PROFESSOR: So today, we're going to talk about the probability that a random variable deviates by a certain amount from its expectation. Now, we've seen examples where a random variable is very unlikely to deviate much from its expectation. For example, if you flip 100 mutually independent fair coins, you're very likely to wind up with close to 50 heads, very unlikely to wind up with 25 or fewer heads, for example.

We've also seen examples of distributions where you are very likely to be far from your expectation, for example, that problem when we had the communications channel, and we were measuring the latency of a packet crossing the channel. There, most of the time, your latency would be 10 milliseconds. But the expected latency was infinite. So you're very likely to deviate a lot from your expectation in that case.

Last time, we looked at the variance. And we saw how that gave us some feel for the likelihood of being far from the expectation-- high variance meaning you're more likely to deviate from the expectation. Today, we're going to develop specific tools for bounding or limiting the probability you deviate by a specified amount from the expectation. And the first tool is known as Markov's theorem.

Markov's theorem says that if the random variable is always non-negative, then it is unlikely to greatly exceed its expectation. In particular, if R is a non-negative random variable, then for all x bigger than 0, the probability that R is at least x is at most the expected value of R, the mean, divided by x.

So in other words, if R is never negative-- for example, say the expected value is smaller. Then the probability R is large will be a small number. Because I'll have a small number over a big number.

So it says that you are unlikely to greatly exceed the expected value. So let's prove that. Now, from the theorem of total expectation that you did in recitation last week, we can compute the expected value of R by looking at two cases-- the case when R is at least x, and the case

when R is less than x. That's from the theorem of total expectation.

I look at two cases. R is bigger than x. Take the expected value there times the probability of this case happening plus the case when R is less than x. OK, now since R is non-negative, this is at least 0. R can't ever be negative. So the expectation can't be negative. A probability can't be negative. So this is at least 0.

And this is trivially at least x. Because I'm taking the expected value of R in the case when R is at least x. So R is always at least x in this case. So its expected value is at least x. So that means that the expected value of R is at least x times the probability R is greater than x, R is greater or equal to x.

And now I can get the theorem by just dividing by x. I'm less than or equal to the expected value of R divided by x. So it's a very easy theorem to prove. But it's going to have amazing consequences that we're going to build up through a series of results today. Any questions about Markov's theorem and the proof?

All right, there's a simple corollary, which is useful. Again, if R is a non-negative random variable, then for all c bigger than 0, the probability that R is at least c times its expected value is at most 1 and c. So the probability you're twice your expected value is at most 1/2.

And the proof is very easy. We just set x to be equal to c times the expected value of R in the theorem. So I just plug in x is c times the expected value of R. And I get expected value of R over c times the expected value of R, which is 1/c. So you just plug in that value in Markov's theorem, and it comes out.

All right, let's do some examples. Let's let R be the weight of a random person uniformly selected. And I don't know what the distribution of weights is in the country. But suppose that the expected value of R, which is the average weight, is 100 pounds. So if I average over all people, their weight is 100 pounds.

And suppose I want to know the probability that the random person weighs at least 200 pounds. What can I say about that probability? Do I know it exactly? I don't think so. Because I don't know what the distribution of weights is.

But I can still get an upper bound on this probability. What bound can I get on the probability that a random person has a weight of 200 given the facts here? Yeah.

AUDIENCE: [INAUDIBLE]

PROFESSOR: Yes, well, it's 100 over 200, right. It's at most the expected value, which is 100, over the x, which is 200. And that's equal to 1/2. So the probability that a random person weighs 200 pounds or more is at most 1/2.

Or I could plug it in here. The expected value is 100. 200 is twice that. So c would be 2 here. So the probability of being twice the expectation is at most 1/2.

Now of course, I'm using the fact that weight is never negative. That's obviously true. But it is implicitly being used here. So what fraction of the population now can weigh at least 200 pounds? Slightly different question. Before I asked you, if I take a random person, what's the probability they weigh at least 200 pounds? Now I'm asking, what fraction of the population can weigh at least 200 pounds if the average is 100? What is it? Yeah?

AUDIENCE: At most 1/2.

PROFESSOR: At most 1/2. In fact, it's the same answer. And why? Why can't everybody weigh 200 pounds, so it would be all the population weighs 200 pounds at least?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Probability would be 1, and that can't happen. And in fact, intuitively, if everybody weighs at least 200 pounds, the average is going to be at least 200 pounds. And we said the average was 100. And this is illustrating this interesting thing that probability implies things about averages and fractions. Because it's really the same thing in disguise. The connection is, if I've got a bunch of people, say, in the country, I can convert a fraction that have some property into a probability by just selecting a random person. Yeah.

AUDIENCE: [INAUDIBLE]

PROFESSOR: No, the variance could be very big. Because I might have a person that weighs a million pounds, say. So you have to get into that. But it gets a little bit more complicated. Yeah.

AUDIENCE: [INAUDIBLE]

PROFESSOR: No, there's nothing being assumed about the distribution, nothing at all, OK? So that's the beauty of Markov's theorem. Well, I've assumed one thing. I assume that there is no negative

values. That's it.

AUDIENCE: [INAUDIBLE]

PROFESSOR: That's correct. They can distribute it any way with positive values. But we have a fact here we've used, that the average was 100. So that does limit your distribution. In other words, you couldn't have a distribution where everybody weighs 200 pounds. Because then the average would be 200, not 100. But anything else where they're all positive and they average 100, you know that at most half can be 200. Because if you pick a random one, the probability of getting one that's 200 is at most 1/2, which follows from Markov's theorem.

And that's partly why it's so powerful. You didn't know anything about the distribution, really, except its expectation and that it was non-negative. Any other questions about this? I'll give you some more examples. All right, here's another example. Is it possible on the final exam for everybody in the class to do better than the mean score? No, of course not. Because if they did, the mean would be higher. Because the mean is the average.

OK, let's do another example. Remember the Chinese appetizer problem? You're at the restaurant, big circular table. There's n people at the table. Everybody has one appetizer in front of them. And then the joker spins the thing in the middle of the table. So it goes around and around. And it stops in a random uniform position.

And we wanted to know, what's the expected number of people to get the right appetizer back? What was the answer? Does anybody remember? One. So you expect one person to get the right appetizer back.

Well, say I want to know the probability that all n people got the right appetizer back. What does Markov tell you about the probability that all n people get the right appetizer back? 1/n. The expected value is 1. And now you're asking the probability that you get R is at least n. So x is n. So it's 1 in n.

And what was the probability, or what is the actual probability? In this case, you know the distribution, that everybody gets the right appetizer back, all n. 1 in n. So in the case of the Chinese appetizer problem, Markov's bound is actually the right answer, right on target, which gives you an example where you can't improve it.

By itself, if you just know the expected value, there's no stronger theorem that way. Because Chinese appetizer is an example where the bound you get, 1/n, of n people getting the right

appetizer is in fact the true probability.

OK, what about the hat check problem? Remember that? So there's n men put the hats in the coat closet. They get uniformly randomly scrambled. So it's a random permutation applied to the hats. Now each man gets a hat back. What's the expected number of men to get the right hat back?

One, same as the other one. Because you've got n men each with a 1 in n chance, so it's 1. Markov says the probability that n men get the right hat back is at most 1 in n, same as before. What's the actual probability that all n men get the right hat back?

AUDIENCE: [INAUDIBLE]

PROFESSOR: 1 in n factorial. So in this case, Markov is way off the mark. It says 1 in n. But in fact the real bound is much smaller. So Markov is not always tight. It's always an upper bound. But it sometimes is not the right answer. And to get the right answer, often you need to know more about the distribution.

OK, what if R can be negative? Is it possible that Markov's theorem holds there? Because I use the assumption in the theorem. Can anybody give me an example where it doesn't work if R can be negative?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Yeah, good, so for example, say probability R equals 1,000 is 1/2, and the probability R equals minus 1,000 is 1/2. Then the expected value of R is 0. And say we asked the probability that R is at least 1,000. Well, that's going to be 1/2. But that does not equal the expected value of R/1,000, which would be 0.

So Markov's theorem really does need that R to be non-negative. In fact, let's see if we saw where we used it in the proof. Anybody see where we use that fact in the proof, that R can't be negative? What is it?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Well, no, because x is positive. We said x is positive. So it's not used there. But that's a good one to look at. Yeah?

AUDIENCE: [INAUDIBLE] is greater than or equal to 0.

PROFESSOR: Yeah, if R can be negative, then this is not necessarily a positive number. It could be a negative number. And then this inequality doesn't hold. OK, good. All right, now it turns out there is a variation of Markov's theorem you can use when R is negative. Yeah.

AUDIENCE: [INAUDIBLE] but would it be OK just to shift everything up?

PROFESSOR: Yeah, yeah, that's great. If R has a limit on how negative it can be, then you make an R prime, which just adds that limit to R, makes it positive or non-negative. And now use Markov's theorem there. And that is now an analogous form of Markov's theorem when R can be negative, but there's a lower limit to it. And I won't stay to improve that here. But that's in the text and something you want to be familiar with.

What I do want to do in class is another case where you can use Markov's theorem to analyze the probability or upper bound the probability that R is very small, less than its expectation. And it's the same idea as you just suggested. So let's state that.

If R is upper bounded, has a hard limit on the upper bound, by u for some u in the real numbers, then for all x less than u, the probability that R is less than or equal to x is at most u minus the expected value of R over u minus x.

So in this case, we're getting a probability that R is less than something instead of R is bigger than something. And we're going to do it using a simple trick that we'll be sort of using all day, really. The probability that R is less than x, this event, R is less than x, is the same as the event u minus R is at least u minus x.

So what have I done? I put negative R over here, subtract x from each side, add u to each side. I've got to put a less than or equal to here. So R is less than or equal to x if and only if u minus r is at least u minus x. It's simple math there.

And now I'm going to apply Markov to this. I'm going to apply Markov to this random variable. And this will be the value I would have had for x up in Markov's theorem. Why is it OK to apply Markov to u minus R?

AUDIENCE: You could just define the new random variable to be u minus R.

PROFESSOR: Yeah, so I got a new random variable. But what do I need to know about that new random

variable to apply Markov?

AUDIENCE: u is always greater than R.

PROFESSOR: u is always greater than R, or at least as big as R. So u minus R is always non-negative. So I can apply Markov now. And when I apply Markov, I'll get this is at most-- maybe I'll go over here. The probability that-- ooh, not R here. This is probability.

The probability that u minus R is at least u minus x is at most the expected value of that random variable over this value. And well now I just use the linearity of expectation. I've got a scalar here. So this is u minus the expected value of R over u minus x. So I've used Markov's theorem to get a different version of it.

All right, let's do an example. Say I'm looking at test scores. And I'll let R be the score of a random student uniformly selected. And say that the max score is 100. So that's u. All scores are at most 100.

And say that I tell you the class average, or the expected value of R, is 75. And now I want to know, what's the probability that a random student scores 50 or below? Can we figure that out? I don't know anything about the distribution, just that the max score is 100 and the average score is 75. What's the probability that a random student scores 50 or less? I want to upper bound that.

So we just plug it into the formula. u is 100. The expected value is 75. u is 100. And x is 50. And that's 25 over 50, is 1/2. So at most half the class can score 50 or below. And state it as a probability question or deterministic fact if I know the average is 75 and the max is 100. Of course, another way of thinking about that is if more than half the class scored 50 or below, your average would have had to be lower, even if everybody else was right at 100. It wouldn't average out to 75. All right, any questions about that?

OK, so sometimes Markov is dead on right, gives the right answer. For example, half the class could have scored 50, and half could have gotten 100 to make it be 75. And sometimes it's way off, like in the hat check problem.

Now, if you know more about the distribution, then you can get better bounds, especially the cases when you're far off. For example, if you know the variance in addition to the expectation, or aside from the expectation, then you can get better bounds on the probability that the random variable is large.

And in this case, the result is known as Chebyshev's theorem. I'll do that over here. And it's the analog of Markov's theorem based on variance. It says, for all x bigger than 0, and any random variable R-- could even be negative-- the probability that R deviates from its expected value in either direction by at least x is at most of the variance of R divided by x squared.

So this is like Markov's theorem, except that we're now bounding the deviation in either direction. Instead of expected value, you have variance. Instead of x, you've got x squared, but the same idea. In fact, the proof uses Markov's theorem.

Well, the probability that R deviates from its expected value by at least x, this is the same event, or happens if and only if R minus expected value squared is at least x squared. I'm just going to square both sides here. OK, I square both sides. And since this is positive and this is positive, I can square both sides and maintain the inequality.

Now I'm going to apply Markov's theorem to that random variable. It's a random variable. It's R minus expected value squared. So it's a random variable. And what's nice about this random variable that lets me apply Markov's theorem? It's a square. So it's always non-negative.

So I can apply Markov's theorem. And my Markov's theorem, this probability is at most the expected value of that divided by this. That's what Markov's theorem says as long as this is always non-negative. All right, what's a simpler expression for this, the expected value of the square of the deviation of a random variable? That's the variance. That's the definition of variance.

So that is just the variance of R over x squared. And we're done. So Chebyshev's theorem is really just another version of Markov's theorem. But now it's based on the variance. OK, any questions?

OK, so there's a nice corollary for this, just as with Markov's theorem. It says the probability that the absolute value, the deviation, is at least c times the standard deviation of R. So I'm looking at the probability that R differs from its expectation by at least some scalar c times the standard deviation.

Well, what's that? Well, that's the variance of R over the square of this thing-- c squared times the standard deviation squared. What's the square of the standard deviation? That's the variance. They cancel, so it's just 1 over c squared. So the probability of more than twice the

standard deviation off the expectation is at most 1/4, for example. All right, let's do some examples of that. Maybe we'll leave Markov up there.

OK, say we're looking at IQs. In this case, we're going to let R be the IQ of a random person. All right, now we're going to assume-- and this actually is the case-- that R is always at least 0, despite the fact that probably most of you have somebody you know who you think has a negative IQ. They can't be negative. They have to be non-zero.

In fact, IQs are adjusted. So the expected IQ is supposed to be 100, although actually the averages may be in the 90's. And it's set up so that the standard deviation of IQ is supposed to be 15. So we're just going to assume those are facts on IQ. And that's what it's meant to be.

And now we want to know, what's the probability a random person has an IQ of at least 250? Now Marilyn, from "Ask Marilyn," has an IQ pretty close to 250. And she thinks that's pretty special, pretty rare. So what can we say about that? In particular, say we used Markov. What could you say about the probability of having an IQ of at least 250? What does Markov tell us?

AUDIENCE: [INAUDIBLE]

PROFESSOR: What is it?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Not quite 1 in 25, but you're on the right track. It's not quite 2/3. It's the expected value, which is 100, over the x value, which is 250. So it's 1 in 2.5, or 0.4. So the probability is at most 0.4, so 40% chance it could happen, potentially, but no bigger than that. What about Chebyshev? See if you can figure out what Chebyshev says about the probability of having an IQ of at least 250.

It's a little tricky. You've got to sort of plug it into the equation there and get it to fit in the right form. Chebyshev says that-- let's get in the right form. I've got probability that R is at least 250. I've got to get it into that form up there. So that's the probability that-- well, first R minus 100 is at least 150. So I've got R minus the expected value. I'm sort of getting it ready to apply Chebyshev here.

And then 150-- how many standard deviations is 150? 10, all right? So this is the probability that R minus the expected value of R is at least 10 standard deviations. That's what I'm asking. I'm not quite there. I'm going to use the corollary there. But I've got to get that absolute

value thing in.

But it's upper bounded by the probability of the absolute value of R minus expected value bigger than or equal to 10 standard deviations. Because this allows for two cases. R is 10 standard deviations high, and R is 10 standard deviations low or more. So this is upper bounded by that.

And now I can plug in Chebyshev in the corollary form. And what's the answer when I do that? 1 in 100-- the probability of being off by 10 standard deviations or more is at most 1 in 100, 1 in 10 squared. So it's a lot better bound. It's 1% instead of 40%. So knowing the variance of the standard deviation gives you a lot more information and generally gives you much better bounds on the probability of deviating from the mean.

And the reason it gives you better bounds is because the variance is squaring deviations. So they count a lot more. All right, now let's look at this step a little bit more. All right, let's say here is a line, and here's the expected value of R. And say here's 10 standard deviations high here. So this will be more than 10 standard deviations. And this will be 10 standard deviations on the low side. So here, I'm low.

Now, this line here with the absolute value is figuring out the probability of being low or high. This is the probability that the absolute value of R minus its expected value is at least 10 standard deviations. What we really wanted to know for bound was just the high side.

Now, is it true that then, since the probability of high or low is 1 in 100, the probability of being high is at most 1 in 200, half? Is that true? Yeah?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Yeah, it is not necessarily true that the high and the low are equal, and therefore the high is half the total. It might be true, but not necessarily true. And that's a mistake that often gets made where you'll take this fact as being less than 100 to conclude that's less than 1 in 200. And that you can't do, unless the distribution is symmetric around the expected value. Then you could do it, if it's a symmetric distribution around the expected value. But usually it's not.

Now, there is something better you can say. So let me tell you what it is. But we won't prove it. I think we might prove it in the text. I'm not sure. If you just want the high side or just want the low side, you can do slightly better than 1 in c squared. That's the following theorem.

For any random variable R, the probability that R is on the high side by c standard deviations is at most 1 over c squared plus 1. So it's not 1 over 2c squared. It's 1 over c squared plus 1, and the same thing for the probability of being on the low side.

Let's see, have I written this right? Hmm, I want to get this as less than or equal to negative c times the standard deviation. So here I'm high by c or more standard deviations. Here I'm low. So R is less than the expected value by at least c standard deviations. And that is also 1 over c squared plus 1.

And it is possible to find distributions that hit these targets-- not both at the same time, but one or the other, hit those targets. So that's the best you can say in general. All right, so using this bound, what's the probability that a random person has an IQ of at least 250? It's a little better than 1 in 100.

AUDIENCE: [INAUDIBLE]

PROFESSOR: Yeah, 1/101. So in fact, the best we can say without knowing any more information about IQs is that it's at most 1/101, slightly better. Now in fact, with IQs, they know more about the distribution. And the probability is a lot less. Because you know more about the distribution than we've assumed here. In fact, I don't think anybody has an IQ over 250 as far as I know. Any questions about this?

OK, all right, say we give the exam. What fraction of the class can score more than two standard deviations, get two standard deviations or more, away from the average, above or below? Could half the class be two standard deviations off the mean? No? What's the biggest fraction that that could happen?

What do I do? What fraction of the class can be two standard deviations or more from the mean? What is it?

AUDIENCE: 1/4.

PROFESSOR: 1/4, because c is 2. You don't even know what the mean is. You don't know what the standard deviation is. You don't need to. I just asked, you're two standard deviations off or more. At most, 1/4. How many could be two standard deviations high or better at most? 1/5-- 1 over 4 plus 1, good. OK, this holds true no matter what the distribution of test scores is. Yeah?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Which one? This one?

AUDIENCE: Yeah.

PROFESSOR: Oh, that's more complicated. That'll take us several boards to do, to prove that. And I forget if we put it in the text or not. It might be in the text, to prove that. Any other questions?

OK so Markov and Chebyshev are sometimes close, sometimes not. Now, for the rest of today, we're going to talk about a much more powerful technique. But it only works in a special case. Now, the good news is this special case happens all the time in practice. And it's the case when you're analyzing a random variable that itself is the sum of a bunch of other random variables. And we've seen already examples like that.

And the other random variables have to be mutually independent. And in this case, you get a bound that's called a Chernoff bound. And this is the same Chernoff who figured out how to beat the lottery. And it's interesting. Long after we started teaching this, originally this stuff was only taught, for Chernoff bounds, for graduate students. And now we teach it here. Because it's so important. And it really is accessible.

It'll be probably the most complicated proof we've done to establish a Chernoff bound. But Chernoff himself, when he discovered this, thought it was no big deal. In fact, he couldn't figure out why everybody in computer science was always writing papers with Chernoff bounds in them.

And that's because he didn't put any emphasis on the bounds in his work. But computer scientists who came later found all sorts of important applications. And we'll see some of those today. So let me tell you what the bound is. And the nice thing is it really is Markov's theorem again in disguise, just a little more complicated.

Theorem-- it's called a Chernoff bound. Let T1, T2, up to Tn be any mutually independent-- that's really important-- random variables such that each of them takes values only between 0 and 1. And if they don't, just normalize them so they do. So we're going to take a bunch of random variables that are mutually independent. And they are all between 0 and 1.

Then we're going to look at the sum of those random variables, call that T. Then for any c at least 1, the probability that the sum random variable is at least c times its expected value. So

it's going to be the high side here-- is at most e to the minus z, and I'll tell you what that is in a minute, times the expected value of T where z is c natural log of c plus 1 minus c. And it turns out if c is bigger than 1, this is positive.

So that's a lot, one of the longest theorems we wrote down here. But what it says is that probability were high is exponentially small. As the expected value is big, the chance of being high gets really, really tiny. Now, I'm going to prove it in a minute. But let's just plug in some examples to see what's going on here.

So for example, suppose the expected value of T is 100. And suppose c is 2. So we expect to have 100 come out of the sum. The probability we get at least 200-- well, let's figure out what that is. c being 2 we can evaluate z now. It's 2 natural log of 2 plus 1 minus 2. And that's close to but a little larger than 0.38.

So we can plug z in, the exponent up there, and find that the probability that T is at least twice its expected value, namely at least 200, is at most e to the minus 0.38 times 100, which is e to the minus 38, which is just really small.

So that's just way better than any results you get with Markov or Chebyshev. So if you have a bunch of random variables between 0 and 1, and they're mutually independent, you add them up. If you expect 100 as the answer, the chance of getting 200 or more-- forget about it, not going to happen.

Now, of course Chernoff doesn't apply to all distributions. It has to be this type. This is a pretty broad class. In fact, it contains the class of all Bernoulli distributions. So I have binomial distributions. Because remember a binomial distribution-- well, remember binomial distributions? That's where T is the sum of Ti's.

In binomial, you have Tj is 0 or 1. It can't be in between. And with binomial, all Tj's have the same distribution. With Chernoff, they can all be different. So Chernoff is much broader than binomial. The individual guys here can have different distributions and attain values anywhere between 0 and 1, as opposed to just one or the other. Any questions about this theorem and what it says in the terms there?

One nice thing about it is the number of random variables doesn't even show up in the answer here. n doesn't even appear. Yeah.

AUDIENCE: [INAUDIBLE]

PROFESSOR: Does not apply to what?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Yeah, when c equals 1, what happens is z is 0. Because I have a log of 1 is 0, and 1 minus 1 is 0. And if z is 0, it says your probability is upper bounded by 1. Well, not too interesting, because any probability is upper bounded by 1. So it doesn't give you any information when c is 0, none at all. But as soon as c starts being-- sorry, if c is 1. As soon as c starts being bigger than 1, which is sort of a case you're interested in, you're bigger than your expectation, then it gives very powerful results. Yeah.

AUDIENCE: [INAUDIBLE]

PROFESSOR: Yeah, you can. It's true for n equals 1 as well. Now, it doesn't give you a lot of information. Because if c is bigger than 1 and n was 1, so it's using one variable, what's the probability that a random variable exceeds its expectation, c times its expectation?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Yeah, let's see now. Maybe it does give you information. Because the random variable has a distribution on 0, 1. That's right, so it does give you some information. But I don't think it gives you a lot. I have to think about that. What happens when there's just one guy?

Because the same thing is true. It's just now for a single random variable on 0, 1 the chance that your twice the expected value. I have to think about that. That's a good question. Does it do anything interesting there? OK, all right, so let's do an example of how you might apply this.

Say that you're playing Pick 4, and 10 million people are playing. And say in this version of Pick 4, you're picking a four digit number, four single digits. And you win if you get an exact match. So the probability of winning, a person winning, well, they've got to get all four digits right. That's 1 in 10,000, 10 to the fourth.

What's the expected number of winners? If I got 10 million people, what's the expected number of winners? What is it? We've got 10 million over 10,000, right? Because what I'm doing here is the number of winners, T, is going to be the sum of 10 million indicator variables. And the probability that any one of these guys wins is 1 in 10,000. So the expected number of winners is 1 in 10,000 added 10 million times, which is this.

Is that OK? Everybody should be really familiar with how to whip these things out. This for sure will have probably at least a couple questions where you're going to need to be able to do that kind of stuff on the final.

All right, say I want to know the probability of getting at least 2,000 winners, and I want to upper bound that just with the information I've given you. Well, any thoughts about an upper bound?

AUDIENCE: [INAUDIBLE]

PROFESSOR: What's that?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Yeah, that's a good upper bound. What did you have to assume to get there? e to the minus 380 is a great bound. Because you're going to plug in expected value is 1,000. And we're asking for more than twice the expected value. So it's e to the minus 0.38 times 1,000. And that for sure is-- so you computed this. And that equals e to the minus 380. So that's really small.

But what did you have to assume to apply Chernoff? Mutual independence. Mutual independence of what?

AUDIENCE: [INAUDIBLE]

PROFESSOR: The numbers people picked. And we already know, if people are picking numbers, they don't tend to be mutually independent. They tend to gang up. But if you had a computer picking the numbers randomly and mutually independently, then you would be e to the minus 380 by Chernoff if mutually independent picks.

Everybody see why we did that? Because it's a probability of twice your expectation. The total number of winners is the sum of 10 million indicator variables. And indicator variables are 0 or 1. So they fit that definition up there. And so we already figured out z is at least 0.38. And you're multiplying by the expected value of 1,000. That's e to the minus 380, so very, very unlikely.

What if they weren't mutually independent? Can you say anything about this, anything at all better than 1, which we know for any probability? Yeah?

AUDIENCE: It's possible that everyone chose the same numbers.

PROFESSOR: Yes, everyone could have chosen the same number. But that number only comes up with a 1 in 10,000 chance. So you can say something.

AUDIENCE: You can use Markov.

PROFESSOR: Use Markov. What does Markov give you? What does Markov give you? 1/2, yeah. Because you've got the expected value is 1,000 divided by the bound threshold, is 2,000, is 1/2 by Markov. And that holds true without any independence assumption.

Now, there is an enormous difference between 1/2 and e to the minus 380. Independence really makes a huge difference in the bound you can compute. OK, now there's another way we could've gone about this. What kind of distribution does T have in this case? It's binomial. Because it's the sum of indicator random variables, 0, 1's. Each of these is 0, 1. And they're all the same distribution. There's a 1 in 10,000 chance of winning for each one of them. So it's a binomial.

So we could have gone back and used the formulas we had for the binomial distribution, plugged it all in, and we'd have gotten something pretty similar here. But Chernoff is so much easier. Remember that pain we would go through with a binomial distribution, the approximation, Stirling's formula, [INAUDIBLE] whatever, the factorials and stuff? And that's a nightmare.

This was easy. e to the minus 380 was very easy to compute. And really at that point it doesn't matter if it's minus 381 or minus 382 or whatever. Because it's really small. So often, even when you have a binomial distribution, well, Chernoff will apply. And that's a great way to go. Because it gives you good bounds generally.

All right, let's figure out the probability of at least 1,100 winners instead of 1,000. So let's look at the probability of at least 100 extra winners over what we expect out of 10 million. We've got 10 million people. You expect 1,000. We're going to analyze the probability of 1,100. What's c in this case? We're going to use Chernoff.

1.1. So this is 1.1 times 1,000. And that means that z is 1.1 times the natural log of 1.1 plus 1 minus 1.1. And that is close to but at least 0.0048. So this probability is at most, by Chernoff, e to the minus 0.0048 times the expected number of winners is 1,000. So that is e to the minus

4.8, which is less than 1%, 1 in 100.

So that's pretty powerful. It says, you've got 10 million people who could win. The chance of even having 100 more than the 1,000 you expect is 1% chance at most-- very, very powerful. It says you really expect to get really close to the mean in this situation.

OK, a lot better-- Markov here gives you, what, 1,000 over 1,100. It says your probability could be 90% or something-- not very useful. Chebyshev won't give you much here either. So if you're in a situation to apply Chernoff, always go there. It gives you the best bounds. Any questions? This of course is why computer scientists use it all the time.

OK, actually, before I do more examples, let me prove the theorem in a special case to give you a feel for what's involved. The full proof is in the text. I'm going to prove it in the special case where the Tj are 0, 1. So they're indicator random variables.

But they don't have to have the same distribution. So it's still more general than you get with a binomial distribution. All right, so we're going to do a proof of Chernoff for the special case where the Tj are either 0 or 1. So they're indicator variables.

OK, so the first step is going to seem pretty mysterious. But we've been doing something like it all day. I'm trying to compute the probability T is bigger than c times its expectation. Well, what I'm going to do is exponentiate both of these guys and compute the probability that c to the T is at least c to the c times the expected value of T.

Now, this is not the first thing you'd expect to do, probably, if you were trying to prove this. So it's one of those divine insights that you'd make this step. And then I'm going to apply Markov, like we've been doing all day, to this. Now, since T is positive and c is positive, these are equal. And this is never non-negative.

So now by Markov, this is simply upper bounded by the expected value of that, expected value of c to the T, divided by this. And that's by Markov. So everything we've done today is really Markov in disguise.

Any questions so far? You start looking at this, you go, oh my god, I got the random variable and the exponent. This is looking like a nightmare. What is the expected value of c to the T, and this kind of stuff? But we're going to hack through it. Because it gives you just an amazingly powerful result when you're done.

All right, so we've got to evaluate the expected value of c to the T. And we're going to use the fact that T is the sum of the Tj's. And that means that c to the T equals c to the T1 times c to the T2 times c to the Tn.

The weird thing about this proof is that every step sort of makes it more complicated looking until we get to the end. So it's one of those that's hard to figure out the first time. All right, that means the expected value of c to the T is the expected value of the product of these things. Now I'm going to use the product rule for expectation.

Now, why can I use the product rule? What am I assuming to be able to do that? That they are mutually independent, that the c to the Tj's are mutually independent of each other. And that follows, because the Tj's are mutually independent.

So if a bunch of random variable are mutually independent, then their exponentiations are mutually independent. So this is by product rule for expectation and mutual independence. OK, so now we've got to evaluate the expected value of c to the Tj.

And this is where we're going to make it simpler by assuming that Tj is just a 0, 1 random variable. So the simplification comes in here. So the expected value of Tj-- well, there's two cases. Tj is 1, or it's 0. Because we made this simplification.

If it's 1, I get c to the 1-- ooh, expected value of c to the Tj. Let's get that right. It could be 1, in which case I get a contribution of c to the 1 times the probability Tj equals 1 plus the case at 0. So I get c to the 0 times the probability Tj is 0.

Well, c to the 1 is just c. c to the 0 is 1. And I'm going to rewrite Tj being 0 as 1 minus the probability Tj is 1. All right, this equals that. And of course the 1 cancels. Now I'm going to collect terms here to get 1 plus c minus 1 times the probability Tj equals 1.

OK, then I'm going to do one more step here. This is 1 plus c minus 1 times the expected value of Tj. Because if I have an indicator random variable, the expected value is the same as the probability that it's 1. Because in the other case it's 0.

And now I'm going to use the trick from last time. Remember 1 plus x is always at most e to the x from last time? None of these steps is obvious why we're doing them. But we're going to do them anyway. So this is at most e to this, c minus 1 expected value of Tj. Because 1 plus anything is at most the exponential of that.

And I'm doing this step because I got a product of these guys. And I want to put them in the exponent so I can then sum them so it gets easy.

OK, now we just plug this back in here. So that means that the expected value of c to the T is at most a product of expected value of e to the cTj is this-- e to the c minus 1 expected value of Tj. And now I can convert this to a sum in the exponent. And this is j equals 1 to n.

And what do I do to simplify that? Linearity of expectation. c minus 1 times the sum j equals 1 to n expected value of Tj. Ooh, let's see, did I? Actually, I used linearity coming out. I already used linearity. I screwed up here.

So here I used the linearity when I took the sum up here inside the expectation. I've already used linearity. What is the sum of the Tj's? T-- yeah, that's what I needed to do here.

OK, we're now almost done. We've got now an upper bound on the expected value of c to the T. And it is this. And we just plug that in back up here. So now this is at most e to the c minus 1 expected value of T over c to the c times the expected value of t. And now I just do manipulation. c to something is the same as e to the log of c times that something. So this is e to the minus c ln c expected value of T plus that.

And then I'm running out of room. That equals-- I can just pull out the expected values of T. I get e to the minus c log of c plus c minus 1 expected value of T. And that's e to the minus z expected value of T.

All right, so that's a marathon proof. It's the worst proof I think. Well, maybe minimum spanning tree was worse. But this is one of the worst proofs we've seen this year. But I wanted to show it to you. Because it's one of the most important results that we cover, certainly in probability, that can be very useful in practice. And it gives you some feel for, hey, this wasn't so obvious to do it the first time, and also some of the techniques that are used, which is really Markov's theorem. Any questions? Yeah.

AUDIENCE: Over there, you define z as 1 minus c.

PROFESSOR: Did I do it wrong?

AUDIENCE: c natural log of c, 1 minus c. Maybe it's plus c?

PROFESSOR: Oh, I've got to change the sign. Because I pulled a negative out in front. So it's got to be

negative c minus 1, which means negative c plus 1. Yeah, good. Yeah, this was OK. I just made the mistake going to there. Any other questions?

OK, so the common theme here in using Markov to get Chebyshev, to get Chernoff, to get the Markov extensions, is always the same. And let me show you what that theme is. Because you can use it to get even other results.

When we're trying to figure out the probability that T is at least c times its expected value, or actually even in general, even more generally than that, the probability that A is bigger than B, even more generally, well, that's the same as the probability that f of A is bigger than f of B as long as you don't change signs. And then by Markov, this is at most the expected value of that as long as it's non-negative over that.

In Chebyshev, what function f did we use for Chebyshev in deriving Chebyshev's theorem? What was f doing in Chebyshev? Actually I probably just erased it. What operation were we doing with Chebyshev?

AUDIENCE: Variance.

PROFESSOR: Variance. And that meant we were squaring it. So the technique used to prove Chebyshev was f was the square function. For Chernoff, f is the exponentiation function, which turns out to be-- in fact, when we did it for Chernoff, that's the optimal choice of functions to get good bounds.

All right, any questions on that? All right, let's do one more example here with numbers. And this is a load balancing application for example you might have with web servers. Say you've got to build a load balancing device, and it's got to balance N jobs, B1, B2, to BN, across a set of M servers, S1, S2, to SN.

And say you're doing this for a decent sized website. So maybe N is 100,000. You get 100,000 requests a minute. And say you've got 10 servers to handle those requests. And say the requests are-- the time for the j-th request is, say, Bj takes the j-th job. The j-th request takes Lj time. And the time is the same on any server. The servers are all equivalent.

And let's assume it's normalized so that Lj is between 0 and 1. Maybe the worst job takes a second to do, let's say. And say that if you sum up the length of all the jobs, you get L. Total workload is the sum of all of them. j equals 1 to N.

And we're going to assume that the average job length is 1/4 second. So we're going to

assume that the total amount of work is 25,000 seconds, say. So the average job length is 1/4 second. And the job is to assign these tasks to the 10 servers so that hopefully every server is doing L/M work, which would be 25,000/10, or 2,500 milliseconds of work, something like that. I don't know.

Because when you're doing load balancing, you want to take your load and spread it evenly and equally among all the servers. Any questions about the problem? You've got a bunch of jobs, a bunch of servers. You want to assign the jobs to the servers to balance the load. Well, what is the simplest algorithm you could think of to do this?

AUDIENCE: [INAUDIBLE]

PROFESSOR: That's a good algorithm to do this. In practice, the first thing people do is, well, take the first N/M jobs, put them on server one, the next N/M on server two. Or they'll use something called round robin-- first job goes here, second here, third here, 10th here, back and start over. And they hope that it will balance the load. But it might well not. Because maybe every 10th job is a big one.

So what's much better to do in practice is to assign them randomly. So a job comes in. You don't even pay attention to how hard it is, how much time you think it'll take. You might not even know before you start the job how long it's going to take to complete. Give it to a random server. Don't even look at how much work that server has. Just give it to a random one.

And it turns out this does very, very well. Without knowing anything, just that simple approach does great in practice. And today, state of the art load balancers do this. We've been doing randomized kinds of thing like this at Akamai now for a decade. And it's just stunning how well it works. And so let's see why that is.

Of course we're going to use the Chernoff bound to do it. So let's let Rij be the load on server Si from job Bj. Now, if Bj is not assigned to Si, it's zero load. Because it's not even doing the work there. So we know that Rij equals the load of Bj if it's assigned to Si. And that happens with probability 1/M. The job picks one of the M servers at random. And otherwise, the load is 0. Because it's not assigned to that server. And that is probability 1 minus 1/M.

Now let's look at how much load gets assigned by this random algorithm to server i. So we'll let Ri be the sum of all the load assigned to server i. So we've got this indicator where the random variables are not 0, 1. They're 0 and whatever this load happens to be for the j-th job, at most

1. And we sum up the value for the contribution to Si over all the jobs.

So now we compute the expected value of Ri, the expected load on the i-th server. So the expected load on the i-th server is-- well, we use linearity of expectation. And the expected value of Rij-- well, 0 or Lj. It's Lj with probability 1/M. This is just now the sum of Lj over M. And the sum of Lj is just L.

So the expected load of the i-th server is the total load divided by the number of servers, which is perfect. It's optimal-- can't do better than that. It makes sense. If you assign all the jobs randomly, every server is expecting to get 1/M of the total load.

Now we want to know the probability it deviates from that, that you have too much load on the i-th server. All right, so the probability that the i-th server has c times the optimal load is at most, by Chernoff, if the jobs are independent, minus zL over M, minus z times the expected load where z is c ln c plus 1 minus c. This is Chernoff now, just straight from the formula of Chernoff, as long as these loads are mutually independent.

All right, so we know that when c gets to be-- I don't know, you pick 10% above optimal, c equals 1.1, well, we know that this is going to be a very small number. L/M is 2,500. And z, in this case, we found was 0.0048. So we get e to the minus 0.0048 times 2,500. And that is really tiny. That's less than 1 in 160,000.

So Chernoff tells us the probability that any server, a particular server, gets 10% load more than you expect is minuscule. Now, we're not quite done. That tells us the probability the first server gets 10% too much load or the problem the second server got 10% too much load, and so forth.

But what we really care about is the worst server. If all of them are good except for one, you're still in trouble. Because the one ruined your day. Because it didn't get the work done. So what do you do to bound the probability that any of the servers got too much load, any of the 10?

So what I really want to know is the probability that the worst server of M takes more than cL over M. Well, that's the probability that the first one has more than cL over M union the second one has more than cL over M union the M-th one. What do I do to get that probability, the probability of a union of events, upper bounded?

AUDIENCE: [INAUDIBLE]

PROFESSOR: Upper bounded by the sum of the individual guys. It's the sum i equals 1 to M probability Ri greater than or equal to cL over M. And so that, each of these is at most 1 in 160,000. This is at most M/160,000. And that is equal to 1 in 16,000.

All right, so now we have the answer. The chance that any server got 10% load or more is 1 in 16,000 at most, which is why randomized load balancing is used a lot in practice. Now tomorrow, you're going to do a real world example where people use this kind of analysis, and it led to utter disaster. And the reason was that the components they were looking at were not independent.

And the example has to do with the subprime mortgage disaster. And I don't have time today to go through it all. But it's in the text, and you'll see it tomorrow. But basically what happened is that they took a whole bunch of loans, subprime loans, put them into these things called bonds, and then did an analysis about how many failures they'd expect to have. And they assumed the loans were all mutually independent.

And they applied their Chernoff bounds. And that concluded that the chances of being off from the expectation were nil, like e to the minus 380. In reality, the loans were highly dependent. When one failed, a lot tended to fail. And that led to disaster. And you'll go through some of the math on that tomorrow in recitation.