

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at [ocw.mit.edu](http://ocw.mit.edu).

**PROFESSOR:** All right, I guess we can get started here. So welcome. Today we're going to do Bayesian updating, which is to say we're going to be using Bayes' Theorem to learn from data. We're going to update our beliefs about various hypotheses based on the data. For this week, we're going to have known priors. We're going to know our belief about the data ahead of time for certain. We'll draw coins out of drawers or pull things out of hats, things like that.

Starting next week, and we discussed this a little last time, we'll have unknown priors, which is you'll have to be making up those priors. And we'll teach you techniques for making it up. So the first slide is just XKCD's view of Bayes' Theorem. Let's go on. We want to start with a clicker question here.

So the question is which treatment would you choose if you needed a treatment? Treatment 1 cured 100% of the patients. Treatment 2 cured 95% in a trial. And treatment 3 cured 90%. So with this information which would you choose?

**AUDIENCE:** With only this information?

**PROFESSOR:** With only this. If this is all I told you. Which would you choose? I recognizing that you might not be too happy with just this information, but if this is their best information, what would you choose? All right, there's some hold outs for the 95% cure. I'm not sure why. Really, that 95% was just, you know, they've cured one person 19 out of 20. They got them 19 out of 20 well, or however you say that.

Yeah. All right, so there's even a few more holdouts. All right, what if I gave you this information? Treatment 1 cured 3 out of 3 patients. That's 100%. Treatment 2 cured 19 of 20 patients. That's 95%. Or, you have a standard treatment which has cured 90,000 out of 100,000 patients, 90%, in clinical practice. Now which one would you

choose?

**PROFESSOR 2:** It's very interesting. Yeah, it totally changed.

**PROFESSOR:** What's that?

**PROFESSOR 2:** It totally shifted.

**PROFESSOR:** It totally shifted, really?

**PROFESSOR 2:** We have a bunch of very conservative people.

**PROFESSOR:** Right. So the majority would choose the third choice, 90 out of 100,000. I think-- well someone tell me, what's the intuition there?

**AUDIENCE:** Because it's been more tested.

**PROFESSOR:** It's been more tested. 3 out of 3 is 100%, but do you really believe 100%? Not yet. Maybe it'll prove out in time. 19 out of 20 is a little more, but for most people, for 79% of you, that's not enough. The 90 out of 100,000 seems like good odds, and it's well tested. 19 out of 20, if the next person tested doesn't get better, now you back to 90%. So that doesn't seem like quite enough. What if were 95 out of 100 patients. We'll let you click in there. Change number 2 to 95 out of 100 and let's see what people would do. Oh, you're very obliging, 95 out of 100--

**PROFESSOR 2:** Now it's about 50-50.

**PROFESSOR:** About 50-50.

**PROFESSOR 2:** Between 2 and 3.

**PROFESSOR:** Realistically, of course, what would you want to do? You'd want to do a little more research and find out what these trials were, how good experiments you thought they were, what people are thinking about it. But with just this data, that's what people would choose. And would people would choose differently. OK.

**PROFESSOR 2:** Let's see. So now I'm going to give you a toy problem that actually mirrors the same

kind of effect we saw on the last slide. So, suppose in this MIT mug, I have dice of two types, 4-sided and 20-sided. All right? So I'm going to reach into the mug. I'm not going to look. Much. I'm going to randomly pull out a die. I'm going to roll it. All right, I got a 1.

OK. I got a 1. Now, just based on this information, which type of die do you think I randomly chose? The 4-sided. Someone want to give an explanation of why they think the 4-sided is more likely?

**AUDIENCE:** It's greater probability.

**PROFESSOR 2:** It's greater probability. What's the probability of getting a 1 on the 4-sided die? And what about a 20-sided die? OK. So we see that it would be more likely to roll a 1 with a 4-sided die than with a 20-sided die.

So, suppose I tell you that I really in this cup only have one die of each type. So does that change your analysis of which die is most likely? Don't hit forward yet.

**PROFESSOR:** I would think of it. My hands are staying.

**PROFESSOR 2:** Or do you want to stick with the same reasoning? All right. Let me show you what's actually in this cup.

**PROFESSOR:** Ouch.

**PROFESSOR 2:** Oh, you can't see it yet. Let's go with the document camera. OK, that was actually what I had in this cup. Now what? So does this change at all of your analysis which die you think I rolled and got a 1 with? What type of die, rather?

**AUDIENCE:** Yes.

**PROFESSOR 2:** Yeah? Now what do you think is more likely? 4-sided or 20-sided?

**AUDIENCE:** How many 20-sided are there?

**PROFESSOR 2:** I think I got 20 20-sided die there. All right, so raise your hand if you like a 4-sided die. Raise your hand if you like the 20-sided die. Great. So that reasoning that

you're doing intuitively is something that we want to formalize this week and next week using Bayes' Theorem to do these calculations in a systematic way, to evaluate hypotheses based on data. And these numbers, for example, how many die I had of one type versus another, we can encode that information in what's called a prior, and we'll come to that today. And you should've seen in your reading this idea.

**PROFESSOR:** OK so I'm going to work through an example to start with. We have 3 types of coins: type A, B, and C. However it happens, the probability of getting heads with type A is 0.5. That's a fair coin. The probability of getting heads with type B is 0.6, and of type C is 0.9. And here's a box that contains two of the fair coins and one of each of the other. So two of type A, one of type B, and one of type C.

Suppose I pick one at random and I get heads. So this is like John picking a die at random, and getting a 1. I get heads. Here's the question: before I flipped it, I have a prior belief in the probability that the coin was of each type based on the number in my drawer. After flipping it, my believe changes. The probability will change based on the data I get. And so a question you could ask: what was learned by flipping the coin?

What we want to teach you to do to answer these questions, we have a nice way of summarizing all of that once in a table. So here the notation, we have this nice scripty H-- that's no good. H is our hypothesis. That is, you're of type A or B or C. We're going to make a conjecture on what we think the coin was.

Before we do anything, we have a prior. So what's the probability of-- how did I label this-- that the coin was of type A? Where all you know is I pulled it out of the drawer. So what's the probability it's of type A.

**PROFESSOR 2:** Let me go back and remind you what the drawer looks like. What is it?

**PROFESSOR:** 0.5. 0.5, or I'll write it like this. 2 out of 4, which is 0.5. And what's the probability-- I'll shorten this-- for just type B. Yeah, 1/4 or 0.25. The probability that we're of type C is 1/4 or 0.25. This was one of our simpler counting problems. In the table, you can

see that's encoded in the line in the column marked prior. Before we take any data, this is what we'd say. This is the sort of odds we'd give if we were going to bet on this.

But now we have some data. So once we have data we can get the likelihood. The likelihood is the probability of seeing the data given a hypothesis. I won't write this out, I'll just point at the slide. John has a-- you were going to make your mouse bigger.

**PROFESSOR 2:** I was.

**PROFESSOR:** OK. So the likelihood column, we use this term when we are talking about maximum likelihood, is the probability of the data given the hypothesis. And this is typically easy, or often easy, to figure out. What's the probability of seeing heads? Our data was I flipped a heads if my coin is a 0.5 probability coin.

So the probability of seeing heads given a type A coin is what? Somebody. It's 0.5. This is particularly easy because that's what we told you. The probability of heads with a type A coin is 0.5. What about for B and C. 0.6 and 0.9. So there right here is the likelihood column in the table.

Now, how do we get the posterior? Bayes' Theorem says the probability of H given D, which is what I want. I want to know, given this data, what do I believe the coin to be? What's its probability? Is the probability of D given H times the probability of H all over the probability of D.

Well notice this numerator right here. That's just our likelihood times our hypothesis, our prior belief in our hypothesis. And so what do we get when I multiply the prior column by the likelihood column? I get this column in red. And I'll call that the unnormalized posterior. Why unnormalized because it doesn't sum to 1. How do I get it to sum to 1? There I have to divide by the probability of the data. And how do I get the probability of the data?

That's called the-- I see someone mouthing it. The Law of Total Probability. And the Law of Total Probability says you multiply for each hypothesis this value times this

value and add them up. Which is in fact just the sum of the red column. So there I've written it 0.625.

And now how do I normalize? I just divide by all my values in my unnormalized column by that normalizing factor, 0.625, and I get these, what I call the posterior.

Let's say-- well now let's analyze this. So now what's your belief? Which coin is the most likely coin? After seeing this data. Still A. But what's happened to the probability from our prior probability? It's dropped. Which one has gained probability.

**AUDIENCE:** C.

**PROFESSOR:** C. Does that make sense? If I flip a heads, which coin is the most likely to flip heads? C. So I've shifted it. Suppose I flipped heads three more times. What would you see in the probability as I kept updating?

**PROFESSOR 2:** Say it again.

**AUDIENCE:** The probability of C gets bigger.

**PROFESSOR 2:** The probability of C would get bigger at the expense of the probabilities of the others. If I flipped heads 9 times in a row, I'd be very convinced that I had a type C coin. One last thing to note about this table. I put totals down here. What does this total of 1 represent in the first column? Emily.

**AUDIENCE:** The total probabilities of anything, so the total space.

**PROFESSOR:** Yeah. When I sum up the total probabilities, they sum up to 1. Same with this last. That's a probability. This I could sum, because they were probabilities. Why didn't I sum this column? Those are the likelihoods, and what's changing as I move down this column? What's changing is H. That's not a probability function. It doesn't sum to 1. There's no reason that the likelihoods have to sum to 1, and here you can see they clearly don't sum to 1.

So I don't sum that. That's an important, maybe slightly subtle point, that the

likelihood function is not a probability function. OK, any questions here?

**PROFESSOR 2:** All right. So, let's go back to this document camera. I got my trusty platonic solids. So, we're going to go through a series of board questions now, three board questions which are going to allow you to practice making these tables and getting a sense for how to do Bayesian Updating, how to iterate it, just how it works.

And our toy problem is going to be involving these 5 platonic dice. You all have them, but let's not break them out today, because it would be very noisy with the cameras I think. We'll use our 4-sided-- here, I got big ones. We've got our 4-sided, 8-sided, 6-sided. No 6-sided. No, I don't have a 6-sided. I have the wrong dice. Hold on. We got 20, we got 4, we got 12, we got 8, ah-- 10, what are you doing there? That's not platonic.

OK, and 6. So now I'm actually going to take these. You can watch this. All right, this is empty cup. OK, they're in there right? OK. How do we do this? How do we convince them?

**PROFESSOR:** Should we choose one at random?

**PROFESSOR 2:** Yeah. All right, choose it. Don't let anyone see. Hand it me. OK. So I have this die. Don't look at the others. I'm going to start rolling it, all right? In this secret bin. Jerry, you can ver-- we can have a student come verify that I'm not making this up.

You ready? OK, here we go. OK, what did I get? 8. All right. So, which die is it so far? What do you think? Could it be 4, the 4-sided die? 6-sided? 8? 12? 20? What's most likely? All right.

Come take a look at another one. Ready? Ah, got a 1.

**AUDIENCE:** It's the same die?

**PROFESSOR 2:** It's the same die, it's the same die. All right, so now what? How does this change, anyone want to tell me how would this now change from before what you believe to be? Does it change what you believe at all, in any way?

You still think it's the 8-sided. Does it make the 8-sided less like, more likely relative to the others? More likely. Wonderful. See you guys all sense this in your bones. It's great. This is going to be so easy. For all of you. All right. We've got to keep going until we figure out with this die.

You, you love to participate. Come on. Ready?

**AUDIENCE:** 3.

**PROFESSOR 2:** 3, oh man. All right. Ready?

**AUDIENCE:** 12.

**PROFESSOR 2:** Ooh. Now what happened? It's a 12-sided die? But it could be a 20-sided die, right? All right, let me just do a few more, I'll tell you it I get. I get 1, I get 10, I get 10, I get 12, I get 3, I get 10. All right. Who wants to bet? What would you be willing to bet?

Suppose I was willing to give you like, I don't know, 10 to 1 odds that it's a 20-sided die. I mean, of course, maybe I know something. I can see the die. Right, so what we want to be able to do is precisely answer that question. All right, and it's just through this Bayesian Updating. It's not so hard, but organizing your work is really important.

So, all right. It was the 12-sided after all. So you all did good. All right, so this is the first board question. All right. I want to practice making the table Jerry showed you. Make one for if it were 13. Make one if I had rolled 5 instead. Or do the same question if I had rolled a 9. In each case, compute the posterior probabilities of each of the 5 types of dice.

**PROFESSOR:** You knew it was a 20-sided die?

**AUDIENCE:** Yeah.

**PROFESSOR:** All right, what about if it's a 5?

**AUDIENCE:** Yeah.

**PROFESSOR:** But that's it. But you see how that captures it very nicely in a table. All of your intuition. You knew it was a--

**PROFESSOR 2:** So right now you're doing the D equals 13 case?

**AUDIENCE:** Yeah, or I guess we could combine the tables.

**PROFESSOR 2:** Oh, I'm sorry. You're not combining tables. This is your likelihood column here?

**AUDIENCE:** Yeah.

**PROFESSOR 2:** And then this is your unnormalized posterior column. Is that right? Or, no this is actually normalized.

**AUDIENCE:** This is for if it were like any number. Like if we didn't know what number it was.

**PROFESSOR 2:** The likelihood?

**AUDIENCE:** Yes.

**PROFESSOR 2:** Absolutely. So the likelihood doesn't depend on-- wait. What are we doing? Rolling a 5 or a 13?

**AUDIENCE:** Rolling a 13, right?

**PROFESSOR 2:** Right, so rolling a 13, for the likelihood you're computing the probability of data that you saw, the 13, given the hypothesis say that it's the 4-sided die.

**AUDIENCE:** Right.

**PROFESSOR 2:** It's the 4-sided die, what's the probability you would roll a 13?

**AUDIENCE:** 0.

**PROFESSOR 2:** Great. Same with this, this and this. But what about here? If it's the 20-sided die, if that's your hypothesis, what's the probability you'd roll a 13.

**AUDIENCE:** Oh no. Oh  $1/20$ .

**PROFESSOR 2:** Exactly. So that's the likelihood column. So this we're going to erase. OK.

**AUDIENCE 2:** Then we multiply this by this and get that.

**PROFESSOR 2:** Exactly.

**AUDIENCE:** [INAUDIBLE] divide by [INAUDIBLE].

**PROFESSOR 2:** Exactly, exactly. Does that make sense?

**AUDIENCE:** Wait, we multiply these together--

**AUDIENCE 2:** Multiply those three together. And then we divide 1 over 100, so it's going to be 0--

**PROFESSOR 2:** So, yeah so you get to just--

**AUDIENCE:** And then HD.

**PROFESSOR 2:** No, no. Oh yeah, you're right. HD. One thing I want to point out is instead of using a comma, all right, this is conditional probability so it's probability of D given H, like this. OK. There you go. So that's absolutely right. At the end, we want to know how likely each die was given the data.

**AUDIENCE:** Because this total is just P.

**PROFESSOR 2:** That's right. Which is 1 over 100. Great. So now, do it again. Yeah, exactly. You don't even need to make a new table.

**PROFESSOR 2:** Awesome

**AUDIENCE:** So this column [INAUDIBLE] to the probability, because [INAUDIBLE].

**PROFESSOR 2:** Yeah, you take each hypothesis times probability of data given that process, which is the same as the probability of D and H.

**AUDIENCE:** Yeah, exactly.

**PROFESSOR 2:** So your partitioning it up, sort of based on hypotheses.

**AUDIENCE:** Because this equals that times probability of D.

**PROFESSOR 2:** That's right.

**AUDIENCE:** D confuses me. What is P of D?

**PROFESSOR 2:** Probability of the data. So, right so a priori, you don't know which die it is, right? But there's still a probability of getting a 13. Right? And what is the probability? Well first, we would have to, in this case, we could only pick the 20-sided die, that has a  $1/5$  chance of happening. And then, given that we picked the 20-sided die, we have to roll a 13, which is a  $1/20$  chance. Which is exactly why this sums to 100. And then more generally, you have the Law of Total Probability.

So here, I think you should keep in mind, which is the D and which is H. Maybe write it above, otherwise it's a little confusing to know what-- great. OK, and so these are the final probabilities you got if you rolled a 9. Is that right? So there's something off here right?

**AUDIENCE:** Sorry, these are our priors. I'm not labeling this. These are priors. This is gonna be [INAUDIBLE].

**PROFESSOR 2:** So actually here, I think you have likelihood. This is likelihood here, exactly. That's likelihood. Your prior should be what?

**AUDIENCE:** 0.2.

**PROFESSOR 2:** What should it be?

**AUDIENCE:** 0.2 for each one.

**PROFESSOR 2:** Yeah,  $1/5$ . Oh, I didn't see the points. Yeah, so that's your prior,  $1/5$  for each. This is your likelihood. Then you multiply them, and that's the unnormalized posterior, right?

Let's have a brief discussion about it, and then we're going to give you something a little more challenging.

So for rolling a 13, right, the first thing you do is ask, what is the prior? And you saw, I had 5 dice. I shook it around. A student reached in randomly, not even me. So we can trust that it was  $1/5$  each die. Equally likely, that's a uniform prior, right, discrete probability mass function. Great.

Likelihood. Well, that's going to depend on data, because it's the probability of the data given the hypotheses. The hypothesis in this case is how many sides on the die. So, if the data's 13, what's the probability that I would roll that data, a 13, given that it was the 4-sided die. 0. Right, it's impossible. Similarly 4, 6, 8 and 12 sides. But given that I had the 20-sided die, there would be a  $1/20$  probability of rolling a 13. That's the likelihood column. All right.

The next column, what we called the unnormalized posterior, we multiply, everything is 0, and then we get 1 out of 100 as the product for the 20-sided die. If we sum this up by the Law of Total Probability, we get the probability of the data. Now, a student asked, what does that mean? We don't know which die it was?

In this case, what it means is that it's reasonable to ask the question, well, if someone reaches in, grabs 1 of these 5 dice randomly and rolls it and gets-- I mean, what's the probability that the result would be a 13. Right? That's a reasonable question. And in this case, it's not so hard to analyze that directly, because the only way that could happen is if they think the 20-sided die and then rolled a 13. Well, given that they picked the 20-sided die.

Well, there's a  $1/5$  chance they picked the 20-sided die, and then a  $1/20$  chance they roll of 13 given that they picked the 20-sided die. So the probability of rolling a 13 overall is 1 in 100. The probability of the data.

**PROFESSOR:** John, let me just say one thing. The way we would have done this before is you would have made a tree. The top branch of the tree would have gone to the 5 types of die, and then each of those die would have gone to all the possibilities. And the

only branch that leads to a 13 is down to the 20-sided die and then down to the 13.

**PROFESSOR 2:** Notice also that this would be a very big tree, right? You have 5 branches, and then this would have 4 and 6, and 8, and 12, and 20 coming off of it. Finally we normalize. Using Bayes Rule, this gives us a posterior probability mass function. It's for everything but the 20-sided hypothesis. Therefore, we get 100% belief, 100% probability given that we rolled a 13, that it was a 20-sided die.

Now of course you all knew that immediately. If you roll a 13, it has to be the 20-sided die. So let's look at a more interesting one that you did, the 5. So what changes? That's the big question. The only thing that changes is the likelihood. I mean, granted the rest changes. But everything these two columns are computed from the first two. The prior stays the same, but the likelihood function changes because we have different data.

In this case, if we roll a 5, we can't get that on the 4-sided die. We have a  $1/6$  chance on the 6-sided,  $1/8$  on the 8-sided,  $1/12$  on the 12-sided and  $1/20$  on the 20-sided. Multiply those columns, get the unnormalize probability, or the unnormalized posterior. That sums to the probability of getting a 5. When you divide by it normalizes the posterior, and what do we see? Well, of course you couldn't have used the 4-sided die, and of the remaining the most likely, the most probable, given the data, is the 6-sided die. It has a probability, that hypothesis has a probability of about 40%, compared to the 20-sided die, which is only about 12%.

OK. Last one. So this is the same deal. Again, the likelihood function changes. The prior's the same. We go through the table. The only two possibilities are the 12 and 20-sided, and the 12-sided, given that you rolled a 9, is about twice as likely as the 20-sided, which should fit with your intuition.

OK. So next, we want you to not erase, but go back to the boards you've already made, and we're going to explore a little bit the idea of repeated trials. Right? Often if you're collecting data, you don't just collect one data point, you collect a series of data points, maybe a series of patients in a clinical trial. You have data coming in, right? And you can update. You can update each time you get more data.

And when you do this, your prior becomes a posterior on the first update, and then you use that posterior as the prior for the next update. Your beliefs are being updated as more data comes in. If you think about it, like when you walk outside this course, if you're like me you'll sort of suddenly realize you're doing this process all the time in your life, and you'll start approaching problems this way. And if you really like me, you'll change your religion on Facebook to Bayesian. But you have to be that extreme.

So great. So here what we want you to do is pretend that I roll the die and then got a 5. And then I rolled the same die, I didn't get a different. I just rolled the same die, just like I did in that demonstration, and then I got a 9. OK, so now we have a sequence of two pieces of data, 5 and 9.

So I want to find the posterior that-- update in two steps. I'm doing this one. I'm doing this one! OK, so let me show you how to do this, and then you'll do it. OK, so first we're going to update for the 5. Then we're going to update again for the 9. Magic!

The prior,  $1/5$  for each, right? That's where we're starting from. We roll a 5. That's our likelihood. Again, you get a 0 for the 4-sided and you get  $1/n$  for each of the other  $n$ -sided possibilities. We multiply. That gives us our unnormalized posterior for the first piece of data. Great. Now that is going to be our new prior.

Now, you may complain, wait a minute. This doesn't add to 1. But it's OK, because we're going to do it again. At the very end we'll normalize, and we'll get the same answer as if we had normalized here and then normalized again. Because we're just multiplying the whole column by the same scalar, by the same real number. OK? So it saves you work to not bother to normalize the first time.

Great. So that's our new prior. We get our next piece of data, the 9. We have 0, 0, 0,  $1/12$  and  $1/20$  as the likelihood. We multiply this new prior by this likelihood, get the new unnormalized posterior. It only has two possible nonzero probabilities in it. The sum, the total probability of rolling a 5 and then a 9 is only 0.0019, OK?

**PROFESSOR:** That's [INAUDIBLE] unnormalized [INAUDIBLE].

**PROFESSOR 2:** Ah! Jerry caught me. I was hoping a student would catch me. OK, maybe not. Maybe I was fooling myself. We haven't normalized. So, it's not quite right. If we had normalized here, then by the time we got to here, we could interpret this as that total probability. But in any case, the point is we can normalize to get to the final stage, this final posterior, by dividing by the sum here.

Now it adds to 1, and at the end of the day, what do we conclude? We conclude that it's almost three times more likely that it was the 12-sided die than that it was the 20-sided die. Right?

So this is very similar to what happened when randomly had the 12-sided and rolled it. After two times, I had data that looked a lot like this, and if we had rigorously computed those probabilities, this is what we would have gotten. OK?

So now, now it's your turn. So, we want to try the same thing, but in the other order. Suppose I first roll the 9 and then I roll the 5. So that's the first question. The second question is, can you do it in one go? Instead of doing it updating for the 9 and then for the 5, can you incorporate both the 9 and the 5 into a single likelihood function and do the updating in one step? All right, so try these and think about what's similar, what's different than what I just did.

You don't have to. You don't have to. You could normalize it, then you would multiply by the likelihood. You'd get something again unnormalized, and then you'd have normalize again, and then you'd get to your final answer. Or, don't bother normalizing at this stage. Just multiply this by the likelihood, and then just only normalize once at the very end. And you'll get the same answer because, you know, normalizing once and then normalizing again is the same as just normalizing at the very end. You're just multiplying a vector, this column, by rescaling it to make it unit length--

**AUDIENCE:** And then you're multiplying it by something else--

**PROFESSOR 2:** Exactly. Exactly.

**AUDIENCE:** Isn't number 1 going to be the same answer as the example we just did, because you're going to be updating-- I mean, it's going different. OK, we just have to do it.

**PROFESSOR 2:** Absolutely. So, yes, absolutely. All of these could give you the same answer because it's the same exact data, right? Starting from the same priors. But what we want you to really see is how the math proves that.

**AUDIENCE:** I mean we're just multiplying in different order.

**PROFESSOR 2:** That's right. So it all comes down to what? What property of multiplication?

**AUDIENCE:** Associative?

**AUDIENCE:** Commutative.

**PROFESSOR 2:** Commutative. I guess you're also using associativity. It's just that multiplication is commutative, right? So it doesn't matter what order you multiply these columns in, the final result's going to be the same. So just about how the order changes depending on the order of the 5 versus the 9, and also, what happens if you think of both at once, how would you then make a table directly for that?

[INTERPOSING VOICES]

**PROFESSOR:** 5. Oh, I see. So you multiplied-- here, is this your likelihood? So you multiplied that. Where's your plus times this? Oh, you even normalized. Right. So did you understand John's statement that you didn't actually have to normalize?

**AUDIENCE:** Right. I mean, if you'd already done it.

**PROFESSOR:** No, I understand. It's a nicer number than  $1/60$ . But if you didn't have it, you could have just used this column and normalized at the end. That make sense? Excellent.

I'm not trying to--

**AUDIENCE:** This is just if you do it in one step.

**PROFESSOR:** I see.

**AUDIENCE:** You just multiply.

**PROFESSOR:** Did you get different answers?

**AUDIENCE:** No, it's the same.

**PROFESSOR:** It's the same. It makes sense. You have the data. That's the evidence you have. Right. So was this all clear? Was John's statement about not only needing the unnormalized posterior in the middle reasonably clear?

**AUDIENCE:** What were talking about is if you were to use the normalized one, would then your unnormalized posterior be the actual probabilities?

**PROFESSOR:** No. No, because in order to use the Law of Total Probability, you have to multiply your likelihood by genuine probabilities. And the unnormalized posterior is not a genuine probability.

**AUDIENCE:** But if you use the normalized one.

**PROFESSOR:** If you used the normalized one, then it would be-- that would be the total probability of that. That's exactly right.

**AUDIENCE:** But, in this case, it's this that's the total probability, right?

**PROFESSOR:** Yeah. This is now your posterior probability. These are now the probability. So, make a distinction. The total probability is about the data, and these probabilities here are about the hypotheses.

**AUDIENCE:** Sort of like the fit of the hypothesis to the data.

**PROFESSOR:** You could think of that. They're genuine probabilities. This is the probability that you chose the 12-sided die given the data. That is, if I did this a million times and every time, I only looked at the times I rolled a 9 and then the 5, you would find this fraction of those times would be-- it would be a 12-sided and this fraction would be a 20-sided. So, I forget what the numbers were, like 3 to 1. Yeah. Makes sense?

Excellent.

**AUDIENCE:** So we multiply them together, but then--

**PROFESSOR 2:** You get the same exact answer, right?

**AUDIENCE:** So we get these, this answer.

**AUDIENCE 2:** That's right, so it's unnormalized.

**PROFESSOR:** And then you normalize it and you get this, right.

**AUDIENCE:** There's no way to normalize it without--

**PROFESSOR 2:** All right, let's come back together and take a look at what happened here. I saw some aha moments. That was great. Some people cited The Commutative Property of Multiplication. That made me very happy. So, what do we do?

So when we do the 9 and then the 5, it looks just like what I did with the 5 and the 9, except the likelihood columns are flipped. So we have our prior, the likelihood for the 9 gives us this column. The product gives us an, first, unnormalized posterior. Then we do the likelihood for rolling a 5. The product of the unnormalized posterior and the second likelihood gives us our final unnormalized posterior.

And notice the sum is still exactly what it was before, 0.019. When we normalize, we get exactly the same answer that we had before. Now, if you contrast this with the table we looked at when we did the 5 and then the 9, what changed is these two columns were flipped around. And so I suppose this column was different and this column was different as a result. And then the final result, however, was exactly the same.

Now, what if we do them both at once? What should that mean? If we do them both at once then, well we start from the same uniform prior, and our likelihood-- now, we should think of the data as both rolls, the 5 and the 9. OK.

Now what's important here is that if we condition on a given hypothesis. For

example, if we condition on the hypothesis that it's a 12-sided die, then the first roll and the second roll are independent events, right? It means that we can figure out the probability of a 5 and the 9 on the 12-sided die. It's just the product of the probability of a 5 and a 12-sided die and the probability of a 9 on a 12-sided die.

That's why we're justified in taking the product. If we condition on the die that it is, on the hypothesis, 12-sided, then the rolls are all independent of each other. In particular, we get  $1/12$  times  $1/12$  for rolling a 5 and then rolling a 9. And for the 20-sided die, we get  $1/20$  times  $1/20$ .

So now in one step, we update. Same exact sum, we normalize, same exact answer. And I mentioned that some groups, maybe when slightly prompted by me, said oh, multiplication is commutative. Right? That's all that's going on here. In the end, all we're doing is taking our prior, the first likelihood column, the second likelihood column, multiplying them together and then normalizing it. Right?

It doesn't matter what order those likelihood columns are in. It doesn't matter if we think of the 5 in the 9, the 9 in the 5, or the 5 and 9 multiplied together to make their own likelihood column. You get the same exact unnormalized result no matter what.

OK. So this is all there is to why it doesn't matter what order you update the data in sequentially, it just matters what the data is. Now I also had a question. Why if you could do it all once would you even bother with doing it once and then doing it again? It's a very good question. Does anyone want to suggest a reason? Maybe other than the group that I talked to specifically about it Yeah.

**AUDIENCE:** Well, I'm just thinking, as you're saying, if you have ongoing patients and you have to like continuously update, then the first data set may be all that you have at that moment, but then you have to keep adding on to it.

**PROFESSOR 2:** That's exactly right. That's exactly right. So, in life and in clinical trials, you don't get all the data necessarily at once. It sort of comes in continuously. When I wake up in the morning, I don't know what temperature it is outside. I have a prior, based on what temperature it was yesterday at about the same time, like I have a memory of

that. So it gives me some prior distribution on the temperature. OK, that might be a continuous one, like maybe a bell curve. We're going to get to that next week, these continuous updating problems. OK.

And then, when I get out of bed, I see that the window is covered in condensation. That's some data. I can update my beliefs. Probably means it's kind of cold. Great. Then I can open the window to let the cat out, and a burst of freezing cold air comes in, like it did this morning-- well, actually not this morning. This morning was great. Yesterday it was cold.

And again, now I have more data and I update my beliefs. All right. You can go through life this way. It'll give you a headache, but you can do it. Now with clinical trials, the same thing can happen. Maybe you do a pilot study. Maybe you have some belief that eating almonds causes something, stomach cramps. Great. And maybe it's just for you. But you want to do a study about that.

So maybe you know for yourself that eating almonds give you stomach cramps. You're at the Harvard School of Public Health and you really want to investigate this further, so you gather a few willing subjects, you feed them some almonds, you see if they get stomach cramps, how many of them. And maybe your prior ahead was you just kind of felt like you had this really strong belief that almonds cause stomach cramps, because it does for you and you really haven't even asked a lot of other people. So you might have a lot of prior probability on that hypothesis being true.

Once you collect a few more people and maybe you ask your friends, maybe you find that most of them don't have the same problem. Suddenly your posterior belief changes and now you're not quite so convinced that it's true. And then, if you're really ambitious and you have a lot of grant money from the NIH, you recruit 1,000 people and you do a randomized controlled trial with placebo and you find out that exactly 10% of the folks complain about the stomach cramps with almonds. Maybe 10% on placebo also, so you're not quite sure what's going on. But the point is, you're getting data a little more, a little more. And you can keep incorporating it, updating your beliefs.

So that's the principal. One last thing I'll point out is, what if you have a prior belief that is so strong, so you are so sure, for example, that almonds don't cause stomach pain, because it doesn't for you and you're just sure that that would be ridiculous, right? So you've put your prior probability at 0. How much evidence will it take, how much data will take for you to be convinced?

Infinite. No amount of data it's going to convince you otherwise. This is why whenever you do this kind of Bayesian Updating, it's generally best to leave a little bit of possibility for Santa Claus, OK. Like a little bit. Just so that should a fat white man come down your chimney and deliver the presents one year, you have a possibility to adjust.

So those are some things to keep in mind. I think we've got one more board question that Jerry's going to lead.

**PROFESSOR:** Right so it's nice to get probabilities for hypotheses, but one thing you also want to do is predict what's going to come next. And so that's what we call probabilistic prediction. If I pick one of the John's 5 die, and I'm about to roll it, you could tell me using the Law of Total Probability the probability that I get a 5.

And we've done that, we did that in the first unit on probability in this class. But suppose he rolls a 5 first. Now I can update my probabilities for which die it is, which is going to update my estimation of the probability of rolling, say a 4 on the next roll. This is what we'll call probabilistic prediction. We can predict what's going to happen next, at least with probabilities, based on the data that we've seen. We update our predictions based on that.

So here's a set up. D1 is the first roll. D2 is the second. I'll do this example. Oh no, this a board question. You're going to do this. You're on your own here. So first, you don't know anything about the die except that 1 of the 5 was chosen at random. What's the probability that the first roll will be a 5?

Now you've rolled that once, and you're going to roll again. And you can ask, well now I've seen some data. I have some information about which die was chosen.

What's the probability that I'll see a 4 on the next die. So that's your job to find that for this problem.

**PROFESSOR 2:** Cool, alright so this is your first answer?

**AUDIENCE:** Right. And then, now we're just multiplying across. You get 0. 1/180, 1/320, 1/740, [INAUDIBLE]. And then add those up.

**PROFESSOR 2:** Yeah, get out your calculator.

What are you dividing by to normalize?

**AUDIENCE:** Is this-- I don't know.

**PROFESSOR 2:** That's right. So this is, when you add these up, exactly the probability of the 5, right? So now if you want to repeat this process, you need to use your normalized posterior instead of the original prior.

**AUDIENCE:** That's what we were--

**PROFESSOR 2:** Yeah. Great. Now, you just need to do division. That I can't help you with. Calculator. Division skills, either way.

[INTERPOSING VOICES]

**AUDIENCE:** Now this is the probability of 4, or the probability of 4 getting 5?

**PROFESSOR 2:** That's right. So tell me what you did to get these.

**AUDIENCE:** So this column [INAUDIBLE] which we calculated for, and then you sum it up to get the total probability of getting a 5. And then [INAUDIBLE] normalized it, and then took the new [INAUDIBLE] times the same--

**PROFESSOR 2:** The same likelihood.

**AUDIENCE:** And then you get the new column and then you add it up. [INAUDIBLE].

**PROFESSOR 2:** Perfect. Now you said, just for the heck of it you normalized it, but it's actually really

important that you normalize it, because if you're going to use the Law of Total Probability again, right, you've got to be working with a probability mass function, something that adds to 1 here. If you didn't normalize, then your final answer would get multiplied by something, right? Whatever that normalization constant was, or its reciprocal. Perfect.

Check these guys out.

**PROFESSOR:** And this right here is not a probability, because it's unnormalized. So what should you normalize this by?

**AUDIENCE:** [INAUDIBLE]

**PROFESSOR:** Yeah. In other words, if you made a normalized probability here, it would be this divided by that, this divided by that.

**AUDIENCE:** So all we need to do is divide this--

**PROFESSOR:** Exactly right.

**PROFESSOR 2:** How's it going?

**AUDIENCE:** It's going fine. So we're just saying--

[INTERPOSING VOICES]

**PROFESSOR 2:** So, what did you get for this? How did you find it? So you summed up this, right? Great, and then what did you do next?

**AUDIENCE:** We do the same thing for 4.

**PROFESSOR 2:** So, starting from scratch.

**AUDIENCE:** Yeah, starting from scratch if you roll a 4.

**AUDIENCE:** OK. So how does that incorporated the fact that you first rolled a 5?

**AUDIENCE:** So we're trying to go back to the original Bayes, but that didn't work out so well

because we don't know how to calculate this.

**PROFESSOR 2:** Exactly, all right. So this is no easier than that. So here's the idea: so you start with this belief, right? And you have your likelihood. You multiply them, Law of Total Probability, and you get the probability of rolling a 5. Now at that point, after you've rolled this 5, your beliefs have changed based on that about the probability that you have each of the 5 dice. So if we take that and you use that as your prior here, then the likelihood, multiply and sum, that should give you--

**AUDIENCE:** I like that.

**PROFESSOR 2:** OK?

**AUDIENCE:** Oh, but now we have decimals.

**AUDIENCE:** Now you have decimals. I can't make your life too easy.

**AUDIENCE:** Use these.

**PROFESSOR:** Right, so you multiply this number by this, this number by this. So in the end, you'll get all of these numbers except they're all multiplied by, divided by. So you should divide this by-- to get a probability. That make sense? So you have to be careful for figuring out probabilities of hypotheses, we can avoid using the unnormalized. We don't have to keep normalizing. But when you want to do posterior prediction, you want real probabilities of data. So you're going to have to normalize these numbers at some point. Either first or at the end. It's probably, if you're going to do it a lot, easier to just keep doing it.

**AUDIENCE:** OK, Thanks.

**AUDIENCE:** How are we doing?

**PROFESSOR:** You tell me. It's a beautiful use of color.

[INTERPOSING VOICES]

**PROFESSOR:** So is this normalized or not normalized?

**AUDIENCE:** This is normalized.

**PROFESSOR:** Normalized. Good, because you are real probabilities there. Yes, thank you.

**AUDIENCE:** [INAUDIBLE]

**AUDIENCE:** And so then we multiply by what are the likelihoods.

**PROFESSOR:** Yeah.

**AUDIENCE:** [INAUDIBLE]

**PROFESSOR:** And now you're going to sum it up.

**AUDIENCE:** Yeah.

**PROFESSOR:** That's perfect.

**AUDIENCE:** Way to go Caitlin!

**PROFESSOR:** Because the Law of Total Probabilities says multiply this by this. That's just what you did with the trees, do you see the trees when I wave my hands? Beautiful.

**AUDIENCE:** [INAUDIBLE]

**AUDIENCE:** Our new prior is given the data we had before.

**PROFESSOR:** Yes. That's good.

**PROFESSOR 2:** You're adding them up? OK, so that looks right. So let's think, how does this number compare to what you would have, if you didn't know this other information? So suppose you knew nothing going in, do you think that the fact that you rolled a 5 the first time makes it more or less likely that you roll a 4 the second time.

**AUDIENCE:** It should be less likely.

**PROFESSOR 2:** Is it less likely?

**AUDIENCE:** Because you get rid of the 4-- the die has only 4 sides.

**PROFESSOR 2:** So it might make you go down, right? On the other hand, you're going to give more probability to the 6-sided, let's say that, than it had originally. So they're sort of competing forces. So I'm actually not sure. I'm not actually sure which is the right answer. But anyway, it's something you can easily check or compute.

**PROFESSOR:** This is a standard day. We always finish ahead of time. I don't need that right here. Like with all the other answers today, we'll use the screens to show the table. So when we go through it, we do the usual thing we did before. We started with a 5, so we'll update our probabilities of the hypotheses based on that 5. So that's the first column is the prior, then the likelihood of getting a 5, which is 0 if you have the 4-sided die:  $1/6$ ,  $1/8$ ,  $1/12$  and  $1/20$ .

We get our unnormalized posterior. And then in this table, I actually got the normalized posterior, because when you're computing predictive probabilities, you need to use normalized probabilities. They should be true probabilities. So we use this column here of the posterior probability. There's different orders you could do this calculation. I think this is one of the straightforward ones.

Then we look at the likelihood of D2 in the next column. Now I put a star for H4 because there's no way you could get a probability there if you have D1 and D2. If I had left off the D1, which most people did and which is OK because of the independence John talked about, given the hypothesis the rolls are independent, I would have had a  $1/4$  where I have the star. But I'm going to multiply it by 0 anyway, so who cares.

So I have my new priors-- excuse me, my new likelihoods for the second bit of data, which are the same as the old ones:  $1/6$ ,  $1/8$ ,  $1/12$ ,  $1/20$ . Now my total probability is done the same way. You multiply in this new language posterior 1, which is your prior for likelihood 2, times likelihood 2 and sum them up, and I get this last column.

So again, once I have my new, my posterior for the first bit of data, I can use that in the Law of Total Probability to make a prediction. So what's happened right here?

What's become more likely? This 0.124 seems a little bit larger than I would have guessed ahead of time. What's happened with rolling a 5 and then a 4 to make that a little bit bigger? Any ideas?

**AUDIENCE:** When you roll a 5 first, it increases the probability that it's one of the lower dice.

**PROFESSOR:** Excellent.

**AUDIENCE:** So then it makes it more likely that you'll get a 4.

**PROFESSOR:** Excellent. That's what I would think too. Seeing a 5-- a 4, what did roll first? The 5-- makes the smaller die more likely. If the smaller die is more likely, then you're more likely to roll a 4 and a 5. For instance, if I had the 6-sided die, what's the probability I'd roll a 4 and a 5 with the 6-sided die? 1 in 36 which is about 3%. So this doesn't get up to 3%, but it's a little bigger than if I had the 20-sided die where would be-- what's 1/400.

**PROFESSOR 2:** Very small.

**PROFESSOR:** Small, John says it's small.

**PROFESSOR 2:** So, this is one argument you can make, right? That rolling the 5 makes it likely that it's a smaller die, and therefore more likely to roll a 4 next. But is there another argue that you can make for why maybe rolling the 5 would have the opposite effect? It might make it less likely to roll a 4 next?

**AUDIENCE:** My initial thought was if you roll a 5, you know it's not the 4-sided die. And that's your die that you're most likely to roll for.

**PROFESSOR 2:** That's right. So I'm not as sure as you as to whether this goes up or down. I think there are two competing arguments. That on the one hand, it makes it more likely it's a 6 or 8-sided, but it also makes it completely impossible that it's a 4-sided.

**PROFESSOR:** I'll give you 3 to 1 odds that it goes up.

**PROFESSOR 2:** Aw crap.

**PROFESSOR:** So what he doesn't know is whether I've computed already or not.

**PROFESSOR 2:** He probably has. But he probably did it at like 3 AM, so I don't trust it.

**PROFESSOR:** I simulated it in [INAUDIBLE].

**PROFESSOR 2:** I see. Are there any questions? That's what we have for today, so are there any questions about Bayesian Updating or doing these computations? All right, great.