An Axiomatic Introduction to Formal Probability Theory

Ishaan Patkar

Probability is a wonderfully intuitive and applicable field of mathematics. Its importance to everday life means that in modern society, some understanding of probability is necessary.

However, there are many flaws within the intuitive notion of probability that everyone is familiar with. The idea of probability commonly taught is the following: given a set of equally likely outcomes, the probability of a desired event A occuring is given by:

$$P(A) = \frac{\text{\# of desired outcomes}}{\text{\# of total outcomes}}.$$

For instance, consider the roll of a typical six-sided die. There are 6 possible outcomes, so the probability of rolling an even number is $\frac{1}{2}$, since there are 3 even numbers to be achieved.

This is the notion of probability that most people are familiar with. But in many cases, this concept of probability falls apart when considering events that are not equally likely. For instance, consider a loaded die. The outcomes are the same, and so by the previous rule, the probability of rolling an even number should be the same. But a loaded die must have different probabilities!

Another form probability takes is geometric. In general, given a point is selected uniformly in a figure B, the probability of selecting a point in figure A is

$$P(A) = \frac{[A]}{[B]}$$

where $[A]$ and $[B]$ are the areas of A and B, respectively.

Consider, for instance, a unit square $ABCD$. Suppose a point is randomly selected from the interior: what is the probability that the point is closer to A than to C? Intuitively, we can divide the square into two halves along BD, with all the points in one half being closer to A than C, and likewise for the other half. Then the probability of a point being closer to A than C is

$$\frac{[ABC]}{[ABCD]}$$

where $[ABC]$ is the area of $\triangle ABC$, and $[ABCD]$ is the area of $\triangle ABCD$. Computing the areas, we have a probability of $\frac{1}{2}$.

But one might notice that the probability that the point is closer to C than A is also equal to $\frac{1}{2}$, which means that the probability of the point being equidistant from A and C is 0—it is impossible. But it is clearly possible to select a point on the line segment BD, and equidistant from A and C.

Not only that, but geometric probability has the same flaw as we saw earlier with the die: we cannot apply the principles of geometric probability when we are not selecting a point uniformly.

The various forms probability can take, and the associated flaws, have led to a more unified theory behind probability.

1 Measures

The first step in studying probability is to ignore *what the probabilities actually are.* Instead, consider every application of probability to be an experiment. This hypothetical experiment is what we call a probability space. Simply put, a probability space is the set of all outcomes of the experiment, the set of events in the experiment, where events are sets of outcomes, and a function that assigns probabilities to the events.

This function is especially important because it has to assign a finite probability to finite sets, countably infinite sets, and uncountably infinite sets. In essence, this function must measure the probabilities of these sets.

Fortunately, there is a field of mathematics devoted to the study of the ways such sets can be measured, in ways that are consistent with boolean operations.

1.1 Measure Spaces

The general idea of a measure is a function that assigns real values to sets that is a measure of the magnitude of a set. The most typical of measures are length, area, and volume, which measure the relative sizes of different sets in $\mathbb{R}$, $\mathbb{R}^2$, and $\mathbb{R}^3$, respectively. The intuitive notion of measures also respects set-theoretic operations: for instance, if we take two disjoint sets in either of the three spaces, the measure of the union of the sets equals the sum of the measures of the sets (for instance, the area of a combined triangle and square is equal to the sum of the areas of the triangle and square).

Thus, the definition of the general measure also has to be consistent with the set-theoretic operations. This means that the space the measure is defined on must be closed under these operations. We call this space a Boolean algebra:

Definition 1 (Boolean algebra)**.** Let X be a set. Then $\mathcal{B}$ is a *Boolean algebra* if it is a set of subsets of X such that:

(i) (Empty Set) $\emptyset \in \mathcal{B}$.

(ii) (Complement) if $E \in \mathcal{B}$, $X \backslash E \in \mathcal{B}$.

(iii) (Finite Unions) if $E, F \in \mathcal{B}$ then $E \cup F \in \mathcal{B}$.

The Boolean algebra is the typical space in which set-theoretic operations take place. Closure under intersection, difference, and symmetric difference all follow from closure under complements and finite unions.

Now, we can define a measure on this Boolean algebra:

Definition 2 (Finitely Additive Measure)**.** Let X be a set, and $\mathcal{B}$ be a Boolean algebra on it. Then $\mu : \mathcal{B} \to [0, +\infty]$ is said to be a *finitely additive measure* if:

(i) (Nondegenerate) $\mu(\emptyset) = 0$

(ii) (Finite Additivity) if E and F are disjoint elements of $\mathcal{B}$, then $\mu(E \cup F) = \mu(E) + \mu(F)$.

A finitely additive measure also has the following properties:

Proposition 1. If $\mathcal{B}$ is a Boolean algebra on a set X, and μ is a finitely additive measure on $\mathcal{B}$, then:

(i) (Monotonicity) If E and F are elements of $\mathcal{B}$, and $E \subseteq F$, then $\mu(E) \leq \mu(F)$.

(ii) (Finite Additivity) If $E_1, E_2, \ldots E_n$ are pairwise disjoint elements of $\mathcal{B}$, then $\mu(E_1 \cup E_2 \cup \cdots \cup E_n) = \mu(E_1) + \mu(E_2) + \cdots + \mu(E_n)$.

(iii) (Principle of Inclusion-Exclusion for Two Sets) If $E, F \in \mathcal{B}$, then $\mu(E \cup F) = \mu(E) + \mu(F) - \mu(E \cap F)$.

(iv) (Finite Subadditivity) If $E_1, E_2, \ldots, E_n \in \mathcal{B}$ then $\mu(E_1 \cup E_2 \cup \cdots \cup E_n) \leq \mu(E_1) + \mu(E_2) + \cdots + \mu(E_n)$.

Intuitively, this definition allows us to combine different sets together while adding their measures. However, in order to deal with infinite sets, we may often require limits and integrals. For instance, regular polygonal approximations of a circle are one way that the area of a circle can be computed (in fact, this is Archimedes' original proof technique). Thus, a strengthening of the Boolean algebra and finitely additive measure is necessary in order to construct a more applicable measure.

Definition 3 (σ-algebra)**.** A *σ-algebra* $\mathcal{B}$ on any set X is a set of subsets of X such that:

(i) (Empty Set) $\emptyset \in \mathcal{B}$.

(ii) (Complements) If $E \in \mathcal{B}$, then $X \backslash E \in \mathcal{B}$.

(iii) (Countable Additivity) If $E_1, E_2, \ldots \in \mathcal{B}$, then $\bigcup_{n=1}^{\infty} E_n \in \mathcal{B}$.

We say that $(X, \mathcal{B})$ is a *measurable space.*

In the same sense, we extend the definition of finitely additive measures:

Definition 4 (Measure)**.** If $(X, \mathcal{B})$ is a measurable space, then $\mu : \mathcal{B} \to [0, +\infty]$ is said to be a *countably additive measure* or simply a *measure* on $\mathcal{B}$ if:

(i) (Nondegenerate) $\mu(\emptyset) = 0$.

(ii) (Countable Additivity) If $E_1, E_2, \ldots$ are pairwise disjoint elements of $\mathcal{B}$, then:

$$\mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mu(E_n).$$

We say that $(X, \mathcal{B}, \mu)$ is a *measure space* if $(X, \mathcal{B})$ is a measurable space, and μ is a measure on $\mathcal{B}$.

Note that countable unions and finite additivity implies finite unions and finite additivity, respectively, since we can assume $E_n = \emptyset$ for $n \geq k$ for some integer k. Thus, a measure inherits all the properties of finitely additive measures.

However, measures have a few additional properties as well:

Proposition 2. If $(X, \mathcal{B}, \mu)$ is a measure space, then:

(i) (Countable Subadditivity) If $E_1, E_2, \ldots \in \mathcal{B}$, then $\mu(E_1 \cup E_2 \cup \cdots) \leq \mu(E_1) + \mu(E_2) + \cdots$.

(ii) (Upwards Monotone Convergence) If $E_1 \subseteq E_2 \subseteq \cdots$ then

$$\mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} \mu(E_n).$$

(iii) (Downwards Monotone Convergence) If $E_1 \supseteq E_2 \supseteq \cdots$ then

$$\mu\left(\bigcap_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} \mu(E_n).$$

Having formally established the idea of a measure, one might wonder why we choose to use an arbitrary σ-algebra instead of the largest possible σ-algebra, the power set of X. The reason is the Banach-Tarski paradox—a decomposition of the sphere into several countable or finite parts that allows one to construct two copies of the sphere, each identical to the original. This breaks the axiom of countable additivity, which means that for a measure to be valid, the sets that make up the decomposed sphere cannot be part of the domain of the measure.

2 Probability Measures

Having discussed measure spaces and a few of their basic properties, we are now ready to formally define probability spaces.

Definition 5. A *probability space* is a measure space $(\Omega, \mathcal{F}, P)$ such that $P(\Omega) = 1$. We call Ω the *sample space*, $\mathcal{F}$ the *event space*, and P the *probability measure.*

Note that as $\emptyset \in \mathcal{F}$, $\Omega \backslash \emptyset = \Omega \in \mathcal{F}$, so this definition simply specifies that the probability of the sample space is 1, or 100%.

Now, consider the example of the unit square. With a formal definition of a probability measure, the apparent contradiction no longer holds. Line segments in the square are degenerate figures and thus have area 0. Thus, the probability that a point is chosen from the line segment is also 0, using the definition of probability in this space (which is the ratio of areas). Then apparent contradiction can be stated as follows. Let L be the set of line segments from $(0, r)$ to $(1, r)$, where $r \in [0, 1]$. Then, if S is the set of points in the square:

$$S = \bigcup_{l \in L} l.$$

The lines in L are pairwise disjoint, so:

$$P(S) = P\left(\bigcup_{l \in L} l\right) = \sum_{l \in L} P(l) = \sum_{l \in L} 0 = 0.$$

However, notice that L is an uncountable set, and so:

$$P\left(\bigcup_{l \in L} l\right) = \sum_{l \in L} P(l)$$

is not necessarily valid because the axiom only holds for countable sets.

In fact, since the union of a countable or finite number of countable or finite sets is always countable or finite, the probabilities of uncountable sets cannot be defined as sums of the probabilities of countable or finite sets. This leads to a bifurcation of probability spaces: probability spaces with a countable or finite sample space are called *discrete probability spaces*, and probability spaces with an uncountable sample space are called *continuous probability spaces*.

With discrete probability spaces, any event $E \in \mathcal{B}$ is countable or finite, so:

$$P(E) = \sum_{\omega \in E} P(\{\omega\}).$$

However, a similar method to compute probability of events requires the concept of a random variable first.

3 Random Variables

If we go back to the comparison between a probability space and an experiment, a random variable is a way we can measure the outcome of the experiment. For instance, if we roll 2 times, a random variable might be the sum of numbers. Formally:

Definition 6. A random variable X on a probability space $(\Omega, \mathcal{F}, P)$ is a function with domain Ω. We say that a random variable is *discrete* if E is finite or countable, and *continuous* if E is uncountable[1].

So for instance, in the example of flipping a coin 3 times, the probability space would be:

$$\Omega = \{1, 2, 3, 4, 5, 6\}^2.$$

Then the sum of the numbers would be the random variable $X : \Omega \to \mathbb{N}$ such that:

$$X(\omega_1, \omega_2) = \omega_1 + \omega_2.$$

If we wanted the probability that the sum is 7, we might write:

$$P(\{\omega \in \Omega \mid X(\omega) = 7\}).$$

For shorthand, we usually simplify the set-builder expression to just the condition, and use X instead of $X(\omega)$:

$$P(X = 7).$$

[1] Typically, E is called the range of X. However, range means something different in probability, so that usage is uncommon.

3.1 Distributions of Random Variables

Formally, the distribution of a random variable is the following:

Definition 7. The *distribution*, or *probability distribution function* for a random variable $X : \Omega \to E$ on probability space $(\Omega, \mathcal{F}, P)$ is a function $f : E \to [0, 1]$ such that, for all $x \in E$:

$$f(x) = P(X = x).$$

Distribution functions are particularly useful for discrete random variables as the following identity holds:

$$P(X \in S) = \sum_{s \in S} P(X = s).$$

But when a random variable is continuous, the identity does not hold due to the countable additivity axiom. However, when a random variable takes on values in the real numbers; that is, if a random variable is real-valued, we can use the powerful concept of a *cumulative distribution function*:

Definition 8. The *cumulative distribution function*, or *CDF* for a real-valued random variable $X : \Omega \to \mathbb{R}$ on a probability space $(\Omega, \mathcal{F}, P)$ is a function $F_X : \mathbb{R} \to [0, 1]$ such that:

$$F_X(x) = P(X \leq x).$$

The power of the CDF is not only that it is applicable for continuous real-valued random variables, but discrete ones as well. In addition, it satisfies the following properties:

Proposition 3. If X is a real-valued random variable, and $F_X(x)$ its CDF, then:

a) $F_X(x)$ is monotonously increasing.

b) $\lim_{x \to -\infty} F_X(x) = 0$.

c) $\lim_{x \to +\infty} F_X(x) = 1$.

d) $F_X(x)$ is right continuous (i.e. $\lim_{x \to a^+} F_X(x) = F_X(a)$).

All four results are left as exercises for the reader (they are corollaries of the monotonicity results for measures).

In some cases, the CDF is differentiable:

Definition 9. If X is a real-valued random variable with CDF $F_X(x)$, and F_X is differentiable, then $f(x) = F'_X(x)$ is the *probability density function*, or *PDF*.

The following proposition thus allows us to compute the CDF given the PDF:

Proposition 4. If X is a real-valued random variable on a probability space $(\Omega, \mathcal{F}, P)$, and X has PDF $f(x)$, then:

$$P(a < X \leq b) = \int_a^b f(x)dx.$$

In addition:

$$F_X(x) = \int_{-\infty}^x f(t)dt.$$

Proof. By definition, $f(x) = F'_X(x)$. By the Fundamental Theorem of Calculus:

$$\int_a^b f(x)dx = F_X(a) - F_X(b) = P(X \leq a) - P(X \leq b) = P(b < X \leq a).$$

Now note that:

$$\int_{-\infty}^x f(t)dt = \lim_{y \to -\infty} P(y < X \leq x).$$

In order to compute the RHS, notice that:

$$P(y < X \leq x) = P(X \leq x) - P(y < X).$$

Thus,

$$\lim_{y \to -\infty} (y < X \leq x) = \lim_{y \to -\infty} (P(X \leq x) - P(X \leq y)) = P(X \leq x) - \lim_{y \to -\infty} P(X \leq y) = P(X \leq x)$$

using the fact that the CDF converges to 0 as x approaches $-\infty$. □

4 Conclusion

One might conclude, after reading this axiomatic formulation of probability theory, that probability theory is simply the study of measures with maximum value 1. While it is true that measure theory and probability theory are closely related, the results of probability are far more directly applicable and varied that the actual axiomatic formulation is not of particular interest most non-mathematicians. But it is my hope that the concepts involved will aid the reader to better understand how to think about probability.