These notes review some essential facts about probability theory and modular arithmetic 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.

0.1 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.)

0.1.1 Probability Distributions

Let’s start with a definition.

Definition 0.1. Let \( \Omega \) be a non-empty countable set. A probability mass function (pmf) or distribution on \( \Omega \) 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 \( \Omega \) 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 \( \Omega \) 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. We are free to choose \( \Omega \) to model our problem.
Note that when $\Omega$ is infinite, the sum is in fact a limit. Since $\Omega$ is countable, the order of elements in the sum does not matter. We caution that when $\Omega$ 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 0.2.** Let $p$ be a distribution on $\Omega$. We say that $p$ is uniform if $p(w) = p(w')$ for all $w, w' \in \Omega$. When $\Omega$ is finite, this implies $p(w) = \frac{1}{|\Omega|}$.

**Example 0.1.** If $\Omega$ is infinite, then there does not exist a uniform distribution on $\Omega$. To prove this, take some $w \in \Omega$ (recall that $\Omega$ 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 $\Omega$ is infinite. Either way, $\sum_{w \in \Omega} p(w) \neq 1$ and $p$ is not a distribution on $\Omega$.

Consider this point of view on uniform probability distributions: If you pick an element of $\Omega$ 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 $\Omega$ is countably infinite, it’s impossible to pick a sample from $\Omega$ 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, a fact that can be explained once the theory of measures is developed.

### 0.1.2 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 $\Omega$. Defining “the probability of a subset” isn’t quite as simple as distributions, which define “the probability of a specific outcome”.

In this following definition, we write $2^\Omega$ to be the collection of all subsets of $\Omega$. When $\Omega$ is finite, $|2^\Omega| = 2^{|\Omega|}$.

**Definition 0.3.** Let $\Omega$ be a non-empty countable set. We say that a function $\Pr : 2^\Omega \to [0, 1]$ is a discrete probability measure on $\Omega$ if the following hold:

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

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

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.\(^1\)

\(^1\)As usual, I caution that when $\Omega$ 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.
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 0.4.** Let $\Omega$ be a non-empty countable set and $p$ be a distribution on $\Omega$. The function $\Pr : 2^\Omega \to [0, 1]$ defined by

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

is called the probability measure (on $\Omega$) induced by $p$.

**Exercise 0.1.** Verify that $\Pr$ is indeed a probability measure according to the definition.

Note that $\Pr$ is a function that takes as input a subset $E$ of $\Omega$, and outputs a number between 0 and 1. Instead of writing $\Pr(E)$ we write $\Pr[E]$, but this has exactly the same meaning. 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.

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 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 0.2.** Fix some non-empty $\Omega$, and suppose you are given a function $\Pr : 2^\Omega \to [0, 1]$ and told that $\Pr$ is induced by some distribution $p$. Show that $\Pr$ determines $p$. 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 0.3.** Show that when $\Omega$ 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 $\Omega$.

**Example 0.2.** Equating the notion of an event with a subset of $\Omega$ 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. $E = \{0 \parallel x : x \in \{0, 1\}^{n-1}\}$. Then

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

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

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

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

- \(\Pr[\Omega] = 1\)
- For disjoint events \(E, F \subseteq \Omega\), \(\Pr[E \cup F] = \Pr[E] \cup \Pr[F]\).
- For any two events \(E, F \subseteq \Omega\), \(\Pr[E \cup F] = \Pr[E] \cup \Pr[F] - \Pr[E \cap F]\).
- 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 \(\Omega\) that is not in \(E\)). Then \(\Pr[E^c] = 1 - \Pr[E]\).

All of the above facts are true for infinite \(\Omega\), but rigorous proofs require some appeal to calculus. 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 \(\Omega\) 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 0.1** (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 0.2** (Law of Total Probability). Let \((\Omega, \Pr)\) be a probability space and let \(E, F \subseteq \Omega\) be events. Then
\[
\Pr[E] = \Pr[E \cap F] + \Pr[E \cap F^c].
\]

### 0.2 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 in [?] if you feel rusty.

**Definition 0.5.** Let \((\Omega, \Pr)\) be a probability space and let \(E, F\) be events with \(\Pr[F] \neq 0\). We define the conditional probability of \(E\) given \(F\) to be
\[
\Pr[E|F] = \frac{\Pr[E \cap F]}{\Pr[F]}.
\]

One application, which will become clear later, is the following.

**Fact 0.3** (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,
\[
\Pr[E] = \Pr[E \cap F] + \Pr[E \cap F^c] \\
\leq \Pr[F] + \Pr[E \cap F^c] \\
= \Pr[F] + \Pr[F^c] \Pr[E|F^c] \\
\leq \Pr[F] + \Pr[E|F^c].
\]
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 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 0.5. Let \(\Omega = \{00, 01, 10, 11\}\) and \(\Pr\) be the uniform measure on \(\Omega\), 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 0.6. 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 \(\Pr_E\)?

0.3 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 0.6. 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 0.3. Let \(\Omega = \{00, 01, 10, 11\}\) and \(\Pr\) but the uniform measure on \(\Omega\). 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 \( \Omega \). 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 \( X : \Omega \to \mathbb{R} \) be a random variable on this space. For \( i \in \mathbb{R} \), we define

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

Note that \( \Pr, X \) are still functions; Prior to this definition, the right-hand side of the equation would not make sense. The left-hand side, however, did already make sense: \( \Pr \) is a function that takes as input subsets of \( \Omega \), 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 0.4.** 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

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

**Definition 0.7.** Let \((\Omega, \Pr)\) be a probability space and let \( X : \Omega \to \mathbb{R} \) be a random variable. The *distribution* of \( X \), denoted \( p_X \), is defined to be

\[
p_X : \mathbb{R} \to [0, 1]
\]

\[
i \mapsto \Pr[X = i]
\].

**Example 0.5.** Let \( X \) be the sum of two fair dice (as in Example 0.4). 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 0.6.** Again using the same probability space as in Example 0.4. 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 $\Omega$ to $\mathbb{R}$.

**Definition 0.8.** 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,

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

**Example 0.7.** One can check that $Z_1$ and $Z_2$ from the previous exercise are independent, but $Z_1$ and $Z_1 + Z_2$ are not.

There is quite a bit more to say about random variables; From here one would usually learn about expectation and variance. However CMSC 28400 will not need these concepts.

### 0.4 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.

**Definition 0.9.** A randomized algorithm $A$ is an algorithm with a distinguished input from some associated finite probability space $\Omega_A$ 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 $x$.”

Note that $A(x; w)$ 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 $A(x)$.

Note that 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.
0.5 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 to 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 to 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.