Ch2 - Probability Theory - 1

(Probability Definitions, Conditional Probability and Independence)

Statistics For Business and Economics - I

Shaikh Tanvir Hossain

East West University, Dhaka

Last Updated May 1, 2025

Outline

Outline

1. **Random Experiment**

2. **Probability Definitions**

3. **Conditional Probability**

4. **Independence**

Random Experiment

Random Experiment

Probability theory starts from Random Experiment. Here is the definition,

Definition 2.1: (Experiment and Event)

A *random experiment* is any process, real or hypothetical, in which *before performing* the experiment we can identify all possible outcomes but we don't know exactly which outcome will come.

- The *set* of all possible outcomes is called *sample space of the experiment*. We will use the notation ω to denote a single outcome and Ω to denote the sample space, this means $\Omega = \{\omega : \omega \text{ is an outcome of the experiment}\}$
- Any subset of the sample space is called an *event of the experiment*.

Random Experiment

- Note that the definition says before the experiment is performed we know all possible outcomes, but we do not know which outcome will come (so there is a lack of information or uncertainty!).
- Also another important thing, usually we can perform the same experiment more than once. When we perform the experiment a single time, we call it a *trial* of the experiment.
- Sidenote: Here both Ω and ω are Greek letters, see https://en.wikipedia.org/wiki/Omega. This is pronounced as "Oh-may-gaa". Ω is the upper-case and ω is the lower-case

Random Experiment

- Here are some examples of Random Experiment.
  - Tossing a coin. The sample space is $\Omega = \{\mathrm{H}, \mathrm{T}\}$
  - Tossing two coins. The sample space is $\Omega = \{(\mathrm{H}, \mathrm{H}), (\mathrm{H}, \mathrm{T}), (\mathrm{T}, \mathrm{H}), (\mathrm{T}, \mathrm{T}\})$ (use multiplication rule to calculate the total number of possible outcomes)
  - Throwing a die - The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - Throwing two dice - The sample space is $\Omega = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 6), (2, 6), (2, 1), \ldots, (6, 6)\}$ (use multiplication rule to calculate the total number of possible outcomes, here total number of possible outcomes is 36.)

Random Experiment

- Another important example of random experiment is sampling, you already know about this.
- Sampling - The current population of Bangladesh is about $168,000,000$. Suppose we *randomly pick a sample* of 100 people so that it is a "good" representative of the population. This is a random experiment, because we don't know which 100 people will come in our sample, but we know the sample space Ω. It is the set of all people in Bangladesh. The sample in this case is called a *random sample*.
- It is important to note that in Statistics the bigger set from which we take our sample is called *population*. This may or may not mean literally population of a country. This could be something else. It depends upon the what problem we are trying to solve.
- In Statistics we are often interested to know about the population, or some characteristics about the population (for example average income of the population) but what happens is we cannot access to the population, so try to get a random sample and then use that sample to say something about the population (we will see more about this later in our course!).
- We will come back to this. But for now just take the lesson that, *random sampling* is a very very important kind of random experiment. In fact most of the data that we analyze is a result of some kind of *random sampling*.

Random Experiment

Figure 1: Throwing dices, tossing a single coin and sampling from a population, all are examples of random experiment!

- Once we know the sample space Ω, we can actually form different subsets of Ω, and think about different *events*. Recall an *events* is simply a subset of the sample space, so in principle everything that we have learned about Sets could be directly applicable when we are talking about Events.

Random Experiment

- For example, for the coin toss experiment, when the sample space is $\Omega = \{\mathrm{H}, \mathrm{T}\}$, can you think about *all possible events* (think about all possible subsets)? Yes, in this case, the answer is easy, it's the power set,

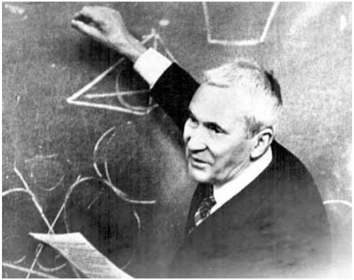
$$\mathcal{P}(\Omega) = \{\{\mathrm{H}\}, \{\mathrm{T}\}, \{\mathrm{H}, \mathrm{T}\}, \emptyset\}$$

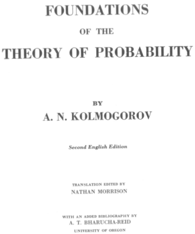
  - $\{\mathrm{H}\}$ is an event since $\{\mathrm{H}\} \subset \Omega$. Here event $\{\mathrm{H}\}$ means only head is appearing.
  - Similarly $\{\mathrm{T}\}$ is an event, it means only tail is appearing.
  - $\{\mathrm{H}, \mathrm{T}\}$ is also an event, since, it satisfies the definition of a subset. Note $\{\mathrm{H}, \mathrm{T}\} = \Omega$
  - Ques- What does the event $\{\mathrm{H}, \mathrm{T}\}$ mean? Ans: It means *any one* of the outcomes will appear, we can write $\{\mathrm{H}, \mathrm{T}\} = \{\mathrm{H}\} \cup \{\mathrm{T}\}$
  - It might look strange why $\emptyset$ is a subset of Ω. The answer is, it satisfies the definition of a subset. Recall, the set A is a subset of the set B if and only if every / all element of A is also an element of B. If A is the empty set then A has no elements and so all of its elements (there are none) belong to B no matter what set B we have. So, the empty set $\emptyset$ is a subset of every set. And in this case $\emptyset \subset \Omega$. Ques- *What does the event $\emptyset$ mean?* Ans: It means, nothing is appearing, so it is an impossible event.

Probability Definitions

Probability Definitions

- Although all of us might have some intuitive understanding of probability, but the history of Mathematics tells us that the modern definition of probability came not so long ago.
- The Russian Mathematician Andrey Nikolaevich Kolmogorov (1903-87) laid the mathematical foundations of probability theory and the theory of randomness.





FOUNDATIONS

OF THE

THEORY OF PROBABILITY

BY

A. N. KOLMOGOROV

Second English Edition

TRANSLATION EDITED BY
NATHAN MORRISON

WITH AN ADDED BIBLIOGRAPHY BY
A. T. BHARUCHA-REID
UNIVERSITY OF OREGON

- His monograph *Grundbegriffe der Wahrscheinlichkeitsrechnung - Foundations of the Theory of Probability**, published in 1933 first introduced the Probability Theory in a rigorous way using fundamental axioms. We will see the *Axiomatic approach* of defining probability later, first let's see the *Classical approach* and *Frequentist approach* of defining probability.

Probability Definitions

Important:

- In all definitions we will calculate probability for events as a function of events such that an event is an input and a probability is the output.
- For example if the set A is an event (i.e., $A \subset \Omega$) then we will calculate the probability $\mathbb{P}(A)$ and this is going to be a number in $[0, 1]$.

Probability Definitions

Definition 2.2: - Classical Definition of Probability

If an experiment has n *equally likely* outcomes, and there is an event A where the number of outcomes is n_A, then the *probability of the event A* is,

$$\mathbb{P}(A) = \frac{n_A}{n} \tag{1}$$

- So when we are thinking about the event A, the classical definition says we can calculate the probability by,

$$\mathbb{P}(A) = \frac{\text{number of outcomes in the event } A}{\text{number of outcomes in the sample space or total number of outcomes}}$$

Probability Definitions

- Let's apply the classical definition and calculate probabilities -

  Example 2.3:
  Suppose our experiment is throwing a *die*, and we want to know that what is the probability that the next outcome will be an even number? Assume the die is a *balanced die*,

  First of all, note that here the word *"balanced"* means all outcomes are *equally likely*, so we can apply the classical definition. The sample space in this case is, $\Omega = \{1, 2, 3, 4, 5, 6\}$, this means $n = 6$. Let A be the event such that *even number appears*, this means

  $$A = \{\omega : \omega \in \Omega \text{ and } \omega \text{ is an even number}\} = \{2, 4, 6\}$$

  We want to calculate $\mathbb{P}(A)$. Here we have three outcomes for the event A so $n_A = 3$, now applying the classical definition we get,

  $$\mathbb{P}(A) = \frac{n_A}{n} = \frac{3}{6} = \frac{1}{2}$$

Example 2.4: Suppose our experiment is tossing 2 coins and assume that all the outcomes are equally likely (i.e., fair coins), now answer the following questions,

- (a) What is the sample space Ω?
- (b) What is the probability of getting at least one head?

Example 2.5:
Now suppose the experiment is tossing 3 coins, and assume that all the outcomes are equally likely (i.e., fair coins), now answer the following questions,

- (a) What is the sample space Ω?
- (b) What is the probability of getting three heads?
- (c) What is the probability of getting exactly one head and two tails?

Example 2.6:(This problem needs the knowledge of permutation or ordering concepts...) Suppose in a city license plates have six characters: 3 letters followed by 3 numbers, now answer the following questions,

- a) How many distinct such plates are possible?
- b) How many distinct plates are possible if the license plate contains no duplicate letters or numbers?
- c) Given that all sequences of six characters are equally likely, what is the probability that a randomly selected license plate for a new car will contain no duplicate letters or numbers?

- Ans of a) We can apply multiplication rule, there are $26^3 = 17,576$ different ways to choose the letters and $10^3 = 1000$ ways to choose the numbers, so we have $26^3 \times 10^3 = 17,576 \times 1000 = 17,576,000$ number of distinct plates. This means Ω consists the set of of all $17,576,000$ possible license plates, so here $n = 17,576,000$

- Ans of b) Let's denote the event with A where we do not have any duplicates with numbers or digits. This means set A has license plates with no duplicate letters or number.

  Now, no duplicate letters means there are $26 \times 25 \times 24 = 15,600$ ways to choose the letters. And then, no duplicate numbers mean there are $10 \times 9 \times 8 = 720$ ways to choose the numbers. From the multiplication principle, the number of outcomes in the event A is $15,600 \times 720 = 11,232,000$, so $n_A = 11,232,000$.

- Ans of c) So now we can calculate the probability of happening the event A,

$$\mathbb{P}(A) = \frac{11,232,000}{17,576,000} = .64$$

Probability Definitions

- We will solve more examples in the practice sheet, now let's discuss the issues with the classical definition.
- There are essentially two major problems with the classical definition of probability
  - Assumption of equally likely outcomes (how do we know this?). For example if we have a biased coin, then how do we calculate probability.
  - Finite sample space issues (sample space can be very large, e.g., $\Omega = \mathbb{R}$)
- Another definition is known as the *Frequency definition* of probability, let's see it now, this is a very important definition and we often use this to explain probability,

Probability Definitions

Definition 2.7: - Frequency Definition of Probability

The probability of an event A is the *relative proportion of outcomes* if we perform the experiment *under identical condition* for a *large number of times*.

- So for example if our experiment is tossing a single coin, the probability of appearing heads is the number of times heads will appear if we perform this experiment almost infinite number of times.
- Any more example...

Probability Definitions

- Frequency definition does not have equally likely outcomes assumption, but the issue is we need to perform the experiment *under identical conditions*, and assume that we are performing many time.....this is often not possible.
- Now we will see the axiomatic way to define probability, which doesn't have these issues. This is the most general way to define probability.
- Important it's an *abstract definition* where we will define probability as a *set function* which will satisfy some rules,
- So we will set the rules that intuitively we think probability should satisfy, ...

Probability Definitions

Definition 2.8: - Axiomatic Definition of Probability

For a random experiment, if we have a sample space Ω and then we can define probability as a *set function* $\mathbb{P}(\cdot)$ such that for any event $A \subset \Omega$

- 1. $\mathbb{P}(A) \geq 0$
- 2. $\mathbb{P}(\Omega) = 1.$
- 3. For *pairwise disjoint but countable number of events* $A_1, A_2, \ldots$ we have

$$\mathbb{P}\left(A_1 \cup A_2 \cup A_3 \ldots\right) = \mathbb{P}\left(A_1\right) + \mathbb{P}\left(A_2\right) + \mathbb{P}\left(A_3\right) + \ldots$$

Sidenote: The idea of disjoint means, $A_1 \cap A_2 = \emptyset$. In this case all possible pairs will be disjoint

- Let's explain each of these axioms (in class discussion).

Probability Definitions

- Note that, unlike the other definition, the Axiomatic definition does not tell us any ways to calculate probabilities, it only defines probability as a function with some rules...
- This means as long as any set function satisfies above three axioms, we will consider that function a probability function. Sometimes Probability function is also called *Probability measure*. We won't talk about *"measure"* in this course, but if some of you take advanced courses in probability theory, you will see this term. In fact probability function is called a measure.

Probability Definitions

- With this definition, now we can show that the following rules of calculating probability

Theorem 2.9: (Probability Calculus or Probability Rules)

$\mathbb{P}(\cdot)$ is a probability function and A and B are any events, then we can show that

- a. $\mathbb{P}(\emptyset) = 0$
- b. $\mathbb{P}(A) \leq 1$
- c. $\mathbb{P}\left(A^c\right) = 1 - \mathbb{P}(A)$
- d. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$;
- e. If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Proof:
c):

First using Venn diagram, we can see that

$$\Omega = A \cup A^c$$

This means A and A^c makes a *partition* of the sample space Ω (What is a partition? It simply means if we take union of disjoint sets will get the whole set) Now we will apply the axioms,

$$
\begin{aligned}
\Omega &= A \cup A^{\mathrm{c}} \\
\Longrightarrow\ \mathbb{P}(\Omega) &= \mathbb{P}\left(A \cup A^{c}\right), \text{ [apply the second axiom]} \\
\Longrightarrow\ \mathbb{P}(\Omega) &= \mathbb{P}(A) + \mathbb{P}\left(A^{\mathrm{c}}\right) \text{ [}A\text{ and }A^{c}\text{ are disjoint, so apply the third axiom]} \\
\Longrightarrow\ 1 &= \mathbb{P}(A) + \mathbb{P}\left(A^{\mathrm{c}}\right) \text{ [Apply first axiom]}
\end{aligned}
$$

So last line means $\mathbb{P}\left(A^{c}\right) = 1 - \mathbb{P}(A)$. So we have shown c

———

b):

Since $0 \leq \mathbb{P}\left(A^{\mathrm{c}}\right) \leq 1$ and we know $1 = \mathbb{P}\left(A^{\mathrm{c}}\right) + \mathbb{P}\left(A\right)$, it means we must have $\mathbb{P}(A) \leq 1$, so this means (b) holds.

———

a):

To prove (a), we use a similar argument like c First note,

$$
\begin{aligned}
\mathbb{P}(\Omega \cup \emptyset) &= \mathbb{P}(\Omega) + \mathbb{P}(\emptyset) \text{ [since } \Omega \text{ and } \emptyset \text{ are disjoint and } \Omega = \Omega \cup \emptyset\text{, we apply third axiom]} \\
\mathbb{P}(\Omega) &= \mathbb{P}(\Omega) + \mathbb{P}(\emptyset) \\
1 &= 1 + \mathbb{P}(\emptyset) \text{ [apply second axiom]}
\end{aligned}
$$

so we have $\mathbb{P}(\emptyset) = 0$.

Probability Definitions

d) and e):

Here we have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

and

$$A \subset B, \text{ then } \mathbb{P}(A) \leq \mathbb{P}(B)$$

We will not prove this in the class, but let's understand them visually. For proof you can look at Casella and Berger (2002) (A bit advanced but a beautiful book!)

$\square$

Probability Definitions

- As a side note here are the formal definitions of disjoint, pairwise disjoint and partition

Definition 2.10: (Disjoint, Pairwise Disjoint and Partition)

- Two events A and B are *disjoint* (or also called *mutually exclusive*) if $A \cap B = \emptyset$.
- The sequence of events $A_1, A_2, \ldots$ are *pairwise disjoint* (or *pairwise mutually exclusive*) if $A_i \cap A_j = \emptyset$ for any $i \neq j$.
- If $A_1, A_2, \ldots$ are pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = \Omega$, then the collection $A_1, A_2, \ldots$ forms a *partition* of Ω.

Partition means it will break the sample space in disjoint parts. These concepts are easy to understand if we draw the Venn Diagrams.

Probability Definitions

- $\mathbb{P}(A \cap B)$ is called the *joint probability*, because this calculates the probability of happening both events. On the other hand $\mathbb{P}(A)$ and $\mathbb{P}(B)$ are called *marginal probabilities*.
- Note that if $\mathbb{P}(A \cap B) = 0$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. But in general we cannot write this, we have to use Theorem 2.9 (d)

Probability Definitions

- When we have a *countable and finite sample space* then there is a nice rule to assign/calculate probability of an event A, following theorem gives us this rule. You have already applied this rule for the equally likely case. But now we don't need "equally likely assumption".

Theorem 2.11: (A rule to assign probabilities of events for a finite sample space)

Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be a finite sample space and let $\mathbb{P}(\{\omega_i\}) = p_i$, for $i = 1, 2, \ldots, n$ such that following two conditions hold

1. $p_i \geq 0$ for all $i = 1, 2, \ldots, n$
2. $\displaystyle\sum_{i=1}^{n} p_i = 1$

If for any event A, we can define $\mathbb{P}(A)$ by

$$\mathbb{P}(A) = \sum_{\{i:\omega_i \in A\}} p_i$$

Also for $\emptyset$ we have $\mathbb{P}(\emptyset) = 0$, then we can show that $\mathbb{P}$ is a probability function (this means all axioms are satisfied).

The above theorem remains true if Ω is a countable set, it means we can apply this theorem when we have $\Omega = \{\omega_1, \omega_2, \ldots\}$

Probability Definitions

- Let's see an application of this theorem. Suppose an experiment has five outcomes: $\omega_1, \omega_2, \omega_3, \omega_4, \omega_5$. Then if we know

$$\begin{aligned}\mathbb{P}(\{\omega_1\}) &= p_1 = 0.2 \\ \mathbb{P}(\{\omega_2\}) &= p_2 = 0.3 \\ \mathbb{P}(\{\omega_3\}) &= p_3 = 0.2 \\ \mathbb{P}(\{\omega_4\}) &= p_4 = 0.1 \\ \mathbb{P}(\{\omega_5\}) &= p_5 = 0.2\end{aligned}$$

  The theorem says we can calculate probabilities for any events. For example, we can calculate $P(\{\omega_1, \omega_2\})$

- First note the sample space is, $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$. Now, let's calculate $P(\{\omega_1, \omega_2\})$. If we apply the theorem we have,

$$P(\{\omega_1, \omega_2\}) = p_1 + p_2 = 0.2 + 0.3 = 0.5$$

- Can you calculate the probability $P(\{\omega_1, \omega_2\})$, if we assume equally likely assumption?

*Go to https://www.kolmogorov.com/Foundations.html to see the scanned version of the English translation.

1. Random Experiment

2. Probability Definitions

3. Conditional Probability

4. Independence

Conditional Probability

Conditional Probability

- Now we will start with an new concept called *Conditioning*.
- *Conditioning is the soul of Statistics* (Joe Blitzstein, Harvard Stat 110)...
- All of the probabilities that we have dealt so far are *unconditional probabilities*. A sample space was defined and all probabilities were calculated with respect to that sample space.
- However in many instances, we have *new information*, then if we calculate the probabilities with updated information, we call it *Conditional Probability*, conditioning on that information....
- Now, when we have new information, ideally we *need to update the sample space*, however in many cases we are not in a position to *update the sample space* and calculate probabilities.
- Hence the need for new tool, where we calculate the probabilities using the original sample space but with new information. Let's see the definition,

Conditional Probability

Definition 2.12: (Conditional Probability)

If A and B are events in Ω, and $\mathbb{P}(B) > 0$, then the *conditional probability* of A given B is defined as

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{2}$$

Conditional Probability

- Let's see an example

  Example 2.13: (Conditional Probability)

- Suppose we have a data of recently admitted 150 students from the four Departments (BBA, EEE, CSE and ECO) of EWU. We also have information whether before taking the admission, they tried to go abroad for undergrad or not (recall this is just a contingency table or cross-tabulation).

	Departments				
	BBA	EEE	CSE	ECO	Total
Tried	18	13	22	24	77
Not Tried	22	25	16	10	73
Total	40	38	38	34	150

Conditional Probability

- Here are couple of questions (For all questions assume, equally likely case, this means all students have same probability of being selected)

  Following questions are for a randomly selected student from the 150 students that we have....

  - a) What is the probability that a randomly selected student had tried abroad?
  - b) What is the probability that a randomly selected student studies at the ECO Department?
  - c) What is the probability that a randomly selected student studies at the ECO Dept. and also had tried abroad?
  - d) *If we know that the student is from ECO*, what is the probability that a randomly selected student had tried abroad?

    Note: In future, often we will omit the phrase *"randomly selected"*... but whenever there is a question about probability, it means we are talking about a random experiment.

Conditional Probability

- Note in this case, the random experiment is selecting a random student from the 150 students
- The sample space is the set of all 150 students
- Now we answer the questions one by one, first let's define
  - Let A be the event that the student had tried abroad.
  - Let B be the event that the student is from the ECO Department.
- The first question asks us to calculate $\mathbb{P}(A)$, the second question asks us to calculate $\mathbb{P}(B)$, the third question asks us to calculate $\mathbb{P}(A \cap B)$, and the fourth question asks us to calculate $\mathbb{P}(A|B)$

Conditional Probability

- We calculate applying classical definition (this is what we will usually do in these problems)

$$\mathbb{P}(A) = \frac{\#\text{ of students who had tried abroad}}{\#\text{ total students}} = \frac{77}{150}$$

$$\mathbb{P}(B) = \frac{\#\text{ of students who studies at ECO}}{\#\text{ total students}} = \frac{34}{150}$$

- So far we calculated the *marginal probabilities* of A and B.
- Now we can also calculate the *joint probability* $\mathbb{P}(A \cap B)$, here $A \cap B$ is the event where a randomly selected student studies at ECO and also had tried abroad.

$$\mathbb{P}(A \cap B) = \frac{24}{150}$$

Conditional Probability

- Now we calculate the *conditional probability*, here we apply the formula for $\mathbb{P}(A|B)$,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{24/150}{34/150} = \frac{24}{34}$$

- Note that there is a difference between the joint event $A \cap B$ and the conditional event $A \mid B$

- Interesting to note is, the conditional probability can also be calculated directly from the cell values using

$$\frac{\#\text{ From the ECO students who had tried abroad}}{\#\text{Total ECO students}} = 24/34$$

this is the calculation with the *updated sample space* that includes only ECO students

- In this case we don't need to apply the formula $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$, we can directly do the calculation as,
- The conditional formula for $\mathbb{P}(A|B)$ is there when we would like to use the original sample space and we want to calculate the conditional probability. ..

Conditional Probability

- We can also calculate $\mathbb{P}(A^c|B) = 1 - \mathbb{P}(A|B) = 10/34$. This is always possible for conditional probability.
- So conditional probability function will act like a probability function. But it will give us an updated calculation with respect to the new sample space.
- So you can say that, the intuition of conditional probability calculation is - our original sample space Ω has been updated to B, and then all further calculations are updated with respect to their relation to B.

Conditional Probability

- **Two Important Questions:**
- **Ques 1**: What happens to conditional probabilities of disjoint sets?
- Ans: Let A and B are disjoint, so $A \cap B = \emptyset$. Note in this case, $\mathbb{P}(A \cap B) = 0$. So $\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A) = 0$. This means conditional probability of disjoint sets is always 0. So conditioning on B will not give any information about A if A and B are disjoint.

- **Ques 2:** When do we have $\mathbb{P}(A \mid B) = \mathbb{P}(A)$
- Ans: Note that this happens when $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$, since

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A) \times \mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$$

- It means conditional and unconditional probabilities are same ... we will come back to this later... actually this happens when A and B are independent events.

Conditional Probability

- Now note that the conditional probability definition gives a way to calculate the probabilities of a joint event, using the formula in (2), we get,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) \tag{3}$$

- This is sometimes called *the multiplication rule of conditional probability* (do not confuse this with the multiplication rule for counting!)
- Now using the same idea in (2) we can also calculate

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \text{ given that } \mathbb{P}(A) \neq 0 \tag{4}$$

- From here using the multiplication idea, we get

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

- So now we have a different way of writing $\mathbb{P}(A|B)$,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \tag{5}$$

- The last formula where we "turned around" the conditional probabilities is called *Bayes' Rule*, this is after the name of Sir Thomas Bayes.

- So the Baye's Rule is

Theorem 2.14: (Bayes' Rule)

Let A and B be two events on the sample space Ω, and assume that $\mathbb{P}(B) > 0$, then we have

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \tag{6}$$

Conditional Probability

- Note for three sets A_1, A_2, A_3 using Conditional Probability we can also calculate,

$$\mathbb{P}(A_3 \mid A_1 \cap A_2) = \frac{\mathbb{P}(A_1 \cap A_2 \cap A_3)}{\mathbb{P}(A_1 \cap A_2)}$$

- Using the multiplication rule of conditional probability, we get

$$\begin{aligned}\mathbb{P}(A_1 \cap A_2 \cap A_3) &= \mathbb{P}(A_1 \cap A_2) \times \mathbb{P}(A_3 \mid A_1 \cap A_2) \\ &= \mathbb{P}(A_1) \times \mathbb{P}(A_2|A_1) \times \mathbb{P}(A_3 \mid A_1 \cap A_2)\end{aligned}$$

- You can extend this formula for more than 2 events, but I am skipping the general version, see DeGroot and Schervish (2012) for details.

- Now we will learn another law, which is called *Law of Total Probability*. This law is very important and it is an application of *partition*.

Theorem 2.15: (Law of Total Probability)

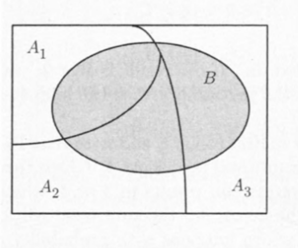
Let $A_1, \ldots, A_n$ be events that form a partition of the sample space Ω and assume that $\mathbb{P}(A_i) > 0$, for all i. Then, for any event B, we have

$$\mathbb{P}(B) = \mathbb{P}(B \mid A_1)\,\mathbb{P}(A_1) + \cdots + \mathbb{P}(B \mid A_n)\,\mathbb{P}(A_n) \tag{7}$$

- We will skip the general proof, but let's understand the theorem for a simpler case.

Conditional Probability

- Consider the following Venn diagram,



- Here A_1, A_2 and A_3 forms a *partition*. Recall a partition is a sequence of sets which splits the entire sample space.
- If A_1, A_2 and A_3 forms a partition of Ω and B is a set, then we can write,

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B) \tag{8}$$

- All sets are disjoint, so using the axiom we have,

$$\mathbb{P}(B) = \mathbb{P}(A_1 \cap B) + \mathbb{P}(A_2 \cap B) + \mathbb{P}(A_3 \cap B)$$

Conditional Probability

- Now using conditional probabilities we have

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(A_1 \cap B) + \mathbb{P}(A_2 \cap B) + \mathbb{P}(A_3 \cap B) \\ &= \mathbb{P}(B \mid A_1)\,\mathbb{P}(A_1) + \mathbb{P}(B \mid A_2)\,\mathbb{P}(A_2) + \mathbb{P}(B \mid A_3)\,\mathbb{P}(A_3)\end{aligned}$$

- So this is the Law of Total Probability, but we explained it for three sets. You can extend the idea generally for n sets,

$$\mathbb{P}(B) = \mathbb{P}(B \mid A_1)\mathbb{P}(A_1) + \cdots + \mathbb{P}(B \mid A_n)\mathbb{P}(A_n)$$

- It is possible to apply the Bayes' rule here. First note applying simple Bayes' rule for set A_1 and B we get,

$$\mathbb{P}\left(A_1 \mid B\right) = \frac{\mathbb{P}\left(A_1\right)\mathbb{P}\left(B \mid A_1\right)}{\mathbb{P}(B)}$$

- Now we apply the law of total probability for $\mathbb{P}(B)$

$$\begin{aligned}\mathbb{P}\left(A_1 \mid B\right) &= \frac{\mathbb{P}\left(A_1\right)\mathbb{P}\left(B \mid A_1\right)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}\left(A_1\right)\mathbb{P}\left(B \mid A_1\right)}{\mathbb{P}\left(B \mid A_1\right)\mathbb{P}\left(A_1\right) + \cdots + \mathbb{P}\left(B \mid A_n\right)\mathbb{P}\left(A_n\right)}\end{aligned}$$

- This is what we call the general version of the Bayes' rule or the Bayes' rule with law of total probability.
- Now we write the general version.

Theorem 2.16: (Bayes' Rule with Law of Total Probability)

Let $A_1, A_2, \ldots, A_n$ be events that form a partition of the sample space Ω, and assume that $\mathbb{P}(A_i) > 0$, for all i. Then, for any event B such that $\mathbb{P}(B) > 0$, we have

$$\begin{aligned}\mathbb{P}(A_i \mid B) &= \frac{\mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(A_i)\,\mathbb{P}(B \mid A_i)}{\mathbb{P}(B \mid A_1)\,\mathbb{P}(A_1) + \cdots + \mathbb{P}(B \mid A_n)\,\mathbb{P}(A_n)}\end{aligned}$$

Independence

Independence

- The idea of independent events is very easy, we need to check whether joint probability is same as the product of marginal probability.

Theorem 2.17: (Independence of two events)

Let A and B be two events from the sample space Ω, we say A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$$

- Here is one interpretation, since multiplication of probabilities will always be smaller, you can think when the events are independent than their joint probability will be very small.
- The idea of independence can also be explained via the conditional probability. This is what we mentioned in page 77. Recall, when A and B are independent we have

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A) \times \mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$$

- So the unconditional probability or the marginal probability of A is same as the conditional probability. So learning B makes no difference to the probability of A.

Independence

- We can easily extend this concept to more than two events, the idea is then we need all subsets of the events to be independent.
- As an example, in order for three events A, B, and C to be independent, the following four relations must be satisfied:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$
$$\mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C)$$
$$\mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C)$$

  and

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$$

- This idea is what we call *mutual independence*.

Definition 2.18: (Independent Events).

The k events $A_1, \ldots, A_k$ are independent (or mutually independent) if, for every subset the joint probability of the events can be written as a product of marginal probabilities.

- You will see some problems in the problem set.

References

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Thomson Learning, Australia ; Pacific Grove, CA, 2nd edition.

DeGroot, M. H. and Schervish, M. J. (2012). *Probability and Statistics*. Addison-Wesley, Boston, 4th edition.