1. Basic rules of probability

So far we have defined the notion of probability space and probability of an event. But most often, we do not calculate probabilities from the definition. This is like in integration, where one defined the integral of a function as a limit of Riemann sums, but that definition is used only to find integrals of x^n, $\sin(x)$ and a few such functions. Instead, integrals of complicated expressions such as $x\sin(x)+2\cos^2(x)\tan(x)$ are calculated by various rules, such as substitution rule, integration by parts, etc. In probability we need some similar rules relating probabilities of various combinations of events to the individual probabilities.

Proposition 1. *Let (Ω, p) be a discrete probability space.*

(1) For any event A, we have $0 \le \mathbf{P}(A) \le 1$. Also, $\mathbf{P}(\emptyset) = 0$ and $\mathbf{P}(\Omega) = 1$.

(2) Finite additivity of probability: *If $A_1, \dots, A_n$ are pairwise disjoint events (i.e., $A_i \cap A_j = \emptyset$ if $i \neq j$), then $\mathbf{P}(A_1 \cup \dots \cup A_n) = \mathbf{P}(A_1) + \dots + \mathbf{P}(A_n)$. In particular, $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$ for any event A.*

(3) Countable additivity of probability: *If $A_1, A_2, \dots$ is a countable collection of pairwise disjoint events, then $\mathbf{P}(\cup A_i) = \sum_i \mathbf{P}(A_i)$.*

All of these may seem obvious, and indeed they would be totally obvious if we stuck to finite sample spaces. But the sample space could be countable, and then probability of events may involve infinite sums which need special care in manipulation. Therefore we must give a proof. In writing a proof, and in many future contexts, it is useful to introduce the following notation.

Notation: Let $A \subseteq \Omega$ be an event. Then, we define a function $\mathbf{1}_A : \Omega \to \mathbb{R}$, called the *indicator function of A*, as follows.

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Since a function from Ω to $\mathbb{R}$ is called a random variable, the indicator of any event is a random variable. All information about the event A is in its indicator function (meaning, if we know the value of $\mathbf{1}_A(\omega)$, we know whether or not ω belongs to A). For example, we can write $\mathbf{P}(A) = \sum_{\omega\in\Omega} \mathbf{1}_A(\omega) p_\omega$.

Now, we prove the proposition.

Proof. (1) By definition of probability space $\mathbf{P}(\Omega) = 1$ and $\mathbf{P}(\emptyset) = 0$. If A is any event, then $\mathbf{1}_\emptyset(\omega)p_\omega \le \mathbf{1}_A(\omega)p_\omega \le \mathbf{1}_\Omega(\omega)p_\omega$. By monotonicity of sums, we get

$$\sum_{\omega\in\Omega} \mathbf{1}_\emptyset(\omega)p_\omega \le \sum_{\omega\in\Omega} \mathbf{1}_A(\omega)p_\omega \le \sum_{\omega\in\Omega} \mathbf{1}_\Omega(\omega)p_\omega.$$

As observed earlier, these sums are just $\mathbf{P}(\emptyset)$, $\mathbf{P}(A)$ and $\mathbf{P}(\Omega)$, respectively. Thus, $0 \leq \mathbf{P}(A) \leq 1$.

(2) It suffices to prove it for two sets (Why?). Let A, B be two events such that $A \cap B = \emptyset$. Let $f(\omega) = p_\omega \mathbf{1}_A(\omega)$ and $g(\omega) = p_\omega \mathbf{1}_B(\omega)$ and $h(\omega) = p_\omega \mathbf{1}_{A\cup B}(\omega)$. Then, the disjointness of A and B implies that $f(\omega) + g(\omega) = h(\omega)$ for all $\omega \in \Omega$. Thus, by linearity of sums, we get

$$\sum_{\omega\in\Omega} f(\omega) + \sum_{\omega\in\Omega} g(\omega) = \sum_{\omega\in\Omega} h(\omega).$$

But, the three sums here are precisely $\mathbf{P}(A)$, $\mathbf{P}(B)$ and $\mathbf{P}(A \cup B)$. Thus, we get $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$.

(3) This is similar to finite additivity, but needs a more involved argument. We leave it as an exercise for the interested reader. ■

Exercise 2. Adapt the proof to prove that for a countable family of events A_k in a common probability space (no disjointness assumed), we have

$$\mathbf{P}(\cup_k A_k) \leq \sum_k \mathbf{P}(A_k).$$

Definition 3 (Limsup and liminf of sets)**.** If $\{A_k, k \geq 1\}$, is a sequence of subsets of Ω, we define

$$\limsup A_k = \bigcap_{N=1}^{\infty} \bigcup_{k=N}^{\infty} A_k, \qquad \text{and} \qquad \liminf A_k = \bigcup_{N=1}^{\infty} \bigcap_{k=N}^{\infty} A_k.$$

In words, $\limsup A_k$ is the set of all ω that belong to infinitely many of the A_ks, and $\liminf A_k$ is the set of all ω that belong to all but finitely many of the A_ks.

Two special cases are of increasing and decreasing sequences of events. This means $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ and $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$. In these cases, the limsup and liminf are the same (so we refer to it as the limit of the sequence of sets). It is $\cup_k A_k$ in the case of increasing events and $\cap_k A_k$ in the case of decreasing events.

Exercise 4. (Monotonicity of $\mathbf{P}$) Events below are all contained in a discrete probability space. Use countable additivity of probability to show that

(1) If A_k are increasing events with limit A, show that $\mathbf{P}(A)$ is the increasing limit of $\mathbf{P}(A_k)$.

(2) If A_k are decreasing events with limit A, show that $\mathbf{P}(A)$ is the decreasing limit of $\mathbf{P}(A_k)$.

Now we re-write the basic rules of probability as follows:

The basic rules of probability:

(1) $\mathbf{P}(\emptyset) = 0$, $\mathbf{P}(\Omega) = 1$ and $0 \leq \mathbf{P}(A) \leq 1$ for any event A.

(2) $\mathbf{P}\left(\bigcup_k A_k\right) \leq \sum_k \mathbf{P}(A_k)$ for any countable collection of events A_k.

(3) $\mathbf{P}\left(\bigcup_k A_k\right) = \sum_k \mathbf{P}(A_k)$ if A_k is a countable collection of pairwise disjoint events.

2. Inclusion-exclusion formula

In general, there is no simple rule for $\mathbf{P}(A\cup B)$ in terms of $\mathbf{P}(A)$ and $\mathbf{P}(B)$. Indeed, consider the probability space $\Omega = \{0,1\}$ with $p_0 = p_1 = \frac{1}{2}$. If $A = \{0\}$ and $B = \{1\}$, then $\mathbf{P}(A) = \mathbf{P}(B) = \frac{1}{2}$ and $\mathbf{P}(A\cup B) = 1$. However, if $A = B = \{0\}$, then $\mathbf{P}(A) = \mathbf{P}(B) = \frac{1}{2}$ as before, but $\mathbf{P}(A\cup B) = \frac{1}{2}$. This shows that $\mathbf{P}(A\cup B)$ cannot be determined from $\mathbf{P}(A)$ and $\mathbf{P}(B)$. Similarly for $\mathbf{P}(A\cap B)$ or other set constructions.

However, it is easy to see that $\mathbf{P}(A\cup B) = \mathbf{P}(A)+\mathbf{P}(B)-\mathbf{P}(A\cap B)$. This formula is not entirely useless, because in special situations we shall later see that the probability of the intersection is easy to compute and hence we may compute the probability of the union. Generalizing this idea to more than two sets, we get the following surprisingly useful formula.

Proposition 5 (Inclusion-Exclusion formula)**.** *Let (Ω, p) be a probability space and let $A_1,\ldots,A_n$ be events. Then,*

$$\mathbf{P}\left(\bigcup_{i=1}^{n} A_i\right) = S_1 - S_2 + S_3 - \cdots + (-1)^{n-1}S_n,$$

where

$$S_k = \sum_{1\le i_1<i_2<\ldots<i_k\le n} \mathbf{P}(A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_k}).$$

We give two proofs, but the difference is only superficial. It is a good exercise to reason out why the two arguments are basically the same.

First proof. For each $\omega\in\Omega$, we compute its contribution to the two sides. If $\omega\notin\bigcup_{i=1}^n A_i$, then p_ω is not counted on either side. Suppose $\omega\in\bigcup_{i=1}^n A_i$ so that p_ω is counted once on the left side. We count the number of times p_ω is counted on the right side by splitting into cases depending on the exact number of A_is that contain ω.

Suppose ω belongs to exactly one of the A_is. For simplicity let us suppose that $\omega\in A_1$, but $\omega\in A_i^c$ for $2\le i\le n$. Then p_ω is counted once in S_1 but not counted in $S_2,\ldots,S_n$.

Suppose ω belongs to A_1 and A_2 but not any other A_i. Then p_ω is counted twice in S_1 (once for $\mathbf{P}(A_1)$ and once for $\mathbf{P}(A_2)$) and subtracted once in S_2 (in $\mathbf{P}(A_1\cap A_2)$). Thus, it is effectively counted once on the right side. The same holds if ω belongs to A_i and A_j but not any other A_ks.

If ω belongs to $A_1,\ldots,A_k$ but not any other A_i, then on the right side, p_ω is added k times in S_1, subtracted $\binom{k}{2}$ times in S_2, added $\binom{k}{3}$ times in S_k, and so on. Thus, p_ω is effectively counted

$$\binom{k}{1} - \binom{k}{2} + \binom{k}{3} - \cdots + (-1)^{k-1}\binom{k}{k}$$

times. By the Binomial formula, this is just the expansion of $1-(1-1)^k$ which is 1. ■

Second proof. Use the definition to write both sides of the statement. Let $A = \cup_{i=1}^{n} A_i$.

$$\text{LHS} = \sum_{\omega \in A} p_\omega = \sum_{\omega \in \Omega} \mathbf{1}_A(\omega) p_\omega.$$

Now, we compute the right side. For any $i_1 < i_2 < \cdots < i_k$, we write

$$\mathbf{P}\left(A_{i_1} \cap \cdots \cap A_{i_k}\right) = \sum_{\omega \in \Omega} p_\omega \mathbf{1}_{A_{i_1} \cap \cdots \cap A_{i_k}}(\omega) = \sum_{\omega \in \Omega} p_\omega \prod_{\ell=1}^{k} \mathbf{1}_{A_{i_\ell}}(\omega).$$

Hence, the right hand side is given by adding over $i_1 < \cdots < i_k$, multiplying by $(-1)^{k-1}$ and then summing over k from 1 to n.

$$\begin{aligned}
\text{RHS} &= \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < \cdots < i_k \le n} \sum_{\omega \in \Omega} p_\omega \prod_{\ell=1}^{k} \mathbf{1}_{A_{i_\ell}}(\omega) \\
&= \sum_{\omega \in \Omega} \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < \cdots < i_k \le n} p_\omega \prod_{\ell=1}^{k} \mathbf{1}_{A_{i_\ell}}(\omega) \\
&= -\sum_{\omega \in \Omega} p_\omega \sum_{k=1}^{n} \sum_{1 \le i_1 < \cdots < i_k \le n} \prod_{\ell=1}^{k} (-\mathbf{1}_{A_{i_\ell}}(\omega)) \\
&= -\sum_{\omega \in \Omega} p_\omega \left(\prod_{j=1}^{n} (1 - \mathbf{1}_{A_j}(\omega)) \; - \; 1 \right) \\
&= \sum_{\omega \in \Omega} p_\omega \mathbf{1}_A(\omega)
\end{aligned}$$

because the quantity $\prod_{j=1}^{n} (1 - \mathbf{1}_{A_j}(\omega))$ equals -1 if ω belongs to at least one of the A_is, and is zero otherwise. Thus the claim follows. ■

As we remarked earlier, it turns out that in many settings it is possible to compute the probabilities of intersections. We give an example now.

Example 6. Place n distinguishable balls in r distinguishable urns at random. Let A be the event that some urn is empty. The probability space is $\Omega = \{\underline{\omega} = (\omega_1, \ldots, \omega_n) : 1 \le \omega_i \le r\}$ with $p_{\underline{\omega}} = r^{-n}$. Let $A_\ell = \{\underline{\omega} : \omega_i \neq \ell\}$ for $\ell = 1, 2 \ldots, r$. Then, $A = A_1 \cup \cdots \cup A_{r-1}$ (as A_r is empty, we could include it or not, makes no difference).

It is easy to see that $\mathbf{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = (r-k)^n r^{-n} = (1 - \frac{k}{r})^n$. We could use the inclusion-exclusion formula to write the expression

$$\mathbf{P}(A) = r\left(1 - \frac{1}{r}\right)^n - \binom{r}{2}\left(1 - \frac{2}{r}\right)^n + \cdots + (-1)^{r-2}\binom{r}{r-1}\left(1 - \frac{r-1}{r}\right)^n.$$

The last term is zero (since all urns cannot be empty). I donot know if this expression can be simplified any more.

We mention two useful formulas that can be proved on lines similar to the inclusion-exclusion principle. If we say "at least one of the events $A_1, A_2, \ldots, A_n$ occurs", we are talking about the union, $A_1 \cup A_2 \cup \cdots \cup A_n$. What about "at least m of the events $A_1, A_2, \ldots, A_n$ occur", how to express it with set operations. It is not hard to see that this set is precisely

$$B_m = \bigcup_{1 \le i_1 < i_2 < \cdots < i_m \le n} (A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}).$$

The event that "exactly m of the events $A_1, A_2, \ldots, A_n$ occur" can be written as

$$C_m = B_m \setminus B_{m+1} = \bigcup_{\substack{S \subseteq [n] \\ |S| = m}} \left(\bigcap_{i \in S} A_i \right) \bigcap \left(\bigcap_{i \notin S} A_i^c \right).$$

Exercise 7. Let $A_1, \ldots, A_n$ be events in a probability space (Ω, p) and let $m \le n$. Let B_m and C_m be as above. Show that

$$\begin{aligned} \mathbf{P}(B_m) &= \sum_{k=m}^{n} (-1)^{k-m} \binom{k-1}{k-m} S_k \\ &= S_m - \binom{m}{1} S_{m+1} + \binom{m+1}{2} S_{m+2} - \binom{m+2}{3} S_{m+3} + \cdots \\ \mathbf{P}(C_m) &= \sum_{k=m}^{n} (-1)^{k-m} \binom{k}{m} S_k \\ &= S_m - \binom{m+1}{1} S_{m+1} + \binom{m+2}{2} S_{m+2} - \binom{m+3}{3} S_{m+3} + \cdots \end{aligned}$$

3. Bonferroni's inequalities

Inclusion-exclusion formula is nice when we can calculate the probabilities of intersections of the events under consideration. Things are not always this nice, and sometimes that may be very difficult. Even if we could find them, summing them with signs according to the inclusion-exclusion formula may be difficult as Example 6 demonstrates. The *idea* behind the inclusion-exclusion formula can however be often used to compute *approximate values of probabilities*, which is very valuable in most applications. That is what we do next.

We know that $\mathbf{P}(A_1 \cup \cdots \cup A_n) \le \mathbf{P}(A_1) + \cdots + \mathbf{P}(A_n)$ for any events $A_1, \ldots, A_n$. This is an extremely useful inequality, often called the *union bound*. Its usefulness is in the fact that there is no assumption made about the events A_is (such as whether they are disjoint or not). The following inequalities generalize the union bound, and gives both upper and lower bounds for the probability of the union of a bunch of events.

Lemma 8 (Bonferroni's inequalities)**.** *Let $A_1, \ldots, A_n$ be events in a probability space (Ω, p) and let $A = A_1 \cup \cdots \cup A_n$. Then, $S_1 - S_2 \le \mathbf{P}(A) \le S_1$. More generally,*

$$\mathbf{P}(A) \le S_1 - S_2 + \cdots + S_m \quad \text{if } m \text{ is odd},$$

$$\mathbf{P}(A) \le S_1 - S_2 + \cdots - S_m \quad \text{if } m \text{ is even}.$$

Proof. We shall write out the proof for the cases $m = 1$ and $m = 2$. When $m = 1$, the inequality is just the union bound

$$\mathbf{P}(A) \le \mathbf{P}(A_1) + \cdots + \mathbf{P}(A_n)$$

which we know. When $m = 2$, the inequality to be proved is

$$\mathbf{P}(A) \ge \sum_k \mathbf{P}(A_k) - \sum_{k<\ell} \mathbf{P}(A_k \cap A_\ell)$$

To see this, fix $\omega \in \Omega$ and count the contribution of p_ω to both sides. Like in the proof of the inclusion-exclusion formula, for $\omega \notin A_1 \cup \cdots \cup A_n$, the contribution to both sides is zero. On the other hand, if ω belongs to exactly r of the sets for some $r \ge 1$, then it is counted once on the left side and $r - \binom{r}{2}$ times on the right side. Note that $r - \binom{r}{2} = \frac{1}{2}r(3-r)$ which is always non-positive (one if $r = 1$, zero if $r = 2$ and non-positive if $r \ge 3$). Hence, we get LHS $\ge$ RHS.

Similarly, one can prove the other inequalities in the series. We leave it as an exercise. The key point is that $r - \binom{r}{2} + \cdots + (-1)^{k-1}\binom{r}{k}$ is non-negative if k is odd and non-positive if k is even (prove this). Here, as always, $\binom{x}{y}$ is interpreted as zero if $y > x$. ■

Here is an application of these inequalities.

Example 9. Return to Example 6. We obtained an exact expression for the answer, but that is rather complicated. For example, what is the probability of having at least one empty urn when

$n = 40$ balls are placed at random in $r = 10$ urns? It would be complicated to sum the series. Instead, we could use Bonferroni's inequalities to get the following bounds.

$$r\left(1-\frac{1}{r}\right)^n - \binom{r}{2}\left(1-\frac{2}{r}\right)^n \leq \mathbf{P}(A) \leq r\left(1-\frac{1}{r}\right)^n.$$

If we take $n = 40$ and $r = 10$, the bounds we get are $0.1418 \leq \mathbf{P}(A) \leq 0.1478$. Thus, we get a pretty decent approximation to the probability. By experimenting with other numbers you can check that the approximations are good when n is large compared to r but not otherwise. Can you reason why?

4. Independence - A first look

We remarked in the context of inclusion-exclusion formulas that often the probabilities of intersections of events is easy to find, and then we can use them to find probabilities of unions, etc. In many contexts, this is related to one of the most important notions in probability.

Definition 10. Let A, B be events in a common probability space. We say that A and B are *independent* is $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$.

Caution: Independence should not be confused with disjointness! If A and B are disjoint, $\mathbf{P}(A \cap B) = 0$ and hence A and B can be independent if and only if one of $\mathbf{P}(A)$ or $\mathbf{P}(B)$ equals 0. Intuitively, if A and B are disjoint, then knowing that A occurred gives us a lot of information about B (that it did not occur!), so independence is not to be expected.

Example 11. Toss a fair coin n times. Then $\Omega = \{\underline{\omega} : \underline{\omega} = (\omega_1, \ldots, \omega_n),\ \omega_i \text{ is } 0 \text{ or } 1\}$ and $p_{\underline{\omega}} = 2^{-n}$ for each $\underline{\omega}$. Let $A = \{\underline{\omega} : \omega_1 = 0\}$ and let $B = \{\underline{\omega} : \omega_2 = 0\}$. Then, from the definition of probabilities, we can see that $\mathbf{P}(A) = 1/2$, $\mathbf{P}(B) = 1/2$ (because the elementary probabilities are equal, and both the sets A and B contain exactly 2^{n-1} elements). Further, $A \cap B = \{\underline{\omega} : \omega_1 = 1, \omega_2 = 0\}$ has 2^{n-2} elements, whence $\mathbf{P}(A \cap B) = 1/4$. Thus, $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$ and hence A and B are independent.

If two events are independent, then the probability of their intersection can be found from the individual probabilities. How do we check if two events are independent? By checking if the probability of the event is equal to the product of the individual probabilities! It seems totally circular and useless! There are many reasons why it is not an empty notion as we shall see.

Firstly, in physical situations dependence is related to a basic intuition we have about whether two events are related or not. For example, suppose you are thinking of betting Rs.1000 on a particular horse in a race. If you get the news that your cousin is getting married, it will perhaps not affect the amount you plan to bet. However, if you get the news that one of the other horses has been injected with undetectable drugs, it might affect the bet you want to place. In other words, certain events (like marriage of a cousin) have no bearing on the probability of the event of interest (the event that our horse wins) while other events (like the injection of drugs) do have an impact. This intuition is often put into the very definition of probability space that we have.

For example, in the above example of tossing a fair coin n times, it is our intuition that a coin does not remember how it fell previous times, and that chance of its falling head in any toss is just $1/2$, irrespective of how many heads or tails occured before[1] And this intuition was used in

[1] It may be better to attribute this to experience rather than intuition. There have been reasonable people in history who believed that if a coin shows heads in ten tosses in a row, then on the next toss it is more likely to show tails (to 'compensate' for the overabundance of heads)! Clearly this is also someone's intuition, and different from ours. Only experiment can decide which is correct, and any number of experiments with real coins show that our intuition is correct, and coins have no memory.

defining the elementary probabilities as 2^{-n} each. Since we started with the intuitive notion of independence, and put that into the definition of the probability space, it is quite expected that the event that the first toss is a head should be independent of the event that the second toss is a tail. That is the calculation shown above.

But, how is independence useful mathematically if the conditions to check independence are the very conclusions we want?! The answer to this lies in the following fact (to be explained later). When certain events are independent, then many other collections of events that can be made out of them also turn out to be independent. For example, if A, B, C, D are independent (we have not yet defined what this means!), then $A \cup B$ and $C \cup D$ are also independent. Thus, starting from independence of certain events, we get independence of many other events. For example, any event depending on the first four tosses is independent of eny event depending on the next five tosses.