Chapter 2

Probability

The main topic of Mathematical Statistics I, probability, is introduced in this chapter.

The topics covered in this chapter are

- Experiments
- Sample spaces
- Events
- Kolmogorov axioms
- Computing or assigning probabilities
- Monte Carlo simulation
- Conditional probability
- The Law of Total Probability
- The Rule of Bayes
- Independence

2.1 Experiments, Sample Spaces, and Events

Definition 2.1. A *random experiment* is one in which the outcome is subject to chance. Every possible outcome can typically be described prior to the execution of the random experiment.

Definition 2.2. The set of all possible outcomes to a random experiment is called the *sample space* and is denoted by S.

Random Experiment	Sample Space
Roll a die and observe the up face	$S = \{1, 2, 3, 4, 5, 6\}$
Roll two dies and observe the ordered pair	$S = \{(1, 1), (1, 2), \dots, (6, 6)\}$
Roll two dies and observe the sum of the up faces	$S = \{2, 3, 4, \dots, 12\}$
Roll a red die and a green die and observe the difference between the red up face and the green up face	$S = \{-5, -4, \dots, 4, 5\}$
Toss a coin twice and observe the sequence of heads (H) and taile (T)	$S = \{HH, HT, TH, TT\}$
Toss a coin twice and observe the number of heads	$S = \{0, 1, 2\}$

Definition 2.3. An *event* A is a subset of the sample space S. If an outcome to a random experiment is in A then event A has occurred.

Set Theory	Probability
universal set	sample space
elements	outcomes of a random experiment
subsets	events

2.2 Probability Axioms

Probability quantifies uncertainty; it measures the likelihood of the occurrence of a future event. As a measure of likelihood it should satisfy at the very least the following conditions:

1. *It should be a function mapping events to nonnegative numbers.*

   Since we are talking about uncertainty of events that are mathematically defined as sets, we should define probability as a function over events[1]. Morever, as a measure of uncertainty it only makes sense that the value of such function is not negative, i.e., nonnegative.

[1] We call such functions over sets as *set functions*.

2. *The probability of the sample space S which is a sure event should be 1.*

   When we talk about likelihood of occurrence we, by assumption, refer to some sample space S that covers all possible events. Since S is the largest event among other events, the occurrence of such event *surely* covers all other possible occurrence of the smaller events. Thus it is only natural that we assign the highest possible probability value that interprets S as *sure* event. By convention, we assign such value to be 1.

3. *The probability of an event that could be partitioned into smaller pairwise disjoint (or mutually exclusive) events is equal to the sum of the probabilities of these smaller events.*

   The same as how "*the whole is equal to the sum of its parts*", the probability of an event could be represented as a sum of the individual probabilities of the smaller events that compose it.

From these three desirable conditions born out what probabilists called as the Kolmogorov Axioms that served as basis for deriving classic results in Probability Theory. Before we present the Kolmogorov Axioms, we first introduce the concept of sigma algebra.

Definition 2.4 (Sigma Algebra)**.** Let $\mathscr{F}$ be collection of events from a sample S. We call $\mathscr{F}$ a *sigma algebra* if the following three conditions are satisfied:

1. $S \in \mathscr{F}$.
2. If $E \in \mathscr{F}$, then $E^c \in \mathscr{F}$.
3. If $E_1, E_2, E_3, ... \in \mathscr{F}$, then $\bigcup_{i=1}^{\infty} E_i \in \mathscr{F}$.

Remark. Recall that events are basically sets. In higher levels of Probability Theory, there are some sets that behave so "weirdly" that could not be considered as events in the common sense of the term. They are so weird that computing their probabilities led to contradictory results. Probabilist call these cases as *pathologies*. With sigma algebra, these pathologies are avoided. Thus, from this point on, we refer to events as sets belonging to a sigma algebra of a given sample space.

2.2.1 Kolmogorov Axioms

Definition 2.5 (Kolmogorov Axioms)**.** Let S be a sample space $\mathscr{F}$ be a sigma algebra from S. Define the set function $\mathbb{P} : \mathscr{F} \to \mathbb{R}$ such that the following three conditions are satisfied:

Axiom 1: $\mathbb{P}(E) \geq 0$ for all $E \in \mathscr{F}$.

Axiom 2: $\mathbb{P}(S) = 1$.

Axiom 3: For every pairwise disjoint sub-collection of events $E_1, E_2, ... \in \mathscr{F}$,

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} E_i\right) = \sum_{k=1}^{\infty} \mathbb{P}(E_i)\,.$$

We call $\mathbb{P}$ a *probability function* or simply *probability* and the triple $(S, \mathscr{F}, \mathbb{P})$ as *probability space.*

Remark. Axioms 1, 2, and 3 are often called respectively as the *nonnegativity axiom*, *unit normatlization axiom* and *additivity axiom.*

2.2.2 Some derived properties

Let $(S, \mathscr{F}, \mathbb{P})$ be probability triple.

Theorem 2.1 (Complementary Property)**.** *For each $A \subset \mathscr{F}$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$*

Theorem 2.2 (Ordering Property)**.** *If $A_1, A_2 \in \mathscr{F}$ such that $A_1 \subset A_2$. Then,*

a. $\mathbb{P}(A_2 \setminus A_1) = \mathbb{P}(A_2 \cap A_1^c) = \mathbb{P}(A_2) - \mathbb{P}(A_1)$ *and*
b. $\mathbb{P}(A_1) \leq \mathbb{P}(A_2)$.

Theorem 2.3. $\mathbb{P}(\emptyset) = 0$.

Theorem 2.4. *For every $A \in \mathscr{F}$, $0 \leq \mathbb{P}(A) \leq 1$.*

Theorem 2.5 (Addition Rule)**.** *If $A_1, A_2 \in \mathscr{F}$, then*

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cup A_2).$$

Theorem 2.6 (General Addition Rule)**.** *If $A_1, A_2, A_3, ..., A_n \in \mathscr{F}$, then for $n \geq 3$,*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} (-1)^{k+1} \mathbb{P}\left(\bigcap_{r=1}^{k} A_{i_r}\right)$$

where $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ means "all combinations of the numbers from 1 to n taken k at a time".

Example 2.1. For $n = 3$, we have

$$\begin{aligned}
&\sum_{k=1}^{n} \sum_{1\le i_1<i_2<\cdots<i_k\le n} (-1)^{k+1}\mathbb{P}\left(\bigcap_{r=1}^{k} A_{i_r}\right) \\
&\quad = \sum_{k=1}^{3} \sum_{1\le i_1<i_2<\cdots<i_k\le 3} (-1)^{k+1}\mathbb{P}\left(\bigcap_{r=1}^{k} A_{i_r}\right) \\
&\quad = \sum_{1\le i_1\le 3} (-1)^{2}\mathbb{P}\left(\bigcap_{r=1}^{1} A_{i_r}\right) \\
&\qquad + \sum_{1\le i_1\le i_2\le 3} (-1)^{3}\mathbb{P}\left(\bigcap_{r=1}^{2} A_{i_r}\right) \\
&\qquad + \sum_{1\le i_1<i_2<i_3\le 3} (-1)^{4}\mathbb{P}\left(\bigcap_{r=1}^{3} A_{i_r}\right) \\
&\quad = \sum_{i_1=1}^{3} \mathbb{P}\left(A_{i_1}\right) \\
&\qquad - \sum_{1\le i_1\le i_2\le 3} \mathbb{P}\left(A_{i_1}\cap A_{i_2}\right) \\
&\qquad + \sum_{1\le i_1<i_2<i_3\le 3} \mathbb{P}\left(A_{i_1}\cap A_{i_2}\cap A_{i_3}\right) \\
&\quad = \mathbb{P}(A_1)+\mathbb{P}(A_2)+\mathbb{P}(A_2) \\
&\qquad - \mathbb{P}(A_1\cap A_2)-\mathbb{P}(A_1\cap A_3)-\mathbb{P}(A_2\cap A_3) \\
&\qquad + \mathbb{P}(A_1\cap A_2\cap A_3).
\end{aligned}$$

Theorem 2.7 (Boole's Inequality)**.** *Let $A_1, A_2, \ldots, A_n \in \mathscr{F}$. Then,*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{i=1}^{n} \mathbb{P}(A_i).$$

Remark. If we replace n with ∞, Boole's inequality still holds.

Theorem 2.8 (Bonferroni's Inequalities)**.** *Let $A_1, A_2, \ldots A_n \in \mathscr{F}$. Then, for odd m,*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{k=1}^{m} \sum_{1\le i_1<i_2<\cdots<i_k\le n} (-1)^{k+1}\mathbb{P}\left(\bigcap_{r=1}^{k} A_{i_r}\right)$$

and for even m,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \ge \sum_{k=1}^{m} \sum_{1\le i_1<i_2<\cdots<i_k\le n} (-1)^{k+1}\mathbb{P}\left(\bigcap_{r=1}^{k} A_{i_r}\right)$$

Definition 2.6 (Sample Space Partition)**.** If a collection of events $\{A_1, A_2, \dots, A_n\}$ satisfies the following two conditions:

1. $A_i \cap A_j = \emptyset$ for every $i \neq j$ and
2. $S = \bigcup_{i=1}^{n} A_i$

then we say that the collection is a *sample space partition* or simply *partition.*

Theorem 2.9. *Let $\{A_1, A_2, \dots, A_n\}$ be a sample space partition. Then for any $B \in \mathscr{F}$,*

1. *the collection $\{C_1, C_2, \dots, C_n\}$ where $C_i = B \cap A_i$ forms a partition of B and*
2. $\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(C_i) = \sum_{i=1}^{n} \mathbb{P}(B \cap C_i)$.

2.3 Probability Assignment

Note that Kolmogorov Axioms and the properties derived thereof is only limited to how we *manipulate* probability but does not provide us with the means to compute actual numeric values of the probability function $\mathbb{P}(\,\cdot\,)$. For a given sample space, one can assign different possible probability values as long as they satisfy that Kolmogorov Axioms.

In this section, we present three viewpoints on how to assign probability values to events. These are, namely, (1) *classical*, (2) *frequency* and (3) *subjective viewpoint.* Each of these viewpoints has assumptions on how they assign probabilities.

2.3.1 Classical Viewpoint

If we could theoretically enumerate all the elements of the sample space, one way to assign probabilities is by assigning equal numeric values. This means that each point in the sample space are equiprobable or are of equal chance to occur.

Example 2.2. For a die-tossing experiment, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Then classical viewpoint assigns equal probabilities to each point in S as follows:

x	1	2	3	4	5	6
$\mathbb{P}(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Notice from the table above that the probabilities should sum to 1.

In general, we have a *finite discrete* sample space $S = \{x_1, x_2, \dots, x_k\}$, classical viewpoint assign probabilities as:

$$\mathbb{P}(x_i) = \frac{1}{k}, \quad \text{for } i = 1, 2, \dots, k.$$

The limitation of the classical viewpoint is that it is only applicable for finite discrete sample spaces and bounded intervals[2]. If you consider sample spaces other than these, assigning equal or uniform values to each point would lead to violation of the Kolmogorov Axiom 2. (Verify this.)

2.3.2 Frequency Viewpoint

Definition 2.7 (Relative Frequency)**.** Perform a random experiment n times. Let x_n be the number of times that the event A occurs. Then ratio $p_n = \frac{x_n}{n}$ is the *relative frequency* of the event A in the n experiments.

Definition 2.8 (Limiting Relative Frequency)**.** Let $p_n = \frac{x_n}{n}$ be the relative frequency of event A in n repetitions of the experiment. Then the *limiting relative frequency* denoted by p_A is defined as the limit of the ratio as n approaches infinity, that is,

$$p_A = \lim_{n\to\infty} p_n = \lim_{n\to\infty} \frac{x_n}{n}.$$

Under the frequency viewpoint, the probability assigned to an event under consideration is its limiting relative frequency.

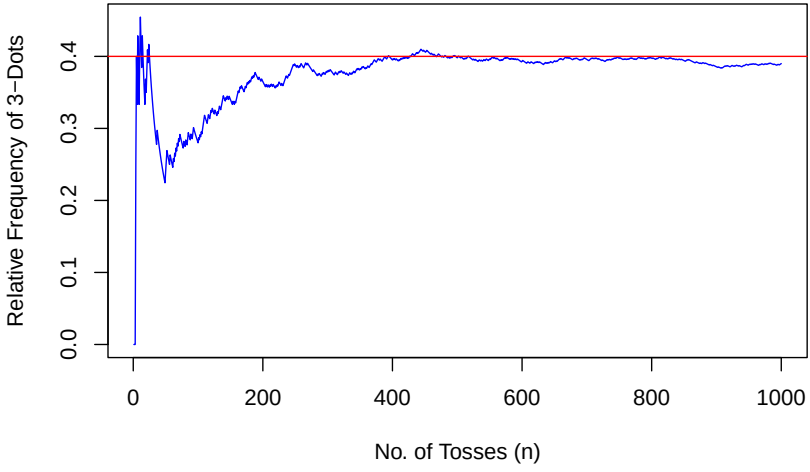
Example 2.3. Consider again the die-tossing experiment. Suppose it is claimed that the die is biased towards the triple-dots up face. To verify the claim, we perform a mini experiment wherein we toss the die 1000 times and tally the relative frequency for each up face dots. The results are shown in the following table.

x	1	2	3	4	5	6
$\mathbb{P}(x)$	$\frac{111}{1000}$	$\frac{134}{1000}$	$\frac{390}{1000}$	$\frac{62}{1000}$	$\frac{211}{1000}$	$\frac{92}{1000}$

Looking at the tally results, the evidence weights in favor towards the claim.

Now, suppose we are able to plot the evolution of the relative frequency $(\frac{x_n}{n})$ for the event resulting to three-dots up face a follows:

[2]We'll consider probability assignments on intervals in the later chapters.



Notice how the relative frequency $\frac{x_n}{n}$ fluctuates closer to $p = 0.4$ as the number of tosses n increases. This behaviour follows in accordance to the limit expression in Definition 2.8.

Remark. Notice that in the above definition, it is not necessary that the random experiment is a synthetic one. In real life applications, probability assignment under the frequency viewpoint often utilizes *historical data* i.e., past information tallies, to compute relative frequencies. Consider the following scenario:

> *There are 12 possible outcomes for the birth month. One can assume that each month is equally likely to occur, but actually in the U.S. population, the number of births during the different months do vary. Using data from the births in the U.S. in 1978, we obtain the following probabilities for the months. We see that August is the most likely birth month with a probability of 0.091 and February (the shortest month) has the smallest probability of 0.075.*

Month	Jan	Feb	Mar	Apr	May	Jun
Probability	0.081	0.075	0.083	0.076	0.082	0.081

Month	Jul	Aug	Sep	Oct	Nov	Dec
Probability	0.088	0.092	0.088	0.087	0.082	0.085

In this case, one could imagine nature "performing a random experiment" of giving birth to babies and noting the birth month as "outcome of the experiment". The tallied relative frequencies *approximates* the limiting relative frequencies given the large number of births.

2.3.3 Subjective Viewpoint

There are some situations wherein

1. *frequency viewpoint is not feasible since performing repetitive random experiments is either impossible or not practical*

   There are cases wherein one cannot perform a process of repeating random experiments just to compute the relative frequencies in favor of the event whose probability we want to compute. Consider for example the presidential election of a country wherein the candidates who won will no longer be able to run the second time after his/her first term. In this case, it is impossible[3] to repeat *a large number of times* the presidential election of the same set of candidates just to compute the probability of winning of a certain candidate.

2. *the equally-likely assumption of the classical viewpoint is so restrictive for practical applications*

   Assigning probabilities using the classical viewpoint is so restrictive that it is impractical to apply in real life problems. Consider the coin toss scenario, suppose someone who knows that nature of the coin tells you that the coin is biased in favor of the Tails. This valuable information is straightway ignored when we use classical viewpoint since under such viewpoint, by assumption, both Head and Tail have equal chance to occur. Aside from the equally likely assumption, there's also the problem of violating Kolmogorov Axiom 2 in the case where the sample space is neither a finite discrete set nor a bounded interval. Such sample spaces are very common in real life.

To answer the above problems, a third viewpoint is introduced which is the *subjective viewpoint*.

[3]Even if it is possible, it is quite impractical due financial and time constraints.

When assigning probabilities, it is not wrong to assume that the person (i.e., the subject), the one doing the assigning holds the right to decide what assumptions to make with regards to the probability values. This is essentially the idea behind subjective viewpoint. The process of probability assignment does not depend on the external object of concern to which we assign values but on us (the subject) together with our personal knowledge and experience about the object.

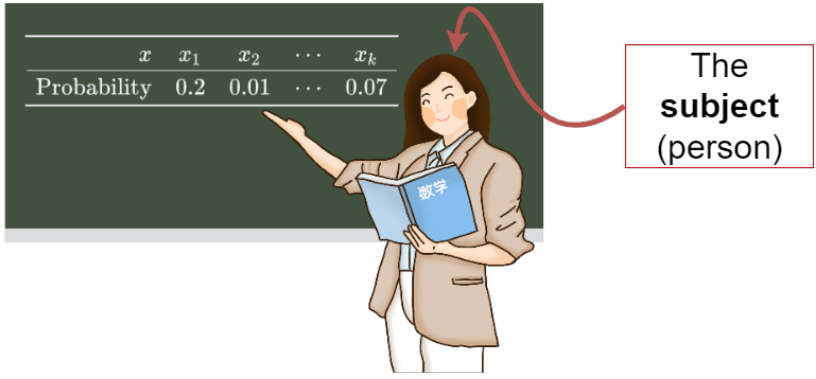


Figure 2.1: Illustrating that a subject is the person doing the process assignment of probability values.

Using the subjective viewpoint, we are able to incorporate both the following two sources of information when we are assigning probabilities:

1. historical data used in frequency viewpoint
2. personal (subjective) knowledge and experience

Moreover, the probability assignments of the classical viewpoint is a special case of the subjective viewpoint when the subject decides to assign equal probability in situations where no knowledge is available.

Consider the following problem:

> *Suppose a girl goes to an ice cream parlor and plans to order a single-dip ice cream cone. This particular parlor has four different ice cream flavors namely Vanilla, Choco, Butter, and Macapuno. Which flavor will the girl order?*

The selection of the flavor is considered a random phenomena here with sample space

$$S = \{\text{Vanilla}, \text{Chocolate}, \text{Butter}, \text{Macapuno}\}$$

The answer to the ice-cream problem is to determine the most likely (highest probability value) flavor to be chosen by the girl. Let's consider the probability in the perspective of three subjects *present* in the scenario.

Subject 1: A bystander who randomly guesses

From the perspective of a bystander who randomly guesses which of the four flavors the girl will choose, classical viewpoint would be the most appropriate assignment of probabilities:

Flavor	Vanilla	Chocolate	Butter	Macapuno
Probability	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

Subject 2: The ice-cream man selling the ice-cream

From the seller's viewpoint, an accumulated history of 1000 ice cream sales in the same area gives the following proportions .

Flavor	Vanilla	Chocolate	Butter	Macapuno
Probability	$\frac{452}{1000}$	$\frac{131}{1000}$	$\frac{310}{1000}$	$\frac{107}{1000}$

This is basically the frequency viewpoint of assigning probabilities using historical data of counts.

Subject 3: The girl's fiance

From the perspective of her fiance who knows her preferences based on experience, the girl is allergic to butter and he has not seen her buy a vanilla before. So he would assign zero probabilities to these two flavors. In addition, he knows that she frequently requests for either chocolate or macapuno but prefers the later if it's available. Thus, he might assign 0.8 probability for latter and 0.2 for the former. Thus, one possible assignment that her fiance would make based on these knowledge or information is the following table:

	Flavor	Vanilla	Chocolate	Butter	Macapuno
Probability		0	0.2	0	0.8

Notice that the three subjects have different probability assignments on what flavor the girl will choose as they have different source of information that influenced or shaped their decisions:

1. *No knowledge.* The bystander, having no knowledge of either the girl's preferences or the seller's historical data, chose to guess randomly indirectly assigning equal probabilities.
2. *Non-personal knowledge.* The seller chose his probability assignment based solely on historical information.
3. *Personal knowledge.* The fiance choice of probability values is shaped by his personal knowledge and experience about girl's preferences.

The assignment done in numbers 1 and 3 (i.e., probabilities shaped from either having no knowledge or with personal knowledge) comprises what we call as *subjective* or *personal knowledge.*

2.4 Computing Probabilities

We already know how to assign probabilities using three viewpoints discussed in the previous section. In this section, we show how to compute probabilities using three approaches:

1. Kolmogorov Axioms together with the derived properties of $\mathbb{P}$ that are presented in Probability Axioms section.
2. Using counting techniques under classical viewpoint
3. Using Monte Carlo simulation under frequency viewpoint.

2.4.1 Using the Kolmogorov Axioms and $\mathbb{P}$ properties

Example 2.4. Consider again the following problem.

> *There are 12 possible outcomes for the birth month. One can assume that each month is equally likely to occur, but actually in the U.S. population, the number of births during the different months do vary. Using data*

from the births in the U.S. in 1978, we obtain the following probabilities for the months. We see that August is the most likely birth month with a probability of 0.091 and February (the shortest month) has the smallest probability of 0.075.

Month	Jan	Feb	Mar	Apr	May	Jun
Probability	0.081	0.075	0.083	0.076	0.082	0.081

Month	Jul	Aug	Sep	Oct	Nov	Dec
Probability	0.088	0.092	0.088	0.087	0.082	0.085

1. Are these values probability values?
2. What is the sample space?
3. Define the following events:

   $L =$ the student is born during the last half of the year
   $E =$ the student is born during a month that is four letters long.

   Compute the following probabilities:

   a. $\mathbb{P}(L)$
   b. $\mathbb{P}(E)$
   c. $\mathbb{P}(E^c)$
   d. $\mathbb{P}(L \cap E)$
   e. $\mathbb{P}(L \cup E)$
   f. $\mathbb{P}(L^c \cap E^c)$
   g. $\mathbb{P}(L \setminus E^c)$

Solutions:

1. If you take the sum of the values, we could see that the sum is equal to 1. Moreover, all values are nonnegative. Therefore, these value are probability values.
2. The sample space is

   $$S = \{\text{Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec}\}.$$

3. By definition of L and E, we have

$$L = \{\text{Jul, Aug, Sep, Oct, Nov, Dec}\}$$
$$E = \{\text{Jun, Jul}\}$$

Now,

a. With L above, we have

$$\begin{aligned}\mathbb{P}(L) &= \mathbb{P}\left(\{\text{Jul, Aug, Sep, Oct, Nov, Dec}\}\right)\\ &= 0.088 + 0.092 + 0.088 + 0.087 + 0.082 + 0.085\\ &= 0.522\end{aligned}$$

b. With E above, we have

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}\left(\{\text{Jun, Jul}\}\right)\\ &= 0.081 + 0.088\\ &= 0.169\end{aligned}$$

c. Since we already have the value of $\mathbb{P}(E)$, we do not need to manually enumerate the elements of E^c but rather use instead the Complementary Property (Theorem 2.1) as follows:

$$\begin{aligned}\mathbb{P}(E^c) &= 1 - \mathbb{P}(E)\\ &= 1 - 0.169\\ &= 0.831\end{aligned}$$

d. Note first that

$$\begin{aligned}L \cap E &= \{\text{Jul, Aug, Sep, Oct, Nov, Dec}\} \cap \{\text{Jun, Jul}\}\\ &= \{\text{Jul}\}\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{P}(L \cap E) &= \mathbb{P}\left(\{\text{Jul}\}\right)\\ &= 0.088\end{aligned}$$

e. Instead of manually enumerating the elements of $L \cup E$, we use the Addition Rule (Theorem 2.5) since we already have computed the probabilities $\mathbb{P}(L)$, $\mathbb{P}(E)$ and $\mathbb{P}(L \cap E)$ from the previous items. Proceeding accordingly,

$$\begin{aligned}\mathbb{P}(L \cup E) &= \mathbb{P}(L) + \mathbb{P}(E) - \mathbb{P}(L \cap E)\\ &= 0.522 + 0.169 - 0.088\\ &= 0.603\end{aligned}$$

f. Instead of manually enumerating the elements of $L^c \cap E^c$, note first that by De Morgan's Laws (Proposition 1.1) $L^c \cap E^c = (L \cup E)^c$. Thus, using the result of the previous item and Complementary Property (Theorem 2.1) again, we have

$$\begin{aligned}\mathbb{P}(L^c \cap E^c) &= \mathbb{P}\left((L \cup E)^c\right) \\ &= 1 - \mathbb{P}(L \cup E) \\ &= 1 - 0.603 \\ &= 0.397\end{aligned}$$

g. First note that by Proposition 1.3 (3a), we can rewrite $L \setminus E^c$ as $L \cap (E^c)^c = L \cap E$. Therefore, using the result of item (d) above, $\mathbb{P}(L \setminus E^c) = \mathbb{P}(L \cap E) = 0.088$.

If you observe from the above exercise, computing probabilities of events utilizes the following:

1. Properties of sets like the ones derived from the previous chapter
2. Properties of the $\mathbb{P}$ function derived from Kolmogorov Axioms as provided above.

2.4.2 Using the Counting Techniques

Under the assumption of a finite discrete sample space $S = \{x_1, x_2, \ldots, x_k\}$, we can compute probabilities without enumerating exhaustively the elements of S by using counting techniques if assign probabilities to each of the element x_i an equal probability.

Example 2.5. A bag contains 15 billiard balls numbered 1 through 15. Five balls are randomly drawn from a bag without replacement[4] Let the event A be exactly two odd-numbered balls are drawn from the bag and they occur on odd-numbered draws. Find $\mathbb{P}(A)$.

Solution. Let $B = \{1, 2, \ldots, 15\}$. Since the selection is done without replacement, every set of 5 balls drawn are distinct. Thus, our sample space in this case is

$$S = \left\{(x_1, x_2, x_3, x_4, x_5) \,|\, x_i \in B \text{ where for } i \neq j,\ x_i \neq x_j\right\}$$

The number points in S which we denote as $|S|$, is equal to $15 \cdot 14 \cdot 13 \cdot 12 \cdot 11$ using Theorem 1.2. Under classical viewpoint, each of the possible outcomes are equally likely to occur because of the assumption that the balls are randomly picked.

[4]When we say "draw without replacement" it means we do not replace/return the ones that are drawn/picked.

To determine the number of outcomes corresponding to exactly two odd-numbered balls drawn from the bag and which occurs on odd-numbered draws from a bag, consider that

- using Theorem 1.5, there are $\binom{8}{2}$ different ways to choose the odd-numbered balls from the bag (order is not relevant),
- using Theorem 1.5 again, there are $\binom{7}{3}$ different ways to choose even-numbered balls from the bag (order is not relevant),
- using Theorem 1.5 again, there are $\binom{3}{1} = 3$ different positions (namely, 1, 3, and 5) for the draw number where the one even draw assumes an odd draw number,
- using Theorem 1.2, there are 3! ways to order the even numbers,
- using Theorem 1.2 again, there are 2! ways to order the odd numbers.

Using the Multiplication Rule (Theorem 1.1), the total number of points in A is

$$|A| = \binom{8}{2} \times \binom{7}{3} \times 3 \times 3! \times 2! = 35,280$$

With the classical view of probability assignment,

$$\mathbb{P}(A) = \frac{|A|}{|S|} = \frac{35,280}{360,360} \approx 0.0979021.$$

2.4.3 Using the Monte Carlo Simulation

Recall that under the frequency viewpoint, if we could perform a random experiment and count the outcomes that are in favor for the event of interest, then the relative frequency can be used as an approximation to the the true probability of the said event. We can use this principle to estimate probabilities if we translate the steps of the experiment into an algorithm that can be perform a computing machine. One such procedure is what we call *Monte Carlo simulation.*

Example 2.6. Consider again the problem in Example 2.5. This time we compute an estimate of the probability using the frequency view point. The steps of the experiment using Monte Carlo simulation are essentially as follows:

1. Draw randomly 5 balls without replacement from the bag. Denote the draw as $(x_1, x_2, x_3, x_4, x_5)$.
2. Check whether the following are all true:

- Exactly two among x_1, x_3 and x_5 are odd
- Both x_2 and x_4 are even

If all are true, add a tally to the count in favor of A.

3. Repeat steps 1 and 2 ten thousand times.
4. Compute the relative frequency by dividing the total tally in favor of A over the total number of repetitions.
5. Repeat Steps 1 to 4 a hundred times.
6. Set the average of all 100 relative frequencies as the estimate of the probability of A.

We are now going to implement the above steps using the statistical software `R`. The raw codes are follows:

```
set.seed(1234) # Some R-related settings

# Set the no. of repetitions (or replications)
Nrep1 <- 10000
Nrep2 <- 100

# Define a container for the 100 ratio estimates
ratio <- rep(NA,Nrep2)

for (i in 1:Nrep2) { # Step 5: Repeat Steps 1 to 4 a hundred times.
  count <- 0
  for (k in 1:Nrep1) { # Step 3: Repeat steps 1 and 2 ten thousand times

    # Step 1: Draw randomly 5 balls w/o replacement from the bag of 15 balls
    x <- sample.int(n = 15, size = 5, replace = FALSE)

    # Step 2: Checking of two conditions and tallying counts in favor of A

      # Check whether exactly two among x1, x3, and x5 are odd
      cond1 <- { (x[1]%%2 + x[3]%%2 + x[5]%%2) == 2 }

      # Check whether both two x2 and x4 are odd
      cond2 <- { (x[2]%%2 + x[4]%%2) == 0 }

      # Tally counts in favor of A
```

```
      if (cond1 && cond2) count <- count + 1
  }
  # Step 4: Divide total tally in favor of A over total repetitions
  ratio[i] <- count/Nrep1
}

# Step 6: Set the average of all 100 relative frequencies as the estimate of
#         the probability of $A$.
pA <- mean(ratio)
pA
#> [1] 0.097887
```

Thus, the estimated probability of event A is 0.097887. If we want to know whether the true value of $\mathbb{P}(A) = 0.0979021$ is within range of possible values of the Monte Carlo esimates, we use the following code:

```
range(ratio)
#> [1] 0.0915 0.1055
```

The above code shows the smallest (0.0915) and largest (0.1055) estimate of $\mathbb{P}(A)$. Clearly, 0.0979021 is within this range.

Below is the same code without the comment lines:

```
set.seed(1234)
Nrep1 <- 10000
Nrep2 <- 100
ratio <- rep(NA,Nrep2)
for (i in 1:Nrep2) {
  count <- 0
  for (k in 1:Nrep1) {
    x <- sample.int(n = 15, size = 5, replace = FALSE)
      cond1 <- { (x[1]%%2 + x[3]%%2 + x[5]%%2) == 2 }
      cond2 <- { (x[2]%%2 + x[4]%%2) == 0 }
      if (cond1 && cond2) count <- count + 1
  }
  ratio[i] <- count/Nrep1
}
pA <- mean(ratio)
pA
range(ratio)
```

2.5 Conditional Probability

This section addresses probability questions in light of other additional information, that is, calculating probabilities conditioned on the fact that another event has occurred. Oftentimes, the probability of an event goes up (or down) when you know whether another event has occurred. In other instances, we are interested in only the outcomes in a subset of S. Applications of conditional probability include:

- *Meteorology:* What is the probability that it rains tomorrow given that it is raining today?
- *Stock market:* What is the probability that a stock market index rises today given that it dropped yesterday?
- *Genetics:* What is the probability that a child will have blue eyes given that one parent has blue eyes?
- *Economics:* What is the probability that government revenue will increase next month given that there is a specified small increase in unemployment this month?

In all cases, we seek the probability of the event given that another event has occurred.

2.5.1 Conditional Probability

Definition 2.9 (Conditional Probability)**.** If A and B are two events in the sample space S, then the probability that A given B is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{2.1}$$

provided that $\mathbb{P}(B) > 0$. We call event as the *conditioning event*.

The condition $\mathbb{P}(B) > 0$ is included in the definition of conditional probability for two reasons:

1. If $\mathbb{P}(B) = 0$, then the definition of $\mathbb{P}(A|B)$ has a zero in the denominator and division by zero is undefined.
2. If $\mathbb{P}(B) = 0$, then the event B is an *impossible* event. If the event B cannot occur, it does not make sense to calculate the probability that A occurs given B occurs.

Example 2.7. Toss a fair die and observe the number of spots on the up face. Let event A correspond to tossing a 1, 2, or 3. Let the even B correspond to tossing an odd number.

1. What is the probability of A?
2. What is the probability of A given has B occurred?

Solution. The sample space and the two events of interest are

$$S = \{1, 2, 3, 4, 5, 6\}$$
$$A = \{1, 2, 3\}$$
$$B = \{1, 3, 5\}$$

Since the die is fair, we use classical viewpoint in assigning probabilities.

x	1	2	3	4	5	6
$\mathbb{P}(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

1. The probability of A is
$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}\left(\{1,\ 2,\ 3\}\right)\\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6}\\ &= 0.5\end{aligned}$$
2. To computed for the probability of A given that B has occurred, we can use Equation (2.1) which requres solving for $\mathbb{P}(A \cap B)$ and $\mathbb{P}(B)$. Proceeding accordingly, we have
$$\begin{aligned}\mathbb{P}(A \cap B) &= \mathbb{P}\left(\{1,\ 2,\ 3\} \cap \{1,\ 3,\ 5\}\right)\\ &= \mathbb{P}\left(\{1,\ 3\}\right)\\ &= \frac{1}{6} + \frac{1}{6}\\ &= 0.3333333\end{aligned}$$
On other hand,
$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}\left(\{1,\ 3,\ 5\}\right)\\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6}\\ &= 0.5\end{aligned}$$
Therefore, by Equation (2.1)
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{0.3333333}{0.5} = 0.6666667.$$

Example 2.8. The results of a random sample of 100 subjects classified by their gender and eye color is given in Table 2.13. If one of the subjects is selected at random,

a. find the probability that they have blue eyes given that they are male,
b. find the probability that they are female given that they have green eyes.

Table 2.13: Gender and eye color for 100 students

	Blue	Green	Other
Male	26	23	24
Female	13	12	2

Solution. For an individual subject, define the events as follows:

$$\begin{aligned} B &= \text{blue eyes} \\ G &= \text{green eyes} \\ O &= \text{other colored eyes} \\ M &= \text{male} \\ F &= \text{female} \end{aligned}$$

	Blue	Green	Other	***Total***
Male	26	23	24	73
Female	13	12	2	27
Total	39	35	26	100

Using Equation (2.1), the conditional probabilities of interest are

a. $\mathbb{P}(B|M) = \dfrac{\mathbb{P}(B \cap M)}{\mathbb{P}(M)} = \dfrac{26/100}{73/100} = \dfrac{26}{73} = 0.3561644$
b. $\mathbb{P}(F|G) = \dfrac{\mathbb{P}(F \cap G)}{\mathbb{P}(G)} = \dfrac{12/100}{35/100} = \dfrac{12}{35} = 0.3428571$

Example 2.9. Consider the events A_1 and A_2 with associated probabilities $\mathbb{P}(A_1) = 0.3$, $\mathbb{P}(A_2) = 0.5$, and $\mathbb{P}(A_1 \cap A_2) = 0.2$. Find $\mathbb{P}(A_1|A_2)$ and $\mathbb{P}(A_2|A_2)$.

Solution. Using Equation (2.1),

$$\mathbb{P}(A_1|A_2) = \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_2)} = \frac{0.2}{0.5} = 0.4$$

and

$$\mathbb{P}(A_2|A_1) = \frac{\mathbb{P}(A_2 \cap A_1)}{\mathbb{P}(A_1)} = \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)} = \frac{0.2}{0.3} = 0.6666667$$

2.5.2 Properties of the Conditional Probability Function

Conditional probability behaves the same way as the usual probability function except the provision that the conditioning event[5] should have non-zero probability.

Like any other probability function, conditional probability satisfies the Kolmogorov Axioms and other associated results derived from them.

Proposition 2.1. *Let A, B be events from a smaple space S such that $\mathbb{P}(B) > 0$. Then*

1. $0 \leq \mathbb{P}(A|B) \leq 1$
2. *(**Complementary Property**)* $\mathbb{P}(A^c|B) = 1 - \mathbb{P}(A|B)$
3. $\mathbb{P}(S|B) = \mathbb{P}(B|B) = 1$
4. *For pairwise disjoint collection of events $E_1, E_2, \ldots$ from S, we have*

$$\mathbb{P}\left[\left(\bigcup_{i=1}^{+\infty} E_i\right)\middle|B\right] = \sum_{i=1}^{+\infty} \mathbb{P}\left(E_i|B\right)$$

5. *(**Multiplication Rule**)* $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$
6. *If $\mathbb{P}(A) > 0$ in addition to the condition that $\mathbb{P}(B) > 0$, then*

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

Theorem 2.10 (General Multiplication Rule)**.** *Let $A_1, A_2, \ldots, A_n$ be events from S. Then*

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathbb{P}(A_n) \times \prod_{i=1}^{n-1} \mathbb{P}\left(A_i|A_{i+1} \cap \cdots \cap A_n\right)$$

provided that all conditioning events have non-zero probability.

[5]Recall from Definition 2.9 that event B in the expression $\mathbb{P}(A|B)$ is called the conditioning event.

Example 2.10. There are 10,000 doctors place on a low daily dose of aspirin and 10,000 other doctors place on a placebo[6] for a year. During this year, 107 of those on aspirin have a heart attack, while 187 of those on the placebo have a heart attack. If a doctor from the study is selected at random, what is the probability that the doctor will have been on aspirin and had a heart attack?

Solution. Define the events

$$\begin{aligned} A &: \text{being on apsirin} \\ P &: \text{being on placebo} \\ H &: \text{with heart attack} \\ H' &: \text{without heart attack} \end{aligned}$$

Approach 1: Using Contingency Tables

Summarizing the information provided, we have

	H	H'	Total
A	107	9,893	10,000
P	187	9,813	10,000
Total	294	19,706	20,000

Thus, the probability that the doctor randomly selected will have been on apsirin and had a heart attack (i.e., the event $A \cap H$) is

$$\mathbb{P}(A \cap H) = \frac{107}{20,000} = 0.00535.$$

Approach 2: Using Multiplication Rule

Instead of using contingency tables, we can directly use the information provided via multiplication rule. Note that by interpretation

Phrase	Implied Notation
"107 of those on aspirin (10,000) have a heart attack"	$\mathbb{P}(H\|A) = \frac{107}{10,000}$

[6] A placebo is a neutral or baseline type of drug or substance which usually doesn't have any effects to patients but included in an experimental studies for the sake of comparison. The group assigned with a placebo is called a *placebo group* or simply *placebo*.

Phrase	Implied Notation
"187 of those on placebo (10,000) have a heart attack"	$\mathbb{P}(H\|P) = \frac{187}{10,000}$

Together with the information that the total number of patients who are in aspirin (A) is $10,000$ out of $20,000$ doctors, the probability that the doctor will have been on aspirin and had a heart attack is

$$\mathbb{P}(A \cap H) = \mathbb{P}(H|A)\mathbb{P}(A) = \frac{107}{10,000} \cdot \frac{10,000}{20,000} = 0.00535.$$

2.5.3 Law of Total Probability

Theorem 2.11 (Law of Total Probability)**.** *Let $A_1, A_2, \ldots, A_n$ be a sample space partition*[7] *of S and $\mathbb{P}(A_i) > 0$ for $i = 1, 2, \ldots, n$. For any event B,*

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$

Example 2.11. According toe Wikipedia, 33% of the births in the US are performed by Caesarean (CS) section. The risk of death for a baby delivered by Caesarean section in the first 28 days is 1.77 deaths per 1000 live births. The risk of death for a baby not delivered by Caesarean section in the first 28 days is 0.62 deaths per 1000 live births. Given these numbers, what is the probability that a baby born alive in the United States will survive its first 28 days?

Solution. Let the event B denote the event that a baby born alive survives its first 28 days. Let the event C denote the event that a baby is delivered by Caesarean section. Our goal is to compute $\mathbb{P}(B)$.

From the information provided, we have

Phrase	Implied Notation
"33% of births are performed by CS section"	$\mathbb{P}(C) = 0.33$
"risk of death by CS section in 1st 28days is 1.77 per 1000"	$\mathbb{P}(B^c\|C) = \frac{1.77}{1,000}$

[7] Recall this from Definition 2.6.

Phrase	Implied Notation
"risk of death not by CS section in 1st 28days is 0.62 per 1000"	$\mathbb{P}(B^c\|C^c) = \frac{0.62}{1{,}000}$

Therefore, by Complementary Property, the probability that a baby survives for the first 28 days after CS section is

$$\mathbb{P}(B|C) = 1 - \mathbb{P}(B^c|C) = 1 - \frac{1.77}{1000} = 0.99823$$

and the probability that a baby survives for the first 28 days born not from a CS section procedure is

$$\mathbb{P}(B|C^c) = 1 - \mathbb{P}(B^c|C^c) = 1 - \frac{0.62}{1000} = 0.99938$$

while the probability that a baby is not born through CS section is

$$\mathbb{P}(C^c) = 1 - \mathbb{P}(C) = 1 - 0.33 = 0.67$$

By the Law of Total Probability (Theorem 2.11) with $A_1 = C$ and $A_2 = C^c$, the probability that a baby born alive in the United States will survive its first 28 days is

$$\begin{aligned}\mathbb{P}(B) &= \sum_{i=1}^{2} \mathbb{P}(B|A_i)\mathbb{P}(A_i) \\ &= \mathbb{P}(B|C)\mathbb{P}(C) + \mathbb{P}(B|C^c)\mathbb{P}(C^c) \\ &= (0.99823)(0.33) + (0.99938)(0.67) \\ &= 0.9990005\end{aligned}$$

A much shorter approach would be use first the Law of Total Probability to compute $\mathbb{P}(B^C)$ then use Complementary Property afterwards. Proceeding accordingly, we have

$$\begin{aligned}\mathbb{P}(B^c) &= \sum_{i=1}^{2} \mathbb{P}(B^c|A_i)\mathbb{P}(A_i) \\ &= \mathbb{P}(B^c|C)\mathbb{P}(C) + \mathbb{P}(B^c|C^c)\mathbb{P}(C^c) \\ &= \frac{1.77}{1000} \cdot 0.33 + \frac{0.62}{1000} \cdot (1 - 0.33) \\ &= 0.0009995.\end{aligned}$$

Thus, be Complementary Property,

$$\begin{aligned}\mathbb{P}(B) &= 1 - \mathbb{P}(B^c) \\ &= 1 - 0.0009995 \\ &= 0.9990005.\end{aligned}$$

2.6 Rule of Bayes

Consider the same information provided in Example 2.11. Suppose a baby born alive for the first 28 days. What is the probability that this baby is born using Caesarean section?

One way to answer this question is by using what we call the *Rule of Bayes*.

Theorem 2.12 (Rule of Bayes)**.** *Let $A_1, A_2, \ldots, A_n$ be a sample space partition*[8] *of S and $\mathbb{P}(A_i) > 0$ for $i = 1, 2, \ldots, n$. For any event B with $\mathbb{P}(B) > 0$,*

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

When $n = 1$, the Rule of Bayes to the following corollary that is commonly called the *Bayes Theorem*.

Corollary 2.1 (Bayes' Theoremm)**.** *Let A and B be events such that $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. Then*

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

Example 2.12. Select a number at random from $1, 2, \ldots, n$. Call your number m. Now select a second number at random from $1, 2, \ldots, m$.

a. Give an expression for the probability that the second number selected is 1.
b. Give an expression for the probability that the first number was n given that the second number selected was $n - 1$.

Solution. Let the events $F_1, F_2, \ldots, F_n$ correspond to the first number selected being $1, 2, \ldots, n$ respectively. Similarly, let $S_1, S_2, \ldots, S_n$ correspond to the second number selected being $1, 2, \ldots, n$ respectively.

[8] Recall this from Definition 2.6.

a. The desired probability is $\mathbb{P}(S_1)$. Using the law of total probability,

$$\begin{aligned}\mathbb{P}(S_1) &= \sum_{i=1}^{n} \mathbb{P}(S_1|F_i)\mathbb{P}(F_i) \\ &= \sum_{i=1}^{n} \frac{1}{i} \cdot \frac{1}{n} \\ &= \frac{1}{n}\left[1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right].\end{aligned}$$

b. The desired probability is $\mathbb{P}(F_n|S_{n-1})$. Using the rule of Bayes,

$$\begin{aligned}\mathbb{P}(F_n|S_{n-1}) &= \frac{\mathbb{P}(S_{n-1}|F_n)\mathbb{P}(F_n)}{\sum_{j=n-1}^{n} \mathbb{P}(S_{n-1}|F_j)\mathbb{P}(F_j)} \\ &= \frac{\mathbb{P}(S_{n-1}|F_n)\mathbb{P}(F_n)}{\mathbb{P}(S_{n-1}|F_{n-1})\mathbb{P}(F_{n-1}) + \mathbb{P}(S_{n-1}|F_n)\mathbb{P}(F_n)} \\ &= \frac{n-1}{2n-1}.\end{aligned}$$

2.7 Independence

One of most important concepts in Mathematics Statistics is the concept of independence. Many theoretical results Statistics are based on the assumption that two events are independent. To discuss this idea, we start by recalling from the definition of conditional probability that

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

Note that if $\mathbb{P}(A|B) = \mathbb{P}(A)$, one can conclude that the occurrence or nonoccurrence of event B has no effect on the probability that event A occurs. In this case,

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

and we say that events A and B are independent. This concept is more formally defined as follows.

Definition 2.10 (Independent Events)**.** Events A and B are *independent* if and only if

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

Remark. The following (mathematical) statements are equivalent.

a. A and B are independent events.

b. $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
c. $\mathbb{P}(A|B) = \mathbb{P}(A)$.
d. $\mathbb{P}(B|A) = \mathbb{P}(B)$.

Remark. The last two statements capture the essence of the independence of two events: the occurrence (or nonoccurrence) of one event does not affect the probability of another event occurring. Events that are not independent are said to be *dependent*.

Example 2.13. A single card is drawn at random from a 52-card deck. Let the event H be that the suit of the card is hearts. Let the event Q be that the rank of the card is a queen. Are the events H and Q independent?

Solution:

Since there are 13 hearts, the probability that the card is a heart is

$$\mathbb{P}(H) = \frac{13}{52} = \frac{1}{4}.$$

Since there are 4 queens, the probability that the card is a queen is

$$\mathbb{P}(Q) = \frac{4}{52} = \frac{1}{13}.$$

Since there is only queen of heaerts, the probability of $H \cap Q$ is

$$\mathbb{P}(H \cap Q) = \frac{1}{52}.$$

The event H and Q are independent because

$$\mathbb{P}(H \cap Q) = \frac{1}{52} = \frac{1}{4} \cdot \frac{1}{13} = \mathbb{P}(H)\mathbb{P}(Q).$$

The indepenence of H and Q are consistent with common sense: the probability that the card is a heart is not altered based on whether the card is a queen, and vice-versa.

Example 2.14 has shown that pairwise independence of events does not imply that three events will satisfy a definition of independence similar to Definition 2.10. So with more than two events, the definition of independence must be a bit more complicated.

Definition 2.11 (Mutually Independent Events)**.** Let $\{A_1, A_2, \dots, A_n\}$ be a collection of events. We say that the events of these collection are *mutually independent* if and only if for every sub-collection $\{A_{i_1}, A_{i_2}, \dots, A_{j_k}\}$, $k = 2, 3, \dots, n$, the following holds

$$\mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} \mathbb{P}\left(A_{i_j}\right).$$

Remark. Note that it's possible that a collection of events of be pairwise independent but not mutually independent as shown in the following example.

Example 2.14. A fair coin is tossed twice. Show that the events

$$A : \text{the first toss yields heads}$$
$$B : \text{the second toss yields heads}$$
$$C : \text{the two tosses yield different results,}$$

as pairwise independent but $\mathbb{P}(A \cap B \cap C) \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$.

Solution:

The probabilities for the individual events are all equal: $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$. The pairwaise probabilities are also all equal:

$$\mathbb{P}(A \cap B) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(B)$$
$$\mathbb{P}(A \cap C) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(C)$$
$$\mathbb{P}(B \cap C) = \frac{1}{4} = \mathbb{P}(B)\mathbb{P}(C)$$

Thus, the three events are pairwise independent. Finally, $\mathbb{P}(A \cap B \cap C) = 0$, which differs from $\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = \frac{1}{8}$.

Example 2.15 (Naive Bayes Classfier)**.** One of the practical applications of the concept of independence and conditional probability is the machine learning model called *Naive Bayes Classifier*. For details of this model, click the following links:

1. Wikipedia - Naive Bayes classifider.
2. MonkeyLearn Blog - A practical explanation of a Naive Bayes classifier

Chapter 3

Random Variables

3.1 Definition of Random Variable

The problems solved in the previous chapter all seemed to have custom solutions tailored specifically for the random experiment in question. The purpose of this chapter is to define a random variable that will force probability problems to look more similar to one another. Random variables are, by their nature, real-valued. This means that the qualitative outcomes such as heads/tails, even/odd, red/blue, Democrat/Republican found in the sample spaces common to problem settings like the previous chapter must be mapped into real values in order to use random variables to describe them. Although this switch requires a bit more mental effort on most problems, we will now be able to use the familiar tools of: Algebra and Calculus to solve probability problems.

The idea behind the use of random variables is to formulate a *rule* or *function* that assigns a real number x to each element of the sample space S.

Definition 3.1 (Random Variable)**.** Given a random experiment with an associated sample space S, a *random variable* is a function X that assigns to each element $s \in S$ one and only one real number $X(s) = x$. The *support* of X is the set of real numbers $\mathcal{A} = \{x \in \mathbb{R} | x = X(s), s \in S\}$.

Remark. From Definition 3.1, we have the following remarks:

1. A random variable *truly* is a function. Its domain is the sample space and its range is a set of real numbers $\mathcal{A}$.
2. A more compact and more intuitive definition for a random variable is that it is a variable whose value is subject to chance.

3. It is common practice to drop the argument s in $X(s)$ and simply write a random variable as just the function name, in this case X.

    a. Although this practice results in a much more compact expression, it is important to remember that the notation $X(s) = x$ from Definition 3.1 indicates that the element of the sample space named s is being mapped by the function X to a real number x.
    b. It is also common practice to use upper-case letters such as X, Y, or Z for random variables.
    c. If there are several of these random variables in a particular probablity problem, then they are often subscripted, for example, $X_1, X_2, \ldots, X_n$.

The sequence from left to right in Figure 3.1 illustrates the concept of a random variable.

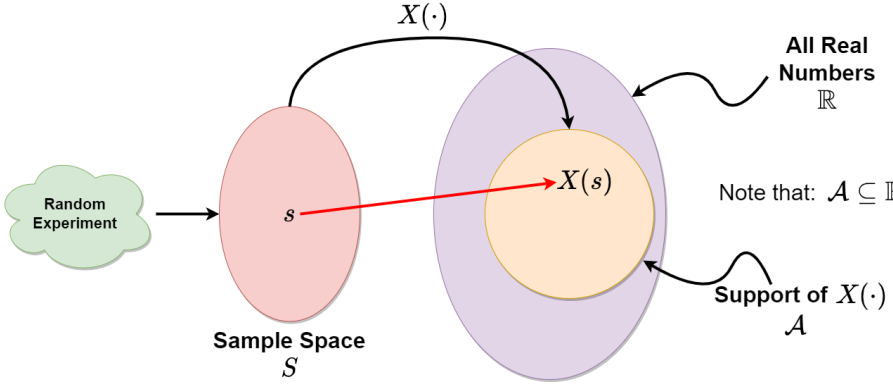


Figure 3.1: A random variable X associated with a random experiment.

The process begins with a random experiment on the left. This could be flipping a coin, rolling a die, asking someone whether they liked the service at a particular restaurant, or crashing a car Into a wall. In any case, this leads to the set of all possible outcomes, the sample space S. Each element s in S will be mapped to a single real value $X(s)$ in the support set $\mathcal{A}$ by the function X, the random variable.

3.2 Events defined by Random Variables

Recall from a remark in Chapter 1 the following sets written in what we call *simplified notation*:

Original	Simplified (assuming ω in context)
$\{\omega : 0 \leq X(\omega) \leq 10\}$	$\{1 \leq X \leq 10\}$
$\{\omega : X(\omega) > 12\}$	$\{X > 12\}$
$\{\omega : 0 \leq X(\omega) < 1,\ -2 \leq Y(\omega) \leq 2\}$	$\{0 \leq X < 1,\ -2 \leq Y \leq 2\}$
$\{\omega : X(\omega) \in (0, 1]\}$	$\{X \in (0, 1]\}$

If we define a sample space S as the set *in context* here, i.e., $\omega \in S$, then we call the sets above as events - the very same terminology we used in Chapter 2 upon which probabilities can be computed. To illustrate this, consider the following example.

Example 3.1. Suppose we toss a fair coin twice and we want to know the probability that the outcome results to heads. The sample space here is

$$S = \{HH, HT, TH, TT\}$$

Solution using Axiomatic Approach:

Since the coin is fair, classical viewpoint assigns equal probability values to each of the sample points. Define the E as the event that the outcome results to two Heads. Thus, the probability the outcome results to both heads is $\mathbb{P}(HH) = \frac{1}{4}$.

Solution using Random Variables:

We can also solve the problem in a different but equivalent way if we let the random variable $X(\cdot)$ be the number of heads, that is

$$X(s) = \begin{cases} 2 & \text{if } s = HH \\ 1 & \text{if } s = HT \text{ or } TH \\ 0 & \text{if } s = TT \end{cases}$$

where the support (range) of X is $\mathcal{A} = \{0, 1, 2\}$

Now observe that as a set, event E is equal to the set $\{HH\}$. However, observe that

$$\{s \in S \,|\, X(s) = 2\} = \{HH\} = E.$$

Therefore,

$$\mathbb{P}(E) = \mathbb{P}\left(\{s \in S \,|\, X(s) = 2\}\right) = \frac{1}{4}.$$

Since we know by context what the sample space is, we can simplify the notation further as

$$\mathbb{P}(E) = \mathbb{P}(\{X = 2\}) = \mathbb{P}(X = 2) = \frac{1}{4}.$$

In summary, the "two yet equivalent worlds" involved are as follows:

$S = \{HH, HT, TH, TT\}$	$\mathcal{A} = \{x \mid x = 0, 1, 2\}$
$E = \{HH\}$	$\{s \in S \mid X(s) = 2\} = \{X = 2\}$
$\mathbb{P}(E) = \frac{1}{4}$	$\mathbb{P}(X = 2) = \frac{1}{4}$

These relationships are illustrated in Figure 3.2

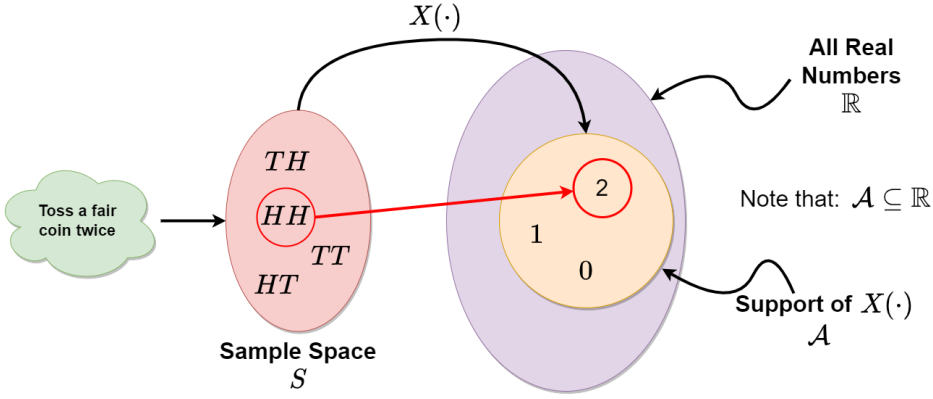


Figure 3.2: The random variable X, the number of heads in two tosses of a fair coin.

3.2.1 Defining your own random variable

When defining random variables, one must consider four things:

1. Know what is the characteristic or value of interest that is to be observed or measured.
2. Know the support (i.e. the range of values the random variable assumes/takes).
3. Know the probability distribution over the support.

Example 3.2. Suppose our random experiment involves tossing a die and observing the dots in the up face. The sample space should have elements that look like the following:

$$S = \left\{ \boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot \right\}$$

Since the value of interest is the *number of dots* on the up face, we could define a random variable X such that

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = \boxdot \\ 2 & \text{if } \omega = \boxdot \\ 3 & \text{if } \omega = \boxdot \\ 4 & \text{if } \omega = \boxdot \\ 5 & \text{if } \omega = \boxdot \\ 6 & \text{if } \omega = \boxdot \end{cases}$$

From this, we could see that the support of X is

$$\mathcal{A} = \{1, 2, 3, 4, 5, 6\}.$$

If the die is fair, we could use the classical viewpoint to assign the following probability distribution:

x	1	2	3	4	5	6
$\mathbb{P}(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

3.3 Discrete Random Variables

Let

3.4 Continuous Random Variables

3.5 Cumulative Distribution Functions

3.6 Expected Values

3.7 Inequalities (Markov, Chebyshev)