<
>

Computational Statistics 3.1: Probability

Probability

Overview

This lesson introduces the basic concepts of probability.

Objectives

After completing this lesson, students should be able to:

1. Calculate simple probabilities.
2. Use the concept of sample spaces to estimate probabilities.
3. Calculate and understand conditional probabilities.
4. Apply Bayes' Theorem to calculate unknown conditional probabilities.

Readings

Schumacker, Chapter 3.

Probability and randomness

Probability is the foundation of statistics, econometrics, and many types of machine learning. Most of what we will cover in the following modules has its foundation in probability. But the fundamental concepts are both straightforward and concrete, even if in the application the algebra can become complex. The key is just to think in terms of *sample spaces*, as we will see shortly.

Calculating probabilities

You flip a coin three times in a row. What is the chance of getting three heads? Well, there's a 1 in 2 chance of getting heads on the first flip, 1 in 2 for the second, and 1 in 2 for the third, so as you probably recall, the answer is

$$\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}=\frac{1}{8}$$

Why do we multiply them though? When two events are independent of each other – when the outcome of one event doesn't affect the other – then the probability of both events is the product of their separate probabilities.

How about if we roll a die. What's the chance of getting a 1 or a 3?

$$\frac{1}{6} + \frac{1}{6} =\frac{1}{3}$$

Again: the answer is simple, but the explanation is a bit harder.

Sample spaces

The easiest way to think about probability is via an event, state, or sample space (all different terms for the same thing). The sample space is just the set of all possible independent outcomes of your event. For a roll of a six-sided

die, the sample space is {1,2,3,4,5,6}, where the key thing is that the sum of the probabilities of each event has to sum to `1`.

We can write this as:

\[p(die=1) + p(die=2) + p(die=3) + p(die=4) + p(die=5) + p(die=6) =\] \ [\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6} = 1\]

As we will see, complex probabilities can be calculated very easily by just counting up the independent events in our sample space.

Let's take a more complicated example: We roll two dice. What's the chance of getting a total of 7? We intuitively know that the basic idea is to just count the number of different ways we could get seven (a 3 and 4; a 2 and a 5; etc.), where each outcome is like getting two heads. The proability of getting a 3 on the first die is \(\frac{1}{6}\) and the probability of getting a 4 on the second is \(\frac{1}{6}\), so the probability of getting the pair as your outcome is the product of the two independent events, or \(\frac{1}{6} \cdot \frac{1}{6} =\frac{1}{36}\). Now all we have to do is count up all the various ways there are of getting 7, and add them up. Why?

Outcomes

Let's create a table of all the possible outcomes of these two rolls – the sample space:

```
die1 = c(1,2,3,4,5,6)
die2 = c(1,2,3,4,5,6)
twodice = outer(die1,die2,"+")
twodice
```

```
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    2    3    4    5    6    7
[2,]    3    4    5    6    7    8
[3,]    4    5    6    7    8    9
[4,]    5    6    7    8    9   10
[5,]    6    7    8    9   10   11
[6,]    7    8    9   10   11   12
```

(As you can see, the `outer` function takes our two vectors and creates a matrix where each element is, in this case, the sum of the vector elements.)

This is what's known as the joint distribution, and each of these cells – the outcome of rolling the pair of dice – is equally likely. To get the total probability of getting seven, we just count up the sevens, and take that as the fraction of the total. There are 36 different outcomes, and 6 different ways of getting 7, so the total probability of getting a seven is just 6 ways out of 36, or \(\frac{6}{36} = \frac{1}{6}\).

Compound events

Once you start thinking in terms of sample spaces, it is easier to conceptualize the various laws of probability and solve probability problems. The challenge is always to just define your outcomes as a set of (usually) equally probable events, and then just add up the ones you are interested in as a fraction of the total.

For instance, what is the chance of getting exactly two heads in a row when flipping a coin three times? Well, we just lay out our sample space first – all the possible outcomes – and then count up the ones with exactly two heads in a row:

Sample space = {TTT, TTH, THT, THH, HTT, HTH, HHT, HHH}

There are 8 equally possible outcomes, and only two of them fulfill our critereon. So the answer is \(\frac{2}{8} = \frac{1}{4}\).

Similarly, if we wanted to answer a harder question about our dice – eg, what is the chance of getting a 5, 7, or 9 – we would again just add up all those outcomes and take it as a fraction of the total.

What is the chance of getting a 5, 7, or 9 when rolling a pair of dice?

(4 + 6 + 4)/36 = 14/36.

Exercise

Much of the rest of probability is just being clever about how you count up

the outcomes you're interested in (the numerator) and the total number of outcomes (the denominator).

Which of the following will use R to correctly count up the numerator (number of ways of getting 5, 7 or 9 on a roll of two dice)?

length(which(twodice==5|twodice==7|twodice==9))

Yes. The `which` gives you a vector of the elements in the matrix which fulfill the critereon; the `length` gives you the number of elements, or 14.

length(twodice[twodice==5|twodice==7|twodice==9])

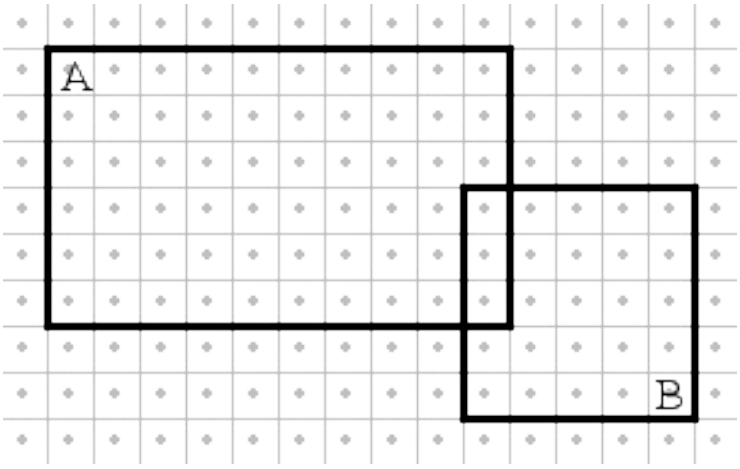
Yes. The inner part gives you the values of the elements of `twodice` which sum to 5, 7 or 9, and the `length` gives you the number of those elements, or 14.

sum(twodice==5|twodice==7|twodice==9)

Yes. If you leave off the `sum` part you will see a matrix with `TRUE` in the right place, which R treats as `1` (while treating `FALSE` as `0`) and sums to 14.

Joint probabilities

We can think a bit more abstractly about sample spaces now. Imagine instead of our 6x6 grid from the dice example, we have a grid with an arbitrary number of squares. Say we have a dartboard like this, and we are the world's worst darts player: we're sure to hit the board, but it's totally random where. In that case, the probability of the dart landing in some portion of the dartboard A is just the area of A as a proportion of the total dartboard.



So the probability of landing in A is the area of A divided by the total area, or in the figure, (6x10)/(10*16) = 60/160 = 0.375. We usually consider the total to always equal 1, so we can just more abstractly say the probability of A, denoted $P(A)$, is equal to the area of A, assuming the total area of the sample space = 1.

More generally A can be any event we're interested in, such as getting a total of 7 on rolling two dice, or getting exactly two heads in a row on flipping a coin three times.

Union / Or

Returning to our figure, what's the chance of our dart landing in A or B – ie, the chance of event A or event B happening? The total area in A is the area in its square, and the same for B; the combination or union of these two outcomes we denote $A \cup B$ (the union of A and B). But $P(A \cup B) \neq P(A) + P(B)$, since they clearly overlap, and we would be double-counting the overlapped area. So in general, we have:

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

Where $P(A \cap B)$ is the probability of landing in the intersection of A and B.

Referring to our previous figure:

What is the probabiltiy of $P(A \cap B)$?
$3/160 = 0.01875$.
What is the probabiltiy of $P(A \cup B)$?
$(6 \cdot 10)/160 + (5 \cdot 5)/160 - 3/160 = 0.375 + 0.15625 - 0.01875 = 0.5125$

Intersection / And

We can also think of $P(A \cap B)$ as the probability of $A \& B$, ie, $P(A \& B)$. If I tell you that our dart has landed somewhere inside A, but we have no idea where, then what is the probability that it has landed in $A \& B$

= A \cap B\)? We can just think of A as our new sample space, and now we just need to figure out the area of \(A \cap B\) as a fraction of A. We write this as:

\[P(A \& B | A) = \frac{P(A \& B)}{P(A)}\]

That is, the probability of getting \(A \& B\) conditional on (\(``|"\)) already getting A is the probability of \(A \& B\) divided by \(P(A)\) – that is, the fraction of A that is \(A \& B\). We could also write the first part as \(P(B|A)\), since given that we are in A, landing in B is necessarily going to be the same as landing in \(A \& B\): \(P(B|A) = P(A \& B | A)\).

So for our figure, we can calculate \(P(B|A)\): \(P(B|A) = \frac{3}{60} = 0.05\), ie, there is a five percent chance of landing in B given that we have landed in A.

Of course, we can also run it the other way, asking what the chance is of landing in A given that we know we landed somewhere in B:

\[P(A \& B | B) = P(A|B) = \frac{P(A \& B)}{P(B)} = \frac{3}{25} = 0.12\]

Bayes' Theorem

So to push this one step farther, what's the chance of landing in \(A \& B\)? The obvious answer is 3 in 160, or 0.01875. But another way to look at it is the chance of first landing in A, and then given you've landed in A, landing in \(A \& B\). So we have \(P(A \& B) = P(B|A) P(A)\), which as you can see, is just a restatement of the first equation from the previous slide.

In this case,

\[P(A \& B) = P(B|A) P(A) = \frac{3}{60} \frac{60}{160} = 0.01875\]

But again, we can run it the other way, putting \(P(A \& B)\) in terms of first landing in B, and then given that we've landed in B, landing in A:

\[P(A \& B) = P(A|B) P(B) = \frac{3}{25} \frac{25}{160} = 0.01875\]

The theorem

Thomas Bayes put these two together (though he wasn't the first):

$$P(A \& B) = P(A|B)\, P(B) = P(B|A)\, P(A)$$

This is often rearranged as:

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

Like many a theorem, this one seems rather obvious. But it turns out to be incredibly useful. In particular, we often are interested in knowing $P(A|B)$, but we only know $P(B|A)$, $P(A)$, and $P(B)$. In fact, people often mix up $P(A|B)$ and $P(B|A)$ altogether, and don't even realize they are calculating the wrong thing.

Bayes Example

For instance, here's an example. Suppose you have a cancer exam, and more specifically you're a woman in your 40s getting a mammogram. The test comes up positive. Uh oh. But because humans are optimistic creatures, you immediately think Well, no exam is infallable, maybe the test was wrong. You inquire, and the doctor tells you that the test has a sensitivity of 80%: 80% of the time, if you have cancer, the test is positive. You despair.

But it turns out you're calculating the wrong thing. What you are interested in is the probability that you have cancer given that your test was positive: P(you have cancer | test was positive). The 80% number is P(test was positive | you have cancer). So how do you figure out the thing you're really interested in, $P(C|T=+)$?

Bayes's Theorem:

$$P(C|T=+) = \frac{P(T=+|C)\, P(C)}{P(T=+)}$$

The numerator on the RHS we know: 80%. $P(C)$ the doctor can tell you, which for breast cancer for your demographic is quite low: 0.004. But what is $P(T=+)$? This can be decomposed into two easier-to-know parts:

Example continued

Since you either have cancer or you don't, we know that:

$$P(T=+) = P(T=+|C)\ P(C) + P(T=+|NC)\ P(NC)$$

That is, if your test is positive, it's either that you got a positive result and have cancer $(P(T=+|C)\ P(C))$ or you got a positive result and don't have cancer $(P(T=+|NC)\ P(NC))$. We know the first quantity, and we know $P(NC) = 1 - P(C) = 0.996$. So we just need to know the "false positive rate", the chance of getting a positive test even when you don't have cancer. This too the doctors know, and in this case it's 10%.

So now we can plug in what we do know to get what we don't know and very much want to know: the chance we have cancer:

$$P(C|T=+) = \frac{P(T=+|C)\ P(C)}{P(T=+)} = \frac{P(T=+|C)\ P(C)}{P(T=+|C)\ P(C) + P(T=+|NC)\ P(NC)}$$

or

$$P(C|T=+) = \frac{0.80 \cdot 0.004}{0.80 \cdot 0.004 + 0.10 \cdot 0.996} = 0.0311$$

That is, even given the positive test and the 80% sensitivity, you still only have a 3% chance of actually having cancer. And that's why they don't give mammograms to 40-year-olds.

The chain rule

$P(A \& B)$ is often also written as $P(A,B)$. We have seen that $P(A,B) = P(A|B)\ P(B)$ $= P(B|A)\ P(A)$. We can continue this process if there are more than two variables. For instance, $P(A,B,C) = P(A|B,C)P(B|C)P(C)$ $= P(C|A,B)P(B|A)P(A) =$ $P(A|B,C)P(C|B)P(B)$. As you can see, it doesn't matter what order we do things in: the probability of getting A, B and C is just the probability of getting C (for instance), times the probability of getting B given that you've first gotten C, then the probability of getting A

given that you've gotten B and C. But you can factor it in any order you prefer. This is known as the *chain rule* of probability.

One final thing to bear in mind: if A and B are independent (eg, two rolls of a die), then $P(A|B) = P(A)$ and $P(B|A) = P(B)$ What's the chance of getting a 3 on the first roll and a 3 on the second? Well, when the two rolls are independent, it's just $1/6 \cdot 1/6$. Ie, when A and B are independent, $P(A,B) = P(A)P(B)$.

Often with complex functions with lots of joint variables, $P(A,B,C...)$, we can approximate $P(A,B,C...)$ by assuming the variables are all independent, $P(A,B,C...) \approx P(A)P(B)P(C)...$ This can be a good approximation if the variables don't intersect very much, ie $P(X,Y) \approx 0$ for all pairs, but it will be very misleading if they do intersect substantially. Much work in probability theory involves turning complicated functions $P(A,B,C...)$ into the product of more mathmatically simple functions using the chain rule and (when it is appropriate) independence assumptions.