### 3 Counting and Probability

Today, we introduce basic concepts in probability theory and we learn about one of its fundamental principles.

**Probability measure.** We begin with a simple example of a probabilistic experiment: throwing two dice and counting the total number of dots. Each die has six sides with 1 to 6 dots. The result of a throw is thus a number between 2 and 12. There are 36 possible outcomes, 6 for each die, which we draw as the entries of a matrix; see Figure 5.

![Figure 5: Left: the two dice give the row index and the column index of the entry in the matrix. Right: the most likely sum is 7, with probability \(\frac{6}{36}\), the length of the diagonal divided by the size of the matrix.](image)

We need definitions to construct a general framework that formalizes this and similar examples. The set of possible outcomes of a probabilistic experiment is the *sample space*, denoted as \(\Omega\). A possible outcome is an element, \(x \in \Omega\). A subset of outcomes is an *event*, \(A \subseteq \Omega\). Subsets with a single element, \(A = \{x\}\), are called *elementary events*. The *probability* or *weight* of an element \(x\) is \(P(x)\), a real number between 0 and 1. For finite sample spaces, the probability of an event is \(P(A) = \sum_{x \in A} P(x)\). For example, in the two dice experiment, we set \(\Omega = \{2, 3, \ldots, 12\}\).

An event could be to throw an even number. The probabilities of the different outcomes are given in Figure 5, and we can compute

\[
P(\text{even}) = \frac{1 + 3 + 5 + 5 + 3 + 1}{36} = \frac{1}{2}.
\]

More formally, we call a function \(P : 2^\Omega \to \mathbb{R}\) a *probability measure* if

1. \(P(x) \geq 0\) for every \(x \in \Omega\);
2. \(P(A \cup B) = P(A) + P(B)\) for all disjoint events \(A \cap B = \emptyset\);
3. \(P(\Omega) = 1\).

A common example is the *uniform probability measure* defined by \(P(x) = P(y)\) for all \(x, y \in \Omega\). Clearly, if \(\Omega\) is finite then \(P(A) = |A|/|\Omega|\) for every event \(A \subseteq \Omega\).

It is very important to notice that the same probabilistic experiment can be modelled on different probability spaces, i.e. the choice of \(\Omega\) is not unique. The model above considered the possible sums as the basic building block of the probability space; this is why \(\Omega\) ended up being the set of integers from 2 to 12. However, one could have recorded the outcomes more meticulously by listing all possible *ordered pairs* \((i, j)\), representing the number of dots on each dice separately. In this case we would choose the probability space as the following set of 36 elements:

\[
\Omega' = \{(i, j) \mid 1 \leq i, j \leq 6\}.
\]

The sum \(i + j\) becomes a simple function on the probability space

\[
f : \Omega' \to \mathbb{R}, \quad f(i, j) = i + j
\]

Functions on a probability space are also called *random variables*.

At first glance (10) looks an over complication; the previous model (9) recorded directly the outcomes we were interested in. But notice that computing the corresponding probabilities anyway required to go back to the matrix in Fig 5, where each little square
anyway represented an element of $\Omega'$ (and the number written in it was the random variable representing the sum). So we secretly used the underlying structure of $\Omega'$. The reason why $\Omega'$ is a better model is that it is tailored to the basic assumption of the model: we said that each outcome of each dice is equally likely (since the dice are fair and the two throws are independent and unbiased). This means that each element of $\Omega'$ has equal probability, thus each element has probability $1/36$. Then the probability that the sum random variable $f(i, j) = i + j$ takes on a fixed value, say 7, will be the sum of the probabilities of all possible pairs that add up to 7, i.e.

$$
\sum_{i+j=7} \frac{1}{36} = \frac{6}{36} = \frac{1}{6}
$$

Here the subscripts in the summation indicates that we sum up for all pairs $(i, j)$ with $i + j = 7$. There are altogether six such pairs. Notice that the problem of computing a probability boiled down to a counting question; this is because we modeled the problem on a probability space where each element has equal probability. In the heart of many probabilistic experiment there is a similar fine resolution of all possible outcomes, which occur with equal chance. It is then often pays off to choose the collection of all these elementary events as the probability space.

Setting up the correct probability space and equipping it with the correct probability measure is an absolutely crucial step in applications. It requires a proper insight into the structure of the practical problem. The resolution of most probabilistic paradoxes hinges on an erroneous choice of the probability space. For example, it would have been disastrous to equip the set $\Omega$ from (9) with the uniform probability, resulting in a probability $1/11$ for every possible outcome for the sum. Very naively one could have argued that the numbers $2, 3, \ldots, 12$ are all possible outcomes, perhaps each should be granted a "fair chance", eventually we do a "totally random" experiment. This would be an obvious misunderstanding of the underlying physical experiment.

Here is a classical paradox to illustrate the importance of the choice of $\Omega$. Consider tossing a coin once and throwing a dice once. Let the set of possible outcomes for the coin be $C = \{\text{head}, \text{tail}\}$ and the possible outcomes for the dice be $D = \{1, 2, 3, 4, 5, 6\}$. Clearly $P(C) = 1$ since all possibilities are listed, and similarly $P(D) = 1$. But clearly $C \cap D = \emptyset$, so by the formula (11) proven below, we have $P(C \cup D) = P(C) + P(D) = 1 + 1 = 2$, but a probability cannot be larger than 1, in fact $P(C \cup D) = 1$ since the probability that $C$ or $D$ happen is clearly one. What’s wrong? [Hint: the probability space is lously defined. Try all sensible possibilities that might fit the description above, i.e. try $\Omega = \{\text{head}, \text{tail}\} \times \{1, 2, 3, 4, 5, 6\}$, $\Omega = \{\text{head}, \text{tail}\}$ or $\Omega = \{1, 2, 3, 4, 5, 6\}$, and conclude that some part of the above argument is wrong in each case.

**Union of non-disjoint events.** Suppose we throw two dice and ask what is the probability that the outcome is even or larger than 7. Write $A$ for the event of having an even number and $B$ for the event that the number exceeds 7. Then $P(A) = \frac{1}{2}$, $P(B) = \frac{5}{6}$, and $P(A \cap B) = \frac{3}{36}$. The question asks for the probability of the union of $A$ and $B$. We get this by adding the probabilities of $A$ and $B$ and then subtracting the probability of the intersection, because it has been added twice:

$$
P(A \cup B) = P(A) + P(B) - P(A \cap B),
$$

which gives $\frac{1}{2} + \frac{5}{6} - \frac{3}{36} = \frac{7}{9}$. If we had three events, then we would subtract all pairwise intersections and add back in the triplewise intersection:

$$
P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).
$$

The reason for adding $P(A \cap B \cap C)$ is that it has been added three times, as part of $P(A)$, $P(B)$, and $P(C)$, but it has also been subtracted three times, as part of $P(A \cap B)$, $P(A \cap C)$, and $P(B \cap C)$. We can generalize the idea which is called the Principle of Inclusion-Exclusion (PIE).

**PIE Theorem (for probability).** The probability of the union of not necessarily disjoint events $A_1$ to $A_n$ is the alternating sum of probabilities over all non-empty subcollections of events:

$$
P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum P(A_{i_1} \cap \ldots \cap A_{i_k}).
$$

**Proof.** Let $x$ be an element in $\bigcup_{i=1}^{n} A_i$ and $H$ the subset of $\{1, 2, \ldots, n\}$ such that $x \in A_i$ iff $i \in H$. The contribution of $x$ to the sum is $P(x)$, for each odd subset of $H$, and $-P(x)$, for each even subset of $H$. If we include $\emptyset$ as an even subset, then the number of
odd and even subsets is the same, as we have proved using the Binomial Theorem. In this application, we do not include the empty set, so we get

\[ \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} = 1. \]

In words, there is a surplus of one odd subset and therefore a net contribution of \( P(x) \). This is true for every element. The claimed equation follows.

Checking hats. Next, we use the Principle of Inclusion-Exclusion to compute the probability of getting hats returned correctly. Suppose \( n \) people check their hats before going to the rock concert, and after the concert, they get their hats returned in random order. What is the chance that at least one gets the correct hat? Let \( A_i \) be the event that person \( i \) gets the correct hat. Then

\[ P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}, \]

Similarly,

\[ P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = \frac{(n-k)!}{n!}. \]

The event that at least one person gets the correct hat is the union of the \( A_i \). Writing \( P = P(\bigcup_{i=1}^{n} A_i) \) for its probability, we have

\[ P = \sum_{k=1}^{n} (-1)^{k+1} \sum P(A_{i_1} \cap \ldots \cap A_{i_k}) \]

\[ = \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} \]

\[ = \sum_{k=1}^{n} (-1)^{k+1} \frac{k!}{k!} \]

\[ = 1 - \frac{1}{2} + \frac{1}{3!} - \ldots + \frac{1}{n!} \]

Recall from the Taylor expansion of real-valued functions that \( e^x = 1 + x + x^2/2 + x^3/3! + \ldots \). Hence,

\[ P = 1 - e^{-1} = 0.632 \ldots \]

in the limit, when \( n \) goes to infinity. For finite numbers, the probability oscillates; that is: the chance that one hat gets returned correctly for an odd number \( n \) of people is higher than for \( n-1 \) people but it is also higher than for \( n+1 \) people. Similarly, for odd \( n \), the probability decreases with increasing \( n \), while for even \( n \), the probability increases with increasing \( n \).

Counting surjective functions. The principle of inclusion-exclusion generally applies to measuring things. Counting elements in finite sets is an example.

PIE Theorem (for counting). The cardinality of the union of not necessarily disjoint sets \( A_1 \) to \( A_n \) is the alternating sum of cardinalities over all non-empty subcollections of the sets:

\[ \sum_{k=1}^{n} (-1)^{k+1} \sum \left| A_{i_1} \cap \ldots \cap A_{i_k} \right| \]

The only difference to the PIE Theorem for Probability is that for each \( x \), we count 1 instead of \( P(x) \), so we do not repeat the proof.

We use the counting version of the principle to count surjective functions. Counting the functions of the form \( f : [m] \to [n] \) is easy. Each \( j \in [m] \) has \( n \) choices for its image, the choices are independent, and therefore the number of functions is \( n^m \). How many of these
functions are surjective? To answer this question, let $A_i$ be the set of functions in which $i \in [n]$ is not the image of any element in $[m]$.

Writing $A$ for the set of all functions and $S$ for the set of all surjective functions, we have

$$S = A - \bigcup_{i=1}^{n} A_i.$$ 

We already know $|A|$. Similarly, $|A_i| = (n - 1)^m$. Furthermore, the size of the intersection of $k$ of the $A_i$ is

$$|A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}| = (n - k)^m.$$ 

We can now use inclusion-exclusion to get the number of functions in the union, namely,

$$|\bigcup_{i=1}^{n} A_i| = \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} (n - k)^m.$$ 

To get the number of surjective functions, we subtract the size of the union from the total number of functions:

$$|S| = \sum_{k=0}^{n} (-1)^k \binom{n}{k} (n - k)^m.$$