Contents

1 How to Count

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:

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

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.

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

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

1.2.1 Permutations with Idential Objects

Theorem 1.2.2 ▶ 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 the number of permutations for these k objects are given by

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

1.3 Combinations

Definition 1.3.1 ▶ Combinations

Let A be a finite set such that $|A| = n$, an r-combination of A is a way to choose r elements from A regardless of the order of selection, denoted as C_r^n and 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. $C_r^n = C_{n-r}^n$.

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

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

1.4 Binomial and Multinomial Coefficients

Consider the expansion of $(x+y)^n$ where $n \in \mathbb{N}$. Note that this expansion is a linear combination of terms in the form of $x^k y^{n-k}$ where $k = 0, 1, 2, \cdots, n$.

Thus, fix any k, to determine how many copies of $x^k y^{n-k}$ there are, it suffices to compute C^n_k. Therefore, in the expanded form of $(x+y)^n$, the coefficient is exactly C^n_r.

Theorem 1.4.1 ▶ Binomial Expansion

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

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

We can extend the idea of binomial coefficients onto multinomial expansions, i.e. expressions in the form of $\left(\sum_{i=1}^{r} x_i\right)^n$.

Note that the binomial coefficient C^n_r is essentially equivalent to dividing n distinct elements into two groups with r and $(n-r)$ members respectively. Now we consider dividing n distinct elements into r groups with $n_1, n_2, \cdots, n_r$ members respectively for each group.

Notice that we can simply permute the n distinct elements and assign them sequentially into the r groups, i.e. the first n_1 elements will go into the first group and so on.

Since the order of elements within each group does not matter, we need to remove repeated selections by dividing by $\prod_{i=1}^{r}(n_i!)$. So we have the following definition:

Definition 1.4.2 ▶ Multinomial Coefficients

The **multinomial coefficient** is defined by

$$\binom{n}{n_1, n_2, \cdots, n_k} = \frac{n!}{\prod_{i=1}^{k}(n_i!)}$$

Theorem 1.4.3 ▶ Multinomial Expansion

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

$$\left(\sum_{i=1}^{r} x_i\right)^n = \sum_{\substack{n_1, n_2, \cdots, n_r \in \mathbb{N} \\ \sum_{j=1}^{r} n_j = n}} \left[\binom{n}{n_1, n_2, \cdots, n_r} \prod_{i=1}^{r} x_i^{n_i} \right]$$

2 Axioms of Probability

2.1 Sample Space and Events

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 $S \neq \varnothing$.

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*.

2.2 Probability

Definition 2.2.1 ▸ 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 2.2.2 ▸ Axioms 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 2.2.3 ▸ 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 2.2.4 ▶ 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

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

□

2.3 Inclusion-Exclusion Principle

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 2.3.1 ▶ 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, $g(x) = 0$. □

Theorem 2.3.2 ▶ 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.

Theorem 2.3.3 ▶ Probability in a Finite Sample Space

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

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

3 Conditional Probability

3.1 Conditional Probability

Given a sample space S, we may wish to find the probability of two events E and F both occurring, $P(EF)$. However, suppose that 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 3.1.1 ▸ 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. We can generalise this for a countable number of events:

Proposition 3.1.2 ▸ 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 3.1.1.

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,

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

□

3.2 Bayes's Formula

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. Generalising this further to 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 3.2.1 ▸ Bayes's Formula

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

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

3.3 Indepedent Events

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 3.3.1 ▸ Independent Events

Let S be a sample space and let $E, F \subseteq S$ be two events. We say that E and F are **independent** if $P(EF) = P(E)P(F)$, and **dependent** otherwise.

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 3.3.2

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

Proof. Since $F = (F^c)^c$, it suffices to prove for one direction.

Notice that $EF \cup EF^{\mathrm{c}} = E(F \cup F^{\mathrm{c}}) = E$, so

$$P(E) = P(EF) + P(EF^{\mathrm{c}}) = P(E)P(F) + P(EF^{\mathrm{c}}).$$

Therefore,

$$P(EF^{\mathrm{c}}) = P(E) - P(E)P(F) = P(E)(1 - P(F)) = P(E)P(F^{\mathrm{c}}),$$

and so E and F^{c} are independent. □

4

Random Variables

4.1 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*.

Definition 4.1.1 ▸ Random Variable

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.

4.2 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 4.2.1 ▸ Discrete Random Variable (DRV)

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

Definition 4.2.2 ▸ 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 4.2.3 ▸ 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 $F_X(a) = \sum_{x \leq a} p_X(x)$ 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 $a_i \leq a < a_{i+1}$,

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

4.2.1 Expectation of Discrete Random Variables

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 4.2.4 ▸ Expectation of Discrete Random Variables

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

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

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

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

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:

Theorem 4.2.5 ▸ Expectation of Functions

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

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

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,

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

□

Two simple corollaries to the above theorem are:

$$\begin{aligned} E[aX + b] &= aE[X] + b \\ E[X + Y] &= E[X] + E[Y]. \end{aligned}$$

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 4.2.6 ▸ Moment

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

Following Theorem 4.2.5, it is easy to see that if X is discrete, then

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

4.2.2 Variance

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 $E[X] = \mu$, we wish to determine $E[|X - \mu|]$ or equivalently $E[(X - \mu)^2]$. This spread is known as the *variance* of a random variable.

Definition 4.2.7 ▸ Variance

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

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

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

The formula for $\text{Var}(X)$ can be derived via Theorem 4.2.5:

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

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

Definition 4.2.8 ▸ Standard Deviation

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

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

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

Proposition 4.2.9 ▸ Variance of Linear Combinations of Random Variables

Let X be a random variable, then

$$\begin{aligned}\text{Var}(aX+b) &= a^2\text{Var}(X) \\ \text{SD(aX + b)} &= |a|\text{SD}(X)\end{aligned}$$

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

Proof. By using Theorem 4.2.5, we have

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

Therefore,

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

□

4.2.3 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 4.2.10 ▶ 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 4.2.11 ▶ 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 $N \approx np$.

Theorem 4.2.12 ▶ Expectation and Variance of Binomial Distribution

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

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

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

□

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 $((n+1)p-1, n]$.

4.2.4 Poisson Random Variable

Suppose that $X \sim \mathrm{B}(n, p)$ such that n is large and p is small. Let $\lambda = np$, then

$$\begin{aligned} p_X(i) &= \binom{n}{i} p^i (1-p)^{n-i} \\ &= \frac{\prod_{j=0}^{i-1}(n-j)}{i!} \left(\frac{\lambda}{n}\right)^i \left(1 - \frac{\lambda}{n}\right)^{n-i} \\ &= \frac{\prod_{j=0}^{i-1}(n-j)}{n^i} \cdot \frac{\lambda^i}{i!} \cdot \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^i} \end{aligned}$$

Therefore,

$$\lim_{n\to\infty} p_X(i) = e^{-\lambda}\frac{\lambda^i}{i!}.$$

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 4.2.13 ▶ 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 $X \sim \mathrm{Po}(\lambda)$.

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

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

Theorem 4.2.14 ▶ Expectation and Variance of Poisson Random Variables

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

Definition 4.2.15 ▶ Weakly Dependent

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

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

Theorem 4.2.16 ▶ Poisson Process

Let E be an event which occurs randomly. Assume that

1. *there are λ occurrences per unit interval;*
2. *no two occurrences happen at the same point;*
3. *numbers of occurrences in disjoint intervals are independent.*

Let $N(t)$ be the number of occurrences of E in an interval of length t, then $N(t) \sim \mathrm{Po}(\lambda t)$.

4.2.5 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 4.2.17 ▶ 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

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

Theorem 4.2.18 ▶ Expectation and Variance of Geometric Random Variables

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

4.2.6 Negative Binomial Random Variable

negBinDRV 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 4.2.19 ▶ 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 $X \sim \mathrm{NB}(r, p)$.

Theorem 4.2.20 ▸ Expection and Variance of Negative Binomial Variables

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

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

4.2.7 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

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

Definition 4.2.21 ▸ Hypergeometric Random Variable

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

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

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

Theorem 4.2.22 ▸ Expection and Variance of Hypergeometric Random Variables

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

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

4.3 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 4.3.1 ▶ 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

$$\begin{aligned} P(a \leq X \leq b) = P(a < X < b) =& P(a \leq X < b) = P(a < X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x, \\ P(X \leq a) =& P(X < a) = \int_{-\infty}^a f_X(x)\,\mathrm{d}x. \end{aligned}$$

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

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

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

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

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

Definition 4.3.2 ▸ Expectation of Continuous Random Variables

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

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

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

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

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

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

4.3.1 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 4.3.3 ▸ 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

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

Theorem 4.3.4 ▸ Expectation and Variance of Uniform Random Variables

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

4.3.2 Normal Random Variable

Definition 4.3.5 ▸ 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 $Z \sim \mathcal{N}\left(\mu, \sigma^2\right)$.

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

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

Theorem 4.3.6 ▸ Expection and Variance of Normal Random Variables

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

4.3.3 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 4.3.7 ▸ Exponential Random Variable

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

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

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

Theorem 4.3.8 ▸ Expection and Variance of Exponential Random Variables

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

Informally, an exponential random variable models the **waiting time** before an event occurs. 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 4.3.9 ▸ 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 $X \sim \mathrm{Exp}(\lambda)$, consider

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

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 4.3.10 ▸ 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 $y \geq 0$,

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

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

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

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