elements. By Theorem 1.1.2, we know that the number of sequences which can be formed is given by $\prod_{i=1}^{r}(n - r + i)$.

Definition 1.2.1 ▸ Permutations

Let A be a finite set such that $|A| = n$, an r-permutation of A is a way to arrange r elements of A, denoted as P_r^n and given by

$$P_r^n = \prod_{i=1}^{r}(n - r + i) = \frac{n!}{(n - r)!}.$$

With some algebraic manipulations, it is easy to derive the following formula, which we,

however, will prove in a combinatorial manner.

Proposition 1.2.2 ▸ An Identity for Permutations

Let $n, r \in \mathbb{N}$ with $r \leq n$, then $P_r^{n+1} = P_r^n + rP_{r-1}^n$.

Proof. Let $S = \{x \in \mathbb{N}^+ : x \leq n+1\}$ represent $(n+1)$ distinct objects. Consider a permutation of S:

If $n+1$ is not inside the permutation, this is equivalent to an r-permutation of $S - \{n+1\}$, so there are P_r^n such permutations.

If $n+1$ is inside the permutation, it means we need to first find an $(r-1)$-permuation of $S - \{n+1\}$, which has P_{r-1}^n ways to do. After that, we need to insert $n+1$ into each of these $(r-1)$-permutations. Note that for each of such permutations, there are r positions into which we can place $n+1$. Therefore, the total number of r-permuations of S derived in this manner is rP_{r-1}^n.

Therefore, there are $P_r^n + rP_{r-1}^n$ r-permutations of S, i.e. $P_r^{n+1} = P_r^n + rP_{r-1}^n$. □

Consider the following scenario: suppose $T_1, T_2, \cdots T_n$ are distinct labels and $a_1, a_2, \cdots, a_r$ are r distinct objects. If we wish to associate each object with one label, then each of the ways to do so is a permutation, the terms of which are taken from the T_i's. Notice that under this setting, the T_i's are allowed to be repeated in a permutation. To reason such problems, we first introduce the notion of a *multi-set*.

Definition 1.2.3 ▸ Multi-set

The **multi-set** is defined to be

$$M := \{r_1 \cdot a_1, r_2 \cdot a_2, \cdots, r_n \cdot a_n\},$$

where a_i has r_i identical copies.

Observe that in a permutation of such a multi-set M, the object a_i is repeated r_i times. This brings us to the following formula:

Proposition 1.2.4 ▸ Generalised Formula for Permutations

Let $k \in \mathbb{N}^+$ and let $A_1, A_2, \cdots, A_k$ be k distinct objects, where A_i occurs $n_i > 0$ times for $i = 1, 2, \cdots, k$, then each permutation for these objects, corresponds to a unique per-

mutation of the multi-set

$$M = \{n_1 \cdot A_1, n_2 \cdot A_2, \cdots, n_k \cdot A_k\},$$

and the total number of permutations is given by

$$\frac{\left(\sum_{i=1}^{k} n_i\right)!}{\prod_{i=1}^{k} (n_i!)}.$$

We may also consider the special case where we want to form an r-permutation with k objects, where each object can repeat an arbitrary number of times. Notice that this is equivalent to finding the permutations of the multi-set

$$M = \{\infty \cdot a_1, \infty \cdot a_2, \cdots, \infty \cdot a_k\}.$$

In this case, we can simply iterate through each of the slots in the r-permutation and choose the object to be placed there. Trivially, there are always k choices for each slot.

Proposition 1.2.5 ▸ A Special Case

Consider the multi-set

$$M = \{\infty \cdot a_1, \infty \cdot a_2, \cdots, \infty \cdot a_n\}.$$

The number of r-permutations formed using the elements from M is given by n^r.

Consider arranging n distinct objects around a circle. If the slots around the circle are uniquely labelled, this is exactly the same as permutations along a straight line.

However, if the slots are identical, i.e. we are arranging n distinct objects around a circle with identical slots, only the **relative positions** of the objects matter.

Let $\boldsymbol{x}_i$ be an arbitrary straight-line permutations of the n objects and let $\boldsymbol{y}_i$ be the corresponding circular permutation of the n objects.

Note that if we translate every element in $\boldsymbol{x}_i$ by k positions, this will result in a different straight-line permutation $\boldsymbol{x}_j$ but does not change the corresponding circular permutation because the relative positions of the objects remain unchanged.

Notice that k can take the values $0, 1, 2, \cdots, n-1$, so for the same set of n distinct objects, every circular permutation is mapped to n straight-line permutations.

Definition 1.2.6 ▶ Circular Permutations

Let A be a finite set such that $|A| = n$, a circular r permutation of A is a way to arrange r elements of A around a circular locus, denoted as Q_r^n and given by

$$Q_r^n = \frac{P_r^n}{r} = \frac{n!}{r(n-r)!}.$$

1.3 Combinations

Beside permutations, there are also occasions where we only care about which elements from a particular set are selected instead of the order of selection. Note that if we want to find a selection of r elements from a set A where the order of selected elements does not matter, it is equivalent to finding a subset of A containing r elements. This motivates us to give the following definition:

Definition 1.3.1 ▶ Combinations

Let A be a finite set such that $|A| = n$, an r-combination of A is a set $B \subseteq A$ with $|B| = r$. The number of combinations of A is given by

$$C_r^n = \frac{P_r^n}{P_r^r} = \frac{n!}{r!(n-r)!} = \binom{n}{r}.$$

Remark. Two obvious results:

1. If $r > n$ or $r < 0$, $C_r^n = 0$;
2. $C_r^n = C_{n-r}^n$.

Similar to permutations, we have the following important identity:

Theorem 1.3.2 ▶ Pascal's Triangle

Let n be an integer with $n \geq 2$ and let r be an integer with $0 \leq r \leq n$, then

$$C_r^{n+1} = C_{r-1}^n + C_r^n.$$

Proof. Let $S = \{x \in \mathbb{N}^+ : x \leq n+1\}$ be $(n+1)$ distinct objects. Consider any r-combination T of S. If $n+1 \notin T$, this is equivalent to an r-combination of $S-\{n+1\}$, so there are C_r^n such permutations. If $n+1 \in T$, it suffices to find an $(r-1)$-combination of $S-\{n+1\}$, which has C_{r-1}^n ways to do. Therefore, there are $C_r^n + C_{r-1}^n$ r-combinations of S, i.e. $C_r^{n+1} = C_r^n + C_{r-1}^n$. □

A useful application of combinations, derived directly from the definition, is to count the number of subsets for a given set which is finite. In other words, if we are given a set A with $|A| = n \in \mathbb{N}$, we wish to find a general formula for $|\mathcal{P}(A)|$. Let A_i be the set of all subsets of A whose cardinality is i, then clearly

$$|\mathcal{P}(A)| = \sum_{i=0}^{n} |A_i| = \sum_{i=0}^{n} C_i^n.$$

We can expand the above expression algebraically and realise that it simplifies to 2^n. However, in a combinatorial perspective, we can prove this result in a more succinct manner:

Theorem 1.3.3 ▸ General Formula for $|\mathcal{P}(A)|$

Let A be a finite set. If $|A| = n$, then $|\mathcal{P}(A)| = 2^n$.

Proof. Let S be an arbitrary subset of A. Consider an arbitrary element $a \in A$, then either $a \in S$ or $a \notin S$. Let $a_i \in A$ for $i = 1, 2, \cdots, n$. For all $S \in \mathcal{P}(A)$, We replace a_i by 1 if $a_i \in S$, and by 0 otherwise. Let B be the set of all binary sequences of length n. It is clear that there exists a bijection between $\mathcal{P}(A)$ and B, and so $|\mathcal{P}(A)| = |B|$. For each binary sequence of length n, each of its digits is either 0 or 1. By Theorem 1.1.2, this means that there are in total 2^n such binary sequences. Therefore,

$$|\mathcal{P}(A)| = |B| = 2^n.$$

□

Given a set of n distinct objects, we wish to find the number of combinations taken from the set where any element is allowed to be selected for multiple times. To reason about such problems, we introduce the following notion of a *multi-subset*.

Definition 1.3.4 ▸ Multi-subset

Let $M = \{\infty \cdot a_1, \infty \cdot a_2, \cdots, \infty \cdot a_n\}$ be a multi-set where $n \in \mathbb{N}$. An r-element **multi-subset** of M is the set

$$\{m_1 \cdot a_1, m_2 \cdot a_2, \cdots, m_n \cdot a_n\}$$

where each of the m_i's is a non-negative integer such that $\sum_{i=1}^{n} m_i = r$. We denote the number of r-element multi-subsets of M by H_r^n.

Intuitively, we can view each of the a_i's as a "box" which holds m_i balls. Therefore, if we can find the ways to distribute a total of r balls into the n "boxes", each of the distribution will then correspond to a multi-subset of M!

Proposition 1.3.5 ▶ A Formula for H_r^n

Let $M = \{\infty \cdot a_1, \infty \cdot a_2, \cdots, \infty \cdot a_n\}$, then the number of r-element multi-subsets of M is given by

$$H_r^n = C_r^{r+n-1}.$$

Proof. Consider a binary string formed using r zeros. Note that we can insert a number of ones into the binary string to partition the zeros into different sections. Now, suppose we insert $(n-1)$ ones into the binary string, it will result in a binary string containing r zeros and $(n-1)$ ones, such that the zeros are partitioned into n sections.

We can then number the sections using $1, 2, 3, \cdots, n$. For the i-th section, the number of zeros is recorded as m_i. Notice that in this case, each of the binary strings will correspond to a multi-subset in the form of

$$\{m_1 \cdot a_1, m_2 \cdot a_2, \cdots, m_n \cdot a_n\}.$$

One can check that this will establish a bijection between the set of all multi-subsets containing n distinct types of objects and the set of all binary strings with r zeros and $(n-1)$ ones. Therefore, the above has proven that $H_r^n = C_r^{r+n-1}$. □

1.4 Binomial and Multinomial Coefficients

In previous sections on combinations, we have already introduced the ways to choose r items from a collection of n distinct items, the number of which is given by

$$C_r^n = \binom{n}{r}.$$

In this section, we will see that there is a clear relation between the number of r-combinations to the *binomial expansion*.

Theorem 1.4.1 ▶ Binomial Expansion

Let $n \in \mathbb{N}$, then

$$(x+y)^n = \sum_{i=0}^{n} \left[\binom{n}{i} x^{n-i} y^i \right].$$

Proof. Note that

$$(x+y)^n = \prod_{i=1}^{n}(x+y).$$

Let the coefficient of $x^{n-r}y^r$ be k, where r is an integer with $0 \leq r \leq n$. Notice that k is exactly the number of ways to choose $(n-r)$ copies of x's from the distinct n terms of $(x+y)$, which is given by

$$\binom{n}{n-r} = \binom{n}{r}.$$

Therefore, summing up the terms, we have

$$(x+y)^n = \sum_{i=0}^{n}\left[\binom{n}{i}x^{n-i}y^i\right].$$

□

There are many nice algebraic properties about the binomial coefficient $\binom{n}{r}$. Beside Theorem 1.3.2, the following can be proven with a few simple algebraic manipulations:

1. $\binom{n}{r} = \frac{n}{r}\binom{n-1}{r-1}$.
2. $\binom{n}{r} = \frac{n-r+1}{r}\binom{n}{r-1}$.
3. $\binom{n}{m}\binom{m}{r} = \binom{n}{r}\binom{n-r}{m-r}$.

One important identity for binomial coefficients is as follows:

Theorem 1.4.2 ▸ Vandermonde's Identity

Let $m, n, r \in \mathbb{N}$, then

$$\sum_{i=0}^{r}\left[\binom{m}{i}\binom{n}{r-i}\right] = \binom{m+n}{r}.$$

Proof. Consider the set

$$X = \{a_1, a_2, \cdots, a_m, b_1, b_2, \cdots, b_n\}.$$

Clearly, $|X| = m+n$. Let $A \subseteq X$ such that $|A| = r$, we consider the number of such A's.

Let i be an integer with $0 \leq i \leq r$. If A contains exactly i elements from the a_j's, then it will contain exactly $(r-1)$ elements from the b_j's. Therefore, the number of A's for each $i = 1, 2, \cdots, r$ is given by

$$\binom{m}{i}\binom{n}{r-i}.$$

However, we know that the total number of such A's is just C_r^{m+n}, so we have

$$\sum_{i=0}^{r}\left[\binom{m}{i}\binom{n}{r-i}\right] = \binom{m+n}{r}.$$

□

Recall that in Theorem 1.3.2, we essentially established a recurrence relation for the binomial coefficients. Thus in theory, it is possible to generate any binomial coefficient recursively. The following theorem illustrates this.

Theorem 1.4.3 ▸ Chu Shih-Chieh (CSC) Identity

Let $r, n, k \in \mathbb{N}$ with $n \geq r$, then

$$\sum_{i=0}^{n-r}\binom{r+i}{r} = \binom{n+1}{r+1}, \qquad \sum_{i=0}^{k}\binom{r+i}{i} = \binom{r+k+1}{k}.$$

Proof. Consider the set

$$X = \{x \in \mathbb{N}^+ : x \leq n+1\}.$$

We will count the number of $(r+1)$-element subsets of X in the following way:

Fix k to be the smallest element in some $Y \subseteq X$, then it suffices to choose r elements from $X - \{x \in \mathbb{N}^+ : x \leq k\}$. Note that the maximum for k is $n-r+1$, and so there are $n-r+1$ disjoint cases. Therefore, it is easy to see that the total number of subsets is given by

$$\sum_{i=0}^{n-r}\binom{r+i}{r}.$$

However, we know that this is equivalent to choosing $(r+1)$ elements directly from

X, so

$$\sum_{i=0}^{n-r} \binom{r+i}{r} = \binom{n+1}{r+1}.$$

Note that $C_r^{r+i} = C_i^{r+i}$ and $C_{r+1}^{n+1} = C_{n-r}^{n+1}$. Therefore, by setting $k = n-r$, it is obvious by symmetry that

$$\sum_{i=0}^{k} \binom{r+i}{i} = \binom{r+k+1}{k}.$$

□

Binomial coefficients can be generalised as *multinomial coefficients*, which has a nice correspondence to a special variant of distribution problems.

Definition 1.4.4 ▸ Multinomial Coefficient

Let $m, n \in \mathbb{N}$, with $m \geq 1$. The **multinomial coefficient**

$$\binom{n}{n_1, n_2, \cdots, n_m} = \frac{n!}{\prod_{i=1}^{m} n_i!}$$

is the number of ways to distribute n distinct objects into m distinct collections such that there are exactly n_i objects in the i-th collection.

Observe that the binomial coefficients are just a special case for multinomial coefficients with $m = 2$ and $n_1 + n_2 = n$. Respectively, multinomial coefficients also generalise binomial expansion.

Theorem 1.4.5 ▸ Multinomial Expansion

For $m, n \in \mathbb{N}+$, we have

$$\left(\sum_{i=1}^{m} x_i\right)^n = \sum_{\sum_{j=1}^{m} n_j = n} \binom{n}{n_1, n_2, \cdots, n_m} \prod_{i=1}^{m} x_i^{n_i}.$$

1.5 Principle of Inclusion and Exclusion (PIE)

Consider the following naïve question:

> If $|A| = a$ and $|B| = b$, what is $|A \cup B|$?

The solution is simple: we first count how many elements are in A and then count how

many elements are in B, but note that there are some elements which might be in both sets and are thus counted twice, so we need to subtract these elements away.

Proposition 1.5.1 ▸ Cardinality of the Union of Two Finite Sets

Let A and B be two finite sets, then

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

In particular, if A and B are disjoint, then $|A \cup B| = |A| + |B|$.

Proposition 1.5.1 can be generalised for a finite sequence of finite sets.

Theorem 1.5.2 ▸ Principle of Inclusion and Exclusion

Let $E_1, E_2, \cdots, E_n$ be a sequence of finite sets. In general, we have

$$\left|\bigcup_{i=1}^{n} E_i\right| = \sum_{j=1}^{n}\left[(-1)^{j+1}\left(\sum_{1\le k_1\le k_2\le\cdots\le k_j\le n}\left|\bigcap_{r=1}^{j} E_{k_r}\right|\right)\right].$$

Proof. Define a function $f_S : S \to \{0,1\}$ by

$$f_S(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{if } x \notin S \end{cases}.$$

Let $E = \bigcup_{i=1}^{n} E_i$. Consider

$$\prod_{i=1}^{n}\left(f_E(x) - f_{E_i}(x)\right). \tag{*}$$

For any x, if $x \in E$, then $x \in E_k$ for some $k \in \{x \in \mathbb{N} : x \le n\}$, which means that $f_E(x) - f_{E_k}(x) = 0$; if $x \notin E$, then $f_E(x) = f_{E_i}(x) = 0$ for all $i \in \{x \in \mathbb{N} : x \le n\}$. In either case, (*) is identically 0.

Note that for any sets A, B, $f_{A\cap B}(x) = f_A(x)f_B(x)$. In particular, $f_{E\cap E_i}(x) = f_{E_i}(x)$ for all $i = 1, 2, \cdots, n$. Expanding (*), we have

$$f_E(x) = \left[f_E(x)\right]^n = \sum_{j=1}^{n}\left\{(-1)^{j+1}\left[\sum_{1\le k_1\le k_2\le\cdots\le k_j\le n}\left(\prod_{r=1}^{j} f_{E_{k_r}}(x)\right)\right]\right\}.$$

Summing up $f_E(x)$ over all x leads to the identity in the original statement. □

We can generalise this further. Note that every set is uniquely determined by some formula ϕ with some parameter p. This means that the statement $x \in X$ is logically equiva-

lent to "x satisfies a property $\phi(x, p)$". This leads to the general principle of inclusion and exclusion.

Theorem 1.5.3 ▸ General Principle of Inclusion and Exclusion

Let S be a finite set with $|S| = n$ and let $\{P_1, P_2, \cdots P_q\}$ be a set of properties. Suppose that $E(m)$ is the number of elements in S that satisfy exactly m of the q properties and that $\omega(P_{i_1} P_{i_2} \cdots P_{i_m})$ is the number of elements in S that satisfy exactly $P_{i_1}, P_{i_2}, \cdots, P_{i_m}$, such that

$$\omega(m) = \sum \omega(P_{i_1} P_{i_2} \cdots P_{i_m}),$$

then

$$E(m) = \sum_{k=m}^{q} \left[(-1)^{k-m} \binom{k}{m} \omega(k) \right].$$

Proof. Let $x \in S$ be an arbitrary element which satisfies exactly t properties. If $t < m$ or $q < t$, then x contributes a count of 0 to both sides. If $t = m$, then x contributes a count of 1 to $E(m)$ and to $\omega(m)$ and contributes a count of 0 to $\omega(k)$ for all $k > m$.

If $q \geq t > m$, then x contributes a count of 0 to the left-hand side. On the right-hand side, x is counted $\binom{t}{k}$ times for $k = m, m+1, \cdots, t$. So the total contribution of x to the right-hand side is

$$\begin{aligned}
\sum_{k=m}^{t} \left[(-1)^{k-m} \binom{t}{k} \binom{k}{m} \right] &= \sum_{k=m}^{t} \left[(-1)^{k-m} \binom{t}{m} \binom{t-m}{k-m} \right] \\
&= \sum_{k=0}^{t-m} \left[(-1)^{k} \binom{t}{m} \binom{t-m}{k} \right] \\
&= \binom{t}{m} \sum_{k=0}^{t-m} \left[(-1)^{k} \binom{t-m}{k} \right] \\
&= 0.
\end{aligned}$$

Therefore, every $x \in S$ contributes an equal count to both sides of the identity, so the identity holds. □

In particular, we see that

$$E(0) = \sum_{i=0}^{q} \left[(-1)^i \omega(i) \right].$$

1.6 Distribution Problems

Another problem in which we are interested is to **distribute** the objects from a collection into a finite number of sub-collections. Specifically, we consider the number of ways to

1. distribute r distinct objects into n identical collections;
2. distribute r distinct objects into n distinct collections;
3. distribute r identical objects into n identical collections;
4. distribute r identical objects into n distinct collections.

These 4 basic models of distribution problems give rise to many variations. We will only discuss the first two types of distribution problems here. The easiest case for this distribution problem is that each collection contains at most one object, i.e., each collection either contains an object, or contains no object at all.

Proposition 1.6.1 ▸ Distinct Objects into Singleton Distinct Collections

The number of ways to distribute r distinct objects into n distinct collections such that each collection contains at most 1 object is given by P_r^n.

We then consider the case where each collection can contain any number of objects.

Proposition 1.6.2 ▸ Distinct Objects into Distinct Collections

The number of ways to distribute r distinct objects into n distinct collections such that each collection can contain any number of objects is given by

$$\frac{(n+r-1)!}{(n-1)!}.$$

Proof. Left to the reader as an exercise. □

For distributing distinct objects into identical collections, the reasoning becomes a bit more complicated. We first consider a scenario where the collections are circular.

Definition 1.6.3 ▸ Stirling Numbers of the First Kind

Let $r, n \in \mathbb{N}$ such that $0 \leq n \leq r$, then the **Stirling Number of the First Kind**, $s(r, n)$, is the number of ways to arrange r **distinct** onjects around n **identical** circles such that no circle is empty.

Remark. Some obvious results:

1. $s(r,0) = 0$ if $r \geq 1$.
2. $s(r,r) = 1$ if $r \geq 0$.
3. $s(r,1) = (r-1)!$ if $r \geq 2$.
4. $s(r,r-1) = C_2^r$ if $r \geq 2$.

One thing to take note of is that there is no general algebraic formula for $s(r,n)$. Instead, all Stirling Numbers of the First Kind follow a recurrence relation in 2 parameters.

Theorem 1.6.4 ▶ A Recurrence Relation for $s(r,n)$

For $r, n \in \mathbb{N}^+$ and $r \leq n$, we have

$$s(r,n) = s(r-1,n-1) + (r-1)s(r-1,n).$$

Proof. Consider the set $\{x_1, x_2, \cdots, x_r\}$ to be the r distinct objects. We consider two cases. If x_r is distributed to a circle such that it is the only objects around that circle, it suffices to find the number of ways to arrange the rest $(r-1)$ distinct objects around the rest $(n-1)$ identical circles, which there are $s(r-1,n-1)$ ways to do.

If x_r is adjacent to some other object, we can first arrange the rest $(r-1)$ distinct objects around the n identical circles, which there are $s(r-1,n)$ ways to do. After that, we can choose any one of the $(r-1)$ spaces between two adjacent objects to slot in x_r. Therefore, there are $(r-1)s(r-1,n)$ ways to distribute the objects. By Theorem 1.1.1, the total number of distributions is

$$s(r,n) = s(r-1,n-1) + (r-1)s(r-1,n).$$

□

Next, we consider the case where the collections arrange their elements linearly.

Definition 1.6.5 ▶ Stirling Number of the Second Kind

The **Stirling Number of the Second Kind**, denoted as $S(r,n)$, is the number of ways to distribute r distinct objects into n indentical collections such that no collection is empty.

Remark. Some trivial results:

- $S(0,0) = 1$.
- $S(r,0) = S(0,n) = 0$ for all $r, n \in \mathbb{N}^+$.
- $S(r,n) > 0$ for all $r, n \in \mathbb{N}$ with $r \geq n \geq 1$.
- $S(r,n) = 0$ if $n > r \geq 1$.
- $S(r,1) = S(r,r) = 1$ for all $r \in \mathbb{N}^+$.

Based on the above trivial results, we can use the following recurrence relation to derive a general formula for $S(r,n)$ for any r and n.

Theorem 1.6.6 ▶ Recurrence Relation of $S(r,n)$

For all $r, n \in \mathbb{N}$ with $r \geq n$,

$$S(r,n) = S(r-1,n-1) + nS(r-1,n).$$

Proof. Consider the first object. If the first object is alone in a collection, it is equivalent to distributing $(r-1)$ distinct objects into $(n-1)$ identical collections such that no collection is empty, which can be done in $S(r-1,n-1)$ ways. If the first object is not alone, we first distribute the rest $(r-1)$ distinct objects into the n identical collections in $S(r-1,n)$ ways, and then choose a collection to place the first object into in n ways, which can be done in $nS(r-1,n)$ ways. Therefore,

$$S(r,n) = S(r-1,n-1) + nS(r-1,n).$$

□

The Stirling numbers of the second kind are closely related to the problem of set partition.

Definition 1.6.7 ▶ Partition

Let A be a set, an n-partition of A is a set $S \subset \mathcal{P}(A)$ such that for all $S_1, S_2 \in S$ with $S_1 \neq S_2$, we have $S_1 \cap S_2 = \varnothing$ and $\bigcup S = A$.

Note that in a set, the elements are pairwise distinct. Therefore, a partition of A into i disjoint subsets is equivalent to a distribution of $|A|$ distinct objects into i identical collections. Therefore, the number of different partitions of A is just

$$\sum_{i=1}^{|A|} S(|A|, i).$$

To find an exact formula for $S(r,n)$, we consider the following proposition:

Proposition 1.6.8 ▸ Number of Surjective Mappings

Let $F(r,n)$ where $r,n \in \mathbb{N}^+$ deonote the number of surjective mappings

$$f : [r] - \{0\} \to [n] - \{0\},$$

then

$$F(r,n) = \sum_{k=0}^{n} \left[(-1)^k \binom{n}{k} (n-k)^r \right].$$

Proof. Note that f is a mapping from a domain of r elements to a co-domain of n elements. Let M_k be the set of such mappings by which at least k elements in the co-domain do not have pre-images. Note that for each element in the domain, its image can be any of the remaining $(n-k)$ elements. Therefore,

$$|M_k| = \binom{n}{k} (n-k)^r.$$

By Theorem 1.5.3,

$$F(r,n) = \left| \left(\bigcup_{k=0}^{n} M_k \right)' \right| = \sum_{k=0}^{n} \left[(-1)^k \binom{n}{k} (n-k)^r \right].$$

□

Note that $F(r,n)$ is just the number of ways to distribute r distinct elements into n collections where the order of the collections matters, so $F(r,n) = n!S(r,n)$. This gives rise to some corollaries, for example:

$$S(r,n) = \frac{1}{n!} \sum_{k=0}^{n} (-1)^k \binom{n}{k} (n-k)^r$$

Since $S(r,n) = 0$ if $n > r \geq 1$, we have

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} (n-k)^r = 0 \quad \text{if } n > r \geq 1.$$

A common application of distribution problems with n identical objects and m identical collections is the partition of an integer.

Definition 1.6.9 ▸ Integer Partition

Let $n \in \mathbb{N}^+$. A **partition** of n is a non-increasing sequence of positive integers whose sum is n.

A useful tool to study such partitions is *Ferrers diagram*.

Definition 1.6.10 ▸ Ferrers Diagram

Let $n \in \mathbb{N}^+$ and consider a partition $P = \{n_1, n_2, \cdots, n_k\}$. The **Ferrers diagram** of the partition, denoted as $\mathcal{F}(P)$, is an array of left-justified asterisks in which the i-th row has the length of n_i.

Just like matrices, a Ferrers diagram can be transposed.

Definition 1.6.11 ▸ Conjugate Partition

Let $\mathcal{F}(P)$ be the Ferrers diagram of an integer partition P. Q is known as the **conjugate partition** of P if $\mathcal{F}(Q) = \mathcal{F}(P)^{\mathrm{T}}$.

Intuitively, conjugate partition is a **symmetric relation**, and if P and Q are conjugate partitions, the number of parts of Q is just the size of the largest part of P. This leads to the following theorem:

Proposition 1.6.12 ▸ Size of Conjugate Partitions

Let $k \leq r \in \mathbb{N}^+$. The number of partitions of r into k parts equals the number of partitions of r with the largest part having size k.

Proof. Let $\mathcal{P}$ be the family of partitions of r into k parts and $\mathcal{Q}$ be the family of partitions of r whose largest part has size k. Define a function $f : \mathcal{P} \to \mathcal{Q}$ such that $\mathcal{F}\big(f(X)\big) = \mathcal{F}(X)^{\mathrm{T}}$, i.e., $f(X)$ and X are conjugates. It is easy to see that f is well-defined and bijective, so $|\mathcal{P}| = |\mathcal{Q}|$. □

Proposition 1.6.12 leads to the following corollary:

Corollary 1.6.13 ▸ Integer Partition into Limited Parts

Let $k \leq r \in \mathbb{N}^+$. The number of partitions of r into at most k parts equals the number of partitions of r with the largest part having size at most k.

The following formula can de derived:

$$\sum_{k=1}^{m} p(n,k) = p(n+m,m).$$

2

Probability Axioms

2.1 Classical and Axiomatic Probability

Definition 2.1.1 ▶ Sample Space

Consider an experiment whose outcome is **not** predictable, then the set of all possible outcomes of the experiment is called the **sample space** of the experiment, denoted by S.

Remark. Note that $S \neq \varnothing$.

Naturally, an *event* is nothing but a collection of outcomes.

Definition 2.1.2 ▶ Events

Let S be a sample space, a set $E \subseteq S$ is known as an **event**.

Remark. S itself is known as the **sure event** and $\varnothing$ is known as the **null event**.

Note that since sample spaces and events are sets, we can apply operations onto events precisely in the same way for sets.

By convention, the intersection of two events E and F is preferably written as EF. Two events which are disjoint are called *mutually exclusive*.

Definition 2.1.3 ▶ Classical Definition of Probability

Let E be any event of an experiment and let $n(E)$ be the number of occurrences of E in the first n repetitions of the experiment, then the **probability** of E is

$$P(E) = \lim_{n\to\infty} \frac{n(E)}{n},$$

if the limit exists.

However, notice that from the above, the notion of probability may not be well-defined as $n(E)$ is not a function, which means that the limit is not defined.

To avoid this problem, we shall use an axiomatic definition instead, i.e., we define probability to be such that if it exists and is well-defined, then it satisfies a series of axioms.

Definition 2.1.4 ▶ Axiomatic Definition of Probability

Let S be a sample space and let $P(E)$ be a real number defined for every $E \subseteq S$. If

- $0 \leq P(E) \leq 1$,
- $P(S) = 1$, and
- for all mutually exclusive E and F, $P(E \cup F) = P(E) + P(F)$,

then $P(E)$ is the **probability** of E.

With induction, one can easily show that if $E_1, E_2, \cdots$ to be any sequence of events in a sample space S, then

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

We now follow up with proofs for two seemingly intuitive results.

Theorem 2.1.5 ▶ The Null Event

Consider the null event $\varnothing$, we have

$$P(\varnothing) = 0.$$

Proof. Let S be a sample space and let $E_1, E_2, \cdots$ be a countably infinite sequence of events such that $E_i = \varnothing$ for all $i \in \mathbb{N}^+$. We can write

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

Note that the countable union of empty sets is empty, so the above is equivalent to

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

This means that $P(\varnothing)$ equals the sum of a countably infinite sequence of itself, so

$$P(\varnothing) = 0.$$

□

Theorem 2.1.6 ▸ Monotonity of Probability

Let E and F be events such that $E \subseteq F$, then

$$P(F) \geq P(E).$$

Proof. Note that E and $F - E$ are mutually exclusive, so

$$P(F) = P(E \cup (F - E)) = P(E) + P(F - E).$$

Note that $P(F - E) \geq 0$, so $P(E) + P(F - E) \geq P(E)$, which means

$$P(F) \geq P(E).$$

□

It is easy to compute the probability of a countable union of mutually exclusive events. However, it may get tricky when an event is the union of events which are not mutually exclusive. Intuitively, we can sum up the probabilities of all individual events and subtract the portions which are double-counted. This approach is rigorously summarised as follows:

Theorem 2.1.7 ▸ Inclusion-Exclusion Principle

Let S be a sample space and let $E_1, E_2, \cdots, E_n$ be a sequence of events. In general, we have

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{j=1}^{n}\left[(-1)^{j+1}\left(\sum_{k_1 \leq k_2 \leq \cdots \leq k_j} P\left(\bigcap_{h=1}^{j} E_{k_h}\right)\right)\right].$$

Proof. Define a function $f_S : S \to \{0, 1\}$ by

$$f_S(x) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{if } x \notin S \end{cases}.$$

Let $E = \bigcup_{i=1}^{n} E_i$. Consider the function $g : S \to \{0, 1\}$ given by

$$g(x) = \prod_{i=1}^{n} \left(f_E(x) - f_{E_i}(x)\right).$$

For any $x \in S$, if $x \in E$, then $x \in E_k$ for some $k \in \{x \in \mathbb{N} : x \leq n\}$, which means that $f_E(x) - f_{E_k}(x) = 0$; if $x \notin E$, then $f_E(x) = f_{E_i}(x) = 0$ for all $i \in \{x \in \mathbb{N} : x \leq n\}$. In either case, $g(x) = 0$. □

One can check that the above principle gives rise to the following inequality:

Theorem 2.1.8 ▶ Boole's Inequality

Let $E_1, E_2, \cdots, E_n, \cdots$ be a countable sequence of events, then

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

In particular, equality is achieved if and only if the E_i's are mutually exclusive.

Proof. Left as an exercise to the reader. □

Lastly, combining everything we have learnt so far, we can derive the following formula for finite sample spaces, which complies with our intuition:

Theorem 2.1.9 ▶ Probability in a Finite Sample Space

Let S be a sample space which is finite and let $E \subseteq S$ be an event, then

$$P(E) = \frac{|E|}{|S|}.$$

Proof. Left as an exercise to the reader. □

2.2 Measure-Theoretic Probability

The above definition for probability, though intuitive, is not very solid when it comes to an infinite sample space and continuous cases. In this section, we would establish the theories of probability using a more modern and rigorous structure.

In a naïve attempt to devise a probability model, if the sample space S is countable, then it suffices to define a *probability mass function* $P: S \to \mathbb{R}$ such that $\sum_{\omega \in S} P(\omega) = 1$. Naturally, the probability for an event $E \subseteq S$ is defined as $P(E) = \sum_{\omega \in E} P(\omega)$. This summation is compatible with the infinite case because if we have countably many pairwise disjoint events $E_1, E_2, \cdots$, we can compute

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) = \lim_{n \to \infty} \sum_{i=1}^{n} E_i,$$

which is clearly convergent by monotone-convergent theorem.

However, when S is uncountable, this construction leads to weird behaviours. For example, suppose S is the sample space for the experiment of tossing a fair coin for uncountably many

times. It is clear that for any $\omega \in S$, we have

$$P(\omega) = \lim_{n\to\infty} \frac{1}{2^n} = 0,$$

but at the same time we must have

$$1 = P(S) = \sum_{\omega\in S} P(\omega) = 0,$$

which is ridiculous. Therefore, we need to find a better way to construct the probability model. Notice that here the incompatibility arises because we build our model by considering the probabilities of individual outcomes. Our next attempt try to bypass this issue by considering the probabilities of events only.

Since the set of all events in a sample space S is simply $\mathcal{P}(S)$, let us instead consider a more generalisable algebraic structure for this collection of subsets.

Definition 2.2.1 ▸ Set Algebra

Let X be a set. A **set algebra** over X is a family $\mathcal{F} \subseteq \mathcal{P}(X)$ such that

- $X \backslash F \in \mathcal{F}$ for all $F \in \mathcal{F}$ (closed under complementation);
- $X \in \mathcal{F}$;
- $X_1 \cup X_2 \in \mathcal{F}$ for any $X_1, X_2 \in \mathcal{F}$ (closed under binary union).

There are several immediate implications from the above definition.

First, by closure under complementation, we know that an algebra over any set X must contain the empty set.

Second, by De Morgan's Law, one can easily check that if the first 2 axioms hold, the closure under binary union is equivalent to

- $X_1 \cap X_2 \in \mathcal{F}$ for any $X_1, X_2 \in \mathcal{F}$;
- $\bigcup_{i=1}^{n} X_i \in \mathcal{F}$ for any $X_1, X_2, \cdots, X_n \in \mathcal{F}$ for all $n \in \mathbb{N}$;
- $\bigcap_{i=1}^{n} X_i \in \mathcal{F}$ for any $X_1, X_2, \cdots, X_n \in \mathcal{F}$ for all $n \in \mathbb{N}$.

$(X, \mathcal{F})$ is known as a *field of sets*, where the elements of X are called *points* and those of $\mathcal{F}$, *complexes* or *admissible sets* of X. In probability theory, what we are interested in is a special type of set algebras known as σ-*algebras*.

Definition 2.2.2 ▸ σ-Algebra

A **σ-Algebra** over a set A is a non-empty set algebra over A that is closed under countable union.

Of course, by the same argument as above, we known that any σ-algebra is closed under countable intersection as well. Roughly speaking, we could now define the probability of an event $E \subseteq S$ as the ratio of the size of E to that of S. The remaining question now is: how do we define the size of a set (and in particular, an infinite set) properly?

Definition 2.2.3 ▸ Measure

Let X be a set and Σ be a σ-algebra over X. A **measure** over Σ is a function

$$\mu : \Sigma \to \mathbb{R} \cup \{-\infty, +\infty\}$$

such that

- $\mu(E) \geq 0$ for all $E \in \Sigma$ (non-negativity);
- $\mu(\varnothing) = 0$;
- $\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu(E_i)$ for any countable collection of pairwise disjoint elements of Σ (countable additivity or σ-additivity).

The triple (X, Σ, μ) is known as a **measure space** and the pair (X, Σ), a **measurable space**.

One thing to note here is that if at least one $E \in \Sigma$ has a finite measure, then $\mu(\varnothing) = 0$ is automatically guaranteed for obvious reasons.

Definition 2.2.4 ▸ Probability Measure

Let $\mathcal{F}$ be a σ-algebra over a sample space S. A **probability measure** over S is a measure $P : \mathcal{F} \to [0, 1]$ such that $P(S) = 1$.

Obviously, the above definition immediately guarantees that

1. $P(A^c) = 1 - P(A)$;
2. $P(A) \leq P(B)$ if $P(A) \subseteq P(A)$;
3. $P(A \cup B) \leq P(A) + P(B)$.

The third result follows from a direct application of the principle of inclusion and exclusion. By induction, one can easily check that

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

for any finitely many events. The following proposition extends this result to countable collections of events:

Proposition 2.2.5 ▸ Union Bound of Countable Collections of Events

Let $(S, \mathcal{F}, P)$ be a probability space and $E_1, E_2, \cdots, E_n, \cdots \in \mathcal{F}$ is any countable sequence of events, then

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

Proof. Define $F_1 := E_1$ and $F_k := E_k \setminus \bigcup_{i=1}^{k-1} E_i$ for $k \geq 2$. Clearly, the F_i's are pairwise disjoint. By Definition 2.2.2, the F_i's are elements of $\mathcal{F}$. Note that $P(F_i) \leq \mathbb{E}_{\mathbb{i}}$ for all $i \in \mathbb{N}^+$, so

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} E_i\right) &= P\left(\bigcup_{i=1}^{\infty} F_i\right) \\ &= \sum_{i=1}^{\infty} P(F_i) \\ &\leq \sum_{i=1}^{\infty} P(E_i). \end{aligned}$$

□

Intuitively, the equality is attained if and only if the events are pairwise disjoint.

2.3 Conditional Probability

In a sample space S, we may wish to find the probability of two events E and F both occurring, $P(EF)$. However, if we already know that event F **has occurred**, then necessarily, the sample space we consider would no longer be S. Essentially, this condition of F having occurred has restricted our sample space to F. Thus, we give the following definition:

Definition 2.3.1 ▸ Conditional Probability

Let S be a sample space and $E, F \subseteq S$ be two events. If $P(F) \geq 0$, then the **conditional probability** is the probability that E occurs given that F has occurred, denoted by

$$P(E|F) = \frac{P(EF)}{P(F)}.$$

In particular, if $E \subseteq F$, we have $P(E|F) = \frac{P(E)}{P(F)}$.

Remark. Note that $P(E|F) = P(EF|F)$.

It is easy to see that $P(EF) = P(E|F)P(F)$, i.e., the probability of E and F both occurring is the product of the probability of F occurring and the probability of E occurring given the occurrence of F. This complies with our intuition in a sense that if we wish both E and F to happen, we just need F to happen first and E to happen given the occurrence of F. We can generalise this for a countable number of events:

Proposition 2.3.2 ▸ Multiplication Rule

Let S be a sample space and let $E_i \subseteq S$ for $i = 1, 2, \cdots, n$ be n events, where $n \geq 2$. Suppose that $P\left(\bigcap_{i=1}^{n-1} E_i\right) > 0$, then

$$P\left(\bigcap_{i=1}^{n} E_i\right) = P(E_1) \prod_{i=2}^{n} P\left(E_i \middle| \bigcap_{j=1}^{i-1} E_j\right).$$

Proof. The case where $n = 2$ is immediate from Definition 2.3.1. Suppose that there is some $k \in \mathbb{N}$ and $k \geq 2$ such that

$$P\left(\bigcap_{i=1}^{k} E_i\right) = P(E_1) \prod_{i=2}^{k} P\left(E_i \middle| \bigcap_{j=1}^{i-1} E_j\right),$$

then we consider

$$\begin{aligned} P\left(E_{k+1} \middle| \bigcap_{i=1}^{k} E_i\right) &= \frac{P\left(\bigcap_{i=1}^{k+1} E_i\right)}{P\left(\bigcap_{i=1}^{k} E_i\right)} \\ &= \frac{P\left(\bigcap_{i=1}^{k+1} E_i\right)}{P(E_1) \prod_{i=2}^{k} P\left(E_i \middle| \bigcap_{j=1}^{i-1} E_j\right)}. \end{aligned}$$

Therefore,

$$\begin{aligned} P\left(\bigcap_{i=1}^{k+1} E_i\right) &= \left[P(E_1) \prod_{i=2}^{k} P\left(E_i \middle| \bigcap_{j=1}^{i-1} E_j\right)\right] P\left(E_{k+1} \middle| \bigcap_{i=1}^{k} E_i\right) \\ &= P(E_1) \prod_{i=2}^{k+1} P\left(E_i \middle| \bigcap_{j=1}^{i-1} E_j\right) \end{aligned}$$

□

Consider a sample space S and two events $E, F \subseteq S$. Suppose that E occurs, then either F has occurred or F has never occurred (i.e. F^c occurred). Therefore, it is easy to see that

$$P(E) = P(EF) + P(EF^c) = P(E|F) + P(E|F^c).$$

We can extend the above argument for more than two events. Suppose that $F_1, F_2, \cdots, F_n$ are n mutually exclusive events such that $\bigcup_{i=1}^{n} F_i = S$, then obviously $\{F_1, F_2, \cdots, F_n\}$ is a *partition* of S.

Consider any event E and let $e \in E$. Clearly, e must be in one and only one of $F_1, F_2, \cdots, F_n$. It then follows that $\{E \cap F_1, E \cap F_2, \cdots, E \cap F_n\}$ is a partition of E. This can be formalised into the following result:

Theorem 2.3.3 ▸ Law of Total Probabilities

Let $F_1, F_2, \cdots, F_n$ be a partition of a sample space S with probability measure P. For any event $A \subseteq S$,

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

For a countably infinite number of mutually exclusive events $F_1, F_2, \cdots$ such that $\bigcup_{i=1}^{\infty} F_i = S$, we arrive at the following formula:

$$P(E) = \sum_{i=1}^{\infty} P(E|F_i)P(F_i).$$

This leads to the *Bayes's Formula*:

Theorem 2.3.4 ▸ Bayes's Formula

Let $F_1, F_2, \cdots$ be a countably infinite sequence of events from a sample space S such that $\bigcup_{i=1}^{\infty} F_i = S$. For any event $E \subseteq S$, we have

$$P(F_j|E) = \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^{\infty} P(E|F_i)P(F_i)}.$$

Proof. Left as an exercise to the reader. □

Note that in general, for two events E and F, $P(E|F) \neq P(E)$, i.e., the occurrence of F may affect the occurrence of E. However, in some cases, we notice that the occurrence of E is *independent* of F, and so we introduce the following definition:

Definition 2.3.5 ▸ Independent Events

Let S be a sample space with probability measure P. Two events $E, F \subseteq S$ are **independent** if $P(E \mid F) = P(E)$, or equivalently, $P(E \cap F) = P(E) P(F)$. A collection of

events $E_1, E_2, \cdots, E_n$ are said to be **jointly independent** if for any $I \subseteq \{1, 2, \cdots, n\}$,

$$P\left(\bigcap_{i\in I} E_i\right) = \prod_{i\in I} P(E_i),$$

or equivalently, $P(E_1 \mid E_2, \cdots, E_n) = P(E_1)$.

Remark. The following results are immediate:

1. If $P(E) = 0$ or $P(F) = 0$, then E and F are independent.
2. If $P(E) > 0$ (respectively, $P(F) > 0$), then E and F are independent if and only if $P(F|E) = P(F)$ (respectively, $P(E|F) = P(E)$).

Intuitively, given independent events E and F, we may believe that if the occurrence of E does not affect the occurrence of F, then naturally the occurrence of E should also not affect the "not-occurring" of F, i.e., the following is true:

Proposition 2.3.6 ▸ Characterisation of Independent Events

E and F are independent events if and only if E and F^c are independent events.

Proof. Notice that $EF \cup EF^c = E(F \cup F^c) = E$, so

$$P(E) = P(EF) + P(EF^c) = P(E)P(F) + P(EF^c).$$

Therefore,

$$P(EF^c) = P(E) - P(E)P(F) = P(E)(1 - P(F)) = P(E)P(F^c),$$

and so E and F^c are independent. The other direction is trivial because $F = (F^c)^c$ □

Note that E and F are independent if and only if E, F, E^c, F^c are pairwise independent. Moreover, for any jointly independent collection of events $E_1, E_2, \cdots, E_n$ and any disjoint index sets $I, J \subseteq \{1, 2, \cdots, n\}$,

$$P\left(\bigcap_{i\in I} E_i \cap \bigcap_{j\in J} E_j^c\right) = \left(\prod_{i\in I} P(E_i)\right)\left(\prod_{j\in J} P\left(E_j^c\right)\right).$$

Remark. Joint independence is a strictly stronger result than pairwise independence, i.e., there exists pairwise independent events E_1, E_2, E_3 such that

$$P(E_1 \cap E_2 \cap E_3) \neq P(E_1)\,P(E_2)\,P(E_3).$$

3

Random Variables

3.1 Random Variables

In many contexts, we might wish to generalise a formula to compute the probability of the occurrence of a certain event. However, in cases where the events are abstract or unquantifiable (e.g. the event "tomorrow is rainy"), it becomes hard to formulate a well-defined mapping from a sample space to $[0,1]$. Thus, to model all events easily using functions and mappings, we introduce the notion of *random variables*.

A random variable can be viewed as a function that maps the outcomes in a sample space to some measurable co-domain. We first introduce a few preliminary definitions.

Definition 3.1.1 ▸ Probability Space

A **probability space** is a tuple $(S, \mathcal{F}, P)$ where S is a sample space, $\mathcal{F}$ is a σ-algebra on S and P is a probability measure on S.

It can be troublesome to consider different sample spaces for different experiments. Therefore, we define the *abstract probability space* as $(\Omega, \mathcal{F}, P)$ with a uniform random variable Z such that for every outcome measured by a random variable X in a sample space S, there exists a function $f : \Omega \to S$ such that $X = f(Z)$. For convenience, we often choose $\Omega = [0,1]$ and P to be the uniform measure on $[0,1]$.

One important property of a probability space is **countable additivity**.

Proposition 3.1.2 ▸ Countable Additivity in Probability Spaces

Let (Ω, P) be a probability space. If $\{A_i\}_{i\in\mathbb{N}^+}$ is a family of subsets of Ω such that $A_i \subseteq A_{i+1}$ for all $i \in \mathbb{N}^+$, then

$$P\left(\bigcup_{i\in\mathbb{N}^+} A_i\right) = \lim_{n\to\infty} P(A_n).$$

If $\{B_i\}_{i\in\mathbb{N}^+}$ is a family of subsets of Ω such that $B_{i+1} \subseteq B_i$ for all $i \in \mathbb{N}^+$, then

$$P\left(\bigcap_{i\in\mathbb{N}^+} B_i\right) = \lim_{n\to\infty} P(B_n).$$

Proof. Define $A_0 := \varnothing$, then

$$\{A_{i+1} \setminus A_i : i \in \mathbb{N}\}$$

is a countable collection of pairwise disjoint subsets of Ω. Therefore,

$$\begin{aligned}
P\left(\bigcup_{i\in\mathbb{N}^+} A_i\right) &= P\left(\bigcup_{i\in\mathbb{N}} (A_{i+1} \setminus A_i)\right) \\
&= \sum_{i=0}^{\infty} P(A_{i+1} \setminus A_i) \\
&= \lim_{n\to\infty} \sum_{i=0}^{n} P(A_{i+1} \setminus A_i) \\
&= \lim_{n\to\infty} P\left(\bigcup_{i=0}^{n} (A_{i+1} \setminus A_i)\right) \\
&= \lim_{n\to\infty} P(A_n).
\end{aligned}$$

Define $C_i := \Omega \setminus B_i$. Since $B_{i+1} \subseteq B_i$ for all $i \in \mathbb{N}^+$, we have $C_i \subseteq C_{i+1}$ for all $i \in \mathbb{N}^+$. Therefore,

$$\begin{aligned}
P\left(\bigcup_{i\in\mathbb{N}^+} B_i\right) &= P\left(\bigcup_{i\in\mathbb{N}^+} (\Omega \setminus C_i)\right) \\
&= \sum_{i=1}^{\infty} P(\Omega \setminus C_i) \\
&= 1 - \sum_{i=1}^{\infty} P(C_i) \\
&= 1 - P\left(\bigcup_{i\in\mathbb{N}^+} C_i\right) \\
&= 1 - \lim_{n\to\infty} P(C_n) \\
&= 1 - \lim_{n\to\infty} P(\Omega \setminus B_n) \\
&= 1 - \lim_{n\to\infty} (1 - P(B_n)) \\
&= \lim_{n\to\infty} P(B_n).
\end{aligned}$$

□

It is important that the co-domain of a random variable is measurable. For this purpose, we construct some structure to generalise open intervals in $\mathbb{R}$.

Definition 3.1.3 ▶ Borel Algebra

Let X be a topological space. A **Borel set** on X is a set which can be formed via countable union, countable intersection and relative complementation of open sets in X. The smallest σ-algebra over X containing all Borel sets on X is known as the **Borel algebra** over X.

Note that the Borel set over $\mathbb{R}$ is just the family of all open intervals. Clearly, the Borel algebra over X contains all open sets in X according to the above axioms from Definition 2.2.2. This helps us define the following:

Definition 3.1.4 ▶ Random Variable

Let $(\Omega, \mathcal{F}, P)$ be the abstract probability space and $(\mathcal{X}, \mathcal{B})$ be a measurable space where $\mathcal{B}$ is the Borel algebra over $\mathcal{X}$. A **random variable** is a function $X : \Omega \to \mathcal{X}$ such that

$$\{\omega \in S : X(\omega) \in B\} \in \mathcal{F}$$

for all $B \in \mathcal{B}$. The probability measure P_X on $\mathcal{X}$ induced by P with

$$P_X(A) \coloneqq P(X \in A) \coloneqq P\big(\{\omega \in \Omega : X(\omega) \in A\}\big)$$

is known as the **distribution** of X.

Remark. Rigorously, such a random variable X is a *measurable function* or *measurable mapping* from $(\Omega, \mathcal{F})$ to $(\mathcal{X}, \mathcal{B})$.

We shall verify that P_X as defined above is indeed a probability measure. Notice that we have $P_X(A) \in [0,1]$ for all $A \subseteq \mathcal{X}$ and that $P_X(\mathcal{X}) = 1$ and $P_X(\varnothing) = 0$. Let $\{A_i\}_{i\in\mathbb{N}}$ be a family of pairwise disjoint events. Consider

$$\begin{aligned} X^{-1}\left(\bigcup_{i\in\mathbb{N}} A_i\right) &= \left\{\omega \in \Omega : X(\omega) \in \bigcup_{i\in\mathbb{N}} A_i\right\} \\ &= \bigcup_{i\in\mathbb{N}} \{\omega \in \Omega : X(\omega) \in A_i\} \\ &= \bigcup_{i\in\mathbb{N}} X^{-1}(A_i). \end{aligned}$$

Therefore,

$$\begin{aligned} P_X\left(\bigcup_{i\in\mathbb{N}} A_i\right) &= P\left(\bigcup_{i\in\mathbb{N}} X^{-1}(A_i)\right) \\ &= \sum_{i\in\mathbb{N}} P\left(X^{-1}(A_i)\right) \\ &= \sum_{i\in\mathbb{N}} P_X(A). \end{aligned}$$

Alternatively, let Ω be a sample space, the random variable

$$X:\ \Omega \to \mathbb{R}$$

is a real-valued function such that for any event $E \subseteq \Omega$,

$$P(E) = P(X[E]) = P(X \in X[E]),$$

where

$$X[E] \coloneqq \{X(\omega):\ \omega \in \Omega\}$$

is the image of the event E under X.

A random variable describes the random outcome of an "experiment" or "phenomenon". When we perform a series of experiments (for example, model the weather for n consecutive days), it is reasonable to use one random variable to capture the outcome for each experiment. In this way, we obtain a collection of random variables denoted as

$$X_i^n \coloneqq (X_1, X_2, \cdots, X_n).$$

Clearly, if X is a real-valued random variable, we have $\{\omega \in S:\ X(\omega) > x\} \in \mathcal{F}$. Moreover, we claim that

$$\{\omega \in S:\ X(\omega) < x\} = \bigcup_{y<x} \{\omega \in S:\ X(\omega) \leq y\}.$$

The proof is quite straightforward and is left to the reader as an exercise. By Definition 2.2.2, this means that

$$\{\omega \in S:\ X(\omega) < x\} \cup \{\omega \in S:\ X(\omega) > x\} \in \mathcal{F}.$$

Therefore, $\{\omega \in S:\ X(\omega) = x\} \in \mathcal{F}$. This argument justifies the probabilities $P(X < x)$ and $P(X = x)$.

When a collection of many random variables is concerned, we may consider their *joint distribution*.

Definition 3.1.5 ▶ Joint Distribution

Let $(\Omega, \mathcal{F}, P)$ be the abstract probability space and $X_i : \Omega \to \mathcal{X}_i$ be random variables for $i = 1, 2, \cdots, n$. The **joint distribution** of $X := (X_1, X_2, \cdots, X_n)$ is the probability measure P_X with domain $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$ such that

$$\begin{aligned} P_X(A_1 \times A_2 \times \cdots \times A_n) &:= P\big(\{\omega \in \Omega : X_1(\omega) \in A_1, X_2(\omega) \in A_2, \cdots, X_n(\omega) \in A_n\}\big) \\ &= P(X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n). \end{aligned}$$

Now we can define independence between random variables in a manner similar to Definition 2.3.5.

In this course, we focus on real-valued random variables, which can be fully determined by their *distribution functions*.

Definition 3.1.6 ▶ Distribution Function

Let X be a real-valued random variable over the abstract probability space, the **distribution function** of X is a function $F_X : \Omega \to [0, 1]$ such that

$$F_X(x) = P(X \leq x).$$

Note that for all $a < b \in \mathbb{R}$,

$$P\big(X \in (a, b]\big) = F_X(b) - F_X(a).$$

It can be shown from here that F_X fully determines the distribution of X (which is non-trivial). By using Proposition 3.1.2, we can prove the following result with some analysis tools:

Proposition 3.1.7 ▶ Properties of Distribution Functions

If F_X is a distribution function of a real-valued random variable X, then

1. *F_X is non-decreasing;*
2. *$\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$;*
3. *for all $x \in \mathbb{R}$, $F_X(x) = \lim_{y \to x^+} F_X(y)$ and $F_X(x^-) := \lim_{y \to x^-} F_X(y)$ exists. In particular, $P(X = x) = F_X(x) - F_X(x^-)$.*

Remark. Conversely, every function $F : \mathbb{R} \to [0, 1]$ satisfying the above properties induces a probability measure P on $\mathbb{R}$ with $P\big((-\infty, x]\big) = F_X(x)$.

In computer programs, a random number is often generated via the uniform random vari-

able Z over $[0,1]$. Z can be used to generate a random variable X associated to any given distribution function F.

Theorem 3.1.8 ▸ Distribution Simulation

Let $F : \mathbb{R} \to [0,1]$ be any distribution function. Define $F' : [0,1] \to \mathbb{R}$ by

$$F'(y) = \inf\{x \in \mathbb{R} : F(x) \geq y\}.$$

Let Z be the uniform random variable on $[0,1]$, then $X := F'(Z)$ is a random variable with distribution function F.

Proof. Notice that for all $x \in \mathbb{R}$, we have $P(X \leq x) = P\left(F^{-1}(Z) \leq x\right)$. One may check that $F^{-1}(z) \leq x$ if and only if $z \leq F(x)$, so

$$P(X \leq x) = P\big(Z \leq F(x)\big) = F(x).$$

□

3.2 Discrete Random Variables

Intuitively, there are certain events whose outcomes are finite or can be enumerated. In such cases, we may associate these events with a *discrete random variable*.

Definition 3.2.1 ▸ Discrete Random Variable (DRV)

Let X be a random variable over a sample space S, if $\text{ran}(X)$ is countable, then X is called a **discrete random variable**.

Among finitely many outcomes, the probability for each outcome is well-defined.

Definition 3.2.2 ▸ Probability Mass Function (PMF)

Let X be a discrete random variable over a sample space S, the function

$$p_X : X[S] \to [0,1]$$

where $p_X(a) = P(X = a)$ is known as the **probability mass function** of X.

Remark. $\sum_{a \in X[S]} p_X(a) = 1$.

For any discrete random variable X, $p_X(a) = 0$ if and only if $\{s \in S : X(s) = a\} = \varnothing$. This essentially means that if p_X evaluates to 0, then the corresponding event is an impossible

event.

Note that p_X essentially gives the probability of **singleton** events in a sample space. Naturally, we can represent the probability of a union of singleton events as a linear combination of the values of p_X.

Definition 3.2.3 ▸ Cumulative Distribution Function (CDF)

Let X be a discrete random variable over a sample space S with PMF p_X, the function

$$F_X : X[S] \to [0,1]$$

where $F_X(a) = \sum_{x \le a} p_X(x)$ is known as the **cumulative distribution function** of X.

Remark. Suppose that $X(s_i) = a_i$ for all $s_i \in S$ such that $a_i < a_j$ whenever $i < j$, then F_X is a non-decreasing *step function*, i.e., for all a such that $a_i \le a < a_{i+1}$,

$$F_X(a) = \sum_{x \le a} p_X(x) = \sum_{k=1}^{i} p_X(a_k).$$

Suppose X is a discrete random variable with range $\{x_1, x_2, \cdots, x_m\}$ and PMF p_X. By repeating an experiment of X for n times, we can approximate the total number of occurrences of $X = x_i$ by $np_X(x_i)$. Therefore, the average value of X can be approximated by

$$\frac{\sum_{i=1}^{m} nx_i p_X(x_i)}{n}.$$

For large n, we have

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{m} nx_i p_X(x_i) x_i}{n} = \lim_{n\to\infty} \sum_{i=1}^{m} x_i p_X(x_i) = \sum_{i=1}^{m} x_i p_X(x_i).$$

Similarly, if the range of X is countably infinite, replacing from the above $\sum_{i=1}^{m} nx_i p_X(x_i) x_i$ with $\sum_{i=1}^{\infty} nx_i p_X(x_i) x_i$ will yield the same limit.

Intuitively, the above limit represents the *expected value* of X when a large number of experiments are conducted, which leads to the following definition:

Definition 3.2.4 ▸ Expectation of Discrete Random Variables

Let X be a discrete random variable. The **expectation** (or **mean**, **expected value**)

of X is defined to be

$$\mathbb{E}[X] = \sum_{i=1}^{m} [x_i p_X(x_i)] = \sum_{i=1}^{m} [x_i P(X = x_i)]$$

if $|\text{ran}(X)| = m$, and

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} [x_i p_X(x_i)] = \sum_{i=1}^{\infty} [x_i P(X = x_i)]$$

if $\text{ran}(X)$ is countably infinite.

By convention, we use μ to denote expectation, so $X[E]$ can be written as μ_X.

For a discrete random variable X, we can define a function $g : \text{ran}(X) \to \mathbb{R}$. It is easy to see that $g(X)$ is also a discrete random variable. Therefore, we may have the following result:

Proposition 3.2.5 ▸ Expectation of Functions

Let X be a discrete random variable and define $Y = g(X)$, then

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_{x} [g(x)P(X = x)] .$$

Proof. Note that for each $y \in \text{ran}(Y)$, $g(x) = y$ for some $x \in \text{ran}(X)$, and so

$$P(Y = y) = \sum_{g(x)=y} P(X = x).$$

Therefore,

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{y} [yP(Y = y)] \\
&= \sum_{y} \left[y \sum_{g(x)=y} P(X = x) \right] \\
&= \sum_{y} \left[\sum_{g(x)=y} g(x)P(X = x) \right] \\
&= \sum_{x} [g(x)P(X = x)] .
\end{aligned}$$

□

Two simple corollaries to the above theorem are:

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$
$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

In later sections, we will prove that the same rule applies to continuous random variables as well. The above theorem gives rise to the following notion of *moments*:

Definition 3.2.6 ▸ Moment

Let X be a random variable. $\mathbb{E}[X^n]$ is called the n-th **moment** of X.

Following Proposition 3.2.5, it is easy to see that if X is discrete, then

$$\mathbb{E}[X^n] = \sum_{i=1}^{\infty} x_i^n p_X(x_i).$$

Note that given two different discrete random variables X and Y, their probability mass functions can be different but they can still have identical expectations. For example, consider p_X to be identically 0 and p_Y to be such that $p_Y(0) = 1$ and $p_Y(y) = 0$ for all $y \neq 0$.

This motivates us to find other properties to classify and characterise random variables. One of these properties is the **spread** of the possible values taken by a random variable with respect to its mean, i.e., consider the random variable X with $\mathbb{E}[X] = \mu$, we wish to determine $\mathbb{E}[|X - \mu|]$ or equivalently $\mathbb{E}[(X - \mu)^2]$. This spread is known as the *variance* of a random variable.

Definition 3.2.7 ▸ Variance

Let X be a random variable with $\mathbb{E}[X] = \mu$, the **variance** of X is defined to be

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

By convention, we use σ^2 to denote variance, so $\mathrm{Var}(X)$ can be written as σ_X^2.

The formula for $\mathrm{Var}(X)$ can be derived via Proposition 3.2.5:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[(X - \mu)^2] \\
&= \mathbb{E}[X^2 - 2\mu X + \mu^2] \\
&= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 \\
&= \mathbb{E}[X^2] - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2 \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2.
\end{aligned}$$

Another term we hear often is *standard deviation*, which is defined as follows:

Definition 3.2.8 ▶ Standard Deviation

Let X be a random variable, the **standard deviation** of X is defined to be

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)} = \sigma_X.$$

Note that we have computed the general formula for any linear combination of discrete random variables. We shall do the same for variance.

Proposition 3.2.9 ▶ Variance of Linear Combinations of Random Variables

Let X be a random variable, then

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$
$$\mathrm{SD}(\mathrm{aX} + \mathrm{b}) = |a|\mathrm{SD}(X)$$

for all $a, b \in \mathbb{R}$.

Proof. By using Proposition 3.2.5, we have

$$\begin{aligned}\mathrm{Var}(aX + b) &= \mathbb{E}[(aX + b)^2] - (\mathbb{E}[aX + b])^2 \\ &= a^2\mathbb{E}[X^2] + 2ab\mathbb{E}[X] + b^2 - \left[a(\mathbb{E}[X])^2 + 2ab\mathbb{E}[X] + b^2\right] \\ &= a^2\left[\mathbb{E}[X^2] - (\mathbb{E}[X])^2\right] \\ &= a^2\mathrm{Var}(X).\end{aligned}$$

Therefore,

$$\mathrm{SD}(aX + b) = \sqrt{\mathrm{Var}(aX + b)} = |a|\mathrm{SD}(X).$$

□

3.2.1 Bernoulli and Binomial Random Variables

Suppose we conduct an experiment. In the most simplistic view, only two outcomes are considered, namely **success** and **failure**. We can model such experiments using a discrete random variable whose range has a cardinality of 2.

Definition 3.2.10 ▸ Bernoulli Random Variable

A random variable X is a **Bernoulli random variable** if

$$p_X(x) = \begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases}$$

for some $p \in [0,1]$.

Now, consider n **independent** trials of an experiment with a probability for success of p. Let X be the number of successes among these n trials, then clearly,

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

Definition 3.2.11 ▸ Binomial Random Variable

A random variable X is a **binomial random variable** if

$$p_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

for some $p \in [0,1]$. X is said to have a **binomial distribution** with parameters (n, p), denoted by $X \sim \mathrm{B}(n, p)$.

Remark. In particular, if X is a Bernoulli random variable, then $X \sim \mathrm{B}(1, p)$.

Suppose $X \sim \mathrm{B}(n, p)$. Let N be the average number of successes in the n trials, it is expected that $p \approx \frac{N}{n}$. Therefore, we may conjure that $\mathbb{E}[X] = np$.

Proposition 3.2.12 ▸ Expectation and Variance of Binomial Distribution

Let $X \sim \mathrm{B}(n, p)$, then $\mathbb{E}[X] = np$ and $\mathrm{Var}(X) = np(1-p)$.

Proof. Note that $iC_i^n = nC_{i-1}^{n-1}$, so

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=0}^{n}\left[i\binom{n}{i} p^i(1-p)^{n-i}\right] \\
&= \sum_{i=1}^{n}\left[n\binom{n-1}{i-1} p^i(1-p)^{n-i}\right] \\
&= n\sum_{j=0}^{n-1}\left[\binom{n-1}{j} p^{j+1}(1-p)^{n-1-j}\right] \\
&= np\sum_{j=0}^{n-1}\left[\binom{n-1}{j} p^{j}(1-p)^{n-1-j}\right] \\
&= np, \\
\mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
&= \sum_{i=0}^{n}\left[i^2\binom{n}{i} p^i(1-p)^{n-i}\right] - n^2p^2 \\
&= n\sum_{i=1}^{n}\left[i\binom{n-1}{i-1} p^i(1-p)^{n-i}\right] - n^2p^2 \\
&= n\sum_{j=0}^{n-1}\left[(j+1)\binom{n-1}{j} p^{j+1}(1-p)^{n-1-j}\right] - n^2p^2 \\
&= np\left\{\sum_{j=0}^{n-1}\left[j\binom{n-1}{j} p^{j}(1-p)^{n-1-j}\right] + 1\right\} - n^2p^2 \\
&= np\left[(n-1)p+1\right] - n^2p^2 \\
&= np - np^2 \\
&= np(1-p).
\end{aligned}$$

□

Let $X \sim \mathrm{B}(n,p)$, consider

$$\begin{aligned}\frac{p_X(i+1)}{p_X(i)} &= \frac{\frac{n!}{(i+1)!(n-i-1)!}p^{i+1}(1-p)^{n-i-1}}{\frac{n!}{i!(n-i)!}p^i(1-p)^{n-i}} \\ &= \frac{\frac{1}{i+1}p}{\frac{1}{n-i}(1-p)} \\ &= \frac{(n-i)p}{(i+1)(1-p)}.\end{aligned}$$

Suppose $p_X(i+1) < p_X(i)$, then $(n-i)p < (i+1)(1-p)$. This implies that $i > (n+1)p-1$, which means that

- $p_X(i)$ is monotonically increasing on $[0,(n+1)p-1]$.
- $p_X(i)$ maximises when $i = \lceil (n+1)p-1 \rceil = \lfloor (n+1)p \rfloor$.
- $p_X(i)$ is monotincally decreasing on $((n+1)p-1, n]$.

3.2.2 Poisson Random Variable

Suppose that we want to count the number of events in unit intervals such that $X(t)$ is the number of events up to time t. Now, each unit interval can be divided infinitesimally such that we can approximate the distribution using binomial random variables.

Theorem 3.2.13 ▸ Law of Rare Events

Let $X_n \sim \mathrm{B}(n,p_n)$ be a binomial random variable where $\lim_{n\to\infty} np_n = \lambda$, then we have $\lim_{n\to\infty} P(X_n = x) \approx \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}$

Proof. Consider

$$\begin{aligned}\lim_{n\to\infty} P(X_n = x) &= \lim_{n\to\infty} \binom{n}{x} p_n^x (1-p_n)^{n-x} \\ &\approx \lim_{n\to\infty} \frac{\prod_{i=1}^{x}(n-i+1)}{x!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^n \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x} \\ &= \lim_{n\to\infty} \frac{\prod_{i=1}^{x}(n-i+1)}{x!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^n \\ &\approx \lim_{n\to\infty} \frac{n^x\lambda^x}{x!n^x}\left(1-\frac{\lambda}{n}\right)^n \\ &= \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}.\end{aligned}$$

□

The term *rare events* refers to the situation where the probability of two events occurring at the same time for some small interval h is $o(h)$, i.e., $\lim_{h\to 0} \frac{p}{h} = 0$.

Note that this means that we can use $e^{-\lambda}\frac{\lambda^i}{i!}$ as a good approximation for $p_X(i)$ when n is large and p is small! In this case, λ is the expected frequency of occurrences of the event corresponding to $X = 1$ within a unit interval.

Definition 3.2.14 ▸ Poisson Random Variable

A random variable X is a **Poisson random variable** if

$$p_X(x) = P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

for some $\lambda > 0$, denoted as $X \sim \text{Po}(\lambda)$.

Note that for $X \sim \text{Po}(\lambda)$, we can find some $Y \sim \text{B}(n, p)$ where n is large and p is small such that $np = \lambda$. Therefore, it is expected that

$$\begin{aligned}\mathbb{E}[X] &\approx \mathbb{E}[Y] = \lambda,\\ \text{Var}(X) &\approx \text{Var}(Y) = np(1-p) \approx \lambda.\end{aligned}$$

Proposition 3.2.15 ▸ Expectation and Variance of Poisson Random Variables

If $X \sim \text{Po}(\lambda)$ *where* $\lambda > 0$, *then* $\mathbb{E}[X] = \text{Var}(X) = \lambda$.

Proof. Left as an exercise to the reader. □

To further examine the properties of Poisson random variables, we introduce the notion of *weak dependence*.

Definition 3.2.16 ▸ Weak Dependence

Let E and F be two events. If $P(E) \approx P(E \mid F)$, we say that E and F are **weakly dependent**.

Let $i = 1, 2, 3, \cdots, n$ and p_i be the probability of event i occurring. If the i's are independent or weakly dependent, then we can approximate for large n that the rate of occurrences of these events is $\sum_{i=1}^{n} p_i$. Let X be the number of events which occur, then

$$X \sim \text{Po}\left(\sum_{i=1}^{n} p_i\right).$$

Definition 3.2.17 ▸ Poisson Process

Poisson process with **rate** or **intensity** $\lambda > 0$ is a sequence of discrete random variables $\{X(t) : t \in \mathbb{N}\}$ such that $X(0) = 0$ and

- for any $t_0, t_1, \cdots, t_n \in \mathbb{N}$ with $t_0 = 0$, $\{X(t_i) - X(t_{i-1}) : i = 1, 2, \cdots, n\}$ are mutually independent;
- for any $s \in \mathbb{N}$ and $t \in \mathbb{N}^+$, we have $X(s+t) - X(s) \sim \text{Po}(\lambda t)$.

3.2.3 Geometric Random Variable

Suppose we perform some experiment with a probability of success of p. Let X be the number of failures before the first success occurs, then clearly,

$$P(X = x) = (1-p)^x p.$$

Additionally, let Y be the number of trials needed to reach the first success, then

$$P(Y = y) = (1-p)^{y-1} p.$$

Note that both $\big(P(X = x)\big)$ and $\big(P(Y = y)\big)$ form geometric sequences.

Definition 3.2.18 ▸ Geometric Random Variable

A random variable X is called a **geometric random variable** with parameter $p \in (0, 1)$, denoted by $X \sim \text{Geo}(p)$, if

$$p_X(n) = (1-p)^{n-1} p.$$

The expectation and variance of geometric random variables are given as follows:

Proposition 3.2.19 ▸ Expectation and Variance of Geometric Random Variables

If $X \sim \text{Geo}(p)$, *then* $\mathbb{E}[X] = \frac{1}{p}$ *and* $\text{Var}(X) = \frac{1-p}{p^2}$.

Proof. Left as an exercise to the reader. □

3.2.4 Negative Binomial Random Variable

Suppose we perform some experiment with a probability of success of p. Let X be the number of trials needed to achieve the r-th success, then clearly, for $X = n$, we need $(r-1)$ successes (i.e., $(n-r)$ failures) in the first $(n-1)$ trials and the r-th trial to be a success.

Therefore,

$$P(X = n) = C_{r-1}^{n-1} p^{r-1}(1-p)^{n-r} p = C_{r-1}^{n-1} p^r (1-p)^{n-r}.$$

Definition 3.2.20 ▶ Negative Binomial Random Variable

A random variable X is called a **negative binomial random variable** if

$$p_X(n) = \binom{n-1}{r-1} p^r (1-p)^{n-r},$$

where $0 < p < 1$ and $n \geq r$, denoted as $X \sim \text{NB}(r, p)$.

The expectation and variance of negative binomial random variables are given as follows:

Theorem 3.2.21 ▶ Expection and Variance of Negative Binomial Variables

Let $X \sim \text{NB}(r, p)$, then

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

Proof. Left as an exercise to the reader. □

3.2.5 Hypergeometric Random Variable

Suppose a collection contains N objects, m of which are of type A. If n objects are selected randomly without replacement and let X be the number of objects of type A selected, then

$$P(X = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}.$$

Definition 3.2.22 ▶ Hypergeometric Random Variable

A random variable X is called a **hypergeometric random variable** with parameters (n, N, m) if

$$p_X(x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}},$$

where $0 \leq m, n \leq N$.

The expectation and variance of hypergeometric random variables are given as follows:

Theorem 3.2.23 ▶ Expection and Variance of Hypergeometric Random Variables

Let X be a hypergeometric random variable with parameters (n, N, m), then

$$\mathbb{E}[X] = np, \qquad \mathrm{Var}(X) = np(1-p)\left(1 - \frac{n-1}{N-1}\right).$$

Proof. Left as an exercise to the reader. □

3.3 Continuous Random Variables

In real life, the outcomes of certain events are infinitely many, and so they cannot be enumerated as discrete cases. Thus, we will need to use *continuous random variables* to model these events.

Definition 3.3.1 ▶ Continuous Random Variable

A random variable X is a **continuous random variable** if there exists some non-negative function f_X such that for all $B \subseteq \mathbb{R}$,

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

The function f_X is known as the **probability density function** of X. The function F_X with $0 \leq F_X(x) \leq 1$ and $F_X'(x) = f_X(x)$ is known as the **cumulative distribution function** of X.

An interesting property of a continuous random variable X is that

$$P(X = x) = \int_x^x f_X(x)\,\mathrm{d}x = 0,$$

which means that the probability of any single outcome of an event is 0, but this does not mean that it is impossible to occur! In particular, it is more meaningful to consider the probability of the occurrence of a range of outcomes. We have

$$P(a \le X \le b) = P(a < X < b) = P(a \le X < b) = P(a < X \le b) = \int_a^b f_X(x)\,\mathrm{d}x,$$
$$P(X \le a) = P(X < a) = \int_{-\infty}^a f_X(x)\,\mathrm{d}x.$$

It is not surprising that a function of a continuous random variable is still a continuous random variable.

Theorem 3.3.2 ▸ Function of Continuous Random Variables

Let X be a continuous random variable and $Y = g(X)$. If g is strictly monotonic and differentiable, then

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right|.$$

Let X be a continuous random variable with probability density function f_X such that for all $x \in \mathbb{R} - [a,b]$, we have $f_X(x) = 0$. We can divide $[a,b]$ into n intervals $[x_{i-1}, x_i]$ for each $i = 1, 2, 3, \cdots, n$ with equal length $\Delta x = \frac{b-a}{n}$. Thus,

$$P(x_{i-1} < X < x_i) \approx \Delta x f_X(x_i).$$

Let Y be a discrete random variable with $P(Y = x_i) = \Delta x f_X(x_i)$, then

$$\mathbb{E}[X] \approx \mathbb{E}[Y] = \sum_{i=1}^n x_i \Delta x f_X(x_i).$$

When $n \to \infty$, i.e., $\Delta x \to 0$, we have

$$\lim_{\Delta x \to 0} \mathbb{E}[Y] = \int_a^b x f_X(x)\,\mathrm{d}x.$$

By letting $a \to -\infty$ and $b \to \infty$, we have arrived at the following definition:

Definition 3.3.3 ▸ Expectation of Continuous Random Variables

Let X be a continuous random variable with probability density function f_X, then

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x.$$

Let Y be a continuous random variable with probability density function f, consider

$$\begin{aligned}\int_0^\infty P(Y > y)\,\mathrm{d}y &= \int_0^\infty \int_y^\infty f(x)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^\infty \int_0^x f(x)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_0^\infty x f(x)\,\mathrm{d}y\,\mathrm{d}x \\ &= \mathbb{E}[Y].\end{aligned}$$

Therefore, set $Y = g(X)$, then similar to discrete random variables, we have

$$\mathbb{E}\left[g(X)\right] = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x.$$

3.3.1 Uniform Random Variable

Intuitively, we may call a random variable X "uniformly" distributed in (a, b) if $P(X = x)$ is a constant for all $x \in (a, b)$.

Definition 3.3.4 ▸ Uniform Random Variable

A continuous random variable X is a **uniform random variable** if

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

denoted by $X \sim \mathrm{U}(a, b)$.

Let $X \sim \mathrm{U}(a, b)$, then the cumulative density function is

$$F_X(x) = \begin{cases} 0, & \text{if } x \le a \\ \frac{x-a}{b-a}, & \text{if } a < x < b \\ 1, & \text{otherwise} \end{cases}.$$

Proposition 3.3.5 ▸ Expectation and Variance of Uniform Random Variables

Let $X \sim \mathrm{U}(a, b)$, *then* $\mathbb{E}[X] = \frac{a+b}{2}$ *and* $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.

Proof. Left as an exercise to the reader. □

3.3.2 Normal Random Variable

Definition 3.3.6 ▶ Normal Random Variable

A continuous random variable Z with probability density function ϕ is a **normal random variable** if

$$\phi(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\mu)^2}{2\sigma^2}},$$

denoted as $Z \sim \mathcal{N}(\mu, \sigma^2)$.

In particular, $Z \sim \mathcal{N}(0,1)$ is known as the *standard normal random variable*. Let Φ be the cumulative density function for Z, then

$$\Phi(z) = P(Z < z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{-\frac{t^2}{2}} \, \mathrm{d}t.$$

Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$, so

$$\Phi_X(x) = P(X < x) = P\left(\frac{X-\mu}{\sigma} < \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

We shall state without proof the expectation and variance of normal random variables as follows:

Proposition 3.3.7 ▶ Expection and Variance of Normal Random Variables

Let $Z \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}[Z] = \mu$ and $\mathrm{Var}(Z) = \sigma^2$.

3.3.3 Exponential Random Variable

Let $N(t) \sim \mathrm{Po}(t\lambda)$ be the number of occurrences of an event in an interval of length t. Suppose X is the time before the first occurrence of the event, then

$$P(X > t) = P(N(t) = 0) = \mathrm{e}^{-\lambda t}.$$

In other words, if F_X is the cumulative distribution function of X, then $F_X(x) = 1 - \mathrm{e}^{-\lambda t}$.

Definition 3.3.8 ▶ Exponential Random Variable

A continuous random variable X is an **exponential random variable** if

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

where $\lambda > 0$, denoted as $X \sim \text{Exp}(\lambda)$.

The expectation and variance of exponential random variables are as follows:

Proposition 3.3.9 ▶ Expection and Variance of Exponential Random Variables

Let $X \sim \text{Exp}(\lambda)$, *then* $\mathbb{E}[X] = \frac{1}{\lambda}$ *and* $\text{Var}(X) = \frac{1}{\lambda^2}$.

Proof. Left as an exercise to the reader. □

Informally, an exponential random variable models the **waiting time** before the first event occurs or between the occurrences of two consecutive events. Suppose we have already waited for s unit of time for the occurrence, we may wish to know the probability of us having to wait for another t unit of time. To solve such questions, we need to understand the *memoryless* property.

Definition 3.3.10 ▶ Memoryless Property

Let X be a random variable, we say that X is **memoryless** if

$$P(X > s + t \mid X > t) = P(X > s).$$

In particular, if $X \sim \text{Exp}(\lambda)$, consider

$$\begin{aligned} P(X > s + t \mid X > t) &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} \\ &= e^{-\lambda s} \\ &= P(X > s). \end{aligned}$$

Therefore, exponential random variables are memoryless. One may also prove that geometric random variables are also memoryless.

Now we introduce another random variable which is closely related to the exponential random variable.

Definition 3.3.11 ▸ Double Exponential Random Variable

A continuous random variable X is a **double exponential variable** if

$$f_X(x) = \frac{1}{2}\lambda x^{-\lambda|x|}$$

for some $\lambda > 0$.

Consider $Y = |X|$ where X is a double exponential random variable. The double exponential random variable is so named because for any $y \geq 0$,

$$\begin{aligned} P(Y > y) &= P(X > y) + P(X < -y) \\ &= 2P(X > y) \\ &= 2\int_y^\infty \frac{1}{2}\lambda e^{-\lambda x}\,dx \\ &= e^{-\lambda y}. \end{aligned}$$

Thus, $Y = |X| \sim \text{Exp}(\lambda)$.

A common application of exponential random variables is to determine the *hazard rate*. Suppose X is the survival time of some object and that the object has already survived for a time t. Consider ϵ to be a small interval, then the probability that the object cannot survive past this small interval is approximately

$$\begin{aligned} P(X < t + \epsilon \mid X > t) &= \frac{P(t < X < t + \epsilon)}{P(X > t)} \\ &\approx \frac{\epsilon f_X(t)}{1 - F_X(t)}. \end{aligned}$$

In general, we have the following definition:

Definition 3.3.12 ▸ Hazard Rate Function

Let X be a positive continuous random variable and define $\overline{F_X}(x) = 1 - F_X(x)$, then the function

$$\lambda(x) = \frac{f_X(x)}{\overline{F_X}(x)}$$

is known as the **hazard rate function** of X.

In particular, if $X \sim \text{Exp}(\lambda)$, then its hazard rate funtion is just $\lambda(x) = \lambda$, which is also known as the *rate* of X.

3.3.4 Gamma Random Variable

We have seen that the exponential random variable can be used to model the waiting time between two consecutive occurrences of an event modelled by a Poisson random variable. We would also like to know the waiting time till the n-th occurrence of an event.

Let $N(t)$ be a Poisson process with rate λ, so $P\big(N(t)=n\big)=\frac{\mathrm{e}^{-\lambda t(\lambda t)^n}}{n!}$. Let T_n be the waiting time till the n-th event with f_k being the probability density function, then

$$\begin{aligned} f_k(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\big(1-P(T_n>t)\big) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\Big(1-P\big(N(t)<k\big)\Big) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\left(1-\mathrm{e}^{-\lambda t}\sum_{i=0}^{k-1}\frac{(\lambda t)^i}{i!}\right) \\ &= \frac{\lambda \mathrm{e}^{-\lambda t}(\lambda t)^{k-1}}{(k-1)!}. \end{aligned}$$

Definition 3.3.13 ▸ Gamma Random Variable

A continuous random variable X is called a **gamma** random variable with parameters (n,λ) where $\lambda>0$ if

$$f(t)=\frac{\lambda \mathrm{e}^{-\lambda t}(\lambda t)^{n-1}}{(n-1)!}$$

for all $t\geq 0$.

The gamma random variable is closely related to the *gamma function*.

Definition 3.3.14 ▸ Gamma Function

Let $\alpha>0$, the **gamma function** is defined as

$$\begin{aligned} \Gamma(\alpha) &= \int_0^\infty \lambda \mathrm{e}^{-\lambda t}(\lambda t)^{\alpha-1}\,\mathrm{d}t \\ &= \int_0^\infty \lambda \mathrm{e}^{-x}x^{\alpha-1}\,\mathrm{d}x. \end{aligned}$$

In particular, if X is a continuous random variable with probability density function

$$f(t)=\frac{\lambda \mathrm{e}^{-\lambda t}(\lambda t)^{\alpha-1}}{\Gamma(\alpha)},$$

where $\lambda>0$ and $t\geq 0$, then X is a gamma random variable with parameters (α,λ).

Remark. One can prove that

1. $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ for all $\alpha > 0$.
2. $\Gamma(n) = (n-1)!$ for all $n \in \mathbb{Z}^+$.

We shall state the following result without proof:

Proposition 3.3.15 ▸ Expectation and Variance of Gamma Random Variables

Let X be a gamma random variable with parameters (α, λ), then

$$\mathbb{E}[X] = \frac{\alpha}{\lambda}, \qquad \mathrm{Var}(X) = \frac{\alpha}{\lambda^2}.$$

3.3.5 Beta Random Variable

In real life, sometimes we may not know the exact probability distribution of a random variable and wish to deduce its probability distribution based on experiments. In other words, suppose there are a successes and b failures of an experiment, we wish to know what is the **most likely** probability of success.

Definition 3.3.16 ▸ Beta Random Variable

A continuous random variable X is a **beta** random variable with parameters (a, b) with $a, b > 0$ if

$$f_X(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1},$$

where

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

If $a = b = 1$, we see that

$$f_X(x) = \frac{\Gamma(2)}{\big(\Gamma(1)\big)^2},$$

so X is uniform on $(0,1)$. In particular, we also see that

$$\frac{B(a+1,b)}{B(a,b)} = \frac{a}{a+b}.$$

We shall state the following result without proof:

Proposition 3.3.17 ▶ Expectation and Variance of Beta Random Variables

Let X be a beta random variable with parameters (a, b), then

$$\mathbb{E}[X] = \frac{a}{a+b}, \qquad \mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

3.4 Jointly Distributed Random Variables

Sometimes, the outcomes of the events we wish to study cannot be expressed with a single random variable. In general, if $X_1, X_2, \cdots, X_n$ are random variables, we may be interested to know

$$P\big((X_1, X_2, \cdots, X_n) \in C\big), \qquad C \subseteq \mathbb{R}^n.$$

Take $C = \prod_{i=1}^{n}(-\infty, x_i]$, then $(X_1, X_2, \cdots, X_n) \in C$ if and only if $X_i \leq x_i$ for $i = 1, 2, \cdots, n$.

Definition 3.4.1 ▶ Joint Cumulative Distribution Function

Let $X_1, X_2, \cdots, X_n$ be random variables, then their **joint cumulative distribution function** is defined as

$$F_X(x_1, x_2, \cdots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n).$$

One may be tempted to think that $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$, but in general this is not true!

> *Remark.* $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$ if and only if X and Y are independent random variables (Definition 3.4.7).

Consider two random variables X and Y jointly distributed, observe that

$$\begin{aligned} P(x_1 \leq X \leq x_2, Y \leq y) &= P(X \leq x_2, Y \leq y) - P(X \leq x_1, Y \leq y) \\ &= F_{X,Y}(x_2, y) - F_{X,Y}(x_1, y). \end{aligned}$$

Similarly,

$$P(X \leq x, y_1 \leq Y \leq y_2) = F_{X,Y}(x, y_2) - F_{X,Y}(x, y_1).$$

Combining the two equations we have

$$\begin{aligned} P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) &= P(x \leq x_2, y_1 \leq Y \leq y_2) - P(x \leq x_1, y_1 \leq Y \leq y_2) \\ &= F_{X,Y}(x_2, y_2) - F_{X,Y}(x_2, y_1) - F_{X,Y}(x_1, y_2) + F_{X,Y}(x_1, y_1). \end{aligned}$$

Note that with this identity we can compute $P\big((X, Y) \in C\big)$ for all Borel sets C. Consider

jointly distributed discrete random variables. We first introduce the following intuitive definition:

Definition 3.4.2 ▸ Joint Probability Mass Function

Let $X_1, X_2, \cdots, X_n$ be discrete random variables. Their **joint probability mass function** is defined to be

$$p_{\boldsymbol{X}}(x_1, x_2, \cdots, x_n) = P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n).$$

In the continuous case, we have a similar definition:

Definition 3.4.3 ▸ Joint Continuity

Let $X_1, X_2, \cdots, X_n$ be continuous random variables. They are said to be **jointly continuous** if there exists a **joint probability density function** f such that

$$P\big((X_1, X_2, \cdots, X_n) \in C\big) = \int \cdots \int_C f_{\boldsymbol{X}}(x_1, x_2, \cdots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

Let X, Y be discrete random variables. It is easy to see that

$$P(X = x) = \sum_{i=1}^{\infty} P(X = x, Y = y_i),$$
$$P(Y = y) = \sum_{i=1}^{\infty} P(X = x_i, Y = y).$$

On the other hand, if X, Y are continuous random variables, we have

$$\begin{aligned} P(X \in A) &= P(X \in A, Y \in \mathbb{R}) \\ &= \int_A \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}y\, \mathrm{d}x. \end{aligned}$$

These are known as *marginal probability density functions*.

Definition 3.4.4 ▸ Marginal Probability Density Function

Let X, Y be discrete random variables. Their **marginal probability density func-**

tions are defined as

$$p_X(x) = \sum_{i=1}^{\infty} p_{X,Y}(x, y_i),$$
$$p_Y(y) = \sum_{i=1}^{\infty} p_{X,Y}(x_i, y).$$

If X, Y are continuous random variables, then

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}y,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}x.$$

Lastly, we state the notion of *joint distribution of functions*.

Proposition 3.4.5 ▸ Joint Distribution of Functions

Let X_1 and X_2 be jointly continuous random variables. Let $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$. If for all y_1, y_2, the linear system

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} g_1(x_1, x_2) \\ g_2(x_1, x_2) \end{bmatrix}$$

has a unique solution, and the Jacobian

$$J(x_1, x_2) = \frac{\partial y_1}{\partial x_1} \cdot \frac{\partial y_2}{\partial x_2} - \frac{\partial y_1}{\partial x_2} \cdot \frac{\partial y_2}{\partial x_1}$$

is continuous and non-zero, then

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1, x_2)|J(x_1, x_2)|^{-1}.$$

In general, we have:

Proposition 3.4.6 ▸ Expectation of Jointly Distributed Random Variables

Let X and Y be jointly distributed random variables. If X and Y are discrete, then

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

If X and Y are continuous, then

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Recall that we have previously defined the notion of independent events. Observe that an event can be precisely expressed as $X \in A$ for some random variable X and set A, which motivates us to define independence of random variables.

Definition 3.4.7 ▶ Independent Random Variables

Two random variables X and Y are **independent** if for any sets A and B,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

Remark. Less rigorously, notice that for Borel sets A and B, since they can be expressed as countable unions and intersections of intervals, we can say that X and Y are independent if and only if $F_{X,Y}(x,y) = F_X(x)F_Y(y)$.

Since the cumulative distribution functions are closely related to the probability mass and probability density, we can then check for independence using them instead. We first consider the discrete case.

Proposition 3.4.8 ▶ Independence of Discrete Random Variables

Let X and Y be discrete random variables with probability mass functions p_X and p_Y, then X and Y are independent if and only if $p_{X,Y}(x,y) = p_X(x)p_Y(y)$.

Proof. Left as an exercise to the reader. □

Based on Proposition 3.4.8, one may check that if X and Y are independent discrete random variables, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y], \qquad \mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

In the continuous case, we have a similar conclusion.

Theorem 3.4.9 ▶ Independence of Continuous Random Variables

Let X and Y be continuous random variables with probability density functions f_X and f_Y. X and Y are independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

Proof. Left as an exercise to the reader. □

The above theorems and definitions can be easily extended to a countable number of independent random variables.

Suppose X and Y are independent **integer-valued** discrete random variables, then clearly

$$\begin{aligned} p_{X+Y}(n) &= P(X+Y=n) \\ &= \sum_i P(X=i, Y=n-i) \\ &= \sum_i p_X(i) p_Y(n-i) \\ &= \sum_{i+j=n} p_X(i) p_Y(j). \end{aligned}$$

On the other hand, if X and Y are independent continuous random variables, we have

$$\begin{aligned} F_{X+Y}(n) &= \iint_{x+y\le n} f_{X+Y}(x+y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{n-y} f_X(x) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} F_X(n-y) f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} F_X(n-y)\,\mathrm{d}F_Y(y). \end{aligned}$$

We say that F_{X+Y} is the *convolution* $F_X * F_Y$.

By Proposition 3.4.6, it is easy to see that if $X_1, X_2, \cdots, X_n$ are random variables,

$$\mathbb{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{E}[X_i].$$

Using the expectation of the sum of random variables, we can also prove Theorem 3.4.10 but for the finite cases:

Theorem 3.4.10 ▸ Boole's Inequality for Finitely Many Events

Let $A_1, A_2, \cdots, A_n$ be events, then

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

Proof. Define

$$X_i = \begin{cases} 1, & \text{if } A_i \text{ occurs} \\ 0, & \text{otherwise} \end{cases}.$$

Let $X = \sum_{i=1}^{n} X_i$ and $Y = \max\{X_1, X_2, \cdots, X_n\}$. Note that $Y = 1$ if and only if at least one of the A_i's occurs, so

$$\mathbb{E}[Y] = P\left(\bigcup_{i=1}^{n} A_i\right).$$

Note that $\mathbb{E}[X] = \sum_{i=1}^{n} P(A_i)$ and that $X \geq Y$, so

$$\sum_{i=1}^{n} P(A_i) = \mathbb{E}[X] \geq \mathbb{E}[Y] = P\left(\bigcup_{i=1}^{n} A_i\right).$$

□

Now, we proceed to discussing some special cases.

Let X and Y be independent uniform random variables on (a, b). Note that $f_Y(y) = \frac{1}{b-a}$ for all $y \in (a, b)$, so

$$\begin{aligned} f_{X+Y}(n) &= \int_a^b f_X(n-y) f_Y(y)\,\mathrm{d}y \\ &= \frac{1}{b-a}\int_a^b f_X(n-y)\,\mathrm{d}y. \end{aligned}$$

Note that $f_X(n-y) = \frac{1}{b-a}$ if and only if $a < n - y < b$, i.e., $n - b < y < n - a$. Otherwise, we have $f_X(n-y) = 0$. If $2a < n < a + b$, we have $(a, b) \cap (n-b, n-a) = (a, n-a)$, so

$$\begin{aligned} \frac{1}{b-a}\int_a^b f_X(n-y)\,\mathrm{d}y &= \frac{1}{b-a}\int_a^{n-a} \frac{1}{b-a}\,\mathrm{d}y \\ &= \frac{n-2a}{(b-a)^2}. \end{aligned}$$

If $a + b < n < 2b$, then $(a, b) \cap (n-b, n-a) = (n-b, b)$, so

$$\begin{aligned} \frac{1}{b-a}\int_a^b f_X(n-y)\,\mathrm{d}y &= \frac{1}{b-a}\int_{n-b}^b \frac{1}{b-a}\,\mathrm{d}y \\ &= \frac{2b-n}{(b-a)^2}. \end{aligned}$$

Therefore,

$$f_{X+Y}(n) = \begin{cases} \frac{n-2a}{(b-a)^2} & \text{if } 2a < n \leq a+b \\ \frac{2b-n}{(b-a)^2} & \text{if } a+b < n < 2b \\ 0 & \text{otherwise} \end{cases}.$$

This is known as a *triangular distribution*.

Let X and Y be independent gamma random variables with parameters (α, λ) and (β, λ). One may check with some computation that

$$\begin{aligned} f_{X+Y}(n) &= \frac{B(\alpha,\beta)}{\Gamma(\alpha)\Gamma(\beta)} \lambda e^{-\lambda n} (\lambda n)^{\alpha+\beta-1} \\ &= \frac{1}{\Gamma(\alpha+\beta)} \lambda e^{-\lambda n} (\lambda n)^{\alpha+\beta-1}. \end{aligned}$$

Therefore, $X + Y$ is a gamma random variable with parameters $(\alpha + \beta, \lambda)$.

Lastly, we shall state the following result without proof:

Proposition 3.4.11 ▸ Sum of Normal Random Variables Is Normal

Let $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

So far we have been discussing the sum of finitely many random variables. It turns out that in the infinite case, the following applies:

Theorem 3.4.12 ▸ Expectation of Infinite Sum of Random Variables

Let $X_1, X_2, \cdots$ be infinitely many random variables, if

- *$X_i \geq 0$ for all $i \in \mathbb{N}^+$, or*
- *the series $\sum_{i=1}^{\infty} \mathbb{E}[|X_i|]$ converges,*

then

$$\mathbb{E}\left[\sum_{i=1}^{\infty} X_i\right] = \sum_{i=1}^{\infty} \mathbb{E}[X_i].$$

3.5 Conditional Distribution

Recall that previously, we have defined the conditional probability

$$P(E \mid F) = \frac{P(EF)}{P(F)}.$$

Let X and Y be discrete random variables, we can similarly see that

$$\begin{aligned} P(X = x \mid Y = y) &= \frac{P(X = x, Y = y)}{P(Y = y)} \\ &= \frac{p_{X,Y}(x, y)}{p_Y(y)}. \end{aligned}$$

Definition 3.5.1 ▶ Conditional Probability Mass Function

Let X and Y be discrete random variables with probability mass functions p_X and p_Y respectively, the **conditional probability mass function** of X given $Y = y$ is defined as

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

Remark. In particular, we have

$$P(X = x \mid X \in A) = \begin{cases} \frac{P(X=x)}{P(X \in A)}, & \text{if } x \in A \\ 0, & \text{otherwise} \end{cases}.$$

We can similar define for the continuous case:

Definition 3.5.2 ▶ Conditional Probability Density Function

Let X and Y be jointly continuous random variables with probability density functions f_X and f_Y respectively, the **conditional probability density function** of X given $Y = y$ is defined as

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

Remark. In particular, we have

$$f_{X|X \in A}(x) = \frac{f_X(x)}{\int_A f_X(x)\,\mathrm{d}x}.$$

Now, we can compute the conditional probability for jointly distributed continuous random variables with

$$P(X \in A \mid Y = y) = \int_A f_{X|Y}(x \mid y)\,\mathrm{d}x.$$

In particular, we could define the *conditional cumulative distribution function* as

$$\begin{aligned} F_{X|Y}(x \mid y) &= P(X \leq x \mid Y = y) \\ &= \int_{-\infty}^{x} f_{X|Y}(x \mid y)\,\mathrm{d}x \end{aligned}$$

What if X is continuous but Y is discrete? In this case, we have

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

3.6 Generating Functions

Capturing the distribution of a random variable could be difficult as the probabilities might be quite irregular. Luckily, a type of tools known as generating functions provide a succinct way to describe such information.

Definition 3.6.1 ▶ Probability Generating Function

Let X be an $\mathbb{N}$-valued discrete random variable, then its **probability generating function** is a map $\phi_X : T \to \mathbb{R}$ defined as

$$\phi_X(t) := \mathbb{E}\left[t^X\right] = \sum_{k \in \mathbb{N}} P(X = k)\, t^k,$$

where $T \subseteq \mathbb{R}$ is the set of all values of t such that the sum converges.

Notice that

$$\phi_X^{(n)}(t) = \sum_{k=n}^{\infty} k! P(X = k)\, t^{k-n},$$

so it is clear that

$$P(X = x) = \frac{\phi_X^{(x)}(0)}{x!},$$

hence the name "probability generating function". Moreover, we have $\phi_X(0) = P(X = 0)$ and $\phi(1) = 1$. Using the properties of expectation, it is easy to prove the following result:

Proposition 3.6.2 ▶ Probability Generating Function of Sum of Random Variables

If X and Y are independent $\mathbb{N}$-valued discrete random variables, then $\phi_{X+Y} = \phi_X \phi_Y$.

Consider n events $A_1, A_2, \cdots, A_n$. Define

$$X_i = \begin{cases} 1, & \text{if } A_i \text{ occurs} \\ 0, & \text{otherwise} \end{cases},$$

then $X = \sum_{i=1}^{n} X_i$ is the number of events which have occurred. Observe that

$$\mathbb{E}[X] = \sum_{i=1}^{n} P(A_i).$$

Define the event

$$E_{m,k} = \bigcap_{i=1}^{k} A_{m_i}$$

where $1 \leq m_i \leq n$ for any $i = 1, 2, \cdots, k$, then the number of such $E_{m,k}$'s which have occurred is given by $\binom{X}{k}$. Note that the event $\bigcap_{i=1}^{k} A_{m_i}$ occurs if and only if $\prod_{i=1}^{k} X_{m_i} = 1$, so

$$\binom{X}{k} = \sum_{m_1 < m_2 < \cdots < m_k} \left(\prod_{i=1}^{k} X_{m_i} \right).$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[\binom{X}{k} \right] &= \mathbb{E}\left[\sum_{m_1 < m_2 < \cdots < m_k} \left(\prod_{i=1}^{k} X_{m_i} \right) \right] \\ &= \sum_{m_1 < m_2 < \cdots < m_k} P\left(E_{m,k}\right). \end{aligned}$$

Notice that $\mathbb{E}\left[\binom{X}{k}\right]$ is a linear combination of the first k-th moments of X. We would like to find a way to quickly compute the n-th moment of a random variable. Let X be a discrete random variable, then

$$\mathbb{E}\left[X^n\right] = \sum_{x} x^n p_X(x).$$

Note that by Maclaurin Series, we have

$$g(t) = \sum_{n=0}^{\infty} \frac{g^{(n)}(0)}{n!} t^n$$

for any function g. A motivation is to construct g such that $g^{(n)}(0) = \mathbb{E}[X^n]$. Therefore,

$$\begin{aligned} g(t) &= \sum_{n=0}^{\infty} \frac{\mathbb{E}[X^n]}{n!} t^n \\ &= \sum_{n=0}^{\infty} \frac{\sum_x x^n p_X(x)}{n!} t^n \\ &= \sum_x \left(p_X(x) \sum_{n=0}^{\infty} \frac{(tx)^n}{n!} \right) \\ &= \sum_x \mathrm{e}^{tx} p_X(x) \\ &= \mathbb{E}\left[\mathrm{e}^{tX}\right]. \end{aligned}$$

Definition 3.6.3 ▸ Moment Generating Function

Let X be an $\mathbb{N}$-valued random variable, then its **moment generating function** is a map $M_X : T \to \mathbb{R}$ defined as

$$M_X(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right] = \sum_{k \in \mathbb{N}} P(X = k)\, \mathrm{e}^{tk},$$

where $T \subseteq \mathbb{R}$ is the set of all values of t such that the sum converges.

By observation, we see that

$$M_X^{(n)}(0) = \sum_{k \in \mathbb{N}} k^n P(X = k)$$

yields the n-th moment of X. However, notice that $M_X(t) = \phi_X\left(e^t\right)$, so actually

$$\mathbb{E}[X^n] = \left. \frac{\mathrm{d}^n \phi\left(\mathrm{e}^t\right)}{\mathrm{d}t^n} \right|_{t=0}.$$

Next, we state a result without proof on the use of generating functions to study the convergence of distributions.

Proposition 3.6.4 ▸ Convergence of Generating Functions

Let $\{X_i\}_{i \in \mathbb{N}}$ be a family of $\mathbb{N}$-valued random variables such that ϕ_i is the probability generating function for X_i and X be an $\mathbb{N}$-valued random variable with probability generating function ϕ. If there exists some $a \in \mathbb{R}$ such that $\{\phi_i\}_{i \in \mathbb{N}}$ converges to ϕ point-wisely,

then

$$\lim_{n\to\infty} P(X_n = k) = P(X = k)$$

for all $k \in \mathbb{N}$.

It can be proven that for any random variable X, M_X is unique. In other words, if two random variables have the same moment generating function, they must be the same random variable.

Theorem 3.6.5 ▶ Uniqueness of Moment Generating Function

Let X and Y be random variables. If

$$\lim_{t\to 0} M_X(t) = \lim_{t\to 0} M_Y(t),$$

then X and Y have the same distribution.

By properties of exponential functions, the following theorem can also be easily proven:

Proposition 3.6.6 ▶ Moments of Sum of Independent Random Variables

Let $X_1, X_2, \cdots, X_n$ be independent random variables, then

$$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t).$$

Proof. Left as an exercise to the reader. □

However, note that the converse of Proposition 3.6.6 is not true in general.

Definition 3.6.7 ▶ Joint Moment Generating Function

Let X and Y be random variables. The **joint moment generating function** of X and Y is defined as

$$M_{X,Y}(s,t) = \mathbb{E}\left[e^{sX+tY}\right].$$

Observe that in the joint case, $M_X(s) = M_{X,Y}(s,0)$ and $M_Y(t) = M_{X,Y}(0,t)$. X and Y are independent if and only if $M_{X,Y}(s,t) = M_X(s)M_Y(t)$. Note that in reality, many random variables are not integer-valued. Therefore, we should extend the notion of generating functions to the general case.

Definition 3.6.8 ▶ Laplace Transform of Random Variables

Let X be a non-negative random variable. The **Laplace transform** of X is defined as a map $\Lambda : \mathbb{R}_0^+ \to \mathbb{R}$ such that $\Lambda(\lambda) \coloneqq \mathbb{E}\left[e^{-\lambda X}\right]$.

One may check that the properties of a Laplace transform is analogous to those of generating functions:

1. $\Lambda_X(\lambda) \in [0,1]$ for all $\lambda \geq 0$;
2. $\Lambda_X(0) = 1$ and Λ_X is decreasing in λ;
3. $\Lambda_X^{(n)}(0) = (-1)^n \mathbb{E}[X^n]$, so Λ_X determines the moments;
4. if $\{\Lambda_n\}_{n\in\mathbb{N}}$ is a family of Laplace transforms for random variables $\{X_n\}_{n\in\mathbb{N}}$ respectively, then $\{\Lambda_n\}_{n\in\mathbb{N}}$ converging to Λ, the Laplace transform of a random variable X, point-wisely on $[0,a]$ for some $a \in \mathbb{R}$ implies that $\{p_{X_n}\}_{n\in\mathbb{N}}$ converges point-wisely to p_X on $\mathbb{R}_0^+$.

Note that we choose $\lambda \geq 0$ which will ensure that $\mathbb{E}\left[e^{-\lambda X}\right]$ is finite if X is non-negative. However, for general random variables which can take negative values, we need to use a different construction.

Definition 3.6.9 ▶ Fourier Transform of Random Variables

Let X be a real-valued random variable. The **Fourier transform** of X is defined as a map $\varphi_X : \mathbb{R} \to \mathbb{R}$ such that $\varphi_X(t) \coloneqq \mathbb{E}\left[e^{itX}\right]$.

One may check that the properties of a Fourier transform is analogous to those of Laplace transforms:

1. $|\varphi_X(t)| \in [0,1]$ for all $t \in \mathbb{R}$;
2. $\varphi_X(0) = 1$;
3. $\varphi_X^{(n)}(0) = i^n \mathbb{E}[X^n]$, so φ_X determines the moments;
4. if $\{\varphi_n\}_{n\in\mathbb{N}}$ is a family of Laplace transforms for random variables $\{X_n\}_{n\in\mathbb{N}}$ respectively, then $\{\varphi_n\}_{n\in\mathbb{N}}$ converging to φ, the Fourier transform of a random variable X, point-wisely on $[-a,a]$ for some $a > 0$ implies that $\{p_{X_n}\}_{n\in\mathbb{N}}$ converges point-wisely to p_X on $\mathbb{R}$.

4 Methodologies for Data Analysis

4.1 Sampling

Note that the expectation and variance are crucial to understand the behaviour of a random variable. However, in the real world, the data based on which we can observe a random variable is always finite. Therefore, it becomes a curious problem as to how we can make some judgement about the behaviour of the random variable based on limited data as accurately as possible.

First, let us define quantitatively what we mean by an "accurate" judgement.

Definition 4.1.1 ▸ Unbiased Estimator

Let x be an unknown quantity to be estimated and X be a random variable representing the observed values obtained for x. We say that X is an **unbiased estimator** of x if $\mathbb{E}[X] = x$.

When it comes to estimating the mean of a random variable, a common practice is to take a large number of samples and compute the average value over the samples.

Definition 4.1.2 ▸ Sample Mean

Let $X_1, X_2, \cdots, X_n$ be independent random variables. The **sample mean** is defined as

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Clearly, if $X_i = \mu$ for all $i = 1, 2, \cdots, n$,

$$\mathbb{E}\left[\overline{X}\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \mu.$$

This means that given a random sample, its sample mean is an *unbiased estimate* of its expectation, so our conventional data collection method does have a theoretical back-up.

Next, let us try to estimate the variance of a random variable. An immediate idea is to take each data point from our samples and compute its distance away from the sample mean.

Definition 4.1.3 ▶ Deviation

Let $X_1, X_2, \cdots, X_n$ be independent random variables with mean μ and variance σ^2. The **deviation** is defined as

$$X_i - \overline{X} = X_i - \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Similar to the sample mean, we would wish the deviation gives an unbiased estimate for the variance. However,

$$\mathbb{E}\left[X_i - \overline{X}\right] = \frac{n-1}{n}\sigma^2.$$

Therefore, we need to eliminate the bias from the deviation.

Definition 4.1.4 ▶ Sample Variance

Let $X_1, X_2, \cdots, X_n$ be independent random variables with mean μ and variance σ^2. The **sample variance** is defined as

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

One may check via some direct computation that the sample variance gives an unbiased estimate for the variance.

4.2 Correlation

We our data set involves multiple random variables, we wish to examine the relationship between them. Suppose that X and Y are independent random variables. We know that

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

This motivates us to consider the quantity $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ in the general case to determine "how dependent" two random variables are.

Definition 4.2.1 ▶ Covariance

Let X and Y be random variables. The **covariance** of X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Remark. We have seen that if X and Y are independent, then $\text{Cov}(X,Y) = 0$. But the converse is not true in general!

Note that covariance has the following properties:

1. $\text{Cov}(X,X) = \text{Var}(X)$.
2. $\text{Cov}(X,Y) = \text{Cov}(Y,X)$.
3. $\text{Cov}(aX,Y) = a\text{Cov}(X,Y)$.
4. $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1,Y) + \text{Cov}(X_2,Y)$.

Remark. Let V be the set of all random variables, then $\text{Cov}(\cdot,\cdot) : V \times V \to \mathbb{R}$ is an inner product over V.

Let X and Y be random variables, then by the above properties, we have

$$\begin{aligned}
\text{Var}(X+Y) &= \text{Cov}(X+Y, X+Y) \\
&= \text{Cov}(X, X+Y) + \text{Cov}(Y, X+Y) \\
&= \text{Cov}(X+Y, X) + \text{Cov}(X+Y, Y) \\
&= \text{Cov}(X,X) + \text{Cov}(Y,X) + \text{Cov}(X,Y) + \text{Cov}(Y,Y) \\
&= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y).
\end{aligned}$$

In general, we have the following formula:

Proposition 4.2.2 ▸ Identity of Variances

Let $X_1, X_2, \cdots, X_n$ be random variables, then

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j).$$

In particular, if all of the X_i's are independent, then

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

Proof. Left as an exercise to the reader. □

Intuitively, we view covariance $\text{Cov}(X,Y)$ as a measure of the degree of spread of the joint distribution of X and Y. Naturally, we can make use of it to measure how related are X and Y.

Definition 4.2.3 ▸ Correlation

Let X and Y be random variables with positive variances. The **correlation** of X and Y is defined as

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

In particular, X and Y are said to be **uncorrelated** if $\rho(X,Y) = 0$.

It is easy to see that for $c > 0$, $\rho(\pm cX, Y) = \pm\rho(X,Y)$. Suppose X and Y have variances σ_X^2 and σ_Y^2 respectively, we have

$$\rho(X,Y) = \rho\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) = \mathrm{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right).$$

The above property leads to the following result:

Proposition 4.2.4 ▸ Boundedness of Correlation

Let X and Y be random variables, then

$$-1 \le \rho(X,Y) \le 1.$$

Proof. Left as an exercise to the reader. □

Note that $\mathrm{Var}(X) = 0$ if and only if X is a constant, i.e., $P(X = c) = 1$ and $P(X = x) = 0$ for all $x \neq c$. Therefore, $\rho(X,Y) = \pm 1$ if and only if $Y = \pm aX + b$ for some constants a, b with $a > 0$.

4.3 Prediction

Often, we wish to predict future behaviour of a random variable. To achieve that, we first study the notions of *condition expectation and variance*.

Definition 4.3.1 ▸ Conditional Expectation

Let X and Y be random variables. The **conditional expectation** of X given $Y = y$ is defined as

$$\mathbb{E}[X \mid Y = y] = \sum_x x p_{X|Y}(x \mid y)$$

if X and Y are discrete, and

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y)}{f_Y(y)}\,\mathrm{d}x$$

if X and Y are jointly continuous.

All properties of expectation are still satisfied by conditional expectation. In particular, notice that $\mathbb{E}[X \mid Y]$ is a function in Y, so it is a random variable by itself. Let $Z = \mathbb{E}[X \mid Y]$, what is the expectation of Z?

Theorem 4.3.2 ▸ Law of Total Expectation

Let X and Y are random variables, then

$$\mathbb{E}\big[\mathbb{E}[X \mid Y]\big] = \mathbb{E}[X].$$

Proof. We will only prove the discrete case. Let $f(Y) = \mathbb{E}[X \mid Y]$, then by Proposition 3.2.5, we have

$$\begin{aligned}
\mathbb{E}\big[\mathbb{E}[X \mid Y]\big] &= \mathbb{E}[f(Y)] \\
&= \sum_{y \in \operatorname{ran}(Y)} f(y) P(Y = y) \\
&= \sum_{y \in \operatorname{ran}(Y)} \mathbb{E}[X \mid Y = y] P(Y = y) \\
&= \sum_{y \in \operatorname{ran}(Y)} \sum_{x \in \operatorname{ran}(X)} x P(X = x \mid Y = y) P(Y = y) \\
&= \sum_{x \in \operatorname{ran}(X)} x \sum_{y \in \operatorname{ran}(Y)} P(X = x \mid Y = y) P(Y = y) \\
&= \sum_{x \in \operatorname{ran}(X)} x P(X = x) \\
&= \mathbb{E}[X].
\end{aligned}$$

The continuous case is similar. □

Similarly, we can define *conditional variance*.

Definition 4.3.3 ▸ Conditional Variance

Let X and Y be random variables. The **condtional variance** of X given Y is defined as

$$\operatorname{Var}(X \mid Y) = \mathbb{E}\big[(X - \mathbb{E}[X \mid Y])^2 \mid Y\big].$$

Notice that $\operatorname{Var}(X \mid Y)$ is also a random variable as it is a function of Y. Since

$$\operatorname{Var}(X \mid Y) = \mathbb{E}\left[X^2 \mid Y\right] - \big(\mathbb{E}[X \mid Y]\big)^2,$$

we have

$$\begin{aligned}\mathbb{E}\left[\mathrm{Var}(X \mid Y)\right] &= \mathbb{E}\Big[\mathbb{E}\left[X^2 \mid Y\right]\Big] - \mathbb{E}\left[\left(\mathbb{E}[X \mid Y]\right)^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}\left[\left(\mathbb{E}[X \mid Y]\right)^2\right], \\ \mathrm{Var}\big(\mathrm{Var}(X \mid Y)\big) &= \mathbb{E}\left[(\mathbb{E}[X \mid Y])^2\right] - \left(\mathbb{E}\big[\mathbb{E}[X \mid Y]\big]\right)^2.\end{aligned}$$

With some manipulations we can prove the following proposition:

Theorem 4.3.4 ▸ Law of Total Variance

Let X and Y be random variables, then

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]).$$

Proof. Left as an exercise to the reader. □

In many situations, we seek to **predict** the outcome of an event. Specifically, given $X = x$, we wish to find a function $g : \mathrm{Range}(X) \to \mathrm{Range}(Y)$ such that $g(x) = \widehat{y}$ is the prediction for Y. In an ideal case, we would want $g(X)$ to be the **closest** to Y. In other words, we would minimise $g(X) - Y$. To eliminate the inconvenience of a negative difference, we would choose a g such that $\mathbb{E}\left[\left(Y - g(X)\right)^2\right]$ is minimised.

Let us consider a simpler case. When $g(x) = c$, to minimise $\mathbb{E}\left[(Y - c)^2\right]$, consider

$$\begin{aligned}\mathbb{E}\left[(Y - c)^2\right] &= c^2 - 2c\mathbb{E}[Y] + \mathbb{E}\left[Y^2\right] \\ &= \left(c - \mathbb{E}[Y]\right)^2 + \mathrm{Var}(Y).\end{aligned}$$

Therefore, we need $c = \mathbb{E}[Y]$. Since $c = g(x)$, this implies that $g(x) = \mathbb{E}[Y \mid X = x]$. This is summarised into the following theorem:

Theorem 4.3.5 ▸ Best Predictor

Let X and Y be random variables. Given X, the best predictor of Y is the function

$$g(X) = \mathbb{E}[Y \mid X].$$

Essentially, this implies that for all function f of X,

$$\mathbb{E}\left[\left(Y - f(X)\right)^2\right] \geq \mathbb{E}\left[\left(Y - \mathbb{E}[Y \mid X]\right)^2\right].$$

However, in some cases we do not know the exact joint distribution of X and Y and so we cannot find $\mathbb{E}[Y \mid X]$. Therefore, we may attempt to predict Y with a linear function

of X which can be easily formulated. Hence, we will need to find constants a and b such that $\mathbb{E}\left[(Y-a-bX)^2\right]$ is minimised.

Theorem 4.3.6 ▸ Best Linear Predictor

Let X and Y be random variables with means μ_X and μ_Y, and variances σ_X^2 and σ_Y^2 respectively. Given X, the best linear predictor of Y is

$$g(X) = \mathbb{E}\left[\left(Y - \mu_Y - \rho(X,Y)\frac{\sigma_Y}{\sigma_X}(X-\mu_X)\right)\right].$$

Proof. Let $a+bX$ be the best linear predictor for Y. Define

$$X' = \frac{X-\mu_X}{\sigma_X}, \qquad Y' = \frac{Y-\mu_Y}{\sigma_Y}.$$

Then, $\mathbb{E}[X'] = \mathbb{E}[Y'] = 0$ and $\mathrm{Var}(X') = \mathrm{Var}(Y') = 1$. Note that

$$Y - a - bX = \sigma_Y\left(Y' - \frac{a+b\mu_X-\mu_Y}{\sigma_Y} - \frac{b\sigma_X}{\sigma_Y}X'\right).$$

Therefore, setting $\frac{a+b\mu_X-\mu_Y}{\sigma_Y} = a'$ and $\frac{b\sigma_X}{\sigma_Y} = b'$, we have

$$\begin{aligned}
\mathbb{E}\left[(Y-a-bX)^2\right] &= \sigma_Y^2\,\mathbb{E}\left[(Y'-a'-b'X')^2\right] \\
&= \sigma_Y^2\,\mathbb{E}\left[Y'^2 + a'^2 + b'^2X'^2 - 2a'Y' + 2a'b'X' - 2b'X'Y'\right] \\
&= \sigma_Y^2\left(1 + a'^2 + b'^2 - 2b'\rho(X,Y)\right) \\
&= \sigma_Y^2\left(a'^2 + \left(b'-\rho(X,Y)\right)^2 + 1 - \rho(X,Y)^2\right).
\end{aligned}$$

Therefore, we need $a' = 0$ and $b' = \rho(X,Y)$, i.e.,

$$a = \mu_Y - b\mu_X, \qquad b = \rho(X,Y)\frac{\sigma_Y}{\sigma_X}.$$

Therefore,

$$\begin{aligned}
a + bX &= \mu_Y - b\mu_X + bX \\
&= \mu_Y + b(X-\mu_X) \\
&= \mu_Y + \rho(X,Y)\frac{\sigma_Y}{\sigma_X}(X-\mu_X).
\end{aligned}$$

□

Remark. In particular, the minimum of $\mathbb{E}\left[(Y-a-bX)^2\right]$ is $\sigma_Y^2\left(1-\rho(X,Y)^2\right)$.

5

Limit Theorems

5.1 Bounding Probabilities

Let X be a non-negative random variable and $a > 0$. Define I to be the indicator of X, i.e.,

$$I = \begin{cases} 0 & \text{if } X < a \\ 1 & \text{if } X \geq a \end{cases}.$$

Observe that if $X \geq a$, $aI = a \leq X$ and if $X < a$, $aI = 0 \leq X$. Therefore, $X \geq aI$ for all a and so $\mathbb{E}[X] \geq a\mathbb{E}[I] = aP(X \geq a)$.

This is generalised to the following result, which gives a way to approximate probabilities by the expectation of a random variable:

Theorem 5.1.1 ▸ Markov's Inequality

If X is a non-negative random variable, then $P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$ for all $a > 0$.

Proof. It suffices to prove for the continuous case. Notice that

$$\begin{aligned} \mathbb{E}[X] &= \int_0^\infty x f_X(x)\, \mathrm{d}x \\ &\geq \int_a^\infty x f_X(x)\, \mathrm{d}x \\ &\geq a \int_0^\infty f_X(x)\, \mathrm{d}x \\ &= P(X \geq a). \end{aligned}$$

Therefore, $P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$. □

Note that the bound given by Markov's inequality is a rather loose bound and we can generalise this result. Suppose X is a random variable with mean μ and variance σ^2. Recall that $\sigma^2 = \mathbb{E}\left[(X-\mu)^2\right]$. Set $Y = (X-\mu)^2 \geq 0$. Now we can apply Theorem 5.1.1 to Y and derive the following:

Theorem 5.1.2 ▸ Chebyshev's Inequality

For any real-valued random variable X with finite variance,

$$P\left(|X - \mathbb{E}[X]| > a\sqrt{\mathrm{Var}(X)}\right) \leq \frac{1}{a^2}$$

for all $a > 0$.

Proof. Define $g(X) \coloneqq (X - \mathbb{E}[X])^2$, which is clearly non-negative. By Theorem 5.1.1, we have

$$P\big(g(X) > a^2\mathrm{Var}(X)\big) \leq \frac{\mathbb{E}[g(X)]}{a^2\mathrm{Var}(X)}.$$

Note that $\mathbb{E}[g(X)] = \mathrm{Var}(X)$, so

$$P\left(|X - \mathbb{E}[X]| > a\sqrt{\mathrm{Var}(X)}\right) = P\big(g(X) > a^2\mathrm{Var}(X)\big) \leq \frac{1}{a^2}.$$

□

Remark. In general,

$$P(|X - \mathbb{E}[X]| > a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

Using Theorem 5.1.2, we can get the following highly intuitive result:

Corollary 5.1.3 ▸ Zero-Variance Random Variables Are Constant

If X is a random variable with $\mathrm{Var}(X) = 0$, then $P(X = \mathbb{E}[X]) = 1$.

Proof. Left as an exercise to the reader. □

5.2 Law of Large Numbers

Recall that if $X_1, X_2, \cdots, X_n$ are independent random variables with identical distribution, then $\mathbb{E}\left[\overline{X}\right] = \mathbb{E}[X_i]$ and $\mathrm{Var}\left(\overline{X}\right) = \frac{\mathrm{Var}(X_i)}{n}$. Applying the above theorems we will have the following conclusion:

Theorem 5.2.1 ▸ Weak Law of Large Numbers

Let $X_1, X_2, \cdots, X_n$ be pairwise independent and identically distributed random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 \in \mathbb{R}$ for every $i \in \mathbb{N}^+$. For every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0.$$

Proof. Note that $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu$ and that

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

By Theorem 5.1.2, we have

$$0 \leq \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

By Squeeze Theorem, this clearly implies that

$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) = 0.$$

□

Alternatively, we may phrase Theorem 5.2.1 as "$\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to μ in probability". When a sequence $\{S_n\}_{n=1}^{\infty}$ converges to b in probability, we write $S_n \xrightarrow{\text{p}} b$.

> *Remark.* Essentially, what Theorem 5.2.1 says is that when n is large, the sample mean from n measurements of the same data converges to the expectation of the distribution.

Under some mild conditions, this convergence occurs exponentially fast, i.e., the probability $\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right)$ decreases at least as fast as $\exp\big(-ng(\epsilon)\big)$ for some real-valued function $g: \mathbb{R}^+ \to \mathbb{R}^+$. In terms of asymptotic analysis, we write this as

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) \leq \exp\big(-ng(\epsilon) + o(n)\big).$$

Equivalently, this would imply that there exists a function $g: \mathbb{R} \to \mathbb{R}$ with $g(\epsilon) > 0$ for every $\epsilon > 0$ such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) \geq g(\epsilon) + o(1).$$

There is a strong version for the law, which shall be stated without proof:

Theorem 5.2.2 ▶ Strong Law of Large Numbers

Let $X_1, X_2, \cdots, X_n$ be pairwise independent and identically distributed random variables with $\mathbb{E}[X_i] = \mu$ and $\operatorname{Var}(X_i) = \sigma^2 \in \mathbb{R}$ for every $i \in \mathbb{N}^+$, then

$$\Pr\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1.$$

Sometimes, Theorem 5.2.1 is known as simply the "Law of Large Numbers" whereas Theorem 5.2.2 is known as the "Law of Truly Large Numbers". By applying Theorem 5.2.1, we can prove a classical result known as *Central Limit Theorem*, which will be stated and the interested ones might want to try to prove it on their own.

Theorem 5.2.3 ▶ Central Limit Theorem

Let $X_1, X_2, \cdots, X_n$ be independent random variables with identical distribution, whose mean is μ and variance σ^2, then the distribution of

$$\frac{\sum_{i=0}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

converges to the standard normal distribution as $n \to \infty$.