Lecture 8: Conditional probability I: definition, independence, the tree method, sampling, chain rule for independent events

Before we begin the lecture on conditional probability, I want to mention two tips:

1. For a counting problem, unless the problem explicitly says *identical* objects, don't assume the objects are identical.

2. For a probability problem, always assume that the objects are distinct, *even if* the problem says they are identical. This is because when computing the probability we are concerned with *the number of ways of reaching an outcome* and not just the outcomes themselves. For example, if a bucket contains 10 identical white balls, and 20 identical red balls, then the probability of picking a red ball (if you pick a ball at random) is 2/3. This is because the number of ways of reaching the outcome of seeing a red ball is 20, and the total number of outcomes is 30. If you just assume that even identical objects are distinct, you don't have to worry about this, since the distinction between "ways" and "outcomes" disappears. This trick works in 99% of the cases.

1   Definition of conditional probability

Let us assume Ω is a sample space for an experiment, and A and B are events. Let us motivate conditional probability through two examples:

Question . I want to draw two cards from a complete deck of shuffled cards, one at a time. If the first card I draw is an ace, what is the probability that the second card is also an ace?

In this case, let B be the event that the first card I drew was an ace, and let A be the event that the second card I draw will be an ace. I have been given the information that the event B has happened/occurred. If I was not given this information, what would be the probability of A? We would define the sample space Ω as set of all sequences of length 2 where every entry of the sequence is one of the 52 cards and no repetition is allowed, and thus $|\Omega| = \binom{52}{2}2!$. A would be the set of sequences of two cards (without repetition) so that the second card is an ace. Thus, $|A| = \binom{4}{1}\binom{51}{1}$ (why?), and thus

$$P(A) = \frac{|A|}{|\Omega|} = \frac{4 \times 51}{52 \times 51} = \frac{4}{52},$$

since all outcomes are equally likely and we use the uniform distribution.

But now you have been given the information that B occured, i.e. the first card that was picked was an ace. Does this affect the probability of A happening? Intuitively, it does because when we

are picking the second card we have one less ace to pick from. We represent this "new" probability, i.e. the probability of A happening *given* that B has happened by $P(A|B)$ ("probability of A given B").

To compute this probability, note that when picking the second card, our "effective" sample space has changed: only outcomes within $B \subseteq \Omega$ are now possible. If this is the case, then only those outcomes in A that are *also* contained in B can happen, and thus, intuitively,

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

It turns out this is indeed how conditional probability of A given B is defined in the uniform case. More generally,

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

We can now use this formula to compute $P(A|B)$ for the above problem. To do this, we need to find $|A \cap B|$ and $|B|$. You can convince yourself that $|B| = 4 \times 51$. As for $|A \cap B|$, i.e. the number of sequences of cards where both the first and second card are aces, it's 4×3. Thus,

$$P(A|B) = \frac{4 \times 3}{4 \times 51} = \frac{3}{51}.$$

This also makes intuitive sense: after you have picked an ace in the first draw, there are 3 aces left (those are the outcomes we want) and the total outcomes is the total number of cards left which is 51, and thus the probability should be 3/51.

Let us consider a different scenario where conditional probabilities can be defined and used.

Question . I drew two cards from a complete deck of shuffled cards, one at a time. If the second card I drew is an ace, what is the probability that the first card was also an ace?

Here we can assume that the experiment has happened, and that the second card that was drawn was an ace, and we want to do a "retrospective analysis" and compute how likely it is that the first card was also an ace *given* the information that the second card was an ace? Let A be the event that the second card is an ace, and let B be the event that the first card is an ace. If you have no information about A happening, then you would say that the probability that B happened is just 4/52 (why? see above calculations). But now that you have been told the information that A did happen, what is the probability that B happened. Again, we denote this conditional probability by $P(B|A)$, and following the same intuition as above, we can say that

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

in the uniform distribution case, and in general

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

You can use the above formulas to conclude that in this case $P(B|A) = \frac{3}{51}$. This means that with this new information (that A did happen) you "updated" your probability that B happened (we will see more about updates later on).

Thus, these are the two scenarios in which conditional probability is useful: the first one is a scenario where there is a multi-step experiment, and you told that some event that involves the previous steps has happened, and you want to "predict" how likely it is that some event that involves future steps will happen, and the second one is where the experiment has already occured and you are doing a retrospective analysis (you are told some event B happened for sure, and want to know the probability that A happened in the light of this information).

1.1   A word of caution

In the above example, it turned out that $P(A|B) = P(B|A)$ but this is not generally true because $P(A|B) = P(A \cap B)/P(B)$ and $P(B|A) = P(A \cap B)/P(A)$, and thus the two are equal if and only if $P(A) = P(B)$.

1.2   Comparing $P(A)$ with $P(A|B)$

In general, $P(A|B)$ could larger or smaller as compared to $P(A)$. Consider the case when $B \subseteq A$. In this case, if I tell that the experiment resulted in an outcome of the event B, then it also resulted in an outcome of the event B, and thus $P(A|B) = 1 \geq P(A)$. In the other extreme, let's say A and B are mutually exclusive events, i.e. if one happens then the other cannot happen. In this case, $P(A|B) = 0$ (why?), and thus $P(A|B) <= P(A)$.

What can we say if $P(A|B) = P(A)$? Intuitively, it means that B happening does not affect the probability of A happening, and thus in some sense the two events are "uncorrelated" (in the colloquial sense).

2   Independent events

If A, B are two events of a sample space Ω such that $P(A|B) = P(A)$, then they are called independent events. Note that if $P(A|B) = P(A)$ then $P(A \cap B)/P(B) = P(A)$ and so

$$P(A \cap B) = P(A)P(B).$$

It turns out we can also define two events to be independent if $P(A \cap B) = P(A)P(B)$ and the two definitions are equivalent (you can derive one assuming the other is true), and the latter is the preferred definition since in certain cases when $P(B) = 0$, $P(A|B)$ might not be defined.

Obviously, if $P(A|B) = P(A)$ it's also true that $P(B|A) = P(B)$ (again, you can derive one from the other).

Question . Suppose we roll a black and a white dice (both fair). Let A be the event that the white dice rolls one, and let B be the event that the sum of the numbers rolled by the dice is 7. Show that A and B are independent.

Proof. In this case Ω is the set of all ordered pairs (w, b) where the first component represents the number rolled by the white dice and the second one represents the number rolled by the black dice.

It is easy to see, then, that

$$B = \{(1,6), (6,1), (2,5), (5,2), (3,4), (4,3)\}$$

and thus $|B| = 6$. Within B, the outcomes that have the white dice rolling a one is just $(1,6)$, and thus $|A \cap B| = 1$, and so

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{1}{6}.$$

It is not hard to show that $|A| = 6$ and $|\Omega| = 36$, and so

$$P(A) = \frac{1}{6} = P(A|B).$$

This completes the proof that A and B are independent. □

2.1 Mutual independence

What does independence look like when we are talking about a whole bunch of events, say $A_1, \ldots, A_n$ in a sample space Ω. We say that $A_1, \ldots, A_n$ are *mutually independent* if for every $1 \leq i \leq n$, the probability of A_i happening is unaffected by the occurence of a subset of the n events (obviously, the subset doesn't include A_i itself). For example, say $n = 7$ then the probability of A_3 happening is the same as the probability of A_3 happening given that A_4, A_6, A_7 have happened, etc.

Formally, we say that $A_1, \ldots, A_n$ are mutually independent if

$$P(A_1 \cap A_2) = P(A_1)P(A_2)$$

$$P(A_1 \cap A_3) = P(A_1)P(A_3)$$

$$\vdots$$

$$P(A_1 \cap A_n) = P(A_1)P(A_n)$$

$$P(A_2 \cap A_3) = P(A_2)P(A_3)$$

$$\vdots$$

$$P(A_{n-1} \cap A_n) = P(A_{n-1})P(A_n)$$

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$$

$$\vdots$$

$$P(A_{i_1} \cap \ldots A_{i_k}) = P(A_{i_1}) \cdot \ldots \cdot P(A_{i_k})$$

$$\vdots$$

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \cdot \ldots \cdot P(A_n).$$

That is, if you take any subset of $A_1, \ldots, A_n$, then the probability of their intersection if the product of the individual probabilities of the events in the subset.

If someone says $A_1, \ldots, A_n$ are independent events, they mean they are mutually independent. We will see other notions of independence in future lectures.

2.2 The chain rule for independent events

Suppose Ω is a sample space, and E is an event of Ω. Furthermore, let's assume that $E = A_1 \cap A_2 \cap \ldots \cap A_n$, where $A_1, \ldots, A_n$ are mutually independent events. In many such situations, it's hard to directly compute $P(E)$, but it's easy to compute each of $P(A_1), \ldots, P(A_n)$. We can exploit the fact that $A_1, \ldots, A_n$ are independent, and say that

$$P(E) = P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2) \ldots P(A_n).$$

Consider the following problem

Question . 20 fair dice are tossed. What is the probability the product of the numbers rolled by the 20 dice is odd?

Proof. Ω in this case is the set of all sequences of 20 coin tosses or the set of all binary strings of length 20 where 1 represents heads and 0 represents tails. E is the event that the product is odd. Note that the product of the numbers rolled by the 20 dice is odd if and only if all 20 dice roll an odd number. Let $A_1, \ldots, A_{20}$ be such that A_i denotes the probability that the ith dice rolls an odd number. Then we can write $E = A_1 \cap A_2 \cap \ldots \cap A_{20}$. Furthermore, $A_1, \ldots, A_n$ are all independent, and so
$$P(E) = P(A_1 \cap \ldots \cap A_{20}) = P(A_1)P(A_2) \ldots P(A_{20}).$$

Note that for every dice i we have that $P(A_i) = 1/2$ since out of the six possible outcomes exactly three lead to an odd number being rolled. This means that $P(E) = (1/2)^{100}$. □

3 The tree method for multi-stage experiments

The tree method is useful for analyzing experiments which have multiple stages: "first do this, if the outcome is ... then do that, if the outcome of that is ... then do ...". For example, consider the following question:

Question . I have two bags. Bag 1 contains two fair coins and Bag 2 contains a fair coin and a coin with two heads. I toss a fair coin. If the outcome is heads, I pick Bag 1, and if the outcome is tails I pick Bag 2. Then from the bag in picked I the previous step, I randomly pick a coin such that both coins in the bag are equally likely to be picked. Finally I toss the coin I picked from the bag. What is the probability that the final outcome is heads? Suppose I tell you that the final outcome of the experiment was heads how likely is it that the final coin toss was done with the coin with two heads?

Proof. Let us assume that Bag 1 contains two fair coins F_1 and F_2 and Bag 2 contains F_3 and B_1, where B_1 is the the coin with two heads. The experiment happens in three stages:

1. You first coin to decide which Bag to go for.

2. Based on what coin you got, you select a bag, and then randomly choose one of the coins in the bag.

3. Finally you toss the coin you took out of the bag.

We can represent the various stages of the experiment and their outcomes using the following tree:
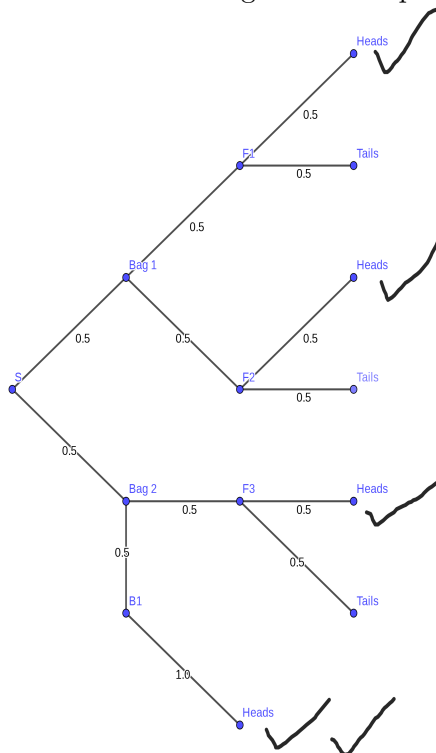
Here S denotes the start of the experiment. The probability of a branch is written along the branch. For example, from S we branch into two possible outcomes with equal probabilities ($1/2$ each): either pick Bag 1 or pick Bag 2.

Once we have the tree drawn along with all the probabilities written on the branches, we can now identify the outcomes we are interested in. Notice that the outcomes of the final stage are the leaves of this tree. Since we are interested in the probability that we see a heads in the final stage, let's first put a check mark on all those leaves which correspond to seeing a heads. The probability of arriving at a particular leaf is the product of all the probabilities along the path from S to that leaf.

For example, consider the heads leaf on the very top of the figure. The probability of arriving at that leaf is $0.5 \times 0.5 \times 0.5 = 0.125$. Similarly, for the second leaf from the top that is labelled heads, the probability is again 0.125, and it's the same for the third leaf from the top that is labelled "heads". For the bottom-most leaf that is labelled heads, the probability is $0.5 \times 0.5 = 0.25$. Thus, the total probability of seeing a heads in the final stage of the experiment is $0.125 + 0.125 + 0.125 + 0.25 = 0.625$. If B is the event of seeing heads in the final coin toss, we have that $P(B) = 0.625$.

Let A denote the probability of using the two-headed coin (i.e., B_1) in the last stage. We want to compute $P(A|B)$. To do this, we first identify the leaves whose root-to-leaf paths pass through $B1$ (this corresponds to the outcome of using a two-headed coin) among those leaves that are labelled "heads" and put another check mark near it. In our case, there is only one such leaf, and the probability of that leaf is $0.5 \times 0.5 \times 1 = 0.25$. This is basically $P(A \cap B)$. We can now compute

$P(A|B)$:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.25}{0.625} = 0.4.$$

$\square$

In general, after putting down the second round of check marks, you would add up the probabilities of all the leafs that have double check marks and divide that by the sum of the probabilities of all the leaves with a single check mark to get the desired conditional probability.

4 Sampling

Suppose you have an urn that contains n balls in it. Each ball has a unique number between 1 and n printed on it. *Sampling* is nothing but picking balls from the urn. There are two kinds of sampling. Let us say you want to sample k balls from the urn:

1. You can take k balls at once, or

2. You can take one ball at a time.

For the latter type, we have to possibilities:

1. Once you pick a ball, you put it back into the urn so that it may be picked again later, or

2. Once you pick a ball, you don't put it back, so that every ball may be picked at most once.

If you sample one at a time, and put the ball back after it is picked, this is called *sampling with replacement*, and if you sample balls one at a time, and don't put back the balls that are already picked, then this is called *sampling without replacement*.

Sometimes it's helpful to recognize that the problem you are dealing with is basically a sampling problems. Once you figure that out, you need to figure out whether you are picking objects at once, or one-by-one. If it's the latter you should figure out whether it's with or without replacement. Once you have nailed down these details, you are in a good shape because the solutions to many sampling problems follow the same template.

Question . Every time you buy a bag of chips from the store you are equally likely to find one of 10 distinct toys in the bag. If you buy 5 bags of chips, what's the probability that at least two of the bags contain the same toy?

Proof. This is a sampling problem in disguise. Every time we buy a bag of chips we are basically sampling from the set of toys, and are equally likely to sample one of 10 distinct toys. Is this with or without replacement? It's with replacement because it is possible sample the same toy more than once.

Let's define our outcomes as sequences of length 5, where every entry can be a number between 1 and 10 (representing the 10 different toys), and the choices for every entry are independent. Thus,

$|\Omega| = (10)^5$. Let A be the set of outcomes that correspond to picking at least one toy twice. Then A^c is the set of all outcomes where we pick 5 different toys. The number of sequences of length 5 where every entry is a number between 1 and 10, and no repetition is allowed, is $\binom{10}{5}5!$ (choose 5 out of 10 and then permute them). Thus, $|A^c| = \binom{10}{5}5!$. This means

$$P(A) = 1 - P(A^c) = 1 - \frac{\binom{10}{5}5!}{(10)^5}.$$

$\square$

Here is another question, but this one deals with *sampling without replacement*:

Question . A box has 6 identical yellow bulbs, 5 identical blue bulbs, and 4 identical white bulbs. Bulbs are sampled one-by-one without replacement. What is the probability that the third bulb that's sampled is not white?

Proof. Let's first use the tip we discussed at the beginning of this lecture: make identical objects distinct! So let's assume we have 15 distinct bulbs: $Y_1, \ldots, Y_6, B_1, \ldots, B_5, W_1, \ldots, W_4$. Since we are sampling one at a time without replacement, a good choice for the sample space if the set of all possible permutations of $Y_1, \ldots, Y_6, B_1, \ldots, B_5, W_1, \ldots, W_4$ – a permutation represents the order in which the bulbs are picked. Thus, $|\Omega| = 15!$[1].

We are interested in the event E that consists of all permutations in which a white bulb *does not* appear in the third position. Then E^c is the set of all permutations in which a white bulb does occur in the third position. There are 4 choices for deciding which white bulb to have in the third position, and there are 14! ways to arrange the remaining bulbs, thus $|E^c| = 4(14!)$. This implies that
$$P(E) = 1 - P(E^c) = 1 - \frac{4(14!)}{15!} = 1 - \frac{4}{15} = \frac{11}{15}.$$

$\square$

Question . A box has 6 identical yellow bulbs, 5 identical blue bulbs, and 4 identical white bulbs. You pick 5 random bulbs from the box. What is the probability that at least one yellow bulb was picked?

Proof. Again, we make all the bulbs distinct: so we have 15 distinct bulbs now (as in the previous problem). This problem can be modeled as sampling 5 bulbs at once from the box, and so the sample space should be the set of all possible subsets of $Y_1, \ldots, Y_6, B_1, \ldots, B_5, W_1, \ldots, W_4$ of size 5. There are $\binom{15}{5}$ such subsets, and so $|\Omega| = \binom{15}{5}$.

Let E denote the event that at least one yellow bulb is picked. Then E^c is the event that no yellow bulb is picked. There are exactly $\binom{9}{5}$ subsets that don't contain any of the yellow bulbs and so $|E^c| = \binom{9}{5}$. This means that

$$P(E) = 1 - P(E^c) = 1 - \frac{\binom{9}{5}}{\binom{15}{5}} = 1 - \frac{9 \times 8 \times \ldots \times 6}{15 \times 14 \times \ldots \times 11}.$$

---

[1]Some of you might have other ways of solving this problem. That's totally fine. For those who still have issues with solving probability problems, I recommend following my proof strategy.

You could also solve the problem by introducing order, and picking 5 bulbs one at a time, without replacement, rather than all 5 at once, and the you would still get the same answer (why? convince yourself). □