Introduction to Probabilities and a Bit of Bandits
CMPUT 365

Csaba Szepesvári

January 12, 2025

Introduction

Probability theory provides a framework to model uncertainty and make informed decisions in the presence of randomness. In this lecture, we explore the foundations of probability, focusing on key concepts such as probability spaces, random variables, and probability distributions. Using stochastic bandits as a motivating example, we aim to build an understanding of how these tools are applied in practice, while continuously asking *why* they are useful and how they relate to real-world scenarios.

Why Use Probabilities?

Uncertainty arises when we lack complete information about a system or process. This could be due to partial knowledge, the complexity of the system, or the inherent randomness of certain phenomena. For instance, when rolling a die, we do not know in advance which number will appear. Similarly, the exact time a bus arrives or how a customer responds to a new product are unpredictable.

An important aspect of uncertainty is that even though individual events may seem "chaotic" (in an everyday sense of the word), they often exhibit predictable patterns when observed across multiple occurrences. This is the concept of *order in chaos*.[1] Consider the roll of a fair die. While the outcome of a single roll is unpredictable, we know that, over many rolls, each face will appear approximately one-sixth of the time. Similarly, shuffled cards or weather patterns, despite their randomness, often follow statistical regularities. Probabilities give us a precise mathematical framework to describe these patterns and reason about them rigorously.

[1] This phenomenon — predictable patterns emerging from seemingly chaotic individual events — is closely tied to the ideas of ergodicity and statistical regularity. In chaotic systems, while individual trajectories may be highly sensitive to initial conditions and appear unpredictable, the system often exhibits predictable behaviour in a statistical sense, such as stable averages or distributions over time or space. Ergodic systems, in particular, ensure that long-term observations of a single trajectory can reveal properties of the entire system, providing a foundation for reasoning about randomness and uncertainty. However, it is important to note, that chaos and ergodicity, while they often go hand in hand, do not imply each other.

Using probabilities to model uncertainty allows us to avoid extreme strategies like assuming the worst case or always expecting the best. By quantifying the likelihood of various outcomes, probabilities enable balanced decision-making.

Using probabilities to model uncertainty allows us to make better decisions by exploiting patterns that emerge from the underlying laws of probability. If the world is subject to the laws of probability, this creates an opportunity to exploit these patterns to our advantage. The *law of large numbers* is a cornerstone of this reasoning: as the number of observations grows, the average of the observed outcomes converges to the expected value. This statistical regularity makes it possible to reason about long-term outcomes, even if individual events are unpredictable.

As an example, consider a coin flip where the probability of heads is 0.6 and tails is 0.4. If you were to repeatedly bet on the outcome, knowing the probabilities allows you to conclude that it is in your best interest to bet on heads every time, regardless of what happened in the past. Without probabilistic reasoning, you might flip the coin a few times, see tails appear twice in a row, and mistakenly believe tails is more likely. If you keep flipping the coin, and you see a changing pattern of tails and heads, you may decide it is impossible to know what the next coin flip will be and give up and just choose to bet on heads, or tails, at random. As we know it well, this would give up on the opportunity of making good money on your betting partner, a suboptimal outcome for you! What is more, with sufficient knowledge of probability theory, you can do well even if you do not know *a priori* the bias of the coins.

Lastly, you may be suspicious that the conclusions that you derive with the help of probability theory may be brittle in the sense that even the tiniest deviations of the rules governing of how a system works from the laws of probability could give disastrous results. In other words, does the world need to obey the precise laws of probability to make probabilistic reasoning useful? Luckily, this does not appear to be the case: Conclusions derived with probabilistic reasoning are (more often than not) robust in the face of all kind of deviations from the laws of probability. However, now we are straying a bit too far from the topic of the course.

In summary, probabilities quantify uncertainty but they also do more than that; they allow us to harness the patterns that emerge from randomness, enabling balanced and informed decision-making. By understanding and leveraging these patterns, we can make better predictions, allocate resources more effectively, and ultimately achieve better outcomes in uncertain environments.

Probability Spaces

At the heart of probability theory lies the concept of a *probability space*, which provides a structured way to model randomness. A probability space consists of three components: the outcome space (Ω), the set of events ($\mathcal{A}$), and the probability measure ($\mathbb{P}$).

The outcome space, Ω, represents all possible outcomes of an experiment. For example, if we roll a die, $\Omega = \{1, 2, 3, 4, 5, 6\}$. If we measure the time until a bus arrives, $\Omega = \mathbb{R}_{\geq 0}$, the set of non-negative real numbers. Oftentimes, we treat the outcome space in a generous way; any sufficiently large set works fine. For example, for dice rolls, we can as well choose $\Omega = \mathbb{R}$, the set of reals.

The *set of events*, $\mathcal{A}$, is a collection of subsets of Ω to which we will assign probabilities (more about this soon). For finite Ω, such as a dice roll, it is often reasonable to include all subsets of Ω in $\mathcal{A}$. However, for infinite Ω, such as $\mathbb{R}$, including all subsets can lead to mathematical inconsistencies. Instead, $\mathcal{A}$ is chosen to be closed under operations like unions, intersections, and complements, ensuring that it forms a mathematically well-behaved collection.[2] The point is that is that we want to choose $\mathcal{A}$ large enough so that we will have an answer for any "reasonable" question concerning the system that we model.

That $\mathcal{A}$ is closed under the above mentioned operations is one way of ensuring this, after we include some basic subset of Ω in $\mathcal{A}$. What these sets are depends on the specific situation we are in, though in every case, we want Ω to be included in $\mathcal{A}$.

Two special cases are worth mentioning: When Ω is finite, and if we include, all the singletons (i.e., sets of the form $\{\omega\}$ where $\omega \in \Omega$) then it is easy to see that the fact that $\mathcal{A}$ needs to be closed under the above-mentioned operations will give that $\mathcal{A} = 2^{\Omega}$. When $\Omega = \mathbb{R}$, the set of reals, we will find it useful if the intervals are included in $\mathcal{A}$. This allows us to talk about the probability that at an outcome is between two numbers.

The *probability measure* (probability distribution, probability distribution function, or just probability function), $\mathbb{P}$, assigns numbers in $[0, 1]$, which we call probabilities, to events in $\mathcal{A}$. So, formally, $\mathbb{P}$ is a function from $\mathcal{A}$ to $[0, 1]$:

$$\mathbb{P} : \mathcal{A} \to [0, 1] .$$

If the identity of $\mathcal{A}$ is clear from the context but we want to name Ω, we often just say, $\mathbb{P}$ is a probability measure (or probability distribution) over Ω.

For finite Ω, when $\mathcal{A} = 2^{\Omega}$, the set of all subsets of Ω, we will see that $\mathbb{P}$ is determined entirely by the probabilities of individual outcomes. For example, if a die is fair, we assign $\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \cdots = \mathbb{P}(\{6\}) = \frac{1}{6}$. However, for continuous Ω, singletons (e.g., $\{x\}$) often have zero probability, as in the case of the normal distribution. In such cases, probabilities are assigned to subsets like intervals, providing meaningful information about ranges of outcomes.

We call a probability distribution $\mathbb{P}$ *discrete*, if there are countably many points $x_1, x_2, \ldots$ of Ω such that for any event $A \in \mathcal{A}$, $\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{I}\{x_i \in A\}\mathbb{P}(\{x_i\})$, where $\mathbb{I}\{\phi\} = 1$ if the logical expression ϕ evaluates to true, and $\mathbb{I}\{\phi\} = 0$ otherwise. In words, the probability of event A under a discrete distribution $\mathbb{P}$ is is the total probability of those outcomes $x_1, x_2, \ldots$ that belong to A. It follows among other things that the probability of any event A such that none of $x_1, x_2, \ldots$ belongs to A, is zero. In other words, under a discrete probability distributions, we can be sure that one of the outcomes of $x_1, x_2, \ldots$ will happen with probability one. (Formally, $\mathbb{P}(\{x_1, x_2, \ldots\}) = 1$.) A die roll has a discrete distribution: Choose Ω to be any set including $\{1, 2, \ldots, 6\}$. Choose $x_i = i$ for $i = 1, 2, \ldots, 6$.

The opposite of discrete probability distributions is a *continuous distribution*. We will not give a general definition for this, but here, you should think about, for example, distributions over the reals, such as the normal distribution. In fact, a continuous probability distribution $\mathbb{P}$ *over the reals* is one such that there exist some function $f : \mathbb{R} \to [0, \infty)$ (called the *density*

[2]Specifically, the standard requirement is that $\mathcal{A}$ is closed under finite intersections, countable unions, and complements. The reason for demanding closedness under countable unions (instead of finite unions) is somewhat subtle and in fact some versions of probability theory do not demand this.

function) such that for any interval $[a, b] \subset \mathbb{R}$, $\mathbb{P}([a, b]) = \int_a^b f(x)dx$. Of course, this also means that these integrals need to be well-defined, etc. We omit the technical details.

The advantage of assigning probabilities to subsets as opposed to not doing that (as is often done in naive approaches to probabilities) is that it allows us to model probability distributions that are neither discrete, nor continuous. Consider for example the case when a reward is generated by the following process: Flip a coin. If the result is heads, the reward is 0. Otherwise, the reward is normally distributed with mean and variance both set to one. *Hybrid distributions* like this arise all the time in reinforcement learning. Assigning probabilities to subsets ensures that we can handle all these cases consistently.

Consistency Properties of Probability Functions

In a probability triplet $(\Omega, \mathcal{A}, \mathbb{P})$, by definition, $\mathbb{P} : \mathcal{A} \to [0, 1]$. However, not all functions of this type are accepted as probability functions. If, say, $\Omega = \mathbb{R}$, $\mathcal{A}$ contains intervals then knowing $\mathbb{P}([1, 2])$, $\mathbb{P}([1, \infty))$ cannot be chosen arbitrarily. Intuitively, the meaning of $\mathbb{P}(A)$ is that the probability that the outcome of our random die rolls (more generally, called, "experiments"[3]) lands in A is the specific number $\mathbb{P}(A)$. We then expect that if $A \subset B$, $A, B \in \mathcal{A}$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$: Probabilities increase if we allow more "favourable outcomes". What is more, we expect that if A and B are disjoint (have no common elements), then the probability of an outcome belonging to $A \cup B$ should be the sum of the probability of the outcome belonging to A, or the probability of the outcome belonging to B: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. This is known as the*additivity* of the probability function. In fact, we require that $\mathbb{P}$ satisfies the stronger requirement that for any sequence of pairwise disjoint events $A_1, A_2, \ldots,$

$$\mathbb{P}(A_1 \cup A_2 \cup \ldots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots .$$

Above, the set $A_1 \cup A_2 \cup \ldots$ is the set of $\omega \in \Omega$, such that ω is in A_i for some $i \in \{1, 2, \ldots\}$. We also denote this set by $\cup_{i=1}^n A_i$ (or, in short, $\cup_i A_i$, when the range of i is clear from the context). Now, the sum on the right-hand side is an infinite sum, which we also write as $\sum_{i=1}^\infty \mathbb{P}(A_i)$. As usual, this is defined as the limit of the partial sums: $\lim_{n\to\infty} \sum_{i=1}^n \mathbb{P}(A_i)$. Thus, the definition presupposes that this sum converges.[4] (For the above constraint to make sense, we need that for any disjoint sequence of events $A_1, A_2, \ldots$ from $\mathcal{A}$, $\cup_i A_i \in \mathcal{A}$ must also hold. Otherwise, $\mathbb{P}(\cup_i A_i)$ on the left-hand side above would not make sense. When $\mathcal{A}$ is closed to complements and finite intersections, it is not hard to see, that the said constraint on $\mathcal{A}$, namely, that it must be closed for taking countable unions for disjoint sequences of events, is equivalent to the same property holding for arbitrary sequences of events, regardless of whether they are disjoint or not. This is the property that probability theory text usually state as the requirement.)

[3]The literature is in fact using the word "experiments" to describe the die rolls. Here, experiment is just to meant in the sense of trying things that results in uncertain things. The origin is of course that probability theory is extremely useful and is very commonly used to analyze the results of scientific experiments. In our case, experiments sounds a bit out of place, though perhaps it helps if we think that the experiments are run by 'mother nature'.

[4]You may recall that an increase sequence of real numbers always converges, where we allow convergence to the value ∞. Since the probability function returns nonnegative numbers, clearly, the partial sums form an increasing sequence of numbers. Thus, the value of the sum takes on a well-defined value in $\mathbb{R} \cup \{+\infty\}$.

A second consistency rule that probability functions need to satisfy is that the probability assigned to Ω must be one:

$$\mathbb{P}(\Omega) = 1\,.$$

(This is where we need that $\Omega \in \mathcal{A}$.)

A little thinking then gives that these rules imply that if $A, B \in \mathcal{A}$ and $A \subset B$ then $\mathbb{P}(A) \subset \mathbb{P}(B)$ indeed. We also get that $\mathbb{P}(\emptyset) = 0$ (note that $\emptyset \in \mathcal{A}$ follows from $\Omega \in \mathcal{A}$ since $\mathcal{A}$ is assumed to be closed under complementation.)

Random Variables

The central questions in probability theory revolve around figuring out probabilities of various events in complicated scenarios. In reinforcement learning (and bandits), where we are users of probability theory, we are interested in the probability of our agent succeeding in achieving its goal (say, receiving a reward of one, after interacting with its environment for a number of time steps). When the robot interacts with its environments, the outcomes of the actions it takes are often uncertain – subject to the laws of probability. One way of thinking of this is that "nature" rolls a die (with many sides, maybe even uncountable many sides, so think about this abstractly), and the outcome of a die roll will influence what happens to the robot and its environment as a result of the robot taking some action. As the robot takes many steps, this involves many die rolls. In addition, the robot itself may also take random actions occasionally (this makes the most sense if the robot is playing a game like rock-paper-scissors, but we will also encounter randomizing algorithms to help the robot collect information). There are many die rolls to keep track off!

Furthermore, which die are rolled, when and how, may depend on the situation. Think of a card game: Initially, you shuffle the deck (that is the roll of a die with 52! sides: all permutations of the 52 cards are equally likely). Next you deal some cards to the players, the game starts, and the players will keep drawing cards from the deck (in addition to doing other things). Now, maybe in the card game, some players can choose an action to reshuffle the deck and the rules are such that this can influence who wins. Such a reshuffling action is again a die roll, but with a die where the number of sides depends on how many cards are left. Now the identity of the die depends on what happened *earlier* in the game. This is getting complicated!

For conceptual clarity, a better approach (and the standard approach since over a century), is to just think about that all these "randomizing action" (die rolls) happen upfront. Use all kind of dies, use infinitely many of them, as you wish. (If we want to model the outcomes of two die rolls (with standard dice), we can use $\Omega = \{1, \dots, 6\} \times \{1, \dots, 6\}$. Now, every element of Ω is in the form of a pair.) With this way, during the game, we will just use the outcomes of these die rolls, instead of actually rolling the die.

Now, conceptually, what this means is that we will have functions mapping Ω to whatever space we need. You need the result of the first die roll when Ω models the result of two die rolls? Then, as discussed $\Omega = \{1, \dots, 6\} \times \{1, \dots, 6\}$ and given (ω_1, ω_2), the function can just return ω_1. Denoting this function by X (for historical reasons we use capital letters to denote functions whose domain is Ω), we have $X((\omega_1, \omega_2)) = \omega_1$.

With this, our approach to probabilistic modeling is as follows: Pick a large enough probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Then choose whatever functions $X_1, X_2, X_3, \ldots$ with domain Ω allow you to describe how things work. Note that these functions can (and often will) be built on the top of each other. For example, X_3 can be defined using X_1: If $X_1 = 1$ then $X_3 = 3$ otherwise $X_2 = X_2$. This approach will let us define the ultimately object of interest, such as the total reward collected by our robot. Call this R, thus, R itself is a map from Ω, in this case to reals. Then, we start working out probabilities associated with R, such as, the probability that R exceeds a threshold. How?

Well, collect all the outcomes ω in Ω, such that $R(\omega) \geq \theta$ (here, θ is our threshold, a real number). This gives us a subset of Ω, call this A. If we did everything well, A is an event: $A \in \mathcal{A}$. Thus, we can look up $\mathbb{P}(A)$. This is the probability that our (imagined) die rolled in such a way that they led to the total reward exceeding θ.

To recap, the fundamental idea in probability theory is to trace back all randomness to a fixed probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Once this randomness is fixed, everything else becomes deterministic.

The functions mentioned above that map outcomes in Ω to some value in some other space are called *random variables* (in some text, *random elements*, as random variables are reserved to random elements where the image space is the set of reals). There is in fact again some technical subtlety that not every function here is allowed, but we will ignore this for now. (The reason for not allowing some functions has to do with that $\mathcal{A}$ may not have all subset of Ω.)

Note that in a mathematical sense, there is no randomness in the random variables themselves. These are just deterministic functions. They also do not change, or "vary". The name reflects the thinking that we think of R as a summary of how the random outcomes ω give rise to specific values. As ω changes, $R(\omega)$ changes. Hence, the value of R is "variable". And we call it random, because we think of an individual ω as the result of a random experiment.

By focusing all randomness on the probability space, this framework simplifies reasoning and analysis; essentially reducing all questions related to probabilities to standard mathematical questions about functions; the only special aspect being that unlike in other parts of mathematics, here we have a probability function $\mathbb{P}$ that allows us to quantify the probability of events of interest. In "probabilistic thinking" we think of what will happen for each of the particular outcome; and this is where "random variables" help to give names of certain consequences of the individual outcomes.

Applying Probability to Stochastic Bandits

The formal definition of a *stochastic bandit problem* is as follows. We are given:

- A finite set of $k > 0$ actions (arms), indexed by $i \in [k] := \{1, 2, \ldots, k\}$.
- Each arm $i \in [k]$ is associated with some probability distribution $\mathcal{P}_i$ over the reals.
- Pulling arm i gives rise to a random reward $R \sim \mathcal{P}_i$.

The definition uses the language of probability theory. In what follows, we go through the parts of this definition to uncover their precise mathematical meaning.

The Probability Distributions $\mathcal{P}_i$

Consider first the sentence:

> "Each arm $i \in [k]$ is associated with some probability distribution $\mathcal{P}_i$ over the reals."

From this we learn that that $\mathcal{P}_i$ is a probability distribution over the reals. Now, thinking back about what was said about probability distributions, we deduce that $\mathcal{P}_i$ must be a map from $\mathcal{A}$ for some $\mathcal{A} \subset 2^{\mathbb{R}}$ to $[0,1]$ that satisfies the requirements we imposed on probability functions. As discussed earlier, in the case of the reals, we choose $\mathcal{A}$ to be a sufficiently rich set so that it contains at least all the intervals.[5] This set, while we won't be too specific about it, is denoted by $\mathcal{A}_{\mathbb{R}}$.[6] Thus,

$$\mathcal{P}_i : \mathcal{A}_{\mathbb{R}} \to [0,1]\,.$$

Thus, we now know what it means that $\mathcal{P}_i$ is a probability distribution over the reals.

Random rewards

Consider now the second sentence:

> "Pulling arm i gives rise to a random reward $R \sim \mathcal{P}_i$."

We already know what $\mathcal{P}_i$ is. How about R? And what is the meaning of $\sim$? And what is the meaning that the reward is random? What is random here?

The precise meaning here is simply as follows: We are given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, which is the source of all the randomness that we will ever encounter when discussing our example. Then, R is a function from Ω to the reals, or, with our previous terminology, it is a random variable. Whenever we want any random quantity, the quantity will be associated with a function mapping Ω to whatever values the quantity can take. Think of R translating any possible outcome $\omega \in \Omega$ (which is the source of all the randomness) to a real, the reward. Formally,

$$R : \Omega \to \mathbb{R}\,.$$

Thus, for *any* $\omega \in \Omega$, $R(\omega)$ is some real value. Is R itself "random"? No, it is a function from Ω. The adjective "random" just means that R maps outcomes (which we can think of

[5] Think about this: if we only demand that closed intervals are contained in $\mathcal{A}$, given that $\mathcal{A}$ needs to satisfy the closedness properties with respect to elementary set theoretic operations, does this mean that $\mathcal{A}$ includes open intervals? Half open, half closed intervals?

[6] There are two choices here, which I write down only for the sake of completeness: The firs optiont is to choose $\mathcal{A}_{\mathbb{R}}$ to be the set of the so-called Borel sets. These are the sets that can be obtained from intervals with finite intersections, complementation, and countable union. The alternative is to choose a slightly bigger set of Lebesgue sets. Normally, we should just go with the Borel sets.

as the result of "random experiments", our "die rolls") to some value. Again, no randomness in R. It is perfectly deterministic. In fact, this is what simplifies all the calculations that we want to do with probability spaces.

Now, it remains to interpret the meaning of $\sim$. In probability texts, $\sim$ is a relation that connects a random variable (like R) and a probability distribution. The random variable has to be on the left-hand side of $\sim$, while the probability distribution needs to be on the right-hand side of $\sim$. Above, the probability distribution is $\mathcal{P}_i$. Moreover, the probability distribution needs to be over the random variable's image space ($\mathbb{R}$ in our case), or $\sim$ would not be used properly. In our case, both are $\mathbb{R}$, hence, $\sim$ is not misused.

Finally, the meaning of $R \sim \mathcal{P}_i$ is that for any $A \in \mathcal{A}_{\mathbb{R}}$,

$$\mathbb{P}(\{\omega \in \Omega : R(\omega) \in A\}) = \mathcal{P}_i(A)\,.$$

In words, for any event A in $\mathcal{A}_{\mathbb{R}}$ (e.g., A could be an interval, such as $[a, b]$), the probability that R takes values in A is given by $\mathcal{P}_i(A)$. Note how all functions involved ($\mathbb{P}$ and $\mathcal{P}_i$) get arguments which they can indeed accept. Writing math is like writing in a typed language: Each function takes some arguments of certain kinds (types) and we have to make sure that the types match.

Since $\mathbb{P}(\{\omega \in \Omega : R(\omega) \in A\})$ takes too much space, a widespread convention is to just write $R \in A$. In the context of probability theory, $R \in A$ (also, $\{R \in A\}$), means exactly just the set $\{\omega \in \Omega : R(\omega) \in A\}$. Moreover, this is generalized to expressions that involve any logical expression that involve any number of random variables. For example, if R_1, R_2 denote random rewards in a bandit problem, then we can write $\mathbb{P}(R_1 > 1, R_2 > 2)$, or even $\mathbb{P}(R_1 + R_2 > 1)$. Note that in the latter expression, $R_1 + R_2$ is an "anonymous function" that gives the value $R_1(\omega) + R_2(\omega)$ for every $\omega \in \Omega$. With this shorthand notation, the meaning of the sentence at the beginning of this section is thus that for any $A \in \mathcal{A}_{\mathbb{R}}$,

$$\mathbb{P}(R \in A) = \mathcal{P}_i(A)\,.$$

Now, what if we want to talk about the distribution of a random variable that is not taking values over the reals, but some other set? For example, when talking about cards in a card game, maybe we have a special set S that lists the 52 cards. Then, if a player gets 4 cards after the deck is shuffled, the cards of the player would be represented by a map $C : \Omega \to S^4$. This is a finite set, and thus, if we want to be able to ask questions like "what is the probability that the player got a Heart Ace?" we better choose as the event space associated with S^4 all subsets of this set. Denote this set by $\mathcal{A}$. Thus, $\mathcal{A} = 2^{S^4}$. Let $\mathcal{P}$ be a probability distribution over $(S^4, \mathcal{A})$ (i.e., $\mathcal{P} : \mathcal{A} \to [0, 1]$). Note that in this and other cases people often would just say that $\mathcal{P}$ is a probability distribution over S^4, because, it is understood that the event space is the power set of S^4 because S^4 is finite. Here, we use a more precise language just to make it clear that the sets considered always come with an event space. Now, the meaning of $C \sim \mathcal{P}$ is that for any $A \in \mathcal{A}$ (with $\mathcal{A} = 2^{S^4}$),

$$\mathbb{P}(C \in A) = \mathcal{P}(A)\,.$$

With full generality, if X is a map from $(\Omega, \mathcal{A}, \mathbb{P})$ to $(\Omega', \mathcal{A}', \mathbb{P}')$, then we write that $X \sim \mathbb{P}'$ when for any $A' \in \mathcal{A}'$, $\mathbb{P}(X \in A') = \mathbb{P}'(A')$. Now, we may notice that this implies that

the sets $\{X \in A'\}$ where $A' \in \mathcal{A}'$ must be in $\mathcal{A}$. Otherwise, $\mathbb{P}(X \in A')$ would not be even defined! So as to avoid this trouble, we thus always assume that any random variable must have this property. Notice that when we do this, we implicitly use that Ω and Ω' come with their associated sets of events and whether X is a random variable or not depends on the choice of these. Note again that this nothing more than being careful about definitions – something that a computer science student practicing programming should be very well aware of.

Math is Like Programming

Staying with the previous metaphor, in mathematics, as in programming, syntax and types are the foundation. Every object has a specific type: $\mathbb{P}$ is a measure on events in $\mathcal{A}$, while R is a random variable mapping outcomes to $\mathbb{R}$ (and both the domain and the image space needs to be associated with appropriate event spaces). Simplified notation, such as $R \sim \mathcal{P}_i$ or $\mathbb{P}(R \in [a, b])$ is powerful, but is a double-edged sword: It is easy to get lost with such a succinct notation. This, when in doubt about the meaning of some notation, go back to the definitions. While some notation may look cryptic at a first sight, once the foundations are cleared up, the meaning of all the expressions should be crystal clear.

What Did We Learn?

One important lesson is that probability theory centralizes randomness into the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, simplifying the study of uncertainty. Then, everything that requires "randomness", will just refer to the outcomes in Ω; i.e., be a random variable. A solid grasp of these foundations is crucial for working with probabilistic models and analyzing their implications.