CMSC 28400 Introduction to Cryptography **Autumn 2021**

Notes #3: Discrete Probability Theory

Instructor: David Cash

Contents

These notes introduce some probability theory that we'll use routinely. As we will quickly see, *probability theory* is central to cryptography: When we want to pick a key that our adversaries do not "know", simply choosing a key at random is the best way to do so meaningfully. Our treatment of probability in CMSC 28400 will not typically be so formal, but I find it very useful to have a precise foundation (i.e. formal definitions and basic theorems) to refer to when things get complicated.

3.1 Distributions and Probability Measures

The type of probability we review now is *discrete* in the sense that all of the sets involved are countable. (By *countable* we mean *finite or countably infinite.*) Intuitively, this means we're concerned with choosing random things from amongst a lumpy set with separated outcomes (maybe a random positive integers), and not a continuous space of outcomes (like a real number between zero and one).

3.1.1 Probability Distributions

Let's start with a definition.

Definition 3.1. *Let Ω be a non-empty countable set. A* probability mass function (pmf) *or* distribution *on Ω is a function $p : \Omega \to [0, 1]$ such that $\sum_{w\in\Omega} p(w) = 1$.*

A distribution p can be seen as a way of assigning to every element of Ω a number between 0 and 1 (a "probability") so that the probabilities sum to one. When modeling the outcome a fair coin, we could take $\Omega = \{0, 1\}$ (representing Heads and Tails as we like) and let $p(0) = p(1) = 1/2$. To model a biased coin, we could take the same Ω but with $p(0) = 1/4, p(1) = 3/4$. For rolling a die, we can take $\Omega = \{1, 2, 3, 4, 5, 6\}$, and so on.

Example 3.1. *Here is an example with an infinite Ω. Suppose we flip a fair coin repeatedly until we get a Heads for the first time, and then let w be the number of times we flipped the coin. To model this, we take $\Omega = \{1, 2, \ldots\}$ to be positive integers, and define $p(w) = 1/2^w$ (we stop at $w = 1$ with probability $1/2$, $w = 2$ with probability $1/4$, and so on). This is a distribution because $0 \le p(w) \le 1$ for every $w \in \Omega$, and moreover,*

$$\sum_{w\in\Omega} p(w) = \sum_{w=1}^{\infty} 1/2^w = 1/2 + 1/4 + 1/8 + \cdots = 1.$$

The infinite sum can be evaluated using a geometric series*.*

Note that when Ω is infinite, the sum is in fact a limit. Since Ω is countable and the values of $p(w)$ are non-negative, the order of elements in the sum does not matter. We caution that when Ω is uncountable, dramatically more complicated definitions are required to give a useful theory. We will not stray into such territory for CMSC 28400, and even the formal details of countable sums will not be important.

Definition 3.2. *Let p be a distribution on Ω. We say that p is* uniform *if $p(w) = p(w')$ for all $w, w' \in \Omega$. When Ω is finite, this implies $p(w) = \frac{1}{|\Omega|}$.*

The following example isn't directly relevant for CMSC 28400, but hopefully it will get you thinking.

Example 3.2. *If Ω is infinite, then there does not exist a uniform distribution on Ω. To prove this, take some $w \in \Omega$ (recall that Ω is non-empty). Then either $p(w) = 0$ or $p(w) \neq 0$. If $p(w) = 0$, then $\sum_{w\in\Omega} p(w) = 0$ since p is uniform. If on the other hand $p(w) = c > 0$, then $\sum_{w\in\Omega} p(w) = \sum_{w\in\Omega} c \to \infty$ because p is uniform and Ω is infinite. Either way, $\sum_{w\in\Omega} p(w) \neq 1$ and p is not a distribution on Ω.*

Consider this point of view on uniform probability distributions: If you pick an element of Ω according the uniform distribution without showing me, then I effectively have "no idea" what you picked. From this perspective, the latter part of the example gives a deep fact: When Ω is countably infinite, it's impossible to pick a sample from Ω so that I have "no idea" what you picked, because it's impossible to pick a *uniform* sample! Even more remarkably, it *is* possible to pick a uniform sample from an *uncountable* set (like, say, real numbers between 0 and 1) in a sense that is justified by a more complete theory.

3.1.2 Probability Measures

The theory of discrete probability could, in principle, begin and end with distributions only. But things get more interesting when we introduce other perspectives on understanding distributions. The first such perspective is *probability measures*, which shift from looking at the probability of individual elements $w \in \Omega$ to the probability of *subsets of* Ω. Defining "the probability of a subset" isn't quite a simple as distributions, which define "the probability of a specific outcome".

In this following definition, we write 2^Ω to be the collection of all subsets of Ω. For example, if $\Omega = \{a, b\}$ then $2^\Omega = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}$. When Ω is finite, $|2^\Omega| = 2^{|\Omega|}$.

Definition 3.3. *Let Ω be a non-empty countable set. We say that a function $\Pr : 2^\Omega \to [0, 1]$ is a* discrete probability measure on Ω *if the following hold:*

- $\Pr[\Omega] = 1$.
- *For any countable sequence $E_1, E_2, \ldots$ of disjoint subsets of Ω, $\Pr[\bigcup_{i=1}^\infty E_i] = \sum_{i=1}^\infty \Pr[E_i]$. This property of $\Pr$ is called* countable additivity.

When $\Pr$ is a probability measure on Ω, the pair $(\Omega, \Pr)$ is called a discrete probability space. *In these notes we will just call it a* probability space.

When a probability space $(\Omega, \Pr)$ has been fixed, we refer to any subset of Ω as an *event*. Thus $\Pr$ is a function that maps events to numbers between zero and one.

Why should this be the definition of a probability measure? It's not because it's the most intuitive definition of what probability should be. As far as I can tell, this definition is used because it is very compact (just two rules!), and it implies that $\Pr$ has *all* of structure that corresponds to anything you'd intuitively expect probability to satisfy.[1]

Note that the additivity condition includes finite sequences $E_1, \ldots, E_n$ of disjoint sets; We can take $E_j = \emptyset$ for all $j > n$, which will technically be a sequence of disjoint sets.

Definition 3.4. *Let Ω be a non-empty countable set and p be a distribution on Ω. The function $\Pr : 2^\Omega \to [0, 1]$ defined by*

$$\Pr[E] = \sum_{w \in E} p(w)$$

is called the probability measure (on Ω) induced by p.

Exercise 3.1. *When Ω is finite, verify that $\Pr$ is indeed a probability measure according to the definition. (Verifying this carefully for infinite Ω requires some appeal to properties of infinite series.)*

Note that $\Pr$ is a function that takes as input a subset E of Ω, and outputs a number between 0 and 1. Instead of writing $\Pr(E)$ we write $\Pr[E]$, but don't be distracted by the notation: $\Pr$ *is* a function in every usual sense. We use the $\Pr[E]$ notation to help us keep straight what is a "probability". It can be easily shown that the sums defining $\Pr$ all converge. Finally, note that when $E = \emptyset$, the sum defining $\Pr[E]$ is trivial and taken to be zero by convention.

[1] As usual, I caution that when Ω is uncountable, another, more complicated, definition is required because often there won't exist a measure satisfying this definition. If you read another mathematical text on probability, they usually refer to this more complicated definition. For CMSC 28400, don't worry about this case.

The dependence on the distribution p in the notation $\Pr[E]$ is implicit. Note that Pr depends on p; If we were ever in a setting with multiple distributions, we would need different notation like $\Pr_p$ to keep this dependence straight. Thankfully we will not to need to do so, but it is important to internalize that Pr is a *specific* function on subsets, defined by context, and *not a universally meaningful symbol* like, say, d/dx from calculus. I have noticed that this custom is often fundamentally confusing for people encountering formal probability theory for the first time.

Exercise 3.2. *Fix some non-empty finite Ω, and suppose you are given a function $\Pr : 2^\Omega \to [0,1]$ and told that Pr is a probability measure. Show that Pr is induced by some unique distribution p. Conclude that there is a one-to-one correspondence between distributions and the probability measures they induce.*

When p is uniform, we also say that the measure Pr induced by p is uniform.

Exercise 3.3. *Show that when Ω is finite and Pr is uniform, we have that*

$$\Pr[E] = \frac{|E|}{|\Omega|}.$$

Thus for uniform measures, calculating probabilities reduces to calculating the sizes of E and Ω.

Example 3.3. *Equating the notion of an event with a subset of Ω gives us a convenient language for connecting intuitive descriptions of outcomes with the formalism. For instance, let $\Omega = \{0,1\}^n$ with the uniform distribution. $E = \{0 \| x \ : \ x \in \{0,1\}^{n-1}\}$. Then*

$$\Pr[\text{“a uniformly random } n\text{-bit string starts with zero”}] = \Pr[E] = \frac{1}{2}.$$

Another example takes $F = \{0^{n/2} \| x \ : \ x \in \{0,1\}^{n/2}\}$. Then

$$\Pr[\text{“a uniformly random } n\text{-bit string starts with } n/2 \text{ zeros”}] = \Pr[F] = \frac{2^{n/2}}{2^n} = \frac{1}{2^{n/2}}.$$

With calculations like this one usually skips the formalism and jumps straight to the answer; In a sense the theory is justified by giving the right answer more than the other way around. But it is good to know what exactly is being formalized, in case a proof makes a more subtle jump.

The following exercise begins to justify the definition of a probability space; From those simple conditions, a lot of intuitively-true properties must also hold.

Exercise 3.4. *Let $(\Omega, \Pr)$ be a probability space. Prove the following:*

- *For disjoint events $E, F \subseteq \Omega$, $\Pr[E \cup F] = \Pr[E] + \Pr[F]$.*
- *For any two events $E, F \subseteq \Omega$, $\Pr[E \cup F] = \Pr[E] + \Pr[F] - \Pr[E \cap F]$.*
- *For any two events $E, F \subseteq \Omega$, if $E \subseteq F$ then $\Pr[E] \leq \Pr[F]$.*
- *For an event E, let $E^c = \Omega \setminus E$ be the compliment of E (i.e. everything in Ω that is not in E). Then $\Pr[E^c] = 1 - \Pr[E]$.*

One could go on generating lists of theorems like in the exercises. The essential idea if the following: Any formula relating the *size* of events as sets remains true when we replace with "$|E|$" with "$\Pr[E]$" everywhere, up to some corner cases that occur when some elements of Ω have probability zero. So for example: $|E \cup F| \leq |E| + |F|$, and thus $\Pr[E \cup F] \leq \Pr[E] + \Pr[F]$. But it might be that $\Pr[E] = 0$ yet $|E| > 0$.

We'll use the following facts later on in the course.

Fact 3.1 (Union Bound)**.** *Let $(\Omega, \Pr)$ be a probability space and let $E_1, \ldots, E_n \subseteq \Omega$ be events. Then*

$$\Pr[E_1 \cup \cdots \cup E_n] \leq \Pr[E_1] + \cdots + \Pr[E_n].$$

Fact 3.2 (Law of Total Probability)**.** *Let $(\Omega, \Pr)$ be a probability space and let $E, F \subseteq \Omega$ be events. Then*

$$\Pr[E] = \Pr[E \cap F] + \Pr[E \cap F^c].$$

More generally, if $F_1, \ldots, F_n \subseteq \Omega$ are disjoint events and $F_1 \cup \cdots \cup F_n = \Omega$, then

$$\Pr[E] = \Pr[E \cap F_1] + \cdots + \Pr[E \cap F_n].$$

3.2 Conditional Probability

Probability starts to get really interesting when you introduction *conditioning*. This brief introduction probably won't be enough to give you full intuition for conditional probability, so I recommend reading up, say, in Prof. Kurtz's notes you feel rusty (`http://cmsc-27100.cs.uchicago.edu/2018-winter/Lectures/14/`).

Definition 3.5. *Let $(\Omega, \Pr)$ be a probability space and let $E, F \subseteq \Omega$ be events with $\Pr[F] \neq 0$. We define the* conditional probability of E given F *to be*

$$\Pr[E|F] = \frac{\Pr[E \cap F]}{\Pr[F]}.$$

The notion of *independence* is tightly connected to conditional probability.

Definition 3.6. *Let $(\Omega, \Pr)$ be a probability space and let $E, F \subseteq \Omega$ be events. We say that E and F are* independent *if*

$$\Pr[E \cap F] = \Pr[E] \Pr[F].$$

Note that if E and F are independent and $\Pr[F] \neq 0$, then $\Pr[E|F] = \Pr[E]$.

The next fact is called the *chain rule*.

Fact 3.3. *Let $(\Omega, \Pr)$ be a probability space and let E, F be events. Then*

$$\Pr[E \cap F] = \Pr[E] \Pr[E|F].$$

More generally, if $E_1, \ldots, E_n$ are events, then

$$\Pr[E_1 \cap \cdots \cap E_n] = \Pr[E_1] \cdot \Pr[E_2|E_1] \cdot \Pr[E_3|E_1 \cap E_2] \cdots \Pr[E_n|E_1 \cap \cdots \cap E_{n-1}].$$

This is proved by simply writing out the definition of conditional probability and canceling.

Example 3.4. *Suppose we deal 5 cards from a standard deck. What is the probability the hand is all spades?*

There are several ways to work this; here's one. For $i = 1, \ldots, 5$ define E_i to be the event that the i-th card is a spade. Then we want $\Pr[E_1 \cap \cdots \cap E_5]$. We can calculate the probabilities in the chain rule intuitively: $\Pr[E_1] = 13/52$. Next, $\Pr[E_2|E_1] = 12/51$, since given E_1, there are 12 spades left amongst the remaining 51 cards. Continuing this pattern, we get

$$\Pr[E_1 \cap \cdots \cap E_5] = \frac{13}{52} \cdot \frac{12}{51} \cdot \frac{11}{50} \cdot \frac{10}{49} \cdot \frac{9}{48}.$$

One application of conditional probably which will be useful later, is the following.

Fact 3.4 (Lazy Cryptographer's Law of Total Probability)**.** *Let $(\Omega, \Pr)$ be a probability space and let E, F be events. Then*

$$\Pr[E] \leq \Pr[F] + \Pr[E|F^c].$$

Proof. By the Law of Total Probability and the Chain Rule,

$$\begin{aligned} \Pr[E] &= \Pr[E \cap F] + \Pr[E \cap F^c] \\ &\leq \Pr[F] + \Pr[E \cap F^c] \\ &= \Pr[F] + \Pr[F^c] \Pr[E|F^c] \\ &\leq \Pr[F] + \Pr[E|F^c]. \end{aligned}$$

The first inequality uses $\Pr[E \cap F] \leq \Pr[F]$. The second inequality uses the simple fact that $\Pr[F^c] \leq 1$. □

Note this is mixing probabilities of two different "types" (conditional and unconditioned), which is not normally recommended. But anyway this inequality is used when one wants to bound the probability of E when the probability that E, given F^c is easy to a analyze and the probability of F is easy to analyze. We will point to examples later in the course, so for now you can treat it as an exercise to understand the proof.

Exercise 3.5. *Let $\Omega = \{00, 01, 10, 11\}$ and $\Pr$ be the uniform measure on Ω, which models choosing two random bits. Let E be the event that the first bit is zero, and F be the event that the chosen bits are the same. Verify that E and F are independent.*

Exercise 3.6. *Let $(\Omega, \Pr)$ be a probability space, and let F be an event with non-zero probability. Show that the function $\Pr_F : 2^\Omega \to [0, 1]$, defined by $\Pr_F[E] = \Pr[E|F]$ is a probability measure. If p is the distribution that induces $\Pr$, what distribution induces $\Pr_F$?*

Finally we discuss how to define independence of several events $E_1, E_2, \ldots E_n$. At first glance, it may seem intuitive to define them to be independent if E_i and E_j are independent for all $i \neq j$; This however turns out to be too weak, as the follow classic example argues.

Exercise 3.7. *Suppose we roll two dice, and define E to be the event that the sum is 7, F to be the event that first die is a 1, and G to be the event that the second die is a 6. Then you can check that each pair is independent. But intuitively all three events should not be considered "independent", for if we know that E and F happened, then we can be certain than G also happened.*

The condition that E_i and E_j be independent for all $i \neq j$ is called *pairwise independence*. The following definition gives a stronger condition, called *mutual independence*.

Definition 3.7. *Let $(\Omega, \Pr)$ be a probability space, and let $E, F, G \subseteq \Omega$ be events. We say E, F, G are* mutually independent *if they are pairwise independent and also*

$$\Pr[E \cap F \cap G] = \Pr[E] \Pr[F] \Pr[G].$$

More generally, if $E_1, \ldots, E_n \subseteq \Omega$ are events, we say they are mutually independent if for all $S \subseteq \{1, \ldots, n\}$,

$$\Pr[\bigcap_{k \in S} E_k] = \prod_{k \in S} \Pr[E_k].$$

You can check that in the example above, the extra condition for mutual independence is violated. In the general case, the definition is saying that all possible subsets of events should "split" when you consider the probability of their intersection. It is a subtle point that you actually need to include *all* of the subsets of all sizes in order to intuitively capture "indpendence."

3.3 Random Variables

We next review *random variables*, which are an abstraction to make sense of informal statements like "Let X and Y be the outcomes of two fair die rolls." By augmenting our theory with a bit more abstraction, we can increase the expressiveness and comprehensibility of the theory of probability similar to how measures (and events) were more powerful and convenient than distributions.

Definition 3.8. *Let $(\Omega, \Pr)$ be a probability space. A random variable on $(\Omega, \Pr)$ with range $\mathcal{R}$ is a function $X : \Omega \to \mathcal{R}$.*

At first glance this is not a very enlightening definition. Let us start with some examples.

Example 3.5. *Let $\Omega = \{00, 01, 10, 11\}$ and $\Pr$ but the uniform measure on Ω. Define $X_1 : \Omega \to \{0, 1\}$ and $X_2 : \Omega \to \{0, 1\}$ by setting $X_1(w)$ to the first bit of w and $X_2(w)$ to the second bit of w, and define $Y : \Omega \to \{0, 1, \ldots, n\}$ by setting $Y(w)$ to be the number of 1 bits of w.*

Then $X_1, \ldots, X_n, Y$ are all random variables. We have for all $w \in \Omega$,

$$Y(w) = X_1(w) + X_2(w)$$

One typically expresses this relationship by writing $Y = X_1 + X_2$, leaving out the w entirely. One sees this type of notation occasionally in calculus, where you might write $f = g + h$ instead of $f(x) = g(x) + h(x)$.

In this example, we can think of the random variables as *measurements* on the outcomes in Ω. Above, we can *think* of Y as representing the outcome of picking a random bit string and then counting the number of 1 bits. Of course, when pressed, we must admit that formally Y is a *function* and not an actual random outcome.

Why should we formalize random variables as functions? The answer will hopefully be clear after we develop some more concepts using random variables. But a first benefit is they give us some language for discussing events compactly, via the following notation.

Notation 1. *Let* $(\Omega, \Pr)$ *be a probability space and let* $X : \Omega \to \mathcal{R}$ *be a random variable on this space. For* $i \in \mathcal{R}$*, we define*

$$\Pr[X = i] = \Pr[\{w \in \Omega \ : \ X(w) = i\}].$$

Note that $\Pr, X$ *are still functions; Prior to this definition, the left-hand side of the equation would not make sense. The right-hand side, however, did already make sense:* Pr *is a function that takes as input subsets of* Ω*, and* $\{w \in \Omega \ : \ X(w) = i\}$ *is such a set.*

The point of this notation is that "$\Pr[X = i]$" is a natural notion to think about: It *should* be the probability that a random variable takes the value i. The notation makes this natural notion precise. Note that this notation is not $\Pr[X(w) = i]$ – It omits the w, being consistent with convention mentioned in the previous example. We remark that this notation is part of why we prefer $\Pr[X = i]$ over $\Pr(X = i)$: Pr is a function, but one we use strangely. (This preference is not universal, particularly not amongst mathematicians).

Example 3.6. *Let* $\Omega = \{(a, b) \ : \ 1 \leq a, b, \leq 6\}$ *with the uniform probability measure (i.e. we model the outcome of rolling a pair of fair dice). Define* X *on this space as* $X(a, b) = a + b$*. Then*

$$\begin{aligned}
\Pr[X = 4] &= \Pr[\{(a, b) \in \Omega \ : \ X(a, b) = 4] \\
&= \Pr[\{(a, b) \in \Omega \ : \ a + b = 4] \\
&= \Pr[\{(1, 3), (2, 2), (3, 1)\}] \\
&= 3/36 = 1/12.
\end{aligned}$$

Definition 3.9. *Let* $(\Omega, \Pr)$ *be a probability space and let* $X : \Omega \to \mathcal{R}$ *be a random variable. The* distribution of X*, denoted* p_X*, is defined to be*

$$\begin{aligned}
p_X : \mathcal{R} &\to [0, 1] \\
i &\mapsto \Pr[X = i]
\end{aligned}$$

.

Example 3.7. *Let* X *be the sum of two fair dice (as in Example 3.6). Then* $\mathcal{R} = \{2, \ldots, 12\}$*, and* $p_X(2) = 1/36, p_X(3) = 2/36$*, etc.*

The next example deserves special attention.

Example 3.8. *Again using the same probability space as in Example 3.6. Define* Z_1 *to be the outcome of the first roll, and* Z_2 *to be the outcome of the second roll. (Formally:* $Z_1(a, b) = a$ *and* $Z_2(a, b) = b$*.) Then* Z_1 *and* Z_2 *have the same distribution: That is,* p_{Z_1} *and* p_{Z_2} *are* exactly the same function*. Both of them map every element of* $\{1, 2, 3, 4, 5, 6\}$ *to* $1/6$*.*

When two (or more) random variables have the same distribution, we say they are identically distributed*.*

This example begins to point to the power of random variables: Z_1 and Z_2 have the same distribution ("uniform on numbers between 1 and 6"), but they are still *different* random variables. Intuitively, this is because they are measuring different dice. Concretely, Z_1 and Z_2 are just different functions from Ω to $\mathcal{R}$.

Definition 3.10. *Let $(\Omega, \Pr)$ be a probability space, and let X and Y be random variables on this space with the same range $\mathcal{R}$. We say that X and Y are independent if, for every $i, j \in \mathcal{R}$*

$$\Pr[X = i, Y = j] = \Pr[X = i] \Pr[Y = j].$$

The notation "$X = i, Y = j$" means $X = i$ and $Y = j$ simultaneously; That is,

$$\Pr[X = i, Y = j] = \Pr[\{w \in \Omega \ : \ X(w) = i \text{ and } Y(w) = j\}].$$

This definition is requiring that the events $X = i$ and $Y = j$ are independent for all i and j.

Example 3.9. *One can check that Z_1 and Z_2 from the previous example are independent, but Z_1 and $Z_1 + Z_2$ are not.*

The notion of mutual independence adapts to random variables as follows.

Definition 3.11. *Let $(\Omega, \Pr)$ be a probability space, and let $X_1, \ldots, X_n$ be random variables on this space with the same range $\mathcal{R}$. We say that $X_1, \ldots, X_n$ are* mutually independent *if for every $S \subset \{1, \ldots, n\}$ and every $i_1, \ldots, i_n$,*

$$\Pr[\bigcap_{k \in S} X_k = i_k] = \prod_{k \in S} \Pr[X_k = i_k].$$

We will revisit the notion of mutual independence when we study perfect secrecy.

3.4 Expectation

Random variables can be very complicated, and a big part of probability theory is finding ways to understand even badly-behaved random variables. One major approach is find features of random variables that we can compute. For example, a notion of "average value" will often be helpful in understanding a random game. Knowing the average value alone doesn't tell the whole story, since very different games can have the same averages (consider a game where you win and lose \$1 with equal probability versus one where you win or lose \$100 with equal probability; the average winnings is \$0 in both cases). But knowing the average is frequently very useful.

Here is a rigorous definition capturing the "average value" of a random variable.

Definition 3.12. *Let $X : \Omega \to \mathbb{R}$ be a random variable on a probability space $(\Omega, \Pr)$. The* expectation *or* expected value of X *is defined to be*

$$\mathbb{E}[X] = \sum_{w \in \Omega} p(w) X(w).$$

The formula in the definition is useful in some proofs, but is rather clunky. We will usually think of the random variable in terms of its distribution p_X, not in terms of its actual values $X(w)$. (See below for an example.) The following lemma gives another very useful formula for expectation.

Lemma 1. *Let $X : \Omega \to \mathbb{R}$ be a random variable on a probability space $(\Omega, \Pr)$. Let $S \subseteq \mathbb{R}$ consist of values that X takes with non-zero probability, i.e. $S = \{s \in \mathbb{R} : \Pr[X = s] > 0\}$. Then the expectation of X can equivalently be written*

$$\mathbb{E}[X] = \sum_{s \in S} s \cdot \Pr[X = s].$$

The set S in the statement of the lemma is called the *support of* X.

Proof. The set Ω is partitioned in sets $\Omega_s = \{w : X(w) = s\}$ as s ranges over S. (More formally, $\Omega = \cup_{s \in S} \Omega_s$, and all the Ω_s are disjoint.) Thus we can break a sum over Ω into a bunch of sums over Ω_s and get the same value, since each w will be accounted for exactly once.

Thus we can calculate

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{w \in \Omega} p(w) X(w) \\
&= \sum_{s \in S} \sum_{w \in \Omega_s} p(w) X(w) \\
&= \sum_{s \in S} \sum_{w \in \Omega_s} p(w) s \\
&= \sum_{s \in S} s \sum_{w \in \Omega_s} p(w) \\
&= \sum_{s \in S} s \Pr[X = s].
\end{aligned}$$

□

Example 3.10. *What is the expected value of rolling a fair six-sided die? The two formulas work out sort of the same here, but we'll use the second. Let X be a random variable modeling the outcome. Then*

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{6} i \cdot \Pr[X = i] \\
&= \sum_{i=1}^{6} i \cdot (1/6) \\
&= 3.5.
\end{aligned}$$

Note that the expected value 3.5 *is never actually taken as an output of* X.

3.4.1 Linearity of Expectation and the Indicator Method

The following simple theorem is frequently useful for computing expectations. It is called the "linearity of expectation", and allows you to split up an expectation of a sum into individual expectations.

Theorem 1 (Linearity of Expectation)**.** *Let X, Y be random random variables on a probability space $(\Omega, \Pr)$ with range $\mathbb{R}$. Then*

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

More generally, if $X_1, \ldots, X_n$ are random variables on a probability space $(\Omega, \Pr)$ then

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

Proof. We prove the first part; The second part is an exercise in induction on n. Let Z be the random variable defined by $Z(w) = X(w) + Y(w)$. We use the original formula for $\mathbb{E}[Z]$:

$$\begin{aligned}
\mathbb{E}[Z] &= \sum_{w \in \Omega} p(w) Z(w) \\
&= \sum_{w \in \Omega} p(w)(X(w) + Y(w)) \\
&= \sum_{w \in \Omega} p(w)X(w) + p(w)Y(w) \\
&= \sum_{w \in \Omega} p(w)X(w) + \sum_{w \in \Omega} p(w)Y(w) \\
&= \mathbb{E}[X] + \mathbb{E}[Y].
\end{aligned}$$

□

To get some intuition for this, consider a course instructor computing the average grade for a quarter. In this class, quarter grades are simply the sum of three exam scores. One way to compute the average is sum each student's scores on three exams, and then average over all the students' quarter scores. This corresponds to computing an expectation using the formula $\mathbb{E}[X_1+X_2+X_3] = \sum_{s \in S} s \Pr[X_1 + X_2 + X_3 = s]$. Another way is to compute the average on each exam, and then sum those averages, with corresponds using the linearity formula.

Example 3.11. *Suppose we roll five fair six-sided dice and take their sum. What is the expected value?*

Let $X_1, \ldots, X_5$ be the outcomes of the five dice. Each is uniform on $\{1, 2, 3, 4, 5, 6\}$, and we can computed $\mathbb{E}[X_i] = 3.5$ above. Using linearity we have

$$\begin{aligned}
\mathbb{E}[X_1 + X_2 + X_3 + X_4 + X_5] &= \mathbb{E}[X_1] + \mathbb{E}[X_2] + \mathbb{E}[X_3] + \mathbb{E}[X_4] + \mathbb{E}[X_5] \\
&= 3.5 + 3.5 + 3.5 + 3.5 + 3.5 \\
&= 17.5.
\end{aligned}$$

Let's pause to see what linearity bought us here. If we tried to use the second formula for expectation, we'd be stuck with computing the probabilities $\Pr[X_1+X_2+X_3+X_4+X_5 = s]$ for s ranging from 5 to 30, which is tractable but annoying. Even better, the linearity approach gives us some intuition for why the sum is 17.5.

The next example highlights the power of linearity: It works even when the random variables involved are not independent, which can be highly counter-intuitive at first.

Example 3.12. *Suppose an urn contains six balls, numbered 1,2,3,4,5,6. We draw two of the balls without replacement and take their sum. What is the expected value?*

We can try to compute this without linearity, but we'll run into the same sort of pain as in the previous example. So let's try linearity. Let X_1 be the first draw and X_2 be the second draw. We are interested in $\mathbb{E}[X_1 + X_2]$, which we know to $\mathbb{E}[X_1] + \mathbb{E}[X_2]$. What are these two expectations? It's pretty intuitive that $\mathbb{E}[X_1] = 3.5$, the same as a die roll, since it's just a random number between 1 and 6. Slightly harder to see is that $\mathbb{E}[X_2]$ is also *3.5. To understand this, imagine covering up the first outcome and only looking at the second. If you do this, the second outcome is also a uniformly*

random number between 1 and 6, and thus its expectation is 3.5. We can conclude that the expected sum is $3.5 + 3.5 = 7$. *This is exactly the same expectation that we'd get by rolling two dice.*

If you find this calculation sneaky, you can resort to working with the actual sample space Ω *and map out what* X_1 *and* X_2 *are doing by hand. The point is that* $\Pr[X_2 = i] = 1/6$ *for each* i*, as a mathematical fact. In reality, the person doing the draw will always know the first outcome, but that doesn't matter in this calculation;* X_2 *only "looks" at the second draw.*

A common trick for computing expectations is to use linearity even when the original random variable is not presented as a sum. One version of this is called the *indicator method*. An *indicator random variable for an event* E is simply a random variable I that takes value one if E happens and zero otherwise[2]. It is easy to compute the expectation of an indicator random variable for an event E:

$$\mathbb{E}[I] = 1 \cdot \Pr[I = 1] + 0 \cdot \Pr[I = 0] = \Pr[I = 1] = \Pr[E].$$

Here is an example of the indicator method.

Example 3.13. *Suppose we deal* 5 *cards from a standard deck. What is the expected number of spades in the hand?*

Let X *be the number spades in the hand. We want to find* $\mathbb{E}[X]$*. A not-so-easy way to do this is to find* $\Pr[X = 0], \Pr[X = 1], \ldots, \Pr[X = 5]$ *and then use the formula for expectation. This gives*

$$\mathbb{E}[X] = \sum_{i=0}^{5} i \cdot \frac{\binom{13}{i}\binom{39}{5-i}}{\binom{52}{5}}.$$

A better way uses indicators. For $i = 1, \ldots, 5$ *define the event* E_i *to be that the* i*-th card is a spade, and let* I_i *be the indicator for* E_i*. The key observation is that*

$$X = I_1 + I_2 + I_3 + I_4 + I_5.$$

That is, we can count the spades by checking if each card is a spade. Using linearity, we only need to calculate $\mathbb{E}[I_i]$ *for each* i *and then add them up. But* $\mathbb{E}[I_i] = \Pr[E_i] = 13/52 = 1/4$*, using the formula for the expectation of an indicator. Thus* $\mathbb{E}[X] = 5 \cdot 1/4 = 1.25$.

Note that this quickly generalizes to a k*-card hand, for* $1 \leq k \leq 52$*; The answer in that case is* $k/4$*. Compare this to awful formula you'd get solving this the straightforward way with the expectation formula. It would be quite difficult to notice that it simplified so nicely!*

3.4.2 Markov's Inequality

We close our discussion of expectation with an application. Suppose you've computed the average score of a set of exams, and assume that negative scores are impossible. Consider the following observation:

"At most half of the class can have double the average score."

Why? Intuitively, if more than half the class had double the average, then they'd pull up the average score above where it is *even if everyone else got zero*! (Note that we need to assume no one got a negative score; otherwise the negatives could pull down the average.) We can extend this reasoning to statements like

[2]More formally, $I(w) = 1$ if $w \in E$ and 0 if $w \notin E$.

"At most one third of the class can have triple the average score."

by the same type of argument.

The following theorem, called Markov's inequality, makes this thinking precise and rigorous.

Theorem 2 (Markov's inequality.)**.** *Let X be a random random variables on a probability space $(\Omega, \Pr)$ with range $\mathbb{R}$ that only takes non-negative values. Then for any $c > 0$,*

$$\Pr[X \geq c] \leq \mathbb{E}[X]/c.$$

Proof. Let S be the support of X. Then

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{s \in S} s \cdot \Pr[X = s] \\
&\geq \sum_{s \in S, s \geq c} s \cdot \Pr[X = s] \\
&\geq \sum_{s \in S, s \geq c} c \cdot \Pr[X = s] \\
&= c \sum_{s \in S, s \geq c} \Pr[X = s] \\
&= c \Pr[X \geq c].
\end{aligned}$$

The first "$\geq$" uses that all $s \in S$ are non-negative to conclude that the sum can only go down when we can omit those s that are less than c. If negative s were allowed, the sum might go up! □

Applications of Markov's inequality are often very "loose", in the sense that $\Pr[X \geq c]$ is actually *way* smaller than $\mathbb{E}[X]/c$ for lots of problems.

3.5 Algorithms and Randomized Algorithms

For this class, you won't need to know what a formal *algorithm* is exactly. But in case you've seen the concept of a Turing Machine or (uniform) circuit family, that's what we mean. If you haven't seen those, or don't recall the definitions, you can think of an *algorithm* as a piece of code that accepts an input, performs some computation than can be counted in discrete steps while consuming some amount of memory, and finally emits an output.

We will at various times consider algorithms that make internal random choices as part of their computation. You can think of these as machines with a special input that should be as many random bits as they need to run. This is akin to reading from the special file `/dev/random` in Unix-like operating systems. For cryptography, we'll be interested in making our algorithms (like selecting a key) randomized in order to achieve certain security goals. We'll also be interested in analyzing the possibility of adversaries using randomized algorithms, just in case they might help break our systems. In either case, randomized algorithms are rather exotic in the "real world", and will hopefully become more motivated as we encounter examples.

Definition 3.13. *A* randomized algorithm *A is an algorithm with a distinguished input from some associated finite probability space Ω_A with the uniform measure. For any "input" x, and $w \in \Omega_A$, we write $A(x; w)$ to mean running A on x with distinguished input w. We define the notation*

$$\Pr[A(x) = y] = \Pr[\{w \in \Omega_A \ : \ A(x; w) = y\}]$$

to formalize "the probability that randomized algorithm A outputs y when given input x."

Here $A(x; w)$ *is a random variable*, as defined in the previous section. And since we like to suppress the placeholder variable w, we will always just write $A(x)$.

Finally, we can speak of running a randomized algorithm A on a random input. In this case, we'd usually take the sample space to include pairs (x, w), and think of $A(x; w)$ as a random variable.

3.6 Probability Theory "In Practice"

Now that you've waded through that very quick review, I will close with a discussion of how discrete probability is used, both in this class and in other domains, from theoretical to applied.

As with many mathematical concepts, it is possible to maintain two modes of thinking about probability theory: The first is intuitive, meaning that when you read "let X be a uniformly random bit-string", you don't have to connect it a sample space, a measure, or a random variable. If you asked "What's the probability that X starts with a zero?", you can say $1/2$ without the help of all this formalism. Similarly, you usually can answer questions of independence intuitively.

This intuitive approach often proceeds without even mentioning a sample space or measure. In a sense, statements like "$\Pr[X = 1]$" use Pr as a symbol to indicate that probability is being modeled, but do not mean to refer a particular measure Pr as we've defined it. This is fine, but sometimes weird steps happen: A sequence of steps in a proof will usually use the symbol Pr everywhere, *even when referring multiple distinct probability measures on different sample spaces.* This is almost always fine, in that one can formalize the true intention if necessary, and in fact it's best to not cloud proofs with too much formalism. Occasionally in this class we'll pause to examine these statements, but the point will only be to understand, at a mathematical level, what the symbols mean. The justification for the steps will almost always be intuitive. We will see examples of this when learning about perfect secrecy.

3.7 Case Study: The Birthday Problem

The material in this section is borrowed from *Introduction to Modern Cryptography* by Bellare and Rogaway[3].

Consider the following question: If a group of n people are in a room, what is the probability that at least two people have the same birthday? A related question is: How large should n be before we have a 50% chance that two people have the same birthday? To idealize the problem, assume birthdays are uniformly random and independent samples from $\{1, \ldots, 365\}$.

Surely 366 is enough, because the pigeonhole principle will guarantee a "collision" with probability 1. But if $n < 366$, then it might not be clear how the probability scales. Maybe $n = 365/2$ are required for a 50% chance? Take a moment to think it over.

To get a handle on the problem, let's compute the expected number of X, the number pairs of people that have the same birthday. (This is a bit artificial; If three people have the same birthday, then we consider this to be three pairs. But it's useful enough to start.)

For each pair $\{i, j\}$ of people, with $1 \leq i < j \leq n$, create an indicator random variable $I_{i,j}$ indicating if people i and j have the same birthday. Then $X = \sum_{1 \leq i < j \leq n} I_{i,j}$.

[3]See Appendix A of `https://web.cs.ucdavis.edu/~rogaway/classes/227/spring05/book/main.pdf`.

We have that $\mathbb{E}[I_{i,j}] = 1/365$, since this is just the probability of the even being indicated. There are $\binom{n}{2} = n(n-1)/2$ pairs (i,j) in the sum defining X. So we get that

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{1\le i<j\le n} I_{i,j}\right] = \sum_{1\le i<j\le n} \mathbb{E}[I_{i,j}] = \frac{n(n-1)}{2}\frac{1}{365}.$$

Okay, so what do we make of that? Try graphing the function $f(x) = x(x-1)/2 \cdot 365$ using a tool like `https://www.desmos.com/calculator`. Zoom out enough to see when the function is greater than 1.

This calculation reveals that the expected value crosses above 1 when $n = 28$! This is remarkable small to most peoples' intuition. It says that in a room of 28 people, we expect about one pair to have the same birthday. The key to understanding this intuitively is that the probability is growing relative $n(n-1) \approx n^2$ and not just n; In other words, including an n-th person has a much larger affect than including the second person.

3.7.1 An Accurate Estimate

Let's go further than the expectation, and get bounds on the probability of at least one collision. Instead of sticking to just birthdays, let's generalize the problem to a level where it will be useful for other problems.

For integers M, n, let $C(M,n)$ be the probability that we get a repeated sample amongst n independent uniform samples from a set of size M. (With birthdays, $M = 365$.) We'd like to estimate $C(M,n)$ as a function of M and n.

Theorem 3. *For all positive integers M, n,*

$$1 - e^{-n(n-1)/2M} \le C(M,n) \le 0.5\frac{n(n-1)}{M}.$$

If $1 \le n \le \sqrt{2M}$ then this implies

$$0.3\frac{n(n-1)}{M} \le C(M,n) \le 0.5\frac{n(n-1)}{M}.$$

Before we prove this, let me point out that in cryptographic contexts, we'll typically care about the latter case, where n is at most $\sqrt{2M}$. So focus on this bound: It tells us that we know $C(M,n)$ to a remarkable accuracy: It's *basically* "some-constant" times $\frac{n(n-1)}{M}$. When M and n are huge numbers, we won't usually even care about the constant!

Example 3.14. *Let's apply this theorem for birthdays. With $M = 365$ and $n = 23$ we get*

$$0.45 \le C(365, 23) \le 0.75.$$

The lower bound is the surprising part; With 23 people, there's at least a 45% chance of getting two people with the same birthday.

See `https://en.wikipedia.org/wiki/Birthday_problem` for finer estimates. The correct answer is actually $n = 23$ for a 50% chance! So our estimate is off by a little bit, but good enough for the cryptography we have planned.

Proof of Theorem 3. For each i, let F_i be the event that there is no repeat amongst the first i samples. Then $C(M,n) = 1 - \Pr[F_n]$. Here's the calculation for the lower bound first; Justification is given afterwards.

$$\begin{aligned}
1 - \Pr[F_n] &= \prod_{i=1}^{n-1} \Pr[F_{i+1}|F_1, \ldots, F_i] \\
&= \prod_{i=1}^{n-1} \Pr[F_{i+1}|F_i] \\
&= \prod_{i=1}^{n-1} \left(1 - \frac{i}{M}\right) \\
&\leq \prod_{i=1}^{n-1} e^{-\frac{i}{M}} \\
&= e^{-\sum_{i=1}^{n-1} \frac{i}{M}} \\
&= e^{-n(n-1)/2M}.
\end{aligned}$$

The first equality uses that $F_n = F_1 \cap \cdots \cap F_n$ and applies the chain rule to this intersection. The next uses this identity again, this time in the form $F_{i+1} = F_1 \cap \cdots \cap F_{i+1}$. For the next equality, $\Pr[F_{i+1}|F_i] = (1 - i/M)$ since given F_i, there are i choices that could cause a repeat. The inequality uses the fact that $1 + x \leq e^x$ for all real x (including negative x). The next-to-last equality is just arithmetic with the product, and the final equality uses the fact that $1+2+\cdots+(n-1) = n(n-1)/2$. Rearranging this inequality gives the lower bound in the theorem.

The upper bound is a lot easier. For each i, let C_i be the event that the i-th sample is a repeat. Then $\Pr[C_i] \leq (i-1)/M$, since there will always be at most $i-1$ choices that cause a repeat. Now use the union bound:

$$\begin{aligned}
C(M,n) &= \Pr[C_1 \cup C_2 \cup \cdots \cup C_n] \\
&\leq \Pr[C_1] + \Pr[C_2] + \cdots + \Pr[C_n] \\
&\leq \frac{0}{M} + \frac{1}{M} + \cdots + \frac{n-1}{M} \\
&= \frac{n(n-1)}{2M}.
\end{aligned}$$

Finally, when $1 \leq n \leq \sqrt{2M}$, we can derive the other form of the lower bound using calculus; In particular we need to the following inequality: For all $0 \leq x \leq 1$,

$$(1 - e^{-1})x \leq 1 - e^{-x}.$$

Then under the assumption $1 \leq n \leq \sqrt{2M}$ we can apply the inequality. (The details are omitted.) □