Contents

Chapter 1

Probability

1.1 Basics

The foundations of mathematical probability lie in set theory, which at this point, we already understand fairly well. In general, when reasoning about probability, we must consider the set of all possible outcomes and the likelihood of each one; the formal definitions are given below.

Definition 1. *A* sample space *is a non-empty countable set. An* outcome *is an element of a sample space, and an* event *is a subset of the sample space (i.e., a set of outcomes). If A is an event of a sample space S, then we let $\overline{A} = S \setminus A$ denote the* complement *of A.*

Now recall that if a, b are real numbers and $a \leq b$, then $[a, b]$ denotes the set $\{x \in \mathbb{R} : a \leq x \leq b\}$.

Definition 2. *If S is a sample space, then a* probability function *on S is a total function $\Pr : S \to [0, 1]$ that satisfies $\sum_{x \in S} \Pr(x) = 1$. If $x \in S$ is an outcome, then $\Pr(x)$ denotes the* probability *of x. If $A \subseteq S$ is an event, then $\Pr(A)$ is defined as $\sum_{x \in A} \Pr(x)$ and denotes the probability of A.*

As we can see, a probability function specifies a value $\Pr(x) \in [0, 1]$ for every outcome $x \in S$. Furthermore, the requirement $\sum_{x \in S} \Pr(x) = 1$ is equivalent to $\Pr(S) = 1$; this formalizes our intuition that no matter what happens, the outcome will definitely be an element of S.

1.1.1 Probability Rules

Now that we have formalized the definitions related to probability, we can start studying some fundamental rules that allow us to reason about probabilistic events. Intuitively, we can reason about these rules as follows: the sample space S is a large dartboard, and an event A is a small region of the dartboard. Then the value of $\Pr(A)$ represents the ratio of the area of A to the area of S. Throughout the following, we let A and B denote events, i.e., subsets of a sample space S.

1. **Sum Rule:** If A and B are disjoint events, then $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

2. **Inclusion-Exclusion:** For any events A and B, $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$. In our dartboard analogy, this corresponds to finding the area of $A \cup B$, and so the rule follows from the principle of inclusion-exclusion that we saw in our lecture on combinatorics. Notice that the sum rule is a special case of inclusion-exclusion: if A and B are disjoint, then $\Pr(A \cap B) = 0$. We now discuss two other special cases of inclusion-exclusion.

(a) **Boole's inequality:** $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$. This follows directly from inclusion-exclusion because $\Pr(A \cap B) \geq 0$.

(b) **Union bound:** For any set of events, the probability of their union is at most the sum of the probability of each event. In other words, if $A_1, A_2, \ldots$ are events, then

$$\Pr(A_1 \cup A_2 \cup \cdots) \leq \Pr(A_1) + \Pr(A_2) + \cdots .$$

Note that this bound holds for a finite set of events, as well as an infinite set.

3. **Difference Rule:** $\Pr(A \setminus B) = \Pr(A) - \Pr(A \cap B)$. This rule follows by observing that any event A can be partitioned into the disjoint events $A \setminus B$ and $A \cap B$.

(a) **Complement Rule:** $\Pr(\overline{B}) = 1 - \Pr(B)$. This rule follows from the difference rule by setting $A = S$, in which case $A \setminus B = \overline{B}$ and $\Pr(A) = \Pr(S) = 1$.

(b) **Monotonicity Rule:** If $A \subseteq B$, then $\Pr(A) \leq \Pr(B)$. This rule is known as monotonicity because it states that the Pr function does not decrease if we add outcomes to the event.

In everyday life, we often use these rules without noticing. However, as we have seen, informal justifications of probability can lead to incorrect results, so a formal mathematical understanding of these rules is critical.

1.2 Conditional probability

In reality, it is common that we want to know the probability of an event A given some information B, written as $\Pr[A|B]$.

Definition 3. $\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}$

Corollary 1. $\Pr[A \cap B] = \Pr[A|B] \cdot \Pr[B] = \Pr[B|A] \cdot \Pr[A]$

Theorem 2. *By definition and Corollary 1,* $\Pr[A|B] = \frac{\Pr[B|A] \cdot \Pr[A]}{\Pr[B]}$. *This is known as* Bayes' Rule.

Corollary 3. $Pr[A|B] = \frac{\Pr[B|A] \cdot \Pr[A]}{\Pr[B|A] \cdot \Pr[A] + \Pr\left[B|\overline{A}\right] \cdot \Pr\left[\overline{A}\right]}$

Proof.

$$\begin{aligned}
&\Pr[A|B] \\
=&\frac{\Pr[B|A] \cdot \Pr[A]}{\Pr[B]} && \text{(By Bayes' Rule)} \\
=&\frac{\Pr[B|A] \cdot \Pr[A]}{\Pr[B \cap A] + \Pr\left[B \cap \overline{A}\right]} \\
=&\frac{\Pr[B|A] \cdot \Pr[A]}{\Pr[B|A] \cdot \Pr[A] + \Pr\left[B|\overline{A}\right] \cdot \Pr\left[\overline{A}\right]} && \text{(By Bayes' Rule)}
\end{aligned}$$

□

Example 1: Let's revisit the Monty Hall problem now that we know the concept of conditional probability. Recall that there are 3 closed doors A, B, and C for the guest to pick and behind one door is a car, while behind the other two are goats. After the guest picks a door, the host opens one of the two remaining doors to reveal a goat. Then, the guest is given the option to switch the door they picked. We saw that the guest is more likely to win by switching. Let's define some events formally.

$$X: \text{ guest wins car by switching}$$
$$Y: \text{ car is at location } A \text{ and there is a goat at } B$$
$$Z: \text{ guest picks door } A \text{ and host reveals a goat at } B$$

Event Y happens when the car is at location A, since both other locations will necessarily have goats. Thus, $\Pr(Y) = 1/3$. Event X and Y both happen when the guest does not originally pick door A, but then switches to door A. Thus, the guest could pick door B or C. This could happen 2 ways with probability $\Pr(X \cap Y) = 2/9$. Now we can compute the probability of the guest winning by switching given that the car is at A:

$$\Pr(X|Y) = \frac{\Pr(X \cap Y)}{\Pr(Y)} = \frac{2/9}{1/3} = 1/3.$$

Let's consider event Z. There are two ways for this event to occur. The location of the car is A and the guest picks A, then the host reveals goat at B is one way with probability $1/18$. The location of the car is at A, the guest picks C, then host reveals goat at B occurs with probability $1/9$. Thus,

$$\Pr(X|Z) = \frac{P(X \cap Z)}{P(Z)} = \frac{1/9}{1/9 + 1/18} = 2/3.$$

The calculation of $\Pr(X|Y)$ includes the outcome that the host opens door C, which included some extraneous outcome in our calculation. In reality, we are interested in the probability of winning by switching given the guest picks door A and the host opens door B. This aligns with what we calculated last time– that the guest's best strategy is to switch.

Let's consider another example of conditional probability leading to a counter-intuitive result.
Example 2: Suppose we have a test that tells us whether a person is sick with a high degree of accuracy. For a person who is healthy, the test will likely be negative and for a person who is sick, the test will likely be positive. We define the following events.

$$A: \text{ test is positive} \qquad \overline{A}: \text{ test is negative}$$
$$B: \text{ person is healthy} \qquad \overline{B}: \text{ person is sick}$$

We have the following tree in Figure 1.1 showing the probability that someone is healthy and the probability of each test result. A positive test result for a healthy person is known as a false positive, and a negative test for a sick person is known as a false negative.

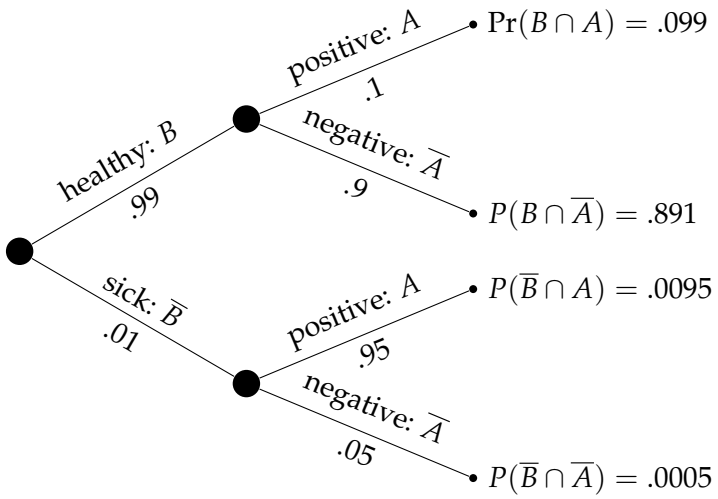


Figure 1.1: Tree Diagram for Example 2.

Now, if we run the test and the test is positive, intuitively we expect that the person is sick. In other words, we expect the number of false positives to be low. That is, we expect $\Pr(B|A)$ to be low and $Pr(\overline{B}|A)$ to be high. Let's see if this is the case.

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{.059}{.1085} = .544$$
$$\Pr(\overline{B}|A) = \frac{\Pr(A \cap \overline{B})}{\Pr(A)} = \frac{.0095}{.1085} = .09$$

What happened? If we were to test a random person and the test is positive, it is more likely that they are healthy than they are sick. This is the result of the fact that the vast majority of people are healthy. Fortunately, medical tests are not usually run on random people, but on those showing symptoms of being sick!

Example 3: Suppose we want to know the probability of Duke winning the NCAA tournament. We define the following events.

A : Duke wins the NCAA tournament
B : Duke beats UNC in the ACC tournament

Suppose we are given that $\Pr(B) = .75$, $\Pr(A|B) = .99$, and $\Pr(A|\overline{B}) = .25$. Now, if we know that Duke won the NCAA tournament, what is the probability they beat UNC? We can use the above formula:

$$\Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A|B)\Pr(B) + \Pr(A|\overline{B})\Pr(\overline{B})} = \frac{(.99)(.75)}{(.99)(.75) + (.25)(.25)} \approx 0.922$$

1.3 Independent Events

Definition 4. *Events A and B are said to be* independent *if* $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$

Corollary 4. *By Bayes' Rule, if A and B are independent,* $\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \Pr[A]$. *Similarly,* $\Pr[B|A] = \Pr[B]$.

We now generalize the definition to more than 2 events.

Definition 5. *Suppose we have events $A_1, \ldots, A_n$ and k is a positive integer such that $k \leq n$. These events are said to be k-*wise independent *if, for any subset with size $\leq k$, the subsets are independent.*

$$\Pr\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_j}\right) = \Pr(A_{i_1}) \Pr(A_{i_2}) \ldots \Pr\left(A_{i_j}\right) \quad \textit{for any } j \leq k.$$

Definition 6. *Suppose we have events $A_1, \ldots, A_n$. These events are said to be* mutually independent *if they are n-wise independence.*

Intuitively, a set of events is mutually independent if the probability of each event is the same no matter which of the other events has occurred.

Example 4: Suppose we have three fair coins, and we toss all of them such that each toss is independent. We will define the following events:

$$\begin{aligned} A_1 &: \text{ Coins 1 and 2 have the same result} \\ A_2 &: \text{ Coins 2 and 3 have the same result} \\ A_3 &: \text{ Coins 1 and 3 have the same result} \end{aligned}$$

We will show that these events are 2-wise independent, often called *pairwise independent*. $\Pr(A_1) = \Pr(A_2) = \Pr(A_3) = 1/2$. Then,

$$\Pr(A_1 \cap A_2) = 1/4$$

However, consider the event $A_1 \cap A_2 \cap A_3$. There are two ways for this to happen: all heads or all tails. So this occurs with probability $1/4$. However, $\Pr(A_1) \Pr(A_2) \Pr(A_3) = 1/8$. Thus, these events are not 3-wise independent. Note that $\Pr(A_3|A_1 \cap A_2) = 1$.

1.4 Permutations and Combinations

In this section, we will introduce the foundations of combinatorics, the branch of mathematics that deals with counting. In particular, we will study permutations and combinations, their relevant formulas, and some basic identities involving these concepts.

Permuting n items: Suppose that we want to place n distinct items in a line, that is, we want to *order* or *permute* the items. The most fundamental principle used in counting is the following: the number of ways to permute n distinct items is the product of the first n positive integers, which is denoted by $n!$ ("n factorial"). This can be proven, somewhat informally, as follows: there are n choices for the first item in line, $(n-1)$ for the second, and so on, until there's only 1 choice left for the last item in line. Since the choices can be made in sequential order, the total number of possible permutations (orderings) is $n(n-1) \cdots 1 = n!$.

Permuting k of n items: Now instead of permuting all n items, suppose we only want to permute k of them (where k is a positive integer less than n). By the same argument above, the number of ways to do this is $n(n-1)\cdots(n-(k-1))$. Notice that there are k terms in this product, because each term corresponds to selecting an item to place next in the final ordering of k items. For convenience, notice that a more concise way of writing this product is the following:

$$n(n-1)\cdots(n-(k-1)) = n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

Choosing k of n items: Now instead of permuting k items, suppose we only want to *choose* k items. In other words, we are finding a *subset* of size k, rather than a *sequence* of length k. We denote this value by $\binom{n}{k}$ ("n choose k"), and we often say a subset is a *combination*. In other words, a combination can be thought of as a permutation in which order doesn't matter.

As we just saw, there are $n!/(n-k)!$ permutations of length k. However, each subset of size k is represented as a permutation $k!$ times. Therefore, we can conclude that:

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

Alternatively, we can reason about these formulas as follows. Let k be any positive integer at most n. Then permuting n items is the same as choosing a subset of size k, ordering them, and then ordering the remaining elements. This line of reasoning results in the following:

$$n! = \binom{n}{k} \cdot k! \cdot (n-k)!$$

which is equivalent to the previous expression.

Remark: If $n = 0$, then the number of ways to order n items is somewhat ambiguous. Similarly, if $k = 0$, then the number of ways to choose k items is somewhat ambiguous. To deal with these ambiguities in a consistent way, we have the following conventions:

$$0! = 1 \qquad \text{and} \qquad \binom{n}{0} = 1.$$

1.4.1 Stars and Bars

The notion of combinations is fundamental to combinatorics.To better familiarize ourselves with combinations, we now look at one application known as "stars and bars".

The setup is the following: suppose there are three children c_1, c_2, c_3, and we must distribute 10 identical candies among these three children. Each child can receive any number of candies, including 0. For example, one possible distribution is $(4, 3, 3)$: in this case, c_1 receives 4 candies, c_2 receives 3, and c_3 receives 3. How many ways can we distribute the candies?

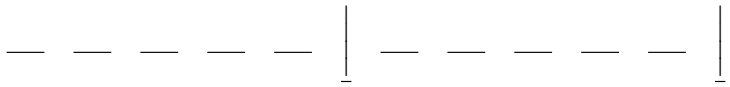
The key observation is the following: we can distribute the candies by arranging them in a line, and then placing two "bars" somewhere along the line. For example, the $(4, 3, 3)$ described above can be modeled by the following:

Each ★ represents a candy, and the two location of the two bars determines the distribution of the candies. Notice that the following distribution is also possible:

The above diagram corresponds to the distribution $(0, 0, 10)$. In general, c_1 receives the candies left of the first bar, c_2 receives the candies between the two bars, and c_3 receives the candies right of the second bar.

So we can see that distributing candies is identical to choosing the location of the two bars. However, the number of ways to do this is *not* $\binom{11}{2}$, because this ignores the possibility of placing two bars next to each other. Instead, we should think of the process as follows: there are 12 empty slots, pictured below.

We must place bars in two of the slots, and the remaining 10 slots will then represent the 10 candies to distribute. For the distribution $(5, 5, 0)$, the diagram becomes the following:



From this perspective, it becomes clear that the number of distributions of 10 candies to 3 children is $\binom{12}{2}$. In general, if there are n candies and k children, then there are $n + k - 1$ slots, and we must place $k - 1$ bars. The remaining n candies, interspersed among the bars, represent a distribution. Thus, the number of distributions is $\binom{n+k-1}{k-1}$.

1.4.2 Combinatorial Identities

We now state a couple of basic identities involving combinations.

Fact 5. *Let n be a positive integer, and k be in $\{1, \ldots, n\}$. Then the following identities hold:*

$$\binom{n}{k} = \binom{n}{n-k} \quad \textit{and} \quad \binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

The latter is known as Pascal's rule*, named after the mathematician Blaise Pascal.*

Remark: One proof of these identities is purely algebraic: if we simply use the formula

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

on every term, then a bit of algebraic manipulation proves the identity. However, in some sense, this proof is not very elegant because it does not take advantage of the combinatorial interpretation of the terms. Therefore, we now prove each identity by giving a combinatorial interpretation of both sides. Under these interpretations, the validity of the equalities becomes clear.

Proof. Let A be a set with n elements, where n is a positive integer. As discussed above, for each equality, we will give a combinatorial interpretation for both the left-hand side (LHS) and the right-hand side (RHS).

In the first equality, the LHS counts the number of subsets of A that contain exactly k items. *Selecting* k items is equivalent to *excluding* $n-k$ items. The number of ways to exclude $n-k$ items from S is precisely the RHS of the equality.

In the second equality, the LHS still counts the number of subsets of A that contain exactly k items. Since n is positive, we can fix a particular element of A which we denote by x. The set of subsets of size k can be partitioned into two sets: those that contain x, and those that do not. The number of subsets of size k that contain x is $\binom{n-1}{k-1}$. This is because once you choose x in the subset, you are left to make $k-1$ choices out of the remaining $n-1$ elements. On the other hand, the number of subsets of size k that exclude x is $\binom{n-1}{k}$. The logic behind this expression is that if you exclude x from the subset, then you are still left to make all the k choices, but only among the remaining $n-1$ elements. Summing these two values gives the number of ways to choose *any* subset of size k, as desired. □

A pictorial representation of Pascal's rule (see Fact 5) is known as *Pascal's triangle*. The triangle is constructed as follows: the top row, which we consider row 0, contains a single 1. Each subsequent row starts and ends with 1, and the internal terms are obtained by summing the closest two terms in the previous row. The first 7 rows are pictured below, where n represents the row index:

$n=0$:						1						
$n=1$:					1		1					
$n=2$:				1		2		1				
$n=3$:			1		3		3		1			
$n=4$:		1		4		6		4		1		
$n=5$:	1		5		10		10		5		1	
$n=6$:	1	6		15		20		15		6		1

Now consider the values in row n. We claim that these values are precisely, in order, the following:

$$\binom{n}{0} \quad \binom{n}{1} \quad \cdots \quad \binom{n}{n-1} \quad \binom{n}{n}.$$

The k-th term of this row (where $k \in \{0, 1, \ldots, n\}$) is precisely $\binom{n}{k-1}$—the proof of this is a simple induction argument that follows easily from Pascal's identity.

We now state and prove another well-known identity; this one gives a combinatorial interpretation of the coefficients of a binomial expression. Due to this connection, the values $\binom{n}{k}$ are often called the *Binomial coefficients*.

Theorem 6 (Binomial Theorem)**.** *For all $n \in \mathbb{Z}^+$ and $a, b \in \mathbb{R}$,*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k.$$

In other words, for every $k \in \{0, 1, \ldots, n\}$, the coefficient of $a^{n-k}b^k$ in the expansion of $(a+b)^n$ is $\binom{n}{k}$.

As an example, let us consider the case when $n = 4$ and $k = 3$. Then the coefficient we seek is that of a^1b^3. There are precisely four ways to obtain a^1b^3 by expanding $(a+b)^4$: $abbb, babb, bbab$, and $bbba$. Each of these ways corresponds to choosing the location of 3 b's from four slots, and the number of ways to do this is precisely $\binom{4}{3}$.

Now we consider the general case. Let $k \in \{0, 1, \ldots n\}$, and consider the term $a^{n-k}b^k$. This term is the product of $(n-k)$ a's and k b's. Furthermore, notice that $(a+b)^n$ is simply

$$(a+b)(a+b)\cdots(a+b),$$

which is the product of n copies of $(a+b)$. Upon expansion of this product, we obtain the sum of a $n+1$ terms, and each term is the product of n variables, distributed as a's and b's.

Thus, we can think of "building" the coefficient of $a^{n-k}b^k$ as selecting k of the $(a+b)$ terms in $(a+b)^n$ that will contribute a b. The number of ways to do this is precisely $\binom{n}{k}$, as desired.

Corollary 7. *For all $n \in \mathbb{Z}^+$, the following equality holds:*

$$2^n = \sum_{k=0}^{n} \binom{n}{k}$$

In other words, the sum of the values in the n-th row of Pascal's triangle is 2^n.

Remark: The above identity simply follows by $a = b = 1$ in Theorem 6. We give an alternate proof below that relates this identity to the set of subsets of a set.

Proof. Let A be a set with n elements, and let A_k denote the subset of the power set 2^A containing the subsets of A of size k. Then the sets $A_0, A_1, \ldots, A_n$ partition 2^A, which means the following equalities hold:

$$\begin{aligned} 2^n = |2^A| &= |A_0| + |A_1| + \cdots + |A_n| \\ &= \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = \sum_{k=0}^{n} \binom{n}{k}. \end{aligned}$$

In other words, every subset of A has a size in $\{0, 1, \ldots, n\}$, so to count the number of subsets of A, we can count the number of subsets of each size over all possible sizes. □

1.5 Probabilistic Reasoning

Now that we have a list of rules, we can state a high-level strategy for reasoning about probabilistic events. Whenever faced with a probability problem, one should consider the following strategy:

1. Determine the sample space (i.e., the set of all possible outcomes).
2. Define the interesting event (i.e., the subset containing all interesting outcomes).
3. Calculate the probability of each outcome.
4. Calculate the probability of the event.

In this section, we will apply this strategy in several scenarios.

Rolling a Fair Die: Suppose we roll a fair 6-sided die once; let X denote the outcome of this roll. Here, the sample space is simply $\{E_1, E_2, E_3, E_4, E_5, E_6\}$, where E_i denotes the outcome "$X = i$". Since the die is fair, each outcome occurs with probability $1/6$; in other words,

$$\Pr(E_1) = \Pr(E_2) = \cdots = \Pr(E_6) = \frac{1}{6}.$$

Now let us formally calculate the probability that X is even. If we let A denote this event, then $A = \{E_2, E_4, E_6\}$. Since each outcome in A occurs with probability $1/6$ and all outcomes are disjoint, we can apply the sum rule to obtain $\Pr(A) = 3/6 = 1/2$, which matches our intuition.

Now consider the event containing outcomes where X is even or prime. Let B denote this event, and notice that there are multiple ways of calculating B:

1. Sum Rule: Since $B = \{E_2, E_3, \ldots, E_6\}$ and all outcomes are disjoint, we can conclude $\Pr(B) = \Pr(E_2) + \Pr(E_3) + \cdots + \Pr(E_6) = 5/6$.

2. Complement Rule: Notice that $\overline{B} = \{E_1\}$, so $\Pr(\overline{B}) = 1/6$. This implies $\Pr(B) = 1 - 1/6 = 5/6$.

3. Inclusion-exclusion: Recall that B must capture the event that X is even or prime. Therefore, $B = V \cup P$ where $V = \{E_2, E_4, E_6\}$ (evens) and $P = \{E_2, E_3, E_5\}$ (primes), and so $V \cap P = \{E_2\}$. By inclusion-exclusion, we have $\Pr(B) = \Pr(V) + \Pr(P) - \Pr(V \cap P) = 3/6 + 3/6 - 1/6 = 5/6$.

Rolling an Unfair Die: Now let's roll another 6-sided die, and let Y denote the outcome of this roll. The sample space is still $\{E_1, \ldots E_6\}$, where E_i denotes the outcome "$Y = i$". However, unlike the previous die, this die is not fair: for every $i \in \{1, 2, \ldots, 5\}$, this die is twice as likely to roll i than $i + 1$. In other words, the die obeys the following probability function:

$$\Pr(E_i) = 2 \cdot \Pr(E_{i+1}) \quad \forall i \in \{1, \ldots, 5\}.$$

Since exactly one of the 6 outcomes must still occur, and all of the outcomes are disjoint, the probability distribution still obeys the following equality:

$$\Pr(E_1) + \Pr(E_2) + \Pr(E_3) + \Pr(E_4) + \Pr(E_5) + \Pr(E_6) = 1.$$

Now let us first calculate the probability of each of the 6 outcomes. Notice that if we let $p = \Pr(E_6)$, then $\Pr(E_5) = 2p$, and similarly, $\Pr(E_4) = 2 \cdot \Pr(E_5) = 4p$. This line of reasoning yields

$$32p + 16p + 8p + 4p + 2p + p = 1,$$

and solving this equation yields $p = 1/63$. Now we can solve for the probability of each outcome. In particular, $\Pr(E_1) = 32/63$, which is much larger than $1/6$. Furthermore, the probability that Y is even is now $\Pr(E_2) + \Pr(E_4) + \Pr(E_6) = 16/63 + 4/63 + 1/63 = 21/63 = 1/3$. Similarly, we can see that the probability that Y is even or prime is (using the complement rule) $1 - \Pr(E_1) = 31/63$.

The Birthday Paradox: We now study a phenomenon known as the *birthday paradox*. This actually isn't a paradox in the strictest sense of the word, because the result we derive will be mathematically rigorous. However, for somebody who has never seen the result, it may sound quite surprising.

The setup is the following: assume that there are n students in a class, and a year has d days. Of course, we know that $d = 365$, but by using the variable d, our analysis can apply to a more general setting (e.g., if we set $d = 30$, then this is solving the problem of only considering students born in April). Assuming that each student is equally likely to have been born on any of the d days, what is the probability that all n birthdays are distinct?

Let's fix an ordering of the students, and count the number of outcomes. In this case, an outcome is a sequence of length n, and each element is one of the d days. Thus, the total number of possible outcomes is d^n. Furthermore, since the birthdays are all independent from each other and identical, every outcome is equally likely. Thus, each outcome occurs with probability $1/d^n$.

Now let D denote the outcome that the birthdays are distinct. For the birthdays to be distinct, the first student can have any one of d birthdays, but then the second birthday only has $(d-1)$ possibilities. This reasoning continues until the n-th student only has $(d-(n-1))$ possibilities. Thus, the total number of outcomes with no repeated birthdays is $d(d-1)(d-2)\cdots(d-(n-1))$.

Since each outcome is equally likely, the probability of our outcome having distinct birthdays is the following:

$$\begin{aligned}\Pr(D) = \frac{d(d-1)(d-2)\cdots(d-(n-1))}{d^n} &= \frac{d}{d}\cdot\frac{d-1}{d}\cdot\frac{d-2}{d}\cdots\frac{d-(n-1)}{d} \\ &= \left(1-\frac{0}{d}\right)\left(1-\frac{1}{d}\right)\left(1-\frac{2}{d}\right)\cdots\left(1-\frac{n-1}{d}\right).\end{aligned}$$

We now make use of the bound $1 - x < e^{-x}$ for any positive real number x. (This can be proved by, say, considering the Taylor series for e^{-x}.) Applying this inequality to each term above yields the following:

$$\begin{aligned}\Pr(D) &< e^{-0/d}\cdot e^{-1/d}\cdot e^{-2/d}\cdots e^{-(n-1)/d} \\ &= e^{-(\sum_{i=1}^{n-1} i/d)} \\ &= e^{-n(n-1)/(2d)}.\end{aligned}$$

Notice that as n increases, this upper bound on $\Pr(D)$ decreases. Intuitively, this makes sense: as the number of students increases, the probability that all birthdays are distinct decreases. In fact, by the pigeonhole principle, if $n \geq d+1$ then $\Pr(D) = 0$.

Equipped with this bound, we can determine the value of n that is large enough to ensure $\Pr(D) < 1/2$. It is straightforward to verify that if $n \geq 25$, then $e^{-n(n-1)/(2\cdot 365)} < 0.44$. This means that in a class of only 25 students, it is more likely than not that two students share a birthday!

1.6 Random Variables

Definition 7. Random variables *are some function f that maps the sample space to a domain, typically a numerical domain.*

Example 1: Suppose we have four outcomes, c_1, c_2, c_3, and c_4 in a sample space, with probabilities $1/4, 1/2, 1/8, 1/8$, respectively. Now let X be a random variable such that $X(c_1) = 1$, $X(c_2) = 2$, $X(c_3) = 3$ and $X(c_4) = 4$. Then $\Pr[X = 1] = 1/4$.

We saw an example in which the random variable is discrete. We can also have continuous random variables, as shown in the next example.

Example 2: Suppose X is random variable that is uniformly chosen random real number between 0 and 1. The domain now is continuous, so we call it a continuous random variable.

Definition 8. Probability distribution *of a discrete random variable X is the probabilities of $X = x$ for all x in the sample space.*

Definition 9. Probability Density Function*, or pdf, of a continuous variable X is the probability that X is between a and b.*

Example 2: Suppose we toss two independent fair coins.

Define random variable $X = 1$ if the outcomes are different and $X = 0$ is they are the same. Then $\Pr[X = 0] = \Pr[X = 1] = 1/2$.

Define another random variable Y which equals to the number of heads. Then $\Pr[Y = 0] = 1/4, \Pr[Y = 1] = 1/2, \Pr[Y = 2] = 1/4$.

1.6.1 Expectation of a Random Variable

Suppose we have a discrete random variable X, and $\Pr[X = x_i] = P_i$ for $i = 1, ..., n$ and $\sum_{i=1}^{n} P_i = 1$. Then the expectation of X is defined as follows:

Definition 10. $\mathbb{E}[X] = \sum_{x \in X} x \cdot \Pr[x]$

Similarly for a continuous random variable X, the expectation of X is defined as follows:

Definition 11. $\mathbb{E}[X] = \int_{x \in D} x \cdot f(x)dx$*, where $f(x)$ is the pdf of X.*

Note that the expectation of a constant is the constant itself. Let $E[X] = \mu$. Since μ is a constant, $E[\mu] = \mu$ and thus $E[E[X]] = \mu$.

We can also define a random variable based on other random variables. For example, given a random variable Y, we can define a new random variable $X = 2Y$. Another example would be, given random variables X and Y, we can define a new random variable $Z = X + Y$. We have following lemmas in terms of their expected value.

Lemma 8. *For random variables X and Y and $X = cY$, where c is a constant, we have $E[X] = cE[Y]$.*

Proof.

$$\begin{aligned} E[X] &= \sum_x x \Pr[X = x] \\ &= 2 \sum_x \frac{x}{2} \Pr[X = x] \\ &= 2 \sum_y y \Pr[Y = y] \\ &= 2E[Y] \end{aligned}$$

□

Lemma 9. *For random variables X, Y, and Z, where $Z = X + Y$, we have $E[Z] = E[X] + E[Y]$.*

Proof.

$$\begin{aligned}
E[Z] &= \sum_x \sum_y (x+y) \Pr[X = x, Y = y] \\
&= \sum_x \sum_y x \cdot \Pr[X = x, Y = y] + \sum_x \sum_y y \cdot \Pr[X = x, Y = y] \\
&= \sum_x x \cdot \sum_y \Pr[X = x, Y = y] + \sum_y y \cdot \sum_x \Pr[X = x, Y = y] \\
&= \sum_x x \cdot \Pr[X = x] + \sum_y y \cdot \Pr[Y = y] \\
&= E[X] + E[Y]
\end{aligned}$$

□

Theorem 10. *$E[\sum_{i=1}^n a_i X_i] = \sum_{i=1}^n a_i E[X_i]$. This is known as the linearity of expectation.*

Proof. By lemma 9, if we have n random variables $X_1, ..., X_n$, then $E[\sum_{i=1}^n a_i X_i] = \sum_{i=1}^n E[a_i X_i]$. By lemma 8, we have $E[a_i X_i] = a_i E[x_i]$. Therefore, $\sum_{i=1}^n E[a_i X_i] = \sum_{i=1}^n a_i E[X_i]$. □

Note that the random variables don't have to be independent. The proof for the continuous case is similar.

Example 3: Suppose we have n coins. Let coin i have the behavior that H appears with probability P_i and T with probability $1 - P_i$. What is the expected number of heads when all coins are tossed?

Define $X_i = 1$ if the ith coin gives H and $X_i = 0$ otherwise. Then the total number of heads is exactly the sum of all X_i, i.e. $\sum_{i=1}^n X_i$. Therefore, $E[\sum_{i=1}^n X_i] = \sum_{i=1}^n E[X_i] = \sum_{i=1}^n (1 \cdot P_i + 0 \cdot (1 - P_i)) = \sum_{i=1}^n P_i$.

1.7 Variance of Random Variables

Recall that a *random variable* is a total function whose domain is a sample space S. The codomain of a random variable is often $\mathbb{R}$ or $\{0, 1\}$, but in general, it can be any set.

Let $X : S \to \mathbb{R}$ be a random variable. Recall that the *expectation* of X represents the "average value" of X, and can be written in the following two ways:

$$\mathbb{E}[X] = \sum_{\omega \in S} X(\omega) \cdot \Pr[\omega] = \sum_{x \in \mathbb{R}} x \cdot \Pr[X = x]. \tag{1.1}$$

Another important quantity associated with X is its *variance*, given as follows:

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}^2[X]. \tag{1.2}$$

(Note that $\mathbb{E}^2[X]$ denotes the quantity $(\mathbb{E}[X])^2$.) Intuitively, the variance of X measures how "spread out" the values of X are: if $\mathrm{Var}[X] = 0$, then X is a constant, and if $\mathrm{Var}[X]$ is high, then X often takes value far from its expectation. Finally, the *standard deviation* of X is the positive square root of its variance.

Theorem 11. *If X and Y are independent random variables, then*

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

Remark: This statement of this theorem is similar to that of linearity of expectation $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. However, for variance, independence of X and Y is required, whereas linearity of expectation holds regardless of whether the random variables in consideration are independent.

Proof. By the equation given in (1.2), we have

$$\text{Var}[X+Y] = \mathbb{E}[(X+Y)^2] - \mathbb{E}^2[X+Y].$$

Consider the first term of the expression above:

$$\mathbb{E}[(X+Y)^2] = \mathbb{E}[X^2 + 2XY + Y^2] = \mathbb{E}[X^2] + 2 \cdot \mathbb{E}[XY] + \mathbb{E}[Y^2], \tag{1.3}$$

and now the second:

$$\mathbb{E}^2[X+Y] = (\mathbb{E}[X+Y])^2 = \mathbb{E}^2[X] + 2 \cdot \mathbb{E}[X] \cdot \mathbb{E}[Y] + \mathbb{E}^2[Y]. \tag{1.4}$$

Since X and Y are independent, we know that $2 \cdot \mathbb{E}[XY] = 2 \cdot \mathbb{E}[X] \cdot \mathbb{E}[Y]$. Thus, subtracting (1.4) from (1.3) yields

$$\begin{aligned} \text{Var}[X+Y] &= \mathbb{E}[X^2] - \mathbb{E}^2[X] + \mathbb{E}[Y^2] - \mathbb{E}^2[Y] \\ &= \text{Var}[X] + \text{Var}[Y], \end{aligned}$$

where the second equality again holds from (1.2). □

Theorem 12. $\text{Var}[aX] = a^2\text{Var}[X]$.

Proof.

$$\begin{aligned} \text{Var}[aX] &= \mathbb{E}[(aX)^2] - \mathbb{E}^2[aX] \\ &= \mathbb{E}[a^2X^2] - (\mathbb{E}[aX])(\mathbb{E}[aX]) \\ &= a^2\mathbb{E}[X^2] - (a\mathbb{E}[X])(a\mathbb{E}[X]) \\ &= a^2(\mathbb{E}[X^2] - \mathbb{E}^2[X]) \\ &= a^2\text{Var}[X] \end{aligned}$$

□

1.7.1 Covariance

Definition 12. *For two random variables X and Y, their covariance $cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.*

Theorem 13. $\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2cov(X,Y)$

Proof.

$$\begin{aligned} &\text{Var}[X] + \text{Var}[Y] + 2cov(X,Y) \\ =&\mathbb{E}[X^2] - \mathbb{E}^2[X] + \mathbb{E}[Y^2] - \mathbb{E}^2[Y] + 2\mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y] \\ =&(\mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2]) - (\mathbb{E}^2[X] + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}^2[Y]) \\ =&\mathbb{E}[X^2 + 2XY + Y^2] - \mathbb{E}^2[X+Y] \\ =&\mathbb{E}[(X+Y)^2] - \mathbb{E}^2[X+Y] \\ =&\text{Var}[X+Y] \end{aligned}$$

□

Corollary 14. $\text{Var}[aX + bY + C] = a^2\text{Var}[X] + b^2\text{Var}[Y] + 2abcov(X, Y)$.

Definition 13. *The* standard deviation σ *of a random variable is defined as the square root of the variance, i.e.* $\sigma = \sqrt{\text{Var}[X]}$.

1.8 Independence of Random Variables

Definition 14. *X and Y are* independent random variables *if* $\Pr[X = x, Y = y] = \Pr[X = x] \cdot \Pr[Y = y], \forall x \in X, y \in Y$.

We then have following theorems.

Theorem 15. *If X and Y are independent random variables, then* $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

Proof.

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x \in D_x} \sum_{y \in D_y} xy \Pr[X = x, Y = y] \\
&= \sum_{x \in D_x} \sum_{y \in D_y} xy \Pr[X = x] \Pr[Y = y] \\
&= (\sum_{x \in D_x} x \Pr[X = x])(\sum_{y \in D_y} y \Pr[Y = y]) \\
&= \mathbb{E}[X] \cdot \mathbb{E}[Y]
\end{aligned}$$

□

Corollary 16. *If X and Y are independent random variables, then* $cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$, *and thus* $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$.

1.9 Common Distributions

1.9.1 Bernoulli Distribution

Definition

If

$$\Pr[X] = \begin{cases} 1 & w.p.p \\ 0 & w.p.1 - p \end{cases}$$

, then $X \sim$Bernoulli(p)

Expectation

$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p$

Variance

$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}^2[X] = (1^2 \cdot p + 0^2 \cdot (1 - p)) - p^2 = p - p^2 = p(1 - p)$

1.9.2 Binomial Distribution

Definition

Binomial(n, p) is the sum of n independent Bernoulli(p) random variables. Therefore, if $X \sim$Binomial(n, p), then $\Pr[X = i] = \binom{n}{i} p^i (1-p)^{n-i}$ for $0 \leq i \leq n$.

Expectation

$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np$ because $X_i \sim$Bernoulli(p).

Variance

$\text{Var}[X] = \sum_{i=1}^{n} \text{Var}[X_i] = np(1-p)$ because $X_i \sim$Bernoulli(p) and they are mutually independent.

1.9.3 Geometric Random Variables

Definition

Recall the following random process: we have a coin that results in H with probability p, and we repeatedly flip this coin until we obtain a head H. Thus, each outcome in the sample space is a sequence of (possibly 0) tails, followed by a single head. We can define a random variable X on this sample space as follows:

$$\forall s \in S.\, X(s) \text{ is equal to the length of } s.$$

Notice that $\Pr[X = x] = (1-p)^{x-1} p$ for every $x \in \mathbb{Z}^+$. Such a random variable X is often known as a *geometric* random variable with parameter p. (Intuitively, the value of p can be interpreted as the probability of "success," i.e., the probability that the experiment ends at each flip.)

Expectation

$\mathbb{E}[X] = 1/p$

Proof. $\mathbb{E}[X] = \sum_{i=1}^{\infty} p \cdot (1-p)^i \cdot i = p \sum_{i=1}^{\infty} i \cdot (1-p)^i$.
Let $s = \sum_{i=1}^{\infty} i \cdot (1-p)^i = 1 + 2(1-p) + 3(1-p)^2 + ...$
Note that $(1-p)s = 1(1-p) + 2(1-p)^2 + ...$
Therefore, $\mathbb{E}[X] = ps = s - (1-p)s = 1 + (1-p) + (1-p)^2 + ... = \frac{1}{1-(1-p)} = \frac{1}{p}$. □

Variance

Recall that the expectation of this random variable is $1/p$. Now we will calculate its variance:

$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}^2[X] = \mathbb{E}[X^2] - \frac{1}{p^2}. \tag{1.5}$$

Let $S = \mathbb{E}[X^2]$, and notice from the right-hand side of (1.2), we can write S as follows:

$$\begin{aligned} S &= 1^2 p + 2^2 (1-p) p + 3^2 (1-p)^2 p + 4^2 (1-p)^3 p + \cdots \\ (1-p) \cdot S &= \quad 1^2 (1-p) p + 2^2 (1-p)^2 p + 3^2 (1-p)^3 p + \cdots \end{aligned}$$

Subtracting the second equation from the first equation yields

$$pS = p + 3(1-p)p + 5(1-p)^2p + 7(1-p)^3p + \cdots,$$

which implies

$$S = 1 + 3(1-p) + 5(1-p)^2 + 7(1-p)^3 + \cdots.$$

and

$$(1-p)S = (1-p) + 3(1-p)^2 + 5(1-p)^3 + \cdots.$$

Again, we subtract $(1-p)S$ from S to obtain the following:

$$\begin{aligned} pS &= 1 + 2(1-p) + 2(1-p)^2 + 2(1-p)^3 + \cdots \\ &= 1 + 2(1-p)\left[1 + (1-p) + (1-p)^2 + \cdots\right] \\ &= 1 + 2(1-p) \cdot \frac{1}{1-(1-p)} \\ &= \frac{2-p}{p}. \end{aligned}$$

Since we initially let $S = \mathbb{E}[X^2]$, substituting this into (1.5) yields

$$\text{Var}[X] = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

1.9.4 Uniform Distribution

Definition

If

$$f(X) = \begin{cases} \frac{1}{b-a} & \forall x \in [a.b] \\ 0 & otherwise \end{cases}$$

, then $X \sim$Uniform(a, b)

Expectation

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x) dx \\ &= \int_a^b \frac{x}{b-a} dx \\ &= \frac{1}{b-a} \frac{x^2}{2}\Big|_a^b \\ &= \frac{1}{b-a} \frac{b^2-a^2}{2} \\ &= \frac{a+b}{2} \end{aligned}$$

Variance

$$\begin{aligned}
\text{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}^2[X] \\
&= \int_a^b x^2 \frac{1}{b-a} - (\frac{a+b}{2})^2 \\
&= \frac{1}{b-a}\frac{x^3}{3}\Big|_a^b - \frac{(a+b)}{4} \\
&= \frac{a^2-ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} \\
&= \frac{a^2+2ab+b^2}{12} \\
&= \frac{(a+b)^2}{12}
\end{aligned}$$

1.9.5 Normal/Gaussian Distribution

Given expectation μ and standard deviation σ, the pdf for Normal distribution is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$. The expectation is then μ and variance σ^2, by definition.