MATH11177

Bayesian Theory

Sagar Udasi

October 7, 2024

Contents

Chapter 1

Foundational Algebra and Analysis

I want you to restart from scratch - I want to change the way you imagine *Mathematics*. I want you to visualise most of its concepts and appreciate its beauty.

Let's start with the most basic question possible - *What is mathematics?* In my opinion, mathematics is the study of abstractions. The first and foremost abstract idea is the *number*. Numbers are the direct consequence of our ability to count. When I mention 3 or 6, you understand the quantity without considering what is being counted; the focus is on the act of counting, making the number itself an abstract entity. Despite it being such an abstract entity, we have been able to work with it and have developed theories of mathematics around it - from algebra to calculus, from analysis to probability theory and what not.

But knowing numbers isn't enough. We have to extend this idea of abstractions before we dive deep into the probability theory.

1.1 The Abstraction of Mathematical Ideas

To explain what I mean by *abstraction* - the word that I have used dozens of times by now, I will present three seemingly unrelated ideas in mathematics and demonstrate how they converge into a single concept when thought abstractly.

1.1.1 Idea 1: Matrix Multiplication and Permutation Matrices

The first example comes from the course on *Linear Algebra*. *Have you ever wondered why you multiply two matrices the way you do?* It seems so bizzare that when you are adding two 3×3 matrices, you are adding elementwise; but when you are multiplying them, then suddenly a weird rule of multiplying rows with columns comes into picture. *Why?*

I expect you are fami liar with the facts that columns of a matrix are *vectors* and n simultaneous linear equations with n variables can be written in the matrix form $AX = B$, where A is the $n \times n$ coefficient matrix, X is the $n \times 1$ matrix of unknowns and B is the $n \times 1$ resultant vector. Geometrically, since each column of the coefficient matrix is a vector in n dimensions, if they are *linearly independent*, we will be able to find a solution i.e. a list of linear combination coefficients X such that a linear combination of those

vectors in A gets us the resultant vector B.

By the nature of the visualization itself in the column picture, if we were to find not just a single linear combination that results in B, but two more linear combinations that result in vectors C and D too, the equation becomes

$$A \cdot \begin{bmatrix} X_1 & X_2 & X_3 \end{bmatrix} = \begin{bmatrix} B & C & D \end{bmatrix}$$

If you put the appropriate elements in this matrix, and try to write three linear combinations i.e. three sets of simultaneous linear equations, you will see the reason - why we multiply the matrices the way we do so. From this idea, it should also get clear that the multiplication $A.B$ has a different meaning than $B.A$. This implies, matrix multiplication is not commutative.

But why do we care so much about matrix multiplications? Because we want to understand the impact on a matrix when it is multiplied by another matrix.

Now that we know matrix multiplication is a collection of linear combinations *(this idea is further refined to be called as **linear transformation**)*, we can easily confirm that - if $AB = B$, then A should have a form where diagonal elements are 1 and rest are 0. It's the *identity matrix*.

Let's do a reverse-engineering exercise. If

$$E. \begin{bmatrix} 1 & 2 & 1 \\ 3 & 8 & 1 \\ 0 & 4 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 2 & -2 \\ 0 & 4 & 1 \end{bmatrix} \tag{1.1}$$

then what is E?

From the first look it is clear that E is very close to the identity matrix as only the second row is different. If we sum three negatives of [1 2 1] and one positive of [3 8 1], we will get [0 2 -2]. Hence,

$$E = \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Now what if I have to nullify the effect of multiplication of E shown above on A? We should add 3 times the first row to the second row, to compensate for subtracting three times the first row from the second. Notice how the elements of the matrix are acting as row-selectors.

$$E^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

And we can confirm that their effects nullify, because $E.E^{-1} = I$. This gives us the concept of *inverses* in matrices. The matrices that you shall encounter will be very complex, but the core principle behind the matrix multiplication remains the same.

Permutation Matrices

Recall, that the matrix equation $AX = B$ is the set of n equations with n variables. The order in which these equations appear should have no impact on the final solution. This implies, that rows can come in any permutation, the final solution remains unchanged. The matrix P whose effect when multiplied on A is to reorder the rows of A is called the *permutation matrix*.

In a $3D$ space, there are 6 permutation matrices, as shown below.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

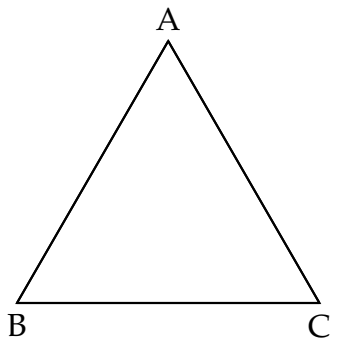
Note: The elements of the above set of permutation matrices are representing some sort of *action*. The effect of the *action* is to reorder the rows of the matrix on which the permutation matrix is operated on. To operate any permutation matrix P on A, you have to left-multiply P on A.

1.1.2 Idea 2: Symmetries of a Triangle

Imagine we have a shape of a triangle. The question is - *what kind of operations can I perform on this shape?* Clearly, this is a shape. It's not a number on which we can perform addition or multiplication or any such operation which we would define for a number. The only two actions that are possible to perform on this shape are: *rotation* and *reflection*. You can rotate this triangle by some angle or you can flip it.

But, *why are we performing these operations? What's our agenda?* The agenda is we want to learn about the symmetries of triangle. Understanding symmetries involves exploring different ways a triangle can be transformed while preserving its appearance. We will examine how a triangle can be rotated or flipped to match its original position.

Consider the triangle ABC as shown below.



Rotations

Rotations involve turning the triangle around its center. The triangle can be rotated by angles that are multiples of $\frac{360^\circ}{3} = 120^\circ$.

- **Rotation by** 0°**:** This is the identity rotation, where the triangle remains unchanged. This is represented by:

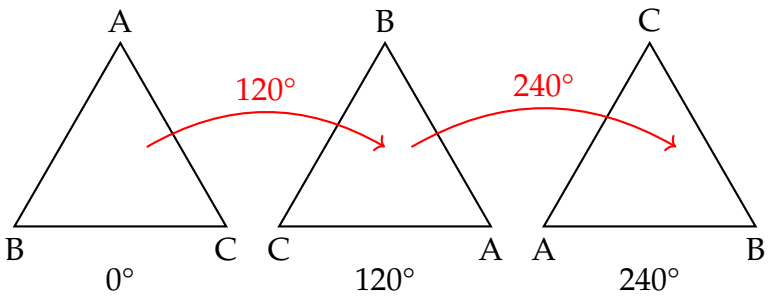
$$\text{Identity: Triangle} \to \text{Triangle}$$

- **Rotation by** 120°**:** Rotating the triangle by 120° counterclockwise about its center maps each vertex to the position of the next vertex in a counterclockwise direction. We denote this rotation as R_{120°. In a triangle with vertices A, B, and C, the effect is:

$$R_{120^\circ} : A \to B,\ B \to C,\ C \to A$$

- **Rotation by** 240°**:** This rotation maps each vertex to the position of the previous vertex in a counterclockwise direction. We denote this rotation as R_{240°. For vertices A, B, and C, the effect is:

$$R_{240^\circ} : A \to C,\ B \to A,\ C \to B$$



Reflections

Reflections involve flipping the triangle over a line of symmetry. A triangle has three lines of symmetry, each passing through a vertex and the midpoint of the opposite side.

- **Reflection over the line through** A **and the midpoint of** BC**:** This reflection maps the triangle to itself by flipping it over the line passing through vertex A and the midpoint of the side BC. We denote this reflection as F_A:

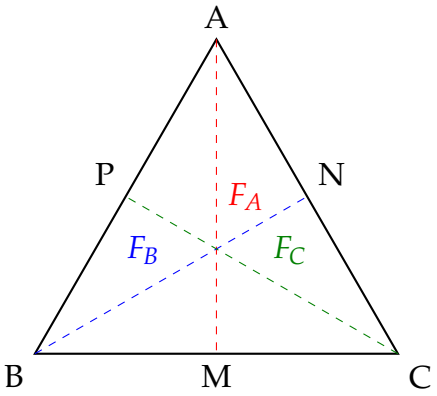
$$F_A : B \leftrightarrow C,\ A \text{ remains fixed}$$

- **Reflection over the line through** B **and the midpoint of** AC**:** This reflection maps the triangle to itself by flipping it over the line passing through vertex B and the midpoint of the side AC. We denote this reflection as F_B:

$$F_B : A \leftrightarrow C,\ B \text{ remains fixed}$$

- **Reflection over the line through** C **and the midpoint of** AB**:** This reflection maps the triangle to itself by flipping it over the line passing through vertex C and the midpoint of the side AB. We denote this reflection as F_C:

$$F_C : A \leftrightarrow B,\ C \text{ remains fixed}$$



Now, it's evident that three rotations in sequence or two flips along the same line of symmetry cancel out as they bring the original configuration of the triangle back. *How many different configurations are possible?* **Six!** The six configurations of the triangle are: $\{ABC, ACB, BAC, BCA, CAB, CBA\}$. Honestly, we don't care much about the configuration of the triangle but rather the operations that are performed on it that change its configuration!

Let r be defined as the operation that rotates the triangle from current configuration by 120°, and f be the operation that flips the triangle along the vertical line of symmetry. When the triangle is not operated by any operation and is in its original state, we represent that by 1.

Let's say we start with the configuration ABC.

A rotation r applied once will result in the configuration CAB. A subsequent rotation, denoted as r^2 from the original configuration, will result in the configuration BCA. A subsequent rotation, denoted by r^3 will bring back the original configuration ABC back, i.e. $r^3 = 1$. If we continue to rotate, we will get the same configurations back because here we are following sequential order of ABC. To reach the other configurations, we must flip.

If we perform f operation, ABC becomes ACB. It means, operation fr will give the configuration BAC and fr^2 will give the configuration CBA. One interesting thing to note is, from the initial configuration, rf operation gives the same effect as fr^2 operation.

Hence, for a triangle, there are 6 operations that do change the configuration of the triangle, but preserve the symmetry of the shape. These can be listed as $\{1, f, fr, r^2, r, rf\}$ where $r^3 = f^2 = 1$ and $rf = fr^2$.

Note: The elements of this set also represent some sort of *action*. The effect of *action* here is to change the configuration of the triangle but maintaining its symmetry. *Why do we care about symmetries?* Operating r on triangle might not have given you enough satisfaction but imagine trying to do the same on a square! Recall a *Rubik's cube*. *Don't you want to find mathematically - irrespective of what the current configuration of the Rubik's cube is, what's the maximum number of steps required in the worst case, to bring it back to the original configuration?* The idea to find an optimal solution for a *Rubik's cube* stems from here.

1.1.3 Idea 3: Permuting any Three Objects

For n objects, the number of permutations are n!. So for our 3-object example, the number of permutations $= 3! = 6$.

Let's connect the dots now. We are interested in *actions* and not the *objects* themselves. *Why?* Because actions allow us to think abstractly. We no longer care about on what we are applying our actions on.

Think it in this way: *numbers* are kind of actions too - the action of counting! It doesn't matter what you count. If something counts 3, we can discuss about its quantity without even worrying about what objects are we talking about.

When we came up with the idea that matrix multiplication is an action where the left matrix is the operator and the right matrix is the operator, we can now discuss what impact the operator matrix will have in general without worrying about on what matrix it is operating. The impact of *permutation matrix* was nothing on the system of equations - it just reorders the rows. Now we don't care about the unknowns - if they are x, y, z or $l, m, n, ...$ or $x_1, x_2, x_3, ...$ We know that the impact of permutation matrix on the system is nothing!

Similarly, in the symmetries of triangle example, we don't care what triangle is in consideration. Is it ABC or XYZ? Be it any - we can now discuss the symmetries of triangle independently. Just like *permutation matrix* had no impact on the system of equations, symmetry actions didn't had the impact on the triangle too!

Lastly, don't think of permutations as rearrangements of items themselves, but *actions* that are performing ordering. {3, 1, 2} is an action in the set of actions - {{1, 2, 3}, {1, 3, 2}, {2, 1, 3}, {2, 3, 1}, {3, 1, 2}, {3, 2, 1}}, which starts by taking the third element, then takes the first element and then takes the second element lastly. What are these elements? We literally don't care!

We can now see how everything is related:

- 3 unknowns $\rightarrow$ 6 permutation matrices (with 1 identity element, I)
- 3-sided triangle $\rightarrow$ 6 symmetry operations (with 1 identity element, 1)
- 3 item permutations $\rightarrow$ 6 arrangement actions (with 1 identity element, {1, 2, 3})

Let's formally ***group*** these ideas. *Pun intended!*

1.1.4 The concept of a Group

Everything starts from the **set**.

Definition 1.1. *A set is a collection of distinct elements.*

In the above examples, we saw three different sets of size six.

Now we want the elements of set to interact with each other. For that, we define a **binary operation**.

Definition 1.2. *A binary operation on a set is a rule that combines any two elements of the set to form another element in the set.*

This idea is a bit tricky to understand. Consider the initial example of permutation matrices. If you take any two permutation matrices and apply in sequence on a 3×3 matrix, you can find another single permutation matrix from the set which has the same effect on that 3×3 matrix. Similarly, if you select any two operations on triangle that preserve symmetry and apply them in sequence, you can find another single operation from the set that takes you to the same resultant configuration. The same thing could be said about permuting three objects.

Formally, if e_1 and e_2 are any two elements of the set S, we define a binary operation $*$ such that $e_1 * e_2 = e_3$ where $e_3 \in S$.

When we have defined the set of elements S and a binary operation $*$, we get an **algebraic structure**.

Definition 1.3. *An algebraic structure is a set S together with a binary operation $*$ defined on it. It is often written as $(S, *)$.*

Mathematicians have created a hierarchy of the algebraic structures depending upon the properties the elements hold under the binary operation $*$. These properties are very intuitive and one must ask these before studying algebraic structures in detail.

Associativity: *If I pick three elements in sequence, do picking the beginning two or ending two first make any difference on the final result?* If no, we call the algebraic structure as **semigroup**.

Definition 1.4. *A semigroup is a set S with an associative binary operation $*$. This means $(a * b) * c = a * (b * c)$ for all $a, b, c \in S$.*

Existence of Identity Element: *Is there an element in the set which has no impact when operated on other elements?* If yes, we call the algebraic structure as **monoid**.

Definition 1.5. *A monoid is a semigroup with an identity element e, such that $a * e = e * a = a$ for all $a \in S$.*

Existence of Inverse Element: *If two elements operated in sequence nullify their effect, they are called inverses of each other. Do all elements have their inverses in the set?* If yes, we call the algebraic structure as **group**.

Definition 1.6. *A group is a monoid where every element has an inverse. For every element $a \in S$, there exists an element $b \in S$ such that $a * b = b * a = e$, where e is the identity element.*

All three algebraic structures (permutation matrices, 3-object permutations and triangle symmetries) represent the same group structure S_3, called **symmetric group** of three distinct elements.

There is a special kind of groups called **Abelian Groups**, where operations are commutative.

Definition 1.7. *An Abelian (or commutative) group is a group where, in addition to the group properties, the operation is commutative: $a \cdot b = b \cdot a$ for all $a, b \in G$.*

1.2 Extension to Fields

The limitation with the *group* is that - it only supports one binary operation. *What if we want two?* In mathematics, the most common operations are addition and multiplication *subtarction and division are just their inverses*. So, we need support for two operations - addition (+) and multiplication ($\cdot$).

This extension brings another set of rules and definitions. We will look at the examples shortly. For now, just understand the terminology.

Definition 1.8. *A ring $(R, +, \cdot)$ is a set equipped with two binary operations: addition (+) and multiplication ($\cdot$). A ring must satisfy the following:*

- ***Additive Group**: $(R, +)$ is an abelian group.*
- ***Multiplicative Closure**: For all $a, b \in R$, $a \cdot b \in R$.*
- ***Distributive Property**: Multiplication is distributive over addition, i.e., for all $a, b, c \in R$:*

  $$a \cdot (b + c) = a \cdot b + a \cdot c$$

  and

  $$(a + b) \cdot c = a \cdot c + b \cdot c.$$

A commutative ring additionally requires that multiplication be commutative: $a \cdot b = b \cdot a$.

Definition 1.9. *A ring with unity (or unital ring) is a ring that has a multiplicative identity element $1 \in R$, such that for all $a \in R$:*

$$1 \cdot a = a \cdot 1 = a.$$

Definition 1.10. *A field $(F, +, \cdot)$ is a commutative ring with unity in which every non-zero element has a multiplicative inverse. A field satisfies:*

- *$(F, +)$ is an abelian group.*
- *$(F \setminus \{0\}, \cdot)$ is an abelian group (with respect to multiplication).*
- *Distributivity holds: $a \cdot (b + c) = a \cdot b + a \cdot c$ for all $a, b, c \in F$.*

Fields have two operations (addition and multiplication), and they allow for division (except by zero).

There is a good news for you - we are done with the abstract definitions for now! You must be wondering - this book was supposed to be on probability, then *why are we studying these concepts?* We are creating a rigorous framework for the concepts in probability. Slowly, we will extend the idea to the *set theory* - where each element is a subset of the sample space S and the binary operations will be union of those sets, their intersection and so on. The field then has a special name called **σ-algebra**. We will get to these ideas, but before that let's strengthen few more concepts from the *real analysis*.

1.3 Understanding Cardinality and Countability

[2]
One theorem, discovered in 500 BC, had lasting impact on the branch of mathematics for the next 2000 years! It was none other than the discovery of $\sqrt{2}$ by Pythagoras.

In 500 BC, Greeks had understanding of the relationship between the geometric length and arithmetic numbers. It was known back then that - given any two line segments AB and CD, we can represent CD as some fractional multiple of AB. At that time, Pythagoras discovered the length of hypotenuse of a right isosceles triangle with sides 1 unit long to be $\sqrt{2}$ units. Greeks for centuries couldn't understand the notion of a *number* because this length cannot be represented as a fractional number - as proved by Pythagoras.

Theorem 1.1. *There is no rational number whose square is 2.*

Proof. Let us assume, for the sake of contradiction, that $\sqrt{2}$ is rational. Then, we can express $\sqrt{2}$ as a fraction of two integers a and b, where a and b have no common factors other than 1 (i.e., the fraction is in its simplest form):

$$\sqrt{2} = \frac{a}{b}$$

Squaring both sides, we get:

$$2 = \frac{a^2}{b^2}$$

Multiplying both sides by b^2, we obtain:

$$2b^2 = a^2$$

This equation implies that a^2 is an even number (since it is equal to $2b^2$, which is even). Therefore, a must also be an even number (because the square of an odd number is odd).
Let $a = 2k$ for some integer k. Substituting this into the equation $2b^2 = a^2$, we get:

$$2b^2 = (2k)^2 = 4k^2$$

Dividing both sides by 2, we get:

$$b^2 = 2k^2$$

This implies that b^2 is also even, and hence b must be even.

Therefore, both a and b are even, which contradicts our assumption that a and b have no common factors other than 1 (since both being even means they are divisible by 2).

Thus, our original assumption that $\sqrt{2}$ is rational must be false. Hence, $\sqrt{2}$ is irrational. □

With our existing knowledge about the numbers, let's try to logically extend the system of numbers from the naturals to the irrationals. The most intuitive set of numbers is the set of counting numbers, i.e. the natural numbers $\mathbb{N}$.

$$\mathbb{N} = \{1, 2, 3, ...\}$$

By focusing on the set of natural numbers $\mathbb{N}$, we can handle addition without any issues. However, to introduce the concept of subtraction, we need to extend our number system to the set of integers $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$, which includes the additive identity (zero) and additive inverses. Next, we turn to multiplication and division. The number 1 serves as the multiplicative identity, but to define division, we require multiplicative inverses. This leads us to extend our system further to the rational numbers

$$\mathbb{Q} = \left\{ \frac{p}{q} \;\middle|\; p, q \in \mathbb{Z},\ q \neq 0 \right\}$$

which includes all fractions.

The properties discussed essentially define what is known as a field. Recall how $\mathbb{Q}$ is a field. A field is any set where addition and multiplication are well-defined operations that satisfy commutativity, associativity, and the distributive law: $a(b + c) = ab + ac$. Additionally, there must be an additive identity, and every element must have an additive inverse. Similarly, there must be a multiplicative identity, and multiplicative inverses must exist for all nonzero elements. Neither $\mathbb{Z}$ nor $\mathbb{N}$ is a field.

The set $\mathbb{Q}$ has a natural ordering. For any two rational numbers r and s, exactly one of the following holds: $r < s$, $r = s$, or $r > s$. This order is transitive, meaning if $r < s$ and $s < t$, then $r < t$. This allows us to visualize $\mathbb{Q}$ as arranged along a number line from left to right. Unlike $\mathbb{Z}$, there are no gaps, since between any two rational numbers $r < s$, the rational number $\frac{r+s}{2}$ lies between them, showing that $\mathbb{Q}$ is dense.

While the field properties of $\mathbb{Q}$ enable us to perform addition, subtraction, multiplication, and division, there are still limitations. By Theorem 1.1.1, not all numbers, such as square roots, can be expressed as rationals. This issue is deeper than just square roots. We can approximate irrational numbers like $\sqrt{2}$ using rational numbers (for example, $1.4142 \approx 1.999396$), but despite better approximations, we realize that there are *gaps* in $\mathbb{Q}$, such as at $\sqrt{2}$. Similar gaps exist at $\sqrt{3}$, $\sqrt{5}$, and other points. This dilemma, faced by the ancient Greeks, reveals the need for a more complete number system. Thus, we extend $\mathbb{Q}$ to the real numbers $\mathbb{R}$, creating the chain $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$.

The process of constructing $\mathbb{R}$ from $\mathbb{Q}$ is quite intricate and is explored later in this section. For now, a simplified view is that $\mathbb{R}$ is formed by filling the gaps in $\mathbb{Q}$. Whenever there is a missing value, a new irrational number is introduced and placed within the existing order of $\mathbb{Q}$. The real numbers consist of both these irrational numbers and the familiar rational numbers. But what characteristics do the irrational numbers possess?

- How do the rational and irrational numbers interrelate? Is there any symmetry between them, or can we argue that one type is more prevalent than the other?
- So far, we have seen examples of irrational numbers through square roots. As expected, other roots like $\sqrt[3]{2}$ or $\sqrt[5]{3}$ are typically irrational as well. Can all irrational numbers be represented as algebraic combinations of roots and rational numbers, or are there irrationals beyond this form?

Answering these questions is not trivial and hence we will take a step-by-step approach - developing concepts from the preliminaries.

1.3.1 Set Theory

[4]
A set is understood as a collection of distinct objects, known as elements. An object either belongs to the set or it does not, which makes the set well-defined. Sets can be described in two ways:

1. **Extensional Definition:** This involves explicitly listing all elements of the set within curly braces. For example, the set of natural numbers from 1 to 5 can be written as $A = \{1, 2, 3, 4, 5\}$.

2. **Intensional Definition:** In this method, a set is described by a property that all its elements satisfy. This is also known as set-builder notation. The set A from above can be expressed as $A = \{x \mid x \leq 5, x \in \mathbb{N}\}$. In general, a set C can be written as $C = \{x \mid P(x)\}$, where $P(x)$ represents some property.

We now introduce the concept of a subset and use it to define when two sets are equal.

Definition 1.11. *1. A set A is called a subset of another set B if every element of A is also an element of B. This is written as $A \subseteq B$, and in this case, B is called a superset of A.*

2. A is a proper subset of B (denoted $A \subset B$) if $A \subseteq B$ and there is at least one element in B that is not in A.

3. Two sets A and B are equal if $A \subseteq B$ and $B \subseteq A$, meaning both sets contain exactly the same elements.

Operations on Sets

1. Complement

Definition 1.12. *For a given set A, its complement is denoted as A^c and is defined as $\{x \mid x \notin A, x \in U\}$, where U represents the universal set containing A.*

The complement is always understood within the context of a larger set U.

2. Union and Intersection

Let I be an index set, and consider a collection of sets $\{A_i \mid i \in I\}$.

Definition 1.13. *The union of the sets $\{A_i \mid i \in I\}$ is defined as:*

$$\bigcup_{i \in I} A_i = \{x \mid x \in A_j \text{ for some } j \in I\}$$

In other words, the union consists of all elements that belong to at least one of the sets A_i.

Definition 1.14. *The intersection of the sets $\{A_i \mid i \in I\}$ is defined as:*

$$\bigcap_{i\in I} A_i = \{x \mid x \in A_j \text{ for every } j \in I\}$$

In simpler terms, the intersection consists of elements common to all sets A_i.

Note: When the index set I is finite, say $I = \{1, 2, 3\}$, the union matches the intuitive understanding, i.e., $\bigcup_{i=1}^{3} A_i = A_1 \cup A_2 \cup A_3$. However, this interpretation does not apply when I is infinite, such as $I = \mathbb{N}$. In this case, $\bigcup_{i=1}^{\infty} A_i$ cannot be thought of as a sequential process of unions. Instead, it should be viewed as described in Definition 1.13: the set of elements in at least one A_i, where $i \in \mathbb{N}$.

The following identities related to unions and intersections can be easily derived:

$$\bigcap_{i\in I}(A_i \cup B) = \bigcap_{i\in I} A_i \cup B \tag{1.2}$$

$$\bigcup_{i\in I}(A_i \cap B) = \bigcup_{i\in I} A_i \cap B \tag{1.3}$$

De Morgan's Laws, which explain the relationship between unions, intersections, and complements, are particularly useful:

Theorem 1.2.

$$\left(\bigcap_{i\in I} A_i\right)^c = \bigcup_{i\in I} A_i^c \tag{1.4}$$

$$\left(\bigcup_{i\in I} A_i\right)^c = \bigcap_{i\in I} A_i^c \tag{1.5}$$

Proof. Let $x \in (\bigcap_{i\in I} A_i)^c$. By the definition of complement, this implies:

$$x \notin \bigcap_{i\in I} A_i$$

which means that there exists some $j \in I$ such that $x \notin A_j$. Therefore, $x \in A_j^c$ for some $j \in I$. This implies:

$$x \in \bigcup_{i\in I} A_i^c$$

Hence, we have shown that:

$$x \in \left(\bigcap_{i\in I} A_i\right)^c \implies x \in \bigcup_{i\in I} A_i^c$$

Now, for the reverse direction, assume $x \in \bigcup_{i\in I} A_i^c$. This means there exists some $j \in I$ such that $x \in A_j^c$, or equivalently, $x \notin A_j$. Therefore, $x \notin \bigcap_{i\in I} A_i$, which implies:

$$x \in \left(\bigcap_{i\in I} A_i\right)^c$$

Thus, we have proven that:

$$x \in \left(\bigcap_{i\in I} A_i\right)^c \iff x \in \bigcup_{i\in I} A_i^c$$

This completes the proof of the first law.

Let $x \in (\bigcup_{i\in I} A_i)^c$. By the definition of complement, this implies:

$$x \notin \bigcup_{i\in I} A_i$$

which means that $x \notin A_i$ for every $i \in I$. Therefore, $x \in A_i^c$ for all $i \in I$. This implies:

$$x \in \bigcap_{i\in I} A_i^c$$

Hence, we have shown that:

$$x \in \left(\bigcup_{i\in I} A_i\right)^c \implies x \in \bigcap_{i\in I} A_i^c$$

Now, for the reverse direction, assume $x \in \bigcap_{i\in I} A_i^c$. This means $x \in A_i^c$ for every $i \in I$, or equivalently, $x \notin A_i$ for all $i \in I$. Therefore, $x \notin \bigcup_{i\in I} A_i$, which implies:

$$x \in \left(\bigcup_{i\in I} A_i\right)^c$$

Thus, we have proven that:

$$x \in \left(\bigcup_{i\in I} A_i\right)^c \iff x \in \bigcap_{i\in I} A_i^c$$

This completes the proof of the second law. □

3. Relative Complement

Definition 1.15. *The relative complement of B in A is defined as:*

$$A \setminus B = \{x \mid x \in A, x \notin B\} = A \cap B^c$$

Similarly, the relative complement of A in B is $B \setminus A = \{x \mid x \in B, x \notin A\} = B \cap A^c$.

4. Cartesian Product

A Cartesian product constructs a set from multiple sets by pairing their elements.

Definition 1.16. *The Cartesian product of two sets A and B is defined as:*

$$A \times B = \{(x, y) \mid x \in A, y \in B\}$$

This represents the set of all ordered pairs where the first element comes from A and the second from B. For instance, if $A = \{1, 2\}$ and $B = \{a\}$, then $A \times B = \{(1, a), (2, a)\}$ and $B \times A = \{(a, 1), (a, 2)\}$. Clearly, the Cartesian product is not commutative.

For n sets $A_1, A_2, \ldots, A_n$, their Cartesian product is:

$$A_1 \times A_2 \times \cdots \times A_n = \{(a_1, a_2, \ldots, a_n) \mid a_i \in A_i\}$$

If all the sets are identical, then we have:

$$A^n = \{(a_1, a_2, \ldots, a_n) \mid a_i \in A\}$$

5. Power Set

Definition 1.17. *The power set of a set A, denoted by $\mathcal{P}(A)$ or 2^A, is the set of all subsets of A, including the empty set and A itself.*

For example, if $A = \{1, 2\}$, then:

$$\mathcal{P}(A) = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$$

1.3.2 Functions

Now that we know enough about sets, it's time to create mappings between two sets using *functions*.

Definition 1.18. *A function f from a set A to a set B is a subset of the Cartesian product $(A \times B)$ such that each element in A appears as the first component in exactly one ordered pair in the subset. Essentially, this means that every element in set A is mapped to a unique element in set B, often denoted as $f : A \to B$. The set A is called the domain, and set B is the codomain.*

If an element $b \in B$ is associated with an element $a \in A$, b is called the image of a, while a is known as the argument or pre-image of b. In this case, we say that f maps a to b and write it as $b = f(a)$. The range of a function is the set of all images of elements in the domain, which forms a subset (not necessarily proper) of the codomain.

Functions are classified into:

1. **Injective (One-to-One):** A function is injective if $a \neq b \implies f(a) \neq f(b)$ for all $a, b \in \text{domain}(f)$. For example, the function $f : \mathbb{N} \to \mathbb{R}$ defined by $f(x) = x$ for all $x \in \mathbb{N}$ is injective.

2. **Serjective (Onto):** A function is surjective if for every $b \in \text{codomain}(f)$, there exists an $a \in \text{domain}(f)$ such that $f(a) = b$.

Examples of surjective functions include:

- Let $A = \{1, 2, 3\}$ and $B = \{0, 1\}$. The function $g : A \to B$ defined by $g(1) = 0$, $g(2) = 0$, and $g(3) = 1$ is surjective.

- The function $h : \mathbb{R} \to \mathbb{R}$ defined by $h(x) = x + 1$ for all $x \in \mathbb{R}$ is also surjective.

A function that is both injective and surjective is called *bijective*. The function h mentioned above is also bijective. In a bijective function, an *inverse function* can be defined as the mapping is unique and covers the entire codomain.

1.3.3 The Axiom of Completeness

What exactly is a *real number?* We got as far as saying that the set $\mathbb{R}$ of real numbers is an extension of the rational numbers $\mathbb{Q}$ in which there are no holes or gaps. We are going to improve this definition.
Let S be a non-empty subset of real numbers $\mathbb{R}$.

Definition 1.19. *The infimum (inf) of S, denoted as* $\inf S$*, is the greatest lower bound of S. That is,* $\inf S = \alpha$ *if:*

- $\alpha \leq x$ *for all* $x \in S$ *(i.e.,* α *is a lower bound of S), and*
- *for every* $\epsilon > 0$*, there exists some* $x \in S$ *such that* $\alpha + \epsilon > x$ *(i.e.,* α *is the greatest such lower bound).*

Similarly,

Definition 1.20. *The supremum (sup) of S, denoted as* $\sup S$*, is the least upper bound of S. That is,* $\sup S = \beta$ *if:*

- $\beta \geq x$ *for all* $x \in S$ *(i.e.,* β *is an upper bound of S), and*
- *for every* $\epsilon > 0$*, there exists some* $x \in S$ *such that* $\beta - \epsilon < x$ *(i.e.,* β *is the least such upper bound).*

Definition 1.21. *The Axiom of Completeness states that every non-empty subset of real numbers that is bounded above has a supremum in* $\mathbb{R}$*. Similarly, every non-empty subset of real numbers that is bounded below has an infimum in* $\mathbb{R}$*.*

In simpler terms, this axiom guarantees that in the set of real numbers, there are no "gaps" — every set that is bounded from above or below has a least upper bound or a greatest lower bound, respectively. But these are different than the *minimum and maximum*. Consider the following example for clarity.

Example 1.1. *Consider the set* $S = (0, 1)$*, the open interval between 0 and 1.*

- *The infimum of S is* $\inf S = 0$*, as 0 is the greatest number less than or equal to all elements of S. However, since 0 is not an element of S, it is* ***not*** *the minimum.*
- *The supremum of S is* $\sup S = 1$*, as 1 is the least number greater than or equal to all elements of S. However, since 1 is not an element of S, it is* ***not*** *the maximum.*

Now consider the set $T = [0, 1]$*, the closed interval between 0 and 1.*

- *The minimum of T is 0, as 0 is the smallest element in T.*
- *The maximum of T is 1, as 1 is the largest element in T.*

Thus, for an open interval like $(0, 1)$*, the infimum and supremum exist but the minimum and maximum do not. In contrast, for a closed interval like* $[0, 1]$*, the infimum is equal to the minimum, and the supremum is equal to the maximum.*

Because of this axiom, we can say that *the real line contains no gaps*.

Theorem 1.3. *The real line contains no gaps.*

Proof. Let $I_n = [a_n, b_n]$ be a sequence of closed intervals such that:

1. Each interval I_n is nested:

$$I_{n+1} \subseteq I_n \quad \text{for all } n \in \mathbb{N}$$

2. The lengths of the intervals tend to zero:

$$b_n - a_n \to 0 \quad \text{as } n \to \infty$$

Let x be a point in the intersection $\bigcap_{n=1}^{\infty} I_n$. Since x lies in every I_n, we have:

$$a_n \leq x \leq b_n \quad \text{for all } n \in \mathbb{N}$$

By the Axiom of Completeness, every bounded set of real numbers has a least upper bound (supremum) and a greatest lower bound (infimum). The sequences (a_n) and (b_n) are bounded, and thus we define:

$$L = \liminf_{n\to\infty} a_n \quad \text{and} \quad U = \limsup_{n\to\infty} b_n$$

Since $b_n - a_n \to 0$, we have:

$$U - L = \lim_{n\to\infty} (b_n - a_n) = 0$$

Thus, $L = U$. Therefore, there exists a unique limit c such that:

$$c = L = U$$

This means that every nested sequence of closed intervals must converge to a point in $\mathbb{R}$, implying that there are no gaps in the real line.

□

If there are no gaps in the real line, it also implies that there exists a real number whose square is 2.

Theorem 1.4. *There exists a real number whose square is 2.*

Proof. Consider the set

$$S = \{x \in \mathbb{R} \mid x^2 < 2\}.$$

By the Axiom of Completeness, since S is non-empty and bounded above, there exists a least upper bound $c = \sup S$.
Show $c^2 \leq 2$: Suppose $c^2 > 2$. Then, there exists some $\epsilon > 0$ such that $c^2 = 2 + \epsilon$. Since c is the least upper bound of S, we can find $x \in S$ such that $c > x$. This implies $x^2 < 2$. We can choose x close enough to c such that

$$x^2 > c^2 - \epsilon = 2,$$

contradicting $x \in S$. Thus, we conclude $c^2 \leq 2$.

Show $c^2 \geq 2$: Suppose $c^2 < 2$. Then there exists some $\epsilon > 0$ such that $c^2 = 2 - \epsilon$. Because c is the least upper bound, there exists $y \in \mathbb{R}$ such that $c < y < c + \delta$ for some small $\delta > 0$. This implies

$$y^2 > c^2 = 2 - \epsilon,$$

which can be made greater than 2 for sufficiently small δ, contradicting the upper bound property of c. Thus, we conclude $c^2 \geq 2$.

Since $c^2 \leq 2$ and $c^2 \geq 2$, we have

$$c^2 = 2.$$

Thus, there exists a real number c such that $c^2 = 2$. □

1.3.4 Cardinality and Countability

In informal language, the cardinality of a set refers to the quantity of elements within that set. To compare the cardinalities of two finite sets A and B, one can simply count the elements in each set and determine whether they have the same cardinality or if one set contains more elements than the other. However, when comparing sets with infinitely many elements (for instance, $\mathbb{N}$ versus $\mathbb{Q}$), this basic method is inadequate. In the late nineteenth century, Georg Cantor proposed a more sophisticated approach for comparing cardinalities based on the types of functions that can be defined from one set to another.

Definition 1.22. *1. Two sets A and B are considered equicardinal (denoted $|A| = |B|$) if there exists a bijective function from A to B.*

2. *Set B has cardinality greater than or equal to that of A (denoted $|B| \geq |A|$) if there is an injective function from A to B.*

3. *Set B has cardinality strictly greater than that of A (denoted $|B| > |A|$) if there exists an injective function but no bijective function from A to B.*

Based on these definitions, the concept of countability of a set is defined as follows:

Definition 1.23. *A set E is said to be countably infinite if it is equicardinal with $\mathbb{N}$. A set is classified as **countable** if it is either finite or countably infinite.*

Example 1.2. *The set of even numbers is equicardinal to $\mathbb{N}$.*

Let $E = \{2n \mid n \in \mathbb{N}\}$ be the set of even numbers. We will construct a bijection $f : \mathbb{N} \to E$ defined by

$$f(n) = 2n.$$

This function is both injective and surjective:

- *Injective: If $f(n_1) = f(n_2)$, then $2n_1 = 2n_2$ implies $n_1 = n_2$.*
- *Surjective: For every $m \in E$, there exists $n \in \mathbb{N}$ such that $m = 2n$.*

Thus, E is equicardinal to $\mathbb{N}$.

It is certainly true that E is a proper subset of $\mathbb{N}$, and for this reason it may seem logical to say that E is a smaller set than $\mathbb{N}$. This is one way to look at it, but it represents a point of view that is heavily biased from an overexposure to finite sets.

Example 1.3. *The set of integers is equicardinal to* $\mathbb{N}$.

Let $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. *We will construct a bijection* $g : \mathbb{N} \to \mathbb{Z}$ *defined by*

$$g(n) = \begin{cases} \frac{n}{2}, & \text{if } n \text{ is even} \\ -\frac{n+1}{2}, & \text{if } n \text{ is odd.} \end{cases}$$

This function is both injective and surjective:

- *Injective: If* $g(n_1) = g(n_2)$, *the cases show that* $n_1 = n_2$.
- *Surjective: For any* $m \in \mathbb{Z}$, *there exists* $n \in \mathbb{N}$ *such that* $g(n) = m$.

Thus, $\mathbb{Z}$ *is equicardinal to* $\mathbb{N}$.

Example 1.4. *The set of all rationals in* $[0, 1]$ ***is countable.***

To show that the set of all rational numbers in the interval $[0, 1]$ *is countable, we consider rational numbers of the form* $\frac{p}{q}$, *where* $q \neq 0$.

We start by incrementing q *in steps of 1, beginning with* $q = 1$. *For each integer* $q \geq 1$, *we consider all integers* p *such that* $0 \leq p \leq q$. *The rational number* $\frac{p}{q}$ *is added to the set if it is not already present.*

The set of rational numbers in $[0, 1]$ *can then be explicitly listed as follows:*

$$\left\{0, 1, \frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \frac{1}{6}, \frac{5}{6}, \ldots\right\}$$

This process demonstrates that we can enumerate all rational numbers in $[0, 1]$.

Next, we can define a bijection from the set of rational numbers $\mathbb{Q} \cap [0, 1]$ *to the natural numbers* $\mathbb{N}$. *Each rational number* $\frac{p}{q}$ *is mapped to its index in the above enumeration.*

Thus, the set of all rational numbers in $[0, 1]$ *is countably infinite, and hence it is countable.*

Theorem 1.5. *Let* I *be a countable index set, and let* E_i *be countable for each* $i \in I$. *Then* $\bigcup_{i \in I} E_i$ *is countable.*

More glibly, it can also be stated as follows: *A countable union of countable sets is countable.*

Proof. Since I is a countable index set, we can enumerate it as $I = \{i_1, i_2, i_3, \ldots\}$. For each i_j, since E_{i_j} is countable, we can enumerate the elements of E_{i_j} as follows:

$$E_{i_j} = \{e_{j,1}, e_{j,2}, e_{j,3}, \ldots\}$$

for $j = 1, 2, 3, \ldots$.

Now, we can construct a new set S by listing the elements of the E_i sets in a systematic way. We can arrange them in a two-dimensional array:

	$e_{1,1}$	$e_{1,2}$	$e_{1,3}$	$\cdots$
E_{i_1}	$e_{2,1}$	$e_{2,2}$	$e_{2,3}$	$\cdots$
E_{i_2}	$e_{3,1}$	$e_{3,2}$	$e_{3,3}$	$\cdots$
E_{i_3}	$\vdots$	$\vdots$	$\vdots$	$\ddots$

We can then define a function $f : \mathbb{N} \to \bigcup_{i \in I} E_i$ that maps each natural number to a unique element of the union. For instance, we can use the following enumeration method: $f(1) = e_{1,1}$, $f(2) = e_{1,2}$, $f(3) = e_{2,1}$, $f(4) = e_{1,3}$, $f(5) = e_{2,2}$, $f(6) = e_{3,1}$, and so on.

This enumeration process ensures that every element in the union $\bigcup_{i \in I} E_i$ will eventually be listed. Therefore, we can conclude that $\bigcup_{i \in I} E_i$ is countable.

□

Example 1.5. *The set of all Rational numbers, $\mathbb{Q}$ is countable.*

*We will now use theorem **1.5** to prove the countability of the set of all rational numbers.*

It has been already proved that the set $\mathbb{Q} \cap [0, 1]$ is countable. Similarly, it can be shown that $\mathbb{Q} \cap [n, n + 1]$ is countable, $\forall n \in \mathbb{Z}$. Let $Q_i = \mathbb{Q} \cap [i, i + 1]$. Thus, clearly, the set of all rational numbers, $\mathbb{Q} = \bigcup_{i \in \mathbb{Z}} Q_i$ – a countable union of countable sets – is countable.

Definition 1.24. *A set F is uncountable if it has cardinality strictly greater than the cardinality of $\mathbb{N}$.*

Theorem 1.6. *The set of all infinite binary strings, $\{0, 1\}^\infty$, is uncountable.*

Proof. Assume for the sake of contradiction that the set of all binary strings, $A = \{0, 1\}^\infty$, is countably infinite. Thus, there exists a bijection $f : A \to \mathbb{N}$. In other words, we can order the set of all infinite binary strings as follows:

$$\begin{array}{cccc} a_{11} & a_{12} & a_{13} & \dots \\ a_{21} & a_{22} & a_{23} & \dots \\ a_{31} & a_{32} & a_{33} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

where a_{ij} is the jth bit of the ith binary string, $i, j \geq 1$.

Consider the infinite binary string given by $\bar{a} = \bar{a}_{11}\bar{a}_{22}\bar{a}_{33}\ldots$, where $\bar{a}_{ij}$ is the complement of the bit a_{ij}.

Since our list contains all infinite binary strings, there must exist some $k \in \mathbb{N}$ such that the string $\bar{a}$ occurs at the kth position in the list, i.e., $f(\bar{a}) = k$. The kth bit of this specific string is $\bar{a}_{kk}$. However, from the above list, we know that the kth bit of the kth string is a_{kk}. Thus, we can conclude that the string $\bar{a}$ cannot occur in any position $k \geq 1$ in

our list, contradicting our initial assumption that our list exhausts all possible infinite binary strings.

Thus, there cannot possibly exist a bijection from $\mathbb{N}$ to $\{0,1\}^\infty$, proving that $\{0,1\}^\infty$ is uncountable.

□

Corollary 1.1. *The sets* $[0,1]$, $\mathbb{R}$ *and* $\{\mathbb{R} \setminus \mathbb{Q}\}$ *are uncountable.*

Proof. Firstly, consider the set $[0,1]$. Any number in this set can be expressed by its binary equivalent, which suggests a bijection from $[0,1]$ to $\{0,1\}^\infty$. However, this is not exactly a bijection due to the issue with the dyadic rationals (i.e., numbers of the form $\frac{a}{2^b}$, where a and b are natural numbers, and a is odd). For example, $0.01000\ldots$ in binary is the same as $0.001111\ldots$.

To address this, we can tweak this "near bijection" to produce an explicit bijection as follows. For any infinite binary string $x = (x_1, x_2, \ldots) \in \{0,1\}^\infty$, let

$$g(x) = \sum_{k=1}^{\infty} x_k 2^{-k}.$$

The function g maps $\{0,1\}^\infty$ "almost bijectively" to $[0,1]$, but the dyadic rationals have two pre-images. For instance, we have $g(1000\ldots) = g(0111\ldots) = \frac{1}{2}$.

To resolve this, let the set of dyadic rationals be given by the list

$$D = \left\{ d_1 = \frac{1}{2}, d_2 = \frac{1}{4}, d_3 = \frac{3}{4}, d_4 = \frac{1}{8}, d_5 = \frac{3}{8}, d_6 = \frac{5}{8}, d_7 = \frac{7}{8}, \ldots \right\}.$$

Note that the dyadic rationals can be enumerated as they are countable.

Next, we define the following bijection $f(x)$ from $\{0,1\}^\infty$ to $[0,1]$:

$$f(x) = \begin{cases} g(x) & \text{if } g(x) \notin D, \\ d_{2n-1} & \text{if } g(x) = d_n \text{ for some } n \in \mathbb{N} \text{ and } x_k \text{ terminates in 1}, \\ d_{2n} & \text{if } g(x) = d_n \text{ for some } n \in \mathbb{N} \text{ and } x_k \text{ terminates in 0}. \end{cases}$$

This defines an explicit bijection from $\{0,1\}^\infty$ to $[0,1]$, proving that the set $[0,1]$ is uncountable.

Next, we can define a bijection from $(0,1)$ to $\mathbb{R}$, for example using the function $\tan\left(\pi x - \frac{\pi}{2}\right)$ for $x \in (0,1)$. Thus, the set of all real numbers, $\mathbb{R}$, is uncountable.

Finally, we write $\mathbb{R} = \mathbb{Q} \cup (\mathbb{R} \setminus \mathbb{Q})$. Since $\mathbb{Q}$ is countable and $\mathbb{R}$ is uncountable, we conclude that $\mathbb{R} \setminus \mathbb{Q}$, the set of all irrational numbers, is also uncountable.

□

Example 1.6. *Prove that* $\mathcal{P}(\mathbb{N})$*, the power set of the natural numbers, is uncountable.*

To do so, we can use the method of contradiction and Cantor's diagonal argument.

Let us assume that $\mathcal{P}(\mathbb{N})$ is countable. This means that we can list all the subsets of $\mathbb{N}$ as follows:

$$S_1, S_2, S_3, \ldots$$

Next, we will associate each subset S_i with an infinite binary string b_i, where the n-th digit of b_i is defined as follows:

$$b_i(n) = \begin{cases} 1 & \text{if } n \in S_i \\ 0 & \text{if } n \notin S_i \end{cases}$$

Thus, each subset S_i corresponds to a binary string that represents whether each natural number is included in the subset or not.

Now, consider the set S of all natural numbers that belong to none of the subsets listed. Specifically, we define S as follows:

$$S = \{n \in \mathbb{N} : n \notin S_n\}$$

By this definition, for each n, the n-th digit of the binary string corresponding to S will be:

$$b(n) = \begin{cases} 1 & \text{if } n \in S \\ 0 & \text{if } n \notin S \end{cases}$$

Now, we need to determine whether S is included in our original list of subsets $S_1, S_2, S_3, \ldots$.

If $S = S_k$ for some k, then by the construction of S, we have:

- *If $k \in S$, then by definition of S, $k \notin S_k$, which is a contradiction.*
- *Conversely, if $k \notin S$, then $k \in S_k$, which again leads to a contradiction.*

Since S cannot be equal to any S_k in our assumed countable list, this means S is a subset of $\mathbb{N}$ that is not included in our original enumeration of subsets.

Thus, we arrive at a contradiction. Therefore, our initial assumption that $\mathcal{P}(\mathbb{N})$ is countable must be false, and we conclude that:

$$\mathcal{P}(\mathbb{N}) \text{ is uncountable.}$$

Example 1.7. *Show that an infinite subset of a countable set is countable.*

Let S be countably infinite set. So, there exists a bijection $f : \mathbb{N} \to S$, meaning we can enumerate the elements of S as $s_1, s_2, s_3, \ldots$.

Assuming A is an infinite subset of S, we can construct a sequence of elements from A:

- *Start with the first element $a_1 \in A$.*

- *Find the next element $a_2 \in A$ such that a_2 is greater than a_1 in the enumeration of S.*
- *Continue this process to find $a_3, a_4, \ldots$, ensuring each a_n is greater than a_{n-1}.*

Since A is infinite, this process will yield an infinite sequence $a_1, a_2, a_3, \ldots$ where each a_n is an element of A.

We have thus constructed a function $g : \mathbb{N} \to A$ that enumerates the elements of A, showing that A is countable.

Chapter 2

Classical Probability

[3]
Whatever we learnt in the previous chapter, will be useful in the next chapter. In this chapter, I want to recap all the basic concepts of probability we have studied previously and at the end demonstrate - why we need a better, a more rigorous framework, for the probability theory.

Just as a point is not defined in elementary geometry, probability theory begins with two entities that are not defined. These undefined entities are a *Random Experiment* and its *Outcome*. These two concepts are to be understood intuitively, as suggested by their respective English meanings. We use these undefined terms to define other entities.

So, let's begin!

2.1 Probability and Counting

The first concept in the probability theory is the *sample space* and the *event*.

Definition 2.1. *The sample space, often denoted by S, is the set of all possible outcomes of a random experiment. Formally, if an experiment can result in one of n distinct outcomes, the sample space is the set containing all these outcomes.*

$$S = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

Example: For a single coin flip, the sample space is:

$$S = \{\text{Heads}, \text{Tails}\}$$

Definition 2.2. *An event is a subset of the sample space. It represents one or more outcomes that may occur as a result of the experiment. Formally, if A is an event, then $A \subseteq S$.*

Example: If we roll a die, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. An event could be rolling an even number:

$$A = \{2, 4, 6\}$$

An event A occurs if the outcome of the experiment is one of the elements of the subset A. Mathematically, this means that if the outcome of the experiment is ω, we say the

event A has occurred if $\omega \in A$.

Example: If we roll a die and the outcome is 4, the event $A = \{2, 4, 6\}$ (rolling an even number) has occurred because $4 \in A$.

The probability of an event occuring is given by the *naive definition of the probability.* The definition is called *naive* because it has some gaps in it, but it works fairly well in many cases and only errs when dealing with *infinitely big sample spaces*. As in real world, most of the experiments have finite number of outcomes - we can progress our theory for now and fix it later.

Definition 2.3. *If all outcomes in the sample space S are equally likely, the probability of an event A, denoted by $P(A)$, is given by:*

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

where $|A|$ is the number of outcomes in the event A, and $|S|$ is the total number of outcomes in the sample space S.

Example: In a fair six-sided die roll, the probability of rolling an even number (event $A = \{2, 4, 6\}$) is:

$$P(A) = \frac{|A|}{|S|} = \frac{3}{6} = \frac{1}{2}$$

Spoiler Alert! What happens when the outcomes are not exactly likely? You can imagine how flawed the argument sounds when I ask *What's the probability of life on Mars?* and someone replies that there are two possibilities - either there is a life or not; hence the probability of life on Mars is 50%. So now you are already seeing why this definition is called *naive*. The theory that fixes this or improves our understanding of probability is the *Bayesian Theory!* There is a lot of prerequisites that we have to cover before we touch that!

2.1.1 Counting

Definition 2.4. *If an event can occur in n_1 ways and another mutually exclusive event can occur in n_2 ways, the total number of ways either of the two events can occur is:*

$$n_1 + n_2$$

This rule applies when we are selecting from two or more disjoint sets.

Example: If you can choose a red shirt in 3 ways and a blue shirt in 5 ways, and you can only choose one shirt, the total number of choices is:

$$3 + 5 = 8$$

Definition 2.5. *If a procedure can be broken down into two stages, where the first stage can occur in n_1 ways and for each of these the second stage can occur in n_2 ways, the total number of ways to perform the procedure is:*

$$n_1 \times n_2$$

This rule applies when events are independent and occur sequentially.

Example: If you have 3 shirts and 2 pants to choose from, the total number of ways to select an outfit is:

$$3 \times 2 = 6$$

Definition 2.6. *A permutation is an arrangement of k objects from n distinct objects where the order of selection matters. The number of permutations is given by:*

$$P(n,k) = \frac{n!}{(n-k)!}$$

Example: The number of ways to arrange 3 students out of 5 is:

$$P(5,3) = \frac{5!}{(5-3)!} = 5 \times 4 \times 3 = 60$$

Definition 2.7. *A combination is a selection of k objects from n distinct objects where the order of selection does not matter. The number of combinations is given by:*

$$C(n,k) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Example: The number of ways to choose 3 students out of 5 is:

$$C(5,3) = \binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5 \times 4 \times 3}{3 \times 2 \times 1} = 10$$

Sampling Table

We can summarize the different ways of selecting items using the following 2x2 table, which distinguishes between sampling with and without replacement, and whether the order of selection matters or not.

Selection Criteria	**With Replacement**	**Without Replacement**
Order Matters	n^k	$\frac{n!}{(n-k)!}$
Order Doesn't Matter	$\binom{n+k-1}{k}$	$\binom{n}{k}$

Scenario: Imagine you're visiting a candy store with 5 different types of candies. You're interested in selecting some candies, and we will explore four different ways you might select these candies based on the two criteria:

- whether you put back the candy after picking (with or without replacement)
- whether the order in which you pick the candies matters

With Replacement, Order Matters

Let's say you want to select 3 candies one by one, and every time you pick a candy, you put it back into the box before picking the next one. Also, you care about the order in which the candies are picked (perhaps because you are arranging them in a special pattern).

Story Proof: The first candy you pick can be any of the 5 types. Since you put the candy back, when you pick the second candy, you again have 5 choices. For the third candy, you once again have 5 choices, because the previous candies were returned. Thus, for each of the 3 choices, there are 5 options, resulting in $5 \times 5 \times 5 = 5^3 = 125$ different possible ways to select the candies.

Without Replacement, Order Matters

Now, imagine you're selecting 3 candies, but this time after selecting a candy, you do **not** put it back into the box. You still care about the order of selection.

Story Proof: For the first candy, you have 5 choices, as all candies are available. After picking the first candy, you're left with 4 candies to choose from for the second selection. Once the second candy is selected, only 3 candies remain for the third selection. Therefore, the total number of ways to arrange these candies is $5 \times 4 \times 3 = 60$.

With Replacement, Order Doesn't Matter

Let's change the scenario. Now, you are still selecting 3 candies, but every time you pick a candy, you put it back into the box, and this time, you don't care about the order in which you pick them.

Story Proof: You are choosing 3 candies, and since you put the candy back after each selection, it's possible to select the same candy more than once. However, because the order doesn't matter, we don't treat "Red, Green, Blue" as different from "Blue, Green, Red." To count how many different groups of candies you can pick, we think of it as a combination with repetition. The formula to calculate this is $\binom{n+k-1}{k}$, where n is the number of candy types, and k is the number of candies you're choosing. In this case, there are $\binom{5+3-1}{3} = \binom{7}{3} = 35$ different combinations of candies.

Without Replacement, Order Doesn't Matter

Finally, let's consider the case where you're selecting 3 candies, but you don't put the candies back after picking, and you don't care about the order in which they are picked.

Story Proof: You are choosing 3 candies from the 5 available types, but once you pick a candy, you don't replace it. Since the order doesn't matter, we only care about which candies were chosen, not the sequence of selection. To count how many different groups of candies you can pick, we use combinations. The number of ways to choose 3 candies from 5 without caring about the order is given by $\binom{5}{3} = 10$.

There are some more story proofs!

Choosing the complement

For nonnegative integers n and k with $k \leq n$, we have:

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

Story Proof: Consider selecting a committee of size k from n people. You can either choose k people to be on the committee or equivalently choose the $n-k$ people who will *not* be on the committee. Both methods count the same outcome, so the two binomial coefficients are equal.

The Team Captain

For positive integers n and k with $k \leq n$, we have:

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

Story Proof: Suppose we are choosing a team of k people from n individuals, one of whom will be the captain. First, choose the team captain, then select the remaining $k-1$ members from the remaining $n-1$ individuals. Alternatively, select the k-member team first, then choose one of the k members to be the captain.

Vandermonde's Identity

Vandermonde's identity states:

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

Story Proof: Consider choosing a committee of k members from a group of m men and n women. If the committee has j men, then the remaining $k-j$ members must be women. The right-hand side sums over all possible values of j, showing different ways to form the committee.

Partnerships

We show:

$$\frac{(2n)!}{2^n \cdot n!} = (2n-1)(2n-3)\cdots 3 \cdot 1.$$

Story Proof: Consider breaking $2n$ people into n partnerships. First, line them up, and pair the first two, the next two, and so on. This overcounts by $n! \cdot 2^n$ because the order of pairs and the order within pairs doesn't matter. Alternatively, pick the first partner from $2n-1$, the next from $2n-3$, and so on, yielding the right-hand side.

2.1.2 Classical Problems on Probability

The Birthday Problem

What is the minimum number of people required to have a 50% chance of at least two people sharing the same birthday?

The total number of ways to assign birthdays to n people (assuming 365 possible days) is 365^n. The probability that no two people share the same birthday can be computed by considering that the first person can have any of the 365 days, the second person can have 364 remaining days, the third person 363, and so on. The probability that all n people have distinct birthdays is:

$$P(\text{no shared birthdays}) = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365 - n + 1}{365}.$$

The probability of at least one birthday clash is:

$$P(\text{at least one shared birthday}) = 1 - P(\text{no shared birthdays}).$$

We seek the smallest n such that $P(\text{at least one shared birthday}) \geq 0.5$. This gives the inequality:

$$1 - \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \geq 0.5.$$

Solving numerically, we find that $n = 23$ gives a probability slightly greater than 0.5.

With just 23 people, the probability of at least two people sharing the same birthday is about 50%. This counterintuitive result occurs because with 23 people, there are many possible pairs, and the chance of any two people sharing a birthday grows rapidly as the number of people increases.

Newton-Pepys Problem

Which event has the highest probability?

- *A: At least one 6 appears when 6 dice are rolled.*
- *B: At least two 6's appear when 12 dice are rolled.*
- *C: At least three 6's appear when 18 dice are rolled.*

The probability of getting at least k sixes when rolling n dice can be computed using the binomial distribution. For each die, the probability of rolling a 6 is $p = \frac{1}{6}$, and the probability of not rolling a 6 is $\frac{5}{6}$. The probability of getting exactly k sixes in n dice is:

$$P(\text{exactly } k \text{ sixes}) = \binom{n}{k} p^k (1-p)^{n-k}.$$

To calculate the probability of getting at least k sixes, we compute:

$$P(\text{at least } k \text{ sixes}) = 1 - P(\text{fewer than } k \text{ sixes}).$$

For each case:

- For $n = 6$, $k = 1$:

$$P(\text{at least one 6}) = 1 - \left(\frac{5}{6}\right)^6 \approx 0.6651.$$

- For $n = 12$, $k = 2$:

$$P(\text{at least two 6's}) = 1 - \sum_{i=0}^{1} \binom{12}{i} \left(\frac{1}{6}\right)^i \left(\frac{5}{6}\right)^{12-i} \approx 0.6187.$$

- For $n = 18$, $k = 3$:

$$P(\text{at least three 6's}) = 1 - \sum_{i=0}^{2} \binom{18}{i} \left(\frac{1}{6}\right)^i \left(\frac{5}{6}\right)^{18-i} \approx 0.5973.$$

De Montmort's Matching Problem

Consider a deck of n cards labeled from 1 to n. You flip over the cards one by one, saying the numbers 1 through n as you do. You win if, at some point, the number you say matches the number on the card. What is the probability of winning?

Let A_i be the event that the i-th card is in position i (i.e., a match for card i). We are interested in the probability of the union of these events:

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

which is the probability that at least one card is in its correct position. To calculate this, we will use the inclusion-exclusion principle.

The inclusion-exclusion formula for the union of n events is:

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1\le i<j\le n} P(A_i \cap A_j) + \sum_{1\le i<j<k\le n} P(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n).$$

Since the problem is symmetric, the probabilities of individual intersections of events A_i are the same for any i, and the expression simplifies considerably.

The probability that the i-th card is in the i-th position (i.e., a match) is:

$$P(A_i) = \frac{1}{n}.$$

This is because there are $n!$ possible orderings of the deck, and in $(n-1)!$ of them, card i is in position i (with the remaining $n-1$ cards arranged freely).

Next, consider the probability that two specific cards, i and j, are both in their correct positions. If both cards i and j are fixed in positions i and j, there are $(n-2)!$ favorable arrangements for the remaining $n-2$ cards, so:

$$P(A_i \cap A_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}.$$

Similarly, the probability that three specific cards, i, j, and k, are all in their correct positions is:

$$P(A_i \cap A_j \cap A_k) = \frac{(n-3)!}{n!} = \frac{1}{n(n-1)(n-2)}.$$

Thus, the inclusion-exclusion formula for the union becomes:

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \frac{1}{n} - \sum_{1\le i<j\le n} \frac{1}{n(n-1)} + \sum_{1\le i<j<k\le n} \frac{1}{n(n-1)(n-2)} - \cdots + (-1)^{n+1}\frac{1}{n!}.$$

This simplifies to:

$$P\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + \frac{(-1)^{n+1}}{n!}.$$

This is the partial sum of the Taylor series expansion of e^{-1}, which is given by:

$$e^{-1} = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots.$$

Thus, for large n, the probability of winning (i.e., having at least one match) is approximately:

$$P\left(\bigcup_{i=1}^{n} A_i\right) \approx 1 - \frac{1}{e}.$$

Since $1 - \frac{1}{e} \approx 0.6321$, the probability of winning the game is about 63.2%.

2.2 Conditional Probability

Conditional probability is a fundamental concept in probability theory that allows us to update our beliefs about an event based on new information or evidence. It helps in understanding how the probability of one event changes in the presence of another event. This concept is particularly useful in various fields such as statistics, machine learning, and decision-making, where we often deal with uncertain events and seek to refine our predictions.

Definition 2.8. *If A and B are events with $P(B) > 0$, then the conditional probability of A given B, denoted by $P(A|B)$, is defined as:*

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

In this definition, A is the event whose uncertainty we want to update, and B is the evidence we observe (or want to treat as given). We refer to $P(A)$ as the prior probability of A and $P(A|B)$ as the posterior probability of A. The term "prior" indicates our initial belief about the event before incorporating the new evidence, while "posterior" reflects our updated belief after taking the evidence into account.

Similarly, we can express the conditional probability of event B given event A as:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \text{for } P(A) > 0.$$

From these two definitions, we can express $P(A \cap B)$ in two different ways:

$$P(A \cap B) = P(A|B) \cdot P(B)$$

$$P(A \cap B) = P(B|A) \cdot P(A)$$

Since both expressions represent the same quantity, we can set them equal to each other:

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

Now, to isolate $P(A|B)$, we divide both sides by $P(B)$ (assuming $P(B) > 0$):

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

This is the formulation of **Bayes' theorem**, which expresses the conditional probability of event A given event B in terms of the conditional probability of event B given event A, the prior probability of A, and the marginal probability of B:

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

Theorem 2.1. *If A and B are events such that $P(B) > 0$, Bayes' theorem states that:*

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

Explanation of the Terms:

- $P(A|B)$: The posterior probability of A given B, which reflects our updated belief about A after observing B.
- $P(B|A)$: The likelihood, or the probability of observing B given that A is true. This measures how well A explains the observed evidence B
- $P(A)$: The prior probability of A, representing our initial belief about A before considering the evidence B.
- $P(B)$: The marginal probability of B, which can be computed using the law of total probability, as described in the next theorem.

Intuition Behind Bayes' Theorem:

Bayes' theorem allows us to revise our beliefs in light of new evidence. For example, suppose we want to determine the probability that a patient has a certain disease (event A) given that they tested positive for it (event B). Using Bayes' theorem, we can incorporate the accuracy of the test (the likelihood $P(B|A)$), our initial belief about the prevalence of the disease (the prior $P(A)$), and the overall rate of positive tests (the marginal $P(B)$) to arrive at a more informed estimate of the probability that the patient actually has the disease.

Theorem 2.2. *The law of total probability provides a way to compute the total probability of an event based on a partition of the sample space. It states that if $B_1, B_2, \ldots, B_n$ are mutually exclusive events that form a complete partition of the sample space, then for any event A:*

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i).$$

Explanation of the Terms:

- $P(A)$: The total probability of the event A.
- $P(A|B_i)$: The conditional probability of A given that the event B_i has occurred.
- $P(B_i)$: The probability of each partition event B_i.

Intuition Behind Law of Total Probability:

The law of total probability helps us compute the probability of an event by considering all possible scenarios (the partition B_i) that could lead to that event. For instance, if we want to determine the probability that it will rain tomorrow (event A), we might partition the sample space based on different weather conditions (e.g., sunny, cloudy, or stormy). By calculating the probability of rain under each condition and weighting these probabilities by the likelihood of each condition occurring, we can obtain the overall probability of rain.

2.2.1 Classical Problems on Conditional Probability

Which coin was tossed?

You have one fair coin and one biased coin that lands Heads with probability $\frac{3}{4}$. You pick one of the coins at random and flip it three times. It lands Heads all three times. Given this information, what is the probability that the coin you picked is the fair one?

Let A be the event that the chosen coin lands Heads three times, and let F be the event that we picked the fair coin. We are interested in $P(F|A)$, but it is easier to find $P(A|F)$ and $P(A|F^c)$ since it helps to know which coin we have; this suggests using Bayes' rule and the law of total probability. Doing so, we have

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

By the law of total probability,

$$P(A) = P(A|F)P(F) + P(A|F^c)P(F^c).$$

Given that $P(F) = P(F^c) = \frac{1}{2}$, we can compute:

$$P(A|F) = \left(\frac{1}{2}\right)^3 = \frac{1}{8},$$

$$P(A|F^c) = \left(\frac{3}{4}\right)^3 = \frac{27}{64}.$$

Thus, we have

$$P(A) = P(A|F)P(F) + P(A|F^c)P(F^c) = \left(\frac{1}{8} \cdot \frac{1}{2}\right) + \left(\frac{27}{64} \cdot \frac{1}{2}\right) = \frac{1}{16} + \frac{27}{128}.$$

Calculating this gives:

$$P(A) = \frac{8}{128} + \frac{27}{128} = \frac{35}{128}.$$

Now, substituting back into Bayes' theorem:

$$P(F|A) = \frac{P(A|F)P(F)}{P(A)} = \frac{\left(\frac{1}{8}\right) \cdot \left(\frac{1}{2}\right)}{\frac{35}{128}} = \frac{\frac{1}{16}}{\frac{35}{128}} = \frac{128}{560} \approx 0.23.$$

Before flipping the coin, we thought we were equally likely to have picked the fair coin as the biased coin: $P(F) = P(F^c) = \frac{1}{2}$. Upon observing three Heads, however, it becomes more likely that we've chosen the biased coin than the fair coin, so $P(F|A)$ is only about 0.23.

Prior vs. Posterior: It would not be correct in the calculation in the above example to say after the first step, $P(A) = 1$ because we know A happened. It is true that $P(A|A) = 1$, but $P(A)$ is the prior probability of A and $P(F)$ is the prior probability of F—both are the probabilities before we observe any data in the experiment. These must not be confused with posterior probabilities conditional on the evidence A.

Testing a rare disease

Consider a rare disease that affects 1% of a population. A diagnostic test is available for this disease, which has the following characteristics:

- The probability of testing positive if the person has the disease (true positive rate) is $P(\text{Positive}|\text{Disease}) = 0.95$.
- The probability of testing positive if the person does not have the disease (false positive rate) is $P(\text{Positive}|\text{No Disease}) = 0.10$.

We want to find the probability that a person has the disease given that they tested positive for the disease, denoted as $P(\text{Disease}|\text{Positive})$.

Using Bayes' theorem, we can express the conditional probability as follows:

$$P(\text{Disease}|\text{Positive}) = \frac{P(\text{Positive}|\text{Disease}) \cdot P(\text{Disease})}{P(\text{Positive})}.$$

To compute this, we need to determine $P(\text{Positive})$ using the law of total probability:

$$P(\text{Positive}) = P(\text{Positive}|\text{Disease}) \cdot P(\text{Disease}) + P(\text{Positive}|\text{No Disease}) \cdot P(\text{No Disease}).$$

Given that $P(\text{Disease}) = 0.01$ and $P(\text{No Disease}) = 1 - P(\text{Disease}) = 0.99$, we can substitute the values:

$$P(\text{Positive}) = (0.95 \cdot 0.01) + (0.10 \cdot 0.99) = 0.0095 + 0.099 = 0.1085.$$

Now we can calculate $P(\text{Disease}|\text{Positive})$:

$$P(\text{Disease}|\text{Positive}) = \frac{0.95 \cdot 0.01}{0.1085} = \frac{0.0095}{0.1085} \approx 0.0875.$$

Thus, the probability that a person has the disease given that they tested positive is approximately 8.75%.

Now, let's consider the situation where the same person tests positive a second time. We need to update our previous result using Bayes' theorem again.

Let B represent the event that the person tests positive again. We want to find $P(\text{Disease}|B \cap \text{Positive})$.

Using the law of total probability, we find $P(B|\text{Disease})$ and $P(B|\text{No Disease})$:

$$P(B|\text{Disease}) = P(\text{Positive}|\text{Disease}) = 0.95,$$
$$P(B|\text{No Disease}) = P(\text{Positive}|\text{No Disease}) = 0.10.$$

Now, we can calculate the probability of a second positive test:

$$P(B) = P(B|\text{Disease}) \cdot P(\text{Disease}|\text{Positive}) + P(B|\text{No Disease}) \cdot P(\text{No Disease}|\text{Positive}).$$

Calculating $P(\text{No Disease}|\text{Positive})$:

$$P(\text{No Disease}|\text{Positive}) = 1 - P(\text{Disease}|\text{Positive}) \approx 1 - 0.0875 = 0.9125.$$

Substituting in the values:

$$P(B) = (0.95 \cdot 0.0875) + (0.10 \cdot 0.9125) \approx 0.083125 + 0.09125 = 0.174375.$$

Now we can find $P(\text{Disease}|B)$:

$$P(\text{Disease}|B) = \frac{P(B|\text{Disease}) \cdot P(\text{Disease}|\text{Positive})}{P(B)} = \frac{0.95 \cdot 0.0875}{0.174375} \approx \frac{0.083125}{0.174375} \approx 0.477.$$

Thus, after a second positive test, the updated probability that the person has the disease is approximately 47.7%.

This illustrates how Bayes' theorem allows us to update our beliefs about the probability of having a disease as new information becomes available. Initially, the probability was low due to the rarity of the disease, but successive positive test results significantly increased the probability of having the disease.

Monty-Hall Problem

The Monty Hall problem is a famous probability puzzle based on a game show scenario. The setup is as follows:

A contestant is presented with three doors: Behind one door is a car (the prize), and behind the other two doors are goats (non-prizes). The contestant picks one door, say Door 1. The host, Monty Hall, who knows what is behind each door, opens another door (say Door 3), revealing a goat. The contestant is then given a choice: stick with their original choice (Door 1) or switch to the remaining closed door (Door 2).

We want to determine whether the contestant should stick with their original choice or switch doors to maximize their chances of winning the car.

Let's define the events:

- Let C_i be the event that the car is behind Door i for $i = 1, 2, 3$.
- Let D be the event that Monty opens Door 3.

We want to find the conditional probability $P(C_1|D)$, the probability that the car is behind Door 1 given that Monty opens Door 3.

According to Bayes' theorem, we have:

$$P(C_1|D) = \frac{P(D|C_1) \cdot P(C_1)}{P(D)}.$$

1. *Prior Probability*:

The probability that the car is behind any specific door is equal:

$$P(C_1) = P(C_2) = P(C_3) = \frac{1}{3}.$$

2. *Likelihood*:

We need to find $P(D|C_1)$, the probability that Monty opens Door 3 given that the car is behind Door 1. If the car is behind Door 1, Monty can open either Door 2 or Door 3 (both have goats). Thus, he will open Door 3 with probability $1/2$:

$$P(D|C_1) = \frac{1}{2}.$$

Now, we calculate $P(D|C_2)$ and $P(D|C_3)$:

(a) If the car is behind Door 2, Monty must open Door 3 (the only door he can open that reveals a goat):

$$P(D|C_2) = 1.$$

(b) If the car is behind Door 3, Monty cannot open Door 3, so:

$$P(D|C_3) = 0.$$

3. *Total Probability*:

Now we need to calculate $P(D)$:

$$P(D) = P(D|C_1) \cdot P(C_1) + P(D|C_2) \cdot P(C_2) + P(D|C_3) \cdot P(C_3).$$

Substituting the values we found:

$$P(D) = \left(\frac{1}{2} \cdot \frac{1}{3}\right) + (1 \cdot \frac{1}{3}) + (0 \cdot \frac{1}{3}) = \frac{1}{6} + \frac{1}{3} = \frac{1}{6} + \frac{2}{6} = \frac{3}{6} = \frac{1}{2}.$$

Now we can use Bayes' theorem:

$$P(C_1|D) = \frac{P(D|C_1) \cdot P(C_1)}{P(D)} = \frac{\left(\frac{1}{2}\right) \cdot \left(\frac{1}{3}\right)}{\frac{1}{2}} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}.$$

Thus, $P(C_1|D) = \frac{1}{3}$.
Next, we calculate $P(C_2|D)$:

$$P(C_2|D) = \frac{P(D|C_2) \cdot P(C_2)}{P(D)} = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

The probabilities are:

$$P(C_1|D) = \frac{1}{3}, \quad P(C_2|D) = \frac{2}{3}.$$

This means that if the contestant switches to Door 2 after Monty reveals Door 3, their chances of winning the car increase to $\frac{2}{3}$. Therefore, it is advantageous for the contestant to switch doors, rather than stick with their original choice.

The intuitive reasoning behind this is: Monty knows where the car is and always opens a door with a goat behind it. This action provides additional information. If your initial choice was wrong (which happens 2/3 of the time), Monty has only one option to reveal a goat. If you switch after Monty reveals a goat, you effectively take advantage of the higher probability that the car is behind one of the doors you didn't initially choose. By switching, you win the car 2/3 of the time.

Gambler's Ruin Problem

Consider a gambler (Player 1) who starts with m dollars and plays a game against another player (Player 2) who has $N - m$ dollars. Player 1 wins each round with a probability p and loses with probability $q = 1 - p$. We aim to find the probability $P(m)$ that Player 1 eventually reaches N dollars before going broke.

Let $P(m)$ be the probability of reaching N dollars starting with m dollars. The boundary conditions are:

$$\begin{aligned} P(0) &= 0 \quad \text{(if Player 1 has no money, they cannot win)} \\ P(N) &= 1 \quad \text{(if Player 1 has } N \text{ dollars, they have won)} \end{aligned}$$

Using the law of total probability, we can express the probability recursively:

$$P(m) = pP(m+1) + qP(m-1)$$

Rearranging the above equation gives:

$$pP(m+1) - P(m) + qP(m-1) = 0$$

This is a second-order linear difference equation. The general solution has the form:

$$P(m) = A + Br^m$$

where $r = \frac{q}{p}$ and A and B are constants determined by the boundary conditions.

1. Using $P(0) = 0$:

$$0 = A + Br^0 \implies A + B = 0 \implies A = -B$$

2. Using $P(N) = 1$:

$$1 = -B + Br^N \implies 1 = B(r^N - 1) \implies B = \frac{1}{r^N - 1}$$

Thus, substituting A:

$$A = -B = -\frac{1}{r^N - 1}$$

We have:

$$P(m) = -\frac{1}{r^N - 1} + \frac{1}{r^N - 1} r^m = \frac{r^m - 1}{r^N - 1}$$

Substituting $r = \frac{q}{p}$:

$$P(m) = \frac{\left(\frac{q}{p}\right)^m - 1}{\left(\frac{q}{p}\right)^N - 1}$$

In the special case where the game is fair, we have $p = \frac{1}{2}$ and $q = \frac{1}{2}$. Substituting these values into our probability expression, we find:

$$r = \frac{q}{p} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1$$

The expression for $P(m)$ becomes indeterminate as it involves division by zero. Therefore, we need to analyze the limits as p approaches $\frac{1}{2}$.

Taking the limit:

$$P(m) = \lim_{p \to \frac{1}{2}} \frac{\left(\frac{q}{p}\right)^m - 1}{\left(\frac{q}{p}\right)^N - 1}$$

This can be rewritten using L'Hôpital's Rule to resolve the $\frac{0}{0}$ indeterminate form:

$$P(m) = \frac{m}{N}$$

Thus, in a fair game, the probability that Player 1 eventually reaches N dollars starting with m dollars is:

$$P(m) = \frac{m}{N}$$

This result intuitively means that in a fair game, the chances of winning are directly proportional to the amount of money the player starts with relative to the total amount.

Prosecutor's Fallacy

The prosecutor's fallacy refers to a common misunderstanding of probability, especially in the context of legal cases. It arises when the prosecution misinterprets statistical evidence, leading to an exaggerated perception of guilt. This fallacy is vividly illustrated in the case of Sally Clark, a British solicitor wrongfully convicted of murdering her two infant sons.

Sally Clark was accused of murdering her two sons, Christopher and Harry, who died suddenly within a span of 16 months. The prosecution's case heavily relied on statistical evidence suggesting that the probability of two children dying of natural causes in a family with no known health issues was extremely low—approximately 1 in 8500.

To understand the prosecutor's fallacy using Bayes' theorem, we define the relevant events:

- Let G be the event that Sally Clark is guilty of murder.
- Let E be the event that two infants die under seemingly natural circumstances.

According to Bayes' theorem, we can express the probability of guilt given the evidence E:

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

Where:

- $P(G|E)$ is the posterior probability that Clark is guilty given the evidence of the two infant deaths.
- $P(E|G)$ is the likelihood of observing two infant deaths if she is guilty.

- $P(G)$ is the prior probability of her being guilty (which is often assumed to be low in cases without strong evidence).
- $P(E)$ is the overall probability of observing two infant deaths in any family.

In the Clark case, the prosecution focused on $P(E|G)$ without properly considering $P(E)$. They asserted that the rarity of two natural infant deaths was strong evidence of guilt. However, this neglects the base rate of infant deaths in the general population and the importance of prior probabilities.

If the base rate $P(E)$ is high due to natural causes, it significantly impacts the posterior probability $P(G|E)$. The correct approach should have considered the conditional probabilities:

$$P(E) = P(E|G) \cdot P(G) + P(E|G^c) \cdot P(G^c)$$

Where G^c is the event that she is not guilty. The prosecution assumed $P(E|G^c)$ was negligible, which is misleading.

The misinterpretation of statistics led to a wrongful conviction. Sally Clark's case demonstrates that statistical evidence can be misleading when not contextualized appropriately. The rarity of an event does not imply that an individual is guilty; instead, it reflects the overall population dynamics.

Defense Attorney's Fallacy

The defense attorney's fallacy is a common misunderstanding in the interpretation of probabilities in legal contexts. It occurs when an attorney asserts that the probability of a defendant being innocent given a piece of evidence is high, based solely on the probability of that evidence occurring in the general population, without considering the prior probability of guilt.

Bayes' theorem provides a mathematical framework for updating probabilities based on new evidence. It states:

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

where:

- H is the hypothesis (e.g., the defendant is guilty).
- E is the evidence (e.g., a DNA match).

In a legal context, a defense attorney might argue that because DNA evidence matches the defendant, the probability of the defendant being innocent is high. This reasoning is flawed as it ignores the prior probabilities of guilt and innocence.

Consider the following scenario:

- There is a rare type of crime where only 1 in 1,000 people commit it, i.e., $P(G) = 0.001$ (the prior probability that the defendant is guilty).

- If the defendant is guilty, the probability that the DNA evidence matches is $P(E|G) = 0.95$.
- If the defendant is innocent, the probability that the DNA evidence matches due to chance is $P(E|I) = 0.01$ (where I represents innocence).

We want to find $P(G|E)$, the probability that the defendant is guilty given the DNA evidence matches.
First, we calculate $P(E)$ using the law of total probability:

$$P(E) = P(E|G) \cdot P(G) + P(E|I) \cdot P(I)$$

where $P(I) = 1 - P(G) = 0.999$.
Substituting the values:

$$P(E) = (0.95 \cdot 0.001) + (0.01 \cdot 0.999)$$

Calculating:

$$P(E) = 0.00095 + 0.00999 = 0.01094$$

Now we apply Bayes' theorem:

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

Substituting the values:

$$P(G|E) = \frac{0.95 \cdot 0.001}{0.01094}$$

Calculating:

$$P(G|E) \approx \frac{0.00095}{0.01094} \approx 0.0869$$

Despite the DNA evidence strongly suggesting a match (with $P(E|G) = 0.95$), the probability that the defendant is guilty given this evidence is only approximately 8.69%. This illustrates the defense attorney's fallacy: simply pointing to the DNA match without considering the base rates of guilt can lead to erroneous conclusions about a defendant's innocence or guilt.

Simpson's Paradox

Simpson's Paradox occurs when a trend evident in several groups reverses when the groups are combined. This paradox highlights the importance of considering the underlying structure of data when interpreting results. Bayes' theorem can be employed to elucidate this phenomenon by examining conditional probabilities.

Consider two doctors, Dr. A and Dr. B, treating two groups of patients (Group 1 and Group 2). Below is the performance data:

Group	Dr. A (Successful)	Dr. A (Total)	Dr. B (Successful)	Dr. B (Total)
Group 1	81	87	234	270
Group 2	192	263	55	80

Table 2.1: Performance of Doctors A and B

Calculating the success rates for each doctor in each group: For Doctor A in Group 1:

$$P(\text{Success}|\text{Doctor A, Group 1}) = \frac{81}{87} \approx 0.9310$$

For Doctor B in Group 1:

$$P(\text{Success}|\text{Doctor B, Group 1}) = \frac{234}{270} \approx 0.8667$$

For Doctor A in Group 2:

$$P(\text{Success}|\text{Doctor A, Group 2}) = \frac{192}{263} \approx 0.7300$$

For Doctor B in Group 2:

$$P(\text{Success}|\text{Doctor B, Group 2}) = \frac{55}{80} = 0.6875$$

In both groups, Doctor A outperforms Doctor B. However, when we combine the results across both groups, we find:

$$\text{Total Successes for Doctor A} = 81 + 192 = 273$$

$$\text{Total Treatments for Doctor A} = 87 + 263 = 350$$

$$P(\text{Success}|\text{Doctor A}) = \frac{273}{350} \approx 0.7800$$

$$\text{Total Successes for Doctor B} = 234 + 55 = 289$$

$$\text{Total Treatments for Doctor B} = 270 + 80 = 350$$

$$P(\text{Success}|\text{Doctor B}) = \frac{289}{350} \approx 0.8257$$

In this case, Doctor A appears better within each group, but the overall success rate for Doctor B is higher when considering the distribution of patients treated.

We have seen this paradox in real-life scenarios multiple times. In the 1970s, the University of California, Berkeley, observed that men were admitted at a higher overall rate than women. However, within most departments, women were admitted at a higher rate than men. Research shows that within any age group, cigarette smokers have higher mortality rates than cigar smokers. However, because cigarette smokers are generally younger than cigar smokers, the overall mortality rate for cigarette smokers can appear lower when not controlling for age. The main reason behind this paradox is that the sizes of groups is different!

Trevsky and Kanheman Problem

The Linda problem, presented by Tversky and Kahneman, illustrates a common cognitive bias known as the conjunction fallacy. It involves a character named Linda and asks respondents to assess the probability of certain scenarios involving her.

Linda is described as a 31-year-old single woman, outspoken, and very bright. She majored in philosophy, and as a student, she was deeply concerned with issues of discrimination and social justice.

The following options are presented:

1. Linda is a bank teller.
2. Linda is a bank teller and is active in the feminist movement.

Many people mistakenly choose the second option, believing it is more probable, despite the fact that the probability of being a bank teller and active in the feminist movement (a conjunction) cannot be greater than the probability of being just a bank teller. Let:

- A: The event that Linda is a bank teller.
- B: The event that Linda is a bank teller and is active in the feminist movement.

We want to find $P(B|A)$, the probability of B given A. By the definition of conditional probability:

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

Where:

- $P(A \cap B)$: The joint probability of both events occurring.
- $P(A)$: The probability that Linda is a bank teller.

Since B is a subset of A:

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

Thus, we have:

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

In this case, since people believe Linda is more likely to be both a bank teller and active in the feminist movement, they might estimate $P(B)$ based on their impression of Linda rather than the actual probabilities.

Assuming:

- $P(A)$ (Linda being a bank teller) is more substantial than $P(B)$ (Linda being a bank teller and a feminist).
- $P(B|A)$ (the probability of being a bank teller and a feminist given that she is a bank teller) can't exceed $P(A)$.

The Linda problem demonstrates the conjunction fallacy where respondents incorrectly assess $P(B|A)$ to be more probable than $P(A)$. In reality, the probability of two events occurring together cannot be greater than the probability of either event occurring alone, which highlights the importance of understanding conditional probabilities and cognitive biases in decision-making.

2.3 Random Variables and Their Distributions

In probability theory, the concept of a random variable provides a powerful way to describe quantities that can change due to random processes. As highlighted in the gambler's ruin problem, instead of cumbersome notation like A_{jk} for gambler A's wealth after k rounds, we can simply denote this quantity as X_k. This simplification not only makes our expressions more manageable but also allows us to perform algebraic manipulations easily.

For instance, if we define Y_k as the wealth of gambler B, we can express the difference in their wealths or convert their wealth into another currency without convoluted notation. This clarity is crucial for deriving properties and relationships of interest.

Definition 2.9. *A **random variable** is a function that associates a real number with each outcome of a random experiment. Formally, if S is the sample space of a random experiment, a random variable X is a mapping:*

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

This mapping allows us to quantify outcomes of random processes in a structured manner.

Random variables are typically classified into two categories:

Definition 2.10. *A **discrete random variable** takes on a countable number of distinct values. These values can be finite or countably infinite. The probability mass function (PMF) for a discrete random variable gives the probability that the variable takes on a particular value x:*

$$P(X = x) = p(x)$$

For example, consider a fair six-sided die. The random variable X representing the outcome of a die roll can take values from the set $\{1, 2, 3, 4, 5, 6\}$ with the PMF:

$$p(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise} \end{cases}$$

Definition 2.11. *A **continuous random variable** can take on an uncountably infinite number of values, often associated with measurements. The probability density function (PDF) for a continuous random variable describes the likelihood of the variable falling within a particular interval:*

$$P(a < X < b) = \int_a^b f(x)\, dx$$

where $f(x)$ is the PDF.

For example, let X represent the height of adult men in a population, which can take any value in the interval $[0, \infty)$. The corresponding PDF $f(x)$ may resemble a normal distribution.

2.3.1 Famous Discrete Distributions

Bernoulli Distribution

The Bernoulli distribution models a single trial with two possible outcomes, typically labeled as "success" (1) and "failure" (0). It is fundamental in probability theory and forms the basis for more complex distributions.

The probability mass function (PMF) for a Bernoulli random variable X is given by:

$$P(X = x) = p^x(1-p)^{1-x} \quad \text{for } x \in \{0, 1\}$$

where p is the probability of success. For example, if $p = 0.7$, then:

$$P(X = 1) = 0.7 \quad \text{and} \quad P(X = 0) = 0.3.$$

If X and Y are independent Bernoulli random variables with the same success probability p, then:

$$X + Y \sim \text{Binomial}(n = 2, p).$$

Bernoulli trials are widely used in quality control, clinical trials, and any scenario requiring binary outcomes, such as whether a patient responds to treatment.

Binomial Distribution

The Binomial distribution models the number of successes in a fixed number of independent Bernoulli trials. It extends the Bernoulli distribution to multiple trials.

The PMF for a Binomial random variable X representing the number of successes in n trials is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, \ldots, n.$$

For instance, if $n = 5$ and $p = 0.6$:

$$P(X = 3) = \binom{5}{3} (0.6)^3 (0.4)^2 \approx 0.2304.$$

If $X \sim \text{Binomial}(n_1, p)$ and $Y \sim \text{Binomial}(n_2, p)$, then:

$$X + Y \sim \text{Binomial}(n_1 + n_2, p).$$

Binomial distributions are utilized in fields such as quality assurance, market research, and genetics, where the outcome of interest is the number of successes in multiple trials.

Hypergeometric Distribution

The Hypergeometric distribution models the probability of successes in draws from a finite population without replacement. This is relevant when the sample size is significant relative to the population.

The PMF for a Hypergeometric random variable X is given by:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad \text{for } k = 0, 1, \ldots, \min(K, n).$$

where N is the population size, K is the number of successes in the population, and n is the number of draws.
For example, in a population of 20 items (10 successes), if we draw 5 items:

$$P(X = 3) = \frac{\binom{10}{3}\binom{10}{2}}{\binom{20}{5}}.$$

If X and Y are independent Hypergeometric random variables from separate populations, their sum does not follow a Hypergeometric distribution.

Hypergeometric distributions are commonly used in quality control and ecological studies where samples are drawn without replacement.

Example 2.1. *Imagine a standard deck of 52 playing cards, which includes 12 face cards (Kings, Queens, Jacks). If you draw 5 cards without replacement, you want to know the probability of getting exactly 3 face cards.*

This scenario can be modeled using the Hypergeometric distribution:

- *Total cards* $N = 52$
- *Face cards* $K = 12$
- *Cards drawn* $n = 5$
- *Face cards drawn* $k = 3$

The probability is given by:

$$P(X = 3) = \frac{\binom{12}{3}\binom{40}{2}}{\binom{52}{5}} = \frac{220 \times 780}{2598960} \approx 0.0674.$$

Poisson Distribution

The Poisson distribution models the number of events occurring in a fixed interval of time or space when these events happen with a known constant mean rate and are independent of the time since the last event.

The PMF for a Poisson random variable X is:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{for } k = 0, 1, 2, \ldots$$

where λ is the average number of events in the interval. For instance, if $\lambda = 4$:

$$P(X = 2) = \frac{4^2 e^{-4}}{2!} \approx 0.1465.$$

If $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$, then:

$$X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2).$$

This is useful in service industries, telecommunications, and event modeling where events occur independently over a fixed interval.

Example 2.2. *A call center receives an average of 3 calls per hour. We want to determine the probability of receiving exactly 5 calls in the next hour.*

The number of calls received can be modeled by a Poisson distribution with $\lambda = 3$:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} = \frac{3^5 e^{-3}}{5!} = \frac{243 \cdot e^{-3}}{120} \approx 0.1008.$$

2.3.2 Connection between Discrete Distributions

Binomial and Hypergeometric

We will explore the connection between the Binomial and Hypergeometric distributions. Specifically, we will prove two theorems that relate the two distributions, first by showing that the conditional distribution of the sum of two independent Binomial random variables results in a Hypergeometric distribution, and second by showing that the Hypergeometric distribution converges to a Binomial distribution under certain conditions.

The connection between the Binomial and Hypergeometric distributions becomes clearer when we think about two different sampling methods from an urn containing w white balls and b black balls.

- **Binomial distribution**: This arises when we sample n balls *with replacement*. Every time we draw a ball, we put it back into the urn before drawing again, so the probability of drawing a white or black ball remains constant in each trial. The probability of drawing a white ball in each trial is $p = \frac{w}{w+b}$.
- **Hypergeometric distribution**: This arises when we sample n balls *without replacement*. Each time we draw a ball, we remove it from the urn, which means the probability of drawing a white or black ball changes after each draw.

As the total number of balls in the urn ($N = w + b$) becomes very large relative to the number of balls drawn (n), the difference between sampling with and without replacement becomes negligible. This is because, when N is very large, removing a ball (in the Hypergeometric case) barely affects the overall composition of the remaining balls. In this situation, the Hypergeometric distribution can be well-approximated by the Binomial distribution.

Theorem 2.3. *Let $X \sim Bin(n, p)$ and $Y \sim Bin(m, p)$, where X and Y are independent random variables. Then, the conditional distribution of X given $X + Y = r$ follows a Hypergeometric distribution:*

$$X \mid (X + Y = r) \sim HGeom(n, m, r)$$

Proof. We will prove this by computing the conditional probability mass function (PMF) of X given $X + Y = r$.

The joint probability that $X = k$ and $X + Y = r$ can be written as:

$$P(X = k, X + Y = r) = P(X = k, Y = r - k)$$

Since X and Y are independent, we have:

$$P(X = k, X + Y = r) = P(X = k)P(Y = r - k)$$

Substitute the Binomial PMFs for X and Y:

$$P(X = k, X + Y = r) = \binom{n}{k} p^k (1 - p)^{n-k} \binom{m}{r-k} p^{r-k} (1 - p)^{m-(r-k)}$$

Now, the conditional probability $P(X = k \mid X + Y = r)$ is given by:

$$P(X = k \mid X + Y = r) = \frac{P(X = k, X + Y = r)}{P(X + Y = r)}$$

First, simplify the numerator:

$$P(X = k, X + Y = r) = \binom{n}{k} \binom{m}{r-k} p^r (1 - p)^{n+m-r}$$

Next, compute the denominator $P(X + Y = r)$ by summing over all possible values of X:

$$P(X + Y = r) = \sum_{k=0}^{r} \binom{n}{k} \binom{m}{r-k} p^r (1 - p)^{n+m-r}$$

Notice that $p^r(1 - p)^{n+m-r}$ factors out from both the numerator and denominator. Thus, the conditional probability simplifies to:

$$P(X = k \mid X + Y = r) = \frac{\binom{n}{k}\binom{m}{r-k}}{\binom{n+m}{r}}$$

This is exactly the PMF of a Hypergeometric distribution with parameters n, m, and r. Hence, we have:

$$X \mid (X + Y = r) \sim \text{HGeom}(n, m, r)$$

□

An example that illustrates the convergence of two Binomial distributions to a Hypergeometric distribution and highlights the fact that the conditional distribution becomes independent of p is the following scenario in medical testing. Suppose we are studying a disease that affects a population in two different regions. Let:

- X be the number of diseased individuals in a random sample of size n taken from Region A, where the disease affects a proportion p of the population.
- Y be the number of diseased individuals in a random sample of size m taken from Region B, where the same disease also affects a proportion p of the population.

Both X and Y are independent and follow Binomial distributions:

$$X \sim \text{Bin}(n, p), \quad Y \sim \text{Bin}(m, p)$$

Now, assume that we know the total number of diseased individuals across both regions, i.e., $X + Y = r$. In this case, the conditional distribution of X given $X + Y = r$ is:

$$X \mid (X + Y = r) \sim \text{HGeom}(n, m, r)$$

Here, the conditional distribution of X does **not depend on** p. Once we know that the total number of diseased individuals is r, we can work directly with the fact that we are drawing a total of r diseased people from a combined population of $n + m$, regardless of the original probability p of being diseased.

This is useful in statistics because it simplifies the problem: even though we started with Binomial distributions that depend on p, once we condition on the total number of diseased individuals, we can work with a Hypergeometric distribution that no longer involves p.

Example 2.3. *Suppose Region A has $n = 10$ people and Region B has $m = 15$ people, and we are testing for a disease that has a probability $p = 0.2$ of infecting each individual. Initially, both X and Y are Binomial with the same p, but suppose we are told that there are exactly $r = 8$ diseased individuals in total across both regions.*

At this point, we no longer need to know the value of p. Instead, the number of diseased individuals X in Region A, given that there are 8 diseased individuals in total across both regions, follows a Hypergeometric distribution:

$$X \mid (X + Y = 8) \sim HGeom(10, 15, 8)$$

This means that we are now selecting 8 diseased individuals from a combined population of 10 in Region A and 15 in Region B, without worrying about the original probability p that generated these numbers.

Theorem 2.4. *Let $X \sim HGeom(w, b, n)$, where w is the number of "white" balls, b is the number of "black" balls, and n is the number of balls drawn. Define $N = w + b$ and suppose that $N \to \infty$ while the ratio $p = \frac{w}{N}$ remains fixed. Then, the PMF of X converges to the PMF of a Binomial random variable:*

$$X \sim Bin(n, p)$$

Proof. The PMF of the Hypergeometric random variable X is given by:

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$$

Substitute $w = pN$ and $b = (1-p)N$ to express the parameters in terms of N and p. The PMF becomes:

$$P(X = k) = \frac{\binom{pN}{k}\binom{(1-p)N}{n-k}}{\binom{N}{n}}$$

We are interested in the limit as $N \to \infty$. First, recall that for large N, the binomial coefficient $\binom{N}{n}$ behaves approximately as:

$$\binom{N}{n} \approx \frac{N^n}{n!}$$

Using this approximation, we can rewrite the PMF as:

$$P(X = k) \approx \frac{(pN)^k}{k!} \frac{((1-p)N)^{n-k}}{(n-k)!} \frac{n!}{N^n}$$

Simplifying this expression:

$$P(X = k) \approx \binom{n}{k} p^k (1-p)^{n-k}$$

This is exactly the PMF of a Binomial distribution with parameters n and p. Hence, as $N \to \infty$, the Hypergeometric distribution converges to the Binomial distribution:

$$X \sim \text{Bin}(n, p)$$

□

Let's consider a practical numerical example to make this clearer.

- $w = 100$ white balls,
- $b = 900$ black balls,
- $N = w + b = 1000$,
- We draw $n = 10$ balls without replacement.

This means that the probability of drawing X white balls (out of 10) follows a Hypergeometric distribution.

If we assume sampling with replacement, the Binomial distribution can approximate the Hypergeometric distribution when $N = 1000$ is large relative to $n = 10$. The probability of drawing a white ball in each trial is $p = \frac{100}{1000} = 0.1$, so we use a Binomial distribution $X \sim \text{Bin}(10, 0.1)$.

Now, let's compare the probabilities for drawing different numbers of white balls (k) under both distributions (refer Table 2.2).

As we can see from this table, the probabilities for different values of k are very similar under both the Hypergeometric and Binomial distributions. Even though one involves sampling without replacement (Hypergeometric) and the other involves sampling with replacement (Binomial), the results are almost the same because $N = 1000$ is large relative to $n = 10$.

k (white balls)	Hypergeometric $P(X = k)$	Binomial $P(X = k)$
0	0.34868	0.34868
1	0.38742	0.38742
2	0.19468	0.19371
3	0.05853	0.05740
4	0.01016	0.01123
5	0.00098	0.00148

Table 2.2: Comparison of Hypergeometric and Binomial Probabilities

Binomial and Poisson

The Binomial and Poisson distributions are closely related, especially in situations where we are dealing with rare events. The key connection lies in the limits of the Binomial distribution as certain parameters change. The Poisson distribution is actually a limiting case of a Binomial distribution when the number of trials, n, gets very large and p, the probability of success, is small. As a rule of thumb, if $n \geq 100$ and $np \leq 10$, the Poisson distribution (taking $\lambda = np$) can provide a very good approximation to the Binomial distribution.

This is particularly useful as calculating the combinations inherent in the probability formula associated with the Binomial distribution can become difficult when n is large.

To better see the connection between these two distributions, consider the Binomial probability of seeing x successes in n trials, with the aforementioned probability of success, p, as shown below:

$$P(x) = \binom{n}{x} p^x q^{n-x}$$

Let us denote the expected value of the Binomial distribution, np, by λ. Note, this means that

$$p = \frac{\lambda}{n}$$

and since $q = 1 - p$,

$$q = 1 - \frac{\lambda}{n}$$

Now, if we use this to rewrite $P(x)$ in terms of λ, n, and x, we obtain

$$P(x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

Using the standard formula for the combinations of n things taken x at a time and some simple properties of exponents, we can further expand things to

$$P(x) = \frac{n(n-1)(n-2)\cdots(n-x+1)}{x!} \cdot \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

Notice that there are exactly x factors in the numerator of the first fraction. Let us swap denominators between the first and second fractions, splitting the n^x across all of the factors of the first fraction's numerator.

$$P(x) = \frac{n^x}{n^x} \cdot \frac{n-1}{n} \cdots \frac{n-x+1}{n} \cdot \frac{\lambda^x}{x!} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

Finally, let us split the last factor into two pieces, noting (for those familiar with Calculus) that one has a limit of $e^{-\lambda}$.

$$P(x) = \frac{n^x}{n^x} \cdot \frac{n-1}{n} \cdots \frac{n-x+1}{n} \cdot \frac{\lambda^x}{x!} \cdot \left(1 - \frac{\lambda}{n}\right)^{n} \left(1 - \frac{\lambda}{n}\right)^{-x}$$

It should now be relatively easy to see that if we took the limit as n approaches infinity, keeping x and λ fixed, the first x fractions in this expression would tend towards 1, as would the last factor in the expression. The second to last factor, as was mentioned before, tends towards $e^{-\lambda}$, and the remaining factor stays unchanged as it does not depend on n. As such,

$$\lim_{n\to\infty} P(x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

Example 2.4. *To illustrate this connection, consider an example of a rare event, such as the occurrence of a specific type of email spam in a large dataset.*

- *Let* $n = 1000$ *represent the number of emails received.*
- *Let* $p = 0.01$ *be the probability that any given email is spam.*

In this case, we can calculate:

$$\lambda = np = 1000 \times 0.01 = 10$$

So, we can model the number of spam emails X *using a Binomial distribution:*

$$X \sim Bin(1000, 0.01)$$

Now, using the Poisson approximation, we can approximate the same scenario using a Poisson distribution:

$$Y \sim Poisson(10)$$

Let's compare the probabilities of receiving exactly $k = 5$ *spam emails:*

For the Binomial distribution:

$$P(X = 5) = \binom{1000}{5}(0.01)^5(0.99)^{995}$$

For the Poisson distribution:

$$P(Y = 5) = \frac{10^5 e^{-10}}{5!}$$

Both probabilities can be calculated, and as n increases and p decreases while keeping λ *constant, the values of* $P(X = k)$ *will closely approximate* $P(Y = k)$.

2.3.3 Famous Continuous Distributions

Uniform Distribution

The Uniform Distribution is a fundamental probability distribution that models situations where all outcomes are equally likely within a specified range. It is characterized by two parameters: the lower bound a and the upper bound b. This distribution is particularly useful in scenarios where we need to represent the likelihood of events uniformly spread over a continuous interval.

Definition 2.12. *The PDF of a continuous Uniform Distribution defined over the interval* $[a, b]$ *is given by:*

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

Proof. To derive the PDF, we start with the concept of total probability. Since all outcomes in the interval $[a, b]$ are equally likely, the area under the curve (which represents the total probability) must equal 1:

$$\int_a^b f(x)\, dx = 1$$

Assuming a constant value c for the PDF within the interval, we have:

$$\int_a^b c\, dx = c(b-a) = 1$$

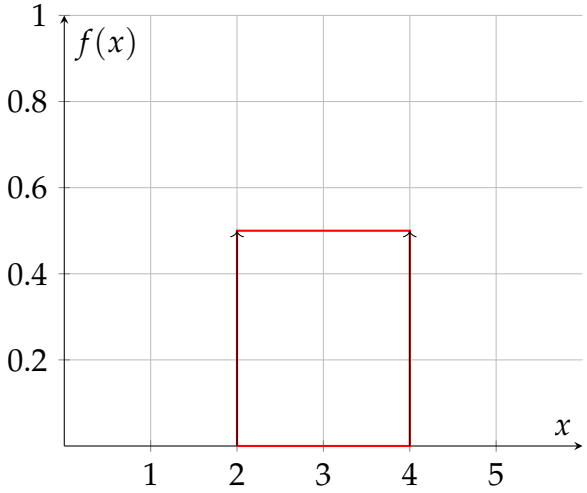
From this, we find:

$$c = \frac{1}{b-a}$$

□

Shape of Uniform Distribution:

The shape of the Uniform Distribution is a rectangle, where the height corresponds to $\frac{1}{b-a}$ and the width spans from a to b.



The width of the rectangle (i.e., $b - a$) determines the range of possible values. A larger range decreases the height of the PDF, thus spreading the probability over a wider area.

If X and Y are independent random variables each following a Uniform Distribution $U(a, b)$, then the distribution of $Z = X + Y$ can be determined. The sum of two independent Uniform random variables results in a triangular distribution over the interval $[2a, 2b]$. Specifically, the PDF of Z is piecewise linear, with its peak at $a + b$.

Exponential Distribution

The Exponential Distribution is a continuous probability distribution that is widely used to model the time until an event occurs, such as the time until failure of a machine, the time between arrivals of customers in a queue, or the time until an earthquake. Its primary motivation lies in its memoryless property, which means that the future probability of an event occurring does not depend on how much time has already elapsed.

The Exponential Distribution is characterized by a single parameter, $\lambda > 0$, known as the rate parameter.

Definition 2.13. *The PDF of the Exponential Distribution is defined as follows:*

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

Proof. To derive this formula, we start from the definition of the cumulative distribution function (CDF):

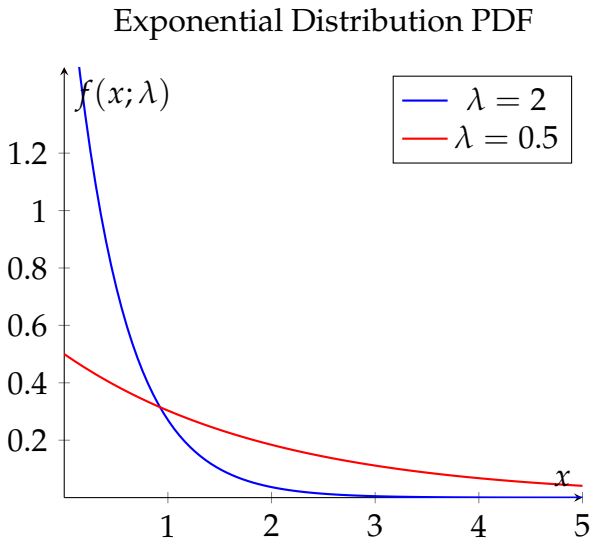
$$F(x; \lambda) = P(X \leq x) = 1 - e^{-\lambda x}, \quad x \geq 0$$

Differentiating the CDF gives us the PDF:

$$f(x; \lambda) = \frac{d}{dx} F(x; \lambda) = \lambda e^{-\lambda x}$$

This derivation shows how the exponential decay characterizes the time until an event occurs. □

The shape of the Exponential Distribution is characterized by its PDF, which is a monotonically decreasing function.



The parameter λ significantly impacts the shape of the distribution:

- A larger λ results in a steeper decline, meaning events occur more frequently (shorter waiting times).
- A smaller λ leads to a slower decline, indicating longer waiting times between events.

For example, with $\lambda = 2$, the average time until the next event is $\frac{1}{2} = 0.5$ units of time, while with $\lambda = 0.5$, it is $\frac{1}{0.5} = 2$ units of time.

If X and Y are independent random variables that follow an Exponential Distribution with the same parameter λ, the distribution of the sum $Z = X + Y$ follows a Gamma Distribution with shape parameter $k = 2$ and scale parameter $\theta = \frac{1}{\lambda}$:

$$Z \sim \text{Gamma}(k = 2, \theta = \frac{1}{\lambda})$$

The PDF of Z is given by:

$$f_Z(z;\lambda) = \frac{\lambda^2 z e^{-\lambda z}}{1!}, \quad z \geq 0$$

We will look at the Gamma Distribution shortly.

Example 2.5. *Consider a scenario where a factory produces light bulbs, and the average lifespan of a bulb is 1000 hours, which means $\lambda = \frac{1}{1000}$ hours^{-1}.*

If we want to model the time until the next bulb fails, we use the Exponential Distribution. The probability that a bulb will last more than 1200 hours is calculated as follows:

$$P(X > 1200) = 1 - P(X \leq 1200) = 1 - F(1200;\lambda) = e^{-\frac{1200}{1000}} \approx e^{-1.2} \approx 0.3012$$

This means there is approximately a 30.12% chance that the bulb will last longer than 1200 hours.

Normal Distribution

The Normal Distribution, also known as the Gaussian distribution, is fundamental in statistics and probability theory. It is crucial for modeling phenomena that tend to cluster around a mean. Many natural occurrences—like heights, test scores, and measurement errors—are often normally distributed.

The Normal Distribution is characterized by two parameters:

- μ: The mean, which indicates the center of the distribution.
- σ: The standard deviation, which measures the dispersion or spread of the distribution.

These parameters define the behavior of the distribution and its shape.

Definition 2.14. *The PDF of a Normal Distribution is given by:*

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

Proof. The PDF must satisfy two properties:

1. It is non-negative: $f(x) \geq 0$ for all x.
2. It integrates to one over the entire space:

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

We want a function $f(x)$ such that it is symmetric on both sides of its mean, allows straightforward integration, decays quickly after the mean - preferably exponentially. The function that fits these criteria is $f(x) = Ae^{-B(x-\mu)^2}$. It's a bell-curve and a good choice because the shape produced by the exponential function is bell-like, which is a natural representation of many real-world phenomena that cluster around a central value. In various fields, such as physics and natural sciences, phenomena like diffusion and heat conduction can be modeled using exponential decay. This physical foundation supports the choice of the exponential function for representing distributions that describe natural processes.

To satisfy the normalization condition, we need:

$$\int_{-\infty}^{\infty} Ae^{-B(x-\mu)^2}\, dx = 1$$

Changing variables to $z = x - \mu$, we have:

$$\int_{-\infty}^{\infty} Ae^{-Bz^2}\, dz = 1$$

This integral is known as the Gaussian integral:

$$\int_{-\infty}^{\infty} e^{-Bz^2}\, dz = \sqrt{\frac{\pi}{B}}$$

Thus, we have:

$$A\sqrt{\frac{\pi}{B}} = 1 \implies A = \sqrt{\frac{B}{\pi}}$$

Let $B = \frac{1}{2\sigma^2}$:

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

This satisfies the normalization condition. Finally, we express the PDF in its standard form:

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

□

Definition 2.15. *The standard normal curve is a special case of the normal distribution where the mean μ is 0 and the standard deviation σ is 1. It is denoted as $N(0,1)$ and is symmetrical about the mean. The equation of the standard normal distribution is given by:*

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

where z represents the z-score.

Definition 2.16. *A z-score indicates how many standard deviations an element is from the mean of the distribution. It is calculated using the formula:*

$$z = \frac{(X - \mu)}{\sigma}$$

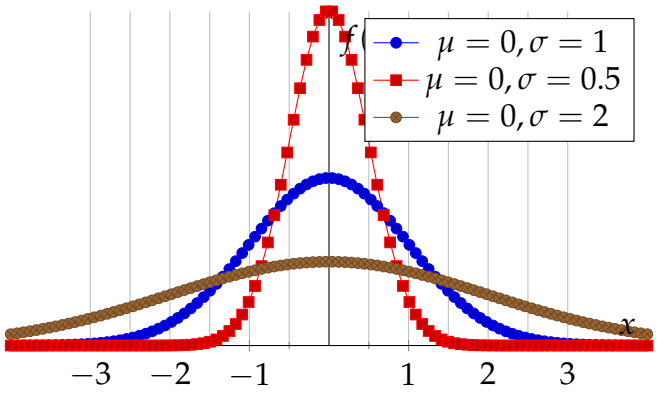
where:

- *X is the value of the observation,*
- *μ is the mean of the distribution,*
- *σ is the standard deviation.*

A positive z-score indicates that the value is above the mean, while a negative z-score indicates that it is below the mean. The z-score is crucial for standardizing values from different normal distributions to the standard normal distribution, allowing for comparison across different datasets.

The Normal Distribution is symmetric and bell-shaped. Its key characteristics include:

- **Mean (μ)**: The peak of the bell curve.
- **Standard Deviation (σ)**: Determines the width of the curve. A smaller σ results in a steeper curve, while a larger σ produces a flatter curve.



The graph illustrates the effect of varying standard deviations on the distribution's shape. For example:

- When $\sigma = 0.5$, the distribution is narrow and peaked.
- When $\sigma = 2$, the distribution is wider and flatter.

If X and Y are independent random variables, both following a Normal Distribution:

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

Then, the sum $Z = X + Y$ is also normally distributed:

$$Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Example 2.6. *Suppose the heights of adult men in a certain city are normally distributed with a mean of $\mu = 175$ cm and a standard deviation of $\sigma = 10$ cm.*

Calculate the following:

1. *The z-score for a man who is 190 cm tall.*
2. *The probability that a randomly selected man from this city is taller than 190 cm.*

The z-score is calculated using the formula:

$$z = \frac{(X - \mu)}{\sigma}$$

where:

- *X is the value for which we want to find the z-score,*
- *μ is the mean,*
- *σ is the standard deviation.*

For a man who is $X = 190$ cm tall:

$$z = \frac{(190 - 175)}{10} = \frac{15}{10} = 1.5$$

To find the probability that a randomly selected man is taller than 190 cm, we look up the z-score in the standard normal distribution table or use a cumulative distribution function (CDF)

calculator.

The CDF gives the probability that a random variable is less than a certain value. Thus, we need to find:

$$P(X > 190) = 1 - P(X \leq 190) = 1 - P(Z \leq 1.5)$$

From standard normal distribution tables, we find:

$$P(Z \leq 1.5) \approx 0.9332$$

Therefore, the probability is:

$$P(X > 190) = 1 - 0.9332 = 0.0668$$

The z-score for a man who is 190 cm tall is 1.5*, and the probability that a randomly selected man from this city is taller than 190 cm is approximately* 0.0668*, or* 6.68%*.*

Log-Normal Distribution

The Log-Normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. It is commonly used to model variables that are multiplicative in nature, such as stock prices, income distributions, and the sizes of living organisms.

The primary parameters involved in modeling a Log-Normal distribution are:

- μ: the mean of the underlying normal distribution (i.e., the logarithm of the variable).
- σ: the standard deviation of the underlying normal distribution.

Definition 2.17. *The PDF of the Log-Normal distribution is defined as follows:*

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \quad x > 0$$

Proof. We start from the fact that if Y is normally distributed, i.e., $Y \sim N(\mu, \sigma^2)$, then the random variable $X = e^Y$ follows a Log-Normal distribution. The transformation of variables provides us with the PDF. Start with the PDF of normal distribution:

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Apply the transformation $x = e^y$, which implies $y = \ln x$ and $dy = \frac{1}{x}dx$.

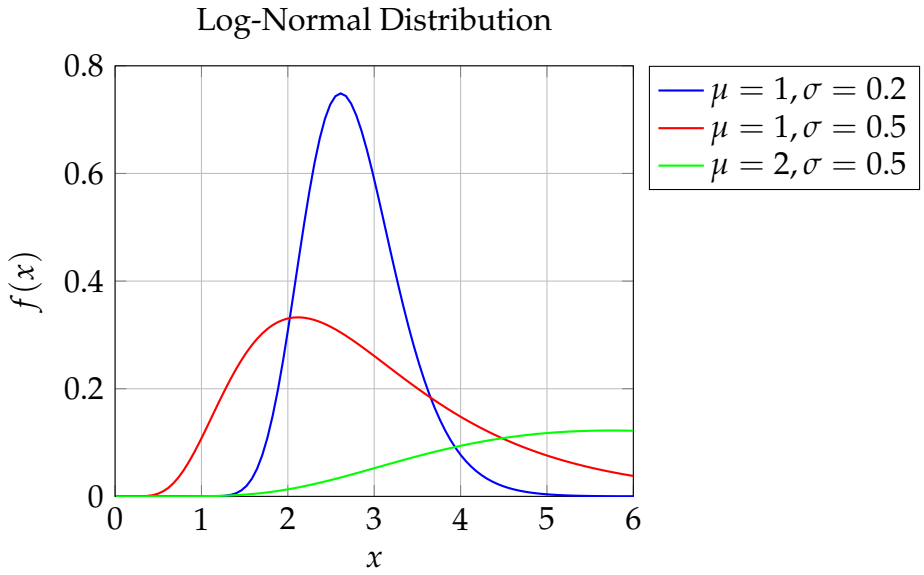
The PDF of X is then obtained using the change of variable:

$$f_X(x) = f_Y(\ln x) \cdot \left| \frac{dy}{dx} \right| = f_Y(\ln x) \cdot \frac{1}{x}$$

Substituting $f_Y(\ln x)$ yields the PDF of the Log-Normal distribution as stated above. □

The shape of the Log-Normal distribution is positively skewed, meaning it has a long tail on the right side. The parameters μ and σ significantly impact its shape:

- As μ increases, the peak of the distribution shifts to the right, indicating a higher central tendency.
- An increase in σ increases the spread of the distribution, resulting in a wider and flatter shape. Conversely, a smaller σ leads to a steeper peak.



For example, with $\mu = 1$ and $\sigma = 0.2$, the distribution is relatively concentrated around the mean, while with $\sigma = 0.5$, it spreads out more, showing the impact of increasing variance.

If X and Y are independent random variables that follow a Log-Normal distribution, specifically $X \sim \text{Log-Normal}(\mu_X, \sigma_X^2)$ and $Y \sim \text{Log-Normal}(\mu_Y, \sigma_Y^2)$, the distribution of the sum $Z = X + Y$ does not follow a Log-Normal distribution.

However, there are approximations to understand the distribution of Z. In some cases, the sum of two Log-Normal random variables can be approximated using a Normal distribution or simulated using numerical methods, depending on the parameters.

Example 2.7. *Consider the modeling of the future price of a stock that currently has a price of $P_0 = 100$. Assuming that the logarithm of the price follows a Normal distribution with parameters $\mu = 0.05$ and $\sigma = 0.1$:*

1. We simulate the future price of the stock after 1 year:

$$P = P_0 \cdot e^Y$$

where $Y \sim N(0.05, 0.1^2)$.

2. If we calculate the expected price and the variance, we can understand the distribution of future prices.

This modeling helps investors understand the risks and potential returns associated with stock investments, providing a probabilistic framework for decision-making.

Weibull Distribution

The Weibull distribution is a continuous probability distribution that is widely used in reliability analysis and failure time analysis. It is particularly useful in modeling life data, where it can describe the time until an event occurs, such as failure of a mechanical component or time to death in survival studies.

Definition 2.18. *The PDF of the Weibull distribution is given by:*

$$f(x;k,\lambda) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

Proof. To derive this PDF, we start with the cumulative distribution function (CDF), defined as:

$$F(x;k,\lambda) = 1 - e^{-(x/\lambda)^k} \quad \text{for } x \geq 0$$

Taking the derivative of the CDF with respect to x yields the PDF:

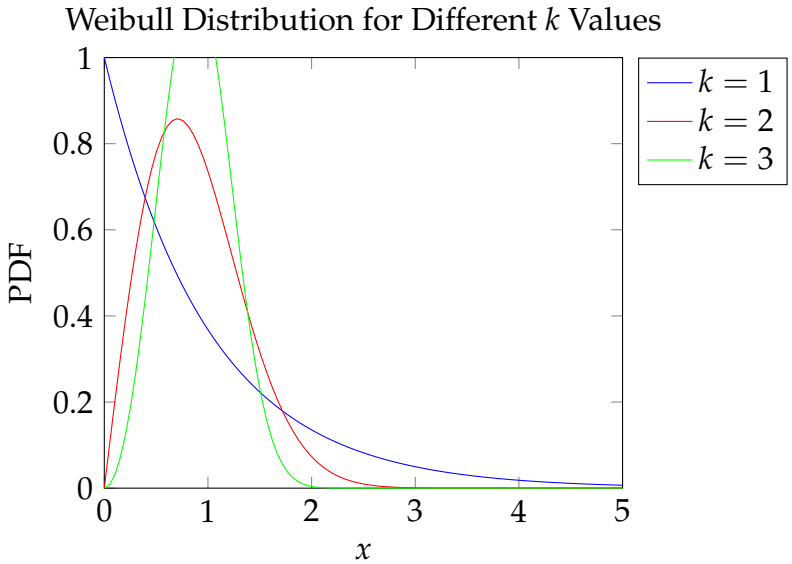
$$f(x;k,\lambda) = \frac{d}{dx}F(x;k,\lambda) = \frac{d}{dx}\left(1 - e^{-(x/\lambda)^k}\right) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$$

□

The shape of the Weibull distribution is heavily influenced by the shape parameter k:

- If $k < 1$: The distribution is decreasing; it indicates that the failure rate decreases over time (e.g., "infant mortality").
- If $k = 1$: The distribution is exponential, implying a constant failure rate.
- If $k > 1$: The distribution is increasing; it indicates that the failure rate increases over time (e.g., "wear-out failures").

The scale parameter λ stretches or compresses the distribution along the x-axis. A larger λ results in a distribution that spreads out more, while a smaller λ makes it steeper.



In this graph, we see the differences in the shapes for $k = 1$, $k = 2$, and $k = 3$.

If X and Y are independent random variables following the Weibull distribution with the same scale parameter λ and shape parameter k, the distribution of $X + Y$ is not generally a Weibull distribution. However, if both are identically distributed, we can use techniques such as convolution to find the PDF of $Z = X + Y$. The resulting distribution would require numerical methods for specific parameters and is generally not expressible in a closed form.

Example 2.8. *Suppose the shape parameter is found to be $k = 1.5$ and the scale parameter is $\lambda = 1000$ hours. This indicates that the bulbs experience increasing failure rates as time progresses. The company can use this model to predict the reliability of their bulbs, estimating that approximately 63% of the bulbs will have failed by 1000 hours.*

Gamma Distribution

The Gamma distribution is a continuous probability distribution that is widely used to model various processes, particularly those that involve waiting times or the time until an event occurs. Its flexibility makes it suitable for a variety of applications, including queuing models, reliability analysis, and Bayesian statistics.

The Gamma distribution is characterized by two parameters:

- α (shape parameter): This parameter dictates the shape of the distribution. It is a positive real number that influences the skewness and the peak of the distribution.
- β (rate parameter): This parameter is also a positive real number, and it controls the scale of the distribution. It is often expressed as the reciprocal of the scale parameter θ (where $\theta = \frac{1}{\beta}$).

Before diving into the Gamma distribution, it is essential to understand the Gamma function, which is defined as:

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

for $n > 0$. The Gamma function generalizes the factorial function to real and complex numbers, with the relationship:

$$\Gamma(n) = (n-1)!$$

for natural numbers n. The derivation of the Gamma function arises from the need to extend the concept of factorial to non-integer values, facilitating calculations in various mathematical fields, particularly in calculus and complex analysis.

Definition 2.19. *The PDF of the Gamma distribution is given by:*

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0$$

where $\Gamma(\alpha)$ is the Gamma function.

Before we dive into the proof of this PDF, we will need some results.

Theorem 2.5. *The PDF of the sum of two independent random variables is the convolution of their PDFs.*

Proof. Let $Z = X + Y$. We want to find the PDF $f_Z(z)$ of the random variable Z.

To derive $f_Z(z)$, we can first find the cumulative distribution function (CDF) $F_Z(z)$:

$$F_Z(z) = P(Z \leq z) = P(X + Y \leq z).$$

Given the independence of X and Y, we can express the probability $P(X + Y \leq z)$ as:

$$F_Z(z) = \int_{-\infty}^{\infty} P(X \leq z - y) f_Y(y)\, dy.$$

Here, $P(X \leq z - y)$ is the CDF of X, denoted by $F_X(z - y)$:

$$F_Z(z) = \int_{-\infty}^{\infty} F_X(z - y) f_Y(y)\, dy.$$

To find the PDF $f_Z(z)$, we differentiate $F_Z(z)$ with respect to z:

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \frac{d}{dz} \int_{-\infty}^{\infty} F_X(z - y) f_Y(y)\, dy.$$

Using the Leibniz rule for differentiating under the integral sign, we have:

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{\partial}{\partial z} F_X(z - y) f_Y(y)\, dy.$$

Since $F_X(z - y)$ is the CDF of X, its derivative is the PDF $f_X(z - y)$:

$$\frac{\partial}{\partial z} F_X(z - y) = f_X(z - y).$$

Thus,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y)\, dy.$$

The expression we derived for $f_Z(z)$ is the convolution of the PDFs f_X and f_Y:

$$f_Z(z) = (f_X * f_Y)(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y)\, dy.$$

Therefore, we have shown that the PDF of the sum of two independent random variables $Z = X + Y$ can be found using the convolution of their PDFs:

$$f_Z(z) = f_X(z) * f_Y(z).$$

□

Now we can prove the PDF of the Gamma Distribution. The Gamma Distribution should be viewed as the Sum of Independent Exponential Random Variables. Let's prove this.

Proof. Let $X_1, X_2, \ldots, X_k$ be k independent random variables, each following an exponential distribution with rate β. The PDF of each X_i is given by:

$$f_{X_i}(x) = \beta e^{-\beta x}, \quad x \geq 0.$$

Define $Y = X_1 + X_2 + \ldots + X_k$. We want to find the PDF of Y.

The PDF of the sum of two independent random variables can be found using the convolution of their PDFs. The convolution of two PDFs $f_X(x)$ and $f_Y(y)$ is defined as:

$$(f_X * f_Y)(z) = \int_0^z f_X(x) f_Y(z-x)\, dx.$$

First, we find the PDF of $Z = X_1 + X_2$:

$$f_Z(z) = \int_0^z f_{X_1}(x) f_{X_2}(z-x)\, dx.$$

Substituting the PDFs of the exponential distributions:

$$f_Z(z) = \int_0^z \beta e^{-\beta x} \cdot \beta e^{-\beta(z-x)}\, dx.$$

This simplifies to:

$$f_Z(z) = \beta^2 e^{-\beta z} \int_0^z dx = \beta^2 e^{-\beta z} \cdot z.$$

Thus,

$$f_Z(z) = \beta^2 z e^{-\beta z}, \quad z \geq 0.$$

Now, let's generalize this for k independent exponential random variables. We can derive it step by step: 1. Assume $Y_k = X_1 + X_2 + \ldots + X_k$. 2. We already have the result for $k = 2$:

$$f_{Y_2}(y) = \beta^2 y e^{-\beta y}.$$

Now, using the result for $k = 2$, we find the PDF of Y_k as follows:

$$f_{Y_k}(y) = \int_0^y f_{Y_{k-1}}(x) f_{X_k}(y-x)\, dx.$$

Substituting $f_{Y_{k-1}}(x)$ and $f_{X_k}(y-x)$:

$$f_{Y_k}(y) = \int_0^y \left(\frac{\beta^{k-1}}{\Gamma(k-1)} x^{k-2} e^{-\beta x} \right) \left(\beta e^{-\beta(y-x)} \right) dx.$$

This simplifies to:

$$f_{Y_k}(y) = \beta^k e^{-\beta y} \int_0^y \frac{x^{k-2}}{\Gamma(k-1)}\, dx.$$

The integral evaluates to:

$$\int_0^y x^{k-2}\, dx = \frac{y^{k-1}}{k-1}.$$

Thus, we have:

$$f_{Y_k}(y) = \beta^k e^{-\beta y} \cdot \frac{y^{k-1}}{\Gamma(k)}.$$

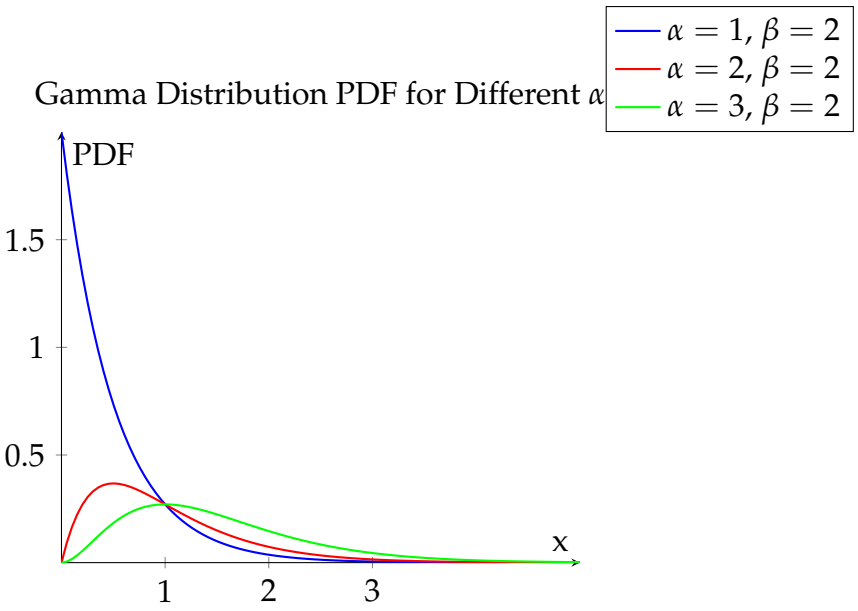
Thus, the PDF of $Y = X_1 + X_2 + \ldots + X_k$ is given by:

$$f_Y(y) = \frac{\beta^k}{\Gamma(k)} y^{k-1} e^{-\beta y}, \quad y \geq 0.$$

This is the PDF of the Gamma distribution with shape parameter k and rate parameter β. Therefore, we have proven that the Gamma distribution can be viewed as the sum of k independent exponentially distributed random variables. □

The shape of the Gamma distribution is highly influenced by its parameters α and β.

- For $\alpha = 1$: The Gamma distribution simplifies to an exponential distribution, characterized by a single peak.
- For $\alpha > 1$: The distribution becomes increasingly bell-shaped, with the peak shifting to the right as α increases.
- For $\alpha < 1$: The distribution is skewed to the right, approaching zero as x increases.



If X and Y are independent random variables each following a Gamma distribution with parameters (α_1, β) and (α_2, β) respectively, then the distribution of $X + Y$ is given by:

$$X + Y \sim \Gamma(\alpha_1 + \alpha_2, \beta)$$

This property is significant in various applications, such as queuing theory, where the total waiting time can be modeled using the sum of Gamma-distributed random variables.

Example 2.9. *Suppose a call center receives calls according to a Poisson process with an average rate of $\lambda = 5$ calls per hour. The waiting time until the k-th call can be modeled using a Gamma distribution.*

If we want to find the distribution of the waiting time for the 3rd call ($X \sim \Gamma(3, \lambda)$), we can determine the expected waiting time as:

$$E[X] = \frac{\alpha}{\beta} = \frac{3}{5} = 0.6 \text{ hours} \quad (\text{or } 36 \text{ minutes})$$

In this scenario, the Gamma distribution helps assess service times and manage resource allocation effectively.

Beta Distribution

The Beta distribution is a continuous probability distribution defined on the interval $[0, 1]$. It is particularly useful in modeling random variables that represent proportions or probabilities. The motivation behind the Beta distribution arises from the need to model outcomes that are constrained within a finite range. The Beta distribution has widespread applications, particularly in Bayesian statistics, project management (e.g., PERT charts), and quality control. For instance, in project management, the Beta distribution can model the completion time of a project where the completion time is uncertain.

The Beta distribution is characterized by two shape parameters, α and β, which control the distribution's shape. The parameters represent the number of successes and failures, respectively, in a binomial distribution. Before delving into the Beta distribution, it's essential to understand the Beta function, which serves as the normalization constant for the Beta distribution.

The Beta function, denoted as $B(x, y)$, is defined as:

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt$$

This integral arises in various areas of mathematics, particularly in calculus and combinatorial analysis. The Beta function can be related to the Gamma function via the identity:

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

The Beta function is used to normalize the Beta distribution, ensuring that the total area under the probability density function (PDF) equals 1.

Definition 2.20. *The PDF of the Beta distribution is given by:*

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \quad \text{for } 0 < x < 1$$

where $\alpha > 0$ and $\beta > 0$.

Proof. The PDF of the Beta distribution, denoted as $f(x; \alpha, \beta)$, can be expressed as a function that needs to be normalized. We propose a form:

$$f(x; \alpha, \beta) = C \cdot x^{\alpha-1}(1-x)^{\beta-1}$$

where C is a normalization constant.

To ensure that the PDF integrates to 1 over the interval $[0, 1]$, we must have:

$$\int_0^1 f(x; \alpha, \beta)\, dx = 1$$

Substituting our proposed form into this equation gives:

$$\int_0^1 C \cdot x^{\alpha-1}(1-x)^{\beta-1}\, dx = 1$$

We recognize the left-hand side as the definition of the Beta function:

$$C \cdot B(\alpha, \beta) = 1$$

Thus, we find the normalization constant C:

$$C = \frac{1}{B(\alpha, \beta)}$$

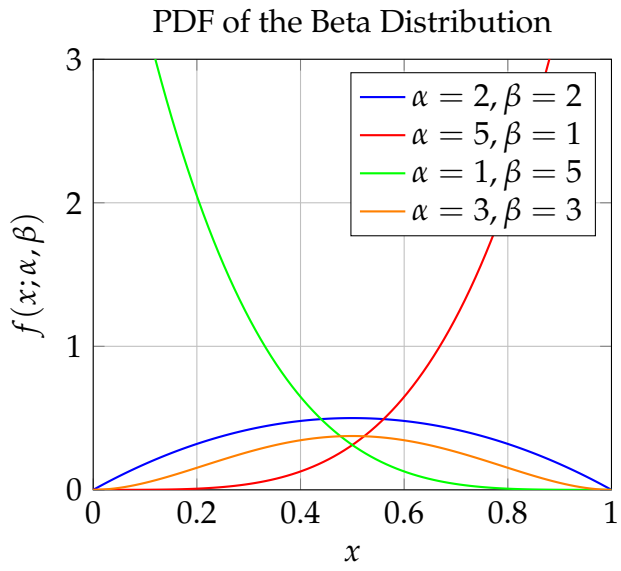
Substituting C back into our proposed PDF gives:

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \cdot x^{\alpha-1}(1-x)^{\beta-1}$$

□

The shape of the Beta distribution can vary significantly depending on the parameters α and β:

- If $\alpha = 1$ and $\beta = 1$, the distribution is uniform on $[0, 1]$.
- If $\alpha > 1$ and $\beta = 1$, the distribution is increasing.
- If $\alpha = 1$ and $\beta > 1$, the distribution is decreasing.
- If $\alpha > 1$ and $\beta > 1$, the distribution is bell-shaped, peaking around $\frac{\alpha-1}{\alpha+\beta-2}$.
- If $\alpha < 1$ and $\beta < 1$, the distribution has U-shape.



For example, when $\alpha = 2$ and $\beta = 2$, the distribution is symmetric around 0.5, while $\alpha = 0.5$ and $\beta = 0.5$ create a U-shaped distribution.

If X and Y are independent random variables that follow a Beta distribution, specifically $X \sim \text{Beta}(\alpha_1, \beta_1)$ and $Y \sim \text{Beta}(\alpha_2, \beta_2)$, the distribution of $X + Y$ does not follow a standard distribution. However, under specific conditions (e.g., when both variables are scaled), the result can be approximated or analyzed using convolution techniques.

In general, the sum of two Beta variables requires numerical methods or simulations for precise analysis.

Example 2.10. *Suppose we have a project that requires three estimates: optimistic time (5 days), pessimistic time (15 days), and most likely time (10 days). We can use the Beta distribution to model the likelihood of completing the project within a certain timeframe.*

Let α and β be determined based on these estimates using the formula:

$$\alpha = \frac{(m-a)(b-m)}{(b-a)^2}, \quad \beta = \frac{(b-m)(m-a)}{(b-a)^2}$$

where m is the most likely time, a is the optimistic time, and b is the pessimistic time.

In this case, the application of the Beta distribution helps project managers evaluate risk and optimize project timelines based on uncertainty.

Chi-squared Distribution

The Chi-squared distribution is a continuous probability distribution that arises in statistical inference, particularly in hypothesis testing and confidence interval estimation for population variance. It is commonly used to model the distribution of the sum of the squares of k independent standard normal random variables.

To understand the motivation behind the Chi-squared distribution, consider that many statistical tests rely on estimating variances from sample data. When you collect data and calculate how far each data point is from the mean, you are effectively measuring variability. The Chi-squared distribution helps us determine whether the observed variability in our sample data is consistent with a certain population variance or if it indicates something unusual.

The parameters involved in this modeling include:

- The number of degrees of freedom (k), which corresponds to the number of independent standard normal variables being squared and summed.

Definition 2.21. *The PDF of the Chi-squared distribution with k degrees of freedom is given by:*

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2}, \quad \text{for } x > 0$$

where Γ is the Gamma function.

Proof. The Chi-squared distribution is defined as the distribution of the sum of the squares of k independent standard normal random variables. This derivation will illustrate how we arrive at the probability density function (PDF) of the Chi-squared distribution.

Let $Z \sim N(0,1)$ be a standard normal random variable, which has the following PDF:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad \text{for } -\infty < z < \infty.$$

Consider k independent standard normal random variables:

$$Z_1, Z_2, \ldots, Z_k \sim N(0, 1.$$

The Chi-squared variable is defined as:

$$Y = Z_1^2 + Z_2^2 + \ldots + Z_k^2.$$

To find the PDF of Y, we start with the joint distribution of the Z_i. The joint PDF of $Z_1, Z_2, \ldots, Z_k$ is given by:

$$f_{Z_1,Z_2,\ldots,Z_k}(z_1, z_2, \ldots, z_k) = \prod_{i=1}^{k} f_Z(z_i) = \left(\frac{1}{\sqrt{2\pi}}\right)^k e^{-\frac{1}{2}\sum_{i=1}^{k} z_i^2}.$$

Next, we make a change of variables from $Z_1, Z_2, \ldots, Z_k$ to Y and the angles in polar coordinates, where:

$$Y = Z_1^2 + Z_2^2 + \ldots + Z_k^2$$

and the Jacobian determinant of the transformation will involve the angles.
The relationship in polar coordinates leads to:

$$z_1^2 + z_2^2 + \ldots + z_k^2 = r^2$$

where $r^2 = Y$. The angle terms introduce a factor of $\frac{(2\pi)^{k/2}}{(r^{k-1})}$ due to the transformation to polar coordinates. We can then express the PDF of Y as follows:

$$f_Y(y) = \int_{\mathbb{R}^{k-1}} f_{Z_1,Z_2,\ldots,Z_k}(z_1, z_2, \ldots, z_k)\, dz_1 dz_2 \ldots dz_{k-1}.$$

Substituting the expression for $f_{Z_1,Z_2,\ldots,Z_k}$ and integrating out the angular parts gives:

$$f_Y(y) = \frac{1}{(2\pi)^{k/2}} \int_0^{2\pi} e^{-\frac{y}{2}} y^{(k/2)-1} \frac{(2\pi)^{k/2}}{(r^{k-1})} dr.$$

After performing the integration, we arrive at:

$$f_Y(y) = \frac{1}{2^{k/2}\Gamma(k/2)} y^{(k/2)-1} e^{-y/2}, \quad y > 0.$$

We can then express the PDF of Y as follows:

$$f_Y(y) = \int_{\mathbb{R}^{k-1}} f_{Z_1,Z_2,\ldots,Z_k}(z_1, z_2, \ldots, z_k)\, dz_1 dz_2 \ldots dz_{k-1}.$$

Substituting the expression for $f_{Z_1,Z_2,\ldots,Z_k}$ and integrating out the angular parts gives:

$$f_Y(y) = \frac{1}{(2\pi)^{k/2}} \int_0^{2\pi} e^{-\frac{y}{2}} y^{(k/2)-1} \frac{(2\pi)^{k/2}}{(r^{k-1})} dr.$$

After performing the integration, we arrive at:

$$f_Y(y) = \frac{1}{2^{k/2}\Gamma(k/2)} y^{(k/2)-1} e^{-y/2}, \quad y > 0.$$

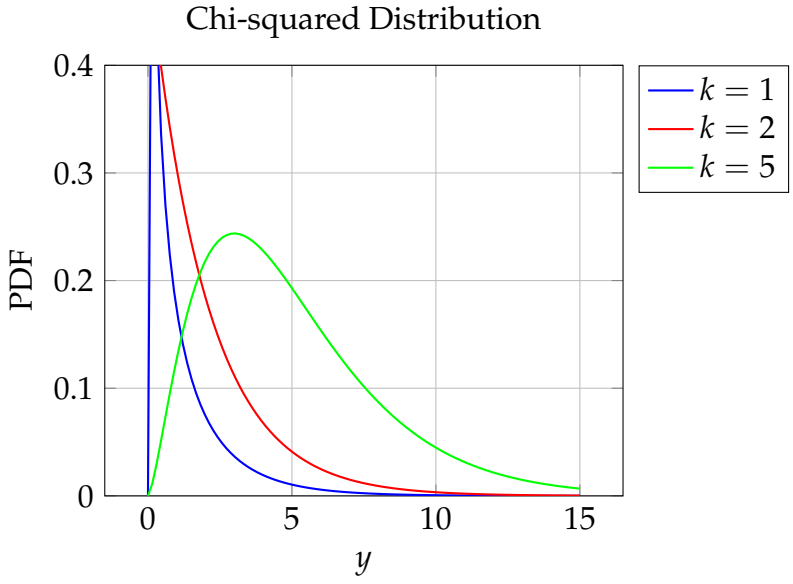
Thus, we have derived the PDF of the Chi-squared distribution with k degrees of freedom:

$$f_Y(y) = \frac{1}{2^{k/2}\Gamma(k/2)} y^{(k/2)-1} e^{-y/2}, \quad y > 0.$$

□

The shape of the Chi-squared distribution depends significantly on the degrees of freedom (k):

- For $k = 1$: The distribution is highly skewed to the right.
- For $k = 2$: The distribution starts to take on a more pronounced bell shape.
- As k increases, the distribution becomes more symmetric and approaches a normal distribution *(due to central limit theorem - will be studied later)*.



For example:

- For $k = 1$, the peak is at $x = 0$, and it falls off steeply.
- For $k = 5$, the distribution is more spread out and has a more pronounced peak, making it resemble a normal distribution.

If $X \sim \chi^2(k_1)$ and $Y \sim \chi^2(k_2)$ are independent Chi-squared random variables, then the sum $Z = X + Y$ follows a Chi-squared distribution with $k_1 + k_2$ degrees of freedom:

$$Z \sim \chi^2(k_1 + k_2)$$

Example 2.11. *Consider a researcher testing whether a six-sided die is fair. They roll the die 60 times and record the frequency of each outcome. If the observed frequencies are significantly different from the expected frequencies (10 for each side), the researcher can use the Chi-squared test to determine whether the die is biased. The test statistic is calculated as:*

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency and E_i is the expected frequency. By comparing this statistic to the Chi-squared distribution with appropriate degrees of freedom, the researcher can draw conclusions about the fairness of the die.

Student's t-Distribution

When we collect data from a population, we often want to estimate the mean and understand the variability of that mean. However, if our sample size is small (typically less than 30), the sampling distribution of the sample mean follows a t-distribution rather than a normal distribution.

The parameters involved in this modeling include:

- The sample size n
- The sample mean $\bar{x}$
- The sample standard deviation s

Definition 2.22. *The PDF of the Student's t-distribution is defined as:*

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}}$$

where v is the degrees of freedom and Γ is the gamma function.

Proof. The Student's t-distribution arises when we estimate the mean of a normally distributed population with an unknown standard deviation. To derive its probability density function (PDF), we start with the following components.

Given a sample of n observations $X_1, X_2, \ldots, X_n$, the sample mean $\bar{X}$ and sample standard deviation S are defined as:

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

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

The t-statistic is defined as:

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

where μ is the true population mean.

If the underlying population is normally distributed, the numerator $\bar{X} - \mu$ follows a normal distribution with mean 0 and variance $\frac{\sigma^2}{n}$, where σ^2 is the population variance. The denominator S is independent of $\bar{X}$ and follows a scaled chi-squared distribution.

Assuming X_i follows a normal distribution $\mathcal{N}(\mu, \sigma^2)$:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

The sample variance S^2 follows a chi-squared distribution:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

The joint distribution of $\bar{X}$ and S^2 can be expressed using their independence:

$$f_{\bar{X},S^2}(x, s^2) = f_{\bar{X}}(x) f_{S^2}(s^2).$$

We need to express the distribution in terms of the t-statistic t. By substituting t in terms of $\bar{X}$ and S:

$$t = \frac{\sqrt{n}(\bar{X} - \mu)}{S}.$$

To derive the PDF of t, we must consider the transformation of variables and compute the Jacobian. The Jacobian of the transformation from (X, S) to (t, S) is given by:

$$J = \frac{\partial(t, S)}{\partial(\bar{X}, S)} = \frac{\sqrt{n}}{S}.$$

The PDF of the t-distribution can be derived using the joint distribution and the transformation:

1. The PDF of $\bar{X}$ is given by:

$$f_{\bar{X}}(x) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-\frac{n(x-\mu)^2}{2\sigma^2}\right).$$

2. The PDF of S^2 follows a chi-squared distribution:

$$f_{S^2}(s^2) = \frac{(n-1)^{(n-1)/2}}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{2\pi s^2}} \left(\frac{s^2}{\sigma^2}\right)^{(n-3)/2} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right).$$

3. Combining these and accounting for the Jacobian, we arrive at the PDF of the Student's t-distribution:

$$f(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n-1}{2}\right)} \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}}.$$

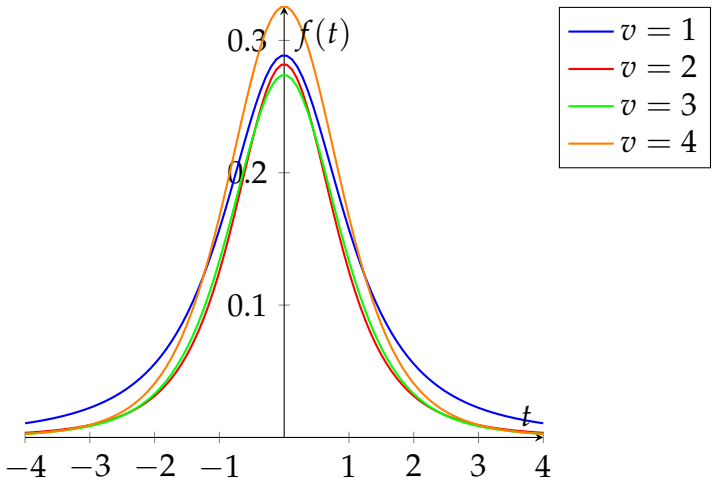
Thus, the PDF of the Student's t-distribution with $n-1$ degrees of freedom is given by:

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}},$$

where $v = n - 1$ is the degrees of freedom. □

The shape of the Student's t-distribution is similar to the normal distribution but has heavier tails. This characteristic allows it to better account for the variability in estimates when sample sizes are small.

Student's t-distribution with different degrees of freedom



The parameter v (degrees of freedom) greatly affects the distribution's shape:

- As v increases, the distribution approaches the normal distribution.
- Smaller values of v lead to heavier tails, indicating a higher likelihood of extreme values.

For example, with $v = 1$, the distribution has very heavy tails, while with $v = 30$, it closely resembles the standard normal distribution.

If X and Y are independent random variables following a Student's t-distribution with v degrees of freedom, then the distribution of $Z = X + Y$ is not straightforward. Generally, the sum of two independent t-distributed variables does not follow a t-distribution unless specific conditions are met. However, if both have the same degrees of freedom, their sum can be approximated by a normal distribution for large v.

Example 2.12. *A common application of the Student's t-distribution is in the context of the two-sample t-test, which is used to determine whether two population means are significantly different from each other.*

For example, suppose a researcher wants to compare the average heights of students in two different classes:

1. *Collect heights from Class A (sample size n_A) and Class B (sample size n_B).*
2. *Calculate the sample means $\bar{x}_A$ and $\bar{x}_B$, and the standard deviations s_A and s_B.*
3. *Use the two-sample t-test to evaluate the hypothesis that $\mu_A = \mu_B$:*

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

4. Compare the calculated t-value with the critical t-value from the t-distribution table with $v = n_A + n_B - 2$ degrees of freedom.

If the calculated t-value exceeds the critical t-value, the null hypothesis is rejected, indicating a significant difference in the average heights of the two classes.

2.4 Moments

Let X be a random variable.

Definition 2.23.

- *The mean, also called the expected value, is the average value of the random variable. It is defined as:*

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

where $f_X(x)$ is the probability density function (pdf) of X in the continuous case, or

$$\mu = \mathbb{E}[X] = \sum_{x \in Range(X)} x P(X = x)$$

in the discrete case.

- *The median is the value that separates the higher half from the lower half of the distribution. Formally, the median m satisfies:*

$$P(X \leq m) = 0.5$$

In other words, 50% of the probability mass is below the median.

- *The mode is the value of X that occurs with the highest frequency, or the value at which the probability density (or mass) function achieves its maximum:*

$$Mode = \arg\max_x f_X(x)$$

where $f_X(x)$ is the pdf in the continuous case or the probability mass function (pmf) in the discrete case.

Definition 2.24. *Let X be a random variable with mean $\mu = E(X)$. The variance of X, denoted as $Var(X)$ or σ^2, is defined as:*

$$Var(X) = E\left[(X - \mu)^2\right] = E(X^2) - (E(X))^2.$$

Variance measures the expected squared deviation of X from its mean μ.

Theorem 2.6. *Let X be a random variable with mean μ, and let m be a median of X.*

- *The value of c that minimizes the mean squared error $E[(X - c)^2]$ is $c = \mu$.*

- *A value of c that minimizes the mean absolute error $E[|X - c|]$ is $c = m$.*

Proof. **(i) Minimizing the Mean Squared Error:**

We want to minimize the mean squared error function:

$$E[(X - c)^2] = \int_{-\infty}^{\infty} (x - c)^2 f_X(x) dx,$$

where $f_X(x)$ is the probability density function of X. Expanding the square, we have:

$$E[(X - c)^2] = \int_{-\infty}^{\infty} \left((x - \mu)^2 + 2(x - \mu)(\mu - c) + (\mu - c)^2 \right) f_X(x) dx.$$

Taking the derivative with respect to c and setting it to zero to find the minimizing value of c:

$$\frac{d}{dc} E[(X - c)^2] = -2E[X - c] = -2(\mu - c).$$

Setting this equal to zero, we obtain $c = \mu$. Therefore, the value of c that minimizes $E[(X - c)^2]$ is the mean μ.

(ii) Minimizing the Mean Absolute Error:

Next, we want to minimize the mean absolute error function:

$$E[|X - c|] = \int_{-\infty}^{\infty} |x - c| f_X(x) dx.$$

To minimize this, we take the derivative of $E[|X - c|]$ with respect to c and analyze the behavior of the function. The derivative is given by:

$$\frac{d}{dc} E[|X - c|] = \int_{-\infty}^{\infty} \text{sign}(x - c) f_X(x) dx,$$

where $\text{sign}(x - c)$ is -1 for $x < c$ and 1 for $x > c$.
For the derivative to be zero, we require that the proportion of values of X less than c equals the proportion of values greater than c, i.e.,

$$P(X \leq c) = P(X \geq c) = \frac{1}{2}.$$

Thus, c must be a median of the distribution of X. Therefore, the value of c that minimizes $E[|X - c|]$ is a median m of X. □

2.4.1 Interpreting Moments

Definition 2.25. *Let X be a random variable with mean μ and variance σ^2. For any positive integer n, we define the following:*

- *The nth moment of X is $E(X^n)$.*
- *The nth central moment of X is $E((X - \mu)^n)$.*

- *The nth standardized moment of X is $E\left(\left(\frac{X-\mu}{\sigma}\right)^n\right)$.*

In the previous sentence, "if it exists" is left implicit.

In probability and statistics, the term *moment* is borrowed from physics, where moments describe the distribution of mass at a distance from a reference point. For a random variable (r.v.) X with mean μ and variance σ^2, moments offer insight into the characteristics of the distribution of X. In particular, the mean is the first moment, and the variance is the second central moment.

Let X be a discrete random variable with distinct possible values $x_1, x_2, \ldots, x_n$, and imagine a system where pebbles with masses $m_j = P(X = x_j)$ are placed at each x_j on a number line. In this analogy, the mean $E(X)$ corresponds to the *center of mass* of the system, and the variance $\mathrm{Var}(X)$ corresponds to the *moment of inertia* about the center of mass.

In physics, the center of mass is given by:

$$E(X) = \sum_{j=1}^{n} m_j x_j,$$

while the moment of inertia about the center of mass is:

$$\mathrm{Var}(X) = \sum_{j=1}^{n} m_j (x_j - E(X))^2.$$

Thus, the mean (first moment) of a random variable corresponds to the center of mass of a system of pebbles, and the variance (second central moment) corresponds to the moment of inertia about this center of mass.

We now introduce the concept of skewness, which provides a single-number summary of asymmetry.

Definition 2.26. *Skewness is based on the third moment and is defined as the third standardized moment of a random variable X with mean μ and variance σ^2. The skewness is given by:*

$$Skew(X) = E\left(\left(\frac{X-\mu}{\sigma}\right)^3\right).$$

Standardizing by σ ensures that the skewness does not depend on the scale or location of X, as μ and σ already provide that information. This makes skewness invariant to the units in which X is measured (e.g., inches versus meters).

To understand how skewness measures asymmetry, we need to first define symmetry in terms of random variables.

Definition 2.27. *We say that a random variable X has a symmetric distribution about μ if $X - \mu$ has the same distribution as $\mu - X$. This implies:*

$$P(X \leq \mu) = P(X \geq \mu).$$

For continuous random variables with probability density function (PDF) f, symmetry about μ can be expressed as:

$$f(x) = f(2\mu - x), \quad \text{for all } x.$$

The third standardized moment is taken as the definition of skewness because the first standardized moment is always zero. Positive skewness indicates a distribution with a long right tail relative to the left tail, while negative skewness indicates the reverse. Although higher standardized moments (like the fifth) could also measure skewness, the third standardized moment is typically easier to calculate and estimate from data.

In addition to skewness, another important characteristic of a distribution is the heaviness of its tails. For a fixed variance, the question arises: is the variability driven by rare extreme events or by more frequent moderate deviations from the mean? This is a key consideration in risk management, especially in finance, where distributions with heavy tails (e.g., returns with heavy left tails due to rare but severe crises) must be accounted for to avoid disastrous consequences, such as those seen in the 2008 financial crisis.

As with measuring skewness, no single measure can perfectly capture tail behavior, but there is a widely used summary based on the fourth standardized moment. This measure is known as *kurtosis*.

Definition 2.28. *The kurtosis of a random variable X with mean μ and variance σ^2 is a shifted version of the fourth standardized moment of X:*

$$Kurt(X) = E\left(\left(\frac{X-\mu}{\sigma}\right)^4\right) - 3.$$

The reason for subtracting 3 is that it adjusts the kurtosis of a normal distribution to 0. In other words, a normal distribution is the benchmark, and distributions with kurtosis greater than 0 are said to be *leptokurtic* (having heavier tails), while those with kurtosis less than 0 are *platykurtic* (having lighter tails).

Example 2.13. *Let $X \sim N(\mu, \sigma^2)$, that is, X follows a normal distribution with mean μ and variance σ^2. We want to compute the kurtosis of X.*

We start by computing the fourth standardized moment:

$$E\left(\left(\frac{X-\mu}{\sigma}\right)^4\right).$$

Since $\frac{X-\mu}{\sigma} \sim N(0,1)$, the random variable $Z = \frac{X-\mu}{\sigma}$ is standard normal. The fourth moment of a standard normal distribution is given by:

$$E(Z^4) = 3.$$

Thus, the fourth standardized moment of X is:

$$E\left(\left(\frac{X-\mu}{\sigma}\right)^4\right) = 3.$$

Now, applying the kurtosis formula:

$$Kurt(X) = 3 - 3 = 0.$$

Therefore, the kurtosis of any normal distribution is 0.

2.4.2 Moment Generating Functions

Before studying *Moment Generating Functions*, I recommend you to first understand the concept of a *generating function*. For that, I recommend watching this amazing video by Grant Sanderson on his YouTube channel, 3Blue1Brown, titled *Olympiad Level Counting*.[1]

Definition 2.29. *Let X be a random variable. The moment generating function (MGF) of X, denoted by $M_X(t)$, is defined as:*

$$M_X(t) = E\left(e^{tX}\right), \quad \textit{for all } t \in \mathbb{R} \textit{ such that } E\left(e^{tX}\right) \textit{ exists.}$$

The MGF can be used to derive the moments of X. In particular, the n-th moment of X is given by:

$$E(X^n) = \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0}.$$

You might be wondering here - *What's the interpretation of 't' in the expression of MGF?* The answer is - nothing! It's just a placeholder. Like you saw in the above example by 3Blue1Brown that x was just a variable and had no interpretation, so is t for us here. Using the concept of moments, we can revisit the ideas discussed above:

The *mean* (or expected value) of the random variable X, denoted by $\mu = E(X)$, is the first moment:

$$\mu = M_X'(0) = \frac{d}{dt} M_X(t)\bigg|_{t=0}.$$

The *variance* of X, denoted by $\sigma^2 = \text{Var}(X)$, is the second central moment and can be expressed as:

$$\sigma^2 = E\left((X-\mu)^2\right) = M_X''(0) - \left(M_X'(0)\right)^2.$$

The *skewness* of X, denoted by $\text{Skew}(X)$, is a measure of the asymmetry of the distribution and is defined using the third standardized moment:

$$\text{Skew}(X) = \frac{E\left((X-\mu)^3\right)}{\sigma^3} = \frac{M_X'''(0) - 3M_X'(0)M_X''(0) + 2(M_X'(0))^3}{\sigma^3}.$$

The *kurtosis* of X, denoted by $\text{Kurt}(X)$, measures the "tailedness" of the distribution and is defined using the fourth standardized moment:

$$\text{Kurt}(X) = \frac{E\left((X-\mu)^4\right)}{\sigma^4} - 3 = \frac{M_X^{(4)}(0) - 4M_X'(0)M_X'''(0) + 6M_X'(0)M_X''(0)^2 - 3(M_X'(0))^4}{\sigma^4} - 3.$$

Binomial Distribution

The moment generating function (MGF) of X is defined as:

$$M_X(t) = E\left(e^{tX}\right) = \sum_{k=0}^{n} e^{tk} P(X = k).$$

Substitute the PMF of the binomial distribution into the definition of the MGF:

$$M_X(t) = \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k}.$$

Rearranging the terms:

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

Notice that this is the binomial expansion of $(1 - p + pe^t)^n$, which can be written as:

$$M_X(t) = \left(1 - p + pe^t\right)^n.$$

Mean: The mean $\mu = E(X)$ is the first derivative of the MGF evaluated at $t = 0$:

$$\mu = M_X'(0) = \frac{d}{dt}\left(1 - p + pe^t\right)^n \Big|_{t=0}.$$

Differentiating the MGF:

$$M_X'(t) = n \cdot \left(1 - p + pe^t\right)^{n-1} \cdot pe^t.$$

Now, evaluate this at $t = 0$:

$$M_X'(0) = n \cdot (1 - p + p \cdot 1)^{n-1} \cdot p \cdot 1 = n \cdot p.$$

Therefore, the mean is:

$$\mu = E(X) = np.$$

Variance: The variance $\sigma^2 = \text{Var}(X)$ is the second central moment, which can be computed as:

$$\text{Var}(X) = M_X''(0) - \left(M_X'(0)\right)^2.$$

First, we compute the second derivative of the MGF:

$$M_X''(t) = n \cdot \left[n - 1 \cdot \left(1 - p + pe^t\right)^{n-2} \cdot (pe^t)^2 + \left(1 - p + pe^t\right)^{n-1} \cdot pe^t\right].$$

Now, evaluate this at $t = 0$:

$$M_X''(0) = n \cdot \left[(n-1) \cdot (1 - p + p \cdot 1)^{n-2} \cdot p^2 + (1 - p + p \cdot 1)^{n-1} \cdot p\right].$$

Simplifying:

$$M_X''(0) = n \cdot \left[(n-1) \cdot p^2 + p\right] = n \cdot p \cdot [(n-1) \cdot p + 1] = np(1 + (n-1)p).$$

Therefore, the variance is:

$$\text{Var}(X) = M_X''(0) - (M_X'(0))^2 = np(1 + (n-1)p) - (np)^2.$$

Simplifying further:

$$\text{Var}(X) = np(1-p).$$

Poisson Distribution

Let X be a random variable that follows a Poisson distribution with parameter $\lambda > 0$, i.e.,

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

The moment generating function (MGF) of X is defined as:

$$M_X(t) = E\left(e^{tX}\right) = \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!}.$$

This is a Taylor series expansion of $e^{\lambda e^t}$, so the MGF becomes:

$$M_X(t) = e^{\lambda(e^t - 1)}.$$

Mean: The mean (or expected value) $\mu = E(X)$ can be calculated as the first derivative of the MGF evaluated at $t = 0$:

$$\mu = M_X'(0) = \frac{d}{dt} M_X(t)\Big|_{t=0}.$$

First, compute the derivative of $M_X(t)$:

$$M_X'(t) = \lambda e^t e^{\lambda(e^t - 1)} = \lambda e^{t + \lambda(e^t - 1)}.$$

Evaluating at $t = 0$:

$$M_X'(0) = \lambda e^{\lambda(1-1)} = \lambda.$$

Thus, the mean of the Poisson distribution is:

$$E(X) = \lambda.$$

Variance: The variance $\sigma^2 = \mathrm{Var}(X)$ is the second central moment and can be calculated as:

$$\sigma^2 = M_X''(0) - (M_X'(0))^2.$$

First, compute the second derivative of $M_X(t)$:

$$M_X''(t) = \lambda e^t \left(\lambda e^t + 1\right) e^{\lambda(e^t - 1)} = \lambda e^{t + \lambda(e^t - 1)} \left(\lambda e^t + 1\right).$$

Evaluating at $t = 0$:

$$M_X''(0) = \lambda \left(\lambda + 1\right).$$

Thus, the variance of the Poisson distribution is:

$$\mathrm{Var}(X) = M_X''(0) - (M_X'(0))^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda.$$

Hypergeometric Distribution

The MGF of the Hypergeometric distribution is not as straightforward to calculate due to the lack of independence in sampling without replacement. However, the mean and variance can still be derived from the combinatorial properties of the distribution as shown above.

Mean:
To compute the mean $E(X)$, we use the fact that the Hypergeometric distribution can be thought of as sampling without replacement. Let X_i be an indicator variable for the i-th trial, such that:

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th draw is a success,} \\ 0 & \text{otherwise.} \end{cases}$$

Then, the total number of successes, X, is the sum of these indicator variables:

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

Since the probability of success in each trial changes because we are sampling without replacement, the expectation of each X_i is:

$$E(X_i) = \frac{K}{N}.$$

Thus, the expected number of successes, or the mean of X, is:

$$E(X) = \sum_{i=1}^{n} E(X_i) = n \cdot \frac{K}{N}.$$

This result can be interpreted as the product of the sample size n and the proportion of successes in the population $\frac{K}{N}$, i.e., the expected number of successes in the sample.

Variance:
To compute the variance $\text{Var}(X)$, we first compute $E(X^2)$. From the definition of variance:

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

First, observe that:

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

so:

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

Expanding the square:

$$E(X^2) = E\left(\sum_{i=1}^{n} X_i^2 + 2 \sum_{1 \le i < j \le n} X_i X_j\right).$$

Since $X_i^2 = X_i$ (because X_i is an indicator variable):

$$E(X^2) = \sum_{i=1}^{n} E(X_i) + 2 \sum_{1 \le i < j \le n} E(X_i X_j).$$

We already know $E(X_i) = \frac{K}{N}$. Now, for $E(X_i X_j)$, the probability that both the i-th and j-th draws are successes is:

$$E(X_i X_j) = \frac{K}{N} \cdot \frac{K-1}{N-1}.$$

Thus:

$$E(X^2) = n \cdot \frac{K}{N} + 2 \cdot \binom{n}{2} \cdot \frac{K}{N} \cdot \frac{K-1}{N-1}.$$

Simplifying the second term:

$$E(X^2) = n \cdot \frac{K}{N} + n(n-1) \cdot \frac{K}{N} \cdot \frac{K-1}{N-1}.$$

Now, using the formula for variance:

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

we substitute $E(X) = \frac{nK}{N}$ and $E(X^2)$ from above:

$$\text{Var}(X) = n \cdot \frac{K}{N} + n(n-1) \cdot \frac{K}{N} \cdot \frac{K-1}{N-1} - \left(\frac{nK}{N}\right)^2.$$

Simplifying the terms:

$$\text{Var}(X) = \frac{nK}{N} \left(\frac{N-K}{N}\right) \cdot \frac{N-n}{N-1}.$$

This expression includes the term $\frac{N-n}{N-1}$, which is a finite population correction factor that accounts for sampling without replacement.

Uniform Distribution

The moment generating function $M_X(t)$ of a uniformly distributed random variable X is defined as:

$$M_X(t) = E\left(e^{tX}\right) = \int_a^b e^{tx} f_X(x)\, dx.$$

Substituting the PDF of X, we get:

$$M_X(t) = \frac{1}{b-a} \int_a^b e^{tx}\, dx.$$

Evaluating this integral:

$$M_X(t) = \frac{1}{b-a} \left[\frac{e^{tx}}{t}\right]_a^b = \frac{1}{b-a} \left(\frac{e^{tb} - e^{ta}}{t}\right), \quad t \neq 0.$$

For $t = 0$, $M_X(0) = 1$ (since the MGF at $t = 0$ must equal 1 for any distribution).

Mean: To find the mean $\mu = E(X)$, we differentiate $M_X(t)$ with respect to t and evaluate at $t = 0$:

$$\mu = M_X'(0) = \frac{d}{dt}\left[\frac{1}{b-a}\left(\frac{e^{tb} - e^{ta}}{t}\right)\right]\Bigg|_{t=0}.$$

Using L'Hopital's rule to evaluate the limit as $t \to 0$, we get:

$$\mu = \frac{b+a}{2}.$$

Thus, the mean of a uniform distribution is:

$$E(X) = \frac{b+a}{2}.$$

Variance: To find the variance $\text{Var}(X) = \sigma^2$, we first compute the second moment $E(X^2)$ by differentiating $M_X(t)$ twice and evaluating at $t = 0$:

$$E(X^2) = M_X''(0) = \frac{d^2}{dt^2}\left[\frac{1}{b-a}\left(\frac{e^{tb} - e^{ta}}{t}\right)\right]\Bigg|_{t=0}.$$

Alternatively, we can compute the second moment directly from the definition:

$$E(X^2) = \int_a^b x^2 f_X(x)\,dx = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{b^3 - a^3}{3(b-a)}.$$

Simplifying, we get:

$$E(X^2) = \frac{a^2 + ab + b^2}{3}.$$

Now, using the formula for variance $\text{Var}(X) = E(X^2) - (E(X))^2$, we find:

$$\text{Var}(X) = \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

Thus, the variance of a uniform distribution is:

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

Exponential Distribution

Let X be a random variable with the Exponential distribution, parameterized by the rate parameter $\lambda > 0$. The probability density function (PDF) of X is given by:

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

The moment generating function (MGF) of X is defined as:

$$M_X(t) = E\left(e^{tX}\right) = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx.$$

This integral simplifies as follows:

$$M_X(t) = \lambda \int_0^\infty e^{x(t-\lambda)}dx.$$

For convergence, we require $t < \lambda$. The integral can be computed:

$$M_X(t) = \lambda \left[\frac{1}{-(t-\lambda)}e^{x(t-\lambda)}\right]_0^\infty = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

Mean:
The mean $\mu = E(X)$ is the first moment and is obtained by differentiating the MGF with respect to t and evaluating at $t = 0$:

$$\mu = M_X'(0) = \frac{d}{dt}\left(\frac{\lambda}{\lambda - t}\right)\Big|_{t=0}.$$

Differentiating, we find:

$$M_X'(t) = \frac{\lambda}{(\lambda - t)^2}.$$

Thus,

$$\mu = M_X'(0) = \frac{1}{\lambda}.$$

Therefore, the mean of the Exponential distribution is:

$$E(X) = \frac{1}{\lambda}.$$

Variance:
The variance $\sigma^2 = \text{Var}(X)$ is the second central moment, which can be calculated from the second derivative of the MGF:

$$M_X''(t) = \frac{2\lambda}{(\lambda - t)^3}.$$

The second moment is given by:

$$E(X^2) = M_X''(0) = \frac{2}{\lambda^2}.$$

Using the formula for variance:

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

Normal Distribution

Let $X \sim N(\mu, \sigma^2)$ be a normally distributed random variable with mean μ and variance σ^2.
The moment generating function $M_X(t)$ of X is defined as:

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

For the normal distribution, the MGF is given by:

$$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

Mean: The mean $E(X)$ is the first moment and can be calculated using the MGF:

$$E(X) = M_X'(0).$$

First, we differentiate the MGF:

$$M_X'(t) = \frac{d}{dt}\left(e^{\mu t + \frac{\sigma^2 t^2}{2}}\right) = e^{\mu t + \frac{\sigma^2 t^2}{2}}\left(\mu + \sigma^2 t\right).$$

Now, we evaluate this at $t = 0$:

$$M_X'(0) = e^{\mu \cdot 0 + \frac{\sigma^2 \cdot 0^2}{2}}\left(\mu + \sigma^2 \cdot 0\right) = e^0 \cdot \mu = \mu.$$

Thus, the mean of X is:

$$E(X) = \mu.$$

Variance: The variance $\mathrm{Var}(X)$ is calculated using the second moment:

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

We first need to find $E(X^2)$, which can be obtained from the MGF:

$$E(X^2) = M_X''(0).$$

Differentiating the MGF again:

$$M_X''(t) = \frac{d^2}{dt^2}\left(e^{\mu t + \frac{\sigma^2 t^2}{2}}\right) = e^{\mu t + \frac{\sigma^2 t^2}{2}}\left(\mu + \sigma^2 t\right)^2 + e^{\mu t + \frac{\sigma^2 t^2}{2}}\left(\mu + \sigma^2 t\right) \cdot \frac{\sigma^2}{2}$$

$$= e^{\mu t + \frac{\sigma^2 t^2}{2}}\left(\mu^2 + 2\mu\sigma^2 t + \sigma^4 t^2 + \sigma^2\right).$$

Now, evaluating at $t = 0$:

$$M_X''(0) = e^{\mu \cdot 0 + \frac{\sigma^2 \cdot 0^2}{2}}\left(\mu^2 + 0 + 0 + \sigma^2\right) = e^0 \cdot \left(\mu^2 + \sigma^2\right) = \mu^2 + \sigma^2.$$

Substituting back into the variance formula:

$$\mathrm{Var}(X) = E(X^2) - (E(X))^2 = \left(\mu^2 + \sigma^2\right) - \mu^2 = \sigma^2.$$

For the remaining distributions, derive *mean* and *variance* as an exercise.

Log-Normal Distribution

Mean:

$$E(X) = e^{\mu + \frac{\sigma^2}{2}}.$$

Variance:

$$\text{Var}(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}.$$

Weibull Distribution

Mean:

$$E(X) = \lambda \Gamma \left(1 + \frac{1}{k}\right).$$

Variance:

$$\text{Var}(X) = \lambda^2 \left(\Gamma \left(1 + \frac{2}{k}\right) - \left(\Gamma \left(1 + \frac{1}{k}\right)\right)^2\right).$$

Gamma Distribution

Mean:

$$E(X) = k\theta.$$

Variance:

$$\text{Var}(X) = k\theta^2.$$

Beta Distribution

Mean:

$$E(X) = \frac{\alpha}{\alpha + \beta}.$$

Variance:

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Chi-squared Distribution

The Chi-squared distribution with k degrees of freedom has the following properties.

Mean:

$$E(X) = k.$$

Variance:

$$\text{Var}(X) = 2k.$$

Student's t-Distribution

The Student's t-distribution with ν degrees of freedom has different properties depending on the degrees of freedom.

Mean:

$$E(X) = \begin{cases} 0 & \text{if } \nu > 1 \\ \text{undefined} & \text{if } \nu \leq 1 \end{cases}$$

Variance:

$$\text{Var}(X) = \begin{cases} \frac{\nu}{\nu-2} & \text{if } \nu > 2 \\ \text{undefined} & \text{if } \nu \leq 2 \end{cases}$$

2.5 Joint Distributions

Definition 2.30. *The joint probability mass function (joint PMF) of two discrete random variables X and Y, denoted as $P(X = x, Y = y)$, gives the probability that X takes on the value x and Y takes on the value y. The joint PMF must satisfy the following properties:*

$$P(X = x, Y = y) \geq 0 \quad \forall x, y,$$
$$\sum_x \sum_y P(X = x, Y = y) = 1.$$

Example 2.14. *Consider two dice rolls X and Y. The joint PMF can be represented as follows:*

$$P(X = x, Y = y) = \frac{1}{36} \quad \text{for } x, y \in \{1, 2, \ldots, 6\}.$$

This means that each combination of dice rolls has an equal probability of $\frac{1}{36}$.

Definition 2.31. *The marginal probability mass function (marginal PMF) for a discrete random variable X can be derived from the joint PMF by summing over all possible values of the other variable Y:*

$$P(X = x) = \sum_y P(X = x, Y = y).$$

Example 2.15. *Continuing with the dice example, the marginal PMF of X is given by:*

$$P(X = x) = \sum_{y=1}^{6} P(X = x, Y = y) = \sum_{y=1}^{6} \frac{1}{36} = \frac{6}{36} = \frac{1}{6} \quad \text{for } x \in \{1, 2, \ldots, 6\}.$$

Definition 2.32. *The conditional probability mass function (conditional PMF) describes the probability of X given that Y has occurred, denoted as $P(X = x \mid Y = y)$. It is calculated using the joint PMF as follows:*

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)},$$

provided that $P(Y = y) > 0$.

Example 2.16. *Using our dice example, if we want to find the conditional PMF of X given $Y = 3$:*

$$P(X = x \mid Y = 3) = \frac{P(X = x, Y = 3)}{P(Y = 3)}.$$

Calculating $P(Y = 3)$:

$$P(Y = 3) = \sum_{x=1}^{6} P(X = x, Y = 3) = \sum_{x=1}^{6} \frac{1}{36} = \frac{6}{36} = \frac{1}{6}.$$

Thus, the conditional PMF is:

$$P(X = x \mid Y = 3) = \frac{P(X = x, Y = 3)}{\frac{1}{6}} = P(X = x, Y = 3) \cdot 6 = \frac{6}{36} = \frac{1}{6} \quad \text{for } x \in \{1, 2, \ldots, 6\}.$$

Definition 2.33. *The joint probability density function (joint PDF) of two continuous random variables X and Y is a function $f_{X,Y}(x, y)$ that describes the likelihood of X and Y simultaneously taking on specific values. The joint PDF must satisfy the following properties:*

1. $f_{X,Y}(x, y) \geq 0$ *for all* x, y.
2. *The integral of the joint PDF over the entire space must equal 1:*

$$\iint_{\mathbb{R}^2} f_{X,Y}(x, y)\, dx\, dy = 1.$$

Consider X and Y representing the height and weight of individuals in a population. The joint PDF might be expressed as:

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}},$$

where μ_x, μ_y are the means and σ_x, σ_y are the standard deviations of the respective distributions.

Definition 2.34. *The marginal probability density function of a random variable is obtained by integrating the joint PDF over the other variable. For example, the marginal PDF of X is given by:*

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

Similarly, the marginal PDF of Y is:

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

Using the joint PDF example above, the marginal PDF of X can be computed as:

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

This gives the distribution of X regardless of Y.

Definition 2.35. *The conditional probability density function describes the probability of one random variable given the value of another. The conditional PDF of Y given $X = x$ is defined as:*

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

provided $f_X(x) > 0$.

Using the same joint PDF, the conditional PDF of Y given $X = x$ is:

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

This expresses how Y behaves when X is fixed at a certain value.

Example 2.17. *Let X and Y be two continuous random variables that follow a bivariate normal distribution with the following parameters:*

- *Means: $\mu_X = 0$, $\mu_Y = 0$*
- *Variances: $\sigma_X^2 = 1$, $\sigma_Y^2 = 1$*
- *Covariance: $Cov(X,Y) = \rho$ (where $-1 < \rho < 1$)*

The joint PDF of X and Y is given by:

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

The conditional PDF of Y given $X = x_0$ is derived from the joint PDF:

$$f_{Y|X}(y|x_0) = \frac{f_{X,Y}(x_0,y)}{f_X(x_0)}.$$

The formula for the conditional PDF of a bivariate normal distribution is:

$$f_{Y|X}(y|x_0) = \frac{1}{\sqrt{2\pi\sigma_Y^2(1-\rho^2)}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(y-\mu_{Y|X})^2}{\sigma_{Y|X}^2}\right)\right),$$

where:

$$\mu_{Y|X} = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x_0 - \mu_X),$$
$$\sigma_{Y|X}^2 = \sigma_Y^2(1-\rho^2).$$

Let's take:

- $\rho = 0.5$
- $x_0 = 1$

We can compute:
1. Mean of Y given $X = 1$:

$$\mu_{Y|X} = 0 + 0.5 \cdot \frac{1}{1}(1 - 0) = 0.5.$$

2. Variance of Y given $X = 1$:

$$\sigma^2_{Y|X} = 1 \cdot (1 - 0.5^2) = 1 \cdot 0.75 = 0.75.$$

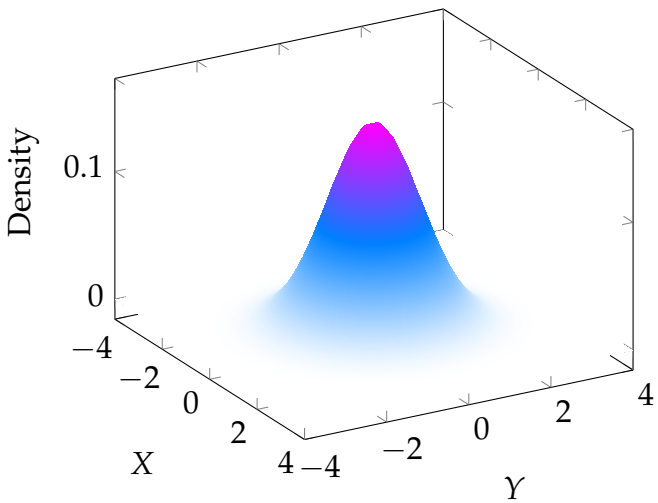
So the conditional PDF of Y given $X = 1$ is:

$$f_{Y|X}(y|1) = \frac{1}{\sqrt{2\pi \cdot 0.75}} \exp\left(-\frac{(y - 0.5)^2}{2 \cdot 0.75}\right).$$

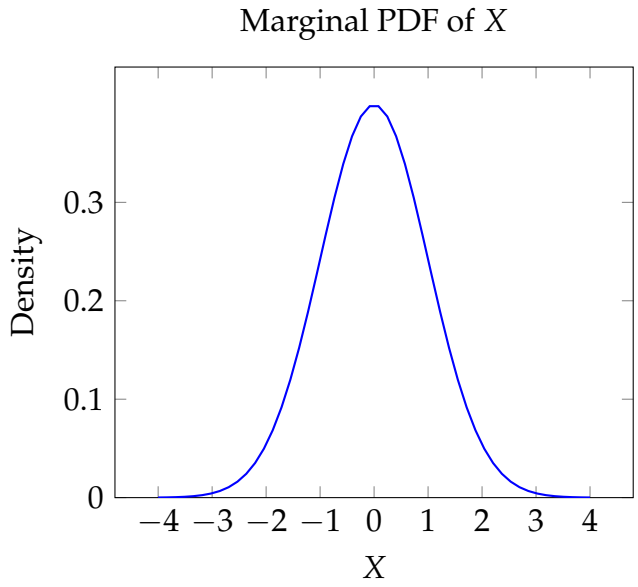
This describes the distribution of Y when X is fixed at 1. The resulting distribution is still normal, with mean 0.5 and variance 0.75.

In the case of a bivariate normal distribution, the joint PDF can be visualized as a 3D surface where the height of the surface at any point (x, y) represents the density of the joint occurrence of X and Y.

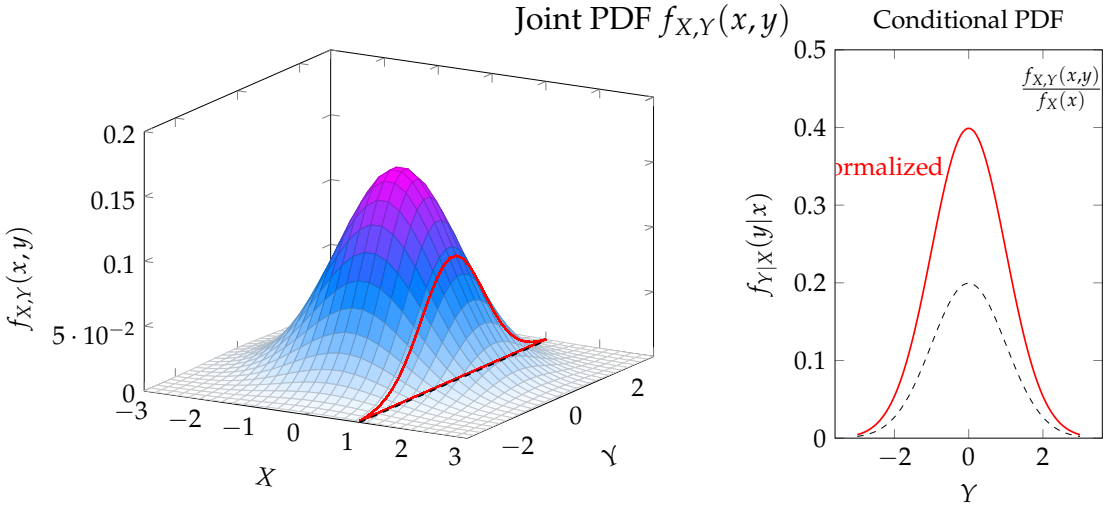
Joint PDF of Bivariate Normal Distribution



The marginal PDF can be visualized as the height of the surface obtained by fixing one variable and observing the distribution of the other. For example, the marginal PDF of X can be visualized as a 2D slice of the joint PDF along the X-axis.



Remember, for conditional probability, we take a vertical slice of the joint PDF corresponding to the observed value of X; since the total area under this slice is $f_X(x)$, we then divide by $f_X(x)$ to ensure that the conditional PDF will have an area of 1.



2.5.1 Independence of Random Variables

Two random variables X and Y are said to be *independent* if the occurrence of one does not affect the probability distribution of the other. This concept can be formally defined in terms of their joint and marginal probability mass functions (PMFs).

Definition 2.36. *The random variables X and Y are independent if and only if the following condition holds for all x and y:*

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

This means that the joint PMF of X and Y can be expressed as the product of their marginal PMFs.

Equivalent Conditions

The independence of random variables can also be expressed in several equivalent forms:

1. **For all Events**: X and Y are independent if for any two events A and B,

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

2. **In terms of Expectations**: If X and Y are independent, then for any functions g and h,

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

Example 2.18. *Consider two dice rolls X and Y, where X represents the outcome of the first die and Y represents the outcome of the second die. Since the outcome of one die does not affect the outcome of the other, we can show their independence.*

The joint PMF is given by:

$$P(X = x, Y = y) = \frac{1}{36} \quad \text{for } x, y \in \{1, 2, \ldots, 6\}.$$

The marginal PMFs are:

$$P(X = x) = \frac{1}{6} \quad \text{and} \quad P(Y = y) = \frac{1}{6}.$$

Verifying independence:

$$P(X = x, Y = y) = \frac{1}{36} = P(X = x)P(Y = y) = \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36}.$$

Thus, X and Y are independent random variables.

Example 2.19. *Consider two independent continuous random variables X and Y with the following PDFs:*

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}, \quad f_Y(y) = \begin{cases} \frac{1}{3} & \text{for } 0 < y < 3 \\ 0 & \text{otherwise} \end{cases}.$$

The joint PDF of X and Y is:

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y) = \begin{cases} \frac{1}{6} & \text{for } 0 < x < 2 \text{ and } 0 < y < 3 \\ 0 & \text{otherwise} \end{cases}.$$

This demonstrates that X and Y are independent, as the joint PDF is indeed the product of the marginal PDFs.

2.5.2 Covariance and Correlation

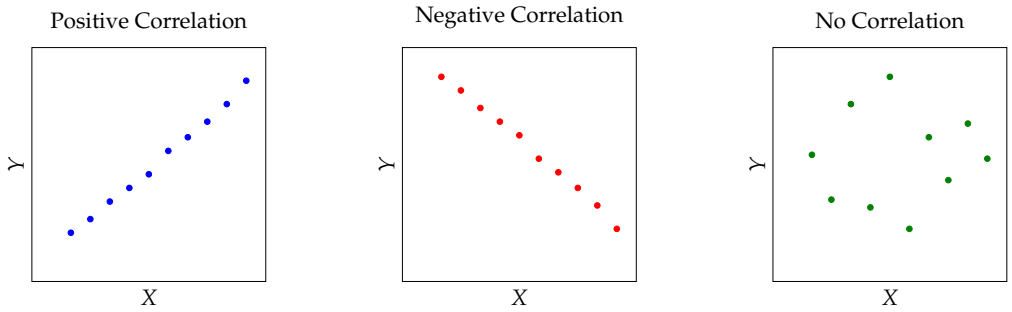
Definition 2.37. *The covariance of two random variables X and Y is a measure of the degree to which the two variables change together. It is defined mathematically as:*

$$Cov(X,Y) = E\left[(X - E[X])(Y - E[Y])\right] = E[XY] - E[X]E[Y],$$

where $E[X]$ and $E[Y]$ are the expected values (means) of X and Y, respectively.

Interpretation:

- If $\text{Cov}(X,Y) > 0$: X and Y tend to increase together.
- If $\text{Cov}(X,Y) < 0$: When X increases, Y tends to decrease, and vice versa.
- If $\text{Cov}(X,Y) = 0$: There is no linear relationship between X and Y.



Example 2.20. *Consider two random variables X and Y with the following joint distribution:*

$X \backslash Y$	1	2	3
1	0.1	0.2	0.1
2	0.2	0.1	0.1
3	0.1	0.1	0.1

First, we calculate the marginal distributions:

$$P(X=1) = 0.4, \quad P(X=2) = 0.4, \quad P(X=3) = 0.2,$$
$$P(Y=1) = 0.4, \quad P(Y=2) = 0.4, \quad P(Y=3) = 0.2.$$

Next, we calculate the expected values:

$$E[X] = 1 \cdot 0.4 + 2 \cdot 0.4 + 3 \cdot 0.2 = 1.8,$$
$$E[Y] = 1 \cdot 0.4 + 2 \cdot 0.4 + 3 \cdot 0.2 = 1.8.$$

Next, we calculate $E[XY]$:

$$E[XY] = 1 \cdot 1 \cdot 0.1 + 1 \cdot 2 \cdot 0.2 + 1 \cdot 3 \cdot 0.1 + 2 \cdot 1 \cdot 0.2 + 2 \cdot 2 \cdot 0.1 + 2 \cdot 3 \cdot 0.1 +$$
$$3 \cdot 1 \cdot 0.1 + 3 \cdot 2 \cdot 0.1 + 3 \cdot 3 \cdot 0.1 = 1.4.$$

Now we calculate the covariance:

$$Cov(X,Y) = E[XY] - E[X]E[Y] = 1.4 - (1.8 \cdot 1.8) = 1.4 - 3.24 = -1.84.$$

This negative covariance indicates that X and Y tend to move in opposite directions.

Properties of Covariance

1. Symmetry:
$$\text{Cov}(X, Y) = \text{Cov}(Y, X).$$

2. Linearity: For any constants a and b,
$$\text{Cov}(aX + b, Y) = a \cdot \text{Cov}(X, Y).$$

3. Independence: If X and Y are independent, then:
$$\text{Cov}(X, Y) = 0.$$

4. Variance Relation: The covariance of a random variable with itself is its variance:
$$\text{Cov}(X, X) = \text{Var}(X).$$

Definition 2.38. *The correlation between two random variables X and Y is a measure of the strength and direction of their linear relationship. It is defined mathematically as:*
$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$
where $Cov(X, Y)$ is the covariance between X and Y, and σ_X and σ_Y are the standard deviations of X and Y, respectively. The correlation coefficient $\rho(X, Y)$ ranges from -1 to 1.

Difference between Covariance and Correlation:

Covariance provides information about the direction of the relationship but is affected by the scale of the variables. Correlation standardizes this relationship, making it easier to interpret the strength and direction of the relationship without the influence of scale.

2.6 Some Inequalities and Limit Theorems

2.7 Markov Chains and Monte Carlo

2.8 The Bertnard's Paradox

Bibliography

[1] Grant Sanderson 3Blue1Brown. Olympiad level counting (generating functions). Technical report, YouTube https://www.youtube.com/watch?v=bOXCLR3Wric, 2022.

[2] Stephen Abbott. *Understanding Analysis*. Springer Publications, 2015.

[3] Joseph Blitzstein. *Introduction to Probability*. CRC Press, 2015.

[4] Krishna Jagannath. *Probability Foundations for Electrical Engineers*. IIT Madras, 2015.