Conditional Probability

Balls and bins. Suppose we toss $m = 3$ balls into $n = 3$ bins: this is a uniform sample space with $3^3 = 27$ points. We already know that the probability the first bin is empty is $(1-\frac{1}{3})^3 = (\frac{2}{3})^3 = \frac{8}{27}$. What is the probability of this event *given that* the second bin is empty? Call these events A, B respectively. In the language of conditional probability we wish to compute the probability $P[A|B]$, which we read to say probability of A given B.

How should we compute $\Pr[A|B]$? Well, since event B is guaranteed to happen, we need to look not at the whole sample space Ω, but at the smaller sample space consisting only of the sample points in B. What should the probabilities of these sample points be? If they all simply inherit their probabilities from Ω, then the sum of these probabilities will be $\sum_{\omega \in B} \Pr[\omega] = \Pr[B]$, which in general is less than 1. So we need to *scale* the probability of each sample point by $\frac{1}{\Pr[B]}$. For each sample point $\omega \in B$, the new probability becomes

$$\Pr[\omega|B] = \frac{\Pr[\omega]}{\Pr[B]}.$$

Now it is clear how to compute $\Pr[A|B]$: namely, we just sum up these scaled probabilities over all sample points that lie in both A and B:

$$\Pr[A|B] = \sum_{\omega \in A \cap B} \Pr[\omega|B] = \sum_{\omega \in A \cap B} \frac{\Pr[\omega]}{\Pr[B]} = \frac{\Pr[A \cap B]}{\Pr[B]}.$$

Definition (conditional probability): For events A, B in the same probability space, such that $\Pr[B] > 0$, the *conditional probability of A given B* is

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

Returning to our example, to compute $\Pr[A|B]$ we need to figure out $\Pr[A \cap B]$. But $A \cap B$ is the event that both the first two bins are empty, i.e, all three balls fall in the third bin. So $\Pr[A \cap B] = \frac{1}{27}$ (why?). Therefore,

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{1/27}{8/27} = \frac{1}{8}.$$

Not surprisingly, $\frac{1}{8}$ is quite a bit less than $\frac{8}{27}$: knowing that bin 2 is empty makes it significantly less likely that bin 1 will be empty.

Let us consider another example:

Dice. Roll two fair dice. Let A be the event that their sum is even, and B the event that the first die is even. By symmetry it's easy to see that $\Pr[A] = \frac{1}{2}$ and $\Pr[B] = \frac{1}{2}$. Moreover, a little counting gives us that $\Pr[A \cap B] = \frac{1}{4}$. What is $\Pr[A|B]$? Well,

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{1/4}{1/2} = \frac{1}{2}.$$

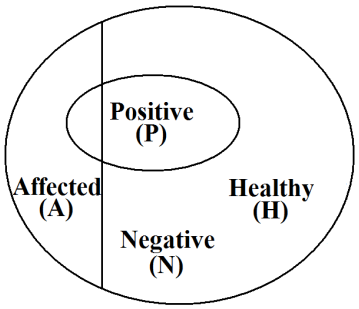
In this case, $\Pr[A|B] = \Pr[A]$, i.e, conditioning on B does not change the probability of A.

Now let us consider a more real World example:

A pharmaceutical company is marketing a new test for a certain medical disorder. According to clinical trials, the test has the following properties:

1. When applied to an affected person, the test comes up positive in 90% of cases, and negative in 10% (these are called "false negatives").
2. When applied to a healthy person, the test comes up negative in 80% of cases, and positive in 20% (these are called "false positives").

Suppose that the incidence of the disorder in the US population is 5%. When a random person is tested and the test comes up positive, what is the probability that the person actually has the disorder? (Note that this is presumably *not* the same as the simple probability that a random person has the disorder, which is just $\frac{1}{20}$.) The implicit probability space here is the entire US population with uniform probabilities.



This is an example of a *conditional probability*: we are interested in the probability that a person has the disorder (event A) *given that* he/she tests positive (event B). In other words $\Pr[A|B]$.

The sample space here consists of all people in the US — denote their number by N (so $N \approx 250$ million). The population consists of four disjoint subsets:

TP: the true positives (90% of $\frac{N}{20} = \frac{9N}{200}$ of them);

FP: the false positives (20% of $\frac{19N}{20} = \frac{19N}{100}$ of them);

TN: the true negatives (80% of $\frac{19N}{20} = \frac{76N}{100}$ of them);

FN: the false negatives (10% of $\frac{N}{20} = \frac{N}{200}$ of them).

Now let A be the event that a person chosen at random is affected, and B the event that he/she tests positive. Note that B is the union of the disjoint sets TP and FP, so

$$|B| = |TP| + |FP| = \frac{9N}{200} + \frac{19N}{100} = \frac{47N}{200}.$$

Thus we have

$$\Pr[A] = \frac{1}{20} \qquad \text{and} \qquad \Pr[B] = \frac{47}{200}.$$

Now when we condition on the event B, we focus in on the smaller sample space consisting only of those $\frac{47N}{200}$ individuals who test positive. To compute $\Pr[A|B]$, we need to figure out $\Pr[A \cap B]$ (the part of A that lies in B). But $A \cap B$ is just the set of people who are both affected and test positive, i.e, $A \cap B = TP$. So we have

$$\Pr[A \cap B] = \frac{|TP|}{N} = \frac{9}{200}.$$

Finally, we conclude from the definition of conditional probability that

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{9/200}{47/200} = \frac{9}{47} \approx 0.19.$$

This seems bad: if a person tests positive, there's only about a 19% chance that he/she actually has the disorder! This sounds worse than the original claims made by the pharmaceutical company, but in fact it's just another view of the same data.

(Incidentally, note that $\Pr[B|A] = \frac{9/200}{1/20} = \frac{9}{10}$; so $\Pr[A|B]$ and $\Pr[B|A]$ can be very different. Of course, $\Pr[B|A]$ is just the probability that a person tests positive given that he/she has the disorder, which we knew from the start was 90%.)

To complete the picture, what's the (unconditional) probability that the test gives a correct result (positive or negative) when applied to a random person? Call this event C. Then

$$\Pr[C] = \frac{|TP|+|TN|}{N} = \frac{9}{200} + \frac{76}{100} = \frac{161}{200} \approx 0.8.$$

So the test is about 80% effective overall, a more impressive statistic.

But how impressive is it? Suppose we ignore the test and just pronounce everybody to be healthy. Then we would be correct on 95% of the population (the healthy ones), and wrong on the affected 5%. I.e, this trivial test is 95% effective! So we might ask if it is worth running the test at all. What do you think?

Bayesian Inference

The medical test problem is a canonical example of an *inference* problem: given a noisy observation (the result of the test), we want to figure out the likelihood of something not directly observable (whether a person is healthy). To bring out the common structure of such inference problems, let us redo the calculations in the medical test example but only in terms of events without explicitly mentioning the sample points of the underlying sample space.

Recall: A is the event the person is affected, B is the event that the test is positive. What are we given?

- $\Pr[A] = 0.05$, (5% of the U.S. population is affected)
- $\Pr[B|A] = 0.9$ (90% of the affected people test positive)
- $\Pr[B|\overline{A}] = 0.2$ (20% of healthy people test positive)

We want to calculate $\Pr[A|B]$. Notice that our formula for conditional probability can be re-arranged to compute $\Pr[A \cap B] = \Pr[B|A]\Pr[A]$. We can proceed as follows:

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{\Pr[B|A]\Pr[A]}{\Pr[B]} \tag{1}$$

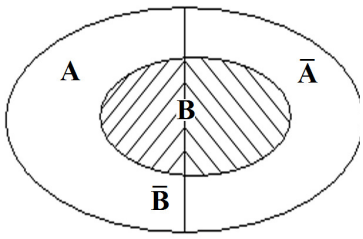
and

$$\Pr[B] = \Pr[A \cap B] + \Pr[\overline{A} \cap B] = \Pr[B|A]\Pr[A] + \Pr[B|\overline{A}](1 - \Pr[A]) \tag{2}$$

Combining equations (1) and (2), we have expressed $\Pr[A|B]$ in terms of $\Pr[A], \Pr[B|A]$ and $\Pr[B|\overline{A}]$:

$$\Pr[A|B] = \frac{\Pr[B|A]\Pr[A]}{\Pr[B|A]\Pr[A] + \Pr[B|\overline{A}](1-\Pr[A])} \tag{3}$$

This equation is useful for many inference problems. We are given $\Pr[A]$, which is the (unconditional) probability that the event of interest A happens. We are given $\Pr[B|A]$ and $\Pr[B|\overline{A}]$, which quantify how noisy the observation is. (If $\Pr[B|A] = 1$ and $\Pr[B|\overline{A}] = 0$, for example, the observation is completely noiseless.) Now we want to calculate $\Pr[A|B]$, the probability that the event of interest happens given we made the observation. Equation (3) allows us to do just that.



Equation (3) is at the heart of a subject called *Bayesian inference*, used extensively in fields such as machine learning, communications and signal processing. The equation can be interpreted as a way to *update knowledge* after making an observation. In this interpretation, $\Pr[A]$ can be thought of as a *prior* probability: our assessment of the likelihood of an event of interest A *before* making an observation. It reflects our prior knowledge. $\Pr[A|B]$ can be interpreted as the *posterior* probability of A after the observation. It reflects our new knowledge.

Of course, equations (1), (2) and (3) are derived from the basic axioms of probability and the definition of conditional probability, and are therefore true with or without the above Bayesian inference interpretation. However, this interpretation is very useful when we apply probability theory to study inference problems.

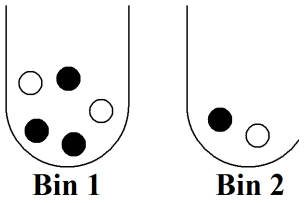
Bayes' Rule and Total Probability Rule

Equations (1) and (2) are very useful in their own right. The first is called **Bayes' Rule** and the second is called the **Total Probability Rule**. Bayes' Rule is useful when one wants to calculate $\Pr[A|B]$ but one is given $\Pr[B|A]$ instead: it allows us to "flip" things around. The Total Probability Rule is an application of the strategy of "dividing into cases" we learned in Note 2 to calculating probabilities. We want to calculate the probability of an event B. There are two possibilities: either an event A happens or A does not happen. If A happens the probability that B happens is $\Pr[B|A]$. If A does not happen, the probability that B happens is $\Pr[B|\overline{A}]$. If we know or can easily calculate these two probabilities and also $\Pr[A]$, then the total probability rule yields the probability of event B.

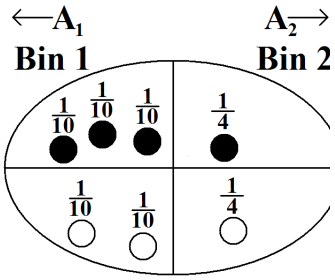
Example (medical test): Let us revisit the medical test described earlier in this reading. Suppose we wish to calculate $\Pr[A|P]$ using Bayes' rule rather than the original definition of conditional probability. We know the following: $\Pr[P|A] = 0.9$, $\Pr[A] = 0.05$, and $\Pr[P|H] = 0.2$ (note that in this case, the event that someone is not affected, $\overline{A}$, is equivalent to the event that they are healthy, H). Now we can simply apply Bayes' rule since we know the value of each term and get: $\Pr[A|P] = \frac{\Pr[P|A]\times\Pr[A]}{\Pr[P|A]\times\Pr[A]+\Pr[P|H]\times\Pr[H]} = \frac{0.9\times0.05}{0.9\times0.05+0.2\times0.95} \approx 0.19$, which is the same value we got earlier using the definition of conditional probability.

Example (balls and bins): Imagine we have two bins containing black and white balls: Bin 1 has three black and two white balls, and Bin 2 has one black and one white ball. Suppose we first pick a bin uniformly at random, and then pick a ball uniformly at random in that bin. What is the probability that we picked Bin 1

given that we picked a white ball, i.e, $\Pr[\text{Bin } 1|\bigcirc]$?



A wrong approach is to say that the answer is clearly $\frac{2}{3}$, since we know there are a total of three white balls, two of which are in bin 1. This picture is misleading because the bins have equal "weight". Instead, what we should do is appropriately scale each sample point as the following picture shows:



In this case, we have two alternatives A_1 and A_2 such that $\Omega = A_1 \cup A_2$ and $A_1 \cap A_2 = \emptyset$. We can use the definition of conditional probability to see that

$$\Pr[\text{Bin } 1|\bigcirc] = \frac{\frac{1}{10}+\frac{1}{10}}{\frac{1}{10}+\frac{1}{10}+\frac{1}{4}} = \frac{\frac{2}{10}}{\frac{9}{20}} = \frac{4}{9}$$

Let us try to achieve this probability using Bayes' rule. The probability that we choose bin 1 is $\Pr[A_1] = \frac{1}{2}$. The chance that we pick a white ball given that we picked bin 1 is $\Pr[\bigcirc|\text{Bin } 1] = \frac{2}{5}$ and the chance that we pick a white ball given that we picked bin 2 is $\Pr[\bigcirc|\text{Bin } 2] = \frac{1}{2}$. Therefore, we can just plug all of these values into Bayes' rule to get the chance that we picked bin 1 given that we picked a white ball:

$$\Pr[\text{Bin } 1|\bigcirc] = \frac{\frac{2}{5}\times\frac{1}{2}}{\frac{2}{5}\times\frac{1}{2}+\frac{1}{2}\times\frac{1}{2}} = \frac{\frac{2}{10}}{\frac{2}{10}+\frac{1}{4}} = \frac{4}{9}.$$

Independent events

Definition 12.1 (independence): Two events A, B in the same probability space are *independent* if $\Pr[A \cap B] = \Pr[A] \times \Pr[B]$.

The intuition behind this definition is the following. Suppose that $\Pr[B] > 0$. Then we have

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

Thus independence has the natural meaning that "the probability of A is not affected by whether or not B occurs." (By a symmetrical argument, we also have $\Pr[B|A] = \Pr[B]$ provided $\Pr[A] > 0$.) For events A, B such that $\Pr[B] > 0$, the condition $\Pr[A|B] = \Pr[A]$ is actually *equivalent* to the definition of independence.

Examples: In the balls and bins example at the start of the note, events A, B are not independent. In the dice example, events A, B are independent.

The above definition generalizes to any finite set of events:

Definition 12.2 (mutual independence): Events $A_1, \ldots, A_n$ are *mutually independent* if for every subset $I \subseteq \{1, \ldots, n\}$,

$$\Pr[\bigcap_{i \in I} A_i] = \prod_{i \in I} \Pr[A_i].$$

Note that we need this property to hold for *every* subset I.

For mutually independent events $A_1, \ldots, A_n$, it is not hard to check from the definition of conditional probability that, for any $1 \leq i \leq n$ and any subset $I \subseteq \{1, \ldots, n\} \setminus \{i\}$, we have

$$\Pr[A_i | \bigcap_{j \in I} A_j] = \Pr[A_i].$$

Note that the independence of every pair of events (so-called *pairwise independence*) does *not* necessarily imply mutual independence. For example, it is possible to construct three events A, B, C such that each *pair* is independent but the triple A, B, C is *not* mutually independent.

Combinations of events

In most applications of probability in Computer Science, we are interested in things like $\Pr[\bigcup_{i=1}^n A_i]$ and $\Pr[\bigcap_{i=1}^n A_i]$, where the A_i are simple events (i.e, we know, or can easily compute, the $\Pr[A_i]$). The intersection $\bigcap_i A_i$ corresponds to the logical AND of the events A_i, while the union $\bigcup_i A_i$ corresponds to their logical OR. As an example, if A_i denotes the event that a failure of type i happens in a certain system, then $\bigcup_i A_i$ is the event that the system fails.

In general, computing the probabilities of such combinations can be very difficult. In this section, we discuss some situations where it can be done.

Intersections of events

From the definition of conditional probability, we immediately have the following *product rule* (sometimes also called the *chain rule*) for computing the probability of an intersection of events.

Theorem 12.1: **[Product Rule]** For any events A, B, we have

$$\Pr[A \cap B] = \Pr[A] \Pr[B|A].$$

More generally, for any events $A_1, \ldots, A_n$,

$$\Pr[\bigcap_{i=1}^n A_i] = \Pr[A_1] \times \Pr[A_2|A_1] \times \Pr[A_3|A_1 \cap A_2] \times \cdots \times \Pr[A_n | \bigcap_{i=1}^{n-1} A_i].$$

Proof: The first assertion follows directly from the definition of $\Pr[B|A]$ (and is in fact a special case of the second assertion with $n = 2$).

To prove the second assertion, we will use induction on n (the number of events). The base case is $n = 1$, and corresponds to the statement that $\Pr[A] = \Pr[A]$, which is trivially true. For the inductive step, let $n > 1$ and assume (the inductive hypothesis) that

$$\Pr[\bigcap_{i=1}^{n-1} A_i] = \Pr[A_1] \times \Pr[A_2|A_1] \times \cdots \times \Pr[A_{n-1} | \bigcap_{i=1}^{n-2} A_i].$$

Now we can apply the definition of conditional probability to the two events A_n and $\bigcap_{i=1}^{n-1} A_i$ to deduce that

$$\begin{aligned} \Pr[\bigcap_{i=1}^n A_i] = \Pr[A_n \cap (\bigcap_{i=1}^{n-1} A_i)] &= \Pr[A_n | \bigcap_{i=1}^{n-1} A_i] \times \Pr[\bigcap_{i=1}^{n-1} A_i] \\ &= \Pr[A_n | \bigcap_{i=1}^{n-1} A_i] \times \Pr[A_1] \times \Pr[A_2|A_1] \times \cdots \times \Pr[A_{n-1} | \bigcap_{i=1}^{n-2} A_i], \end{aligned}$$

where in the last line we have used the inductive hypothesis. This completes the proof by induction. ■

The product rule is particularly useful when we can view our sample space as a sequence of choices. The next few examples illustrate this point.

1. **Coin tosses.** Toss a fair coin three times. Let A be the event that all three tosses are heads. Then $A = A_1 \cap A_2 \cap A_3$, where A_i is the event that the ith toss comes up heads. We have

$$\begin{aligned}\Pr[A] &= \Pr[A_1] \times \Pr[A_2|A_1] \times \Pr[A_3|A_1 \cap A_2] \\ &= \Pr[A_1] \times \Pr[A_2] \times \Pr[A_3] \\ &= \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = \tfrac{1}{8}.\end{aligned}$$

   The second line here follows from the fact that the tosses are mutually independent. Of course, we already know that $\Pr[A] = \frac{1}{8}$ from our definition of the probability space in the previous lecture note. The above is really a check that the space behaves as we expect[1].

   If the coin is biased with heads probability p, we get, again using independence,

$$\Pr[A] = \Pr[A_1] \times \Pr[A_2] \times \Pr[A_3] = p^3.$$

   And more generally, the probability of any sequence of n tosses containing r heads and $n-r$ tails is $p^r(1-p)^{n-r}$. This is in fact the reason we defined the probability space this way in the previous lecture note: we defined the sample point probabilities so that the coin tosses would behave independently.

2. **Balls and bins.** Let A be the event that Bin 1 is empty. We saw in the previous lecture note (by counting) that $\Pr[A] = (1-\frac{1}{n})^m$, where m is the number of balls and n is the number of bins. The product rule gives us a different way to compute the same probability. We can write $A = \bigcap_{i=1}^m A_i$, where A_i is the event that ball i misses Bin 1. Clearly $\Pr[A_i] = 1 - \frac{1}{n}$ for each i. Also, the A_i are mutually independent since ball i chooses its bin regardless of the choices made by any of the other balls. So

$$\Pr[A] = \Pr[A_1] \times \cdots \times \Pr[A_m] = \left(1 - \frac{1}{n}\right)^m.$$

3. **Card shuffling.** We can look at the sample space as a sequence of choices as follows. First the top card is chosen uniformly from all 52 cards, i.e, each card with probability $\frac{1}{52}$. Then (conditional on the first card), the second card is chosen uniformly from the remaining 51 cards, each with probability $\frac{1}{51}$. Then (conditional on the first two cards), the third card is chosen uniformly from the remaining 50, and so on. The probability of any given permutation, by the product rule, is therefore

$$\frac{1}{52} \times \frac{1}{51} \times \frac{1}{50} \times \cdots \times \frac{1}{2} \times \frac{1}{1} = \frac{1}{52!}.$$

   Reassuringly, this is in agreement with our definition of the probability space in the previous lecture note, based on counting permutations.

4. **Poker hands.** Again we can view the sample space as a sequence of choices. First we choose one of the cards (note that it is not the "first" card, since the cards in our hand have no ordering) uniformly from all 52 cards. Then we choose another card from the remaining 51, and so on. For any given poker hand, the probability of choosing it is (by the product rule):

$$\frac{5}{52} \times \frac{4}{51} \times \frac{3}{50} \times \frac{2}{49} \times \frac{1}{48} = \frac{1}{\binom{52}{5}},$$

[1] Strictly speaking, we should really also have checked from our original definition of the probability space that $\Pr[A_1]$, $\Pr[A_2|A_1]$ and $\Pr[A_3|A_1 \cap A_2]$ are all equal to $\frac{1}{2}$.

just as before. Where do the numerators 5, 4, 3, 2, 1 come from? Well, for the given hand the first card we choose can be any of the five in the hand: i.e, five choices out of 52. The second can be any of the remaining four in the hand: four choices out of 51. And so on. This arises because the order of the cards in the hand is irrelevant.

Let's use this view to compute the probability of a flush in a different way. Clearly this is $4 \times \Pr[A]$, where A is the probability of a Hearts flush. And we can write $A = \bigcap_{i=1}^{5} A_i$, where A_i is the event that the ith card we pick is a Heart. So we have

$$\Pr[A] = \Pr[A_1] \times \Pr[A_2|A_1] \times \cdots \times \Pr[A_5|\bigcap_{i=1}^{4} A_i].$$

Clearly $\Pr[A_1] = \frac{13}{52} = \frac{1}{4}$. What about $\Pr[A_2|A_1]$? Well, since we are conditioning on A_1 (the first card is a Heart), there are only 51 remaining possibilities for the second card, 12 of which are Hearts. So $\Pr[A_2|A_1] = \frac{12}{51}$. Similarly, $\Pr[A_3|A_1 \cap A_2] = \frac{11}{50}$, and so on. So we get

$$4 \times \Pr[A] = 4 \times \frac{13}{52} \times \frac{12}{51} \times \frac{11}{50} \times \frac{10}{49} \times \frac{9}{48},$$

which is exactly the same fraction we computed in the previous lecture note.

So now we have two methods of computing probabilities in many of our sample spaces. It is useful to keep these different methods around, both as a check on your answers and because in some cases one of the methods is easier to use than the other.

5. **Monty Hall.** Recall that we defined the probability of a sample point by multiplying the probabilities of the sequence of choices it corresponds to; thus, e.g,

$$\Pr[(1,1,2)] = \tfrac{1}{3} \times \tfrac{1}{3} \times \tfrac{1}{2} = \tfrac{1}{18}.$$

The reason we defined it this way is that we knew (from our model of the problem) the probabilities for each choice *conditional on* the previous one. Thus, e.g, the $\frac{1}{2}$ in the above product is the probability that Carol opens door 2 conditional on the prize door being door 1 and the contestant initially choosing door 1. In fact, we used these conditional probabilities to define the probabilities of our sample points.

Unions of events

You are in Las Vegas, and you spy a new game with the following rules. You pick a number between 1 and 6. Then three dice are thrown. You win if and only if your number comes up on at least one of the dice.

The casino claims that your odds of winning are 50%, using the following argument. Let A be the event that you win. We can write $A = A_1 \cup A_2 \cup A_3$, where A_i is the event that your number comes up on die i. Clearly $\Pr[A_i] = \frac{1}{6}$ for each i. Therefore,

$$\Pr[A] = \Pr[A_1 \cup A_2 \cup A_3] = \Pr[A_1] + \Pr[A_2] + \Pr[A_3] = 3 \times \frac{1}{6} = \frac{1}{2}.$$

Is this calculation correct? Well, suppose instead that the casino rolled six dice, and again you win iff your number comes up at least once. Then the analogous calculation would say that you win with probability $6 \times \frac{1}{6} = 1$, i.e, certainly! The situation becomes even more ridiculous when the number of dice gets bigger than 6.

The problem is that the events A_i are *not disjoint*: there are some sample points that lie in more than one of the A_i. (We could get really lucky and our number could come up on two of the dice, or all three.) So if we add up the $\Pr[A_i]$ we are counting some sample points more than once.

Fortunately, there is a formula for this, known as the *Principle of Inclusion/Exclusion*:

Theorem 12.2: **[Inclusion/Exclusion]** For events $A_1, \ldots, A_n$ in some probability space, we have

$$\Pr[\bigcup_{i=1}^n A_i] = \sum_{i=1}^n \Pr[A_i] - \sum_{\{i,j\}} \Pr[A_i \cap A_j] + \sum_{\{i,j,k\}} \Pr[A_i \cap A_j \cap A_k] - \cdots \pm \Pr[\bigcap_{i=1}^n A_i].$$

(In the above summations, $\{i, j\}$ denotes all unordered pairs with $i \neq j$, $\{i, j, k\}$ denotes all unordered triples of distinct elements, and so on.)

To compute $\Pr[\bigcup_i A_i]$, we start by summing the event probabilities $\Pr[A_i]$, then we *subtract* the probabilities of all pairwise intersections, then we *add* back in the probabilities of all three-way intersections, and so on.

We won't prove this formula here; but you might like to verify it for the special case $n = 3$ by drawing a Venn diagram and checking that every sample point in $A_1 \cup A_2 \cup A_3$ is counted exactly once by the formula. You might also like to prove the formula for general n by induction (in similar fashion to the proof of the Product Rule above).

Taking the formula on faith, what is the probability we get lucky in the new game in Vegas?

$$\Pr[A_1 \cup A_2 \cup A_3] = \Pr[A_1] + \Pr[A_2] + \Pr[A_3] - \Pr[A_1 \cap A_2] - \Pr[A_1 \cap A_3] - \Pr[A_2 \cap A_3] + \Pr[A_1 \cap A_2 \cap A_3].$$

Now the nice thing here is that the events A_i are mutually independent (the outcome of any die does not depend on that of the others), so $\Pr[A_i \cap A_j] = \Pr[A_i]\Pr[A_j] = (\frac{1}{6})^2 = \frac{1}{36}$, and similarly $\Pr[A_1 \cap A_2 \cap A_3] = (\frac{1}{6})^3 = \frac{1}{216}$. So we get

$$\Pr[A_1 \cup A_2 \cup A_3] = \left(3 \times \tfrac{1}{6}\right) - \left(3 \times \tfrac{1}{36}\right) + \tfrac{1}{216} = \tfrac{91}{216} \approx 0.42.$$

So your odds are quite a bit worse than the casino is claiming!

When n is large (i.e, we are interested in the union of many events), the Inclusion/Exclusion formula is essentially useless because it involves computing the probability of the intersection of every non-empty subset of the events: and there are $2^n - 1$ of these! Sometimes we can just look at the first few terms of it and forget the rest: note that successive terms actually give us an overestimate and then an underestimate of the answer, and these estimates both get better as we go along.

However, in many situations we can get a long way by just looking at the *first* term:

1. **Disjoint events.** If the events A_i are all *disjoint* (i.e, no pair of them contain a common sample point — such events are also called *mutually exclusive*), then

   $$\Pr[\bigcup_{i=1}^n A_i] = \sum_{i=1}^n \Pr[A_i].$$

   (Note that we have already used this fact several times in our examples, e.g, in claiming that the probability of a flush is four times the probability of a Hearts flush — clearly flushes in different suits are disjoint events.)

2. **Union bound.** Always, it is the case that

   $$\Pr[\bigcup_{i=1}^n A_i] \leq \sum_{i=1}^n \Pr[A_i].$$

   This merely says that adding up the $\Pr[A_i]$ can only *over*estimate the probability of the union. Crude as it may seem, in the next lecture note we'll see how to use the union bound effectively in a Computer Science example.