

MA2116 Probability

Thang Pang Ern

These notes are based off **Prof. Chin Chee Whye's** MA2116 Probability materials. Additional references are cited in the bibliography. On top of the these, 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.

Also, any question taken from S. M. Ross '*A First Course in Probability*' is from the 10th edition text unless stated otherwise. Note that this course is also known as **ST2131 Probability**, but I prefer to call it MA2116.

This set of notes was last updated on **December 13, 2025**. If you would like to contribute a nice discussion to the notes or point out a typo, please send me an email at `thangpangern@u.nus.edu`. Thanks to Ng Kang Zhe, Malcolm Tan Jun Xi, and Ryan Au Eong for contributing.

Contents

Contents	i
1 Combinatorial Analysis	1
1.1 The Basic Principle of Counting	1
1.2 Permutations and Combinations	2
1.3 Multinomial Coefficients	8
1.4 Distribution Problems	11
2 Axioms of Probability	19
2.1 The Kolmogorov Axioms	19
2.2 Probability Properties	25
2.3 Probability as a Continuous Set Function	32

3 Conditional Probability and Independence	35
3.1 Conditional Probabilities	35
3.2 Bayes' Theorem	40
3.3 Independent Events	47
4 Discrete Random Variables	57
4.1 Discrete Random Variables	57
4.2 Expectation and Variance	59
4.3 Bernoulli Distribution	68
4.4 Binomial Distribution	70
4.5 Geometric Distribution	79
4.6 Negative Binomial Distribution	86
4.7 Poisson Distribution	87
4.8 Hypergeometric Distribution	94
5 Continuous Random Variables	97
5.1 Continuous Random Variables	97
5.2 Expectation and Variance	99
5.3 Continuous Uniform Distribution	101
5.4 Normal Distribution	107
5.5 Exponential Distribution	114
5.6 Gamma Distribution	121
5.7 Beta Distribution	123
5.8 Cauchy Distribution	124
5.9 Order Statistics	125
6 Joint Probability Distribution	127
6.1 Joint Distribution Functions	127
6.2 Independent Random Variables	134
6.3 Conditional Probability Distribution	149
6.4 Joint Probability Distribution Function of Functions of Several Variables .	155
7 Expectation Properties	161
7.1 Expectation of Sums of Random Variables	161
7.2 Covariance, Variance and Correlation	167
7.3 Conditional Expectation	176
7.4 Moment Generating Function	189
8 Limit Theorems	203
8.1 Statistical Inequalities	203
8.2 Laws of Large Numbers (LLN)	206
8.3 Central Limit Theorem (CLT)	207
Bibliography	213

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 chapter have already been covered in high school or Olympiad Mathematics. 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 $1 \leq k \leq r$, 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.

$$n_1 + n_2 + \dots + n_r.$$

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 .

Naturally, we can extend Proposition 1.2. That is to say, it can be applied to 2 or more tasks performed independently in succession. If the k^{th} task can be performed in m_k ways, where $1 \leq k \leq r$, then the number of ways of performing the r tasks in succession is

$$m_1 m_2 \dots m_n.$$

Example 1.1 (ST2131 AY24/25 Sem 1 Lecture 1). A 4-digit code is to be formed using the digits 0, 1, 2, ..., 9. By the multiplication principle (Propostion 1.2), the number of codes that can be formed is $10^4 = 10000$. If the digits may not be repeated, then we can form $10 \cdot 9 \cdot 8 \cdot 7 = 5040$ codes.

Example 1.2 (handshaking lemma; 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. First, we label the 20 people as P_1, P_2, \dots, P_{20} . Then, P_1 can shake the hands of P_2, P_3, \dots, P_{20} , so there are 19 ways to do so. P_2 can shake the hands of P_3, P_4, \dots, P_{20} , so there are 18 ways to do so. Repeat this till P_{19} , who can only shake P_{20} 's hand. By the addition principle (Proposition 1.1), the required number of ways is

$$19 + 18 + \dots + 1 = \frac{19 \cdot 20}{2} \quad (1.1)$$

so there are 190 ways to do so. Note that (1.1) used the formula for the sum to n terms of an arithmetic progression. \square

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. For example, in the complete graph on 6 vertices (denoted by K_6), every vertex is adjacent to the other 5, so the graph has 15 interior edges (Figure 1.1).

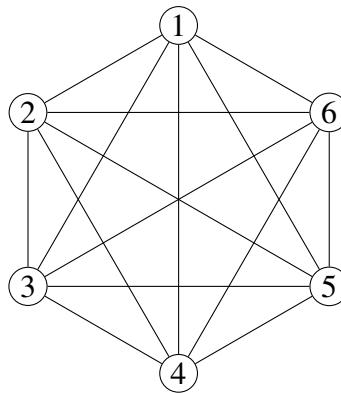


Figure 1.1: Complete graph K_6

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.

Example 1.3 (ST2131 AY24/25 Sem 1 Lecture 1). 6 boys and 4 girls compete in a running race (no tie). If the boys and the girls run together, there are $10! = 3628800$ different finishing orders. If the boys and the girls run separately, there are $6! \cdot 4! = 17280$ different finishing orders.

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. \square

Proposition 1.4 (permutations involving identical objects). Given n objects, of which n_1 are identical, n_2 are identical, and all the way up to n_r are identical, the number of different permutations of the n objects is

$$\frac{n!}{n_1!n_2!\dots n_r!} \quad \text{where } n_1 + \dots + n_r = n.$$

We then 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, so we divide the number of linear permutations by the number of people.

Definition 1.2 (combination). 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 1.5 (symmetry of binomial coefficients). For any $n \in \mathbb{Z}_{\geq 0}$,

$$\binom{n}{r} = \binom{n}{n-r} \quad \text{for all } 0 \leq r \leq n. \quad (1.2)$$

A special case of Proposition 1.2 is the following identity:

$$\binom{n}{0} = \binom{n}{n} = 1, \quad (1.3)$$

which can be obtained by considering the formula for r combinations out of n objects (Definition 1.2). In particular, setting $r = 0$ and $r = n$ will yield (1.3). As for the general case of (1.2), the algebraic proof is simple but not as meaningful as its combinatorial counterpart. As such, we provide a proof for the latter.

Proof. There are two ways to select a group of r items from a group of n , which picking the r items that you are 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. \square

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

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

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. \square

We now discuss Example 1.4, which demonstrates how the same counting problem can be approached from three distinct combinatorial perspectives. Each part of the example showcases a classic technique from basic counting principles, providing an elegant illustration of how different decompositions of the same problem lead to equivalent algebraic expressions. Note that we previously encountered an example of a combinatorial proof in the proof of the symmetry of binomial coefficients (1.3).

Example 1.4 (Ross p. 30 Question 12). From a group of n people, suppose that we want to choose a committee of k , where $k \leq n$, 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 (i), (ii), and (iii) — it is just the same event viewed from three different perspectives.

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 (Proposition 1.2).
- (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 (Proposition 1.2).

- (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 (Proposition 1.2). \square

Example 1.5 (modified Ross p. 33 Question 17). Give an analytic verification of

$$k \binom{n}{k} = (n - k + 1) \binom{n}{k-1} = n \binom{n-1}{k-1} \quad \text{where } 1 \leq k \leq n.$$

Now, give a combinatorial proof for this identity.

Solution. We first give an analytic representation. We have

$$k \binom{n}{k} = k \cdot \frac{n!}{k!(n-k)!} = (n - k + 1) \cdot \frac{n!}{(k-1)!(n-k+1)!} = (n - k + 1) \binom{n}{k-1}.$$

We then prove the second equality. We have

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

which shows that the identity holds.

We then give a combinatorial proof of the identity. Say we have a group of n persons. Say we wish to choose k students to be part of a committee, and elect 1 person to be the President. There are $\binom{n}{k} \binom{k}{1} = k \binom{n}{k}$ ways to do so.

Another way of formulating this is to choose the President first in $\binom{n}{1} = n$ ways, then form the committee (which requires $k - 1$ persons) in $\binom{n-1}{k-1}$ ways. It follows that

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

Alternatively, we choose the non-Presidents first in $\binom{n}{k-1}$ ways, then from the remaining $n - (k - 1)$ persons, we choose the President, which shows that

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

The result follows. \square

Example 1.6 (Ross p. 33 Question 17). Give an analytic verification of

$$\binom{n}{2} = \binom{k}{2} + k(n - k) + \binom{n - k}{2} \quad \text{where } 1 \leq k \leq n.$$

Now give a combinatorial argument for this identity.

Solution. The analytic verification is trivial. We now give a combinatorial interpretation. The binomial coefficient $\binom{n}{2}$ represents the number of ways to choose 2 numbers in $1 \leq k \leq n$. We then classify a chosen pair according to how many of the two elements lie in the first k numbers and the last $n - k$ numbers.

- **Case 1:** Suppose both numbers chosen are in $[1, k]$. Then, there are $\binom{k}{2}$ ways.
- **Case 2:** Suppose both numbers chosen are in $[k+1, n]$. Then, there are $\binom{n-k}{2}$ ways.
- **Case 3:** Suppose the first number is chosen in $[1, k]$ and the second number is chosen in $[k+1, n]$. Then, there are $k(n-k)$ ways.

The three cases are disjoint so the result follows by the addition principle (Proposition 1.1). \square

We then introduce the binomial theorem¹.

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).

$$\sum_{k=0}^n \binom{n}{k} = 2^n \quad (1.4)$$

Proof. For an analytic proof, we set $x = y = 1$ in the binomial theorem formula (Theorem 1.2) and we are done. Combinatorially, we can also view (1.4) as follows. If a set has n elements, then the number of subsets, including the empty set \emptyset 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. \square

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). \square

Example 1.7 (ST2131 AY22/23 Sem 2 Tutorial 2). Consider the combinatorial identity (1.5). 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.

$$\sum_{k=1}^n k \binom{n}{k} = n 2^{n-1} \quad (1.5)$$

¹Pascal's triangle is a triangular array of the binomial coefficients that arises in Probability Theory. 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 1.5) as mentioned, and others involving the triangular numbers and Fibonacci numbers. You can find some fascinating patterns here.

Solution. We focus on the right side of (1.5) 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, 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 (1.5).

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, the two quantities in (1.5) are equal. \square

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

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

Proof. Use induction or repeatedly apply Pascal's identity (Theorem 1.1). \square

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

$$\sum_{k=0}^r \binom{m}{k} \binom{n}{r-k} = \binom{m+n}{r}. \quad (1.6)$$

We provide a combinatorial proof of this result².

Proof. 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 (Proposition 1.2), we obtain the left side of (1.6). The right side can be obtained by considering the fact that we have $m+n$ men and women and we wish to form a group of size r , thus there are $\binom{m+n}{r}$ ways to do so. \square

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). \square

²This question also appeared in ST2131 AY24/25 Sem 2 Tutorial 1.

Example 1.8 (Ross p. 33 Question 21). Simplify

$$n - \binom{n}{2} + \binom{n}{3} - \dots + (-1)^{n+1} \binom{n}{n}.$$

Solution. We see that the sum can be written as

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

By the binomial theorem (Theorem 1.2),

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

We can set $a = -1$ and $b = 1$ to obtain

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

Multiplying both sides by -1 , then subtracting the term when $k = 0$, we see that the original sum is equal to 1. \square

1.3 Multinomial Coefficients

Many real-world situations require dividing a set into multiple distinct groups or analysing outcomes involving more than two types of events. While the binomial theorem handles expansions of expressions like $(x+y)^n$, it falls short when more than two variables are involved. This is where the multinomial theorem comes into play — it generalises the binomial expansion to sums with any number of terms. In particular, multinomial coefficients arise naturally when counting the number of ways to distribute n identical items into r distinguishable boxes, or when assigning labels or roles across multiple partitions (see Chapter 1.4).

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

$$(x_1 + \dots + x_r)^n = \sum_{k_1+\dots+k_r=n} \binom{n}{k_1, \dots, k_r} x_1^{k_1} \dots x_r^{k_r} \quad (1.7)$$

where $\binom{n}{k_1, \dots, k_r}$ 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 $n = 2$. Combinatorially, the multinomial coefficient $\binom{n}{k_1, \dots, k_r}$ in (1.7) counts the number of ways to divide a set of n distinct objects into r distinct groups, where the first group has k_1 objects, the second has k_2 objects, and so on, such that $k_1 + \dots + k_r = n$.

Example 1.9 (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.
- (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 $\frac{1}{3!} \binom{9}{3,3,3}$. \square

Example 1.10. A group of 15 students — 5 in Mathematics, 5 in Physics, and 5 in Computer Science, are divided into 5 teams of 3.

- (a) What is the probability that all teams have exactly 1 student from each field?
- (b) If we randomly pick a group of n students, what is the minimal value of n such that there is $> 75\%$ chance to have 1 student from each field?

Solution.

- (a) The probability is
- $$\frac{(5!)^3 \cdot \frac{1}{5!}}{\binom{15}{3,3,3,3,3} \cdot \frac{1}{5!}} = \frac{72}{7007}.$$
- (b) Perform casework and see that the smallest value of n is 6. \square

Example 1.11 (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. The total number of ways to deal 52 distinct cards evenly to 4 players (13 each) is given by the multinomial coefficient $\binom{52}{13,13,13,13}$. This counts all possible ways to assign 13 cards to each of the 4 players, treating the players as distinct (labelled).

To satisfy the condition that each player gets exactly 3 picture cards, observe that there are 12 picture cards, and we want to distribute them so that each player gets exactly 3. As such, we divide 12 distinct picture cards into 4 labelled players, each getting 3. The number of such ways is $\binom{12}{3,3,3,3}$. The remaining 40 cards (non-picture cards) must also be

dealt evenly: 10 to each player. Again, the number of such ways is $\binom{40}{10,10,10,10}$. Hence, the answer is

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

□

Example 1.12 (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) We are asked to find the probability that exactly two of these pairs are mixed-gender (i.e. consist of 1 man and 1 woman), while the other two pairs are same-gender (i.e. one pair of two men and one pair of two women). We first compute the total number of ways to divide 8 distinct people into 4 unlabelled pairs. The number of such partitions is $\frac{1}{4!} \binom{8}{2,2,2,2}$. To see why, the multinomial coefficient counts the number of ways to divide 8 people into ordered groups of 2 (i.e. 4 labelled pairs), and dividing by $4!$ corrects for the fact that the order among the 4 pairs does not matter.

We leave it to the reader to count the number of favourable outcomes and eventually conclude that the desired probability is

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

- (b) There are a total of $\frac{1}{2!} \binom{8}{4,4}$ outcomes. For the favourable outcomes, it is a simple calculation of $\binom{4}{2}^2$. The desired probability is

$$\frac{\binom{4}{2}^2}{\frac{1}{2!} \binom{8}{4,4}}.$$

□

1.4 Distribution Problems

Many classical counting problems can be framed as the distribution of objects into boxes under various constraints — such as whether the objects or boxes are distinct, and whether each box must receive at least one object. These problems are best understood using tools such as the stars and bars method, multinomial coefficients, and Stirling numbers.

A surprising and powerful link emerges when we observe that many of these distribution problems are equivalent to solving linear Diophantine equations — equations where the variables must take integer values. For example, finding the number of non-negative integer solutions to the equation $x_1 + \dots + x_n = r$ is the same as counting the number of ways to distribute r identical objects into n distinct boxes, possibly with or without restrictions like non-emptiness.

Proposition 1.6 (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 1.13. 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:

$$\star \mid \star \mid \star \star \mid \star \star \mid \star \star$$

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. Note that each bin is non-empty,

so we distribute 5 of the 8 objects into each bin, so each bin receives one object. Now, this becomes the usual stars and bars problem, so the required answer is $\binom{7}{4} = 35$.³ \square

Proposition 1.7 (distribution of distinct objects into distinct boxes). We consider three cases. Let $r, n \in \mathbb{N}$.

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

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

- (ii) **Case 2:** To distribute r distinct objects into n distinct objects 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 $r \geq n$ 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 1.4).

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 (Proposition 1.2), the number of ways is

$$n(n-1)(n-2)\dots(n-r+1)$$

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). \square

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 which involves an asymptotic formula for $n!$ for large values of n . That is,

$$n! \sim \sqrt{2n\pi} \left(\frac{n}{e}\right)^n.$$

Definition 1.3 (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.

³Alternatively, we can directly apply Case 2 of Proposition 1.6.

Proposition 1.8. We have 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 1.9. A useful recurrence relation for $s(r, n)$ 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 (Proposition 1.1). \square

Definition 1.4 (Stirling numbers of the second kind). Given $r, n \in \mathbb{Z}_{\geq 0}$ where $0 \leq n \leq r$, the Stirling 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 1.10. 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. \square

Similar to the Stirling numbers of the first kind, we have a similar recurrence relation for the Stirling numbers of the second kind (Proposition 1.11), which is slightly easier to derive since we are not considering circular permutations.

Proposition 1.11. A useful recurrence relation for $S(r, n)$ 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 1.1). \square

Proposition 1.12 (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

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

Before we discuss the distribution of identical objects into identical boxes, we first define a partition of an integer (Definition 1.5).

Definition 1.5 (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 .

Note that the ordering of the integers in the collection in Definition 1.5 is immaterial since the integers are regarded as identical objects. We denote the number of such distinct partitions by $P(r, n)$.

Proposition 1.13 (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 1.14. 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 (Proposition 1.1), the result follows. \square

We now discuss some applications to linear Diophantine equations. 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 $1 \leq i \leq n$,

$$a_1x_1 + a_2x_2 + \dots + a_nx_n = b.$$

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

$$x_1 + x_2 + \dots + x_r = n \quad \text{where } x_i > 0 \text{ for } 1 \leq i \leq r.$$

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 (recall Proposition 1.6).

Proposition 1.15. There are $\binom{r+n-1}{r-1}$ distinct non-negative integer-valued vectors (x_1, x_2, \dots, x_r) that satisfy the equation

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

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

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

is the same as the number of positive solutions to

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

or equivalently,

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

which is $\binom{r+n-1}{r-1}$. \square

Here is a relatively easy problem.

Example 1.14 (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

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

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

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

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 $11 + 2 \cdot 10 + 2 \cdot 9 + \dots + 2 \cdot 1 = 121$. \square

Example 1.15 (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 $\binom{24}{4} = 10626$. \square

Example 1.16 (Ross p. 31 Question 15). Let $H_k(n)$ be the number of vectors (x_1, \dots, x_k) for which each x_i is a positive integer satisfying $x_1 \leq x_2 \leq \dots \leq x_k \leq n$.

(i) Prove that $H_1(n) = n$, and

$$H_k(n) = \sum_{i=1}^n H_{k-1}(i) \quad \text{for } k \geq 2.$$

(ii) Give a direct combinatorial proof that $H_k(n) = \binom{n+k-1}{k}$.

Solution.

(i) $H_1(n)$ denotes the number of vectors (x_1) for which $1 \leq x_1 \leq n$. There are n choices, so $H_1(n) = n$. We then establish the recurrence relation. Suppose $x_k = i$. Then, (x_1, \dots, x_{k-1}) satisfies $1 \leq x_1 \leq x_2 \leq \dots \leq x_{k-1} \leq i$. Hence, there are $H_{k-1}(i)$ choices for (x_1, \dots, x_{k-1}) . Summing over the possible values of $x_k = i$ from 1 to n gives

$$H_k(n) = \sum_{i=1}^n H_{k-1}(i).$$

- (ii) Observe that specifying a vector $1 \leq x_1 \leq \dots \leq x_k \leq n$ is equivalent to choosing a multiset of size k from $\{1, \dots, n\}$. The number of k -element multisets selected from an n -element ground set is

$$\binom{n+k-1}{k},$$

which yields the required result. \square

CHAPTER 2

Axioms of Probability

2.1 The Kolmogorov Axioms

The basic terminologies of Probability Theory, including experiment, outcomes, sample space, events, have already been 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 (Axiom 2.1) are named after Russian mathematician Andrey Kolmogorov. There are numerous Russian mathematicians who contributed to Probability Theory and Statistics. Some include Andrey Markov, who is known for Markov's inequality and Markov chains, Nikolai Smirnov, for which the Kolmogorov-Smirnov test¹, as well as Pafnuty Chebyshev. Chebyshev's inequality (Theorem 8.2) and the Chebyshev polynomials of the first kind and the second kind are named after him. He also contributed to the much celebrated prime number theorem.

Axiom 2.1 (Kolmogorov axioms).

- (i) **Axiom 1:** For any event A , $0 \leq P(A) \leq 1$.
- (ii) **Axiom 2:** Let S be the sample space. Then, $P(S) = 1$.
- (iii) **Axiom 3 (countable additivity on disjoint sets):** For any sequence of mutually exclusive events A_1, A_2, \dots (i.e. $A_i \cap A_j = \emptyset$ 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 six-sided die 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.

¹This non-parametric test may be covered in ST2132 Mathematical Statistics.

- (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. Since the die ranges from 1-6, by the Kolomogorov axioms (Axiom 2.1),

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

Solving the above system of equations 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. Some students might recognise this setup modelling that of a binomial random variable (see Chapter 4.4).

- (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, the required probability is

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

□

Example 2.3 (Ross p. 68 Question 3). A deck of cards is dealt out.

- (a) What is the probability that the 14th card dealt is an ace?
- (b) What is the probability that the first ace occurs on the 14th card?

Solution.

- (a) Note that a regular deck of cards has 52 cards. We first compute the probability that the 14th card dealt is an ace. By symmetry, any card is equally likely to be dealt, so the required probability is $\frac{4}{52} = \frac{1}{13}$.
- (b) The probability that the first ace occurs on the 14th card is

$$\frac{48}{52} \cdot \frac{47}{51} \cdot \dots \cdot \frac{36}{40} \cdot \frac{4}{39} = \frac{48!/35!}{52!/38!} \cdot 4 = \frac{38 \cdot 37 \cdot 36}{52 \cdot 51 \cdot 50 \cdot 49} \cdot 4 \approx 0.312.$$

□

Example 2.4 (Ross p. 68 Question 5). An ordinary deck of 52 cards is shuffled. What is the probability that the top four cards have

- (a) different denominations?
- (b) different suits?

Solution. We will first compute the total outcomes, which is

$$\binom{52}{4} = 270725.$$

- (a) Let A be the event the top 4 cards will be of different domination, then we have

$$P(A) = \frac{\binom{13}{4} \cdot 4^4}{270725} \approx 0.676.$$

- (b) Let B be the event the top 4 cards will be of different suits, then we have

$$P(B) = \frac{13^4}{270725} \approx 0.105.$$

□

Example 2.5. 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. The probability that player 1 receives exactly three picture cards is

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

The probability that player 2 receives exactly three picture cards is

$$\frac{\binom{9}{3} \binom{30}{10}}{\binom{39}{13}}$$

and so on. As such, the desired probability is

$$\frac{\binom{12}{3} \binom{9}{3} \binom{6}{3} \binom{3}{3} \cdot \binom{40}{10} \binom{30}{10} \binom{20}{10} \binom{10}{10}}{\binom{52}{13} \binom{39}{13} \binom{26}{13} \binom{13}{13}} = \frac{\frac{12!}{(3!)^4} \cdot \frac{40!}{(10!)^4}}{\frac{52!}{(13!)^4}} \approx 0.0324.$$

□

Example 2.6 (Ross p. 69 Question 19). Ten cards are randomly chosen from a deck of 52 cards that consists of 13 cards of each of 4 different suits. Each of the selected cards is put in one of 4 piles, depending on the suit of the card.

- (a) What is the probability that the largest pile has 4 cards, the next largest has 3, the next largest has 2 and the smallest has 1 card?
- (b) What is the probability that two of the piles have 3 cards, one has 4 cards and one has no cards?

Solution.

- (a) Define the sample space to be all ways to choose 10 cards out of 52, which has cardinality $\binom{52}{10}$. We must assign the counts $\{4, 3, 2, 1\}$ to the 4 suits. There are $4! = 24$ ways to assign the counts to the suits. Letting A denote the desired event, one can deduce that

$$P(A) = \frac{\binom{13}{4} \binom{13}{3} \binom{13}{2} \binom{13}{1} \cdot 4!}{\binom{52}{10}}.$$

- (b) Let B denote the desired event. We are now assigning the counts $\{4, 3, 3, 0\}$ to the 4 suits. Since the two 3's are indistinguishable, the number of ways to assign is $\frac{4!}{2!}$. Then, one can deduce that

$$P(B) = \frac{\binom{13}{3} \binom{13}{3} \binom{13}{4} \binom{13}{0} \cdot \frac{4!}{2!}}{\binom{52}{10}}.$$

□

Example 2.7 (Ross p. 69 Question 18). 4 red, 8 blue, and 5 green balls are randomly arranged in a line.

- (a) What is the probability that the first 5 balls are blue?
- (b) What is the probability that none of the first 5 balls is blue?
- (c) What is the probability that the final 3 balls are of different colours?
- (d) What is the probability that all the red balls are together?

Solution.

- (a) The number of arrangements without restriction is $\frac{17!}{4!8!5!}$. We then fix 5 blue balls in front in $\binom{8}{5}$ ways, then arrange the remaining balls in $\frac{12!}{4!3!5!}$ ways. The required probability is

$$\frac{12!8!}{3!17!} = \frac{2}{221}.$$

- (b) To have no blue there, you must choose all five from the 9 non-blue balls (4 red and 5 green), which can be done in $\binom{9}{5}$ ways. So, the desired probability is

$$\frac{\binom{9}{5}}{\binom{17}{5}} = \frac{9}{442}.$$

- (c) The number of arrangements is $3! \cdot \frac{14!}{3!7!4!}$. So, the desired probability is $\frac{4}{17}$.
- (d) The number of ways to arrange such that the red balls are together is $\frac{14!}{8!5!}$. The desired probability is

$$\frac{14!4!8!5!}{8!5!17!} = \frac{1}{170}$$

□

Example 2.8 (Ross p. 69 Question 17). Five balls are randomly chosen without replacement from an urn that contains 5 red, 6 white and 7 blue balls. Find the probability that at least one ball of each colour is chosen.

Solution. The partitions of 5 into three distinct parts are

$$(R, W, B) = (3, 1, 1), (2, 2, 1), (1, 2, 2), (2, 1, 2), (1, 3, 1), (1, 1, 3).$$

So the probability is

$$\frac{\binom{5}{3}\binom{6}{1}\binom{7}{1}}{\binom{18}{5}} + \dots + \frac{\binom{5}{1}\binom{6}{1}\binom{7}{3}}{\binom{18}{5}} \approx 0.707.$$

□

Example 2.9. A committee of 4 persons is to be formed from a group of 5 men and 4 women, among which there is a husband-wife couple.

- (i) How many committees are possible?
- (ii) If the committee must have 2 men and 2 women, how many committees are possible?
- (iii) If the committee must have 2 men and 2 women, and the couple is not allowed to serve together, how many committees are possible?

Solution.

- (i) The number of committees is $\binom{9}{4} = 126$.
- (ii) By the multiplication principle (Proposition 1.2), the number of possible committees is $\binom{5}{2}\binom{4}{2} = 60$.
- (iii) We use the principle of complementation. Consider the case when we choose 2 men and 2 women, where one of the men-women pairs is the husband-wife couple. To do this, there are $\binom{4}{1}\binom{3}{1} = 12$ ways, so the desired number of ways is $60 - 12 = 48$, where we used (ii). □

Example 2.10 (Ross p. 69 Question 8). From a group of 3 first-year students, 4 sophomores, 4 juniors, and 3 seniors, a committee of size 4 is randomly selected. Find the probability that the committee will consist of

- (a) 1 from each class

- (b) 2 sophomores and 2 juniors
- (c) only sophomores or juniors

Solution.

- (a) We will first compute the total outcomes, which is

$$\binom{14}{4} = 1001.$$

Let A be the event 1 from each class is chosen, then we have

$$P(A) = \frac{\binom{3}{1} \cdot \binom{4}{1} \cdot \binom{4}{1} \cdot \binom{3}{1}}{1001} \approx 0.144.$$

- (b) Let B be the event 2 sophomores and 2 juniors are chosen, then we have

$$P(B) = \frac{\binom{4}{2} \cdot \binom{4}{2}}{1001} \approx 0.0360.$$

- (c) Let C be the event only sophomores or juniors chosen, then we have

$$P(C) = \frac{\binom{8}{4}}{1001} = 0.0699$$

□

Example 2.11 (ST2131 AY24/25 Sem 2 Tutorial 2). A closet contains 10 pairs of shoes. If 8 shoes are randomly selected, what is the probability that there will be

- (a) no complete pair;
- (b) exactly 1 complete pair?

Solution.

- (a) From the 10 pairs of shoes, we choose 8. Within each pair, we choose one of the shoes each time so by the multiplication principle (Definition 1.2), there are

$$\binom{10}{8} \cdot 2^8$$

ways to obtain no complete pairs. As such, the desired probability is

$$\frac{\binom{10}{8} \cdot 2^8}{\binom{20}{8}}.$$

- (b) We first fix the pair of shoes that is a complete pair. There are $\binom{10}{1}$ ways to choose such a pair. Thereafter, we have 9 pairs of shoes left and we wish to choose 6.

Within each pair, we choose one of the shoes each time so by the multiplication principle (Definition 1.2), there are

$$\binom{10}{1} \binom{9}{6} \cdot 2^6 \text{ ways}$$

to obtain exactly one complete pair. The desired probability is

$$\frac{\binom{10}{1} \binom{9}{6} \cdot 2^6}{\binom{20}{8}}.$$

□

Example 2.12 (Ross p. 69 Question 12). A basketball team consists of 6 frontcourt and 4 backcourt players. If the players are divided into roommates at random, what is the probability that there will be exactly two roommate pairs made up of a backcourt and a frontcourt player?

Solution. The possible number of pairs is

$$\binom{10}{2} \binom{8}{2} \cdots \binom{2}{2} \cdot \frac{1}{5!} = 945.$$

To have exactly two roommate pairs made up of a backcourt and a frontcourt player, the number of ways is

$$\binom{6}{1} \binom{4}{1} \binom{5}{1} \binom{3}{1} \cdot \binom{4}{2} \binom{2}{2} \binom{2}{2} \cdot \frac{1}{2!2!} = 540.$$

The required probability is $\frac{540}{945} = \frac{4}{7}$.

□

2.2 Probability Properties

Using Kolmogorov's axioms (Axiom 2.1), 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 such that $P(A) + P(A') = 1$. In particular, we shall discuss the principle of inclusion and exclusion (Proposition 2.1).

Proposition 2.1 (principle of inclusion and exclusion). If we have n events A_1, A_2, \dots, A_n ,

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

One would generally be more familiar with the principle of inclusion and exclusion for two events. Say we have two events A and B . Then,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \quad (2.1)$$

This can be illustrated using a Venn diagram (Figure 2.1).

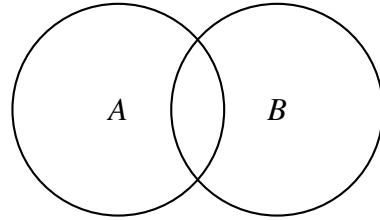


Figure 2.1: Principle of inclusion and exclusion for 2 sets

If we have three events A , B and C , then as shown in Figure 2.2,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C). \quad (2.2)$$

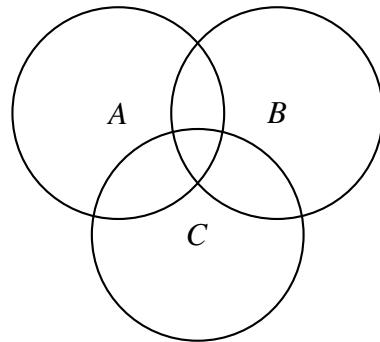


Figure 2.2: Principle of inclusion and exclusion for 3 sets

Example 2.13 (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. We first calculate the probability that a member plays either badminton or squash or both. This can be done using the principle of inclusion and exclusion (Proposition 2.1), which tells us that the probability of a member playing either badminton or squash is equal to the sum of the probabilities of playing each sport individually, minus the probability of playing both sports. By (2.1), the answer is $1 - (0.4 + 0.35 - 0.1) = 0.35$. \square

Example 2.14 (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. It is valid. Let E and F denote the first and second events written above respectively. By the principle of inclusion and exclusion (Proposition 2.1),

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

which implies

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

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

Example 2.15 (ST2131 AY24/25 Sem 1 Lecture 1). Similar to Example 2.14, 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. It is invalid. Again, let E and F denote the first and second events written above respectively. Then, by the principle of inclusion and exclusion (Proposition 2.1),

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

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

Example 2.16. In a large language school where students take classes in Chinese, Japanese, Korean, and other languages,

- 51% of students are enrolled in the Chinese class;
- 40% of students are enrolled in the Japanese class;
- 32% of students are enrolled in the Korean class;
- 14% of students are enrolled in both Chinese and Japanese classes;
- 17% of students are enrolled in both Chinese and Korean classes;
- 10% of students are enrolled in both Japanese and Korean classes;
- 3% of students are enrolled in Chinese, Japanese, and Korean classes

A student of the school is randomly selected. What is the probability that the selected student is enrolled in none of the three languages mentioned above?

Solution. Let C, J, K denote the events that a student is enrolled in the Chinese, Japanese, and Korean class respectively. The trick is to first invoke de Morgan's law before using the principle of inclusion and exclusion. By de Morgan's law,

$$P(C' \cap J' \cap K') = 1 - P(C \cup J \cup K).$$

We then use the principle of inclusion and exclusion (Proposition 2.1), or rather (2.2), so

$$P(C \cup J \cup K) = P(C) + P(J) + P(K) - P(C \cap J) - P(C \cap K) - P(J \cap K) + P(C \cap J \cap K).$$

At this juncture, we realise that the incorporation of the principle of inclusion and exclusion usually involves de Morgan's law right before it. Substituting the relevant probabilities yields

$$P(C \cup J \cup K) = 0.51 + 0.40 + 0.32 - 0.14 - 0.17 - 0.10 + 0.03 = 0.85$$

so the desired probability is 0.15. \square

We then discuss Example 2.17, which involves a classic counting problem in Combinatorics. It is a problem on enumerating surjective functions.

Example 2.17 (enumerating surjective functions; H3 Mathematics 2020). Let

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

and f be a function mapping from X to Y . 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

$$|A'_1 \cap A'_2 \cap \dots \cap A'_n|,$$

which is equivalently, by de Morgan's law,

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

Note that

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

and so on. It is clear that $n(S) = m^n$. Using the general form of the principle of inclusion and exclusion (Proposition 2.1),

$$\left| \bigcup_{i=1}^n A_i \right| = \sum_{i=1}^n |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cap A_j| + \sum_{1 \leq i < j < k \leq n} |A_i \cap A_j \cap A_k| + \dots + \binom{m}{m}(m-m)^n$$

so

$$\left| \bigcup_{i=1}^n A_i \right| = \sum_{r=0}^m (-1)^{r+1} \binom{m}{r} (m-r)^n$$

This implies that

$$|A'_1 \cap A'_2 \cap \dots \cap A'_n| = \sum_{r=0}^m (-1)^r \binom{m}{r} (m-r)^n$$

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 2.18 (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:

$$D_n = (n-1)(D_{n-1} + D_{n-2}) \quad \text{with initial conditions } D_1 = 0 \text{ and } D_2 = 1. \quad (2.3)$$

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. \square

Let us ponder over Example 2.18 further. By repeatedly applying the recurrence relation (2.3) or simply using induction, we can establish that

$$D_n = nD_{n-1} + (-1)^n \quad \text{for } n \geq 2.$$

We can find a formula for D_n in terms of a sum. This involves ingeniously considering a new expression, namely $D_n = n!P_n$. Thus, $n!P_n = n!P_{n-1} + (-1)^n$. By the method of difference for summing telescoping series,

$$P_n = \sum_{i=2}^n \frac{(-1)^i}{i!}$$

In conclusion,

$$D_n = n! \sum_{i=0}^n \frac{(-1)^i}{i!}. \quad (2.4)$$

We can also prove this result by the principle of inclusion and exclusion (Proposition 2.1). In fact, see Examples 2.19 and 2.20 for examples using this method, though the latter may be slightly more complicated. Note that for large values of n , P_n tends to e^{-1} . This can be proven by the series expansion of e^x , namely

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

Example 2.19 (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 $A = E'_1 \cap \dots \cap E'_5$. The complement of A is $A' = E_1 \cup \dots \cup E_5$. By the principle of inclusion and exclusion,

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

\square

Example 2.20 (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 $A = E_1 \cup \dots \cup E_4$. By the principle of inclusion and exclusion (Proposition 2.1), we have

$$P(A) = \sum_{i=1}^4 P(E_i) - \sum_{i < j} P(E_i \cap E_j) + \sum_{i < j < k} P(E_i \cap E_j \cap E_k) - P(E_1 \cap E_2 \cap E_3 \cap E_4).$$

In general, we have

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

Hence,

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

□

We can generalise Example 2.20. 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 $n = 1, \dots, 8$ 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 . We now recall Example 2.17 and explore a nice application of derangements.

Example 2.21 (H3 Mathematics 2020). Let

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

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 . 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 (Proposition 2.1). Let A_i be the set of permutations in which the i^{th} element goes into the correct position, where $1 \leq i \leq 5$. Note that $|A_i| = 4!$, $|A_i \cap A_j| = 3!$ and so on. Using the principle of inclusion and exclusion (Proposition 2.1), the number of derangements, D_5 , is

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

Alternatively, using the derangement formula (2.4) will yield the same result. □

Example 2.22 (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.

To see why, 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,

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

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

$$p(n) = 1 - \frac{365 \cdot 364 \cdot 363 \cdots (366 - n)}{365^n} = 1 - \frac{365!}{365^n (365 - n)!} = 1 - \frac{n!}{365^n} \binom{365}{n}.$$

Example 2.23 (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,

$$P(A'_2) = \frac{7 \cdot 6}{7^2} \quad \text{and} \quad P(A'_3) = \frac{7 \cdot 6 \cdot 5}{7^3}.$$

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

$$P(A'_n) = \frac{7 \cdot \dots \cdot (7 - n + 1)}{7^n}.$$

We accept this without proof (recall our discussion in Example 2.22 for a proof). Then by computation, we find that $n \geq 4$ satisfies the inequality $1 - P(A'_n) > 0.6$. \square

Example 2.24 (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 2.23), 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,

$$P(A'_n) = \frac{12 \cdot \dots \cdot (12 - n + 1)}{12^n} \quad \text{for any } n \leq 13.$$

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

2.3 Probability as a Continuous Set Function

In Probability Theory, events are often organised into sequences that either grow larger (increasing) or shrink (decreasing) as n progresses. Studying such sequences is important because many real-world phenomena are modelled by limits of events. For example, the occurrence of an event eventually or infinitely often. The continuity property of probability ensures that probabilities behave consistently with these limits, i.e. probabilities of sequences converge to the probability of the limiting event. This principle is foundational for more advanced results such as the Borel–Cantelli lemma (you would encounter it in courses like MA4262 Measure and Integration) and the laws of large numbers (Theorems 8.4 on the weak law and 8.5 on the strong law).

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

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

whereas it is a decreasing sequence if

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

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

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

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

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

Example 2.25 (increasing sequence of events). Suppose we repeatedly flip a fair coin, and define E_n to be the event that at least one head appears among the first n coin tosses. Then, $E_1 = \{H \text{ in toss 1}\}$, $E_2 = \{H \text{ in toss 1 or toss 2}\}$ and so on. We can construct an increasing sequence $E_1 \subseteq E_2 \subseteq \dots$. The expression

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

denotes the event that a head eventually appears at some point in the infinite sequence of coin tosses. So,

$$\lim_{n \rightarrow \infty} E_n = \text{at least one head occurs in the infinite sequence.}$$

Since the probability of getting all tails forever is 0, we conclude that

$$P\left(\lim_{n \rightarrow \infty} E_n\right) = 1.$$

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

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

Theorem 2.1 (Boole's inequality). For a countable set of events A_1, A_2, \dots ,

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

The generalisation of Boole's inequality (Theorem 2.1) is Bonferroni's inequality.

Proof. For each $n \geq 1$, the trick is to define

$$B_1 = A_1 \quad \text{and} \quad B_n = A_n \setminus \bigcup_{k=1}^{n-1} A_k \text{ for } n \geq 2.$$

Then the sets B_n are pairwise disjoint and

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i.$$

By countable additivity on disjoint sets (third Kolmogorov axiom in Axiom 2.1), we have

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

But $B_i \subseteq A_i$ for each i , so by monotonicity, $P(B_i) \leq P(A_i)$. Hence,

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

This proves the claim. □

Conditional Probability and Independence

3.1 Conditional Probabilities

In many real-world situations, we often have some partial information about the outcome of a random experiment, and we want to update our beliefs based on this information. This naturally leads to the notion of conditional probability, which tells us how likely an event is to occur given that another event has already occurred.

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

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

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

Example 3.1 (Ross p. 128 Question 6). An urn contains b black balls and r red balls. One of the balls is drawn at random, but when it is put back in the urn, c additional balls of the same colour are put in with it. Now, suppose that we draw another ball. Show that the probability that the first ball was black, given that the second ball drawn was red is

$$\frac{b}{b+r+c}.$$

Solution. Let A denote the event that the first ball was black and B denote the event that the second ball drawn was red. By Definition 3.1,

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

We have

$$P(A \cap B) = \frac{b}{b+r} \cdot \frac{r}{b+r+c}.$$

Next,

$$P(B) = \frac{b}{b+r} \cdot \frac{r}{b+r+c} + \frac{r}{b+r} \cdot \frac{r+c}{b+r+c} = \frac{r}{b+r}.$$

The result follows. \square

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

$$P(A_1 \cap \dots \cap A_n) = P(A_1)P(A_2 | A_1)P(A_3 | A_1 \cap A_2) \dots P(A_n | A_1 \cap \dots \cap A_{n-1}). \quad (3.1)$$

Proof. Apply the definition of conditional probability (Definition 3.1) to the right side of (3.1) to get the following expression:

$$P(A_1) \cdot \frac{P(A_2 \cap A_1)}{P(A_1)} \cdot \frac{P(A_3 \cap A_2 \cap A_1)}{P(A_2 \cap A_1)} \cdot \dots \cdot \frac{P(A_1 \cap \dots \cap A_n)}{P(A_1 \cap \dots \cap A_{n-1})}$$

which is clear that it is equal to the left side of (3.1) after some cancellation. \square

Example 3.2 (ST2131 AY24/25 Sem 2 Tutorial 3). Suppose that an ordinary deck of 52 cards is shuffled and the cards are then turned over one at a time until the first ace appears. Given that the first ace is the 20th card to appear, what is the conditional probability that the card following it is the ace of spades?

Solution. Let B be the event that the 21st card is the ace of spades and A be the event that the 20th card is the first ace. We wish to compute $P(B | A)$. By Definition 3.1, this is equal to

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

$P(B \cap A)$ is the probability that the 20th card is the first ace and the 21st card is the ace of spades. We proceed with casework. If the 20th card is the ace of spades, then this contributes 0 to the probability. On the other hand, if the 20th card is some other type of ace (i.e. diamonds, clubs, hearts), then the first 19 cards must be non aces. Hence, this contributes a probability of

$$\frac{48}{52} \cdot \frac{47}{51} \cdot \dots \cdot \frac{30}{34}.$$

For the 20th card to be an ace but of some other type, we then multiply by $3/33$. Lastly, for the 21st card to be the ace of spades, we multiply by $1/32$. As such,

$$P(B \cap A) = 0 + \left(\frac{48}{52} \cdot \frac{47}{51} \cdot \dots \cdot \frac{30}{34} \right) \cdot \frac{3}{33} \cdot \frac{1}{32}.$$

$P(A)$ is the probability that the 20th card is the first ace, which is

$$\left(\frac{48}{52} \cdot \dots \cdot \frac{30}{34} \right) \cdot \frac{4}{33}.$$

Hence, $P(B | A) = 3/128$. \square

Example 3.3. From a standard deck of cards, we take the twelve picture cards (i.e. the J, Q, K of each suit) and leave out the other cards. These 12 cards are then shuffled at random and dealt out to 4 players (each getting 3 cards).

- (a) What is the probability that the king of spades and the king of clubs are with different players?
- (b) What is the probability that the king of spades, the king of clubs and the king of hearts are with different players?
- (c) What is the probability that the four king cards are with different players?

Solution.

- (a) Think of the 12 cards arranged as 4 hands of 3 cards each as follows:

$$(A_1, A_2, A_3), \dots, (D_1, D_2, D_3)$$

Place K_{\spadesuit} anywhere, say in a slot belonging to A . Then, K_{\spadesuit} can go into any of B_1, \dots, D_3 , which has 9 favourable outcomes. So, the required probability is $\frac{9}{11}$.

- (b) We first keep K_{\spadesuit} and K_{\clubsuit} in different hands with probability $\frac{9}{11}$. Given that there are 10 slots left, the two used hands have $2 + 2 = 4$ slots, so the other By Proposition 3.1),

$$P(K_{\heartsuit} \text{ in a third hand} \mid \text{first two different}) = \frac{6}{10} = \frac{3}{5}.$$

Hence, the probability required is

$$\frac{9}{11} \cdot \frac{3}{5} = \frac{27}{55}.$$

- (c) Repeat this one last time — with three kings all in distinct hands, the remaining slots are 9 total, which the one untouched hand has 3. Thus, by Proposition 3.1),

$$P(K_{\diamondsuit} \text{ in fourth hand} \mid \text{first three different}) = \frac{3}{9} = \frac{1}{3}.$$

The required probability is

$$\frac{9}{11} \cdot \frac{3}{5} \cdot \frac{1}{3} = \frac{9}{55}.$$

□

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, \dots be a sequence of mutually exclusive events. Then,

$$P(B_1 \cup B_2 \cup \dots \mid A) = \sum_{i=1}^{\infty} P(B_i \mid A).$$

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 Definition 3.1 that

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

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

Lastly, for (iii),

$$P(B_1 \cup B_2 \cup \dots | A) = \frac{1}{P(A)} \cdot P(A \cap (B_1 \cup B_2 \cup \dots)) = \frac{1}{P(A)} \sum_{i=1}^{\infty} P(B_i \cap A)$$

which simplifies to the desired result. \square

We now discuss an interesting concept in Probability known as Penny's game. This refers to a two-player combinatorial game in which one player, X , selects a sequence of heads and tails of a given length (i.e. three coin flips), and the other player, B , selects a different sequence of the same length. A fair coin is then flipped repeatedly, and the first player whose sequence appears in the sequence of flips wins.

This game is particularly interesting because it demonstrates a surprising *non-transitive property*. That is, for any sequence chosen by A, B can always pick a sequence that has a higher probability of appearing first, despite both sequences having the same length. The optimal strategy in Penny's game relies on overlapping patterns in sequences.

Example 3.4 (Penny's game). Say we are given the following 8 patterns:

$$HHH, HHT, HTH, THH, HTT, THT, TTH, TTT$$

Player X picks one of the patterns, and Player Y then picks one of the remaining 7 patterns. A fair coin is tossed repeatedly until either Player X 's pattern appears (X wins) or Player Y 's pattern appears (Y wins).

- (i) Find $P(X \text{ wins})$ if X picks HHT and Y picks THH .
- (ii) Find $P(X \text{ wins})$ if X picks HTH and Y picks HHT .

Solution. Note that the solution to this problem implicitly uses the law of total probability (Proposition 3.3), but we will eventually state it in due course.

(i) Let

$$p_X = \text{probability } X \text{ wins}$$

$$p_H = \text{probability } X \text{ wins given that previous outcome was } H$$

$$p_{HT} = \text{probability } X \text{ wins given that previous outcomes were } H \text{ then } T$$

$$\vdots = \vdots$$

We note that

$$p_X = \frac{1}{2}p_H + \frac{1}{2}p_T \quad \text{and} \quad p_H = \frac{1}{2}p_{HH} + \frac{1}{2}p_{HT}.$$

Also, $p_{HT} = p_T$. To see why, the first two outcomes that X and Y pick are HH are TH respectively. Since neither outcome is HT , then we can effectively *ignore* the first outcome, which is H , and analyse the remaining sequence from that point onward. Hence,

$$p_H = \frac{1}{2}p_{HH} + \frac{1}{2}p_T = \frac{1}{2} + \frac{1}{2}\left(\frac{1}{2}p_{TH} + \frac{1}{2}p_{TT}\right) \quad \text{since } p_{HH} = 1$$

so

$$p_H = \frac{1}{2} + \frac{1}{4}p_{TH} + \frac{1}{4}p_{TT}. \quad (3.2)$$

We shall talk our way through these coloured probabilities. Before that, note we mentioned that $p_{HH} = 1$. This is because

$$p_{HH} = \frac{1}{2}p_{HHH} + \frac{1}{2}p_{HHT} = \frac{1}{2}p_{HH} + \frac{1}{2} \cdot 1 = \frac{1}{2} + \frac{1}{2}p_{HH}.$$

Here, we used the fact that $p_{HHT} = 1$. As such, $p_{HH} = 1$. Intuitively, once X obtains the sequence HH , in the long run, no matter the subsequent outcomes, there exists a sequence of HHT .

- We have $p_{TH} = \frac{1}{2}p_{THT} + \frac{1}{2}p_{THH}$. Since Y wins (so X loses) upon obtaining the sequence THH , then $p_{THH} = 0$. Also, $p_{THT} = p_{TH} = p_T$. As such, $p_{TH} = \frac{1}{2}p_T$.
- We have

$$p_{TT} = \frac{1}{2}p_{TTT} + \frac{1}{2}p_{TTH} = \frac{1}{2}p_T + \frac{1}{2}p_{TH} = \frac{1}{2}p_T + \frac{1}{4}p_T = \frac{3}{4}p_T$$

Substituting these into (3.2),

$$p_H = \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2}p_T + \frac{1}{4} \cdot \frac{3}{4}p_T = \frac{1}{2} + \frac{5}{16}p_T.$$

We leave it to the reader to prove that Note that

$$p_T = \frac{1}{2}p_{TH} + \frac{1}{2}p_{TT} = \frac{5}{8}p_T$$

which involves many intermediate steps. So, $p_T = 0$. As such,

$$p_H = \frac{1}{2} + \frac{5}{16} \cdot 0 = \frac{1}{2}$$

so we conclude that $p_X = 1/4$.

- (ii) We use the same notation in (i). Similarly, we have

$$p_X = \frac{1}{2}p_H + \frac{1}{2}p_T.$$

Hence,

$$p_X = \frac{1}{2} \left(\frac{1}{2} p_{HH} + \frac{1}{2} p_{HT} \right) + \frac{1}{2} \left(\frac{1}{2} p_{TH} + \frac{1}{2} p_{TT} \right)$$

Note that $p_{HH} = 0$ since no matter what sequence of outcomes appear after HH , there eventually exists a sequence of HHT , making Y win. So,

$$\begin{aligned} p_X &= \frac{1}{4} p_{HT} + \frac{1}{4} p_{TH} + \frac{1}{4} p_{TT} \\ &= \frac{1}{4} \left(\frac{1}{2} p_{HTH} + \frac{1}{2} p_{HTT} \right) + \frac{1}{4} \left(\frac{1}{2} p_{THH} + \frac{1}{2} p_{THT} \right) + \frac{1}{4} \left(\frac{1}{2} p_{TTH} + \frac{1}{2} p_{TTT} \right) \\ &= \frac{1}{4} \left(\frac{1}{2} \cdot 1 + \frac{1}{2} p_T \right) + \frac{1}{4} \left(\frac{1}{2} \cdot 0 + \frac{1}{2} p_{HT} \right) + \frac{1}{4} \left(\frac{1}{2} p_H + \frac{1}{2} p_T \right) \\ &= \frac{1}{8} + \frac{1}{4} p_T + \frac{1}{8} p_{HT} + \frac{1}{8} p_H \end{aligned}$$

Note that

$$p_{HT} = \frac{1}{2} p_{HTH} + \frac{1}{2} p_{HTT} = \frac{1}{2} p_H + \frac{1}{2} \cdot 1 = \frac{1}{2} p_H + \frac{1}{2}.$$

So,

$$p_X = \frac{1}{8} + \frac{1}{4} p_T + \frac{1}{8} \left(\frac{1}{2} p_H + \frac{1}{2} \right) + \frac{1}{8} p_H.$$

In the process, we deduced that $p_H = p_T = 1/3$ (check this). These imply $p_H = 2/7$ and $p_T = 1/7$. As such, $p_X = 1/3$. \square

3.2 Bayes' Theorem

In many practical problems, we are often interested in reversing conditional probabilities. That is, rather than computing the probability of an event B given that another event A has occurred, we wish to determine the probability of A given that B has been observed. Situations of this nature arise frequently in medical testing, quality control, etc. where the observable evidence is the effect and the underlying cause is unknown.

Bayes' theorem (Theorem 3.1) provides a systematic way to update probabilities when new information becomes available. It relates forward probabilities such as $P(B | A)$ to reverse probabilities such as $P(A | B)$.

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. This is a direct application of Definition 3.1. \square

We introduce the notion of the partition of a sample space S . We say that A_1, A_2, \dots, A_n are partitions of S if they are mutually exclusive and collectively exhaustive. The term *mutually exclusive* is studied at both O-Level and A-Level Mathematics. It simply means

$$A_i \cap A_j = \emptyset \quad \text{for all } i \neq j.$$

In relation to probabilities,

$$P(A \cup B) = P(A) + P(B) \quad \text{or equivalently } P(A \cap B) = 0.$$

The term *collectively exhaustive* means that

$$\bigcup_{i=1}^n A_i = S.$$

Example 3.5 (Ross p. 128 Question 8).

(a) Show that

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

(b) Suppose that before new evidence is observed, the hypothesis H is three times as likely to be true as is the hypothesis G . If the new evidence is twice as likely when G is true than it is when H is true, which hypothesis is more likely after the evidence has been observed?

Solution. (a) is trivial by repeatedly applying the definition of conditional probability (Definition 3.1). We will only solve (b). It is given that

$$\frac{P(H)}{P(G)} = 3 \text{ before new evidence} \quad \text{and} \quad \frac{P(E | H)}{P(E | G)} = \frac{1}{2} \text{ after new evidence.}$$

By (a),

$$\frac{P(H | E)}{P(G | E)} = \frac{P(H)}{P(G)} \cdot \frac{P(E | H)}{P(E | G)} = 3 \cdot \frac{1}{2} > 1$$

which implies that H is still more likely after observing the evidence. □

We now introduce an important result known as the law of total probability (Proposition 3.3), which allows us to compute the probability of an event B by conditioning on a set of mutually exclusive and collectively exhaustive events. The idea is that if the sample space can be partitioned into disjoint events A_1, A_2, \dots, A_n — each representing a possible underlying cause, then the total probability of B can be computed by summing over the conditional probabilities of B given each A_i , weighted by the probabilities of each A_i occurring. This result is especially useful when the direct computation of $P(B)$ is difficult, but the conditional probabilities $P(B | A_i)$ are more accessible.

Proposition 3.3 (law of total probability). Suppose the events A_1, A_2, \dots, 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,

$$P(B) = P(B | A_1)P(A_1) + P(B | A_2)P(A_2) + \dots + P(B | A_n)P(A_n).$$

Example 3.6 (Ross p. 128 Question 11). A type C battery is in working condition with probability 0.7, whereas a type D battery is in working condition with probability 0.4. A battery is randomly chosen from a bin consisting of 8 type C and 6 type D batteries.

- (a) What is the probability that the battery works?
- (b) Given that the battery does not work, what is the conditional probability that it was a type C battery?

Solution.

- (a) Let W be the event that the battery works. By the law of total probability (Proposition 3.3),

$$P(W) = P(W | C)P(C) + P(W | D)P(D) = 0.7 \cdot \frac{8}{14} + 0.4 \cdot \frac{6}{14} = \frac{4}{7}.$$

- (b) By Definition 3.1, we have

$$P(C | W') = \frac{P(C \cap W')}{1 - P(W)} = \frac{P(W' | C)P(C)}{1 - 4/7}.$$

□

Example 3.7. Identical twins are always of the same sex. Fraternal twins are equally likely to be of the same sex as to be of different sex. Data collected in a city shows that 55% of twins are of the same sex. What is the probability that a randomly selected twin in the city is a pair of identical twins?

Solution. Let A be the event that the twins are of same sex, and I be the event that the twins are identical. So, $P(I) = p$, $P(I') = 1 - p$, $P(A) = 0.55$, $P(A | I) = 1$, $P(A | I') = 0.5$. By the law of total probability (Proposition 3.3),

$$P(A) = P(A | I)P(I) + P(A | I')P(I').$$

Substituting the probabilities yields $p = 0.1$. That is, there is a 0.1 probability that a randomly selected twin in the city is a pair of identical twins. □

Example 3.8 (Ross p. 128 Question 9). You ask your neighbour to water a sickly plant while you are on vacation. Without water, it will die with probability 0.8; with water, it will die with probability 0.15. You are 90 percent certain that your neighbour will remember to water the plant.

- (a) What is the probability that the plant will be alive when you return?
- (b) If the plant is dead upon your return, what is the probability that your neighbour forgot to water it?

Solution:

- (a) Let event A be the event that the plant is alive and event B denote that the plant was watered. Then, by the law of total probability (Proposition 3.3), we have

$$P(A) = P(A | B)P(B) + P(A | B')P(B') = 0.05 \cdot 0.9 + 0.2 \cdot 0.1 = 0.785.$$

(b) By Definition 3.1, we have

$$P(B' | A') = \frac{P(A' | B')}{P(A')} = \frac{P(A' | B') P(B')}{1 - P(A)} = \frac{0.8 \cdot 0.1}{1 - 0.785}.$$

□

To apply Bayes' theorem effectively, it is often helpful to break the sample space into simpler, mutually exclusive scenarios. As such, we state Corollary 3.1, which can be seen as a consequence of Bayes' theorem (Theorem 3.1) and the law of total probability (Proposition 3.3).

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

$$P(A_i | B) = \frac{P(B | A_i) P(A_i)}{P(B | A_1) P(A_1) + \dots + P(B | A_n) P(A_n)}. \quad (3.3)$$

Example 3.9. The police in a city administer a breath analyzer test to catch drunk drivers. A drunk driver taking the test will always get a positive result (for alcohol). However, 5% of sober drivers taking the test will still get a positive result. One in a thousand drivers in that city drive while they are drunk. The police stops a driver at random and administers the test. The test gives a positive result. What is the probability that this driver is drunk?

Solution. Let A be the event that the driver is drunk and B be the event that we have a positive result. By applying Bayes' theorem (Theorem 3.1) and the law of total probability (Proposition 3.3) in succession, we have

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

Observe that this matches the expression in (3.3). Substituting the respective probabilities, we have

$$\frac{1 \cdot \frac{1}{1000}}{1 \cdot \frac{1}{1000} + 0.05 \cdot \frac{999}{1000}} = 0.02.$$

□

Example 3.10 (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 programme?
- (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),

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

- (b) We use the definition of conditional probability on the event $P(S | \star)$ (Definition 3.1). Hence,

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

□

Example 3.11 (Ross p. 129 Question 14). A coin having probability 0.8 of landing on heads is flipped. A observes the result — either heads or tails, and rushes off to tell B . However, with probability 0.4, A will have forgotten the result by the time he reaches B . If A has forgotten, then rather than admitting this to B , he is equally likely to tell B that the coin landed on heads or that it landed tails. If he does remember, then he tells B the correct result.

- (a) What is the probability that B is told that the coin landed on heads?
- (b) What is the probability that B is told the correct result?
- (c) Given that B is told that the coin landed on heads, what is the probability that it did in fact land on heads?

Solution:

- (a) Let H be the event the coin is heads, T be the event the coin is tails and F be the event A forgets and R be the event B is told heads. Then, $P(H) = 0.8$, $P(T) = 0.2$, $P(F) = 0.4$, and $P(F') = 0.6$. By the law of total probability (Proposition 3.3),

$$\begin{aligned} P(R) &= P(R | H)P(H) + P(R | T)P(T) \\ &= (0.6 \cdot 1 + 0.4 \cdot 0.5)(0.8) + (0.6 \cdot 0 + 0.4 \cdot 0.5)(0.2) \end{aligned}$$

which evaluates to 0.68.

- (b) Let C be the event that B is told the correct result. By the law of total probability (Proposition 3.3),

$$\begin{aligned} P(C) &= P(C | H)P(H) + P(C | T)P(T) \\ &= (0.6 \cdot 1 + 0.4 \cdot 0.5)(0.8) + (0.6 \cdot 1 + 0.4 \cdot 0.5)(0.8) \end{aligned}$$

which evaluates to 0.8.

(c) By Bayes' theorem (Theorem 3.1),

$$P(H | R) = \frac{P(R | H)P(H)}{P(R)} = \frac{(0.6 \cdot 1 + 0.4 \cdot 0.5)(0.8)}{0.68}$$

which evaluates to 0.941. \square

Example 3.12 (Ross p. 125 Question 4). A ball is in any one of n boxes and is in the i^{th} box with probability P_i . If the ball is in box i , a search of that box will uncover it with probability α_i . Show that the conditional probability that the ball is in the box j , given that a search of box i did not uncover it, is

$$\frac{P_j}{1 - \alpha_i P_i} \text{ if } j \neq i \quad \text{and} \quad \frac{(1 - \alpha_i) P_i}{1 - \alpha_i P_i} \text{ if } j = i.$$

Solution. Let B_j be the event that ball is in box j and N_i be the event that search box i does not uncover the ball. Then

$$P(N_i | B_j) = \begin{cases} 1 - \alpha_i & \text{if } j = i; \\ 1 & \text{if } j \neq i. \end{cases}$$

By the law of total probability (Proposition 3.3), we have

$$P(N_i) = \sum_{j=1}^n P(N_i | B_j) P(B_j) = (1 - \alpha_i) P_i + \sum_{j \neq i} P(N_i | B_j) P(B_j)$$

Note that

$$\sum_{j \neq i} P(N_i | B_j) P(B_j) = \sum_{j \neq i} P(B_j) = 1 - P_i.$$

Hence,

$$P(N_i) = (1 - \alpha_i) P_i + 1 - P_i = 1 - \alpha_i P_i.$$

We then use Bayes' theorem (Theorem 3.1) which states that

$$P(B_j | N_i) = \frac{P(N_i | B_j) P(B_j)}{P(N_i)}.$$

If $j \neq i$, then

$$P(B_j | N_i) = \frac{P_j}{1 - \alpha_i P_i}.$$

On the other hand, if $j = i$, then

$$P(B_j | N_i) = \frac{(1 - \alpha_i) P_i}{1 - \alpha_i P_i}$$

which completes the proof. \square

Example 3.13 (Ross p. 129 Question 19). Three players simultaneously toss coins. The coin tossed by A, B, C turns up heads with respective probabilities P_1, P_2, P_3 . If one person gets an outcome different from those of the other two, then he is the odd man out. If there is no odd man out, the players flip again and continue to do so until they get an odd man out. What is the probability that A will be the odd man?

Solution. At each stage, everyone can obtain heads with probability $P_1P_2P_3$ or tails with probability $Q_1Q_2Q_3$, where $Q_i = 1 - P_i$. Let $k \in \mathbb{N}$ be arbitrary. Consider a finite run of length $3k$ consisting of blocks of the form $\boxed{P_1P_2P_3}$ and $\boxed{Q_1Q_2Q_3}$. We can then arrange the blocks as follows:

$$\underbrace{\boxed{P_1P_2P_3} \boxed{Q_1Q_2Q_3} \boxed{Q_1Q_2Q_3} \dots \boxed{Q_1Q_2Q_3} \boxed{P_1P_2P_3}}_{\text{run of length } 3n}$$

So, the probability required is

$$\sum_{k=0}^{\infty} (P_1P_2P_3 + Q_1Q_2Q_3)^k (Q_1P_2P_3 + P_1Q_2Q_3) = \frac{Q_1P_2P_3 + P_1Q_2Q_3}{1 - P_1P_2P_3 - Q_1Q_2Q_3}.$$

□

Example 3.14 (Ross p. 129 Question 21). If A flips $n+1$ fair coins and B flips n fair coins, what is the probability that A gets more heads than B ? A hint is to condition on which player has more heads after each has flipped n coins.

Solution. We shall proceed using brute force — that is we will ignore the hint. If A obtains k heads, then we wish to find the probability that B gets either $1, \dots, k-1$ heads, where $1 \leq k \leq n+1$. We have

$$P(A \text{ obtains } k \text{ heads}) = \binom{n+1}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n+1-k} \binom{n+1}{k} \frac{1}{2^{n+1}}.$$

Then,

$$P(B \text{ obtains } 1, \dots, k-1 \text{ heads}) = \sum_{j=1}^{k-1} \binom{n}{j} \left(\frac{1}{2}\right)^j \left(\frac{1}{2}\right)^{n-j} = \sum_{j=1}^{k-1} \binom{n}{j} \frac{1}{2^n}.$$

By independence and taking the sum over all $1 \leq k \leq n+1$, the required probability is

$$\sum_{k=1}^{n+1} \sum_{j=1}^{k-1} \binom{n+1}{k} \binom{n}{j} \frac{1}{2^{2n+1}} = \frac{1}{2^{2n+1}} \sum_{k=1}^{n+1} \binom{n+1}{k} \sum_{j=1}^{k-1} \binom{n}{j} \quad (3.4)$$

Recall Pascal's identity (Theorem 1.1), which states that

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

This can be easily proven by writing each binomial coefficient out using factorials. Hence, (3.4) becomes

$$\frac{1}{2^{2n+1}} \sum_{k=1}^{n+1} \left(\binom{n}{k} + \binom{n}{k-1} \right) \sum_{j=0}^{k-1} \binom{n}{j}.$$

Upon expansion, we obtain

$$\frac{1}{2^{2n+1}} \sum_{k=1}^{n+1} \sum_{j=0}^{k-1} \binom{n}{k} \binom{n}{j} + \frac{1}{2^{2n+1}} \sum_{k=1}^{n+1} \sum_{j=0}^{k-1} \binom{n}{k-1} \binom{n}{j}. \quad (3.5)$$

Note that $\binom{n}{n+1} = 0$. We write the sum in teal as

$$\sum_{0 \leq j < k \leq n} a_k a_j \quad \text{where } a_r = \binom{n}{r} \text{ for } 0 \leq r \leq n.$$

Note the identity

$$\sum_{0 \leq j < k \leq n} a_j a_k = \frac{1}{2} \left(\left(\sum_{r=0}^n a_r \right)^2 - \sum_{r=0}^n a_r^2 \right).$$

Now

$$\sum_{r=0}^n \binom{n}{r} = 2^n \quad \text{and} \quad \sum_{r=0}^n \binom{n}{r}^2 = \binom{2n}{n}.$$

Here, we used (1.4) and Corollary 1.3 respectively. Hence,

$$\sum_{k=1}^{n+1} \sum_{j=0}^{k-1} \binom{n}{k} \binom{n}{j} = \frac{1}{2} \left(4^n - \binom{2n}{n} \right).$$

In a similar fashion,

$$\sum_{k=1}^{n+1} \sum_{j=0}^{k-1} \binom{n}{k-1} \binom{n}{j} = \frac{1}{2} \left(4^n + \binom{2n}{n} \right).$$

Substituting everything into (3.5), we see that the required probability is

$$\frac{1}{2^{2n+1}} \left(\frac{2 \cdot 4^n}{2} \right) = \frac{1}{2}.$$

□

3.3 Independent Events

Independence is a fundamental concept that describes the absence of influence between events. Intuitively, two events are said to be independent if the occurrence of one does not alter the likelihood of the other. For instance, when tossing two fair coins, the result of the first toss does not affect the outcome of the second; the events are independent. We now formally introduce the idea of independent events (Definition 3.2).

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

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

If equality does not hold, then A and B are dependent.

In relation to conditional probability, suppose $P(B) > 0$. If A and B are independent, then $P(A | B) = P(A)$ if knowledge that B has occurred does not change the probability that A occurs. Again, this follows by the definition of conditional probability (Definition 3.1).

Example 3.15 (Ross p. 125 Question 9). In each of n independent tosses of a coin, the coin lands on heads with probability p . How large does n need to be so that the probability of obtaining at least one head is at least $\frac{1}{2}$?

Solution. We must have

$$P(\text{at least one head}) \geq \frac{1}{2} \quad \text{so} \quad P(\text{all tails}) \leq \frac{1}{2}.$$

So, $(1-p)^n \leq \frac{1}{2}$, which implies $n \geq -\frac{\ln 2}{\ln(1-p)}$. \square

Example 3.16 (Ross p. 192 Question 59). How many people are needed so that the probability that at least one of them has the same birthday as you is greater than $\frac{1}{2}$?

Solution. Assume that the birthdays are independent and uniform over 365 days (ignoring leap years). Let n be the sample size and consider the probability that none shares the same birthday as me, which is

$$\left(\frac{364}{365}\right)^n.$$

As such, we want

$$1 - \left(\frac{364}{365}\right) > \frac{1}{2} \quad \text{so} \quad n > \frac{\ln \frac{1}{2}}{\ln \frac{364}{365}} = 252.65$$

Hence, $n = 253 + 1 = 254$ (including myself). \square

Example 3.17 (Ross p. 129 Question 20). Suppose that there are n possible outcomes of a trial, with outcome i resulting with probability p_i for $i \in \{1, \dots, n\}$ and

$$\sum_{i=1}^n p_i = 1.$$

If two independent trials are observed, what is the probability that the result of the second trial is larger than that of the first?

Solution. Let X_1 and X_2 denote the two events. We have

$$P(X_2 > X_1) = P(X_1 > X_2)$$

by symmetry. Then, note that

$$P(X_1 > X_2) = \sum_{i < j} P(X_1 = j \text{ and } X_2 = i) = \sum_{i=1}^n \sum_{j=1}^n p_i p_j.$$

Here, we used the fact that X_1 and X_2 are independent events. Moreover,

$$P(X_1 = X_2) = \sum_{i=1}^n P(X_1 = i \text{ and } X_2 = i) = \sum_{i=1}^n p_i^2.$$

Since

$$P(X_1 > X_2) + P(X_2 > X_1) + P(X_1 = X_2) = 1,$$

then

$$P(X_2 > X_1) = \frac{1}{2} \left(1 - \sum_{i=1}^n p_i^2\right).$$

\square

Example 3.18 (ST2131 AY24/25 Sem 2 Tutorial 4). Prove that if E_1, \dots, E_n are independent events, then

$$P(E_1 \cup \dots \cup E_n) = 1 - \prod_{i=1}^n [1 - P(E_i)].$$

Solution. Recall de Morgan's law, which states that

$$P(A_1 \cup A_2) = 1 - P(A'_1 \cap A'_2).$$

We can generalise this to n sets, i.e.

$$P(E_1 \cup \dots \cup E_n) = 1 - P(E'_1 \cap \dots \cap E'_n) = 1 - \prod_{i=1}^n P(E'_i)$$

and the result follows, where we used the fact that E_1, \dots, E_n are independent (to be precise, mutually independent; see Definition 3.3) in the last equality. \square

We shall now discuss pairwise independence and mutual independence (Definition 3.3).

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

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

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.

Example 3.19 (Ross p. 129 Question 23). Let A and B be events having positive probability. State whether each of the following statements is necessarily true, necessarily false, or possibly true.

- (a) If A and B are mutually exclusive, then they are independent
- (b) If A and B are independent, then they are mutually exclusive
- (c) $P(A) = P(B) = 0.6$, and A and B are mutually exclusive
- (d) $P(A) = P(B) = 0.6$ and A and B are independent

Solution. Suppose $P(A) > 0$ and $P(B) > 0$.

- (a) We are given that A and B are mutually exclusive, so $P(A \cup B) = P(A) + P(B)$. By the principle of inclusion and exclusion (Proposition 2.1), $P(A \cap B) = 0$. For A and B to be independent, we must have $P(A \cap B) = P(A)P(B)$, but this forces either $P(A)$ or $P(B)$ to be zero, which is a contradiction. So, the statement is necessarily false.

- (b) We have $P(A \cap B) = P(A)P(B)$. Same as (a), this leads to a contradiction so the statement is necessarily false.
- (c) By the principle of inclusion and exclusion (Proposition 2.1), $P(A \cup B) = 1.2 - P(A \cap B)$. For A and B to be mutually inclusive, $P(A \cap B) = 0$, but this implies that $P(A \cup B) = 1.2$, which is a contradiction. Hence, A and B are not mutually exclusive. So, the statement is necessarily false.
- (d) Let A and B both denote the event that a biased coin, which shows heads with probability 0.6, lands on heads. Then $P(A) = P(B) = 0.6$ but the events are dependent. Hence, the statement is necessarily false. \square

Example 3.20 (Ross p. 125 Question 9). Consider two independent tosses of a fair coin. Let A be the event that the first toss results in heads, let B be the event that the second toss results in heads, and let C be the event that in both tosses, the coin lands on the same side. Show that the events A, B, C are pairwise independent — that is A and B are independent, A and C are independent, and B and C are independent, but A, B, C are not independent.

Solution. We have $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$, and $P(C) = \frac{1}{2}$. For $A \cap B$, is the event that the first and second toss result in heads, which has probability $\frac{1}{4}$. So, A and B are independent. Next, $A \cap C$ is again the event that the first and second toss result in heads, which has probability $\frac{1}{4}$. Lastly, $B \cap C$ is also the event that the first and second toss result in heads, which has probability $\frac{1}{4}$. Indeed, A and B are independent, A and C are independent, and B and C are independent.

Note that $A \cap B \cap C$ denotes the event that both the first and second toss result in heads, which has probability $\frac{1}{4}$. However, $P(A \cap B \cap C)$ is not equal to $P(A)P(B)P(C)$. \square

Example 3.21 (Ross p. 129 Question 22). Prove or give counterexamples to the following statements:

- (a) If E is independent of F and E is independent of G , then E is independent of $F \cup G$
- (b) If E is independent of F and E is independent of G and $FG = \emptyset$, then E is independent of $F \cup G$
- (c) If E is independent of F , and F is independent of G , and E is independent of FG , then G is independent of EF

Solution.

- (a) The statement is false. Let $S = \{1, 2, 3, 4\}$. Define $E = \{1, 2\}$, $F = \{1, 3\}$, and $G = \{1, 4\}$. Hence, $E \cap F = \{1\}$ and $E \cap G = \{1\}$. By computation, we see that

$$P(E \cap F) = P(E)P(F) \quad \text{and} \quad P(E \cap G) = P(F \cap G).$$

However, $F \cup G = \{1, 3, 4\}$, which implies $E \cap (F \cup G) = \{1\}$. This shows that

$$P(E \cap (F \cup G)) \neq P(E)P(F \cup G).$$

- (b)** The statement is true. Suppose E is independent of F , and E is independent of G , and $F \cap G = \emptyset$. By distributivity of union over intersection, we have

$$\begin{aligned} P(E \cap (F \cup G)) &= P((E \cap F) \cup (E \cap G)) \\ &= P(E \cap F) + P(E \cap G) - P(E \cap F \cap G) \\ &= P(E)P(F) + P(E)P(G) \\ &= P(E)(P(F) + P(G) - P(F \cap G)) \end{aligned}$$

which is equal to $P(E)P(F \cup G)$. This implies that E is independent of $F \cup G$.

- (c)** Suppose E is independent of F , F is independent of G , and E is independent of $F \cap G$. By the associativity and commutativity of intersection, we have

$$\begin{aligned} P(G \cap (E \cap F)) &= P(E \cap (F \cap G)) \\ &= P(E)P(F \cap G) \\ &= P(E)P(F)P(G) \end{aligned}$$

which is equal to $P(G)P(E \cap F)$. This shows that G is independent of $E \cap F$. \square

The gambler's ruin problem (Example 3.22) states that a gambler playing a game with negative expected value will eventually go broke, regardless of their betting system.

Example 3.22 (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 win¹?

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

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

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

$$px_{j+1} - (p+q)x_j + qx_{j-1} = 0 \quad \text{where } x_0 = 0 \text{ and } x_N = 1 \tag{3.6}$$

This is known 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

¹The interested reader can visit this link for more information on this interesting concept of the gambler's ruin problem.

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 in (3.6) 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. \square

Example 3.23 (gambler's ruin problem). The gambler's ruin problem is a fundamental question in Probability Theory and it can be stated as follows. A person is playing a game where they repeatedly win or lose \$1, stopping to play either when they have no money left or when they reach their pre-determined goal. We assume here that the probability to win or lose \$1 is the same, equal to $\frac{1}{2}$.

- (a) Assuming that they start with \$2 and have a goal of \$4, how long on average do they play this game?
- (b) What happens if they instead start with \$4 and have a goal of \$8?

Solution.

- (a) Let E_i denote the expected number of steps until absorption starting from i , where $i = 0, 1, 2, 3, 4$. Note the boundary conditions $E_0 = 0$ and $E_4 = 0$. Next, for the non-absorbing states, we have

$$E_i = 1 + \frac{1}{2}E_{i+1} + \frac{1}{2}E_{i-1}.$$

As such,

$$\begin{aligned} E_1 &= 1 + \frac{1}{2}E_0 + \frac{1}{2}E_2 \\ E_2 &= 1 + \frac{1}{2}E_1 + \frac{1}{2}E_3 \\ E_3 &= 1 + \frac{1}{2}E_2 + \frac{1}{2}E_4 \end{aligned}$$

By symmetry, $E_1 = E_3$. So, $E_2 = 1 + E_1$. By taking the sum of $E_1 = 1 + \frac{1}{2}E_0 + \frac{1}{2}E_2$ and $E_3 = 1 + \frac{1}{2}E_2 + \frac{1}{2}E_4$, we have

$$2E_1 = 2 + E_2 = 2 + 1 + E_1.$$

Hence, $E_1 = 3$ so $E_2 = 4$.

(b) We now provide a faster method. Note that

$$E_i = 1 + \frac{1}{2}E_{i+1} + \frac{1}{2}E_{i-1} \quad \text{with boundary conditions } E_0 = 0 \text{ and } E_8 = 0.$$

The solution to the homogeneous recurrence relation $E_{i+1} - 2E_i + E_{i-1} = 0$ is $E_i = Ai + B$. Suppose a particular solution is $E_i = Ci^2$. Then, one can deduce that $C = -1$ by substituting this into the recurrence relation. Hence,

$$E_i = Ai + B - i^2.$$

When $i = 0$, we have $E_0 = 0$ so $B = 0$. Since $E_8 = 0$, then $8A - 64 = 0$ so $A = 8$.

As such, $E_i = 8i - i^2$. To conclude, $E_4 = 16$. \square

Example 3.24 (Markov chain). Consider the 3 possible weather options in Singapore: hot and comfortable, hot and humid, and hot and rainy. Looking at the weather for every hour, we observe the following pattern: if it is comfortable, it becomes humid in the next hour with probability $\frac{1}{4}$, rainy with probability $\frac{1}{5}$, and stays comfortable otherwise; if it is humid, it becomes rainy with probability $\frac{1}{2}$ and remains humid otherwise; and if it is rainy, it becomes comfortable with probability $\frac{2}{3}$, humid with probability $\frac{1}{4}$, and remains rainy otherwise.

If it is currently humid, on average how long should we wait before it becomes comfortable again?

Solution. Let t_C , t_H and t_R denote the expected number of hours to hit comfortable starting from C , H and R respectively. From the transition rules, we have

$$t_H = 1 + \frac{1}{2}t_H + \frac{1}{2}t_R \quad \text{and} \quad t_R = 1 + \frac{1}{4}t_H + \frac{1}{12}t_R.$$

Solve the system to obtain $t_H = \frac{17}{4}$. \square

Example 3.25 (ST2131 AY24/25 Sem 2 Tutorial 4). If $0 \leq a_i \leq 1$, where $i = 1, 2, \dots$, show that

$$\sum_{i=1}^{\infty} \left[a_i \prod_{j=1}^{i-1} (1 - a_j) \right] + \prod_{i=1}^{\infty} (1 - a_i) = 1.$$

Hint: Suppose that an infinite number of coins are to be flipped. Let a_i be the probability that the i^{th} coin lands heads, and consider when the first head occurs.

Solution. Let a_i denote the probability that the i^{th} coin lands heads, so $1 - a_i$ denotes the probability that the i^{th} coin lands tails. We first investigate this problem by constructing some example, i.e. *replace* ∞ by some positive integer, say 2 (the term on the left is defined if the **upper index in the sum ≥ 2**). So, we wish to justify that

$$\sum_{i=1}^2 \left[a_i \prod_{j=1}^{i-1} (1 - a_j) \right] + \prod_{i=1}^2 (1 - a_i) = 1.$$

Equivalently, we have

$$\underbrace{a_1}_H + \underbrace{a_2(1 - a_1)}_{TH} + \underbrace{(1 - a_1)(1 - a_2)}_{TT} = 1.$$

Note that

$$\color{red}a_1 = a_1 \prod_{j=1}^0 (1 - a_j) = \color{red}a_1 \cdot 1 = a_1.$$

We mentioned the respective outcomes too. For example, $a_2(1 - a_1)$ denotes the probability that the first coin lands heads and the second coin lands tails. In fact, although ‘ $\color{red}a_1$ denotes heads’ feels like a lack of information, it actually implies that the second outcome can be either heads or tails, i.e.

$$a_1 = \underbrace{a_1 a_2}_{HH} + \underbrace{a_1 (1 - a_2)}_{HT}.$$

So, for the case when the **mentioned upper index in the sum** is 2, it means that in a sequence of two tosses, the probability of obtaining either HH, HT, TH, TT is 1. When the **upper index** is some $N \in \mathbb{N}$, it means that in a sequence of N tosses, the probability of obtaining either of the 2^N outcomes is 1. Naturally, when we let $N \rightarrow \infty$, we obtain the desired identity.

We shall prove this formally. In the expression

$$\sum_{i=1}^{\infty} \left[a_i \prod_{j=1}^{i-1} (1 - a_j) \right],$$

the product of $1 - a_j$ over all $1 \leq j \leq i - 1$ denotes the probability that the first $i - 1$ flips are tails. Further multiplying this by a_i , we see that for the case when $i = 1$,

$$a_1 \prod_{j=1}^{i-1} (1 - a_j)$$

denotes the probability of the first head occurring on the 1st trial. In general,

$$a_i \prod_{j=1}^{i-1} (1 - a_j)$$

denotes the probability that the first head occurs on the i^{th} trial. Summing over all $i \in \mathbb{N}$, we obtain the probability that the first head occurs on some arbitrary trial! Since

$$\prod_{i=1}^{\infty} (1 - a_i)$$

denotes the probability that all tosses land on tails, summing the mentioned probabilities yields 1. \square

Example 3.26 (ST2131 AY24/25 Sem 2 Tutorial 4). Let $S = \{1, 2, \dots, n\}$ and suppose that A and B are, independently, equally likely to be any of the 2^n subsets (including the null set and S itself) of S .

Show that

$$P(A \subseteq B) = \left(\frac{3}{4}\right)^n.$$

Solution. We condition on the number of elements of B . Suppose $|B| = 0$, i.e. $B = \emptyset$. Then, $A = \emptyset$ trivially. On the other hand, if $|B| = 1$, then $B = \{1\}$. Then, there are two possibilities for A — \emptyset or $\{1\}$. Continuing this process, if $|B| = 2$, then $B = \{1, 2\}$. Then, there are four possibilities for A — $\emptyset, \{1\}, \{2\}, \{1, 2\}$.

In general, if $|B| = k$, where $k \leq n$, i.e. $B = \{1, \dots, k\}$, there are 2^k possible ways to form our subset $A \subseteq B$ since each of the k elements can either be chosen or not. Summing over all $0 \leq k \leq n$, the total number of ways to form sets $A \subseteq B \subseteq S$ is

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

Here, the term $\binom{n}{k}$ is the number of ways to form the sets B , i.e. k -element subsets. This is equivalent to the number of ways to choose k objects out of n distinct ones. By the binomial theorem,

$$\sum_{k=0}^n \binom{n}{k} 2^k = (2+1)^n = 3^n.$$

We divide this quantity by 4^n to obtain the desired probability. In fact, $4^n = 2^n \cdot 2^n$ can be interpreted as the number of ways to form the n -element subsets $A, B \subseteq S$ without restrictions. \square

Example 3.27 (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 (Figure 3.1). 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?

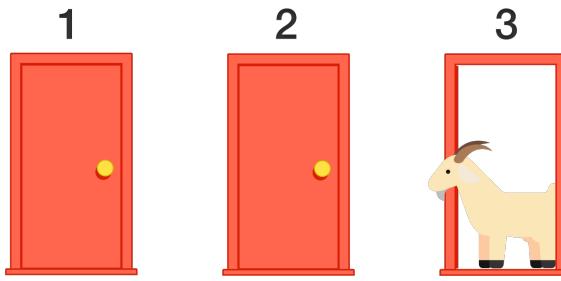


Figure 3.1: The Monty hall problem

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. By Corollary 3.1, we have

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

A is the event that Door No. 1 has a car behind it. $B | A$ 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 $P(B | A) = P(B | A') = 1$. Putting everything together, $P(A | B) = 1/3$. Hence, the probability that the car is behind Door No. 3 is $2/3$ and so you, the contestant, should make the switch.

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 $f : X \rightarrow \mathbb{N}$. Examples of discrete random variables include the binomial distribution (Chapter 4.4) covered in H2 Mathematics and the geometric distribution and the Poisson distribution (Chapters 4.5 and 4.7 respectively) covered in H2 Further Mathematics.

Here, we will study a few more distributions. For example, the Bernoulli distribution (Chapter 4.3), named after Jacob Bernoulli, who came from an academically gifted family that produced eight notable mathematicians and physicists. This can be seen as a special case of the binomial distribution. Also, we will study the negative binomial distribution (Chapter 4.6), which is closely related to the geometric distribution, and the last addition is the hypergeometric distribution (Chapter 4.8), 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, \dots . Then, the probability mass function of X is

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

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 $i = 1, 2, \dots$
- (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.

Example 4.1 (Ross p. 200 Question 31). Let X be the i^{th} smallest number in a random sample of n of the numbers $1, \dots, n+m$. Find the probability mass function of X .

Solution. For k to be the i^{th} smallest number, we must have $i-1$ numbers in $\{1, \dots, k-1\}$, then let the i^{th} number be k , and $n-i$ numbers in $\{k+1, \dots, m+n\}$. Hence,

$$P(X = k) = \frac{\binom{k-1}{i-1} \binom{m+n-k}{n-i}}{\binom{m+n}{n}} \quad \text{for } k = i, i+1, \dots, N - (n-i).$$

Note that this looks similar to the probability mass function of a hypergeometric random variable (Definition 4.9). \square

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

$$F_X : \mathbb{R} \rightarrow \mathbb{R} \quad \text{where } F_X(x) = P(X \leq x).$$

Note that if $x_1 < x_2 < x_3 < \dots$, then F is a *step function*. That is, F is constant in the interval $[x_{i-1}, x_i)$. Also, note that for any discrete random variable, the probability mass function and cumulative distribution function are related as follows:

$$P(X = x) = P(X \leq x) - P(X \leq x-1) \tag{4.1}$$

Example 4.2 (Ross p. 189 Question 19). If the distribution function of the random variable X is given by

$$F(x) = \begin{cases} 0 & x < 1 \\ \frac{1}{4} & 1 \leq x < 3 \\ \frac{5}{8} & 3 \leq x < 4 \\ \frac{3}{4} & 4 \leq x < 6 \\ \frac{7}{8} & 6 \leq x < 7 \\ 1 & x \geq 7 \end{cases}$$

calculate the probability mass function of X .

Solution. Given the distribution function $F(x)$ of the random variable X , we compute the probability mass function $p(x) = P(X = x)$ by evaluating the jump in $F(x)$ at each point.

For example, $P(X = 1) = F(1) - F(1^-)$ which is equal to $\frac{1}{4}$. Repeat this for the other probabilities to obtain the probability mass function as follows:

x	1	3	4	6	7
$P(X = x)$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

□

Example 4.3 (Ross p. 199 Question 10). An urn contains n balls numbered 1 through n . If you withdraw m balls randomly in sequence, each time replacing the ball selected previously, find $P(X = k)$ where $k = 1, \dots, n$, where X is the maximum of the m chosen numbers. A hint is to first find $P(X \leq k)$.

Solution. $X \leq k$ means every one of the m draws lies in $\{1, \dots, k\}$. Each ball hits the set with probability $\frac{k}{n}$. Hence, we have

$$P(X \leq k) = \left(\frac{k}{n}\right)^m.$$

Using (4.1), we see that

$$P(X = k) = P(X \leq k) - P(X \leq k - 1) = \left(\frac{k}{n}\right)^m - \left(\frac{k-1}{n}\right)^m.$$

□

4.2 Expectation and Variance

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

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

Sometimes, we will use $\mathbb{E}(X)$ or $\mathbb{E}[X]$ to denote the expectation of the random variable 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, \dots, 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 what is known as the *probability density function* of $g(X)$ first before proceeding to compute $E(g(X))$ by definition. 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 in Proposition 4.2 easily by definition. Next, we see that 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 (see Definition 4.4 for the definition of the moment of a random variable). 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 (Definition 7.7). As such,

$$E(X^n) = \sum_{\text{all } x} x^n P_X(x) \quad \text{for all } n \in \mathbb{N}.$$

We can also define quantities like

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

Example 4.4 (Ross p. 198 Question 2). Suppose that X takes on one of the values 0, 1 and 2. If for some constant c , $P(X = i) = cP(X = i - 1)$ for $i = 1, 2$, find $E(X)$.

Solution. We have the following probability distribution table.

i	0	1	2
$P(X = i)$	p	cp	c^2p

As such, we have $E(X) = cp + 2c^2p$. □

Example 4.5 (Ross p. 195 Question 6). Let X be a random variable such that $P(X = 1) = 1 - P(X = -1)$. Find $c \neq 1$ such that $E(c^X) = 1$.

Solution. Trivial! □

Example 4.6 (Ross p. 189 Question 24). A and B play the following game: A writes down either number 1 or 2, and B guesses which one. If A has written i and B also guesses i , then B receives i dollars from A. If B makes a wrong guess, B pays \$0.75 to A. Suppose B randomizes his decision by guessing 1 with probability p and 2 with probability $1 - p$.

Let X and Y be the amount that B gains if A has written 1 and 2 respectively. Find $E[X]$ and $E[Y]$ and the value of p such that the smaller value of $E[X]$ and $E[Y]$ attains the maximum value.

Solution. We have

$$E(X) = p - 0.75(1-p) = 1.75p - 0.75 \quad \text{and} \quad E(Y) = 2(1-p) - 0.75p = 2 - 2.75p.$$

We wish to find the value of p that maximises $\min\{1.75p - 0.75, 2 - 2.75p\}$, which is the p -coordinate of the intersection point of the graphs of $y = 1.75p - 0.75$ and $y = 2 - 2.75p$. One checks that it is $p = 11/18$. \square

Example 4.7 (Ross p. 189 Question 23). You have \$1000, and a certain commodity presently sells for \$2 per ounce. Suppose that after one week the commodity will sell for either \$1 or \$4 an ounce, with these two possibilities being equally likely.

- (a) If your objective is to maximize the expected amount of money that you possess at the end of the week, what strategy should you employ?
- (b) If your objective is to maximize the expected amount of the commodity that you possess at the end of the week, what strategy should you employ?

Solution.

- (a) Let X be the random variable denoting the price next week. Say we buy q ounces now, leaving us with $1000 - 2q$ cash remaining. Hence, our wealth next week is

$$W(q) = 1000 - 2q + qX.$$

Hence,

$$E(W) = 1000 - 2q + qE(X) = 1000 - 2q + q\left(\frac{1+4}{2}\right) = 1000 + \frac{q}{2}$$

which is increasing in q . To maximise the expectation, we should set $q = 0$.

- (b) We now wish to maximise the expected amount of the commodity that we possess at the end of the week. Let q be the number of ounces bought today. After one week, say the price of the commodity is X , so define

$$C(q) = q + \frac{1000 - 2q}{X}.$$

The expectation is $\frac{5}{4}q + \frac{5000}{8}$. To maximise the expected commodity, take $q = 500$. \square

Example 4.8 (ST2131 AY18/19 Sem 1 Midterm). Three coins each show heads with probability $3/5$ (and tails otherwise). However, for each coin, you receive a different number of points according to whether it shows head or tail. Here is the table of points allocation for each coin:

Coin	Points when Heads	Points when Tails
C_1	10	2
C_2	4	4
C_3	3	20

Suppose you and an opponent play a game. You will choose a coin each. Each of you will then toss your coin *once*; the person with the greater number of points wins. Your opponent gets to choose his/her coin first, and chooses C_3 . You then have two options before you toss your coins: either pick C_2 or C_1 . What do you do? Explain your decision clearly. *Hint:* Should we think in terms of expected values or in terms of the probability with which you can defeat his/her coin?

Solution. Suppose our opponent picks C_3 . The expected number of points our opponent would obtain is

$$3 \cdot \frac{3}{5} + 20 \cdot \frac{2}{5} = 9.8.$$

If we pick C_1 , our expected score is 6.8; if we pick C_2 , our expected score is 4. In both cases, our expected score is less than our opponent, so using expectation is not a good quantifier.

We use an alternative quantifier using the notion of condition probability.

$$P(\text{I win} \mid \text{I picked } C_1) = P(\text{me } C_1 \text{ heads and opponent } C_3 \text{ heads}) = \left(\frac{3}{5}\right)^2. \quad (4.2)$$

On the other hand,

$$\begin{aligned} P(\text{I win} \mid \text{I picked } C_2) &= P(\text{me } C_2 \text{ heads and opponent } C_3 \text{ heads}) \\ &\quad + P(\text{me } C_2 \text{ tails and opponent } C_3 \text{ heads}) \\ &= \left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)\left(\frac{3}{5}\right) \end{aligned}$$

which is obviously greater than the probability in (4.2), so picking C_2 gives me a better chance. \square

Example 4.9 (Ross p. 195 Question 4). Suppose

$$P(X = n) = \frac{4}{n(n+1)(n+2)} \quad \text{where } n \geq 1. \quad (4.3)$$

- (a) Show that (4.3) is actually a probability mass function.
- (b) Show that $E(X) = 2$.
- (c) Show that $E(X^2) = \infty$.

Solution.

- (a) Use partial fraction decomposition. Then, recognise that the sum over all $n \geq 1$ can be interpreted as a telescoping series, so the sum of probabilities is equal to 1. That is,

$$\sum_{n=1}^{\infty} \frac{4}{n(n+1)(n+2)} = 1$$

so indeed, (4.3) is indeed a probability mass function.

(b) We have

$$E(X) = \sum_{n=1}^{\infty} nP(X=n) = \sum_{n=1}^{\infty} \frac{4}{(n+1)(n+2)}$$

for which by partial fraction decomposition again yields $E(X) = 2$.

(c) We have

$$E(X^2) = \sum_{n=1}^{\infty} \frac{4n}{(n+1)(n+2)} \geq \sum_{n=1}^{\infty} \frac{1}{n}$$

where the sum on the right, known as the harmonic series, diverges. The result follows. This exercise shows that $E(X^2)$ is infinite, thus $\text{Var}(X)$ is also infinite. \square

Example 4.10 (modified from Ross p. 198 Question 4). There are four buses, carrying 40, 35, 25, 50 students respectively (not counting the drivers).

- (a)** From the 150 students, one of them is selected at random, and X denotes the number of students on his/her bus. Find $E(X)$.
- (b)** From the 4 bus drivers, one of them is selected at random, and Y denotes the number of students on his/her bus. Find $E(Y)$.

In a completely new setup, suppose we have buses $1, \dots, r$ with bus i having n_i many students. Let X be the random variable denoting the number of students in the bus when the student is randomly selected, and Y is the random variable denoting the number of students in the bus when the driver is randomly selected. Prove or disprove whether $E(Y) \leq E(X)$.

Solution.

- (a)** Picking a student uniformly from the 150 size-biases towards bigger buses. We have

$$E(X) = \sum_x xP(X=x) = 40 \cdot \frac{40}{150} + \dots + 50 \cdot \frac{50}{150} = \frac{119}{3}.$$

- (b)** Picking a driver uniformly from the 4 drivers gives a plain average. We have

$$E(Y) = \sum_y yP(Y=y) = 40 \cdot \frac{1}{4} + \dots + 50 \cdot \frac{1}{4} = 0.375.$$

As for the last (interesting) part where we are asked to prove or disprove whether $E(Y) \leq E(X)$, let the total number of students be S . Then, $n_1 + \dots + n_r = S$. We have

$$E(X) = n_1 \cdot \frac{n_1}{S} + \dots + \frac{n_r}{S} = \frac{1}{S} \sum_{i=1}^r n_i^2.$$

Next,

$$E(Y) = n_1 \cdot \frac{1}{r} + \dots + n_r \cdot \frac{1}{r} = \frac{1}{r} \sum_{i=1}^r n_i = \frac{S}{r}.$$

By the Cauchy-Schwarz inequality,

$$r \left(\sum_{i=1}^r n_i^2 \right) \geq \left(\sum_{i=1}^r n_i \right)^2 \quad \text{so} \quad r \left(\sum_{i=1}^r n_i^2 \right) \geq S^2.$$

Equivalently, $E(X) \geq E(Y)$, so the statement holds. \square

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 ,

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

Proof. Let X be a non-negative integer-valued random variable, so $X \in \{0, 1, 2, \dots\}$. Then by definition of expectation (Definition 4.3), we have

$$E(X) = \sum_{k=0}^{\infty} k \cdot P(X = k).$$

Noting that we can start the sum at $k = 1$ since $0 \cdot P(X = 0) = 0$ contributes nothing. Now, for each k , write

$$k = \sum_{i=1}^k 1.$$

Hence, we can write

$$E(X) = \sum_{k=1}^{\infty} \left(\sum_{i=1}^k 1 \right) \cdot P(X = k) = \sum_{k=1}^{\infty} \left(\sum_{i=1}^k P(X = k) \right).$$

By changing the order of summation (justified because all terms are non-negative), we obtain

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

Since

$$\sum_{k=i}^{\infty} P(X = k) = P(X \geq i),$$

this proves the desired result. □

Example 4.11 (ST2131 AY18/19 Midterm). Let X_1, X_2, \dots be independent and identically distributed discrete random variables with

$$p_i = \Pr(X_j = i) \quad \text{and} \quad \sum_{i=1}^{\infty} p_i = 1.$$

For a given *pattern* of length n , denoted by (i_1, i_2, \dots, i_n) , let T be the number of variables we need to observe until the pattern appears. For instance, if the pattern is $(3, 5, 1)$, then $T = 8$ if the following sequence is observed:

$$5, 3, 1, 3, 5, 3, 5, 1, 6, 2, \dots$$

Assume the pattern has *no overlaps* (i.e. no proper prefix of the pattern equals a suffix).

(a) Show that, for $j = 0, 1, \dots$,

$$\{T = j + n\} \quad \text{if and only if} \quad \{T > j, (X_{j+1}, \dots, X_{j+n}) = (i_1, \dots, i_n)\}.$$

(b) Show that

$$E(T) = \frac{1}{p_{i_1} p_{i_2} \cdots p_{i_n}}.$$

Solution.

(a) By definition, $\{T = j + n\}$ if and only if the pattern first occurs ending at time $j + n$. This means that the block $(X_{j+1}, \dots, X_{j+n}) = (i_1, \dots, i_n)$. Note that the pattern does not occur earlier. That is, $T > j$. So, we obtain the forward inclusion.

For the reverse inclusion, $T > j$ implies that the pattern has not occurred in the first j positions. The tuple equivalence $(X_{j+1}, \dots, X_{j+n}) = (i_1, \dots, i_n)$ means that we see the pattern occurring at $j + 1, \dots, j + n$, thus ending at $j + n$.

(b) We have

$$\sum_{j=0}^{\infty} P(T = j + n) = 1.$$

Equivalently by **(a)**,

$$\sum_{j=0}^{\infty} P(T > j \text{ and } (X_{j+1}, \dots, X_{j+n}) = (i_1, \dots, i_n)) = 1.$$

By the law of total probability (Proposition 3.3),

$$\sum_{j=0}^{\infty} P(T > j \mid (X_{j+1}, \dots, X_{j+n}) = (i_1, \dots, i_n)) \cdot p_{i_1} \cdots p_{i_n} = 1.$$

Note that the sum in red is $E(T)$ so the result follows. \square

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

$$\text{Var}(X) = E(X^2) - [E(X)]^2. \tag{4.4}$$

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. \square

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 $\text{Var}(X)$ in Definition 4.5), 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 (4.4), 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 Chapter 8.1, and to prove this result, we need to use a famous inequality called Jensen's inequality (Theorem 8.3). It turns out that the answer is true as shown in Corollary 8.1.

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, \dots, X_n are independent random variables, then

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

Example 4.12 (Ross p. 244 Question 8). Let X be a random variable that takes on values between 0 and c . That is, $P(0 \leq X \leq c) = 1$. Show that

$$\text{Var}(X) \leq \frac{c^2}{4}.$$

Solution. Assume that X is a discrete random variable (note that our proof will still hold even if X is a continuous random variable since an integral can be regarded as a continuous analogue of a sum). Hence,

$$E(X) = \sum_{0 \leq x \leq c} xP(X = x) \quad \text{and} \quad E(X^2) = \sum_{0 \leq x \leq c} x^2P(X = x).$$

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

$$\text{Var}(X) = \sum_{0 \leq x \leq c} x^2P(X = x) - \left[\sum_{0 \leq x \leq c} xP(X = x) \right]^2.$$

Since we wish to find an upper bound for $\text{Var}(X)$, we need to find an upper bound for $E(X^2)$. Since $y = x^2$ is convex, then for any $0 \leq x \leq c$, we have $0 \leq x^2 \leq cx$ so

$$\text{Var}(X) = c \sum_{0 \leq x \leq c} xP(X=x) - \left[\sum_{0 \leq x \leq c} xP(X=x) \right]^2.$$

Let $E(X) = k$ so $\text{Var}(X) = ck - k^2$. The variance can be interpreted as a function in terms of k , which is maximised at $k = c/2$. So, the maximum value for the variance is

$$c \cdot \frac{c}{2} - \left(\frac{c}{2} \right)^2 = \frac{c^2}{4}$$

which completes the proof. \square

Example 4.13 (ST2131 AY18/19 Sem 1 Midterm). A biased coin is tossed n times. On each toss, it shows heads with probability p and tails with probability $q = 1 - p$. Each toss is independent of other tosses. A run is defined to be a sequence of tosses which result in the same outcome. For instance, the sequence $HHTHTTH$ contains five runs. Let I_j be the indicator function of the event that the outcome of the $(j+1)^{\text{th}}$ toss is different from the outcome of the j^{th} toss. Then the number of distinct runs, R , can be written as

$$R = 1 + \sum_{j=1}^{n-1} I_j.$$

- (a) Find $E(R)$.
- (b) Show that $E[(R-1)^2] = 2(2n-3)pq + 4(n-2)(n-3)p^2q^2$.
- (c) Hence, find $\text{Var}(R)$.

Solution.

- (a) By the linearity of expectation, we have

$$E(R) = 1 + E(I_1) + \dots + E(I_{n-1}).$$

Two successive tosses differ if the first is heads and the next tails, or the first is tails and the next heads. So, the probability of obtaining different outcomes is $pq + qp = 2pq$. So, $E(I_j) = 2pq$ for all $1 \leq j \leq n-1$. Hence,

$$E(R) = 1 + 2(n-1)pq.$$

- (b) We have

$$R - 1 = \sum_{j=1}^{n-1} I_j.$$

So,

$$(R-1)^2 = \left(\sum_{j=1}^{n-1} I_j^2 \right) + 2 \sum_{1 \leq i \leq j \leq n-1} I_i I_j.$$

By property of indicator random variables, we have $I_j^2 = I_j$ so

$$E \left(\sum_{j=1}^{n-1} I_j^2 \right) = E \left(\sum_{j=1}^{n-1} I_j \right) = 2(n-1)pq.$$

Next, we compute $E(I_i I_j) = E(I_i)E(I_j)$. For non-adjacent indicators, that is $|j - k| \geq 2$, the pairs involve disjoint tosses. So, we obtain four different coin flips. By independence,

$$E(I_i)E(I_j) = (2pq)^2 = 4p^2q^2.$$

For adjacent indicators, that is, $k = j + 1$, the events share the $(j+1)^{\text{th}}$ toss so they are now not independent. For both $I_j = 1$ and $I_{j+1} = 1$, we need toss j to be $\neq j+1$, and toss $j+1$ to be $\neq j+2$. This happens when the sequence of 3 consecutive tosses alternates, i.e. *HTH* or *THT*. So,

$$P(I_j = 1 \text{ and } I_{j+1} = 1) = P(\text{HTH}) + P(\text{THT}) = p^2q + q^2p = pq.$$

Putting everything together yields the desired result.

(c) We computed

$$E[(R-1)^2] = E(R^2) - 2E(R) + 1$$

in (b). Also, recall we obtained a formula for $E(R)$ in (a). All that is left is to use the standard formula for variance. \square

4.3 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

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

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

Proof. $E(X) = 0 \cdot q + 1 \cdot p = p$ \square

Next, we prove the result for variance.

Proof. $E(X^2) = 0^2 \cdot q + 1^2 \cdot p = p$. As proven earlier, we have $[E(X)]^2 = p^2$. The result follows. \square

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 Chapter 4.4.

Example 4.14 (Ross p. 200 Question 21). Suppose

$$P(X = a) = p \quad \text{and} \quad P(X = b) = 1 - p.$$

Here, $a, b \in \mathbb{R}$ are distinct. One can easily prove that $\frac{X-a}{a-b}$ is a Bernoulli random variable, but we omit the details. Find $\text{Var}(X)$.

Solution. We have

$$E(X) = ap + b - bp \quad \text{and} \quad E(X^2) = a^2p + b^2 - b^2p.$$

Then,

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = a^2p + b^2 - b^2p - (ap + b - bp)^2.$$

□

Example 4.15. Suppose X_1, \dots, X_9 are Bernoulli random variables with parameters

$$\frac{1}{1 \cdot 2}, \frac{1}{2 \cdot 3}, \dots, \frac{1}{9 \cdot 10}$$

respectively.

- (a) Find the value of $E(X_1 + \dots + X_9)$.
- (b) Find the value of $E(X_1 + X_2^2 + X_3^3 + X_4^4)$.

Solution.

- (a) By the linearity of expectation,

$$\begin{aligned} E(X_1 + \dots + X_9) &= E(X_1) + \dots + E(X_9) \\ &= \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \dots + \frac{1}{9 \cdot 10} \\ &= \sum_{n=1}^9 \frac{1}{n(n+1)} \end{aligned}$$

By partial fraction decomposition, one can see this as a telescoping series, which evaluates to 0.9.

- (b) Again by the linearity of expectation,

$$\begin{aligned} E(X_1 + X_2^2 + X_3^3 + X_4^4) &= E(X_1) + E(X_2^2) + E(X_3^3) + E(X_4^4) \\ &= \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \frac{1}{4 \cdot 5} \end{aligned}$$

which evaluates to 0.8. □

To summarise,

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

4.4 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 $0 \leq k \leq n$,

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

We can write it as $X \sim B(n, p)$ 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 probability mass function for the binomial random variable.

Examples 4.16 and 4.17 give some elementary examples of the binomial distribution.

Example 4.16 (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 $X \sim B(20, \frac{1}{4})$.

Example 4.17 (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 $X \sim B(50, \frac{1}{2})$.

Example 4.18 (Ross p. 191 Question 50). When coin 1 is flipped, it lands on heads with probability 0.4; when coin 2 is flipped, it lands on heads with probability 0.7. One of these coins is randomly chosen and flipped 10 times

- (a) What is the probability that the coin lands on heads on exactly 7 of the 10 flips?
- (b) Given that the first of these 10 flips lands heads, what is the conditional probability that exactly 7 of the 10 flips land on heads?

Solution. Let X denote the number of times coin 1 lands on heads and Y denote the number of times coin 2 lands on heads. Then, $X \sim B(10, 0.4)$ and $Y \sim B(10, 0.7)$.

- (a) Note that the probability of choosing either coin 1 or coin 2 is $\frac{1}{2}$. By the law of total probability (Proposition 3.3), the desired probability is

$$\frac{1}{2}P(X = 7) + \frac{1}{2}P(Y = 7) = \frac{1}{2}\binom{10}{7}(0.4)^7(0.6)^3 + \frac{1}{2}\binom{10}{y}(0.7)^7(0.3)^3 = 0.155.$$

- (b) Let A be the event that the first of the 10 flips lands heads, and B be the event that exactly 7 of the 10 flips land on heads. Then, we wish to find the value of $P(B | A)$. Then, $A \cap B$ is the event that we obtain the sequence

$$H \quad \underbrace{THHHHHHTT}_{\text{some permutation of Hs and Ts}} \quad ,$$

which occurs with probability

$$\frac{1}{2} \cdot 0.4 \cdot \binom{9}{6} (0.4)^6 (0.6)^3 + \frac{1}{2} \cdot 0.7 \cdot \binom{9}{6} (0.7)^6 (0.3)^3.$$

Here, we used the law of total probability (Proposition 3.3). Next, and again using the law of total probability (Proposition 3.3), we have

$$P(A) = \frac{1}{2} \cdot 0.4 \sum_{k=0}^9 \binom{9}{k} (0.4)^k (0.6)^{9-k} + \frac{1}{2} \cdot 0.7 \sum_{k=0}^9 \binom{9}{k} (0.7)^k (0.3)^{9-k}.$$

By the definition of conditional probability (Definition 3.1),

$$P(B | A) = \frac{P(B \cap A)}{P(A)} = \frac{0.108253341}{0.55} = 0.197.$$

□

Example 4.19. A biased coin has a probability of 75% of showing head. The coin is flipped 10 times independently.

- (a) Given that a total of 6 heads appeared among the 10 flips, what is the conditional probability that the first 3 flips are head, tail, tail?
- (b) Given that a total of 6 heads appeared among the 10 flips, what is the conditional probability that exactly 3 heads appear in the first 4 flips?

Solution.

- (a) Let X be the random variable denoting the outcome of the biased coin. Then, $X \sim B(10, 0.75)$, where heads denotes a positive outcome. The required probability is

$$\frac{0.75 \cdot 0.25^2 \cdot \binom{7}{5} 0.75^5 \cdot 0.25^2}{P(X = 6)} = 0.0996.$$

- (b) The probability is

$$\frac{\binom{4}{3} 0.75^3 \cdot 0.25 \cdot \binom{6}{3} 0.75^3 \cdot 0.25^3}{P(X = 6)} = 0.380.$$

□

Example 4.20 (Ross p. 199 Question 13). Each of the members of a 7-judge panel independently makes a correct decision with probability 0.7. If the panel's decision is made by majority rule, what is the probability that the panel makes the correct decision? Given that 4 of the judges agreed, what is the probability that the panel made the correct decision?

Solution. Let X denote the number of judges that make the correct decision. So, $X \sim B(7, 0.7)$. The panel is correct if and only if $X \geq 4$. One works out that $P(X \geq 4) = 0.874$. Next, let A be the event that the panel is correct and B be the event that 4 of the judges agreed. By Definition 3.1, we have

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(X = 4)}{P(X = 3) + P(X = 4)} = 0.7.$$

□

Example 4.21. There are 2 coins. Coin A has a 40% probability of showing heads, and coin B has a 70% probability of showing heads. One of the coins is chosen at random and flipped 10 times.

- (a) What is the probability that we see 9 or more heads?
- (b) If we do see 9 or more heads, what is the probability that coin B was the chosen one?

Solution.

- (a) By the law of total probability (Proposition 3.3),

$$P(\text{9 or more heads}) = P(\text{9 or more heads} | A)P(A) + P(\text{9 or more heads} | B)P(B)$$

which is equal to

$$\left[\binom{10}{9} 0.4^9 \cdot 0.6 + \binom{10}{10} 0.4^{10} \right] 0.5 + \left[\binom{10}{9} 0.7^9 \cdot 0.3 + \binom{10}{10} 0.7^{10} \right] 0.5 = 0.0755.$$

- (b) From (a), we know that

$$P(\text{9 or more heads} | B) = 0.14930 \quad \text{and} \quad P(B) = 0.5.$$

Also, $P(\text{9 or more heads}) = 0.075493$. By Bayes' theorem (Theorem 3.1),

$$P(B | \text{9 or more heads}) = \frac{0.14930 \cdot 0.5}{0.075493}$$

which evaluates to 0.989. □

Proposition 4.6 (binomial complement; Ross p. 200 Question 30). Let $X \sim B(n, p)$ be a binomial random variable. Define $Y = n - X$. Then, $Y \sim B(n, 1 - p)$.

Proof. Note that

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

By the symmetry of binomial coefficients (1.2), this is equal to

$$\binom{n}{y} (1-p)^y p^{n-y}$$

which confirms our assertion that $Y \sim B(n, 1 - p)$. □

Example 4.22 (Ross p. 199 Question 8). Let $B(n, p)$ represent a binomial random variable with parameters n and p . Argue that

$$P(B(n, p) \leq i) = 1 - P(B(n, 1-p) \leq n-i-1) \quad (4.5)$$

Solution. Let $Y = n - X$. By Proposition 4.6, we know that $Y \sim B(1, p)$. We wish to prove that

$$P(X \leq i) = 1 - P(Y \leq n-i-1).$$

As such, we see that (4.5) is equivalent to

$$P(X \leq i) = 1 - P(n-X \leq n-i-1)$$

which holds because $P(X \leq i) + P(X \geq i+1) = 1$. \square

Example 4.23 (Ross p. 195 Question 16). Suppose that n independent tosses of a coin having probability p of coming up heads are made. Show that the probability that an even number of heads results is $\frac{1}{2}[1 + (q-p)^n]$, where $q = 1-p$.

Solution. Let $X \sim B(n, p)$ be the number of heads and $q = 1-p$. Then,

$$P(\text{even number of heads}) = \sum_{i=0}^{\lfloor n/2 \rfloor} \binom{n}{2i} p^{2i} q^{n-2i}.$$

We consider the binomial expansions

$$(p+q)^n = \sum_{j=0}^n \binom{n}{j} p^j q^{n-j} \quad \text{and} \quad (q-p)^n = \sum_{j=0}^n \binom{n}{j} (-1)^j p^j q^{n-j}.$$

Adding them yields

$$(p+q)^n + (q-p)^n = 2 \sum_{\substack{0 \leq j \leq n \\ j \text{ even}}} \binom{n}{j} p^j q^{n-j}.$$

Dividing both sides by 2 and recognising that $p+q = 1$, the result follows. \square

Proposition 4.7. Let $X \sim B(n, p)$. Then,

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

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

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

which is 1, and we are done. \square

Alternatively, we can prove the expectation formula by considering it as a sum of independent Bernoulli trials (recall our discussion in Chapter 4.3 and the linearity of expectation). For the variance proof, I will leave it as an exercise 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

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

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

Example 4.24 (Ross p. 199 Question 9). If X is a binomial random variable with expectation 6 and variance 2.4, find $P(X = 5)$.

Solution. Suppose $X \sim B(n, p)$. By Proposition 4.7, we have $np = 6$ and $np(1-p) = 2.4$. By elimination, we have $1-p = 0.4$, so $p = 0.6$. So, $n = 10$. Hence, $P(X = 5) = 0.201$. \square

Example 4.25 (Ross 9th edition p. 181 Question 10). Let X be a binomial random variable with parameters n and p . Show that

$$E\left[\frac{1}{X+1}\right] = \frac{1-(1-p)^{n+1}}{(n+1)p}$$

Solution. Let $X \sim B(n, p)$. Then,

$$E\left(\frac{1}{X+1}\right) = \sum_{k=0}^n \frac{P(X=k)}{k+1}$$

which is equal to

$$\sum_{k=0}^n \frac{1}{k+1} \binom{n}{k} p^k (1-p)^{n-k} = \frac{1}{n+1} \sum_{k=0}^n \binom{n+1}{k+1} p^k (1-p)^{n-k}.$$

The result follows with some manipulation. \square

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

Proposition 4.8. 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 probability mass function 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. \square

Example 4.26. A biased coin has a 55% probability of showing heads. The coin is flipped 200 times independently. What is the number of heads that is most likely to appear?

Solution. We can model this using a binomial distribution, say $X \sim B(200, 0.55)$. We are essentially finding the mode of the distribution. Suppose the mode is k . Then,

$$P(X = k) \geq P(X = k-1) \quad \text{and} \quad P(X = k) \geq P(X = k+1).$$

By the mass formula for a binomial distribution,

$$\binom{200}{k} (0.55)^k (0.45)^{200-k} \geq \binom{200}{k-1} (0.55)^{k-1} (0.45)^{201-k}$$

and

$$\binom{200}{k} (0.55)^k (0.45)^{200-k} \geq \binom{200}{k+1} (0.55)^{k+1} (0.45)^{199-k}.$$

One can show, using algebraic manipulation any mode k must satisfy

$$201 \cdot 0.55 - 1 \leq k \leq 201 \cdot 0.55.$$

Since $201 \cdot 0.55 = 110.55$, the only integer k in this interval is 110. \square

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

Proposition 4.9 (additivity). The sum of two independent binomial random variables with the same probability of success, p , still follows a binomial distribution. That is, if $X \sim B(m, p)$ and $Y \sim B(n, p)$, then $X + Y \sim B(m+n, p)$.

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

$$P(X + Y = k) = \sum_{i=0}^k P(\{X = i\} \cap \{Y = k - i\}) = \sum_{i=0}^k P(X = i)P(Y = k - i)$$

where we used the fact that X and Y are independent. By substituting the respective probability mass function formulae, the above expression is equal to

$$\sum_{i=0}^k \binom{m}{i} p^i q^{m-i} \binom{n}{k-i} p^{k-i} q^{n-k+i} = p^k q^{m+n-k} \sum_{i=0}^k \binom{m}{i} \binom{n}{k-i} = \binom{m+n}{k} p^k q^{m+n-k}.$$

Here, the last equality involved the use of Vandermonde's identity (Theorem 1.4). \square

Example 4.27 (Ross p. 200 Question 24). Ten balls are to be distributed among 5 urns, with each ball going into urn i with probability p_i , where $p_1 + \dots + p_5 = 1$. Let X_i denote the number of balls that go into urn i . Assume that events corresponding to the locations of different balls are independent.

- (a) What type of random variable is X_i ? Be as specific as possible.
- (b) For $i \neq j$, what type of random variable is $X_i + X_j$?
- (c) Find $P\{X_1 + X_2 + X_3 = 7\}$.

Solution.

- (a) X_i denotes the number of balls that go into urn i , where $1 \leq i \leq 5$. We have $X_1 + \dots + X_5 = 10$. Hence,

$$P(X_1 = x_1, \dots, X_5 = x_5) = \frac{10!}{x_1! \dots x_5!} p_1^{x_1} \dots p_5^{x_5}$$

so

$$P(X_i = x_i) = \frac{10!}{x_i!(10-x_i)!} p_i^{x_i} (1-p_i)^{10-x_i}$$

We see that $X_i \sim B(10, p_i)$, for which we would see in due course this is actually the marginal distribution of X_i (Definition 6.2).

- (b) We have $X_i + X_j \sim B(10, p_i + p_j)$ by additivity property of binomial random variables (Proposition 4.9).
- (c) From (b), we infer that

$$X_1 + X_2 + X_3 \sim B(10, p_1 + p_2 + p_3).$$

We must have $X_1 + X_2 + X_3 = 7$, so the probability of this event is

$$\binom{10}{7} (p_1 + p_2 + p_3)^7 (1 - p_1 - p_2 - p_3)^3.$$

\square

Example 4.28 (Ross p. 195 Question 12). A random walk S_n consists of sums of successive steps X_i , each of which can be ± 1 with probability p for $X_1 = 1$ such that

$$S_n = \sum_{i=1}^n X_i.$$

Show that $\frac{1}{2}(S_n + n)$ is binomially distributed and work out its mean and variance.

Solution. We have

$$P(X_i = 1) = p \quad \text{and} \quad P(X_i = -1) = 1 - p.$$

The trick is to define

$$Y_i = \frac{X_i + 1}{2}.$$

So,

$$\frac{1}{2}(S_n + n) = \frac{X_1 + \dots + X_n + n}{2} = Y_1 + \dots + Y_n.$$

Since $P(Y_i = 1) = p$, then Y_i follows a Bernoulli distribution with parameter p . Also, the Y_i 's are independent random variables. Since every Bernoulli random variable is a binomial random variable, by the additivity property of binomial random variables (Proposition 4.9), $\frac{1}{2}(S_n + n)$ follows a binomial distribution with mean np and variance $np(1-p)$. \square

Example 4.29 (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. \square

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 4.30 (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. \square

4.5 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 $1, 2, 3, \dots$ 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 probability mass function of X :

$$P(X = k) = pq^{k-1} \quad (4.6)$$

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.

For a situation to be modelled using a geometric distribution, independent trials are carried out, the outcome of each trial is deemed either a success or a failure, and the probability of a successful outcome p is the same for every trial.

The formula for $P(X = k)$ in (4.6) captures the intuition that we must fail the first $k - 1$ times, each with probability $q = 1 - p$, before succeeding on the k^{th} try. Also, note that the distribution is called *geometric* because the successive probabilities p, pq, pq^2, \dots form a geometric sequence with first term p and common ratio q .

Example 4.31. 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.

Example 4.32. A biased coin has 30% probability of showing heads. Assume that the outcomes of the tosses are independent.

- (a) What is the probability that exactly 3 tosses are needed to see the first tail?
- (b) What is the probability that at least 8 tosses are needed to see the first tail?

Solution. Let N be the number of occurrences up to and including the occurrence of the first tail. Then, N follows a geometric distribution with parameter 0.7. That is,

$$P(N = n) = 0.7 \cdot 0.3^{n-1}.$$

For (a), $P(N = 3) = 0.063$; for (b), $P(N \geq 8) = 0.0002187$. \square

Proposition 4.10. 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,

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

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

Example 4.33 (Ross p. 200 Question 22). Each game you play is a win with probability p . You plan to play 5 games, but if you win the fifth game, then you will keep on playing until you lose.

- (a) Find the expected number of games that you play.
- (b) Find the expected number of games that you lose.

Solution.

- (a) Note that we always play at least 5 games. If we lose the 5th game, we stop and this occurs with probability $q = 1 - p$. If we win the 5th game, we continue playing until our first loss. After the 5th game, the number of games is geometric with parameter being the probability of success.

We find $P(X = k)$. Note that $P(X = 5) = q$. On the other hand, for $k > 5$, we win the 5th, win $k - 6$ games, and lose the $(k - 5)^{\text{th}}$ game, which occurs with probability $P(X = k) = p \cdot p^{k-6} \cdot q = qp^{k-5}$. If $X = 5$, then the expected number of games is $5q$. If $X = k > 5$, then the expected number of games is

$$5q + \sum_{k=6}^{\infty} k \cdot qp^{k-5} = 5 + \frac{p}{1-p}.$$

- (b) We can lose on the k^{th} game, where $k \in \mathbb{N}$. If $k = 1, 2, 3, 4$, we lose with probability q . We then discuss the case where $k \geq 5$. We lose on the 5th game with probability q , lose on the 6th game with probability pq , and so on. In general, we lose on the k^{th} game with probability $p^{k-5}q$.

So, the expected number of losses is

$$4q + \sum_{k=5}^{\infty} p^{k-5}q = 5 - 4p.$$

\square

We give an alternative perspective to Example 4.33. For (a), let X denote the number of games we play. Then, $X - 4 \sim \text{Geo}(1 - p)$. Hence,

$$E(X) = E(X - 4) + 4 = \frac{1}{1-p} + 4.$$

As for (b), let Y denote the number of games we lose. Then, Y is the sum of the **number of games lost among the first 4 games** and the **number of games lost starting from the first 5 games**. Note that the **number of games lost among the first 4 games** follows a binomial distribution $B(4, 1 - p)$. By linearity of expectation,

$$E(Y) = 4(1 - p) + 1.$$

Example 4.34 (Ross p. 200 Question 26). Let α be the probability that a geometric random variable X with parameter p is an even number.

- (a) Find α by using the identity

$$\alpha = \sum_{i=1}^{\infty} \Pr\{X = 2i\}.$$

- (b) Find α by conditioning on whether $X = 1$ or $X > 1$.

Solution.

- (a) Suppose $X \sim \text{Geo}(p)$. Then,

$$P(X = k) = pq^{k-1} \quad \text{so} \quad \alpha = \sum_{i=1}^{\infty} pq^{2i-1} = \frac{pq}{1-q^2}.$$

Using the fact that $q = 1 - p$, we see that $\alpha = \frac{1-p}{2-p}$.

- (b) By the law of total probability (Proposition 3.3),

$$\begin{aligned} \alpha &= P(X \text{ even}) \\ &= P(X \text{ even} \mid X = 1)P(X = 1) + P(X \text{ even} \mid X > 1)P(X > 1) \end{aligned}$$

Since $P(X = 1) = 0$, then

$$\alpha = P(X \text{ even} \mid X > 1)P(X > 1) = P(X \text{ odd})(1 - p).$$

Since $P(X \text{ even}) = \alpha$, then $P(X \text{ odd}) = 1 - \alpha$, so the result follows. \square

Example 4.35 (Ross p. 199 Question 20). Show that if X is a geometric random variable with parameter p , then

$$E\left(\frac{1}{X}\right) = -\frac{p \log p}{1-p}.$$

Solution. We have

$$E\left(\frac{1}{X}\right) = \sum_{k=1}^{\infty} \frac{1}{k} \cdot p(1-p)^{k-1} = \frac{p}{1-p} \sum_{k=1}^{\infty} \frac{(1-p)^k}{k}. \quad (4.7)$$

The trick is to recall that (4.7) resembles the series expansion of $\ln(1+x)$ but with a slight twist. From

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots,$$

we have

$$\ln(1-x) = -\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \dots\right).$$

Setting $x = q$ and then recognising that $q = 1-p$ yields the desired result. \square

What makes the geometric distribution especially intriguing is its *memoryless property*. Say we have a biased coin with a probability p of landing heads. If we have already flipped 3 tails without success, the probability of needing 5 more flips to get a head is exactly the same as it was at the very start — past failures do not change the game.

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,

$$P(X > m+n \mid X > m) = P(X > n). \quad (4.8)$$

We apply the definition of conditional probability (Definition 3.1) to the left side of (4.8). 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 $P(X = k) = P(X > k-1) - P(X > k)$ (relationship between the cumulative distribution function and the probability mass function) to get

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

Setting $p = 1 - P(X > 1)$,¹ we obtain $P(X = k) = p(1 - p)^{k-1}$, which is indeed the probability mass function of the geometric distribution with parameter p . \square

To summarise,

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

In Proposition 4.9, 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 (see Chapter 4.6)! That is, $X + Y \sim \text{NB}(2, p)$.

Proposition 4.11 (additivity). If $X, Y \sim \text{Geo}(p)$, then $X + Y \sim \text{NB}(2, p)$.

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

$$P(X + Y = k) = \sum_{i=1}^{k-1} P(\{X = i\} \cap \{Y = k-i\}) = \sum_{i=1}^{k-1} P(X = i) P(Y = k-i)$$

where we used the fact that X and Y are independent. Substituting the probability mass function formulae, the above is equal to

$$\sum_{i=1}^{k-1} pq^{i-1} p q^{k-i-1} = (k-1) p^2 q^{k-2} = \binom{k-1}{1} p^2 q^{k-2}$$

which is the probability mass function of a negative binomial random variable with parameters 2 and p . \square

As mentioned, we will formally introduce the negative binomial distribution in Chapter 4.6, but Proposition 4.11 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 in this chapter is the Poisson random variable (Chapter 4.7). We will see in Proposition 4.16 that the sum of two Poisson random variables also follows a Poisson distribution.

Example 4.36 (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*. We leave the details to the reader. \square

¹Note that $q = P(X > 1)$, which is clear because we claimed 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, $p = P(X = 1)$.

Example 4.37 (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?

We shall investigate this. Let 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,

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

Note that t_i follows a geometric distribution with parameter p_i and $T = t_1 + t_2 + \dots + t_n$. We shall prove two interesting results, which are expressions for $E(T)$ and $\text{Var}(T)$ (Theorem 4.2), and they are related to the harmonic numbers and the famous Basel problem respectively:

Theorem 4.2.

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

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,

$$E(T) = E(t_1 + t_2 + \dots + t_n) = E(t_1) + E(t_2) + \dots + E(t_n)$$

which is equal to

$$\sum_{i=1}^n \frac{n}{n-i+1} = n \sum_{i=1}^n \frac{1}{n-i+1} = n \sum_{i=1}^n \frac{1}{i} = nH_n.$$

□

Next, we prove the result for variance.

Proof. We have

$$\text{Var}(T) = \text{Var}(t_1 + \dots + t_n) = \text{Var}(t_1) + \dots + \text{Var}(t_n) = \sum_{i=1}^n \frac{1-p_i}{p_i^2} = n \sum_{i=1}^n \frac{i-1}{(n-i+1)^2}.$$

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. \square

At this juncture, we give a motivation for studying the distribution of the minimum or maximum of geometric random variables. The geometric distribution naturally arises whenever we model the number of independent Bernoulli trials (each with success probability p) needed until the first success. Its memoryless property and simple form make it a canonical example of a waiting-time distribution in discrete time. However, in many practical situations we observe not just one such waiting time, but several *in parallel* or *in repetition*, and we care about extremes. For example, say we have a store with n identical checkout counters. Time is slotted, and in each slot, each cashier independently sees a customer with probability p . For cashier i , the number of slots until their first customer X follows a geometric distribution with parameter p . How long do we have to wait till any cashier gets their first customer? In other words, what is the distribution of $\min\{X_1, \dots, X_n\}$?

Proposition 4.12 (distribution of the minimum). If $X_i \sim \text{Geo}(p_i)$ for $1 \leq i \leq n$ and the X_i 's are independent, then

$$W = \min\{X_1, \dots, X_n\} \sim \text{Geo}\left(1 - \prod_{i=1}^n (1-p_i)\right).$$

Proof. Note that $P(W \leq r) = 1 - P(W > r)$. By definition of the minimum,

$$W = \min\{X_1, \dots, X_n\} > r \quad \text{implies} \quad X_i > r \text{ for all } 1 \leq i \leq n.$$

Hence,

$$P(W \leq r) = 1 - P(X_1 > r) \dots P(X_n > r).$$

Since $P(X_i > r) = (1-p_i)^r$, it follows that

$$P(W \leq r) = 1 - (1-p_1)^r \dots (1-p_n)^r = 1 - \left(\prod_{i=1}^n (1-p_i)\right)^r.$$

Hence, W follows a geometric distribution with parameter $1 - \prod_{i=1}^n (1-p_i)$. \square

Corollary 4.1. For the case where the X_i 's are also identically distributed, i.e. $X_i \sim \text{Geo}(p)$ for all $1 \leq i \leq n$, then

$$W = \min\{X_1, \dots, X_n\} \sim \text{Geo}\left(1 - (1-p)^n\right).$$

Proposition 4.13 (distribution of the maximum). If $X_i \sim \text{Geo}(p_i)$ for $1 \leq i \leq n$ and the X_i 's are independent, then defining $M = \max\{X_1, \dots, X_n\}$, we have

$$P(M = r) = \prod_{i=1}^n (1 - (1-p_i)^r) - \prod_{i=1}^n \left(1 - (1-p_i)^{r-1}\right).$$

Proof. The proof is similar to that of Proposition 4.12. Note that

$$P(M \leq r) = P(X_1 \leq r) \dots P(X_n \leq r) = (1 - q_1^r) \dots (1 - q_n^r) = \prod_{i=1}^n (1 - q_i^r).$$

By using the formula $P(M = r) = P(M \leq r) - P(M \leq r-1)$, the result follows. \square

4.6 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 $X \sim \text{NB}(r, p)$. The probability mass function 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 $\text{Geo}(p)$ as $\text{NB}(1, p)$ since for the geometric distribution, we are interested in the number of tries up to and including the first success.

Proposition 4.14. The expectation and variance of the negative binomial distribution $X \sim \text{NB}(r, p)$ are

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

To summarise,

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

We move on to discuss Banach's matchbox problem (Example 4.38). This problem is named after the Mathematician Stefan Banach, who is known for the Banach-Tarski paradox.

Example 4.38 (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

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

□

4.7 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 $0, 1, 2, \dots$ with probabilities

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}. \quad (4.9)$$

We say that $X \sim \text{Po}(\lambda)$. Interestingly, the parameter $\lambda > 0$ in (4.9) represents the expected number of occurrences in a fixed interval. We will see in Proposition 4.15 that if $X \sim \text{Po}(\lambda)$, then $E(X)$ and $\text{Var}(X)$ are also equal to λ .

Some examples where the Poisson distribution can be used are as follows. Take note that time plays a critical role when defining a Poisson random variable.

Example 4.39 (calls per hour). Call centres use the Poisson distribution to model the number of expected calls per hour that they will receive so they know how many call centre reps to keep on staff. For example, suppose a given call centre receives 10 calls per hour. Then, $X \sim \text{Po}(10)$.

Example 4.40 (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, $X \sim \text{Po}(100)$.

Example 4.41 (Ross p. 199 Question 14). On average, 5.2 hurricanes hit a certain region in a year. What is the probability that there will be 3 or fewer hurricanes hitting this year?

Solution. Let X denote the number of times a hurricane hits in a year. So, $X \sim \text{Po}(5.2)$. Hence, $P(X \leq 3) = 0.238$. \square

Example 4.42 (Ross p. 196 Question 17). Let X be a Poisson random variable with parameter λ . One can use (4.9) to show that $P(X = i)$ increases monotonically and then decreases monotonically as i increases, reaching its maximum when i is the largest integer not exceeding λ .

Example 4.43 (Ross p. 196 Question 18). Let X be a Poisson random variable with parameter λ . Show that

$$P(X \text{ is even}) = \frac{1}{2} (1 + e^{2\lambda}).$$

Solution. Suppose X is even. Then, there exists $k \in \mathbb{Z}_{\geq 0}$ such that $x = 2k$. The required probability is

$$P(X \text{ is even}) = \sum_{k=0}^{\infty} P(X = 2k) = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^{2k}}{(2k)!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!}.$$

Expanding the power series yields

$$1 + \frac{\lambda^2}{2!} + \frac{\lambda^4}{4!} + \frac{\lambda^6}{6!} + \frac{\lambda^8}{8!} + \dots \quad (4.10)$$

Recall that

$$e^\lambda = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \frac{\lambda^4}{4!} + \dots \quad \text{and} \quad e^{-\lambda} = 1 - \lambda + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} + \frac{\lambda^4}{4!} - \dots$$

So, the required power series expansion in (4.10) can be written as

$$\frac{e^\lambda + e^{-\lambda}}{2}$$

so the result follows. In fact, the mentioned power series in (4.10) is that of the hyperbolic cosine function $\cosh \lambda$. \square

Proposition 4.15. If $X \sim \text{Po}(\lambda)$, then

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

We shall prove the result for expectation. As an added bonus, one can refer to Example 4.44 for the derivation of $E(X^3)$ when $X \sim \text{Po}(\lambda)$.

Proof.

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

□

Example 4.44. The random variable X follows a Poisson distribution with parameter λ . Find $E[X^2]$ and $E[X^3]$.

Solution. Recall that $\text{Var}(X) = E(X^2) - (E(X))^2$, and for any Poisson random variable with parameter λ , it has mean λ and variance λ . Hence, $E(X^2) = \lambda + \lambda^2$.

Next,

$$E(X^3) = \sum_{k=0}^{\infty} k^3 \cdot \frac{e^{-\lambda} \lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} k^2 \cdot \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{r=0}^{\infty} (r+1)^2 \cdot \frac{\lambda^r}{r!}$$

Note that for

$$\sum_{r=0}^{\infty} (r+1)^2 \cdot \frac{\lambda^r}{r!},$$

we consider the standard series

$$\sum_{r=0}^{\infty} \frac{\lambda^r}{r!} = e^{\lambda} \quad \text{and} \quad \sum_{r=0}^{\infty} r \frac{\lambda^r}{r!} = \lambda e^{\lambda} \quad \text{and} \quad \sum_{r=0}^{\infty} r(r-1) \frac{\lambda^r}{r!} = \lambda^2 e^{\lambda}.$$

Hence, $E(X^3) = \lambda^3 + 3\lambda^2 + \lambda$. □

Example 4.45 (Ross p. 199 Question 15). The number of eggs laid on a tree leaf by an insect of a certain type is a Poisson random variable with parameter λ . However, such a random variable can be observed only if it is positive since if it is 0, then we cannot know that such an insect was on the leaf. If we let Y denote the observed number of eggs, then

$$P(Y = i) = P(X = i | X > 0)$$

where X is Poisson with parameter λ . Find $E(Y)$.

Solution. By Definition 3.1,

$$P(Y = i) = \frac{P(X = i \text{ where } i \in \mathbb{N})}{P(X > 0)}. \quad (4.11)$$

Since $X \sim \text{Po}(\lambda)$, then

$$P(X > 0) = 1 - e^{-\lambda}.$$

Also,

$$P(X = i \text{ where } i \in \mathbb{N}) = \frac{e^{-\lambda} \lambda^i}{i!}.$$

Substituting these into (4.11), we have

$$P(Y = i) = \frac{e^{-\lambda} \lambda^i}{i! (1 - e^{-\lambda})}.$$

Hence,

$$E(Y) = \sum_{i=1}^{\infty} i \cdot \frac{e^{-\lambda} \lambda^i}{i! (1 - e^{-\lambda})} = \frac{\lambda e^{-\lambda}}{1 - e^{-\lambda}} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = \frac{\lambda e^{-\lambda}}{1 - e^{-\lambda}} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = \frac{\lambda}{1 - e^{-\lambda}}.$$

□

To summarise,

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

Proposition 4.16 (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. We have

$$P(X + Y = n) = \sum_{k=0}^n P(\{X = k\} \cap \{Y = n - k\}) = \sum_{k=0}^n P(X = k) P(Y = n - k) \quad (4.12)$$

where we used the fact that X and Y are independent. Substituting the respective probability mass function formulae, (4.12) can be written as

$$\sum_{k=0}^n \frac{e^{-\lambda} \lambda^k}{k!} \cdot \frac{e^{-\mu} \mu^{n-k}}{(n-k)!} = \frac{e^{-(\lambda+\mu)} \mu^n}{n!} \sum_{k=0}^n \frac{n!}{k! (n-k)!} \left(\frac{\lambda}{\mu}\right)^k = \frac{e^{-(\lambda+\mu)} \mu^n}{n!} \sum_{k=0}^n \binom{n}{k} \left(\frac{\lambda}{\mu}\right)^k.$$

By the binomial theorem, the above simplifies to

$$\frac{e^{-(\lambda+\mu)} \mu^n}{n!} \left(1 + \frac{\lambda}{\mu}\right)^n = \frac{e^{-(\lambda+\mu)} (\lambda + \mu)^n}{n!}$$

and the result follows. □

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 (appears in MA3220 Ordinary Differential Equations).

Example 4.46 (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. Let $X(t)$ denote the total number of α -particles emitted in t seconds from the combined sample. Since emissions from the samples can be regarded as independent events (more formally, these are known as Poisson processes and you may look at its definition in Definition 5.8), then $X(t) \sim \text{Po}(\lambda t)$, where $\lambda = 0.3 + 0.5 = 0.8$. Thus, $X(t) \sim \text{Po}(0.8t)$.

- (a) This is equivalent to computing

$$P(X(10) \geq 3) = 1 - P(X(10) = 2)$$

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. \square

Example 4.47 (Ross 9th Edition p. 178 Question 63). People enter a gambling casino at a rate of 1 every 2 minutes.

- (a) What is the probability that no one enters between 12.00 and 12.05?
(b) What is the probability that at least 4 people enter the casino during that time?

Solution.

- (a) We should expect 2.5 persons entering the casino during this time period. Say $N \sim \text{Po}(2.5)$. Hence, $P(N = 0) = e^{-2.5}$.
(b) We wish to find the probability that at least 4 people enter the casino from 12.00 to 12.05, which is $P(N \geq 4) = 0.242$. \square

Proposition 4.17 (conditional of Poisson distribution is binomial). If X and Y are independent Poisson random variables such that $X \sim \text{Po}(\lambda)$ and $Y \sim \text{Po}(\mu)$, then

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

where

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

Proof. By Definition 3.1),

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

By applying the respective mass 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 $\lambda / (\lambda + \mu)$. \square

Example 4.48 (Ross p. 192 Question 63). The number of times that a person contracts a cold in a given year is a Poisson random variable with parameter $\lambda = 5$. Suppose that a new wonder drug (based on large quantities of vitamin C) has just been marketed that reduces the Poisson parameter to $\lambda = 3$ for 75 percent of the population. For the other 25 percent of the population, the drug has no appreciable effect on colds. If an individual tries the drug for a year and has 2 colds in that time, how likely is it that the drug is beneficial for him or her?

Solution. Let X denote the number of times an individual who tries the drugs has a cold. Let B be the event the drug is beneficial and N be the event the drug has no effect such that we have $X \sim \text{Po}(3)$ with $P(B) = 0.75$ and $X \sim \text{Po}(5)$ with $P(N) = 0.25$. By Bayes' theorem (Theorem 3.1) and the law of total probability (Proposition 3.3), $P(B | X = 2) = 0.889$. \square

Example 4.49 is adapted from a past year A-Level Mathematics Special Paper question dated back to 2004. It is the equivalent of the current H3 Mathematics.

Example 4.49 (A-Level 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. One should draw some parallelism with Proposition 4.17. 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 $s \leq r$, is

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

Hence,

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

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

$$P(S = s) = \frac{e^{-8}(1.5)^s}{s!} \sum_{r=s}^{\infty} \frac{(3.2)^r}{(r-s)!} = \frac{e^{-8}(1.5)^s}{s!} \sum_{j=0}^{\infty} \frac{(3.2)^{j+s}}{j!}$$

where we set $r - s = j$. Hence,

$$P(S = s) = \frac{e^{-8}(4.8)^s}{s!} \sum_{j=0}^{\infty} \frac{(3.2)^j}{j!} = \frac{e^{-4.8}(4.8)^s}{s!}.$$

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

Example 4.50 (Ross p. 197 Question 26). Suppose that the number of events that occur in a specified time is a Poisson random variable with parameter λ . If each event is occurred with probability p , independently of every other count, show that the number of events counted is a Poisson random variable with parameter λp .

Solution. Let the number of events be N . Then, $N \sim \text{Po}(\lambda)$. Each event is counted with probability p . Let M be the number of events counted. We wish to show that $M \sim \text{Po}(\lambda p)$. Condition on $N = n$. Suppose n events have occurred. Note that

$$M \mid N = n \sim \text{Binomial}(n, p).$$

Since

$$P(M = k \mid N = n) = \binom{n}{k} p^k (1-p)^{n-k},$$

then by the law of total probability (Proposition 3.3),

$$P(M = k) = \sum_{n=k}^{\infty} P(M = k \mid N = n) P(N = n) = \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \cdot \frac{e^{-\lambda} \lambda^n}{n!}$$

which simplifies to

$$\begin{aligned}
 e^{-\lambda} \sum_{n=k}^{\infty} \frac{1}{k! (n-k)!} (\lambda p)^k (\lambda (1-p))^{n-k} &= \frac{e^{-\lambda} (\lambda p)^k}{k!} \sum_{n=k}^{\infty} \frac{(\lambda (1-p))^{n-k}}{(n-k)!} \\
 &= \frac{e^{-\lambda} (\lambda p)^k}{k!} \cdot e^{\lambda(1-p)} \\
 &= \frac{e^{-\lambda p} (\lambda p)^k}{k!}
 \end{aligned}$$

which is precisely a Poisson random variable with parameter λp . \square

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

Theorem 4.3 (law of rare 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 $X \sim B(n, p)$ and let $\lambda = np$. Then, by first using the binomial probability mass function formula,

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

For large n and a moderate-sized λ ,

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

Hence, we conclude that

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

□

4.8 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 4.18. The sum of probabilities is indeed equal to 1. That is,

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

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

Example 4.51. 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. We have the following cases: *RRB*, *RBR* and *BRR*. For the first case, the probability is

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

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

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

respectively. 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! \square

In solving Example 4.51, it turns out that our O-Level method actually has a potential limitation when n and k are large because the total number of permutations will also be large and many cases will arise².

Proposition 4.19. 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	PMF	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	$\frac{nK}{N}$	$\frac{nK(N-K)(N-n)}{N^2(N-1)}$

To summarise the main components of discrete random variables,

²When 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 am writing this which is 4 July 2022.

Random Variable	PMF	Parameter(s)	$E(X)$	$\text{Var}(X)$
Bernoulli(p)	$X = \begin{cases} 1 & \text{if it is a success;} \\ 0 & \text{if it is a failure} \end{cases}$	p	p	pq
B(n, p)	${n \choose k} p^k q^{n-k}$	n and p	np	npq
Geo(p)	pq^{k-1}	p	$\frac{1}{p}$	$\frac{q}{p^2}$
NB(r, p)	${k-1 \choose r-1} p^r q^{k-r}$	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)	$\frac{{K \choose k} {N-K \choose n-k}}{{N \choose n}}$	N, K and n	$\frac{nK}{N}$	$\frac{nK(N-K)(N-n)}{N^2(N-1)}$

Continuous Random Variables

5.1 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 (Chapter 5.5) and the amount of rainfall obtained in a region during the dry season might be modelled by a continuous uniform distribution (Chapter 5.3). 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 probability density function (PDF) of X .

By letting $B = [a, b]$, we obtain

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

Sometimes, we may write $f_X(x)$ simply as $f(x)$ for convenience.

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

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

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 for discrete random variables, expectations and probabilities naturally take the form of finite or countable sums. In contrast, for continuous random variables, the corresponding expressions are written as integrals. This difference is not accidental: the continuous theory can be obtained as a limiting refinement of the discrete one.

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

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

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 have 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 $a, b \in \mathbb{R}$,

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$

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

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

Example 5.1 (Ross p. 245 Question 2). Fix a non-negative integer n . For some constant c , the random variable X has the probability density function

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

- (a) Find c .
- (b) Find $P(X > x)$ where $0 < x < 1$.

Solution.

(a) We have

$$\int_0^1 cx^n \, dx = 1 \quad \text{so} \quad c = n + 1.$$

(b) We have $f(x) = (n+1)x^n$ for $0 < x < 1$. Then,

$$P(X > x) = \int_x^1 (n+1)t^n \, dt = 1 - x^{n+1}.$$

□

5.2 Expectation and Variance

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

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

Note that these are analogous to the formulae for expectation and variance for the discrete counterpart in Chapter 4.2, 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,

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

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

Proposition 5.2. If X is a continuous random variable with probability density function $f(x)$, then for any real-valued function g ,

$$E[g(X)] = \int_{-\infty}^{\infty} f(x)g(x) \, dx. \quad (5.1)$$

Example 5.2 (Ross p. 246 Question 3). For some constant c , the random variable X has the probability density function

$$f(x) = \begin{cases} cx^4 & \text{if } 0 < x < 2; \\ 0 & \text{otherwise.} \end{cases}$$

Find $E(X)$ and $\text{Var}(X)$.

Solution. We first find the value of c . Note that

$$\int_0^2 cx^4 \, dx = 1 \quad \text{so} \quad c = \frac{5}{32}.$$

Hence,

$$E(X) = \int_0^2 xf(x) dx = \frac{5}{32} \int_0^2 x^5 dx$$

and we leave the computation of $\text{Var}(X)$ as a simple integration exercise as I am lazy. \square

Example 5.3 (Ross p. 246 Question 4). The random variable X has the probability density function

$$f(x) = \begin{cases} ax + bx^2 & \text{if } 0 < x < 1; \\ 0 & \text{otherwise.} \end{cases}$$

If $E(X)$, find $P(X < \frac{1}{2})$ and $\text{Var}(X)$.

Solution. Since

$$\int_0^1 ax + bx^2 dx = 1 \quad \text{then} \quad \frac{a}{2} + \frac{b}{3} = 1.$$

Since $E(X) = 0.6$, then

$$\int_0^1 x(ax + bx^2) dx = 0.6 \quad \text{so} \quad \frac{a}{3} + \frac{b}{4} = 0.6.$$

Solving both equations yields $a = 3.6$ and $b = -2.4$. So, $f(x) = 3.6x - 2.4x^2$ for $0 < x < 1$. We only find $P(X < \frac{1}{2})$, which is

$$\int_0^{1/2} 3.6x - 2.4x^2 dx = 0.35.$$

So, we leave it to the reader to compute the variance. \square

Example 5.4. A random variable X has a probability density function given by

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

Here, a is a constant.

(a) Find the value of a .

(b) Compute $E(X)$.

Solution.

(a) We have

$$\int_{-a}^a e^{-x} dx = 1.$$

Eventually, one can show that

$$a = \ln\left(\frac{1 + \sqrt{5}}{2}\right).$$

(b) We have

$$E(X) = \int_{-a}^a xe^{-x} dx = -0.0760.$$

\square

5.3 Continuous Uniform Distribution

Definition 5.4 (continuous uniform distribution). For $a < b$, a random variable X is said to be 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 $X \sim U(a, b)$.

Example 5.5 (Ross p. 246 Question 6). Your company must make a sealed bid for a construction project. If you succeed in winning the contract (by having the lowest bid), then you plan to pay another firm \$100,000 to do the work. If you believe that the minimum bid (in thousands of dollars) of the other participating companies can be modelled as the value of a random variable that is uniformly distributed on $(70, 140)$, how much should you bid to maximise your expected profit?

Solution. Let b be your bid, and M be the random variable denoting the minimum competing bid. So, $M \sim U(70, 140)$. We have

$$P(M > b) = \int_b^{140} \frac{1}{70} dt = \frac{140 - b}{70}.$$

The expected profit is

$$E(b) = (b - 100) \cdot \frac{140 - b}{70} = \frac{240 - b^2 - 14000}{70}.$$

We need $E'(b) = 0$ so $b = 120$. One checks that $E(120)$ is a maximum value. To maximise the expected profit, we should bid \$120,000. \square

Example 5.6 (Ross p. 246 Question 5). The random variable X is said to be a discrete uniform random variable on the integers $1, 2, \dots, n$ if

$$P(X = i) = \frac{1}{n} \quad \text{where } i = 1, 2, \dots, n.$$

For any non-negative real number x , let $\text{Int}(x)$ (sometimes written as $[x]$ or the more well-known floor function $\lfloor x \rfloor$) be the largest integer that is less than or equal to x . Show that if U is a uniform random variable on $(0, 1)$, then $X = \text{Int}(nU) + 1$ is a discrete uniform random variable on $1, \dots, n$.

Solution. We use the definition of the floor function. Note that

$$P(X = i) = P(\text{Int}(nU) = i - 1) = P\left(\frac{i-1}{n} \leq U < \frac{i}{n}\right) = \frac{1}{n}.$$

Hence, X is also a discrete uniform random variable on the same support. \square

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.

Proof. See Example 6.20 for a derivation of the density formula for $X + Y$ given that $X \sim U(0, 1)$ and $Y \sim U(0, 1)$. \square

Proposition 5.3. 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. As

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

then

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

which concludes the proof. \square

Example 5.7 (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 $x \leq 1/5$. This means $1-x \geq 4/5$. Since the line segment is symmetrical, the desired probability is $2 \cdot (1/5 \div 4/5) = 2/5$. \square

Example 5.8 (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?

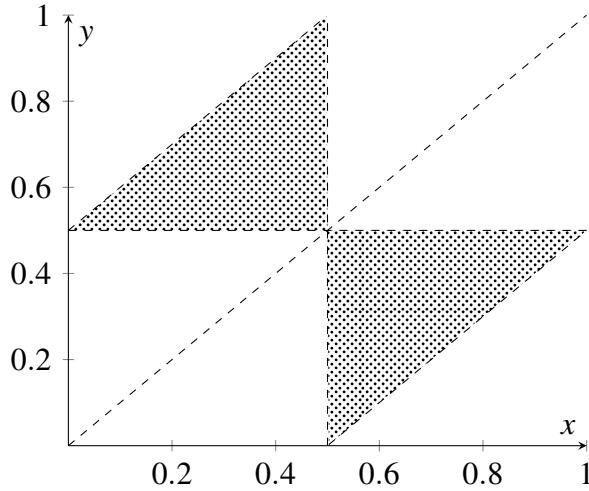
Solution. Suppose we cut the line segment at points x and y , where $x < y$. Then, we have three pieces of length $x, y-x, 1-y$. By the triangle inequality, we must have the following:

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

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?



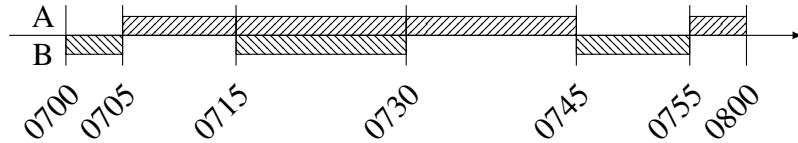
One can deduce that the required probability is $1/4$. □

Example 5.9 (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) 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?
- (b) 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.

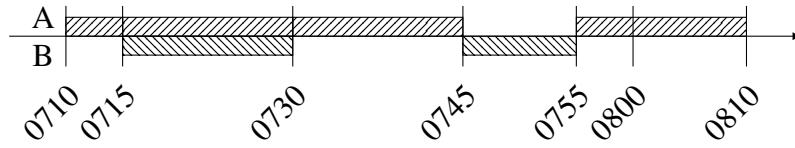
- (a) Consider the following diagram:



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,

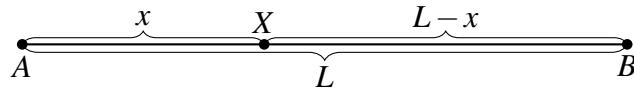
$$P(\text{man goes to destination A}) = \frac{10+15+5}{60} + \frac{1}{2} \cdot \frac{15}{60} = \frac{1}{3} + \frac{1}{8}.$$

(b) Similarly for the woman,



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

Example 5.10. A point is chosen at random (in a uniformly distributed way) on a line segment of length L , thus dividing the line segment into two pieces. What is the probability that the longer piece is at least 3 times as long as the shorter piece?



Solution. Let the length of the line segment be L , and X be the distance from one end of the point to an arbitrary point on the line. Then, $X \sim U(0, L)$. The lengths of the two pieces are X and $L - X$, and the length of the longer piece is $\max\{X, L - X\}$. We investigate when

$$\max\{X, L - X\} \geq 3 \min\{X, L - X\}.$$

Naturally, we consider two cases. For the first case, suppose $\max\{X, L - X\} = X$. Then, $\min\{X, L - X\} = L - X$. We must have

$$X \geq 3(L - X) \quad \text{so} \quad X \geq \frac{3L}{4}.$$

On the other hand, if $\max\{X, L - X\} = L - X$, then we must have

$$L - X \geq 3X \quad \text{so} \quad X \leq \frac{L}{4}.$$

The required probability is

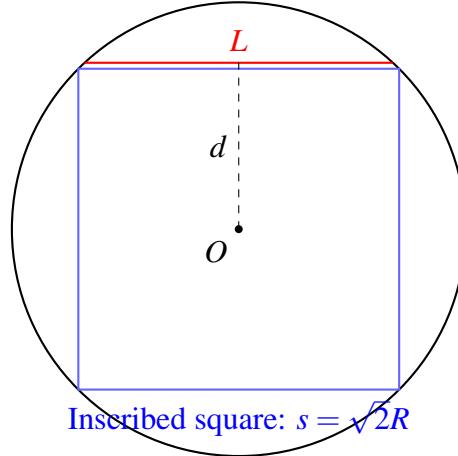
$$P\left(X \geq \frac{3L}{4} \text{ or } X \leq \frac{L}{4}\right) = 1 - P\left(\frac{L}{4} \leq X \leq \frac{3L}{4}\right) = 1 - \int_{L/4}^{3L/4} \frac{1}{L} dt = \frac{1}{2}.$$

\square

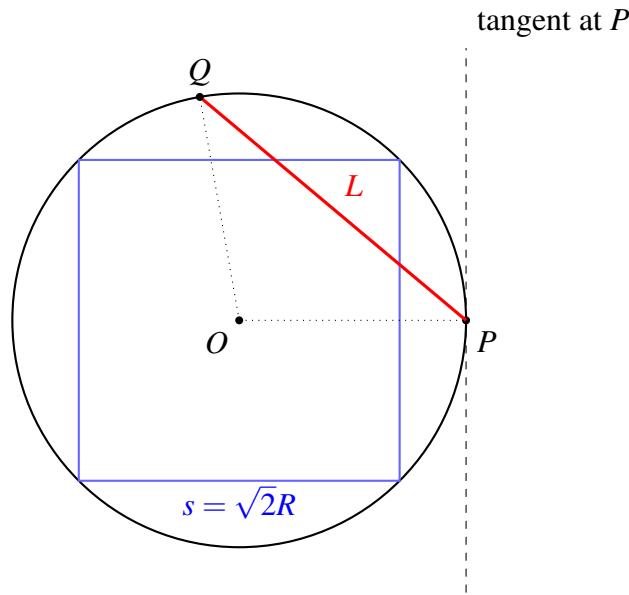
Example 5.11. Consider a circle of radius R , together with an inscribed square in the circle.

(a) Suppose a chord of the circle is picked randomly in such a way that the distance of the chord from the centre of the circle is uniformly distributed between 0 and R .

What is the probability that the chord is shorter than the side length of the square?



- (b) Suppose now that the chord of the circle is picked randomly in such a way that the angle between the chord and the tangent to the circle at one end is uniformly distributed between 0° and 180° . What is the probability that the chord is shorter than the side length of the square?



Solution.

- (a) Note that the length of the diagonal of the square is $2R$ since it is the diameter of the circle. By Pythagoras' theorem, the length of the square is $R\sqrt{2}$. The distance of the chord from the circle follows a uniform distribution $d \sim U(0, R)$. Let X be the random variable denoting

$$X = \text{length of square} - \text{length of chord} = R\sqrt{2} - \text{length of chord}.$$

By Pythagoras' theorem,

$$\text{length of chord} = 2\sqrt{R^2 - d^2}.$$

We wish to find $P(X > 0)$, or equivalently,

$$P\left(2\sqrt{R^2 - d^2} < R\sqrt{2}\right) = P\left(\frac{R}{\sqrt{2}} < d < R\right) = \int_{R/\sqrt{2}}^R \frac{1}{R} dt = 1 - \frac{1}{\sqrt{2}}.$$

- (b) The length of the side length of the square is $\sqrt{2}R$ as seen in (a). Suppose the angle between the chord and the tangent to the circle at one end follows a uniform distribution $\Theta \sim U(0, \pi)$. Note that $\angle QPO = \frac{\pi}{2} - \theta$ so $\angle QOP = 2\theta$. Let L denote the length of the chord. By the sine rule,

$$\frac{L}{\sin 2\theta} = \frac{R}{\sin\left(\frac{\pi}{2} - \theta\right)} \quad \text{so} \quad L = 2R \sin \theta.$$

We need $L < \sqrt{2}R$ so

$$P(L < \sqrt{2}R) = P(2R \sin \theta < \sqrt{2}R) = P\left(\sin \theta < \frac{1}{\sqrt{2}}\right) = P\left(\theta < \frac{\pi}{4}\right) = \frac{1}{2}.$$

□

We then introduce the inverse transform sampling method (Theorem 5.2). 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 Theorem 5.2.

Theorem 5.2 (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. Applying F to both sides, we have

$$P[F^{-1}(Y) \leq x] = P[Y \leq F(x)].$$

Since $Y \sim U(0, 1)$, the above is equal to $F(x)$. □

Example 5.12 (Ross p. 245 Question 29). Let X be a continuous random variable having cumulative distribution function F . Assume that F is strictly increasing. Find the distribution of $Y = F(X)$.

Solution. We have

$$P(Y \leq y) = P(F(X) \leq y) = P(X \leq F^{-1}(y)) \quad \text{since } F \text{ is strictly increasing.}$$

So,

$$P(X \leq F^{-1}(y)) = \int_0^{F^{-1}(y)} f(x) dx = y - F(0).$$

As such, $P(Y \leq y) = y$ so $f_Y(y) = 1$, implying that $Y \sim U(0, 1)$. □

To summarise,

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

5.4 Normal Distribution

The normal distribution is one of the most important probability distributions in Mathematics and Statistics. It is often called the bell curve because of its characteristic symmetric shape, with most of the probability mass concentrated around the mean and tapering off smoothly in both directions. This distribution arises naturally in many real-world situations: the heights of people in a population, measurement errors in experiments, IQ scores, and even fluctuations in stock prices often follow an approximately normal distribution. Its ubiquity is explained in part by the central limit theorem (CLT), which states that the sum (or average) of many independent random variables tends to follow a normal distribution, regardless of the original distributions of those variables (Theorem 8.6).

Definition 5.5. A random variable X is normally distributed with parameters μ and σ^2 , 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 (Theorem 5.3).

Theorem 5.3 (Gaussian integral).

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi} \quad (5.2)$$

We give a proof of this amazing result and note that we are jumping the gun to Chapter 6. I will provide some preliminaries in the proof anyway.

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

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

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

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

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

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

We will use polar coordinates. Let I be the original integral in (5.2). Then,

$$I = \int_{-\infty}^{\infty} e^{-x^2} dx = \int_{-\infty}^{\infty} e^{-y^2} dy.$$

Hence,

$$I^2 = \left(\int_{-\infty}^{\infty} e^{-x^2} dx \right) \left(\int_{-\infty}^{\infty} e^{-y^2} dy \right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-x^2} e^{-y^2} dxdy$$

where we used Fubini's theorem. As such,

$$I^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dxdy. \quad (5.3)$$

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

$$dxdy = r dr d\theta$$

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 $dxdy = \det(\mathbf{J}) dr d\theta$, then the result follows¹. Hence, (5.3) can be transformed to

$$I^2 = \int_0^{2\pi} \int_0^{\infty} r e^{-r^2} dr d\theta = \pi.$$

We conclude that $I = \sqrt{\pi}$. □

Proposition 5.4. The expectation and variance of a normal random variable $X \sim N(\mu, \sigma^2)$ are

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

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

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

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

¹One should refer to Algorithm 6.1 for a complete discussion.

by Z . That is, $Z \sim N(0, 1)$. Its PDF and 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 5.5. 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 5.13 (Ross p. 246 Question 8). A randomly chosen IQ test taker obtains a score that is approximately a normal random variable with mean 100 and standard deviation 15. What is the probability that the score of such a person is (a) more than 125; (b) between 90 and 110?

Solution. Consider $X \sim N(100, 15^2)$. For (a), we have $P(X > 125) = 0.0478$, for (b), we have $P(90 < X < 100) = 0.495$. \square

Example 5.14 (Ross p. 246 Question 9). Suppose that the travel time from your home to your office is normally distributed with mean 40 minutes and standard deviation 7 minutes. If you want to be 95% certain that you will not be late for an office appointment at 1 pm, what is the latest time that you should leave home?

Solution. Let t_0 be the minutes before 1 pm to leave home. Then we have $P(X < t_0) \geq 0.95$. By standardisation, we deduce that $t_0 \approx 52$ so we should leave home latest by 12.08pm. \square

Example 5.15 (Ross p. 246 Question 10). The life of a certain type of automobile tire is normally distributed with mean 34,000 miles and standard deviation 4000 miles.

- (a) What is the probability that such a tire lasts more than 40,000 miles?
- (b) What is the probability that it lasts between 30,000 and 35,000 miles?
- (c) Given that it has survived 30,000 miles, what is the conditional probability that the tire survives another 10,000 miles?

Solution. Let $X \sim N(34000, 4000^2)$ denote the life of the automobile.

- (a) We have $P(X > 40000) = 0.066807 \approx 0.0668$.

(b) We have $P(30000 < X < 35000) = 0.44005 \approx 0.440$.

(c) By Definition 3.1, we have

$$P(\text{another } 10000 \mid \text{survived } 30000) = \frac{P(X \geq 40000)}{P(X \geq 30000)} = \frac{0.066807}{0.84134} \approx 0.0794.$$

□

Example 5.16. The scores of students taking an exam are assumed to follow a normal distribution. It is known that 25% of the students scored less than 45 points, and 25% of the students scored more than 70 points.

- (i) What is the probability that a randomly chosen student scores less than 50 points?
- (ii) What is the probability that a randomly chosen student scores more than 90 points?

Solution. Let X denote the score of a randomly chosen student. Then, $X \sim N(\mu, \sigma^2)$.

- (i) We know that $P(X \leq 45) = 0.25$ and $P(X \geq 70) = 0.25$. By standardisation,

$$\frac{45 - \mu}{\sigma} = -0.67449 \quad \text{and} \quad \frac{70 - \mu}{\sigma} = 0.67449.$$

Solving yields $\mu = 57.5$ and $\sigma = 18.546$. So, $P(X < 50) = 0.343$.

- (ii) $P(X > 90) = 0.0399$.

□

Example 5.17 (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 (Definition 5.6), we have

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

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 5.18 (Ross p. 244 Question 10). Let $f(x)$ denote the probability density function of a normal random variable with mean μ and variance σ^2 . Show that $\mu - \sigma$ and $\mu + \sigma$ are points of inflection of this function. That is, show that $f''(x) = 0$ when $x = \mu - \sigma$ or $x = \mu + \sigma$.

Solution. Suppose $X \sim N(\mu, \sigma^2)$. Then,

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

Hence, we have

$$f'(x) = -\frac{x-\mu}{\sigma^2} f(x) \quad \text{so} \quad f''(x) = \left(\frac{(x-\mu)^2 - \sigma^2}{\sigma^4}\right) f(x).$$

Setting $f''(x) = 0$, we have $x = \mu \pm \sigma$, so these are the inflection points of the probability density function. \square

Example 5.19 (Ross p. 244 Question 11). Let Z be a standard normal random variable and g be a differentiable function with derivative g' .

- (a) Prove that $E(g'(Z)) = E(Zg(z))^2$
- (b) Show that $E(Z^{n+1}) = nE(Z^{n-1})$
- (c) Find $E(Z^4)$.

Solution.

- (a) Let f denote the probability density function of Z . Recall that

$$E(Z) = \int_{-\infty}^{\infty} xf(x) dx \quad \text{and} \quad f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

By (5.1), we have

$$E(g'(Z)) = \int_{-\infty}^{\infty} g'(x) f(x) dx \tag{5.4}$$

and

$$E(Zg(z)) = \int_{-\infty}^{\infty} xg(x) f(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-\frac{x^2}{2}} g(x) dx.$$

Using integration by parts, this is equal to

$$\frac{1}{\sigma\sqrt{2\pi}} \left[-g(x) e^{-\frac{x^2}{2}} \right]_{-\infty}^{\infty} - \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} -e^{-\frac{x^2}{2}} g'(x) dx$$

and upon simplification, we have

$$E(Zg(z)) = -\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} -e^{-\frac{x^2}{2}} g'(x) dx.$$

This line refers to the integral of $g'(x) f(x)$ and the proof is complete.

- (b) We have

$$E(Z^{n+1}) = \int_{-\infty}^{\infty} x^{n+1} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^n \cdot xe^{-\frac{x^2}{2}} dx.$$

Using integration by parts, the result follows.

²This is known as Stein's lemma.

- (c) By repeatedly applying (b), we have $E(Z^4) = 3E(Z^2)$. Recall that $\text{Var}(Z) = E(Z^2) - (E(Z))^2$ so $E(Z^2) = 1$. Hence, $E(Z^4) = 3$. \square

Example 5.20 (Ross p. 244 Question 11). Let Z be the standard normal random variable. Show that

$$E(Z^{n+2}) = (n+1)E(Z^n).$$

Solution. Use integration by parts. In fact, the idea is pretty much the same as (b) of Example 5.19. \square

Proposition 5.6 (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

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

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.

We then discuss the de Moivre-Laplace theorem (Theorem 5.4). Recall that the binomial distribution $B(n, p)$ describes the number of successes in n independent Bernoulli trials, each with success probability p . While exact binomial probabilities can be computed for small n , they become computationally intensive for large n . A powerful and widely used method to approximate the binomial distribution in such cases is through the normal approximation, justified by the central limit theorem.

Theorem 5.4 (de Moivre-Laplace theorem). Suppose $X \sim B(n, p)$. Then, for any $a < b$, we have

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

as $n \rightarrow \infty$. 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 approximation in Theorem 5.4 becomes increasingly accurate as n grows, and is generally considered good when the product $npq \geq 10$. However, one must be cautious when approximating discrete distributions using a continuous distribution. A key refine-

ment to improve the accuracy of the approximation is the use of the so-called *continuity correction* (Proposition 5.7).

Proposition 5.7 (continuity correction). If $X \sim B(n, p)$, then

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

Example 5.21. A fair dice is rolled 15 times. Assume the outcomes are independent. The random variable X denotes the number of times we see 1 or 2 among the 15 dice rolls. Let p denote the exact probability of $X = 5$ and let q denote the probability of $X = 5$ computed using a normal approximation. What is the relative error of $\left|\frac{p-q}{p}\right|$ expressed as a percentage?

Solution. We have $X \sim B(15, \frac{1}{3})$. Then, $p = P(X = 5)$. Consider $Y \sim N(5, \frac{10}{3})$ so X approximately follows a normal distribution with mean 5 and variance $\frac{10}{3}$. So, $p = 0.21430$ and $q = 0.21580$ (we used continuity correction for the latter). Hence, the required relative error is $\left|\frac{p-q}{p}\right| = 0.704\%$. \square

Example 5.22. In a country with a large population, a new law is being considered to ban smoking. Suppose 70% of all citizens support this new law.

- (i) What is the probability that in a random sample of 100 citizens, at least 65% of them support this new law?
- (ii) What is the minimum size of our random sample if we want to be 95% sure that at least 65% of the sampled citizens support this new law?

Solution.

- (i) We can model the situation using $X \sim B(100, 0.7)$. Then, we wish to compute $P(X \geq 65)$. We turn to normal approximation. So, X approximately follows a normal distribution, say Y , with mean 70 and variance 21, i.e. $Y \sim N(70, 21)$ approximately. Then, by continuity correction (Proposition 5.7),

$$P\left(Y \geq 65 - \frac{1}{2}\right) = 0.885.$$

- (ii) Suppose $X \sim B(n, 0.7)$ Then, we wish to find the smallest value of n such that

$$P(X \geq 0.65n) \geq 0.95.$$

Suppose X approximately follows a normal distribution, say Y , with mean $0.7n$ and variance $0.21n$, i.e. $Y \sim N(0.7n, 0.21n)$ approximately. Then, by continuity

correction (Proposition 5.7),

$$P\left(Y \geq 0.65n - \frac{1}{2}\right) \geq 0.95.$$

By standardisation,

$$P\left(Z \geq \frac{0.65n - \frac{1}{2} - 0.7n}{\sqrt{0.21n}}\right) \geq 0.95.$$

Hence,

$$\frac{0.65n - \frac{1}{2} - 0.7n}{\sqrt{0.21n}} \leq -1.645$$

so the smallest value of n is 207. \square

Example 5.23. A certain school has a maximum enrolment capacity of 200 students per year. Historical records indicate that 65% of the offers of admission made to applicants of this school were accepted.

- (i) If the school decides to make 300 offers of admissions this year, what is the probability that more than 200 students accept their offers?
- (ii) The school is only willing to take a risk of 1% chance that more than 200 students accept their offers. What is the maximum number of offers the school can make this year?

Solution.

- (i) Suppose X is the random variable denoting the number of students who accept offers. Then, $X \sim B(300, 0.65)$. We approximate this using a normal distribution with mean $300 \cdot 0.65$ and variance $300 \cdot 0.65 \cdot 0.35$. Hence, $X \sim N(195, 68.25)$ approximately. By continuity correction (Proposition 5.7),

$$P(X > 200) \approx P\left(X \geq 200 - \frac{1}{2}\right) = 0.293.$$

- (ii) Let Y be the random variable denoting the number of students who accept offers. Then, $Y \sim B(n, 0.65)$, where n is the number of offers the school can make this year. We want $P(Y > 200) < 0.01$. Again, we approximate this using a normal distribution with mean $0.65n$ and variance $0.65 \cdot 0.35n$. By continuity correction (Proposition 5.7), the maximum number of offers is 279. \square

5.5 Exponential Distribution

The exponential distribution is one of the most fundamental continuous probability distributions, often used to model the time between events in a Poisson process (see Definition 5.8). It is particularly useful when dealing with scenarios where events occur randomly and independently over time, such as waiting times, lifetimes of components, or inter-arrival times.

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 $X \sim \text{Exp}(\lambda)$.

Proposition 5.8. If $X \sim \text{Exp}(\lambda)$, then the expectation and variance are

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

Example 5.24 (Ross p. 247 Question 13). At a certain bank, the amount of time that a customer spends being served by a teller is an exponential random variable with mean 5 minutes. If there is a customer in service when you enter the bank, what is the probability that he or she will still be with the teller after an additional 4 minutes?

Solution. Suppose $X \sim \text{Exp}(\lambda)$. We are given that $1/\lambda = 5$ so $\lambda = \frac{1}{5}$. So,

$$\int_4^\infty \frac{1}{5} e^{-\frac{1}{5}x} dx = 0.449.$$

□

Example 5.25 (Ross p. 247 Question 19). Evidence concerning the guilt or innocence of a defendant in a criminal investigation can be summarised by the value of an exponential random variable X whose mean μ depends on whether the defendant is guilty. If innocent, $\mu = 1$; if guilty, $\mu = 2$. The deciding judge will rule the defendant guilty if $X > c$ for some suitably chosen value of c .

- (a) If the judge wants to be 95 percent certain that an innocent man will not be convicted, what should be the value of c ?
- (b) Using the value of c found in (a), what is the probability that a guilty defendant will be convicted?

Solution.

- (a) Let $X \sim \text{Exp}(\lambda)$ with mean $\mu = 1/\lambda$. The judge convicts if $X > c$. We have

$$P(\text{convict} \mid \text{innocent}) = P(X > c \mid \mu = 1) = e^{-c} = 0.05$$

so $c = \ln 20$.

- (b) With the value of c in (a),

$$P(\text{convict} \mid \text{guilty}) = P(X > c \mid \mu = 2) = e^{-\frac{1}{2}c} = \frac{1}{\sqrt{20}}.$$

□

Example 5.26 (Ross p. 244 Question 19). If X is an exponential random variable with mean $1/\lambda$, show that

$$E(X^k) = \frac{k!}{\lambda^k} \quad \text{where } k = 1, 2, 3, \dots$$

Solution. We have $X \sim \text{Exp}(\lambda)$. So,

$$E(X^k) = \int_0^\infty x^k \cdot \lambda e^{-\lambda x} dx.$$

Using integration by parts,

$$E(X^k) = \lambda \left[x^k \left(-\frac{1}{\lambda} e^{-\lambda x} \right) \right]_0^\infty - \lambda \int_0^\infty -\frac{1}{\lambda} e^{-\lambda x} \cdot kx^{k-1} dx = \frac{k}{\lambda} \int_0^\infty x^{k-1} \lambda e^{-\lambda x} dx.$$

This shows that $E(X^k) = kE(X^{k-1})$. Applying this repeatedly yields the desired result. \square

Proposition 5.9 (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 $F(m) = \frac{1}{2}$. Hence, $e^{-\lambda m} = \frac{1}{2}$. This implies that $m = \frac{\ln 2}{\lambda}$. \square

The expression $\ln 2/\lambda$ in Proposition 5.9 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 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 5.10 (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. Hence, for any $m, n > 0$, we have

$$P(X > m + n) = P(X > m)P(X > n). \quad (5.5)$$

We note that $P(X \leq x) = F(x)$ by definition of the CDF. Hence, (5.5) becomes

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

Using the substitution $G(x) = 1 - F(x)$ for all $x > 0$, we have

$$G(m + n) = G(m)G(n), \quad (5.6)$$

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 of differentiation, we have

$$G'(x) = \lim_{\delta x \rightarrow 0} \frac{G(x + \delta x) - G(x)}{\delta x} = \lim_{\delta x \rightarrow 0} \frac{G(x)G(\delta x) - G(x)}{\delta x}$$

where we used the functional equation (5.6). Hence,

$$G'(x) = G(x) \lim_{\delta x \rightarrow 0} \frac{G(\delta x) - G(0)}{\delta x} = G(x)G'(0).$$

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

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

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

Example 5.27. When an MRT (subway) line breaks down, the time (in hours) until the resumption of operations is an exponentially distributed random variable with parameter $1/3$.

- (i) What is the probability that more than 3 hours is needed to fix a broken down MRT line?
- (ii) One of the MRT lines has broken down 6 hours ago. What is the probability that it will get fixed within the next 3 hours?

Solution.

- (i) Let T be the time taken until the resumption of operations of a broken MRT line. Then, $T \sim \text{Exp}(1/3)$. Then,

$$P(T > 3) = \int_3^\infty \frac{1}{3} e^{-\frac{1}{3}t} dt = 0.368.$$

- (ii) We wish to find $P(T > 9 | T > 6)$, which is equal to $P(T > 3)$ by the memoryless property of the exponential distribution (Proposition 5.10). So, the answer is 0.368.

□

Example 5.28 (Ross p. 247 Question 18). There are two types of batteries in a bin. When in use, type i batteries last (in hours) an exponentially distributed time with rate λ_i , where $i = 1, 2$. A battery that is randomly chosen from the bin will be a type i battery with probability p_i , where $p_1 + p_2 = 1$. If a randomly chosen battery is still operating after t hours of use, what is the probability that it will still be operating after an additional s hours?

Solution. Let T_i denote the lifespan of a type i battery, where $i = 1$ or $i = 2$. Then, $T_i \sim \text{Exp}(\lambda_i)$. Let T denote the lifespan of a randomly chosen battery. By Definition 3.1,

$$P(T > s+t | T > t) = \frac{P(T > s+t)}{P(T > t)}.$$

By the law of total probability (Proposition 3.3),

$$P(T > t) = P(T_1 > t)p_1 + P(T_2 > t)p_2 = e^{-\lambda_1 t}p_1 + e^{-\lambda_2 t}p_2.$$

Hence, the required probability is

$$\frac{e^{-\lambda_1(s+t)}p_1 + e^{-\lambda_2(s+t)}p_2}{e^{-\lambda_1 t}p_1 + e^{-\lambda_2 t}p_2}.$$

□

Example 5.29 (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 driven 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

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

are probabilities of interest. Both probabilities are actually equal by the memoryless property of the exponential random variable (Proposition 5.10). That is,

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

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

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

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. \square

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, $T \sim \text{Exp}(\lambda)$. Note that in most cases, we usually denote an exponential random variable by T since it encompasses the essence of time.

Proof. Note that

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

Hence, $P(T \leq t) = 1 - e^{-\lambda t}$, so $f(t) = \lambda e^{-\lambda t}$. To conclude, $T \sim \text{Exp}(\lambda)$. \square

Theorem 5.5 (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. Then, W also follows an exponential distribution. That is,

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

Proof.

$$P(W \leq t) = 1 - P(W > t) = 1 - P(T_1 > t)P(T_2 > t) \dots P(T_n > t)$$

where we used the fact that the T_i 's are independent. By considering the CDF of each T_i , we see that

$$P(W \leq t) = 1 - e^{-\lambda_1 t}e^{-\lambda_2 t} \dots e^{-\lambda_n t} = 1 - e^{-(\lambda_1 + \lambda_2 + \dots + \lambda_n)} = 1 - \exp\left(-\sum_{i=1}^n \lambda_i t\right).$$

Differentiating both sides yields

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

asserting that our claim is true. \square

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

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

Example 5.30 (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 $P(X \leq x) = e^{-x}$ for $x \geq 0$ by definition of the exponential distribution. Thus,

$$P(Y \leq y) = P\left(\frac{1}{X^2} \leq y\right) = P\left(\frac{1}{y} \leq X^2\right) = P\left(X \leq -\frac{1}{\sqrt{y}} \text{ or } X \geq \frac{1}{\sqrt{y}}\right)$$

which is equal to

$$P\left(X \leq -\frac{1}{\sqrt{y}}\right) + P\left(X \geq \frac{1}{\sqrt{y}}\right) = 0 + \int_{1/\sqrt{y}}^{\infty} e^{-x} dx = \exp\left(-\frac{1}{\sqrt{y}}\right).$$

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

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

Here, $\exp(x) = e^x$. Next, we find the support of Y . Since X is defined for $x \geq 0$, then Y is defined for $y \geq 0$. To conclude,

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

□

Previously in Theorem 5.5, 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, \dots, X_n are independent random variables with common CDF F and PDF f . Let

$$U = \max\{X_1, \dots, X_n\} \quad \text{and} \quad V = \min\{X_1, \dots, X_n\}.$$

The CDF of U is

$$F_U(u) = P(U \leq u) = \prod_{i=1}^n P(X_i \leq u) = [F(u)]^n,$$

and the PDF of U is

$$f_U(u) = nf(u)[F(u)]^{n-1}.$$

Similarly,

$$F_V(v) = 1 - [1 - F(v)]^n \quad \text{and} \quad f_V(v) = nf(v)[1 - F(v)]^{n-1}.$$

The results in Theorem 5.6 are easy to establish. In particular, the respective PDFs can be easily derived by differentiating the CDF and we make use of $F' = f$.

To summarise,

Random Variable	PDF	Parameter(s)	$E(X)$	$\text{Var}(X)$
$\text{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	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 5.31 (ST2131 AY24/25 Sem 1 Lecture 12). The random variable X follows the Laplace distribution with parameter $1/\pi$. Find $P(X > \pi)$, $P(-1 < X < 2)$, $E(X)$, $\text{Var}(X)$.

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

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{1}{2\pi} \int_{-\infty}^0 x^2 e^{-x/\pi} dx + \frac{1}{2\pi} \int_0^{\infty} x^2 e^{-x/\pi} dx = 2\pi^2.$$

Here, the evaluation of each integral is quite simple — use integration by parts. \square

5.6 Gamma Distribution

The gamma distribution is a flexible family on $[0, \infty)$ that generalises the exponential law and naturally models accumulated waiting time: in a Poisson process with rate λ , the time to observe α events is $\Gamma(\alpha, \lambda)$. It is widely used for lifetimes, service times, and positive-valued data.

Definition 5.10 introduces the gamma distribution. The shape is controlled by the parameter α . For example, $\alpha = 1$ gives the exponential distribution and any integer $\alpha = n$

yields the Erlang distribution (sum of n independent and identically distributed exponentials). The rate parameter $\lambda > 0$ sets the time scale (units 1/time). Many texts use the scale $\theta = 1/\lambda$ instead; then $X \sim \Gamma(\alpha, \theta)$ has $E[X] = \alpha\theta$ and $\text{Var}(X) = \alpha\theta^2$.

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 recurrence relation $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$. One can use integration by parts to deduce the recurrence relation. Hence, it is easy to establish that for integer values of α , say $\alpha = n$, we have $\Gamma(n) = (n - 1)!$.

Observe that $\Gamma(1, \lambda) = \text{Exp}(\lambda)$, 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 5.3). \square

Proposition 5.11. If $X \sim \Gamma(\alpha, \lambda)$, then the expectation and variance are

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

To summarise,

Random Variable	PDF	Parameter(s)	$E(X)$	$\text{Var}(X)$
$\Gamma(\alpha, \lambda)$	$\frac{\lambda e^{-\lambda x} (\lambda x)^{\alpha-1}}{\Gamma(\alpha)}$	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 $N(t) \sim \text{Po}(\lambda t)$. 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. \square

5.7 Beta Distribution

Definition 5.11 (beta distribution). A random variable X is said to follow a beta distribution with parameters (a, b) , denoted by $X \sim \text{Beta}(a, b)$, 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,

$$\Gamma(a)\Gamma(b) = \left(\int_0^\infty e^{-u} u^{a-1} du \right) \left(\int_0^\infty e^{-v} v^{b-1} dv \right) = \int_0^\infty \int_0^\infty e^{-(u+v)} u^{a-1} v^{b-1} du dv.$$

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

$$\begin{aligned} \Gamma(a)\Gamma(b) &= \int_0^\infty \int_0^1 e^{-z} (zt)^{a-1} (z(1-t))^{b-1} z dt dz \\ &= \left(\int_0^\infty e^{-z} z^{a+b-1} dz \right) \left(\int_0^1 t^{a-1} (1-t)^{b-1} dt \right) \\ &= \Gamma(a+b)B(a, b) \end{aligned}$$

which asserts that the statement is true. \square

Proposition 5.12. 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	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$

5.8 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!

Example 5.32 (Ross p. 247 Question 16). A standard Cauchy random variable X has probability density function

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty.$$

Prove that $1/X$ is also a standard Cauchy random variable.

Solution. Let $Y = 1/X$. Then,

$$P(Y \leq y) = P\left(\frac{1}{X} \leq y\right) = P\left(\frac{1}{y} \leq X\right) = \int_{1/y}^{\infty} \frac{1}{\pi(1+x^2)} dx = \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{1}{y}\right).$$

Differentiating both sides yields the desired result. □

To summarise,

Random Variable	PDF	Parameter(s)	$E(X)$	$\text{Var}(X)$
$U(a, b)$	$\frac{1}{b-a}$	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
$\text{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	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	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
$\text{Beta}(a, b)$	$\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$	a and b	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$
$\text{Cauchy}(\theta)$	$\frac{1}{\pi} \cdot \frac{1}{1+(x-\theta)^2}$		θ	

5.9 Order Statistics

Suppose we have a sample of n random variables, X_1, X_2, \dots, X_n , drawn from some distribution. To study the ordered values, we sort the sample in increasing order as follows:

$$X_{(1)} < X_{(2)} < \dots < X_{(n)}$$

Here, $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ 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 $1 \leq i < j \leq n$ 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 5.33. Suppose X_1, X_2, \dots, 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 $F_{X_{(2)}}(x) = P(X_{(2)} \leq x)$. This means that at least two of the random variables $X_{(1)}, X_{(2)}, X_{(3)}$ have values $\leq x$. We shall consider two cases.

- **Case 1:** Suppose two random variables have values $\leq x$. Then, the contribution of this event is

$$\binom{3}{2} [F_X(x)]^2 [1 - F_X(x)].$$

- **Case 2:** Suppose all three random variables have values $\leq x$. The contribution of this event is

$$\binom{3}{3} [F_X(x)]^3.$$

The desired CDF is the sum of contributions of the two cases, so,

$$F_{X_{(2)}}(x) = \binom{3}{2} [F_X(x)]^2 [1 - F_X(x)] + \binom{3}{3} [F_X(x)]^3.$$

□

Joint Probability Distribution

6.1 Joint Distribution Functions

In many real-world situations, we are interested in studying the relationship between two or more random variables simultaneously. For instance, in Probability Theory, we may wish to understand the joint behaviour of height and weight, demand and supply, or the time and location of an event. This leads naturally to the concept of joint distributions, which capture the full probabilistic structure between two (or more) random variables defined on the same sample space.

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

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) \quad \text{for } x, y \in \mathbb{R}.$$

Note that $\{X \leq x, Y \leq y\}$ is equivalently $\{X \leq x\} \cap \{Y \leq y\}$.

Definition 6.2 (marginal distribution). The distribution function of X can be obtained from the joint density function of X and Y via

$$F_X(x) = \lim_{y \rightarrow \infty} F_{X,Y}(x,y) \quad \text{where } F_X \text{ is the marginal distribution of } X.$$

Similarly,

$$F_Y(y) = \lim_{x \rightarrow \infty} F_{X,Y}(x,y) \quad \text{where } F_Y \text{ is the marginal distribution of } Y.$$

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) $P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) + F_{X,Y}(a_1, b_1) - F_{X,Y}(a_2, b_1)$

$$F_{X,Y}(a_2, b_1)$$

We will only prove (i).

Proof. We set $A = \{X \leq a\}$ and $B = \{Y \leq b\}$. 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 (Proposition 2.1), 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. \square

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 as shown in

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.

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

$$P((X,Y) \in C) = \iint_{(x,y) \in C} f_{X,Y}(x,y) dx dy.$$

Example 6.2. A committee of 4 persons is to be randomly selected from a group of 12 professors consisting of 6 physicists, 4 chemists and 2 biologists. Let X and Y respectively denote the number of physicists and chemists in the committee.

- (a) Determine the joint probability mass function p of X and Y
- (b) Determine the marginal probability mass function p_X of X
- (c) Determine the marginal probability mass function p_Y of Y

Solution.

- (a) We have

$$p(x,y) = \frac{\binom{6}{x} \binom{4}{y} \binom{2}{4-x-y}}{\binom{12}{4}}.$$

- (b) We have

$$p_X(x) = \frac{\binom{6}{x} \binom{6}{4-x}}{\binom{12}{4}}.$$

- (c) We have

$$p_Y(y) = \frac{\binom{4}{y} \binom{8}{4-y}}{\binom{12}{4}}.$$

□

Example 6.3 (Ross p. 312 Question 5). Suppose X, Y, Z are independent random variables that are each equally likely to be either 1 or 2. Find the probability mass function of

- (a) XYZ

- (b) $XY + XZ + YZ$
(c) $X^2 + YZ$

Solution.

- (a) We have $X, Y, Z \sim U(1, 2)$, where U is a discrete uniform random variable. Let K denote the number of 2's among X, Y, Z so $K \sim B(3, 0.5)$. Then, $XYZ = 2^K$. Hence,

$$P(XYZ = 1, 2, 4, 8) = \left(\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}\right).$$

- (b) Recall K in (a). If $K = 0$, then $XY + XZ + YZ = 3$; if $K = 1$, then we get 5; if $K = 2$, we get 8; if $K = 3$, we get 12, so

$$P(XY + XZ + YZ = 3, 5, 8, 12) = \left(\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}\right).$$

- (c) We first condition on X . If $X = 1$, then the values are 2, 3, 5 with respective probabilities $1/4, 1/2, 1/4$; if $X = 2$, then the values are 5, 6, 8 with respective probabilities $1/4, 1/2, 1/4$. The result follows by considering $P(X = 1) = P(X = 2) = \frac{1}{2}$ and then applying the law of total probability (Proposition 3.3). \square

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,

$$P(X \in A, Y \in B) = \int_A \int_B f_{X,Y}(x, y) dy dx.$$

- (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

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x, y) dy dx.$$

- (iii) Let $a, b \in \mathbb{R}$. Then,

$$F_{X,Y}(a, b) = P(X \leq a, Y \leq b) = \int_{-\infty}^a \int_{-\infty}^b f_{X,Y}(x, y) dy dx.$$

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 6.4. 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:

$$\int_1^\infty \int_0^1 2e^{-x}e^{-2y} dy dx$$

The answer is $e^{-1}(1 - e^{-2})$.

- (ii) As $0 < x < y$ and $0 < y < \infty$, the required probability is

$$\int_0^\infty \int_0^y 2e^{-x}e^{-2y} dx dy.$$

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

$$\int_0^\infty \int_x^\infty 2e^{-x}e^{-2y} dy dx = \frac{1}{3}.$$

It yields the same conclusion as before!

- (iii) Recall from Definition 6.5 that the formula for the marginal probability density function of X is

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

As such,

$$f_X(x) = \int_0^\infty 2e^{-x}e^{-2y} dy = e^{-x}.$$

Hence, for $x > 0$, the marginal probability density function is $f_X(x) = e^{-x}$.

- (iv) Note that $P(X \leq x)$ is the marginal distribution function of x , so

$$F_X(x) = \int_0^x e^{-t} dt = 1 - e^{-x} \quad \text{where } x > 0.$$

- (v) The marginal distribution function of Y , for $y > 0$, is $F_Y(y) = 1 - e^{-2y}$. \square

Example 6.5 (Ross p. 311 Question 3). The joint density of X and Y is given by

$$f(x, y) = C(y - x)e^{-y} \quad \text{where } -y < x < y \text{ and } 0 < y < \infty.$$

- (a) Find C
- (b) Find the density function of X
- (c) Find the density function of Y
- (d) Find $E(X)$
- (e) Find $E(Y)$

Solution.

- (a) We know that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1 \quad \text{so} \quad \int_0^{\infty} \int_{-y}^y C(y - x)e^{-y} dx dy = 1.$$

One can use integration by parts to deduce that $C = \frac{1}{4}$.

- (b) We have

$$f_X(x) = \begin{cases} \frac{1}{4}e^{-x} & \text{if } x \geq 0; \\ \frac{1}{4}e^{-x}(1 - 2x) & \text{if } x < 0 \end{cases}.$$

- (c) We have

$$f_Y(y) = \begin{cases} 0 & \text{if } y < 0; \\ \frac{1}{2}y^2e^{-y} & \text{if } y \geq 0. \end{cases}$$

- (d) We have

$$E(X) = \int_0^{\infty} \frac{1}{4}xe^{-x} dx + \int_{-\infty}^0 \frac{1}{4}xe^{-x}(1 - 2x) dx.$$

- (e) We have

$$E(Y) = \int_0^{\infty} \frac{1}{2}y^3e^{-y} dy.$$

We leave it to the reader to integrate by parts and obtain an expression for $E(Y)$. \square

Example 6.6. The random variables X and Y are jointly continuous with a joint probability density function given by

$$f(x, y) = \begin{cases} c(x^2 + \frac{xy}{2}) & \text{if } 0 < x < 1 \text{ and } 0 < y < 2; \\ 0 & \text{otherwise.} \end{cases}$$

Here, c is a constant. Find the following.

- (a) Find c .
- (b) Find $P(X < \frac{1}{2} \text{ and } Y > \frac{1}{2})$
- (c) Determine the marginal density function f_X of X .
- (d) Determine the marginal density function f_Y of Y .

(e) Determine the probability density function g of Y/X .

Solution.

(a) We have

$$\int_0^2 \int_0^1 c \left(x^2 + \frac{xy}{2} \right) dx dy = 1 \quad \text{so} \quad c = \frac{6}{7}.$$

(b) From (a), we had $c = \frac{6}{7}$. So,

$$P \left(X < \frac{1}{2} \text{ and } Y > \frac{1}{2} \right) = \int_{1/2}^2 \int_0^{1/2} \frac{6}{7} \left(x^2 + \frac{xy}{2} \right) dx dy = \frac{69}{448}.$$

(c) We have

$$f_X(x) = \int_0^2 \frac{6}{7} \left(x^2 + \frac{xy}{2} \right) dy = \frac{6}{7} (2x^2 + x).$$

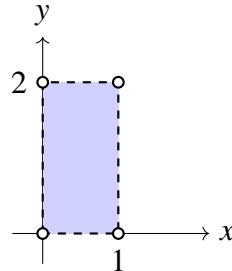
(d) We have

$$f_Y(y) = \int_0^1 \frac{6}{7} \left(x^2 + \frac{xy}{2} \right) dx = \frac{3y+4}{14}.$$

(e) Let $T = \frac{Y}{X}$. Then,

$$P(T \leq t) = P(Y \leq tX).$$

Consider the rectangle governed by the inequalities $0 < x < 1$ and $0 < y < 2$ as shown. We shall consider two different cases depending on the value of t .



First, if $0 < t \leq 2$, then the region $Y \leq tX$ represents the interior of the right-angled triangle with vertices $(0,0)$, $(1,0)$, $(1,2)$. So,

$$P(Y \leq tX) = \int_0^{tx} \frac{3y+4}{14} dy = \frac{3t^2x^2}{28} + \frac{2tx}{7}.$$

Differentiating both sides yields

$$f_t = \frac{3tx^2}{14} + \frac{2x}{7}.$$

Integrating both sides with respect to x from 0 to 1 yields

$$\int_0^1 \frac{3tx^2}{14} + \frac{2x}{7} dx = \frac{3}{14} + \frac{3}{28}t.$$

If $t > 2$, then the region $Y \leq tX$ represents the interior of the trapezium with vertices $(0,0), (1,0), (1,2), (2/t, 2)$. Note that the line $Y = tX$ has slope $2/t$ so the region below the line has slope $\leq 2/t$. Then,

$$\begin{aligned} P(Y \leq tX) &= \int_0^{tX} \int_0^{2/t} f(x,y) \, dxdy + \int_{2/t}^1 \int_0^2 f(x,y) \, dydx \\ &= \frac{6}{7} \left(\frac{7}{6} - \frac{1}{t^2} - \frac{4}{3t^3} \right) \end{aligned}$$

Differentiating gives

$$g(t) = \frac{12}{7t^3} + \frac{24}{7t^4}.$$

□

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.

In contrast to Definition 6.6, some might be more familiar with the following definition. For both discrete and continuous random variables, X and Y are independent if and only if

there exist functions $g, h : \mathbb{R} \rightarrow \mathbb{R}$ such that for all $x, y \in \mathbb{R}$, $f_{X,Y}(x,y) = g(x)h(y)$.

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

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 6.7. Suppose X and Y are jointly continuous random variables whose joint density function is given by $f(x,y)$, where

$$f(x,y) = \begin{cases} xe^{-(x+y)} & \text{if } x > 0 \text{ and } y > 0; \\ 0 & \text{otherwise.} \end{cases}$$

Are X and Y independent?

Solution. We can write

$$f_X(x) = \int_0^\infty xe^{-(x+y)} dy = xe^{-x} \quad \text{and} \quad f_Y(y) = \int_0^\infty xe^{-(x+y)} dx = e^{-y}.$$

Observe that $f(x,y) = f_X(x)f_Y(y)$ so X and Y are independent. \square

Example 6.8. Suppose X and Y are jointly continuous random variables whose joint density function is given by $f(x,y)$, where

$$f(x,y) = \begin{cases} 2 & \text{if } 0 < x < y < 1; \\ 0 & \text{otherwise.} \end{cases}$$

Are X and Y independent?

Solution. We have

$$f_X(x) = \int_x^1 2 dy = 2 - 2x \quad \text{and} \quad f_Y(y) = \int_0^y 2 dx = 2y.$$

Note that $f_X(x)f_Y(y) = 2y(2 - 2x) \neq 2$ so X and Y are not independent. \square

Example 6.9. Suppose X and Y are jointly continuous random variables whose joint density function is given by $f(x,y)$, where

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

Are X and Y independent?

Solution. We have

$$f_X(x) = \int_0^1 x+y dy = x + \frac{1}{2} \quad \text{and} \quad f_Y(y) = \int_0^1 x+y dx = y + \frac{1}{2}.$$

So, X and Y are not independent. \square

Example 6.10 (Ross p. 312 Question 7). The joint density function of X and Y is

$$f(x,y) = \begin{cases} xy & \text{if } 0 < x < 1, 0 < y < 2; \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Are X and Y independent?
- (b) Find the density function of X .
- (c) Find the density function of Y .
- (d) Find the joint distribution function.
- (e) Find $E(Y)$.
- (f) Find $P(X+Y < 1)$.

Solution.

- (a) Observe that

$$f(x,y) = xy = x \cdot \frac{y}{2} \quad \text{where } 0 < x < 1 \text{ and } 0 < y < 2$$

so X and Y are independent random variables.

- (b) We have

$$f_X(x) = \int_0^2 xy \, dy = 2x.$$

- (c) We have

$$f_Y(y) = \int_0^1 xy \, dx = \frac{y}{2}.$$

- (d) The joint distribution function is

$$F_{X,Y}(x,y) = \int_0^y \int_0^x uv \, du \, dv$$

- (e) There are five cases to consider.

- If $x < 0$ and $y < 0$, then $F_{X,Y}(x,y) = 0$.
- If $0 < x < 1$ and $0 < y < 2$, then

$$F_{X,Y}(x,y) = \int_0^y \int_0^x uv \, du \, dv = \frac{1}{4}x^2y^2.$$

If $x \geq 1$ and $0 < y < 2$, then

$$F_{X,Y}(x,y) = \int_0^y \int_0^1 uv \, du \, dv = \frac{1}{4}y^2.$$

- If $0 < x < 1$ and $y \geq 2$, then

$$F_{X,Y}(x,y) = \int_0^2 \int_0^x uv \, du \, dv = x^2.$$

- If $x \geq 1$ and $y \geq 2$, then $F_{X,Y}(x,y) = 1$.

(f) We have

$$P(X+Y < 1) = P(Y < 1-X) = \int_0^1 \int_0^{1-x} xy \, dy \, dx = \frac{1}{24}.$$

□

Example 6.11 (Ross p. 312 Question 6). Let X and Y be continuous random variables with joint density function

$$f(x,y) = \begin{cases} \frac{x}{5} + cy & \text{if } 0 < x < 1 \text{ and } 1 < y < 5; \\ 0 & \text{otherwise.} \end{cases}$$

where c is a constant.

- (a) What is the value of c ?
- (b) Are X and Y independent?
- (c) Find $P(X+Y > 3)$
- (d) Let $S = X+Y$. What is the density function of S ?

Solution.

- (a) We have

$$\int_0^1 \int_1^5 \frac{x}{5} + cy \, dy \, dx = 1 \quad \text{so} \quad c = \frac{1}{20}.$$

- (b) We have

$$f_X(x) = \int_1^5 \frac{x}{5} + \frac{1}{20}y \, dy = \frac{4x+3}{5} \quad \text{and} \quad f_Y(y) = \int_0^1 \frac{x}{5} + \frac{1}{20}y \, dx = \frac{2+y}{20}.$$

Note that

$$f_X(x)f_Y(y) = \frac{(4x+3)(2+y)}{100} \neq f(x,y)$$

so X and Y are not independent.

- (c) We have

$$P(X+Y > 3) = \int_0^1 \int_{3-x}^5 \frac{x}{5} + \frac{1}{20}y \, dy \, dx = \frac{11}{15}.$$

- (d) Since $S = X+Y$, then $s = x+y$ so $1 < s < 6$. As such,

$$f_S(s) = \int_L^U f(x, s-x) \, dx \quad \text{where } L = \max\{0, s-5\} \text{ and } U = \min\{1, s-1\}.$$

So,

$$f_S(s) = \int_L^U \frac{x}{5} + \frac{s-x}{20} \, dx.$$

As such,

$$f_S(s) = \begin{cases} \frac{5s^2-8s+3}{40} & 1 < s < 2; \\ \frac{2s+3}{40} & 2 \leq s \leq 5; \\ \frac{-5s^2+42s-72}{40} & 5 < s < 6; \\ 0 & \text{otherwise.} \end{cases}$$

□

Example 6.12. A box contains 10 balls, some of which are red and the others are green. A sample of 100 balls is drawn at random from the box with replacement. Let X denote the number of red balls from this sample. Another sample of 100 balls is drawn at random from the box with replacement. Let Y denote the number of red balls from this sample. Are the random variables X and Y independent? Why?

Solution. Let the probability of choosing a red ball in the first sample be p and the probability of choosing a red ball in the second sample be q . Then, $X \sim B(100, p)$ and $Y \sim B(100, q)$. Since X only depends on the first sample and Y only depends on the second sample, the samples are independent and count X and Y as independent random variables. As such, we have

$$P(X = x \text{ and } Y = y) = P(X = x)P(Y = y).$$

□

Example 6.13 (Ross p. 313 Question 17). Find the probability that X_1, X_2, \dots, X_n is a permutation of $1, 2, \dots, n$, when X_1, X_2, \dots, X_n are independent and

- (a) each is equally likely to be any of the values $1, \dots, n$
- (b) each has the probability mass function

$$\Pr\{X_i = j\} = p_j \quad \text{where } j = 1, \dots, n.$$

Solution.

- (a) The number of favourable outcomes is $n!$ (in fact, this is precisely the order of the symmetric group S_n), whereas the sample space has n^n elements since any of the X_i 's can take any of the values from $\{1, \dots, n\}$. Hence, the probability is

$$\frac{n!}{n^n}.$$

- (b) By independence of the random variables, then for any fixed outcome (x_1, \dots, x_n) , we have

$$P(X_1 = x_1, \dots, X_n = x_n) = p_{x_1} \dots p_{x_n}.$$

The event that (X_1, \dots, X_n) is a permutation of $(1, \dots, n)$ means that each of the numbers $1, \dots, n$ appears exactly once among the coordinates. As such, we must sum the probability over all $n!$ permutations σ of $\{1, \dots, n\}$. Let S_n denote the set of all bijections σ from $\{1, \dots, n\}$ to itself. Hence,

$$P((X_1, \dots, X_n) \text{ is a permutation}) = \sum_{\sigma \in S_n} P((X_1, \dots, X_n) = (\sigma(1), \dots, \sigma(n))).$$

So the desired probability is

$$\sum_{\sigma \in S_n} \prod_{i=1}^n p_{\sigma(i)} = \sum_{\sigma \in S_n} p_1 \dots p_n \quad \text{which does not depend on } \sigma.$$

Hence, the answer is $n! \cdot p_1 \dots p_n$.

□

Example 6.14 (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 $D = L \cos \theta$,

$$f_X(x) = \begin{cases} \frac{2}{x} & \text{if } 0 \leq x \leq \frac{D}{2}; \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad f_\theta(\theta) = \begin{cases} \frac{2}{\pi} & \text{if } 0 \leq \theta \leq \frac{\pi}{2}; \\ 0 & \text{otherwise.} \end{cases}$$

We thus obtain the joint probability density function

$$f_X(x) f_\theta(\theta) = \begin{cases} \frac{4}{\pi D} & \text{if } 0 \leq x \leq \frac{D}{2} \text{ and } 0 \leq \theta \leq \frac{\pi}{2}; \\ 0 & \text{otherwise.} \end{cases}$$

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 6.15 (Buffon's needle problem). A table is ruled with equidistant parallel lines a distance 1 cm apart. A needle of length 2 cm is randomly thrown on the table.

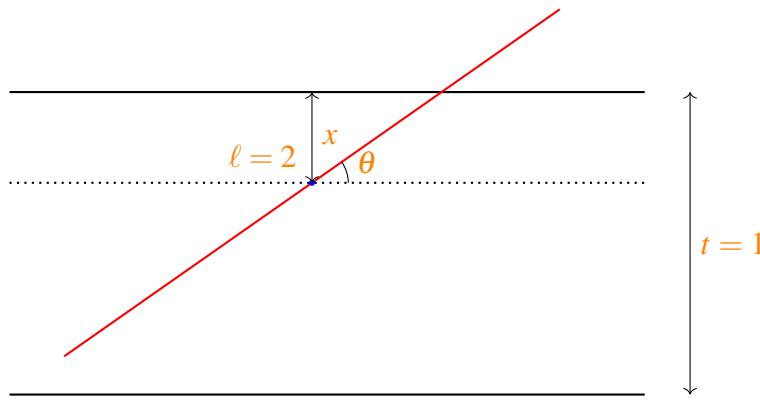


Figure 6.1: Buffon's needle problem

- (a) What is the probability that the needle intersects at least one line on the table?
- (b) What is the probability that the needle intersects two lines on the table simultaneously?

Solution. We consider the setup shown in Figure 6.1, where $\ell = 2$ and $t = 1$ denote the length of the needle and the distance between the parallel lines respectively. Let X be the random variable denoting the distance between the centre of the needle and the nearest line, and let Θ be the random variable denoting the acute angle the needle makes with the line.

- (a) We see that

$$X \sim U\left(0, \frac{t}{2}\right) = U\left(0, \frac{1}{2}\right) \quad \text{and} \quad \Theta \sim U\left(0, \frac{\pi}{2}\right).$$

Here, X and Θ are independent random variables. Hence,

$$f_X(x) = 2 \text{ for } 0 \leq x \leq \frac{1}{2} \quad \text{and} \quad f_\Theta(\theta) = \frac{2}{\pi} \text{ for } 0 \leq \theta \leq \frac{\pi}{2}.$$

Hence, their joint density is

$$f_{X,\Theta}(x, \theta) = \frac{4}{\pi} \quad \text{where } 0 \leq x \leq \frac{1}{2} \text{ and } 0 \leq \theta \leq \frac{\pi}{2}.$$

For a fixed angle θ , the needle crosses at least one line if and only if $X \leq \frac{\ell}{2} \sin \theta = \sin \theta$. Since $X \leq \frac{1}{2}$, this condition is automatically true when $\sin \theta \geq \frac{1}{2}$, i.e. when $\theta \geq \frac{\pi}{6}$. Hence,

$$P(\text{cross} \mid \Theta = \theta) = \begin{cases} 2 \sin \theta & \text{if } 0 \leq \theta < \frac{\pi}{6}; \\ 1 & \text{if } \frac{\pi}{6} \leq \theta \leq \frac{\pi}{2}. \end{cases}$$

By the law of total probability (Proposition 3.3), the probability of at least one intersection is

$$\int_0^{\pi/6} 2 \sin \theta \cdot \frac{2}{\pi} d\theta + \int_{\pi/6}^{\pi/2} 1 \cdot \frac{2}{\pi} d\theta = \frac{2}{\pi} (2 - \sqrt{3}) + \frac{2}{3} \approx 0.8367.$$

- (b)** For $\ell = 2$ and $t = 1$, the needle hits two lines if and only if both endpoints straddle the two neighbouring lines, i.e.

$$X \leq \frac{\ell}{2} \sin \theta \quad \text{and} \quad X + \frac{\ell}{2} \sin \theta \geq t.$$

With $\ell = 2$ and $t = 1$, this becomes

$$\sin \theta \geq 1 - X \quad \text{where } 0 \leq X \leq \frac{1}{2} \text{ and } 0 \leq \theta \leq \frac{\pi}{2}.$$

Since $X \sim U(0, \frac{1}{2})$ and $\Theta \sim U(0, \frac{\pi}{2})$ are independent, the required probability is

$$\int_0^{1/2} \int_{\sin^{-1}(1-x)}^{\pi/2} \frac{4}{\pi} d\theta dx = \frac{2\sqrt{3}}{\pi} - \frac{2}{3} \approx 0.436.$$

□

Example 6.16 (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, $X \sim U(0, \sqrt{3}/2)$. 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

$$\{(X, \theta) \in [0, \sqrt{3}/2] \times [0, \pi/2] : \sin \theta > X\}.$$

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.

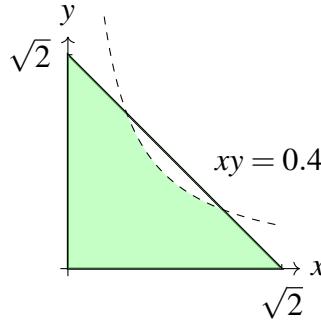
The area of the feasible region is

$$\frac{\pi}{2} \times \frac{\sqrt{3}}{2} - \int_0^{\sqrt{3}/2} \sin^{-1}(x) dx = 0.4069$$

and take this divided by the area of the rectangle to give 70%. □

Example 6.17. Consider the right-angled triangle on the plane with vertices at $(0, 0)$, $(\sqrt{2}, 0)$ and $(0, \sqrt{2})$. A point is chosen randomly within this triangle.

- (a) The random variable A denotes the area of the rectangle bounded by the x - and y -axes and the horizontal and vertical lines through the chosen point. Find $P(A < 0.4)$.



- (b) The random variable D denotes the distance of the chosen point from the origin. Find $P(D > 1.2)$.

Solution.

- (a) Note that $y \leq \frac{0.4}{x}$ denotes the region below the hyperbola $xy = 0.4$. By comparing the y -values, the region is represented by

$$0 \leq y \leq \min \left\{ \frac{0.4}{x}, \sqrt{2} - x \right\}.$$

We note that

$$\frac{0.4}{x} \leq \sqrt{2} - x \quad \text{if} \quad \alpha = \frac{1}{\sqrt{2}} - \frac{1}{\sqrt{10}} \leq x \leq \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{10}} = \beta.$$

So,

$$P(A < 0.4) = \int_{\alpha}^{\beta} \int_0^{\frac{0.4}{x}} dy dx + \int_0^{\alpha} \int_0^{\sqrt{2}-x} dy dx + \int_{\beta}^{\sqrt{2}} \int_0^{\sqrt{2}-x} dy dx = 0.938.$$

- (b) We have $D = \sqrt{X^2 + Y^2}$. Consider the complement of the event, which is $X^2 + Y^2 < 1.2^2$. However, we need to restrict this to the interior of the right-angled triangle. Let

$$R = \left\{ (x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1.2^2, x + y \leq \sqrt{2}, x \geq 0, y \geq 0 \right\}.$$

Using polar coordinates, the region is equivalent to

$$R = \left\{ (r, \theta) \in \mathbb{R}^2 : 0 \leq r \leq \min \left\{ \frac{\sqrt{2}}{\cos \theta + \sin \theta}, 1.2 \right\} \text{ and } 0 \leq \theta \leq \frac{\pi}{2} \right\}.$$

Hence,

$$P(D \leq 1.2) = \frac{1}{\text{area of triangle}} \cdot \frac{1}{2} \int_0^{\frac{\pi}{2}} \min \left\{ \frac{\sqrt{2}}{\cos \theta + \sin \theta}, 1.2 \right\} d\theta$$

$$\text{so } P(D > 1.2) = 0.049. \quad \square$$

Example 6.18 (ST2131 AY17/18 Sem 1; Poisson thinning property). Let N be a Poisson random variable with mean $\lambda > 0$. We place N balls one by one randomly into one of r boxes. For $1 \leq i \leq r$, let X_i be the number of balls in box i . Prove that X_1, \dots, X_r are independent Poisson random variables with mean λ/r .

Solution. We have $N \sim \text{Po}(\lambda)$. For any n in the support of N , let $x_1 + \dots + x_r = n$. If we fix in advance that ball 1 goes to some box, ball 2 to another, and so on, then the probability of this exact arrangement is $1/r^n$ because each ball independently has probability $1/r$ of going into the chosen box. Hence, we have

$$P(X_1 = x_1, \dots, X_r = x_r \mid N = n) = \frac{n!}{x_1! \dots x_r!} \cdot \frac{1}{r^n}.$$

By the law of total probability (Proposition 3.3),

$$\begin{aligned} P(X_1 = x_1, \dots, X_r = x_r) &= \sum_{n=0}^{\infty} P(X_1 = x_1, \dots, X_r = x_r \mid N = n) P(N = n) \\ &= \frac{(x_1 + \dots + x_r)!}{x_1! \dots x_r!} \cdot \frac{1}{r^{x_1+\dots+x_r}} \cdot \frac{e^{-\lambda} \lambda^{x_1+\dots+x_r}}{(x_1 + \dots + x_r)!} \\ &= \frac{e^{-\lambda}}{x_1! \dots x_r!} \cdot \left(\frac{\lambda}{r}\right)^{x_1+\dots+x_r} \\ &= \prod_{i=1}^r e^{-\lambda/r} \cdot \frac{(\lambda/r)^{x_i}}{x_i!} \end{aligned}$$

This is the product of the joint probability mass function of r Poisson random variables, hence the result follows. \square

Example 6.19 (Ross p. 312 Question 11). Let X_1, X_2, \dots be a sequence of independent uniform $(0, 1)$ random variables. For a fixed constant c , define the random variable N by

$$N = \min \{n : X_n > c\}.$$

Is N independent of X_N ? That is, does knowing the value of the first random variable that is greater than c affect the probability distribution of when this random variable occurs? Give an intuitive explanation for your answer.

Solution. Yes — N and X_N are independent. Note that for any set $A \subseteq (c, 1]$, where $0 < c < 1$, we have

$$\{N = n : X_N \in A\} = \{X_1 \leq c, \dots, X_{n-1} \leq c : X_n \in A\}.$$

Hence,

$$P(N = n : X_n \in A) = [P(X \leq c)]^{n-1} P(X \in A) = c^{n-1} (1 - c).$$

By the law of total probability (Proposition 3.3),

$$P(X_N \in A) = \sum_{n=1}^{\infty} P(N = n \text{ and } X_N \in A) = \sum_{n=1}^{\infty} c^{n-1} (1 - c)$$

so indeed, N and X_N are independent.

We give some intuition to this problem. Think of each draw as a Bernoulli success if $X_n > c$. Then N is geometric with success probability $1 - c$, depending only on the success/failure pattern — not on the actual values. When the first success occurs, that successful value is just a fresh uniform draw conditioned to exceed c , i.e. uniform on $(c, 1)$. Because the trials are independent and identically distributed, the point at which the first success happens tells you nothing about the value of that success, and vice versa. \square

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

$$F_{X+Y}(x) = P(X + Y \leq x) = \iint_{s+t \leq x} f_{X,Y}(s, t) ds dt$$

which simplifies to

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

\square

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 6.20. Recall from Theorem 6.20 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 $x \leq 0$ and $x \geq 2$, 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 $0 < x \leq 1$ and $1 < x < 2$.

For $0 < x \leq 1$,

$$f_{X+Y}(x) = \int_0^1 f_X(x-t) dt + \int_x^1 f_X(x-t) dt = \int_0^x f_X(x-t) dt = \int_0^x dt = x.$$

In a similar fashion, it can be shown that for $1 < x < 2$,

$$f_{X+Y}(x) = 2-x.$$

Hence,

$$f_{X+Y} = \begin{cases} x & \text{if } 0 < x \leq 1; \\ 2-x & \text{if } 1 < x < 2; \\ 0 & \text{otherwise.} \end{cases}$$

The density function has the shape of a triangle, so $X+Y$ follows a triangular distribution.

□

More generally, we have the following result:

Proposition 6.6. Assume that X and Y are independent uniform random variables on intervals $X \sim U(\alpha, \beta)$ and $Y \sim U(\gamma, \delta)$. Let $Z = X + Y$. Then,

$$f_Z(z) = \frac{1}{(\beta - \alpha)(\delta - \gamma)} \max \{0, \min \{\beta, z - \gamma\} - \max \{\alpha, z - \delta\}\}$$

Example 6.21 (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. One can use the formula given in Proposition 6.6 to tackle this problem straight-away.

- (a) The probability is $P(-0.5 + 2Y < 3X < 0.5 + 2Y)$, which is

$$\begin{aligned} \iint_{-0.5+2y < 3x < 0.5+2y} f(x,y) dx dy &= \iint_{-0.5+2y < 3x < 0.5+2y} f_X(x)f_Y(y) dx dy \\ &= \int_0^1 \int_{(2y-0.5)/3}^{(2y+0.5)/3} \left(\frac{1}{1}\right)^2 dx dy \\ &= \frac{1}{3} \end{aligned}$$

- (b) In a similar fashion, the required probability is

$$\iint_{2y < 3x < 2.5} f(x,y) dx dy = \int_0^1 \int_{2y/3}^{2.5/3} dx dy = \frac{1}{2}.$$

□

Example 6.22. Let X and Y be independent random variables uniformly distributed on the unit interval $[0, 1]$. Determine the probability density function f of $3X - 2Y$.

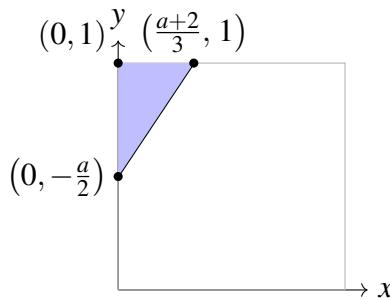
Solution. Define $A = 3X - 2Y$. Then, the range of values of A is $[-2, 3]$. Clearly, if $a \leq -2$ or $a \geq 3$, then $f_A(a) = 0$. We consider the region $3x - 2y \leq a$ for different values of a , where $0 \leq x, y \leq 1$. First, note that $3x - 2y = a$ is equivalent to $y = \frac{3}{2}x - \frac{1}{2}a$, which has gradient $\frac{3}{2}$ and y -intercept $-\frac{1}{2}a$.

- **Case 1:** If $-2 \leq a \leq 0$, the region $3x - 2y \leq a$ restricted to $0 \leq x, y \leq 1$ represents a right-angled triangle. It has an area of

$$\frac{1}{2} \left(1 + \frac{a}{2}\right) \left(\frac{a+2}{3}\right).$$

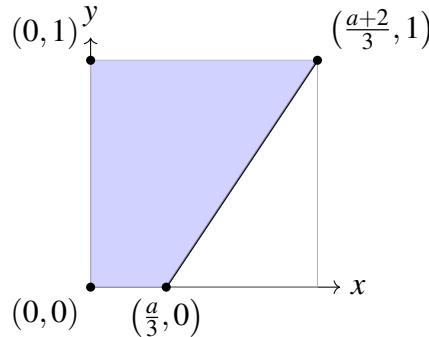
So,

$$F_A(a) = \frac{1}{2} \left(1 + \frac{a}{2}\right) \left(\frac{a+2}{3}\right) \quad \text{so} \quad f_A(a) = \frac{1}{6}(a+2).$$



- If $0 \leq a \leq 1$, the region represents a trapezium. Similarly, we find its area so

$$F_A(a) = \frac{1}{2} \cdot 1 \left(\frac{a+2}{3} + \frac{a}{3}\right) \quad \text{so} \quad f_A(a) = \frac{1}{3}.$$



- If $1 \leq a \leq 3$, we have a pentagonal region. One can sketch the region, then show that

$$f_A(a) = \frac{1}{6}(3-a).$$

We are done!

□

Example 6.23 (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.

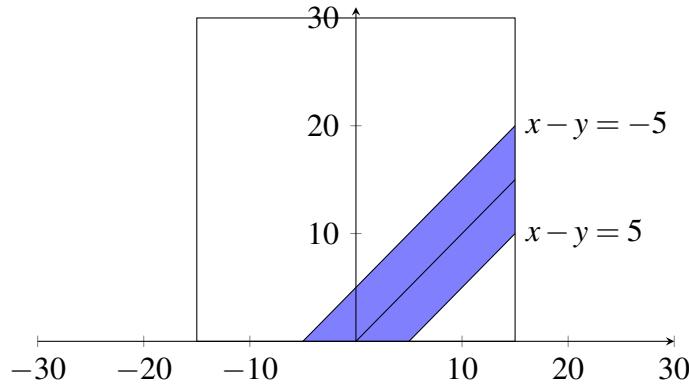
- (a) What is the probability that the first person to arrive waits less than 5 minutes for the second person?
- (b) What is the probability that the man arrives first?

Solution.

- (a) Let X and Y be the number of minutes the man and woman arrive with respect to 12 pm. We have $X \sim U(-15, 15)$ and $Y \sim U(0, 30)$. It suffices to find

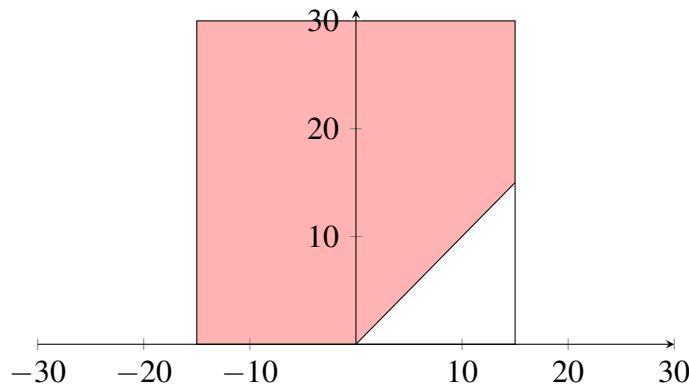
$$P(|X - Y| < 5) \quad \text{or equivalently} \quad P(-5 < Y - X < 5).$$

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,

$$P(Y > X) = \frac{\text{shaded area}}{\text{total area}} = 1 - \frac{\frac{1}{2} \cdot 15^2}{30^2} = 0.88.$$

□

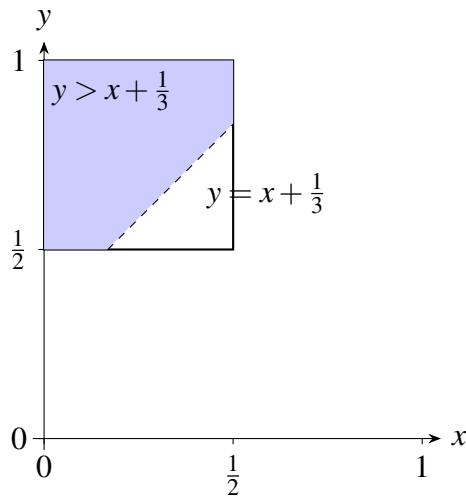
Example 6.24. One point is randomly selected on the interval $[0, \frac{1}{2}]$. Another point is randomly selected on the interval $[\frac{1}{2}, 1]$. What is the exact probability that the distance between the two points is more than $\frac{1}{3}$?

Solution. Define

$$X \sim U\left(0, \frac{1}{2}\right) \quad \text{and} \quad Y \sim U\left(\frac{1}{2}, 1\right).$$

Then, the distance between the two points is $Y - X$. For each fixed $x \in [0, \frac{1}{2}]$, we need $y > x + \frac{1}{3}$. Note that the joint density is uniform on the rectangle bounded by $0 \leq x \leq \frac{1}{2}$ and $\frac{1}{2} \leq y \leq 1$. We need $y > x + \frac{1}{3}$, so the answer is

$$\frac{\text{area of blue pentagon}}{\text{area of rectangle bounded by } 0 \leq x \leq \frac{1}{2} \text{ and } \frac{1}{2} \leq y \leq 1}.$$



The area of the blue pentagon is

$$\frac{1}{2} \left(\frac{1}{6} + \frac{1}{2} \right) \left(\frac{1}{3} \right) + \frac{1}{6} \cdot \frac{1}{2}.$$

Next, the area of the rectangle is $\frac{1}{4}$, so the answer is $\frac{7}{9}$. \square

Example 6.25 (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 $2/6 = 1/3$. \square

Example 6.26 (Ross p. 305 Question 26). 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, its discriminant $B^2 - 4AC \geq 0$. Since A, B, C are independent, we can condition on $C = c$ and integrate over c from 0 to 1 to obtain

$$P(B^2 \geq 4AC) = \int_0^1 P(B^2 \geq 4Ac) dc.$$

Note that this involved the law of total probability (Proposition 3.3). Now, fix $c \in [0, 1]$. We evaluate the probability that $B^2 \geq 4Ac$ over the square $A, B \in [0, 1]^2$. This corresponds to computing the area of the region

$$\{(a, b) \in [0, 1]^2 : b^2 \geq 4ac\}.$$

To proceed, we split into two cases based on the value of c . We shall consider when $4ac < 1$ and when $4ac > 1$. Since $0 \leq a \leq 1$, it is natural for us to consider the cases $0 < c < \frac{1}{4}$ and $c \geq \frac{1}{4}$.

- **Case 1:** If $0 < c < \frac{1}{4}$, it is possible for b^2 to be greater than $4ac$ because $4ac < 1$. For a fixed a , we need $b \geq \sqrt{4ac}$, so

$$P(B^2 \geq 4Ac) = \int_0^1 \int_{\sqrt{4ac}}^1 db da = \int_0^1 (1 - \sqrt{4ac}) da.$$

We integrate to obtain

$$P(B^2 \geq 4Ac) = 1 - \frac{4}{3}\sqrt{c}.$$

- **Case 2:** If $c \geq \frac{1}{4}$, then $4ac$ can be greater than 1, especially when a is large. For each b , we find the upper limit of a such that $b^2 \geq 4ac$, i.e. $a \leq \frac{b^2}{4c}$. So,

$$P(B^2 \geq 4Ac) = \int_0^1 \int_0^{b^2/4c} dadb = \int_0^1 \frac{b^2}{4c} db = \frac{1}{4c} \int_0^1 b^2 db = \frac{1}{12c}.$$

Now, we put everything together by integrating over c and we obtain.

$$P(B^2 \geq 4AC) = \int_0^{1/4} \left(1 - \frac{4}{3}\sqrt{c}\right) dc + \int_{1/4}^1 \frac{1}{12c} dc.$$

which yields $\frac{1}{36}(5 + 3\ln 4)$. □

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

$$p_{X|Y}(x | y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \quad \text{for all } y \text{ such that } p_Y(y) > 0. \quad (6.1)$$

Similarly, the conditional distribution function of X given that $Y = y$ is defined by

$$F_{X|Y}(x | y) = P(X \leq x | Y = y) \quad \text{for all } y \text{ such that } p_Y(y) > 0.$$

It follows that

$$F_{X|Y}(x | y) = \sum_{a \leq x} p_{X|Y}(a | y).$$

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 $p_Y(y) > 0$.

Example 6.27. Suppose X and Y are discrete random variables whose joint probability mass function is given by $p(1,1) = \frac{1}{8}$, $p(1,2) = \frac{1}{4}$, $p(2,1) = \frac{1}{8}$, $p(2,2) = \frac{1}{2}$. Determine the conditional probability mass function $p_{X|Y}(x|y)$ of X given $Y = y$.

- (a) $p_{X|Y}(1|1)$
- (b) $p_{X|Y}(2|1)$
- (c) $p_{X|Y}(1|2)$
- (d) $p_{X|Y}(2|2)$

Solution. Using the conditional density formula in (6.1), we can derive the following:

- (a) So,

$$p_{X|Y}(1|1) = \frac{p_{X,Y}(1,1)}{p_Y(1)} = \frac{\frac{1}{8}}{\frac{1}{8} + \frac{1}{8}} = \frac{1}{2}.$$

- (b) Next,

$$p_{X|Y}(2|1) = \frac{p_{X,Y}(2,1)}{p_Y(1)} = \frac{\frac{1}{8}}{\frac{1}{8} + \frac{1}{8}} = \frac{1}{2}.$$

- (c) We have

$$p_{X|Y}(1|2) = \frac{p_{X,Y}(1,2)}{p_Y(2)} = \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{2}} = \frac{1}{3}.$$

- (d) In a similar fashion, one can show that this probability is $\frac{2}{3}$. □

Example 6.28. A standard die is thrown before flipping as many fair coins as given by the die value. We then count the number of heads counted by the coins.

- (a) What is the probability that no head is flipped?
- (b) Given that at least 1 head was flipped, what is the probability that exactly 1 head was flipped?

Solution.

- (a) Let X be the number of heads obtained. Then, $X \sim B(N, 0.5)$, where N follows a discrete uniform distribution over the integers from 1 through 6 inclusive. Then,

$$P(X=0) = \sum_{i=1}^6 \frac{1}{6} \cdot 0.5^i = \frac{21}{128}.$$

- (b) We wish to find $P(X=1|X \geq 1)$. By Definition 3.1, this is precisely

$$\frac{P(X=1)}{1 - P(X=0)},$$

where

$$P(X=1) = \sum_{i=1}^6 \frac{1}{6} \cdot i(0.5)^i = \frac{5}{16}.$$

We computed $P(X=1)$ in (a) so the answer is $\frac{40}{107}$. □

Example 6.29. A fair dice is rolled. Let N denote the number shown on the dice. A fair coin is then tossed N times. Let X denote the number of heads seen. Find the following probabilities:

- (a) $P(X = 4 | N = 5)$
- (b) $P(X = 4)$
- (c) $P(N = 5 | X = 4)$

Solution.

- (a) By Definition 3.1,

$$P(X = 4 | N = 5) = \frac{P(X = 4 \text{ and } N = 5)}{P(N = 5)} = \frac{\binom{5}{4} 0.5^4 \cdot 0.5 \cdot \frac{1}{6}}{\frac{1}{6}} = \frac{5}{32}.$$

- (b) By the law of total probability (Proposition 3.3),

$$P(X = 4) = \sum_{k=1}^6 P(X = 4 | N = k) P(N = k) = \frac{1}{6} \sum_{k=4}^6 P(X = 4 | N = k)$$

We see that

$$P(X = 4 | N = k) = \binom{k}{4} 0.5^k$$

so the answer is

$$\frac{1}{6} \sum_{k=4}^6 \binom{k}{4} 0.5^k = \frac{29}{384}.$$

- (c) By Bayes' theorem (Theorem 3.1),

$$P(N = 5 | X = 4) = \frac{P(X = 4 | N = 5) P(N = 5)}{P(X = 4)}.$$

We computed $P(X = 4 | N = 5)$ in (a) and $P(X = 4)$ in (b). Note that $P(N = 5) = \frac{1}{6}$. Substituting all the probabilities yields the answer, which is $\frac{10}{29}$. \square

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

$$f_{X|Y}(x | y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad \text{for all } y \text{ such that } f_Y(y) > 0. \quad (6.2)$$

For $A \subseteq \mathbb{R}$ and y such that $f_Y(y) > 0$,

$$P(X \in A | Y = y) = \int_A f_{X|Y}(x | y) dx.$$

The conditional distribution of X given that $Y = y$ is defined by

$$F_{X|Y}(x | y) = P(X \leq x | Y = y) = \int_{-\infty}^x f_{X|Y}(t | y) dt.$$

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 $f_Y(y) > 0$.

Example 6.30. The random variables X and Y are jointly continuous, with a joint probability density function given by

$$f(x, y) = \begin{cases} \frac{1}{8} (x^2 - y^2) e^{-x} & \text{if } x > 0 \text{ and } -x < y < x; \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Determine the conditional density function $f_{X|Y}(x | y)$ of X given $Y = y$. Given that $Y = 1$, what is the probability that $2 < X < 3$?
- (b) Determine the conditional density function $f_{Y|X}(y | x)$ of Y given $X = x$. Given that $X = 2$, what is the probability that $0 < Y < 1$?

Solution.

- (a) Recall from (6.2) that

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

We shall compute the marginal density

$$f_Y(y) = \int_y^\infty \frac{1}{8} (x^2 - y^2) e^{-x} dx = \frac{y+1}{4e^y}.$$

So,

$$f_{X|Y}(x | y) = \frac{1}{8e^y} (x^2 - y^2) \cdot \frac{4e^y}{y+1} = \frac{e^{y-x} (x^2 - y^2)}{2(y+1)}.$$

Next,

$$P(2 < X < 3 | Y = 1) = \int_2^3 \frac{e^{1-x} (x^2 - 1)}{4} dx = \frac{1}{4} \left(\frac{9}{e} - \frac{16}{e^2} \right).$$

- (b) Again by (6.2), we have

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

We then compute the marginal density

$$f_X(x) = \int_{-x}^x \frac{1}{8} (x^2 - y^2) e^{-x} dy = \frac{x^3 e^{-x}}{6}.$$

So,

$$f_{Y|X}(y | x) = \frac{1}{8} (x^2 - y^2) e^{-x} \cdot \frac{6}{x^3 e^{-x}} = \frac{3 (x^2 - y^2)}{4x^3}.$$

Next,

$$P(0 < Y < 1 | X = 2) = \int_0^1 \frac{3 (4 - y^2)}{4 \cdot 2^3} dy = \frac{11}{32}.$$

□

Example 6.31. We are given a coin. It can be tossed any number of times to generate independent outcomes (heads vs tails). Its probability X of showing heads is assumed to have a continuous distribution on the interval $(0, 1)$ with density function given by

$$f(x) = \begin{cases} cx^3(1-x)^2 & \text{if } 0 < x < 1; \\ 0 & \text{otherwise.} \end{cases}$$

Here, c is some real constant.

- (a) What is the probability that when we toss the coin 10 times, we see exactly 7 heads among them?
- (b) We toss the coin 10 times and we see exactly 7 heads among them. What is the probability that $0.65 < X < 0.75$?

Solution. First, note that

$$\int_0^1 cx^3(1-x)^2 dx = 1 \quad \text{so} \quad c = 60.$$

- (a) We have

$$\begin{aligned} P(\text{7 heads out of 10}) &= \int_0^1 P(\text{7 heads} \mid X = x) f(x) dx \\ &= \int_0^1 \binom{10}{7} x^7 (1-x)^3 \cdot 60x^3 (1-x)^2 dx \\ &= 60 \binom{10}{7} \int_0^1 x^{10} (1-x)^5 dx \end{aligned}$$

Recall a nice formula for the beta function (Lemma 5.2) where we mentioned how it is closely related to the gamma function. It states that

$$\int_0^1 t^{a-1} (1-t)^{b-1} dt = \frac{(a-1)!(b-1)!}{(a+b-1)!} \quad \text{for any } a, b \in \mathbb{N}, \quad (6.3)$$

where one should also recall that $\Gamma(a) = (a-1)!$. Set $a = 11$ and $b = 6$ in (6.3) to obtain the required answer which is 0.14985.

- (b) By the conditional probability formula (Definition 3.1),

$$P(0.65 < X < 0.75 \mid \text{7 heads out of 10})$$

is equal to

$$\frac{P(\text{7 heads out of 10 and } 0.65 < X < 0.75)}{P(\text{7 heads out of 10})}. \quad (6.4)$$

The numerator is equal to

$$\int_{0.65}^{0.75} \binom{10}{7} x^7 (1-x)^3 \cdot 60x^3 (1-x)^2 dx$$

which is equal to 0.048009. Substituting everything into (6.4), the required answer is $\frac{0.048009}{0.14985} = 0.320$. \square

Example 6.32. Suppose U is a uniform random variable on $(0, 1)$, and $X | U \sim B(n, p)$. Prove that X is a discrete uniform random variable on $\{0, 1, \dots, n\}$.

Solution. First, condition on $U = u$. Say $X | U = u \sim B(n, u)$. We have

$$P(X = x | U = u) = \binom{n}{k} u^k (1-u)^{n-k} \quad \text{where } 0 \leq k \leq n.$$

Hence,

$$P(X = x) = \int_0^1 \binom{n}{k} u^k (1-u)^{n-k} f_U(u) du = \binom{n}{k} \int_0^1 u^k (1-u)^{n-k} du. \quad (6.5)$$

Note that the integral in (6.5) represents the beta function. Recall from Definition 5.11 and Lemma 5.2 that

$$B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt \quad \text{and} \quad B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Hence,

$$P(X = x) = \binom{n}{k} \cdot \frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1}$$

and the result follows. \square

Example 6.33 (Ross p. 313 Question 14). Let N be a geometric random variable with parameter p . Suppose that the conditional distribution of X given that $N = n$ is the gamma distribution with parameters n and λ . Find the conditional probability mass function of N given that $X = x$.

Solution. We have $N \sim \text{Geo}(p)$ and $X | N \sim \text{Gamma}(n, \lambda)$. So,

$$f_{X|N}(x | n) = \frac{(\lambda x)^{n-1} \lambda e^{-\lambda x}}{\Gamma(n)}.$$

As such,

$$\frac{f_{X,N}(x, n)}{P(N = n)} = \frac{(\lambda x)^{n-1} \lambda e^{-\lambda x}}{\Gamma(n)}.$$

Since $N \sim \text{Geo}(p)$, then $P(N = n) = p(1-p)^{n-1}$. As such,

$$f_{X,N}(x, n) = \frac{p(\lambda x(1-p))^{n-1} \lambda e^{-\lambda x}}{\Gamma(n)}.$$

Hence,

$$P(N | X = x) = \frac{f_{X,N}(x, n)}{f_X(x)} = \frac{p(\lambda x(1-p))^{n-1} \lambda e^{-\lambda x}}{\Gamma(n) f_X(x)}. \quad (6.6)$$

By definition of the gamma function, we have $\Gamma(n) = (n-1)!$. We then perform a change of variable from n to k to obtain

$$f_X(x) = \sum_{n=1}^{\infty} \frac{p(\lambda x(1-p))^{n-1} \lambda e^{-\lambda x}}{\Gamma(n)} = p \lambda e^{-\lambda x} \sum_{k=0}^{\infty} \frac{(\lambda x(1-p))^k}{k!}.$$

Upon simplification, we see that $f_X(x) = p \lambda e^{-p \lambda x}$. Substituting this into (6.6), we have

$$P(N | X = x) = \frac{f_{X,N}(x, n)}{f_X(x)} = \frac{(\lambda x(1-p))^{n-1}}{\Gamma(n)} \cdot e^{-(1-p)\lambda x}.$$

As such, $N - 1 \sim \text{Po}(\lambda x(1-p))$. \square

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 . 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 in Algorithm 6.1.

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 6.34 (Ross p. 294). 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,

$$f_{R,\Theta}(r, \theta) = \frac{f_{X,Y}(x, y)}{\det(J(x, y))} = \sqrt{x^2 + y^2} \cdot \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right) = \frac{1}{2\pi} r e^{-\frac{r^2}{2}}$$

which is the joint probability density function of R and Θ .

- (ii) They are independent. □

In Example 6.34, 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 6.35 (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 $P(U \geq 1)$.

Solution.

- (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

$$f_{U,V}(u,v) = \frac{f_{X,Y}(x,y)}{|J(x,y)|} \quad \text{then} \quad f_{U,V}(u,v) = \frac{e^{-y}}{x+y} \quad (6.7)$$

but it is not in terms of u and v ! As such, consider $Y = U + X$, so $Y = U + \frac{V}{Y}$. This 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 $f_{U,V}(u,v)$ in (6.7), we have

$$f_{U,V}(u,v) = \frac{\exp\left[-(U + \sqrt{U^2 + 4V})/2\right]}{\sqrt{U^2 + 4V}} \quad \text{where } u \geq -1, v \geq 0.$$

- (b) We first find the marginal density of U . That is, finding $f_U(u)$ from $f_{U,V}(u,v)$. So,

$$f_U(u) = \int_0^\infty \frac{\exp\left[-(u + \sqrt{u^2 + 4v})/2\right]}{\sqrt{u^2 + 4v}} dv.$$

We use the substitution $t = -\frac{u+\sqrt{u^2+4v}}{2}$, so $\frac{dt}{dv} = -\frac{1}{\sqrt{u^2+4v}}$. The integral becomes

$$\int_{-u}^{-\infty} -e^t dt = e^{-u}$$

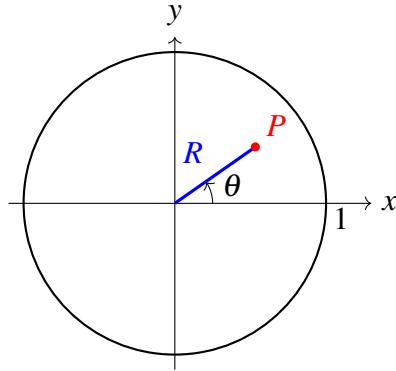
so $f_U(u) = e^{-u}$. So,

$$P(U \geq 1) = \int_1^\infty e^{-u} du = 1/e.$$

□

Example 6.36. A point P is picked at random from the unit disk in the plane centred at the origin, i.e. from the set

$$\{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}.$$



- (a) What is the probability that the line joining P to the origin makes an angle (in either clockwise or anticlockwise direction) of less than 10° with the x -axis?
- (b) What is the probability that the point P lies within the disk of radius $1/3$ centred at the origin?

Solution. We parametrise the position of P using (R, Θ) , where R is the random variable denoting the distance P makes with the origin and Θ is the angle P makes with the x -axis. We first find the joint density function of R and Θ . Since the point is uniformly chosen in the unit disk, then

$$f_{X,Y}(x,y) = \frac{1}{\pi}.$$

Perform the usual change of variables involving the Jacobian to obtain

$$f_{R,\Theta}(r, \theta) = \frac{r}{\pi}.$$

Hence, the joint density function is

$$f_{R,\Theta}(r, \theta) = \frac{2r}{\pi} \quad \text{where } 0 \leq r \leq 1, -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}.$$

By independence of R and Θ , we have

$$f_R(r) = 2r \text{ where } 0 \leq r \leq 1 \quad \text{and} \quad f_\Theta(\theta) = \frac{1}{\pi} \text{ where } -\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}.$$

- (a) The probability is

$$\int_{-\pi/18}^{\pi/18} \frac{1}{\pi} d\theta = \frac{1}{9}.$$

- (b) The probability is

$$\int_0^{1/3} 2r dr = \frac{1}{9}.$$

□

Example 6.37 (Ross p. 313 Question 15). Let X and Y be independent uniform $(0, 1)$ random variables. Find the joint density of $U = X$ and $V = X + Y$.

Solution. Note that $X = U$ and $Y = V - U$. The Jacobian matrix is

$$\mathbf{J} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \quad \text{which has determinant 1.}$$

Since $0 \leq x \leq 1$ and $0 \leq y \leq 1$, then $0 \leq v - u \leq 1$ so $u \leq v \leq u + 1$ and $0 \leq u \leq 1$. Hence,

$$f_{U,V}(u, v) = 1 \quad \text{where } u \leq v \leq u + 1 \text{ and } 0 \leq u \leq 1.$$

□

Example 6.38. Consider the random variables X and Y with joint probability density function given by

$$f(x, y) = \begin{cases} \frac{1}{8} (x^2 - y^2) e^{-x} & \text{if } x > 0 \text{ and } -x < y < x; \\ 0 & \text{otherwise.} \end{cases}$$

Compute the following:

- (a) $E(XY)$
- (b) $E((XY)^k)$ for any $k \geq 2$
- (c) $E(e^{XY})$
- (d) $E\left(\frac{1}{XY^2}\right)$

Solution.

- (a) The trick is to realise that $f_{Y|X}$ is symmetric so $E(XY) = 0$. Alternatively, using integration, one can see that

$$E(XY) = \int_0^\infty \int_{-x}^x xy \cdot \frac{1}{8} (x^2 - y^2) e^{-x} dy dx = 0.$$

- (b) If k is odd, one can infer that the expectation is equal to 0. On the other hand, if k is even, one can deduce that

$$E((XY)^{2k}) = \int_0^\infty \int_{-x}^x (xy)^{2k} \cdot \frac{1}{8} (x^2 - y^2) e^{-x} dy dx$$

which is equal to

$$\frac{(2k+3)-(2k+1)}{4(2k+1)(2k+3)} \int_0^\infty e^{-x} x^{4k+3} dx = \frac{1}{2(2k+1)(2k+3)} \cdot (4k+3)!$$

where we used the definition of the gamma function.

- (c) Note that

$$E(e^{XY}) = E\left(\sum_{n=0}^{\infty} \frac{(XY)^n}{n!}\right) = \sum_{n=0}^{\infty} \frac{(4n+3)!}{(2n)!(2n+1)(2n+3)}$$

which diverges.

(d) We have

$$E\left(\frac{1}{XY^2}\right) = \int_0^\infty \int_{-x}^x \frac{1}{xy^2} \cdot \frac{1}{8} (x^2 - y^2) e^{-x} dy dx$$

which goes to infinity. \square

Example 6.39. Consider the variables X and Y with joint probability density function given by

$$f(x,y) = \begin{cases} \frac{2}{x^2y^3} & \text{if } x > 1 \text{ and } y > 1; \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute $E(X^2Ye^{-X})$.

(b) Compare it to $E(X^2e^{-X})$ and $E(Y)$.

Solution.

(a) We have

$$E(X^2Ye^{-X}) = \int_1^\infty \int_1^\infty x^2ye^{-x} \cdot \frac{2}{x^2y^3} dx dy = \frac{2}{e}.$$

(b) We have

$$E(X^2e^{-X}) = \int_1^\infty \int_1^\infty x^2e^{-x} \cdot \frac{2}{x^2y^3} dx dy = \frac{1}{e}$$

and

$$E(Y) = \int_1^\infty \int_1^\infty y \cdot \frac{2}{x^2y^3} dx dy = 2.$$

This shows that

$$E(X^2Ye^{-X}) = E(X^2e^{-X}) E(Y)$$

because the joint density function can be factorised as an expression of the form $g(x)h(y)$. \square

Example 6.40. Consider the random variables X and Y with joint probability density function given by

$$f(x,y) = \begin{cases} \frac{6}{7}(x^2 + \frac{xy}{2}) & \text{if } 0 < x < 1 \text{ and } 0 < y < 2; \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute $E(XY)$.

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

Solution.

(a) We have

$$E(XY) = \int_0^2 \int_0^1 xy \cdot \frac{6}{7} \left(x^2 + \frac{xy}{2}\right) dx dy = \frac{17}{21}.$$

(b) We have

$$E(X^2Y^2) = \int_0^2 \int_0^1 x^2y^2 \cdot \frac{6}{7} \left(x^2 + \frac{xy}{2}\right) dx dy = \frac{31}{35}$$

so $\text{Var}(XY) = \frac{508}{2205}$. \square

Expectation Properties

7.1 Expectation of Sums of Random Variables

We start off this chapter with the following result: if $a \leq X \leq b$, then $a \leq E(X) \leq b$. It is not difficult to see why this is true — we will only 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. We have

$$E(X) = \sum_{\text{all } x} xp(x) \geq \sum_{\text{all } x} ap(x) = a.$$

In a similar fashion, we can use the same technique to show that $E(X) \leq b$.

Example 7.1 (Ross p. 390 Question 6). A fair die is rolled 10 times. Calculate the expected sum of the 10 rolls.

Solution. Let X_i denote the value of the i^{th} roll. Note that $X_1 = \dots = X_{10}$. Define $S = X_1 + \dots + X_{10}$. Then, Then,

$$E(S) = E(X_1 + \dots + X_{10}) = 10E(X) = 10 \cdot 3.5 = 35.$$

□

Proposition 7.1. The following hold:

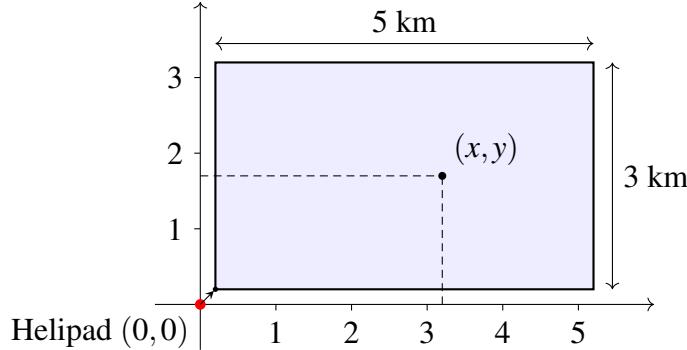
- (i) If X and Y are jointly discrete random variables with joint probability density function $p_{X,Y}$, then

$$E[g(X,Y)] = \sum_{\text{all } y} \sum_{\text{all } x} g(x,y) p_{X,Y}(x,y)$$

- (ii) If X and Y are jointly continuous random variables with joint probability density function $f_{X,Y}$, then

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

Example 7.2 (Ross p. 390 Question 5). A city in the shape of a rectangle stretches 5 kilometres from west to east and 3 kilometres from north to south. A rescue helicopter waits in a helipad just outside the city near the southwestern corner, with coordinates $(0, 0)$. A rescue call, which follows a uniform distribution, can arrive at any point (x, y) in the city. Find the expected distance covered by the helicopter in travelling to this point.



Solution. The distance D is given by the expression

$$D = \sqrt{X^2 + Y^2} \quad \text{where } 0 \leq x \leq 5 \text{ and } 0 \leq y \leq 3.$$

So,

$$E(D) = \frac{1}{15} \int_0^5 \int_0^3 \sqrt{x^2 + y^2} dy dx.$$

□

Example 7.3 (Ross p. 390 Question 4). If X and Y have joint density function

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{3}(x+y), & \text{if } 0 < x < 1, 0 < y < 2, \\ 0, & \text{otherwise} \end{cases}$$

find:

- (a) $E[XY]$
- (b) $E[X]$
- (c) $E[Y^2]$

Solution.

- (a) We have

$$E(XY) = \int_0^2 \int_0^1 xy f_{X,Y}(x,y) dx dy = \frac{2}{3}.$$

- (b) We have

$$E(X) = \int_0^1 \int_0^2 xf_{X,Y}(x,y) dy dx = \frac{5}{9}.$$

Alternatively, we can compute the marginal density

$$f_X(x) = \int_0^2 f_{X,Y}(x,y) dy = \frac{2}{3}x + \frac{2}{3}$$

so

$$E(X) = \int_0^1 x \left(\frac{2}{3}x + \frac{2}{3} \right) dx = \frac{5}{9}.$$

(c) We have

$$E(Y^2) = \int_0^1 \int_0^2 y^2 f_{X,Y}(x,y) dy dx = \frac{16}{9}.$$

□

Corollary 7.1. Some important consequences are as follows:

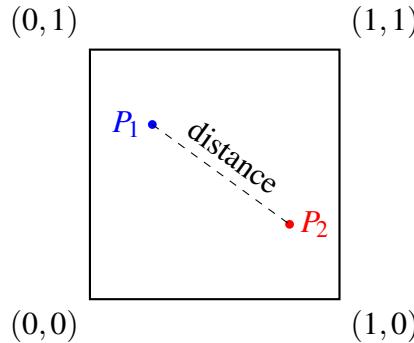
- (i) **Non-negativity:** If $g(x,y) \geq 0$ whenever $p_{X,Y}(x,y) > 0$, then $E[g(X,Y)] \geq 0$
- (ii) **Linearity:** $E[g(X,Y) + h(X,Y)] = E[g(X,Y)] + E[h(X,Y)]$
- (iii) **Linearity:** $E[g(X) + h(Y)] = E[g(X)] + E[h(Y)]$
- (iv) **Monotonicity:** If jointly distributed random variables X and Y satisfy $X \leq Y$, then $E(X) \leq E(Y)$. Of course, this result can be easily extended to

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

The formula for the expectation of the sample mean, \bar{X} , can be derived from (iv) in Corollary 7.1. It is clear that $E(\bar{X}) = \mu$, 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 7.4. 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 $E(X) = np$.

Example 7.5 (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?



This is a very interesting problem. The answer is actually not $\frac{1}{2}$, but rather close to it! The mean distance is

$$\frac{2 + \sqrt{2} + 5 \ln(1 + \sqrt{2})}{15} \approx 0.52140.$$

Let us prove this result.

Solution. Let $P_1 = (U_1, V_1)$ and $P_2 = (U_2, V_2)$, where $U_1, U_2, V_1, V_2 \sim U(0, 1)$ are independent uniform random variables. The Euclidean distance between P_1 and P_2 is

$$D = \sqrt{(U_1 - U_2)^2 + (V_1 - V_2)^2}.$$

We wish to compute $E(D)$. Define

$$X = U_1 - U_2 \quad \text{and} \quad Y = V_1 - V_2.$$

Here, X and Y denote the difference of two independent $U(0, 1)$ random variables, and it is a known fact that X and Y follow a triangular distribution with density

$$f_X(x) = 1 - |x| \quad \text{where } -1 < x < 1 \quad (7.1)$$

and similarly for Y . To briefly see why, we first need to compute $F_X(x) = P(X \leq x) = P(U_1 - U_2 \leq x)$. Fix $u_2 \in [0, 1]$. The condition is $0 \leq u_1 \leq 1$ and $u_1 \leq u_2 + x$. So, the range of values for u_1 is $0 \leq u_1 \leq \min\{1, u_2 + x\}$ provided that $u_2 + x \geq 0$, otherwise the range is empty. Thus, the inner integral is

$$\int_0^{\min\{1, u_2 + x\}} du_1 = \begin{cases} 0 & \text{if } u_2 + x \leq 0; \\ u_2 + x & \text{if } 0 < u_2 + x < 1; \\ 1 & \text{if } u_2 + x \geq 1. \end{cases}$$

For the outer integral, we integrate this over $u_2 \in [0, 1]$. We proceed with casework. If $x \leq -1$, then $u_2 + x \leq 0$ for all $u_2 \in [0, 1]$, implying that $F_X(x) = 0$. Next, if $-1 < x < 0$, for $u_2 \in [0, -x]$, we have $u_2 + x \leq 0$ so the inner integral is 0; for $u_2 \in [-x, 1]$, we have $0 < u_2 + x < 1$ so the inner integral is $u_2 + x$. Thus,

$$F_X(x) = \int_{-x}^1 u_2 + x \, du_2 = \frac{1}{2} + x + \frac{1}{2}x^2.$$

Next, if $0 \leq x < 1$, if $u_2 \in [0, 1-x]$, we have $u_2 + x \leq 1$ so the inner integral becomes $u_2 + x$; if $u_2 \in [1-x, 1]$, we have $u_2 + x \geq 1$ so the inner integral becomes 1. Thus,

$$F_X(x) = \int_0^{1-x} u_2 + x \, du_2 + \int_{1-x}^1 du_2 = \frac{1}{2} + x - \frac{1}{2}x^2.$$

For $x \geq 1$, we have $F_X(x) = 1$. After obtaining the distribution function $F_X(x)$, differentiate it to get (7.1). Since X and Y are independent random variables, then

$$f_{X,Y}(x, y) = (1 - |x|)(1 - |y|) \quad \text{where } -1 < x, y < 1.$$

The expected distance is

$$E(D) = \iint_{[-1,1]^2} \sqrt{x^2 + y^2} (1 - |x|)(1 - |y|) \, dxdy$$

by using the formula

$$E(D) = E(g(X, Y)) = \iint_{\mathbb{R}^2} g(x, y) f_{X,Y}(x, y) \, dxdy.$$

By symmetry, we can restrict the domain to the first quadrant then multiply the integral by 4, so we obtain

$$E(D) = 4 \int_0^1 \int_0^1 \sqrt{x^2 + y^2} (1-x)(1-y) dx dy.$$

Naturally, we would turn to polar coordinates. Let $x = r \cos \theta$ and $y = r \sin \theta$, where $0 \leq r \leq 1$ and $0 \leq \theta \leq \frac{\pi}{2}$. The Jacobian matrix is

$$\mathbf{J}(x, y) = \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix} \quad \text{so} \quad \text{determinant of Jacobian matrix is } = r.$$

Hence, the double integral becomes $4(A + B)$, where

$$A = \int_0^{\pi/4} \int_0^{1/\cos \theta} r^2 (1 - r \cos \theta)(1 - r \sin \theta) dr d\theta$$

$$B = \int_{\pi/4}^{\pi/2} \int_0^{1/\sin \theta} r^2 (1 - r \cos \theta)(1 - r \sin \theta) dr d\theta$$

This eventually yields the desired result. \square

Example 7.6 (Erdős–Rényi model). A group of 10 students is chosen at random from a class. Students are friends with each other with probability p , and this independently of the other friendships within the class.

- (a) On average, how many pairs of friends can we form?
- (b) What is the minimal value of p such that, on average, we can find five students who all know each other?

We now assume that if two students have one friend in common, they will also eventually become friends.

- (c) What is the expected number of *eventual* groups of three friends?

Solution.

- (a) There are $\binom{10}{2} = 45$ unordered pairs, each an edge with probability p . By the linearity of expectation, the expected number of pairs is $45p$.
- (b) For any fixed set of 5 students, all $\binom{5}{2} = 10$ edges must be present, and this occurs with probability p^{10} . There are $\binom{10}{5} = 252$ such sets, so $E(X) = 252p^{10}$. We need $E(X) \geq 1$ so $p \geq 0.575$.
- (c) Under the rule that ‘if two students have one friend in common, they eventually become friends’, any triple that initially spans at least two edges will close up to a triangle, and a triple that already has 3 edges is already a triangle. For one fixed triple, the three possible edges are independent, so if $Y \sim B(3, p)$ counts the present edges, then the probability that we have an eventual triangle on this triple is

$$P(Y \geq 2) = \binom{3}{2} p^2 (1-p) + p^3 = 3p^2 (1-p) + p^3.$$

As there are $\binom{10}{3} = 120$ triples, by the linearity of expectation, it follows that the answer is $120(3p^2 - 2p^3)$. \square

Let us give some insight into Example 7.6. We can model the situation using a *random graph*. Let the 10 students be represented by vertices $V = \{1, 2, \dots, 10\}$, and for each unordered pair (i, j) we place an edge between i and j with probability p , independently of all other pairs. This is exactly the Erdős–Rényi random graph model $G(n, p)$ with $n = 10$. For (a), we are counting the expected number of edges in the graph $G(10, p)$. In $G(10, p)$, there are $\binom{10}{2} = 45$ possible edges, each present with probability p . If we write X for the total number of edges, we may decompose

$$X = \sum_{1 \leq i < j \leq 10} I_{ij},$$

where I_{ij} is the indicator that the edge (i, j) is present. Hence, $E(I_{ij}) = p$ so $E(X) = 45p$. This is the standard formula for the expected number of edges in $G(n, p)$, which is the expected number of edges being equal to $\binom{n}{2}p$.

In (b), we are counting the number of complete graphs with 5 vertices, also known as K_5 . A set of 5 students who all know each other corresponds to a copy of the complete graph K_5 in $G(10, p)$. For any fixed 5-element subset $S \subseteq V$, the induced subgraph on S is a K_5 exactly when all $\binom{5}{2} = 10$ edges among these vertices are present. Since edges appear independently with probability p , this event has probability p^{10} . There are $\binom{10}{5} = 252$ possible 5-element subsets S , so if Y denotes the number of K_5 's in $G(10, p)$, then $E(Y) = \binom{10}{5}p^{10} = 252p^{10}$.

In summary, Example 7.6 is naturally phrased in the language of Erdős–Rényi model random graphs $G(n, p)$. (a) and (b) compute expected numbers of edges and cliques in $G(10, p)$, while (c) studies how a simple local rule (closing wedges into triangles) modifies the structure of such a random graph and increases its clustering. Figure 7.1 shows a sample friendship network on 10 students. Each edge is present independently with probability p , so this is an instance of the Erdős–Rényi model random graph $G(10, p)$. The highlighted triangle is one possible group of three mutual friends.

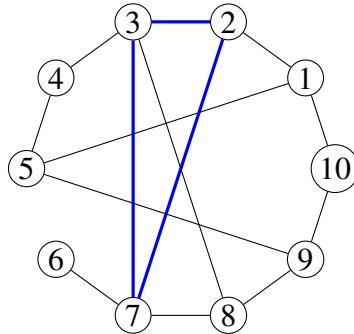


Figure 7.1: A sample friendship network on 10 students

Example 7.7. A class is composed of n students, whose birthdays are independently and uniformly distributed.

- (a) On average, how many pairs of students are born on the same day?

- (b) What is the minimal value of n such that we expect 3 students to be born on the same day?

Solution.

- (a) Let X be the number of pairs of students born on the same day. Then,

$$x_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ on the same day;} \\ 0 & \text{otherwise.} \end{cases}$$

Define

$$X = \sum_{i,j} x_{ij}.$$

Note that x_{ij} is a Bernoulli random variable with parameter $p = \frac{1}{365}$. So,

$$E(X) = \sum_{i,j} E(X_{ij}) = \sum_{i,j} \frac{1}{365} = \binom{n}{2} \cdot \frac{1}{365} = \frac{n(n-1)}{730}.$$

- (b) Define

$$Y_{i,j,k} = \begin{cases} 1 & \text{if } i = j = k; \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$E(Y) = E\left(\sum_{i,j,k} Y_{i,j,k}\right) = \sum_{i,j,k} E(Y_{i,j,k}) = \sum_{i,j,k} \frac{1}{365^2} = \binom{n}{3} \cdot \frac{1}{365^2}.$$

So, the smallest value of n is 94. □

7.2 Covariance, Variance and Correlation

Definition 7.1 (covariance). The covariance of jointly distributed random variables X and Y , denoted by $\text{cov}(X, Y)$, is defined by

$$\text{cov}(X, Y) = E(X - \mu_X)(Y - \mu_Y),$$

where μ_X and μ_Y denote the means of X and Y respectively.

Definition 7.2 (correlation). If $\text{cov}(X, Y) \neq 0$, we say that X and Y are correlated, but if $\text{cov}(X, Y) = 0$, we say that X and Y are uncorrelated.

An alternative formula for covariance is

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y).$$

As a result, if X and Y are independent, it is clear that $\text{cov}(X, Y) = 0$. However, the converse is not true. Correlation does not imply causation¹.

¹I 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 7.2. If X and Y are independent random variables, then for any functions $g, h : \mathbb{R} \rightarrow \mathbb{R}$, we have

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

Some other properties of covariance are as follows:

(i) $\text{Var}(X) = \text{cov}(X, X)$

(ii) **Symmetry:** $\text{cov}(X, Y) = \text{cov}(Y, X)$

(iii)

$$\text{cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{cov}(X_i, Y_j)$$

(iv)

$$\text{cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{cov}(X_i, Y_j)$$

(v)

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

We only prove (iv).

Proof.

$$\begin{aligned} \text{cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j\right) &= E\left(\sum_{i=1}^n \sum_{j=1}^m a_i b_j X_i Y_j\right) - E\left(\sum_{i=1}^n a_i X_i\right) E\left(\sum_{j=1}^m b_j Y_j\right) \\ &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j E(X_i Y_j) - \sum_{i=1}^n \sum_{j=1}^m a_i b_j E(X_i) E(Y_j) \\ &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{cov}(X_i, Y_j) \end{aligned}$$

□

Let X_1, X_2, \dots, X_n be independent random variables. Recall from H2 Mathematics that

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

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

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

be the sample mean. The quantities $X_i - \bar{X}$, for $1 \leq i \leq n$, are called deviations as they equal to the differences between the individual data and the sample mean. The random variable

$$S^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$$

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

$$\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n}.$$

Proof. Note that $X_i - \bar{X} = X_i - \mu + \mu - \bar{X}$. Hence,

$$\begin{aligned} (X_i - \bar{X})^2 &= (X_i - \mu + \mu - \bar{X})^2 \\ &= (X_i - \mu)^2 + (\bar{X} - \mu)^2 - 2(\bar{X} - \mu)(X_i - \mu) \end{aligned}$$

When we take the sum of i from 1 to n , note that \bar{X} is unaffected by the index. Hence,

$$\begin{aligned} S^2 &= \frac{1}{n-1} \left[\sum_{i=1}^n (X_i - \mu)^2 + \sum_{i=1}^n (\bar{X} - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) \right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (X_i - \mu)^2 + n(\bar{X} - \mu)^2 - 2n(\bar{X} - \mu)^2 \right] \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \right] \end{aligned}$$

As such,

$$E(S^2) = \frac{1}{n-1} \left[\sum_{i=1}^n E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2] \right].$$

By the definition of variance, as $E[(X - \mu)^2] = \text{Var}(X)$, then it is clear that

$$\sum_{i=1}^n E[(X_i - \mu)^2] = n\sigma^2.$$

The term $E[(\bar{X} - \mu)^2]$ 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

$$\sum_{i=1}^n E[(\bar{X} - \mu)^2] = \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Putting everything together,

$$E(S^2) = \frac{1}{n-1} \left(n\sigma^2 - n \left(\frac{\sigma^2}{n} \right) \right) = \sigma^2.$$

□

Example 7.8. Let X_1, \dots, X_n be a sequence of independent and identically distributed random variables with mean μ and variance σ^2 . Consider the random variable

$$Z = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu).$$

We note that $E(X_i X_j) = E(X_i)E(X_j)$ for any $i \neq j$. Compute the mean and variance of Z .

Solution. We have

$$E(Z) = \frac{1}{\sigma\sqrt{n}} E\left(\sum_{i=1}^n (X_i - \mu)\right) = 0$$

by considering the linearity of expectation.

Next,

$$Z^2 = \frac{1}{n\sigma^2} \left[\sum_{i=1}^n (X_i - \mu) \right]^2.$$

Expanding the square yields

$$\left[\sum_{i=1}^n (X_i - \mu) \right]^2 = \sum_{i=1}^n (X_i - \mu)^2 + 2 \sum_{i < j} (X_i - \mu)(X_j - \mu).$$

Using the fact that $E(X_i X_j) = E(X_i)E(X_j)$ for distinct i and j , it follows that

$$E(Z^2) = \frac{1}{n\sigma^2} \cdot n\sigma^2 = 1$$

so $\text{Var}(Z) = 1$. This normalisation shows that Z is the standardised sum of independent and identically distributed random variables, which is the foundation for the central limit theorem (Theorem 8.6). It implies that Z converges in distribution to $N(0, 1)$ as n approaches infinity. \square

Example 7.9 (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 $1, \dots, n$. 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,

$$E(N) = 1 + \sum_{i=1}^{n-1} E(Y_i) = 1 + (n-1)E(Y_i) \quad \text{for any } 1 \leq i \leq n.$$

For two adjacent magnets i and $i+1$, they are broken if and only if they assume one of the following orientations:

$$\text{NS} | \text{SN} \quad \text{or} \quad \text{SN} | \text{NS}$$

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

$$\begin{aligned} \text{Var}(N) &= \text{Var}\left(\sum_{i=1}^{n-1} Y_i\right) \\ &= \sum_{i=1}^n \text{Var}(Y_i) + 2 \sum_{1 \leq i < j \leq n-1} \text{cov}(Y_i, Y_j) \\ &= \sum_{i=1}^n (E(Y_i^2) - (E(Y_i))^2) + 2 \sum_{1 \leq i < j \leq n-1} \text{cov}(Y_i, Y_j) \\ &= \sum_{i=1}^n (E(Y_i) - (E(Y_i))^2) + 2 \sum_{1 \leq i < j \leq n-1} \text{cov}(Y_i, Y_j) \\ &= \frac{n}{4} - 2 \sum_{1 \leq i < j \leq n-1} \text{cov}(Y_i, Y_j) \end{aligned}$$

Note that $\text{cov}(Y_i Y_j) = E(X_i X_j) - E(X_i)E(X_j)$. 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 $j+1$.

Since Y_i depends only on the orientations of magnets i and $i+1$, it follows that Y_i and Y_j are independent whenever $|i-j| > 1$. Hence, $\text{Cov}(Y_i, Y_j) = 0$ for $|i-j| > 1$. It remains to consider the case $j = i+1$. In this situation the random variables share the orientation of magnet $i+1$, so they are not independent, although, as we shall see, they are uncorrelated. We compute

$$\text{Cov}(Y_i, Y_{i+1}) = E(Y_i Y_{i+1}) - E(Y_i)E(Y_{i+1}).$$

Since $E(Y_i) = E(Y_{i+1}) = \frac{1}{2}$, we have $E(Y_i)E(Y_{i+1}) = \frac{1}{4}$. To compute $E(Y_i Y_{i+1})$, note that the triple of orientations $(i, i+1, i+2)$ has $2^3 = 8$ equally likely configurations. A

configuration contributes to $Y_i Y_{i+1} = 1$ exactly when both adjacent bonds are broken. A direct check shows that exactly 2 of the 8 possibilities break both bonds. Hence,

$$E(Y_i Y_{i+1}) = \frac{2}{8} = \frac{1}{4}.$$

Consequently,

$$\text{Cov}(Y_i, Y_{i+1}) = \frac{1}{4} - \frac{1}{4} = 0.$$

Thus all covariances vanish, and we obtain

$$\text{Var}(N) = \sum_{i=1}^{n-1} \text{Var}(Y_i) = (n-1) \cdot \frac{1}{4} = \frac{n-1}{4}.$$

□

Definition 7.3 (correlation). The correlation of random variables X and Y , denoted by $\rho(X, Y)$, is defined by

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

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).

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\left[\sqrt{\sum (x - \bar{x})^2} \right] \left[\sqrt{\sum (y - \bar{y})^2} \right]}$$

The two quantities ρ and r are of course equivalent. We can show that $-1 \leq \rho(X, Y) \leq 1$.

Proof. Note that $\text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$, $\text{Var}(X) = E[(X - \mu_X)^2]$ and $\text{Var}(Y) = E[(Y - \mu_Y)^2]$. 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

$$f(t) = E(U^2)t^2 + 2tE(UV) + E(V^2).$$

Note that $f(t) \geq 0$ since $\text{Var}(X) \geq 0$ 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 give a nice remark that the inequality $(E(UV))^2 \leq E(U^2)E(V^2)$ is the famous Cauchy-Schwarz inequality. \square

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.

Example 7.10. Consider the random variables X and Y with joint probability density function given by

$$f(x, y) = \begin{cases} \frac{1}{8}(x^2 - y^2)e^{-x} & \text{if } x > 0 \text{ and } -x < y < x; \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute their covariance and their correlation.

(b) Are X and Y independent?

Solution.

(a) By symmetry, $E(Y) = 0$ and $E(XY) = 0$. Since

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

then $\text{cov}(X, Y) = 0$. So, the correlation of X and Y is 0.

(b) Observe that $P(X > 1) > 0$ and $P(Y < 1) > 0$ but $P(X > 1 \text{ and } Y < 1) = 0$ so X and Y are not independent. \square

Example 7.11. Consider the random variables X and Y with joint probability density function given by

$$f(x, y) = \begin{cases} \frac{6}{7}\left(x^2 + \frac{xy}{2}\right) & \text{if } 0 < x < 1 \text{ and } 0 < y < 2; \\ 0, & \text{otherwise.} \end{cases}$$

(a) Compute their covariance.

(b) Compute their correlation.

Solution.

(a) We have

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y).$$

Note that

$$E(XY) = \int_0^2 \int_0^1 xy \cdot \frac{6}{7} \left(x^2 + \frac{xy}{2} \right) dx dy$$

and

$$E(X) = \int_0^1 x \cdot \frac{6}{7} \left(x^2 + \frac{xy}{2} \right) dx \quad \text{and} \quad E(Y) = \int_0^2 y \cdot \frac{6}{7} \left(x^2 + \frac{xy}{2} \right) dy.$$

Hence, $\text{cov}(X, Y) = -\frac{1}{147}$.

(b) The correlation coefficient is

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}} = -0.0561.$$

□

Example 7.12. A set of 12 dice are thrown at once. A first player takes all the dice whose value is 6, and the second player gets all dice with value ≤ 3 . Write X and Y respectively for the total value of their dice.

- (a)** Should we expect the covariance between X and Y to be positive or negative?
- (b)** Compute $\text{cov}(X, Y)$.

Solution.

- (a)** We would expect it to be negative.

- (b)** Let

$$X = \sum_{i=1}^{12} X_i \quad \text{and} \quad Y = \sum_{i=1}^{12} Y_i.$$

Here, each X_i and each Y_i are indicator random variables so $X_i = 6$ if the outcome of the i^{th} die is 6; Y_i is equal to the outcome of the i^{th} die if the outcome of the i^{th} die is ≤ 3 , otherwise the random variable takes on the value 0.

We have

$$\text{cov}(X, Y) = \text{cov} \left(\sum_{i=1}^{12} X_i, \sum_{i=1}^{12} Y_i \right) = \sum_{i=1}^{12} \text{cov}(X_i, Y_i).$$

We have

$$E(X_i) = 6P(X_i = 6) = 1 \quad \text{and} \quad E(Y_i) = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} = 1.$$

Next, we compute $E(X_i Y_i)$. If the outcome of the i^{th} die is 6, then $X_i = 6$ and $Y_i = 0$ so $X_i Y_i = 0$. On the other hand, if the outcome of the i^{th} die is ≤ 3 , then Y_i takes on that value but $X_i = 0$. So, $X_i Y_i = 0$. For other cases, $X_i Y_i = 0$ too. Hence, $\text{cov}(X, Y) = 12 \cdot (-1) = -12$. □

Example 7.13. Let U be a number between 2 and 9 chosen at random. Let

$$X = \left\lfloor \frac{U}{3} \right\rfloor \quad \text{and} \quad Y = \begin{cases} 1 & \text{if } 4 \leq U \leq 7; \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Identify the distribution of X and Y .
- (b) Compute $\text{cov}(X, Y)$.
- (c) Are X and Y independent?

Solution.

- (a) We have

$$f_U(u) = \begin{cases} \frac{1}{7} & \text{if } 2 \leq u \leq 9; \\ 0 & \text{otherwise.} \end{cases}$$

By casework, it is easy to see that X is a discrete random variable where $P(X = 0) = \frac{1}{7}$, $P(X = 1) = \frac{3}{7}$, and $P(X = 2) = \frac{3}{7}$. Next, $P(Y = 1) = \frac{3}{7}$ and $P(Y = 0) = \frac{4}{7}$ so Y is a Bernoulli random variable with parameter $\frac{3}{7}$.

- (b) We have

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y).$$

Note that $E(X) = \frac{9}{7}$ and $E(Y) = \frac{3}{7}$. It remains to compute $E(XY)$. Use the formula

$$E(XY) = \sum_{\text{all } x,y} xy \cdot P(X = x, Y = y).$$

When $Y = 1$, then $4 \leq U \leq 7$ so $X = 1$ or $X = 2$. So,

$$E(XY) = P(X = 1, Y = 1) + 2P(X = 2, Y = 1) = \frac{2}{7} + 2 \cdot \frac{1}{7}.$$

Substituting everything, we have $\text{cov}(X, Y) = \frac{1}{49}$.

- (c) If X and Y were independent, we would have

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

However, observe that $P(Y = 1 | X = 0) = 0$ but $P(Y = 1 | X = 2) > 0$ so Y depends on X . So, X and Y are dependent. \square

Example 7.14. In the game ‘‘president’’, the best card is the 2. Two players each pick 5 cards at random from the same deck (without replacement). Let X and Y be the number of 2’s each player has.

- (a) Should we expect the covariance between X and Y to be positive or negative?
- (b) Compute $\text{cov}(X, Y)$.

Solution.

- (a) We should expect the covariance to be negative because if one player gets more 2's, fewer are left for the other so the counts move in the opposite direction.
- (b) Let the four 2's be c_1, \dots, c_4 . For each j , define

$$I_j = \mathbf{1}\{c_j \text{ is in player one's hand}\} \quad \text{and} \quad J_j = \mathbf{1}\{c_j \text{ is in player two's hand}\}.$$

Then,

$$X = \sum_{j=1}^4 I_j \quad \text{and} \quad Y = \sum_{j=1}^4 J_j.$$

So,

$$E(I_j) = E(J_j) = \frac{5}{52} \quad \text{so} \quad E(X) = E(Y) = \frac{5 \cdot 4}{52} = \frac{5}{13}.$$

Also, for the same card j , $I_j J_j = 0$ because a card cannot be in both hands. For distinct cards (i.e. $j \neq k$), we have

$$P(I_j = 1, J_k = 1) = \frac{5}{52} \cdot \frac{5}{51}.$$

Hence,

$$E(XY) = \sum_{j \neq k} E(I_j J_k) = 4 \cdot 3 \cdot \frac{5}{52} \cdot \frac{5}{51} = \frac{25}{221}.$$

$$\text{So, } \text{cov}(X, Y) = -\frac{100}{2873}. \quad \square$$

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.

Proposition 7.3 (law of total expectation).

$$E(X) = E(E(X | Y))$$

Example 7.15 (Ross p. 390 Question 1). A player throws a fair die and simultaneously flips a fair coin. If the coin lands heads, then she wins twice, and if tails, then she wins one-half of the value that appears on the die. Determine her expected winnings.

Solution. Let the die outcome be D and the coin outcome be C . The winnings W depend on both — if C is heads, then $W = 2D$; if C is tails, then $W = \frac{1}{2}D$. So

$$\begin{aligned} E(W) &= E(W | C = \text{heads}) P(C = \text{heads}) + E(W | C = \text{tails}) P(C = \text{tails}) \\ &= E(2D) \cdot \frac{1}{2} + E\left(\frac{1}{2}D\right) \cdot \frac{1}{2} \end{aligned}$$

which is equal to $\frac{5}{4}E(D)$. Since $E(D) = 3.5$, then the result follows. \square

Example 7.16. 30 fair dice are thrown.

- (a) What is the expected number of 6's?
- (b) On average, how many *Yahtzees* (i.e. 5 dice with the same value) do we obtain?
- (c) Is the number of Yahtzees a binomial random variable?
- (d) Is the number of Yahtzees of 6's a binomial random variable?

Solution.

- (a) Let N denote the number of 6's. Then, $N \sim B(30, \frac{1}{6})$ so $E(N) = \frac{1}{6} \cdot 30 = 5$.
- (b) For each 5-subset of the 30 dice, the probability that all 5 show the same face is $\frac{6}{6^5} = \frac{1}{6^4}$. The expected number of Yahtzees is

$$\binom{30}{5} \cdot \frac{1}{6^4} = 109.96.$$

- (c) No. If you index each 5-subset with an indicator for ‘this 5-subset is a Yahtzee’, these indicators are not independent as overlapping subsets share dice. So the total is not binomial.
- (d) Again, the answer is no. Whether two different 5-subsets are all 6's is again dependent when the subsets overlap. Equivalently, this count equals to $\binom{X_6}{5}$, where $X_6 \sim B(30, \frac{1}{6})$ which does not follow a binomial distribution. \square

Example 7.17. Consider the pair of integer-valued random variables (N, K) with joint distribution given by

$$P(N = n, K = k) = \begin{cases} \frac{(e-1)e^{-2n}n^k}{k!} & \text{if } n \geq 1 \text{ and } k \geq 0; \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Compute $E(N | K = 0)$.
(b) What is the distribution of $E(K | N)$?
(c) Compute $E(NK)$.

Solution.

- (a) We have

$$P(K = k) = \sum_{n=1}^{\infty} \frac{(e-1)e^{-2n}n^k}{k!} = \frac{e-1}{k!} \sum_{n=1}^{\infty} \frac{n^k}{e^{2n}}.$$

So,

$$P(K = 0) = (e-1) \sum_{n=1}^{\infty} \frac{1}{e^{2n}} = \frac{1}{e+1}.$$

As such,

$$E(N | K = 0) = \sum_{n=1}^{\infty} n \cdot \frac{P(N = n \text{ and } K = 0)}{P(K = 0)}$$

which is equal to

$$(e^2 - 1) \sum_{n=1}^{\infty} \frac{n}{e^{2n}} = \frac{1}{1 - e^{-2}}.$$

- (b) We have

$$P(N = n) = \sum_{k=0}^{\infty} \frac{(e-1)e^{-2n}n^k}{k!} = \frac{e-1}{e^n}.$$

So,

$$P(K = k | N = n) = \frac{(e-1)e^{-2n}n^k}{k!} \cdot \frac{e^n}{e-1} = \frac{e^{-n}n^k}{k!}.$$

As such, $K | N = n \sim \text{Po}(n)$. As such, $E(K | N) = N$.

- (c) By the law of total expectation (Proposition 7.3),

$$E(NK) = E(E(NK | N)) = E(NE(K | N)) = E(N^2).$$

Recall that

$$P(N = n) = \frac{e-1}{e^n}$$

so

$$E(N^2) = (e-1) \sum_{n=1}^{\infty} \frac{n^2}{e^n} = \frac{e(e+1)}{(e-1)^2}.$$

□

Example 7.18. Consider the pair of integer-valued random variables (X, Y) with joint distribution given by

$$P(X = k, Y = l) = \begin{cases} \frac{1}{e^2} \cdot \frac{1}{l!(k-l)!} & \text{if } 0 \leq l \leq k; \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Compute $E(X | Y = 0)$.
(b) What is the distribution of $2E(Y | X)$?

(c) Compute $E(XY)$.

Solution.

(a) We have

$$P(Y = 0) = \sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{1}{e^2} \cdot \frac{1}{k!} = \frac{1}{e}.$$

Hence,

$$E(X | Y = 0) = \sum_{x=0}^{\infty} x \cdot P(X = x | Y = 0) = \sum_{x=0}^{\infty} x \cdot \frac{P(X = x \text{ and } Y = 0)}{P(Y = 0)}.$$

Substituting the relevant probabilities, we have

$$E(X | Y = 0) = \sum_{k=0}^{\infty} k \cdot \frac{\frac{1}{e^2} \cdot \frac{1}{k!}}{\frac{1}{e}} = 1.$$

(b) We have

$$E(Y | X) = \sum_y y \cdot P(Y = y | X = x).$$

Hence,

$$E(Y | X) = \sum_{l=0}^k l \cdot \frac{\frac{1}{e^2} \cdot \frac{1}{l!(k-l)!}}{P(X = k)}.$$

Note that

$$P(X = k) = \sum_{l=0}^k \frac{1}{e^2} \cdot \frac{1}{l!(k-l)!} = \frac{1}{e^2} \cdot \frac{2^k}{k!}.$$

As such,

$$E(Y | X = k) = \frac{k}{2} \quad \text{so} \quad 2E(Y | X) = X.$$

Since

$$P(X = k) = \frac{e^{-2} \cdot 2^k}{k!},$$

then $X \sim \text{Po}(2)$ so $2E(Y | X)$ follows a Poisson distribution with parameter 2.

(c) The trick is to observe that if $Y = U$ and $X = U + V$, then

$$P(U = l, X = k - l) = \frac{e^{-1}}{l!} \cdot \frac{e^{-1}}{(k-l)!}.$$

Hence, $U \sim \text{Po}(1)$ and $V \sim \text{Po}(1)$. So,

$$E(XY) = E(U^2) + E(UV) = \text{Var}(U) + [E(U)]^2 + E(U)E(V)$$

which is equal to $1 + 1 + 1 \cdot 1 = 3$. \square

Example 7.19. Consider the pair of discrete random variables (N, K) with joint probability mass function

$$P(N = n, K = k) = \begin{cases} \frac{c}{\binom{n+2}{3}} & \text{if } 1 \leq k \leq n; \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Find the value of c .
- (b) Compute $E(N | K = k)$.
- (c) Compute $E(2^K | N = n)$.
- (d) Compute $P(E(K | N = 3) \geq 5)$ and $P(E(K | N) \geq 5)$.

Solution.

- (a) We have

$$\sum_{n=1}^{\infty} \sum_{k=1}^n \frac{c}{\binom{n+2}{3}} = 1.$$

So, one can deduce that $c = \frac{1}{3}$.

- (b) We have

$$P(N = n | K = k) = \frac{P(N = n \text{ and } K = k)}{P(K = k)}.$$

Note that

$$P(K = k) = \sum_{n=k}^{\infty} \frac{1/3}{\binom{n+2}{3}} = \frac{3}{k(k+1)}.$$

Hence,

$$E(N | K = k) = \sum_{n=k}^{\infty} n \cdot P(N = n | K = k) = 2k.$$

- (c) We have

$$E(2^K | N = n) = \sum_{k=1}^n 2^k \cdot P(K = k | N = n).$$

Note that

$$P(N = n) = \sum_{k=1}^n \frac{1/3}{\binom{n+2}{3}} = \frac{2}{(n+2)(n+1)}.$$

Hence,

$$E(2^K | N = n) = \frac{1}{n} \sum_{k=1}^n 2^k = \frac{2(2^n - 1)}{n}.$$

- (d) For the first probability, we have

$$E(K | N = n) = \frac{n+1}{2}.$$

So, the first probability is equal to 0. One can then work out that the second probability is equal to $P(N \geq 9) = 0.2$. \square

Example 7.20. Consider the pair of discrete random variables (N, K) with joint probability density function

$$P(N = n, K = k) = \begin{cases} \frac{c}{n} & \text{if } 1 \leq k \leq n \leq 10; \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Find the value of c .
- (b) Compute $E(N | K = k)$.

- (c) Compute $E(K^2 | N = n)$.
 (d) Compute $P(E(K^2 | N = 5) \leq 10)$ and $P(E(K^2 | N) \leq 10)$.

Solution.

- (a) We have

$$\sum_{n,k} \frac{c}{n} = 1 \quad \text{so} \quad \sum_{n=1}^{10} \sum_{k=1}^n \frac{c}{n} = 1.$$

Hence, $c = 0.1$.

- (b) We have

$$P(N = n | K = k) = \frac{P(N = n \text{ and } K = k)}{P(K = k)}.$$

We have

$$P(K = k) = \sum_{n=k}^{10} \frac{c}{n} \quad \text{so} \quad P(N = n | K = k) = \frac{1}{11 - k}.$$

Hence,

$$E(N | K) = \sum_{n=k}^{10} n \cdot P(N = n | K = k) = \sum_{n=k}^{10} \frac{n}{11 - k} = \frac{110 - k^2 + k}{2(11 - k)}.$$

- (c) We have $P(N = n) = c$ so $P(K = k | N = n) = \frac{1}{n}$. Hence,

$$E(K^2 | N) = \sum_{k=1}^n k^2 \cdot P(K = k | N = n) = \sum_{k=1}^n \frac{k^2}{n} = \frac{(n+1)(2n+1)}{6}.$$

- (d) We have

$$E(K^2 | N = 5) = 11 > 10$$

so the first probability is equal to 0. For the second probability, since the support of N takes on integers, then

$$\begin{aligned} P(E(K^2 | N) \leq 10) &= P(E(K^2 | N) < 11) \\ &= P(2N^2 + 3N - 65 < 0) \\ &= P\left(-\frac{13}{2} < N < 5\right) \end{aligned}$$

Since $P(N = n) = \frac{1}{10}$, the second probability is equal to 0.4. \square

Example 7.21. Consider the pair of continuous random variables (X, Y) with joint distribution given by

$$f(x, y) = \begin{cases} c(2x+y) & \text{if } 0 < x < y < 1; \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Find the value of c .

- (b) Compute $E(X | Y = y)$. For which value of y is this properly defined?

(c) Compute

$$P\left(E(Y|X) \leq \frac{1}{2}\right).$$

Solution.

(a) We have

$$\int_0^1 \int_0^y c(2x+y) dy dx = 1$$

$$\text{so } c = 1.5.$$

(b) We have

$$E(X|Y=y) = \int_0^y x \cdot \frac{f(x,y)}{f_Y(y)} dx = \frac{7}{12}y.$$

(c) We have

$$f_X(x) = \int_x^1 \frac{3}{2}(2x+y) dy = \frac{3}{4}(1-5x)(1+x).$$

So,

$$E(Y|X=x) = \int_x^1 \frac{y \cdot \frac{3}{2}(2x+y)}{\frac{3}{4}(1-5x)(1+x)} dx.$$

The rest of the working is trivial! □

Example 7.22. Consider the pair of continuous random variables (X, Y) with joint distribution given by

$$f(x,y) = \begin{cases} c(2x+y)e^{-y} & \text{if } 0 < x < y; \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the value of c .

(b) Compute $E(X|Y=y)$ for $y > 0$.

Solution.

(a) We have

$$\int_0^\infty \int_0^y c(2x+y)e^{-y} dx dy = 1$$

$$\text{so } c = \frac{1}{4}.$$

(b) We have

$$E(X|Y=y) = \int_0^y x \cdot \frac{f_{X,Y}(x,y)}{f_Y(y)} dy.$$

Note that

$$f_Y(y) = \int_0^y \frac{1}{4}(2x+y)e^{-y} dx = \frac{1}{2}y^2e^{-y}$$

Substituting everything yields

$$E(X|Y=y) = \int_0^y x \cdot \frac{\frac{1}{4}(2x+y)e^{-y}}{\frac{1}{2}y^2e^{-y}} dx = \frac{7}{12}y.$$

□

Example 7.23. Consider the pair of continuous random variables (X, Y) with joint distribution

$$f(x,y) = \begin{cases} c(2x+y^2)e^{-2y} & \text{if } 0 < x < y; \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Find the value of c .
- (b) Compute $E(X | Y)$.
- (c) Compute $E(X(1+Y))$.

Solution.

- (a) We have

$$c \int_0^\infty \int_0^y (2x+y^2)e^{-2y} dx dy = 1$$

$$\text{so } c = 1.6.$$

- (b) We have

$$f_{X|Y}(x | y) = \frac{2x+y^2}{y^2+y^3}.$$

Hence,

$$E(X | Y) = \int_0^y x \cdot \frac{2x+y^2}{y^2+y^3} dx = \frac{\frac{2}{3}Y + \frac{1}{2}Y^2}{1+Y}.$$

- (c) We have

$$E(X(1+Y)) = E(E(X(1+Y) | Y)) = E(1+Y)E(X | Y)$$

which is equal to

$$E\left(\frac{2}{3}Y + \frac{1}{2}Y^2\right) = 3.3.$$

□

Example 7.24. Consider the pair of continuous random variables (X, Y) with joint density function

$$f(x,y) = \begin{cases} c(x+y)^2 & \text{if } -1 < x < 1 \text{ and } -1 < y < 1; \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Find the value of c .
- (b) Compute $E(X | Y)$. What is $E(Y | X)$?
- (c) Compute $E(XY^2)$.

Solution.

- (a) We have

$$\int_{-1}^1 \int_{-1}^1 c(x+y)^2 dx dy = 1$$

$$\text{so } c = \frac{3}{8}.$$

(b) We have

$$f_Y(y) = \int_{-1}^1 \frac{3}{8} (x+y)^2 dx = \frac{3}{8} \left(\frac{2}{3} + 2y^2 \right).$$

So,

$$E(X | Y) = \frac{\int_{-1}^1 \int_{-1}^1 x f_{X,Y}(x,y) dx dy}{f_Y(y)} = \frac{2y}{1+3y^2}.$$

By symmetry, $E(Y | X) = \frac{2x}{1+3x^2}$.

(c) We have

$$E(XY^2) = \int_{-1}^1 \int_{-1}^1 xy^2 \cdot \frac{3}{8} (x+y)^2 dx dy = 0.$$

□

Example 7.25. When playing *Pokémon Red or Blue*, capturing a Chansey is one of the ultimate challenges. Indeed, they only appear with 4% chance, and for each turn, get captured with 5% chance and otherwise flee with 25% chance.

- (a)** If you encountered n Pokemon, on average how many Chansey did you capture?
- (b)** Pokémons actually spawn randomly following a Poisson distribution, with an average of 2 Pokémons per minute. On average, how many Chansey did you capture after playing for one hour?

Solution.

- (a)** Let X be the number of Chansey captured. Then, $X \sim B(n, p)$, where p is to be found. The probability that a Chansey is eventually captured is a geometric sum. That is,

$$\sum_{k=1}^{\infty} 0.70^{k-1} \cdot 0.05 = \frac{1}{6}.$$

The probability p that a random encounter results in a captured Chansey is

$$P(\text{encounter Chansey}) \cdot P(\text{eventually capture} | \text{encounter Chansey})$$

which is $0.04 \cdot \frac{1}{6} = \frac{1}{150}$. Hence, $p = \frac{1}{150}$. As such, $E(X) = \frac{n}{150}$.

- (b)** Let Y denote the number of Pokémons spawning per hour. Then, $Y \sim \text{Po}(120)$. By considering our notation in **(a)**, we have

$$X | Y \sim B\left(Y, \frac{1}{150}\right) \quad \text{and} \quad Y \sim \text{Po}(120).$$

We need to find $E(X)$. By the law of total expectation (Proposition 7.3),

$$E(X) = E(E(X | Y)) = E\left(\frac{Y}{150}\right) = \frac{1}{150} \cdot 120 = 0.8.$$

□

Example 7.26. When playing Pokémon Red or Blue, our first opportunity to capture a Pikachu is in the Viridian forest, where they appear with 5% chances, and are captured throwing a Pokéball with probability 0.1865.

- (a) If we encountered 10 Pokémons, what is the probability that none of them were Pikachu?
- (b) Assuming that whenever we encounter a Pikachu we throw Pokéballs until capture. After encountering 10 Pokémons, how many Pokéballs did we throw on average?

Solution.

- (a) Let N denote the number of Pikachu encountered. Then, $N \sim B(10, 0.05)$ so $P(N = 0) = 0.95^{10}$.
- (b) For every Pikachu encountered, let B_i denote the number of Pokéballs used to capture it. Then, define $B = B_1 + \dots + B_N$. We wish to find $E(B)$. We have

$$\begin{aligned} E(B) &= E(B_1 + \dots + B_N) \\ &= E(E(B_1 + \dots + B_N | N)) \\ &= E(E(B_1 | N)) + \dots + E(E(B_N | N)) \\ &= E(B_i)E(N) \end{aligned}$$

Note that $N \sim B(10, 0.05)$ and $B_i \sim \text{Geo}(0.1865)$ so $E(B) = 2.68$. □

Example 7.27. In the Texas hold'em variant of poker, players start by receiving 2 cards and have to place an initial bet accordingly. A first player has a strategy where they add the total T of their two cards (seeing J, Q, K as 11, 12, 13) and places a (continuous) uniform bet X between 0 and T . The second player uses a similar strategy and places a (continuous) uniform bet Y between 0 and $30N$ where N is the number of faced values (J, Q , or K) they have in their hands.

- (a) Which player places the highest average bet?
- (b) What is the covariance of the 2 bets? Do we expect it to be positive or negative?

Solution.

- (a) Let the first player's two-card total be T . Then, $X \sim U(0, T)$. Next, let the second player's number of face cards be N . Then, $Y \sim U(0, 30N)$. Hence,

$$E(X | T) = \frac{T}{2} \quad \text{and} \quad E(Y | N) = 15N.$$

By the law of total expectation (Proposition 7.3),

$$E(X) = E(E(X | T)) = E\left(\frac{T}{2}\right) = \frac{1}{2}E(T)$$

and

$$E(Y) = 15E(N).$$

Note that

$$E(T) = \frac{1+2+\dots+13}{13} = 14.$$

Also, $E(N) = 2 \cdot \frac{3}{13}$ so $E(X) > E(Y)$.

(b) Recall from **(a)** we had $E(X) = 7$ and $E(Y) = \frac{90}{13}$. We have

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y).$$

We have

$$\text{cov}(X, Y) = \text{cov}(E(X | T), E(Y | N)) = \text{cov}\left(\frac{T}{2}, 15N\right) = \frac{15}{2} \text{cov}(T, N).$$

We shall compute $\text{cov}(T, N)$ using indicator random variables. We index the cards $i = 1, \dots, 52$ with values v_i and let A_i be the indicator that card i is in player 1's hand and B_i for player 2's hand; let F be the 12 face cards. Then,

$$T = \sum_i v_i A_i \quad \text{and} \quad N = \sum_{j \in F} B_j.$$

Hence,

$$\text{cov}(T, N) = \sum_i \sum_{j \in F} (E(A_i B_j) - E(A_i)E(B_j)).$$

Uniformly dealing two disjoint 2-card hands gives $E(A_i) = \frac{1}{26}$ and $E(B_j) = \frac{1}{26}$.

Also,

$$E(A_i B_j) = \begin{cases} 0 & \text{if } i = j; \\ \frac{1}{650} & \text{if } i \neq j. \end{cases}$$

Hence for $i \neq j$, the difference is $\frac{1}{650} - \frac{1}{676} = \frac{1}{16900}$, while for $i = j$, it is $-\frac{1}{676}$. Using the fact that

$$\sum_{i \in F} v_i = 4(11 + 12 + 13) = 144 \quad \text{and} \quad \sum_{i \notin F} v_i = 4(1 + \dots + 10) = 220,$$

then one can deduce that $\text{cov}(T, N) = \frac{12}{325}$ so $\text{cov}(X, Y) = \frac{18}{65}$. \square

Example 7.28 (stopping-time expectation problem). You play a simplified slot machine where, every time you place your bet, you receive an integer number uniform between 0 and 9 (we are not interested in the reward here). You then opt for the following strategy: start by betting \$1 for the first game, then, for each subsequent game, bet the value displayed by the machine. Thus, you stop playing whenever you get a 0.

(a) On average, how long do you play this game?

(b) On average, how much do you bet?

Solution. Let X_1, X_2, \dots be independent and identically distributed uniform random variables on $0, 1, \dots, 9$, where X_i is the value shown after the i^{th} game. Also, we define $T = \min\{t \geq 1 : X_t = 0\}$.

- (a) Since $P(X_i = 0) = 0.1$, then the stopping time T follows a geometric distribution with parameter 0.1. Then, $E(T) = 10$.
- (b) Let B_k denote the amount we bet on the k^{th} game. Then, for $k \geq 2$, we have

$$B_1 = 1 \quad \text{and} \quad B_k = \begin{cases} X_{k-1} & \text{if } T \geq k; \\ 0 & \text{otherwise.} \end{cases}$$

The condition $T \geq k$ means that $X_1, \dots, X_{k-1} \neq 0$. Define

$$S = \sum_{k=1}^{\infty} B_k = 1 + \sum_{k=2}^{\infty} B_k.$$

We wish to compute $E(S)$. Note that for $k \geq 2$, we have

$$E(B_k) = E(X_{k-1} \mathbf{1}_{\{T \geq k\}}) = P(T \geq k) E(X_{k-1} \mid T \geq k)$$

which is equal to

$$(0.9)^{k-1} \cdot \frac{1+2+\dots+9}{9} = 5(0.9)^{k-1}.$$

Hence, the answer is 46. \square

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 7.29. A standard dice is thrown before flipping as many fair coins as given by the dice value. We then count the number of heads counted by the coins.

- (a) What is the expected number of heads?
- (b) What is the variance of the number of heads?

Solution.

- (a) Let X be the random variable denoting the number on the die and Y denoting the number of heads. Then, X is a discrete uniform random variable on $\{1, \dots, 6\}$ and $Y \sim B(X, \frac{1}{2})$. So, $E(Y | X) = \frac{1}{2}X$. Hence, by the law of total expectation (Proposition 7.3),

$$E(Y) = E(E(Y | X)) = E\left(\frac{1}{2}X\right) = \frac{7}{4}.$$

(b) We use the law of total variance (Proposition 7.4), so

$$\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X)).$$

We have

$$\text{Var}(Y|X) = X \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}X.$$

Hence,

$$\text{Var}(Y) = E\left(\frac{1}{4}X\right) + \text{Var}\left(\frac{1}{2}X\right) = \frac{77}{48}.$$

□

Example 7.30. A game is played where we flip a coin until it lands on head, and for any time we flip a tail, we get to throw a standard die. We then add all the values given by the different dice (this total value possibly being 0 if we flipped a head first).

- (a)** What is the expected total value of the dice?
- (b)** What is the variance of the total value of the dice?

Solution.

- (a)** Let N denote the total number of tails before the first head. Then, $P(N = n) = \frac{1}{2^{n+1}}$. For each tail, we roll a standard die and its outcome is X_i , where X_i follows a discrete uniform random variable on $\{1, \dots, 6\}$. The total score is

$$S = \sum_{i=1}^N X_i.$$

Then, by the law of total expectation (Proposition 7.3),

$$E(S) = E(E(S|N)) = n \cdot E(X_i) = 3.5n.$$

So, $E(S) = 3.5E(N) = 3.5$.

- (b)** By the law of total variance (Proposition 7.4),

$$\text{Var}(S) = E(\text{Var}(S|N)) + \text{Var}(E(S|N=n)).$$

Recall that

$$E(S|N) = 3.5n \quad \text{and} \quad \text{Var}(S|N) = n \text{Var}(X_i) = \frac{35}{12}n.$$

Hence,

$$\text{Var}(S) = E\left(\frac{35}{12}N\right) + \text{Var}(3.5N) = \frac{35}{12}E(N) + 3.5^2 \text{Var}(N)$$

which is equal to $\frac{329}{12}$.

□

Example 7.31 (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

$$X := \sum_{i=1}^N \xi_i = \xi_1 + \xi_2 + \dots + \xi_N.$$

Find the mean and variance of X .

Solution. By the law of total expectation (Proposition 7.3),

$$\begin{aligned} E(X) &= \sum_{i=1}^n E(X|N=i)P(N=i) \\ &= \sum_{i=1}^n E(X|N=i) \cdot \frac{e^{-1}}{i!} \\ &= E(\xi_1) \cdot \frac{e^{-1}}{1!} + E(\xi_1 + \xi_2) \cdot \frac{e^{-1}}{2!} + \dots + E(\xi_1 + \xi_2 + \dots + \xi_n) \cdot \frac{e^{-1}}{n!} \\ &= 0 \cdot \frac{e^{-1}}{1!} + 0 \cdot \frac{e^{-1}}{2!} + \dots + 0 \cdot \frac{e^{-1}}{n!} \\ &= 0 \end{aligned}$$

For the variance, by the law of total variance (Proposition 7.4),

$$\text{Var}(X) = E(\text{Var}(X|N)) + \text{Var}(E(X|N)) = E(N\text{Var}(X)) = E(N) = 1.$$

□

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

$$M_X(t) = E(e^{tX}). \quad (7.2)$$

If X is a discrete random variable with PDF p_X , then

$$M_X(t) = \sum_{\text{all } x} e^{tx} p_X(x). \quad (7.3)$$

If X is a continuous random variable with PDF f_X , then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx. \quad (7.4)$$

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 . \square

Example 7.32. For example,

$$M(t) = \frac{e^{3t}}{3 - e^t}$$

is not a proper moment generating function because $M_2(0) \neq 1$.

Example 7.33. For example,

$$M(t) = \frac{2e^{3t}(2 - e^t)}{3 - e^t}$$

is not a proper moment generating function because $M'_3(0) = 2.5$ and $M''_3(0) = 5$ so the variance of the distribution is negative.

Example 7.34. For example,

$$M(t) = \frac{e^{2t}(3 - e^t)^2}{5 - e^{4t}}$$

is not a proper moment generating function. To see why, although $M(0) = 1$, but we see that the denominator $5 - e^{4t}$ tends to 0 as t tends to $\frac{1}{4} \ln 5$, so for $t > \frac{1}{5} \ln 4$, the denominator becomes negative, giving $M(t) < 0$, contradicting (7.2) in relation to the exponential function.

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, X and Y have the same distribution, i.e. $f_X = f_Y$.

Example 7.35. Let N be the discrete random variable with probability mass function

$$P(N = n) = \frac{1}{n(n+1)} \quad \text{for all } n \geq 1.$$

- (a) Find the moment generating function of N .
- (b) Deduce the value of $E(N)$.

Solution.

- (a) We have

$$M_N(t) = \sum_{n=1}^{\infty} e^{tn} \cdot \frac{1}{n(n+1)} = \sum_{n=1}^{\infty} \frac{e^{tn}}{n(n+1)} = \sum_{n=1}^{\infty} e^{tn} \left(\frac{1}{n} - \frac{1}{n+1} \right).$$

By considering series expansion, observe that

$$\sum_{n=1}^{\infty} \frac{e^{tn}}{n} = -\ln(1 - e^t)$$

so

$$M_N(t) = 1 - (1 - e^{-t}) \ln(1 - e^{-t}).$$

- (b) $E(N) = M'(0)$ so one can show that this goes to infinity. In fact, this is not surprising since $E(N)$ is just a translated sum of the harmonic series so it will diverge. \square

Example 7.36. Let N be the discrete random variable with probability mass function

$$P(N = n) = c \left(\frac{1}{3^n} + \frac{1}{n!} \right) \quad \text{for } n \geq 1.$$

- (a) Find the value of c .
- (b) Find the moment generating function of N .
- (c) Compute $E(N)$.

Solution.

- (a) We have

$$\sum_{n=1}^{\infty} c \left(\frac{1}{3^n} + \frac{1}{n!} \right) = 1$$

$$\text{so } c = \frac{2}{2e-1}.$$

- (b) We have

$$M_N(t) = \sum_{n=1}^{\infty} e^{nt} \cdot \frac{2}{2e-1} \left(\frac{1}{3^n} + \frac{1}{n!} \right)$$

which is equal to

$$\frac{2}{2e-1} \sum_{n=1}^{\infty} \left(\frac{e^t}{3} \right)^n + \frac{2}{2e-1} \sum_{n=1}^{\infty} \frac{e^{nt}}{n!}.$$

Note that

$$\sum_{n=1}^{\infty} \left(\frac{e^t}{3} \right)^n = \frac{e^t}{3 - e^t} \quad \text{which holds for } t < \ln 3$$

and

$$\sum_{n=1}^{\infty} \frac{e^{nt}}{n!} = \exp(e^t) - 1$$

so the moment generating function is equal to

$$M_N(t) = \frac{2}{2e-1} \left[\frac{e^t}{3-e^t} + \exp(e^t) - 1 \right].$$

(c) One can show that

$$E(N) = M'_N(0) = \frac{2e + \frac{3}{2}}{2e-1}.$$

□

Example 7.37. We say that $n \in \mathbb{Z}$ is congruent to 2 mod 4 if and only if there exists $k \in \mathbb{Z}$ such that $n - 2 = 4k$. Let N be the discrete random variable with distribution

$$P(N=n) = \begin{cases} \frac{c}{2^{\frac{n}{2}}} & \text{if } n \equiv 2 \pmod{4}; \\ 0 & \text{otherwise} \end{cases}.$$

- (a) Compute the moment generating function of N .
- (b) What should the value of c be?
- (c) Compute $E(N)$.
- (d) Compute $\text{Var}(N)$.

Solution.

- (a) We have

$$M_N(t) = \sum_{n \equiv 2 \pmod{4}} e^{tn} \cdot \frac{c}{2^{\frac{n}{2}}}$$

which is equal to

$$c \left(\frac{e^{2t}}{2^{\frac{2}{2}}} + \frac{e^{6t}}{2^{\frac{6}{2}}} + \frac{e^{10t}}{2^{\frac{10}{2}}} + \dots \right) = \frac{ce^{2t}}{2^{\frac{2}{2}}} \left(1 + \frac{e^{4t}}{2^2} + \frac{e^{8t}}{2^4} + \dots \right) = \frac{2ce^{2t}}{4 - e^{4t}}.$$

- (b) We have

$$c \left(\frac{1}{2^{\frac{2}{2}}} + \frac{1}{2^{\frac{6}{2}}} + \frac{1}{2^{\frac{10}{2}}} + \dots \right) = 1$$

so $c = 1.5$.

- (c) We have $M'_N(0) = \frac{10}{3}$.

- (d) We have $E(N^2) = M''_N(0)$ so $\text{Var}(N) = \frac{64}{9}$. □

Example 7.38. Let X be the continuous random variable with probability density function

$$f(x) = \begin{cases} 6x(1-x) & \text{if } 0 < x < 1; \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Find the moment generating function of X .
 (b) Deduce the value of $E(X)$.

Solution.

- (a) We have

$$M_X(t) = \int_0^1 e^{tx} \cdot 6x(1-x) dx = \frac{6}{t^3} [e^t(t-2) + t + 2].$$

- (b) We have

$$E(X) = M'_X(0) = 0.5.$$

□

Example 7.39. Let X be the continuous random variable with probability density function

$$f(x) = \begin{cases} c(2+x)e^{-x} & \text{if } x > 0; \\ 0 & \text{otherwise.} \end{cases}$$

Note that

$$\int_0^\infty c(2+x)e^{-x} dx = 1 \quad \text{so} \quad c = \frac{1}{3}.$$

- (a) Find the moment generating function of X .
 (b) Compute $E(X)$.

Solution.

- (a) We have

$$M_X(t) = \int_0^\infty e^{tx} f(x) dx = \frac{1}{3} \int_0^\infty (2+x)e^{tx} dx = \frac{3-2t}{3(1-t)^2}$$

which holds for $t < 1$.

- (b) We have

$$M'_X(t) = \frac{2}{3}(1-t)^{-3}(2-t)$$

$$\text{so } E(X) = M'_X(0) = \frac{4}{3}.$$

□

Example 7.40. Let X be the random variable with probability density function

$$f(x) = \begin{cases} \sum_{k=1}^4 ck^3 e^{-kx} & \text{if } x > 0; \\ 0 & \text{otherwise.} \end{cases}$$

- (a) Compute the moment generating function of X .
 (b) What should the value of c be?
 (c) Compute $E(X)$.
 (d) Compute $\text{Var}(X)$.

Solution.

(a) The moment generating function is

$$M_X(t) = \int_0^\infty e^{tx} \sum_{k=1}^4 ck^3 e^{-kx} dx = c \sum_{k=1}^4 k^3 \int_0^\infty e^{(t-k)x} dx$$

which is equal to

$$c \left(\frac{1}{1-t} + \frac{8}{2-t} + \frac{27}{3-t} + \frac{64}{4-t} \right).$$

(b) Since $M_X(0) = 1$, then $c = \frac{1}{30}$.

(c) Note that

$$M_X(t) = \frac{1}{30} \left(\frac{1}{1-t} + \frac{8}{2-t} + \frac{27}{3-t} + \frac{64}{4-t} \right)$$

We have $E(X) = M'_X(0) = \frac{1}{3}$.

(d) Note that $E(X^2) = M''_X(0) = \frac{4}{15}$ so using the formula $\text{Var}(X) = E(X^2) - (E(X))^2$, we have $\text{Var}(X) = \frac{7}{45}$. \square

We now 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. By (7.3), $M(t) = e^{t(0)}P(X=0) + e^{t(1)}P(X=1) = (1-p) + pe^t$. \square

Proposition 7.7 (MGF of binomial random variable). If $X \sim \text{B}(n, p)$, then

$$M(t) = (1 - p + pe^t)^n.$$

Proof. By (7.3) and writing it in sigma notation,

$$\sum_{k=0}^n e^{kt} \binom{n}{k} p^k q^{n-k} = q^n \sum_{k=0}^n \binom{n}{k} \left(\frac{pe^t}{q} \right)^k = q^n \left(1 + \frac{pe^t}{q} \right)^n = (1 - p + pe^t)^n.$$

\square

Proposition 7.8 (MGF of geometric random variable). If $X \sim \text{Geo}(p)$, then

$$M(t) = \frac{pe^t}{1 - qe^t}.$$

Proof. By (7.3),

$$M(t) = \sum_{k=1}^n e^{kt} pq^{k-1} = \frac{p}{q} \sum_{k=1}^n (qe^t)^k = \frac{pe^t}{1 - qe^t}.$$

\square

Example 7.41. If $X \sim \text{Geo}(p)$, find an expression for $E(X^3)$.

Solution. The third moment, or $E(X^3)$, is the coefficient of t^3 divided by $3! = 6$ in the series expansion of $M(t)$. \square

Proposition 7.9 (MGF of Poisson random variable). If $X \sim \text{Po}(\lambda)$, then

$$M(t) = \exp(\lambda(e^t - 1)).$$

Proof. By (7.3),

$$M(t) = \sum_{k=0}^n \frac{e^{kt} e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^n \frac{(\lambda e^t)^k}{k!} = e^{-\lambda} (e^{\lambda e^t}) = \exp(\lambda(e^t - 1))$$

□

Example 7.42. Let P_1, P_2, \dots be a sequence of independent Poisson random variables with parameters $\lambda_1, \lambda_2, \dots$. Let $X_n = P_1 + \dots + P_n$.

- (a) Find the moment generating function of X_1 .
- (b) Find the moment generating function of X_2 .
- (c) Find the moment generating function of X_n for any $n \geq 1$.
- (d) What is the distribution of X_n ?

Solution.

- (a) We have $X_1 \sim \text{Po}(\lambda_i)$ so

$$P(X_1 = k) = \frac{e^{-\lambda_1} \lambda_1^k}{k!}.$$

So (alternatively, we can use Proposition 7.9 directly),

$$M_{X_1}(t) = \sum_{k=0}^{\infty} e^{tk} P(X_1 = k) = e^{-\lambda_1} \sum_{k=0}^{\infty} \frac{e^{tk} \lambda_1^k}{k!} = \exp(\lambda_1(e^t - 1)).$$

- (b) By the additivity property of the Poisson random variable (Proposition 4.16), $X_2 \sim \text{Po}(\lambda_1 + \lambda_2)$ so

$$P(X_2 = k) = \frac{e^{-(\lambda_1 + \lambda_2)} (\lambda_1 + \lambda_2)^k}{k!}.$$

So,

$$M_{X_2}(t) = \sum_{k=0}^{\infty} e^{tk} P(X_2 = k) = \exp((\lambda_1 + \lambda_2)(e^t - 1)).$$

- (c) In a similar fashion,

$$M_{X_n}(t) = \exp((\lambda_1 + \lambda_2 + \dots + \lambda_n)(e^t - 1)).$$

- (d) From (c), we see that X_n represents a Poisson random variable with parameter $\lambda_1 + \dots + \lambda_n$, where we used the multiplicativity of MGFs (Proposition 7.5). □

Example 7.43. Let N be a Poisson random variable with parameter λ and X be a binomial random variable with parameters N and p .

- (a) Compute the moment generating function of X .

- (b) Identify the distribution of X .
- (c) What is the distribution of $N - X$?
- (d) Are X and $N - X$ independent?

Solution.

- (a) We have $N \sim \text{Po}(\lambda)$. By the law of total expectation (Proposition 7.3), we have $E(X) = E(E(X | N))$. Note that

$$M_X(t) = E(e^{tX}) = E(E(e^{tX} | N)).$$

We wish to find

$$E(e^{tX} | N = n) \quad \text{where } N \sim B(n, p).$$

We have

$$E(e^{tX} | N = n) = \sum_{k=0}^{\infty} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} = (pe^t + 1 - p)^n.$$

Hence,

$$E(e^{tX} | N) = (pe^t + 1 - p)^N.$$

This implies that

$$M_X(t) = E((pe^t + 1 - p)^N) = \exp(\lambda p(e^t - 1)).$$

- (b) From the moment generating function of X in (a), we see that $X \sim \text{Po}(\lambda p)$.
- (c) Given $N = n$, we know that $X \sim B(n, p)$ so $N - X \sim B(n, 1 - p)$. However, $N \sim \text{Po}(\lambda)$. Using the same technique as in (a), the moment generating function of $N - X$ is

$$M_{N-X}(t) = E\left(e^{t(N-X)}\right) = E\left(E\left(e^{t(N-X)}\right) | N\right).$$

We already know that

$$E(e^{-tX} | N = n) = (1 - p + pe^{-t})^n$$

so

$$M_{N-X}(t) = E\left((e^t(1 - p + pe^{-t}))^N\right)$$

which is equal to $\exp(\lambda(1 - p)(e^t - 1))$. Hence, $N - X \sim \text{Po}(\lambda(1 - p))$.

- (d) We just showed that $X \sim \text{Po}(\lambda p)$ and $N - X \sim \text{Po}(\lambda(1 - p))$. As such, the joint distribution is

$$P(X = k, N = n) = P(N = k) \cdot P(X = k | N = n) = \frac{e^{-\lambda} \lambda^n}{n!} \binom{n}{k} p^k (1 - p)^{n-k}.$$

Hence,

$$P(X = k, N - X = m) = P(X = k, N = k + m).$$

So,

$$P(X = k, N - X = m) = \frac{e^{-\lambda} \lambda^{k+m}}{(k+m)!} \binom{k+m}{k} p^k (1-p)^m = e^{-\lambda} \frac{\lambda^k p^k}{k!} \cdot \frac{\lambda^m (1-p)^m}{m!}.$$

This factors as

$$P(X = k, N - X = m) = P(X = k) \cdot P(N - X = m).$$

Hence, X and $N - X$ are independent. \square

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. By (7.4),

$$M(t) = \int_{-\infty}^a \frac{e^{kt}}{b-a} dk + \int_a^b \frac{e^{kt}}{b-a} dk + \int_b^\infty \frac{e^{kt}}{b-a} dk = \int_a^b \frac{e^{kt}}{b-a} dk = \frac{e^{bt} - e^{at}}{t(b-a)}$$

\square

Example 7.44. For example,

$$M(t) = \frac{e^{\pi t} - e^{et}}{t(\pi - e)}$$

is a proper moment generating function because by Proposition 7.10, it is the moment generating function of a continuous uniform random variable on $[e, \pi]$.

Proposition 7.11 (MGF of normal random variable). If $X \sim N(\mu, \sigma^2)$, then

$$M(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

Proof. By (7.4),

$$\begin{aligned} M_X(t) &= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(tk - \frac{(k-\mu)^2}{2\sigma^2}\right) dk \\ &= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(k-(\mu+\sigma^2 t))^2 + \frac{(\mu+\sigma^2 t)^2 - \mu^2}{2\sigma^2}\right) dk \\ &= \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right) \cdot \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(k-(\mu+\sigma^2 t))^2\right) dk \\ &= \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right), \end{aligned}$$

since the integral equals 1 (it is the total mass of a $N(\mu + \sigma^2 t, \sigma^2)$ density). \square

Example 7.45. Let X be the standard normal random variable and N be a Poisson random variable with parameter λ . So,

$$X \sim N(0, 1) \quad \text{and} \quad N \sim \text{Po}(\lambda).$$

- (a) What is the moment generating function of X ?
- (b) What is the moment generating function of N ?
- (c) What is the moment generating function of $N - \lambda$?
- (d) Show by using the moment generating function method that

$\frac{N - \lambda}{\sqrt{\lambda}}$ is similar to X as $\lambda \rightarrow \infty$.

Solution.

- (a) From Proposition 7.11, we have

$$M_X(t) = \exp\left(\frac{1}{2}t^2\right) \quad \text{where } t \in \mathbb{R}.$$

- (b) From Proposition 7.9, we have

$$M_N(t) = \exp(\lambda(e^t - 1)) \quad \text{where } t \in \mathbb{R}.$$

- (c) We have

$$M_{N-\lambda}(t) = E\left(e^{t(N-\lambda)}\right) = E\left(e^{tN}\right)E\left(e^{-t\lambda}\right)$$

which is equal to

$$e^{-t\lambda} \exp(\lambda(e^t - 1)) = \exp(\lambda(e^t - 1 - t\lambda)).$$

- (d) Let $Y = \frac{N-\lambda}{\sqrt{\lambda}}$. Then,

$$M_Y(t) = E\left[\exp\left(\frac{t(N-\lambda)}{\sqrt{\lambda}}\right)\right] = e^{-\frac{t\lambda}{\sqrt{\lambda}}} E\left(e^{\frac{t}{\sqrt{\lambda}}N}\right) = e^{-t\sqrt{\lambda}} M_N\left(\frac{t}{\sqrt{\lambda}}\right).$$

Then,

$$M_Y(t) = \exp\left(\lambda\left(e^{t/\sqrt{\lambda}} - 1 - \frac{t}{\sqrt{\lambda}}\right)\right).$$

Expand $e^{t/\sqrt{\lambda}}$ as $\lambda \rightarrow \infty$ to obtain

$$e^{t/\sqrt{\lambda}} = 1 + \frac{t}{\sqrt{\lambda}} + \frac{t^2}{2\lambda} + \frac{t^3}{6\lambda^{3/2}} + O\left(\frac{1}{\lambda^2}\right).$$

The result follows. □

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. By (7.4),

$$M(t) = \int_0^\infty e^{tk} \lambda e^{-\lambda k} dk = \frac{\lambda}{\lambda - t}$$

$M(t)$ 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. \square

Example 7.46. For example,

$$M(t) = \frac{6}{(1-t)(2-t)(3-t)}$$

is a proper moment generating function. We see that $M_3(0) = 1$. Recall from Proposition 7.12 that the moment generating function of an exponential random variable with parameter λ is $\frac{\lambda}{\lambda - t}$. As

$$M_3(t) = \frac{1}{1-t} \cdot \frac{2}{2-t} \cdot \frac{3}{3-t},$$

then by the multiplicativity property of moment generating functions (Proposition 7.5), we see that M_3 is the moment generating function of $E_1 + E_2 + E_3$, where E_1, E_2, E_3 are independent exponential random variables, where $E_k \sim \text{Exp}(k)$.

Example 7.47. Let X be a random variable with moment generating function given by

$$M_X(t) = \frac{e^{2t}(2+e^{2t})}{3(2-e^t)}.$$

- (a) Is X discrete or continuous?
- (b) Compute $E(X)$.
- (c) Characterise the distribution of X .

Solution.

- (a) Observe that

$$M_X(t) = \frac{e^{2t}(2+e^{2t})}{3} \cdot \frac{1}{2-e^t} = \frac{e^{2t}(2+e^{2t})}{3} \sum_{k=0}^{\infty} \left(\frac{e^t}{2}\right)^k.$$

Hence,

$$M_X(t) = \frac{1}{6} \sum_{k=0}^{\infty} [2e^{(2+k)t} + e^{(4+k)t}] \cdot \frac{1}{2^k}$$

so X is a discrete random variable.

- (b) Note that

$$M'_X(t) = \frac{e^{2t}(8+8e^{2t}-2e^t-3e^{3t})}{3(2-e^t)^2}$$

so $M'_X(0) = E(X) = \frac{11}{3}$.

(c) From (a), we know that X is a discrete random variable. As

$$M_X(t) = \sum_{k=0}^{\infty} e^{kt} P(X = k),$$

then we see from (a) that

$$M_X(t) = \frac{e^{2t}}{3} \sum_{k=0}^{\infty} \frac{e^{kt}}{2^k} + \frac{e^{4t}}{6} \sum_{k=0}^{\infty} \frac{e^{kt}}{2^k}.$$

It follows that $P(X = x) = 0$ if $x < 2$, $P(X = 2) = \frac{1}{3}$, $P(X = 3) = \frac{1}{6}$ and $P(X = x) = 2^{-(x-2)}$ if $x \geq 4$. \square

Example 7.48. Let N be the discrete random variable with moment generating function

$$M(t) = 1 + (1 - e^{2t}) e^{et}.$$

- (a) Find $E(N)$.
- (b) Characterise the distribution of N .

Solution.

- (a) We have $E(N) = M'(0) = -2e$.
- (b) One sees that $M''(0) = -8e < 0$ which is impossible since this represents $E(N^2)$. As such, $M(t)$ cannot be the moment generating function of a random variable. \square

Example 7.49. Using the moment generating function, how that the following relationships are correct as $n \rightarrow \infty$.

- (a) A discrete uniform random variable on $\{1/n, 2/n, \dots, n/n\}$ is similar to a continuous uniform on $[0, 1]$.
- (b) A binomial random variable with parameters n and λ/n is similar to a Poisson random variable with parameter λ .
- (c) A geometric random variable with parameter λ/n is similar to an exponential random variable with parameter λ .

Solution.

- (a) One can show that the moment generating function of a discrete uniform random variable on $\{1/n, 2/n, \dots, n/n\}$ is

$$\frac{e^{t/n}(1 - e^t)}{n(1 - e^{t/n})} \quad \text{for } t \neq 0.$$

Then, consider the limit as $n \rightarrow \infty$ for which we would obtain some function, and because moment generating functions uniquely characterise the distribution, the result follows.

(b) Let $X \sim B(n, p)$, where $p = \lambda/n$ so $X \sim B\left(n, \frac{\lambda}{n}\right)$. So,

$$M_X(t) = \left[1 + \frac{\lambda}{n} (e^t - 1)\right]^n.$$

Next, let $Y \sim \text{Po}(\lambda)$. Then,

$$M_Y(t) = \exp(\lambda(e^t - 1)).$$

Let n go to infinity, then the result follows.

(c) Again, find the MGF of each then the result follows. \square

Limit Theorems

8.1 Statistical Inequalities

Probability Theory is often used to describe the behaviour of random phenomena. However, in many practical situations, we are interested not in the outcome of a single trial, but in the cumulative or average behaviour of many random variables such as the total number of successes in repeated trials, or the average outcome of a random experiment performed many times. This is where *limit theorems* come in: they describe how sums, averages, and distributions of random variables behave in the long run.

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.

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx \int_0^{\infty} xf(x) dx \quad \text{since } X \text{ is non-negative.}$$

We can bound this below by

$$E(X) \geq \int_a^{\infty} xf(x) dx \geq \int_a^{\infty} af(x) dx = aP(X \geq a)$$

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),

$$P(|X - \mu| \geq a) = P((X - \mu)^2 \geq a^2) \leq \frac{E[(X - \mu)^2]}{a^2} = \frac{\text{Var}(X)}{a^2}.$$

□

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)

$$1 + n + \frac{n^2}{2!} + \cdots + \frac{n^{2n}}{(2n)!} \geq \left(1 - \frac{1}{n}\right) e^n \quad (8.1)$$

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 (Theorem 8.2),

$$\begin{aligned} P(|X - 20n| > 4n) &\leq \frac{16n}{(4n)^2} \\ 1 - P(|X - 20n| \leq 4n) &\leq \frac{1}{n} \\ 1 - \frac{1}{n} &\leq P(|X - 20n| \leq 4n) \\ \alpha &\geq 1 - \frac{1}{n} \end{aligned}$$

and the result follows.

(ii) This is quite interesting. Observe that the left side of the inequality in (8.1) 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 $Y \sim$

$\text{Po}(n)$. 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

$$P(Y > 2n) \leq \frac{1}{n} \quad \text{so} \quad 1 - \frac{1}{n} \leq P(Y \leq 2n).$$

Hence,

$$1 - \frac{1}{n} \leq \sum_{i=0}^{2n} \frac{e^{-n} n^i}{i!} \quad \text{so} \quad \sum_{i=0}^{2n} \frac{n^i}{i!} \geq \alpha e^n$$

which concludes our proof. \square

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,

$$0 = \lim_{n \rightarrow \infty} P\left(|X - \mu| > \frac{1}{n}\right) = P\left(\lim_{n \rightarrow \infty} \left\{ |X - \mu| > \frac{1}{n} \right\}\right) = P(X \neq \mu).$$

This asserts that $P(X = \mu) = 1$. \square

8.2 Laws of Large Numbers (LLN)

One of the central goals of Probability Theory is to understand how random behaviour averages out over time. In many applications, we rely on repeated sampling to estimate some unknown quantity. A natural question then arises: as we increase the number of observations, does the average outcome become closer to the true value? The laws of large numbers (LLN) provide a rigorous answer to this question.

Theorem 8.4 (weak law of large numbers (WLLN)). Let X_1, X_2, \dots, X_n be a sequence of independent and identically distributed random variables, with a common mean μ . We define the sample mean to be

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n).$$

Then, for any $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|\bar{X} - \mu| \geq \varepsilon) = 0.$$

In other words, the sample mean converges to the expected value as $n \rightarrow \infty$.

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 $E(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$, then by Chebyshev's inequality (Theorem 8.2),

$$P(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

As $n \rightarrow \infty$, the expression on the right side of the inequality tends to 0. \square

Theorem 8.5 (strong law of large numbers (SLLN)). Let X_1, X_2, \dots, 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 $n \rightarrow \infty$,

$$\bar{X} \rightarrow \mu.$$

In probabilistic terms,

$$P\left(\left\{\lim_{n \rightarrow \infty} \bar{X} = \mu\right\}\right) = 1.$$

The weak law states that for a specified large n , the average \bar{X} is likely to be near μ . Thus, it leaves open the possibility that $|\bar{X} - \mu| > \varepsilon$ 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 $|\bar{X} - \mu| < \varepsilon$ 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 con-

verse 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.

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*. A fun fact is that if you were to go to Wikipedia, you will find three other types of CLT which are namely the Lyapunov CLT, Lindeberg 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, \dots, X_n be a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2 . Then, the distribution of

$$\frac{X_1 + X_2 + \dots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \rightarrow \infty$.

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)

$$X_1 + X_2 + \dots + X_n \sim N(n\mu, n\sigma^2) \quad \text{approximately}$$

(ii)

$$\frac{1}{n}(X_1 + X_2 + \dots + X_n) \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{approximately}$$

Example 8.2. Consider the continuous random variable X with probability density function

$$f(x) = \begin{cases} (1+2x)e^{-2x} & \text{if } x > 0; \\ 0 & \text{otherwise.} \end{cases}$$

We sample independent and identically distributed samples X_1, \dots, X_n with the same distribution as X .

- (a) State the central limit theorem satisfied by $X_1 + \dots + X_n$.

- (b)** What is the probability that $X_1 + \dots + X_{100}$ is larger than 70?

Solution.

- (a)** First, note that $E(X) = 0.75$ and $\text{Var}(X) = 0.4375$. Let $S_n = X_1 + \dots + X_n$. By the central limit theorem (Theorem 8.6),

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - 0.75n}{\sqrt{0.4375n}} \sim N(0, 1).$$

approximately.

- (b)** We wish to find $P(X_1 + \dots + X_{100}) > 70$. Equivalently, we need $P(S_{100} > 70)$ so

$$P\left(\frac{S_{100} - 0.75 \cdot 100}{\sqrt{0.4375 \cdot 100}}\right) > \frac{70 - 0.75 \cdot 100}{\sqrt{0.4375 \cdot 100}} = -0.755928946018.$$

So, the answer is 0.776. □

Example 8.3. Consider the discrete random variable X with probability mass function

$$P(X = n) = \frac{2+n}{2^{n+2}} \quad \text{for } n \in \mathbb{N}.$$

We sample independent and identically distributed random variables X_1, \dots, X_n with the same distribution X .

- (a)** State the central limit theorem satisfied by $X_1 + \dots + X_n$.
(b) What is the probability that $X_1 + \dots + X_{100} > 100$?

Solution.

- (a)** We note that

$$E(X) = 2.5 \quad \text{and} \quad \text{Var}(X) = 3.25.$$

Define $S_n = X_1 + \dots + X_n$. By the central limit theorem (Theorem 8.6),

$$\frac{S_n - 2.5n}{\sqrt{3.25n}} \sim N(0, 1) \text{ approximately.}$$

- (b)** Equivalently, we wish to find $P(S_{100} > 100)$. By the central limit theorem (Theorem 8.6),

$$P\left(\frac{S_{100} - 2.5 \cdot 100}{\sqrt{3.25 \cdot 100}} > \frac{100 - 2.5 \cdot 100}{\sqrt{3.25 \cdot 100}}\right)$$

which is approximately 1. □

Example 8.4. We want to connect two cities 1 km from each other using electric cables. The cable company provides us cable parts with a random length following the distribution (in meters)

$$f(x) = \begin{cases} c(x+x^2) & \text{if } 1 < x < 2; \\ 0 & \text{otherwise.} \end{cases}$$

We then attach these cables together to create a longer cable and connect the two cities.

- (a) What is the value of c ? Compute $E(X)$ and $\text{Var}(X)$.
- (b) Using the central limit Theorem, how many cables should we order to be at least 95% confident that we have enough cables?
- (c) If we order 630 cables, what is the probability that we can connect the two cities?

Solution.

- (a) $c = \frac{6}{23}$. Consequently,

$$E(X) = \frac{73}{46} = 1.5869 \quad \text{and} \quad \text{Var}(X) = 0.077221.$$

- (b) Let $S_n = X_1 + \dots + X_n$. Then, by the central limit theorem (Theorem 8.6),

$$\frac{S_n - 1.5869n}{\sqrt{0.077221n}} \sim N(0, 1) \quad \text{approximately.}$$

We need to find the smallest value of n such that $P(X_1 + \dots + X_n > 1000) > 0.95$.

Hence,

$$P\left(S_n > \frac{1000 - 1.5869n}{\sqrt{0.077221n}}\right) > 0.95.$$

As such, $n = 638$.

- (c) We wish to find

$$P(X_1 + \dots + X_{630} > 1000)$$

which is equal to

$$P\left(Z > \frac{1000 - 1.5869 \cdot 630}{\sqrt{0.077221 \cdot 630}}\right) = P(Z > 0.036272)$$

which is equal to 0.512. □

Example 8.5. When playing a slot machine at the casino, the probability of our gain is given by the random variable N with distribution

$$P(N = n) = \begin{cases} \frac{1}{2} & \text{if } n = 0, \\ \frac{1}{5} & \text{if } n = 1, \\ \frac{1}{5} & \text{if } n = 2, \\ \frac{1}{10} & \text{if } n = 5, \\ 0 & \text{otherwise.} \end{cases}$$

We keep playing until our total earnings are \$100.

- (a) Compute $E(N)$ and $\text{Var}(N)$.
- (b) Using the central limit theorem, how many times should we play in order to be at least 95% confident that we earned enough?
- (c) If we play 100 times, what is the probability that we did not reach our goal yet?

Solution.

- (a) We have $E(N) = 1.1$ and $\text{Var}(N) = 2.29$.
- (b) Let the number of times be n . Define $S_n = N_1 + \dots + N_n$, where the N_i 's are independent and identically distributed. By the central limit theorem (Theorem 8.6),

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - 1.1n}{\sqrt{2.29n}} \sim N(0, 1).$$

We need to find n such that $P(S_n > 100) > 0.95$. Hence,

$$P\left(Z > \frac{100 - 1.1n}{\sqrt{2.29n}}\right) > 0.95.$$

Hence, $n = 116$.

- (c) We wish to find $P(S_{100} < 100)$, which is equal to

$$P\left(Z < \frac{99.5 - 100 \cdot 1.1}{\sqrt{2.29 \cdot 100}}\right) = 0.244$$

by continuity correction. \square

Example 8.6. Let X be a normal random variable with mean 0 and variance σ^2 and N be a binomial random variable with parameters n and p .

- (a) What is the moment generating function of X ?
- (b) What is the moment generating function of N ?
- (c) What is the moment generating function of $N - np$?
- (d) Show by using the moment generating function that

$\frac{N - np}{\sqrt{n}}$ is similar to X with $\sigma^2 = p(1-p)$ where $n \rightarrow \infty$.

- (e) Explain why this is a direct application of the central limit theorem.

Solution.

- (a) By Proposition 7.11, we have

$$M_X(t) = \exp\left(\frac{\sigma^2 t^2}{2}\right).$$

- (b) By Proposition 7.7, we have

$$M_N(t) = (1 - p + pe^t)^n.$$

- (c) Let $Y = N - np$. Then,

$$M_Y(t) = E(e^{tY}) = E\left(e^{tN} e^{-t(np)}\right) = e^{-np} e^{tN} = e^{-np} M_N(t)$$

so we conclude that

$$M_Y(t) = e^{-np} (1 - p + pe^t)^n.$$

(d) Let $Z_n = \frac{N-np}{\sqrt{n}}$. We see that

$$M_{Z_n}(t) = E\left(e^{t \cdot \frac{N-np}{\sqrt{n}}}\right) = E\left(e^{\frac{t}{\sqrt{n}} \cdot (N-np)}\right) = M_{N-np}\left(\frac{t}{\sqrt{n}}\right) = M_Y\left(\frac{t}{\sqrt{n}}\right).$$

Hence,

$$M_{Z_n}(t) = e^{-np \frac{t}{\sqrt{n}}} \left(1 - p + pe^{\frac{t}{\sqrt{n}}}\right)^n.$$

Taking natural logarithm on both sides yields

$$\ln M_{Z_n}(t) = -ptn\sqrt{n} \ln \left(1 - p + pe^{\frac{t}{\sqrt{n}}}\right).$$

Thereafter, apply Taylor series expansion which would yield the desired result.

(e) This result is a direct application of the central limit theorem because the quantity $\frac{N-np}{\sqrt{n}}$ is exactly

$$\frac{N-np}{\sqrt{n}} = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n X_i - np \right) = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n X_i - p \right).$$

By the central limit theorem (Theorem 8.6), for independent and identically distributed random variables with finite mean μ and variance σ^2 ,

$$\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu)$$

converges in distribution to the standard normal $N(0, 1)$. In our case, set $\mu = p$ and $\sigma^2 = p(1-p)$, and the result follows. One should recall that this is precisely the de Moivre-Laplace theorem (Theorem 5.4), which is a special case of the central limit theorem for Bernoulli random variables. \square

Bibliography

- [1] S. M. Ross. *A First Course in Probability*. 10th Edition. Pearson Education Limited, Harlow, 2019. ISBN: 9781292269207.
- [2] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye. *Probability and Statistics for Engineers and Scientists*. 9th Edition. Pearson Education Limited, Harlow, 2016. ISBN: 9781292161365.
- [3] C. C. Chen and K.-M. Koh. *Principles and Techniques in Combinatorics*. World Scientific, 1992. ISBN: 9789810211394.