Events and Probability

Piyush Patil

January 27, 2017

In this chapter, we'll introduce the basics of probability in a ground up, semi-rigorous manner that references our intuitive understanding of probability and chance.

1 Axioms of Probability

Let's start with a mathematical treatment of basic probability. Any probabilistic statement implicitly refers to some underlying probability space, which provides the basis for what probability actually is. First, the definition of a probability function, followed by the definition of a probability space. Note that we use $\mathcal{P}(\cdot)$ to denote the power set (ie set of subsets) of a set.

(Definition) Probability space: *A **probability space** is composed of*

1. *A **sample space** Ω, the set of all possible outcomes.*
2. *A set of allowable **events**, $F \subseteq \mathcal{P}(\Omega)$.*
3. *A **probability function** (defined below) over F.*

(Definition) Probability function: *A **probability function** $\Pr : F \to [0, 1]$, where $F \subseteq \mathcal{P}(\Omega)$ for some set Ω, such that $Pr(\Omega) = 1$ and for any $\{E_n\}_{n\in\mathbb{N}}$ which are pairwise disjoint,*

$$Pr\left(\bigcup_{n\in\mathbb{N}} E_n\right) = \sum_{n\in\mathbb{N}} Pr(E_n)$$

For the remainder of the chapter, unless otherwise stated we use Ω to refer to a sample space, F to refer to the set of allowable events, and
Pr to denote a probability function. Events are thus sets of outcomes, and we say an event has occurred if any of the outcomes in the event occurs. Trivial events are the singleton sets which consist only of an element in Ω are called *simple* events. We'll mainly consider discrete probability spaces, ie spaces in which Ω is finite or countable. We adapt set theoretic notation to express basic operations on events. Namely, we write $E_1 \cap E_2$ to denote the occurrence of both E_1 and E_2, and we write $E_1 \cup E_2$ to denote the occurrence of either one.
Notice that our definition of a probability function guarantees that the function splits over the union of disjoint events, but doesn't necessarily say anything about the union of two events in general. It's not to difficult to prove a simple relation between probability functions and the union of two events.

(Theorem) Probability function over event union: *For events E_1, E_2,*

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)$$

- *Proof*: The proof easily follows if we separate each of the events from their intersection:

$$\begin{aligned}\Pr(E_1) &= \Pr(E_1 - (E_1 \cap E_2)) + \Pr(E_1 \cap E_2)\\ \Pr(E_2) &= \Pr(E_2 - (E_1 \cap E_2)) + \Pr(E_1 \cap E_2)\end{aligned}$$

from which it follows that

$$\begin{aligned}\Pr(E_1 \cup E_2) &= \Pr(E_1 - (E_1 \cap E_2)) + \Pr(E_2 - (E_1 \cap E_2)) + \Pr(E_1 \cap E_2)\\ &= \Pr(E_1) - \Pr(E_1 \cap E_2) + \Pr(E_2) - \Pr(E_1 \cap E_2) + \Pr(E_1 \cap E_2)\\ &= \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)\end{aligned}$$

□

We can generalize this principle to the union of any number of events.

(Theorem) Inclusion-Exclusion Principle: *Let $E_1, \cdots, E_n$ be events. Then*

$$\Pr\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \Pr(E_i) - \sum_{i=1}^{n}\sum_{j=1}^{i-1} \Pr(E_i \cap E_j) + \sum_{i=1}^{n}\sum_{j=1}^{i-1}\sum_{k=1}^{j-1} \Pr(E_i \cap E_j \cap E_k) - \cdots$$

or more compactly,

$$\Pr\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{k=1}^{n} \left((-1)^{k+1} \sum_{1 \le i_1 < \cdots < i_k \le n} \Pr\left(\bigcap_{j=1}^{k} E_{i_j}\right)\right)$$

• *Proof*: The proof follows by induction on n, where the previous theorem concerning $\Pr(E_1 \cup E_2)$ serves as the base case. For the inductive step, assume that the inclusion-exclusion principle holds for events $E_1, \cdots, E_{n-1}$. Then

$$\begin{aligned}\Pr\left(\bigcup_{i=1}^{n} E_i\right) &= \Pr\left(E_n \cup \bigcup_{i=1}^{n-1} E_i\right) = \Pr(E_n) + \Pr\left(\bigcup_{i=1}^{n-1} E_i\right) - \Pr\left(\bigcup_{i=1}^{n-1} (E_n \cap E_i)\right) \\ &= \Pr(E_n) + \sum_{k=1}^{n-1} \left((-1)^{k+1} \sum_{1 \le i_1 < \cdots < i_k \le n-1} \left(\Pr\left(\bigcap_{j=1}^{k} E_{i_j}\right) - \Pr\left(\bigcap_{j=1}^{k} (E_{i_j} \cap E_n)\right)\right)\right)\end{aligned}$$

TODO

• *Intuition*: This principle references the more general inclusion-exclusion principle concerning the cardinality of sets, which intuitively states to first sum up the individual cardinalities, yielding an overestimate which double counts all the pairwise intersections. Subtract out all the pairwise intersections to yield an underestimate since we want to count the intersections exactly once rather than not at all. Add back all the triple-wise intersections, again yielding an overestimate which double counts the triple estimates, and so on.

Recall that the definition of a probability function had the notion of disjoint events "built in" to it, in the sense that we take

$$E_1, \cdots, E_n \text{ disjoint } \rightarrow \Pr(E_1 \cup \cdots \cup E_n) = \Pr(E_1) + \cdots + \Pr(E_n)$$

as an axiom, where as usual disjoint means empty intersection. The set theoretic concept of disjointness is often confused with the probabilistic concept of *independence*, which we use to refer to the informal but intuitive idea of events which don't affect each other; that is, the occurrence of one event doesn't affect the probability of the other occurring. To make this a formal definition, let's first explore *conditional probability*, which defines what it means for the probability of one event to depend on the probability of another. This is abundant in the real world conception of probability and chance, of course - for example, the probability of it raining depends dramatically whether there are clouds in the sky.

(Definition) Conditional probability: *The **conditional probability** of event E_1 occurring given the occurrence of E_2 is defined*

$$\Pr(E_1|E_2) = \frac{\Pr(E_1 \cap E_2)}{\Pr(E_2)}$$

• *Motivation*: Intuitively, to find the probability of E_1 given E_2, we first assume that E_2 has occurred, which means that our sample space of possible outcomes now shrinks from the full Ω to the subset E_2, since the only outcomes which may occur, by assumption, are those in E_2. Thus, when asking about the probability of E_1 occurring, we restrict our attention to the outcomes of E_1 which are in E_2 (and hence have the possiblity of occurring), which is why we consider $\Pr(E_1 \cap E_2)$. Because this refers to the probability of $E_1 \cap E_2$ in the context of the full sample space and not the restricted one, we normalize by dividing by $\Pr(E_2)$.

We now define independent events as follows.

(Definition) Independent: *Events E_1, E_2 are **independent** if*

$$\Pr(E_1 \cap E_2) = \Pr(E_1)\Pr(E_2)$$

This is a good definition because, as we'd expect, independent events aren't conditioned on each other: if E_1 and E_2 are independent, then

$$\Pr(E_1|E_2) = \frac{\Pr(E_1 \cap E_2)}{\Pr E_2} = \frac{\Pr(E_1)\Pr(E_2)}{\Pr(E_2)} = \Pr(E_1)$$

so knowing whether or not E_2 occurred tells us nothing about the probability of E_1 occurring. One important subltety to address here is that just because $E_1, \cdots, E_n$ are pairwise independent doesn't mean they're independent together, and the same goes for any subset of the events being independent. To define what it means for all the events to be simultaneously independent with each other, we clearly define the following.

(Definition) Mutually Independent: *Events $E_1, \cdots, E_n$ are **mutually independent** if for any subset $I \subseteq \{1, \cdots, n\}$,*

$$\Pr\left(\bigcap_{i \in I} E_i\right) = \prod_{i \in I} \Pr(E_i)$$

Next, let's look at the law of total probability.

(Theorem) Law of Total Probability: *Let $E_1, \cdots, E_n$ be mutually disjoint events whose union is Ω. Then for any event E,*

$$\Pr(E) = \sum_{i=1}^{n} \Pr(E \cap E_i) = \sum_{i=1}^{n} \Pr(E|E_i) \Pr(E_i)$$

- *Proof*: Since E_i are pairwise disjoint, so are the events $E \cap E_i$, and since the E_i cover Ω, it follows that

$$\Pr(E) = \sum_{i=1}^{n} \Pr(E \cap E_i)$$

The last part of the theorem follows from the definition of conditional probability. $\square$

- *Intuition*: All this theorem is really stating is that since the E_i are disjoint, and cover the whole sample space, their intersections with E divide E up into n sections, and the total probability of E is the sum of probabilities of these sections, because the sections are guaranteed to be themselves disjoint.

By far the most common application of conditional probability is *Bayes' Law*, which states that

$$\Pr(A|B) = \frac{\Pr(B|A) \Pr(A)}{\Pr(B)}$$

When we have mutually disjoint events $E_1, \cdots, E_n$ which cover B, we can use the law of total probability to write

$$\Pr(E_j|B) = \frac{\Pr(B|E_j) \Pr(E_j)}{\sum_{i=1}^{n} \Pr(B|E_i) \Pr(E_i)}$$

This spawns the field of Bayesian statistics, which in general is the study of probability as a system of beliefs with a certain degree of certainty. In other words, from a Bayesian perspective, the probability of an event is metaphyscally already fixed, but in practice depends on how much information we have. The more information we have about the event, the more we can iteratively update our degree of certainty, ie our guess for the probability. We start with a *prior* probability, based on no information, and every time we get more information, we update it with Bayes' law, yielding a *posterior* probability that accounts for the new information.