CS109A Week 7 Notes

Ian Tullis

February 15, 2022

I. Beta: setting the scene

I was introduced to the beta distribution in a (frankly pretty boring) stats class long ago. All I remembered about it was that – unlike the normal distribution – it is supported only within a limited range $[0, 1]$, and so it can be a good model for things like exam scores that can't be negative.[1] But there are many other distributions with that property, and the math behind the beta is scary – it has a whole separate gamma function in it! What even is that? So I mostly forgot about it.

Watch out, Cloud! This giant snake knows probability!

Big mistake! The beta distribution is AWESOME[2], and it, along with its multinomial cousin the Dirichlet distribution, has applications all over the place when we need to quantify uncertainty about uncertainty. If you take CS238 (Decision Making Under Uncertainty), which I recommend, you will get much more practice with those.

[1] To use the beta to model scores on an exam out of 120, for instance, we can just multiply it by 120.

[2] You may recall that the last thing I said was AWESOME was linearity of expectation, and I meant it!

Suppose we are the main character in a spy movie. It's still early on in the film, so for some reason we're gambling with the supervillain in a Monte Carlo casino. We'll be fighting them later in the movie, probably on top of a train on a different continent, but for now we have to be superficially polite while trying to crush them at this game.

As is often the case with CS109A casino games, the rules are rather silly. In each round of the game, the dealer – who is in a tuxedo and wearing white gloves – flips a fancy coin made of platinum or something. If it comes up heads, we have to give $10000 to the supervillain. If it comes up tails, the supervillain has to give $10000 to us.[3]

We have been playing for a little while, and so far we have lost more rounds than we have won. We don't trust this game *or* the supervillain. They keep smirking at us, and they seem awfully smug about their chances. As they sip from a six-figure glass of 100-year-old port, they assure us that our losses are simply due to chance. And, irritatingly, they have a point: we can't completely rule out that possibility. But we think the coin might be unfair – that is, it has some probability p of coming up heads, and we think p is not 0.5. How do we express our uncertainty about this in some way other than punching? (since it is too early in the movie for that)

II. A frequentist approach

We could keep track of the number N of rounds played and the number H of heads seen, and then declare that $p = \frac{H}{N}$. This is the "maximum likelihood" estimate, as we will see in class, and surely that's the only sensible estimate, right? And if that's not precise enough, we can just play more rounds of the game to get more information, our pocketbook be damned.

A shortcoming of this method is that it only gives us a single value as an estimate. What if we want some notion of confidence in that estimate? That is, if we think $p = 0.54$, for example, are we 95% sure that the value is within 0.02 of that, or is there still a decent chance that it could be farther away, maybe even *below* 0.5? (Then we would just look like sore losers!)

Before we dive into the beta distribution, let's investigate the situation through the lens of frequentist statistics – which, broadly speaking, asks "what is the probability of the observed data, given the hypothesis?" In this case, our hypothesis is that the coin is unfair, i.e., $p \neq 0.5$. Specifically, we think it is unfair in the supervillain's favor, i.e., $p > 0.5$. But just showing that the coin is unfair (without saying in whose favor) should embarrass the supervillain (and the casino) publicly. So we declare a *null hypothesis* – basically, that the coin is fair and nothing fishy is going on – and then proceed to argue that that null

[3] The real game of baccarat is awfully close to this.

hypothesis is unlikely to be true, given what we actually saw. And if the null hypothesis is unlikely to be true, then all that remains are the unsavory alternatives...

Problem 1. Suppose that so far, we have played 10 rounds of the game, and we have seen 8 heads and 2 tails.

(a) For a fair coin ($p = 0.5$), what is the probability of seeing 8 heads in 10 flips? (Use Python / Wolfram Alpha / etc. to find the value to 3 decimal places or so.)

(b) This value is pretty small – less than 0.05. We remember from reading scientific journal articles (in our downtime between spy missions) that, by convention, results are thought to be "significant" when $P < 0.05$, i.e. when they have a less than 5% probability of occurring purely by chance. Why is it not a convincing argument to just point to your answer from (a) and say that it is less than 0.05, and so it is very unlikely to have happened by chance?

(Hint: imagine, instead, that we had flipped 1000 coins and seen 500 heads.)

(c) How could we modify the approach in (b) to argue more convincingly? (That is, we want to calculate something other than $P(X = 8)$. What other outcomes should we include?)

(d) Using this new method, what is the probability of seeing the observed data, given that the coin is fair? Does it fall under the 0.05 threshold?

(e) If we had to make a guess at the value of p, based on the results that you have seen so far, what would we say? Why?

Solutions to Problem 1.

(a) The number of heads in 10 flips has the distribution $Bin(10, 0.5)$, and $P(X = 10) = \binom{10}{8}(0.5)^8(1-0.5)^{10-8}$. This comes out to $\boxed{\approx 0.044}$.

(b) One problem is that just because an event is low-probability doesn't mean that it can't happen. The supervillain can always counter with this, no matter how we argue, and indeed, they're not wrong. But we have to draw a line somewhere, where a reasonable person would find it implausible that chance was the only explanation. (If this makes you uncomfortable because it feels subjective, well, that's statistics for you! We can try to make the subjective feel more objective, but all the math in the world can never make it completely objective.)

But there is a more subtle issue here, which is that our approach is kind of unfair to the poor supervillain. Suppose we had flipped 1000 coins and gotten 500 heads. The probability of this (for a fair coin) is around 0.025, which is less than 0.05, so we could claim that this was not due to chance. But this is an absurd claim – if we flip 1000 times, what result could be *more* like a fair coin than 500 heads? In fact, in this case, there is *no* outcome that occurs with probability ≥ 0.05. So no matter what happens, we will accuse the supervillain of cheating, even if the setup is fair!

(c) Therefore, in our 10-flip case, we should instead be finding the probability of seeing *at least* 8 heads. That is, our argument will be: even if the coin were fair, the probability of seeing at least this extreme a result is very small.

(d) The values for $P(X = 9)$ and $P(X = 10)$ turn out to be about 0.010 and 0.001. So the overall $P(X \geq 8)$ is $\boxed{\approx 0.055}$, which is over the 0.05 threshold (and therefore not "statistically significant"). So maybe we shouldn't go accusing the supervillain just yet!

(e) It seems most sensible to conclude that $p = \frac{8}{10}$, and we will make this more rigorous in a future CS109 lecture on maximum likelihood. But we might have a bad feeling about this. Do we really feel confident concluding, on the basis of a small amount of data, that p is so large? As an extreme case, what if we had seen three heads in three flips? Does it really make sense to conclude that $p = 1$?

III. A Bayesian approach

Our frequentist methods above gave us an estimate of the coin's probability p of coming up heads (0.8), and a way to argue that $p \neq 0.5$. But what if we want to explicitly quantify our beliefs about, say, how much more likely p is to be 0.8 than 0.79? What if we want to see a distribution of these beliefs, to get a sense of whether they are tightly centered around 0.8 (in which case we can be more confident) or more diffuse (in which case we probably need more information)? We can do this with frequentist methods as well, but this is where Bayesian methods really shine.

As we've already seen in CS109, Bayesian methods involve bringing in a prior set of beliefs. We might believe that the person who just walked into our ice cream shop has an 0.7 probability of buying some ice cream, just based on overall trends about customers (maybe yesterday we saw 70 out of 100 customers make a purchase). But then we see this particular customer look long and hard at the mint chocolate chip, and maybe our posterior probability goes up to 0.95. Notice that Bayesian methods ask "what is the probability of the hypothesis, given the observed data?", whereas frequentist methods ask the opposite.

There have been acrimonious disputes between frequentists and Bayesians in the literature and elsewhere. To oversimplify a couple of the points of contention:

- Frequentists think that explicitly bringing your own beliefs into an analysis makes everything hopelessly subjective.
- Bayesians argue that what we really want to know is the probability of the hypothesis given the observed data, and doing it the other way around is artificial and misguided.

My own position is that – as much as I usually dislike lazy both-sides-ism – both approaches really do have their merits. There is no canonically correct way to do statistics, because uncertainty can only be described, not eliminated. Probably the contemporary ethos is to use whatever drives progress forward and brings in the most sweet, sweet cash from venture capitalist investors. Of course, it is dangerous to uncritically just try everything and see what works – this makes it more likely that some approach will only *appear*, by chance, to work well! But I think it's best to understand both frequentist and Bayesian methods well enough to be able to use them in situations that seem appropriate.

So, back to the beta distribution and our spy movie example. The beta distribution itself is not *inherently* Bayesian, but the way we use it in CS109 is. Specifically, we have some set of beliefs about the value of p, and then each time we make an observation (the outcome of one round), we update those beliefs.

What set of beliefs should we have even before the game begins? In this case we might reasonably come in *expecting* the supervillain to be a cheater. Or, we might be unusually magnanimous (for a secret agent), and come in very confident that the coin is fair. A third approach (which is maybe less objectionable to frequentists) is to come in giving every possible value of p equal likelihood, i.e. "flat priors", and this is what we will often do when working with betas.

Let's derive the beta distribution, like Chris did in class...

Problem 2. Say we came into the game believing that every possible value of p was equally likely. Then we saw 8 heads and 2 tails.

(a) First of all, we need to turn "every possible value of p is equally likely" into a PDF – call it $f(x)$ or $f(p = x)$. That is, if we want the probability density of our belief that $p = 0.9$, we evaluate $f(0.9)$. Here we use x to avoid confusion with p, which is a single fixed (but unknown) value representing how unfair the coin really is.

This PDF should just look like a horizontal line floating somewhere above the x-axis; it should have the same value c everywhere. It should also only be supported on the range [0, 1], since p is a probability and can only take on values in that range. What is $f(x)$?

(b) Using Bayes' Rule, we can express $f(p = x|8 \text{ heads out of } 10)$ as

$$\frac{P(8 \text{ heads out of } 10|p = x)f(p = x)}{P(8 \text{ heads out of } 10)}$$

One of those terms is your distribution from part (a) of this problem. Another is very similar to what you found in part (a) of problem 1. Replace both of those terms with expressions in terms of x and/or constants.

(c) We need to use a version of the Law of Total Probability in the denominator, but here we can't write out a finite sequence of terms that look like the numerator. What integral should we use instead? (Use your numerator from this problem, but introduce a new letter like y as the integration variable, to avoid confusion with the x in the numerator).

(d) Evaluate this integral to get a constant, then plug it into our expression from (b) to get our final Bayesian posterior distribution.

(e) Look at the Wolfram Alpha result for `beta distribution with alpha = 9, beta = 3`. You should see that it gives the same PDF. Visually find the value of x for which $f(x)$ is maximized; is it what you expect?

(f) What is the CDF of the beta distribution that you just found? (Feel free to use Wolfram Alpha to do the integral.) Using this CDF, what is the probability that the coin is quite unfair – i.e. has $p \geq 0.55$, for example?

(g) Why could we not have answered part (f) directly using frequentist methods? (Put differently, what did we include in our Bayesian method that let us answer that question?)

(h) A uniform distribution is $Beta(1, 1)$. How different would our posterior beta distribution look if we had instead started with Laplace smoothing, i.e., $Beta(2, 2)$? (Don't go through all the steps of the problem again – just consider what the final beta distribution's parameters α and β would be in this case. Then use Wolfram Alpha to plot that.)

Solutions to Problem 2.

(a) This uniform distribution is a PDF, so it has to have $\int_0^1 c dx = 1$. Integrating $\int_0^1 c dx$, we get $[cx]_0^1 = c(1) - c(0) = c$. So $c = 1$, and therefore $\boxed{f(x) = 1}$.

(b) For $P(8 \text{ heads out of } 10)|p = x)$, we again use the binomial distribution, but now with a success probability of x. So this evaluates to $\binom{10}{8}x^8(1-x)^2$, which is $45x^8(1 - 2x + x^2) = 45x^8 - 90x^9 + 45x^{10}$. (This form will be easier to integrate later on.)

Plugging in that expression and our prior from part (a) (which is just 1, so it goes away), we now have

$$f(p = x|8 \text{ heads out of } 10) = \frac{45x^8 - 90x^9 + 45x^{10}}{P(8 \text{ heads out of } 10)}$$

(c) The integral is

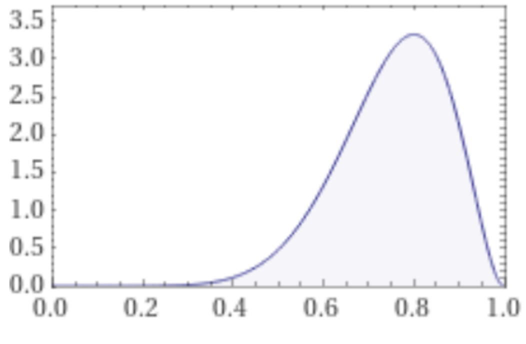
$$\int_0^1 (45y^8 - 90y^9 + 45y^{10})dy$$

(d) Evaluating this, we get

$$[5y^9 - 9y^{10} + \frac{45}{11}y^{11}]_0^1 = 5 - 9 + \frac{45}{11} = \boxed{\frac{1}{11}}$$

$$f(p = x|8 \text{ heads out of } 10) = \frac{45x^8 - 90x^9 + 45x^{10}}{\frac{1}{11}} = \boxed{495x^8 - 990x^9 + 495x^{10}}$$

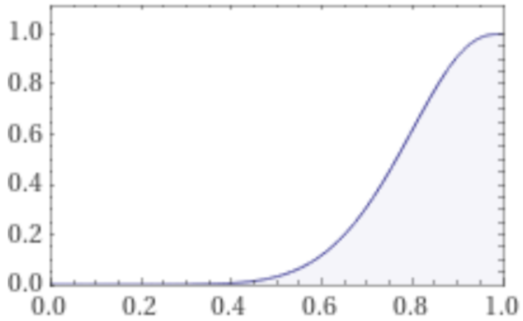
(e) The PDF looks like this:



and the maximum is at 0.8, which is what we would expect. (It is exactly 0.8, as you could check directly via calculus.)

(f) As usual, the CDF is the integral of the PDF from the lower end of the support range to some stopping point y.

$$\int_0^y (495x^8 - 990x^9 + 495x^{10})dx = [55x^9 - 99x^{10} + 45x^{11}]_0^y = \boxed{55y^9 - 99y^{10} + 45y^{11}}$$

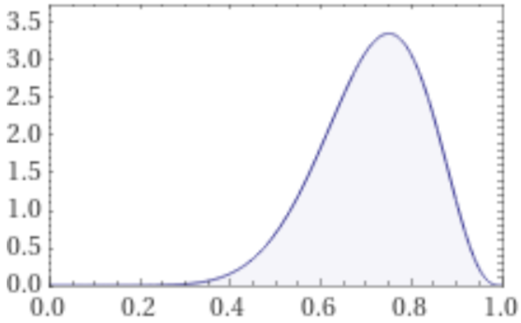


To find the area of the part of the PDF that is *below* 0.55, we evaluate the CDF at 0.55 to get $55(0.55)^9 - 99(0.55)^{10} + 45(0.55)^{11} \approx 0.065$. Then the area that is above 0.55 is $\approx 1 - 0.065 \boxed{\approx 0.935}$. That is, we have a pretty strong belief that the true probability p of the coin coming up heads is 0.55 or higher.

(g) The critical piece of information that the Bayesian method brought in was the assumption that, in the absence of other information, all values of p were equally likely. Without such an assumption, it doesn't even make sense to talk about the *probability* of p taking on any particular value. It depends on how evil the supervillain was feeling this morning, what fake coins they own, and so on. This seems hopelessly complicated to quantify. So maybe it's not so bad to make a very basic assumption (flat priors) and then iterate from there?

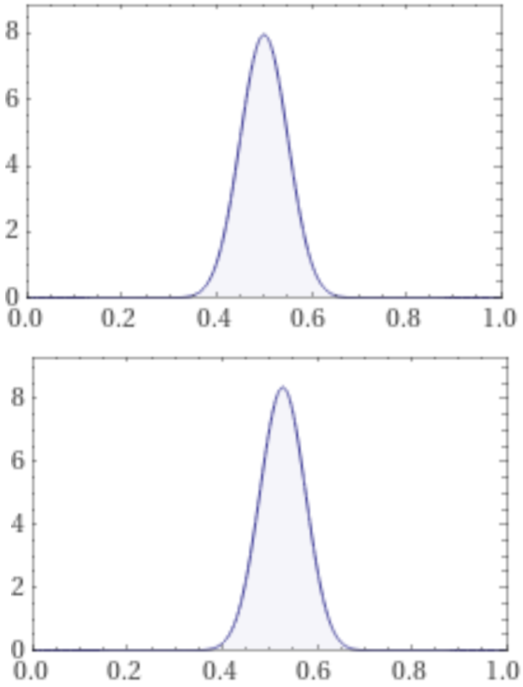
However, we shouldn't lose sight of the fact that our probability estimate includes an assumption. That is, we have not objectively determined that the true probability of $p \geq 0.55$ is 0.935. If we had used a non-flat prior distribution, we would have gotten very different results...

(h) Starting with $Beta(2, 2)$ and adding 8 "successes" to α and 2 "failures" to β, in this case we would end up with $Beta(10, 4)$:



This is fairly close to our original $Beta(9,3)$, but now our best estimate of p is slightly below 0.8 – it has been dragged down a bit because the Laplace smoothing essentially adds one "bonus" head and one "bonus" tail into the mix.

On the other hand, suppose that we come in naively believing that the coin is very fair, i.e., something like $Beta(50,50)$. Then after our 10 flips, we believe $Beta(58,52)$. Our prior beliefs were so strong that the new data barely changes them. So – as a frequentist would point out – the choice of prior distribution can *dramatically* change the results!



IV. Bayesian Networks

We have been cooped up inside all day coding, and we are about to leave Stanford to head into San Francisco on 101. We check the traffic on Google Maps. Oh no, 101 is bright red! Why is the traffic so much worse than usual? We come up with a couple of theories:

- There might be a Warriors game tonight.
- It might be raining, and as we all know, in general, coastal Californians can't drive in any kind of weather.[4]

We also know that on days when there is a Warriors game, people are more likely to wear Warriors jerseys.

Of course, traffic is sometimes bad for other reasons, and it might just happen that more people wear Warriors jerseys on some days. So we can't rule out the effects of chance.

Problem 3. Let:

- W be the event that there is a Warriors game,
- R be the event that it is raining,
- T be the event that there is unusually bad traffic, and
- J be the event that more people than usual are wearing Warriors jerseys.

(a) Translate the information given on this page into a Bayesian network with circles for W, R, T, and J, and arrows between them as appropriate. (Assume, rather unrealistically, that there are no other specific factors involved besides chance. Note that this model assumes that R and W are independent, which makes sense for an indoor sport!)

(b) In each of the following situations, assume that the person operates under the belief system above, but unless otherwise stated, they start off knowing nothing – i.e., they do not know whether there is a Warriors game tonight, whether it is raining, whether there is unusually bad traffic, or whether more people than usual are wearing Warriors jerseys. or whether there is a Warriors game tonight. For this part, try to think intuitively and not in terms of specific probabilities. Remember that events T and J could each happen, with at least *some* probability, by chance; that is, it is possible that more people than usual are wearing Warriors jerseys, even though there is no Warriors game.

(i) Klay already knows that traffic is unusually bad. He checks the weather and sees that it is not raining. Does this new information change his belief about whether there is a Warriors game tonight?

[4] I include myself in this statement!

(ii) Draymond sees that more people than usual are wearing Warriors jerseys. Does this change his belief about whether traffic is unusually bad? Does this change his belief about whether it is raining?

(iii) Steph already knows that there is no Warriors game tonight. Then he sees that more people than usual are wearing Warriors jerseys. Does this change his belief about whether traffic is unusually bad?

(iv) Ayesha already knows there is a Warriors game tonight. Then she sees that traffic is unusually bad. Does this change her belief about whether it is raining?

(c) (Under construction – please skip this one for now.) Now suppose that the actual underlying model is:

- $P(R) = 0.2$. (Hey, we can dream.)
- $P(W) = 0.1$.
- $P(T|R \cap W) = 0.9$; $P(T|R^c \cap W) = 0.8$; $P(T|R \cap W^c) = 0.7$; $P(T|R^c \cap W^c) = 0.1$.
- $P(J|W) = 0.6$; $P(J|W^c) = 0.1$.

(i) What are $P(R \cap W), P(R^c \cap W), P(R \cap W^c)$, and $P(R^c \cap W^c)$?

(ii) What are $P(T|W)$, $P(T)$, and $P(W|T)$?

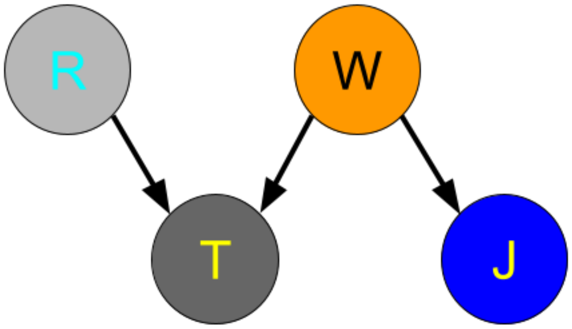
(iii) What is $P(W|T, R)$? Comparing this to $P(W|T)$, does this fit your answer to (b)(i)?

(iv) Time permitting, try to check some of the other statements as well. It may be easiest to make a table with the probabilities of all 16 of the possible scenarios, although the whole point of these Bayesian networks is to avoid having to do this explicitly. Make sure you know how to do this yourself, but the results are provided here.

W	R	T	J	Prob.
0	0	0	0	0.5832
0	0	0	1	0.0648
0	0	1	0	0.0648
0	0	1	1	0.0072
0	1	0	0	0.0486
0	1	0	1	0.0054
0	1	1	0	0.1124
0	1	1	1	0.0136
1	0	0	0	0.0064
1	0	0	1	0.0096
1	0	1	0	0.0256
1	0	1	1	0.0384
1	1	0	0	0.0008
1	1	0	1	0.0012
1	1	1	0	0.0072
1	1	1	1	0.0108

Solutions to Problem 3.

(a) The Bayesian network looks like this:



(b) (a) $\boxed{\text{Yes.}}$ If rain and the Warriors game are the two possible explanations (other than chance) for the bad traffic, and rain is ruled out, then it is much more likely that a Warriors game is to blame.

(b) $\boxed{\text{Yes; No.}}$ Given that more people than usual are wearing jerseys, Draymond's belief that there is a Warriors game is strengthened. Because this is known to result in bad traffic, his belief that there is bad traffic is also strengthened.

However, intuitively, this all has nothing to do with whether it is raining. Even though rain might make already bad traffic worse, Draymond knows that whether or not it rains is independent of whether or not there is a Warriors game, so feeling more confident that there is a Warriors game shouldn't tell him anything about rain.

We might fool ourselves with an argument like "well, we think traffic is bad, and bad traffic is associated with rain", but our new belief about the traffic is already fully explained by our new belief about the Warriors game. We really haven't learned anything about the rain.

(c) $\boxed{\text{No.}}$ First of all, since Steph knows that there is no Warriors game, and that is the only factor (besides chance) that influences whether more people than usual are wearing Warriors jerseys, he can safely conclude that the latter is due to chance.

Now, usually, the only value in knowing whether more people are wearing Warriors jerseys is that it makes Steph more likely to believe there is a game, but given that he already knows there isn't one, he learns nothing.

(d) $\boxed{\text{Yes.}}$ This one is tricky! At first it might seem that Ayesha's knowledge of the Warriors game fully explains the bad traffic. But that would only be true if a Warriors game *surely* resulted in bad traffic. Otherwise, there is some chance that the bad traffic is *not* due to the Warriors game, and in that case, as usual, bad traffic increases our suspicion that rain may be involved.

There is a set of rules, involving something called "d-separation", to make this type of analysis easier. This is not in scope for CS109, but you might enjoy learning more: `https://bayes.cs.ucla.edu/BOOK-2K/d-sep.html`. I personally always have trouble remembering these rules, and I prefer to think through example situations.