Probability Notes (A) 18.310, Fall 2010

We are going to be spending around four lectures on probability theory this year. These notes cover approximately the first three lectures on it.

Probability theory is, surprisingly, an area that people generally do not have a very good intuition for. One reflection of this might be the recent financial crisis, which seems to have arisen in part because of substantial widespread misjudgment about the risk of certain financial positions.

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 point x of the sample space, we assign a probability, which will be a positive number between 0 and 1, which we will call $p(x)$. We will 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.

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 events would be x_1, x_2, ..., x_6, corresponding to rolling a 1, 2, ..., 6. Each element would have probability 1/6. If we consider tossing a fair coin, the outcomes would be H (heads) and T (tails), with the probability of each one being 1/2.

We will 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 $P(A)$, is the sum of the probabilities of the corresponding elements in the sample space. For rolling an even number, we have

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

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 call this event $\neg A$. Since all the points in a sample space S add to 1, we see that

$$P(A) + P(\neg A) = \sum_{x \in A} p(x) + \sum_{x \notin A} p(x)$$

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

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

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 $A \wedge B$, and the event that at least one happens A or B, denoted by $A \vee B$.

2 Conditional Probability and Independence

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. These cases cover the sample space, accounting for each element in it exactly once, so we get

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

We define the probability of event 2, conditioned on event 1, and denoted $P(B|A)$, to be

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

Why is this a reasonable thing to do? Let's give an example. Suppose we roll a die, and we ask what is the probability of getting an even number, conditioned on our having rolled a number that is at most 3? If we know that our roll is 1, 2, or 3, then there are three possible numbers it could be. If they are equally likely (what else could happen in this case?) then the probability of each of them must be $\frac{1}{3}$. Formally, we have

$$P(\text{even}| \leq 3) = \frac{P(\text{even})}{P(\leq 3)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

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

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

The first term is $P(A \wedge B)$ and the second term $P(A \wedge \not B)$. Adding these together, we get

$$P(A \wedge B) + P(A \wedge \not B) = P(A).$$

If we have two events A and B, we will 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

$$P(A \wedge B) = P(A) \cdot P(B).$$

We now explain why this condition makes sense. Let's assume that we toss two coins (not necessarily fair coins). The sample space is HH, HT, TH, TT, and let us represent the probabilities of these events by p_{HH}, p_{HT}, p_{TH}, p_{TT}. Let us denote the probability of the first coin being a tail is $p_{\mathrm{T}\circ} = p_{\mathrm{TH}} + p_{\mathrm{TT}}$, 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

$$P(\text{2nd is T}|\text{1st is T}) = P(\text{2nd is T})$$

but by our definition of conditional probability,

$$P(\text{2nd is T}|\text{1st is T}) = \frac{p_{TT}}{p_{T\circ}}$$

so substituting this into the abbove equation gives

$$\frac{p_{TT}}{p_{T\circ}} = p_{\circ T}$$

or

$$p_{TT} = p_{\circ T} p_{T\circ},$$

which is just

$$P(\text{1st is T} \wedge \text{2nd is T}) = P(\text{1st is T})P(\text{2nd is T})$$

which was our condition for independence.

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

We define k events $A_1 \ldots A_k$ to be independent if

a) Any subset of $k-1$ of these events is independent, and

b) $P(A_1 \wedge A_2 \wedge \ldots \wedge A_k) = P(A_1)P(A_2)\ldots P(A_k)$

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 contains elements $(x_1, x_2, \ldots, x_k)$ where $x_i \in S_i$, and has a probability distribution defined by

$$P(x_1, x_2, \ldots, x_k) = \Pi_{i=1}^{k} P(x_i).$$

For example, if you roll k dice, you will obtain the product distribution. The probability of rolling a one and a two will be

$$P(1,2) + P(2,1) = \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} = 1/18,$$

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

It is possible to have a set of three events such that any two of them are independent, but all three are not independent. This will be an exercise.

3 Conditional Probability and 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 $P(A)$ so that they will still add to 1. For instance, the example given at the end of the previous section can be obtained by flipping a fair coin k times, and then conditioning on the number of 1's being even.

Let us now consider 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 is $\frac{1}{3}$.

A similar probability puzzle goes as follows: *You meet a woman who has two children, one of whom is a girl. What is the probability that the other is a girl?* 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 [(either her older child or her younger child can be the 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 $\frac{1}{3}$.

This may be contrary to your intuition. Let me remark that this also depends on where you meet the woman. Suppose you meet her at a party for parents of first grade girls (or, maybe more likely, at a birthday party for a first grader to which only girls are invited). Then the probability calculation goes as follows. You know she has two children, one a first-grader and the other very likely not a first-grader. The first-grader is a girl, and the other one is equally likely to be a boy or a girl, so the probability is $\frac{1}{2}$. [Of course, this ignores the possibility of her having twins.]

Suppose that A and B are independent. In this case, we have

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

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.

Bayes' rule says that if we have two events, A and B, then

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

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

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

an identity.

We now give one of the canonical examples of an application of Bayes' rule., Suppose we have some disease, which we will call disease D. 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:

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

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

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

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 D. Let's define event A as testing positive and B as having the disease. Then Bayes' rule tells us that

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

or

$$P(\text{disease}|\text{test } +) = P(\text{test } + |\text{disease})\frac{P(\text{disease})}{P(\text{test } +)}.$$

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

$$P(A) = P(A|B)P(B) + P(A|\text{not}B)P(\text{not}B)$$

This gives

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

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

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

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

Whether it is a good idea to test everyone for disease D 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.

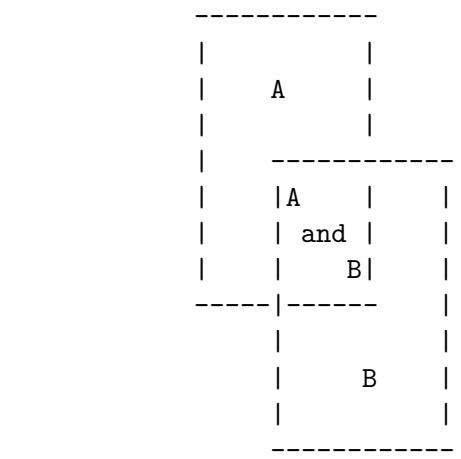
This is not just a theoretical problem. Recently, a number of medical clinics have been advertising whole body CAT scans for apparently healthy people, on the chance that they will detect some cancer or other serious illness early enough to cure it. The FDA and some other medical organizations are questioning 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

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

We will now derive a formula for $P(A \vee B)$, the probability of at least one of A or B happening, by looking at the Venn diagram below.

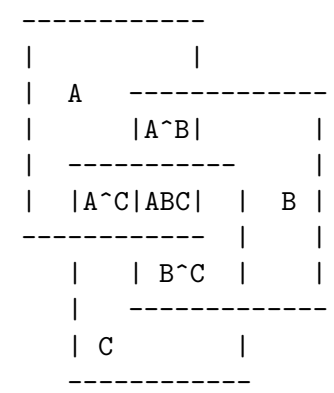


This diagram divides the plane into four parts, each of which represents one of the four subsets the events A and B divide the sample space into.

We see that if we take $P(A) + P(B)$, 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 $P(A \wedge B)$. We then get

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

Now, if we have three events, A, B and C, then we get the Venn diagram



We want to obtain a formula for $P(A \vee B \vee C)$. If we take $P(A) + P(B) + P(C)$, 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 $P(A \wedge B) + P(B \wedge C) + P(A \wedge C)$. 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

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

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 in a straightforward way. It is awkward to draw Venn diagrams for more than three events, and also trying to generalize the Venn diagram proof becomes unwieldly for an arbitrary number n of events. 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 name our events $A_1, \ldots, A_n$. What we would like to show is that

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

We will prove this by induction. Let us divide the right-hand-side up this equation into three parts. The first part will consist of all probabilities that do not contain

the nth event A_n. The second part will consist of all probabilities that contain both the nth event and at least one other event. The third part will be the one remaining probability: $P(A_n)$.

The first collection of probabilities is

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

By the induction hypothesis, this is just

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

The second collection of probabilities is the negative of

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

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

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

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,

Now, we have that

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

We let B be the event $A_1 \wedge A_2 \wedge \ldots \wedge A_{n-1}$, then the right-hand-side of the above equation is

$$P(B) - P(B \wedge A_n) + P(A_n)$$

By the inclusion-exclusion formula for two events, this is equal to $P(B \vee A_n)$, which is what we wanted to show.

5 An example

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

$$1 - P(A_1 \vee A_2 \vee A_3 \vee \ldots \vee A_n)$$

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

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

The first term is

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

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

The second term is

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

There are $\binom{n}{2} = n(n-1)/2$ terms. The event here is that both the ith and the jth letter go into the correct envelopes. The probability that we put the ith letter into the correct envelope is, as before, $\frac{1}{n}$. Given that we have put the ith letter into the correct envelope, the probability that we put the jth letter into the correct envelope is $\frac{1}{n-1}$, since there are $n-1$ letters other than the ith one, and they are all equally likely to go into the jth 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

$$\pm \sum_{1 \le j_1 < j_2 < \ldots < j_t \le n} P(A_{j_1} \wedge A_{j_2} \wedge \ldots \wedge A_{j_t})$$

There are

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

terms in this sum, and each term is the probability

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

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

$$1 - \frac{1}{2!} + \frac{1}{3!} - \dots \pm \frac{1}{n!}$$

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

$$1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \dots \mp \frac{1}{n!}.$$

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 goes to $\frac{1}{e}$.

6 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.

Suppose that we have an event A. We define a random variable, called an *indicator variable*, I_A for this event as follows:

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

We can see that

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

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

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

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

Another expression for the expectation is

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

This is straightforward to verify using the fact that

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

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

$$\mathrm{E}(f + g) = \mathrm{E}f + \mathrm{E}g$$

and if we have a constant α,

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

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

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

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

The expectation of a product of random variables, however, is not necessarily equal to the product of the expectations. For example, for an indicator random variable I_A (thus taking values only 0 or 1), we have that $E(I_A) = P(A)$ while $E(I_A \cdot I_A) = E(I_A^2) = E(I_A) = P(A)$ which is not $P(A)^2$. An important case in

which we have equality is when the two random variables f and g are independent. In this case, we have

$$\begin{aligned}
\mathrm{E}(f \cdot g) &= \sum_{x \in S} p(x) f(x) g(x) \\
&= \sum_{\alpha} \sum_{\beta} \sum_{x \in S : f(x) = \alpha, g(x) = \beta} p(x) \alpha \beta \\
&= \sum_{\alpha} \sum_{\beta} \alpha \beta P(f(x) = \alpha \wedge g(x) = \beta) \\
&= \sum_{\alpha} \sum_{\beta} \alpha \beta P(f(x) = \alpha) P(g(x) = \beta) \\
&= \left(\sum_{\alpha} \alpha P(f(x) = \alpha) \right) \left(\sum_{\beta} \beta P(g(x) = \beta) \right) \\
&= \mathrm{E}(f) \mathrm{E}(g),
\end{aligned}$$

where the 4th equality follows from the independence of f and g.