Chapter 1

Probability basics

1.1 Introduction

Probability is the most important concept in modern science, especially as nobody has the slightest notion of what it means. ——Bertrand Russell

What is probability? *Chance. Luck. Coincidence. Fortune. Randomness. Coincidence...* We all talk about probabilities in everyday life, but mostly in vague languages. This course is to introduce probability as a logical framework for quantifying uncertainty and randomness.

Mathematics is the logic of certainty; probability is the logic of uncertainty.

The earliest development of probability is rooted in gambling. For instance, the renowned Monte Carlo method in statistics, invented by Stanislaw Ulam in the late 1940s, takes its name from the *Monte Carlo Casino* in Monaco, where Ulam's uncle would borrow money from relatives to gamble. Probability theories still apply today to analyze gambling odds, but their applications have expanded to nearly every field. It is the foundation of statistics, machine learning, and artificial intelligence. It also plays a crucial role in everyday decision-making, from stock investments to effective strategies to combat an infectious disease.

Probability is a concept that is intuitive to understand but very hard to define formally. Perhaps, the first formal definition of probability is often attributed to Pierre-Simon Laplace in the 18th century. In his work "Théorie analytique des probabilités," published in 1812,

> The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible.

This definition is outdated, as we will soon discover. But before we explore the modern definition of probability, let's first clarify some preliminary concepts (based on sets), which is the mathematical language we use to describe probabilistic events.

1.2 Events and sample spaces

The mathematical framework for probability is built around sets (like the cases in other math subjects as well).

Definition 1.1. The **sample space** S of an experiment is the set of all possible outcomes of the experiment. An **event** A is a subset of the sample space S. We say A occurred if the actual outcome is in A.

An experiment can be understood loosely. Anything (a gamble, an exam, a financial year, ...) can be an experiment. The sample space can be finite, countably infinite, or uncountably infinite. It is convenient to visualize events in a **Venn diagram**.

Set theory provides a rich language for expressing and working with events. Set operations, especially unions, intersections, and complements, make it easy to build new events in terms of already-defined events. For example, let S be the sample space of an experiment and let $A, B \subseteq S$ be events. Then the union $A \cup B$ is the event that occurs if and only if at least one of A and B occurs, the intersection $A \cap B$ is the event that occurs if and only if both A and B occur, and the complement A^c is the event that occurs if and only if A does not occur.

Example 1.1 (Coin flips)**.** A coin is flipped twice. We write 'H' if a coin lands Head, and 'T' if a coin lands Tail. The sample space is the set of all possible outcomes. Therefore, $S = \{HH, HT, TH, TT\}$. Let's look at some events:

1. Let A_1 be the event that the first flip is Heads. Then $A_1 = \{HH, HT\}$. Let A_2 be the event that the second flip is Heads. Then $A_2 = \{HH, TH\}$.

2. Let B be the event that at least one flip is Heads. Then $B = A_1 \cup A_2$.

3. Let C be the event that all the flips are Heads. Then $C = A_1 \cap A_2$.

4. Let D be the event that no flip is Heads. Then $D = B^c$.

Here is a list of events described in both English and set notations.

English	Sets
sample space	S
s is a possible outcome	$s \in S$
A is an event	$A \subseteq S$
A occurred	$s_{\text{actual}} \in A$
A or B	$A \cup B$
A and B	$A \cap B$
not A	A^c
at least one of $A_1, \ldots, A_n$	$A_1 \cup \cdots \cup A_n$
all of $A_1, \ldots, A_n$	$A_1 \cap \cdots \cap A_n$
A implies B	$A \subseteq B$
A and B are mutually exclusive (disjoint)	$A \cap B = \phi$
$A_1, \ldots, A_n$are a partition of S	$A_1 \cup \cdots \cup A_n = S$ and $A_i \cap A_j = \phi$ for $i \neq j$

1.3 Classical probability

Definition 1.2. Naive definition of probability:

$$P(A) = \frac{|A|}{|S|} = \frac{\text{number of outcomes favorable to A}}{\text{total number of outcomes in A}}$$

assuming the outcomes are *finite* and *equally likely*.

Example 1.2. Flip a coin twice. Find the probability of landing two heads.

Solution: There are four possible outcomes: {HH, HT, TH, TT}, each with equal probability. Therefore, $P(\text{HH}) = \frac{1}{4}$.

The naive definition is very restrictive. It has often been misapplied by people who assume equally likely outcomes without justification. Besides, it is easy to

conceive examples of probabilities that do not fit into this formula, e.g. probability of rain. By saying it is "naive", it is definitely not the preferred definition in this course.

Nonetheless, we do some examples using the naive definition as a warm-up. Calculating the naive probability of an event A often involves counting the number of outcomes in A and the number of outcomes in the sample space S, which usually involve some counting methods. We now review some of the counting methods (multiplications, factorials, permutations, combinations) that was introduced in high schools.

Multiplications. Consider a compound experiment consisting of two sub-experiments, Experiment A and Experiment B. Suppose that Experiment A has a possible outcomes, and for each of those outcomes Experiment B has b possible outcomes. Then the compound experiment has $a \times b$ possible outcomes.

Exponentiations. Consider n objects and making k choices from them, one at a time with replacement. Then there are n^k possible outcomes.

Factorials. Consider n objects $1, 2, \ldots, n$. A permutation of $1, 2, \ldots, n$ is an arrangement of them in some order, e.g., $3, 5, 1, 2, 4$ is a permutation of $1, 2, 3, 4, 5$. The are $n!$ permutations of $1, 2, \ldots, n$.

Permutations. Consider n objects and making k choices from them, one at a time without replacement. Then there are $P_n^k = n(n-1)\cdots(n-k+1)$ possible outcomes, for $k \leq n$. (Ordering matters in this case, e.g. $1, 2, 3$ is considered different from $2, 3, 1$)

Combinations. Consider n objects and making k choices from them, one at a time without replacement, without distinguishing between the different orders in which they could be chosen (e.g. $1, 2, 3$ is considered no different from $2, 3, 1$). Then there are $C_n^k = \frac{n(n-1)\cdots(n-k+1)}{k!}$ possible outcomes. It literally counts the number of subsets of size k for a set of size n.

C_n^k is known as the Binomial coefficient, also denoted as $\binom{n}{k}$, read as "n choose k". As it is related to the Binomial theorem, which states that

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

The following table summarizes the counting methods.

	order matters	order doesn't matter
with replacement	n^k	C^k_{n+k-1}
non-replacement	P^k_n	C^k_n

We don't explain the upper-right corner case C^k_{n+k-1} as it is not relevant for our purpose here. Feel free to figure it out yourself if you are interested.

Example 1.3. Find the probability of a "full house" in a five-card hand.

Solution:

$$P(\text{Full House}) = \frac{13C^3_4 \cdot 12C^2_4}{C^5_{52}} = 0.14\%.$$

Example 1.4 (Birthday problem)**.** Suppose there are k people. Find the probability that two of them have the same birthday.

Solution: Assuming there are 365 days in a year, ignoring leap years. If $k > 365$, the probability is 1. If $k \leq 365$,

$$P(\text{no match}) = \frac{365 \cdot 365 \cdots (365 - k + 1)}{365^k};$$

$$P(\text{match}) = \begin{cases} 50.7\% & k = 23 \\ 70.6\% & k = 30 \\ 97\% & k = 50 \\ 99.999\% & k = 100 \end{cases}.$$

Example 1.5 (Newton-Pepys problem)**.** Isaac Newton was consulted about the following problem by Samuel Pepys, who wanted the information for gambling purposes. Which of the following events has the highest probability?

A: At least one 6 appears when 6 fair dice are rolled.

B: At least two 6's appear when 12 fair dice are rolled.

C: At least three 6's appear when 18 fair dice are rolled.

1.4 Axiomatic probability

We have now seen several methods for counting outcomes in a sample space, allowing us to calculate probabilities if the naive definition applies. But the naive

definition can only take us so far, since it requires equally likely outcomes and can't handle an infinite sample space. To generalize the notion of probability, we'll use the best part about math, which is that you get to *make up your own definitions*. What this means is that we write down a short wish list of how we want probability to behave (in math, the items on the wish list are called axioms), and then we define a probability function to be something that satisfies the properties we want.

Definition 1.3. A **probability space** consists of S and P, where S is a sample space, and P is a function which takes an event $A \subseteq S$ as input and returns $P(A) \in [0, 1]$ such that

1. $P(\phi) = 0$,
2. $P(S) = 1$,
3. $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ if $A_1, A_2, \ldots, A_n$ are disjoint.

Note that this Definition does not imply any particular interpretation of probability. In fact, any function P that satisfies the axioms are valid "probabilities". Thus, the theories of probability do not depend on any particular interpretation. It is purely axiomatic. From the three axioms, we can derive any property of probabilities. The interpretation also matters, but it is more of a philosophical debate. Basically, there are two views in this regard.

- The *frequentist* view of probability is that it represents a long-run frequency over a large number of repetitions of an experiment: if we say a coin has probability $1/2$ of Heads, that means the coin would land Heads 50% of the time if we tossed it over and over and over.
- The *Bayesian* view of probability is that it represents a degree of belief about the event in question, so we can assign probabilities to hypotheses like "candidate A will win the election" or "the defendant is guilty" even if it isn't possible to repeat the same election or the same crime over and over again.

Theorem 1.1. *Probability has the following properties. For any events A and B, we have*

1. $P(A^c) = 1 - P(A)$

2. If $A \subseteq B$, then $P(A) \leq P(B)$.

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

Proof.

1. Since A and A^c are disjoint and their union is S, apply the third axiom:

$$P(S) = P(A \cup A^c) = P(A) + P(A^c);$$

By the second axiom, $P(S) = 1$. So $P(A) + P(A^c) = 1$.

2. The key is to break up the set into disjoint sets. If $A \subseteq B$, then $B = A \cup (B \cap A^c)$ where A and $B \cap A^c$ are disjoint (draw a Venn diagram for intuition). By the third axiom, we have

$$P(B) = P(A \cup (B \cap A^c)) = P(A) + P(B \cap A^c) \geq P(A).$$

3. We can write $A \cup B$ as the union of the disjoint set A and $B \cap A^c$. Then by the third axiom,

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

It suffices to show that $P(B \cap A^c) = P(B) - P(A \cap B)$. Since $B \cap A$ and $B \cap A^c$ are disjoint, we have

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

So $P(B \cap A^c) = P(B) - P(A \cap B)$ as desired. □

The last property is a very usueful formula for finding the probability of a union of events when the events are not necessarily disjoint. We have showed that for two events A and B. A natural question is to generalize it for three or more events. For three events,

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

We skip the proof. It can be easily justified by showing a Venn diagram. For the n-events case, we state it as the following theorem.

Theorem 1.2 (Inclusion-exclusion)**.** *For any events $A_1, A_2, \ldots, A_n$, it holds that*

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

This formula can be proved by induction using the axioms. Below is a famous application (known as de Montmort's problem, named after French mathematician Pierre Remond de Montmort) of the inclulsion-exclusion theorem.

Example 1.6 (Matching problem)**.** Given n cards, labeled $1, 2, ..., n$. Let A_j be the event "j-th card matches"(the j-th card is numbered as j). Find the probability of at least one match, i.e. $P(A_1 \cup A_2 \cup \cdots \cup A_n) =$?

Solution: Since all position are equally likely, $P(A_j) = \frac{1}{n}$. The probability of there being two matches is: $P(A_1 \cap A_2) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$. Similarly, the probability of there being k matches is: $P(A_1 \cap \cdots \cap A_k) = \frac{(n-k)!}{n!} = \frac{1}{n(n-1)\cdots(n-k+1)}$. Using the property of the union of events,

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_n) =& n \cdot \frac{1}{n} - \binom{n}{2} \frac{1}{n(n-1)} + \binom{n}{3} \frac{1}{n(n-1)(n-2)} - \cdots \\ =& 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots + (-1)^{n+1} \frac{1}{n!} \approx 1 - \frac{1}{e}. \end{aligned}$$

1.5 Conditional probability

Abraham Wald, the renowned statistician, was hired by the Statistical Research Group (SRG) at Columbia University to figure out how to minimize the damage to bomber aircraft. The data they had comprised aircraft returning from missions with bullet holes on their bodies. If asked which parts of the aircraft should be armored to enhance survivability, the obvious answer seemed to be to armor the damaged parts. However, Wald suggested the exact opposite—to armor the parts that were not damaged. Why? Because the observed damage was conditioned on the aircraft returning. If an aircraft had been damaged on other

parts, it likely would not have returned. Thinking conditionally completely changes the answer![1]

The probability of A **conditioned on** B is the updated probability of event A after we learn that event B has occurred. Since events contain information, the occurring of a certain event may change our believes on probabilities of other relevant events. The updated probability of event A after we learn that event B has occurred is the conditional probability of A given B.

Definition 1.4. If A and B are events with $P(B) > 0$, then the **conditional probability** of A given B is defined as

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

Exercise 1.1. Prove that conditional probabilities are probabilities. (Hint: using the three axioms.)

Theorem 1.3. *Properties of conditional probability:*

- $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$
- $P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_n|A_1 \ldots A_{n-1})$
- $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$ *(Bayes' rule)*

The last property, Bayes' rule, quantifies how to update probabilities based on new evidence. It is named after Thomas Bayes in the 18th century. It gained prominence posthumously through Richard Price's publication of Bayes' work in 1763. The rule calculates the probability of a hypothesis based on prior knowledge and new data, foundational for Bayesian statistics.

Historically, Bayes studied the problem in order to prove David Hume wrong. Hume argued that we cannot directly observe causation; instead, we infer it from patterns of events. Bayes' rule allows for a systematic way to update our beliefs about causal relationships as new evidence emerges, thereby bridging the gap between empirical observation and theoretical inference. This approach counters Hume's skepticism by providing a method for rationally assessing the likelihood of causes based on observed effects.[2]

[1] See an interesting talk by Professor Joseph Blitzstein: "The Soul of Statistics". Available on `http://www.youtube.com/watch?v=dzFf3r1yph8`

[2] See `https://faculty.som.yale.edu/jameschoi/bayes-theorem-began-as-a-defense-of-christianity`.

Theorem 1.4 (Law of total probability)**.** *Let $A_1, ..., A_n$ be a partition of the sample space S (i.e., the A_i are disjoint events and their union is S), with $P(A_i) > 0$ for all i. Then*

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

Example 1.7. Get a random 2-card hand from a standard deck. Find the probability of getting another ace conditioned on (a) having one ace, or (b) having the ace of spade.

Solution: The example shows the subtleness of conditional probabilities. The seemingly indifferent probabilities are in fact different:

$$P(\text{another ace} \mid \text{one ace}) = \frac{P(\text{both aces})}{P(\text{one ace})} = \frac{C_4^2/C_{52}^2}{1 - C_{48}^2/C_{52}^2} = \frac{1}{33};$$
$$P(\text{another ace} \mid \text{ace of spade}) = \frac{P(\text{ace of spade \& another ace})}{P(\text{ace of spade})} = \frac{C_3^1/C_{52}^2}{C_{51}^1/C_{52}^2} = \frac{1}{17}.$$

In the first case, the denominator is interpreted as "at least one ace"; whereas in the second case, it is "ace of space + another card".

Example 1.8. The pandemic afflicted roughly 1/3 of the world population. The PCR test is 98% accurate. (this means if you have been infected, the test reports positive 98% of the time.) Find the probability of being infected when a test is positive.

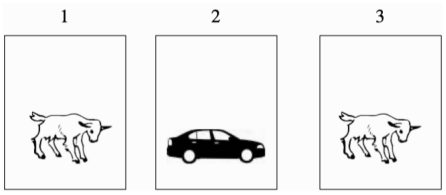
Solution: Let D: actually infected, T: test positive. The test accuracy means: $P(T|D) = 98\%$. It also means $P(T|D^C) = 2\%$. We also know that $P(D) = 1/3$. We want to find $P(D|T)$. Apply the Bayes' rule:

$$\begin{aligned} P(D|T) =& \frac{P(T|D)P(D)}{P(T)} \\ =& \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^C)P(D^C)} \\ =& \frac{0.98 \times 1/3}{0.98 \times 1/3 + 0.02 \times 2/3} \approx 96\%. \end{aligned}$$

Note that how $P(T|D)$ is different from $P(D|T)$, though confusing the conditionality is quite common in daily life. The difference is even pronounced if the disease is rare. Suppose $P(D) = 10\%$. Then $P(D|T) = 84\%$. A large difference

from the test accuracy rate 98%!

Example 1.9 (Monty Hall problem)**.** Suppose you are on Monty Hall's TV show. There are three doors. One of them has a car behind it. The other two doors have goats. Monty knows which one has the car. Monty now asks you to pick one door. You will win whatever is behind the door. After you pick one door. Monty opens another door that shows a goat. Monty then asks you if you want to switch. Is it optimal to switch?



We present two solutions to the problem. The first one is using the law of total probability. Let S: succeed assuming switch; D_j: door j has the car, $j \in 1, 2, 3$. Without loss of generality, assume the initial pick is Door 1. Monty will always open the door with a goat. By the law of total probability,

$$\begin{aligned} P(S) = & \underbrace{P(S|D_1)}_{\text{switch from initial pick}} P(D_1) + \underbrace{P(S|D_2)}_{\text{Monty opens door 3}} P(D_2) + \underbrace{P(S|D_3)}_{\text{Monty opens door 2}} P(D_3) \\ =& 0 + 1 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{2}{3}. \end{aligned}$$

The problem can also be solved using the Bayes' rule. Let D_j: door j has the car; M_j: Monty opens door j, $j \in 1, 2, 3$. Assume the initial pick is Door 1. If Monty opens door 3, the probability of winning the car assuming switching is

$$\begin{aligned} P(D_2|M_3) =& \frac{P(M_3|D_2)P(D_2)}{P(M_3)} \\ =& \frac{P(M_3|D_2)P(D_2)}{P(M_3|D_1)P(D_1) + P(M_3|D_2)P(D_2) + P(M_3|D_3)P(D_3)} \\ =& \frac{1 \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0} = \frac{2}{3}. \end{aligned}$$

Note that, if door 1 has the car, Monty will open door 2 and 3 with equal probability, thus $P(M_3|D_1) = \frac{1}{2}$. And Monty will never open the door with

the car, therefore $P(M_3|D_3) = 0$. Similarly, if Monty opens door 2, we have $P(D_3|M_2) = \frac{2}{3}$. Therefore, the optimal choice is always to switch. Intuitively, because Monty knows which door has the car, the fact that he always opens the door without the car gives additional information regarding the choice of the door.

Example 1.10 (Simpson's paradox)**.** There are two doctors, Dr. Lee and Dr. Wong, performing two types of surgeries — heart surgery (hard) and band-aid removal (easy). Dr. Lee has higher overall surgery success rate. Is Dr. Lee necessarily a better doctor than Dr. Wong?

Solution: No. Consider the following example:

	Dr. Lee			Dr. Wong		
	Heart	Band-Aid	Total	Heart	Band-Aid	Total
Success	2	81	83	70	10	80
Failure	8	9	17	20	0	20
Success rate	20%	90%	83%	78%	100%	80%

The truth is Dr. Lee has overall higher success rate because he only does easy surgeries (band-aid removal). Dr. Wong does mostly hard surgeries and thus has a lower overall success rate. Yet, he is better at each single type of surgery. To formalize the argument, let S: successful surgery; D: treated by Dr. Lee, D^c: treated by Dr. Wong; E: heart surgery, E^c: band-aid removal. Dr. Wong is better at each type of surgery,

$$P(S|D, E) < P(S|D^c, E)$$
$$P(S|D, E^c) < P(S|D^c, E^c);$$

But, Dr. Lee has a higher overall successful rate,

$$P(S|D) > P(S|D^c).$$

This is because there is a "confounder" E:

$$P(S|D) = \underbrace{P(S|D,E)}_{<P(S|D^c,E)}\underbrace{P(E|D)}_{\text{weight}} + \underbrace{P(S|D,E^c)}_{<P(S|D^c,E^c)}\underbrace{P(E^c|D)}_{\text{weight}}.$$

A **confounder** is a variable that influences with both explanatory variable and the outcome variable, which therefore "confounds" the correlation between the two. In our example, the type of surgery (E) is associated with both the doctor and the outcome. Without the confounder being controlled, it is impossible to draw valid conclusions from the statistics.

In general terms, Simpson's paradox refers to the paradox in which a trend that appears across different groups of aggregate data is the reverse of the trend that appears when the aggregate data is broken up into its components. It is one of the most common sources of statistical misuse. Here is another example.[3]

Example 1.11 (UC Berkeley gender bias)**.** One of the best-known examples of Simpson's paradox comes from a study of gender bias among graduate school admissions to University of California, Berkeley. The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance.

	Male		Female	
	Applicants	Admitted	Applicants	Admitted
Total	8,442	44%	4,321	35%

However, when taking into account the information about departments being applied to, the conclusion turns to the opposite: in most departments, the admission rate for women is higher than men. The lower overall admission rate is caused by the fact that women tended to apply to more competitive departments with lower rates of admission, whereas men tended to apply to less competitive departments with higher rates of admission.

Department	Male		Female	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
Total	2691	45%	1835	30%

[3] See https://setosa.io/simpsons for a really good illustration of the Simpson's paradox.

1.6 Independence

Definition 1.5. If event B's occurrence does not change the probability of A, then we say A and B are independent. That is to say A and B are **independent** if

$$P(A|B) = P(A) \text{ when } P(B) > 0.$$

Or more generally, A and B are **independent** if

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

(A definition including cases where A or B has zero probability.)

Theorem 1.5. *If events A and B are independent, then*

- *A and B^c are independent;*
- *A^c and B^c are independent.*

A and B are independent means they do not provide information to each other in the sense that conditional probability is not different from the unconditional probability. It is not an intuitive idea as it seems. It will become clearer when we discuss random variables in later chapters. Here we clarify some likely confusions.

Remark 1.1. Independence is not the same as disjointness.

A and B are disjoint means if A occurs, B cannot occur. But independence means A occurs has nothing to do with B.

Remark 1.2. Pairwise independence does not imply independence.

Definition 1.6. Events A, B, and C are said to be **(mutually) independent** if all of the following equations hold:

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

If the first three conditions hold, we say that A, B, and C are **pairwise independent**. Pairwise independence does not imply independence. Convince yourself with the following example.

Example 1.12. Consider two fair, independent coin tosses, and let A be the event that the first is Heads, B the event that the second is Heads, and C the event that both tosses have the same result. Show that A, B, and C are pairwise independent but not independent.

Solution: For each event, $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$. Consider the two events together, there are four possible outcomes: HH, HT, TH, TT. $P(C) = P(HH) + P(TT) = \frac{1}{2}$. Thus,

$$P(A \cap B) = P(HH) = \frac{1}{4} = P(A)P(B)$$
$$P(A \cap C) = P(HH) = \frac{1}{4} = P(A)P(C)$$
$$P(B \cap C) = P(HH) = \frac{1}{4} = P(B)P(C)$$

But A, B, C are not independent, because

$$P(A \cap B \cap C) = P(HH) = \frac{1}{4} \neq P(A)P(B)P(C).$$

Definition 1.7. For n events $A_1, A_2, \ldots, A_n$ to be **(mutually) independent**, we require any pair to satisfy $P(A_i \cap A_j) = P(A_i)P(A_j)$ (for $i \neq j$), any triplet to satisfy $P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k)$ (for i, j, k distinct), and similarly for all quadruplets, quintuplets, and so on.

Definition 1.8. Events A and B are **conditional independent** given C if

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

Remark 1.3. Conditional independence does not apply independence.

Consider an example of playing chess games. Conditioned on the strength of your opponents, the outcome of each game is reasonably independent (ignoring the psychology and fatigues of the players). But the outcomes are not unconditionally independent, because stronger player has higher chances of winning each game.

Remark 1.4. Independence does not apply conditional independence.

Consider an example of fire alarm. Suppose there are two potential causes to trigger the fire alarm: (1) there is fire; (2) someone smoking. Assume the two events are independent. But they are not conditional independent if conditioning on the alarm beeping. Because if the alarm is on, but no one smokes, we definitely know there is fire. So there they are not conditional independent.