Contents

Preface

This set of notes is aligned to the NUS lecture notes, and my reference books, namely 'Probability and Statistics for Engineers and Scientists' by Sharon L. Myers, Raymond Myers, Ronald E. Walpole and Keying Ye and 'Principles and Techniques in Combinatorics' by Chuan Chong Chen and Koh Khee Meng.

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.

1 Combinatorics

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

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

1.1 Basic Principles of Counting

1.1.1 Addition Principle (AP)

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.

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

The different choices cannot occur at the same time.

1.1.2 Multiplication Principle (MP)

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

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

$$m_1 m_2 \ldots m_n.$$

1.2 Permutations and Combinations

1.2.1 Permutations

A permutation is an ordered arrangement of objects. In permutations, order matters, i.e. the 3-letter arrangement of ABC and ACB are considered. Given n distinct objects, the total number of ways of arranging all these n objects in is simply

$$n(n-1)(n-2)\ldots 3 \cdot 2 \cdot 1 = n!.$$

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, we have only 1 way to put the last object there. □

If we consider the permutation of n objects of which n_1 are identical, n_2 are identical,..., n_r are identical, there are

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

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

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

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

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

1.2.2 Combinations

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

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

We establish two results.
(1):

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

Proof: By considering the formula for r combinations out of n objects, setting $r = 0$ and $r = n$ will yield this result.

(2): Symmetry of Binomial Coefficients

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

The algebraic proof is simple and not as meaningful as its combinatorial counterpart. As such, we provide a proof for the latter.

Proof: Actually, the combinatorial proof is simple. There are two ways to select a group of r items from a group of n, which are namely picking the r items that you're going to include or picking the nr 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. □

1.2.3 Combinatorial Identities

We establish some combinatorial identites in this section.

(1): Pascal's Identity
For $n, k \in \mathbb{N}$,

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

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

(2): The Binomial Theorem
Let n be a non-negative integer. 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. By setting $x = y = 1$, we obtain

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

If a set has n elements, then the number of subsets, including the null set and itself, is 2^n.
Proof: Every element can be chosen or not chosen during the selection process. Since there are n elements, the result follows. □

Also, setting $x = 1$ and $y = -1$, we have

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

(3): Hockey-Stick Identity
For $n, r \in \mathbb{N}$, the Hockey-Stick Identity states that

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

Proof: This can be proven via induction or repeatedly applying Pascal's Identity.

(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,

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

By setting $m = n = p$, we obtain

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

1.2.4 Secrets of the Pascal's Triangle

Pascal's Triangle is a triangular array of the binomial coefficients that arises in Probability Theory, Combinatorics, and Algebra. There are interesting patterns which arise due to the features of the triagle such as the Pascal's Identity and the binomial coefficients aformentioned, and others including the Triangular Numbers and Fibonacci Numbers. You can find a document containing some fascinating patterns and the link to it is here.

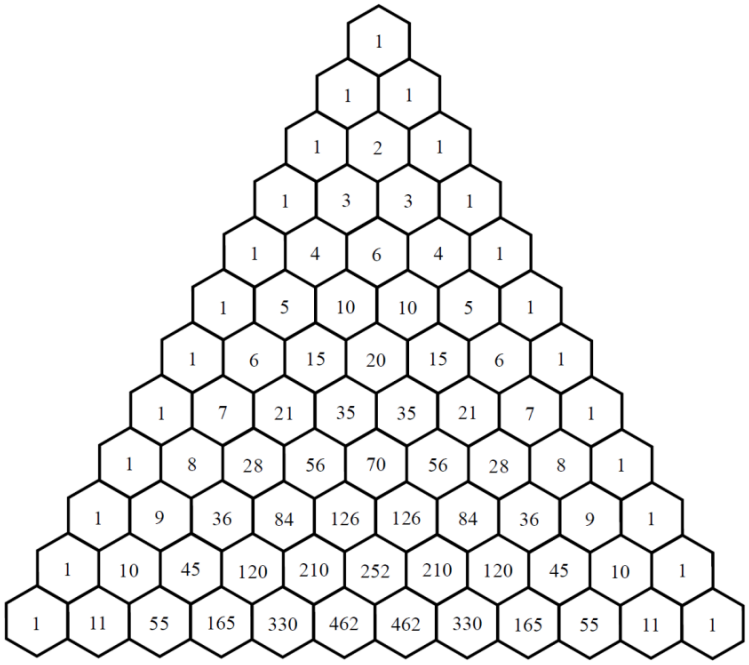

Figure 1: Pascal's Triangle

1.2.5 Multinomial Theorem

For any positive integer r and non-negative integer n, the multinomial theorem describes how a sum with m terms expands when raised to an arbitrary power n.

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

where $\binom{n}{k_1,k_2,\ldots,k_r}$ is a multinomial coefficient. Note that the binomial theorem is a special case of the multinomial theorem and its formula can be obtained by setting $n = 2$.

1.3 Distribution Problems

1.3.1 Identical Objects into Distinct Boxes

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

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

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 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: 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: Modeling the problem as stars and bars, it would start off by looking like this:

$$\star \vert \star \vert \star\star \vert \star\star \vert \star\star$$

Figure 2: Example of Stars and Bars Approach

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

1.3.2 Distinct Objects into Distinct Boxes

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

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

Proof: The first object goes into the first box. There are n ways to do this. The second object goes into the second box and there are $n-1$ ways to do so. Repeating to the r^{th} object, there are $n-r+1$ ways for it to go into the n^{th} box. By the Multiplication Principle, the required number of ways is $n(n-1)(n-2)\dots(n-r+1)$, which yields the above expression. □

Case 2: To distribute r distinct objects into n distinct objects, where $r, n \in \mathbb{N}$ such that each box can hold any number of objects, the number of ways is n^r.

Proof: 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. □

Case 3: Read the section on Stirling Numbers of the Second Kind, which is one of the later section in Combinatorics, before proceeding with this case. To distribute r distinct objects into n distinct objects, where $r, n \in \mathbb{N}$ and $r \geq n$ such that no box is empty, the number of ways is

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

The proof relies on the Principle of Inclusion and Exclusion (PIE) but we shall not discuss it as of now.

1.3.3 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. These numbers $s(r,n)$ are called the Stirling Numbers of the First Kind, named after Scottish Mathematician James Stirling. He is also known for

Stirling's Approximation, which involves an asymptotic formula for $n!$ for large values of n.

We state some obvious results.

$$\begin{aligned} s(r,0) &= 0 \text{ if } r \geq 1 \\ s(r,r) &= 1 \text{ if } r \geq 0 \\ s(r,1) &= (r-1)! \text{ for } r \geq 2 \\ s(r,r-1) &= \binom{r}{2} \text{ for } r \geq 2 \end{aligned}$$

We can form a recurrence relation for $s(r,n)$. That is,

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

Proof: We 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 the first case, 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 the second case, 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. Thus, the result follows.

As the two cases are mutually exclusive, the result follows by the Addition Principle. □

1.3.4 Distinct Objects into Identical Boxes and Stirling Numbers of the Second Kind

Case 1: Given non-negative integers r and n 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.

We state some obvious results.

$$\begin{aligned} S(r,1) &= S(r,r) = 1 \\ S(r,k) &= 0 \text{ if } 1 \leq r < k \\ S(r,0) &= S(0,k) = 0 \text{ if } r \geq 1 \text{ and } k \geq 1 \\ S(r,2) &= 2^{r-1} - 1 \text{ for } r \geq 1 \\ S(r,r-1) &= \binom{r}{2} \text{ for } r \geq 1 \end{aligned}$$

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

Similar to the Stirling Numbers of the First Kind, we have a similar recurrence relation for the Stirling Numbers of the Second Kind, which is slightly easier to derive since we are not considering circular permutations. The recurrence relation is

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

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

For the first case, 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 the second case, 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. □

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

$$S(r,1)+S(r,2)+\ldots+S(r,n).$$

Proof: This follows by the fact that we allow the boxes to be empty and the Addition Principle. □

1.3.5 Identical Objects into Identical Boxes

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 is immaterial since the integers are regarded as identical objects. Let the number of different partitions of n be denoted by $P(r,n)$.

This is the same as distributing r identical objects into n identical boxes.

We have a 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 the first case, 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 the second case, we place one object into each of the n boxes. Then we distribute the remaining $r-n$ objects into the n boxes such that each box has at least two objects. The number of ways this can be done is $P(r-n,n)$. By the Addition Principle, the result follows. □

1.3.6 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 $1\leq i\leq n$,

$$a_ix_i+a_2x_2+\ldots+a_nx_n=b.$$

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

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

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

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

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

where $x_i\geq 0$ for $1\leq i\leq r$.

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

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

is the same as the number of positive solutions to

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

or equivalently,

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

which is $\binom{r+n-1}{r-1}$. □

Example: Here is a relatively easy problem from the Singapore Mathematical Olympiad (SMO) 2022 Open Section Problem 18, which asks the solver to find the number of integer solutions to the equation $x_1 + x_2 - x_3 = 20$ with $x_1 \geq x_2 \geq x_3 \geq 0$.

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

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

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

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

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

Example: This question is from the Singapore Mathematical Olympiad (SMO) 2007 Open Section Problem 6, which asks the solver to find the non-negative solutions to the following inequality:

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

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

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

which is $\binom{24}{4} = 10626$. □

2 Probability

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

2.1 Kolmogorov Axioms

The Kolmogorov Axioms are named after Russian Mathematician Andrey Kolmogorov. There are numerous Russian Mathematicians who contributed to the field of Probability and Statistics. Some include Andrey Markov, who is known for Markov's Inequality and Markov Chains, Nikolai Smirnov, for which the Kolmogorov-Smirnov Test, a non-parametric test, is named after him and Kolmogorov, as well as Pafnuty Chebyshev. Chebyshev's Inequality and the Chebyshev Polynomials of the First Kind and the Second Kind are named after him. Not to mention, he also contributed to the much celebrated Prime Number Theorem.

First Axiom
For any event A, $0 \leq P(A) \leq 1$.

Second Axiom
Let S be the sample space. Then, $P(S) = 1$.

Third Axiom
For any sequence of mutually exclusive events $A_1, A_2, \ldots$ (that is, $A_i A_j =$ whenever $i \neq j$), then

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

2.2 Properties of Probability

Using Kolmogorov's Axioms, we can derive a few useful properties such as De Morgan's Laws and the probability of the complement of an event, where the complement is usually denoted by A' or A^c, where $P(A) + P(A^c) = 1$. In particular, we shall discuss the Principle of Inclusion and Exclusion.

2.2.1 Principle of Inclusion and Exclusion

We start off simple with two events, A and B. Then,

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

This can be illustrated using a Venn Diagram.

If we have three events, A, B and c, then

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

In general, if we have n events $A_1, A_2, \ldots, A_n$, then

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

Example: This is a question from the H3 Mathematics 2020 paper involving the Principle of Inclusion and Exclusion. Let $X = \{1, 2, \ldots, m\}$ and $Y = \{1, 2, \ldots, n\}$ be sets of positive integers and f be a function mapping from X to Y. f is called one-to-one if no two elements of X map to the same element of Y, and f is called onto if each element of Y is the image of an element of X. For $m \geq n$, we wish to find an expression for the number of functions mapping X to Y which are onto.

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

$$\left|A_1^c \cap A_2^c \cap \ldots \cap A_n^c\right|,$$

which is equivalently, by De Morgan's Law,

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

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 Principle of Inclusion and Exclusion,

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

which is the required expression. □

2.2.2 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 $!n$.

The Hat-Check Problem involves a group of n men entering a restaurant and checking 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. Then, for $n \geq 3$, D_n satisfies the following recurrence relation:

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

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

Proof: 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 the first case, 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 the second case, treating the i^{th} person as the first person, this is equivalent to arranging the $n-1$ hats among $n-1$ people such that no one received his own hat. There are D_{n-1} ways to do so. The result follows. □

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

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

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

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

by the method of difference. In conclusion,

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

We can also prove this result by the Principle of Inclusion and Exclusion.

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

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

Example: We recall the example from the H3 Mathematics 2020 paper. Let $X = \{1, 2, \ldots, m\}$ and $Y = \{1, 2, \ldots, n\}$ be sets of positive integers and f be a function mapping from X to Y. f is called one-to-one if no two elements of X map to the same element of Y, and f is called onto if each element of Y is the image of an element of X. Now, for $m = n = 5$, we wish to find the number of one-to-one functions mapping from X to Y which map no element to itself.

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

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

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

2.2.3 Birthday 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 $n \leq 365$ because if $n \geq 366$, then we obtain a contradiction by the Pigeonhole Principle. As $p(22) = 0.47569$ and $p(23) = 0.50729$, it asserts that the statement is true. We provide a proof for this.

Proof: A person can have his/her birthday on any of the 365 days. There is 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^c 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,

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

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

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

□

2.3 Sequence of Events

A sequence of events E_n, $n \geq 1$, is said to be an increasing sequence if

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

whereas it is said to be a decreasing sequence if

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

If E_n, $n \geq 1$, is an increasing sequence of events, then we define a new event, denoted by $\lim_{n\to\infty} E_n$, as

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

Similarly, if E_n, $n \geq 1$, is a decreasing sequence of events, then we define a new event, denoted by $\lim_{n\to\infty} E_n$, as

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

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

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

2.4 Conditional Probability and Independence

2.4.1 Conditional Probability

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

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

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 A has occurred, we can now think of A as our new, or reduced sample space.

We establish the general multiplication rule. If $A_1, A_2, \ldots, A_n$ are events, then

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

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

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

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

Let A be an event such that $P(A) > 0$. Then, the following three conditions hold:
(1): For any event B,

$$0 \leq P(B|A) \leq 1.$$

Proof: As $P(A \cap B) \geq 0$ and $P(A) > 0$, we prove the lower bound for $P(B|A)$. To prove $P(B|A) \leq 1$, note that $B|A \subset A$, implying that $P(A \cap B) \leq P(A)$, and the result follows. □

(2):

$$P(S|A) = 1$$

Proof: This follows from the fact that

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

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

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

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

Proof:

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} B_i \middle| A\right) &= \frac{P\left(A \cap \bigcup\limits_{i=1}^{\infty} B_i\right)}{P\left(A\right)} \\ &= \frac{P\left(A \cap \bigcup\limits_{i=1}^{\infty} B_i\right)}{P\left(A\right)} \\ &= \frac{\sum\limits_{i=1}^{\infty} P\left(B_i \cap A\right)}{P\left(A\right)} \\ &= \sum_{i=1}^{\infty} \frac{P\left(B_i \cap A\right)}{P\left(A\right)} \\ &= \sum_{i=1}^{\infty} P\left(B_i|A\right) \end{aligned}$$

□

2.4.2 Bayes' Theorem

Bayes' Theorem states that if A and B are two different events with $P\left(B\right) \neq 0$, then

$$P\left(A|B\right) P\left(B\right) = P\left(B|A\right) P\left(A\right).$$

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

2.4.3 Law of Total Probability

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

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

We are now ready to state the Law of Total Probability. Suppose the events $A_1, A_2, \ldots, A_n$ are partitions of S. Assume further that $P(A_i) > 0$ for all $1 \leq i \leq n$. Let B be any event. Then,

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

COROLLARY: From the above equation, we have, for $1 \leq i \leq n$,

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

2.4.4 Monty Hall Problem

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

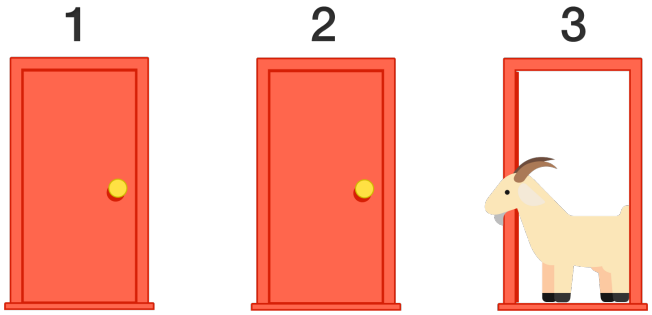

Figure 3: Illustration of the Monty Hall Problem

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

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

Then,

$$P\left(B|A\right)P\left(A\right) + P\left(B|A^c\right)P\left(A^c\right) = P\left(B\right)$$

Proof:

$$P\left(B|A\right)P\left(A\right) + P\left(B|A^c\right)P\left(A^c\right) = P\left(A \cap B\right) + P\left(B \cap A^c\right).$$

As A and B are independent events,

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

□

It is clear that

$$P\left(A|B\right) = \frac{P\left(B|A\right)P\left(A\right)}{P\left(B|A\right)P\left(A\right) + P\left(B|A^c\right)P\left(A^c\right)}.$$

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 $P\left(A\right) = \frac{1}{3}$ and $P\left(B|A\right) = P\left(B|A^c\right) = 1$. Putting everything together, $P\left(A|B\right) = \frac{1}{3}$. Hence, the probability that the car is behind Door No. 3 is $\frac{2}{3}$ and so, you, the contestant, should make the switch.

2.4.5 Independent Events

Two events A and B are said to be independent if

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

If

$$P(A \cap B) \neq P(A)P(B),$$

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

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

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

Proof: By the definition of conditional probability, $P(A|B) = \frac{P(A \cap B)}{P(B)}$. Since A and B are independent, $P(A \cap B) = P(A)P(B)$. The rest is simple algebra. □

We shall now discuss pairwise independence and mutual independence.

Given three events A, B and C, we say that A, B and C are pairwise independent if

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

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

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

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

2.4.6 Gambler's Ruin Problem

The Gambler's Ruin Problem states that a gambler playing a game with negative expected value will eventually go broke, regardless of their betting system.

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

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

$$P(\text{win}) = p,\ P(\text{lose}) = q \text{ and } P(\text{draw}) = r$$

and thus, $p + q + r = 1$. By employing the first-step analysis, we can set up a second-order linear homogeneous recurrence relation. That is,

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

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

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

Proof: To transit from the x_j to x_{j+1}, the player needs to win, hence we multiply by the associated probability p. The same can be said for the transition from x_j to x_{j-1}, where the player needs to lose, implying that we

multiply by q. For the player to remain at the same state, he needs to obtain a draw. That is, multiplying x_j by r. As such,

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

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

One can solve the recurrence relation to obtain the required probability

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

Proof: 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

$$\begin{aligned} x_N &= 1 \\ A - A\left(\frac{q}{p}\right)^N &= 1 \\ A &= \frac{1}{1 - \left(\frac{q}{p}\right)^N} \end{aligned}$$

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

The interested reader can visit this link for more information on this interesting concept of the Gambler's Ruin Problem.

3 Discrete Random Variables

A random variable is said to be discrete if the range of X is either finite or countably infinite. Examples of discrete random variables include the binomial distribution covered in H2 Mathematics and the Geometric Distribution and the Poisson Distribution covered in H2 Further Mathematics.

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

Suppose a random variable X is discrete, taking values $x_1, x_2, \ldots$. Then, the probability density function (or in some older textbooks, probability mass function) of X is

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

The probability density function is abbreviated as PDF, whereas the probability mass function is abbreviated as PMF. Some properties of the probability density function are
(1): $p_X(x_i) \geq 0$ for $i = 1, 2, \ldots$
(2): $p_X(x) = 0$ for all other values of x
(3): 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.

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

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

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

3.1 Expectation

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

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

We state some properties of expectation. Let X and Y be random variables and a and b be constants. Then,
(1): $E(aX) = aE(X)$
(2): $E(a) = a$
(3): $E(aX \pm b) = aE(X) \pm b$
(4): $E(aX \pm bY) = aE(X) \pm bE(Y)$
(5): If $X_1, X_2, \ldots, X_n$ are independent random variables, then

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

3.1.1 Expectation of a Function of a Random Variable

Given a random variable X, we are often interested about $g(X)$ and $E(g(X))$. The question is how do we compute the latter? One method is to find the PDF of $g(X)$ first before proceeding to compute $E(g(X))$ by

definition. We have the following proposition that if X is a discrete random variable that takes values x_i, where $i \geq 1$, with respective probabilities $p_X(x_i)$, then for any real-valued function g,

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

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

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

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

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

In general, for $n \geq 1$, $E(X^n)$ is called 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.

3.1.2 Tail Sum Formula for Expectation

For a non-negative integer-valued random variable X,

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

3.2 Variance and Standard Deviation

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

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

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

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

Proof:

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

and the result follows. □

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

We say that $\text{Var}(X) = 0$ if and only if X is a degenerate random variable. Moreover, from the formula for variance, it follows that

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

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

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

Hmm ... I wonder ... We will discuss in one of the final sections, and to prove this conjecture, we need to use a famous inequality called Jensen's Inequality.

We state some properties of variance. Let X be a random variable and a and b be constants. Then,
(1): $\text{Var}(aX) = a^2 \text{Var}(X)$
(2): $\text{Var}(a) = 0$
(3): $\text{Var}(aX + b) = a^2 \text{Var}(X)$
(4): If $X_1, X_2, \ldots, X_n$ are independent random variables, then

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

3.3 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 $P(X = 0) = 1 - p = q$.

3.3.1 Expectation and Variance

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

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

We first prove the result for expectation, which is obvious.
Proof: $E(X) = 0(q) + 1(p) = p$ □

Next, we prove the result for variance.
Proof: $E(X^2) = 0^2(q) + 1^2(p) = p$. As $[E(X)]^2 = p^2$, then $\text{Var}(X) = p - p^2 = p(1 - p) = pq$. □

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

To summarise,

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

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

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

- **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 could be represented as $X \sim \mathrm{B}\left(20, \frac{1}{4}\right)$.
- **Coin toss:** Suppose a coin is tossed 50 times and we wish to find out how many heads we obtain. Here, X is the number of successes. That is the number of times heads occurs. The probability of getting a head is $\frac{1}{2}$. The binomial distribution could be represented as $X \sim \mathrm{B}\left(50, \frac{1}{2}\right)$.

3.4.1 Expectation and Variance

The expressions for expectation and variance are

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

We first prove the formula for expectation.
Proof:

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

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

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

$\sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j}$ is the sum of probabilities of the binomial random variable, which is 1, and we are done. □

Alternatively, we can prove the expectation formula by considering it as a sum of independent Bernoulli Trials.

For the variance proof, I'll leave it as an exericse as it is not too complicated and the technique is, of course, similar to that for the expectation. When proving the formula for variance, note that

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

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

3.4.2 Mode

The mode is the value k at which the PDF takes its maximum value. In other words, it is the value that is most likely to be sampled. For a binomial distribution with parameters n and p, the mode is

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

Proof: Note that

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

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

To summarise,

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

3.4.3 Additivity

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

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

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

From the second last line to the last line, we used Vandermonde's Identity. □

3.5 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 $1, 2, 3, \ldots$ because the minimum number of tries required to obtain the first success is 1. As such, it is easy to derive the following formula, which is the PDF of X:

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

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

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.

3.5.1 Expectation and Variance

The formulae for expectation and variance are

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

We will only prove the formula for expectation.
Proof:

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

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

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

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

3.5.2 Memoryless Property

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 $m, n \in \mathbb{N}$,

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

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

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

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

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

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

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

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

It is clear that

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

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

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

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

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

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

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

To summarise,

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

3.5.3 Additivity

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

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

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

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

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

3.5.4 The 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?

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

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

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

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

and the result follows. □

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

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

where H_n is the n^{th} harmonic number. For $\text{Var}(T)$, it is rather interesting that we do not have an explicit formula but only an upper bound for it.

We shall first prove the result for expectation.
Proof: Assume that the t_i's are independent. Then,

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

□

Next, we prove the result for variance.
Proof:

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

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

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

Thus, it suffices to prove

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

This is true because $i - 1 < n \iff i < n + 1$. Hence, the result follows. □

3.6 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 PDF of a negative binomial random variable is

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

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

The geometric distribution is a special case of the negative binomial distribution. We can view the geometric distribution $\text{Geo}(p)$ 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.

3.6.1 Expectation and Variance

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

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

To summarise,

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

3.6.2 Banach's Matchbox Problem

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

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

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

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

□

3.7 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 $0, 1, 2, \ldots$ with probabilities

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

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

Some examples where the Poisson Distribution can be used are as follows:

- **Calls per hour at a call centre:** Call centers use the Poisson Distribution to model the number of expected calls per hour that they'll receive so they know how many call center reps to keep on staff. For example, suppose a given call center receives 10 calls per hour. Then, $X \sim \text{Po}(10)$.
- **Number of arrivals at a restaurant:** 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, $X \sim \text{Po}(100)$.

Time plays a critical role when defining a Poisson Random Variable.

3.7.1 Expectation and Variance

If $X \sim \text{Po}(\lambda)$, then

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

We shall prove the result for expectation.
Proof:

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

□

To summarise,

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

3.7.2 Additivity

The additivity property of the Poisson Distribution states that if X and Y are independent Poisson Random Variables where $X \sim \text{Po}(\lambda)$ and $Y \sim \text{Po}(\mu)$, then $X + Y \sim \text{Po}(\lambda + \mu)$.

Proof:

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

□

Even though $X+Y$ follows a Poisson Distribution, $X-Y$ actually does not follow a Poisson Distribution. In general, the difference of two Poisson Random Variables is said to follow a *Skellam Distribution*. Its PDF is rather complicated to compute as it involves the Modified Bessel Function of the First Kind (related to differential equations).

3.7.3 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

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

where

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

Proof:

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

which is indeed the PDF of a binomial random variable with n tries and probability of success $\frac{\lambda}{\lambda+\mu}$.

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

Fish comes to the surface of a stretch of river randomly and independently at a mean rate of 8 per minute.

When a fish comes to the surface, the probability that it catches a fly is 0.6. If S is the number of flies caught in a randomly chosen minute, show that

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

and deduce that S follows a Poisson Distribution.

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

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

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

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

Hence,

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

To prove that S follows a Poisson Distribution, we manipulate with the given PDF formula.

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

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

3.7.4 Law of Rare Events

The Poisson Distribution has a variety of applications in diverse areas because it caan 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. We provide a proof for this.

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

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

For large n and a moderate-sized λ,

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda},\ \left(1-\frac{\lambda}{n}\right)^k \approx 1 \text{ and } \frac{n(n-1)(n-2)\dots(n-k+1)}{n^k} \approx 1.$$

Hence, we conclude that

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

3.8 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.

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

By Vandermonde's Identity, the sum of probabilities is indeed equal to 1. That is,

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

Example: 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 PDF formula of a hypergeometric distribution. Note that $N = 14$, $K = 6$, $n = 3$ and $k = 2$. Substituting everything into the formula yields

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

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

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

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

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

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

Our *O-Level method* actually has a potential limitation, which is that if n and k are large, the total number of permutations will also be large and many cases will arise (just like how the number of COVID-19 cases there are as of now when I'm writing this which is 4 July 2022).

3.8.1 Expectation and Variance

The expectation and variance of a hypergeometric random variable are

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

To summarise,

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

3.9 Summary of Discrete Random Variables

To summarise the main components of discrete random variables,

Random Variable	PDF	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)	$\binom{n}{k} p^k q^{n-k}$	n and p	np	npq
Geo(p)	pq^{k-1}	p	$\frac{1}{p}$	$\frac{q}{p^2}$
NB(r, p)	$\binom{k-1}{r-1} p^r q^{k-r}$	r and p	$\frac{r}{p}$	$\frac{rq}{p^2}$
Po(λ)	$\frac{e^{-\lambda}\lambda^k}{k!}$	λ	λ	λ
Hypergeometric(N, K, n)	$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$	N, K and n	$\frac{nK}{N}$	$\frac{nK(N-K)(N-n)}{N^2(N-1)}$

4 Continuous Random Variables

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

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

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

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

The function f_X is called the PDF of X. Recall that PDF stands for probability density function. By letting $B = [a, b]$, we obtain

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

We define the cumulative distribution function, or CDF, of X by

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

for $x \in \mathbb{R}$. Note that the definition of the distribution function is the same for both discrete and continuous random variables. Therefore, in the context of continuous random variables,

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

By the Fundamental Theorem of Calculus,

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

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

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

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

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.

We establish some properties in relation to continuous random variables.
(1): $P(X = x) = 0$
(2): The CDF, that is F_X, is continuous
(3): For any $a, b \in \mathbb{R}$,

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

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

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

4.1 Expectation and Variance

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

Let X be a continuous random variable with PDF $f(x)$. Then,

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

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

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

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

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

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

4.2 Continuous Uniform Distribution

A random variable X is said to be uniformly distributed over the interval $(0, 1)$ if its PDF is

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

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

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

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

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.

4.2.1 Expectation and Variance

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

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

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:

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

Hence,

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

□

To summarise,

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

4.3 Normal Distribution

A random variable X is said to be normally distributed with parameters μ and σ, where μ is the mean and σ^2 is the variance, if its PDF is

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

where $x \in \mathbb{R}$. We say that $X \sim N(\mu, \sigma^2)$. Even though the PDF formula looks very complicated, one can verify that the integral from $-\infty$ to ∞ is indeed 1 (i.e. sum of probabilities is 1). To the interested, this uses a well-known result, known as the Gaussian Integral. I will attach the proof here, but one needs to have some pre-requisites to matrices and Multivariable Calculus to understand the 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, $\int_c^d f(x, y)\ dy$ is a function of x and we can integrate it with respect to x from $x = a$ to $x = b$. The resulting integral

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

is known as an *iterated integral*.

The Fubini-Tonelli 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 $R = [a, b] \times [c, d]$,

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

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

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

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

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

involves Fubini's Theorem.

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

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

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

$$dxdy = rdrd\theta$$

using the Jacobian of a suitable matrix. That is,

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

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

$$\begin{aligned} I^2 &= \int_0^{2\pi} \int_0^{\infty} re^{-r^2}\, drd\theta \\ &= \pi \\ I &= \sqrt{\pi} \end{aligned}$$

□

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

4.3.1 Expectation and Variance

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

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

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)	$E(X)$	$\mathrm{Var}(X)$
$N(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ and σ	μ	σ^2

4.3.2 Standard Normal Random Variable

A normal random variable is called a standard normal random variable when $\mu = 0$ and $\sigma = 1$. This is denoted by SZ. That is, $Z \sim N(0, 1)$. Its PDF and CDF are usually denoted by ϕ and Φ respectively. That is,

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

Some properties of the standard normal distribution are as follows:
(1): $P(Z \geq 0) = P(Z \leq 0) = 0.5$ due to symmetry
(2): $-Z \sim N(0, 1)$
(3): $P(Z \leq x) = 1 - P(Z > x)$ for $x \in \mathbb{R}$
(4): $P(Z \leq -x) = P(Z \geq x)$ for $x \in \mathbb{R}$
(5): If $X \sim N(\mu, \sigma^2)$, then,

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

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

4.3.3 The 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.

4.3.4 de Moivre-Laplace Theorem

Suppose $X \sim B(n,p)$. Then, for any $a < b$, we have

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

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

$$\frac{X-np}{\sqrt{npq}} \approx Z,$$

where $Z \sim N(0,1)$.

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

4.3.5 Continuity Correction

If $X \sim B(n,p)$, then

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

4.4 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 $X \sim \text{Po}(\lambda)$.

For non-negative integers x, the CDF is $F(x) = 1 - e^{-\lambda x}$.

4.4.1 Mean and Variance

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

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

To summarise,

Random Variable	PDF	Parameter(s)	$E(X)$	$\text{Var}(X)$
$\text{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

4.4.2 Median and Exponential Decay

If $T \sim \text{Exp}(\lambda)$, then the median m is $\frac{\ln 2}{\lambda}$.
Proof: This is easy to prove by considering the CDF formula. Substituting $t = m$, we have

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

□

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

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

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

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

4.4.3 Memoryless Property

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

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

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

We provide a proof for this statement.
Proof: We apply the definition of conditional probability to the left side. Hence,

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

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

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

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

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

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

By First Principles,

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

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

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

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

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

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

4.4.4 Poisson Process

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

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

Proof: Note that

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

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

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

4.4.5 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 and claim that W also follows an exponential distribution. That is,

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

Proof:

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

Differentiating both sides yields

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

asserting that our claim is true. □

COROLLARY: If $T_i \sim \text{Exp}(\lambda)$ for $1 \leq i \leq n$, and that all the T_i's are identically distributed, then

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

4.4.6 Inverse Transform Sampling

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: specifically, if Y has a uniform distribution on $(0, 1)$ and if X has a cumulative distribution F_X, then the random variable $F_X^{-1}(Y)$ has the same distribution as X.

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

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

□

4.5 Gamma Distribution

A random variable X is said to follow a gamma distribution with parameters α and λ, and is denoted by $X \sim \Gamma(\alpha, \lambda)$. The PDF only exists for $x \geq 0$ and its formula is

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

where $\alpha, \lambda > 0$ and $\Gamma(\alpha)$, called the gamma function, is defined by

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

It is easy to prove that $\Gamma(1) = 1$ and that the Gamma Function satisfies the following recurrence relation:

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$$

The recurrence relation can be proven using integration by parts. Hence, it is easy to establish that for integer values of α, say $\alpha = n$, we have $\Gamma(n) = (n-1)!$.

Observe that $\Gamma(1, \lambda) = \text{Exp}(\lambda)$, which implies that the exponential distribution is a special case of the gamma distribution.

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

$$\int_0^\infty e^{-t} t^{-\frac{1}{2}}\ dt = \int_{-\infty}^\infty e^{-u^2}\ du = \sqrt{\pi}$$

This follows from the Gaussian Integral. □

4.5.1 Expectation and Variance

If $X \sim \Gamma(\alpha, \lambda)$, then the expectation and variance are

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

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	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$

4.5.2 Gamma Process

Similar to the Poisson Process, we have a similar result, known as a 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. □

4.6 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

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx.$$

4.6.1 Relation to the Gamma Function

The beta function is related to the gamma function by the following relationship:

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

Proof: We first consider $\Gamma(a)\Gamma(b)$ and write it as an integral. Then,

$$\Gamma(a)\Gamma(b) = \left(\int_0^\infty e^{-u}u^{a-1}\, du\right)\left(\int_0^\infty e^{-v}v^{b-1}\, dv\right) = \int_0^\infty \int_0^\infty e^{-(u+v)}u^{a-1}v^{b-1}\, dudv.$$

We use the change of variables $u = zt$ and $v = z(1-t)$. Hence, $v = -z(t-1)$. Recall that $u, v \geq 0$, which implies that $0 \leq t \leq 1$ and $z \geq 0$. Upon change of variables, we have

$$\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_0^\infty \int_0^1 e^{-z}(zt)^{a-1}(z(1-t))^{b-1} z\, dtdz \\
&= \left(\int_0^\infty e^{-z}z^{a+b-1}\, dz\right)\left(\int_0^1 t^{a-1}(1-t)^{b-1}\, dt\right) \\
&= \Gamma(a+b)B(a,b)
\end{aligned}$$

which asserts that the statement is true. □

4.6.2 Expectation and Variance

If $X \sim \text{Beta}(a, b)$, then

$$E(X) = \frac{a}{a+b} \text{ and } \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

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	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$

4.7 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!

4.8 Summary of Continuous Random Variables

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 Joint Probability Distribution

5.1 Joint Distribution Functions

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

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$$

for $x, y \in \mathbb{R}$. Note that $\{X \leq x, Y \leq y\}$ is equivalently $\{X \leq x\} \cap \{Y \leq y\}$.

The distribution function of X can be obtained from the joint density function of X and Y via

$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x,y).$$

We call F_X the marginal distribution function of X.

Similarly,

$$F_Y(y) = \lim_{x \to \infty} F_{X,Y}(x,y)$$

and F_Y is the marginal distribution function of Y.

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,

(1):

$$P(X > a, Y > b) = 1 - F_X(a) - F_Y(b) + F_{X,Y}(a,b)$$

Proof: We set $A = \{X \leq a\}$ and $B = \{Y \leq b\}$. Then, the required event is $A^c \cap B^c$, which is the same as $(A \cap B)^c$, and by considering the complement of it, it is equivalently $n(S) - (A \cap B)$. By the Principle of Inclusion and Exclusion, the required probability is $1 - P(A \cup B)$. Hence,

$$\begin{aligned} P(X > a, Y > b) &= 1 - P(A \cup B) \\ &= 1 - P(A) - P(B) + P(A \cap B) \\ &= 1 - F_X(a) - F_Y(b) + F_{X,Y}(a,b) \end{aligned}$$

which concludes the proof. □

(2):

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = F_{X,Y}(a_2,b_2) - F_{X,Y}(a_1,b_2) + F_{X,Y}(a_1,b_1) - F_{X,Y}(a_2,b_1)$$

5.1.1 Jointly Discrete Random Variables

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

$$p_X(x) = P(X = x) = \sum_{y\in\mathbb{R}} p_{X,Y}(x,y) \text{ and } p_Y(y) = P(Y = y) = \sum_{x\in\mathbb{R}} p_{X,Y}(x,y).$$

p_x and p_y are the marginal probability density function of X and Y respectively.

Example: 3 balls are randomly selected from an urn containing 3 red, 4 white and 5 blue balls. If we let R and W denote the number of red and white balls chosen respectively, then we can construct a joint probability density function table of R and W. It is shown below.

Solution:

White (right); Red (bottom)	0	1	2	3	$P(R = r)$
0	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$	$\frac{84}{220}$
1	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$	0	$\frac{108}{220}$
2	$\frac{15}{220}$	$\frac{12}{220}$	0	0	$\frac{27}{220}$
3	$\frac{1}{220}$	0	0	0	$\frac{1}{220}$
$P(W = w)$	$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$	

It should be clear as to how these probabilities are computed. □

Some useful formulae are as follows:

(1):

$$P(a_1 < X \le a_2, b_1 < Y \le b_2) = \sum_{a_1<x\le a_2} \sum_{b_1<y\le b_2} p_{X,Y}(x,y)$$

(2):

$$F_{X,Y}(a,b) = P(X \le a, Y \le b) = \sum_{x\le a} \sum_{y\le b} p_{X,Y}(x,y)$$

(3):

$$P(X > a, Y > b) = \sum_{x>a} \sum_{y>b} p_{X,Y}(x,y)$$

5.1.2 Jointly Continuous Random Variables

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 $C \subset \mathbb{R}^2$, we have

$$P((X,Y) \in C) = \int\int_{(x,y\in C)} f_{X,Y}(x,y)\ dxdy.$$

We state some useful formulae.

(1): Let $A, B \subset \mathbb{R}$. Set $C = A \times B$ (i.e. C is the Cartesian product of A and B). Then,

$$P(X \in A, Y \in B) = \int_A \int_B f_{X,Y}(x,y)\ dydx.$$

(2): 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

$$P(a_1 < X \le a_2, b_1 < Y \le b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x,y)\ dydx.$$

(3): Let $a, b \in \mathbb{R}$. Then,

$$F_{X,Y}(a,b) = P(X \le a, Y \le b) = \int_{\infty}^{a} \int_{\infty}^{b} f_{X,Y}(x,y)\ dydx.$$

Hence,

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

Example: 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

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

The joint probability density function of X and Y is

$$f_{X,Y}(x,y) = \begin{cases} 2e^{-x}e^{-2y} & x > 0, y > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Suppose we wish to compute the following probabilities:
(i) $P(X > 1, Y < 1)$
(ii) $P(X < Y)$
(iii) the marginal probability density function of X
(iv) $P(X \leq x)$
(v) the marginal distribution function of Y

Solution:
(i) This probability can be expressed by the following integral:

$$\int_1^{\infty} \int_0^1 2e^{-x}e^{-2y}\ dydx$$

and this is a very simple problem in Multivariable Calculus. The answer is $e^{-1}(1 - e^{-2})$. □

(ii) As $0 < x < y$ and $0 < y < \infty$, the required probability is

$$\int_0^{\infty} \int_0^y 2e^{-x}e^{-2y}\ dxdy.$$

The answer is $\frac{1}{3}$. I omit the integration process because it is simple. I believe the only issues readers might have is setting up the double integral. We have an alternative representation for it. That is, we set $x < y < \infty$ and $0 < x < \infty$. Hence, the integral is just

$$\int_0^{\infty} \int_x^{\infty} 2e^{-x}e^{-2y}\ dydx = \frac{1}{3}.$$

It yields the same conclusion as before! □

(iii) Recall that the formula for the marginal probability density function of X is

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

Substituting everything in yields

$$f_X(x) = \int_{-\infty}^{\infty} 2e^{-x}e^{-2y}\ dy = e^{-x}.$$

Hence, for $x > 0$, the marginal probability density function is $f_X(x) = e^{-x}$. □

(iv) Note that $P(X \leq x)$ is the marginal distribution function of x, so

$$F_X(x) = \int_{\infty}^{x} e^{-t}\ dt = 1 - e^{-x},$$

for $x > 0$. □

(v) The marginal distribution function of Y, for $y > 0$, is $F_Y(y) = 1 - e^{-2y}$. □

5.2 Independent Random Variables

Two random variables X and Y are said to be independent if

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

for any $A, B \subset \mathbb{R}$. Random variables that are not independent are said to be dependent.

For jointly discrete random variables, we have the following three equivalent statements:
(1): X and Y are independent
(2): For all $x, y \in \mathbb{R}$, $p_{X,Y}(x, y) = p_X(x)p_Y(y)$
(3): 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.
(1): X and Y are independent
(2): For all $x, y \in \mathbb{R}$, $f_{X,Y}(x, y) = f_X(x)f_Y(y)$
(3): For all $x, y \in \mathbb{R}$, $F_{X,Y}(x, y) = F_X(x)F_Y(y)$

For both discrete and continuous random variables, X and Y are independent if and only if there exist functions $g, h : \mathbb{R} \to \mathbb{R}$ such that for all $x, y \in \mathbb{R}$, $f_{X,Y}(x, y) = g(x)h(y)$.

In many applications, we either know or assume that X and Y are independent. Then the joint probability density function of X and Y can be obtained by multiplying the individual probability density functions.

Independence is said to be 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.

5.2.1 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 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 $\frac{L}{2}$. That is,

$$\frac{X}{\cos\theta} < \frac{L}{2} \implies X < \frac{L}{2}\cos\theta.$$

As X varies between 0 and $\frac{D}{2}$ and θ between 0 and $\frac{\pi}{2}$, it is reasonable to assume that they are independent and uniformly distributed random variables over these respective ranges. Note that $D = L\cos\theta$, and for $0 \leq x \leq \frac{D}{2}$, $f_X(x) = \frac{2}{x}$ and for $0 \leq \theta \leq \frac{\pi}{2}$, $f_\theta(\theta) = \frac{2}{\pi}$. We thus obtain the joint probability density function

$$f_X(x)f_\theta(\theta) = \frac{4}{\pi D}$$

for $0 \leq x \leq \frac{D}{2}, 0 \leq \theta \leq \frac{\pi}{2}$ and 0 elsewhere.

Hence,

$$P\left(X < \frac{L}{2}\cos\theta\right) = \int_0^{\frac{\pi}{2}} \int_0^{\frac{L}{2}\cos\theta} \frac{4}{\pi D}\, dxd\theta = \frac{2L}{\pi D}.$$

□

5.2.2 Sums of Independent Random Variables

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.

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:

$$\begin{aligned} F_{X+Y}(x) &= P(X+Y \le x) \\ &= \int\int_{s+t\le x} f_{X,Y}(s,t)\, dsdt \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{x-t} f_X(s)f_Y(t)\, dsdt \\ &= \int_{-\infty}^{\infty} F_X(x-t)f_y(t)\, dt \end{aligned}$$

□

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: Recall that the sum of two independent uniform distributions follows a triangular distribution. Let us prove this result! Suppose X and Y are independent random variables with a common uniform distribution over $(0,1)$. That is, $X \sim U(0,1)$ and $Y \sim U(0,1)$. We wish to find the probability density function of $X+Y$.

Solution: $X+Y$ takes values in $(0,2)$. For $x \le 0$ and $x \ge 2$, it follows that $f_{X+Y}(x) = 0$. For $0 < x < 2$,

$$\begin{aligned} f_{X+Y}(x) &= \int_{-\infty}^{\infty} f_X(x-t)f_Y(t)\, dt \\ &= \int_0^1 f_X(x-t)\, dt \end{aligned}$$

$f_X(x-t) > 0$ if and only if $0 < x-t < 1$. Note that x is fixed and t varies. We split this into two cases, namely $0 < x \le 1$ and $1 < x < 2$.

For $0 < x \le 1$,

$$\begin{aligned} f_{X+Y}(x) &= \int_0^1 f_X(x-t)\, dt + \int_x^1 f_X(x-t)\, dt \\ &= \int_0^x f_X(x-t)\, dt \\ &= \int_0^x dt \\ &= x \end{aligned}$$

In a similar fashion, it can be shown that for $1 < x < 2$,

$$f_{X+Y}(x) = 2 - x.$$

Hence,

$$f_{X+Y}(x) = \begin{cases} x & 0 < x \leq 1 \\ 2 - x & 1 < x < 2 \\ 0 & \text{elsewhere} \end{cases}.$$

The density function has the shape of a triangle, so $X + Y$ follows a triangular distribution. □

5.3 Conditional Probability Distribution

5.3.1 Conditional Discrete Probability Distribution

The conditional probability density function of X given that $Y = y$ is defined by

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

for all values of y such that $p_Y(y) > 0$. Similarly, the conditionl distribution function of X given that $Y = y$ is defined by

$$F_{X|Y}(x|y) = P(X \leq x|Y = y)$$

for y 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 $p_Y(y) > 0$.

5.3.2 Conditional Continuous Probability Distribution

Suppose X and Y are jointly continuous random variables. We define the conditional probability density function of X given $Y = y$ to be

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

for all values of y such that $f_Y(y) > 0$.

For $A \subset \mathbb{R}$ 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 $f_Y(y) > 0$.

5.4 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 $U = g(X,Y)$ and $V = h(X,Y)$ for some functions g 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 $V = \frac{X}{X+Y}$. It is clear that

$$g(x, y) = x + y \text{ and } h(x, y) = \frac{x}{x + y}.$$

In general, to find the joint probability density function of U and V, we state some conditions first.

Formulation of the Joint Probability Density Function

We assume that the following conditions are satisfied:
(1): Let X and Y be jointly continuously distributed random variables with a known joint probability density function.
(2): 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, $x = a(u, v)$ and $y = b(u, v)$.
(3): The functions g and h have continuous partial derivatives and

$$J(x, y) = \det \begin{pmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \\ \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \end{pmatrix} = \frac{\partial g}{\partial x}\frac{\partial h}{\partial y} - \frac{\partial g}{\partial y}\frac{\partial h}{\partial x} \neq 0.$$

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: Let X and Y be jointly distributed with the joint probability density function

$$f_{X,Y}(x, y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2 + y^2}{2}\right).$$

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 $\theta = h(x, y) = \tan^{-1}\left(\frac{y}{x}\right)$. Hence, $x = r\cos\theta$ and $y = r\sin\theta$, which is essentially the conversion formulae from polar to Cartesian coordinates.

I omit the differentiation process in this case, but anyway, $J(x, y) = (x^2 + y^2)^{-\frac{1}{2}}$. Hence,

$$\begin{aligned} f_{R,\Theta}(r, \theta) &= \frac{f_{X,Y}(x, y)}{\det(J(x, y))} \\ &= \sqrt{x^2 + y^2} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2 + y^2}{2}\right) \\ &= \frac{1}{\sqrt{2\pi}} r e^{-\frac{r^2}{2}} \end{aligned}$$

which is the joint probability density function of R and Θ. □

(ii) They are independent. □

In the above example, R is actually a special continuous random variable. We say that R follows a Rayleigh Distribution. A Rayleigh Distribution is often observed when the overall magnitude of a vector is related to its directional components. One example where the Rayleigh distribution naturally arises is when wind velocity is analysed in two dimensions. Assuming that each component is uncorrelated, normally distributed with equal variance, and zero mean, then the overall wind speed (vector magnitude) will be characterised by a Rayleigh Distribution.

If $X \sim \text{Rayleigh}(\sigma)$, where $\sigma > 0$, then

$$f(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right).$$

We call σ the scale parameter. Not only is the Rayleigh Distribution related to the normal distribution, but it is also related to the exponential distribution! That is, if $Y \sim \text{Exp}(\lambda)$, then

$$X = \sqrt{Y} \sim \text{Rayleigh}\left(\frac{1}{\sqrt{2\lambda}}\right).$$

6 Expectation Properties

We start off this section with the property that if $a \leq X \leq b$, then $a \leq E(X) \leq b$.

Proof: We will prove for the case where X is a discrete random variable. The proof for the continuous counterpart is similar, but we simply change the sum to an integral.

$$E(X) = \sum_{\text{all } x} xp(x) \geq \sum_{\text{all } x} ap(x) = a.$$

In a similar fashion, we can use the same technique to show that $E(X) \leq b$. □

6.1 Expectation of Sums of Random Variables

We state the following two propositions:
(a): If X and Y are jointly discrete random variables with joint probability density function $p_{X,Y}$, then

$$E[g(X,Y)] = \sum_{\text{all } y} \sum_{\text{all } x} g(x,y)p_{X,Y}(x,y)$$

(b): If X and Y are jointly continuous random variables with joint probability density function $f_{X,Y}$, then

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

Some important consequences are as follows:
(1): If $g(x,y) \geq 0$ whenever $p_{X,Y}(x,y) > 0$, then $E[g(X,Y)] \geq 0$
(2): $E[g(X,Y) + h(X,Y)] = E[g(x,Y)] + E[h(X,Y)]$
(3): $E[g(X) + h(Y)] = E[g(X)] + E[h(Y)]$
(4): 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

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i),$$

which was covered in H2 Mathematics. The formula for the expectation of the sample mean, $\overline{X}$, can be derived from the fourth point. It is clear that $E\left(\overline{X}\right) = \mu$, so the expected value of the sample mean is μ, the mean of the distribution. Hence, when μ is unknown, the sample mean is often used to estimate it.

Example: Recall that the binomial distribution is closely linked to the Bernoulli Distribution. Suppose we perform an experiment n times and the probability of success for each trial is p. We define X to be the number of successes in n Bernoulli(p) trials. Since the expectation of each Bernoulli random variable is p and there are n Bernoulli trials, by the linearity property of expectation, we can use this method to derive that $E(X) = np$.

6.1.1 Mean Line Segment Length

This involves a concept known as the mean line segment length. Suppose we have a unit square with vertices at $(0,0), (0,1), (1,0)$ and $(1,1)$. What is the mean distance between any two points in the square?

Solution: The above is a very interesting problem. The answer is definitely not $\frac{1}{2}$, but actually, rather close to it. The mean distance is approximately 0.52140, or in exact form,

$$\frac{2 + \sqrt{2} + 5\ln(1 + \sqrt{2})}{15}.$$

Let us prove this result. Let U and V be independent uniform random variables as such: $U \sim U(0,1)$ and $V \sim U(0,1)$. We wish to find the distribution of $W = |U - V|$. We find the CDF of W first, before differentiating

to find its PDF.

$$\begin{aligned}
P(W \le w) &= 1 - P(W > w) \\
&= 1 - P(|U - V| > w) \\
&= 1 - P(U - V < -w) - P(U - V > w) \\
&= 1 - P(V > U + w) - P(U > V + w) \\
&= 1 - \int_0^{1-w} P(V > U + w) f_U(u)\ du - \int_0^{1-w} P(U > V + w) f_V(v)\ dv \\
&= 1 - \int_0^{1-w} 1 - (u + w)\ du - \int_0^{1-w} 1 - (v + w)\ dv \\
&= 1 - (1 - w)^2
\end{aligned}$$

Upon differentiation yields $f_W(w) = 2(1 - w)$, where $0 < w < 1$. We use the formula

$$E[g(U, V)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u, v) f_{U,V}(u, v)\ dudv.$$

Since we are interested in the mean distance, or rather the expected distance, then $g(U, V) = \sqrt{U^2 + V^2}$, so $g(u, v) = \sqrt{u^2 + v^2}$. Note that $f_{U,V}(u, v) = 2(1 - u) \cdot 2(1 - v)$ due to independence. Therefore,

$$E\left(\sqrt{U^2 + V^2}\right) = 4 \int_0^1 \int_0^1 \sqrt{u^2 + v^2}(1 - u)(1 - v)\ dudv.$$

Using polar coordinates, $u = r \cos\theta$ and $v = r \sin\theta$. We need to find the bounds for r and θ too. By considering the lower half of the region, $0 \le r \le \sec\theta$ and $0 \le \theta \le \frac{\pi}{4}$. The integral becomes

$$8 \int_0^{\frac{\pi}{4}} \int_0^{\sec\theta} r^2(1 - \cos\theta)(1 - \sin\theta)\ drd\theta = 8 \int_0^{\frac{\pi}{4}} \frac{\sec^3\theta}{12} - \frac{\sec^3\theta \tan\theta}{20}\ d\theta.$$

The integral of $\sec^3\theta \tan\theta$ is a standard one because the derivative of $\sec\theta$ is $\sec\theta \tan\theta$. To integrate $\sec^3\theta$, we need to use integration by parts.

$$\begin{aligned}
\int_0^{\frac{\pi}{4}} \sec^3\theta\ d\theta &= \int_0^{\frac{\pi}{4}} \sec\theta \sec^2\theta \\
&= [\sec\theta \tan\theta]_0^{\frac{\pi}{4}} - \int_0^{\frac{\pi}{4}} \tan\theta \sec\theta \tan\theta\ d\theta \\
&= \sqrt{2} - \int_0^{\frac{\pi}{4}} \sec\theta(\sec^2\theta - 1)\ d\theta \\
&= \sqrt{2} - \int_0^{\frac{\pi}{4}} \sec^3\theta\ d\theta + \int_0^{\frac{\pi}{4}} \sec\theta\ d\theta \\
2 \int_0^{\frac{\pi}{4}} \sec^3\theta\ d\theta &= \sqrt{2} + [\ln|\sec\theta + \tan\theta|]_0^{\frac{\pi}{4}} \\
\int_0^{\frac{\pi}{4}} \sec^3\theta\ d\theta &= \frac{1}{\sqrt{2}} + \frac{1}{2} \ln\left(\sqrt{2} + 1\right)
\end{aligned}$$

The rest of the working is left as a simple exercise. □

6.1.2 Boole's Inequality

For a countable set of events $A_1, A_2, \ldots$, Boole's Inequality states that

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

The generalisation of this is Bonferroni's Inequality.

6.2 Covariance, Variance and Correlation

6.2.1 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. 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 causation. I strongly recommend a video by Zach Star which illustrates how easy it is to lie with Statistics.

Example: 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.

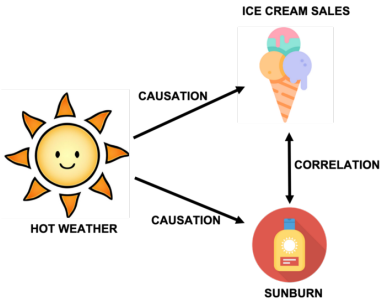

Figure 4: Correlation and Causation

If X and Y are independent random variables, then for any functions $g, h : \mathbb{R} \to \mathbb{R}$, we have

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

Some other properties of covariance are as follows:
(1): $\text{Var}(X) = \text{cov}(X, X)$
(2): $\text{cov}(X, Y) = \text{cov}(Y, X)$ (symmetrical property)
(3):

$$\text{cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \,\text{cov}\left(X_i, Y_j\right)$$

Proof:

$$\begin{aligned}
\text{cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) &= E\left(\sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j X_i Y_j\right) - E\left(\sum_{i=1}^{n} a_i X_i\right) E\left(\sum_{j=1}^{m} b_j Y_j\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j E\left(X_i Y_j\right) - \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j E\left(X_i\right) E\left(Y_j\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \,\text{cov}\left(X_i, Y_j\right)
\end{aligned}$$

□

(4):

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

Let $X_1, X_2, \ldots, X_n$ 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

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

be the sample mean. The quantities $X_i - \overline{X}$, 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

$$S^2 = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n-1}$$

is called the *sample variance*. We shall prove that $E(S^2) = \sigma^2$. That is, S^2 is used as an estimator for σ^2 instead of the more natural choice of

$$\sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n}.$$

Proof: Note that $X_i - \overline{X} = X_i - \mu + \mu - \overline{X}$. Hence,

$$\begin{aligned}\left(X_i - \overline{X}\right)^2 &= \left(X_i - \mu + \mu - \overline{X}\right)^2 \\ &= \left(X_i - \mu\right)^2 + \left(\overline{X} - \mu\right)^2 - 2\left(\overline{X} - \mu\right)\left(X_i - \mu\right)\end{aligned}$$

When we take the sum of i from 1 to n, note that $\overline{X}$ is unaffected by the index. Hence,

$$\begin{aligned}S^2 &= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(X_i - \mu\right)^2 + \sum_{i=1}^{n}\left(\overline{X} - \mu\right)^2 - 2\left(\overline{X} - \mu\right)\sum_{i=1}^{n}\left(X_i - \mu\right)\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(X_i - \mu\right)^2 + n\left(\overline{X} - \mu\right)^2 - 2n\left(\overline{X} - \mu\right)^2\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(X_i - \mu\right)^2 - n\left(\overline{X} - \mu\right)^2\right] \\ E\left(S^2\right) &= \frac{1}{n-1}\left[\sum_{i=1}^{n} E\left[\left(X_i - \mu\right)^2\right] - nE\left[\left(\overline{X} - \mu\right)^2\right]\right]\end{aligned}$$

By the definition of variance, as $E\left[\left(X - \mu\right)^2\right] = \text{Var}(X)$, then it is clear that

$$\sum_{i=1}^{n} E\left[\left(X_i - \mu\right)^2\right] = n\sigma^2.$$

The term $E\left[\left(\overline{X} - \mu\right)^2\right]$ is called the *variance of the sample mean*. We wish to find the sum of it from $i = 1$ to $i = n$. This is straightforward because

$$\sum_{i=1}^{n} E\left[\left(\overline{X} - \mu\right)^2\right] = \text{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}.$$

Putting everything together,

$$E\left(S^2\right) = \frac{1}{n-1}\left(n\sigma^2 - n\left(\frac{\sigma^2}{n}\right)\right) = \sigma^2.$$

□

6.2.2 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, is

$$r = \frac{\sum (x-\overline{x})\sum (y-\overline{y})}{\left[\sqrt{\sum (x-\overline{x})^2}\right]\left[\sqrt{\sum (y-\overline{y})^2}\right]}.$$

The two 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 $\text{Var}(X) \geq 0 \iff E(X^2) \geq (E(X))^2 \geq 0$. Hence, the discriminant of $f(t)$, Δ must satisfy $\Delta \leq 0$. That is,

$$(2E(UV))^2 - 4(E(U^2))(E(V^2)) \leq 0.$$

Rearranging yields the formula

$$(E(UV))^2 \leq E(U^2)E(V^2),$$

which implies that $-1 \leq \rho(X,Y) \leq 1$. To conclude, we remark that the inequality $(E(UV))^2 \leq E(U^2)E(V^2)$ is the famous Cauchy-Schwarz Inequality. □

The correlation coefficient is a measure of the degree of linearity between X and Y. A value of $\rho(X,Y)$ near $+1$ or -1 indicates a high degree of linearity between X and Y, whereas a value near 0 indicates a lack of such linearity. A positive value of $\rho(X,Y)$ indicates that Y tends to increase as X does, whereas a negative value indicates that Y tends to decrease as X increases. If $\rho(X,Y) = 0$, then X and Y are said to be uncorrelated. If X and Y are independent, then $\rho(X,Y) = 0$. However, the converse is not true.

6.3 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,

$$E(g(X)|Y=y) = \sum_{\text{all } x} g(x) p_{X|Y}(x|y) \text{ for the discrete case;}$$
$$E(g(X)|Y=y) = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\ dx \text{ for the continuous case}$$

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.

6.4 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

$$\text{Var}(X) = E(\text{Var}(X|Y)) + \text{Var}(E(X|Y)).$$

6.5 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.

The MGF of a random variable X, denoted by M_X, is defined as

$$M_X(t) = E\left(e^{tX}\right).$$

If X is a discrete random variable with PDF p_X, then

$$M_X(t) = \sum_{\text{all } x} e^{tx} p_X(x).$$

If X is a continuous random variable with PDF f_X, then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\ dx.$$

We call such a function a moment generating function because it generates all the moments of this random variable X. Indeed, for $n \geq 0$,

$$E\left(X^n\right) = M_X^{(n)}(0),$$

where

$$M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t) \text{ when } t \text{ is evaluated at } 0.$$

Proof: Using series expansion,

$$\begin{aligned} E(e^{tX}) &= E\left(\sum_{k=0}^{\infty} \frac{(tX)^k}{k!}\right) \\ &= \sum_{k=0}^{\infty} \frac{E(X^k) t^k}{k!} \\ &= \sum_{k=0}^{\infty} \frac{M_X^{(k)}(0) t^k}{k!} \end{aligned}$$

The result follows by equating the coefficient of t^n. □

The MGF of a random variable satisfies two properties. We state them.

(1): Multiplicativity

If X and Y are independent, then

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

(2): Uniqueness

Let X and Y be random variables with MGFs being M_X and M_Y respectively. If there exists $h > 0$ such that

$$M_X(t) = M_Y(t)$$

for all $-h < t < h$. Then, it implies that X and Y have the same distribution, meaning that $f_X = f_Y$. We state and prove the MGFs for some random variables.

(i): MGF of Bernoulli Random Variable

If $X \sim \text{Bernoulli}(p)$, then

$$M(t) = 1 - p + pe^t.$$

Proof: Using the formula, $M(t) = e^{t(0)}P(X = 0) + e^{t(1)}P(X = 1) = (1 - p) + pe^t$. □

(ii): MGF of Binomial Random Variable

If $X \sim \text{B}(n, p)$, then

$$M(t) = (1 - p + pe^t)^n.$$

Proof: Using the formula, and writing it in sigma notation,

$$\sum_{k=0}^{n} e^{kt}\binom{n}{k}p^k q^{n-k} = q^n \sum_{k=0}^{n}\binom{n}{k}\left(\frac{pe^t}{q}\right)^k = q^n\left(1 + \frac{pe^t}{q}\right)^n = (1 - p + pe^t)^n.$$

□

(iii): MGF of Geometric Random Variable

If $X \sim \text{Geo}(p)$, then

$$M(t) = \frac{pe^t}{1 - qe^t}.$$

Proof: Using the formula,

$$M(t) = \sum_{k=1}^{n} e^{kt}pq^{k-1} = \frac{p}{q}\sum_{k=1}^{n}(qe^t)^k = \frac{pe^t}{1 - qe^t}.$$

□

(iv): MGF of Poisson Random Variable

If $X \sim \text{Po}(\lambda)$, then

$$M(t) = \exp(\lambda(e^t - 1)).$$

Proof:

$$M(t) = \sum_{k=0}^{n}\frac{e^{kt}e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{n}\frac{(\lambda e^t)^k}{k!} = e^{-\lambda}\left(e^{\lambda e^t}\right) = \exp(\lambda(e^t - 1))$$

□

(v): MGF of Uniform Random Variable

If $X \sim U(a, b)$, then

$$M(t) = \frac{e^{bt} - e^{at}}{t(b - a)}.$$

Proof:

$$M(t) = \int_{-\infty}^{a}\frac{e^{kt}}{b - a}\,dk + \int_{a}^{b}\frac{e^{kt}}{b - a}\,dk + \int_{b}^{\infty}\frac{e^{kt}}{b - a}\,dk = \int_{a}^{b}\frac{e^{kt}}{b - a}\,dk = \frac{e^{bt} - e^{at}}{t(b - a)}$$

□

(vi): MGF of Normal Random Variable
If $X \sim N(\mu, \sigma^2)$, then

$$M(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

The proof will be left as an exercise.

(vii): MGF of Exponential Random Variable
If $X \sim \text{Exp}(\lambda)$, then

$$M(t) = \frac{\lambda}{\lambda - t}$$

for $t < \lambda$.

Proof:

$$M(t) = \int_0^\infty e^{tk}\lambda e^{-\lambda k}\, dk = \frac{\lambda}{\lambda - t}$$

□

$M(t)$ is only defined for $t < \lambda$ because the expectation of an exponential random variable is always positive.

Proof: To justify that the expectation of an exponential random variable is always positive, 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. □

7 Limit Theorems

7.1 Statistical Inequalities

7.1.1 Markov's Inequality

Let X be a non-negative random variable. For $a > 0$, we have

$$P(X \geq a) \leq \frac{E(X)}{a}.$$

Proof: We only prove this for the continuous random variable X. The discrete case is very similar, just that the integral is replaced by summation.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)\ dx \\ &= \int_{0}^{\infty} xf(x)\ dx \ \because X \text{ is non-negative} \\ &\geq \int_{a}^{\infty} xf(x)\ dx \\ &\geq \int_{a}^{\infty} af(x)\ dx \ \because f(x) \text{ is non-negative} \\ &= aP(X \geq a) \end{aligned}$$

which concludes the proof. □

7.1.2 Chebyshev's Inequality

Let X be a random variable with finite mean μ aand 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,

$$P(|X - \mu| \geq a) = P\left((X - \mu)^2 \geq a^2\right) \leq \frac{E\left[(X - \mu)^2\right]}{a^2} = \frac{\text{Var}(X)}{a^2}.$$

□

Example: The following question is from Question 12 of the STEP 3 2016 paper. The probability of a biased coin landing heads up is 0.2. It is thrown $100n$ times, where n is an integer greater than 1. Let α be the probability that the coin lands heads up N times, where $16n \leq N \leq 24n$. We can use Chebyshev's Inequality to prove the following two results:

(i)

$$\alpha \geq 1 - \frac{1}{n}$$

(ii)

$$1 + n + \frac{n^2}{2!} + \cdots + \frac{n^{2n}}{(2n)!} \geq \left(1 - \frac{1}{n}\right) e^n$$

However, in the test, their form of Chebyshev's Inequality is slightly different. That is for $k > 0$,

$$P(|X - \mu| > k\sigma) \leq \frac{1}{k^2}.$$

Solution:

(i) We first recognise that this is a setup modelling a binomial distribution. Let X be the random variable denoting the number of times the coin lands heads up, out of $100n$. Then, $\alpha = P(|X - 20n| \leq 4n)$. Note

that $E(X) = 20n$, $\text{Var}(X) = 16n$ and $|X - 20n| \leq 4n$. Removing the modulus, $16n \leq X \leq 24n$, which indeed satisfies the original inequality that $16n \leq N \leq 24n$. By Chebyshev's Inequality,

$$\begin{aligned} P(|X - 20n| > 4n) &\leq \frac{16n}{(4n)^2} \\ 1 - P(|X - 20n| \leq 4n) &\leq \frac{1}{n} \\ 1 - \frac{1}{n} &\leq P(|X - 20n| \leq 4n) \\ \alpha &\geq 1 - \frac{1}{n} \end{aligned}$$

and the result follows. □

(ii) This is quite interesting. Observe that the left side of the inequality is the partial sum of the Maclaurin Series of e^n. If we can prove that

$$1 + n + \frac{n^2}{2!} + \cdots + \frac{n^{2n}}{(2n)!} \geq \alpha e^n,$$

then we are done. Recall that the only discrete random variable we studied which contains the exponential function is the Poisson random variable. Suppose $Y \sim \text{Po}(n)$. Then, $\mu = \sigma^2 = n$. Set $a = n$. Substituting these into Chebyshev's Inequality yields

$$P(|Y - n| > n) \leq \frac{1}{n}.$$

We consider the modulus inequality first. This is equivalent to $Y - n \geq n$ or $Y - n \leq -n$, which implies that $Y \geq 2n$ or $Y \leq 0$ respectively. The latter does not make sense because the support of Y is the non-negative integers. Thus, the inequality becomes

$$P(Y > 2n) \leq \frac{1}{n}.$$

With some simple algebraic manipulation,

$$1 - \frac{1}{n} \leq P(Y \leq 2n).$$

Hence,

$$\begin{aligned} 1 - \frac{1}{n} &\leq \sum_{i=0}^{2n} \frac{e^{-n} n^i}{i!} \\ \sum_{i=0}^{2n} \frac{n^i}{i!} &\geq \alpha e^n \end{aligned}$$

which concludes our proof. □

The importance of Markov's and Chebyshev's Inequalities 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 could be exactly computed and we would not need to resort to bounds.

7.1.3 Jensen's Inequality

If X is a random variable and ϕ is a convex function, then Jensen's Inequality states that

$$\phi(E(X)) \leq E(\phi(X)).$$

COROLLARY: For $x \geq 0$, the graph of $\phi(x) = x^n$, where $n \in \mathbb{N}$, is convex. Hence, we establish that

$$E(X^n) \geq (E(X))^n$$

for $n \in \mathbb{N}$.

COROLLARY: 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, for any $n \geq 1$,

$$0 \leq P\left(|X - \mu| > \frac{1}{n}\right) \leq \frac{\text{Var}(X)}{\frac{1}{n^2}} = 0.$$

By the squeeze theorem, it implies that

$$P\left(|X - \mu| > \frac{1}{n}\right) = 0.$$

Taking limits on both sides and using the continuity property of probability,

$$0 = \lim_{n\to\infty} P\left(|X - \mu| > \frac{1}{n}\right) = P\left(\lim_{n\to\infty}\left\{|X - \mu| > \frac{1}{n}\right\}\right) = P(X \neq \mu).$$

This asserts that $P(X = \mu) = 1$. □

7.2 Laws of Large Numbers (LLN)

7.2.1 Weak Law of Large Numbers (WLLN)

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables, with a common mean μ. We define the sample mean to be

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n).$$

Then, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\overline{X} - \mu\right| \geq \varepsilon\right) = 0.$$

In other words, the sample mean converges to the expected value as $n \to \infty$.

Proof: We shall prove this theorem only under the additional assumption that the random variables have a finite variance σ^2. As It is clear that $E(\overline{X}) = \mu$ and $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}$, then by Chebyshev's Inequality,

$$P\left(\left|\overline{X} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

As $n \to \infty$, the expression on the right side of the inequality tends to 0. □

7.2.2 Strong Law of Large Numbers (SLLN)

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables, each having a finite mean $\mu = E(X_i)$. Recall how the sample mean is defined when we introduced the Weak Law of Large Numbers. Then, the Strong Law of Large Numbers states that as $n \to \infty$,

$$\overline{X} \to \mu.$$

In probabilistic terms,

$$P\left(\left\{\lim_{n\to\infty} \overline{X} = \mu\right\}\right) = 1.$$

The weak law states that for a specified large n, the average $\overline{X}$ is likely to be near μ. Thus, it leaves open the possibility that $|\overline{X} - \mu| > \varepsilon$ happens an infinite number of times, although at infrequent intervals.

In contrast, the strong law shows that this almost surely will not occur. Note that it does not imply that with probability 1, we have that for any $\varepsilon > 0$, the inequality $|\overline{X} - \mu| < \varepsilon$ holds for all large enough n since the convergence is not necessarily uniform on the set where it holds.

Almost sure convergence implies convergence in probability, but the converse is not true. The proof is out of the scope of our discussion as it is with reference to Probability Theory at a higher level. It uses a complex branch of Pure Mathematics called Measure Theory and we use a lemma, called the Borel-Cantelli Lemma, to prove the aforementioned statement. This is why there is a distinction between the weak law and the strong law.

7.3 Central Limit Theorem (CLT)

The Central Limit Theorem (CLT) is one of the most remarkable results in Probability Theory. It states that the sum of a large number of independent random variables has a distribution that is approximately normal. Hence, it not only provides a simple method for computing approximate probabilities for sums of independent random variables, but it also helps explain the remarkable fact that the empirical frequencies of so many natural populations exhibit bell-shaped curves.

We will only study one form of the CLT and it is known as the *Classical CLT*. Fun fact, if you were to go to Wikipedia, you will find that there are three other types of CLT, namely the Lyapunov CLT, Lindenberg CLT and the Multidimensional CLT. All these will be out of scope of our discussion.

Without further ado, we state the simplest form of the CLT - the Classical CLT.

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2. Then, the distribution of

$$\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$.

We have two results, one of which is related to the sum of X_i's, and one is related to the sample mean.
(1):

$$X_1 + X_2 + \ldots + X_n \sim N(n\mu, n\sigma^2)$$

approximately

(2):

$$\frac{1}{n}(X_1 + X_2 + \ldots + X_n) \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

approximately