LECTURE 2
CONDITIONAL PROBABILITY, INDEPENDENCE, BAYES' RULE

1 Conditional probability

The probability model is concerned with evaluating the likeliness of events. However often when evaluating the likeliness of events, researchers need to incorporate some prior information. Suppose one is interested in the measure of likeliness of an event A when it is known that another event B has occurred. In this situation it seems unwise to use $P(A)$ since that measure ignores the information on the event B. The revised measure that takes into account the additional information (the B has occurred) is called the conditional probability of A given B.

Definition 1. Suppose that events A and B are defined on the same probability space, and the event B is such that $P(B) > 0$. The *conditional probability* of A given that B has occurred is given by $P(A|B) = P(A \cap B)/P(B)$.

To understand the intuition behind the concept of conditional probability, note that if B has occurred then we know that the selected ω belongs to B. Therefore, when evaluating A, one should ignore all ω's that are not in B. Thus, a revised measure of the likeliness of A should depend on $P(A \cap B)$. Intuitively this is equivalent to reducing the effective sample space from Ω to $B \subset \Omega$. However, according to the second axiom of probability, the total probability measure must be equal to one. Hence, we re-scale $P(A \cap B)$ by $P(B)$ to obtain $P(\Omega|B) = P(B|B) = 1$.

Example. Consider rolling a dice example: $\Omega = \{1, 2, \ldots, 6\}$. Let $A = \{6\}$ and $B = \{2, 4, 6\}$. In this example, the unconditional probability of rolling a six is $P(A) = 1/6$. Now suppose you know that a player rolled an even number, however, you do not know what number exactly. In that case, you can update the probability of rolling a six to $P(A|B) = P(A \cap B)/P(B) = P(\{6\})/P(\{2, 4, 6\}) = 1/3$.

Theorem. *The conditional probability function is a probability function, i.e. it satisfies the axioms of probability.*

Proof. Let B be such that $P(B) > 0$. Since $P(\cdot)$ is a probability function and satisfies the axioms of probability, $P(A \cap B) \geq 0$. It follows that $P(A|B) = P(A \cap B)/P(B) \geq 0$. Hence, the first axiom holds.

The second axiom has be already shown to hold above. To verify the third axiom, suppose

that $A_1, A_2, \ldots$ are mutually exclusive. Then, for $i \neq j$

$$
\begin{aligned}
(A_i \cap B) \cap (A_j \cap B) &= A_i \cap A_j \cap B \\
&= \emptyset \cap B \\
&= \emptyset,
\end{aligned}
$$

where the first equality holds by the associative property of the union. Thus, the events $(A_i \cap B)$ are mutually exclusive for $i = 1, 2, \ldots$. Next,

$$
\begin{aligned}
P(A_1 \cup A_2 \cup \ldots | B) &= P((A_1 \cup A_2 \cup \ldots) \cap B)/P(B) \\
&= P((A_1 \cap B) \cup (A_2 \cap B) \cup \ldots)/P(B) \\
&= (P(A_1 \cap B) + P(A_2 \cap B) + \ldots)/P(B) \\
&= P(A_1 \cap B)/P(B) + P(A_2 \cap B)/P(B) + \ldots \\
&= P(A_1|B) + P(A_2|B) + \ldots,
\end{aligned}
$$

where the second equality holds do to the distributive property, the third equality holds since the events $(A_i \cap B)$ are mutually exclusive, and the last equality holds by the definition of conditional probability. $\qquad \square$

Note that once it has been established that conditional probability satisfies the axioms of probability, other properties such as those discussed in Theorem 7 in Lecture 1 follow immediately.

The definition of conditional probability implies that

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

In many cases, the computation of conditional probability is very natural and straightforward, which makes it much easier to calculate the probability of an intersection using the conditional probability approach rather than the counting technique.

Example. (Based on Wonnacott & Wonnacott) Suppose that there are 10 domestic and 10 international students in a class. Two students are selected randomly one after another. What is the probability that they are both international? Let I_1 be the event that the first selected student is international. Let I_2 denote the event that the second selected student is international. We are interested in $P(I_1 \cap I_2)$. Assuming that the selection procedure is purely random, $P(I_1) = 1/2$. Next, consider $P(I_2|I_1)$. Since we condition on I_1, we assume that the first student was international. Thus, after the first round, there are 10 domestic and 9 international students left. Hence, $P(I_2|I_1) = 9/19$, and therefore, $P(I_1 \cap I_2) = 9/38$.

The same answer can be obtained by the counting technique. We can use the permutations

formula or ignore the order of selection and use the combinations formula. Both approaches will give the same answer. The number of possible permutations of choosing 2 students out of 20 is

$$P(20, 2) = \frac{20!}{18!} = 20 \times 19.$$

The number of possible permutations with 2 international students (out of 10 international students) is

$$P(10, 2) = \frac{10!}{8!} = 10 \times 9.$$

Hence, the probability that the two selected students are international is

$$\frac{10 \times 9}{20 \times 19} = \frac{9}{38}.$$

2 Independence

Independence is one of the central concepts in probability and statistics. It will be used repeatedly in this course and its continuations. Intuitively, the independence of events A and B means that the information about B is useless for predicting the occurrence of A:

Definition 2. Suppose that $P(B) > 0$. Two events A and B are *independent* if

$$P(A|B) = P(A).$$

This definition of independence together with the formula for conditional probability in Definition 1, imply:

$$
\begin{aligned}
P(A) &= P(A|B) \\
&= P(A \cap B)/P(B)
\end{aligned}
$$

or

$$P(A \cap B) = P(A)P(B). \tag{1}$$

Suppose that $P(A) > 0$ and $P(B) > 0$. Then (1) implies that $P(A|B) = P(A)$ and $P(B|A) = P(B)$. Thus, equation (1) can be used that as the definition of independence between A and B.

Definition 3. Two events A and B are *independent* if $P(A \cap B) = P(A)P(B)$.

The advantage of Definition (2) is that it clearly shows the intuition behind the concept of independence: conditioning on B does not change the probability of occurrence of A. However,

3

it can only be applied with events B that satisfy $P(B) > 0$. On the other hand, Definition (3) applies to all events including events including events that occur with probability zero. For events that occur with non-zero probabilities, the two definitions are equivalent as they connected by the definition of conditional probability (Definition 1).

Example. Consider two events: {to be a hockey player} and {to live in Canada}. Are they independent?

Example. Can disjoint events be independent? The intuition suggests not: if A and B are disjoint, they have no common outcomes, and therefore if B occurred we can immediately deduce that A did not occur. Hence, the information contained in B can be used to revise the probability of occurrence of A. Nevertheless, disjoint events can be independent if one also considers events that occur with probability zero. We have $P(A \cap B) = P(\emptyset) = 0$. If $P(A) = 0$ or $P(B) = 0$, then $P(A)P(B) = 0$ and, therefore, A and B satisfy the criterion for independence (according to Definition 3). Thus, mutually exclusive events can be independent.

Example. Can an event be independent from itself?

Next, we will discuss certain properties and implications of independence. First, note that independence is a property of the probability function $P(\cdot)$ rather than events themselves. For the same sample space and a collection of events one can define many probability functions. Thus, it is possible that A and B are independent according to $P(\cdot)$, but not independent according to some other probability function Q: i.e $P(A \cap B) = P(A)P(B)$ but $Q(A \cap B) \neq Q(A)Q(B)$.

Theorem. *Suppose that A and B are independent (with respect to some probability function P). Then the following statements are true (with respect to the same probability function P):*
  (a) A^c and B are independent.
  (b) A^c and B^c are independent.
  (c) A and B^c are independent.

Proof. To show (a), we need to evaluate $P(A^c \cap B)$. Recall from Lecture 1 that $B = (A \cap B) \cup (A^c \cap B)$ or $P(B) = P(A \cap B) + P(A^c \cap B)$. After re-arranging the terms, we obtain:

$$
\begin{aligned}
P(A^c \cap B) &= P(B) - P(A \cap B) \\
&= P(B) - P(A)P(B) \\
&= P(B)P(1 - P(A)) \\
&= P(A^c)P(B),
\end{aligned}
$$

where the equality in the second line holds by independence of A and B, and the last equality holds because $P(A^c) = 1 - P(A)$.

Parts (b) and (c) follow immediately from (a). $\qquad\square$

The concept can be extended to more than two events. The following definition applies in the case of three events.

Definition 4. Events A_1, A_2, and A_3 are *mutually independent* if:

(a) $P(A_1 \cap A_2) = P(A_1)P(A_2)$, $P(A_1 \cap A_3) = P(A_1)P(A_3)$, $P(A_2 \cap A_3) = P(A_2)P(A_3)$, and

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

Thus, in the case of three events, they all must be *pairwise independent* (condition (a)), but also the probability of the intersection of the three events should be equal to the product of the individual probabilities.

The concept can be generalized to an arbitrary number of events:

Definition 5. Events $A_1, A_2, A_3, \ldots$ are *mutually independent* if $P(A_{i_1} \cap A_{i_2} \cap \ldots A_{i_j}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_j})$ for any *finite* collection of indices $\{i_1, i_2, \ldots, i_j\}$, $j = 2, 3, \ldots$.

Sometimes experiments consist of multiple stages or trials. We say that the trials are independent if events from different trials are mutually independent. This occurs when the outcome of one trial does not help with predicting the outcomes in other trials

Definition 6. Suppose that an experiment consists of n trials. Let $\mathcal{A}_i$ denote the collection of events associated with trial $i = 1, 2, \ldots, n$. The trials are said to be independent if $P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2) \ldots P(A_n)$ for any combination of $A_1 \in \mathcal{A}_1$, $A_2 \in \mathcal{A}_2$, $\ldots, A_n \in \mathcal{A}_n$.

Example. (Hockey example revisited) Consider the hockey example from Lecture 1. Recall that for a complete lineup one has to select 3 forwards out of 12, 2 defencemen out of 6, and 1 goalie out of 2. Suppose that selection of forwards does not affect selection of defencemen or the goalie, and etc. In other words, suppose that the three stages, selection of forwards, of defencemen, and of the goalie are independent. Then the probability that Higgins (forward) and Bieksa (defencemen) will be in the same lineup is given by

$$\frac{3}{12} \times \frac{2}{6} = \frac{1}{12} \approx 0.083333.$$

The probability that Higgins and Bieksa will both play L on the same shift is

$$\frac{1}{12} \times \frac{1}{6} \approx 0.013889.$$

3   Bayes' rule

Baye's rule allows one to deduce $P(B|A)$ given $P(A|B)$ and the probabilities of A and B. The next theorem presents the simplest version of the rule.

Theorem 7. *Let A and B be two events such that $P(A) > 0$ and $P(B) > 0$. Then*

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

Proof. $P(A|B)P(B)/P(A) = P(A \cap B)/P(A) = P(B|A)$, where the first equality is by the definition of conditional probability. □

Next, we will present a more general version of the result.

Definition 8. A collection of events $\{B_1, B_2, \ldots, B_n\}$ is called a *partition* of the sample space Ω if:

   (a) The events B's are mutually exclusive, i.e. $P(B_i \cap B_j) = \emptyset$ for all $i \neq j$.
   (b) $\cup_{i=1}^n B_i = \Omega$.

The following result is know as *the law of total probability*.

Theorem 9. *Let A be an event on Ω, and let $\{B_1, B_2, \ldots, B_n\}$ be a partition of Ω such that $P(B_i) > 0$ for all $i = 1, 2, \ldots, n$. Then*

$$P(A) = \sum_{i=1}^n P(A \cap B_i) = \sum_{i=1}^n P(A|B_i)P(B_i).$$

Proof. Since B's partition Ω,

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_n) = \cup_{i=1}^n (A \cap B_i).$$

Since B's are mutually exclusive, the events $(A \cap B_i)$ and $(A \cap B_j)$ are also mutually exclusive for $i \neq j$. Hence,

$$\begin{aligned} P(A) &= P(\cup_{i=1}^n (A \cap B_i)) \\ &= \sum_{i=1}^n P(A \cap B_i). \end{aligned}$$

Lastly, by the definition of conditional probability,

$$P(A \cap B_i) = P(A|B_i)P(B_i).$$

□

Now we can state a more general version of Bayes' theorem.

Theorem 10. (Bayes' Theorem) *Let A be an event on Ω, and let $\{B_1, B_2, \ldots, B_n\}$ be a partition of Ω such that $P(B_i) > 0$ for all $i = 1, 2, \ldots, n$. Suppose further that $P(A) > 0$.*

Then

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}$$

for all $i = 1, 2, \ldots, n$.

Proof. By Theorem 9, the denominator is equal to $P(A)$. By the definition of conditional probability, the numerator is equal to $P(B_i \cap A)$. The result follows from the definition of conditional probability. □

Example. Suppose that 30% of Canadians play hockey. Suppose further that 10% of people in the USA and 1% of people in Mexico play hockey, and that for the rest of North American countries and territories the probability of playing hockey is zero. If a randomly selected North American person plays hockey, what is the probability that he is from Canada? What is the probability that he is from the US? Note that Canada has 6% of the population in North America, the USA has 57%, and Mexico has 21%.

Answer: Let A denote hockey, and B_1, B_2, and B_3 denote Canada, USA, and Mexico respectively. Let $B_4, \ldots, B_{41}$ denote the rest of North American countries and territories. We have

$$
\begin{aligned}
P(A|B_1) &= 0.30, \\
P(A|B_2) &= 0.10, \\
P(A|B_3) &= 0.01,
\end{aligned}
$$

and $P(A|B_j) = 0$ for $j = 3, \ldots, 41$. Applying Bayes' Theorem, the probability that a randomly selected North American hockey player is from Canada is

$$P(B_1|A) = \frac{0.30 \times 0.06}{0.30 \times 0.06 + 0.10 \times 0.57 + 0.01 \times 0.21} = \frac{0.018}{0.0771} \approx 0.23.$$

The probability that he is from the US is

$$P(B_2|A) = \frac{0.10 \times 0.57}{0.30 \times 0.06 + 0.10 \times 0.57 + 0.01 \times 0.21} = \frac{0.057}{0.0771} \approx 0.74.$$