

Chapter 2

Preliminaries

“I could be bounded in a nutshell and count myself king of infinite space.” -Hamlet Act II, Scene ii

In this Chapter we will first cover the basic terminology and discuss the naïve definitions of probability. Then we will cover some basic counting methods. We will next give the formal definition of probability. We will prove some basic properties and cover the inclusion-exclusion principle. Finally, we will discuss the important concept of conditional probability.

2.1 Fundamental Terms

An **experiment** is some action with distinct **outcomes**. Each of these outcomes occurs with some **probability**.

Sample Space: A sample space denoted by S is the set of all possible outcomes of an experiment.

Note 1: The term *experiment* has a very general interpretation. It could be tossing a coin or measuring the protein level in a blood sample to determine malignancy.

Non-determinism: While we know the sample space, i.e., all the outcomes that comprises the sample space, for a specific experiment we do not know what will be the outcome. When we measure the protein level in the blood, we know the range within which the measured value could lie but we do not know what will be specific value for a given blood sample. **Event:** An event is a subset of the sample space, and as such, it is a set of outcomes. All the typical properties of set theory still apply here. Note that an event could be one outcome or a set of outcomes. Don't confuse the idea of “event” with experiment: one experiment could have multiple events apply to its outcome. For example, when drawing a single card, the card could be a Red Card and a Number Card. Both of these could be represented by different events resulting from the same physical experiment.

Figure 2.1 shows a sample space S . The sample space consists of outcomes represented by the dots. It shows 2 events, Event A and Event B . Note that an event can be a single outcome as it is the case for Event B .

The sample space can be finite, countably infinite, or uncountably infinite. Countably infinite corresponds to the set of integers and uncountably infinite corresponds to the real numbers. We consider examples of each type.

2.1.1 Finite Sample Space

Example: Tossing a Coin What is the sample space for the experiment of tossing a fair coin?

Solution: The sample space $S = \{H, T\}$ where H denotes a Head and T denotes a Tail.

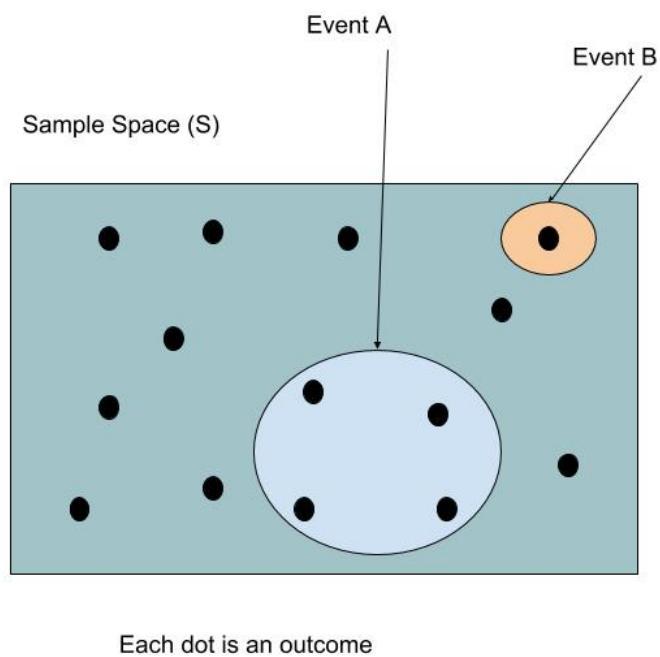


Figure 2.1: A sample space s consisting of outcomes and events. The figure shows a sample space consisting of discrete and finite set of outcomes. Event A consists of 1 outcome. Event B consists of 4 outcomes.

Example: Rolling Dice What is the samples space for rolling 2 dice of different colors?

Solution: The sample space is 36 pairs of values, i.e., $S = \{(6, 6), (6, 5), \dots (1, 1)\}$. Because the dice are different colors, we consider outcomes as ordered pairs. As a result, outcomes (5, 6) and (6, 5) are different.

2.1.2 Countably Infinite Sample Space

Example: The Drowsy Programmer Program code written by a student is compiled. Let C represent the outcome that the program compiles successfully and let E represent the outcome that the compiler encounters an error. Suppose an experiment consists of selecting and compiling programs until a program is encountered that compiles on the first run.

1. Define the sample space.

Solution: The sample space is a countably infinite set $S = \{C, EC, EEC, EEE, \dots\}$

2. What is the event that exactly one, three, or five programs are examined?

Solution: Let A be the event that exactly one, three, or five programs are examined. Then $A = \{C, EEC, EEEEC\}$.

Recall that we are assuming that there is a countably infinite number of program code files. We could also state the problem in which there is a single program code which is compiled. If it fails, the programmer makes changes to the code and then compiles again. This is repeated until the program successfully compiles. Theoretically, this can go on forever and hence countably infinite. Practically, the programmer should retake the appropriate programming language course if it goes on for too long!!

2.1.3 Uncountably Infinite Sample Space

Example: Stick Cutting Suppose that two points (x_1 and x_2) are marked (in some nondeterministic way) on a stick of length 1 meter.

1. Define the sample space for this experiment.

Since the sample space is portion of the two dimensional real plane, we know it is uncountably infinite. This will be the case for just about any infinitely variable real quantity.

Solution: Let x_1 and x_2 denote the two points. Since x_1 and x_2 are marked randomly, sample space S is given by

$$S = \{x_1, x_2 : 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1\}$$

2. Identify the event, "The second point is to the left of the first point."

Solution: Let G denote the event: "The second point is to the left of the first point." Then the event G is given by

$$G = \{x_1, x_2 \in S \text{ s.t. } x_1 > x_2\}$$

3. Suppose that the stick is cut at the marked points. Identify the event, "A triangle can be formed with the resulting three pieces."

Solution: Let Z denote the event that "A triangle can be formed with the resulting three pieces." The first thing to note is that in any triangle the sum of any two sides must be greater than the third side. For our stick problem this implies that none of the three pieces should be greater than 0.5 meters. There can be two cases: 1) $x_1 > x_2$ and 2) $x_1 \leq x_2$.

Let's consider one case (intuitively, the other case will be symmetrical). If $x_1 > x_2$ then the three pieces will be of lengths x_2 , $x_1 - x_2$, and $1 - x_1$. Due to the constraints each of these lengths must be less than 0.5. Thus, if we define 3 events A, B, and C

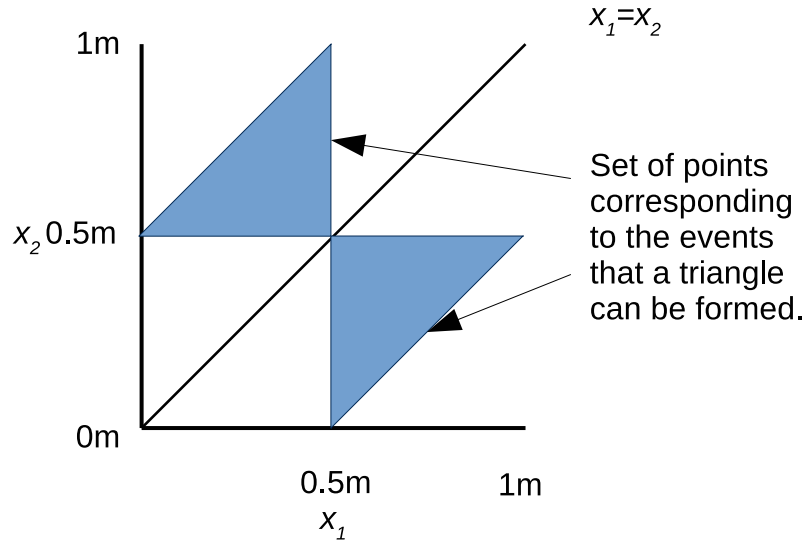


Figure 2.2: The sample space and the event for which the triangle can be formed.

$$\begin{aligned} A &: x_2 < 0.5 \\ B &: x_1 - x_2 < 0.5 \\ C &: 1 - x_1 < 0.5 \end{aligned}$$

Then, the event that the three pieces will form a triangle is $A \cap B \cap C$. This is shown by the shaded area below the 45 degree line ($x_1 = x_2$) in Figure 2.2.

Note that the case when $x_1 \leq x_2$ is symmetrical to the case $x_1 > x_2$. We can define events D, E and F which defined as follows

$$\begin{aligned} D &: x_1 < 0.5 \\ E &: x_2 - x_1 < 0.5 \\ F &: 1 - x_2 < 0.5 \end{aligned}$$

and for this case, the shaded area above the 45 degree line in Figure 2.2 shows the set of points for which the triangle can be formed.

In terms of the algebra of events, if we let Z denotes the set of points that can form a triangle, then

$$Z = (A \cap B \cap C) \cup (D \cap E \cap F)$$

2.2 Naïve Definition Of Probability

If A is an event in the sample space S , then following is the naïve definition for the probability of event A , denoted by $P(A)$,

$$P(A) = \frac{\text{Number of outcomes favorable to event } A}{\text{Total number of outcomes}} \quad (2.1)$$

Example: Double Coin Toss

1. What is the sample space S for flipping a coin twice?

Solution:

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

2. What is the probability of the event $A = \text{Both Tails}$ using the naïve definition of probability?

Solution:

$$P(A) = \frac{1}{4}$$

There are two significant limitations of this naïve definition for general use of calculating probability of events. The naïve definition of probability is only applicable for

1. finite sample space and
2. equally likely outcomes.

Despite the limitations, many problems can be solved using this naïve definition. And do not hesitate to use this wherever applicable.

Example: The Birthday Paradox

We will discuss the Birthday Paradox. See Chapter 3.

2.3 Basic Counting Techniques

In order to be able to apply the naïve definition Equation @ref(eq:naïve-definition) we need to be able to count outcomes. For this we now look at some basic counting rules.

2.3.1 Multiplication Rule

Consider a sequence of r experiments E_1, E_2, \dots, E_r . If experiment E_1 has n_1 possible outcomes and for each outcome of E_1 , there are n_2 outcomes of E_2 , and for each outcome of E_2 , there are n_3 outcomes of E_3 , ..., for each out of E_{r-1} there are n_r outcomes of E_r , then there are $n_1 \times n_2 \times \dots \times n_r$ total possible outcomes.

Example: Choosing Coffee Customers of a café can customize their coffee in a number of different ways by selecting the type of coffee and the type of milk. We are considering only two features. In fact there are many more. Suppose there are two types of coffee caffeinated (caf), decaffeinated (decaf) and three types milk 1 percent, 2 percent, soy. The total number of outcomes (types of coffee) is $2 \times 3 = 6$. In writing so we consider that the customer first chooses the type of coffee E_1 and then the type of milk E_2 . This is shown in The solution would still be the same if we reversed the order.

From Figure 2.3, we can see that as the number of experiments increase, growth in the number of outcomes is very fast (exponential).

Example: Generating Strings How many 5 letter strings can we generate with the 26 letters of English alphabet?

Solution: The first letter in the string can be one of 26. For each of the first letter, the second letter can be one of 26, and so on. Usig the multiplication rule, we have 26^5 .

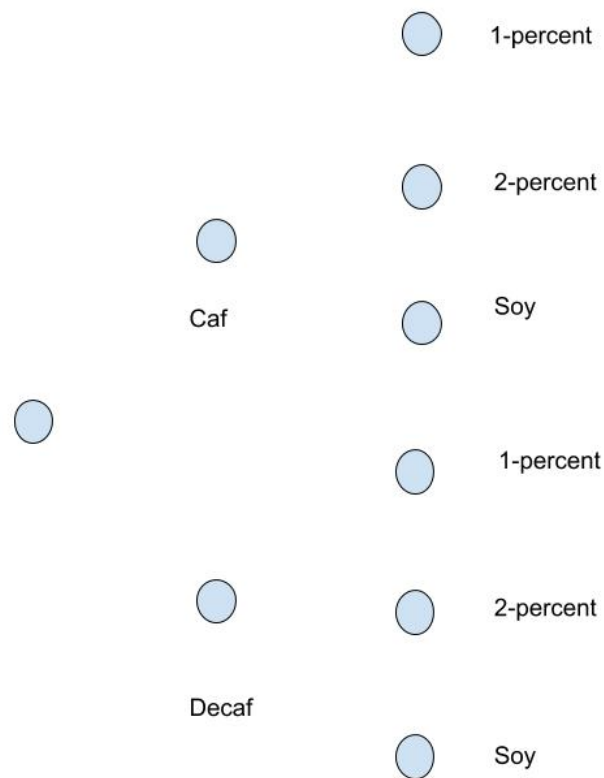


Figure 2.3: Example of multiplication rule in terms of the number of different types of coffee requested by customers.

2.3.2 Multinomial Coefficients

A set of n distinct objects is to be divided into r distinct groups of sizes n_1, n_2, \dots, n_r such that $\sum_{i=1}^r n_i = n$. The total number of ways T is given by

$$T = \underbrace{\binom{n}{n_1}}_{\text{no. ways of choosing 1st group}} \underbrace{\binom{n-n_1}{n_2}}_{\text{no. ways of choosing 2nd group}} \dots \underbrace{\binom{n-n_1-\dots-n_{r-1}}{n_r}}_{\text{no. ways of choosing rth group}} \quad (2.2)$$

$$= \frac{n!}{n_1!n_2!\dots n_r!} \quad (2.3)$$

Example: Bridge In the game of bridge, how many ways can 52 cards be divided into 4 equal parts?

Solution: Using the multinomial theorem we have:

$$T = \frac{52!}{13!13!13!13!}$$

Example: Choosing Library Pages The library employs 10 undergraduate students for different functions like shelving books, managing the checkout desk, help desk. The library has decide to have 5 students shelving books, 3 students working fulltime at the book checkout desk, and 2 students at the help desk. How many divisions (T) of the 10 students into the three possible groups are possible.

Solution: We can apply the multinomial coefficients

$$\begin{aligned} T &= \frac{10!}{5!3!2!} \\ &= 252 \end{aligned}$$

2.3.2.1 Notes

1. Consider $(x_1 + x_2 + \dots + x_r)^n$. The multinomial theorem states that

$$(x_1 + x_2 + \dots + x_r)^n = \sum_{(k_1, k_2, \dots, k_r): n_1 + n_2 + \dots + n_r} \binom{n}{n_1, n_2, \dots, n_r} x_1^{n_1} x_2^{n_2} \dots x_r^{n_r}$$

The sum is taken over all possible values of (n_1, n_2, \dots, n_r) such that their sum is equal to n . The coefficient of the different terms of the expression are the multinomial coefficients.

2. If we consider 2 groups with size n_1, n_2 , then $n_2 = n - n_1$. Then the total number of ways T

$$\begin{aligned} T &= \frac{n!}{n_1!n_2!} \\ &= \frac{n!}{n_1!n - n_1!} \\ &= \binom{n}{n_1} \end{aligned}$$

which are the coefficients of the Binomial Theorem.

Example: Arranging Balls

How many linear arrangements (T) are possible of n balls of which r balls are black and $n - r$ balls are white?

Solution: There are n positions. If we choose r of those positions to place black balls, the remaining $n - r$ positions will have white balls. Thus

$$T = \binom{n}{r}$$

2.3.3 Sampling

Suppose there are n distinguishable objects. We want to find the number of ways we can pick k objects out of n . We can consider two additional features:

- a) whether order matters or not and
- b) with or without replacement.

Let us first understand these two features: When we say *Order Does Not Matter*, what this means is that different permutations of the same k objects are not considered to be different outcomes. For example, ABC, ACB, BAC, BCA, BAC, CAB, CBA are not considered to be different outcomes. Whereas if Order Matters they are different outcomes. As an aside, we know that if there are k distinguishable objects, there are $k!$ different permutations/arrangements.

When we say pick *Without Replacement*, what we mean is that when an object is chosen it is put aside and the next object is chosen from the remaining objects. On the other hand, when we replace the object before picking again, we sample *With Replacement*.

Given the above two features, we have the following 4 cases:

2.3.3.1 Case 1: Order matters, With replacement

In this case the first object can be chosen in n ways and since we are replacing the object, the second object can also be chosen in n ways. Using the multiplication rule we have a total of n^k outcomes.

Example: Generating Airline Reservation Identifiers

A domestic airline wishes to generate reservation identifiers for its new “speak to look up” system. The airline previously used 6-character boarding pass numbers with all the letters of the alphabet except ‘O,’ and all 10 unique digits. They want to move to a system with just 3 words. The words will be chosen from 3000 commonly-used English words. Would the proposed system provide the same amount or more reservation identifiers compared to the old system?

Solution:

1. The old system generated 35^6 or 1,838,265,625 identifiers.
2. The new system would generate (3000^3) or 27 billion identifiers, so the new system would be more than adequate.

2.3.3.2 Case 2: Order matters, Without Replacement

This is similar to Case 1, except that once we pick the first object in n ways, the second object can be picked in $n - 1$ ways since the first object is not replaced. Hence, the total number of outcomes in picking k objects out of n is $n(n - 1) \dots (n - k + 1)$.

2.3.3.3 Case 3: Order does not matter, With Replacement

This is not so straightforward to determine but it can be shown that the total number of outcomes is $\binom{n+k-1}{k}$. In order to prove this, the problem can be cast into the problem discussed in the following example.

Example: RAID Array Failure Modes

Consider a set of n disks in RAID array. Suppose that m have become defective and hence $n - m$ are still operational. Assume that the defective and functional disks are indistinguishable. How many linear orderings are there in which no two defectives are consecutive? (Why is this important? The manner in which data is replicated and stored in the disks, failure of two adjacent disks will result in data loss.)

Solution:

Consider that $n - m$ functional disks are arranged in a linear order. Since it is required that no two defective disks can be consecutive, then the positions of the m defective disks must be chosen from $n - m + 1$ positions. There are $n - m - 1$ positions in between the defective disks and 2 positions at the end.

Thus, the positions of m defective disks can be chosen from $n - m + 1$ positions which gives the total number of orderings T as

$$T = \binom{n - m + 1}{m}$$

2.3.3.4 Case 4. Order does not matter, Without Replacement

This follows from Case 1. Since, for each specific choice of k objects, there are $k!$ different permutations, we need to scale down the number of outcomes (T) in Case 1 by $k!$. This is easily shown to be $\binom{n}{k}$.

$$\begin{aligned} T &= \frac{n(n-1) \dots (n-k+1)}{k!} \\ &= \frac{n(n-1) \dots (n-k+1)}{k!} \times \frac{(n-k)!}{(n-k)!} \\ &= \binom{n}{k} \end{aligned}$$

Example: A person has 12 friends and will invite 7 to a party. However, Alice and Bob are feuding, and refuse to go together.

1. How many ways are there to choose the friends to invite to the party such that either Alice or Bob is not invited?
2. How many choices of invitations are possible if Alice and Betty say that neither of them will go if one of them isn't invited?

Solution:

1. This is choosing without replacement. First, we enumerate all the possible choices, which would be $\binom{12}{7}$. Second, we can subtract the cases in which they would both be going, which is $\binom{10}{5}$, since we've removed Alice and Bob from the sampling. This gives us:

$$\begin{aligned} T &= \binom{12}{7} - \binom{10}{7} \\ &= \frac{12!}{7!5!} - \frac{10!}{5!5!} \end{aligned}$$

2. This is also choosing without replacement. However, we will now add the sample spaces of the two different cases. If they were both invited, then it would be $\binom{10}{5}$. If neither of them are invited, then we still have to choose 7, but from a smaller set of 10, yielding $\binom{10}{7}$. This gives us:

$$\begin{aligned}
T &= \binom{10}{5} + \binom{10}{7} \\
&= \frac{10!}{5!5!} + \frac{10!}{7!3!}
\end{aligned}$$

2.4 Formal Definition of Probability

Give the probability space consisting of P and S where S is the sample space and P is a function that takes as an input an event A and returns $P(A) \in [0, 1]$ such that

1. [Axiom 1:] $P(\Phi) = 0$ 2. [Axiom 2:] $P(S) = 1$, 3. [Axiom 3:] If A_1, A_2, \dots , are disjoint non-overlapping events then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

2.4.1 Properties

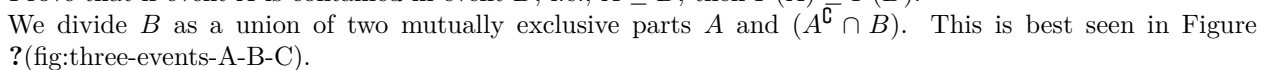
1. Property 1: If A^c denote the complement of event A , then $P(A^c) = 1 - P(A)$.
2. Property 2: If event A is contained in event B , i.e., $A \subseteq B$, then $P(A) \leq P(B)$
3. Property 3: Consider two events A, B , then

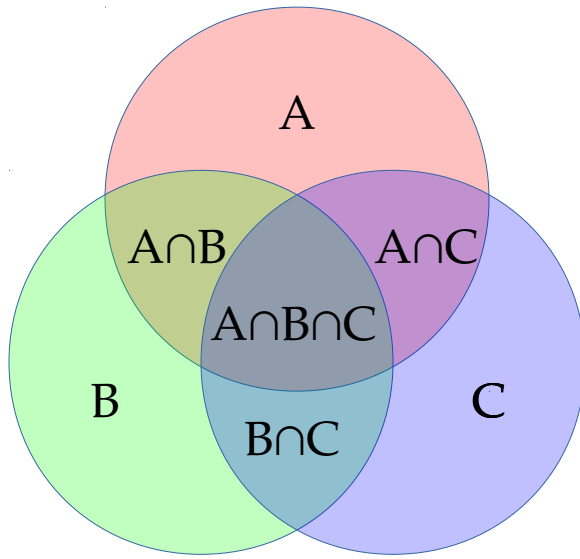
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

All of the above properties can be proven from the axioms. Below we show the proof of Property 2. The proofs of the other two properties are left as exercises.

2.4.1.1 Proof of Property 2

Prove that if event A is contained in event B , i.e., $A \subseteq B$, then $P(A) \leq P(B)$.

We divide B as a union of two mutually exclusive parts A and $(A^c \cap B)$. This is best seen in Figure  (fig:three-events-A-B-C).



Thus,

$$B = A \cup (A^c \cap B)$$

which implies that

$$P(B) = P(A) + P(A^c \cap B)$$

since $P(A^c \cap B) \geq 0$, $P(B) \geq P(A)$

2.5 Inclusion-Exclusion Principle

First let us consider three events A_1, A_2 , and A_3 . Consider a figure similar to that shown in Figure ?? with A replaced by A_1 , B by A_2 and C by A_3 .

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) \\ &- P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\ &+ P(A_1 \cap A_2 \cap A_3) \end{aligned}$$

1. When we add (include) $P(A_1) + P(A_2) + P(A_3)$ we have added the pairwise intersection area once too many. So in the second line we subtract (exclude). But in doing so we subtracted $A \cap B \cap C$ out completely. So we add (include) $A \cap B \cap C$. Due to the alternating addition and subtraction, the process is called the inclusion-exclusion principle.
2. Collectively, the terms in the second line can be written as $\sum_{i < j} P(A_i \cap A_j)$ for $i \in \{1, 2, 3\}$ and $j \in \{1, 2, 3\}$.
3. Also, note that the cardinality of the set $\{A_i \cap A_j\}$ such that $i < j$ is $\binom{3}{2}$.

Now we consider the general case. If A_1, A_2, \dots, A_n are n events, then

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3 \cup \dots \cup A_n) &= \sum_{i=1}^{i=n} P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} A_{i_2}) \\ &+ (-1)^{r+1} \sum_{i_1 < i_2 < \dots < i_r} P(A_{i_1} A_{i_2} \dots A_{i_r}) + \dots \\ &+ (-1)^{n+1} P(A_1 A_2 A_3 \dots A_n) \end{aligned}$$

where $i_1 < i_2 < \dots < i_r$ refers to all possible r -wise combinations of $(1, 2, \dots, n)$. What the above equation states is that the probability of the union of n events is equal to the sum of the probabilities of the events taken one at a time, minus the probability of the events taken two at a time, plus the probability of the events taken three at a time, and so on.

2.6 Proof of the Inclusion-Exclusion Principle

The inclusion-exclusion identity can be proved by induction. Below we show one step of the induction. In particular, assuming that the equation is true for $n = 2$, we derive the expression for $n = 3$. That is derive the expression for $P(A_1 \cup A_2 \cup A_3)$ given the $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$. In doing this step we will convince ourselves that the above general Equation (2.4) is true.

We are given that for any two events E_1 and E_2 , $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 E_2)$. Now consider three events E_1 , E_2 , and E_3 . $E_1 \cup E_2 \cup E_3$ can be written as $E_{12} \cup E_3$ where E_{12} is $E_1 \cup E_2$.

$$P(E_1 \cup E_2 \cup E_3) = P(E_{12} \cup E_3) \tag{2.4}$$

$$= P(E_{12}) + P(E_3) - P(E_{12} E_3) \tag{2.5}$$

$$= P(E_1) + P(E_2) - P(E_1 E_2) + P(E_3) - P(E_{12} E_3) \tag{2.6}$$

Now, $E_{12} E_3 = (E_1 \cup E_2) \cap E_3$ which using the distributive law is $(E_1 \cap E_3) \cup (E_2 \cap E_3)$. Now we can use the inclusion-exclusion identity for two events to obtain the following

$$\begin{aligned} P(E_{12} E_3) &= P((E_1 \cap E_3) \cup (E_2 \cap E_3)) \\ &= P(E_1 E_3) + P(E_2 E_3) - P(E_1 E_3 E_2 E_3) \\ &= P(E_1 E_3) + P(E_2 E_3) - P(E_1 E_3 E_2) \end{aligned}$$

The last step is because $E_3 \cap E_3 = E_3$. Now replace $P(E_{12} E_3)$ in Equation (2.6) to obtain the final result.

Example: Network Receive Buffer A server has 10 incoming packet flows (flows are distinct and uniquely identified by a 5-tuple flow-id). Packets from each flow are stored in a buffer. Suppose there are two packets from each flow (packets have sequence numbers and hence are distinguishable) which are randomly assigned to locations in the buffer of size 20. What is the probability that no two packets from the same flow are in adjacent locations in the buffer?

Solution: Let $E_i, i = 1, 2, 3, \dots, 10$, be the event that packets from the i th flow are placed in adjacent locations in the buffer. Then the required probability α is given by

$$\alpha = 1 - P\left(\bigcup_{i=1}^{10} E_i\right)$$

Using the Inclusion-Exclusion principle, we have

$$P\left(\bigcup_{i=1}^{10} E_i\right) = \sum_1^{10} P(E_i) + \dots + (-1)^n \sum_{i_1 < i_2 < \dots < i_n} P(E_{i_1} E_{i_2} \dots E_{i_n}) + \dots - P(E_1 E_2 \dots E_{10})$$

Now we have to find the individual probabilities, i.e., $P(E_{i_1} E_{i_2} \dots E_{i_n})$ which is the probability that packets from n flows are placed adjacent to each other. Now the total number of ways in which 20 packets can be arranged in the buffer locations is $20!$ and these are equally likely outcomes. If we consider the packets from the flows that are adjacent as single elements, then the total number of possible orderings is $(20 - n)!$. Thus,

$$P(E_{i_1} E_{i_2} \dots E_{i_n}) = \binom{10}{n} 2^n \frac{(20 - n)!}{20!}$$

The first term in the RHS is the number of ways we can choose the n flows. The second term comes from the fact that there are two ways in arranging the two packets in each flow that are together. Thus we get

$$\begin{aligned} P\left(\bigcup_{i=1}^{10} E_i\right) &= \binom{10}{1} 2^1 \frac{19!}{20!} - \binom{10}{2} 2^2 \frac{18!}{20!} + \binom{10}{3} 2^3 \frac{17!}{20!} - \dots - \binom{10}{10} 2^{10} \frac{10!}{20!} \\ &\approx 0.6416608 \end{aligned}$$

```
sum = 0
n = 10
for (i in 1:n)
{
    sum = sum + (-1)^(i+1)*(factorial(n)/(factorial(n - i)*factorial(i)))*(2^i)*factorial(2*n - i)/factorial(20)
}
print(sum)
```

```
## [1] 0.6416608
```

Thus, $\alpha = 1 - 0.6416608 = 0.3583392$

Example: The Matching Problem We will discuss the Matching Problem in Chapter 3 (Classic Problems)

2.7 Conditional Probability

We started with the naïve definition of probability and solved some challenging and important problems. We discussed the limitations of the naïve definition and then gave a formal definition of probability. Now we take the next step to increase the power of probabilistic modeling. In particular, we discuss how we update the evaluation of the uncertainty (risk) when we are given new data. In a sense this is what we as humans are constantly doing—re-evaluating risk based on data that we sense from the environment. Conditional probability is also considered the foundation for statistics.

Definition of Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad P(B) > 0$$

In Figure 2.4 given that event B has occurred the relevant portion of event A is $A \cap B$. Given that B has occurred, outcomes that are not in B (B^c) are no longer relevant. The sample space has now collapsed to B since only outcomes in B are possible.

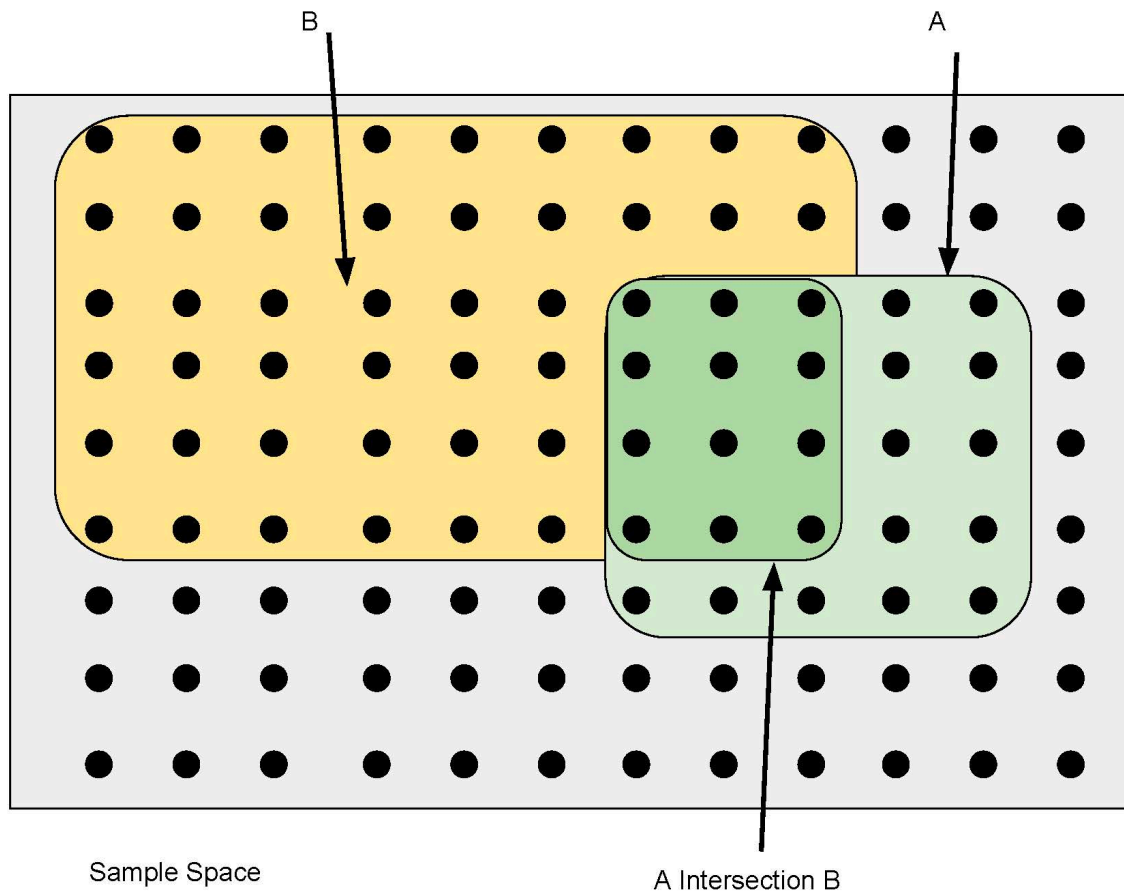


Figure 2.4: Conditional Probability illustrated using a discrete sample space. Note that once B has occurred, the sample space collapses to B .

2.7.1 Few Theorems

1. Theorem 1

$$P(A \cap B) = P(B) \times P(A|B) \quad (2.7)$$

$$= P(A) \times P(B|A) \quad (2.8)$$

2. Multiplication Rule

$$P(A \cap B \cap C) = P(A) \times P(B|A) \times P(C|A \cap B)$$

3. Bayes' Rule

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

$P(A|B)$ is called the posterior probability of A and $P(A)$ is called the prior. Note that $P(B)$ can also be written as

$$P(B) = P(A) \times P(B|A) + P(A^c) \times P(B|A^c) \quad (2.9)$$

4. The Law of Total Probability: If A_1, A_2, \dots, A_n are mutually exclusive and exhaustive events, then

$$P(B) = P(B|A_1) \times P(A_1) + P(B|A_2) \times P(A_2) + \dots + P(B|A_n) \times P(A_n)$$

The above theorems can be proved using the axiom of probability and the definition of conditional probability.

2.7.2 Examples

Example: Rolling Two Dice

Two fair six-sided dice are rolled. What is the conditional probability that one of them lands on a 6, given that the dice land on different numbers?

The restricted sample space consists of 30 outcomes. $S = \{(i, j), i \neq j, i, j \in \{1, 2, 3, 4, 5, 6\}\}$. This can be seen by considering how many outcomes have both dies equal – there are only 6. There are then ten outcomes in which at least one lands on 6. Thus, the required probability is $1/3$.

Example: Testing for a Rare Disease

Patients get tested for a disease that affects 1% of the population. The test is 95% accurate. If a patient tests positive what is the probability that the patient has the disease?

A common (wrong) answer is 95%.

Let

1. T : event that a patient tests positive
2. D : event that a patient has the disease

We will assume that the test is 95% accurate implies that

$$P(T|D) = 0.95$$

$$P(T^c|D) = 0.05$$

$$P(T^c|D^c) = 0.95$$

$$P(T|D^c) = 0.05$$

The above is a formal re-statement of the properties of the test. Recall that “95%” accurate implies that the test is 95% accurate in both directions—for testing the presence or absence of the disease. Hence, in this scenario, $P(T|D)$ can be thought of as the “true positive” conditional probability. For example, suppose that the person tested positive for the disease and then later presented symptoms in a clinical setting. $P(T^c|D)$ is the “false negative conditional probability. Suppose in this case that the person tested negative for the disease, but then presented symptoms later in a clinical setting. $P(T^c|D^c)$ is the “true negative” conditional probability—the odds that someone tested negative and does not, in fact, have the disease. Finally, $P(T|D^c)$ is the “false positive” conditional probability. This patient tested positive for the disease, but never presented symptoms.

We now use Bayes’ Theorem

$$\begin{aligned}
 P(D|T) &= \frac{P(D) \times P(T|D)}{P(T)} \\
 &= \frac{P(D) \times P(T|D)}{P(D) \times P(T|D) + (D^c) \times P(T|D^c)} \\
 &= \frac{0.01 \times 0.95}{0.95 \times 0.01 + 0.05 \times 0.99} \\
 &= 0.16
 \end{aligned}$$

Here is another way of arriving at the solution: Let say we have a population of 10,000. Since 1% of the population has the disease, this implies that about 100 people will have the disease and 9900 people will not have the disease. Even though the test is reasonably accurate, 5% of 9900 which is 495 people who do not have the disease will test positive. And 95 of the hundred people who have the disease will test positive. Now if a person test positive, the probability that she has the disease is $\frac{95}{95+495}$ which is 16%. The important point to note here is that even though the test is reasonably accurate, a significant number who do not have the disease will test positive. You can also think of this as a “signal to noise” problem. The “signal” is the fact that this only affects 1% of the population, but the noise is 5% of inaccuracy of the test.

This contraindicates the common fallacy that $P(A|B) = P(B|A)$ with the introduction of new information. Notice how the fact that the disease only affects 1% of the population was subtly conveyed, but indeed was key to calculating $P(D)$.

We will continue this is more details in the Chapter on Classification.

2.7.3 The Law of Total Probability

In this case we use conditional probability to determine the probability $P(B)$ of an event B. If A_1, A_2, \dots, A_n are mutually exclusive and exhaustive events, then

$$\begin{aligned}
 P(B) &= P(B \cap A_1) \cup P(B \cap A_2) \cup \dots \cup P(B \cap A_n) \\
 &= P(B \cap A_1) + P(B \cap A_2) + \dots \cup P(B \cap A_n) \\
 &= P(B|A_1) \times P(A_1) + P(B|A_2) \times P(A_2) + \dots + P(B|A_n) \times P(A_n)
 \end{aligned}$$

This is best illustrated in Figure 2.5.

It is important to note that A_1, A_2, \dots, A_n are mutually exclusive and exhaustive events. Note that a simple case is that we can use A and A^c .

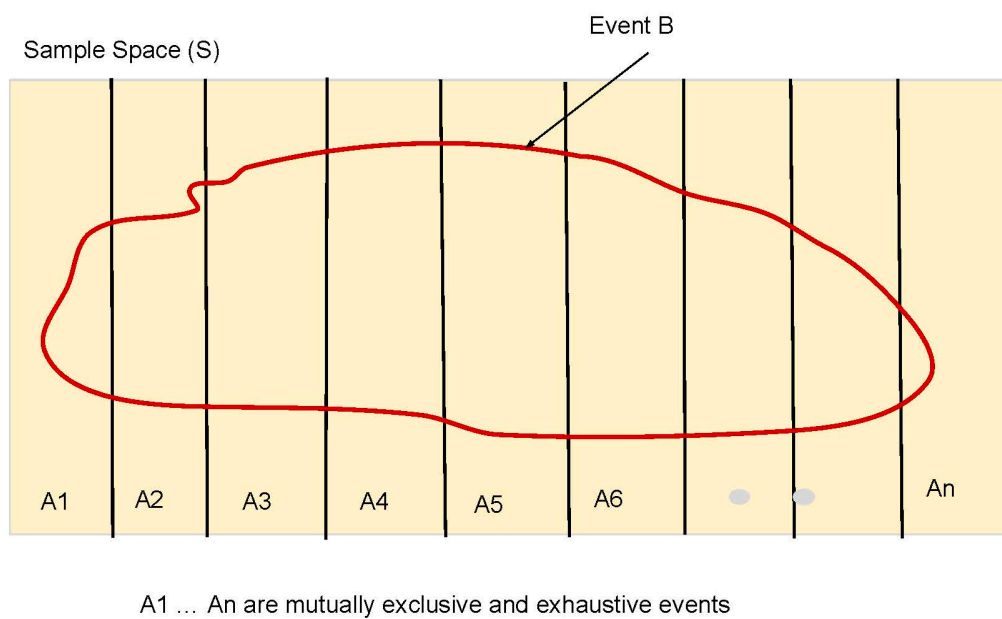


Figure 2.5: The figure shows an event B and sample space S partitioned into n events A_1, A_2, \dots, A_n .

$$\begin{aligned}
P(B) &= P(B \cap A) \cup P(B \cap A^c) \\
&= P(B \cap A) + P(B \cap A^c) \\
&= P(B|A) \times P(A) + P(B|A^c) \times P(A^c)
\end{aligned}$$

Example: A Conditional Urn Problem

This following problem is difficult and shows the use of Law of Total Probability and Bayes' Theorem. As you go through it, remember that $P(*|F)$ is complete probability space and all axioms of probability applies. Only the sample space is now changed.

There is a one ball in a bag. It is equally likely that the ball is either red or blue. A red ball is now added to the bag and then a random ball is taken out. If the ball withdrawn is red what is the probability that the remaining ball is also red?

We define the following events

1. A: event that the initial marble is red
2. B: event that the removed marble is red
3. C: event that the remaining marble is red

We need to find $P(C|B)$. To find this by conditioning whether the initial marble is red.

$$\begin{aligned}
P(C|B) &= P(C|B \cap A)P(A|B) + P(C|B \cap A^c)P(A^c|B) \\
&= 1 \times P(A|B) + 0 \times P(A^c|B) \\
&= P(A|B)
\end{aligned}$$

To find $P(A|B)$ we use Bayes' Rule

$$\begin{aligned}
P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\
&= \frac{\frac{1}{2}}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\
&= \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{4}} \\
&= \frac{2}{3}
\end{aligned}$$

Thus the required probability $P(C|B) = 2/3$.

2.7.4 Prosecutor's Fallacy

Police in a city of 1 million (10^6) are working on a crime. Police are given some description of the perpetrator (such as height, limp, etc) and based on statistics 1 in 10,000 people satisfy the description. Police on a routine patrol see a person fitting the description and they make the arrest. The prosecutor makes the following case

1. **S1:** Only 1 in 10,000 fit the description;
2. **S2:** It is highly unlikely that an innocent person fits the description;
3. **S3:** It is highly unlikely that the defendant who fits the description is innocent.

Clearly, since 1 in 10,000 fit the description, with 10^6 people, there will be 100 people who will fit the description. So the probability that the defendant is guilty is 1 in 100 which is small. We want to clearly understand where the prosecutor is making wrong statement.

Let D denote the event the person fits the description and let I denote the event that a person is innocent. Clearly, S1 which restate the statistics that 1 in 10,000 people fit the description is true. Statement S2 is calculating $P(D|I)$ probability that an innocent person fits the description.

$$\begin{aligned} P(D|I) &= \frac{P(D \cap I)}{P(I)} \\ &= \frac{99}{999,999} \\ &= 9.9 \times 10^{-5} \end{aligned}$$

This is indeed small and hence statement S2 is true.

Statement S3 is asking the probability that the defendant is innocent given that the defendant fits the description which is $P(I|D)$. The prosecutor is claiming that this is really small and hence the defendant is guilty. We can calculate it

$$\begin{aligned} P(I|D) &= \frac{P(D \cap I)}{P(D)} \\ &= \frac{99}{100} \\ &= 0.99 \end{aligned}$$

which is very close to 1.

Hence, statement S3 is wrong. The prosecutor in essence is wrongly equating $P(I|D)$ to be the same as $P(D|I)$ but they are not.

Read about the real case of Sally Clark.

2.8 Independence of Events

We consider two events A and B to be independent if

$$P(A \cap B) = P(A) \times P(B) \quad (2.10)$$

What this intuitively means is that that the occurrence of Event A does not affect the occurrence of Event B. It is important to note that this is different from Events A and B being mutually exclusive. Infact, if two events are mutually exclusive they are not independent.

Example: Two Randomly-Drawn Cards

Suppose a card is drawn at random from a deck of cards. Let A be the event that the card is a spade and B be the event that card is an Ace. These two events are independent. $P(A) = 13/52 = 1/4$ and $P(B) = 4/52 = 1/13$ and $P(A \cap B) = 1/52 = P(A) \times P(B)$

Example: Odds of Having a Baby Boy or Girl Consider a family of two children. Let A be the event that the family has children of both sexes and B be the event that there is at most one girl. Are the events A and B independent?

The sample space $S = \{bb, bg, gb, gg\}$. Now $P(A) = 2/4 = 1/2$ and $P(B) = 3/4$. $P(A \cap B) = 2/4 = 1/2$ which is not equal to $P(A) \times P(B)$