Chapter 1

Fundamentals

1.1 Brief Background

The study of the mathematical theory of probability started off as an analysis of games of chance in the 16th century by Gerolamo Cardano, and then by Pierre de Fermat and Blaise Pascal in the 17th century. The initial investigation of Probability Theory is on discrete events using mostly combinatorial methods, like the studying of coin tosses and the chances for taking certain coloured balls from a bag. Later on, analytical tools were included in the discipline as people began to study more complicated events.

The study of Probability Theory does not stop here. By using the language of Measure Theory, Soviet Mathematician Andrey Kolmogorov, along with many others, built the foundation of modern Probability Theory using his axiom system in 1933. This work is published in his book *Foundations of the Theory of Probability*. It is also due to this rigorous construction of the foundation that has caused Mathematicians to view Probability Theory as a proper branch of Mathematics. Unfortunately, we will not be studying much of these more advanced work in this notebook, since this is supposed to focus primarily on the more elementary side of things. The good news, however, is that I am working on another notebook *Probability Theory and Statistics Notebook II* which will be on the measure theoretic Probability Theory and Statistics.

1.2 Probability Space

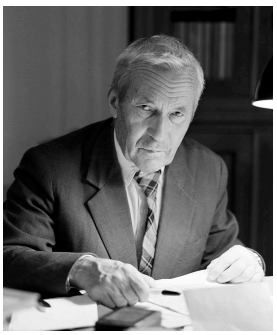
In the theory of probability, we will be working 'in' a **probability space**. A probability space is a triple $(\Omega, \mathcal{F}, P)$ consisting of three objects - sample space Ω, events $\mathcal{F}$ and probability measure P.

The **sample space** is the set of all possible outcomes of an experiment. For example, the sample space of a single-time coin tossing experiment will be the set $\{H, T\}$ where H

represents the result head and T represents tail. The sample space is sometimes called the universe as it encapsulates every possible outcomes.

An **event**, one element of $\mathcal{F}$, is a subset of the sample space. We can understand events $\mathcal{F}$ as a set consisting of all possible outcomes of the experiment[1]. Using the previous example, a possible event of the experiment is $\{H\}$ where the coin lands head.

The **probability measure**[2] is a function that maps an event to a numerical value under certain restrictions.



Andrey Kolmogorov

The restrictions are known as the **probability axioms**. They are introduced by Kolmogorov, so sometimes it is also known as the **Kolmogorov axioms**[3]. We will do some minor adjustments and rephrasing so that they are more understandable at this level. A **probability law** will be the function that satisfies the probability axioms. Now given a probability measure P, we have the following axioms:

1. (**Nonnegativity**) $P(A) \geq 0$ for all event A.

2. (**Countable Additivity**) If the sample space contains countably many[4] disjoint events $A_1, A_2, \ldots$, then the probability of their union satisfies

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

When we let some of the events to be empty set, we will get finite additivity. For example, if all events except A_1 and A_2 are null sets, we have

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

3. (**Normalisation**) The probability of the entire sample space Ω is equal to 1, which is $P(\Omega) = 1$.

From the axioms, we can derive many useful properties of probability measure. One such property is that $P(\varnothing) = 0$ since $1 = P(\Omega \cup \varnothing) = P(\Omega) + P(\varnothing) = 1 + P(\varnothing)$.

More properties of probability measure and events will be listed below without proof. We expect the readers to have some familiarities with common set operations like union, intersection and complement. They will not be defined and heavily discussed in this note.

[1] Here, all elements of the power set of the sample space are defined as events. This is true when we are working with finite sample space - which is the assumed context of this notebook unless stated otherwise.

[2] Probability measure is the formal name. The discussion of measure and measurable will not be carried out in this notebook. The study of measure is known as Measure Theory, categorised under Analysis.

[3] Although the axioms of probability are named after Kolmogorov, that does not imply that he is the one who created the entire field. Instead, the naming exists because Kolmogorov restated the axioms in a rigorous and powerful manner that no one else did before. In addition, we should remember that this is not the only way to construct the theory of probability.

[4] Set S is countable if S can be put in 1-1 correspondence with the set of all positive integers.

The proofs are expected to be completed by the readers and this will not likely to be a difficult task.

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ P(B) &= P(A) + P(A^c \cap B) \\ P(A \cup B) &\leq P(A) + P(B) \\ (\cup_i A_i) \cap B &= \cup_i (A_i \cap B) \\ (\cap_i A_i) \cup B &= \cap_i (A_i \cup B) \end{aligned}$$

–

To give some more examples of a probability space, we look at the discrete cases when the sample space consists of a finite number of possible outcomes. We have, according to the aforementioned axioms, the probability of any event $\{A_1, A_2, \ldots, A_n\}$ is the sum of the probabilities of its elements

$$P(\{A_1, A_2, \ldots, A_n\}) = P(\{A_1\}) + P(\{A_2\}) + \cdots + P(\{A_n\})$$

where each A_i represents an outcome.

If the probability of each of the element is uniform, we have

$$P(A) = \frac{\text{number of elements of } A}{n}$$

when there is a total of n possible outcomes. This will be the probability in the discrete uniform case.

–

This section talks about probability space. In Mathematics, a lot of 'spaces' are defined and used. These definitions, along with many other, are used mostly for expository purposes. This technique allows us to better group objects into more abstract but general categories, since 'abstractness is the price of generality'. With this, we can find common properties for each category and apply them to all objects under it. This habit, originated from Euclid, allows us to formalise and simplify the writing, remove repeated conditions, and more importantly to emphasise the significance of structure and the 'big picture'. We will see more of this in the study of, for example, Functional Analysis with things like Hilbert Space and Banach Space.

1.3 Conditional Probability and Independence

1.3.1 Conditional Probability

Conditional probability is an important concept, just like almost everything else in this notebook. There are two ways to see its importance. One is that we can calculate

probabilities when it is conditional, meaning some partial information are given. The other is that even when we have no partial information, conditional probabilities can be used to compute the desired probabilities more easily using the law of total probability.

–

Let us start with defining what conditional probability is. Given two events A and B, the **conditional probability**, denoted by $P(A|B)$, is the probability of event A happening when we know event B has already happened, or equivalently when $P(B) > 0$.

In order to compute that probability, we will have the formula

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

This is not too hard to understand. If we think about it in the discrete uniform case, the conditional probability of A given B is basically the number of elements in both A and B, or in $A \cap B$, divided by the number of elements in B. Another way of looking at this is by using the Venn Diagram, which will be quite obvious once the diagram is drawn.

–

The conditional probability satisfies the three axioms of probability and is therefore a probability law. We can verify it pretty easily. First, it is obviously nonnegative. Then, we notice that

$$P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1,$$

which gives us the normalisation. We can see that the condition actually changes the universe from Ω to B and all conditional probabilities of B will be concentrated within this event. Lastly, given two disjoint events A_1 and A_2, we have

$$\begin{aligned} P(A_1 \cup A_2|B) &= \frac{P\big((A_1 \cup A_2) \cap B\big)}{P(B)} \\ &= \frac{P\big((A_1 \cap B) \cup (A_2 \cap B)\big)}{P(B)} \\ &= \frac{P(A_1 \cap B) + P(A_2 \cap B)}{P(B)} \\ &= \frac{P(A_1 \cap B)}{P(B)} + \frac{P(A_2 \cap B)}{P(B)} \\ &= P(A_1|B) + P(A_2|B) \end{aligned}$$

where the second equality is valid since $A_1 \cap B$ and $A_2 \cap B$ are disjoint sets. This can be generalised to countably infinite disjoint events, which shows countable additivity.

–

We can manipulate the formula slightly. By some multiplication, we will have, under the same conditions,

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

meaning the probability that both A and B occur is equal to the probability that B occurs multiplied by the conditional probability of A given B occurred.

We can extend the manipulation further. We used two events A and B just now. What if we use more than two events? Given events $A_1, A_2, \ldots, A_n$, we have the **multiplication rule**

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

This can be proved by applying the definition of conditional probability n times.

–

Previously, when we want to find the probability of an event, we do it by comparing that event with the entire sample space. Now, since we know conditional probability allows us to switch sample space from one to another, we can use that to come up with a different way of computing. By dividing the event we want to find the probability of into sections according to the partition of the universe, we will get the chance of the event occurring under each pieces of the universe using conditional probability. By multiplying the sectional probability of event by the probability of that piece of partition and summing all sections up, we will get the total probability of the event.

This method is called the **total probability theorem**, or the law of total probability. The formal definition is the following. Let $A_1, \ldots, A_n$, each with possible probability of occurring, be a partition of the sample space, the probability of any event B will be

$$\begin{aligned} P(B) &= P(A_1 \cap B) + \cdots + P(A_n \cap B) \\ &= P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n). \end{aligned}$$

1.3.2 Bayes' Rule

Bayes' rule, named after Reverend Thomas Bayes, is the next thing we will be looking at in this section. The statement of the rule can be obtained easily by a small manipulation of the formula for conditional probability, but it has a deeper meaning under.

Conditional probability tells us that $P(A \cap B) = P(B)P(A|B)$. We know $A \cap B$ and $B \cap A$ are two identical sets, so we have $P(B \cap A) = P(A)P(B|A)$. Combining the two, we get $P(B)P(A|B) = P(A \cap B) = P(A)P(B|A)$, which implies the Bayes' rule

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

Here, although we have never mentioned explicitly before, we are interpreting probability measure as "a proportion of outcomes". It is easy to understand if we think about the discrete uniform case. This way of thinking is known as the **frequentist** interpretation.

The importance of Bayes' rule does not end here. Other than a "proportion of outcomes", we can also interpret probability measures as a "degree of belief". A 50% success rate means that we believe with 50% certainty that the result is going successful. The term "belief" is very important as it suggests the view can be very subjective.

The formula of Bayes' rule can be rearranged into $P(A|B) = P(A) \cdot \frac{P(B|A)}{P(B)}$. Here, $P(A)$, the *prior*, is the initial degree of belief of A. The fraction represents an adjustment of our belief using B. $P(A|B)$, the *posterior*, is the degree of belief after the adjustment. Our original proposition A is therefore been improved by evidence B. The Philosophy of Bayes is that we can never understand how things go exactly, but by having more and more evidence, we can have a better understanding of the truth by updating with each new piece of evidence[5]. This is known as the **Bayesian** interpretation.

Bayes' rule, or more specifically the underneath philosophy behind it, is the foundation of **Bayesian statistics** , which is a different school of thought in Statistics other than the frequentist one. We will elaborate more in Part II of this notebook.

Earlier on, we derived the total probability theorem which is often used in conjunction with Bayes' rule. Let $A_1, \ldots, A_n$, each with possible probability of occurring, be a partition of the sample space. For any event B, we have

$$\begin{aligned} P(A_i|B) &= \frac{P(A_i)P(A_i|B)}{P(B)} \\ &= \frac{P(A_i)P(A_i|B)}{P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n)}. \end{aligned}$$

1.3.3 Independence

The second half of this section will be on independence. From the dictionary definition of the term "independence", we know that it roughly describes a state where objects are isolated and do not affect each other. This is similar to what **independence** means in Probability Theory. When we say two events A and B are independent, or A and B are independent events, we mean the occurrence of one does not affect the probability of occurrence of the other, denoted by the following formula

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

The formula says that the probability of both events happening is the same as the probability of the first event occurring then the second one occurring when the two events are independent.

[5]The readers are encouraged to check out 3Blue1Brown's video "Bayes Theorem" on YouTube for a graphical illustration.

Another way to understand independence is by using conditional probability. Conditional probability $P(A|B)$ captures the partial information that event B provides to event A. The special case arises when the occurrence of B provides no information and does not alter the probability of A occurring, i.e. $P(A|B) = P(A)$.

When objects do not have independence, we will say they are **dependent**.

Independence does not restrict to two events. A series of events can be independent of each other. However, the conditions will get more complicated when we move from two to many events. Given three events A, B, and C, they are said to be independent if

$$\begin{aligned} P(A \cap B \cap C) &= P(A)P(B)P(C) \\ P(A \cap B) &= P(A)P(B) \\ P(A \cap C) &= P(A)P(C) \\ P(B \cap C) &= P(B)P(C) \end{aligned}$$

In general, for a collection of independent events $A_1, A_2, \ldots A_n$, for any $I \subset \{1, 2, \ldots, n\}$, we have

$$P(\cap_{i \in I} A_i) = \prod_{i \in I} P(A_i).$$

For multiple events, they may not be independent but **pairwise independence**. Pairwise independence means that for any two events from the list of events are independent of each other. This means, for a collection of events $A_1, A_2, \ldots A_n$, they are pairwise independent if $P(A_i \cap A_j) = P(A_i)P(A_j)$ for all $i, j \in \{1, 2, \ldots, n\}$ and $i \neq j$.

It is not hard to see that independence implies pairwise independence. This relationship can be observed if I are all 2-element subsets of $\{1, 2, \ldots, n\}$. Pairwise independence, on the other hand, does not imply independence.

So far, the discussion of independence is restricted to events. Independence can occur between random variables and sample spaces too. We should also remember that the independence of sample spaces, independence of random variables, and independence of events are equivalent. Most of the detailed explanation will not be included as it is beyond the scope of this notebook. They will, however, be in the second notebook *Probability Theory and Statistics Notebook II*.