Discrete Probability

SIT192 - Discrete Mathematics

By

Brandon Smith

TABLE OF CONTENTS

1 INTRODUCTION

While I could certainly dedicate an extensive portion to establishing the importance of the subject, I'll opt to touch upon it only briefly.

In the swirling vortex of uncertainty that is life, there are few tools as elegant and powerful as Probability Theory. This report aims to illuminate its beauty, showcasing the fundamental concepts and tools that form the backbone of this discipline. Life is just filled with uncertainties, and things rarely happen exactly as planned. This ever-present uncertainty means we need effective ways to think about and deal with it. That's where Probability Theory comes in, providing us models that help us manage uncertainty in a structured manner.

Suppose you're a machine learning engineer working with big data. You're constantly battling against noise, which is unpredictable by nature. How do you conceptualise it? How do you tackle it? Or perhaps you're a business owner, routinely grappling with market demand, a factor that's often erratic. Or maybe you're navigating the stock market, where randomness is the only guarantee. Indeed, numerous aspects of life are random.

Various fields employ Probability Theory to handle this randomness. While each field adapts these concepts to their specific needs, they all originate from the basic principles we'll be discussing today. It's my hope that these concepts will not only be useful to you, but also inspire you to dive deeper into this fascinating subject.

1.2 LEARNING OBJECTIVES

Upon completion of this module, you, the reader, should be able to do the following:

1. explain and define probability models and axioms, such as sample space and events,
2. compute conditional probability directly and using Bayes' rule, on real world problems,
3. recognise and apply discrete distributions to find the probability of various events.

1.3 PRELIMINARY CONCEPTS

Throughout this module, we'll be using several concepts which are not explained here, predominantly relating to Set Theory.

A set is a collection of distinct items, known as elements, while an empty set, is a set without any elements, which is denoted as Ø. Sets can be expressed in various ways. Let's say we have a finite set S that contains the elements a_1, $a_2$up to a_n. We could list these elements within curly braces like this:

$$S = \{a_1, a_2 \ldots, a_n\}$$

Subsets on the other hand, which means if every element in one set (we'll call this E) is also in another set (we'll call this S), then we say E is a subset of S, and this is written as $E \subseteq S$. In other words, a subset is a set *composed* entirely of elements from another set.

Let's say we have a set S that contains the objects A, B, and C:

$$S = A, B, C$$

In this case, sets like {A}, {B}, {C}, {A, B}, {A, C}, {B, C}, and {A, B, C} are all subsets of S. The empty set Ø is considered a subset of every set. This idea of subsets forms the backbone for many discussions in probability theory, especially when we're talking about event spaces within a particular sample space.

For further reading, I recommend reading Naïve Set Theory by Paul R. Halmos. It is a friendly and fun to read introduction to Set Theory. We will also be using Venn Diagrams to illustrate some ideas of Probability Theory, I have provided a resource here for which you can use to get yourself familiar.

2 PROBABILITY MODELS

Probability models serve as mathematical representations of uncertain situations or experiments, helping us navigate the unpredictable seas of randomness that affect our lives. They can be thought of as blueprints, that guide us through uncertainty and allow us to make informed decisions.

The creation of a probability model involves a few important steps. Firstly, we identify all possible outcomes of an experiment or situation - we establish what's known as a sample space. This comprehensive listing of potential results provides a broad canvas on which we begin to establish our model.

Once we've defined our sample space, we establish a probability law. This assigns a likelihood to each individual outcome or groups of outcomes within our sample space, enabling us to understand which results are more likely to transpire. However, these assigned probabilities need to adhere to a set of rules or principles, ensuring a logical coherence within our model. These principles form the foundational pillars of probability theory, known as the axioms. While only a few, these axioms serve as the bedrock upon which we construct many important insights and theorems.

In the next few points in this report will provide a detailed exploration of these concepts, breaking down their mechanisms and highlighting their significance. We will also shine a light on the practical applications of these probability models, demonstrating their substantial impact in real-world situations.

2.1 PROBABILITY FUNCTION

This is just a quick explanation of a simple probability function. How it works is it assigns probabilities to events, or specific sets of outcomes from a random experiment. This function abides by Kolmogorov's axioms – which we will discuss in detail later.

1. The *domain* of this function is the event space, a collection of subsets from the sample space.

2. The *codomain* of a probability function falls between 0 and 1, inclusive, reflecting the probability of an event's occurrence. It's important to note that this will differ from the functions we'll explore in later discussions on discrete distributions.

Let's assume the probability function is denoted by P, with the domain being the event space E, and the codomain falling between 0 and 1 inclusive. This can be represented as:

$$P: E \rightarrow [0,1]$$

2.2 SAMPLE SPACE AND EVENTS

Let us imagine that we're conducting an experiment with an unpredictable outcome. We can't foresee the exact result, but we know all the potential outcomes that could occur. This collection of all conceivable results is what we call the sample space of the experiment, and in this module, we will denote it as S. However, you may find that it is can also be denoted as either Ω, or U.

Sample Space

We begin with the idea of a sample space. This is the full set of potential outcomes for a given scenario or experiment.

For example, let's consider an experiment where a coin is tossed twice, and we record the face that shows after each toss. We can represent this as H for heads and T for tails. The sample space of this experiment, denoted S, would be:

$$S = \{HH, HT, TH, TT\}$$

In more general terms, an experiment can be thought of as a procedure or action that could be repeated under the same conditions, leading to uncertain outcomes. Take rolling a pair of dice or flipping a coin or drawing a card from a deck as an example. In each of these experiments, the specific outcomes remain unknown until the experiment is conducted, and the collective set of all possible outcomes constitutes the sample space.

Event Space

An event is a specific outcome or a collection of outcomes from the sample space, also known as the subset of the sample space. In the coin-tossing example, an event could be getting a head followed by a tail, or a tail followed by a head. This can be denoted like so:

$$E = \{HT, TH\}$$

This event represents the proposition that when the coin is tossed twice, it will show one head and one tail. Events are collections of individual sample points that make up some interesting statement or proposition in the experiment.

Remark 1. To extend on the definition of sample space and events, I would like to point out that it is important to understand how certain principles of probability theory operate in these contexts. While it is human to often associate a probability of zero with impossibility, this isn't always the case. We need to be able to differentiate between an *"impossible"* event and those which are *"improbable"*.

In Kolmogorov's second remark from Foundations of the theory of probability, he states that if an event has no potential outcomes within the sample space, we can safely consider it *"impossible"*, and hence it has a probability of zero.

An example of such an event would be flipping a fair coin and landing on both heads and tails at the same time — this is physically impossible.

However, having zero probability doesn't necessarily make an event *"impossible"*. A zero-probability event may seem practically impossible within a single instance or a finite event space, but when considering infinite trials or an infinite event space, such an event can indeed occur. This points to the event's extreme "*improbability"*, not its absolute "*impossibility"*.

Anyways, let us bring this concept to life with a practical example. Suppose we are organising a pizza party for the students enrolled in SIT192, a unit at Deakin University. To ensure we order the right types of pizzas, we conduct a survey asking each student about their pizza preferences, specifically:

1. Do you prefer Meat Lovers?
2. Do you prefer Hawaiian?

Out of a total of 100 students, we receive varying responses, providing us with four potential outcomes. Each student's response defines an outcome within the sample space S, representing the entire group of students who responded to the survey. These outcomes are:

1. Prefers Meat Lovers only.
2. Prefers Hawaiian only.
3. Prefers both Meat Lovers and Hawaiian.
4. Prefers neither Meat Lovers nor Hawaiian.

We find that:

- Total number of students: $\mathrm{N} = 100$
- Students preferring Meat Lovers: $\mathrm{M} = 35$
- Students preferring Hawaiian: $\mathrm{H} = 20$
- Students preferring both Meat Lovers and Hawaiian: $\mathrm{M\ and\ H} = 15$
- Students not preferring either Meat Lovers or Hawaiian: $\mathrm{N} - (\mathrm{M\ or\ H}) = 30$

Therefore, the sample space S, of the students' pizza preferences can be defined as:

$$S = \{Meat\ and\ Not\ Hawaiian, Not\ Meat\ and\ Hawaiian, Both, Neither\}$$

This way, we can categorise the preferences of the entire sample space into distinct events, providing us a more clear and systematic understanding of the students' pizza preferences.

Definition 2.2 Consider S to be the sample space for a given experiment. Any subset E within S, which can range from the empty set to the entire sample space S is defined as an event.

2.3 OPERATIONS ON EVENTS

Given an experiment and its associated sample space, we can perform several operations on events, which are subsets of the sample space. Using **definition 2.2** it defines an event as any subset within the sample space, which could range from the empty set to the entire sample space itself. From here, we can dive into several operations on these events within the sample space.

Definition 2.3.1 (**Union of Events):** Denoted as A ∪ B, this represents the occurrence of either event A, event B, or both.

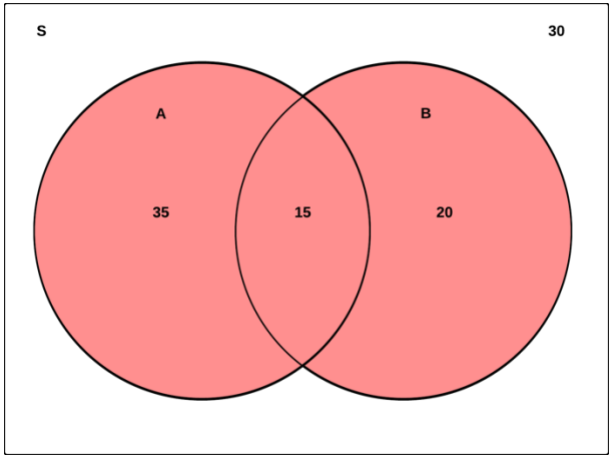
Definition 2.3.2 (Intersection of Events): Represented as A ∩ B, this signifies that events A and B occur simultaneously.

Definition 2.3.3 (**Complement of an Event):** Denoted as A' or ¬A, this signifies the non-occurrence of event A.

Union of Events

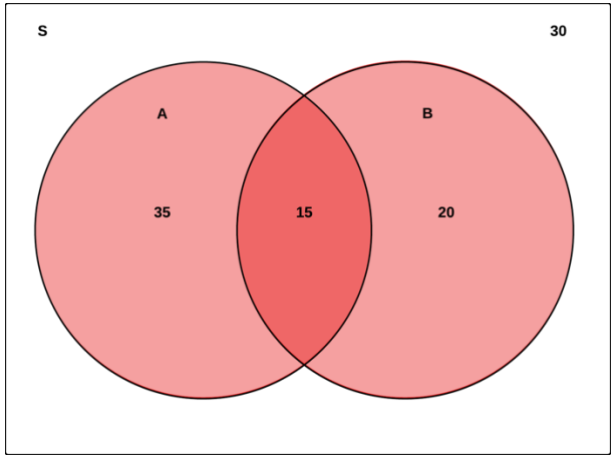
Let us go back to our pizza survey, let A be the event representing students who prefer Meat Lovers pizza and B the event representing students who prefer Hawaiian pizza. The union of A and B, denoted as A ∪ B, is the event representing all students who prefer either Meat Lovers, Hawaiian, or both.

This group includes students who enjoy either pizza individually or also those who like both, covering all possibilities related to these two pizza types.



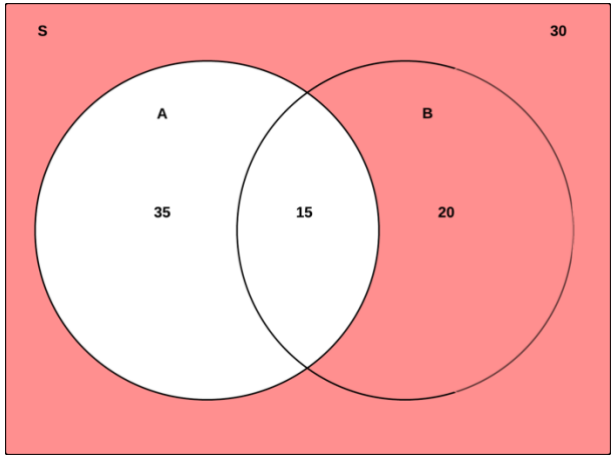
Intersection of Events

Still using our pizza example, the intersection of events A and B, $A \cap B$, represents students who enjoy both Meat Lovers and Hawaiian pizzas. This is a specific group of students who prefer both types of pizza.



Complement of an Event

In the context of our pizza survey, if A is the event of students preferring Meat Lovers pizza, then $\neg A$ or A' represents the group of students who do not prefer Meat Lovers pizza. This group would include students who like Hawaiian only and those who like neither.



2.4 PRODUCT AND SUM RULE

This portion of the module is dedicated to defining the product and sum rules. These widely accepted mathematical principles form the basis for the subsequent sections of this module.

Rule 2.4.1 (Product Rule): For independent events A and B, the product rule determines the probability of both events happening concurrently, denoted as $P(A \cap B)$. It is defined as:

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

Rule 2.4.2 (Sum Rule): The sum rule calculates the probability of either or both events A and B occurring, denoted as $P(A \cup B)$, as:

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

It corrects for the double counting of A and B occurring together in the individual probabilities $P(A)$ and $P(B)$.

2.5 KOLMOGOROV AXIOMS

In this section we will introduce Kolmogorov's axioms. The axioms described by Kolmogorov in his publication – Foundations of the theory of probability, he writes:

"The theory of probability, as a mathematical discipline, can and should be developed from axioms in exactly the same way as Geometry and Algebra."

What this means is, after we've identified the basic elements and their relationships, and laid out the rules (axioms) guiding these relationships, everything that follows should be based strictly on these rules. This approach suggests that we should set aside real-world interpretations for a bit and focus on the logic and reasoning that these rules provide.

But why do we take such an approach? The purpose is not merely abstraction for abstraction's sake, but to strive for generality in our understanding and application of the theory. This point resonates with Kolmogorov's assertion that an abstract theory admits an unlimited number of concrete interpretations beyond those from which it was originally derived.

This means by focusing on the abstract structure and logic provided by the axioms, we are creating a framework that can be adapted to a wide list of different scenarios and disciplines.

Since we are discussing Kolmogorov's axioms, I would like to share his notes on terminology from Foundations of the theory of probability which will help with our understanding as we progress through the module.

Set Theory	Random Events
1. A and B are disjoint sets, i.e., $A \cap B = \emptyset$.	**1.** Events A and B are mutually exclusive.
2. The intersection of sets $A, B, \ldots, N$ is represented as $AB \ldots . N$, and if $AB \ldots . N = \emptyset$, it means that sets $A, B, \ldots, N$ are mutually exclusive.	**2.** If events $A, B, \ldots, N$ are mutually exclusive, it means that they cannot occur simultaneously.
3. If $AB \ldots . N = X$, it signifies that X represents the intersection of sets $A, B, \ldots, N$.	**3.** Event X happens if and only if events $A, B, \ldots, N$ all occur simultaneously or concurrently.
4. $A + B + \ldots + N = X$ means that X represents the union or combination of $A, B, \ldots, N$. It indicates that X occurs if at least one of $A, B, \ldots, N$ occurs.	**4.** Event X is defined as the occurrence of at least one of the events $A, B, \ldots, N$. It means that X happens if any of the events occur.
5. The complementary set of A, denoted as $\overline{A}$. Which means the the opposite of A.	**5.** The event A', representing the non-occurrence of event A.
6. $A = \emptyset$.	**6.** Event A is impossible within the given sample space.
7. $A = E$	**7.** Event A represents the entire sample space E, which indicates that all outcomes in the sample space is certain to occur.
8. B is a subset of A (which means included in A) $A : B \subseteq A$	**8.** If event B happens, it guarantees that event A will happen without a doubt.

THE AXIOMS

Axiom 1: The probability of any event E within the sample space S can never be negative.

$$P(E) \geq 0 \; \forall \; E \in S$$

Axiom 2: The probability of at least one event occurring within the sample space is certain, hence equal to 1.

$$P(S) = 1$$

Axiom 3: If events are mutually exclusive then the probability of the union of these events equals the sum of their individual probabilities.

$$P(A1 \cup A2 \cup A3 \cup \ldots) = P(A1) + P(A2) + P(A3) + \ldots$$

Kolmogorov's set theory terminology gives us some essential definitions for understanding his axioms. In particular, he uses the notation $A = \emptyset$ to indicate that event A is impossible within the given sample space. This means that the probability of an impossible event is $P(\emptyset) = 0$.

However, this fact is not explicitly mentioned in Kolmogorov's axioms (and it doesn't need to be). That is, because it is a consequence that can be derived directly from the axioms.

PROOF

Let's consider the empty set $\emptyset$, an event that represents impossibility as it contains no outcomes. Even though it's not explicitly stated, the probability of the empty set is embedded within these axioms.

By Axiom 3, if we consider two mutually exclusive events which are both the empty set, we have:

$$P(\emptyset \cup \emptyset) = P(\emptyset) + P(\emptyset)$$

But the union of the empty set with itself remains the empty set, so we end up with:

$$P(\emptyset) = P(\emptyset) + P(\emptyset)$$

Which simplifies to (by subtracting $P(\emptyset)$ from both sides)

$$0 = P(\emptyset)$$

Therefore, by using the axioms we can prove that $P(\emptyset) = 0$.

2.6 PROBABILITIES OF UNIONS AND INTERSECTIONS

Theorem 2.6: The probability of the union of two events A and B, denoted as $P(A \cup B)$, is calculated using the sum rule **(rule 2.4.2)**.

PROOF

We start by defining our events:

1. $E_1 = A \cap B'$
2. $E_2 = A \cap B$
3. $E_3 = A' \cap B$

Here, A' and B' denotes the complements of A and B, respectively, representing all outcomes not in A or B. Also, $E_1, E_2,$ and E_3 are pairwise mutually exclusive events. This means that if one of these events happens, the others cannot. It's an either/or situation, not both.

- E_1 and E_2 are mutually exclusive because if E_1 happens then E_2 cannot.
- E_2 and E_3 are mutually exclusive because if E_2 happens then E_3 cannot.
- E_1 and E_3 are mutually exclusive because if E_1 happens then E_3 cannot.

Therefore, by applying axiom 3, we have:

$$P(A) = P(E_1) + P(E_2)$$

$$P(B) = P(E_2) + P(E_3)$$

Looking at it now, it's clear that $A \cup B = E_1 \cup E_2 \cup E_3$. This means by applying axiom 3 again we have:

$$P(A \cup B) = P(E_1) + P(E_2) + P(E_3)$$

Since $P(A) = P(E_1) + P(E_2)$ and $P(B) = P(E_2) + P(E_3)$, we have:

$$P(A \cup B) = P(A) + P(B) - P(E_2)$$

Since $E_2 = A \cap B$, we have:

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

Therefore, we have successfully proven **Theorem 2.6**.

From the theorem, we were also to derive the following:

1. **Inclusion-Exclusion Principle:** The theorem shows that the probability of either A or B occurring isn't just the sum of the probabilities of A and B. The intersection is double counted when we add these probabilities, so we need to subtract it out.
2. **Upper Bound:** The theorem also provides an upper bound on the probability of the union of two events. The probability of either A or B happening is at most the sum of the probabilities of A and B. This keeps probabilities from going above 1, respecting the boundaries of the probability space.

Ok, now going back to our pizza party example for students enrolled in SIT192. Remember, we surveyed 100 students about their pizza preferences. Let A represent the event that a student prefers Meat Lovers and B represent the event that a student prefers Hawaiian.

We conducted the survey again because we didn't believe that 10 students did not like pizza, in the new survey we found that:

- 50 students prefer Meat Lovers only.
- 30 students prefer Hawaiian only.
- 20 students enjoyed both.

Using the **Probability of Union**, we get:

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

$$= 0.50 + 0.30 - 0.20$$

$$= 0.60$$

So, the probability that a randomly chosen student prefers either Meat Lovers or Hawaiian (or both) is 0.6 or 60%.

The reason I have chosen this example is that it highlights the usage of the theorem and its proof. From the survey, we were able to systematically understand the students' pizza preferences and make an informed decision about the pizza order for SIT192 party. Not only have we seen this proof in action, but we also see the consequences of the theorem at play in this example:

When we calculated the probability of a student liking either of the pizzas, we had to subtract the probability of students liking both to avoid double-counting. This was using the inclusion-exclusion principle.

We also see the upper bound consequence, the probability of a student liking either type of pizza (or liking both) is less than or equal to the sum of the individual probabilities of liking Meat Lovers and Hawaiian.

Probability of Intersection: The probability of the intersection of two events A and B, denoted as $P(A \cap B)$, represents the probability that both events A and B happen at the same time.

We apply the product rule **(rule 2.4.1)** when two events are independent of each other – which means two events are independent if the occurrence of one does not affect the probability of the occurrence of the other.

We will investigate this further when I introduce the concept of independent events in the outcoming section.

3 CONDITIONAL PROBABILITY USING BAYES' RULE

Here's probability theory starts to get a little more interesting. However, that is not to say the previous concepts “wasn’t interesting”, it just gets more... exciting. We're about to dive into areas where probability is not just about counting outcomes and assigning equal chances, but where we consider the context and specific conditions.

In my opinion, applying the fundamentals isn't as intimidating as one might initially believe. Take the example of the SIT192 pizza party example; the application of probability concepts appears straightforward, does it not?

However, as we dive deeper, we start to see that the application of these concepts is only the tip of the iceberg. The real challenge lies beneath the surface, in the ability to apply logical thinking and problem-solve. That's where things become less trivial because suddenly, the principles that seemed intuitive don't fit as neatly into the problems we're trying to solve. That, in my opinion, is the true allure of probability.

The core of probability lies in translating a logical problem into a quantitative one, turning abstract ideas into numbers. You will also start to find that it is not just about inserting values into formulas to arrive at an answer. To quote Laplace from 1819 "Probability theory is nothing but common sense reduced to calculation.". I found this a nice way to think of probability. I first read this quote in *Probability Theory: The Logic of Science by E. T. Jaynes*, which is a very interesting read and definitely a book I recommend.

Think about it like this, imagine yourself as a data scientist working on predicting a customer's likelihood of buying a particular product. You find that 10% of customers buy this product. Ok, so what if you also find that there are customers who have bought similar products in the past? How does this new information change the model? It's not just about the overall probability anymore, it's about the probability given this new piece of information - given that the customer has a history of buying similar products. The odds of that customer buying this product are likely to be higher than 10%.

3.1 INDEPENDENT EVENTS

Definition 3.1 (Independent Events): events A and B are said to be independent if the fact that one event happens does not affect the probability that the other event will happen. If whether event B happens does not affect the probability of event A, then event A is independent of event B.

When two events are independent such as A and B we can express it as:

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

Let me introduce **definition 3.1** with an entirely new example. Suppose we have a fair six-sided die and a fair coin. We are interested in the event A which is rolling a 6 and event B which is flipping heads.

We know that these two events are independent of each other. Rolling a 6 has nothing to do with flipping a head. If you don't believe me, I suggest you try conduct this experiment to see for yourself.

Since rolling a 6 is one out of six possible outcomes, the probability of $\mathrm{A} = 1/6$ and since flipping heads is one out of two possible outcomes, the probability of $\mathrm{B} = 1/2$.

Using the product rule **(rule 2.4.1)**, we can now figure out the probability of getting a 6 and a head:

$$P(A \cap B) = P(1/6) \cdot P(1/2)$$

$$P(A \cap B) = 1/12$$

$$\approx 8.3\%$$

Let's think more deeply about this. We know that rolling a 6 and landing heads are independent events. How do we know this? If we refer back to Kolmogorov's notes, he defines an event X as one that occurs if and only if events A, B, and so forth, all occur simultaneously or concurrently.

In this case, event X is the outcome where we roll a 6 and flip heads, while event A is rolling a 6 and event B is flipping heads. Event X cannot happen unless both A and B happen.

With this understanding, we can apply the product rule, a direct consequence of the definition of independence, to calculate the probability of the intersection of two independent events.

3.2 CONDITIONAL PROBABILITY

The axioms of probability provided by Kolmogorov form a cornerstone for the concept of conditional probability. It's based on the natural understanding of the revised likelihood of an event given that another event has already occurred.

Definition 3.2 (Conditional Probability): If we consider two events A and B within a given sample space, with P(B) > 0, the conditional probability of A given B is defined as follows:

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

In the definition above, B acts as our new sample space, and we are interested in the likelihood of event A within this new space. It's important to remind ourselves that $P(B) > 0$ is required to avoid division by zero.

Suppose you are planning to go for a surf tomorrow. There are two possible events in your case:

- **Event A**: It storms tomorrow.
- **Event B:** You go surfing.

You have access to a pretty accurate weather app and feel confident in using it to conclude on a decision.

1. Based on the weather forecast and historical data, you estimate there's a 30% chance it will storm tomorrow. So, $P(A) = 0.30$.

2. You're a brave soul who loves to surf storm or shine, so there's a 90% chance you'll surf. P(B) = 0.90.

3. Based on past experiences, you know that 20% of your surfs have occurred on stormy days. This is the intersection of events A and B. $P(A \cap B) = 0.20$.

We want to find $P(A|B)$, which is the probability it will storm given that you'll surf. By using the formula of conditional probability **(Definition 3.2):**

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.20}{0.90} \approx 0.22 \text{ or } 22\%$$

So, given that you decide to surf, there's a 22% chance that it will storm.

Now, let's also calculate $P(B|A)$, the probability that you will surf given that it storms:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.20}{0.30} \approx 0.67 \text{ or } 67\%$$

So, if it storms, there's a 67% chance that you will surf.

What does this mean? Well… Given that you decide to surf, there's a 22% chance that it will storm. This tells you the risk of encountering a storm if you go surfing. However, if it storms, there's a 67% chance that you will surf. This tells you how likely you are to stick with your plan to surf, even if it storms.

So really, it depends on your attitude toward risk and how much you dislike surfing in a storm.

If you're not deterred by the prospect of a storm, the high probability of surfing (90%) means you're likely to go regardless of the weather. Knowing that there's a 22% chance of a storm shouldn't change this.

On the other hand, if you're considering whether to change your plans when you know there's a storm coming, the 67% figure is relevant. This is pretty high, which means you're fairly committed to surfing, even when a storm is brewing.

Of course, you might also consider other factors, like how severe the storm is likely to be, whether there are safe places to shelter, etc. But from a purely probabilistic standpoint, if you're okay with a 22% chance of surfing in a storm, then you should go for it!

3.3 LAW OF TOTAL PROBABILITY

The Law of Total Probability is like a recipe that helps you find the overall likelihood of an event by considering all the different ways it could happen. This "recipe" or law is especially handy when you have a situation that can be divided into several distinct or non-overlapping scenarios, called partitions. Each of these partitions represents a different way the event of interest can happen.

The law states that the total probability of an event A is the sum of the probabilities of A happening across all possible scenarios. If you think of the event A as a pizza, then the law of total probability tells us that to find the total "probability pizza", we need to add up all the "probability slices" from each different scenario.

Definition 3.3 (Partition of a Sample Space): A collection of events $\{B_1, B_2, \dots, B_n\}$ is said to be a partition of the sample space S if the following conditions are satisfied:

1. The events B_i are pairwise mutually exclusive and exhaustive – for any pair of unique events in the collection B_i and B_j, we have $B_i \cap B_j = \emptyset$.

2. The union of all the events in the collection forms the sample space - $B_1 \cup B_2 \cup \dots B_n = S$.

3. If $\mathrm{P}(\mathrm{B_i}) = 0$ for some i, then B_i is an event that never occurs, but this doesn't affect the partition conditions. However, for computing conditional probabilities like $P(A|B_i)$, we need $P(B_i) > 0$ to avoid division by zero.

Theorem 3.3 (The Law of Total Probability): Let $B_1, B_2, \dots, B_n$ be a partition of the sample space S but together cover all possible outcomes. Let A be an event. Then the probability of A can be calculated as:

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

PROOF

To prove the theorem, we can use definition of conditional probability defined in **definition 3.2**, the Kolmogorov's third axiom, and **definition 3.3**.

Let us assume that $B_1, B_2, \ldots, B_n$ forms a partition of the sample space S. This means that the sets B_i are mutually exclusive and exhaustive — no two sets share elements and their union is the entire sample space. Therefore, we can represent Event A as a union of mutually exclusive events and by applying the third axiom:

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n)$$

Then by using the definition of conditional probability, we have:

$$P(A \cap B_i) = P(A|B_i) \cdot P(B_i), \text{ for each } i$$

To understand why, let me first illustrate quickly:

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

If we multiply both sides by $P(B_i)$, we can then cancel out the like terms to get:

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

Now we apply this definition to $P(A)$, which gives:

$$P(A) = P(A \mid B_1) \cdot P(B_1) + P(A \mid B_2) \cdot P(B_2) + \cdots + P(A \mid B_n) \cdot P(B_n)$$

Which now can be rewritten as:

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

Which completes the proof of **theorem 3.3**.

3.4 BAYES' THEOREM

Bayes' theorem presents an approach for computing the conditional probability of an occurrence A, given occurrence B. It operates on the premise that we're already aware of the probabilities of A, B considering A, and B in the absence of A.

Theorem 3.4 (Bayes' theorem): For events A and B, For events A and B, with P(A) > 0 and P(B) > 0, the conditional probability of A given B, denoted as P(A | B), is given by:

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

PROOF

We will prove Bayes' theorem using the definition of conditional probability and the Law of Total Probability, defined in **definition 3.2** and **theorem 3.3**. Firstly, we start by using the definition of conditional probability:

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

and also:

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

While proving **theorem 3.3** I pointed out that: $(A \cap B) = P(B|A) \cdot P(A)$. Therefore, we can replace $(A \cap B)$ with this expression.

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

Now, Bayes' theorem is **proved**. However, let me introduce this by revisiting our previous example in conditional probability. Previously we calculated that:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.30}{0.60} = 0.50$$

But what if now we want to find the conditional probability $P(A|B)$, which represents the probability that a student understands Propositional Logic given that they already understand Sets and Functions.

Well, using Bayes' theorem, we can calculate this as:

$$P(A|B) = \frac{\big(P(B|A) \cdot P(A)\big)}{P(B)} = \frac{(0.50 \cdot 0.60)}{0.50} = 0.60$$

Understanding this theorem is important, especially in probability, Naïve Bayes used for classification problems is built of this theorem. Ever wonder how your emails get re-directed to spam or not spam? – I suggest looking into the Naïve Bayes classifier.

I believe a firm understanding of Bayes' Theorem is important not just in theory, but also in understanding how practical applications like email spam filtering work. It's a great example of how seemingly abstract mathematical concepts can have a big impact on everyday technologies – which is always evolving. Also, sometimes, it is much easier to find P(B|A) then it is P(A|B).

4 DISCRETE DISTRIBUTIONS

In this final section of the module, we will explore a set of discrete probability distributions. These distributions are used for describing various types of random phenomena, particularly those which involve discrete outcomes or events.

Now, as we move forward, we will explore five fundamental probability distributions. These distributions are fundamental tools for describing various types of random phenomena.

First, we start with the Bernoulli distribution, a probability distribution that models the outcome of a single binary experiment.

We'll then introduce the binomial distribution, which is an extension of the Bernoulli distribution, but instead of a single trial, we consider multiple independent and identical Bernoulli trials. Here, the outcome of one trial does not influence the others. But, when we count the number of successes across these trials, we see the binomial distribution presenting a comprehensive depiction of how randomness behaves over repeated trials.

Following this, we examine the geometric and hypergeometric distributions. The geometric distribution describes the number of trials required to achieve the first success, which is helpful in situations like product testing or clinical trials where the focus is on achieving that first breakthrough. In contrast, the hypergeometric distribution considers scenarios of success and failure without replacement, making it suitable for quite important situations like quality control in manufacturing.

Finally, we'll encounter the Poisson distribution. This powerful tool is used to estimate the frequency of events within a specific period.

These discrete distributions and their corresponding Probability Mass Functions (PMFs) make up a nice toolkit for handling randomness. Their applications span a wide variety of fields, a few too many to mention.

4.1 BERNOULLI DISTRIBUTION

Picture this: you're at a roulette table, placing bets on either red or black. Each bet you place, each spin of the wheel, is an independent trial - a Bernoulli trial, to be exact.

So, what exactly is a Bernoulli trial? Well, it's quite simple, yet very important. In a Bernoulli trial, you're dealing with something random, a situation that has exactly two outcomes. It could be red or black on the roulette wheel, heads, or tails on a coin flip but more formally we would say success, or failure.

Key properties of a Bernoulli trial:

1. **Independence:** The outcomes of each trial are independent events, meaning the occurrence of one outcome does not influence the occurrence of another outcome.
2. **Two possible outcomes:** The sample space S contains exactly two distinct outcomes, $\{0, 1\}$, representing failure and success respectively.
3. **Constant probabilities:** The probabilities of success and failure, p and q, remain fixed throughout the experiment. These probabilities are not dependent on previous trials or outcomes.
4. **Probability distribution:** Where there are two outcomes the probabilities can be defined as $P(0) = q$ and $P(1) = p$, satisfying the conditions $P(0) + P(1) = q + p = 1$.

Now that we understand what a Bernoulli trial is, what exactly is the Bernoulli distribution? Well, it is the probability distribution of a random variable which takes a binary, Boolean output: 1 (success) or 0 (failure) – which sounds familiar to a Bernoulli trial, correct?

Well, that is because it is. Essentially, a Bernoulli distribution represents the outcome of a *single* Bernoulli trial, we can model this by using the Probability Mass Function (PMF) of the Bernoulli distribution:

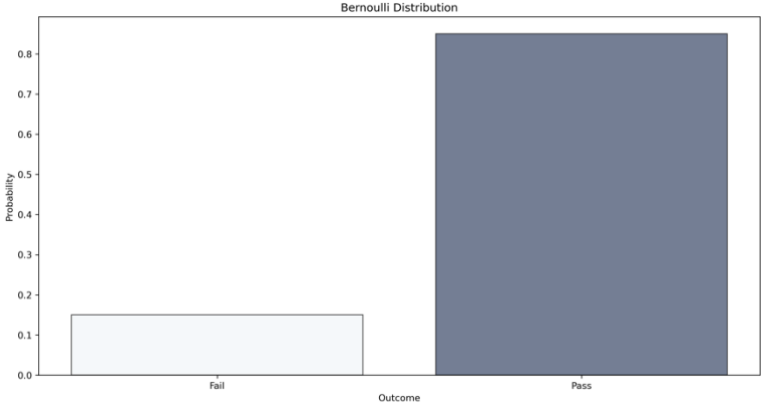
$$P(X = k) = p^k(1-p)^{1-k} \text{ for } k \in \{0,1\}$$

Suppose there are 150 students in a course. Based on previous data, we know that 85% of students pass the course while the rest fails 15%. Here, a Bernoulli trial can be considered as a single student either passing or failing the course: 1 – passes and 0 – fails. So, we have:

When $k = 1$: $P(X = 1) = 0.85^1 \cdot (1 - 0.85)^{1-1} = 0.85$ or 85%

When $k = 0$: $P(X = 0) = 0.85^0 \cdot (1 - 0.85)^{1-0} = 0.15$ or 15%

Visually, this would look like so



4.2 PROBABILITY MASS FUNCTIONS (PMF)

Here we discuss what a probability mass function (PMF) is, which was first introduced prior when introducing the Bernoulli distribution.

The reason it was structured this way, is because it helps us prepare for a broader understanding of PMF. By understanding binary outcomes and their associated probabilities, we can more intuitively understand the concept of a PMF, which can allow variables to take on more than just two outcomes.

However, before I provide a more general definition of this, we need to understand what a random variable is.

A random variable, which we typically denote as X, is a variable representing the numeric outcomes of a random phenomenon. Random variables are either discrete or continuous. However, we will only be focusing on discrete probabilities.

So then, what is a discrete random variable? Well, it is a random variable which has the distinct feature of assuming a set of specific, countable values. For further reading on this, I recommend visiting the link here.

Imagine you're at a basketball game, and we're observing Stephen Curry who's renowned for his shooting accuracy. Let's define a variable Y to represent the number of successful shots he makes in a game. Here, Y isn't just a typical variable. Instead, its value relies on the unpredictable performance of the player, thereby making Y a random variable.

Considering the random characteristic of Y, it wouldn't be logical to ask outright if $Y = 3$, for example, but it's entirely reasonable to ask about the probability that $Y = 3$, or if $Y < 2$. This is because $Y = 3$ and $Y < 2$ correlate to events within our sample space, set out by the player's shooting outcomes. The likelihoods of these specific outcomes, or the $Y = 3$ and $Y < 2$ events, are formally characterised by the Probability Mass Function (PMF).

Definition 4.1 (Probability Mass Function (PMF)): If Y is a discrete random variable with a possible range of values represented as $\{y_1, y_2, y_3, \dots\}$, then the Probability Mass Function of Y, denoted as $P(Y = y_i)$ for all i in the index set, assigns probabilities to each of these feasible values of Y

One important property of the PMF is that the sum of probabilities for all possible values of the discrete random variable equals 1. This is because the probabilities for each outcome collectively represent all possible events within the sample space. While I used Y in this example, it's more common to see X used to represent a random variable in PMF definitions. However, this is a matter of convention, and any letter can be used to denote the random variable.

4.3 BINOMIAL DISTRIBUTION

The Binomial distribution models the number of successes in a fixed number of independent Bernoulli trials. Recall from our previous discussions on Bernoulli trials that they consist of only two outcomes, labelled as “success” and “failure”, with probabilities p and $q = 1 - p$, respectively. The distribution of such trials is governed by the Bernoulli distribution.

Now, suppose we conduct n independent Bernoulli trials, each with a success probability p. Let's denote the number of successes as k. We are interested in finding the probability of obtaining exactly k successes, written as:

$$P(X = k)$$

To derive this, we'll use the concept of the Bernoulli distribution which was explained in the section prior and expand it to n trials.

For any sequence of k successes and $n - k$ failures, the probability will be the product of individual probabilities. This is due to the property of independence in Bernoulli trials. Therefore, for a specific sequence, we have:

$$p^k(1-p)^{n-k}$$

However, there are multiple ways of obtaining k successes in n trials. The number of such combinations is given by the binomial coefficient $C(n,k)$, representing the number of ways to choose k successes from n trials.

So, the total probability $P(X = k)$ for k successes in n trials is given by multiplying the probability of a single sequence by the number of such sequences. This gives us the Probability Mass Function (PMF) of the Binomial distribution:

$$P(X = k) = C(n,k) \cdot p^k(1-p)^{n-k}$$

This shows that the Binomial distribution is indeed a generalisation of the Bernoulli distribution to a fixed number of n independent trials. It also gives a comprehensive model for representing the number of successes in n trials. It encapsulates the core of Bernoulli trials with two possible outcomes, constant probabilities, and independence of trials while allowing for multiple trials and more varied outcomes.

Using the same pizza example from the section of this module, let's consider each student's preference for Hawaiian Bernoulli trial. Each student either prefers Hawaiian (success) or doesn't (failure).

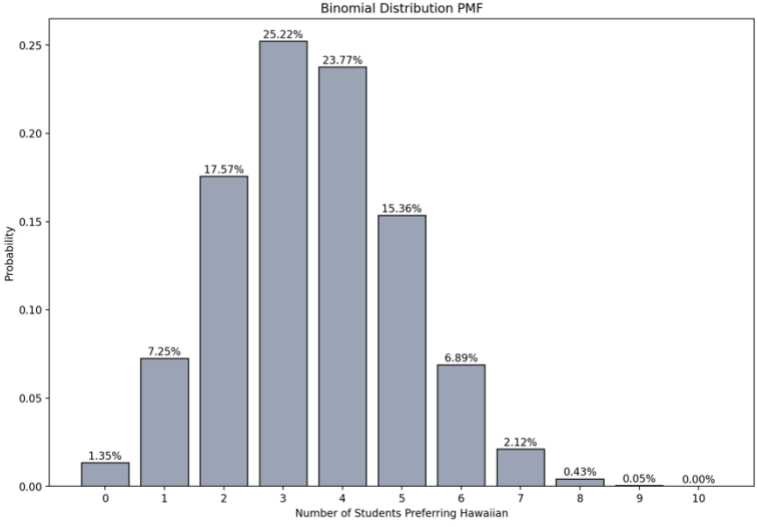
Suppose we take a random sample of 10 students out of 100. What is the probability that exactly k students prefer Hawaiian? We can start by using the Binomial distribution PMF:

$$P(X = k) = C(10,k) \cdot 0.35^k(1-0.35)^{10-k}$$

So, if $k = 3$, what would the probability be?

$$P(X = 3) = C(10,3) \cdot 0.35^3(1 - 0.35)^{10-3} \approx 0.252 \text{ or } 25.2\%$$

This means the probability of finding exactly 3 students in a random sample of 10 – meaning they were randomly chosen from the 100 students, would be approximately 25.2%. How about for other k values in the sample of 10? Well, we can use a visualisation to show us each.



How about we make this a little more interesting and we include the visualisation of the number of students picking Meat Lovers as well? The math stays the same, the only thing that changes is $\boldsymbol{p}$, like so:

$$P(X = k) = C(10,k) \cdot 0.50^k(1 - 0.50)^{10-k}$$

So, now what is the probability we find exactly 3 students picking Meat Lovers? Well, we input $k = 3$ and…

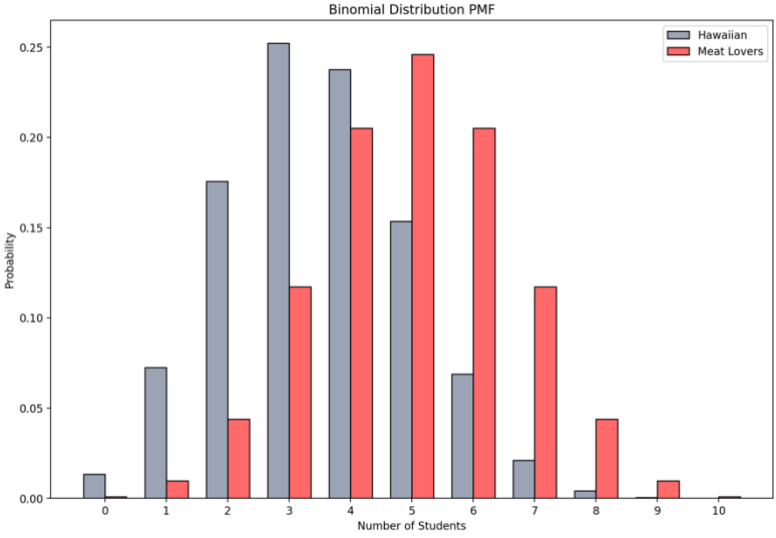
$$P(X = 3) = C(10,3) \cdot 0.50^3(1 - 0.50)^{10-3} \approx 0.117 \text{ or } 11.7\%$$

What is interesting is that even though the probability of finding a student who likes Meat Lovers is significantly higher, the probability of finding exactly 3 out of 10 students who prefer Meat Lovers is significantly lower than finding 3 who prefer Hawaiian. Why is that?

It is because the binomial distribution considers not just the total number of trials and the number of successes, but also considers the underlying probability of success.

In our pizza example, since more students prefer Meat Lovers, we're more likely to see a higher number of students preferring Meat Lovers in a random sample of 10 students. Conversely, fewer students prefer Hawaiian, so it's more probable to see a lower number of students preferring Hawaiian in our sample.

To finish off, I've provided a visualisation of the binomial distributions for both Meat Lovers and Hawaiian preferences below. This gives us some interesting insights into how different population proportions can influence the probabilities of different outcomes in a binomial distribution.



4.4 GEOMETRIC DISTRIBUTION

The geometric distribution is a probability distribution that models the number of trials needed to get the first success in repeated, independent Bernoulli trials. Here's an easy way to remember it: the geometric distribution effectively answers the question, "How many times do I need to try before I succeed?"

Just like the Binomial distribution, we can prove that the PMF of the geometric distribution using the concept of independent Bernoulli trials.

Suppose we denote the number of trials required as k. The event of interest is that the first success happens on the k-th trial.

Like in the binomial example, the Bernoulli trials are independent, and each trial has a success probability of p. Therefore, any specific sequence of $k - 1$ failures followed by a success has a probability of:

$$(1 - p)^{k-1} \cdot p$$

This represents the probability of getting $k - 1$ failures in a row, and p represents the probability of getting a success on the k-th trial.

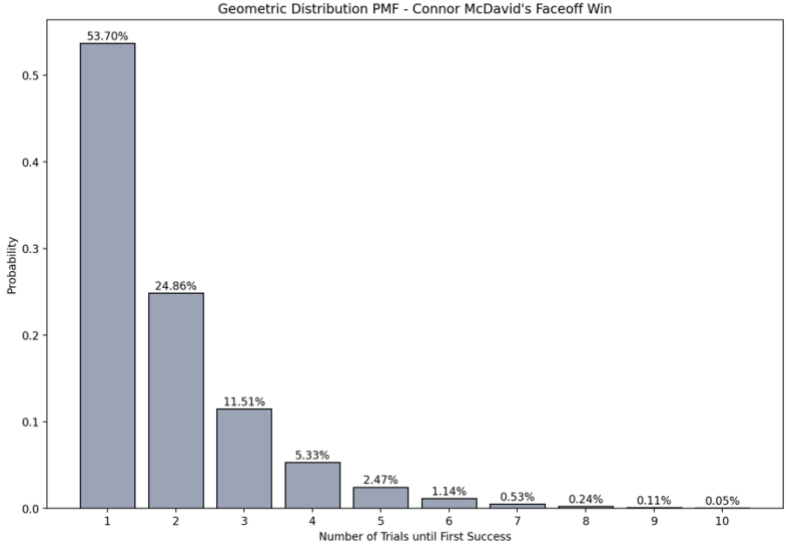
However, unlike the binomial example, there's only one "ordering" of outcomes that results in the first success occurring on the k-th trial. That is, you must observe $k - 1$ failures followed by a success. This means that there are no combinations to consider, and the probability of the first success occurring on the k-th trial is just the probability of this single sequence of outcomes.

Therefore, we have the PMF of the geometric distribution as follows:

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

Now, consider ice hockey player Connor McDavid who has a faceoff win percentage of 53.7%. Let's suppose you and a mate make a wager about whether the Connor will succeed on their first faceoff after initially missing the first two. Considering the player's success rate, you may feel confident about the bet, but let's determine the actual probability of success on the third attempt.

$$\mathrm{P(X = 3)} = (1 - 0.537)^{(3-1)} \cdot 0.537 \approx 0.115 \text{ or } 11.5\%$$



4.5 HYPERGEOMTRIC DISTRIBUTION

Hypergeometric distribution differs from the first two distributions (Bernoulli and geometric) because the trials are no longer independent, and the probabilities do not remain constant. This is because we are dealing with a finite population and selections are made without replacement. This changes the probabilities with each trial.

Suppose we have a finite population of N items, where K items are considered "successes" and $N-K$ items are "failures". We draw n items without replacement from this population.

Any specific sequence of k successes and $n-k$ failures has a certain likelihood. However, unlike in the Bernoulli or the geometric distributions, the probability is not simply:

$$p^k(1-p)^{n-k}$$

This is because the events are not independent and the probability changes with each draw.

However, like the Bernoulli and geometric scenarios, there are many ways that k successes can occur in n draws. We need to consider all possible orderings of these successes and failures.

- The number of ways that we can choose k successes from the K available is given by the binomial coefficient $\binom{K}{k}$ – which is the equivalent to $C(n,k)$ which we used prior.
- The number of ways we can choose $n-k$ failures from the $N-K$ failures available is given by $\binom{N-K}{n-k}$.

From this, we know that the total number of ways we can obtain a sequence of k successes and $n-k$ failures is by n draws in the given product $\binom{K}{k}\binom{N-K}{n-k}$.

Lastly, the total number of ways one can draw n items from the population N is given by the binomial coefficient $\binom{N}{n}$.

Therefore, with this understanding, we can conclude that the hypergeometric distribution is obtained by taking the ratio of the number of favourable outcomes to the total number of outcomes can be given by:

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

Let me introduce this with an example. Suppose that you are a consultant, and a high school has hired you to help them with a problem they are currently facing. They have around 2500 students, and they have recently found that around 150 of them are secretly using ChatGPT to solve their schoolwork. They know this only because of the similarities between each other work and because of the incorrect answers which the model produces.

They have an exam coming up which will provide a scholarship to a top University in Mount Sasquatch. There are 10 students who were selected for this exam, and they want to know the probability of finding exactly 3 students cheating in the exam.

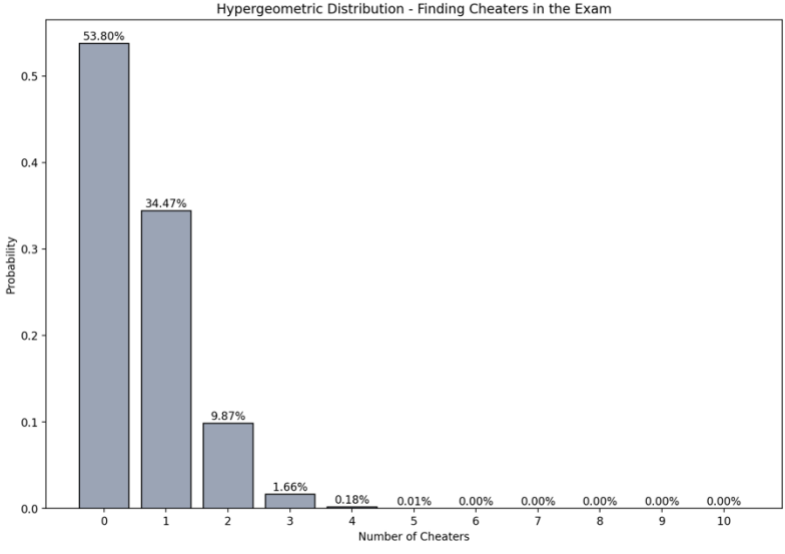
For this problem we can define the following:

- $N = 2500$
- $K = 150$
- $n = 10$
- $k = 3$

Putting this into the PMF function for Hypergeometric distributions:

$$P(X = 3) = \frac{\binom{150}{3} \cdot \binom{2350}{7}}{\binom{2500}{10}} \approx 0.0166$$

You find that the result of finding exactly 3 cheaters in the exam is approximately 0.0166 or 1.66%. You go back to the high school and provide them with the results. To assist them further you decided to compute the hypergeometric distribution for them, so they can view the different probabilities in the distribution of 10 students.



4.6 POISSON DISTRIBUTION

A very interesting distribution and quite different to the previous distributions in terms of the type of events it models and the parameter it uses.

The Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time or space, given a fixed average rate of occurrence. These events must be independent, meaning they don't affect each other.

The PMF of a Poisson distribution is given by:

$$P(k;\lambda) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

Where we have:

- $P(k;\lambda)$ is the Poisson probability, which asks the question "what is the probability of k events happening in an interval?"
- λ is the rate parameter (also known as the rate or mean number of events).
- k is the actual number of events that result.

Now, let's say you got another job, remember you're a consultant... the job is to assist Student Connect at Deakin University by providing insights into the number of phone calls they receive per hour. Suppose that they receive an average of 12 calls per hour. The Poisson distribution can then be used to calculate the probability of a certain number of calls (let's say 15 calls) happening in the next hour.

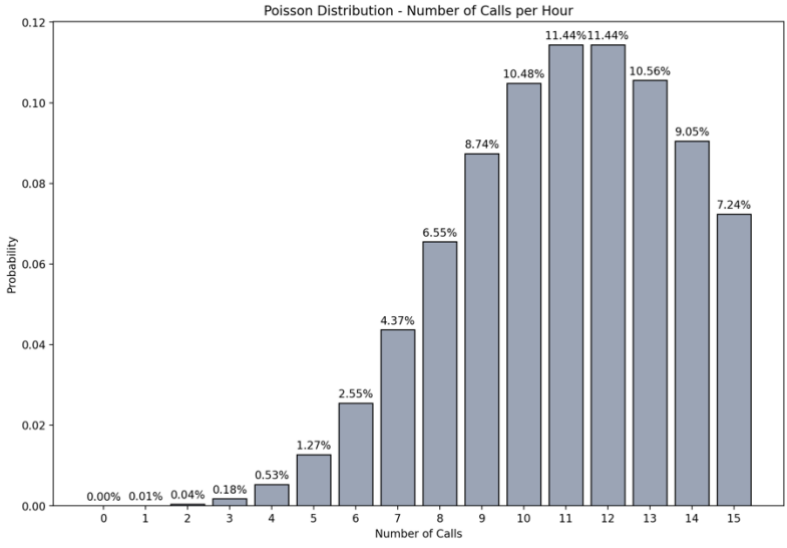
So here, we define the following:

- $\mathrm{P}(\mathrm{k};\lambda) = \mathrm{P}(15;12)$
- $\lambda = 12$
- $\mathrm{k} = 15$

You then plug in the numbers to the PMF:

$$\mathrm{P}(15;12) = \frac{12^{15} \cdot \mathrm{e}^{-12}}{15!} \approx 0.0724 \text{ or } 7.24\%$$

After performing the calculations, you find that there is approximately a 7.24% chance that Deakin Connect will receive exactly 15 calls per hour on any given day. They were happy with your result, and you also provided the visualisation of the Poisson distribution of 15 calls an hour with an average rate of 12 per hour.


Poisson Distribution - Number of Calls per Hour
Probability
Number of Calls
0.00%
0.01%
0.04%
0.18%
0.53%
1.27%
2.55%
4.37%
6.55%
8.74%
10.48%
11.44%
11.44%
10.56%
9.05%
7.24%

5 RECAP

We have finally reached the end of this module, to quickly recap. You were introduced to some core concepts in probability theory and their applications. It started off with probability models, which are mathematical representations that help us navigate and understand uncertain scenarios. We then explored conditional probability using Bayes' Rule, where we examined how prior knowledge or conditions can affect our probability calculations. Then finally, we looked at discrete distributions, which provide a way to model the probability of outcomes for discrete random variables.

We learnt that probability is more than just counting outcomes and assigning chances. It's about understanding the context, the conditions and making sense of it all through logic and quantitative reasoning. From the basic axioms of probability to complex concepts like Bayes' theorem and probability mass functions, each topic builds upon the previous, leading to a deeper understanding of this fascinating field.

Probability theory is an incredibly powerful tool. It underpins many practical applications, from data science and predictive modelling to everyday technologies. Understanding these fundamental concepts doesn't just deepen your mathematical knowledge; it helps you see the world in a different way. As a quote mentioned prior by Laplace, "Probability theory is nothing but common sense reduced to calculation".

But understanding the theory is only half the journey. The other half is applying it. Whether you're predicting customer behaviour, modelling outcomes of a basketball game, or navigating the odds in a game of chance, the principles and methods learned in this module are valuable tools.

I do recommend to check out E. T. Jaynes's "Probability Theory: The Logic of Science" and definitely recommend to look through Foundations of the theory of probability by Kolmogorov, A. N. Remember, probability isn't just a branch of mathematics, it's a way of thinking, a way of making sense of the world around us, one uncertainty at a time.