[SQUEAKING]

[RUSTLING]

[CLICKING]

SARA ELLISON: OK. So let's get started. I'm going to start with a few definitions.

So a sample space is a collection of all possible outcomes of an experiment. And an event-- we'll denote that with capital letters, A, B, C, things like that-- is any collection of outcomes, including individual outcomes, the entire sample space, the null set, et cetera, OK?

So we have some kind of an experiment-- and I'll do some examples in a few minutes to add some concreteness to this. But we have some experiment, and an event is a collection of outcomes of that experiment. Yep.

So if the outcome of the experiment is a member of an event, then we say that the event is said to-- or we say the event has occurred. And event A is contained in event B if every outcome in B also belongs in A. So here we're just establishing sort of a bunch of terminology and definitions.

And as you have probably gathered, these things that I'm defining are sets. And so we're going to use many of the same definitions and results from set theory in probability theory because we're sort of setting everything out, or the foundation that we're using is-- the mathematical foundation is the one from set theory.

So all the results from set theory are going to apply-- the associative, commutative, distributive properties, et cetera. And here I'm just going to post some useful results without much discussion of them. If you've done some set theory or some probability theory, these things should be pretty familiar to you.

If an event A is contained in another event B, then the union of those two events is equal to event B.

Oh, I should point out a couple of notes about notation. OK, so the sign that looks kind of like a C is how we indicate "contained in." And that piece of notation is straight from set theory. And this capital U kind of thing is how we indicate union. And, again, that notation is straight from set theory.

The second useful result of A is contained in B, and B is contained in A. Then the events A and B are equal. And if A is contained in B, then AB is equal to A.

Now, here is actually one place where the notation between set theory and probability theory deviates a little bit. Let's see. Oh, actually, let me skip ahead.

In set theory, typically, we indicate the intersection between two sets with a symbol that looks like this. And in probability, we usually just leave that out. So instead of talking about or writing the intersection of AB like this, we usually just write the intersection of AB this way. And I'll follow that convention most of the time, just leave out the intersection.

And let's see. I'll go back for a second. And then, finally, the union between A and A complement is equal to the sample space. So, again, these are mostly definitions taken straight from set theory. You can check standard set theory references if you need a little refresher. But for the most part, these are just-- we're going to rely on the apparatus of set theory initially.

OK, two other important definitions where probability theory uses slightly different terminology than set theory is the following two. So we say that A and B are mutually exclusive if they have no outcomes in common. So we have two events. And if they don't share any outcomes, we say they're mutually exclusive.

We also sometimes say they're disjoint. And in set theory, we'll typically talk about them being disjoint. In probability theory, you use the term "mutually exclusive" more commonly.

A and B-- events A and B are exhaustive if their union is equal to the entire sample space. And, again, in set theory, we'll typically use the term "complementary" instead of "exhaustive." And in probability theory, we'll usually use "exhaustive." They mean the same thing.

So oftentimes, we'll talk about, and we'll talk about later today groups of events that are mutually exclusive and exhaustive. And that just means that their union is equal to the sample space, but they have no events in common. And another term for that is a partition. They form a partition of the sample space.

OK. So now that we have a few of those definitions and some of the terminology and notation under our belt, let me define probability. So we can take-- we're going to take every event A and assign it a number, P of A. And so this is basically a function that goes from the sample space to the real line.

And we're going to require a few things of this function, P of A. First of all, we're going to require that it's greater than or equal to 0 for all events in the sample space. The second thing that we're going to require is that the function P of the sample space, the entire sample space, is equal to 1.

And the third thing we're going to require is for any sequence of disjoint sets-- I'm going to call those A1, A2, dot, dot, dot. So it could be a finite sequence. It could be an infinite sequence.

But for any sequence of disjoint sets, the probability of the union of that sequence is equal to the sum of the probabilities of those events. Is this notation familiar to most of you, this sort of union and summation notation?

OK, so it's important to point out that this thing, this function P, this probability, isn't necessarily a unique object. We're just saying we have a sample space. And if we have some function that satisfies these three properties, then we're going to call it a probability. And it's just simply any function defined on a sample space that satisfies these three properties.

Just a little comment about terminology-- so different textbooks use different terms for this object here. So sometimes, you'll see it called a probability distribution or a probability function. I'm trying to use the most standard terminology that I can find, but there is no standard terminology for all of probability theory. So just keep that in mind that sometimes the terms are actually a little ambiguous, and feel free to ask if you ever have any questions about exactly what I'm referring to.

OK, so we've got this thing now. We call it a probability. And you can prove a lot of useful things about probabilities using set theory.

So if we were doing a course entirely in probability, we'd probably spend the rest of the lecture just using the sort of mechanics of set theory and proving lots of things about probabilities. We don't have that luxury here. We want to get to everything that we're excited about in data analysis in the rest of the class.

And so what I'm going to do is just list a lot of the useful facts and useful results about probabilities that we could prove. And I'll mention a couple of them quickly. First of all-- and we'll actually use this in an example either today or next-- probably today. The probability of A complement-- so the event that contains all of the outcomes that are not in event A-- the probability of A complement is just equal to 1 minus the probability of A.

So why is that useful? Well, it's useful in particular because sometimes, the probability of A complement could be difficult to compute, and the probability of A could be very easy to compute. And so if you want to get at the probability of A complement, this result could be the easiest way for you to do it.

The probability of the empty set is equal to 0. If A is contained in B, then the probability of A is less than or equal to the probability of B. We can also prove that for any event, the probability of that event is between 0 and 1.

Probability of A union B is just equal to the sum of the probabilities of those two events minus the probability of their intersection. And finally, the probability of A times B complement is equal to the probability of A minus the probability of the intersection. So just keep in mind we'll refer back to a number of these results in the examples that we're doing coming up. So keep those in the back of your mind.

So now I'd like to introduce an important special case. Suppose that we have a sample space, and suppose that that sample space is finite. So there's a finite number of outcomes in this sample space.

Let's define a function n that gives the number of elements in a set and then define a probability-- well, we haven't checked to make sure it's a probability-- but define a function P of A which is just equal to n of A over n of S.

So what is this? We're defining a function. We're defining something that we hope is going to be a probability. We're going to say, all we do is we count the number of outcomes in an event and divide by the number of outcomes in the sample space. And that gives us our probability. All we're doing is counting.

This is called-- this has a name. It's called a simple sample space. And we can check, in fact, that it is a probability. So how do we check that it's a probability?

Well, I told you before, if we have a function and it satisfies those three axioms, it's a probability. So all we have to do is check to make sure this particular function P of A satisfies the three axioms. So I did it right here.

First of all, it's always going to be nonnegative because it's a count. So we're fine with axiom 1. Axiom 2, probability of the whole sample space is going to equal to 1. That's essentially by definition because we're just counting-- we've defined P of A as n of A over n of S.

So just P of S is equal to n of S over n of S. So we're fine with axiom 2. And axiom 3, the probability of the union between A and B is just equal to the number of elements in the union of A and B. And since they're mutually exclusive, that means that it's just equal to the number of elements in A plus the number of elements in B. And then that's just equal to, by definition, probability of A plus B, or probability of A plus probability of B.

Questions about-- is this OK? OK. So you can just go back and look at the axioms with this particular function and double-check to convince yourself that it satisfies all the axioms. OK?

So why do we have this special case, and why do we consider it? Well, it's a really powerful notion. If you think about it, if you can take an experiment that you're interested in and put it into this framework of a simple sample space. In other words, create-- define the experiment or sort of-- yeah, basically, put it into the framework where each outcome is equally likely-- that's what the simple sample space implies-- all we need to do is count to compute probabilities of events.

So basically, if we're interested in the probability of a particular event, and we can describe the experiment in the framework of a sample space, all we have to do is count, and we can compute the probability. Is that clear? OK.

So let me do a very simple example first. Let's suppose I roll two fair dice, and I want to compute the probability that the sum on the faces of the two dice come up 4 when I roll them. Well, if I can put this in the framework of the simple sample space-- so in other words, have every outcome be equally likely-- then all I have to do is count.

All I have to do is count all the possible outcomes and count all the outcomes that satisfy my particular event, and I'm done. I've computed my probability. So that's exactly what I do here.

So what are all of the possible outcomes of this experiment? Well, if I have two dice, and they're fair dice, that means each face is equally likely to come up. And so then each one of these pairs that I've indicated is an equally likely outcome of this experiment. So I could get 1 on the first die and 1 on the second die. I could get 1 on the first die and 2 on the second die, et cetera. So if I count up all of those, I get 36.

And the event that I'm interested in is that the sum of the two dice is equal to 3. Oops. Oh, no, no. That's right-- is equal to 4. Sorry, is equal to 4.

And how many different outcomes give me that the sum of the two dice is equal to 4? There are three different outcomes, exactly-- so 1 on the first die, 3 on the second die, 2 on both of them, or 3 and 1.

And so then, to compute this probability, that's all I did is I counted the number of outcomes in the sample space, I counted the number of outcomes in the event, and I divided. So that's what makes simple sample spaces such a powerful notion. And a lot of times, when we want to calculate probabilities, we can put the experiment into this kind of framework. It makes our lives easier.

One more example-- slightly less simple, I guess-- so let's say the state of Massachusetts issues license plates with six different characters on them. And for each character, you can use one of 26 letters or one of 10 digits. And they're just chosen randomly on the license plate.

So my question is, What's the probability that I'll receive an all-digit license plate? Well, we have 36 possibilities for the first character, 36 possibilities for the second character and so forth. So just intuitively, if you were to start enumerating all of the different possibilities, you'd figure out pretty quickly-- you don't have to actually count all of these possibilities. You don't have to write all of them down. You just have to realize that there are 36 possibilities for the first character, 36 for the second, and so forth. So the total number of possibilities is 36 to the sixth, OK?

And it turns out if I typed it into my calculator correctly, that's 2.176 billion. So we have a lot of possible license plate numbers. So what's the probability that I'm going to receive a license plate with all digits just by chance?

Well, there's only 10 possibilities for the first character in that license plate-- 10 to the second and so forth. And so that gives me 10 to the sixth possibilities for my all-digit license plate, or a million. And so then if we divide, we do the division, we get the probability of A is equal to 0.0005.

Questions? Is this clear? Does this make sense so far? Good. OK, so this example-- and we'll come back to this. This will arise again throughout the semester. But this is an example of a sampling with replacement.

What happens, however, if Massachusetts, in their making of the license plate, decided they didn't want to reuse a letter or a digit? OK. Now, in this sample space, there are 36 possibilities for the first character. But they're not going to reuse whatever they use in the first character. So there's only 35 for the second character.

And they're not going to reuse what they're using for the second character and so forth. So there's only 34 for the third character. And then we can multiply those, the number of possibilities for each one of the characters. And we come up with something-- well, we come up with 36 times 35 times 34 and so forth.

And we can write this in a somewhat more compact way as 36 with an exclamation point after it-- that's called factorial-- divided by 30 factorial. OK? OK.

So we can write the number of possible license plates here as 36 factorial over 30 factorial. OK? And then, in counting the number of possibilities in the event, there are 10 possibilities for the first character, 9 left for the second, and so forth.

And so then I can express the number of possibilities in the event as 10 factorial over 4 factorial. And so the probability here is 0.0001 when I do the division. And this is an example of sampling without replacement. And, again, this is something that we'll see later on in the semester as well.

OK, questions before I go on? Nope? Oh, I guess I just mentioned that. So this is just a note about the factorial notation.

So to compute probabilities in these examples that I just gave you, all I did was count. The counting was a little fancy sometimes. But it was just counting.

So let me actually be a little bit more concrete, a little bit more organized, about the counting rules that I used. And another term to describe these counting rules is combinatorics. So the first counting rule that I used in one of the examples-- or, actually, all the examples-- if an experiment has two parts, the first one has m possibilities and the second one has n possibilities, then the experiment has m times n possible outcomes.

So this is something-- we're just being complete by including it here. But this is something you would do intuitively anyhow.

Two other counting rules-- an ordered arrangement of objects is called a permutation. The number of different permutations of n objects is n factorial. So I use this counting rule on the second license plate example. And more generally, the number of different permutations of little n objects taken from big N objects is a big N factorial over a big N minus little n factorial. So that's exactly the counting rule that I used in the license plate example.

And then, sometimes, we want to count-- we want to count outcomes in a sample space. But we don't want to count-- so in the license plate example, order was important. So a license plate that was ABC123 is different from a license plate that's BCA231 or something.

And so there are going to be problems, though, where order doesn't matter. And we don't want to count different orderings as different outcomes. And so then we have a third counting rule, which is called the counting rule of combinations. An unordered arrangement of objects is called a combination.

And the number of different combinations of little n objects taken from big N objects is just the same as the number of permutations, but we also divide by little n. And I think of that in some sense as taking out-- making sure we're not double-counting all of the different orderings. So you're taking out the orderings by dividing by little n factorial.

We typically denote this using notation like this. And we call it-- the terminology we use is big N choose little n. OK.

Let's go through a few more examples. Let's suppose all the candidates for the Republican presidential nomination gather on stage for an event. How many handshakes are exchanged if all of them are exchanging handshakes with all of the others?

By the way, who knew? Apparently, he got 12 votes. I checked it on Wikipedia. He got 12 votes in Iowa. Did anyone know Jim Gilmore was still in the race? No. You did, OK. I had no idea. Anyhow, I checked it on Wikipedia-- 12 votes in Iowa. I think he's still in there.

OK, so let's suppose all nine of these men-- gentlemen and one woman-- or eight gentlemen and one woman-- gather on stage, and they're all exchanging handshakes. How many handshakes are exchanged? So this isn't a probability problem. This is just using sort of the counting rules. Any guesses? Yep?

AUDIENCE: 9 choose 2.

SARA ELLISON: 9 choose 2? OK. So do you want to explain-- would you like to take a shot at explaining why that would be?

AUDIENCE: So there's nine candidates, and we want to find the number of ways to pick any two of them. So therefore, we want to use 9 choose 2.

SARA ELLISON: That's right. That's right. So a handshake involves two of the nine candidates. And order doesn't matter because Jim Gilmore shaking Carly Fiorina's hand is exactly the same as Carly Fiorina shaking Jim Gilmore's hand.

And so we don't want to use a permutation counting rule. We want to use the combination counting rule. So you're exactly right.

Another example-- so we've got this new building. You guys should come visit it if you haven't already. So the economics department just moved back into a renovated E52. It is spectacular.

And all of the faculty-- we have all these beautiful, newly renovated faculty offices. And all the faculty got assigned to one of the offices. So I think we have 40-- maybe we have 41 faculty offices. I'm not sure. Let's say we have 40 faculty offices in the renovated E52.

Now, for the purposes of this problem, let's assume that they're in a continuous line. They're actually not. They're in two different floors. Let's assume they're in a continous line. And if 40 faculty members are placed randomly in the offices, what is the probability that Esther and I are next to each other in the new E52?

By the way, this is actually where we are. This is Esther up here on the upper left, and I'm over there on the-- but what's the probability that Esther and I are next to each other? Does anyone want to take a stab at this?

So maybe we should go step by step. How many different ways are there to put 40 faculty members into 40 offices? I think I heard it. 40 factorial. Right, OK.

And how many different ways are there to do that where Esther and I are sitting next to each other? I hear little murmurings, but I'm not sure if I heard the right one.

AUDIENCE: [INAUDIBLE]

SARA ELLISON: Oh, OK.

AUDIENCE: [INAUDIBLE]

SARA ELLISON: Yeah.

AUDIENCE: [INAUDIBLE]

SARA ELLISON: So the way that I think about this problem is that I think about-- oh.

AUDIENCE: 39 [INAUDIBLE]?

SARA ELLISON: Yes, sorry, 39 factorial times 2? Is that what you said? Yeah. Would you like to try to explain it if you--

AUDIENCE: No.

SARA ELLISON: No? OK. OK, I'll explain it then. So the way I think about this is that someone ties Esther and I together. And now we're one unit. We have to come together, right?

And so now there's only 39 factorial ways to do the arrangement since the two of us are tied together. Oh, but there's one extra thing we have to consider, which is that she could be tied to my left or she could be tied to my right. So then we have to multiply by 2, OK? So there we have it.

OK. Yes?

AUDIENCE: For your previous example, does that-- tied to your left versus tied to your right, does that account for like if one of you has the corner?

SARA ELLISON: So I'm assuming that the offices are in one continuous line, just to make the question simpler. Does that answer your question?

AUDIENCE: Kind of.

SARA ELLISON: OK. OK. Another example-- maybe this will make you hungry because it's right before lunchtime. Area Four-- has anyone been to Area Four? You guys know Area Four? Yeah. OK.

So I went on to their website as I was writing my lecture notes. I was getting hungry. So I went on to Area Four's website to see what toppings they offer on their pizzas. So it turns out they have six vegetarian pizza toppings-- caramaliezed onions, pickled banana peppers, mushrooms, green olives, arugula, and eggs. I think that's it.

And then they have five nonvegetarian toppings-- soppressata, sausage, bacon, chicken, and anchovies. So let's suppose I want fate to decide what pizza I order instead of just choosing my toppings.

So I write each topping on a piece of paper, and I put them in a hat. And I randomly choose out two toppings from two pieces of paper. What's the probability that I end up with a pizza that has one vegetarian and one nonvegetarian topping?

AUDIENCE: [INAUDIBLE]

SARA ELLISON: Sorry.

AUDIENCE: 6 over [INAUDIBLE].

SARA ELLISON: Oh, so I think you're on the right track. But let me put up the explanation. So the idea is, first of all, we have to count the number of different possibilities in the sample space. And that's just going to be-- sorry, that's going to be 11 choose 2.

And then the number of possibilities of one vegetarian and one nonvegetarian topping-- we have six nonvegetarian possibilities and five-- sorry, I have it the wrong way around. I think six vegetarian and five nonvegetarian-- so we have 30 possibilities. And I think this is what you were getting at-- 30 possibilities for a pizza with one vegetarian and one nonvegetarian topping. And so then we just divide the number of outcomes in that event by the total number of outcomes, the total number of possibilities. So we end up with 30 over 66.

AUDIENCE: [INAUDIBLE] think about it because you have everything in one [INAUDIBLE]. And then so you have 5 and 6, [INAUDIBLE]. So you have a total of 11, right?

SARA ELLISON: Yes.

AUDIENCE: So it's 6 over 11. And so that's my first topping [INAUDIBLE]. And then I have another probability, which is 6 over 10, which is the remaining 10 because I already have one [? extra. ?]

SARA ELLISON: So you're absolutely right that that is an alternative way that you can think about-- as opposed to this way, and you get exactly the same answer. So you're absolutely right about that.

I think a lot of times, when we see probability problems, there are different ways to approach them that give you the same answer. And the reason why-- I actually thought about giving the two alternative ways to solve this problem, but we don't have quite all of the machinery to solve it the way that you suggest. But you're absolutely right. Yeah.

So there's another way that you can think about this problem in a slightly more general fashion. So let's suppose instead of just choosing two toppings, I was choosing n toppings. And I ask what the probability that my pizza had n1 vegetarian toppings and n2 nonvegetarian toppings was.

So this is basically just a generalization of the problem that I just did. Instead of having one vegetarian and one nonvegetarian, we're having n1 and n2. And note that there would be 6 choose n1 possibilities for the vegetarian topping, and there would be 5 choose n2 possibilities for the nonvegetarian topping.

So we multiply those two together in the numerator to count the number of outcomes in the experiment that satisfy our event. And then the denominator is just 11 choose n for the total number of n toppings that we're putting on the pizza. OK? And I want to note that we're going to refer back to this example as the basis for a special distribution, the hypergeometric distribution.

So going forward it's going to be important for us to be able to talk about the relationship between these probabilistic events, also known as stochastic events. And the most fundamental of the relationships that we're going to be concerned with is something called independence. So let me first give you the definition of independence in probability.

Events A and B are independent if the probability of their intersection is equal to the product of their probabilities. So I never find this definition very intuitive. I feel like probably a lot of you have sort of an intuitive sense of what independence means in probability. But this is not a very intuitive definition, or at least in my opinion it's not.

There is, though-- I should give a word of warning, though, that sometimes, our intuition about what independence is could lead us astray. So let me give you an example to illustrate that.

Suppose I toss one die. Now consider two different events. A is the event that I roll a number less than 5, and B is the event that I roll an even number. Are these events independent?

Well, your intuition might make you first think, well, how could they be? They rely on a single roll of a single die. So how could these two events actually be independent? But in fact, they are.

If we use the definition, if we check using the definition of independence, we find that-- we write down the probability of the first event is equal to 2/3, the probability of the second event is equal to 1/2, and the probability of their intersection, the intersection being an even number less than 5, rolling an even number less than 5, the probability of that is equal to 1/3. And that is the product of the two probabilities of the individual events. So they are. It does, in fact, satisfy the definition of independence.

And I think that this example sort of illustrates the way in which we need to be careful about our intuition and, in fact, refine our intuition based on this. So, really, the proper intuition about independent events is that if I know that one occurred, it doesn't tell me anything about the probability that the other occurred. OK? Yes?

AUDIENCE: So this is basically an issue of using independence with mutual exclusivity, right?

SARA ELLISON: Yes. Well, so you're right that independence and mutual exclusivity are not the same. And it could be that that's the way your intuition was going. So, yes, in that sense, you're absolutely right. And so, yeah, yeah, it's important to draw that distinction.

Another sort of useful thing that we can prove about independence is that if we have two independent events, A and B, then A and B complement are also independent. And I've listed the proof there in case you want to convince yourself that that's true.

And furthermore, if we have more than two events, the definition that we use for independence kind of generalizes in the way you think it would. So the events are independent if the probability of their intersection is equal to the product of their probabilities for more than two events.

OK, who's this?

AUDIENCE: Steph Curry.

SARA ELLISON: OK, does anyone want to give the class sort of a two-sentence skinny on who Steph Curry is for those people who are not NBA fans? Yeah?

AUDIENCE: The best shooter in NBA history.

SARA ELLISON: The best shooter-- that was a less-than-one-sentence description, the best shooter in NBA-- and in particular a 3-point shooter. Right? So--

AUDIENCE: [INAUDIBLE]

SARA ELLISON: OK, fine, fine. So anyhow, he's having this absolutely phenomenal season, and not only are NBA fans really excited about it, but sort of data geeks are really excited about it, too. And we'll actually see some data on Steph Curry's shooting pretty soon, in a couple of lectures, and we'll get to play around with it. But let me just do an example sort of, I don't know, inspired by Steph Curry's performance first.

So I looked up his 3-point field goal percentage. Oh, maybe I should take a step back. If you really don't know anything about basketball, a 3-point shot is one taken from-- how many feet away from the basket it is? 23 feet or more away from the basket. And these are, depending on the shooter, can be pretty low-percentage shots. For Steph Curry, they're not low-percentage shots. He makes 44% of them.

For the purposes of this problem let's assume that every 3-point shot he attempts, he has the same probability of making it. Realistically, that's not true, but let's assume that for now. And let's also assume that all of these shots are independent. So we're ruling out things like him being in a slump or him having a hot hand or something like that.

So for the purposes of this problem, we have a guy. He makes 3-point shots 44% of the time he takes them. And they're independent. There's independents across shots.

What's the probability that he misses the next three shots that he takes and then makes the three after that? Question?

AUDIENCE: Would it be 0.56 cubed times 0.44 cubed?

SARA ELLISON: Exactly, 0.56 cubed times 0.44 cubed. And how did we get to that? Well, we just take these are independent events. So to get the probability of the intersection between the event that he misses, misses, misses, then makes, makes, makes, we just take the sum of the probabilities of each of those events.

So the probability of miss, miss, miss times, or-- yeah, the product of all those probabilities is equal to 0.56 cubed times 0.44 cubed, or 0.015. Now, keep in mind that the order that I had in this particular problem doesn't matter. I could have asked you what's the probability of any particular sequence of three misses and three shots made. And your answer would have been exactly the same.

So what is the probability that he misses three and makes three of the next six shots he takes? So how do we think about this? So now I'm not asking about the probability of any particular sequence of three misses and three completed shots, but just that there are three misses and three completed shots.

AUDIENCE: Because now we have the combination of those.

SARA ELLISON: Exactly. So we know that we know the probability of each particular sequence. And so now what we have to do is we have to add-- we have to count up the number of such sequences there are. What's the number of such sequences that there are?

AUDIENCE: 6 choose 3.

SARA ELLISON: 6 choose 3. Exactly. So there we go.

Why are there 6 choose 3 sequences of misses and makes in the next six shots he takes? Does anyone want to-- yep?

AUDIENCE: He has six shots to take, and you want to 6 and choose 3 in which he, we'll say, makes them.

SARA ELLISON: That's right. OK. OK, and so if we multiply 6 choose 3 by the probability of any one such sequence, we get 0.30. So that's the probability of any sequence-- or, sorry, that's the probability that he makes the next-- he makes three of the next six shots he takes. Yeah?

AUDIENCE: If there are more than two options or two outcomes, that framework wouldn't work, right?

SARA ELLISON: It gets more complicated. Yes, that's right. OK, final question-- what is the probability that he makes at least one shot in the next six that he takes?

AUDIENCE: 1 minus the probability that he misses all of them.

SARA ELLISON: Yes, 1 minus the probability that he misses all of them. So as I said earlier in the lecture, we could calculate this probability by calculating the probability he makes one shot, calculating the probability he makes two shots, calculating the probability he makes three shots, and adding those all together. That's a perfectly fine way to do it.

There's an easier way, and that's using one of the facts about probability that I had up earlier. And all we have to do is calculate the probability he misses all six of them, which is 0.56 raised to the sixth, and subtract that from 1. And we get the probability. Yes? Question?

AUDIENCE: I have a question. So wouldn't-- for example, let's say he starts on this six-shot penalty [INAUDIBLE]. And he misses the first one. Wouldn't that change his ratings upon how likely he is to make a 3-pointer as a whole?

SARA ELLISON: Well, we're sort of-- if I understand your question correctly, we're sort of ruling that out in this question. So what I said is that we're going to assume that the probability for any particular shot is 0.44.

AUDIENCE: OK, but I realize you have to update in real time, right?

SARA ELLISON: Well, and in particular, in real life, he's going to take higher and lower percentage 3-point shots. And if the shot clock is about to expire, he's likely to take a lower-percentage shot just because that's the last chance he's going to get and so forth. So, yeah, we're kind of abstracting away from that in this particular problem. But, yeah, you're right. OK?

OK. Now, this example is one we will refer back to later on in the semester as the basis for a special distribution, the binomial distribution. OK.

So now on to conditional probability-- so recall that just a minute ago, when we were talking about independence, knowing that two events are independent means that the occurrence or nonoccurrence of one event doesn't tell you anything about the other. So if I know that A occurred, and A and B are independent, then the probability that B is going to occur or did occur doesn't change.

But what if we have two events where the occurrence of one event actually tells us something relevant about the probability of another event? How can we alter the probability of the second event appropriately? How can we, in some sense, update what-- using the information from the first event, how can we sort of update the probability of the second event?

Well, we use this by using something called conditional probability. So the probability of A conditional on B, which we denote this way-- probability of A with a vertical line B-- is just the probability of the intersection of A and B divided by the probability of B. Now, this definition only holds if the probability of B is greater than 0. We can't condition on an event that has 0 probability. So we don't even define it if the probability of B is equal to 0. OK?

So what I-- let me see if I-- yeah. So the way I think about conditional probability-- and I hope this was helpful-- is I think about the numerator of the conditional probability as essentially redefining both the event-- or, sorry, redefining the event as only the part of the event that's now relevant given that we know B occurred.

And then I think about the denominator as redefining the sample space, as only considering the part of the sample space that's now relevant, given that we know B occurred. So it might actually be useful for me to draw a Venn diagram to illustrate this.

OK. So let's say we have a sample space S. We have an event and another event B. And we know that B occurred. So basically, we're going to be conditioning on B. Then the only part of A that's relevant is this part, right? The intersection. Actually, I could have drawn that-- colored that in better. That's the only part of A that's relevant. The only part of the sample space that's relevant-- we know that B occurred-- the only part of the sample space that's relevant is this. OK?

So essentially, the conditional probability formula is basically redefining both the numerator and the denominator based on this new information we have. So as you might suspect, there is a relationship between independence and conditional probability.

So suppose we have two events, A and B, and they're independent. And we have to assume that the probability of B is greater than 0. So we can talk about conditional probability. Then the probability of A conditional on B is just equal to, by definition, the probability of the intersection divided by the probability of B.

But since they're independent, that's equal to-- the probability of the intersection is equal to the product of the probabilities. So it's probability of A times probability of B over probability of B. The probabilities of B cancel, and we get probability of A.

So this is exactly what we would have expected happening. Given our refined intuition about what independence means, independence means B occurs. It doesn't tell us anything about the probability of A. And here we have-- if A and B are independent, the probability of A conditional on B is equal to the probability of A.

And note that this implication goes both ways. If probability of A conditional on B is equal to probability of A, then that also implies independence.

So I go back to the American political system once again for another example. It's such fertile ground these days. So one sort of interesting part, as an observer of the American-- an observer and a participant in the American political system-- I always think it's very interesting that there's this sort of tension between the primaries and the general election.

It's interesting to watch how the candidates are always trying to position themselves to appeal to the party faithful in the primaries. But then, when it comes to the general election, they have to somehow convince the electorate that they're more centrist or something like that. And so there's this sort of interesting tension that's always, I think, fun to watch, or interesting to watch, at least.

So let me use that tension to inspire an example on conditional probability. So let's suppose that we have four candidates with the following probabilities of winning the nomination. So we have the probability of event A1, which is that Donald Trump wins the nomination, is 0.4. The probability of event A2, which is Cruz winning, is 0.3, and so forth.

And then let's suppose that conditional on winning the nomination, these candidates have the following probabilities of winning the general election. So Donald Trump might have the highest probability of winning the Republican nomination, but he doesn't have a particularly high probability of winning the general election, conditional on winning the nomination.

And maybe perhaps-- by the way, I've made up these numbers. I didn't get these numbers from anywhere. But I don't know. They seem at least somewhat plausible to me.

And perhaps Marco Rubio, among this group, has the highest probability of winning the general election, conditional on winning the Republican nomination. So I gave him a 0.6.

ESTHER DUFLO: By the way, I think you could get this number [INAUDIBLE] doing it from the betting markets.

SARA ELLISON: Yes. In fact, Esther is absolutely right. I could have gone to the political betting markets and gotten actual numbers. So, yeah, that is correct.

OK, so this is what I was talking about. The tension is embodied in the fact that the candidates with the higher probability of winning the nomination might not have the higher probability of winning the general election. OK, so how can we compute the probability of a Republican win in the general election?

So we've got all this information. And from this information, we should be able to compute the probability of a Republican win in the general election? And in order to do that, I'm going to do a little side calculation.

So let me first start by saying the probability of a Republican win is just equal to the probability of the intersection between a Republican win and the sample space. That's kind of almost a tautology, but it's going to be helpful for me in the calculation.

And then I can write the sample space as the union between these four events, A1 through A4. And why is that? Because A1 through A4 are mutually exclusive and exhaustive events and therefore form a partition. So the union of all of them is equal to the sample space.

So then I write this as the probability of W intersected with A1 union A2, so forth. Then I can write this-- I can distribute the W out and write it as the probability of W intersected with A1 union W intersected with A2 and so forth.

And since the A1's are mutually exclusive, the intersection of the A1's with anything is also going to be-- that's going to be a mutually exclusive set. So I can write this as the sum of the probabilities. And then, using the definition of conditional probability, I can write each of these as the probability of Republicans winning conditional on A1 times the probability of A1. And this first term I can write this way, just using the definition of conditional probability. And I can do that with all four terms. And so I get this very nice formula.

And we know-- I gave you all of these. I gave you the set of four of these, and I gave you the set of four of those. So all we have to do is plug in to get the probability of a Republican win in the general election. OK?

OK. So I'm actually-- I don't know if this is good news or bad news. I'm actually going faster than I thought I would, and I did not think I would get to Bayes' theorem today. But let's go ahead and go with it.

So we have seen-- oh, by the way, this is the Reverend Thomas Bayes, who was a Presbyterian minister in England or-- Scotland? I'm not sure. Anyhow, he did sort of probability as a hobby. And he came up with this enormously useful result called Bayes' theorem, which we will see throughout the semester and on which an entire branch of statistical analysis is based.

So probably most of you have heard about Bayesian analysis, and this is all-- the entire branch of Bayesian analysis is based on Bayes' theorem and sort of the ideas of Reverend Thomas Bayes. OK.

So we've seen that the probability of the intersection of A and B is equal to the probability of B conditional on A times the probability of A. That's just the definition of conditional probability. Alternatively, we could also write that as a probability of A conditional on B times the probability of B-- again, just the definition of conditional probability.

And this is provided both A and B have positive probabilities. So putting those two together, we can write the probability of A conditional on B as being equal to the probability of B conditional on A times the probability of A divided by the probability of B.

We also saw a slightly more complicated version of this two slides ago. But we also saw that we can write the probability of B as the probability of B conditional on A times the probability of A plus the probability of B conditional on A complement times the probability of A complement. So let me just go back and make sure you understand where that came from.

So that is just simply a simpler version of this with only two events, A and B-- or, sorry, A and A complement instead of A1 through A4.

OK, so we put these two facts together, and we get Bayes' theorem. So the probability of A conditional on B is equal to the probability of B conditional on A times the probability of A divided by that same term plus the probability of B conditional on A complement times the probability of A complement.

So we could do this because A and A complement form a partition of S. You can do this with any partition of S. So basically, you can write down Bayes' theorem, a more complicated version of Bayes' theorem, as long as the thing that's playing the role of A and A complement is a partition of S. So we could have more than just sort of two terms in the denominator. Is that clear? OK?

OK. So let me now go to an example using Bayes' theorem. So why is it that Bayes' theorem is such a powerful idea? Why is it that this algebraic manipulation of conditional probabilities has ended up being so important, such an important part of probability?

Well, I think this example will at least give you some notion of how powerful Bayes' theorem can be. So let's suppose a pregnant woman lives in an area where the Zika virus is fairly rare. So 1 in 1,000 people have it in the area where she lives.

But still, she's concerned. So she gets tested. There's a good test for the Zika virus, suppose, but it's not a perfect test. And it gives a positive leading with probability 0.99 if the person has the virus and a positive leading with probability 0.05 if the person does not have the virus.

So we've got, in the world of medical testing, that would be considered a pretty reliable test. There are some false positives, some false negatives. But the rate of false negatives is 1%. The rate of false positives is 5%. So it's a pretty reliable test.

So she goes in and has the test, and her reading is positive. How concerned should she be now?

So you have-- you think you might have a virus. You take a test that's really quite reliable, and it comes out positive. So what do you imagine the probability that she actually has the virus is? Does anyone want to take a guess?

AUDIENCE: She shouldn't be that worried because I think the objective probability of her is very, very low.

SARA ELLISON: Yeah. So the answer was she shouldn't be that worried. I'm sure you can't do the calculation in your head. But in fact, even updating on the information in the test, the probability that she has the virus is pretty low because the initial probability was only 1 in 1,000, and the test is not perfect.

So let me go through the specifics. So we can use Bayes' theorem to calculate the probability that she actually has the virus conditional on her positive test. The probability-- we'll call this the prior probability, or the unconditional probability of her having the Zika virus-- is 0.001. That's because she lives in an area where not that many people have it.

That means also that we can write down the probability of the complement of Z as 0.99. The probability of getting a positive test result, conditional on having the Zika virus, is 0.99. So there's only 1% of false negatives on the test.

And the probability of getting a positive if you don't have the Zika virus is 0.05. So we plug all of these into Bayes' theorem, and we get that the probability that she has the virus is still less than 2%.

And basically, I think some people probably find this result surprising. And maybe I can give you some intuition to help you understand how this can be. So the unconditional or prior probability of having the virus was quite low, 1 in 1,000. We updated the probability based on the results of this test, but the test was imperfect.

So if you think about two counterfactuals, one is that the test was perfect-- no, wait, how do I want to say this? Yeah, the test was perfect. The probability of her testing positive would be 1 in 1,000.

And the other counterfactual is if the test was as stated, but there was no Zika virus, the probability of her testing positive would be 50 out of 1,000. And so basically, we're updating based on valuable but not perfect information. And the probability doesn't get updated that much as a result. OK? Yep?

AUDIENCE: What if you have more than one variable [INAUDIBLE] than just one [INAUDIBLE]?

SARA ELLISON: More than one set depending on each other?

AUDIENCE: For example, so here you have it that-- the condition of the test being positive. What if you have another condition?

SARA ELLISON: Yes, yes, yes. So I think what you're asking is that you have several pieces of information, not just the test. You have whether she has symptoms, for instance.

AUDIENCE: What if she has a fever?

SARA ELLISON: Exactly, or you have-- yeah.

AUDIENCE: [INAUDIBLE]

SARA ELLISON: Yeah, exactly. So you may have a whole series of different pieces of information. And if we can write down-- if we can write down the probabilities that she has the Zika virus conditional on-- sorry, if we can write down the probabilities of her having a fever conditional on having Zika virus, the probability of her having a fever conditional on her not having the Zika virus, then we can also update the probabilities based on that information.

And there are basically two ways we can do it. We can do it sequentially, or we could write down a formula that lets us update based on all that information simultaneously. And we'll get the same answers. So-- OK, I think now would be a good place to stop, even though we're a few minutes early. The next topic we're going to cover is random variables. And we will do that on Wednesday.