MA2116 MA2116T ST2131 Probability

Thang Pang Ern and Ng Kang Zhe

Reference books:

(1) Ross, S. (2020). *A First Course in Probability 10th Edition*. Pearson.

(2) Myers, S. L., Myers, R., Walpole, R. E. and Ye, K. (2011). *Probability and Statistics for Engineers and Scientists 9th Edition*. Pearson.

(3) Koh, K. M. and Chuan, C. C. (1992). *Principles and Techniques in Combinatorics*. World Scientific.

Sixth Term Examination Paper (STEP) Mathematics is a well-established Mathematics examination designed to test candidates on questions that are similar in style to undergraduate mathematics. You can visit their question database for some interesting problems related to Combinatorics and various probability distributions.

Contents

1. Combinatorial Analysis

1.1. *The Basic Principle of Counting*

Many problems in probability theory can be solved simply by counting the number of different ways that a certain event can occur. Effective methods for counting would then be useful in our study of probability. The mathematical theory of counting is formally known as Combinatorial Analysis.

Most of the concepts in this section are taught in Junior College. Hopefully they would be a breeze.

Proposition 1.1 (addition principle). If there are r choices for performing a particular task, and the number of ways to carry out the k^{th} choice is n_k, for $1 \le k \le r$, the total number of ways of performing the particular task is equal to the sum of the number of ways for all the r different choices, i.e.

$$n_1 + n_2 + \ldots + n_r.$$

The different choices cannot occur at the same time.

Proposition 1.2 (multiplication principle). If one task can be performed in m ways, and following this, a second task can be performed in n ways (regardless of which way the first task was performed), then the number of ways of performing the 2 tasks in succession is mn.

This can be applied to 2 or more tasks performed independently in succession. In general, if the k^{th} task can be performed in m_k ways, where $1 \le k \le r$ then the number of ways of performing the r tasks in succession is

$$m_1 m_2 \ldots m_n.$$

Example 1.1 (ST2131 AY24/25 Sem 1 Lecture 1). A 4-digit code is to be formed using the digits $0, 1, 2, \ldots, 9$.

(a) How many codes can be formed?

(b) If the digits may not be repeated, how many codes can be formed

Solution.

(a) $10^4 = 10000$

(b) $10 \times 9 \times 8 \times 7 = 5040$ □

Example 1.2 (ST2131 AY22/23 Sem 2 Tutorial 1). Consider a group of 20 people. If everyone shakes hands with everyone else, how many handshakes take place?

Solution. Label the 20 people as $P_1, P_2, \ldots, P_{20}$. Then, P_1 can shake the hands of $P_2, P_3, \ldots, P_{20}$, so there are 19 ways to do so. P_2 can shake the hands of $P_3, P_4, \ldots, P_{20}$, so there are 18 ways to do so. Repeat this till P_{19} who can only shake P_{20}'s hand. The required number of ways is

$$19 + 18 + \ldots + 1 = \frac{19(20)}{2}$$

so there are 190 ways to do so. □

We can formulate a similar question (compare with Example 1.2) with an identical line of reasoning. Given a regular n-sided polygon, how many ways are there to connect the vertices? This is closely tied to a mathematics branch known as Graph Theory. In general, if there are n people, there are a total of

$$\frac{n(n-1)}{2} = \binom{n}{2}$$

handshakes. In Graph Theory, there is a similar idea to this known as the handshaking lemma.

1.2. *Permutations and Combinations*

Definition 1.1 (permutation). A permutation is an ordered arrangement of objects.

When we are dealing with permutations, order matters, i.e. the 3-letter arrangement of ABC and ACB are considered.

Proposition 1.3. Given n distinct objects, the total number of ways of arranging all these n objects in a line is $n!$.

Proof. There are n ways to put the first object in the first slot, $n-1$ ways to put the second object in the second slot. Repeating this process up to the last slot, by the multiplication principle (Proposition 1.2), we have only 1 way to put the last object there. □

Proposition 1.4 (permutations involving identical objects). Given n objects of which n_1 are identical, n_2 are identical,..., n_r are identical, there are

$$\frac{n!}{n_1!n_2!\ldots n_r!}$$

different permutations of the n objects, where $n_1+n_2+\ldots+n_r=n$.

Example 1.3 (ST2131 AY24/25 Sem 1 Lecture 1). 6 boys and 4 girls compete in a running race (no tie).

(a) If the boys and the girls run together, how many different finishing orders are possible?

(b) If the boys and the girls run separately, how many different finishing orders are possible?

Solution.

(a) $10! = 3628800$

(b) $6! \times 4! = 17280$ □

Now, we will discuss circular permutations. If we have n people sitting in a circle, there are

$$\frac{n!}{n} = (n-1)!$$

ways to arrange them. A simple way to understand this is that a circle has no beginning and no end.

Proposition 1.5. If there are n distinct objects, of which we choose a group of r items, the number of groups, denoted by $\binom{n}{r}$, can be written as

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

Proposition 1.6. The following hold for any $n \in \mathbb{Z}_{\geq 0}$:

(i)

$$\binom{n}{0} = \binom{n}{n} = 1$$

(ii) Symmetry of binomial coefficients:

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

Proof. For **(i)**, by considering the formula for r combinations out of n objects, setting $r = 0$ and $r = n$ will yield this result.

As for **(ii)**, the algebraic proof is simple but not as meaningful as its combinatorial counterpart. As such, we provide a proof for the latter. There are two ways to select a group of r items from a group of n, which are namely picking the r items that you're going to include or picking the $n-r$ items that you are going to leave out. Either way, the number of ways of forming the collection using the first method must be equal to the number of ways of forming the collection using the second. □

Theorem 1.1 (Pascal's identity). For $n, k \in \mathbb{N}$,

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Proof. Consider picking one fixed object out of n objects. Then, we can choose k objects including that one in $\binom{n-1}{k-1}$ ways. As our final group of objects either contains the specified one or doesn't, we can choose the group in $\binom{n-1}{k-1} + \binom{n-1}{k}$ ways. However, we already know they can be picked in $\binom{n}{k}$ ways, so the result follows. □

Example 1.4 (recurrence relation induction). If $p_{m,n}$ satisfies the recurrence relation

$$p_{m,n} = \frac{1}{2}\left(p_{m,n-1} + p_{m-1,n}\right),$$

prove that

$$p_{m,n} = \frac{1}{2^{m+n-1}} \sum_{k=m}^{m+n-1} \binom{m+n-1}{k}.$$

This recurrence relation appears in a classic problem in Probability Theory known as the problem of points. It asks how to fairly divide the stakes of a game of chance that is interrupted before its conclusion, given that each player has a certain probability of winning if the game were to continue. The problem was famously debated in 1654 between two mathematical giants, Blaise Pascal and Pierre de Fermat, through a series of letters.

Solution. Note the initial conditions $p_{m,0} = 0$ and $p_{0,n} = 1$. Set $q = m+n-1$. Let $P(q)$ be the proposition that

$$p_{m,n} = \frac{1}{2^q} \sum_{k=m}^{q} \binom{q}{k}$$

for $q \in \mathbb{Z}_{\geq 0}$. The case where $q = -1$, which is obtained when $m = n = 0$, is neglected. The base case $q = 0$ is true. To see why, we have $m+n = 1$ so we have either $(m,n) = (1,0)$ or $(m,n) = (0,1)$. Then, use the formula given in the proposition. Assume $P(r)$ to be true for some $r \in \mathbb{Z}_{\geq 0}$. Then, we wish to prove that $P(r+1)$ is true.

We have $r+1=m+n$ so we consider the cases $p_{m+1,n}$ and $p_{m,n+1}$ while assuming the validity of $p_{m,n}$. We have

$$\begin{aligned}
p_{m+1,n} &= \frac{1}{2}\left(p_{m+1,n-1}+p_{m,n}\right)\\
&= \frac{1}{2^{m+n}}\sum_{k=m+1}^{m+n-1}\binom{m+n-1}{k}+\frac{1}{2^{m+n}}\sum_{k=m}^{m+n-1}\binom{m+n-1}{k}\\
&= \frac{1}{2^{m+n}}\left[\sum_{k=m+1}^{m+n-1}\binom{m+n-1}{k}+\sum_{k=m}^{m+n-1}\binom{m+n-1}{k}\right]
\end{aligned}$$

It now suffices to prove

$$\sum_{k=m+1}^{m+n-1}\binom{m+n-1}{k}+\sum_{k=m}^{m+n-1}\binom{m+n-1}{k}=\sum_{k=m+1}^{m+n}\binom{m+n}{k}.$$

This is trivial by Pascal's identity (Theorem 1.1). The reader can prove the case for $p_{m,n+1}$ similarly. Hence, the result follows by induction. □

The recurrence relation in Example 1.4 is rather interesting. It is in fact directly related to the following probability problem: in a sequence of independent tosses of a fair coin, what is the probability that the first to occur is m heads before n tails? *This is just for you to ponder over* — we will discuss such problems in due course.

Example 1.5 (ST2131 AY24/25 Sem 2 Tutorial 1)**.** From a group of n people, suppose that we want to choose a committee of k, where $k \le n$, one of whom is designated as chairperson.

(i) By focusing first on the choice of the committee and then on the choice of the chair, argue that there are $\binom{n}{k}k$ possible choices.

(ii) By focusing first on the choice of the non-chair committee members and then the choice of the chair, argue that there are $\binom{n}{k-1}(n-k+1)$ possible choices.

(iii) By focusing first on the choice of the chair and then the choice of the committee members, argue that there are $n\binom{n-1}{k-1}$ possible choices.

There was originally a **(iv)** to this problem which asked the reader to deduce that

$$\binom{n}{k}k=\binom{n}{k-1}(n-k+1)=n\binom{n-1}{k-1}$$

which follows from the first three parts **(i)**, **(ii)**, and **(iii)** — it is just the same event viewed from three different points-of-view.

Solution.

(i) There are $\binom{n}{k}$ ways to form the committee of k, and $\binom{k}{1}=k$ ways to assign a chairperson. Then, apply the multiplication principle.

(ii) We form the non-chair committee members, so there are $\binom{n}{k-1}$ ways. Then, we choose one chairperson from the remaining $n-(k-1)$ people, for which there are $\binom{n-k+1}{1}=n-k+1$ ways. Lastly, apply the multiplication principle.

(iii) We first choose the chairperson, for which there are $\binom{n}{1}=n$ ways. Then, we choose the $k-1$ committee members from $n-1$ persons, for which there are $\binom{n-1}{k-1}$ ways. Lastly, apply the multiplication principle. □

We then introduce the binomial theorem[†].

[†]Pascal's triangle is a triangular array of the binomial coefficients that arises in Probability theory and Combinatorics. There are interesting patterns which arise due to the features of the triangle such as the Pascal's identity (Theorem 1.1) and the binomial coefficients (Proposition 1.6) aforementioned, and others including the triangular numbers and Fibonacci numbers. You can find a document containing some fascinating patterns and the link to it is here.

Theorem 1.2 (binomial theorem). Let $n \in \mathbb{Z}_{\geq 0}$. Then,

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

The term $\binom{n}{k}$ is referred to as the binomial coefficient.

Corollary 1.1 (sum of binomial coefficients).

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

Proof. Set $x = y = 1$ in the binomial theorem formula (Theorem 1.2). Combinatorially, we can also view this result as follows. If a set has n elements, then the number of subsets, including the null set and itself, is 2^n. This is because every element can be chosen or not chosen during the selection process. Since there are n elements, the result follows. □

Corollary 1.2 (alternating sum of binomial coefficients).

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

Proof. Set $x = 1$ and $y = -1$ in the binomial theorem formula (Theorem 1.2). □

Example 1.6 (ST2131 AY22/23 Sem 2 Tutorial 2). Consider the following combinatorial identity:

$$\sum_{k=1}^{n} k \binom{n}{k} = n2^{n-1}$$

Present a combinatorial argument for this identity by considering a set of n people, in two ways, the number of possible selections of a committee of any size and a chairperson for the committee.

Solution. We focus on the right side of the equation first. Say we wish to form a committee of size 3 (start with a small number other than 1 or 2 to illustrate) and one of them is the chairperson. There are $\binom{n}{3}$ ways to form the committee, and thereafter, $\binom{3}{1}$ ways to choose one of the 3 persons to be the chairperson.

Hence, it is easy to see that the number of ways to form a committee of size k, where $1 \leq k \leq n$, such that 1 of the persons in the committee of k is the chairperson, is

$$\binom{n}{k}\binom{k}{1} = k\binom{n}{k}.$$

We take the sum as k runs from 1 to n, which yields the left side of the equation.

On the right side of the equation, we select the chairperson first. There are $\binom{n}{1} = n$ ways to do so. We then choose any subset of the remaining $n-1$ people in 2^{n-1} ways.

As such, we have established a bijection, so the two quantities must be equal. □

Theorem 1.3 (hockey-stick identity). For any $n, r \in \mathbb{N}$ where $r \le n$,

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

Proof. This can be proven via induction or repeatedly applying Pascal's identity (Theorem 1.1). □

Theorem 1.4 (Vandermonde's identity). Any combination of r objects from a group of $m+n$ objects must have some $0 \le k \le r$ objects from group m and the remaining from group n. That is,

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

Proof. We provide a combinatorial proof of this result[†]. Consider a group of m men and n women. Suppose we wish to form a group of size r. Then, we can select k men out of the r, and consequently, $r-k$ women will be chosen. We can vary k from 0 to r inclusive, and by combining this fact with the multiplication principle, we obtain the LHS. The RHS can be obtained by simply considering the fact that we have $m+n$ men and women and we wish to form a group of size r, there are $\binom{m+n}{r}$ ways to do so. □

Corollary 1.3.

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

Proof. Set $m = n = p$ in Vandermonde's identity (Theorem 1.4). □

1.3. *Multinomial Coefficients*

Theorem 1.5 (multinomial theorem). For any $r \in \mathbb{Z}^+$ and $n \in \mathbb{Z}_{\ge 0}$, the multinomial theorem describes how a sum with m terms expands when raised to an arbitrary power n, i.e.

$$(x_1 + x_2 + \cdots + x_r)^n = \sum_{k_1+k_2+\cdots+k_r=n} \binom{n}{k_1, k_2, \ldots, k_r} \prod_{t=1}^{r} x_t^{k_t},$$

where $\binom{n}{k_1,k_2,\ldots,k_r}$ is a multinomial coefficient.

Note that the binomial theorem (Theorem 1.2) is a special case of the multinomial theorem — the former can be obtained by setting $n = 2$.

Example 1.7 (ST2131 AY24/25 Sem 1 Lecture 2). A group of 9 gamers are playing computer games.

(a) The first game consists of three different tasks presented at the same time. The gamers divide themselves into three groups of 3 to work on the problems simultaneously. How many divisions are possible?

(b) The second game requires three teams to play simultaneously, each team against the other two. The gamers divide themselves into three groups of 3 to play this game. How many divisions are possible?

Solution.

(a) Distribute the 9 people into 3 ordered groups, which yields $\binom{9}{3,3,3}$. This is precisely the definition of the multinomial coefficient.

[†]This question also appeared in ST2131 AY24/25 Sem 2 Tutorial 1

(b) Now that the tasks are not involved, the order does not matter, i.e. the groups are the *same*. The possible number of divisions is

$$\frac{1}{3!}\binom{9}{3,3,3}.$$

□

Example 1.8 (ST2131 AY24/25 Sem 1 Lecture 2). A standard deck of 52 cards is dealt out randomly to 4 players, each getting 13 cards. The picture cards are the J, Q, K of each suit. What is the probability that each player receives exactly three picture cards?

Solution.

$$\frac{\binom{12}{3,3,3,3}\binom{40}{10,10,10,10}}{\binom{52}{13,13,13,13}}$$

□

Example 1.9 (ST2131 AY24/25 Sem 1 Lecture 3). In an office of workers, there are 4 men and 4 women.

(a) If the 8 workers are randomly divided into 4 pairs, what is the probability that exactly 2 pairs are of mixed gender?

(b) If the 8 workers are randomly divided into 2 teams of four, what is the probability that in every team the four workers are of the same gender?

Solution.

(a)

$$\frac{\binom{4}{2}^2\times 2!\times\binom{2}{2}^2}{\binom{8}{2,2,2,2}/4!}$$

(b) In the probability space, there are a total of $\binom{8}{4,4}/2!$ outcomes. For the favourable outcomes, it is a simple calculation of $\binom{4}{4}\binom{4}{4}$. The desired probability is

$$\frac{\binom{4}{4}^2}{\binom{8}{4,4}/2!}.$$

□

1.4. *Distribution Problems and an Application to Linear Diophantine Equations*

Proposition 1.7 (distribution of identical objects into distinct boxes). We consider two cases.

(i) **Case 1:** To distribute r identical objects into n distinct boxes, where $r,n\in\mathbb{N}$, the number of ways is

$$\binom{r+n-1}{n-1}.$$

(ii) **Case 2:** To distribute r identical objects into n distinct boxes, where $r,n\in\mathbb{N}$, such that no box is empty, the number of ways is

$$\binom{r-1}{n-1}.$$

Proof. We only prove **(ii)**. We distribute 1 object into each of the n boxes. In total, we distribute n objects and have $r-n$ objects left. Now, the problem translates to distributing $r-n$ identical objects into n distinct boxes without restrictions, which is simply

$$\binom{r-n+n-1}{n-1} = \binom{r-1}{n-1}.$$

□

Example 1.10. Consider a problem in which we are attempting to find the number of distributions of 8 identical objects among 5 distinct bins, and bins cannot be left empty. How many ways are there to do this?

Solution. Modelling the problem as stars and bars, it would start off by looking like as follows:

$$\star \;\Big|\; \star \;\Big|\; \star\;\star \;\Big|\; \star\;\star \;\Big|\; \star\;\star$$

The objects are represented by the stars and the gaps between the bars are represented by the bins. In other words, we regard the bars as a partition. As such, the required answer is $12!/(8!5!) = 495$. Note that $12!$ is simply $(8+5-1)!$. □

Proposition 1.8 (distribution of distinct objects into distinct boxes). We consider two cases.

- **(i)** **Case 1:** To distribute r distinct objects into n distinct boxes, such that each box can hold at most 1 object, where $r,n \in \mathbb{N}$ and $r \leq n$, the number of ways is

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

- **(ii)** **Case 2:** To distribute r distinct objects into n distinct objects, where $r,n \in \mathbb{N}$ such that each box can hold any number of objects, the number of ways is n^r.
- **(iii)** **Case 3:** To distribute r distinct objects into n distinct objects, where $r,n \in \mathbb{N}$ and $r \geq n$ such that no box is empty, the number of ways is

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

$S(r,n)$ is known as the Stirling numbers of the second kind (Definition 1.3)

Proof. We first prove **(i)**. The first object goes into the first box. There are n ways to do this. The second object goes into the second box and there are $n-1$ ways to do so. Repeating to the r^{th} object, there are $n-r+1$ ways for it to go into the n^{th} box. By

the multiplication principle the required number of ways is $n(n-1)(n-2)\ldots(n-r+1)$,

which yields the desired expression.

As for **(ii)**, the first object can go into the first box and there are n ways to do it. The same can be said for the remaining objects. We will not discuss the proof of **(iii)** but anyway, it would rely on the principle of inclusion and exclusion (Proposition 2.1). □

Before we discuss the Stirling numbers of the second kind, we shall start with the Stirling numbers of the first kind! We denote the latter by $s(r,n)$ and the former by $S(r,n)$. These types of numbers are named after the Scottish mathematician James Stirling. He is known for Stirling's approximation

$n! \sim \sqrt{2n\pi}\left(\frac{n}{e}\right)^n$ which involves an asymptotic formula for $n!$ for large values of n.

Definition 1.2 (Stirling numbers of the first kind). Given $r, n \in \mathbb{Z}$ such that $0 \leq n \leq r$, let $s(r,n)$ be the number of ways to arrange r distinct objects around n indistinguishable circles such that each circle has at least one object.

Proposition 1.9. Here are some obvious results.

(i) For $r \leq 1$, $s(r,0) = 0$

(ii) For $r \geq 0$, $s(r,r) = 1$

(iii) For $r \geq 2$, $s(r,1) = (r-1)!$

(iv) For $r \geq 2$, $s(r,r-1) = \binom{r}{2}$

Proposition 1.10. Here is a useful recurrence relation for $s(r,n)$. That is,

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

Proof. Fix an object a_1. Then, we have two cases, which are namely **(i)** a_1 is the only object in a circle and **(ii)** a_1 is mixed with other objects.

For **(i)**, we shift our focus to the remaining $r-1$ objects. We distribute these objects around the remaining $n-1$ objects, and there are $s(r-1,n-1)$ ways to do so by definition. For **(ii)**, we have $r-1$ objects left to distribute around n tables. a_1 can be placed in either one of the $r-1$ distinct spaces to the immediate right of the corresponding $r-1$ distinct objects.

As the two cases are mutually exclusive, the result follows by the addition principle. □

Definition 1.3 (Stirling numbers of the second kind). Given $r, n \in \mathbb{Z}_{\geq 0}$ where $0 \leq n \leq r$, the Sirtling numbers of the second kind, $S(r,n)$, is defined as the number of ways of distributing r objects into n identical boxes such that no box is empty.

Proposition 1.11. Here are some obvious results.

(i) $S(r,1) = S(r,r) = 1$

(ii) For $1 \leq r < k$, $S(r,k) = 0$

(iii) For $r, k \geq 1$, $S(r,0) = S(0,k) = 0$

(iv) For $r \geq 1$, $S(r,2) = 2^{r-1} - 1$

(v) For $r \geq 1$, $S(r,r-1) = \binom{r}{2}$

Proof. We will only prove **(iv)**. The complement of the case where no box is empty is that one box is empty. The number of ways to distribute r distinct objects into 1 box is 1. Since each object can go into either box, there are 2^r ways to distribute, but we also have to consider that the boxes are identical, so we have to divide by 2. Thus, there are 2^{r-1} ways to distribute r distinct objects into 2 identical boxes without restrictions. The result follows. □

Similar to the Stirling numbers of the first kind, we have a similar recurrence relation for the Stirling numbers of the second kind, which is slightly easier to derive since we are not considering circular permutations.

Proposition 1.12. Here is a useful recurrence relation for $S(r,n)$. That is,

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

Proof. Fix an object a_1. Then, we have two cases, which are namely **(i)** a_1 is the only object in a box and **(ii)** a_1 is mixed with other objects.

For (i), we shift our focus to the remaining $r-1$ objects. We distribute the remaining $r-1$ objects into the $n-1$ boxes, and there are $S(r-1,n-1)$ ways to do so by definition. For **(ii)**, we have $r-1$ objects left to distribute into n identical boxes. a_1 can be distributed into either of the boxes, and so there are a total of $nS(r-1,n)$ ways to do so.

As the two cases are mutually exclusive, the result follows by the addition principle. □

Proposition 1.13 (distribution of distinct objects into identical boxes)**.** Suppose $r,n \in \mathbb{Z}_{\geq 0}$, where $0 \leq n \leq r$. We have two cases.

(i) **Case 1:** $S(r,n)$ is defined as the number of ways of distributing r objects into n identical boxes such that no box is empty.

(ii) **Case 2:** If we have r objects and n boxes and each box can hold any number of objects, the number of ways to distribute the objects is

$$S(r,1) + \ldots + S(r,n).$$

Before we discuss the distribution of identical objects into identical boxes, we first define a partition of an integer.

Definition 1.4 (partition)**.** We define a partition of a positive integer r into n parts to be a set of n positive integers whose sum is r. The ordering of the integers in the collection is immaterial since the integers are regarded as identical objects.

Let the number of different partitions of n be denoted by $P(r,n)$.

Proposition 1.14 (distribution of identical objects into identical boxes)**.** Given $r,n \in \mathbb{Z}_{\geq 0}$, where $0 \leq n \leq r$, $P(r,n)$ is the number of ways of r identical objects into n identical boxes.

Proposition 1.15. Here is a useful recurrence relation for $P(r,n)$, which is

$$P(r,n) = P(r-1,n-1) + P(r-n,n),$$

where $r,n \in \mathbb{N}$, $1 < n \leq r$ and $r \geq 2n$.

Proof. We consider two cases, namely **(i)** at least one box has exactly one object and **(ii)** all the boxes have more than one object.

For **(i)**, we place one object in one box. Then we distribute the remaining $r-1$ objects into the remaining $n-1$ boxes such that no boxes are empty. The number of ways this can be done is $P(r-1,n-1)$. For **(ii)**, we place one object into each of the n boxes. Then we distribute the remaining $r-n$ objects into the n boxes such that each box has at least two objects. The number of ways this can be done is $P(r-n,n)$. By the addition principle, the result follows. □

A Diophantine equation is a polynomial equation, usually involving two or more unknowns, such that the only solutions of interest are the integer ones. A linear Diophantine equation equates to a constant the sum of two or more monomials, each of degree one. That is, for constants a_i and b and variables x_i, where $1 \leq i \leq n$,

$$a_i x_i + a_2 x_2 + \ldots + a_n x_n = b.$$

There are $\binom{n-1}{r-1}$ distinct positive integer-valued vectors $(x_1, x_2, \ldots, x_r)$ that satisfy the equation

$$x_1 + x_2 + \ldots + x_r = n,$$

where $x_i > 0$ for $1 \leq i \leq r$. Note that this is the equivalent of the distribution of r identical objects into n distinct boxes, where $r, n \in \mathbb{N}$, such that no box is empty.

Proposition 1.16. There are $\binom{r+n-1}{r-1}$ distinct non-negative integer-valued vectors $(x_1, x_2, \ldots, x_r)$ that satisfy the equation

$$x_1 + x_2 + \ldots + x_n = r \quad \text{where } x_i \geq 0 \text{ for } 1 \leq i \leq r.$$

Proof. Let $y_i = x_i + 1$, then each of the y_i's is positive, implying that the number of non-negative solutions to

$$x_1 + x_2 + \ldots + x_n = r$$

is the same as the number of positive solutions to

$$(y_1 - 1) + (y_2 - 1) + \ldots + (y_r - 1) = n,$$

or equivalently,

$$y_1 + y_2 + \ldots + y_r = n + r,$$

which is $\binom{r+n-1}{r-1}$. □

Here is a relatively easy problem.

Example 1.11 (SMO Open 2022 Question 18). Find the number of integer solutions to the equation $x_1 + x_2 - x_3 = 20$ with $x_1 \geq x_2 \geq x_3 \geq 0$.

Solution. We proceed with some casework. First, set $x_3 = 0$. Then, we have $x_1 + x_2 = 20$, where $x_1 \geq x_2 \geq 0$. There are 10 solutions for this case, namely

$$(x_1, x_2) = (20, 0), (19, 1), \ldots, (10, 10).$$

For the second case, set $x_3 = 1$. Then, we have $x_1 + x_2 = 21$. There are 10 solutions for this case, namely

$$(x_1, x_2) = (20, 1), (19, 2), \ldots, (11, 10).$$

We repeat this process until $x_3 = 20$, which implies that $x_1 + x_2 = 40$, where $x_1 \geq x_2 \geq 20$. There is only one solution for this. If one considers the cases in between these, you can spot a pattern, which implies that the total number of solutions is $11 + 2 \cdot 10 + 2 \cdot 9 + \ldots + 2 \cdot 1 = 121$. □

Example 1.12 (SMO Open 2007 Question 6)**.** Find the number of non-negative solutions to the following inequality:

$$x+y+z+u \leq 20$$

Solution. Using the substitution $v = 20-(x+y+z+u)$, then $v \geq 0$ if and only if $x+y+z+u \leq 20$. The required answer is the number of non-negative integer solutions to the equation

$$x+y+z+u+v = 20,$$

which is $\binom{24}{4} = 10626$. □

2. Axioms of Probability

2.1. *Axioms*

The basic terminologies of Probability Theory, including experiment, outcomes, sample space, events, should be covered in secondary school so we shall not discuss them here. We will define the probability of an event and show how it is computed using a variety of examples.

The Kolmogorov axioms are named after Russian mathematician Andrey Kolmogorov. There are numerous Russian mathematicians who contributed to the field of Probability and Statistics. Some include Andrey Markov, who is known for Markov's Inequality and Markov chains, Nikolai Smirnov, for which the Kolmogorov-Smirnov test, a non-parametric test (may be covered in ST2132), is named after him and Kolmogorov, as well as Pafnuty Chebyshev. Chebyshev's inequality and the Chebyshev polynomials of the first kind and the second kind are named after him. Not to mention, he also contributed to the much celebrated Prime number theorem.

Axiom 2.1 (Kolmogorov axioms).

(i) **Axiom 1:** For any event A, $0 \le P(A) \le 1$.

(ii) **Axiom 2:** Let S be the sample space. Then, $P(S) = 1$.

(iii) **Axiom 3:** For any sequence of mutually exclusive events $A_1, A_2, \ldots$ (i.e. $A_iA_j =$ whenever $i \neq j$),

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

Example 2.1 (ST2131 AY24/25 Sem 1 Lecture 1). A dice is biased, with the even numbers being equally likely to appear but each odd number is twice as likely to appear as any of the even numbers.

(a) Find the probability of obtaining a 3.

(b) Find the probability of obtaining a 1 or 6.

Solution.

(a) Let P_o and P_e denote the probability of an odd and an even value showing up respectively. Assuming that the die ranges from 1-6, by the Kolomogrov axioms (Axiom 2.1),

$$P_o = 2P_e \quad \text{and} 3P_o + 3P_e = 1.$$

Solving the above yields $P_o = 2/9$.

(b) We have $P_o + P_e = 1/9 + 2/9 = 1/3$. □

Example 2.2 (ST2131 AY24/25 Sem 1 Lecture 1).

(a) Three fair coins are tossed. What is the probability that exactly two heads appear?

(b) Four fair coins are tossed. What is the probability that at least two heads appear?

Solution.

(a) There are a total of $2^3 = 8$ possible outcomes. Choose two of the three coins to be fixed as heads, then the last coin must be a tail. Since any two of the three coins could be chosen to be heads, there are $\binom{3}{2}$ possible cases. The required probability is $\binom{3}{2}/2^3 = 3/8$.

(b) The total number of outcomes is 2^4. The desired outcomes are $2, 3$ or 4 heads. Hence,

$$P = \frac{\binom{4}{2} + \binom{4}{3} + \binom{4}{4}}{2^4} = \frac{11}{16}.$$

□

2.2. *Probability Properties*

Using Kolmogorov's axioms, we can derive a few useful properties such as de Morgan's laws and the probability of the complement of an event, where the complement is usually denoted by A' or A^c, where $P(A)+P(A')=1$. In particular, we shall discuss the principle of inclusion and exclusion.

Proposition 2.1 (principle of exclusion and exclusion)**.** If we have n events $A_1, A_2, \ldots, A_n$,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1\le i<j\le n} P(A_i \cap A_j) + \sum_{1\le i<j<k\le n} P(A_i \cap A_j \cap A_k) + \ldots + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} A_i\right).$$

One would generally be more familiar with the principle of inclusion and exclusion for two events.

Example 2.3. Say we have two events A and B. Then,

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

This can be illustrated using a Venn diagram.

Example 2.4. If we have three events A, B and C, then

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

Example 2.5 (ST2131 AY24/25 Sem 1 Lecture 1)**.** In a large sports club,

- 40% of members play badminton
- 35% of members play squash
- 10% of members play both

Find the probability that a randomly selected member plays neither of the two sports mentioned above.

Solution. By the principle of inclusion and exclusion, the answer is $1-(0.4+0.35-0.1)=0.35$. □

Example 2.6 (ST2131 AY24/25 Sem 1 Lecture 1)**.** Suppose you assess that there is more than 85% chance that the weather will be nice tomorrow, and there is more than 65% chance that the weather will be nice the day after tomorrow. Is it valid to infer that there is more than a fair chance that the weather will be nice on both days?

Solution. Valid. Let E and F denote the first and second events written above respectively. Then,

$$P(E\cup F) = P(E)+P(F)-P(E\cap F) \quad \text{which implies} \quad P(E\cap F) = P(E)+P(F)-P(E\cup F).$$

Since $P(E)>0.85$, $P(F)>0.65$, and $P(E\cup F)\le 1$, it follows that $P(E\cap F)>0.5$. □

Example 2.7 (ST2131 AY24/25 Sem 1 Lecture 1)**.** Similar to Example 2.6, suppose you assess that there is more than 80% chance that the weather will be nice tomorrow and there is more than 80% chance that the weather will be nice the day after tomorrow. Is it valid to infer that there is more than 80% chance that the weather will be nice on both days?

Solution. Invalid. Again, let E and F denote the first and second events written above respectively. Then,

$$P(E\cap F) > 0.8+0.8-1 = 0.6.$$

However, we cannot conclude that there is more than an 80% chance that the weather will be nice on both days. □

Example 2.8 (H3 Mathematics 2020)**.** Let

$$X = \{1,2,\ldots,m\} \text{ and } Y = \{1,2,\ldots,n\} \quad \text{be sets of positive integers}$$

and f be a function mapping from X to Y. f is called one-to-one if no two elements of X map to the same element of Y, and f is called onto if each element of Y is the image of an element of X. For $m \geq n$, we wish to find an expression for the number of functions mapping X to Y which are onto.

Solution. Let A_i be the event denoting the element $n_i \in Y$ which does not get mapped from any element in X, where $1 \leq i \leq n$. We wish to find

$$\left|A_1' \cap A_2' \cap \ldots \cap A_n'\right|,$$

which is equivalently, by de Morgan's law,

$$n(S) - \left|\bigcup_{i=1}^{n} A_i\right|.$$

Note that

$$\begin{aligned}
\sum_{i=1}^{n} |A_i| &= \binom{m}{1}(m-1)^n \\
\sum_{1 \leq i < j \leq n} \left|A_i \cap A_j\right| &= \binom{m}{2}(m-2)^n \\
\sum_{1 \leq i < j < k \leq n} \left|A_i \cap A_j \cap A_k\right| &= \binom{m}{3}(m-3)^n
\end{aligned}$$

and so on. It is clear that $n(S) = m^n$. Using the principle of inclusion and exclusion,

$$\begin{aligned}
\left|\bigcup_{i=1}^{n} A_i\right| &= \sum_{i=1}^{n} |A_i| - \sum_{1 \leq i < j \leq n} \left|A_i \cap A_j\right| + \sum_{1 \leq i < j < k \leq n} \left|A_i \cap A_j \cap A_k\right| + \ldots + \binom{m}{m}(m-m)^n \\
\left|\bigcup_{i=1}^{n} A_i\right| &= \sum_{r=0}^{m} (-1)^{r+1} \binom{m}{r}(m-r)^n \\
\left|A_1' \cap A_2' \cap \ldots \cap A_n'\right| &= \sum_{r=0}^{m} (-1)^{r} \binom{m}{r}(m-r)^n
\end{aligned}$$

which is the required expression. □

Definition 2.1 (derangement)**.** A derangement is a permutation of the elements of a set, such that no element appears in its original position. If a set has n elements, then its derangement is denoted by D_n or $!n$.

Example 2.9 (hat-check problem)**.** A group of n men enter a restaurant and check in their hats at the reception. The hat-checker is absent-minded, and upon leaving, he redistributes the hats to the men randomly. Suppose D_n is the number of ways such that no men get his own hat. For $n \geq 3$, prove that D_n satisfies the following recurrence relation:

$$D_n = (n-1)(D_{n-1} + D_{n-2})$$

with initial conditions $D_1 = 0$ and $D_2 = 1$.

Solution. Suppose the first person receives the i^{th} person's hat, where $i \neq 1$. There are $n-1$ ways to do so. We consider two cases, namely **(i)** the i^{th} person received hat 1 and **(ii)** the i^{th} person received a hat that is not hat 1.

For **(i)**, ignoring the first and i^{th} person, there are D_{n-2} ways to arrange the $n-2$ hats among the $n-2$ people such that no one received his own hat. For **(ii)**, treating the i^{th} person as the first person, this is equivalent to arranging the $n-1$ hats among $n-1$ people such that no one received his own hat. There are D_{n-1} ways to do so. The result follows. □

By repeatedly applying the recurrence relation or simply using induction, we can establish that

$$D_n = nD_{n-1} + (-1)^n$$

for $n \geq 2$. We can find a formula for D_n in terms of a sum. This involves considering a new expression, namely $D_n = n!P_n$. Thus,

$$\begin{aligned} n!P_n &= n!P_{n-1} + (-1)^n \\ P_n &= \sum_{i=2}^{n} \frac{(-1)^i}{i!} \end{aligned}$$

by the method of difference. In conclusion,

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

We can also prove this result by the principle of inclusion and exclusion.

For large values of n, P_n tends to e^{-1}. This can be proven by the Maclaurin Series of e^x, namely

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}.$$

□

We recall the example from the H3 Mathematics 2020 paper (Example 2.8).

Example 2.10 (derangement). Five couples are being seated at a long table. The five women are seated first, along one side of the table. The five men are then assigned seats along the other side, at random. What is the probability that none of the couples end up facing each other?

Solution. Let E_i denote the event where the i^{th} couple end up facing each other. Let A denote the event where none of the couples face each other, then $A = E_1' \cap \ldots \cap E_5'$. The complement of A is $A' = E_1 \cup \ldots \cup E_5$. By the principle of inclusion and exclusion,

$$P(A) = 1 - \left(\frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \frac{1}{5!} \right).$$

□

Example 2.11 (ST2131 AY24/25 Sem 1 Lecture 3). Four couples are seated randomly at a round table. What is the probability that at least one of the couples end up sitting next to each other?

Solution. There are a total of $(8-1)! = 7!$ possible outcomes. Define E_i to be event where the i^{th} couple sits next to each other. Let $A = E_1 \cup \ldots \cup E_4$. By the principle of inclusion and exclusion, we have

$$P(A) = \sum_{i=1} P(E_i) - \sum_{i<j} P(E_i \cap E_j) + \sum_{i<j<k} P(E_i \cap E_j \cap E_k) - P(E_1 \cap E_2 \cap E_3 \cap E_4).$$

In general, we have

$$P(E_{i_1} \cap \ldots \cap E_{i_n}) = \frac{(7-n)! \times 2^n}{(8-1)!}.$$

By the principle of inclusion and exclusion,

$$P(A') = 1 - \left(\sum_{i=1}^{4} (-1)^{i+1} \binom{4}{i} \frac{(7-i)! \times 2^i}{(8-1)!} \right) = \frac{31}{105}.$$

□

We can generalise Example 2.11. The number of ways c_n to seat n couples around a round table with no spouses next to each other is given by

$$c_n = \sum_{i=0}^{n} (-1)^i 2^i \binom{n}{i} (2n-i-1)!.$$

The first few values of c_n for $n = 1, \ldots, 8$ are

$0, 2, 32, 1488, 112\,512, 12\,771\,840$ which appears as sequence A129348 in the OEIS.

This sequence is known as the number of directed Hamiltonian circuits in the cocktail party graph (appears in the handshake lemma) of order n.

Example 2.12 (H3 Mathematics 2020). Let

$X = \{1, 2, \ldots, m\}$ and $Y = \{1, 2, \ldots, n\}$ be sets of positive integers

and f be a function mapping from X to Y. f is called one-to-one if no two elements of X map to the same element of Y, and f is called onto if each element of Y is the image of an element of X. Now, for $m = n = 5$, find the number of one-to-one functions mapping from X to Y which map no element to itself.

Solution. We shall use the principle of inclusion and exclusion to assist us. Let A_i be the set of permutations in which the i^{th} element goes into the right position, where $1 \le i \le 5$. Note that $|A_i| = 4!$, $|A_i \cap A_j| = 3!$ and so on. Hence, using the principle of inclusion and exclusion, the number of derangements, D_5, is

$$\begin{aligned}
D_5 &= 5! - \left| \bigcup_{i=1}^{5} A_i \right| \\
&= 5! - \sum_{i=1}^{5} |A_i| + \sum_{1 \le i < j \le 5} |A_i \cap A_j| + \ldots + (-1)^5 \left| \bigcap_{i=1}^{5} A_i \right| \\
&= 5! - \binom{5}{1} 4! + \binom{5}{2} 3! - \binom{5}{3} 2! + \binom{5}{4} 1! - \binom{5}{5} 0! \\
&= 44
\end{aligned}$$

Alternatively, using the derangement formula will yield the same result. □

Example 2.13 (birthday problem/paradox). The birthday problem asks for the probability that, in a set of n randomly chosen people, at least two will share a birthday. The birthday paradox is that, counter-intuitively, the probability of a shared birthday exceeds 50% in a group of only 23 people. The probability that at least two of the n persons share the same birthday, denoted by $p(n)$, can be expressed as

$$p(n) = 1 - \frac{365!}{365^n (365-n)!} = 1 - \frac{n!}{365^n} \binom{365}{n}.$$

Note that $n \leq 365$; if $n \geq 366$, we obtain a contradiction by the pigeonhole principle. As $p(22) = 0.47569$ and $p(23) = 0.50729$, it asserts that the statement is true. We provide a proof for this.

Solution. A person can have his/her birthday on any of the 365 days. There are a total of 365^n outcomes. Let A denote the event that there at least two people among the n people sharing the same birthday. Then, A' is the event that none of them shares the same birthday. Without a loss of generality, treating each person as an object and each date as a box, the *first person can go into the first day* in 365 ways. The *second person can go into the second day* in 364 ways and so on, till the n^{th} *person goes into the* n^{th} *day* in $366 - n$ ways. Hence,

$$P(A') = \frac{365 \cdot 364 \cdot 363 \cdot \ldots \cdot (366-n)}{365^n}.$$

Since $1 - P(A') = P(A) = p(n)$, then

$$\begin{aligned} p(n) &= 1 - \frac{365 \cdot 364 \cdot 363 \cdot \ldots \cdot (366-n)}{365^n} \\ &= 1 - \frac{365!}{365^n (365-n)!} \\ &= 1 - \frac{n!}{365^n} \binom{365}{n} \end{aligned}$$

□

Example 2.14 (ST2131 AY24/25 Sem 1 Lecture 2; modified birthday paradox)**.** Assume that the students in a large class are equally likely to have their birthdays fall on any of the 7 days of the week.

What is the smallest integer n such that, in a group of n students randomly selected from this class, there is more than 60% chance for at least two of them to have their birthdays fall on the same day of the week?

Solution. We have a trivial upper bound of $n \leq 8$ because there are only 7 days in a week. We find $P(A_n')$, which is the probability that each student in a group of n students has a different day in a week for their birthday.

For $n = 1$, it is trivial. Also,

$$P(A_2') = \frac{7 \times 6}{7^2} \quad \text{and} \quad P(A_3') = \frac{7 \times 6 \times 5}{7^3}.$$

One can generalise the above — for any $n \leq 8$,

$$P(A_n') = \frac{7 \times \ldots \times (7-n+1)}{7^n}.$$

We accept this without proof for now. Then by computation, we find that $n \geq 4$ satisfies the inequality $1 - P(A_n') > 0.6$. □

Example 2.15 (ST2131 AY24/25 Sem 1 Lecture 2; modified birthday paradox)**.** Assume that the students in a large class are equally likely to have their birthdays fall on any of the 12 months of the year.

What is the smallest integer n such that, in a group of n students randomly selected from this class, there is more than 60% chance for at least two of them to have their birthdays fall on the same month for their birthday?

Solution. Likewise (compare with Example 2.14), we should have an upper bound of $n \leq 13$. We use a similar strategy — consider the event where n students all have different months of their birthday. Then,

$$P(A_n') = \frac{12 \times \ldots \times (12-n+1)}{12^n} \quad \text{for any } n \leq 13.$$

Solving the inequality $1 - P(A_n') > 0.6$ gives us $n \geq 5$. □

2.3. *Probability as a Continuous Set Function*

Definition 2.2 (increasing and decreasing sequences)**.** A sequence of events E_n, $n \geq 1$, is an increasing sequence if

$$E_1 \subseteq E_2 \subseteq \ldots \subseteq E_n \subseteq E_{n+1} \subseteq \ldots$$

whereas it is a decreasing sequence if

$$E_1 \supseteq E_2 \supseteq \ldots \supseteq E_n \supseteq E_{n+1} \supseteq \ldots$$

If E_n, $n \geq 1$, is an increasing sequence of events, then we define the following new event:

$$\lim_{n\to\infty} E_n = \bigcup_{i=1}^{\infty} E_i$$

Similarly, if E_n, $n \geq 1$, is a decreasing sequence of events, then we define the following new event:

$$\lim_{n\to\infty} E_n = \bigcap_{i=1}^{\infty} E_i$$

Proposition 2.2. If E_n, where $n \geq 1$, is either an increasing or a decreasing sequence of events, then

$$\lim_{n\to\infty} P(E_n) = P\left(\lim_{n\to\infty} E_n\right).$$

3. Conditional Probability and Independence

3.1. *Conditional Probabilities*

Definition 3.1 (conditional probability). Let A and B be two events. The conditional probability of A given B is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \text{provided that } P(B) \neq 0.$$

$P(A \mid B)$ can also be read as the conditional probability of A occurring given that B has occurred. Since we know that A has occurred, we can now think of A as our new, or reduced sample space.

Proposition 3.1 (generalised multiplication rule). If $A_1, A_2, \ldots, A_n$ are events, then

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\ldots P\left(A_n \middle| \bigcap_{i=1}^{n-1} A_i\right).$$

Proof. Use the definition of conditional probability to the right side of the equation to get the following expression:

$$P(A_1) \cdot \frac{P(A_2 \cap A_1)}{P(A_1)} \cdot \frac{P(A_3 \cap A_2 \cap A_1)}{P(A_2 \cap A_1)} \cdot \ldots \cdot \frac{P\left(\bigcap_{i=1}^{n} A_i\right)}{P\left(\bigcap_{i=1}^{n-1} A_i\right)}$$

which is clear that it is equal to the left side. □

Proposition 3.2. Let A be an event such that $P(A) > 0$. Then, the following three statements hold:

(i) For any event B, $0 \leq P(B \mid A) \leq 1$.

(ii) $P(S \mid A) = 1$

(iii) Let $B_1, B_2, \ldots$ be a sequence of mutually exclusive events. Then,

$$P\left(\bigcup_{i=1}^{\infty} B_i \middle| A\right) = \sum_{i=1}^{\infty} P(B_i|A).$$

Proof. We first prove **(i)**. As $P(A \cap B) \geq 0$ and $P(A) > 0$, we prove the lower bound for $P(B \mid A)$. To prove $P(B \mid A) \leq 1$, note that $B \mid A \subseteq A$, implying that $P(A \cap B) \leq P(A)$, and the result follows.

For **(ii)**, this follows from the fact that

$$P(S \mid A) = \frac{P(A \cap S)}{P(A)}$$

and since $P(A \cap S) = P(A)$, the result follows.

Lastly, for **(iii)**,

$$P\left(\bigcup_{i=1}^{\infty} B_i \middle| A\right) = \frac{P\left(A \cap \bigcup_{i=1}^{\infty} B_i\right)}{P(A)} = \frac{P\left(A \cap \bigcup_{i=1}^{\infty} B_i\right)}{P(A)} = \frac{\sum_{i=1}^{\infty} P(B_i \cap A)}{P(A)} = \sum_{i=1}^{\infty} \frac{P(B_i \cap A)}{P(A)}$$

which simplifies to

$$\sum_{i=1}^{\infty} P(B_i \mid A).$$

□

3.2. *Bayes' theorem*

Theorem 3.1 (Bayes' theorem). If A and B are two different events with $P(B) \neq 0$, then

$$P(A|B)P(B) = P(B|A)P(A).$$

Proof. By the definition of conditional probability, $P(A \mid B)P(B) = P(A \cap B)$. Also, as $P(A \cap B) = P(B \mid A)P(A)$, the result follows. □

We introduce the notion of the partition of a sample space S. We say that $A_1, A_2, \ldots, A_n$ are partitions of S if they are mutually exclusive and collectively exhaustive.

The term *mutually exclusive* is studied in both secondary school and junior college. It simply means $A_i \cap A_j =$ for all $i \neq j$. In relation to probabilities, $P(A \cup B) = P(A) + P(B)$, i.e. $P(A \cap B) = 0$. The term *collectively exhaustive* means that $\bigcup_{i=1}^{n} A_i = S$.

We are now ready to state the law of total probability.

Proposition 3.3 (law of total probability). Suppose the events $A_1, A_2, \ldots, A_n$ are partitions of S. Assume further that $P(A_i) > 0$ for all $1 \leq i \leq n$. Let B be any event. Then,

$$P(B) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \ldots + P(B \mid A_n)P(A_n).$$

Example 3.1 (ST2131 AY24/25 Sem 1 Lecture 7). A student applying to a graduate program asks his professor for a letter of recommendation. He estimates that his chances of getting a strong, average, weak recommendation are 30%, 20%, and 10%, and lastly a 40% chance of not receiving a recommendation letter.

He also estimates that his chances of getting accepted by the graduate programme would be 90%, 40%, and 10% if the recommendation is strong, average, and weak respectively.

(a) Based on these estimates, what is his probability of getting accepted by the graduate program?

(b) If he gets accepted by the graduate programme, what is the probability that his letter of recommendation was a strong one?

Solution.

(a) We can split the event into disjoint events. Let S, A, W, N denote the event that he gets a strong, average, weak, and no recommendation letter respectively. Also, let $\star$ denote the event where he gets accepted. Then, by the law of total probability (Proposition 3.3),

$$\begin{aligned} P(\star) &= P(\star \cap S) + P(\star \cap A) + P(\star \cap W) \\ &= P(\star \mid S)P(S) + P(\star \mid A)P(A) + P(\star \mid W)P(W) \\ &= 0.9 \cdot 0.3 + 0.4 \cdot 0.2 + 0.1 \cdot 0.1 = 0.36 \end{aligned}$$

(b) We use the definition of conditional probability on the event $P(S \mid \star)$. We have

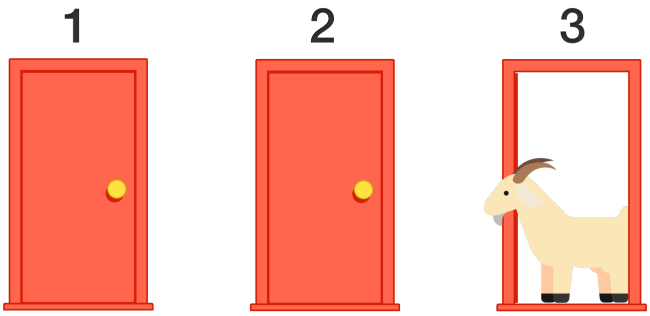
$$P(S \mid \star) = \frac{P(\star \cap S)}{P(\star)} = \frac{P(\star \mid S)P(S)}{P(\star)} = \frac{0.9 \cdot 0.3}{0.36} = 0.75.$$

□

Corollary 3.1. For $1 \leq i \leq n$,

$$P(A_i \mid B) = \frac{P(B \mid A_i) P(A_i)}{P(B \mid A_1) P(A_1) + \ldots + P(B \mid A_n) P(A_n)}.$$

Example 3.2 (Monty Hall problem). Suppose you are on a game show, and given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?



The answer is yes! Initially, the probability of winning a car is $1/3$. After the host opens Door 3, the probability of winning a car is surprisingly not $1/2$, but instead $2/3$! We can prove this result using a tree diagram or in a more elegant manner, Bayes' theorem.

Let A be the event that Door No. 1 has a car behind it and B be the event that the host has revealed a door with a goat behind it.

Then,

$$P(B \mid A) P(A) + P(B \mid A') P(A') = P(B).$$

To see why,

$$P(B \mid A) P(A) + P(B \mid A') P(A') = P(A \cap B) + P(B \cap A').$$

As A and B are independent events,

$$\begin{aligned} P(A \cap B) + P(B \cap A') &= P(A) P(B) + P(B) P(A') \\ &= P(B) \left[P(A) + P(A)' \right] \\ &= P(B) \end{aligned}$$

Now, it is clear that

$$P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B \mid A) P(A) + P(B \mid A') P(A')}.$$

A is the event that Door No. 1 has a car behind it. $B \mid A$ is the event that the host shows a door with nothing behind, given that there is a car behind Door No. 1. Note that $P(A) = 1/3$ and $P(B \mid A) = P(B \mid A') = 1$. Putting everything together, $P(A \mid B) = 1/3$. Hence, the probability that the car is behind Door No. 3 is $2/3$ and so you, the contestant, should make the switch.

3.3. *Independent Events*

Definition 3.2 (independent events)**.** Two events A and B are independent if

$$P(A\cap B)=P(A)P(B).$$

If $P(A\cap B)\neq P(A)P(B)$, then A and B are dependent.

In relation to conditional probability, suppose $P(B)>0$. Then, if A and B are independent,

$$P(A\mid B)=P(B).$$

In other words, A is independent of B if knowledge that B has occurred odes not change the probability that A occurs.

Proof. By the definition of conditional probability (Definition 3.1),

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

Since A and B are independent, $P(A\cap B)=P(A)P(B)$. The rest is simple algebraic manipulation. □

We shall now discuss pairwise independence and mutual independence.

Definition 3.3 (pairwise independence and mutual independence)**.** Given three events A, B and C, we say that A, B, and C are pairwise independent if

$$P(A\cap B)=P(A)P(B),\ P(A\cap C)=P(A)P(C)\text{ and }P(B\cap C)=P(B)P(C).$$

We say that A, B, and C are mutually independent if

$$P(A\cap B\cap C)=P(A)P(B)P(C).$$

If A, B, and C are pairwise independent, it does not necessarily imply that they are mutually independent. The converse, however, is true. That is, if A, B and C are mutually independent, then they are necessarily also pairwise independent. It follows that mutual independence is a stronger condition than pairwise independence.

The gambler's ruin problem (Example 3.3) states that a gambler playing a game with negative expected value will eventually go broke, regardless of their betting system.

Example 3.3 (gambler's ruin problem)**.** Consider a gambler's situation, where his starting fortune is \$$j$, in every game, the gambler bets \$1 and the gambler decides to play until he either loses it all (i.e. fortune is 0) or his fortune reaches \$$N$ and he quits. What is the probability to win?

Solution. We use the gambler's ruin equation to help us. However, we have to set up the equation first! Let A_j be the event that the gambler wins if he starts with a fortune of \$$j$. Then, we can let $x_j=P(A_j)$. For every game, suppose

$$P(\text{win})=p,\ P(\text{lose})=q\text{ and }P(\text{draw})=r\quad\text{which implies}\quad p+q+r=1.$$

By using first-step analysis, we can set up a second-order linear homogeneous recurrence relation. That is,

$$px_{j+1}-(p+q)x_j+qx_{j-1}=0,$$

where $x_0 = 0$ and $x_N = 1$ and the aforementioned equation is referred to as the gambler's ruin equation. We shall prove this result. It suffices to show that

$$(p+q)x_j = px_{j+1} + qx_{j-1}$$

To transit from the x_j to x_{j+1}, the player needs to win, hence we multiply by the associated probability p. The same can be said for the transition from x_j to x_{j-1}, where the player needs to lose, implying that we multiply by q. For the player to remain at the same state, he needs to obtain a draw. That is, multiplying x_j by r. As such,

$$\begin{aligned} x_j &= px_{j+1} + qx_{j-1} + rx_j \\ (1-r)x_j &= px_{j+1} + qx_{j-1} \\ (p+q)x_j &= px_{j+1} + qx_{j-1} \end{aligned}$$

The initial conditions $x_0 = 0$ and $x_N = 1$ are obvious because when he has a fortune of \$0, it is impossible for him to win, and similarly, when he reaches \$$N$, he already won, implying that $P(A_N) = x_N = 1$

One can solve the recurrence relation to obtain the required probability

$$x_j = \frac{1-\left(\frac{q}{p}\right)^j}{1-\left(\frac{q}{p}\right)^N} \quad \text{if } p \neq q.$$

To see why the above equation holds, given the gambler's ruin equation

$$px_{j+1} - (p+q)x_j + qx_{j-1} = 0,$$

we first find the auxiliary equation. That is, $pm^2 - (p+q)m + q = 0$. Solving yields $m = 1$ or $m = \frac{q}{p}$. The solution to the recurrence relation is of the form

$$x_j = A + B\left(\frac{q}{p}\right)^j.$$

Setting $j = 0$ gives $A = -B$. Setting $j = N$ gives $x_N = 1$, which implies

$$A - A\left(\frac{q}{p}\right)^N = 1 \quad \text{so} \quad A = \frac{1}{1-\left(\frac{q}{p}\right)^N}$$

Once we have found A, we can find B, and the rest is simple algebraic manipulation. □

The interested reader can visit this link for more information on this interesting concept of the gambler's ruin problem.

4. Discrete Random Variables

4.1. *Discrete Random Variables*

A random variable is discrete if the range of X is either finite or countably infinite. The latter means that there exists a bijection $f : X \to \mathbb{N}$. Examples of discrete random variables include the binomial distribution covered in H2 Mathematics and the geometric distribution and the poisson distribution covered in H2 Further Mathematics.

Here, we will study a few more distributions. For example, the Bernoulli distribution, named after Jacob Bernoulli, who came from an academically gifted family that produced eight notable mathematicians and physicists. Also, we will study the negative binomial distribution, which is closely related to the geometric distribution, and the last addition to this series is the hypergeometric distribution, which is implicitly covered since one's O-Level days.

Definition 4.1 (discrete random variable). Suppose a random variable X is discrete, taking values $x_1, x_2, \ldots$. Then, the probability mass function of X is

$$P_X(x) = \begin{cases} P(X = x) & \text{if } x = x_1, x_2, \ldots; \\ 0 & \text{otherwise} \end{cases}$$

The probability mass function is abbreviated as PMF.

Proposition 4.1. Some properties of the probability density/mass function are as follows:

(i) $p_X(x_i) \geq 0$ for $i = 1, 2, \ldots$

(ii) $p_X(x) = 0$ for all other values of x

(iii) Since X must take on one of the values of x_i, then

$$\sum_{i=1}^{\infty} p_X(x_i) = 1.$$

We use uppercase letters to denote random variables and use lowercase letters to denote the values of random variables.

Definition 4.2 (cumulative distribution function). The cumulative distribution fufnction of X, or CDF in short and denoted by F_X, is defined as $F_X : \mathbb{R} \to \mathbb{R}$, where

$$F_X(x) = P(X \leq x) \quad \text{for } x \in \mathbb{R}.$$

If $x_1 < x_2 < x_3 < \ldots$, then, F is a step function. That is, F is constant in the interval $[x_{i-1}, x_i)$.

4.2. *Expectation*

Definition 4.3 (expectation). The expected value of X, or the expectation of X, denoted by $E(X)$ or μ_X, is defined by

$$E(X) = \sum_{\text{all } x} x P(X = x).$$

Proposition 4.2. We state some properties of expectation. Let X and Y be random variables and a and b be constants.

(i) $E(aX) = aE(X)$

(ii) $E(a) = a$

(iii) $E(aX \pm b) = aE(X) \pm b$

(iv) $E(aX \pm bY) = aE(X) \pm bE(Y)$

(v) If $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = nE(X).$$

Given a random variable X, we are often interested about $g(X)$ and $E(g(X))$. The question is how do we compute the latter? One method is to find the PDF of $g(X)$ first before proceeding to compute $E(g(X))$ by definition. We have the following proposition that if X is a discrete random variable that takes values x_i, where $i \geq 1$, with respective probabilities $p_X(x_i)$, then for any real-valued function g,

$$E(g(X)) = \sum_{\text{all } x} g(x) P_X(x).$$

We can derive the four properties of expectation stated above, as well as by setting $g(x) = x^2$, we have

$$E(X^2) = \sum_{\text{all } x} x^2 P_X(x).$$

We call this the second moment of X. We can generalise this result to $E(X^n)$ for $n \in \mathbb{N}$, which is of interest when we discuss the moment generating function of a random variable, as well as $E(1/X)$. As such,

$$E(X^n) = \sum_{\text{all } x} x^n P_X(x) \quad \text{and} \quad E\left(\frac{1}{X}\right) = \sum_{\text{all } x} \frac{1}{x} \cdot P_X(x).$$

Definition 4.4 (moment). In general, for $n \geq 1$, $E(X^n)$ is the n^{th} moment of X. The expected value of a random variable X, $E(X)$, is also referred to as the first moment or the mean of X.

Next, we define

$$E(X-\mu)^n$$

to be the n^{th} central moment of X. Hence, the first central moment is 0 and the second central moment is $E(X-\mu)^2$, which is called the variance of X.

Proposition 4.3 (tail sum formula for expectation). For a non-negative integer-valued random variable X,

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

4.3. *Variance and Standard Deviation*

Definition 4.5 (variance and standard deviation). If X is a random variable with mean μ, then the variance of X, denoted by $\text{Var}(X)$, is defined by

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

The standard deviation of X, denoted by σ_X, is defined by $\sqrt{\text{Var}(X)}$.

An alternative formula for variance is

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

Proof.

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

and the result follows. □

Note that $\text{Var}(X) \geq 0$ since it is the square of the standard deviation. Since standard deviation is defined as the spread of data about the mean, then the result follows. Alternatively, we can think of it in a more mathematical way. By the definition of $\text{Var}(X)$, we have $\text{Var}(X) = E(X-\mu)^2$. Note that the right side of the equation is non-negative, and hence the result follows too.

Definition 4.6 (degenerate random variable). We say that $\text{Var}(X) = 0$ if and only if X is a degenerate random variable.

Moreover, from the formula for variance, it follows that

$$E(X^2) \geq [E(X)]^2 \geq 0.$$

Is it true that for all $n \in \mathbb{N}$,

$$E(X^n) \geq [E(X)]^n?$$

We will discuss this in one of the final sections, and to prove this conjecture, we need to use a famous inequality called Jensen's inequality (Theorem 8.3).

Proposition 4.4. We state some properties of variance. Let X be a random variable and a and b be constants. Then,

(i) $\text{Var}(aX) = a^2\,\text{Var}(X)$

(ii) $\text{Var}(a) = 0$

(iii) $\text{Var}(aX+b) = a^2\,\text{Var}(X)$

(iv) If $X_1, X_2, \ldots, X_n$ are independent random variables, then

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

4.4. *Bernoulli Distribution*

Let us discuss the first special discrete random variable, known as the Bernoulli Distribution. If $X \sim \text{Bernoulli}(p)$,

$$X = \begin{cases} 1 & \text{if it is a success} \\ 0 & \text{if it is a failure} \end{cases}.$$

We refer p to be the probability of success and q to be the probability of failure. In particular, we say that p is the *parameter* of the distribution since it is the only term within the bracket. As such, $p+q=1$. Then, $P(X=1)=p$ and $P(X=0)=1-p=q$.

Proposition 4.5. The expectation and variance of a Bernoulli random variable with parameter p is

$$E(X)=p \quad \text{and} \quad \mathrm{Var}(X)=pq.$$

We first prove the result for expectation, which is obvious.

Proof. $E(X)=0\cdot q+1\cdot p=p$ □

Next, we prove the result for variance.

Proof. $E\left(X^2\right)=0^2\cdot q+1^2\cdot p=p$. As $[E(X)]^2=p^2$, then $\mathrm{Var}(X)=p-p^2=p(1-p)=pq$. □

Even though the Bernoulli distribution is rather *new* in this context, it is actually not new because it is closely related to the binomial distribution. We will discuss this in the next section.

To summarise,

Random Variable	PDF	Parameter(s)	$E(X)$	$\mathrm{Var}(X)$
Bernoulli(p)	$X=\begin{cases}1 & \text{if it is a success}\\ 0 & \text{if it is a failure}\end{cases}$	p	p	pq

4.5. *Binomial Distribution*

Suppose we perform an experiment n times and the probability of success for each trial is p. We define X to be the number of successes in n Bernoulli(p) trials. Then, X takes values between 0 and n inclusive and for $0\le k\le n$,

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

We can write it as $X\sim \mathrm{B}(n,p)$ and we the values of k the random variable can take are referred to as the *support* of X. Recall that there are k successes and hence, $n-k$ failures. The probability of success and probability of failure are p and q respectively. Thus, we obtain the PDF formula for the binomial random variable.

Some examples where the binomial distribution can be used are as follows:

Example 4.1 (number of correct answers from multiple-choice questions)**.** The probability of getting right answers out of 20 multiple-choice questions when one out of four options were chosen arbitrarily. Here, X denotes the number of right answers. The probability of an answer being right is $\frac{1}{4}$. The binomial distribution can be represented as $X\sim \mathrm{B}\left(20,\frac{1}{4}\right)$.

Example 4.2 (coin toss)**.** Suppose a coin is tossed 50 times and we wish to find out how many heads we obtain. Here, X is the number of successes. That is the number of times heads occurs. The probability of getting a head is $\frac{1}{2}$. The binomial distribution could be represented as $X\sim \mathrm{B}\left(50,\frac{1}{2}\right)$.

Proposition 4.6. Let $X \sim \mathrm{B}(n,p)$. Then,

$$E(X) = np \quad \text{and} \quad \mathrm{Var}(X) = npq.$$

We will only prove the formula for expectation.

Proof.

$$E(X) = \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k} = \sum_{k=1}^{n} k\binom{n}{k}p^k q^{n-k} = \sum_{k=0}^{n} n\binom{n-1}{k-1}p^k q^{n-k}$$

Using the substitutions $m = n-1$ and $j = k-1$,

$$np\sum_{k=0}^{n}\binom{n-1}{k-1}p^{k-1}q^{(n-1)-(k-1)} = np\sum_{j=0}^{m}\binom{m}{j}p^j q^{m-j}.$$

We then note that

$$\sum_{j=0}^{m}\binom{m}{j}p^j q^{m-j} \quad \text{is} \quad \text{the sum of probabilities of the binomial random variable,}$$

which is 1, and we are done. □

Alternatively, we can prove the expectation formula by considering it as a sum of independent Bernoulli trials. For the variance proof, I will leave it as an exericse as it is not too complicated and the technique is, of course, similar to that for the expectation. When proving the formula for variance, note that

$$\mathrm{Var}(X) = E\left[X(X-1)\right] + E(X) - \left[E(X)\right]^2$$

and a classic trick to proving this result is by finding an expression for $E[X(X-1)]$.

Definition 4.7 (mode). The mode is the value k at which the PDF takes its maximum value. In other words, it is the value that is most likely to be sampled.

Proposition 4.7. For a binomial distribution with parameters n and p, the mode is

$$k = \lfloor (n+1)p \rfloor \text{ or } k = \lceil (n+1)p \rceil - 1.$$

Proof. Note that

$$\frac{P(X=k+1)}{P(X=k)} = \frac{p(n-k)}{(1-p)(k+1)}$$

which can be easily derived via the PDF of the binomial distribution. For convenience sake, we set $P(X=k) = f(k)$. There are three cases to consider, namely $f(k) > f(k+1)$, $f(k) < f(k+1)$ and $f(k) = f(k+1)$. For the last case, where $f(k) = f(k+1)$, the graph of the binomial distribution has two peaks, or two maximum points. Such a distribution is *bimodal*.

For the first case,

$$\frac{p(n-k)}{(1-p)(k+1)} < 1,$$

which implies that $(n+1)p < k+1$. Since we know that k is the mode, by definition of the floor function, the result follows. It is not difficult to prove the modal result for the other two cases. I shall leave this as an exercise. □

Random Variable	PDF	Parameter(s)	$E(X)$	$\mathrm{Var}(X)$
$\mathrm{B}(n,p)$	$\binom{n}{k}p^kq^{n-k}$	n and p	np	npq

Proposition 4.8 (additivity). The sum of two independent binomial random variables with the same probability of success, p, still follows a binomial distribution. That is, if $X \sim \mathrm{B}(m,p)$ and $Y \sim \mathrm{B}(n,p)$, then $X+Y \sim \mathrm{B}(m+n,p)$.

Proof: Note that $X+Y$ takes values between 0 and $m+n$ inclusive. Hence,

$$\begin{aligned}
P(X+Y=k) &= \sum_{i=0}^{k} P(\{X=i\}\cap\{Y=k-i\}) \\
&= \sum_{i=0}^{k} P(X=i)P(Y=k-i) \\
&= \sum_{i=0}^{k} \binom{m}{i} p^i q^{m-i} \binom{n}{k-i} p^{k-i} q^{n-k+i} \\
&= p^k q^{m+n-k} \sum_{i=0}^{k} \binom{m}{i}\binom{n}{k-i} \\
&= \binom{m+n}{k} p^k q^{m+n-k}
\end{aligned}$$

From the second last line to the last line, we used Vandermonde's identity (Theorem 1.4). □

Example 4.3 (ST2131 AY24/25 Sem 1 Lecture 6). A fair coin is tossed repeatedly. The outcomes of the tosses are assumed to be independent.

(a) Let p be the probability of getting 30 heads before 10 tails. Let q be the probability of getting 30 tails before 10 heads. Is $p=q$?

(b) Let p be the probability of getting 30 heads before 10 tails. Let q be the probability of getting 10 heads before 30 tails. Is $p=q$?

(c) Let p be the probability of getting 30 heads before 10 tails. Let q be the probability of getting 10 heads before 30 tails. Is $p+q=1$?

Solution.

(a) This is true. The coin is fair, so by symmetry it is true.

(b) False. Intuitively, it is easier to reach 10 heads first before 30 heads. We can do some calculations to verify this.

$$p = \sum_{k=30}^{39}\binom{39}{k}\left(\frac{1}{2}\right)^k\left(1-\frac{1}{2}\right)^{39-k} = \sum_{k=30}^{39}\binom{39}{k}\left(\frac{1}{2}\right)^{39}$$

On the other hand,

$$q = \sum_{k=10}^{39}\binom{39}{k}\left(\frac{1}{2}\right)^k\left(1-\frac{1}{2}\right)^{39-k} = \sum_{k=10}^{39}\binom{39}{k}\left(\frac{1}{2}\right)^{39}.$$

It follows that $q>p$.

(c) True. Observe that both events are complements, so they must add up to 1. □

Theorem 4.1 (serve-and-rally)**.** To find the probability that player A wins a serve-and-rally match where A serves first and $2n-1$ rallies are played, where the win condition is n points, we have

$$\begin{aligned} P(A \text{ wins match}) &= P(A \text{ wins at least } n \text{ points in } 2n-1 \text{ rallies}) \\ &= \sum_{m=n}^{2n-1} P(A \text{ wins exactly } m \text{ points in } 2n-1 \text{ rallies}) \end{aligned}$$

where if p_A and p_B are the probabilities that A wins when A and B serve respectively,

$$P(A \text{ wins exactly } m \text{ points in } 2n-1 \text{ rallies})$$

is equal to the following:

$$\begin{aligned} &= \sum_{k=1}^{n} P(A \text{ wins } k \text{ points he serves}) \cdot P(A \text{ wins } m-k \text{ points } B \text{ serves}) \\ &= \sum_{k=1}^{n} \binom{n}{k} (p_A)^k (1-p_A)^{n-k} \binom{n-1}{m-k} (p_B)^{m-k} (1-p_B)^{n-m+k-1} \end{aligned}$$

Example 4.4 (ST2131 AY24/25 Sem 1 Lecture 7; serving protocol)**.** A quick badminton serve-and-rally match is played following the alternative serve protocol, with player A serving first. The rules are changed so that the match ends when a player wins 2 points (rallies), and that player is declared the winner of the match.

If player A has a 60% chance of winning a rally when he serves, but only 30% chance of winning when player B serves, what is the probability of player A winning the match?

Solution. If we want player A to win the match, it means we want A to win at least 2 points in the first 3 games. That is, we find

$$P(A \text{ wins}) = P(A \text{ wins 2 rallies in first 3}) + P(A \text{ wins 3 rallies in first 3})$$

A natural question is why do we consider the case that A wins 3 rallies, since he would have already won with 2. This assumption works because no matter how that game ends, we still put artificial rallies at the back such that we can always assume that 3 rallies are always played. Since the outcome will always be the same, it does not matter that we append rallies afterwards.

- **Case 1:** Suppose A serves 1 and B serves 1. Then, A wins 1 rally he serves, A wins 1 rally that B serves, and the remaining rally is lost by A, which yields a probability of
$$\binom{2}{1}(0.6)\binom{1}{1}(0.3)(1-0.6).$$
- **Case 2:** Suppose A serves 2. Then, A wins the 2 rallies he serves. The remaining rally is served by B, so A has to lose the game. This yields a probability of
$$\binom{2}{2}(0.6)^2\binom{1}{1}(1-0.3).$$
- **Case 3:** Suppose A wins all 3 rallies. So, A serves 2 rallies with both winning, and B serves one rally. This yields a probability of $(0.6)^2(0.3)$.

The desired answer is the sum of probabilities in all three cases, which is approximately 0.5. □

4.6. *Geometric Distribution*

Let X be the random variable denoting the number of Bernoulli trials required to obtain the first success, where the probability of success is p. Here, the support of X is the positive integers $1,2,3,\ldots$ because the minimum number of tries required to obtain the first success is 1. As such, it is easy to derive the following formula, which is the PDF of X:

$$P(X=k)=pq^{k-1}$$

We say that $X \sim \text{Geo}(p)$. In certain textbooks, the geometric distribution is defined to be the number of failures in the Bernoulli trials in order to obtain the first success. However, we will stick to the former definition.

Example 4.5. An example where the geometric distribution can be used includes the number of tries up to and including finding a defective item on a production line.

Proposition 4.9. Suppose $X \sim \text{Geo}(p)$. Then,

$$E(X)=\frac{1}{p} \quad \text{and} \quad \text{Var}(X)=\frac{q}{p^2}.$$

Again, we will only prove the formula for expectation.

Proof. By definition,

$$E(X)=\sum_{k=1}^{\infty} kpq^{k-1}.$$

□

Suppose $f(q)=q^k$. Then, $f'(q)=kq^{k-1}$. Hence,

$$\sum_{k=1}^{\infty} kpq^{k-1}=p\sum_{k=1}^{\infty}\frac{d}{dq}\left(q^k\right)=p\frac{d}{dq}\left(\sum_{k=1}^{\infty}q^k\right)=p\frac{d}{dq}\left(\frac{q}{1-q}\right)=\frac{1}{p}.$$

This proof uses the technique of using a derivative to replace the summand. □

Definition 4.8 (memorylessness)**.** A probability distribution is said to have a memoryless property if the probability of some future event occurring is not affected by the occurrence of past events. If a random variable X satisfies the memoryless property, then for $m,n \in \mathbb{N}$,

$$P(X>m+n \mid X>m)=P(X>n).$$

In particular, the geometric distribution is the only distribution that exhibits the memoryless property. For the continuous counterpart, the exponential distribution is the only one which exhibits memorylessness. We will discuss this in due course. It is easy to verify that the geometric distribution has the memoryless property but to prove that it is the only one, it is slightly more complicated.

Proof. Suppose X is a random variable which satisfies the memoryless property. Then,

$$P(X>m+n|X>m)=P(X>n).$$

We apply the definition of conditional probability to the left side. Hence,

$$P(X>m+n)=P(X>m)P(X>n).$$

Note that $P(X > 0) = 1$ since the support of X is the positive integers. Then,

$$\begin{aligned}
P(X > 1) &= [P(X > 0)]^2 = 1 \\
P(X > 2) &= P(X > 1)P(X > 1) = [P(X > 1)]^2 \\
P(X > 3) &= P(X > 1)P(X > 2) = [P(X > 1)]^3
\end{aligned}$$

It is clear that

$$P(X > k) = [P(X > 1)]^k.$$

To compute $P(X = k)$, we use the formula $P(X = k) = P(X > k-1) - P(X > k)$ to get

$$\begin{aligned}
P(X = k) &= [P(X > 1)]^{k-1} - [P(X > 1)]^k \\
&= [P(X > 1)]^{k-1}[1 - P(X > 1)]
\end{aligned}$$

Setting $p = 1 - P(X > 1)$, we obtain

$$P(X = k) = p(1-p)^{k-1},$$

which is indeed the PDF of the geometric distribution with parameter p. □

In the above proof, note that $q = P(X > 1)$, which is clear because we claim that p is the probability of success, or in relation to attempts, p is the probability of attaining a success on the first try. That is, $p = P(X = 1)$.

To summarise,

Random Variable	PDF	Parameter(s)	$E(X)$	$\text{Var}(X)$
$\text{Geo}(p)$	pq^{k-1}	p	$\frac{1}{p}$	$\frac{q}{p^2}$

Previously, we claimed that the sum of two independent binomial random variables with the same probability of success p will still follow a binomial distribution. However, if we have the sum of two geometric distributions (namely X and Y) with the same probability of success p, then $X+Y$ actually follows a negative binomial distribution! That is, $X+Y \sim \text{NB}(2,p)$.

Proposition 4.10 (additivity). If $X,Y \sim \text{Geo}(p)$, then $X+Y \sim \text{NB}(2,p)$.

Proof. Note that $X+Y$ takes values $2,3,\ldots$. We set $k \geq 2$. Then,

$$\begin{aligned}
P(X+Y=k) &= \sum_{i=1}^{k-1} P(\{X=i\} \cap \{Y=k-i\}) \\
&= \sum_{i=1}^{k-1} P(X=i)\, P(Y=k-i) \\
&= \sum_{i=1}^{k-1} pq^{i-1} pq^{k-i-1} \\
&= (k-1)\, p^2 q^{k-2} \\
&= \binom{k-1}{1} p^2 q^{k-2}
\end{aligned}$$

which is the PDF of a negative binomial random variable with parameters $(2,p)$. We will formally introduce this distribution in due course, but this is just to illustrate that the sum of identical distributions with the same parameters may not result in the new distribution to be of the same kind as the original. □

One of the random variables which we would encounter under discrete random variables is the Poisson distribution. Later, we will see that the sum of two Poisson random variables also follows a Poisson distribution.

Example 4.6 (ST2131 AY24/25 Sem 1 Lecture 6). A coin is tossed repeatedly. The outcomes of the tosses are assumed to be independent. The coin is biased with each toss showing head with probability 60%.

What is the probability of getting a run of 3 consecutive heads before a run of 2 consecutive tails?

Solution. Observe that this process is memoryless. That is, if we get some string, say HHT, then everything is invalidated, since we need 3 heads in a row. Hence, it suffices to only keep track of the *effective states*. □

Example 4.7 (coupon collector's problem). The coupon collector's problem describes "collect all coupons and win" contests. It asks the following question: if each box of a brand of cereals contains a coupon, and there are n different types of coupons, what is the probability that more than t boxes need to be bought to collect all n coupons?

Solution. By letting T be the number of draws needed to collect all n coupons and t_i be the time to collect the i^{th} coupon after $i-1$ coupons have been collected and regarding them as random variables, then the probability of collecting a new coupon, denoted by p_i, can be written as

$$p_i = \frac{n-i+1}{n}.$$

To see why, the i^{th} coupon must be different from all the previous collected. The probability of obtaining a coupon that is of the same type as any one of the i coupons previously collected is $\frac{i-1}{n}$. Hence,

$$p_i = 1 - \frac{i-1}{n}$$

and the result follows. $\square$

We remark that t_i follows a geometric distribution with parameter p_i and $T = t_1 + t_2 + \ldots + t_n$. We shall prove two interesting results, which are expressions for $E(T)$ and $\text{Var}(T)$, and they are related to the harmonic numbers and the famous Basel problem respectively:

Theorem 4.2.

$$E(T) = nH_n \text{ and } \text{Var}(T) < \frac{n^2\pi^2}{6},$$

where H_n is the n^{th} harmonic number.

For $\text{Var}(T)$, it is rather interesting that we do not have an explicit formula but only an upper bound for it. We shall first prove the result for expectation.

Proof. Assume that the t_i's are independent. Then,

$$\begin{aligned} E(T) &= E(t_1 + t_2 + \ldots + t_n) \\ &= E(t_1) + E(t_2) + \ldots + E(t_n) \\ &= \sum_{i=1}^{n} \frac{n}{n-i+1} \\ &= n\sum_{i=1}^{n} \frac{1}{n-i+1} \\ &= n\sum_{i=1}^{n} \frac{1}{i} \\ &= nH_n \end{aligned}$$

$\square$

Next, we prove the result for variance.

Proof.

$$\begin{aligned} \text{Var}(T) &= \text{Var}(t_1 + t_2 + \ldots + t_n) \\ &= \text{Var}(t_1) + \text{Var}(t_2) + \ldots + \text{Var}(t_n) \\ &= \sum_{i=1}^{n} \frac{1-p_i}{p_i^2} \\ &= n\sum_{i=1}^{n} \frac{i-1}{(n-i+1)^2} \end{aligned}$$

The Basel problem, proved by Leonhard Euler in 1734, states that

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}.$$

Thus, it suffices to prove

$$\sum_{i=1}^{n} \frac{i-1}{(n-i+1)^2} < \sum_{i=1}^{n} \frac{n}{(n-i+1)^2} < \sum_{i=1}^{\infty} \frac{n}{i^2} = \frac{n\pi^2}{6}.$$

This is true because $i-1<n$ if and only if $i<n+1$. Hence, the result follows. □

4.7. *Negative Binomial Distribution*

Define the random variable X to be the number of Bernoulli trials, with parameter p, required to obtain r successes. Here, the support of X is $k \geq r$ and we say that the distribution is negative binomial with parameters r and p. We write $X \sim \mathrm{NB}(r, p)$. The PDF of a negative binomial random variable is

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

We can think of the negative binomial distribution as such: for the first $k-1$ trials, we wish to have $r-1$ successes. As such, there are $k-r$ failures. Then, we ensure that the k^{th} trial is a success and we are done.

The geometric distribution is a special case of the negative binomial distribution. We can view the geometric distribution $\mathrm{Geo}(p)$ as $\mathrm{NB}(1, p)$ since for the geometric distribution, we are interested in the number of tries up to and including the first success.

Proposition 4.11. The expectation and variance of the negative binomial distribution $X \sim \mathrm{NB}(r, p)$ are

$$E(X) = \frac{r}{p} \text{ and } \mathrm{Var}(X) = \frac{rq}{p^2}.$$

To summarise, We move on to discuss Banach's matchbox problem. This problem is named after the

Random Variable	PDF	Parameter(s)	$E(X)$	$\mathrm{Var}(X)$
$\mathrm{NB}(r, p)$	$\binom{k-1}{r-1} p^r q^{k-r}$	r and p	$\frac{r}{p}$	$\frac{rq}{p^2}$

mathematician Stefan Banach, who is known for the Banach-Tarski paradox, a problem encompassing the elements of Set Theory and Geometry (Vsauce made a video on this in 2015 so do check it out if you are interested).

Example 4.8 (Banach's Matchbox Problem)**.** We state Banach's Matchbox Problem. Suppose a mathematician carries two matchboxes at all times — one in his left pocket and one in his right. Each time he needs a match, he is equally likely to take it from either pocket. Suppose he reaches into his pocket and discovers for the first time that the box picked is empty. If it is assumed that each of the matchboxes originally contained N matches, what is the probability that there are exactly k matches in the other box?

Solution. Let E be the event that the mathematician first discovers that the right pocket matchbox is empty and there are k matches in the left pocket matchbox at that instant. E will occur if and only if the $(N+1)^{\text{th}}$ choice of the right pocket matchbox is made at the $(N+1+N-i)^{\text{th}}$ trial. We see that this setup is essentially using a negative binomial distribution model with parameters $r = N+1$ and $p = 1/2$. Here, $k = 2N - i + 1$. As such,

$$P(E) = \binom{2N-i}{N}\left(\frac{1}{2}\right)^{2N-i+1}.$$

As there is an equal probability that the left pocket matchbox is the first to be discovered to be empty and there are k matches in the right pocket matchbox at that time, the desired result is simply $2P(E)$, or

$$\binom{2N-i}{N}\left(\frac{1}{2}\right)^{2N-i}.$$

□

4.8. *Poisson Distribution*

A random variable X is said to follow a Poisson distribution with parameter λ if the support of X is the non-negative integers $0, 1, 2, \ldots$ with probabilities

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

We say that $X \sim \mathrm{Po}(\lambda)$.

Some examples where the Poisson distribution can be used are as follows.

Example 4.9 (calls per hour)**.** Call centers use the Poisson distribution to model the number of expected calls per hour that they'll receive so they know how many call center reps to keep on staff. For example, suppose a given call center receives 10 calls per hour. Then, $X \sim \mathrm{Po}(10)$.

Example 4.10 (arrivals)**.** Restaurants use the Poisson distribution to model the number of expected customers that will arrive at the restaurant per day. Suppose a restaurant receives an average of 100 customers per day. Then, $X \sim \mathrm{Po}(100)$.

Time plays a critical role when defining a Poisson random variable.

Proposition 4.12. If $X \sim \mathrm{Po}(\lambda)$, then

$$E(X) = \lambda \quad \text{and} \quad \mathrm{Var}(X) = \lambda.$$

We shall prove the result for expectation.

Proof.

$$E(X) = \sum_{k=0}^{\infty} k \cdot \frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} k \cdot \frac{e^{-\lambda}\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda$$

□

To summarise,

Random Variable	PDF	Parameter(s)	$E(X)$	$\mathrm{Var}(X)$
$\mathrm{Po}(\lambda)$	$\frac{e^{-\lambda}\lambda^k}{k!}$	λ	λ	λ

Proposition 4.13 (additivity). The additivity property of the Poisson distribution states that if X and Y are independent Poisson random variables where $X \sim \text{Po}(\lambda)$ and $Y \sim \text{Po}(\mu)$, then $X+Y \sim \text{Po}(\lambda+\mu)$.

Proof.

$$\begin{aligned}
P(X+Y=n) &= \sum_{k=0}^{n} P(\{X=k\} \cap \{Y=n-k\}) \\
&= \sum_{k=0}^{n} P(X=k)P(Y=n-k) \quad \text{since } X \text{ and } Y \text{ are independent} \\
&= \sum_{k=0}^{n} \frac{e^{-\lambda}\lambda^k}{k!} \cdot \frac{e^{-\mu}\mu^{n-k}}{(n-k)!} \\
&= \frac{e^{-(\lambda+\mu)}\mu^n}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{\mu}\right)^k \\
&= \frac{e^{-(\lambda+\mu)}\mu^n}{n!} \sum_{k=0}^{n} \binom{n}{k} \left(\frac{\lambda}{\mu}\right)^k \\
&= \frac{e^{-(\lambda+\mu)}\mu^n}{n!} \left(1+\frac{\lambda}{\mu}\right)^n \\
&= \frac{e^{-(\lambda+\mu)}(\lambda+\mu)^n}{n!}
\end{aligned}$$

□

Even though $X+Y$ follows a Poisson distribution, $X-Y$ actually does not follow a Poisson distribution. In general, the difference of two Poisson Random Variables is said to follow a Skellam distribution. Its probability mass function is rather complicated to compute as it involves the modified Bessel function of the first kind (related to differential equations).

Example 4.11 (ST2131 AY24/25 Sem 1 Lecture 16). A sample of radioactive substance is observed to emit 0.3 α-particles per second, on average. A sample of another substance is observed to emit 0.5 α-particles per second, on average. The two samples are now combined.

(a) What is the probability that at least three α-particles are emitted from the combined sample in a ten second interval?

(b) What is the longest time interval we need to wait in order that the probability of the combined sample emitting any α-particle in that time interval is $> 90\%$?

Solution.

(a) This is a Poisson process. Denote the first substance with $M(t) \sim \text{Po}(0.3t))$ and the second substance with $N(t) \sim \text{Po}(0.5t)$. Then $X(t) = M(t)+N(t) \sim \text{Po}(0.3t+0.5t) = \text{Po}(0.8t)$. So, $P(X(10) > 3) = 1-P(X(10)=0)-P(X(10)=1)-P(X(10)=2)$ which is equal to

$$1-e^{-8}\left(1+8+\frac{8^2}{2!}\right).$$

(b) Let T be this said time interval. We have

$$P(X(T) \geq 1) > 0.9 \quad \text{if and only if} \quad P(X(T)=0) < 0.1$$

So, $e^{-0.8T} < 0.1$, which implies $T > \ln 10/0.8$ as desired. □

Proposition 4.14 (conditional of Poisson distribution is binomial). If X and Y are independent Poisson random variables such that $X \sim \mathrm{Po}(\lambda)$ and $Y \sim \mathrm{Po}(\mu)$, then

$$P(X = k \mid X + Y = n) = P(J = k),$$

where

$$J \sim \mathrm{B}\left(n, \frac{\lambda}{\lambda + \mu}\right).$$

Proof.

$$\begin{aligned} P(X = k \mid X + Y = n) &= \frac{P(\{X = k\} \cap \{Y = n - k\})}{P(X + Y = n)} \\ &= \frac{P(X = k) P(Y = n - k)}{P(X + Y = n)} \qquad \text{since } X \text{ and } Y \text{ are independent} \end{aligned}$$

By applying the respective density formulae, the above simplifies to

$$\begin{aligned} \frac{e^{-\lambda}\lambda^k}{k!} \cdot \frac{e^{-\mu}\mu^{n-k}}{(n-k)!} \cdot \frac{n!}{e^{-(\lambda+\mu)}(\lambda+\mu)^n} &= \frac{\mu^n}{(\lambda+\mu)^n} \cdot \binom{n}{k}\left(\frac{\lambda}{\mu}\right)^k \\ &= \binom{n}{k}\left(\frac{\lambda}{\lambda+\mu}\right)^k\left(\frac{\mu}{\lambda+\mu}\right)^{n-k} \end{aligned}$$

which is indeed the probability mass function of a binomial random variable with n tries and probability of success $\lambda/(\lambda+\mu)$. □

This is from a past year A-Level Mathematics Special Paper dated back to 2004. It is the equivalent of the current H3 Mathematics.

Example 4.12 (Special Paper 2004). Fish comes to the surface of a stretch of river randomly and independently at a mean rate of 8 per minute. When a fish comes to the surface, the probability that it catches a fly is 0.6. If S is the number of flies caught in a randomly chosen minute, show that

$$P(S = s) = \sum_{r=s}^{\infty} \frac{e^{-8}8^r}{r!}\binom{r}{s}(0.6)^s(0.4)^{r-s}$$

and deduce that S follows a Poisson distribution.

Solution. Let R be the random variable denoting the number of fish coming to the surface in a minute. The probability that r fish come to the surface in a randomly chosen minute is

$$P(R = r) = \frac{e^{-8}8^r}{r!}.$$

The probability that s flies are caught during a period of a randomly chosen minute in which r fish come to the surface, where $s \le r$, is

$$\frac{e^{-8}8^r}{r!}\binom{r}{s}(0.6)^s(0.4)^{r-s}.$$

Hence,

$$\begin{aligned} P(S = s) &= P(\{R = s\} \cap \{S = s\}) + P(\{R = s+1\} \cap \{S = s\}) + \ldots \\ &= \sum_{r=s}^{\infty} P(\{S = s\} \cap \{R = r\}) \\ &= \sum_{r=s}^{\infty} P(R = r) P(S = s | R = r) \\ &= \sum_{r=s}^{\infty} \frac{e^{-8}8^r}{r!}\binom{r}{s}(0.6)^s(0.4)^{r-s} \end{aligned}$$

To prove that S follows a Poisson distribution, we manipulate with the given probability mass function formula.

$$\begin{aligned}
P(S=s) &= \sum_{r=s}^{\infty} \frac{e^{-8}8^r}{r!}\binom{r}{s}(0.6)^s(0.4)^{r-s} \\
&= \frac{e^{-8}(1.5)^s}{s!}\sum_{r=s}^{\infty}\frac{(3.2)^r}{(r-s)!} \\
&= \frac{e^{-8}(1.5)^s}{s!}\sum_{j=0}^{\infty}\frac{(3.2)^{j+s}}{j!} \quad \text{by setting } r-s=j \\
&= \frac{e^{-8}(4.8)^s}{s!}\sum_{j=0}^{\infty}\frac{(3.2)^{j}}{j!} \\
&= \frac{e^{-4.8}(4.8)^s}{s!}
\end{aligned}$$

This asserts that S indeed follows a Poisson distribution with parameter 4.8. That is, $S \sim \mathrm{Po}(4.8)$. □

The Poisson distribution has a variety of applications in diverse areas.

Theorem 4.3 (law of raw events)**.** The Poisson distribution can be used as an approximation for a binomial random variable with parameters (n,p) when n is large and p is small enough so that np is of moderate size.

Proof. Suppose $X \sim \mathrm{B}(n,p)$ and let $\lambda = np$. Then, by first using the binomial PDF formula,

$$\begin{aligned}
P(X=k) &= \frac{n!}{k!(n-k)!}p^k q^{n-k} \\
&= \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k}\cdot\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^{n-k}
\end{aligned}$$

For large n and a moderate-sized λ,

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda} \qquad \left(1-\frac{\lambda}{n}\right)^k \approx 1 \qquad \frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k} \approx 1.$$

Hence, we conclude that

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

□

4.9. *Hypergeometric Distribution*

The hypergeometric distribution describes the probability of k successes in n draws, without replacement, from a finite population of size N that contains K objects with that feature, wherein each draw is either a success or a failure. In contrast, the binomial distribution describes the probability of k successes in n draws with replacement.

Definition 4.9 (hypergeometric distribution)**.** If a random variable follows a hypergeometric distribution with parameters N, K and n, then

$$P(X=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

We say that $X \sim \text{Hypergeometric}(N,K,n)$.

Proposition 4.15. The sum of probabilities is indeed equal to 1. That is,

$$\sum_{0\le k\le n} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} = 1.$$

Proof. Use Vandermonde's identity (Theorem 1.4). □

Example 4.13. Michael has a box of 8 blue balls and 6 red balls. He draws 3 balls from the box without replacement. Calculate the probability that 2 balls are red.

Solution. We can use the probability mass function formula of a hypergeometric distribution. Note that $N = 14$, $K = 6$, $n = 3$ and $k = 2$. Substituting everything into the formula yields

$$P(X=2) = \frac{\binom{6}{2}\binom{8}{1}}{\binom{14}{3}} = \frac{30}{91}.$$

However, we can think of it from an O-Level student's perspective. I believe questions of this type were covered in Secondary Four. We have the following cases: *RRB*, *RBR* and *BRR*. For the first case, the probability is

$$\frac{6}{14}\times\frac{5}{13}\times\frac{8}{12} = \frac{10}{91}.$$

Observe that the probabilities for the other two cases are the same, namely

$$\frac{6}{14}\times\frac{8}{13}\times\frac{5}{12} \text{ and } \frac{8}{14}\times\frac{6}{13}\times\frac{5}{12}$$

respectively. Hence, the answer we obtain is $\frac{30}{91}$ too, yielding the same conclusion as before. So, it appears that the hypergeometric distribution is not something exactly new! □

Our O-Level method actually has a potential limitation, which is that if n and k are large, the total number of permutations will also be large and many cases will arise[†].

Proposition 4.16. The expectation and variance of a hypergeometric random variable are

$$E(X) = \frac{nK}{N} \quad \text{and} \quad \text{Var}(X) = \frac{nK(N-K)(N-n)}{N^2(N-1)}.$$

To summarise,

Random Variable	PDF	Parameter(s)	$E(X)$	$\text{Var}(X)$
$\text{Hypergeometric}(N,K,n)$	$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$	N,K and n	$\frac{nK}{N}$	$\frac{nK(N-K)(N-n)}{N^2(N-1)}$

[†]When I first wrote this set of notes in 2022, I had an analogy regarding the number of COVID-19 cases. I mentioned that the 'number of cases will arise' just like how the number of COVID-19 cases there are as of *now* (back then) when I'm writing this which is 4 July 2022.

To summarise the main components of discrete random variables,

Random Variable	PDF	Parameter(s)	$E(X)$	$\mathrm{Var}(X)$
Bernoulli(p)	$X = \begin{cases} 1 & \text{if it is a success} \\ 0 & \text{if it is a failure} \end{cases}$	p	p	pq
$\mathrm{B}(n,p)$	$\binom{n}{k} p^k q^{n-k}$	n and p	np	npq
Geo(p)	pq^{k-1}	p	$\frac{1}{p}$	$\frac{q}{p^2}$
NB(r,p)	$\binom{k-1}{r-1} p^r q^{k-r}$	r and p	$\frac{r}{p}$	$\frac{rq}{p^2}$
Po(λ)	$\frac{e^{-\lambda}\lambda^k}{k!}$	λ	λ	λ
Hypergeometric(N,K,n)	$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$	N, K and n	$\frac{nK}{N}$	$\frac{nK(N-K)(N-n)}{N^2(N-1)}$

5. Continuous Random Variables

In discrete random variables, our support, or the set of possible values, is countable. The support can be finite (i.e. binomial distribution) or infinite (i.e. geometric distribution). In this section, we wish to study the continuous counterpart, and the property of such random variables is that their set of possible values is uncountable.

In this case, elements like time, a person's height etc. come into play. For example, the lifetime of an electrical appliance might follow an exponential distribution and the amount of rainfall obtained in a region during the dry season might be modelled by a continuous uniform distribution. Such scenarios are examples which make use of continuous random variables.

Definition 5.1 (continuous random variable). We say that X is a continuous random variable if there exists a non-negative function f_X, defined for all real $x \in \mathbb{R}$, having the property that for any set B of real numbers,

$$P(X \in B) = \int_B f_X(x)\,dx.$$

The function f_X is called the PDF of X.

Recall that PDF stands for probability density function. By letting $B = [a,b]$, we obtain

$$P(a \le X \le b) = \int_a^b f_X(x)\,dx.$$

Definition 5.2 (cumulative distribution function). We define the cumulative distribution function, or CDF, of X by

$$F_X(x) = P(X \le x) \quad \text{for } x \in \mathbb{R}.$$

Note that the definition of the distribution function is the same for both discrete and continuous random variables. Therefore, in the context of continuous random variables,

$$F_X(x) = \int_{-\infty}^x f_X(t)\,dt.$$

By the fundamental theorem of calculus,

$$F_X'(x) = f_X(x).$$

Observe that in continuous random variables, so far, we have been dealing with integrals. However, for discrete random variables, we only talked about sums. This is not surprising because the extension from discrete to continuous random variables involves Riemann integration. To further justify this, each partition gets finer and hence, the limit of the Riemann sums is equivalent to an integral.

Going back, the PDF is regarded as the derivative of the CDF, or the cumulative distribution function. More intuitively, we have

$$P\left(x - \frac{\varepsilon}{2} \le X \le x + \frac{\varepsilon}{2}\right) = \int_{x-\frac{\varepsilon}{2}}^{x+\frac{\varepsilon}{2}} f_X(x)\,dx \approx \varepsilon f(x).$$

This occurs when ε is small and when f is continuous at x. The probability that X will be contained in an interval of length ε around the point x is approximately $\varepsilon f(x)$. Hence, we see that $f(x)$ is a measure of how likely that the random variable would be near x.

Proposition 5.1. We establish some properties in relation to continuous random variables.

(i) $P(X = x) = 0$

(ii) The CDF, that is F_X, is continuous

(iii) For any $a, b \in \mathbb{R}$,

$$\begin{aligned} P(a \le X \le b) &= P(a < X \le b) \\ &= P(a \le X < b) \\ &= P(a < X < b) \end{aligned}$$

(iv) Since the sum of probabilities is equal to 1, then

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

5.1. *Expectation and Variance*

We shall write $f_X(x)$ simply as $f(x)$ for convenience sake.

Definition 5.3 (expectation and variance)**.** Let X be a continuous random variable with PDF $f(x)$. Then,

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx \text{ and } \operatorname{Var}(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x)\, dx.$$

Note that these are analogous to the formulae for expectation and variance for the discrete counterpart, just that for continuous random variables, the sum is changed to an integral. We can manipulate the expression for variance till it resembles that of $E(X^2) - [E(X)]^2$. That is,

$$\operatorname{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \left(\int_{-\infty}^{\infty} x f(x)\, dx \right)^2.$$

The linearity properties for expectation and variance also apply here. That is, $E(aX \pm b) = aE(X) \pm b$ and $\operatorname{Var}(aX \pm b) = a^2 \operatorname{Var}(X)$.

If X is a continuous random variable with PDF $f(x)$, then for any real-valued function g,

$$E[g(X)] = \int_{-\infty}^{\infty} f(x) g(x)\, dx.$$

Definition 5.4 (continuous uniform distribution)**.** A random variable X is uniformly distributed over the interval $(0,1)$ if its PDF is

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1; \\ 0 & \text{otherwise.} \end{cases}$$

We denote it by $X \sim U(0,1)$. In general, for $a < b$, we say that a random variable X is uniformly distributed over the interval (a,b) if its PDF is

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

This is denoted by $X \sim U(a,b)$.

Theorem 5.1 (triangular distribution and Irwin-Hall distribution)**.** The sum of two independent, equally distributed, uniform distributions yields a symmetric triangular distribution. In general, if we have n independent and identically distributed (i.i.d.) uniform distributions $U(0,1)$, the new distribution is said to follow an Irwin-Hall distribution.

Proposition 5.2. The expectation and variance of a uniform distribution $X \sim U(a,b)$ are

$$E(X) = \frac{a+b}{2} \quad \text{and} \quad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

One can prove the formula for expectation using integration, but observe since $f(x)$ is a constant, then the expectation should be the x-coordinate of the mean (to be more precise, arithmetic mean) of a and b.

We shall prove the formula for variance only.

Proof.

$$E(X^2) = \int_a^b \frac{x^2}{b-a}\,dx = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{b^3-a^3}{b-a} = a^2+ab+b^2$$

Hence,

$$\text{Var}(X) = a^2+ab+b^2 - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

□

Example 5.1 (ST2131 AY24/25 Sem 1 Lecture 9)**.** A point is chosen at random on a line segment of length 1, thus dividing the line segment into two pieces. What is the probability that the longer piece is at least four times as long as the shorter piece?

Solution. Suppose the shorter segment is of length x, so $x \le 1/5$. This means $1-x \ge 4/5$. Since the line segment is symmetrical, the desired probability is $2\cdot(1/5 \div 4/5) = 2/5$. □

Example 5.2 (ST2131 AY24/25 Sem 1 Lecture 10; triangle inequality)**.** Two points are chosen at random on a line segment of length 1, thus dividing the line segment into three pieces. What is the probability that we can form a triangle?

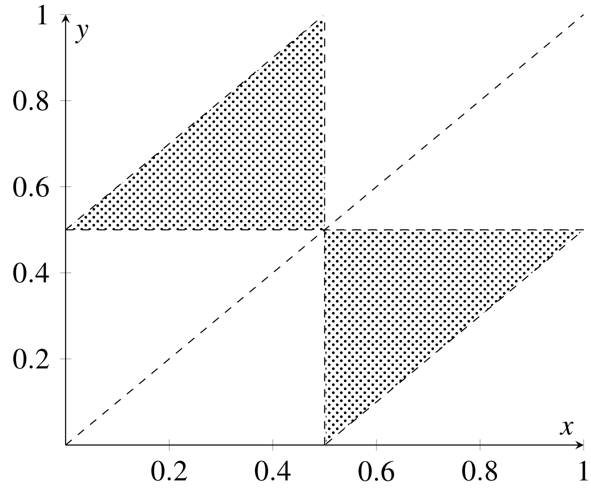
Solution. Suppose we cut the line segment at points x and y, where $x < y$. Then, we have three pieces of length $x, y, 1-y$. By the triangle inequality, we must have the following:

$$x + y - x > 1 - y \quad \text{and} \quad x + 1 - y > y - x \quad \text{and} \quad y - x + 1 - y > x$$

Upon simplification, we obtain

$$y > \frac{1}{2} \quad \text{and} \quad x < \frac{1}{2} \quad \text{and} \quad y > x + \frac{1}{2}.$$

By symmetry, suppose we had $y < x$ instead. Then, all the inequalities will still hold via substitution. By plotting a graph for both cases with y against x, we see that we can translate the problem to be such that if we pick a point $(x, y) \in [0, 1]^2$, what is the probability that it falls in the regions satisfying both inequalities?



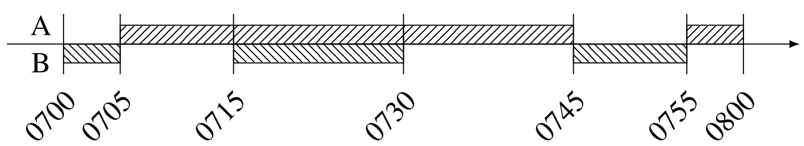
One can deduce that the required probability is $1/4$. □

Example 5.3 (ST2131 AY24/25 Sem 1 Lecture 10). Trains heading to destination A arrive at the station at 15 minutes interval starting at 7:00 am. Trains heading to destination B arrive at the station at 25 minutes interval starting at 7:05 am.

A man arrives at the station at a time uniformly distributed between 7:00 am and 8:00 am, and takes the first train that arrives. What is the probability that he goes to destination A?

A woman arrives at the station at a time uniformly distributed between 7:10 am and 8:10 am, and takes the first train that arrives. What is the probability that she goes to destination A?

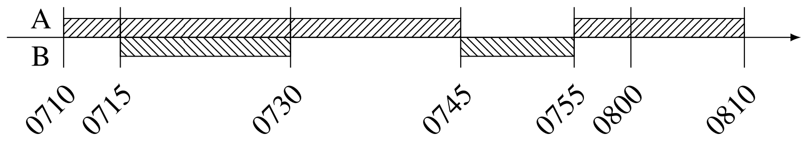
Solution. Consider the following diagram:



Then, the intervals that he goes to A is the intervals that he definitely goes to A, plus the interval that both A and B arrive at the same time (7.15–7.30), with half the probability of choosing the trains. That is,

$$P(\text{man goes to destination A}) = \frac{10 + 15 + 5}{60} + \frac{1}{2} \cdot \frac{15}{60} = \frac{1}{3} + \frac{1}{8}.$$

Similarly for the woman,



and we can calculate the probability in a similar fashion. We will get $25/60 + 1/8$. $\square$

To summarise,

Random Variable	PDF	Parameter(s)	$E(X)$	$\mathrm{Var}(X)$
$U(a,b)$	$\frac{1}{b-a}$	a and b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

5.3. *Normal Distribution*

Definition 5.5. A random variable X is normally distributed with parameters μ and σ, where μ is the mean and σ^2 is the variance, if its PDF is

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

where $x \in \mathbb{R}$. We say that $X \sim N(\mu, \sigma^2)$.

Even though the PDF formula looks very complicated, one can verify that the integral from $-\infty$ to ∞ is indeed 1 (i.e. sum of probabilities is 1). To the interested, this uses a well-known result, known as the Gaussian integral. I will attach the proof here, but one needs to have some pre-requisites regarding matrices and Multivariable Calculus to understand the proof.

Theorem 5.2 (Gaussian integral).

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}$$

Proof. Let $f(x,y)$ be a function defined on $R = [a,b] \times [c,d]$. The integral

$$\int_c^d f(x,y)\ dy$$

means that x is regarded as a constant and $f(x,y)$ is integrated with respect to y from $y = c$ to $y = d$. Thus, this integral is a function of x and we can integrate it with respect to x from $x = a$ to $x = b$. The resulting integral

$$\int_a^b \int_c^d f(x,y)\ dydx$$

is known as an *iterated integral*.

The Fubini theorem allows the order of integration to be changed in certain iterated integrals. It states that if $f(x,y)$ is *absolutely convergent* and continuous on $R = [a,b] \times [c,d]$, then

$$\iint_R f(x,y)\ dA = \int_a^b \int_c^d f(x,y)\ dydx = \int_c^d \int_a^b f(x,y)\ dxdy.$$

As mentioned earlier, for Fubini's theorem to be applied, f must be an absolutely convergent integral. Similar to the absolute convergence of series, if an integral is absolutely convergent, then

$$\int_R |f(x)| \, dx < \infty.$$

One of the ways to evaluate the famous Gaussian integral, which is

$$\int_{-\infty}^{\infty} e^{-x^2} \, dx = \sqrt{\pi},$$

involves Fubini's theorem.

We will use polar coordinates. Let I be the original integral. Then,

$$\begin{aligned} I &= \int_{-\infty}^{\infty} e^{-x^2} \, dx = \int_{-\infty}^{\infty} e^{-y^2} \, dy \\ I^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2} \, dx \right) \left(\int_{-\infty}^{\infty} e^{-y^2} \, dy \right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-x^2} e^{-y^2} \, dxdy \text{ by Fubini's Theorem} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\left(x^2+y^2\right)} \, dxdy \end{aligned}$$

We will do a change of variables from the Cartesian world to the polar world. We will establish the following result

$$dxdy = rdrd\theta$$

using the Jacobian of a suitable matrix. That is,

$$\mathbf{J} = \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix}.$$

Since $dxdy = \det(\mathbf{J}) \, drd\theta$, then the result follows. Hence, the integral can be transformed to

$$I^2 = \int_0^{2\pi} \int_0^{\infty} re^{-r^2} \, drd\theta = \pi.$$

We conclude that $I = \sqrt{\pi}$. □

The central limit theorem, or CLT, as well as the de Moivre-Laplace theorem, will be covered in due course. The central limit theorem should be no stranger to you if you still recall it from H2 Mathematics.

Proposition 5.3. The expectation and variance of a normal random variable $X \sim N(\mu, \sigma^2)$ are

$$E(X) = \mu \quad \text{and} \quad \operatorname{Var}(X) = \sigma^2.$$

One interesting property is that the mean, median and mode of a normal random variable are the same, which is μ.

To summarise,

Definition 5.6 (standard normal random variable)**.** A normal random variable is called a standard normal random variable when $\mu = 0$ and $\sigma = 1$. This is denoted by SZ. That is, $Z \sim N(0, 1)$. Its PDF and

Random Variable	PDF	Parameter(s)	$E(X)$	$\text{Var}(X)$
$N(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ and σ	μ	σ^2

CDF are usually denoted by ϕ and Φ respectively. That is,

$$\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \quad \text{and} \quad \Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,dt.$$

Proposition 5.4. Some properties of the standard normal distribution are as follows:

(i) $P(Z \geq 0) = P(Z \leq 0) = 0.5$ due to symmetry

(ii) $-Z \sim N(0,1)$

(iii) $P(Z \leq x) = 1 - P(Z > x)$ for $x \in \mathbb{R}$

(iv) $P(Z \leq -x) = P(Z \geq x)$ for $x \in \mathbb{R}$

(v) If $X \sim N(\mu,\sigma^2)$, then,

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

(vi) If $Z \sim N(0,1)$, then $X = aZ + b \sim N(b, a^2)$ for $a, b \in \mathbb{R}$

Example 5.4 (ST2131 AY21/22 Sem 1)**.** Let Z be a standard normal random variable. For any real number $a \in \mathbb{R}$, define X_a by

$$X_a := \begin{cases} Z & \text{if } Z > a; \\ 0 & \text{otherwise.} \end{cases}$$

Find $E(X_0)$ and $E(X_1)$.

Solution. Note that $X_0 = Z$ for $Z > 0$ and 0 otherwise. By definition of the probability density function of the standard normal random variable,

$$E(X_0) = \int_0^\infty x \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx \approx 0.40.$$

As for X_1, it is equal to Z for $Z > 1$ and 0 otherwise. In a similar fashion,

$$E(X_1) = \int_1^\infty x \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx \approx 0.24.$$

□

Example 5.5 (ST2132 AY18/19 Sem 2 Tutorial 2)**.** Let $X_1, \ldots, X_n$ be independent variables, with $E(X_i) = \mu_i$ and $\text{Var}(X_i) = \sigma_i^2$. For each i, let a_i and b_i be constants such that $Y_i = a_i X_i + b_i$ is standardised.

(a) Express a_i and b_i in terms of μ_i and σ_i.

(b) Are Y_i independent?

(c) Are Y_i identically distributed?

(d) Repeat **(c)**, if each X_i has a Bernoulli distribution.

(e) Repeat **(c)**, if each X_i has a normal distribution.

Solution.

(1). Since $E(Y_i) = 0$ and $\text{Var}(Y_i) = 1$, then $a_i\mu_i + b = 0$ and $a_i^2\sigma_i^2 = 1$, so $a_i = 1/\sigma_i$ and $b_i = -\mu_i/\sigma_i$. Here, we assume that $a_i > 0$.

(2). Yes, since the X_i's are independent.

(3). Unable to tell from the given information.

(4). $X_i \sim \text{Bernoulli}(p)$, so $E(X_i) = p_i$ and $\text{Var}(X_i) = p_i(1-p_i)$. Hence, $a_i = 1/p_iq_i$ and $b_i = -1/q_i$, where $q_i = 1-p_i$. We prove that the statement is not true in general. Suppose $X_1 \sim \text{Bernoulli}(0.5)$ and $X_2 \sim \text{Bernoulli}(0.1)$, so $Y_1 = 4X_1 - 2$ and $Y_2 = 90X_2 - 9$.

(5). Since Y_i is the standard normal random variable, then each of the Y_i's is identically distributed. □

Proposition 5.5 (68-95-99.7 rule)**.** The 68–95–99.7 rule, also known as the empirical rule, is a shorthand used to remember the percentage of values that lie within an interval estimate in a normal distribution: 68%, 95% and 99.7% of the values lie within one, two, and three standard deviations of the mean, respectively. That is, for a random variable X

$$
\begin{aligned}
P(\mu-\sigma \le X \le \mu+\sigma) &\approx 0.6827 \\
P(\mu-2\sigma \le X \le \mu+2\sigma) &\approx 0.9545 \\
P(\mu-3\sigma \le X \le \mu+3\sigma) &\approx 0.9973
\end{aligned}
$$

In the empirical sciences, the so-called three-sigma rule of thumb (or 3σ rule) expresses a conventional heuristic that nearly all values are taken to lie within three standard deviations of the mean, and thus it is empirically useful to treat 99.7% probability as near certainty.

Theorem 5.3 (de Moivre-Laplace theorem)**.** Suppose $X \sim B(n,p)$. Then, for any $a < b$, we have

$$P\left(a < \frac{X-np}{\sqrt{npq}} < b\right) \to \Phi(b) - \Phi(a)$$

as $n \to \infty$. That is, $B(n,p) \approx N(np,npq)$. Equivalently,

$$\frac{X-np}{\sqrt{npq}} \approx Z \quad \text{where } Z \sim N(0,1).$$

The normal approximation will generally be good for values of n satisfying $npq \ge 10$. The approximation is further improved if we incorporate continuity correction.

Proposition 5.6 (continuity correction)**.** If $X \sim B(n,p)$, then

$$
\begin{aligned}
P(X=k) &= P\left(k-\frac{1}{2} < X < k+\frac{1}{2}\right) \\
P(X \ge k) &= P\left(X \ge k-\frac{1}{2}\right) \\
P(X \le k) &= P\left(X \le k+\frac{1}{2}\right)
\end{aligned}
$$

5.4. *Exponential Distribution*

Definition 5.7 (exponential distribution). A random variable X is said to follow an exponential distribution with parameter $\lambda > 0$ if its PDF is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0; \\ 0 & \text{if } x < 0. \end{cases}$$

We say that $X \sim \text{Exp}(\lambda)$. For $x \in \mathbb{Z}_{\geq 0}$, the CDF is $F(x) = 1 - e^{-\lambda x}$.

Proposition 5.7. If $X \sim \text{Exp}(\lambda)$, then the expectation and variance are

$$E(X) = \frac{1}{\lambda} \quad \text{and} \quad \text{Var}(X) = \frac{1}{\lambda^2}.$$

Proposition 5.8 (median and exponential decay). If $T \sim \text{Exp}(\lambda)$, then the median m is $\ln 2/\lambda$.

Proof. This is easy to prove by considering the CDF formula. Substituting $t = m$, we have

$$\begin{aligned} F(m) &= \frac{1}{2} \\ e^{-\lambda m} &= \frac{1}{2} \\ m &= \frac{\ln 2}{\lambda} \end{aligned}$$

□

The expression $\ln 2/\lambda$ is of great significance. It is known as *half-life* and it plays an important role in the exponential decay of an object.

A quantity is subject to exponential decay if it decreases at a rate proportional to its current value. Symbolically, this process can be expressed by the following differential equation:

$$\frac{dN}{dt} = -\lambda N,$$

where N is the quantity and λ is a positive rate called the exponential decay constant. The solution to the equation is $N = N_0 e^{-\lambda t}$, where $N_0 = N(0)$ is the initial quantity at time $t = 0$.

Recall that the if a random variable X satisfies the memoryless property, then for $m, n \in \mathbb{N}$,

$$P(X > m + n \mid X > m) = P(X > n).$$

Previously, we claimed and proved that the geometric distribution is the only discrete random variable exhibiting the memoryless property. Here, we set $m, n \in \mathbb{R}^+$ since we are dealing with continuous random variables.

Proposition 5.9 (memorylessness). For the continuous counterpart, only the exponential distribution has the memoryless property.

We provide a proof for this statement.

Proof. We apply the definition of conditional probability to the left side. Hence,

$$P(X > m+n) = P(X > m)P(X > n).$$

We note that $P(X \leq x) = F(x)$ by definition of the CDF. Hence, the equation becomes

$$[1 - F(m+n)] = [1 - F(m)][1 - F(n)].$$

Using the substitution $G(x) = 1 - F(x)$ for all $x \in \mathbb{R}^+$, we have

$$G(m+n) = G(m)G(n),$$

which is a functional equation involving two variables. Setting $m = n = 0$ yields $G(0) = [G(0)]^2$, and so $G(0)[1 - G(0)] = 0$. Hence, $G(0) = 0$ or $G(0) = 1$.

By first principles,

$$\begin{aligned} G'(x) &= \lim_{\delta x \to 0} \frac{G(x+\delta x) - G(x)}{\delta x} \\ &= \lim_{\delta x \to 0} \frac{G(x)G(\delta x) - G(x)}{\delta x} \\ &= G(x) \lim_{\delta x \to 0} \frac{G(\delta x) - G(0)}{\delta x} \\ &= G(x)G'(0) \end{aligned}$$

Note that $G'(0)$ is a constant, say c, so we end up with a first-order separable differential equation, namely $G'(x) = cG(x)$. This is easy to solve. We get $G(x) = e^{cx+d}$, where c and d are both constants. By setting $A = e^d$, the solution is just

$$G(x) = Ae^{cx}.$$

Hence, $F(x) = 1 - Ae^{cx}$ and since $f(x)$ is the derivative of the CDF, then

$$f(x) = F'(x) = -Ace^{cx}.$$

By setting $c = -\lambda$ and $-Ac = \lambda$, we have $A = 1$, and the result follows. □

Example 5.6 (ST2131 AY24/25 Sem 1 Lecture 12)**.** Used cars are sold at a garage. The total lifetime mileage that a car from the garage can be drive before it breaks down is assumed to have an exponential distribution. You and I both bought a car from the garage.

Your car has been driven 100 thousand kilometres. My car has been driven 150 thousand kilometres. Which of the two cars is more likely to be driven for a longer distance before breaking down?

Solution. Let Y be the lifetime of your car and M be the lifetime of my car. Then

$$P(Y > t + 100 \mid Y > 100) \quad \text{and} \quad P(M > t + 150 \mid M > 150)$$

are probabilities of interest. Both probabilities are actually equal by the memoryless property of the exponential random variable. That is,

$$P(Y > t + 100 \mid Y > 100) = P(Y > t) = P(M > t) = P(M > t + 150 \mid M > 150).$$

So, both cars are equally likely to break down. Well, to further justify, let X be the lifetime of the car, which can be modelled as $X \sim \exp(\lambda)$. Define a random process

$$S(t) = \begin{cases} 1 & \text{if } X \leq t; \\ 0 & \text{otherwise} \end{cases} \qquad \text{which is a memoryless process.}$$

Consider $P(S(s) = 0 \mid S(t) = 0)$. This is equal to $P(S(s-t) = 0)$, i.e. when you arrive at point t, the process forgets the history and refreshes itself. This is known as the Markov property. □

Definition 5.8 (Poisson process). A homogeneous Poisson point process can be defined as a counting process, which can be denoted by $\{N(t), t \geq 0\}$. A counting process represents the total number of occurrences or events that have happened up to and including time t. A counting process is a homogeneous Poisson counting process with rate $\lambda > 0$ if it has the properties $N(0) = 0$, has independent increments and the number of events in any interval of length t is a Poisson random variable with parameter (or mean) λt.

We shall prove that if $N(t) \sim \text{Po}(\lambda t)$, then the inter-arrival time T, follows an exponential distribution with parameter λ. That is, $T \sim \text{Exp}(\lambda)$.

Proof. Note that

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

Hence, $P(T \leq t) = 1 - e^{-\lambda t}$, which implies that $f(t) = \lambda e^{-\lambda t}$. Therefore, $T \sim \text{Exp}(\lambda)$. □

In most cases, we usually denote an exponential random variable by T since it encompasses the essence of time.

Example 5.7 (ST2131 AY24/25 Sem 1 Lecture 12). When an MRT line breaks down, the time in hours until the resumption of operations is an exponentially distributed random variable with parameter $1/4$.

(a) What is the probability that more than four hours is needed to fix a broken down MRT line?

(b) One of the MRT lines has broken down five hours ago. What is the probability that it will get fixed within the next four hours?

Solution.

(a) Let X denote the number of hours required to fix the MRT line. Find $P(X > 4)$, which is given by

$$\int_4^\infty \frac{1}{4} e^{-x/4}\, dx = 1/e.$$

(b) This is finding $P(X \leq 4 + 5 \mid X \geq 5)$. Recall the memoryless property of the exponential variable, so

$$1 - P(X > 4 + 5 \mid X > 5) = 1 - P(X > 4) = 1 - \frac{1}{e}.$$

□

Theorem 5.4 (distribution of the minimum). Suppose $T_i \sim \text{Exp}(\lambda_i)$ for $1 \leq i \leq n$ and the T_i's are independent exponential random variables. We define W to be the minimum of all the T_i's and claim that W also follows an exponential distribution. That is,

$$W = \min\{T_1, T_2, \ldots, T_n\} \sim \text{Exp}\left(\sum_{i=1}^n \lambda_i\right).$$

Proof.

$$
\begin{aligned}
P(W \le t) &= 1 - P(W > t) \\
&= 1 - P(T_1 > t)P(T_2 > t)\dots P(T_n > t) \text{ since the } T_i\text{'s are independent} \\
&= 1 - e^{-\lambda_1 t} e^{-\lambda_2 t} \dots e^{-\lambda_n t} \\
&= 1 - e^{-(\lambda_1 + \lambda_2 + \dots + \lambda_n)} \\
&= 1 - \exp\left(-\sum_{i=1}^{n} \lambda_i t\right)
\end{aligned}
$$

Differentiating both sides yields

$$
f_W(t) = \left(\sum_{i=1}^{n} \lambda_i\right) \exp\left(-\sum_{i=1}^{n} \lambda_i t\right),
$$

asserting that our claim is true. □

Corollary 5.1. If $T_i \sim \text{Exp}(\lambda)$ for $1 \le i \le n$, and that all the T_i's are identically distributed, then

$$
W = \min\{T_1, T_2, \dots, T_n\} \sim \text{Exp}(n\lambda_i).
$$

We then introduce the inverse transform sampling method. The probability integral transform states that if X is a continuous random variable with cumulative distribution function F_X, then the random variable $Y = F(X)$ has a uniform distribution on $(0,1)$. The inverse probability integral transform is just the inverse of this. To be specific, we have the following result:

Theorem 5.5 (inverse transform sampling)**.** If $Y \sim U(0,1)$ and if X has a cumulative distribution F_X, then the random variable $F_X^{-1}(Y)$ has the same distribution as X.

Proof. First, note that the CDF is an increasing function so from the first step to the second step, the inequality sign will not change.

$$
\begin{aligned}
P\left[F^{-1}(Y) \le x\right] &= P\left[Y \le F(x)\right] \text{ by applying } F \text{ to both sides} \\
&= F(x) \text{ since } Y \text{ is uniform on } (0,1)
\end{aligned}
$$

□

Example 5.8 (ST2131 AY19/20 Sem 2)**.** Let X be an exponential random variable with mean 1. Find the probability density function of $Y = 1/X^2$.

Solution. We have $P(X \le x) = e^{-x}$ for $x \ge 0$ by definition of the exponential distribution. Thus,

$$\begin{aligned} P(Y \le y) &= P\left(\frac{1}{X^2} \le y\right) \\ &= P\left(\frac{1}{y} \le X^2\right) \\ &= P\left(X \le -\frac{1}{\sqrt{y}} \text{ or } X \ge \frac{1}{\sqrt{y}}\right) \\ &= P\left(X \le -\frac{1}{\sqrt{y}}\right) + P\left(X \ge \frac{1}{\sqrt{y}}\right) \\ &= 0 + \int_{1/\sqrt{y}}^{\infty} e^{-x}\, dx \\ &= \exp\left(-\frac{1}{\sqrt{y}}\right) \end{aligned}$$

Differentiating $P(Y \le y)$ with respect to y yields $f_Y(y)$, which is the probability density function of Y, so

$$f(y) = \frac{\exp\left(-1/\sqrt{y}\right)}{2y^{3/2}}.$$

Next, we find the support of Y. Since X is defined for $x \ge 0$, then Y is defined for $y \ge 0$. To conclude,

$$f(y) = \begin{cases} \dfrac{\exp\left(-1/\sqrt{y}\right)}{2y^{3/2}} & \text{if } y \ge 0; \\ 0 & \text{otherwise.} \end{cases}$$

□

Previously in Theorem 5.4, we talked about the distribution of the minimum of independent random variables. In general, we have the following result:

Theorem 5.6 (distribution of the maximum and minimum). Assume that $X_1, X_2, \ldots, X_n$ are independent random variables with common CDF F and PDF f. Let

$$U = \max\{X_1, \ldots, X_n\} \quad \text{and} \quad V = \min\{X_1, \ldots, X_n\}.$$

The CDF of U is

$$F_U(u) = P(U \le u) = \prod_{i=1}^{n} P(X_i \le u) = [F(u)]^n,$$

and the PDF of U is

$$f_U(u) = nf(u)[F(u)]^{n-1}.$$

Similarly, the CDF of V is

$$F_V(v) = 1 - [1 - F(v)]^n,$$

and the PDF of V is

$$f_V(v) = nf(v)[1 - F(v)]^{n-1}.$$

The results in Theorem 5.6 are easy to established. In particular, the respective PDFs can be easily derived by differentiating the CDF and we make use of $F' = f$. Also, since V is the minimum, $V \ge v$ if and only if for all $1 \le i \le n$, $X_i \ge v$.

To summarise,

Random Variable	PDF	Parameter(s)	$E(X)$	Var(X)
Exp(λ)	$\lambda e^{-\lambda x}$	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Definition 5.9 (Laplace distribution). The definition of the Laplace distribution is

$$f(x) = \frac{1}{2}\lambda e^{-\lambda |x|}.$$

Realise that the Laplace distribution is a natural extension of the exponential distribution.

Example 5.9 (ST2131 AY24/25 Sem 1 Lecture 12). The random variable X follows the Laplace distribution with parameter $1/\pi$. Find $P(X > \pi), P(-1 < X < 2), E(X), \mathrm{Var}(X)$.

Solution. So

$$P(X > \pi) = \int_{\pi}^{\infty} \frac{1}{2} \cdot \frac{1}{\pi} e^{-x/\pi}\, dx = \frac{1}{2e}.$$

For $P(-1 < X < 2)$, we find

$$P(-1 < X < 2) = \int_{-1}^{0} \frac{1}{2} \cdot \frac{1}{\pi} e^{-x/\pi}\, dx + \int_{0}^{2} \frac{1}{2} \cdot \frac{1}{\pi} e^{-x/\pi}\, dx.$$

Observe that this is a symmetric function, so we can compute

$$P(-1 < X < 2) = \int_{0}^{1} \frac{1}{2} \cdot \frac{1}{\pi} e^{-x/\pi}\, dx + \int_{0}^{2} \frac{1}{2} \cdot \frac{1}{\pi} e^{-x/\pi}\, dx \approx 0.37.$$

For $E(X)$, note that this is a symmetric distribution. So, $E(X) = 0$. Lastly, the variance is

$$\mathrm{Var}(X) = E\left(X^2\right) - [E(X)]^2 = \frac{1}{2\pi} \int_{-\infty}^{0} x^2 e^{-x/\pi}\, dx + \frac{1}{2\pi} \int_{0}^{\infty} x^2 e^{-x/\pi}\, dx = 2\pi^2.$$

Here, the evaluation of each integral is quite simple — use integration by parts. □

5.5. *Gamma Distribution*

Definition 5.10 (gamma distribution). A random variable X is said to follow a gamma distribution with parameters α and λ, and is denoted by $X \sim \Gamma(\alpha, \lambda)$. The PDF only exists for $x \geq 0$ and its formula is

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

where $\alpha, \lambda > 0$ and $\Gamma(\alpha)$, called the gamma function, is defined by

$$\Gamma(\alpha) = \int_{0}^{\infty} e^{-t} t^{\alpha - 1}\, dt.$$

It is easy to prove that $\Gamma(1) = 1$ and that the gamma function satisfies the following recurrence relation:

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$$

Proof. Use integration by parts. □

Hence, it is easy to establish that for integer values of α, say $\alpha = n$, we have $\Gamma(n) = (n-1)!$.

Observe that $\Gamma(1, \lambda) = \mathrm{Exp}(\lambda)$, which implies that the exponential distribution is a special case of the gamma distribution.

Lemma 5.1. A very interesting result states that

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty e^{-t} t^{-\frac{1}{2}}\, dt = \sqrt{\pi}.$$

Proof. Using the substitution $u = \sqrt{t}$, we have

$$\int_0^\infty e^{-t} t^{-\frac{1}{2}}\, dt = \int_{-\infty}^\infty e^{-u^2}\, du$$

This follows from the Gaussian integral (Theorem 5.2). □

Proposition 5.10. If $X \sim \Gamma(\alpha, \lambda)$, then the expectation and variance are

$$E(X) = \frac{\alpha}{\lambda} \quad \text{and} \quad \mathrm{Var}(X) = \frac{\alpha}{\lambda^2}.$$

To summarise,

Random Variable	PDF	Parameter(s)	$E(X)$	$\mathrm{Var}(X)$
$\Gamma(\alpha,\lambda)$	$\frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}$	α and λ	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$

Similar to the Poisson process, we have a similar result, known as a gamma process.

Theorem 5.7 (gamma process). If events are occurring randomly and in accordance with the axioms required for a situation to be modelled by a Poisson process, then the amount of time one has to wait until a total of n events has occurred will be a gamma random variable with parameters (n, λ).

Proof. Let T_n denote the time at which the n^{th} event occurs, and $N(t)$ equal to the number of events in $[0,t]$. Note that $N(t) \sim \mathrm{Po}(\lambda t)$. Hence, $\{T_n \leq t\} = \{N(t) \geq n\}$. Therefore,

$$P(T_n \leq t) = P(N(t) \geq n) = \sum_{j=n}^\infty P(N(t) = j) = \sum_{j=n}^\infty \frac{e^{-\lambda t}(\lambda t)^j}{j!}.$$

To get the PDF of T_n, we differentiate both sides with respect to t. This should be straightforward and will be left as an exercise. □

5.6. *Beta Distribution*

Definition 5.11 (beta distribution). A random variable X is said to follow a beta distribution with parameters (a,b), denoted by $X \sim \mathrm{Beta}(a,b)$, if its PDF is

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

where the support of x is $0 < x < 1$. The expression $B(a,b)$ is known as the beta function, where

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

Lemma 5.2 (relationship with gamma function).

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

Proof. We first consider $\Gamma(a)\Gamma(b)$ and write it as an integral. Then,

$$\Gamma(a)\Gamma(b) = \left(\int_0^\infty e^{-u}u^{a-1}\,du\right)\left(\int_0^\infty e^{-v}v^{b-1}\,dv\right) = \int_0^\infty \int_0^\infty e^{-(u+v)}u^{a-1}v^{b-1}\,dudv.$$

We use the change of variables $u = zt$ and $v = z(1-t)$. Hence, $v = -z(t-1)$. Recall that $u, v \geq 0$, which implies that $0 \leq t \leq 1$ and $z \geq 0$. Upon change of variables, we have

$$\begin{aligned}\Gamma(a)\Gamma(b) &= \int_0^\infty \int_0^1 e^{-z}(zt)^{a-1}(z(1-t))^{b-1}z\,dtdz \\ &= \left(\int_0^\infty e^{-z}z^{a+b-1}\,dz\right)\left(\int_0^1 t^{a-1}(1-t)^{b-1}\,dt\right) \\ &= \Gamma(a+b)B(a,b)\end{aligned}$$

which asserts that the statement is true. □

Proposition 5.11. If $X \sim \text{Beta}(a,b)$, then

$$E(X) = \frac{a}{a+b} \quad \text{and} \quad \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

We shall prove the formula for expectation only.

Proof. It is clear that

$$E(X) = \frac{1}{B(a,b)}\int_0^1 x^a(1-x)^{b-1}\,dx.$$

By definition of the beta function and using the relationship between the beta function and the gamma function, we can rewrite the above integral as

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

□

To summarise,

Random Variable	PDF	Parameter(s)	$E(X)$	$\text{Var}(X)$
$\text{Beta}(a,b)$	$\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$	a and b	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$

5.7. *Cauchy Distribution*

Definition 5.12 (Cauchy distribution). A random variable X is said to follow a Cauchy distribution with parameter θ, where $\theta \in \mathbb{R}$, denoted by $X \sim \text{Cauchy}(\theta)$, if its PDF is

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

It is also the distribution of the ratio of two independent normally distributed random variables with mean zero. Interestingly, the expectation and variance of a Cauchy random variable do not exist!

To summarise,

Random Variable	PDF	Parameter(s)	$E(X)$	$\mathrm{Var}(X)$
$U(a,b)$	$\frac{1}{b-a}$	a and b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$N(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ and σ	μ	σ^2
$\mathrm{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$\Gamma(\alpha,\lambda)$	$\frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}$	α and λ	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
$\mathrm{Beta}(a,b)$	$\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$	a and b	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$
$\mathrm{Cauchy}(\theta)$	$\frac{1}{\pi}\cdot\frac{1}{1+(x-\theta)^2}$	θ		

5.8. *Order Statistics*

Suppose we have a sample of n random variables, $X_1,X_2,\ldots,X_n$, drawn from some distribution. To study the ordered values, we sort the sample in increasing order as follows:

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)}$$

Here, $X_{(1)},X_{(2)},\ldots,X_{(n)}$ are called the order statistics.

- $X_{(1)}$ is the smallest value in the sample, also known as the minimum
- $X_{(n)}$ is the largest value in the sample, also known as the maximum
- $X_{(i)}$ is the i^{th} smallest value in the sample (also called the i^{th} order statistic)

The notation $X_{(i)} < X_{(j)}$ for $1 \le i < j \le n$ indicates that the order statistics are arranged in strictly increasing order. The index (i) specifies the position in the ordered sequence, not the original order of X_i.

Theorem 5.8. The CDF of $X_{(r)}$ (where r specifies the order statistic) is

$$F_{X_{(r)}}(x) = \sum_{j=r}^{n}\binom{n}{j}[F_X(x)]^j[1-F_X(x)]^{n-j}$$

and the corresponding PDF is

$$f_{X_{(r)}}(x) = \frac{n!}{(r-1)!(n-r)!}f_X(x)[F_X(x)]^{r-1}[1-F_X(x)]^{n-r}.$$

Example 5.10. Suppose $X_1,X_2,\ldots,X_n$ are independent and identically distributed random variables, each following a uniform distribution on $[0,1]$. We are interested in the second smallest value, $X_{(2)}$, in a sample of size $n=3$. Obtain its CDF.

Solution. In fact, the second smallest value is the median. We have $F_{X_{(2)}}(x) = P\left(X_{(2)} \le x\right)$. This means that at least two of the random variables $X_{(1)},X_{(2)},X_{(3)}$ have values $\le x$. We shall consider two cases.

- **Case 1:** Suppose two random variables have values $\le x$. Then, the contribution of this event is

$$\binom{3}{2}\left[F_X(x)\right]^2\left[1-F_X(x)\right].$$

- **Case 2:** Suppose all three random variables have values $\leq x$. The contribution of this event is

$$\binom{3}{3}\left[F_X(x)\right]^3.$$

The desired CDF is the sum of contributions of the two cases, so,

$$F_{X_{(2)}}(x) = \binom{3}{2}\left[F_X(x)\right]^2\left[1 - F_X(x)\right] + \binom{3}{3}\left[F_X(x)\right]^3.$$

$\square$

6. Joint Probability Distribution

6.1. *Joint Distribution Functions*

Definition 6.1 (joint distribution)**.** For any two random variables X and Y defined on the same sample space, we define the joint distribution function of X and Y by

$$F_{X,Y}(x,y) = P(X \le x, Y \le y) \quad \text{for } x,y \in \mathbb{R}.$$

Note that $\{X \le x, Y \le y\}$ is equivalently $\{X \le x\} \cap \{Y \le y\}$.

Definition 6.2 (marginal distribution)**.** The distribution function of X can be obtained from the joint density function of X and Y via

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y) \quad \text{where} \quad F_X \text{ is the marginal distribution of } X.$$

Similarly,

$$F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x,y) \quad \text{where} \quad F_Y \text{ is the marginal distribution of } Y.$$

Proposition 6.1. We present two formulae which are useful in some calculations. Let a,b be real numbers, where $a_1 < a_2$ and $b_1 < b_2$. Then, the following hold:

(i)

$$P(X > a, Y > b) = 1 - F_X(a) - F_Y(b) + F_{X,Y}(a,b)$$

(ii)

$$P(a_1 < X \le a_2, b_1 < Y \le b_2) = F_{X,Y}(a_2,b_2) - F_{X,Y}(a_1,b_2) + F_{X,Y}(a_1,b_1) - F_{X,Y}(a_2,b_1)$$

We will only prove **(i)**.

Proof. We set $A = \{X \le a\}$ and $B = \{Y \le b\}$. Then, the required event is $A' \cap B'$, which is the same as $(A \cap B)'$, and by considering the complement of it, it is equivalently $n(S) - (A \cap B)$. By the principle of inclusion and exclusion, the required probability is $1 - P(A \cup B)$. Hence,

$$\begin{aligned} P(X > a, Y > b) &= 1 - P(A \cup B) \\ &= 1 - P(A) - P(B) + P(A \cap B) \\ &= 1 - F_X(a) - F_Y(b) + F_{X,Y}(a,b) \end{aligned}$$

which concludes the proof. □

Definition 6.3 (joint density function)**.** In the case where X and Y are discrete random variables, the joint probability density function of X and Y is

$$p_{X,Y}(x,y) = P(X = x, Y = y).$$

We can recover the probability density function of X and Y using

$$p_X(x) = P(X = x) = \sum_{y\in\mathbb{R}} p_{X,Y}(x,y) \quad \text{and} \quad p_Y(y) = P(Y = y) = \sum_{x\in\mathbb{R}} p_{X,Y}(x,y).$$

p_x and p_y are the marginal probability density function of X and Y respectively.

Example 6.1. 3 balls are randomly selected from an urn containing 3 red, 4 white and 5 blue balls. If we let R and W denote the number of red and white balls chosen respectively, then we can construct a joint probability density function table of R and W. It is shown below.

Solution.

white (right); red (bottom)	0	1	2	3	$P(R=r)$
0	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$	$\frac{84}{220}$
1	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$	0	$\frac{108}{220}$
2	$\frac{15}{220}$	$\frac{12}{220}$	0	0	$\frac{27}{220}$
3	$\frac{1}{220}$	0	0	0	$\frac{1}{220}$
$P(W=w)$	$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$	

It should be clear as to how these probabilities are computed. □

Example 6.2 (ST2132 AY18/19 Sem 2 Tutorial 2). Let X_1 be uniformly distributed on $\{0,1,2\}$. Given that $X_1 = 0$, let X_2 be 0, 1 or 2 with probabilities 1/2, 1/4 and 1/4 respectively. Given that $X_1 = 1$, let X_2 be 0, 1 or 2 with probabilities 1/2, 1/4 and 1/4 respectively. Given that $X_1 = 2$, let X_2 be 0, 1 or 2 with probabilities 0, 1/2 and 1/2.

(a) Find the joint distribution of (X_1,X_2), displaying the probabilities in a table, with the realisations of X_1 as rows.

(b) Display the joint distribution of (X_2,X_1) in another table with realisations of X_2 as rows.

(c) What is the relationship between the two tables?

(d) True or false, and explain: (X_1,X_2) and (X_2,X_1) have the same joint distribution.

Solution. Nothing special about this question. For **(c)**, the tables are transposes of each other. For **(d)**, as the tables are not identical, the answer is false. □

Proposition 6.2. Some useful formulae are as follows:

(i)

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = \sum_{a_1<x\leq a_2} \sum_{b_1<y\leq b_2} p_{X,Y}(x,y)$$

(ii)

$$F_{X,Y}(a,b) = P(X \leq a, Y \leq b) = \sum_{x\leq a} \sum_{y\leq b} p_{X,Y}(x,y)$$

(iii)

$$P(X > a, Y > b) = \sum_{x>a} \sum_{y>b} p_{X,Y}(x,y)$$

Definition 6.4 (jointly density function). We say that X and Y are jointly continuous random variables if there exists a function, denoted by $f_{X,Y}$ and known as the joint probability density function of X and Y if for every set $C \subseteq \mathbb{R}^2$, we have

$$P((X,Y) \in C) = \int\int_{(x,y\in C)} f_{X,Y}(x,y)\, dxdy.$$

Proposition 6.3. We state some useful formulae.

(i) Let $A, B \subseteq \mathbb{R}$. Set $C = A \times B$ (i.e. C is the Cartesian product of A and B). Then,

$$P(X \in A, Y \in B) = \int_A \int_B f_{X,Y}(x,y)\, dydx.$$

(ii) In particular, we can set $a_1, a_2, b_1, b_2 \in \mathbb{R}$, where $a_1 < a_2$ and $b_1 < b_2$, and so

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x,y)\, dydx.$$

(iii) Let $a, b \in \mathbb{R}$. Then,

$$F_{X,Y}(a,b) = P(X \leq a, Y \leq b) = \int_{\infty}^{a} \int_{\infty}^{b} f_{X,Y}(x,y)\, dydx.$$

Hence,

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

Definition 6.5 (marginal density function)**.** The marginal probability density function of X is

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

Similarly, the marginal probability density function of Y is

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

Example 6.3. The joint probability density function of X and Y is

$$f_{X,Y}(x,y) = \begin{cases} 2e^{-x}e^{-2y} & \text{if } x, y > 0; \\ 0 & \text{otherwise.} \end{cases}$$

Suppose we wish to compute the following probabilities:

(i) $P(X > 1, Y < 1)$

(ii) $P(X < Y)$

(iii) the marginal probability density function of X

(iv) $P(X \leq x)$

(v) the marginal distribution function of Y

Solution.

(i) This probability can be expressed by the following integral:

$$\int_1^{\infty} \int_0^1 2e^{-x}e^{-2y}\, dydx$$

and this is a very simple problem in Multivariable Calculus. The answer is $e^{-1}(1 - e^{-2})$.

(ii) As $0 < x < y$ and $0 < y < \infty$, the required probability is

$$\int_0^{\infty} \int_0^{y} 2e^{-x}e^{-2y}\, dxdy.$$

The answer is $1/3$. I omit the integration process because it is simple. I believe the only issue readers might have is setting up the double integral. We have an alternative representation for it. That is, we set

$x < y < \infty$ and $0 < x < \infty$. Hence, the integral is just

$$\int_0^\infty \int_x^\infty 2e^{-x}e^{-2y}\, dydx = \frac{1}{3}.$$

It yields the same conclusion as before!

(iii) Recall that the formula for the marginal probability density function of X is

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

Substituting everything in yields

$$f_X(x) = \int_{-\infty}^{\infty} 2e^{-x}e^{-2y}\, dy = e^{-x}.$$

Hence, for $x > 0$, the marginal probability density function is $f_X(x) = e^{-x}$.

(iv) Note that $P(X \le x)$ is the marginal distribution function of x, so

$$F_X(x) = \int_{\infty}^{x} e^{-t}\, dt = 1 - e^{-x} \quad \text{where} x > 0.$$

(v) The marginal distribution function of Y, for $y > 0$, is $F_Y(y) = 1 - e^{-2y}$. □

6.2. *Independent Random Variables*

Definition 6.6 (independent random variables)**.** Two random variables X and Y are independent if

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

for any $A, B \subseteq \mathbb{R}$. Random variables that are not independent are dependent.

Proposition 6.4. For jointly discrete random variables, we have three equivalent statements:

(i) X and Y are independent

(ii) For all $x, y \in \mathbb{R}$, $p_{X,Y}(x,y) = p_X(x)p_Y(y)$

(iii) For all $x, y \in \mathbb{R}$, $F_{X,Y}(x,y) = F_X(x)F_Y(y)$

For jointly continuous random variables, we also have three equivalent statements.

(i) X and Y are independent

(ii) For all $x, y \in \mathbb{R}$, $f_{X,Y}(x,y) = f_X(x)f_Y(y)$

(iii) For all $x, y \in \mathbb{R}$, $F_{X,Y}(x,y) = F_X(x)F_Y(y)$

For both discrete and continuous random variables, X and Y are independent if and only if

$$\text{there exist functions } g, h : \mathbb{R} \to \mathbb{R} \quad \text{such that} \quad \text{for all } x, y \in \mathbb{R}, \quad f_{X,Y}(x,y) = g(x)h(y).$$

In many applications, we either know or assume that X and Y are independent. Then, the joint probability density function of X and Y can be obtained by multiplying the individual probability density functions.

Independence is a *symmetric relation*. To say that X is independent of Y is equivalent to saying that Y is independent of X, or simply saying that X and Y are independent. In considering whether X is independent of Y in situations where it is not at all intuitive that knowing the value of Y will not change the probabilities concerning X, it can be beneficial to interchange the roles of X and Y and ask instead whether Y is independent of X.

Example 6.4 (Buffon's needle problem)**.** A table is ruled with equidistant parallel lines with distance D apart from one another. A needle of length L, where $L \leq D$, is randomly thrown onto the table. The Buffon's needle problem asks for the probability that the needle will intersect one of the lines.

Solution. The answer is a surprising

$$\frac{2L}{\pi D}.$$

This shows that when $L \approx D$, we can find a good estimate of the value of π. However, the approximation is not powerful until we toss the needle over 3400 times, which allows us to get the value of π to 6 decimal places.

We determine the position of the needle by specifying the distance X from the midpoint of the needle to the nearest parallel line, and the angle θ between the needle and the projected line of length X. The needle will intersect a line if the hypotenuse of the right triangle is less than $L/2$. That is,

$$\frac{X}{\cos\theta} < \frac{L}{2} \quad \text{which implies} \quad X < \frac{L}{2}\cos\theta.$$

As X varies between 0 and $D/2$ and θ between 0 and $\pi/2$, it is reasonable to assume that they are independent and uniformly distributed random variables over these respective ranges. Note that $D = L\cos\theta$, and for $0 \leq x \leq D/2$, $f_X(x) = 2/x$ and for $0 \leq \theta \leq \pi/2$, $f_\theta(\theta) = 2/\pi$. We thus obtain the joint probability density function

$$f_X(x)f_\theta(\theta) = \frac{4}{\pi D}$$

for $0 \leq x \leq D/2, 0 \leq \theta \leq \pi/2$ and 0 elsewhere.

Hence,

$$P\left(X < \frac{L}{2}\cos\theta\right) = \int_0^{\frac{\pi}{2}} \int_0^{\frac{L}{2}\cos\theta} \frac{4}{\pi D}\, dx d\theta = \frac{2L}{\pi D}.$$

□

Example 6.5 (ST2131 AY24/25 Sem 1 Lecture 14; Buffon's needle problem)**.** A table is ruled with equidistant parallel lines a distance $\sqrt{3}$ cm apart. A needle of length 2 cm is randomly thrown on the table. What is the probability that the needle will intersect with (at least) one of the lines?

Solution. Let X denote the minimum distance between the center of the needle and the ruled lines. Then, $X \sim U\left(0, \sqrt{3}/2\right)$. Let θ denote the acute angle between the needle and the lines. Then, $\theta \sim U(0, \pi/2)$.

For the needle to intersect with one of the lines, we must have $\sin\theta > X$. We then now find the area of

$$\left\{(X,\theta) \in \left[0, \sqrt{3}/2\right] \times [0, \pi/2] : \sin\theta > X\right\}.$$

Note that X is bounded by $\sqrt{3}/2$, so we would have to be careful in carrying out the integration. The probability that the needle intersects with the ruled lines is given by the ratio of the feasible area over the total area.

The area of the feasible region is

$$\frac{\pi}{2} \times \frac{\sqrt{3}}{2} - \int_0^{\sqrt{3}/2} \sin^{-1}(x)\, dx = 0.4069$$

and take this divided by the area of the rectangle to give 70%. □

Very often, we are interested in the sums of independent random variables. For example, when two dice are rolled, we are interested in the sum of the two numbers.

Proposition 6.5. Suppose we have two independent random variables X and Y. Then, for $x, y \in \mathbb{R}$,

$$f_{X,Y}(x,y) = f_X(x) f_Y(y).$$

It follows that

$$F_{X+Y}(x) = \int_{-\infty}^{\infty} F_X(x-t) f_y(t) \, dt.$$

Proof. We have

$$F_{X+Y}(x) = P(X+Y \le x) = \iint_{s+t \le x} f_{X,Y}(s,t) \, dsdt$$

which simplifies to

$$\int_{-\infty}^{\infty} \int_{-\infty}^{x-t} f_X(s) f_Y(t) \, dsdt = \int_{-\infty}^{\infty} F_X(x-t) f_y(t) \, dt.$$

□

Similarly,

$$F_{X+Y}(x) = \int_{-\infty}^{\infty} F_Y(x-t) f_x(t) \, dt.$$

By differentiation, it can be shown that

$$f_{X+Y}(x) = \int_{-\infty}^{\infty} f_X(x-t) f_Y(t) \, dt = \int_{-\infty}^{\infty} f_X(t) f_Y(x-t) \, dt.$$

Example 6.6. Recall that the sum of two independent uniform distributions follows a triangular distribution. Let us prove this result! Suppose X and Y are independent random variables with a common uniform distribution over $(0,1)$. That is, $X \sim U(0,1)$ and $Y \sim U(0,1)$. We wish to find the probability density function of $X+Y$.

Solution. $X+Y$ takes values in $(0,2)$. For $x \le 0$ and $x \ge 2$, it follows that $f_{X+Y}(x) = 0$. For $0 < x < 2$,

$$f_{X+Y}(x) = \int_{-\infty}^{\infty} f_X(x-t) f_Y(t) \, dt = \int_0^1 f_X(x-t) \, dt$$

$f_X(x-t) > 0$ if and only if $0 < x-t < 1$. Note that x is fixed and t varies. We split this into two cases, namely $0 < x \le 1$ and $1 < x < 2$.

For $0 < x \le 1$,

$$\begin{aligned} f_{X+Y}(x) &= \int_0^1 f_X(x-t) \, dt + \int_x^1 f_X(x-t) \, dt \\ &= \int_0^x f_X(x-t) \, dt \\ &= \int_0^x dt \\ &= x \end{aligned}$$

In a similar fashion, it can be shown that for $1 < x < 2$,

$$f_{X+Y}(x) = 2 - x.$$

Hence,

$$f_{X+Y} = \begin{cases} x & \text{if } 0 < x \le 1; \\ 2-x & \text{if } 1 < x < 2; \\ 0 & \text{otherwise.} \end{cases}$$

The density function has the shape of a triangle, so $X+Y$ follows a triangular distribution. □

Example 6.7 (ST2131 AY24/25 Sem 1 Lecture 14). A man and a woman agreed to meet at the location at 12 pm. The man arrives at the location at the time uniformly distributed between 11:45 am and 12:15 pm. The woman arrives at the location at a time uniform distributed between 12 pm and 12:30 pm.

(a) What is the probability that the first person to arrive waits less than 5 minutes for the second person?

(b) What is the probability that the man arrives first?

Solution.

(a) Let X and Y be the number of minutes the man and woman arrive with respect to 12 pm.

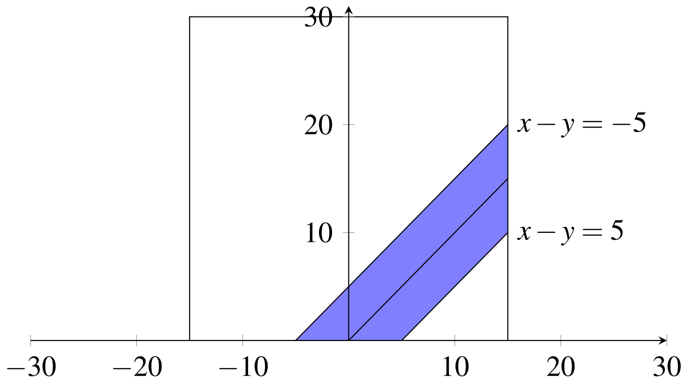
We have

$$X \sim U(-15, 15) \quad \text{and} \quad Y \sim Y(0, 30).$$

It suffices to find

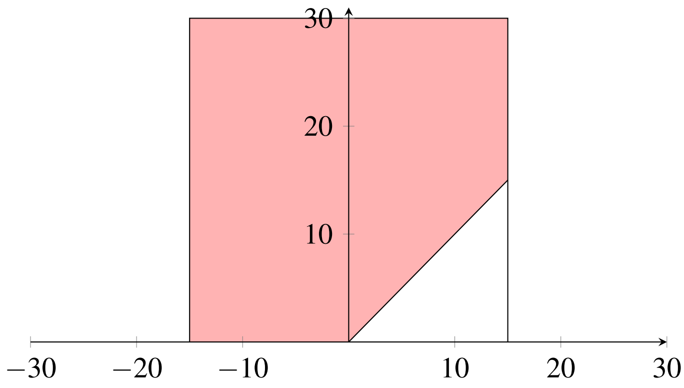
$$P(|X - Y| < 5) \quad \text{or equivalently} \quad P(-5 < Y - X < 5).$$

Let us construct a diagram as follows:



Plot Y against X, and we use this to find the area bounded between $y = 5 + x$ and $y = x - 5$ in the rectangle $[-15, 15] \times [0, 30]$, and divide it by the area of the rectangle. Computation yields us 17%.

(b) For this, we are finding $P(X < Y)$. The diagram is given as follows:



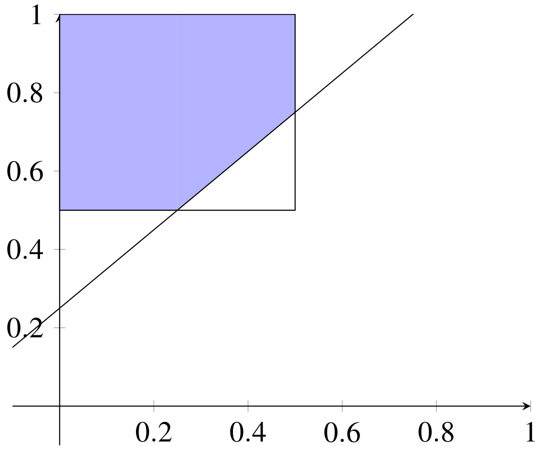
With this,

$$P(Y > X) = \frac{\text{shaded area}}{\text{total area}} = 1 - \frac{\frac{1}{2} \cdot 15^2}{30^2} = 0.88.$$

□

Example 6.8 (ST2131 AY24/25 Sem 1 Lecture 14). One point is randomly selected on the interval $[0, 1/2]$. Another point is randomly selected on the interval $[1/2, 1]$. What is the probability that the distance between the two points is greater than $1/4$?

Solution. Let $X \sim U(0,1/2)$ and $Y \sim U(1/2,1)$ be the two points selected. It suffices to find $P(Y > X + 1/4)$. Let us plot the desired region.



We find that this shaded area is $7/8$. □

Example 6.9 (ST2131 AY24/25 Sem 1 Lecture 14). Three points X, Y, Z are selected independently at random from the interval $[0,1]$. What is the probability that Y lies between X and Z?

Solution. There are 2 permutations where Y is between X and Z. The total number of permutations is $3! = 6$. So, the desired answer is $2/6 = 1/3$. □

Example 6.10 (ST2131 AY21/22 Sem 2). Three numbers A, B, C are selected independently at random from the unit interval $[0,1]$. What is the probability that both roots of the equation $Ax^2 + Bx + C = 0$ are real?

Solution. For the roots to be real, $B^2 - 4AC \geq 0$. By the total law of probability,

$$P(B^2 \geq 4AC) = \int_0^1 P(B^2 \geq 4AC | C = c) f(c)\, dc.$$

Due to independence, the above can be written as

$$\int_0^1 P(B^2 \geq 4Ac)\, dc.$$

Since $b^2 \in [0,1]$, we consider two cases, namely when $c \in (0,1/4)$ and $c \in [1/4,1)$.

For $c \in (0,1/4)$,

$$P(B^2 \geq 4Ac) = \int_0^1 \int_{\sqrt{4ac}}^1 dbda = 1 - \frac{4}{3}\sqrt{c}.$$

For $c \in [1/4,1)$,

$$P(B^2 \geq 4Ac) = \int_0^1 \int_0^{b^2/4c} dadb = \frac{1}{12c}.$$

Putting everything together,

$$P(B^2 \geq 4AC) = \int_0^{1/4} 1 - \frac{4}{3}\sqrt{c}\, dc + \int_{1/4}^1 \frac{1}{12c}\, dc = \frac{5 + 3\ln 4}{36}.$$

□

Example 6.11 (ST2131 AY21/22 Sem 2). Let X and Y be independent random variables uniformly distributed on the unit interval $[0,1]$. Find

(a) $P(-0.5 < 3X - 2Y < 0.5)$

(b) $P(0 < 3X - 2Y < 2.5)$

Solution.

(a) The probability is $P(-0.5 + 2Y < 3X < 0.5 + 2Y)$, which is

$$\begin{aligned}\iint_{-0.5+2y<3x<0.5+2y} f(x,y)\ dxdy &= \iint_{-0.5+2y<3x<0.5+2y} f_X(x) f_Y(y)\ dxdy \\ &= \int_0^1 \int_{(2y-0.5)/3}^{(2y+0.5)/3} \left(\frac{1}{1}\right)^2\ dxdy \\ &= \frac{1}{3}\end{aligned}$$

(b) In a similar fashion, the required probability is

$$\iint_{2y<3x<2.5} f(x,y)\ dxdy = \int_0^1 \int_{2y/3}^{2.5/3}\ dxdy = \frac{1}{2}$$

□

6.3. *Conditional Probability Distribution*

Definition 6.7 (conditional discrete probability density function)**.** The conditional probability density function of X given that $Y = y$ is defined by

$$P_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \quad \text{for all } y \text{ such that } p_Y(y) > 0.$$

Similarly, the conditional distribution function of X given that $Y = y$ is defined by

$$F_{X|Y}(x \mid y) = P(X \le x \mid Y = y) \quad \text{for all } y \text{ such that } p_Y(y) > 0.$$

It follows that

$$F_{X|Y}(x \mid y) = \sum_{a \le x} p_{X|Y}(a|y).$$

If X is independent of Y, then the conditional probability density function of X given $Y = y$ is the same as the marginal probability density function of X for every y such that $p_Y(y) > 0$.

Definition 6.8 (conditional continuous probability density function)**.** Suppose X and Y are jointly continuous random variables. We define the conditional probability density function of X given $Y = y$ to be

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad \text{for all } y \text{ such that } f_Y(y) > 0.$$

For $A \subseteq \mathbb{R}$ and y such that $f_Y(y) > 0$,

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

The conditional distribution of X given that $Y = y$ is defined by

$$F_{X|Y}(x \mid y) = P(X \le x \mid Y = y) = \int_{-\infty}^{x} f_{X|Y}(t \mid y)\ dt.$$

If X is independent of Y, then the conditional probability density function of X given $Y = y$ is the same as the marginal probability density function of X for every y such that $f_Y(y) > 0$.

6.4. *Joint Probability Distribution Function of Functions of Several Variables*

Let X and Y be jointly distributed random variables with joint probability density function $f_{X,Y}$. It is sometimes necessary to obtain the joint distribution of the random variables U and V, which arise as functions of X and Y. Suppose

$$U = g(X,Y) \text{ and } V = h(X,Y) \quad \text{for some functions } g \text{ and } h.$$

We wish to find the joint probability function of U and V in terms of the joint probability density function $f_{X,Y}$, g and h.

Example 6.12. For example, say X and Y are independent exponentially distributed random variables. We are interested in the joint probability density function of $U = X+Y$ and $V = X/(X+Y)$. It is clear that

$$g(x,y) = x+y \quad \text{and} \quad h(x,y) = \frac{x}{x+y}.$$

In general, to find the joint probability density function of U and V, we state some conditions first.

Algorithm 6.1 (formulation of the joint probability density function). We assume that the following conditions are satisfied:

- **(i)** Let X and Y be jointly continuously distributed random variables with a known joint probability density function.
- **(ii)** Let U and V be given functions of X and Y of the form $U = g(X,Y)$ and $V = h(X,Y)$ and we can uniquely solve X and Y in terms of U and V. That is,

$$x = a(u,v) \quad \text{and} \quad y = b(u,v).$$

- **(iii)** The functions g and h have continuous partial derivatives and

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

  We call the matrix the Jacobian matrix and J the determinant of the Jacobian.

Hence, the joint probability density function of U and V is

$$f_{U,V}(u,v) = \frac{f_{X,Y}(x,y)}{J},$$

where $x = a(u,v)$ and $y = b(u,v)$ as mentioned.

Example 6.13. Let X and Y be jointly distributed with the joint probability density function

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

Note that X and Y are independent standard normal random variables and $\exp(x) = e^x$. If the term in the exponent is complicated, we usually use the former expression. Let R and θ denote the polar coordinates of the point (x,y). That is,

$$R = \sqrt{X^2+Y^2} \text{ and } \Theta = \tan^{-1}\left(\frac{Y}{X}\right).$$

Θ is the uppercase version of θ.

- **(i)** Find the joint probability density function of R and Θ.
- **(ii)** Show that R and Θ are independent.

Solution.

(i) Note that the random variables R and Θ take values in the respective intervals $(0,\infty)$ and $(0,2\pi)$. We set $r = g(x,y) = \sqrt{x^2+y^2}$ and $\theta = h(x,y) = \tan^{-1}(y/x)$. Hence, $x = r\cos\theta$ and $y = r\sin\theta$, which is essentially the conversion formulae from polar to Cartesian coordinates.

I omit the differentiation process in this case, but anyway, $J(x,y) = (x^2+y^2)^{-\frac{1}{2}}$. Hence,

$$\begin{aligned} f_{R,\Theta}(r,\theta) &= \frac{f_{X,Y}(x,y)}{\det(J(x,y))} \\ &= \sqrt{x^2+y^2}\cdot\frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right) \\ &= \frac{1}{2\pi}re^{-\frac{r^2}{2}} \end{aligned}$$

which is the joint probability density function of R and Θ.

(ii) They are independent. □

In the above example, R is actually a special continuous random variable. We say that R follows a Rayleigh distribution. A Rayleigh distribution is often observed when the overall magnitude of a vector is related to its directional components. One example where the Rayleigh distribution naturally arises is when wind velocity is analysed in two dimensions. Assuming that each component is uncorrelated, normally distributed with equal variance, and zero mean, then the overall wind speed (vector magnitude) will be characterised by a Rayleigh distribution.

If $X \sim \text{Rayleigh}(\sigma)$, where $\sigma > 0$, then

$$f(x) = \frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right).$$

We call σ the scale parameter. Not only is the Rayleigh distribution related to the normal distribution, but it is also related to the exponential distribution! That is, if $Y \sim \text{Exp}(\lambda)$, then

$$X = \sqrt{Y} \sim \text{Rayleigh}\left(\frac{1}{\sqrt{2\lambda}}\right).$$

Example 6.14 (ST2131 AY19/20 Sem 2)**.** Let X be a uniform random variable on $[0,1]$ and let Y be an independent exponential random variable with parameter 1.

(a) Find the joint p.d.f. of $U = Y - X$ and $V = XY$.

(b) Find $P(U \geq 1)$.

(a) The joint density function of X and Y, due to independence, is

$$f_{X,Y}(x,y) = \frac{1}{1}\cdot e^{-y} = e^{-y}.$$

Let $g(x,y) = y - x$ and $h(x,y) = xy$. The Jacobian determinant is

$$J(x,y) = \det\begin{pmatrix} \partial g/\partial x & \partial g/\partial y \\ \partial h/\partial x & \partial h/\partial y \end{pmatrix} = \det\begin{pmatrix} -1 & 1 \\ y & x \end{pmatrix} = -(x+y).$$

As

$$f_{U,V}(u,v) = f_{X,Y}(x,y)/|J(x,y)|,$$

then

$$f_{U,V}(u,v) = \frac{e^{-y}}{x+y},$$

but it is not in terms of u and v! As such, consider

$$Y = U + X \implies Y = U + \frac{V}{Y},$$

which yields the quadratic equation $Y^2 - UY - V = 0$. Thus,

$$Y = \frac{U \pm \sqrt{U^2 + 4V}}{2}.$$

Note that for the $\pm$ sign, we reject the negative. Suppose otherwise. Since $y \geq 0$, we have $u - \sqrt{u^2+4v} \geq 0$, so $u^2 \geq u^2 + 4v$, and thus, $0 \geq 4v$, which is a contradiction since $V = XY$, so $v \geq 0$. Hence,

$$Y = \frac{U + \sqrt{U^2 + 4V}}{2}.$$

In a similar fashion,

$$X = \frac{-U + \sqrt{U^2 + 4V}}{2}.$$

Substituting these into $f_{U,V}(u,v)$, we have

$$f_{U,V}(u,v) = \frac{\exp\left[-\left(U + \sqrt{U^2+4V}\right)/2\right]}{\sqrt{U^2+4V}}, \ u \geq -1, v \geq 0.$$

(b) We first find the marginal density of U. That is, finding $f_U(u)$ from $f_{U,V}(u,v)$. So,

$$f_U(u) = \int_0^\infty \frac{\exp\left[-\left(u + \sqrt{u^2+4v}\right)/2\right]}{\sqrt{u^2+4v}} \, dv.$$

Let $t = -\dfrac{u+\sqrt{u^2+4v}}{2}$, so $\dfrac{dt}{dv} = -\dfrac{1}{\sqrt{u^2+4v}}$. The integral becomes

$$\int_{-u}^{-\infty} -e^t \, dt = e^{-u}$$

so $f_U(u) = e^{-u}$. So, $P(U \geq 1) = \displaystyle\int_1^\infty e^{-u} \, du = 1/e$. □

7. Expectation Properties

7.1. *Expectation of Sums of Random Variables*

We start off this chapter with the following proposition.

Proposition 7.1. If $a \le X \le b$, then $a \le E(X) \le b$.

Proof. We will prove for the case where X is a discrete random variable. The proof for the continuous counterpart is similar, but we simply change the sum to an integral.

$$E(X) = \sum_{\text{all } x} xp(x) \ge \sum_{\text{all } x} ap(x) = a.$$

In a similar fashion, we can use the same technique to show that $E(X) \le b$. $\square$

Proposition 7.2. The following hold:

(i) If X and Y are jointly discrete random variables with joint probability density function $p_{X,Y}$, then

$$E[g(X,Y)] = \sum_{\text{all } y} \sum_{\text{all } x} g(x,y) p_{X,Y}(x,y)$$

(ii) If X and Y are jointly continuous random variables with joint probability density function $f_{X,Y}$, then

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

Corollary 7.1. Some important consequences are as follows:

(i) **Non-negativity:** If $g(x,y) \ge 0$ whenever $p_{X,Y}(x,y) > 0$, then $E[g(X,Y)] \ge 0$

(ii) **Linearity:** $E[g(X,Y) + h(X,Y)] = E[g(x,Y)] + E[h(X,Y)]$

(iii) **Linearity:** $E[g(X) + h(Y)] = E[g(X)] + E[h(Y)]$

(iv) **Monotonicity:** If jointly distributed random variables X and Y satisfy $X \le Y$, then $E(X) \le E(Y)$. Of course, this result can be easily extended to

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

which was covered in H2 Mathematics.

The formula for the expectation of the sample mean, $\overline{X}$, can be derived from **(iv)** in Corollary 7.1. It is clear that $E\left(\overline{X}\right) = \mu$, so the expected value of the sample mean is μ, the mean of the distribution. Hence, when μ is unknown, the sample mean is often used to estimate it.

Example 7.1. Recall that the binomial distribution is closely linked to the Bernoulli distribution. Suppose we perform an experiment n times and the probability of success for each trial is p. We define X to be the number of successes in n Bernoulli(p) trials. Since the expectation of each Bernoulli random variable is p and there are n Bernoulli trials, by the linearity property of expectation, we can use this method to derive that $E(X) = np$.

Example 7.2 (mean line segment length)**.** This involves a concept known as the mean line segment length. Suppose we have a unit square with vertices at $(0,0), (0,1), (1,0)$ and $(1,1)$. What is the mean distance between any two points in the square?

Solution. The above is a very interesting problem. The answer is definitely not $1/2$, but actually, rather close to it. The mean distance is approximately 0.52140, or in exact form,

$$\frac{2+\sqrt{2}+5\ln(1+\sqrt{2})}{15}.$$

Let us prove this result. Let U and V be independent uniform random variables as such: $U \sim U(0,1)$ and $V \sim U(0,1)$. We wish to find the distribution of $W = |U-V|$. We find the CDF of W first, before differentiating to find its PDF.

$$\begin{aligned}
P(W \le w) &= 1 - P(W > w) \\
&= 1 - P(|U-V| > w) \\
&= 1 - P(U-V < -w) - P(U-V > w) \\
&= 1 - P(V > U+w) - P(U > V+w) \\
&= 1 - \int_0^{1-w} P(V > U+w) f_U(u)\, du - \int_0^{1-w} P(U > V+w) f_V(v)\, dv \\
&= 1 - \int_0^{1-w} 1-(u+w)\, du - \int_0^{1-w} 1-(v+w)\, dv \\
&= 1-(1-w)^2
\end{aligned}$$

Upon differentiation yields $f_W(w) = 2(1-w)$, where $0 < w < 1$. We use the formula

$$E[g(U,V)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(u,v) f_{U,V}(u,v)\, dudv.$$

Since we are interested in the mean distance, or rather the expected distance, then $g(U,V) = \sqrt{U^2+V^2}$, so $g(u,v) = \sqrt{u^2+v^2}$. Note that $f_{U,V}(u,v) = 2(1-u)\cdot 2(1-v)$ due to independence. Therefore,

$$E\left(\sqrt{U^2+V^2}\right) = 4\int_0^1\int_0^1 \sqrt{u^2+v^2}(1-u)(1-v)\, dudv.$$

Using polar coordinates, $u = r\cos\theta$ and $v = r\sin\theta$. We need to find the bounds for r and θ too. By considering the lower half of the region, $0 \le r \le \sec\theta$ and $0 \le \theta \le \frac{\pi}{4}$. The integral becomes

$$8\int_0^{\frac{\pi}{4}}\int_0^{\sec\theta} r^2(1-\cos\theta)(1-\sin\theta)\, dr d\theta = 8\int_0^{\frac{\pi}{4}} \frac{\sec^3\theta}{12} - \frac{\sec^3\theta\tan\theta}{20}\, d\theta.$$

The integral of $\sec^3\theta\tan\theta$ is a standard one because the derivative of $\sec\theta$ is $\sec\theta\tan\theta$. To integrate $\sec^3\theta$, we need to use integration by parts.

$$\begin{aligned}
\int_0^{\frac{\pi}{4}} \sec^3\theta\, d\theta &= \int_0^{\frac{\pi}{4}} \sec\theta\sec^2\theta \\
&= [\sec\theta\tan\theta]_0^{\frac{\pi}{4}} - \int_0^{\frac{\pi}{4}} \tan\theta\sec\theta\tan\theta\, d\theta \\
&= \sqrt{2} - \int_0^{\frac{\pi}{4}} \sec\theta(\sec^2\theta - 1)\, d\theta \\
&= \sqrt{2} - \int_0^{\frac{\pi}{4}} \sec^3\theta\, d\theta + \int_0^{\frac{\pi}{4}} \sec\theta\, d\theta \\
2\int_0^{\frac{\pi}{4}} \sec^3\theta\, d\theta &= \sqrt{2} + [\ln|\sec\theta+\tan\theta|]_0^{\frac{\pi}{4}} \\
\int_0^{\frac{\pi}{4}} \sec^3\theta\, d\theta &= \frac{1}{\sqrt{2}} + \frac{1}{2}\ln\left(\sqrt{2}+1\right)
\end{aligned}$$

The rest of the working is left as a simple exercise. □

Theorem 7.1 (Boole's inequality). For a countable set of events $A_1, A_2, \ldots$, Boole's inequality states that

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

The generalisation of Boole's inequality (Theorem 7.1) is Bonferroni's inequality.

7.2. *Covariance, Variance and Correlation*

Definition 7.1 (covariance). The covariance of jointly distributed random variables X and Y, denoted by $\operatorname{cov}(X,Y)$, is defined by

$$\operatorname{cov}(X,Y) = E(X-\mu_X)(Y-\mu_Y),$$

where μ_X and μ_Y denote the means of X and Y respectively.

Definition 7.2 (correlation). If $\operatorname{cov}(X,Y) \neq 0$, we say that X and Y are correlated, but if $\operatorname{cov}(X,Y) = 0$, we say that X and Y are uncorrelated.

An alternative formula for covariance is

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

As a result, if X and Y are independent, it is clear that $\operatorname{cov}(X,Y) = 0$. However, the converse is not true. Correlation does not imply causation. I strongly recommend a video by Zach Star which illustrates how easy it is to lie with Statistics. For example, an increase in ice cream sales, as well as cases of sunburn, are caused by the hot weather, whereas there is a correlation between the number of ice cream sales and the number of sunburn cases.

Proposition 7.3. If X and Y are independent random variables, then for any functions $g,h : \mathbb{R} \to \mathbb{R}$, we have

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

Some other properties of covariance are as follows:

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

(ii) Symmetry: $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$

(iii)

$$\operatorname{cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \operatorname{cov}(X_i, Y_j)$$

(iv)

$$\operatorname{cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \operatorname{cov}(X_i, Y_j)$$

(v)

$$\operatorname{Var}\left(\sum_{k=1}^{n} X_k\right) = \sum_{k=1}^{n} \operatorname{Var}(X_k) + 2\sum_{i<j} \operatorname{cov}(X_i, X_j)$$

We only prove **(iv)**.

Proof.

$$\begin{aligned}\operatorname{cov}\left(\sum_{i=1}^{n}a_iX_i,\sum_{j=1}^{m}b_jY_j\right)&=E\left(\sum_{i=1}^{n}\sum_{j=1}^{m}a_ib_jX_iY_j\right)-E\left(\sum_{i=1}^{n}a_iX_i\right)E\left(\sum_{j=1}^{m}b_jY_j\right)\\&=\sum_{i=1}^{n}\sum_{j=1}^{m}a_ib_jE\left(X_iY_j\right)-\sum_{i=1}^{n}\sum_{j=1}^{m}a_ib_jE\left(X_i\right)E\left(Y_j\right)\\&=\sum_{i=1}^{n}\sum_{j=1}^{m}a_ib_j\operatorname{cov}\left(X_i,Y_j\right)\end{aligned}$$

□

Let $X_1,X_2,\ldots,X_n$ be independent random variables. Recall from H2 Mathematics that

$$\operatorname{Var}\left(\sum_{k=1}^{n}X_k\right)=\sum_{k=1}^{n}\operatorname{Var}(X_k).$$

Under independence, the variance of a sum is the sum of variances. We provide more information about the random variables. Suppose each of the X_i's has an expected value of μ and variance σ^2. We let

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

be the sample mean. The quantities $X_i-\overline{X}$, for $1\le i\le n$, are called deviations as they equal to the differences between the individual data and the sample mean. The random variable

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

is called the *sample variance*. We shall prove that $E(S^2)=\sigma^2$. That is, S^2 is used as an estimator for σ^2 instead of the more natural choice of

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

Proof. Note that $X_i-\overline{X}=X_i-\mu+\mu-\overline{X}$. Hence,

$$\begin{aligned}\left(X_i-\overline{X}\right)^2&=\left(X_i-\mu+\mu-\overline{X}\right)^2\\&=(X_i-\mu)^2+\left(\overline{X}-\mu\right)^2-2\left(\overline{X}-\mu\right)(X_i-\mu)\end{aligned}$$

When we take the sum of i from 1 to n, note that $\overline{X}$ is unaffected by the index. Hence,

$$\begin{aligned}S^2&=\frac{1}{n-1}\left[\sum_{i=1}^{n}(X_i-\mu)^2+\sum_{i=1}^{n}\left(\overline{X}-\mu\right)^2-2\left(\overline{X}-\mu\right)\sum_{i=1}^{n}(X_i-\mu)\right]\\&=\frac{1}{n-1}\left[\sum_{i=1}^{n}(X_i-\mu)^2+n\left(\overline{X}-\mu\right)^2-2n\left(\overline{X}-\mu\right)^2\right]\\&=\frac{1}{n-1}\left[\sum_{i=1}^{n}(X_i-\mu)^2-n\left(\overline{X}-\mu\right)^2\right]\\E\left(S^2\right)&=\frac{1}{n-1}\left[\sum_{i=1}^{n}E\left[(X_i-\mu)^2\right]-nE\left[\left(\overline{X}-\mu\right)^2\right]\right]\end{aligned}$$

By the definition of variance, as $E\left[(X-\mu)^2\right] = \mathrm{Var}(X)$, then it is clear that

$$\sum_{i=1}^{n} E\left[(X_i-\mu)^2\right] = n\sigma^2.$$

The term $E\left[\left(\overline{X}-\mu\right)^2\right]$ is called the *variance of the sample mean*. We wish to find the sum of it from $i=1$ to $i=n$. This is straightforward because

$$\sum_{i=1}^{n} E\left[\left(\overline{X}-\mu\right)^2\right] = \mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}.$$

Putting everything together,

$$E\left(S^2\right) = \frac{1}{n-1}\left(n\sigma^2 - n\left(\frac{\sigma^2}{n}\right)\right) = \sigma^2.$$

□

Example 7.3 (ST2132 AY18/19 Sem 2 Tutorial 1). Let a population consist of nine 0's and one 1. Make 2 random draws without replacement. Let X_1 and X_2 denote the outcomes of these 2 draws.

(a) Find the distribution of X_1.

(b) Find the mean and variance of X_1.

(c) Find the conditional distribution of X_2 given $X_1 = 0$, and the similar distribution given $X_1 = 1$. Hence, find the joint distribution of X_1 and X_2. Hence, find the distribution of X_2.

(d) Find $\mathrm{cov}(X_1, X_2)$.

Solution.

(a) We have $P(X_1 = 0) = 0.9$ and $P(X_1 = 1) = 0.1$

(b) $E(X_1) = 0.1$ and $\mathrm{Var}(X_1) = 0.1 \cdot 0.9 = 0.09$

(c) We have

$$P(X_2 = 0|X_1 = 0) = \frac{8}{9}$$
$$P(X_2 = 1|X_1 = 0) = \frac{1}{9}$$

and

$$P(X_2 = 0|X_1 = 1) = 1$$
$$P(X_2 = 1|X_1 = 1) = 0$$

For the joint distribution, $P(\{X_2 = 0\} \cap \{X_1 = 0\}) = P(X_2 = 0|X_1 = 0) \cdot P(X_1 = 0) = 8/9 \cdot 9/10 = 0.8$. The other probabilities can be computed. Namely, finding $P(\{X_2 = i\} \cap \{X_1 = j\}$ for $i = 1,2$ and $j = 1,2$.

(d) We have

$$P(\{X_2 = 0\} \cap \{X_1 = 0\}) = 0.8$$
$$P(\{X_2 = 0\} \cap \{X_1 = 1\}) = 0.1$$
$$P(\{X_2 = 1\} \cap \{X_1 = 0\}) = 0.1$$
$$P(\{X_2 = 1\} \cap \{X_1 = 1\}) = 0$$

We have

$$\begin{aligned}\operatorname{cov}(X_1,X_2) &= E(X_1X_2) - E(X_1)E(X_2)\\ &= \sum_{\text{all } x} x_1x_2P(x_1x_2 = x) - 0.1\cdot 0.1\\ &= 0\cdot 0\cdot 0.8 + 0\cdot 1\cdot 0.1 + 1\cdot 0\cdot 0.1 + 1\cdot 1\cdot 0 - 0.01\\ &= -0.01\end{aligned}$$

so $\operatorname{cov}(X_1,X_2) = -0.01$. □

Example 7.4 (ST2132 AY18/19 Sem 2 Tutorial 1)**.** Let $X_1,\ldots,X_n$ be random variables, with $E(X_i) = \mu_i$ and $\operatorname{Var}(X_i) = \sigma_i^2$. Their variance matrix (or variance-covariance matrix) V is an $n\times n$ matrix with

$$V_{ij} = \operatorname{cov}(X_i,X_j) = E[(X_i-\mu_i)(X_j-\mu_j)].$$

(a) What are the diagonal entries of V?
(b) Is V symmetric?
(c) Let X be distributed as $B(n,p)$ and $Y = n-X$. What is the variance matrix of X and Y?
(d) If $X_1,\ldots,X_n$ are independent, what can you comment about V?
(e) Repeat **(d)**, if $X_1,\ldots,X_n$ have the same distribution.
(f) Repeat **(d)**, if $X_1,\ldots,X_n$ are IID.

Solution.

(a)

$$\begin{aligned}V_{ij} &= E(X_iX_j) - E(X_i)\mu_j - E(X_j)\mu_i + \mu_i\mu_j = E(X_iX_j) - \mu_i\mu_j\\ V_{ii} &= E(X_i^2) - \mu_i^2 = E(X_i^2) - [E(X_i)]^2\end{aligned}$$

so the diagonal entries of V represent the variances.

(b) Yes, since $V_{ij} = V_{ji}$.

(c) We have

$$V_{11} = \operatorname{Var}(X) = np(1-p) \quad \text{and} \quad V_{22} = \operatorname{Var}(n-X) = np(1-p).$$

Also, as $E(X) = np$ and $E(Y) = n - np$, then

$$\begin{aligned}V_{12} &= E\left((X-np)(Y-n+np)\right)\\ &= E\left((X-np)(np-X)\right)\\ &= -E\left((X-np)^2\right)\\ &= -\operatorname{Var}(X)\\ &= -np(1-p)\end{aligned}$$

Note that $V_{12} = V_{21}$ due to symmetry.

(d) If X and Y are independent random variables, then $\operatorname{cov}(X,Y) = 0$. Using this fact, we see that $\operatorname{cov}(X_i,X_j) = 0$ for all $1\le i<j\le n$. Thus, V is now a diagonal matrix, with each diagonal entry representing the variance of X_i.

(e) The diagonal entries are the same.

(f) V is just a constant multiple of the $\mathbf{I}$, where the constant is the common variance. □

Example 7.5 (MA3238 AY13/14 Sem 2 Homework 3). A total of n bar magnets are placed end to end in a line on the table, where the orientation of the south and north poles of each magnet is randomly chosen from the two possibilities with equal probability. Adjacent magnets with opposite poles facing each other join to form a block. Find the mean and variance of the number of blocks of joined magnets.

Solution. Denote the bar magnets by $1, \ldots, n$. Call the bond between magnet i and magnet $i+1$ broken if they are not joined together. The key observation is that if there are a total of k broken bonds for $1 \leq k \leq n-1$, then $k+1$ disjoint blocks. Let N be the number of blocks and B be the number of broken bonds. Then,

$$N = B + 1.$$

As B denotes the number of broken bonds, then

$$B = \sum_{i=1}^{n-1} Y_i,$$

where Y_i is the indicator random variable denoting the event that the bond between magnet i and $i+1$ is broken. By the linearity of expectation,

$$\begin{aligned} E(N) &= 1 + \sum_{i=1}^{n-1} E(Y_i) \\ &= 1 + (n-1)E(Y_i) \text{ for any } 1 \leq i \leq n \end{aligned}$$

For two adjacent magnets i and $i+1$, they are broken if and only if they assume one of the following orientations:

NS | **SN** or **SN** | **NS**

As there are only a total of four orientations, the probability that any adjacent magnets are disjoint is $1/2$. Thus, $E(Y_i) = 1/2$ for all $1 \leq i \leq n$. Therefore,

$$E(N) = 1 + \frac{n-1}{2} = \frac{n+1}{2}.$$

Note that the variance of N is

$$\begin{aligned} \operatorname{Var}(N) &= \operatorname{Var}\left(\sum_{i=1}^{n-1} Y_i\right) \\ &= \sum_{i=1}^{n} \operatorname{Var}(Y_i) + 2 \sum_{1 \leq i < j \leq n-1} \operatorname{cov}(Y_i, Y_j) \\ &= \sum_{i=1}^{n} \left(E(Y_i^2) - (E(Y_i))^2\right) + 2 \sum_{1 \leq i < j \leq n-1} \operatorname{cov}(Y_i, Y_j) \\ &= \sum_{i=1}^{n} \left(E(Y_i) - (E(Y_i))^2\right) + 2 \sum_{1 \leq i < j \leq n-1} \operatorname{cov}(Y_i, Y_j) \\ &= \frac{n}{4} - 2 \sum_{1 \leq i < j \leq n-1} \operatorname{cov}(Y_i, Y_j) \end{aligned}$$

Note that $\operatorname{cov}(Y_i Y_j) = E(X_i X_j) - E(X_i)E(X_j)$. We claim that if $j = i+1$, the covariance is non-zero. If $i < j$, then Y_i depends on the orientation of magnets i and $i+1$, whereas Y_j depends on the orientation of magnets j and $j+1$. $\square$

Definition 7.3 (correlation). The correlation of random variables X and Y, denoted by $\rho(X,Y)$, is defined by

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

One would be more familiar with the formula given in the List of Formulae (MF26) during his/her A-Level days. That is, the product moment correlation coefficient, r.

Definition 7.4 (product moment correlation coefficient).

$$r=\frac{\sum(x-\bar{x})\sum(y-\bar{y})}{\left[\sqrt{\sum(x-\bar{x})^2}\right]\left[\sqrt{\sum(y-\bar{y})^2}\right]}$$

The two quantities ρ and r are of course equivalent. We can show that $-1\leq\rho(X,Y)\leq 1$.

Proof. Note that $\operatorname{cov}(X,Y)=E[(X-\mu_X)(Y-\mu_Y)]$, $\operatorname{Var}(X)=E[(X-\mu_X)^2]$ and $\operatorname{Var}(Y)=E[(Y-\mu_Y)^2]$. Hence, the original equation for ρ becomes

$$\rho(X,Y)=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]E[(Y-\mu_Y)^2]}}.$$

Using the substitution $U=X-\mu_X$ and $V=Y-\mu_Y$,

$$\rho(X,Y)=\frac{E(UV)}{\sqrt{E(U^2)E(V^2)}}.$$

Assume that X and Y are continuous random variables, which would imply that U and V are continuous random variables too. The proof will be the same for the discrete case, just that the integrals become sums. We define $f(t)$ to be the following polynomial in t:

$$f(t)=E[(tU+V)^2]$$

Then, expanding the right side yields

$$f(t)=E\left(U^2\right)t^2+2tE\left(UV\right)+E\left(V^2\right).$$

Note that $f(t)\geq 0$ since $\operatorname{Var}(X)\geq 0$ if and only if $E(X^2)\geq(E(X))^2\geq 0$. Hence, the discriminant of $f(t)$, Δ must satisfy $\Delta\leq 0$. That is,

$$(2E(UV))^2-4(E(U^2))(E(V^2))\leq 0.$$

Rearranging yields the formula

$$(E(UV))^2\leq E(U^2)E(V^2),$$

which implies that $-1\leq\rho(X,Y)\leq 1$. To conclude, we remark that the inequality $(E(UV))^2\leq E(U^2)E(V^2)$ is the famous Cauchy-Schwarz inequality. □

The correlation coefficient is a measure of the degree of linearity between X and Y. A value of $\rho(X,Y)$ near $+1$ or -1 indicates a high degree of linearity between X and Y, whereas a value near 0 indicates a lack of such linearity. A positive value of $\rho(X,Y)$ indicates that Y tends to increase as X does, whereas a negative value indicates that Y tends to decrease as X increases. If $\rho(X,Y)=0$, then X and Y are uncorrelated. If X and Y are independent, then $\rho(X,Y)=0$. However, the converse is not true.

7.3. *Conditional Expectation*

Definition 7.5 (conditional expectation). If X and Y are jointly distributed discrete random variables, then if $p_Y(y) > 0$,

$$E(X|Y=y) = \sum_{\text{all } x} x p_{X|Y}(x|y).$$

If X and Y are jointly distributed continuous random variables, then if $f_Y(y) > 0$,

$$E(X|Y=y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx.$$

Note that for both the discrete and continuous cases, we can replace X with $g(X)$ and the formula will just have minor tweaks to it. That is,

$$E(g(X)|Y=y) = \sum_{\text{all } x} g(x) p_{X|Y}(x|y) \text{ for the discrete case} \quad \text{and}$$
$$E(g(X)|Y=y) = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\, dx \text{ for the continuous case}$$

Hence,

$$E\left(\sum_{i=1}^{n} X_i|Y=y\right) = \sum_{i=1}^{n} E(X_i|Y=y).$$

We can compute expectations and probabilities by conditioning.

Definition 7.6 (conditional variance). The conditional variance of X given $Y=y$ is defined as

$$\text{Var}(X|Y) = E((X-E(X|Y))^2|Y).$$

A useful relationship between $\text{Var}(X)$ and $\text{Var}(X|Y)$, called the law of total variance, is as follows:

Proposition 7.4 (law of total variance).

$$\text{Var}(X) = E(\text{Var}(X|Y)) + \text{Var}(E(X|Y))$$

Example 7.6 (ST2131 AY19/20 Sem 2). Let N be a Poisson random variable with mean 1. Let $(\xi_i)_{i\in\mathbb{N}}$ be i.i.d. standard normal random variables. Define

$$X := \sum_{i=1}^{N} \xi_i = \xi_1 + \xi_2 + \ldots + \xi_N.$$

Find the mean and variance of X.

Solution. By the law of total expectation,

$$\begin{aligned}
E(X) &= \sum_{i=1}^{n} E(X|N=i)P(N=i) \\
&= \sum_{i=1}^{n} E(X|N=i) \cdot \frac{e^{-1}}{i!} \\
&= E(\xi_1) \cdot \frac{e^{-1}}{1!} + E(\xi_1+\xi_2) \cdot \frac{e^{-1}}{2!} + \ldots + E(\xi_1+\xi_2+\ldots+\xi_n) \cdot \frac{e^{-1}}{n!} \\
&= 0 \cdot \frac{e^{-1}}{1!} + 0 \cdot \frac{e^{-1}}{2!} + \ldots + 0 \cdot \frac{e^{-1}}{n!} \\
&= 0
\end{aligned}$$

For the variance, by the law of total variance (Proposition 7.4),

$$\mathrm{Var}(X) = E(\mathrm{Var}(X|N)) + \mathrm{Var}(E(X|N)) = E(N\,\mathrm{Var}(X)) = E(N) = 1.$$

□

7.4. *Moment Generating Function*

The moment generating function (MGF) of a real-valued random variable, X, is an alternative specification of its probability distribution. It provides the basis of an alternative route to analytical results compared with working directly with PDFs or CDFs. There are particularly simple results for the MGFs of distributions defined by the weighted sums of random variables. However, not all random variables have MGFs.

Definition 7.7 (moment generating function). The MGF of a random variable X, denoted by M_X, is defined as

$$M_X(t) = E\left(e^{tX}\right).$$

If X is a discrete random variable with PDF p_X, then

$$M_X(t) = \sum_{\text{all } x} e^{tx} p_X(x).$$

If X is a continuous random variable with PDF f_X, then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx.$$

We call such a function a moment generating function because it generates all the moments of this random variable X. Indeed, for $n \geq 0$,

$$E(X^n) = M_X^{(n)}(0),$$

where

$$M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t) \quad \text{when } t \text{ is evaluated at } 0.$$

Proof. Using series expansion,

$$E(e^{tX}) = E\left(\sum_{k=0}^{\infty} \frac{(tX)^k}{k!}\right) = \sum_{k=0}^{\infty} \frac{E(X^k)t^k}{k!} = \sum_{k=0}^{\infty} \frac{M_X^{(k)}(0)t^k}{k!}.$$

The result follows by equating the coefficient of t^n. □

Proposition 7.5. The MGF of a random variable satisfies two properties. We state them.

(i) Multiplicativity: If X and Y are independent, then

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

(ii) Let X and Y be random variables with MGFs being M_X and M_Y respectively. If there exists $h > 0$ such that

$$M_X(t) = M_Y(t) \quad \text{for all } -h < t < h.$$

Then, it implies that X and Y have the same distribution, meaning that $f_X = f_Y$.

We state and prove the MGFs for some random variables.

Proposition 7.6 (MGF of Bernoulli random variable). If $X \sim \text{Bernoulli}(p)$, then

$$M(t) = 1 - p + pe^t.$$

Proof. Using the formula, $M(t) = e^{t(0)}P(X=0) + e^{t(1)}P(X=1) = (1-p) + pe^t$. □

Proposition 7.7 (MGF of binomial random variable). If $X \sim \mathrm{B}(n,p)$, then

$$M(t) = (1 - p + pe^t)^n.$$

Proof. Using the formula, and writing it in sigma notation,

$$\sum_{k=0}^{n} e^{kt}\binom{n}{k}p^k q^{n-k} = q^n \sum_{k=0}^{n}\binom{n}{k}\left(\frac{pe^t}{q}\right)^k = q^n\left(1+\frac{pe^t}{q}\right)^n = (1-p+pe^t)^n.$$

□

Proposition 7.8 (MGF of geometric random variable). If $X \sim \text{Geo}(p)$, then

$$M(t) = \frac{pe^t}{1-qe^t}.$$

Proof. Using the formula,

$$M(t) = \sum_{k=1}^{n} e^{kt}pq^{k-1} = \frac{p}{q}\sum_{k=1}^{n}(qe^t)^k = \frac{pe^t}{1-qe^t}.$$

□

Example 7.7. If $X \sim \text{Geo}\,(p)$, find an expression for $E\left(X^3\right)$.

Solution. The moment generating function, $M\,(t)$, for X is $E\left(e^{tX}\right)$. If $X \sim \text{Geo}(p)$, then

$$M(t) = \sum_{k=1}^{n} e^{kt}pq^{k-1} = \frac{p}{q}\sum_{k=1}^{n}(qe^t)^k = \frac{pe^t}{1-qe^t},$$

where $q = 1-p$. Hence, the third moment, or $E(X^3)$, is the coefficient of t^3 divided by $3! = 6$ in the series expansion of $M(t)$. □

Proposition 7.9 (MGF of Poisson random variable). If $X \sim \text{Po}(\lambda)$, then

$$M(t) = \exp(\lambda(e^t - 1)).$$

Proof.

$$M(t) = \sum_{k=0}^{n}\frac{e^{kt}e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{n}\frac{(\lambda e^t)^k}{k!} = e^{-\lambda}\left(e^{\lambda e^t}\right) = \exp(\lambda(e^t-1))$$

□

Proposition 7.10 (MGF of uniform random variable). If $X \sim U(a,b)$, then

$$M(t) = \frac{e^{bt}-e^{at}}{t(b-a)}.$$

Proof.
$$M(t) = \int_{-\infty}^{a} \frac{e^{kt}}{b-a}\,dk + \int_{a}^{b} \frac{e^{kt}}{b-a}\,dk + \int_{b}^{\infty} \frac{e^{kt}}{b-a}\,dk = \int_{a}^{b} \frac{e^{kt}}{b-a}\,dk = \frac{e^{bt}-e^{at}}{t(b-a)}$$

□

Proposition 7.11 (MGF of normal random variable)**.** If $X \sim N(\mu, \sigma^2)$, then

$$M(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

The proof will be left as an exercise.

Proposition 7.12 (MGF of exponential random variable)**.** If $X \sim \text{Exp}(\lambda)$, then

$$M(t) = \frac{\lambda}{\lambda - t}$$

for $t < \lambda$.

Proof.
$$M(t) = \int_{0}^{\infty} e^{tk} \lambda e^{-\lambda k}\,dk = \frac{\lambda}{\lambda - t}$$

$M(t)$ is only defined for $t < \lambda$ because the expectation of an exponential random variable is always positive. To justify this, if $f(x) = \lambda e^{-\lambda x}$, then $E(X) = \frac{1}{\lambda}$. By the definition of the exponential distribution, as $\lambda > 0$, the result follows. □

8. Limit Theorems

8.1. *Statistical Inequalities*

Theorem 8.1 (Markov's inequality)**.** Let X be a non-negative random variable. For $a > 0$,

$$P(X \geq a) \leq \frac{E(X)}{a}.$$

Proof. We only prove this for the continuous random variable X. The discrete case is very similar, just that the integral is replaced by summation.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)\ dx \\ &= \int_{0}^{\infty} xf(x)\ dx \text{ since } X \text{ is non-negative} \\ &\geq \int_{a}^{\infty} xf(x)\ dx \\ &\geq \int_{a}^{\infty} af(x)\ dx \text{ since } f(x) \text{ is non-negative} \\ &= aP(X \geq a) \end{aligned}$$

which concludes the proof. □

Theorem 8.2 (Chebyshev's inequality)**.** Let X be a random variable with finite mean μ and variance σ^2. Then, for $a > 0$, we have

$$P(|X-\mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Proof. Applying Markov's inequality (Theorem 8.1),

$$P(|X-\mu| \geq a) = P\left((X-\mu)^2 \geq a^2\right) \leq \frac{E\left[(X-\mu)^2\right]}{a^2} = \frac{\text{Var}(X)}{a^2}.$$

□

Example 8.1 (STEP 3 2016 Question 12)**.** The probability of a biased coin landing heads up is 0.2. It is thrown $100n$ times, where n is an integer greater than 1. Let α be the probability that the coin lands heads up N times, where $16n \leq N \leq 24n$. We can use Chebyshev's inequality to prove the following two results:

(i) $\alpha \geq 1 - \frac{1}{n}$

(ii) $1 + n + \frac{n^2}{2!} + \cdots + \frac{n^{2n}}{(2n)!} \geq \left(1 - \frac{1}{n}\right) e^n$

However, in the test, their form of Chebyshev's inequality (Theorem 8.2) is slightly different. That is for $k > 0$,

$$P(|X-\mu| > k\sigma) \leq \frac{1}{k^2}.$$

Solution.

(i) We first recognise that this is a setup modelling a binomial distribution. Let X be the random variable denoting the number of times the coin lands heads up, out of $100n$. Then, $\alpha = P(|X-20n| \leq 4n)$. Note that $E(X) = 20n$, $\text{Var}(X) = 16n$ and $|X-20n| \leq 4n$. Removing the modulus, $16n \leq X \leq 24n$, which indeed satisfies the original inequality that $16n \leq N \leq 24n$. By Chebyshev's Inequality,

$$\begin{aligned} P(|X-20n| > 4n) &\leq \frac{16n}{(4n)^2} \\ 1 - P(|X-20n| \leq 4n) &\leq \frac{1}{n} \\ 1 - \frac{1}{n} &\leq P(|X-20n| \leq 4n) \\ \alpha &\geq 1 - \frac{1}{n} \end{aligned}$$

and the result follows.

(ii) This is quite interesting. Observe that the left side of the inequality is the partial sum of the Maclaurin Series of e^n. If we can prove that

$$1 + n + \frac{n^2}{2!} + \cdots + \frac{n^{2n}}{(2n)!} \geq \alpha e^n,$$

then we are done. Recall that the only discrete random variable we studied which contains the exponential function is the Poisson random variable. Suppose $Y \sim \mathrm{Po}(n)$. Then, $\mu = \sigma^2 = n$. Set $a = n$. Substituting these into Chebyshev's inequality (Theorem 8.2) yields

$$P(|Y-n| > n) \leq \frac{1}{n}.$$

We consider the modulus inequality first. This is equivalent to $Y - n \geq n$ or $Y - n \leq -n$, which implies that $Y \geq 2n$ or $Y \leq 0$ respectively. The latter does not make sense because the support of Y is the non-negative integers. Thus, the inequality becomes

$$P(Y > 2n) \leq \frac{1}{n}.$$

With some simple algebraic manipulation,

$$1 - \frac{1}{n} \leq P(Y \leq 2n).$$

Hence,

$$\begin{aligned} 1 - \frac{1}{n} &\leq \sum_{i=0}^{2n} \frac{e^{-n} n^i}{i!} \\ \sum_{i=0}^{2n} \frac{n^i}{i!} &\geq \alpha e^n \end{aligned}$$

which concludes our proof. □

The importance of Markov's and Chebyshev's inequalities (Theorems 8.1 and 8.2 respectively) is that they enable us to derive bounds on probabilities when only the mean, or both the mean and the variance, of the probability distribution are known. Of course, if the actual distribution were known, then the desired probabilities can be exactly computed and we would not need to resort to bounds.

Theorem 8.3 (Jensen's inequality). If X is a random variable and ϕ is a convex function, then

$$\phi(E(X)) \leq E(\phi(X)).$$

Corollary 8.1. For $x \geq 0$, the graph of $\phi(x) = x^n$, where $n \in \mathbb{N}$, is convex. Hence,

$$E(X^n) \geq (E(X))^n \quad \text{for } n \in \mathbb{N}.$$

Corollary 8.2. If $\text{Var}(X) = 0$, then X is a constant. In other words, $P(X = E(X)) = 1$. We say that X is a degenerate random variable.

Proof. By Chebyshev's inequality (Theorem 8.2), for any $n \geq 1$,

$$0 \leq P\left(|X-\mu| > \frac{1}{n}\right) \leq \frac{\text{Var}(X)}{1/n^2} = 0.$$

By the squeeze theorem, it implies that

$$P\left(|X-\mu| > \frac{1}{n}\right) = 0.$$

Taking limits on both sides and using the continuity property of probability,

$$0 = \lim_{n\to\infty} P\left(|X-\mu| > \frac{1}{n}\right) = P\left(\lim_{n\to\infty}\left\{|X-\mu| > \frac{1}{n}\right\}\right) = P(X \neq \mu).$$

This asserts that $P(X = \mu) = 1$. □

8.2. *Laws of Large Numbers (LLN)*

Theorem 8.4 (weak law of large numbers (WLLN))**.** Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables, with a common mean μ. We define the sample mean to be

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n).$$

Then, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\overline{X} - \mu\right| \geq \varepsilon\right) = 0.$$

In other words, the sample mean converges to the expected value as $n \to \infty$.

Proof. We shall prove this theorem only under the additional assumption that the random variables have a finite variance σ^2. As It is clear that $E(\overline{X}) = \mu$ and $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$, then by Chebyshev's inequality (Theorem 8.2),

$$P\left(|\overline{X} - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

As $n \to \infty$, the expression on the right side of the inequality tends to 0. □

Theorem 8.5 (strong law of large numbers (SLLN))**.** Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables, each having a finite mean $\mu = E(X_i)$. Recall how the sample mean is defined when we introduced the WLLN. Then, the SLLN states that as $n \to \infty$,

$$\overline{X} \to \mu.$$

In probabilistic terms,

$$P\left(\left\{\lim_{n\to\infty} \overline{X} = \mu\right\}\right) = 1.$$

The weak law states that for a specified large n, the average $\overline{X}$ is likely to be near μ. Thus, it leaves open the possibility that $|\overline{X}-\mu|>\varepsilon$ happens an infinite number of times, although at infrequent intervals.

In contrast, the strong law shows that this almost surely will not occur. Note that it does not imply that with probability 1, we have that for any $\varepsilon>0$, the inequality $|\overline{X}-\mu|<\varepsilon$ holds for all large enough n since the convergence is not necessarily uniform on the set where it holds.

Almost sure convergence implies convergence in probability, but the converse is not true. The proof is out of the scope of our discussion as it is with reference to Probability Theory at a higher level. It uses a complex branch of Pure Mathematics called Measure Theory and we use a lemma, called the Borel-Cantelli lemma, to prove the aforementioned statement. This is why there is a distinction between the weak law and the strong law.

8.3. *Central Limit Theorem (CLT)*

The central limit theorem (CLT) is one of the most remarkable results in Probability Theory. It states that the sum of a large number of independent random variables has a distribution that is approximately normal. Hence, it not only provides a simple method for computing approximate probabilities for sums of independent random variables, but it also helps explain the remarkable fact that the empirical frequencies of so many natural populations exhibit bell-shaped curves.

We will only study one form of the CLT and it is known as the *Classical CLT*. Fun fact, if you were to go to Wikipedia, you will find that there are three other types of CLT, namely the Lyapunov CLT, Lindenberg CLT and the multidimensional CLT. All these will be out of scope of our discussion.

Without further ado, we state the simplest form of the CLT — the classical CLT.

Theorem 8.6 (central limit theorem)**.** Let $X_1,X_2,\ldots,X_n$ be a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2. Then, the distribution of

$$\frac{X_1+X_2+\ldots+X_n-n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n\to\infty$.

Lemma 8.1. We have two results, one of which is related to the sum of X_i's, and one is related to the sample mean.

(i)

$$X_1+X_2+\ldots+X_n\sim N(n\mu,n\sigma^2)\quad\text{approximately}$$

(ii)

$$\frac{1}{n}(X_1+X_2+\ldots+X_n)\sim N\left(\mu,\frac{\sigma^2}{n}\right)\quad\text{approximately}$$

Example 8.2 (ST2132 AY18/19 Sem 2 Tutorial 1)**.** Let $X_1,\ldots,X_n$ be IID with the Bernoulli(p) distribution.

(a) Let x_i be a realisation of X_i. As $n\to\infty$, what can you say about $x_1,\ldots,x_n$?

(b) Let $S_n=X_1+\ldots+X_n$. Derive the distribution of S_n, using the following fact: There are

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

ways to arrange k 1's and $(n-k)$ 0's in a row.

(c) Find the expectation and standard deviation of S_n.

(d) What does the central limit theorem say about S_n as $n \to \infty$?

Solution.

(a) The mean converges to p and the variance converges to $p(1-p)$.

(b) Suppose $P(X_i = 0) = p$ and $P(X_i = 1) = 1-p$ as defined by the Bernoulli distribution. Then, $P(S_1 = 0) = p$ and $P(S_1 = 1) = 1-p$. At this stage, it is apparent that S_n resembles a binomial distribution, so

$$P(S_n = k) = P(X_1 + X_2 + \ldots + X_n = k) = \binom{n}{k} p^k (1-p)^k.$$

(c) $E(S_n) = np$ and $\mathrm{Var}(S_n) = np(1-p)$, so $S_n \sim B(np, np(1-p))$.

(d) The distribution converges to the standard normal distribution $N(0,1)$. $\square$