Probability

Benjamin Cosman, Patrick Lin and Mahesh Viswanathan

Fall 2020

TAKE-AWAYS

- Probability space:
  - *Sample space* S, the set of possible outcomes
  - *Probability distribution*: $\Pr : S \to [0,1]$ so that $\sum_{x \in S} \Pr[x] = 1$
  - For an *event* $E \subseteq S$, $\Pr[E] = \sum_{x \in E} \Pr[x]$
- "Counting" rules:
  - Sum Rule: For *disjoint* $E_1, \ldots, E_n$, $\Pr\left[\bigcup_{i=1}^{n} E_i\right] = \sum_{i=1}^{n} \Pr[E_i]$
  - Difference Rule: $\Pr[A \setminus B] = \Pr[A] - \Pr[A \cap B]$
  - Union Bound: $\Pr\left[\bigcup_{i=1}^{n} E_i\right] \leq \sum_{i=1}^{n} \Pr[E_i]$
  - Monotonicity Rule: If $A \subseteq B$, then $\Pr[A] \leq \Pr[B]$
- Conditional probability:
  - $\Pr[x|B]$: the probability of outcome x, *given* event B
  - Kolmogorov's Rule: $\Pr[A \cap B] = \Pr[A|B] \Pr[B]$
  - Bayes' Rule: $\Pr[A|B] = \frac{\Pr[B|A] \Pr[A]}{\Pr[B]}$
  - Events A and B are *independent* if and only if $\Pr[A|B] = \Pr[A]$ if and only if $\Pr[A \cap B] = \Pr[A] \Pr[B]$

Suppose you are developing some algorithm for solving some problem. A malicious attacker with seemingly unlimited resources figures out that your algorithm is much slower on some inputs than others.[1] This attacker is spamming your program with the worst-case inputs, and you are losing your mind because every time you change your implementation, the attacker figures out what the worst-case inputs for your new implementation is. What are you going to do?

Well, computer scientists figured out a number of decades ago that the thing to do is to randomize. If you randomly pick a strategy every time, then the malicious attacker cannot predict which input will be worst-case for that strategy every time.

Randomization doesn't just help prevent against malicious attackers, but it can also provide algorithms that are significantly faster in practice than the best known deterministic algorithms, even though

[1] For example, a naïve deterministic implementation of the QuickSort algorithm that chooses the first element as the pivot every time is fast for most inputs, but is much slower for an array that is already sorted.

randomized algorithms are not always guaranteed to be correct or to finish quickly. The most useful algorithm for checking if a large number is prime, for example, is a randomized algorithm; the fastest deterministic algorithm for accomplishing the same task is significantly slower.

Well, what does randomization mean, mathematically? How do we analyze a randomized strategy? The answer lies in the study of Probability Theory.

Finite Probability with Equal Likelihoods

Recall the following example from the Notes on Counting:

Example 1. Two standard six-sided dice,[2] one orange and one blue, are rolled. How many ways can the sum of the two rolls be even? The number of ways to roll two odd numbers is $3 \cdot 3 = 9$, and the number of ways to roll two even numbers is also $3 \cdot 3 = 9$. Adding the numbers together gives the total number of ways to roll an even sum as $9 + 9 = 18$.

[2] Standard here means the sides are labeled 1 through 6.

One natural question to ask is not the how many even-sum outcomes there are, but rather what percentage of the set of total outcomes are even-sum outcomes. For the preceding example, the computation is simple: each die has six possibilities, for a total of $6 \cdot 6 = 36$ total outcomes, so half of the possible outcomes are even-sum outcomes.

We formalize this in terms of *probability*, and specifically, because this is a course on Discrete Structures, we will be focused on *discrete* probability.

Formally, we have a ***sample space*** S, which is a (finite) set of possible ***outcomes***. For example, in the case of our blue and orange dice, the $S = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$. An ***experiment*** is some procedure that generates some outcome of the sample space. In the case of the blue and orange dice, an experiment would simply be rolling the dice. For now, we will work with the assumption that every possible outcome from the sample space happens with *uniform probability*, i.e., that each outcome is equally likely. We write this as $\Pr[x] = \frac{1}{|S|}$ for all $x \in S$.

As we have seen before, we are commonly not interested in the probability of one specific outcome, but rather the probability of some set of outcomes occuring. In the language of probability, subsets of the sample space are called ***events***. The ***probability of an event*** E is denoted by $\Pr[E]$, and under our current working assumption that each outcome is equally likely, $\Pr[E] = \frac{|E|}{|S|}$. The complement event $\overline{E} = S \setminus E$ is simply the event that E does not happen.

In this situation, trying to compute the probability of an event can simply be done by counting, which presumably we now know how to do.[3]

[3] Now would be a good time to practice more counting if you are still nervous about last week's material :)

Example 2. Consider a coin whose sides are labeled Heads (H) and Tails (T). Suppose the coin is flipped twice. What is the probability that both flips land on Heads? Well, the set of possible outcomes is $S = \{HH, HT, TH, TT\}$,[4] and the event that both flips is heads is $E = \{HH\}$. Comparing the cardinalities of the two sets, we find $\Pr[\{HH\}] = \mathbf{\frac{1}{4}}$.

[4] To be precise, the elements of S should really be pairs, e.g., (H, T), but the parentheses and commas mostly serve to add clutter.

Example 3. Two standard six-sided dice, one orange and one blue, are rolled. We will compute the probability that the blue die rolls at least a 5. That is, we want to compute $\Pr[\{5,6\} \times \{1,2,3,4,5,6\}]$. $|\{5,6\} \times \{1,2,3,4,5,6\}| = 2 \cdot 6 = 12$, so the probability is $\frac{12}{36} = \mathbf{\frac{1}{3}}$.

Example 4. A random binary sequence of length n is picked. What is the probability that exactly k entries have value 1? The number of length n binary sequences is 2^n, but how do we count the number of sequences with exactly k 1s? Recall that there is a bijection $f : \{0,1\}^n \to \mathcal{P}(\{1,2,\ldots,n\})$ given by $f((x_1,\ldots,x_n)) = \{i \mid x_i = 1\}$. Under this bijection the binary sequences with exactly k 1s correspond to the k-element subsets of $\{1,2,\ldots,n\}$, and we know that there are $\binom{n}{k}$ such subsets. So the probability is $\mathbf{\binom{n}{k}/2^n}$.

Discrete Probability

Previously we considered the situation where every outcome happens with equal likelihood. In general, this is not the case. Consider, for example, a coin that has been weighted so that it lands on one side with much higher likelihood than the other, and so we wish to have some way to quantify that.

The solution will be to define a ***probability distribution*** over our sample space S, which is a function $\Pr : S \to [0,1]$ such that $\sum_{x \in S} \Pr[x] = 1$. The requirement that $\sum_{x \in S} \Pr[x] = 1$ ensures that the probability that *some* outcome occurs is 1, which is consistent with the uniform case. In general, the ***probability of an event*** E will be $\Pr[E] = \sum_{x \in E} \Pr[x]$. [5]

[5] In *continuous* probability, the sample space is infinite and comes with a "measure" over which we can integrate. The probability of a specific outcome is meaningless, and only ever work with probabilities of events, as computed via the integral $\Pr[E] = \int_E p(x)\,dx$, for some "probability density" $p : S \to [0,1]$. Doing this precisely is a messy business. In particular, we cannot allow all subsets of S to be valid events, due to subtleties of measure theory that are well beyond the scope of this course. As a result, the axioms for continuous probability usually require a declaration of a family of subsets that are valid events.

Example 5. Suppose we roll a die that has been manufactured so that 1 is rolled with probability $\frac{1}{2}$, 6 is rolled with probability $\frac{1}{6}$, and the remaining four numbers are each rolled with probability $\frac{1}{12}$. The probability of rolling an even number is $\Pr[\{2,4,6\}] = \Pr[2] + \Pr[4] + \Pr[6] = \frac{1}{12} + \frac{1}{12} + \frac{1}{6} = \mathbf{\frac{1}{3}}$.

As we saw previously, when working under the uniform distribution, the fact that $\Pr[E] = \frac{|E|}{|S|}$ allows us to make use of the Counting

rules from last week for computing $|E|$. For non-uniform distributions, we can no longer use said rules in their original form, because $\Pr[E]$ is no longer computed via only cardinalities. However, we can recover similar rules.

Proposition 6 (Sum Rule). [6] *For* disjoint *events* $E_1, \ldots, E_n$,
$\Pr\left[\bigcup_{i=1}^{n} E_i\right] = \sum_{i=1}^{n} \Pr[E_i]$.

[6] Some sources declare this as an axiom of Probability instead of defining $\Pr[E] = \sum_{x \in E} \Pr[x]$. This is especially important in continuous probability, as in that setting the probability of an individual outcome is meaningless. For discrete probability, either choice is acceptable: if we assume the Sum Rule we easily conclude that $\Pr[E] = \sum_{x \in E} \Pr[\{x\}]$.

Proof. By definition, $\Pr\left[\bigcup_{i=1}^{n} E_i\right] = \sum_{x \in \bigcup_{i=1}^{n} E_i} \Pr[x]$. If $E_1, \ldots, E_n$ are disjoint, then each element $x \in \bigcup_{i=1}^{n} E_i$ appears in exactly one E_i, so $\sum_{x \in \bigcup_{i=1}^{n} E_i} \Pr[x] = \sum_{i=1}^{n} \sum_{x \in E_i} \Pr[x] = \sum_{i=1}^{n} \Pr[E_i]$. □

Corollary 7 (Difference Rule). $\Pr[A \setminus B] = \Pr[A] - \Pr[A \cap B]$. *In particular,* $\Pr\left[\overline{B}\right] = 1 - \Pr[B]$.

Proof. Recall that $A \setminus B = \{x \in A \mid x \notin B\}$. Since $A \cap B = \{x \in A \mid x \in B\}$, $A \setminus B$ and $A \cap B$ are disjoint, and $A = (A \setminus B) \cup (A \cap B)$. So $\Pr[A] = \Pr[A \setminus B] + \Pr[A \cap B]$. □

As nice as the sum rule is, we will often find ourselves in situations where we cannot guarantee that the events are disjoint, and furthermore we will find it difficult to figure out by how much we have overcounted. In such a situation, we will often settle for being able to compute upper and lower bounds on the actual probability.

Corollary 8 (Union Bound). [7] $\Pr\left[\bigcup_{i=1}^{n} E_i\right] \leq \sum_{i=1}^{n} \Pr[E_i]$.

[7] The Union Bound also appears in the literature under the names "Boole's inequality", "the Bonferroni inequality", and "subadditivity of measures."

Corollary 9 (Monotonicity Rule). *If* $A \subseteq B$ *then* $\Pr[A] \leq \Pr[B]$.

Example 10. A task in the recently popular game *Among Us*[8] is to correctly connect a set of four wires. Consider a more general version of the problem, where the player needs to connect n wires. A player, trying to speedrun this task, will fail to connect each wire with probability p. That is, letting E_i be the event that the i-th wire is correctly connected, $\Pr[E_i] = 1 - p$ for each $1 \leq i \leq n$. Let E be the event that the task is completed correctly. By definition, $E = \bigcap_{i=1}^{n} E_i$, because the task is completed correctly exactly when all n wires are correctly connected. What is $\Pr[E]$?

[8] Given how quickly trends come and go, this pop culture reference is probably[9] already stale.

[9] Pun intended.

Without further information, it will be difficult to find an exact answer, but we can still compute some useful bounds.

On the one hand, we have $E \subseteq E_1$, so by the monotonicity rule, $\Pr[E] \leq \Pr[E_1] = 1 - p$. On the other hand, $\overline{E} = \bigcup_{i=1}^{n} \overline{E_i}$, [10] so by the union bound, $\Pr[\overline{E}] \leq \sum_{i=1}^{n} \Pr[\overline{E_i}] = np$, so applying the difference rule gives us $\Pr[E] = 1 - \Pr[\overline{E}] \geq 1 - np$.

[10] This is an application of DeMorgan's law for sets, $\overline{A \cap B} = \overline{A} \cup \overline{B}$.

In summary, $1 - np \leq \Pr[E] \leq 1 - p$. For example, a good player who fails each wire with probabilty 5% will succeed on four wires with probability between 80% and 95%. Not bad!

Conditional Probability

A common situation that occurs in probability is when we know something that can reduce the sample space.

Example 11. A gambler walks up to a table, and is told that they can bet on the value of the sum of two standard six-sided dice, one orange and one blue. Suppose the gambler puts their money on the sum being (strictly) less than 7. If these are standard dice, then we can fall back to counting to find that there are fifteen possible outcomes that lead to this sum, out of thirty-six, for a probability of $\frac{15}{36} = \frac{5}{12}$ of the gambler winning. However, to generate excitement and tension, the dealer first rolls the orange die, waits for everyone to gasp with excitement, and only then rolls the blue one. If the orange die rolled a 6, then we know that the gambler has lost, for there is no way for the blue die to roll a 0. But if the orange die rolls a 1, then we know that as long as the blue die does not roll a 6, then the gambler has won; in short, the probability of the gambler winning, *conditioned on the orange die rolling a 1*, is $\frac{5}{6}$, whereas the probability of winning, *conditioned on the orange die rolling a 6*, is 0.

Formally, we define conditional probability as follows.

Definition 12. For an outcome x, the probability of x ***given*** event B, written $\Pr[x|B]$, is defined as

$$\Pr[x|B] = \begin{cases} \frac{\Pr[x]}{\Pr[B]} & \text{if } x \in B, \\ 0 & \text{otherwise.} \end{cases}$$

For an event A, $\Pr[A|B] = \sum_{x \in A} \Pr[x|B]$.

Proposition 13 (Kolmogorov's Rule)**.** [11] $\Pr[A|B] \Pr[B] = \Pr[A \cap B]$.

[11] For continuous probability, this is often taken as the definition of conditional probability, since, as before, the probability of an individual outcome is meaningless.

Proof. By definition, $\Pr[A|B] = \sum_{x \in A} \Pr[x|B]$. We can split this sum into $\sum_{x \in A} \Pr[x|B] = \sum_{x \in A \cap B} \Pr[x|B] + \sum_{x \in A \cap \overline{B}} \Pr[x|B]$, but recalling that $\Pr[x|B] = 0$ for $x \notin B$, the second term cancels to 0. Furthermore, $\sum_{x \in A \cap B} \Pr[x|B] = \sum_{x \in A \cap B} \frac{\Pr[x]}{\Pr[B]} = \frac{1}{\Pr[B]} \sum_{x \in A \cap B} \Pr[x] = \frac{\Pr[A \cap B]}{\Pr[B]}$. Rearranging the resulting equation $\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}$ gives the stated rule. □

Applying Kolmogorov's rule in two different ways, once to $\Pr[A|B]$ and once to $\Pr[B|A]$, gives us $\Pr[A|B] \Pr[B] = \Pr[A \cap B] = \Pr[B|A] \Pr[A]$. Rearranging the two ends of this equation gives us the following:

Corollary 14 (Bayes' Rule)**.** $\Pr[A|B] = \frac{\Pr[B|A] \Pr[A]}{\Pr[B]}$.

One final thing to note about the earlier example of the gambler betting on the sum of the rolls being less than 7: in our calculation for the conditional probability based on the blue die's outcome, each

outcome of the orange die's roll had equal likelihood, no matter what the blue die rolled beforehand. In the language of probability, the orange die's outcome and the blue die's outcome are *independent*.

Definition 15. Events A and B are ***independent*** if A does not depend on B, i.e., $\Pr[A|B] = \Pr[A]$. Equivalently,[12] $\Pr[A \cap B] = \Pr[A] \Pr[B]$.

[12] via Kolmogorov's rule

Example 16. Recall that in the earlier example of the wires task from *Among Us*, we were unable to produce an exact bound. One reason is that we could not say that the events E_i were all mutually independent, i.e., that $\Pr[E] = \Pr\left[\bigcap_{i=1}^{n} E_i\right] = \prod_{i=1}^{n} \Pr[E_i]$, and we had to settle for somewhat loose upper bounds and lower bounds instead.

If we could say that the events E_i were all independent, i.e., success on any one wire does not affect the probability of success any other wire, then we could simply compute $\Pr[E]$ as $\prod_{i=1}^{n} \Pr[E_i] = (1-p)^n$. In reality, messing up one wire is likely to affect a player on subsequent wires. If we knew the conditional probabilities, we would be able to apply Kolmogorov's rule (repeatedly). For example, for $n = 4$, $\Pr[E] = \Pr[E_1] \Pr[E_2|E_1] \Pr[E_3|E_1 \cap E_2] \Pr[E_4|E_1 \cap E_2 \cap E_3]$.

More Examples

Example 17 (Binomial Distribution)**.** Consider a coin that has been weighted so that it flips Heads with probability p (and so flips Tails with probability $1-p$). Suppose we flip the coin independently[13] n times. What is the probability of flipping Heads exactly k times out of n?

[13] Each coin flip is called a *Bernoulli trial*.

By independence, the probability of any sequence of flips that includes exactly k Heads is $p^k(1-p)^{n-k}$: we need to flip Heads exactly k times, and we need to flip Tails all the other $n-k$ times; each Heads is flipped with probability p and each Tails is flipped with probability $1-p$.

We also know that there are $\binom{n}{k}$ ways to choose k out of n flips to be Heads, and each of these events are disjoint, so applying the sum rule tells us that the probability of flipping exactly k Heads out of n is $\binom{n}{k} p^k (1-p)^{n-k}$.

As a sanity check, we should be able to verify that summing this probability over all k should lead to a value of 1. Applying the Binomial Theorem,[14] we get $\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^k = (p + (1-p))^n = 1^n = 1$.

[14] Example 20 of the Counting Notes

Example 18 (Birthday Paradox)**.** For $n \leq 365$, we consider n people, each of whom were born on one of the 365 dates in the standard non-leap-year calendar. We will assume that the each person is equally likely to be born on each day,[15] and that these probabilities are all independent. We want to compute the probability that at least two people have the same birthday.

[15] This assumption is not entirely representative of reality. According to one study, between the years of 1994 and 2004, births in September were slightly more common in United States than births in other months.

It will turn out to be easier to compute the probability of the complement event, that everyone has a different birthday.

Since $n \leq 365$, each way of having n people, each of whom has a different birthday, corresponds to an ordering of n dates out of 365. The number of such orderings is $P(365, n) = \frac{365!}{(365-n)!}$; since each person's birthdays are assumed to be independently chosen, each of these orderings occurs with equal probability. The total number of possible sequences of birthdays is 365^n, so the probability that everyone has a different birthday is $\frac{P(365,n)}{365^n}$, and so the probability that at least two people share a birthday is $\mathbf{1 - \frac{P(365,n)}{365^n}}$. It turns out that for $n \geq 23$, the probability that two people share a birthday is over 50%; for $n \geq 57$ the probability is over 99%.

Example 19 (Birthday Attacks)**.** The Birthday Paradox has ramifications in cryptographic hashing. When sending secure messages, often a *digital signature* is used to verify that the message is legitimate: given a message m, a *hash* $h(m)$ of the message is computed and signed; the message and signature are both sent together. In the verification process, one checks that the hash that was signed matches the original message. The trouble arises when multiple messages are found that hash to the same value, and a fraudulent message is passed along with a legitimate signature. Replacing people with messages and birthdates with hashes in the previous example shows the importance of hashing into a large number of possible values.

Example 20. Suppose we independently throw n balls into n bins uniformly at random.[16] Suppose we are instead interested in the probability that at least k of the balls land in the same bin. Our intuition might tell us that for large values of k, this probability should be small, but let us see how to justify this.

[16] this means that each ball falls into each bin with equal probability

For a subset S of the balls, let E_S be the event that all those balls fall into the same bin. If *at least* k balls fall into the same bin, then there is *some* subset containing *exactly* k balls wherein all k balls fall into the same bin. Thus letting T be the set of all k-element subsets of the balls, the probability we are trying to compute can be written as $\Pr[\bigcup_{S \in T} E_S]$. Computing this probability exactly is difficult since the various E_S events are *not* independent, but we can easily compute the union bound: $\Pr[\bigcup_{S \in T} E_S] \leq \sum_{S \in T} \Pr[E_S]$.

The number of different k-element subsets, i.e. $|T|$, is $\binom{n}{k}$. For any fixed $S \in T$, we can compute $\Pr[E_S]$ as follows: for each of the n bins, the probability that all k balls fall into that specific bin is $\frac{1}{n^k}$,[17] so the probability that all k balls fall in the same bin is $n \cdot \frac{1}{n^k} = \frac{1}{n^{k-1}}$. In sum: $\Pr[\bigcup_{S \in T} E_S] \leq \sum_{S \in T} \Pr[E_S] = \binom{n}{k} \frac{1}{n^{k-1}} = \frac{n!}{k!(n-k)!} \cdot \frac{1}{n^{k-1}} \leq \frac{n^k}{k!} \cdot \frac{1}{n^{k-1}} = \frac{n}{k!}$, so $\frac{n}{k!}$ is a simple (though not necessarily very tight) upper bound on the probability that at least k balls end up in the same bin.

[17] This is the part where we used independence.