Lecture 7: Probability Theory

Anup Rao

April 15, 2019

We begin talking about probability theory. We learn about the famous Bayes' rule, and start talking about Random Variables.

Probability

A PROBABILITY SPACE GIVES a very useful way to generalize the idea of counting the size of sets. A *probability space* is defined by a set Ω, usually called the *domain* or *sample space*, and a *distribution*. A distribution is a function $p : \Omega \to \mathbb{R}$ mapping the elements of Ω to real numbers so that

There are many commonly used notations for distributions. Sometimes people write Pr or *Prob* instead of p.

- For all $x \in \Omega$, $p(x) \geq 0$.
- $\sum_{x \in \Omega} p(x) = 1$.

A probability space captures the concept of a random process happening. The elements x encode all the possible outcomes of the process, and $p(x)$ represents the chance that x is the outcome.

How would you encode a single coin toss as a porbability space?

An *event* in the probability space is a subset $E \subseteq \Omega$. The probability of the event is $p(E) = \sum_{x \in E} p(x)$.

For example, suppose we toss a fair coin twice. Then the sample space is $\Omega = \{HH, TT, HT, TH\}$. The distribution puts equal weight on all outcomes, so we have $p(x) = 1/4$ for every $x \in \Omega$. If we consider the subset $E \subseteq \Omega$ where the first coin toss is heads, then it is of size 2 so $p(E) = 2/4 = 1/2$.

A very common situation is when the distribution is uniform over the sample space, meaning that $p(x) = p(y)$ for all $x, y \in \Omega$. In this case, the probability of an event E is exactly $p(E) = \frac{|E|}{|\Omega|}$—it is just the ratio of the size of E to the size of Ω. However, the nice thing that probabilities make sense even when the sets Ω and E are of infinite size.

Example

Suppose we toss a fair coin n times. What is the probability that the number of heads is even?

Here the sample space consists of all 2^n possible coin tosses. If E denotes the event that the number of heads is even, then we have

$$|E| = \binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \dots$$

So, the probability that the number of heads is even is

$$\frac{|E|}{|\Omega|} = \frac{\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \dots}{2^n}.$$

This looks like a complicated expression, but we can simplify it using facts that we have proved about binomial coefficients. First, we know that $2^n = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \dots$, and we also know that all the even coefficients sum to exactly the same value as all the odd coefficients sum to. So, we have

$$\frac{|E|}{|\Omega|} = \frac{\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \dots}{\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \dots} = \frac{\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \dots}{2(\binom{n}{0} + \binom{n}{2} + \dots)} = \frac{1}{2}.$$

See Lecture 4 from January 10.

You can use a very similar argument to prove that the probability that the number of heads is odd is $1/2$.

Example

Suppose you take a uniformly random walk in a grid, starting from the point $(0, 0)$ and ending at the point (n, n), and always moving either up or right, what is the probability that you cross the diagonal?

In previous lectures, we calculated the total number of such walks to be $|\Omega| = \binom{2n}{n}$. We also calculated the number that cross the diagonal as $|E| = \frac{n}{n+1} \cdot \binom{2n}{n}$. So, the probability of crossing the diagonal is $\frac{\frac{n}{n+1} \cdot \binom{2n}{n}}{\binom{2n}{n}} = \frac{n}{n+1}$.

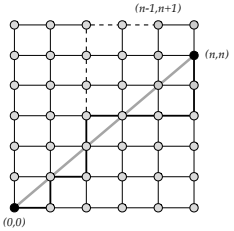

Figure 1: We discussed walking on the grid in Lecture 3 on April 5.

Conditional Probability

Something very nice happens to probability spaces when you *zoom in* to a particular event. For example, consider the two events A, B in the sample space Ω shown below. As usual, let p denote the distribution of the outcomes in Ω.

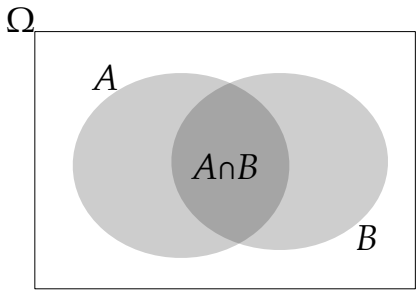

Figure 2: Two events and their intersection in a probability space.

Let us think about the event $A \cap B$. This event corresponds to a subset of the sample space Ω, but it is also a subset of A. In a sense,

we could think of A itself as a sample space, and $A \cap B$ as an event in that sample space.

This view is particularly useful to modify our view of the probability space when some partial information has been revealed to us. If we have a probability space as above, and we know that the event A has happened, then the probability that B also happens, *given* that A has happened is

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

In particular, this definition gives:

$$p(B|A) \cdot p(A) = p(A \cap B) = p(A|B) \cdot p(B),$$

which implies that

$$p(B|A) = \frac{p(A|B) \cdot p(B)}{p(A)}.$$

This last equation is called *Bayes' rule*.

Example: Two Dice

Suppose you roll two dice at the same time. There are $6 \times 6 = 36$ possible outcomes of these rolls, and all are equally likely. What is the probability that the dice add up to 8?

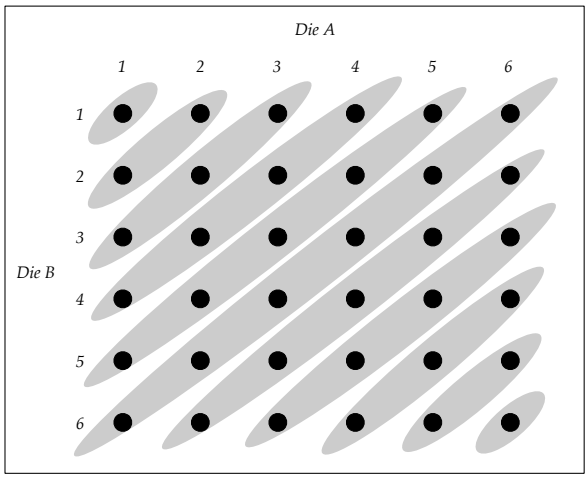

Figure 3: The sample space when two dice are rolled.

There are 11 possible values for the sum of the two dice, from $2 - 12$. Figure 4 shows the 11 corresponding events where the sum of the dice is fixed to a value. So, for example, the probability that the sum of the dice is 2 is only $1/36$, but the probability that the sum is 7 is the largest: $6/36 = 1/6$. If F denotes the event that the sum of the dice is 8, we have

$$p(F) = 5/36.$$

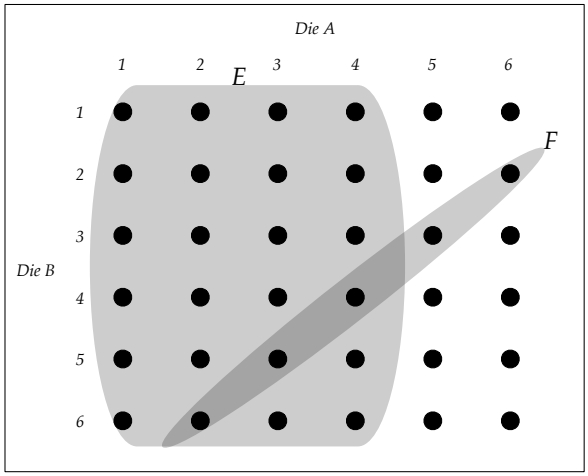

Figure 4: The events corresponding to the sum of the dice being 8 and the first die being ≤ 4 are shown.

Now, let us consider a different kind of question. What is the probability that the sum of the dice is 8, given that the first die gives a value that is ≤ 4?

If E denotes the event that the first die gives a value ≤ 4 and F denotes the event that the sum of the dice is 8, then we see that

$$p(E) = 4/6 = 2/3,$$

$$p(F) = 5/36,$$

and

$$p(E \cap F) = 3/36 = 1/12.$$

So,

$$p(F|E) = \frac{p(E \cap F)}{p(E)} = \frac{1/12}{4/6} = 1/8.$$

You can also calculate $p(F|E)$ directly from the picture: it corresponds to picking 3 out of 4×6 outcomes, so the probability is $3/24 = 1/8$. From the picture, we see that

$$p(E|F) = \frac{3}{5},$$

and we can verify Bayes' rule:

$$p(F|E) = \frac{1}{8} = \frac{(3/5) \cdot (5/36)}{2/3} = \frac{p(E|F) \cdot p(F)}{p(E)}.$$

If H is the event that the sum of the dice is 6, we see that $p(H|E) = 4/24 = 1/6$. So, even though the probablity of the sum being 6 is the same as the probability that the sum is 8, once we know that the first die roll is at most 4, the lower sum values get a boost: the probablity that the sum is 5 conditioned on E is larger than the probability that the sum is 8 conditioned on E.

The Conditional Probabilities give a Probability Space

It is easy to check that the probabilities $p(x|E) = \frac{p(x \cap E)}{p(E)}$ satisfy all the axioms that a probability space is supposed to satisfy. It is clear that $p(x|E) \geq 0$. Moreover, we have:

$$\sum_{x \in E} p(x|E) = \sum_{x \in E} \frac{p(x \cap E)}{p(E)} = \frac{p(E)}{p(E)} = 1.$$

So we can think of a new distribution $q(x)$ given by $q(x) = p(x|E)$. This distribution gives 0 weight to the points in the sample space outside of E, and is a valid distribution supported on Ω. It can also be viewed as a distribution supported on E, since it does not assign any weight to the points outside E.

Random Variables

IT CAN BE QUITE CUMBERSOME to talk about events in a probability space, because there are so many of them. One piece of notation that really helps is the concept of a *random variable*. A random variable is just a function $X : \Omega \to S$ that maps the points in the sample space to some other set.

Random variables can be thought of as a partition of the entire sample space into disjoint events, namely those sets where the random variables is constant.

For example, in the case that we are throwing two dice, we can define the random variable X to be the value of the first die, and Y to be the value of the second die.

Then the event E is the same as the event that $X \leq 4$, and the event F is the same as the event that $X + Y = 8$.

The distribution p on the probability space induces a distribution $p(X)$ on the values in $[6]$ taken by the random variable X, a distribution $p(Y)$ on the values in $[6]$ taken by the random variable Y and a *joint distribution* $p(X, Y)$ on the values in $[6] \times [6]$ taken by both values. Often, if $p(X, Y)$ is the distribution of X and Y, then $p(X)$ is referred to as the *marginal distribution* on X.