Last Updated: January 7, 2024

Chapter 1

An Introduction to Elementary Probability

1.1 Introduction

The idea of probability is to be able to quantify the "chance of something taking place." The chance of obtaining a "tail" on tossing a fair coin is 0.5. What does this mean? On tossing a fair coin 10 times, are precisely 5 "tails" observed each time? It has been observed in many fields that on performing an experiment a sufficient number of times, eventually, certain averages approach constant values. What are these averages? What are these experiments? The probability of an event E is a real number P(E) associated with it. How do we define the event E? What exactly is P(E)? Are there any constraints on its range of values? Why do we even need this? At this point, the concept of probability is too vague for it to be useful mathematically. Thus, we must first formalize this intuition.

1.2 The Axiomatic Definition of Probability

Before we formally define the probability space, we must recap a few mathematical notions.

A σ-algebra of subsets of a set X

A σ-algebra of subsets of a set X is a collection $\mathcal{F}$ of subsets of X that satisfy the following conditions:

1. $\phi \in \mathcal{F}$
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
3. If $A_1, A_2, \cdots$ is a countable collection of sets in $\mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Generated σ-algebras

Let X be a set and $\mathcal{B}$ a non-empty collection of subsets of X. The smallest σ–algebra containing all the sets of $\mathcal{B}$ is denoted by $\sigma(\mathcal{B})$. We use smallest to indicate that any σ-algebra containing the sets of $\mathcal{B}$ will also contain ALL the sets of $\sigma(\mathcal{B})$ as well.

Consider the set $X = \mathcal{R}$. The smallest σ-algebra that contains all intervals of the form (a, b), where $a, b \in \mathcal{R}$ is called a Borel Sigma Algebra.

1.2.1 The Probability Space (Ω, F, P)

Sample Space (Ω)

The sample space of an experiment is the set of all possible outcomes of the experiment.

Event Space (F)

The event space is a σ-algebra of subsets of Ω (We sometimes simply write this as a σ-algebra of Ω). The subsets of Ω are called events.
Let's consider an example to illustrate this idea.

Example 1: Consider the sample space of a dice being rolled. What are the smallest possible event spaces? What is the smallest event space containing { {1}, {1,2} }?
The smallest event space is { ϕ, {1,2,3,4,5,6} } since ϕ must be present in every event space. The complement of ϕ is Ω, which is {1,2,3,4,5,6}.
The smallest event space containing {1} and {1,2} is { {1}, {1,2}, ϕ, Ω, {2,3,4,5,6}, {3,4,5,6}, {1,3,4,5,6}, {2} }. The first two elements are required according to the demands of the example. As discussed, ϕ and Ω must be present. The complement of {1} must be present; thus, we have {2,3,4,5,6}. The same goes for {1,2}, so we have {3,4,5,6}. Now, the union of {1} and {3,4,5,6} gives {1,3,4,5,6} while its complement is {2}. Notice that using the rules, we can determine the event space under any set of constraints.

Probability Measure (P)

The probability measure is a function from the event space to the set of real numbers between 0 and 1 (both inclusive), i.e., $P : F \to [0, 1]$ such that:

1. $P(\Omega) = 1$
2. For a countable collection of disjoint events $(A_1, A_2, \cdots)$:
   $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

So what does this mean? The critical point to note here is that when we assign real numbers, when we use the term probability of an event, it denotes a number corresponding to an event taken from the event space F, not from the sample space Ω. Therefore, when we assign a probability, it is a number corresponding to an element of a subset of the sample space. We now see the importance of the event space being a σ-algebra. Say this wasn't the case. Say, condition number 2 in the definition for σ-algebras wasn't compulsory: what would happen? There could then be an event space containing just one subset and not its complement. The union of probability measures may not be 1. This would violate the condition regarding the probability of a countable union of disjoint subsets.

An obvious question arises, however. Why do we assign a probability measure to events? Why not assign them to outcomes?

1.2.2 Outcomes and Events

An outcome is the result of a random experiment (one can say that an outcome is an "elementary event"). An event is a set of outcomes to which a probability measure is assigned. Let's consider an example to illustrate this idea.

Example 2: Consider the set of real numbers between 0 and 1, i.e., $[0, 1]$. What is the probability of randomly choosing a number in the range $[0.2, 0.3]$- the answer seems obvious- 0.1, and is indeed correct. However, what is the probability of choosing a number between 0 and 1? It is 0! This is because there are uncountably many numbers between 0 and 1. Similarly, there are uncountably many numbers between 0.2 and 0.3 as well. Despite the fact that each real number in the interval is unique (i.e., you can't simultaneously obtain two real numbers at the same time), the disjoint union rule discussed above is not applicable here. This is because the set under scrutiny is uncountable! However, every real number in the interval is a valid outcome of the random experiment.
This is precisely why we don't assign probability measures to outcomes. To avoid exactly these kinds of issues, we assign them to events- something you inherently did in the illustration above!

‘

We will now analyze why this axiomatic definition was needed in the first place.

1.3 Alternate Definitions of Probability

1.3.1 Relative Frequency Definition

The probability of an event A is defined as $P(A) = \lim_{n\to\infty} \frac{n_A}{n}$. where n_A is the number of occurrences of event A and n is the number of trials. What is the issue here? Well, infinity is only a concept. The number of trials can be large, very large in fact, but never quite infinity. After all, the number of trials will always be finite. Thus the ratio of occurrences to trials cannot be approximated as a limit. Therefore, this is only a hypothesis. It isn't something that can be determined experimentally. This definition can be useful in determining patterns or to obtain an idea of what the probability of an event might be, but never the exact value.

1.3.2 Classical Definition

The probability of an event A is defined as $P(A) = \frac{N_A}{N}$ where N_A is the number of "favorable" outcomes (corresponding to event A) and N is the number of possible outcomes. Here, the primary assumption is that all events are equally likely. Only then does this definition work. Consider the case where we use 2 dice. What's the probability of obtaining a sum of 6? Well, the possible sums go from 2 through to 12. Thus, the answer should be $\frac{1}{11}$. That doesn't seem right. How aboubt we try a different approach? Consider the possible pairs $(1,5), (2,4), (3,3)$- all these give a sum of 6. Notice we have distinguished the two dice. This gives a probability of $\frac{3}{21} = \frac{1}{7}$. However, had we considered them as distinct, it would've been $\frac{5}{36}$, which is the correct answer.

Another issue here is with the definition of "favorable". It is simply too open to interpretation. A classical paradox has been discussed below to show the lack of efficacy of the above experiment:

Bertrand's Paradox

Consider a circle C with a radius of r. What is the probability that the length of a chord drawn in this circle is more than $r\sqrt{3}$? Depending on how you attempt to solve this question, you could obtain an answer of $\frac{1}{2}, \frac{1}{3}$ or $\frac{1}{4}$, all of which seem justified and reasonable. This illustrates the vagueness in "favorable."

1.3.3 Other Paradoxes, Interesting Problems and Motivation

The 100 prisoners problem

There are a 100 prisoners, each with a number tag, an integer between 1 and 100 (both inclusive). There are a 100 boxes, each labelled with a number between 1 and 100 (both inclusive), and each containing a piece of paper with a number written on it. Once again, this number is between 1 and 100 (both inclusive). The prisoners are called one by one into a room containing the boxes. They must find the box containing the box containing their number tag. They are allowed to open at most 50 boxes each. If they succeed in finding their number, they win. In order for the prisoners to escape, all 100 prisoners must win. What is the probability of this happening if an ideal strategy is followed? What is this strategy?

Solution: If each prisoner chooses a number at random, the chance of escape is $(0.5)^{100} = 8 \times 10^{-31}$. This isn't the right way to go about it. The ideal strategy is to open the box labelled with your tag number. If a prisoner finds their own ticket, they have won. If not, the prisoner must then open that box with the label of the number contained inside the previous one. For example, say your number is 15. So, open box 15. If it contains the number 15, you are done. If not, say it contains some number x. Now open box x and continue till you have opened either 50 boxes or you have won. What's the probability of success in this case?

Every box contains one ticket, and the tickets are unique. This means that either a ticket is in the correct box or it points to another box. As the tickets are distinct, there is only one ticket pointing to each box (and only one way to get to any box). The boxes will form circular chains. A box can only be part of one chain because there is only one pointer in and one pointer out of any box (think of linked lists). If a box doesn't have its own ticket, it will be in a chain. Because the prisoner starts on the box of their own number they are, by definition, on the chain that contains their ticket (there is only one ticket that points to that box). By following the chain around, they are guaranteed to end up at their ticket by following the chain. If the chain length of a given number is less than or equal to 50, a prisoner will win. Thus, for all prisoners to escape, the idea is to have the maximum chain length less than or equal to 50. What is this probability?
P(winning)=1-P(losing)=1-P(at least 1 chain of length greater than 50)
The probability that there is a chain of length l is $\frac{1}{l}$ (why?). Thus the required result is $1 - \sum_{i=51}^{100} 1/i \approx 0.31$

A very important idea is illustrated here. We can always express probabilities in terms of complements in terms of which entity is easier to calculate. This concept will be used a lot in future problems, in this course as well as many others.

The Birthday Problem

In a group of 23 people, what is the probability that at least two people share their birthdays?

Solution: We use the concept that we have discussed above. P(at least two people share their birthday)=1-(no two people share their birthday)$=1 - \frac{364}{365}\frac{363}{365}\cdots\frac{343}{365} \approx 0.5$
To simplify such expressions, we often use exponential approximations, such as $e^x \approx 1 + x$ for small values of x. So, in a class of about 100 students, the probability that at least two people share the same birthday is 0.9999996927510721! (Note: Be careful

while using exponential approximations. On using the approximation, the answer won't be 1, even for 366 people, which is clearly misleading (we aren't considering leap years here)- this is because x should b small for the approximation, and as the number of people increases, this property withers away). This idea can then be generalized and is very useful in real-life applications.

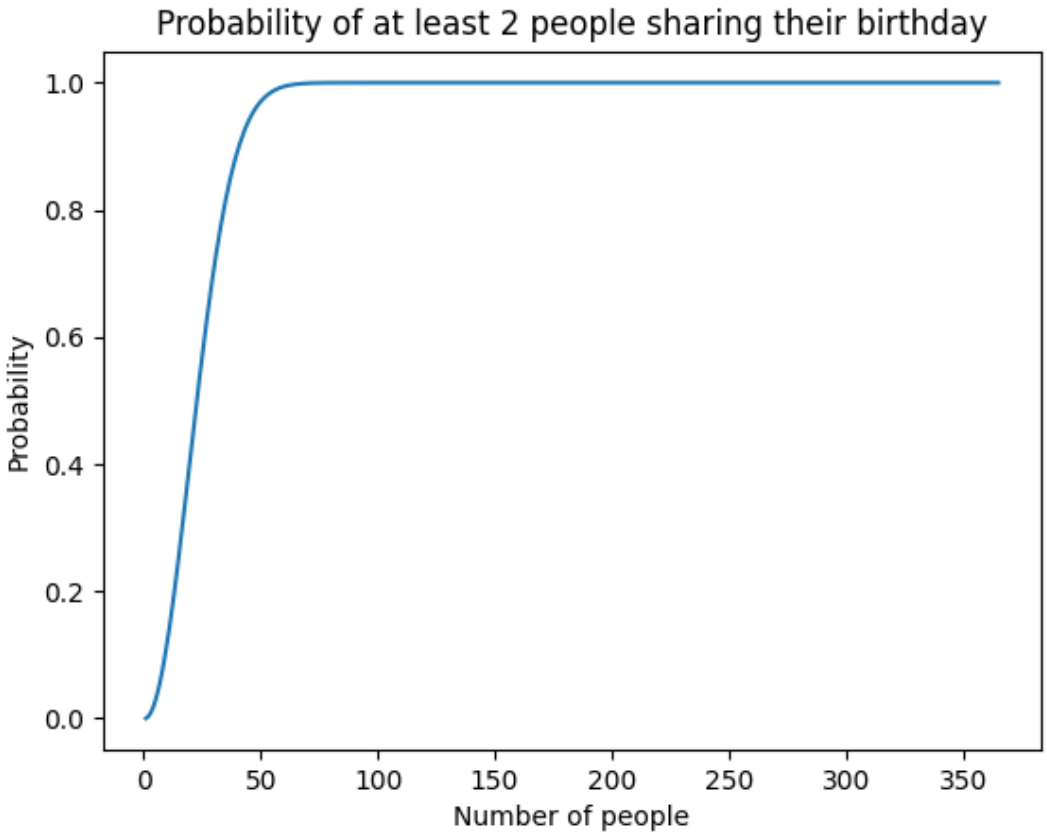

Figure 1.1: Simulation of probabilities for the birthday paradox

Bertrand's Box Paradox

There are three boxes. One of them contains two gold coins, one of them contains two silver coins, and the third one contains 1 gold and 1 silver coin. Now, you pick a box at random and pick up a coin. It happens to be a gold coin. What is the probability that the other coin in the same box is also gold?

Incorrect Solution: Since there are only two boxes containing gold coins, if I have picked up a gold coin, I must have picked up one of the two boxes containing a gold coin. Thus, the other coin could either be gold or silver. Thus, the probability of the other coin being gold and thus the box containing two gold coins is $\frac{1}{2}$

Correct Solution: Assume that you pick up a box but do not choose a coin. Label the coins from 1 to 6. Say the gold coins are numbered 1,2 and 3, whilst the silver ones are numbered 4,5 and 6. Say that one of the boxes contains the coins 1 and 2, while the other two contain 3 and 4, and 5 and 6, respectively. The probability of choosing a gold coin with a particular number (say, the one with the label 1) is $\frac{1}{6}$. Now, let us analyze what cases are useful: Since the first coin is gold, we must have picked 1, 2 or 3. Three possible cases thus arise: We have picked up gold coin 1, and thus, we have picked the first box, and the next coin we pick will be number 2. Another case is that we have picked up gold coin 2, and thus, we have picked the first box, and the next coin we pick will be number 1. Finally, we might have picked gold coin number 3, and this means we have chosen the second box and the next coin we pick will be silver. Thus, 2 out of the 3 cases are favorable, and the required probability is $\frac{2}{3}$ and not $\frac{1}{2}$. This might seem a bit counterintuitive. One might also feel that we are missing out on several cases and not accounting for the probability in a meticulous fashion.

The solution to this problem is much more clear when we apply Bayes' Theorem, a very important concept and the motivation behind discussing this problem.

The Boy-Girl Paradox

A parent has two children. At least one of them is a boy. What is the probability that both of them are boys?

Solution: First things first, the answer is NOT $\frac{1}{2}$. Consider the possible cases: {Boy-Boy}, {Boy-Girl}, {Girl-Boy} and {Girl-Girl}. Notice the 4th case need not be considered. In only one of the three cases is our outcome favorable. Thus, the required probability is $\frac{1}{3}$. The second and third case are indeed different based on which child is older. If the question is rephrased as "the elder child is a boy", then the probability would indeed be $\frac{1}{2}$.

The Monty Hall Problem

There are three doors, behind which are two goats and a car. You first pick a door (but are not allowed to open it). Monty Hall, the host of the game show, examines the other two doors and opens the door behind which there's a goat. Your objective is to get the car. Should you stick with your choice or change to the other door?

Solution: Yet again, it is not a 50-50 chance; the mathematics is slightly more involved. There is a $\frac{1}{3}$ chance that you have chosen the door with the car. In the remaining $\frac{2}{3}$ cases, Monty Hall, the host, can only open one door since you have chosen the door with the goat. Thus, he is indirectly telling you where the car is; it is in the door you haven't chosen! Of course, this happens in only two-thirds of the cases, but it is still your best bet. There is only a one-in-three chance that you have guessed correctly initially-make the switch !! In the plots, for each simulation, the result for switching your guess and staying with your original decision is shown. Notice how the experiment backs up the theory.

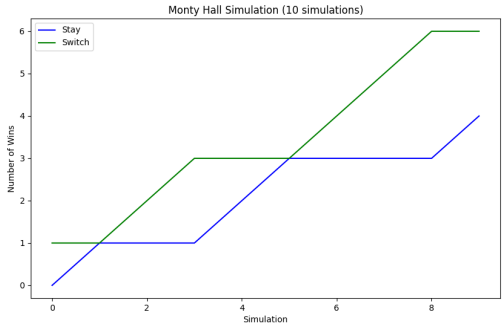

Figure 1.2: Simulation of the Monty Hall Problem (10)

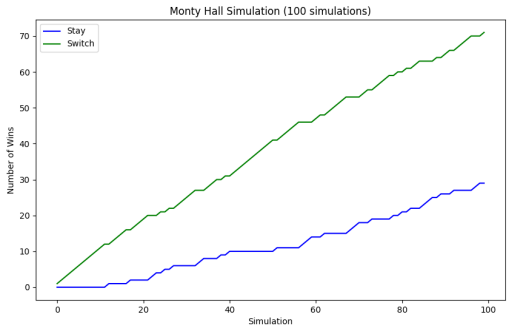

Figure 1.3: Simulation of the Monty Hall Problem (100)

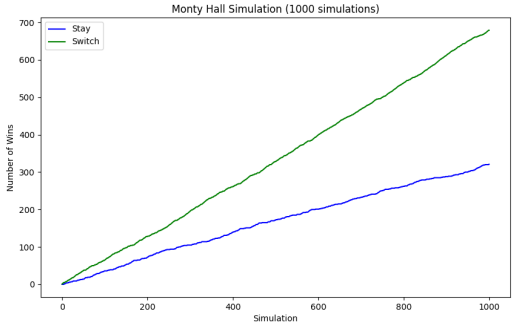

Figure 1.4: Simulation of the Monty Hall Problem (1000)

Generalizing the idea: What if there are d doors, c cars and o doors are opened (you still get to pick exactly 1 door)? What is the probability of getting a car by switching doors? The answer is $\frac{(d-1)c}{d(d-o-1)}$. Try to get to this result! What if the player can open p doors at first and then switch k of them? What would the new probability be?

Simpson's Paradox

Consider a new drug to cure an illness. Say it is used on 1 woman and 4 men. Say the woman survives, and only 1 of the 4 men survives. Consider 4 other women who had the illness but were not given the drug. Say 3 of them survive. Finally, consider 1 man who had the illness but wasn't given the drug, and he didn't survive. Thus, 75% of untreated women survive while 100% of treated ones do and 0% of untreated men survive but 25% of treated men do. Thus, a possible conclusion would be that the drug increases the chances of survival.
However, on aggregating the data, only 40% of the people who received the drug survived, while 60% of the people who didn't receive the drug survived. Thus, the drug isn't actually very effective. This is Simpson's paradox and this has to do with causality and data aggregation.

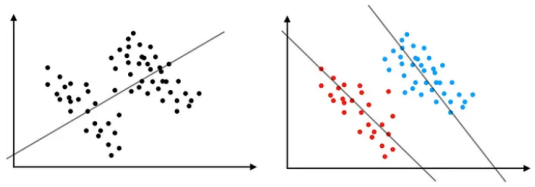

Figure 1.5: The Simpson's Paradox- individual trends v/s overall trends

The Wine/Water Paradox

A mixture contains wine and water. Say the ratio of wine and water is x, and x lies between $\frac{1}{3}$ and 3. What is the probability that, say, x is lesser than 2? The objective here is to find out apriori distributions- this will be introduced in upcoming chapters.

1.4 Conditional Probability

Till now, we have assumed no prior information while determining the probability of an event. In this section, we will discuss how to deal with scenarios when specific information is already available to us and we then try to caclulate the probabilities of certain events. The conditional probability of an event A assuming another event E is defined as follows:

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

We assume here that $P(E) \neq 0$. $P(A,E) = P(A \cap E) = P(AE)$. Can you show that this, indeed, is a valid probability measure?

1.4.1 Total Probability

Consider a set S, which can be partitioned into (disjoint, as the word partition of a set is defined) subsets as follows: $U = [A_1, A_2, \cdots A_n]$. Consider an event B, then we have:

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

This is known as the total probability theorem. The simple proof relies on the fact that S can be written in terms of a partition, and the union of probabilities of subsets is the probability of the union of subsets.

$$B = B \cap S = B \cap (\bigcup_{i=1}^{n} A_i) = \bigcup_{i=1}^{n} (B \cap A_i)$$

$$\implies P(B) = P(\bigcup_{i=1}^{n} (B \cap A_i)) = \sum_{i=1}^{n} P(BA_i) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

(using the definition of conditional probability)

1.4.2 Bayes Theorem

By rearranging the result of the total probability theorem, we get the following:

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

$P(A_i)$ is known as the apriori probability. $P(A_i|B)$ is known as the aposteriori probability. $P(B|A_i)$ is the probability of the event B given A_i is true. From a machine learning perspective, this is known as Likelihood and $P(B)$ is effectively marginalization, something which you will encounter later.

1.5 Independence

Two events are independent if the occurrence of one of them does not depend on the occurrence of the other. Mathematically:

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

Three events are independent if:

$$P(A \cap B \cap C) = P(A)P(B)P(C) \text{ and } P(AB) = P(A)P(B),\ P(BC) = P(B)P(C),\ P(CA) = P(C)P(A)$$

From this, we can inductively derive a result regarding the independence of n events: n events are said to be independent if *any* k of them ($k < n$) are also independent, given the definition of independence of two events as cited above.

Practice Questions

*Note: Some of the questions here are difficult and require a considerable amount of time to be solved. Solutions to such problems will eventually be uploaded. However, it is recommended that you do try all problems, since it will help you enhance your grasp on fundamental concepts in Probability Theory.

1. A box contains m red balls and n blue balls. Balls are drawn at random, one at a time, from the box. Find the probability of encountering a red ball by the k^{th} draw.

2. Two players, A and B, draw balls one at a time, without replacement, from a box containing m white balls and n black balls. What is the probability that the player who starts first picks the first white ball?

3. A box contains balls of only two colours- white and black. When two balls are drawn without replacement, the probability that both are white is $\frac{1}{3}$. What is the minimum possible number of balls in the box?

4. Two friends, A and B, plan to meet each other at 8 p.m. today. They decide to meet at a plaza between 8 and 8:20. A, once he gets there, will wait for only 4 minutes, and B once he gets there will wait for only 5 minutes. Assuming that they arrive independently, what is the probability that they will meet?

5. In terms of n, how many equations are required to establish the independence of n events?

6. Two players, A and B, play a set of games such that A is declared a winner if he succeeds in winning m games before B wins n games. The probability that A wins a game is p. There are no draws. What is the probability that B wins the set of games?

7. Lottery tickets are numbered from 1 through to 20. A player picks 3 numbers. He wins the lottery if the 3 lucky numbers are the ones he has chosen. What is the probability of winning the lottery?

8. Two players, A and B, start with a certain amount of money, say a and b respectively. They play a series of games. The probability of A winning a game is p, and there are no draws. The winner of each game in the series receives 1 unit of money from the loser. Find the probability that A wins the series- if the loser is the first one to go bankrupt.

9. Consider the previous problem, but with the loser paying k units of money instead of 1. How does your result change?

10. A pair of dice are thrown. A player loses if she throws a total of 2, 12, or 11. She wins if the first throw is 6 or 9. If anything else is thrown, the game continues. From every turn hereon, she wins if she throws the same value on two consecutive turns, and she loses if she throws a 6. What is the probability that she wins?

11. Two players, A and B, play a set of games, where the probability of A winning is $p = \frac{1}{2} - \epsilon$. A can decide the total number of games in the set under the constraint that this must be even. A person is declared the winner if they have won more than half the games. How many games should A choose to play to maximize his odds of winning?