Pre-reading 9, Part 2

The first few questions continue on likelihoods: the probability of some event happening given that some state of affairs is true.

1. Now imagine a weird dice, that comes up with a 6 on average half the time you roll it, and the rest of the time produces 1, 2, 3, 4 or 5 with equal probability. We can specify the probability distribution of rolls of that dice - which if the following is the correct probability distribution?

Note: following the notation in the reading, I am going to write p(6 | weird-dice) to signify the probability that I roll a 6 given that I am rolling the weird dice. The probability distribution then just specifies a probability for each of the 6 possible outcomes.

The correct answer is: p(1 | weird-dice)=p(2 | weird-dice)=p(3 | weird-dice)=p(4 | weird-dice)=p(5 | weird-dice)=1/10, p(6 | weird-dice) = 1/2

In words, that says: the probability of rolling a 1 given that I am rolling the weird dice is 1/10 (one in ten; a 10% chance); the probability of rolling a 2 given that I am rolling the weird dice is 1/10; then the same for rolling a 3, 4, 5; the probability of rolling a 6 given that I am rolling the weird dice is 1 in 2 (i.e. this will happen one time in 2, or 50% of the time, or half the time).

The first option of the three was wrong because it was the probability distribution over possible outcomes for a fair dice: all outcomes have probability 1/6.

The last option of the three offered was incorrect because, among other things, the probabilities don't sum to 1, which they should: whenever I roll the weird dice one of the 6 possible outcomes will happen, so the probabilities of the individual outcomes should sum to 1.

How did I work out that the probability of rolling a 1 (or 2, or 3, or 4, or 5) was 1/10? Well, the probability that the dice rolls a 6 is 1/2, so the probability that it rolls anything other than a 6 is 1/2 too (remember, the probability that it rolls *something* must equal 1). So there is 1/2 probability mass, to be divided up equally between 5 possibilities (1, 2, 3, 4 or 5 - the five other possible rolls of the dice). 1/2 divided by 5 = 1/10.

2 **What is the probability that this weird dice will roll a 6 then a 2?**

The correct answer is 1/20: to find the probability of two independent events happening, we multiply their probabilities, so this sequence of events has probability 1/2 (the probability that we roll a 6 on the weird dice) * 1/10 (the probability that we roll a 2 on the weird dice) = 1/20.

3 What is the probability that, on two rolls, it will generate at least one 6?

The correct answer here is 3/4, basically following the same logic as question 10 in part 1. There are four possible outcomes of rolling the dice twice: we get a non-6 then a non-6; we get a 6 then a non-6; we get a non-6 then a 6; we get a 6 then a 6. For our weird dice, the probability of each of these possibilities is 1/4 (because our weird dice rolls a 6 with

probability 1/2, and a non-6 with probability 1/2, so each of these combinations involve multiplying 1/2 x 1/2, which makes 1/4). For three of these 4 possible outcomes, we get at least one 6 (the first roll is a 6, the second roll is a 6, both rolls are a 6), so we want to know the probability of having non6-6 *or* 6-non6 *or* 6-6 - since we are taking the or-rule for combining probabilities, we add these together, and get 3/4.

Another way to think about is that the only time we get no 6s on two rolls is if both rolls come up as non-6s - the probability of both rolls come up as non-6s is 1/4 (see above), so the probability of this *not* happening is 1-1/4 = 3/4.

The remaining questions in this section are about priors.

4. Imagine that tomorrow you will meet someone new. Before you meet them, you have some prior knowledge of what they will be like. In Bayesian models this is captured by the prior - in the case of people, the prior for meeting a new person is presumably based on your experiences in the past (but it needn't be - see the question about innate priors below), but it is still a prior because it captures your knowledge about this person prior to meeting them. What is the prior probability that this new person will be female?

The correct answer is: Approximately 1/2. I know that roughly half the population of Edinburgh is female, so the probability of picking someone at random from this population and them being female is 1/2.

5 **What is the probability that they will have two arms?**

Approximately 1. I do know that some people have less than 1 arm, but I have only met a few people like that in my whole life, so I am guessing it is rare. So the probability of picking a random person from the population and them having two arms is nearly 1 - you *could* get someone with less than two arms, but it's quite unlikely.

6 What is the probability that they will have red hair?
I think the correct answer is: approximately 1/10. According to wikipedia (!), 13% of the population of Scotland are red-headed, so that means if you select a random person from the population of Scotland there is a probability of 0.13 that they are red-headed. I bet redheads are rarer in Edinburgh than in Scotland as a whole (lots of non-Scots here), so I adjusted this down a little bit to 10%, or 0.1, or 1.10.

7 **We will also want to consider cases where prior probability is innate, i.e. not derived from experience, but built-in to an individual. Depending on your theoretical persuasion, you will find it easy to imagine or hard to imagine these priors! Imagine for a moment that you are a linguist who believes that the Extended Projection Principle (roughly: sentences must have subjects) is innate: children know, prior to encountering any linguistic data, that the language they are learning will obey the EPP. What is the prior probability of languages which follow the EPP for children?**

Well, if they are certain that the language will obey the EPP, then languages which obey the EPP must have prior probability of 1.

8 **What is the prior probability for languages that *don't* follow the EPP?**

Assuming all languages either do or don't obey the EPP, they have to share the probability mass of 1 between those two options. If the prior probability of a language obeying the EPP is 1, the prior probability of languages *not* obeying the EPP is 1 - 1 = 0.

Note that this means that a learner with this prior could *never* learn a language that violates the EPP: Bayes Rule says that posterior probability is proportional to prior times likelihood, so if the prior probability of non-EPP languages is 0, then the learner will never select a language of that type, because its posterior probability will be 0 (0 times anything is 0). So Bayes' Rule provides a nice neat way to think of strong innate constraints on learning: the options that are ruled out by Universal Grammar have prior probability 0.

9 Now imagine you are casting Christiansen & Devlin's connectionist model of sequence learning in Bayesian terms: languages which exhibit recursive inconsistency are harder to learn than languages that are recursively consistent. We can capture this in Bayesian terms in the prior: languages with low prior probability are 'hard' to learn, in the sense that they require more data to outweigh their low prior probability. How would you capture the Christiansen & Devlin model?

The correct answer is: The prior for recursively consistent languages is higher than the prior for recursively inconsistent languages. For a Bayesian learner, low prior probability = hard to learn - following the logic of the previous question, the extreme case of this is a language with prior probability 0, which can never be learnt. In Christiansen & Devlin's case, I don't think they are saying the prior probability of recursively inconsistent languages is 0 - in fact, such languages do exist, and I guess if given enough data, their network could learn them. But they are harder to learn than recursively consistent languages, so they must have lower prior probability.

Note that thinking about language learning therefore gives us a nice way of capturing absolute constraints (some logically-possible language types have prior probability of 0) but also weaker biases: some languages have lower prior probability than others, and can be learned given enough data but are harder to learn. What is more, the way the bias works in Bayesian models is in principle really clear - it's all in the prior probability distribution, which is something we can be really explicit about. That's a really nice feature of these models - and it's much simpler even that the weight matrix / network model we have been working with, where the bias is hiding in a slightly complicated way in the weight update rules.