PROFESSOR: Last time we defined the expected value of a random variable. And we talked about a lot of ways it could be computed. We proved all sorts of equivalent definitions.

Today, we're going to keep talking about expectation. And we're going to start with an example that talks about the expected number of events that you expect to have occur. And it's a generalization of what we did with Chinese appetizer and hat check from last time.

We're going to call this theorem 1. If you have a probability space, s, and you've got a collection of n events, let's call them A1 through A n, and these are just subsets of s, then the expected number of events to occur, of these events, is simply the sum of i equals 1 to n of the probability of the i-th event occurring. So you just sum up the probabilities that the events occur. And that tells you the expected number of events that will occur. So a very simple formula.

So for example, A i might be the event if the i-th man gets the right hat back from last time. Or it could be the event if the i-th person gets the right appetizer back at the Chinese restaurant after we spin the wheel in the center. So we're going to prove this. And the proof is very similar to what we did last time when we figured out the expected number of people to get the right hat back.

In particular, we're going to start by setting up an indicator variable, T sub i, that tells us whether or not the i-th event, A sub i occurs. So we define T sub i-- and it's a function of a sample point-- to be 1 if the sample point is in the i-th event, meaning the i-th event occurs, and 0 otherwise. And this is just another way of saying that T sub i is 1 if and only if A sub i happens or occurs.

Now what we care about is the number of events that occur. And we get that just by summing up the T sub i. So we'll let T be T1 plus T2 plus T n. And that'll count because we'll get a 1 every time an event occurs. By adding those up, we'll get the number of events that occur.

All right. Now we care about the expected value of T, the expected number of events to occur. And I claim that's just the sum i equals 1 to n of the expected value of T i. Why is that true? Why is the expected value of T the sum of the expected values of the T sub i?

Linearity of expectations. Now did we need the T sub i's to be independent events for that? No. OK, very good.

Now the expected value of T i is really easy to evaluate. It's just a 0, 1 variable. So it's just the probability that T i is 1 because it's 1 times this plus 0 times the probability of 0, and that cancels out. And the event that T i equals 1 is just the situation where the i-th event occurs because T i equals 1 is the case that A i occurs. That's what it is.

So we've shown that the expected number of events to occur is simply the sum of the probabilities that the events occur. So a very simple formula, very handy. And you don't need independence for that. Any questions about that? We're going to use that theorem a lot today.

As a simple example, suppose we flip n fair coins. And we let A i be the event that the i-th coin is heads. And suppose we want to know the expected number of heads in the n flips. Well, we can use this theorem. The expected number of heads is just going to be the sum of the probabilities that each coin is a heads.

So let's do that. T is the number of heads. We want to know the expected value of T. And from theorem one, that's just the probability the first coin is heads plus the probability the second coin is heads. And the same out to the probability the n-th coin is heads.

What's the probability the first coin is heads? And 1/2. The probability the second coin is heads is a 1/2. The probability the last coin is heads is 1/2. And so the expected number of heads-- we add up 1/2 n times-- is just n/2.

Of course, you all knew that. If you flip n fair coins, the expected number of heads is half of them. But that's a very simple way to prove it.

Did we need the coin tosses to be mutually independent to conclude that? No. I could've glued them together in some weird way. In fact, I could have glued some face up and some face down and done weird things, and you still expect n/2 heads even if they were glued together in strange ways. Because I don't need independence for linearity of expectation to prove this.

Now that's the easy way to evaluate the expected number of heads. There is a hard way to do it. Let me set that up.

We could start from the definition, a different definition, namely, that the expected value of T is the sum from i equals 0 to n. i times the probability that you have i heads. This would be a natural way to compute the expected number of heads. You add up the case where there's zero heads times the probability of 0, 1 times the probability of one head, 2 times probably two heads, and so forth. That's one of the first definitions of expectation.

So let's keep trying to do this. What is the probability of getting i heads? And now I'm going to have to assume mutual independence actually. Now I'm going to need mutual independence.

So already, this method isn't as good because I had to make that assumption to answer this question. If you don't make that assumption, you can't answer that question. What's the probability of getting i heads out of n. Yeah?

AUDIENCE: s to the n-th power times n [INAUDIBLE] i.

PROFESSOR: Yes, because if you look at the sample space, there are 2 to the n sample points all equally likely. They're all probability 2 the minus n. And there's n choose i of them that have i heads.

And now you'd have to evaluate that sum which is sort of a pain. That one won't come to mind readily. So you might say I reached sort of a dead end here. But in fact, the answer is easy to use, easy to get using this method.

In fact, we've actually proved an identity here because we know the answer is n/2. We've just proved that this messy thing is n/2. In fact, you can multiply by 2 to the n here. We have proved, using probability theory and theorem 1 over there, the sum of i n choose i equals n 2 to the n minus 1. Just multiply by 2 to the n on each side.

So we've given a probability based proof of this identity which is sort of hard to do otherwise, could be harder to do. Any questions about that? So again, if it comes time for a homework problem or a test problem, if it naturally divides up in this way where you can take a random variable when you got a computer's expectation to make it a sum of indicator variables, if that is a natural thing to do, do it that way. Because it's so much easier than trying to go from the definition because you might encounter nasty things like that that you got to evaluate.

So in this case, we flipped n coins. We expect n/2 heads. In the hat check, in Chinese appetizer problems-- we had n hats or n appetizers-- we expected to get one back to the right person-- so a smaller expected value.

For some problems, the expected value is even less. The expected number of events to occur is less than 1. In fact, it might be much less than 1.

Now in those cases it turns out that the expected value is an upper bound on the probability that one or more events occur. We're going to state this as theorem 2. The probability that at least one event occurs is always upper bounded by the expected number of events to occur. Now this theorem is pretty useless if the expected number of events to occur is bigger than 1 because all probabilities are at most 1. But if the expected number of events to occur is small, something much less than 1, this is a pretty useful bound.

So let's prove that. The expected value of T-- and in this case we'll use one of the definitions we have of expected value, name of the one where you sum from i equals 1 to infinity of the probability T is greater than or equal to i.

Now what did we have to know about T to use that definition? That doesn't work for all random variables T. What condition do I have on T to be able to use this one? Anybody remember? Yeah.

AUDIENCE: Must be defined on the natural numbers?

PROFESSOR: T must defined on the natural numbers. If it is, I can use this very simple definition. And is T defined on the natural numbers here? Is the range of T natural numbers?

Well, I'm counting the number of events that occurred. Could be 0, 1, 2, 3, 4. Has to be a natural number. So it's OK. So I can use this definition.

Now this is summing up probability T is at least 1 plus probability T is at least 2 and so forth. I'm just going to use the first term. This is at least the size of the first term because probabilities are non-negative. So I'm just going to throw away all the terms after the first and conclude that this is at least the probability T is bigger than or equal to 1. And I'm done.

I just look at it in reverse. The probability of at least one event occurring is at most the expected value. Very simple. There's a very quick corollary here.

The probability at least one event occurs is at most the sum of the probabilities of the events.

And the proof there is just plugging in theorem 1. Because the expected value is the sum of the probabilities.

So we just plug-in theorem 1 for the expected value because it's just that. Any questions about the proof? Very simple.

Now this theorem is very useful in situations where you're trying to upper bound the probability of some kind of disaster or something bad happening. For example, suppose you want to compute the probability that a nuclear plant melts down. Now actually, the government does this. They got to figure this thing out because if it's a high probability, well, we're not going to allow anybody to build them.

And the way they do it is they convene a panel of experts. They get some people from various good universities and they bring them down to Washington. And they have them figure out every way they can think of that a meltdown could occur, every possible event that would lead to a meltdown. And then they'd have them figure out the probability for each one of those events.

For example, A1 could be the event that the operator goes crazy and makes it meltdown. A2 could be the event that an earthquake hits and the cooling pipes are ruptured, and then you got a meltdown. A3 is the event that terrorists shoot their way in and cause a meltdown.

So you've got a lot of possibilities for how the thing can melt down. And then they compute the probabilities. And then they add them up just using this result. And they say, well, the probability that a meltdown-causing event or one or more occurs is at most this small number. And hopefully it's small.

So for example, suppose there's 100 ways that a meltdown could occur, 100 things that could cause a meltdown. And each one happens with probability one in a million. What can you say about the probability that a meltdown occurs?

You got a hundred ways it could happen only. Each is a one in a million chance. What's the probability a meltdown occurs?

1 in 10,000 because you're adding up one in a million 100 times. n is 100. Each of these is one in a million.

So you get 100 over a million. There's 1 in 10,000. And so then they publish a report that says the chance of this reactor melting down is 1 in 10,000.

Now what if I've got 100 reactors? What's the chance that at least one of them melts down? 1 in 100 because I got 100 over 10,000-- same theorem. So there's a 1 in 100 chance something melts down somewhere, at most. Hopefully, the numbers are better than that.

Same thing if you bought 100 lottery tickets, each a one in a million chance, you got a 1 in 10,000 chance of winning, at most. So simple fact but powerful and used a lot in practice. And this is sort of the good case when the expected number of events that are bad to happen is small, like a lot less than 1.

But what if the expected number of events to happen is big. Say it's 10. Say this government panel gets together and they add up all the probabilities and it comes out to be 10.

Well, it doesn't sound so good if that's the case. But is it necessarily bad? Does it necessarily mean that you're going to have a meltdown.

So for example, let's say there's 1,000 ways you could melt down. And let's say that the probability of each one is 1 in 100. So the expected number of things that could happen to cause a meltdown is 10.

Am I guaranteed we're going to melt down? No. Can anybody think of a way where it's unlikely we're going to melt down but respect these values here, hypothetically? Is there any chance that it's still unlikely to have a meltdown? We're going to think of a way? Yeah.

AUDIENCE: They all happen at the same time.

PROFESSOR: They all happen at the same time. Now the examples I gave you-- the terrorists, the earthquake, and the crazy operator-- put that on the side. If they all happen together, when any one happens, the others have to happen. So we can express that as for all ij, probability of the i-th event happening given the j-th event happening is one, so total dependence.

What's the probability of a meltdown in that scenario? What's the probability one of those meltdown-inducing events occurs? They all happen at once.

AUDIENCE: 1 in 100.

PROFESSOR: 1 in 100. Because it's the same as the probability of the first event happening which, by definition, was 1 in 100.

So it could be that the probability of a meltdown is small. But it might not be as well. There's no way, given this, to know.

What if I chain-- in fact, this is like Chinese appetizer, right? If one person gets their appetizer back, everybody does. So there are circumstances where you have the total dependence like that.

Let's say I change a little bit and I don't do this scenario. In fact, say I tell you the events are mutually independent, but you expect 10 to occur. Do you sleep at night now? Of course, 1% is still a pretty high number.

But how many people think that if they're mutually independent and you expect 10 that, no matter what, there's a least a 50% chance of a meltdown? Anybody? OK. In fact, if you expect 10 and they're mutually independent, a meltdown is a virtual certainty. The chance you don't melt down is less than 1 in 22,000.

For sure something will occur that's bad. And this is a theorem that we call Murphy's law. And Murphy's law, probably you've all heard of it, it says-- it's a famous saying. If something can go wrong, it will go wrong. And we're going to see why that's true, at least, in our circumstances here.

That's a pretty powerful theorem.

If you have mutually independent events A1 through A n, then the probability that none of them occur, t equals 0, is upper bounded by e to the minus expected number of events to occur. So if I expect 10 to occur, the chance that none do is upper bounded by e to the minus 10, which is very small, which means almost surely one of the events or more will occur. And that's bad news in this case. So let's prove that.

Well, the probability that t equals 0 is the same as the probability that A1 does not occur and A2 does not occur and all the way to A n does not occur. And I claim this [INAUDIBLE] is the product of the probabilities they don't occur. So I'm taking the product i equals 1 to n of the probability that A i does not occur.

Now why can I make that step? Yeah. Independence. This is the product rule for independent events.

Now the probability that A i does not occur is simply 1 minus the probability it does occur. And now I'm going to use a simple fact, which is that for any number x, 1 minus x is at most e to the minus x. Just a simple fact from algebra.

So I've got 1 minus-- I'm going to treat this as the x. So this is at most e to the minus probability of A i using that fact. And now I'll take the product and put it into a sum in the exponent.

And then some of the probabilities of the events is just the expected value. That was theorem 1 that I just erased. So this is e to the minus expected number of events to occur.

So not too hard a proof. We had to use that fact. But that gets the expected number of events to occur in the exponent. So a simple corollary is the case when we expect 10 events to occur.

So if we expect 10 or more mutually independent events to occur, the probability that no event occurs is at most e to the minus 10, which is less than 1 over 22,000. Now there's not even any dependence on n here. It had nothing to do with a number of possible events, just that if you expect 10 of them to occur, you're pretty sure one of them will.

And this explains why you see weird coincidences. Or people sometimes see what they think are miracles. Because out in the real world, there's billions of possible weird things that could happen. You just can create all sorts of crazy possibilities.

And each one might be one in a billion chance of actually happening. But you got billions that could've. And if they're all mutually independent-- because you made up all these different things-- than you expect some of them to happen. And so you should-- in fact, you're going to know that for sure some of those weird things are going to happen. At least the chance that no weird thing happens is 1 in 22,000.

And so this can be why somebody will go along and say, oh my goodness. You won't believe what happened, a coincidence. And it's like, wow, the chance of that happening was one in a billion. It must've been a miracle or an act of God that this happened. But you're not thinking about the other 10 billion things that didn't happen. So for sure some of those things are going to happen.

It's not likely that I'm going to win megabucks next week. But somebody's going to win. If enough people play and it's more than 1 over the probability that you're going to win, than it's very likely somebody will win if everybody is guessing randomly. Any questions about what we're doing?

So this is amazingly powerful, this result. In fact, it's so powerful that it's going to let me read somebody's mind in the class. We're going to do a little card trick here.

Now the way this card trick works-- it's a little complicated. I'm going to need a volunteer, probably one of you guys down front. We'll get you. And one of your buddies is going to keep you honest for me here.

I'm going to reveal-- first I'm going to let you shuffle the deck. So go ahead and shuffle it, do whatever you want. It's a normal deck. It's got 52 cards and two jokers. And I don't care what order they're in.

I'm going to turn over the cards one at a time. Now I'm going to ask you to pick a number from 1 to 9 ahead of time. Don't tell me or anybody else. In fact, I'm going to want you guys to play along too. And we're going to see where we all end up here.

And that's your starting number. And as I turn over the cards one at a time-- say you started with a 3 was the number you had in mind-- on the third card I show, that becomes your card. You don't tell me or jump up and down or anything. But that's your card.

And say it's a 4 of diamonds. Now a 4 replaces the three in your mind and you count 4 more cards, then that becomes your card. Now let's say that's a jack or a face card or a 10 or a joker. 10, face card, and jokers all count as 1 just like an ace counts as 1. And so then the next card would be your card because you count 1.

And we keep on going until you have a card, maybe it's a 7. But there's only four cards left in the deck. And so you don't get a new one. And your last card is the 7.

And then you're going to write that down here, not showing me. And you're going to do this, maybe do this with a friend over there. And you're going to make sure you count right on the deck because if you screw up the counting, it's going to be hard for me to read your mind.

So just to make sure we all understand this, let me write the rules down here because I want the whole class to pick a number from 1 to 9 and play the same game. And we're going to see what happens. So let me show you the rules again just to make sure everybody understands.

So say the deck starts out like this. I got a 4, a 5. So my first few cards of the deck go like this.

10 equals a 1. Then I got a queen equals a 1. 3, 7, 6, 4, 2. Say it's a small deck. I'm going to use 54 cards.

And say you're chosen number to start, you start with a three. As I show the cards, you're going to count 1, 2, 3. That becomes your new card.

Then you're going to count 1, 2. That becomes your card. It's a 10, so you convert it to a 1 because we're only doing single digit numbers.

Go to 1, that becomes your card. Queen converts to a 1. You go 1, that becomes your card. 3, 1, 2, 3. That becomes your card.

And you can't get 4, so you remember the final card. Does everybody understand what you're supposed to do? Because we're going to do 54 cards of this.

Maybe we get the TAs to play along here. And as you do it, maybe you want to talk to your buddy, make sure you got it worked out there. And if I could read your mind maybe we'll have a gift certificate or something.

So you shuffle the deck? Got it good? All right.

So I'm going to start revealing the cards one at a time. So you guys play along quietly in your mind. And we'll see if we can concentrate long enough.

Aces are 1.

Jacks are 1.

10's are 1.

10's are 1.

We're halfway done.

Jokers are 1.

OK. That's the last card. So remember the last one that was yours.

And you got to go check with your buddy to make sure you guys agree on the counting there. And then write it down. Don't tell me because I'm going to read your mind. I'm going to tell you.

This is not good. They're arguing over the last card. I'll have to read one of your minds.

What's that? The 11 of clubs. That's hard one to predict.

Make your best guess and write it down. Don't tell me. Write it down.

You got two? Well, write them both. I'll predict one of them.

I've never had a dispute on what the-- because if you started with the same position, you've got to wind up in the same position. You wrote it down? Now think about your number really hard.

We'll take yours. I'll trust you there. Think about it really hard because I need the brain waves to come over and read the mind here.

Yeah, yeah. I'm getting a really strong signal on the last card. Maybe I don't know. Maybe it's something-- it's really powerful.

I'm going to say it's the queen of hearts. Is that right? Both were the queen of-- oh, you were trying to screw me up, mess with me. So you both got the queen of hearts. Oh, you did.

So how many people got the queen of hearts? Oh wow. How many people did not get the queen of hearts. Somebody. OK.

Now there's a chance you did it legitimately. But usually, with a deck, we're all going to get the same last card. Now in this case, it happened to be the very last card. That is typically not the case. So very good.

So I read your mind. So you guys get the gift certificates here. Very good. One for you and your sponsor there.

So it's clear how I read his mind because I got the same number everybody did. And somehow it doesn't matter where we started. We all had the same card at the end.

How is that possible? There's nine different starting points. Why don't we wind up in nine different places?

And why isn't there a one in nine chance that I guess his card? Why do we all wind up in the same place? Any thoughts? Yeah?

AUDIENCE: Get the same card. After that you stay [INAUDIBLE].

PROFESSOR: That's right. If ever we had the same card, then we're going to track forever and finish in the same card. But why should we ever get the same card? What are the chances of that, that we land on the same card?

Why don't we just keep missing each other? It's a 1 in 9 chance or something. I don't know. Why don't we keep missing?

AUDIENCE: It seems like there are enough low cards that you just move slowly along and, eventually, you're going to intersect.

PROFESSOR: Yeah. I did make a lot of 1's in the deck. If I would've made all these face cards be 10's, the chances of my reading your mind go down.

Why do they go down? What does it have to do? Why did I put a lot of 1's in the deck?

AUDIENCE: It goes on longer.

PROFESSOR: What's that?

AUDIENCE: The game goes on longer?

PROFESSOR: The game goes on longer. So there's more chances to hit together. Because at any given time-- you've got your card. I've got mine.

If mine is behind you, I got a chance to land on you. And if you're behind me in the deck, you've got a chance to land on me with your number. And if the numbers are smaller, there's more chances to land on each other.

And it is true that on any given chance, the chances are low that we land on the same card.

But there's a lot of chances. And if there's a lot more chances than the probability of landing on each other, we've got Murphy's law. If you've got a lot of chances and they're not less likely than the number of chances or the inverse of that, then we expect to have a certain bunch of times that we're going to land on each other. And therefore, a very high probability we do.

Now that was a little hand-wavy. And in fact, there's a reason it was hand-wavy. Why doesn't Murphy's law really apply in this case, really mathematically apply? Yeah.

AUDIENCE: They're not mutually independent. Once you draw one card, it's not coming back.

PROFESSOR: That's correct. And it means that the knowledge that we haven't collided yet tells me something about the cards we've seen-- not a lot, but something maybe. And it's a finite deck which tells me something about the cards that are coming.

And it might influence the probability that we land together, the next person who's jumping on the deck. And so the events of-- like for example, in this case, we let A i be the event of a collision on the i-th jump. And there's about 20 jumps in this game, 10 for each of us expected. So A i is the event that we collide on the i-th jump. These events are not necessarily mutually independent.

Now if I had an infinite deck or a deck with replacement so every card is equally likely to come next no matter what's come in the past, now you can start getting some mutual independence here. And then you could start really applying the theorem. Now in this case, you don't expect 10 things to happen. You expect a few.

But that's good enough that, in fact-- so we did a computer simulation once and I got about a 90% chance that we'll all be on the same card. So I have a pretty good chance that I'm going to guess right. And so far, I haven't guessed wrong. But it will happen some day that we'll start with a different first number, and we will miss at the end because they'll be two possible outcomes. Just the way it works out with 52 cards.

Now of course, if we have more cards or I made more things be 1's instead of 9's, say, my odds go up because the number of events I've got, the number of chances to collide, increases. And the chance of hitting when I jump also increases. Any questions on that game?

So the point of all this is that if the expected number of events to occur is small, then it's an upper bound on the probability that something happens, whether they're independent or not. If the expected number of events to occur is bigger than 1, large, and if the events are mutually independent, then you can be sure one of those events is going to occur-- very, very likely one of them will occur. And that's Murphy's law. Any questions about numbers of events to occur?

We'll talk more about the probability in the numbers of events that occur next time. Before we do that, I want to talk about some more useful facts about expectation. Now we know from linearity of expectation that the expected value of a sum of random variables is the sum of the expected values of the random variables.

Now we're going to look at the expected value of a product of random variables. And it turns out there's a very nice rule for that. Theorem 4, and it's the product rule for expectation. And it says that for-- if your random variables are independent, R1 and R2 are independent, then the expected value of their product, also a random variable, is simply the product of the expected values.

So it's sort of the equivalent thing to linearity of expectation, except we're doing products. And you need independence. Now the proof of this is not too hard, and it's in the book. So we're not going to do it in class. But we can give an example.

Say we roll two six-sided fair and independent dice. And I want to know what's the expected product of the dice.

So we're going to let R1 be the value on the first die, and R2 would be the value on the second one. And the expected value of the product is the product of the expectations. And we already know the expected value of a single die is 7/2. So we get 7/2 times 7/2 is 49/4 or 12 and 1/4.

So it's easy to compute the expected product of two dice. Any questions about that? Much easier than looking at all 36 outcomes to use this rule.

Now what if the dice we're rigged, glued together somehow so they always came up the same? Would the expected product B 12 and 1/4 then? No? Why not?

Why wouldn't it be the case? Why isn't the expected value of R1 squared the square of the expected value of R1? Isn't that what this says?

AUDIENCE: Independent.

PROFESSOR: They're not independent. R1 is not independent of R1 In fact, it's the same thing. And you need independence for that to be the case.

So a non example, the expected value of R1 times R1 is the expected value of R1 squared. And to do that, we got to go back to the basics. We're taking the six possible values of R1.

i equals 1 to 6. i squared, because we're squaring it, times the probability R1 equals i. And each of those probabilities is 1/6. So we get 1/6 times 1 plus 4 plus 9 plus 16 plus 25 plus 36.

And if you add all that up you get 15 and 1/6, which is not 3 and 1/2 squared, which is the expected value of R1 squared. So the expected value of the square is not necessarily the square of the expectation. Because a random variable is not independent of itself generally.

OK. Any questions there? There's a couple of quick corollaries.

The first is you take this rule and apply it to many random variables as long as they're mutually dependent. So if R1 R2 out to R n are mutually independent, then the expected value of their product is the product of the expected values. And the proof is just by induction on the number of random variables. So that's pretty easy.

There's another easy corollary. And that says, for any constants, constant values, a and b, and any random variable R, the expected value of a times R plus b is simply a times the expected value of R plus b. And the reason that's true-- well, the sum works because of linearity of expectation for the sum. You can think of b as a random variable that just always has the value b.

And the a comes out in front because you can think of it as a random variable that always has a value a. And that's independent of any other random variable because it never changes. So by the product rule, the a can come out.

Now you've got to prove all that. But it's not too hard and not especially interesting. So we won't do that here. Any questions about those?

So we've got a rule for computing the sum of random variables, a rule for the product of random variables. What about a rule for the ratio of random variables? Let's look at that.

So is this the corollary? In fact, let's take the inverse of random variable. Is the expected value of 1/R equal to 1 over the expected value of R for any random variable R? Some folks saying yes. Some saying no.

What do you think? Is that true? Oh, got a mix. How many say yes? How many say no?

Oh, more no's. Somebody tell me why that's not true. Who would like to give me an example? Give us an example there that'll be very convincing. Yeah?

AUDIENCE: I don't think it's one that would be immediately obvious, but I think if R is the result of the roll of a die, I don't think it works out.

PROFESSOR: So it's 50 chance of-- oh, I see. So I take the average of 1/i-- that's sort of hard to compute. I got to do [INAUDIBLE] the sixth harmonic number and then invert it. There's an easier way to show that this is false. Yeah?

AUDIENCE: The expected value equals 0?

PROFESSOR: Yeah. The expected value equals 0 which could happen if R is plus 1 or minus 1 equally likely. So here's an example here. So R equals 1 with probability 1/2, and minus 1 with probability 1/2. So the expected value of R is 0.

So that blows up. That's infinity. What's the expected value of 1/R?

Well, 1/1 and 1 over minus 1, it's the same. It equals 0. And this would say 0 equals 1/0.

That's not true. So this is false. It is not true for every random variable.

So once you see this example, just obviously not true. In fact, there's very few random variables for which this is true, even an indicator random variable. So it's 1 with probability 1/2 and 0 with probability 1/2.

Then the expected value of 1/R is infinite. 1 over the expected value of R is 2. So it's clearly not true. Let's do another one.

What about this potential corollary? Given independent random variables R and T, if the expected value or R/T is bigger than 1, then the expected value of R is bigger than the expected value of T. And let me even give you a potential proof of this, see if you like this proof.

Well, let's assume the expected value of R/T is bigger than 1. And let's multiply both sides by the expected value of T. And well, the product rule says that this is just the expected value of R/T times T, which is just the expected value of R because the T's cancel.

So I gave you a proof. Anybody have any quibbles with this proof? Yeah?

That's a big problem. R/T is not independent of T. [INAUDIBLE] if T is very big, likely that R/T is small.

So we can't do that step. We can't use the independence here to go from here to here. That's wrong.

There's actually another big problem with this proof. Anybody see another problem? Yeah?

Yeah. If the expected value of T is negative, I would end up doing that. So that step's wrong.

So this is a pretty lousy proof. Every step is wrong. So this is not a good one to use.

And in fact, the theorem is wrong. Not only is the proof wrong, but the result is wrong. It's not true. And we can see examples. We'll do some examples in a minute.

Now the amazing thing is that despite the fact that this is just blatantly wrong, it is used all the time in research papers. And let me give you a famous example. This is a case of, actually, a pretty well-known paper written by some very famous computer science professors at Berkeley. And let me show you what they did. And this is so that you will never do this.

They were trying to compare two instruction sets way back in the day. And they were comparing the RISC architecture to something called the Z8002. And they were proponents of RISC. And they were using this to prove that it was a better way to do things.

So they had a bunch of benchmark problems, probably 20 or so in the paper. And I'm not going to do all 20 here. I'm going to give you a flavor for what the data showed. And then they looked at the code size for RISC and the other guys, Z8002. And then they took the ratio.

So the first problem was called E-string search, whatever that is. But it was some benchmark problem out there at the time. And the code length on RISC was 150 say-- I've changed these numbers a little bit to make them simpler. Code length here and the Z8002 is 120.

The ratio is 0.8. So for this problem, you're trying to get low, short code. So this was a better way to go to support that.

And they had something called F-bit test. And here you have 120 lines. Here's 180. So in this case, risk is better. So the ratio of that way to this way would be 1.5.

And they had computing and Ackermann function. And that was 150 and 300. So a big win for RISC, ratio of 2.

And then they had a thing called recursive sorting problem. This is a hard problem. There's 2,800 lines on RISC. 1400 on the old way. Ratio of 0.5.

And there was a bunch more which I'm not going to go through. But their analysis, what they did is they took the ratio, and then they averaged it. And so when you do this, you get an average of, well, 2.3, 4.3, 4.8/4 is 1.2.

So the conclusion is that on average code in this framework is 20% longer than the code on RISC. Therefore, clearly, RISC is a better way to go. Your code on average will be shorter. Using the Z8002 on average, the code will be 20% longer. Makes perfect sense, right?

In fact, this is one of the most common things that is done when people are comparing two systems. Now just one problem with this approach, and that's that it's completely bogus, completely bogus. You cannot conclude-- let's make this conclusion. So their conclusion, they concluded that Z8002 programs are 20% longer on average.

Everybody understands the reasoning why, right? Take the ratio of all the test cases, average them up. Then you get the average ratio.

Now there could be some hint why this is bogus. If I just looked at-- I took and summed these numbers, if I add all those numbers up, I get 3,220. And all these, I get 2,000. RISC code is not looking shorter if I do that. Looking longer.

But all that gain, all the loss of RISC is in this one problem. And maybe it's not fair to do that. And that's why when people have data like this, they just take the ratios. Because now it would be-- if I just took the average code length and took the ratio of that, it's not fair because one problem just wiped out the whole thing.

I might as well not even do it. And they want every problem to count equally. And that's why they take the ratio, to make them all count equally.

Let's do one more thing here. Let's look at what happens if we take the ratio of RISC to the Z8002. Make some room for that. So this is-- this column is Z8002 over RISC.

What if I just did this-- RISC over the Z8002 I mean the answer should come out to 1/1.2, right? That's what we expect, because I've just been turning it upside down. Well, I get 1.25 here. These are just being inverted.

Here I've got 2/3, 0.67. Here I get 1/2. Here I get 2.

Let's add those up. I get 1.92, 2.42, 4.42. Divide by 4-- wow, I get 1.1 something which says, well, that on average, RISC is 10% longer than the other one.

So same analysis says that RISC programs are 10% longer on average. Now the beauty of this method is you can make any conclusion you want seem reasonable, typically. You could have the exact same data. And if you want RISC to look better you do it this way. If you want RISC to look worse, you do it that way.

You see? Is that possible? Is it possible for one to be 20% longer than the other on average, but the other be 10% longer on average?

How many people think that's possible? We had some weird things happen in this class, but that's not possible. That can't happen. These conclusions are both bogus.

Now I'm not teaching you this so that later when you're doing your PhD thesis and it's down to the wire and you need a conclusion to be proved, good news. You can prove it. No matter what your conclusion is, you can prove it.

That's not why we're doing this. We're doing this so you can spot the flaw in this whole setup and that you'll never do this. And you'll see it when other people do because people do it all the time.

So let's try to put some formality under this in terms of probability. Because when you start talking about averages, really think about expectations of random variables and stuff. So let's try to view this as a probability problem and see if we can shed some light on what's going on here because it sure seemed reasonable.

So let's let x be the benchmark. And maybe that'll be something in the sample space, an outcome in the sample space. Let's let R x be the code length for RISC on x and Z x be the code length for the other processor on x. And then we'll define a probability of seeing x.

That's our problem we're looking at. And typically, you might assume that it's uniform, the distribution there. We need this to be able to define an expected value for R and for Z.

Now what they're doing in the paper, what really is happening here, is instead of this, they have the expected value of Z/R is 1.2. That is what they can conclude. That does not mean that the expected value of Z is 1.2, the expected value of R, which is what they conclude that the Z8002 code is 20% longer than RISC code. This is true. That is not implied.

And why not? That's just, actually, what this corollary was doing. Really, it's just what we were-- they made the same false assumption as happened in the corollary.

You can't multiply both sides here by the expected value of R and then get the expected value Z. Of course, if you ask them, they would have known that. But they don't even think through that, they just used the standard method of taking the expected value of a ratio.

So this is fair to conclude. But as we saw, the expected value of R/Z was 1.1. So both of these can be true at the same time. That's fine. But you can't make the conclusions that they tried to make.

Here's another-- in fact, in this case, if we had a uniform distribution, the expected value of R is like 805 for uniform. And the expected value of Z is 500. And that's all you can conclude if you're taking uniform distribution. In which case, of course, if they're promoting RISC, well, you don't like that conclusion.

So it's better to get this one's. I don't think it was intentional of course. But it's nice that it came out that way.

Here's another example that really makes it painfully clear why you never want to do this. So a really simple case, just generic variables R and Z. And I got two problems only-- problem one, problem two.

R is 2 for problem 1, and z is 1. And they reverse on problem 2. Z/R is 2 and 1/2. R/Z, just the reverse.

Now the expected value of R/Z here is 2 plus 1/2 divided by 2 is 1 and 1/4. And what's the expected value of Z/R? The average of these is 1 and 1/4, what's the average of these? Same thing, 1 and 1/4.

So never, ever take averages of ratios without really knowing what you're doing. Any questions? Yeah.

AUDIENCE: What would be the word explanation of the expected value of Z/R? What is that?

PROFESSOR: That is the average of the ratio. It is not the ratio of the average. They are very different things.

And you can see how you get caught up in that. You could see how you have linearity of expectation, you got the product rule for expectation. You do not have a rule that says this implies that.

AUDIENCE: [INAUDIBLE]

PROFESSOR: Which two?

AUDIENCE: [INAUDIBLE] Z/R?

PROFESSOR: Well, in this case, they're one. I don't think that'll be true in general.

AUDIENCE: Does that give you information?

PROFESSOR: They give you information. That may not be the information you want. It wouldn't imply that which is what you're after in some sense. But it gives you some information. It's the expected average ratio.

The problem is the human brain goes right from there to here. It's just you do. It's hard to help yourself from doing it. And it's not true. That's the problem.

We have a version of this in the one in the homework questions which is true. But it's a special version of it where you can say something positive. Any questions about this?

So anybody ever shows you an average of ratios, you want the light to go off and say, danger, danger. Think what's happening here. Or if you're ever analyzing data to compare two systems.

So we talked a lot about expectation, seen a lot of ways of computing it. We've done a lot of examples. For the rest of today and for next time, we're going to talk about deviations from the expected value.

Now for some random variables, they are very likely to take on values that are near their expectation. For example, if I flip 100 coins. And say they're fair and mutually independent. We know that the expected number of heads is 50.

Does anybody remember the probability of getting far from that, namely having 25 or fewer heads or 75 or more heads? Remember, we did that? It was a couple of weeks ago? Is it likely to have 25 or fewer heads?

AUDIENCE: It's less than 1 in a million?

PROFESSOR: Less than 1 in a million. Yeah. It was 1 in 5 million or something. I don't know, some horribly small number.

So if I flip 100 coins, I expect to get 50 heads. And I'm very likely to get close to 50 heads. I'm not going to be 25 off.

And then the example we had in recitation, you got a noisy channel, and you expect an error rate, 1% of your 10,000 bits to be corrupted. The chance of getting 2% corrupted was like-- what was it, 2 to the minus 60 or something? Extremely unlikely to be far from the expected value.

But there's other cases where you are likely-- you could well be far from the expected value. Can anybody remember an example we've done where you are almost surely way off your expected value for a random variable? Anybody remember an example we did that has that feature?

AUDIENCE: The appetizer I think.

PROFESSOR: The appetizer. Let's see. Appetizers, you expect 1, but you're almost certain to be 0.

Or actually, you're almost certain to be 0, and you have a chance of being n. So if you count 0 as being close to 1, you're likely to be close to your expectation. Because you're likely to be 0, and you expect 1.

Remember that noisy channel problem-- not the noisy channel, the latency problem across the channel? And we show that the expected latency was infinite? But 99% of the time you had 10 milliseconds, something like that? There's an example where almost all the time you are far from your expectation which is infinite. So there are examples that go both ways.

Now let's look at another couple of examples that'll motivate the definition that measures this. I'd say that we've got a simple Bernoulli random variable where the probability that R is 1,000 is 1/2, and the probability that R is minus 1,000 is 1/2. Then the expected value of R is 0.

Similarly, we could have another one, S, where the probability that S equals 1 is 1/2, and the probability that S equals minus 1 is 1/2. And the expected value of S is 0. Now if this was a betting game and we're talking about dollars-- here's where you're wagering $1,000, fair game. Here's where you're wagering $1.

Now in this game-- both games are fair. Expected value is 0. But here you're likely to end up near your expected value. Here you're certain to be far by some measure from your expected value.

And in fact, if you were offered to play a game, you might have a real decision as to which game you played. If you like risk, you might play that game. If you're risk averse, maybe you stick with this game because what you could lose would be less.

Now this motivates the definition of the variance because it helps mathematicians distinguish between these two cases with a simple statistic.

The variance of a random variable R-- we'll denote it by var, V-A-R, of R-- is defined as the expected value of the random variable minus this expected value squared. That's sort of a mouthful there. So let's break it down.

This is the expected value of R. This is the deviation from the expected value. So this is the deviation from the mean. Then we square it.

So that equals the square of the deviation. And then we take the expected value of the square. So the variance equals the expected value of the square of the deviation. In other words, the variance gives us the average of the squares of the amount by which the random variable deviates from its mean.

Now the idea behind this is that if a random variable is likely to deviate from its mean, the variance will be high. And if it's likely to be near its mean, the variance will be low. And so variance can tell us something about the expected deviation.

So let's compute the variance for R and S and see what happens. So with R minus the expected value of R, well, that is going to be 1,000, because you expect the value of 0, with probability 1/2, and minus 1,000 with probability 1/2. Then I square that.

Well, I square 1,000, I get a million with probability 1/2. And I square minus 1,000, I get a million again with probability 1/2. And so therefore, the variance of R, well, it's the expected value of this, which is-- well, it's always a million. So it's just a million. Big.

Now if I were to do this with S, S minus the expected value of S is 1 with probability 1/2, minus 1 with probability 1/2. If I square that, well, I get 1 squared is 1, minus 1 squared is 1. And so the variance of S is the expected value of this. And that's just 1. So a big difference in the variance.

So the variance being different tells us these random variables are-- the distributions are very different even though their expected values are the same. And the guy with big variance says, hey, we're likely to deviate from the mean here. And so risk averse people stay away from strategies when they're investing that have high variance.

Now does anybody have any idea why we square the deviation? Why don't we just-- why didn't mathematicians when they figured out this stuff I don't know how many centuries ago, why didn't they just take the expected deviation? Why do the stupid squaring thing? That only is going to complicate it?

Why don't we instead compute the expected value of R minus the mean? Why didn't they do that and call that the variance? Yeah?

That's zero. Yeah. Because by linearity of expectation, that corollary [? 4-2 ?] or whatever, this is just the expected value of R minus the expected value of the expected value of R. The expected value of a scalar is just that scalar. And that is 0.

So the expected deviation from the mean is 0 because of how the mean is defined. It's the midpoint, the weighted midpoint. The times you're high cancel out the times you're low if you got the mean right. And so this is a useless definition. It's always 0.

So mathematicians had to do something to capture this. Now what would have been the more logical thing to do that is the next step. This doesn't work, but what would you think the mathematicians would've done?

Absolute value would have made a lot of sense here. Why they didn't do that? Well, you could do that, but it's hard to work with mathematically. You can't prove nice theorems, it turns out.

If you put the square in there and make that be the variance, you can prove a theorem about linearity of variance. And if the random variables are independent, then the variance of the sum is the sum of the variances. And mathematicians like that kind of thing. It makes it easier to work with and do things with.

Now there are also other choices like, in fact, there's a special name for a weird case where you take the fourth power. You could do that. As long as an even power, you could do it.

And that's actually called the kurtosis. Sounds like a foot disease. But it's the kurtosis of the random variable. Now we're not going to worry about that in this class.

But we are going to worry about variance. And let me do one more definition, then we'll talk about variance a lot more tomorrow. That square is a bit of a pain. And to get rid of it, they made another definition after the fact called the standard deviation.

And standard deviation is defined as follows. For a random variable R, the standard deviation of R is denoted by a sigma of R. And it's just the square root of the variance, undoing that nasty square root after the fact. So it turns out to be the square root of the expectation of the deviation squared.

Another name for this you've probably seen, it's the root of the mean of the square of the deviations. And so you get this thing called root-mean-square, which if any of you ever done curve fitting or any of those kinds of things in statistics or whatever, this is what you're talking about. And so that's why that expression were to come about.

So for the standard deviation of R-- what's the standard deviation of R? 1,000? In effect, that's pretty close to what you expect the deviation to be.

What's the standard deviation of S? 1. Square root of 1 is 1, and that's what you expect its deviation to be.

So we'll do more of this tomorrow on recitation.