Lecture 7: Probability basics: sample space, events, probability distributions, axioms of probability, the uniform distribution

Discrete Structures II (Summer 2018)
Rutgers University
Instructor: Abhishek Bhrushundi

References: Relevant parts of chapter 17 of the Math for CS book.

We will now begin the second part of the course which deals with discrete probability theory. Probability theory is the study of random/uncertain phenomenon and provides us with a framework to model these phenomenon mathematically and make predictions. Let us introduce the basic concepts of probability theory using an example:

Question . Suppose you toss a fair/unbiased coin 10 times. What is the probability that you see exactly 5 heads?

1 Experiment and sample space

Before we begin solving the problem, we need to model the random phenomenon we are dealing with. We will refer to these random phenomenon as *experiments*. The very first step is to identify what the different *outcomes* of the experiment are.

In the case of the above question, the experiment is tossing a fair coin 10 times, and so the possible outcomes are all possible sequences of length 10 where every entry in the sequence is either "H" (heads) or "T" (tails). For example, one possible outcome is $HHHHHHHHHH$ which simply denotes the outcome in which we see a heads every single time, and another example is $HTTHHHHTTT$, which says that we see a heads in the 1st, 4th, 5th, 6th, and 7th coin toss and tails in the 2nd, 3rd, 8th, 9th, and 10th coin toss.

Note the problem seems to suggest that we are interested only in the total number of heads we see during the 10 coin tosses, so we can also define an outcome to be simply the number of heads we see during the 10 trials. While this is a totally legitimate choice, it does *miss out* on other potentially useful details about how the experiment *unfolded* which might be useful or even critical to compute probabilities later on.
For example, in the 10 coin tosses case, if you define an outcome to be just the total number of heads you see during the trials, then you miss out on the information about where (i.e., during which trials) the heads came up! Thus, **it's always safer to define an outcome in a way that it captures all the important details about how the experiment unfolded**.

Anyway, the set of all possible outcomes is called the *sample space*, and is denoted using the Greek letter Ω. So in the case of the coin toss problem, we have

$$\Omega = \{H, T\}^{10},$$

where $\{H, T\}^{10}$ denotes the set of all sequences of length 10 consisting of "H" and "T". The size of the sample space, thus, is 2^{10}. Often it's useful to represent a sequence of coin tosses using binary

strings, usiing 1 to represent heads and 0 to represent tails. So in this case, we could have defined the sample space to the set of all binary strings of length 10.

2 Events

An *event* is a subset of the sample space, i.e. $E \subseteq \Omega$. This means that an event is just a set containing *some of* the outcomes (Ω, the sample space, is the set of ALL possible outcomes) If an event E is a *singleton set*, i.e. it contains exactly one outcome from Ω, the sample space, then it's called an *atomic event* or an *elementary event*. For example, consider the event E that corresponds to seeing 10 heads during the 10 trials, then $E = \{HHHHHHHHHH\}$, and E is an atomic/elementary event.

Events that contain more than one outcome are called *compound events*. For example, consider the event E of seeing a heads in the first coin toss, then E is the set consisting of all sequences of length 10 consisting of H and T that begin with an H, and $|E| = 2^9$ (why?). In this case, E is a compound event.

We say an event *takes place or happens* if the experiment results/ends up in any one of the outcomes in that event.

What is the set of all possible events? It's the powerset of the sample space, i.e. 2^{Ω} or $powerset(\Omega)$. (Why?)

2.1 Operations on events

Since events are nothing but subsets of the sample space Ω, we can use the usual set operations on them. Let Ω be some sample space, then

- if $E \subseteq \Omega$ is an event, the complement of the event E is simply the set $E^c = \Omega - E$. For example, going back to the 10 coin tosses example, let E be the event that you see at least one heads during the 10 trials. As in the case of counting, it's much easier to deal with the complement of E, i.e. the event that you see no heads and thus all tails. Thus, in this case, $E^c = \{TTTTTTTTTT\}$, and $|E| = |\Omega| - |E^c| = 2^{10} - 1$ (think difference method!). Another way of to think of E^c is to think of it as the set of all outcomes where the event E doesn't "happen".
- if E_1 and E_2 are two events in Ω, then we can define the event $E_1 \cup E_2$, i.e. the set of all outcomes that are either in E_1 or in E_2 or in both. For example, if E_1 is the event of seeing at least one heads and E_2 is the event of seeing no heads, then $E_1 \cup E_2$ is the event of seeing zero or more heads which is basically Ω, the entire sample space.
- Similarly, one can define $E_1 \cap E_2$, the set of all outcomes that are in both E_1 and E_2. Suppose E_1 was the set of outcomes in which you see at least one heads, and E_2 was the set of outcomes where you see at least one tails, then $E_1 \cap E_2$ is the set of all outcomes where you see at least one heads AND at least one tails. (Can you compute $|E_1 \cap E_2|$?)
- Finally, we can talk about $E = \emptyset$, the empty event that doesn't contain any outcome and *never* happens (why?), and $E = \Omega$, the trivial event that always *happens* (why?).

2.2 Mutually exclusive events/disjoint events

If we have a bunch of events $E_1, \ldots, E_k$ such that for every pairs of sets in this bunch E_i, E_j we have that $E_i \cap E_j = \emptyset$, then we say that $E_1, \ldots, E_k$ are *mutually exclusive or disjoint* events. This means that if an event E_i happens/takes place, then none of the other events among the remaining $k-1$ events take place or happen. For example, if E_1 is the event of seeing at least two heads during the 10 trials, and E_2 is the event of seeing exactly one heads during the trials, and E_3 is the set of all outcomes where you don't see any heads, then E_1, E_2, E_3 are mutually exclusive events, and in fact $E_1 \cup E_2 \cup E_3 = \Omega$ (why?).

3 Probability

Let's go back to the question we are trying to solve. The question asks what is the probability of seeing exactly 5 heads. We have identified what the outcomes are and what the sample space is. We now want to define a probability distribution on our sample space. What is a probability distribution? It's a function that assigns a real number to every event of our sample space. The real number is supposed to be a "measure of likelihood" of that event.

More formally the probability distribution is a function $P : powerset(\Omega) \to \mathbb{R}$. Why is the domain of the function P $powerset(\Omega)$? It's because we want the function P is supposed to associate a real value with every event, and its domain, thus, must be the set of all events which is $powerset(\Omega)$ or 2^{Ω}.

What behavior do we expect the probability function to have? We want it to model our intuitive notion of "likelihood" or "chance".

3.1 Axioms of probability

Here are some common sense properties we want the probability distribution function P to satisfy:

1. Obviously, it doesn't make sense to use negative numbers to denote the likelihood of an event or the chance of the event happening (What would it even mean to say there is a -50% chance of raining tomorrow?). So the first condition we want is that for every event E,

$$P(E) \geq 0.$$

2. Remember, when we defined the event $E = \Omega$ and said it's the event that *always happens*. Thus, we want to have that

$$P(\Omega) = 1.$$

   Similarly, remember the event $E = \emptyset$, the event that has no outcomes and thus should never happen? It makes sense to have $P(\emptyset) = 0$. (it turns out we can *derive* this fact from the other axioms — we will see this later)

3. Suppose E_1 is the event of seeing at least two heads during the 10 trials and E_2 is the event of seeing exactly 1 heads during the trial. First note that E_1 and E_2 are disjoint. Also, if the chance of E_1 happening is p_1 and the chance of E_2 happening is p_2, we *expect* the chance of

either E_1 happening or E_2 happening to be $p_1 + p_2$. We can formalize this to the following: If $E_1, \ldots, E_k$ are disjoint events/mutually exclusive events, then

$$P(E_1 \cup \ldots \cup E_k) = P(E_1) + P(E_2) + \ldots + P(E_k).$$

We can all this axiom the *sum rule of probability.*

3.2 Uniform distribution for equally likely outcomes

Ok, so now we understand that we need to define a probability distribution on the sample space Ω which in the case of the coin toss problem is the set of all binary strings of length 10 (or the set of all sequences of length 10 made using H and T). How should out probability distribution look, or in other words what numbers should we associate with the all the events?

Let's try to first define the probability for the atomic/elementary events. There are 2^{10} atomic events, let's denote them by $E_1, \ldots, E_{2^{10}}$. Intuitively, since the coins are all unbiased and fair, we expect all the outcomes, and thus all the elementary events, to be equally likely! (convince yourself!). And so we want:

$$P(E_1) = P(E_2) = \ldots = P(E_{2^{10}}).$$

We know that $P(\Omega) = 1$, and also that $\Omega = E_1 \cup \ldots \cup E_{2^{10}}$. The last fact is true because the sample space Ω is made up of these elementary/atomic events and nothing else, so we can write Ω as a union of these events. Moreover all these elementary events are disjoint because remember: each of these elementary events contain exactly one outcome, and no two events contain the same outcome (otherwise they would be the same event!). We can thus use the sum rule of probability:

$$1 = P(\Omega) = \sum_{i=1}^{2^{10}} P(E_i).$$

Let $p \geq 0$ be such that

$$P(E_1) = P(E_2) = \ldots = P(E_{2^{10}}) = p.$$

Then, we have using the above equation that

$$1 = \sum_{i=1}^{2^{10}} P(E_i) = 2^{10} \times p \implies p = \frac{1}{2^{10}},$$

and thus,

$$P(E_1) = P(E_2) = \ldots = P(E_{2^{10}}) = p = \frac{1}{2^{10}}.$$

Thus, our probability distribution should assign $1/2^{10}$ to every elementary/atomic event!

But so far we have just assigned probabilities for atomic events! How about compound events? Let E be an arbitrary compound event of size k that consists of the outcomes $o_1, \ldots, o_k$, i.e, $E = \{o_1, \ldots, o_k\}$. Then we can write

$$E = \{o_1\} \cup \{o_2\} \cup \ldots \{o_k\},$$

where $\{o_1\}, \ldots, \{o_k\}$ are nothing but elementary/atomic events! Using the sum rule for probability

$$P(E) = \sum_{i=1}^{k} P(\{o_i\}).$$

Recall that every elementary event was given a probability of $1/2^{10}$, and so

$$P(E) = \sum_{i=1}^{k} P(\{o_i\}) = \frac{k}{2^{10}} = \frac{|E|}{2^{10}}.$$

So we have now completely specified the probability distribution P. It's not hard to verify that the P defined above satisfies the axioms of probability:

1. For every event E, since $|E| \geq 0$, we have that $P(E) = \frac{|E|}{2^{10}} \geq 0$
2. Also, $P(\Omega) = \frac{|\Omega|}{2^{10}} = \frac{2^{10}}{2^{10}} = 1$, and similarly $P(\emptyset) = \frac{|\emptyset|}{2^{10}} = 0$.
3. Let E_1 and E_2 be disjoint events of Ω, then we know using the sum rule for sets that $|E_1 \cup E_2| = |E_1| + |E_2|$. It follows that

$$P(E_1 \cup E_2) = \frac{|E_1 \cup E_2|}{2^{10}} = \frac{|E_1| + |E_2|}{2^{10}} = \frac{|E_1|}{2^{10}} + \frac{|E_2|}{2^{10}} = P(E_1) + P(E_2).$$

This sort of a probability distribution that assigns the same probability to every elementary event/outcome is called the *uniform distribution*.

In general, for any experiment and any sample space Ω associated with it, the uniform distribution on Ω is the one that assigns probability $\frac{1}{|\Omega|}$ to all the elementary/atomic events of Ω, and for a compound event E assigns the probability $\frac{|E|}{|\Omega|}$.

Whenever you are dealing with an experiment where all the outcomes should be equally likely, use the uniform distribution!

3.3 Solution to the question

We are now in perfect shape to answer the question: what is the probability of seeing exactly 5 heads during the 10 trials? Well, our Ω was defined to be the set of all binary strings of length 10, and we used the uniform probability distribution on Ω. We are interested in the probability of the event E that we see exactly 5 heads. What outcomes does E contain? E contains all the binary strings that have exactly 5 ones, and thus $|E| = \binom{10}{5}$. Thus, the uniform distribution would give us

$$P(E) = \frac{|E|}{|\Omega|} = \frac{\binom{10}{5}}{2^{10}}.$$

Here is another question to think about:

Question . If we roll two fair dice, one black and one white, what is the probability that both dice turn up 3? What is the probability that both dice turn up 3 or both dice turn up 1?

It's not hard to see that Ω is $\{1,\ldots,6\}\times\{1,\ldots,6\}$, i.e. the set of all 2-tuples where every entry is a number between 1 and 6, and the first entry in the tuple represents the number that the black dice rolls, and the second entry is the number that the white dice rolls. Thus $\Omega = 36$. Let $E_1 = (3,3)$ and $E_2 = (1,1),(3,3)$. Then E_1 is the set of outcomes where both dice roll 3, and E_2 is the event where both dice roll 3 or both dice roll 1. It follows from out discussion above that

$$P(E_1) = \frac{1}{36}$$

and

$$P(E_2) = \frac{2}{36}.$$

3.4 Some more properties of probability distributions

Let us fix an experiment and associate a fixed sample space Ω with it. Furthermore, let us say that we have a probability distribution P on Ω. Let $E \subseteq \Omega$ be an event. Suppose we know the value of $P(E)$. What is $P(E^c)$?

Recall that $E^c = \Omega - E$, and thus we have that E and E^c are disjoint events and $E \cup E^c = \Omega$. It follows from the sum rule of probability that

$$P(\Omega) = P(E \cup E^c) = P(E) + P(E^c).$$

We also know that $P(\Omega) = 1$, and so we get that

$$P(E) + P(E^c) = 1 \implies P(E^c) = 1 - P(E).$$

Thus: the probability of the event not happening is one minus the probability of the event happening.

Using this, we can also conclude the sort-of-obvious fact that for every event E, $P(E) \leq 1$. This is because we know that $P(E) + P(E^c) = 1$, and we also know that $P(E^c) \geq 0$, and so $P(E) \leq 1$. Thus, combining this with one of the axioms we saw earlier, we have that for every event E,

$$0 \leq P(E) \leq 1.$$

We know how to compute $P(E_1 \cup E_2)$ when E_1 and E_2 are disjoint, but how about the case when they are not? Turns out one can prove an inclusion-exclusion formula for probability distributions:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2).$$

To see why this is true, observe that $E_1 - E_2$, $E_2 - E_1$, and $E_1 \cap E_2$ are all disjoint sets, and we can write

$$E_1 \cup E_2 = (E_1 - E_2) \cup (E_2 - E_1) \cup (E_1 \cap E_2).$$

Thus using the sum rule of probabilities, we get

$$P(E_1 \cup E_2) = P(E_1 - E_2) + P(E_2 - E_1) + P(E_1 \cap E_2).$$

We also know that $P(E_1) = P(E_1 - E_2) + P(E_1 \cap E_2)$ (why?) and $P(E_2) = P(E_2 - E_1) + P(E_1 \cap E_2)$, and thus we can substitute $P(E_1 - E_2) = P(E_1) - P(E_1 \cap E_2)$ and $P(E_2 - E_1) = P(E_2) - P(E_1 \cap E_2)$ in the above expression for $P(E_1 \cup E_2)$. This gives us

$$P(E_1 \cup E_2) = P(E_1) - P(E_1 \cap E_2) + P(E_2) - P(E_1 \cap E_2) + P(E_1 \cap E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2).$$

Once we have the inclusion exclusion formula for the probability of union of two events, using the same ideas as in the case of cardinality of union of sets we can find a general formula for the probability of union of k events.

Using the inclusion-exclusion formula, we can derive an important ineqality:

$$P(E_1 \cup E_2) \leq P(E_1) + P(E_2).$$

This is called the *union bound*. The proof is really simple: we know that $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$, and that $P(E_1 \cap E_2) \geq 0$ (why?), and thus $P(E_1 \cup E_2) \leq P(E_1) + P(E_2)$.

Suppose that we have events E_1, E_2 such that $E_1 \subseteq E_2$. How does $P(E_1)$ compare with $P(E_2)$? Intuitively, whenever E_1 "happens", E_2 also happens, and thus it should be the case that $P(E_1) \leq P(E_2)$. We can prove this formally: we can write $P(E_2) = P(E_1) + P(E_2 - E_1)$ (why?), and observe that $P(E_2 - E_1) \geq 0$, and this gives us that $P(E_2) \geq P(E_1)$!