Probability Theory

Lecturer: Michel Goemans

These notes cover the basic definitions of discrete probability theory, and then present some results including Bayes' rule, inclusion-exclusion formula, Chebyshev's inequality, and the weak law of large numbers.

1 Sample spaces and events

To treat probability rigorously, we define a *sample space* S whose elements are the possible outcomes of some process or experiment. For example, the sample space might be the outcomes of the roll of a die, or flips of a coin. To each element x of the sample space, we assign a probability, which will be a non-negative number between 0 and 1, which we will denote by $p(x)$. We require that

$$\sum_{x \in S} p(x) = 1,$$

so the total probability of the elements of our sample space is 1. What this means intuitively is that when we perform our process, exactly one of the things in our sample space will happen.

Example. The sample space could be $S = \{a, b, c\}$, and the probabilities could be $p(a) = 1/2$, $p(b) = 1/3$, $p(c) = 1/6$.

If all elements of our sample space have equal probabilities, we call this the *uniform probability distribution* on our sample space. For example, if our sample space was the outcomes of a die roll, the sample space could be denoted $S = \{x_1, x_2, \ldots, x_6\}$, where the event x_i correspond to rolling i. The uniform distribution, in which every outcome x_i has probability $1/6$ describes the situation for a fair die. Similarly, if we consider tossing a fair coin, the outcomes would be H (heads) and T (tails), each with probability $1/2$. In this situation we have the uniform probability distribution on the sample space $S = \{H, T\}$.

We define an *event* A to be a subset of the sample space. For example, in the roll of a die, if the event A was rolling an even number, then $A = \{x_2, x_4, x_6\}$. The probability of an event A, denoted by $\mathbb{P}(A)$, is the sum of the probabilities of the corresponding elements in the sample space. For rolling an even number, we have

$$\mathbb{P}(A) = p(x_2) + p(x_4) + p(x_6) = \frac{1}{2}$$

Given an event A of our sample space, there is a complementary event which consists of all points in our sample space that are *not* in A. We denote this event by $\neg A$. Since all the points in a sample space S add to 1, we see that

$$\mathbb{P}(A) + \mathbb{P}(\neg A) \;=\; \sum_{x \in A} p(x) + \sum_{x \notin A} p(x) \;=\; \sum_{x \in S} p(x) \;=\; 1,$$

and so $\mathbb{P}(\neg A) = 1 - \mathbb{P}(A)$.

Note that, although two elements of our sample space cannot happen simultaneously, two events can happen simultaneously. That is, if we defined A as rolling an even number, and B as rolling a small number (1,2, or 3), then it is possible for both A and B to happen (this would require a roll of a 2), neither of them to happen (this would require a roll of a 5), or one or the other to happen. We call the event that both A and B happen "A and B", denoted by $A \wedge B$ (or sometimes $A \cap B$), and the event that at least one happens "A or B", denoted by $A \vee B$ (or sometimes $A \cup B$).

Suppose that we have two events A and B. These divide our sample space into four disjoint parts, corresponding to the cases where both events happen, where one event happens and the other does not, and where neither event happens, see Figure 1. These cases cover the sample space, accounting for each element in it exactly once, so we get

$$\mathbb{P}(A \wedge B) + \mathbb{P}(A \wedge \neg B) + \mathbb{P}(\neg A \wedge B) + \mathbb{P}(\neg A \wedge \neg B) = 1.$$

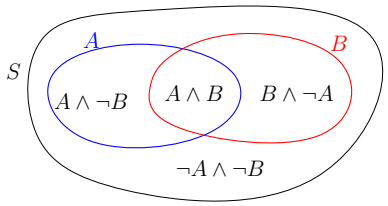

Figure 1: Two events A and B as subsets of the state space S.

2 Conditional probability and Independence

Let A be an event with non-zero probability. We define the *probability of an event B conditioned on event A*, denoted by $\mathbb{P}(B|A)$, to be

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \wedge B)}{\mathbb{P}(A)}.$$

Why is this an interesting notion? Let's give an example. Suppose we roll a fair die, and we ask what is the probability of getting an odd number, conditioned on having rolled a number that is at most 3? Since we know that our roll is 1, 2, or 3, and that they are equally likely (since we started with the uniform distribution corresponding to a fair die), then the probability of each of these outcomes must be $\frac{1}{3}$. Thus the probability of getting an odd number (that is, of getting 1 or 3) is $\frac{2}{3}$. Thus if A is the event "outcome is at most 3" and B is the event "outcome is odd", then we would like the mathematical definition of the "probability of B conditioned on A" to give $\mathbb{P}(B|A) = 2/3$. And indeed, mathematically we find

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \wedge A)}{\mathbb{P}(A)} = \frac{2/6}{1/2} = \frac{2}{3}.$$

The intuitive reason for which our definition of $\mathbb{P}(B|A)$ gives the answers we wanted is that the probability of every outcome in A gets multiplied by $\frac{1}{\mathbb{P}(A)}$ when one conditions on the event A.

It is a simple calculation to check that if we have two events A and B, then

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|\neg B)\mathbb{P}(\neg B).$$

Indeed, the first term is $\mathbb{P}(A \wedge B)$ and the second term $\mathbb{P}(A \wedge \neg B)$. Adding these together, we get

$$\mathbb{P}(A \wedge B) + \mathbb{P}(A \wedge \neg B) = \mathbb{P}(A).$$

If we have two events A and B, we say that they are *independent* if the probability that both happen is the product of the probability that the first happens and the probability that the second happens, that is, if

$$\mathbb{P}(A \wedge B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

Example. For a die roll, the events A of rolling an even number, and B of rolling a number less or equal to 3 are not independent, since $\mathbb{P}(A) \cdot \mathbb{P}(B) \neq \mathbb{P}(A \wedge B)$. Indeed, $\frac{1}{2} \cdot \frac{1}{2} \neq \frac{1}{6}$. However, if you define C to be the event of rolling a 1 or 2, then A and C are independent, since $\mathbb{P}(A) = \frac{1}{2}$, $\mathbb{P}(C) = \frac{1}{3}$, and $\mathbb{P}(A \wedge C) = \frac{1}{6}$.

Let us now show on an example that our mathematical definition of independence does capture the intuitive notion of independence. Let's assume that we toss two coins (not necessarily fair coins). The sample space is $S = \{HH, HT, TH, TT\}$ (where the first letter represents the result of the first coin). Let us denote the event of the first coin being a tail by $T\circ$, and the event of the second coin being a tail by $\circ T$ and so on. By definition, we have $\mathbb{P}(T\circ) = p(TH) + p(TT)$ and so on. Suppose that knowing that the first coin is a tail doesn't change the probability that the second coin is a tail. This gives

$$\mathbb{P}(\circ T \mid T\circ) = \mathbb{P}(\circ T).$$

Moreover, by definition of conditional probability,

$$\mathbb{P}(\circ T \mid T\circ) = \frac{\mathbb{P}(TT)}{\mathbb{P}(T\circ)}.$$

Combining these equations gives

$$\mathbb{P}(TT) = \mathbb{P}(T\circ)\mathbb{P}(\circ T),$$

or equivalently

$$\mathbb{P}(T\circ \wedge \circ T) = \mathbb{P}(T\circ)\mathbb{P}(\circ T).$$

Which is the condition we took to define the independence. Conclusion: knowing that the first coin is a tail doesn't change the probability that the second coin is a tail is the same as what we defined as "independence" between the events $T\circ$ and $\circ T$.

More generally, suppose that A and B are independent. In this case, we have

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \wedge B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(B).$$

That is, if two events are independent, then the probability of B happening, conditioned on A happening is the same as the probability of B happening without the conditioning. It is straightforward to check that the reasoning can be reversed as well: if the probability of B does not change when you condition on A, then the two events are independent.

We define k events $A_1 \ldots A_k$ to be *independent* if the intersection of any subset of these events is equal to the product of their probability, that is, if for all $1 \leq i_1 < i_2 < \cdots < i_s \leq k$,

$$\mathbb{P}(A_{i_1} \wedge A_{i_2} \wedge \cdots \wedge A_{i_s}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_s}).$$

It is possible to have a set of three events such that any two of them are independent, but all three are not independent. It is an interesting exercise to try to find such an example.

If we have k probability distributions on sample spaces $S_1 \ldots S_k$, we can construct a new probability distribution called the *product distribution* by assuming that these k processes are independent. Our new sample space is made of all the k-tuples $(s_1, s_2, \ldots, s_k)$ where $s_i \in S_i$. The probability distribution on this sample space is defined by

$$p(s_1, s_2, \ldots, s_k) = \prod_{i=1}^{k} p(s_i).$$

For example, if you roll k dice, your sample space will be the set of tuples $(s_1, \ldots, s_k)$ where $s_i \in \{x_1, x_2, \ldots, x_6\}$. The value of s_i represents the result of the i-th die (for instance $s_i = x_3$ means that the i-th die rolled 3). For 2 dice, the probability of rolling a one and a two will be

$$p(x_1, x_2) + p(x_2, x_1) = \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} = 1/18,$$

because you could have rolled the one with either the first die or the second die. The probability of rolling two ones is $\mathbb{P}(x_1, x_1) = 1/36$.

3 Bayes' rule

If we have a sample space, then conditioning on some event A gives us a new sample space. The elements in this new sample space are those elements in event A, and we normalize their probabilities by dividing by $\mathbb{P}(A)$ so that they will still add to 1.

Let us consider an example. Suppose we have two coins, one of which is a trick coin, which has two heads, and one of which is normal, and has one head and one tail. Suppose you toss a random one of these coins. You observe that it comes up heads. What is the probability that the other side is tails? I'll tell you the solution in the next paragraph, but you might want to first test your intuition by guessing the answer.

To solve this puzzle, let's label the two sides of the coin with two heads: we call one of these H_1 and the other H_2. Now, there are four possibilities for the outcome of the above process, all equally likely. They are as follows:

coin 1	coin 2
H_1	H
H_2	T

If you observe heads, then you eliminate one of these four possibilities. Of the remaining three, the other side will be heads in two cases (if you picked coin 1) and tails in only one case (if you picked coin 2). Thus, the probability the other side is tails is equal to $\frac{1}{3}$.

A similar probability puzzle goes as follows: *You meet a woman who has two children, at least one of whom is a girl. What is the probability that the two children are girls?* The intended answer

is that if you choose a woman with two children randomly, with probability $\frac{1}{4}$, she has two boys, with probability $\frac{1}{2}$ she has one boy and one girl and with probability $\frac{1}{4}$, she has two girls. Thus the conditional probability that she has two girls, given that she has at least one, is $\frac{1}{3}$.[1] Note that the above calculation does not take into account the possibility of twins.

Exercise. *A woman lives in a school district where, one-fifth of women with two children have twins, and of these, one-fifth are identical twins (with both children being of the same sex). Now what is the probability that a woman with two children she meets at the party for parents of first grade girls has two daughters? [Assume that all mothers are equally likely to go to the party, that two siblings are in the same grade if and only if they are twins, and that children are equally likely to be boys or girls.]*

We now state Bayes' rule, which is a simple but useful identity.

Bayes' rule. For any two events A and B, one has

$$\mathbb{P}(B|A) = \mathbb{P}(A|B)\frac{\mathbb{P}(B)}{\mathbb{P}(A)}.$$

The proof of Bayes' rule is straightforward. Replacing the conditional probabilities in Bayes' rule by their definition, we get

$$\frac{\mathbb{P}(A \wedge B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B \wedge A)}{\mathbb{P}(B)}\frac{\mathbb{P}(B)}{\mathbb{P}(A)},$$

which is easily checked.

We now give a canonical application of Bayes' rule. Suppose there are some disease, which we will call disease L. Now, let us suppose that the incidence of the disease in the general population is around one in a thousand. Now, suppose that there is some test for the disease which works most of the time, but not all. There will be a false positive rate:

$$\mathbb{P}(\text{positive test}|\text{no disease}).$$

Let us assume that this probability of a false positive is 1/30. There will also be some false negative rate:

$$\mathbb{P}(\text{negative test}|\text{disease}).$$

Let us assume that this probability of a false negative is 1/10.

Now, is it a good idea to test everyone for the disease? We will use Bayes' rule to calculate the probability that somebody in the general population who tests positive actually has disease L. Let's define event A as testing positive and B as having the disease. Then Bayes' rule tells us that

$$\mathbb{P}(B|A) = \mathbb{P}(A|B)\frac{\mathbb{P}(B)}{\mathbb{P}(A)}.$$

[1] This might be contrary to your intuition. Indeed if you meet a woman with a girl, and have never seen her other child, this second child has probability 1/2 (and not 1/3) of being a girl. A way of making the question confusing so as to trick people is to ask: *You meet a woman who has two children, one of whom is a girl. What is the probability that the other is a girl?*.

What are these numbers. $\mathbb{P}(A|B)$ is the chance you test positive, given that you have disease L, which we find is $0.9 = 1 - 1/10$ by using the false negative rate. $\mathbb{P}(B) = 1/1000$ is the incidence of the disease. $\mathbb{P}(A)$ is a little harder to calculate. We can obtain it by using the formula

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|\neg B)\mathbb{P}(\neg B)$$

This gives

$$\mathbb{P}(A) = \frac{9}{10}\frac{1}{1000} + \frac{1}{30}\frac{999}{1000} \approx 0.0342.$$

You can see that this calculation is dominated by the rate of false positives. Then, using Bayes' rule, we find that

$$\mathbb{P}(B|A) = \mathbb{P}(A|B)\frac{B}{A} = 0.9\frac{0.001}{0.0342} \approx 0.0265.$$

That is, even if you test positive, the chance that you have disease L is only around 2.65 percent.

Whether it is a good idea to test everyone for disease L is a medical decision, which will depend on the severity of the disease, and the side effects of whatever treatment they give to people who test positive. However, anybody deciding whether it is a good idea should take into account the above calculation.

This is not just a theoretical problem. Recently, a number of medical clinics have been advertising whole body CAT scans for apparently healthy people, on the chance that they will detect some cancer or other serious illness early enough to cure it. The FDA and some other medical organizations are questioning whether the benefits outweigh the risks involved with investigating false positives (which may involve surgery) that ultimately turn out to be no threat to health.

4 The Inclusion-Exclusion Formula

Recall that if we have two events, A and B, then they divide the sample space into four mutually exclusive subsets. This corresponds to the formula

$$\mathbb{P}(A \wedge B) + \mathbb{P}(A \wedge \neg B) + \mathbb{P}(\neg A \wedge B) + \mathbb{P}(\neg A \wedge \neg B) = 1.$$

We will now derive a formula for $\mathbb{P}(A \vee B)$, the probability of at least one of A or B happening, by looking at the Venn diagram represented in Figure 2. This diagram divides the plane into four parts, each of which represents one of the four subsets the events A and B divide the sample space into.

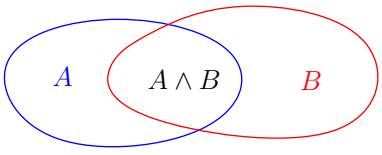

Figure 2: Venn diagram of two events A and B.

We see that if we take $\mathbb{P}(A) + \mathbb{P}(B)$, we have double counted all points of the sample space that are in both A and B, so we need to subtract their probabilities. This can be done by subtracting $\mathbb{P}(A \wedge B)$. We then get

$$\mathbb{P}(A \vee B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \wedge B).$$

Now, if we have three events, A, B and C, then we get the Venn diagram represented in Figure 3.

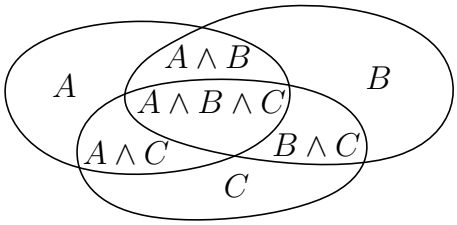

Figure 3: Venn diagram of three events A, B and C. (To clear any confusion, event A corresponds to 4 of the 8 regions in the Venn diagram, and $A \wedge B$ corresponds to two of them.)

We want to obtain a formula for $\mathbb{P}(A \vee B \vee C)$. If we take $\mathbb{P}(A)+\mathbb{P}(B)+\mathbb{P}(C)$, we have counted every point in the pairwise intersections twice, and every point in the triple intersection $A \wedge B \wedge C$ three times. Thus, to fix the pairwise intersections, we must subtract $\mathbb{P}(A \wedge B)+\mathbb{P}(B \wedge C)+\mathbb{P}(A \wedge C)$. Now, if we look at points in $A \wedge B \wedge C$, we started having counted every point in this set three times, and we then subtracted each of these points three times, so we have to add them back in again once. Thus, for three events, we get

$$\mathbb{P}(A \vee B \vee C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \wedge B) - \mathbb{P}(B \wedge C) - \mathbb{P}(A \wedge C) + \mathbb{P}(A \wedge B \wedge C).$$

Thus, to get the probability that at least one of these three events occurs, we add the probability of all events, subtract the intersection of all pairs of events, and add back the probability of the intersection of all three events.

The inclusion-exclusion formula can be generalized to n events. Here is the general result:

Theorem 1. *Let A_1, ..., A_n be events. Then the probability of their union is*

$$\begin{aligned} \mathbb{P}(A_1 \vee A_2 \vee \ldots \vee A_n) &= \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{1 \le i < j \le n} \mathbb{P}(A_i \wedge A_j) \\ &\quad + \sum_{1 \le i < j < k \le n} \mathbb{P}(A_i \wedge A_j \wedge A_k) - \ldots + (-1)^{n+1} \mathbb{P}(A_1 \wedge A_2 \wedge \ldots \wedge A_n). \end{aligned} \tag{1}$$

The last term has a + sign if n is odd and a − sign if n is even.

We will now prove Theorem 1. It is awkward to draw Venn diagrams for more than three events, and also trying to generalize the Venn diagram proof becomes unwieldy for an arbitrary number n of events. Instead we will prove the formula for n by induction. We have already shown it for $n = 2$ and 3, so we will assume that it holds for all numbers of events between 2 and $n - 1$, and prove that it holds for n. Let us divide the right-hand-side of Equation (1) into three parts. The first part will consist of all probabilities that do not contain explicitly the n-th event A_n. The second part will consist of all probabilities that contain both the n-th event and at least one other event. The third part will be the one remaining probability: $\mathbb{P}(A_n)$.

The first collection of probabilities is

$$\sum_{i=1}^{n-1} \mathbb{P}(A_i) - \sum_{1 \le i < j \le n-1} \mathbb{P}(A_i \wedge A_j) + \sum_{1 \le i < j < k \le n-1} \mathbb{P}(A_i \wedge A_j \wedge A_k) - \ldots \pm \mathbb{P}(A_1 \wedge A_2 \wedge \ldots \wedge A_{n-1}).$$

By the induction hypothesis, this is just

$$\mathbb{P}(A_1 \vee A_2 \vee \ldots \vee A_{n-1}).$$

The second collection of probabilities is the negative of

$$\sum_{i=1}^{n-1} \mathbb{P}(A_i \wedge A_n) - \sum_{1 \le i < j \le n-1} \mathbb{P}(A_i \wedge A_j \wedge A_n) + \ldots \pm \mathbb{P}(A_1 \wedge A_2 \wedge \ldots \wedge A_n).$$

This is the same as the right side of the inclusion-exclusion formula for $n-1$, except that every term has an additional $\wedge A_n$ included in it. We claim that this sum is

$$P\Big(\ (A_1 \vee A_2 \vee \ldots \vee A_{n-1}) \ \wedge A_n \Big).$$

There are two ways to prove this. The first is to let

$$\tilde{A}_i = A_i \wedge A_n.$$

Then, by induction, we have that the second collection of probabilities sums to

$$\mathbb{P}(\tilde{A}_1 \vee \tilde{A}_2 \vee \ldots \vee \tilde{A}_{n-1}) = P\Big(\ (A_1 \vee A_2 \vee \ldots \vee A_{n-1}) \ \wedge A_n \Big).$$

The other, which we won't go into the details of, is to consider the sample space obtained by conditioning on the event A_n.

Summarizing, we have shown that the right-hand-side of (1) is equal to

$$\mathbb{P}(A_1 \vee A_2 \vee \ldots \vee A_{n-1}) - P\Big(\ (A_1 \vee A_2 \vee \ldots \vee A_{n-1}) \ \wedge A_n \Big) + \mathbb{P}(A_n),$$

or

$$\mathbb{P}(B) - \mathbb{P}(B \wedge A_n) + \mathbb{P}(A_n)$$

if we define B to be the event $A_1 \vee A_2 \vee \ldots \vee A_{n-1}$. By the inclusion-exclusion formula for two events, this is equal to $\mathbb{P}(B \vee A_n)$, which is what we wanted to show (as this is precicely the left-hand-side of (1)). This completes the proof of Theorem 1.

Example of an application of the inclusion-exclusion formula.
Suppose that I have addressed n envelopes, and written n letters to go in them. My young child, wanting to be helpful, puts all the letters in the envelopes and gives them to the mailman. Unfortunately, it turns out that he has put them in at random. What is the probability that none of the letters goes into the correct envelopes?

We can solve this using the inclusion-exclusion formula. Let A_i be the event that the correct letter goes into the i-th envelope. Then, the probability that at least one letter has been addressed correctly is

$$\mathbb{P}(A_1 \vee A_2 \vee A_3 \vee \ldots \vee A_n)$$

and all we need do is calculate this probability using the inclusion-exclusion formula, and subtract it from 1. The inclusion exclusion formula says that this is

$$\sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{1 \le i < j \le n} \mathbb{P}(A_i \wedge A_j) + \sum_{1 \le i < j < k \le n} \mathbb{P}(A_i \wedge A_j \wedge A_k) \ldots \pm \mathbb{P}(A_1 \wedge A_2 \wedge \ldots \wedge A_n)$$

The first term is

$$\sum_{i=1}^{n} \mathbb{P}(A_i)$$

By symmetry, each of these n events has the same probability. Since A_i is the probability the right letter goes into the i-th envelope, and a random letter is inserted into the i-th envelope, these probabilities are all $\frac{1}{n}$, and the sum is 1.

The second term is

$$-\sum_{1 \le i < j \le n} \mathbb{P}(A_i \wedge A_j)$$

There are $\binom{n}{2} = n(n-1)/2$ terms. The event here is that both the i-th and the j-th letter go into the correct envelopes. The probability that we put the i-th letter into the correct envelope is, as before, $\frac{1}{n}$. Given that we have put the i-th letter into the correct envelope, the probability that we put the j-th letter into the correct envelope is $\frac{1}{n-1}$, since there are $n-1$ letters other than the i-th one, and they are all equally likely to go into the j-th envelope. This probability is then $\frac{1}{n(n-1)}$. The second term thus is (remembering the minus sign)

$$-\binom{n}{2}\frac{1}{n(n-1)} = -\frac{1}{2}.$$

The t-th term is

$$(-1)^{t+1} \sum_{1 \le j_1 < j_2 < \ldots < j_t \le n} \mathbb{P}(A_{j_1} \wedge A_{j_2} \wedge \ldots \wedge A_{j_t})$$

There are

$$\binom{n}{t} = \frac{n(n-1)\ldots(n-t+1)}{t!}$$

terms in this sum, and each term is the probability

$$\frac{1}{n(n-1)(n-2)\ldots(n-t+1)}.$$

Multiplying these quantities, we find that the t-th sum is $(-1)^{t+1}\frac{1}{t!}$. We thus have that the probability that at least one of the n letters goes into the right envelope is

$$1 - \frac{1}{2!} + \frac{1}{3!} - \ldots + (-1)^{n+1}\frac{1}{n!}$$

and subtracting this from 1, we get that the probability that none of these letters goes into the right envelope is

$$1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \ldots + (-1)^{n}\frac{1}{n!}.$$

This can be rewritten as

$$\sum_{k=0}^{n}(-1)^k\frac{1}{k!}.$$

You may recognize this as the first $n+1$ terms of the Taylor expansion of e^x, with the substitution $x = -1$. Thus, as n goes to ∞, the probability that none of the letters go into the correct envelope tends to $\frac{1}{e}$.

5 Expectation

So far we have dealt with events and their probabilities. Another very important concept in probability is that of a random variable. A *random variable* is simply a function f defined on the points of our sample space S. That is, associated with every $x \in S$, there is a value $f(x)$. For the time being, we will only consider functions that take values over the reals $\mathbb{R}$, but the range of a random variable can be any set.

We say that two random variables f and g are *independent* if the events $f(x) = \alpha$ and $g(x) = \beta$ are independent for any choice of values α, β in the range of f and g.

We define the *expected value* of a random variable f to be

$$\mathbb{E}(f) = \sum_{x \in S} p(x) f(x).$$

The expectation is also sometimes denoted by $\bar{f}$.

Another expression for the expectation is

$$\mathbb{E}(f) = \sum_{\alpha \in \text{range}(f)} \alpha \mathbb{P}(f = \alpha).$$

This is straightforward to verify using the fact that

$$\mathbb{P}(f = \alpha) \quad = \sum_{x \in S : f(x) = \alpha} p(x).$$

Suppose that we have an event A. There is an important random variable I_A associated with A, called an *indicator variable* for A:

$$I_A(x) = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A. \end{cases}$$

We can see that

$$\mathbb{E}(I_A) = \sum p(x) I_A(x) = \sum_{x \in A} p(x) = \mathbb{P}(A).$$

At first glance, there might not seem like much point in using such a simple random variable. However, it can be very useful, especially in conjunction with the following important fact:

Linearity of expectation. A very useful fact about expectation is that it is linear. That is, if we have two functions, f and g, then

$$\mathbb{E}(f + g) = \mathbb{E}(f) + \mathbb{E}(g),$$

and if we have a constant $\alpha \in \mathbb{R}$,

$$\mathbb{E}(\alpha f) = \alpha \mathbb{E}(f).$$

This is straightforward to prove. The proof of the first of these equations is as follows:

$$\mathbb{E}(f + g) = \sum_{x \in S} p(x)(f(x) + g(x)) = \sum_{x \in S} p(x) f(x) + \sum_{x \in S} p(x) g(x) = \mathbb{E}(f) + \mathbb{E}(g).$$

The proof of the second is essentially similar, and we will not give it.

It it tempting to hope that $\mathbb{E}(f) \cdot \mathbb{E}(g) = \mathbb{E}(f \cdot g)$, but this is **false** in general. For example, for an indicator random variable I_A (thus taking values only 0 or 1), we have that $\mathbb{E}(I_A) = \mathbb{P}(A)$ while $\mathbb{E}(I_A \cdot I_A) = \mathbb{E}(I_A^2) = \mathbb{E}(I_A) = \mathbb{P}(A)$ which is not $\mathbb{P}(A)^2$ (unless $\mathbb{P}(A)$ is 0 or 1). However, if the two random variables f and g are independent, then equality does hold:

$$\begin{aligned}
\mathbb{E}(f \cdot g) &= \sum_{x \in S} p(x) f(x) g(x) \\
&= \sum_{\alpha} \sum_{\beta} \sum_{x \in S : f(x) = \alpha, g(x) = \beta} p(x) \alpha \beta \\
&= \sum_{\alpha} \sum_{\beta} \alpha \beta \mathbb{P}(f(x) = \alpha \wedge g(x) = \beta) \\
&= \sum_{\alpha} \sum_{\beta} \alpha \beta \mathbb{P}(f(x) = \alpha) \mathbb{P}(g(x) = \beta) \quad \text{(by independence)} \\
&= \Big(\sum_{\alpha} \alpha \mathbb{P}(f(x) = \alpha) \Big) \Big(\sum_{\beta} \beta \mathbb{P}(g(x) = \beta) \Big) \\
&= \mathbb{E}(f) \mathbb{E}(g).
\end{aligned}$$

So to summarize: we *always* have $\mathbb{E}(\alpha f + \beta g) = \alpha \mathbb{E}(f) + \beta \mathbb{E}(g)$, and if in addition f and g are independent, then $\mathbb{E}(f) \cdot \mathbb{E}(g) = \mathbb{E}(fg)$.

Exercise. *It is however possible for $\mathbb{E}(f) \cdot \mathbb{E}(g)$ to equal $\mathbb{E}(fg)$ for f and g that are not independent; given an example of such a pair of random variables.*

6 Variance

Another quantity associated with a random variable is its *variance*. This is defined as

$$\mathrm{Var}(f) = \mathbb{E}[(f - \bar{f})^2].$$

That is, the variance is the expectation of the square of the difference between the value of f and the expected value $\bar{f} = \mathbb{E}(f)$ of f. We can also expand this into:

$$\mathrm{Var}(f) = \sum_{x \in S} p(x)(f(x) - \bar{f})^2.$$

The variance tells us how closely the value of a random variable is clustered around its expected value. You might be more familiar with the *standard deviation* σ; the standard deviation of f is defined to be the square root of $\mathrm{Var}(f)$.

We can rewrite the definition of the variance as follows:

$$\begin{aligned}
\mathrm{Var}(f) &= \mathbb{E}[(f - \bar{f})^2] \\
&= \mathbb{E}(f^2 - 2\bar{f}f + \bar{f}^2) \\
&= \mathbb{E}[f^2] - 2\bar{f}\mathbb{E}[f] + \bar{f}^2 \\
&= \mathbb{E}[f^2] - \bar{f}^2
\end{aligned}$$

so the variance of f is the expectation of the square of f minus the square of the expectation of f. We get from the second line to the third line by using linearity of expectation, and the third to the fourth by using the definition $\mathbb{E}(f) = \bar{f}$. Notice that the variance is always nonnegative, and that it is equal to 0 if f is constant.

Let us compute the variance and standard deviation of the roll of a die. Let the number on the die be the random variable X. We have that each of the numbers 1 through 6 are equally likely, so

$$\mathbb{E}(X) = \sum_{i=1}^{6} p(i)i = \sum_{i=1}^{6} \frac{1}{6} i = \frac{21}{6}$$

and

$$\mathbb{E}[X^2] = \sum_{i=1}^{6} p(i)i^2 = \sum_{i=1}^{6} \frac{1}{6} i^2 = \frac{91}{6}.$$

So

$$\mathrm{Var}(X) = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = \frac{35}{12}$$

and the standard deviation

$$\sigma = \sqrt{\frac{35}{12}} = 1.7078.$$

One can show that by the Cauchy-Schwartz inequality that

$$\mathbb{E}[|f - \bar{f}|] \leq \sqrt{\mathrm{Var}(f)}.$$

Recall that the Cauchy Schwartz inequality says that for two vectors $\vec{s}$ and $\vec{t}$, the inner product $\sum_i s_i t_i$ is at most the product of their lengths. For s_i choose $\sqrt{p(x_i)}|f(x_i) - \bar{f}|$ and for t_i choose $\sqrt{p(x_i)}$. Then their inner product is the expected value of $|f - \bar{f}|$, the length of $\vec{s}$ is the standard deviation, and the length of $\vec{t}$ is 1. For the die example above, we have that the expected value of $|f - \bar{f}|$ is 3/2, which is slightly less than the standard deviation of 1.7078.

If we a random variable f and a real c then $\mathrm{Var}(cf) = c^2\mathrm{Var}(f)$. Suppose we have two random variables f and g, and want to compute the variance of their sum. We get

$$\begin{aligned}
\mathrm{Var}(f+g) &= \mathbb{E}[(f+g)^2 - (\bar{f}+\bar{g})^2] \\
&= \mathbb{E}[f^2] + 2\mathbb{E}[fg] + \mathbb{E}[g^2] - (\bar{f}^2 + 2\bar{f}\bar{g} + \bar{g}^2) \\
&= \mathbb{E}[f^2] - \bar{f}^2 + 2\mathbb{E}[g^2] - \bar{g}^2 + 2\mathbb{E}[fg] - 2\bar{f}\bar{g} \\
&= \mathrm{Var}(f) + \mathrm{Var}(g) + 2\mathbb{E}[fg] - 2\bar{f}\bar{g}
\end{aligned}$$

This last quantity, $\mathbb{E}[fg] - \bar{f}\bar{g}$, is called the *covariance*. Recall from the previous section that if f and g are independent, $\mathbb{E}[fg] = \bar{f}\bar{g}$, and so the covariance is 0. Thus we have that if f and g are independent,

$$\mathrm{Var}(f+g) = \mathrm{Var}(f) + \mathrm{Var}(g);$$

variance is linear *if* we have independence (whereas expectation is linear even *without* independence).

Finally, suppose we have k random variables, $f_1, \ldots, f_k$, which are *pairwise independent*: for any $i \neq j$, f_i and f_j are independent. What is the variance of the random variable $f = f_1 + f_2 + f_3 + \ldots + f_k$? We have, using the same reasoning as above,

$$\begin{aligned}
\mathrm{Var}(f) &= \mathbb{E}(f_1 + \ldots + f_k)^2 - (\bar{f}_1 + \ldots + \bar{f}_k)^2 \\
&= \sum_{i=1}^{k} \mathbb{E}[f_i^2] - \sum_{i=1}^{k} \bar{f}_i^{\,2} + 2 \sum_{1 \leq i < j \leq k} \mathbb{E}[f_i f_j] - 2 \sum_{1 \leq i < j \leq k} \bar{f}_i \bar{f}_j \\
&= \sum_{i=1}^{k} \mathrm{Var}(f_i),
\end{aligned}$$

using independence in the last step. This is a very useful fact!

This lets us calculate the variance of the number of heads if you flip a coin n times, say. Suppose a biased coin has probability p of coming up heads and probability $q = 1 - p$ of coming up tails. Then define f to be 1 if the coin comes up heads and 0 if the coin comes up tails. The variance is

$$\mathrm{Var}(f) = \mathbb{E}[f^2] - (\mathbb{E}f)^2 = p - p^2 = p(1-p).$$

If you flip the coin n times, the results of each coin flip are independent, Thus the variance is n times the variance of a single coin flip, or $np(1-p)$.

7 Chebyshev's inequality

A very useful inequality, that can give bounds on probabilities, can be proved using the tools that we have developed so far. This is Chebyshev's inequality, and it is used to get an upper bound on the probability that a random variable takes on a value that is too far away from its mean.

Suppose you know the mean and variance of a random variable f. Is there some way that you can put a bound on the probability that the random variable is a long way away from its mean? This is exactly what Chebyshev's inequality does. Let's first derive Chebyshev's inequality intuitively, and then figure out how to turn this into a mathematically rigorous proof.

We will turn the probability around. Suppose we fix the probability p that the random variable is farther than $c\sigma$ away (for some given c) from the mean $\bar{f}$ (recall σ is the standard deviation $\sqrt{\mathrm{Var}(f)}$). Let's see how small the variance can be. Let's divide the sample space into two events. The first (which happens with probability say $1 - p$) is that

$$|f - \bar{f}| < c\sigma.$$

In this case, the way to minimize their contribution to the variance is to set $f(x) = \bar{f}$, and the contribution of this case to the variance is 0. The second event is when

$$|f - \bar{f}| \geq c\sigma.$$

In this case, the way to minimize the variance is to set $|f(x) - \bar{f}| = c\sigma$, and the contribution of this case to the variance is $pc^2\sigma^2$. Since the variance is σ^2, we have

$$\begin{aligned}
\sigma^2 &= pc^2\sigma^2, \\
p &= c^{-2}.
\end{aligned}$$

Since this was the case that minimized the variance, in any other case, the variance has to be greater than this. This gives us Chebyshev's inequality, namely

$$\mathbb{P}(\,|f-\bar{f}|\geq c\sigma\,)\leq 1/c^2$$

or, letting $y=c\sigma$,

$$\mathbb{P}(\,|f-\bar{f}|\geq y\,)\leq \sigma^2/y^2.$$

Now, let's turn this derivation into rigorous mathematical formulas. We first write down the formula for the variance

$$\sigma^2=\sum_{x\in S}p(x)|f(x)-\bar{f}|^2.$$

Now, let's divide it into the two cases we talked about above.

$$\sigma^2=\sum_{x:|f(x)-\bar{f}|<c\sigma}p(x)|f(x)-\bar{f}|^2+\sum_{x:|f(x)-\bar{f}|\geq c\sigma}p(x)|f(x)-\bar{f}|^2.$$

For the first sum, we have $\sum p(x)=1-p$ and for the second sum, we have $\sum p(x)=p$, where p is the probability that $|f-\bar{f}|\geq c\sigma$. Similarly, for the first sum, we have $|f(x)-\bar{f}|^2\geq 0$ and for the second sum, we have $|f(x)-\bar{f}|^2\geq c^2\sigma^2$. Putting these facts together, we have

$$\sigma^2\geq pc^2\sigma^2,$$

which gives

$$p\leq 1/c^2,$$

and proves Chebyshev's inequality.

8 Weak law of large numbers

Using Chebyshev's inequality, we can now show the so-called *weak law of large numbers*. This law says that if we have a random variable f (say the value resulting from the roll of a die) and take many independent copies of it, the average value of all these copies will be very close to the expected value of f. More formally, the weak law of large numbers is the following.

Theorem 2 (Weak law of large numbers)**.** *Fix $\epsilon>0$. Let $f_1,\cdots,f_n$ be n independent copies of a random variable f. Let*

$$g_n=\frac{1}{n}(f_1+f_2+\cdots f_n).$$

Then

$$\mathbb{P}\left[|g_n-\bar{f}|\geq\epsilon\right]\to 0$$

as $n\to\infty$.

In plain English, the probability that g_n deviates from the expected value of f by at least ϵ becomes arbitrarily small as n grows arbitrarily large.

The weak law of large numbers can be proved by using Chebyshev's inequality applied to g_n. For this, we need to know $\mathbb{E}[g_n]$ and $\mathrm{Var}(g_n)$. By linearity of expectations, we have

$$\mathbb{E}[g_n] = \mathbb{E}\left[\frac{1}{n}(f_1 + \cdots + f_n)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[f_i] = \frac{n}{n}\mathbb{E}[f] = \bar{f}.$$

For the variance, we get

$$\begin{aligned}
\mathrm{Var}[g_n] &= \mathrm{Var}\left[\frac{1}{n}(f_1 + \cdots + f_n)\right] \\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} f_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[f_i] \\
&= \frac{1}{n}\mathrm{Var}[f],
\end{aligned}$$

the third equality being true since the f_i's are *independent*. Thus, as n tends to infinity, $\mathbb{E}[g_n]$ remains constant while $\mathrm{Var}[g_n]$ tends to 0. For example, we saw that the roll of a fair die gives a variance of $\frac{35}{12}$. If we were to roll the die 1000 times and average all 1000 values, we would get a random variable whose expected value is still 3.5 but whose variance is much smaller, it is only $\frac{35}{12{,}000}$.

Now that we know the expected value and variance of g_n, we can simply use Chebyshev's inequality on g_n to get:

$$\mathbb{P}[|g_n - f| \geq \epsilon] \leq \frac{\mathrm{Var}[g_n]}{\epsilon^2} = \frac{\mathrm{Var}[f]}{n\epsilon^2},$$

and indeed this probability tends to 0 as n tends to infinity. This proves the weak law of large numbers.