JOHN TSITSIKLIS: We're going to start today a new unit. so we will be talking about limit theorems. So just to introduce the topic, let's think of the following situation. There's a population of penguins down at the South Pole. And if you were to pick a penguin at random and measure their height, the expected value of their height would be the average of the heights of the different penguins in the population. So suppose when you pick one, every penguin is equally likely. Then the expected value is just the average of all the penguins out there.

So your boss asks you to find out what that the expected value is. One way would be to go and measure each and every penguin. That might be a little time consuming. So alternatively, what you can do is to go and pick penguins at random, pick a few of them, let's say a number n of them. So you measure the height of each one. And then you calculate the average of the heights of those penguins that you have collected. So this is your estimate of the expected value.

Now, we called this the sample mean, which is the mean value, but within the sample that you have collected. This is something that's sort of feels the same as the expected value, which is again, the mean. But the expected value's a different kind of mean. The expected value is the mean over the entire population, whereas the sample mean is the average over the smaller sample that you have measured.

The expected value is a number. The sample mean is a random variable. It's a random variable because the sample you have collected is random.

Now, we think that this is a reasonable way of estimating the expectation. So in the limit as n goes to infinity, it's plausible that the sample mean, the estimate that we are constructing, should somehow get close to the expected value. What does this mean? What does it mean to get close? In what sense? And is this statement true?

This is the kind of statement that we deal with when dealing with limit theorems. That's the subject of limit theorems, when what happens if you're dealing with lots and lots of random variables, and perhaps take averages and so on.

So why do we bother about this? Well, if you're in the sampling business, it would be reassuring to know that this particular way of estimating the expected value actually gets you close to the true answer. There's also a higher level reason, which is a little more abstract and mathematical. So probability problems are easy to deal with if you're having in your hands one or two random variables. You can write down their mass functions, joints density functions, and so on. You can calculate on paper or on a computer, you can get the answers.

Probability problems become computationally intractable if you're dealing, let's say, with 100 random variables and you're trying to get the exact answers for anything. So in principle, the same formulas that we have, they still apply. But they involve summations over large ranges of combinations of indices. And that makes life extremely difficult.

But when you push the envelope and you go to a situation where you're dealing with a very, very large number of variables, then you can start taking limits. And when you take limits, wonderful things happen. Many formulas start simplifying, and you can actually get useful answers by considering those limits. And that's sort of the big reason why looking at limit theorems is a useful thing to do.

So what we're going to do today, first we're going to start with a useful, simple tool that allows us to relates probabilities with expected values. The Markov inequality is the first inequality we're going to write down. And then using that, we're going to get the Chebyshev's inequality, a related inequality.

Then we need to define what do we mean by convergence when we talk about random variables. It's a notion that's a generalization of the notion of the usual convergence of limits of a sequence of numbers. And once we have our notion of convergence, we're going to see that, indeed, the sample mean converges to the true mean, converges to the expected value of the X's. And this statement is called the weak law of large numbers.

The reason it's called the weak law is because there's also a strong law, which is a statement with the same flavor, but with a somewhat different mathematical content. But it's a little more abstract, and we will not be getting into this. So the weak law is all that you're going to get.

All right. So now we start our digression. And our first tool will be the so-called Markov inequality. So let's take a random variable that's always non-negative. No matter what, it gets no negative values. To keep things simple, let's assume it's a discrete random variable. So the expected value is the sum over all possible values that a random variable can take. The values of the random variables that can take weighted according to their corresponding probabilities.

Now, this is a sum over all x's. But x takes non-negative values. And the PMF is also non-negative. So if I take a sum over fewer things, I'm going to get a smaller value. So the sum when I add over everything is less than or equal to the sum that I will get if I only add those terms that are bigger than a certain constant.

Now, if I'm adding over x's that are bigger than a, the x that shows up up there will always be larger than or equal to a. So we get this inequality. And now, a is a constant. I can pull it outside the summation. And then I'm left with the probabilities of all the x's that are bigger than a. And that's just the probability of being bigger than a.

OK, so that's the Markov inequality. Basically tells us that the expected value is larger than or equal to this number. It relates expected values to probabilities. It tells us that if the expected value is small, then the probability that x is big is also going to be small. So it's translates a statement about smallness of expected values to a statement about smallness of probabilities.

OK. What we actually need is a somewhat different version of this same statement. And what we're going to do is to apply this inequality to a non-negative random variable of a special type. And you can think of applying this same calculation to a random variable of this form, (X minus mu)-squared, where mu is the expected value of X.

Now, this is a non-negative random variable. So, the expected value of this random variable, which is the variance, by following the same thinking as we had in that derivation up to there, is bigger than the probability that this random variable is bigger than some-- let me use a-squared instead of an a times the value a-squared.

So now of course, this probability is the same as the probability that the absolute value of X minus mu is bigger than a times a-squared. And this side is equal to the variance of X. So this relates the variance of X to the probability that our random variable is far away from its mean. If the variance is small, then it means that the probability of being far away from the mean is also small.

So I derived this by applying the Markov inequality to this particular non-negative random variable. Or just to reinforce, perhaps, the message, and increase your confidence in this inequality, let's just look at the derivation once more, where I'm going, here, to start from first principles, but use the same idea as the one that was used in the proof out here.

Ok. So just for variety, now let's think of X as being a continuous random variable. The derivation is the same whether it's discrete or continuous. So by definition, the variance is the integral, is this particular integral. Now, the integral is going to become smaller if I integrate, instead of integrating over the full range, I only integrate over x's that are far away from the mean. So mu is the mean. Think of c as some big number.

These are x's that are far away from the mean to the left, from minus infinity to mu minus c. And these are the x's that are far away from the mean on the positive side. So by integrating over fewer stuff, I'm getting a smaller integral.

Now, for any x in this range, this distance, x minus mu, is at least c. So that squared is at least c squared. So this term over this range of integration is at least c squared. So I can take it outside the integral. And I'm left just with the integral of the density. Same thing on the other side.

And so what factors out is this term c squared. And inside, we're left with the probability of being to the left of mu minus c, and then the probability of being to the right of mu plus c, which is the same as the probability that the absolute value of the distance from the mean is larger than or equal to c. So that's the same inequality that we proved there, except that here I'm using c. There I used a, but it's exactly the same one.

This inequality was maybe better to understand if you take that term and send it to the other side and write it this form. What does it tell us? It tells us that if c is a big number, it tells us that the probability of being more than c away from the mean is going to be a small number. When c is big, this is small.

Now, this is intuitive. The variance is a measure of the spread of the distribution, how wide it is. It tells us that if the variance is small, the distribution is not very wide. And mathematically, this translates to this statement that when the variance is small, the probability of being far away is going to be small. And the further away you're looking, that is, if c is a bigger number, that probability also becomes small.

Maybe an even more intuitive way to think about the content of this inequality is to, instead of c, use the number k, where k is positive and sigma is the standard deviation. So let's just plug k sigma in the place of c. So this becomes k sigma squared. These sigma squared's cancel. We're left with 1 over k-square.

Now, what is this? This is the event that you are k standard deviations away from the mean. So for example, this statement here tells you that if you look at the test scores from a quiz, what fraction of the class are 3 standard deviations away from the mean? It's possible, but it's not going to be a lot of people. It's going to be at most, 1/9 of the class that can be 3 standard deviations or more away from the mean.

So the Chebyshev inequality is a really useful one. It comes in handy whenever you want to relate probabilities and expected values. So if you know that your expected values or, in particular, that your variance is small, this tells you something about tailed probabilities.

So this is the end of our first digression. We have this inequality in our hands. Our second digression is talk about limits. We want to eventually talk about limits of random variables, but as a warm up, we're going to start with limits of sequences.

So you're given a sequence of numbers, a1, a2, a3, and so on. And we want to define the notion that a sequence converges to a number. You sort of know what this means, but let's just go through it some more. So here's a. We have our sequence of values as n increases.

What do we mean by the sequence converging to a is that when you look at those values, they get closer and closer to a. So this value here is your typical a sub n. They get closer and closer to a, and they stay closer. So let's try to make that more precise.

What it means is let's fix a sense of what it means to be close. Let me look at an interval that goes from a - epsilon to a + epsilon. Then if my sequence converges to a, this means that as n increases, eventually the values of the sequence that I get stay inside this band. Since they converge to a, this means that eventually they will be smaller than a + epsilon and bigger than a - epsilon.

So convergence means that given a band of positive length around the number a, the values of the sequence that you get eventually get inside and stay inside that band. So that's sort of the picture definition of what convergence means. So now let's translate this into a mathematical statement.

Given a band of positive length, no matter how wide that band is or how narrow it is, so for every epsilon positive, eventually the sequence gets inside the band. What does eventually mean? There exists a time, so that after that time something happens. And the something that happens is that after that time, we are inside that band.

So this is a formal mathematical definition, which actually translates what I was telling in the wordy way before, and showing in terms of the picture. Given a certain band, even if it's narrow, eventually, after a certain time n0, the values of the sequence are going to stay inside this band.

Now, if I were to take epsilon to be very small, this thing would still be true that eventually I'm going to get inside of the band, except that I may have to wait longer for the values to get inside here. All right, that's what it means for a deterministic sequence to converge to something.

Now, how about random variables. What does it mean for a sequence of random variables to converge to a number? We're just going to twist a little bit of the word definition.

For numbers, we said that eventually the numbers get inside that band. But if instead of numbers we have random variables with a certain distribution, so here instead of a_n we're dealing with a random variable that has a distribution, let's say, of this kind, what we want is that this distribution gets inside this band, so it gets concentrated inside here. What does it means that the distribution gets inside this band?

I mean a random variable has a distribution. It may have some tails, so maybe not the entire distribution gets concentrated inside of the band. But we want that more and more of this distribution is concentrated in this band. So that -- in a sense that -- the probability of falling outside the band converges to 0 -- becomes smaller and smaller.

So in words, we're going to say that the sequence random variables or a sequence of probability distributions, that would be the same, converges to a particular number a if the following is true. If I consider a small band around a, then the probability that my random variable falls outside this band, which is the area under this curve, this probability becomes smaller and smaller as n goes to infinity. The probability of being outside this band converges to 0. So that's the intuitive idea.

So in the beginning, maybe our distribution is sitting everywhere. As n increases, the distribution starts to get concentrating inside the band. When a is even bigger, our distribution is even more inside that band, so that these outside probabilities become smaller and smaller.

So the corresponding mathematical statement is the following. I fix a band around a, a +/- epsilon. Given that band, the probability of falling outside this band, this probability converges to 0. Or another way to say it is that the limit of this probability is equal to 0.

If you were to translate this into a complete mathematical statement, you would have to write down the following messy thing. For every epsilon positive -- that's this statement -- the limit is 0.

What does it mean that the limit of something is 0? We flip back to the previous slide. Why? Because a probability is a number. So here we're talking about a sequence of numbers convergent to 0.

What does it mean for a sequence of numbers to converge to 0? It means that for any epsilon prime positive, there exists some n0 such that for every n bigger than n0 the following is true -- that this probability is less than or equal to epsilon prime.

So the mathematical statement is a little hard to parse. For every size of that band, and then you take the definition of what it means for the limit of a sequence of numbers to converge to 0. But it's a lot easier to describe this in words and, basically, think in terms of this picture. That as n increases, the probability of falling outside those bands just become smaller and smaller. So the statement is that our distribution gets concentrated in arbitrarily narrow little bands around that particular number a.

OK. So let's look at an example. Suppose a random variable Yn has a discrete distribution of this particular type. Does it converge to something? Well, the probability distribution of this random variable gets concentrated at 0 -- there's more and more probability of being at 0.

If I fix a band around 0 -- so if I take the band from minus epsilon to epsilon and look at that band-- the probability of falling outside this band is 1/n. As n goes to infinity, that probability goes to 0. So in this case, we do have convergence. And Yn converges in probability to the number 0. So this just captures the facts obvious from this picture, that more and more of our probability distribution gets concentrated around 0, as n goes to infinity.

Now, an interesting thing to notice is the following, that even though Yn converges to 0, if you were to write down the expected value for Yn, what would it be? It's going to be n times the probability of this value, which is 1/n. So the expected value turns out to be 1. And if you were to look at the expected value of Yn-squared, this would be 0. times this probability, and then n-squared times this probability, which is equal to n. And this actually goes to infinity.

So we have this, perhaps, strange situation where a random variable goes to 0, but the expected value of this random variable does not go to 0. And the second moment of that random variable actually goes to infinity. So this tells us that convergence in probability tells you something, but it doesn't tell you the whole story. Convergence to 0 of a random variable doesn't imply anything about convergence of expected values or of variances and so on.

So the reason is that convergence in probability tells you that this tail probability here is very small. But it doesn't tell you how far does this tail go. As in this example, the tail probability is small, but that tail acts far away, so it gives a disproportionate contribution to the expected value or the expected value squared.

OK. So now we've got everything that we need to go back to the sample mean and study its properties. So the sad thing is that we have a sequence of random variables. They're independent. They have the same distribution. And we assume that they have a finite mean and a finite variance. We're looking at the sample mean.

Now in principle, you can calculate the probability distribution of the sample mean, because we know how to find the distributions of sums of independent random variables. You use the convolution formula over and over. But this is pretty complicated, so let's not look at that. Let's just look at expected values, variances, and the probabilities that the sample mean is far away from the true mean.

So what is the expected value of this random variable? The expected value of a sum of random variables is the sum of the expected values. And then we have this factor of n in the denominator. Each one of these expected values is mu, so we get mu. So the sample mean, the average value of this Mn in expectation is the same as the true mean inside our population.

Now here, this is a fine conceptual point, there's two kinds of averages involved when you write down this expression. We understand that expectations are some kind of average. The sample mean is also an average over the values that we have observed.

But it's two different kinds of averages. The sample mean is the average of the heights of the penguins that we collected over a single expedition. The expected value is to be thought of as follows, my probabilistic experiment is one expedition to the South Pole. Expected value here means thinking on the average over a huge number of expeditions.

So my expedition is a random experiment, I collect random samples, and they record Mn. The average result of an expedition is what we would get if we were to carry out a zillion expeditions and average the averages that we get at each particular expedition. So this Mn is the average during a single expedition. This expectation is the average over an imagined infinite sequence of expeditions. And of course, the other thing to always keep in mind is that expectations give you numbers, whereas the sample mean is actually a random variable.

All right. So this random variable, how random is it? How big is its variance? So the variance of a sum of random variables is the sum of the variances. But since we're dividing by n, when you calculate variances this brings in a factor of n-squared. So the variance is sigma-squared over n.

And in particular, the variance of the sample mean becomes smaller and smaller. It means that when you estimate that average height of penguins, if you take a large sample, then your estimate is not going to be too random. The randomness in your estimates become small if you have a large sample size. Having a large sample size kind of removes the randomness from your experiment.

Now let's apply the Chebyshev inequality to say something about tail probabilities for the sample mean. The probability that you are more than epsilon away from the true mean is less than or equal to the variance of this quantity divided by this number squared. So that's just the translation of the Chebyshev inequality to the particular context we've got here. We found the variance. It's sigma-squared over n. So we end up with this expression.

So what does this expression do? For any given epsilon, if I fix epsilon, then this probability, which is less than sigma-squared over n epsilon-squared, converges to 0 as n goes to infinity. And this is just the definition of convergence in probability. If this happens, that the probability of being more than epsilon away from the mean, that probability goes to 0, and this is true no matter how I choose my epsilon, then by definition we have convergence in probability.

So we have proved that the sample mean converges in probability to the true mean. And this is what the weak law of large numbers tells us. So in some vague sense, it tells us that the sample means, when you take the average of many, many measurements in your sample, then the sample mean is a good estimate of the true mean in the sense that it approaches the true mean as your sample size increases. It approaches the true mean, but of course in a very specific sense, in probability, according to this notion of convergence that we have used.

So since we're talking about sampling, let's go over an example, which is the typical situation faced by someone who's constructing a poll. So you're interested in some property of the population. So what fraction of the population prefers Coke to Pepsi? So there's a number f, which is that fraction of the population. And so this is an exact number. So out of a population of 100 million, 20 million prefer Coke, then f would be 0.2.

We want to find out what that fraction is. We cannot ask everyone. What we're going to do is to take a random sample of people and ask them for their preferences. So the ith person either says yes for Coke or no. And we record that by putting a 1 each time that we get a yes answer.

And then we form the average of these x's. What is this average? It's the number of 1's that we got divided by n. So this is a fraction, but calculated only on the basis of the sample that we have. So you can think of this as being an estimate, f_hat, based on the sample that we have.

Now, even though we used the lower case letter here, this f_hat is, of course, a random variable. f is a number. This is the true fraction in the overall population. f_hat is the estimate that we get by using our particular sample.

Ok. So your boss told you, I need to know what f is, but go and do some sampling. What are you going to respond? Unless I ask everyone in the whole population, there's no way for me to know f exactly. Right? There's no way.

OK, so the boss tells you, well OK, then that'll me f within an accuracy. I want an answer from you, that's your answer, which is close to the correct answer within 1 % point. So if the true f is 0.4, your answer should be somewhere between 0.39 and 0.41. I want a really accurate answer.

What are you going to say? Well, there's no guarantee that my answer will be within 1 %. Maybe I'm unlucky and I just happen to sample the wrong set of people and my answer comes out to be wrong. So I cannot give you a hard guarantee that this inequality will be satisfied.

But perhaps, I can give you a guarantee that this inequality will be satisfied, this accuracy requirement will be satisfied, with high confidence. That is, there's going to be a smaller probability that things go wrong, that I'm unlikely and I use a bad sample. But leaving aside that smaller probability of being unlucky, my answer will be accurate within the accuracy requirement that you have.

So these two numbers are the usual specs that one has when designing polls. So this number is the accuracy that we want. It's the desired accuracy. And this number has to do with the confidence that we want. So 1 minus that number, we could call it the confidence that we want out of our sample. So this is really 1 minus confidence.

So now your job is to figure out how large an n, how large a sample should you be using, in order to satisfy the specs that your boss gave you. All you know at this stage is the Chebyshev inequality. So you just try to use it. The probability of getting an answer that's more than 0.01 away from the true answer is, by Chebyshev's inequality, the variance of this random variable divided by this number squared. The variance, as we argued a little earlier, is the variance of the x's divided by n. So we get this expression. So we would like this number to be less than or equal to 0.05.

OK, here we hit a little bit off a difficulty. The variance, (sigma_x)-squared, what is it? (Sigma_x)-squared is, if you remember the variance of a Bernoulli random variable, is this quantity. But we don't know it. f is what we're trying to estimate in the first place. So the variance is not known, so I cannot plug in a number inside here.

What I can do is to be conservative and use an upper bound of the variance. How large can this number get? Well, you can plot f times (1-f). It's a parabola. It has a root at 0 and at 1. So the maximum value is going to be, by symmetry, at 1/2 and when f is 1/2, then this variance becomes 1/4.

So I don't know (sigma_x)-squared, but I'm going to use the worst case value for (sigma_x)-squared, which is 4. And this is now an inequality that I know to be always true. I've got my specs, and my specs tell me that I want this number to be less than 0.05.

And given what I know, the best thing I can do is to say, OK, I'm going to take this number and make it less than 0.05. If I choose my n so that this is less than 0.05, then I'm certain that this probability is also less than 0.05.

What does it take for this inequality to be true? You can solve for n here, and you find that to satisfy this inequality, n should be larger than or equal to 50,000. So you can just let n be equal to 50,000. So the Chebyshev inequality tells us that if you take n equal to 50,000, then by the Chebyshev inequality, we're guaranteed to satisfy the specs that we were given.

Ok. Now, 50,000 is a bit of a large sample size. Right? If you read anything in the newspapers where they say so much of the voters think this and that, this was determined on the basis of a sample of 1,200 likely voters or so. So the numbers that you will typically see in these news items about polling, they usually involve sample sizes about the 1,000 or so. You will never see a sample size of 50,000. That's too much.

So where can we cut some corners? Well, we can cut corners basically in three places. This requirement is a little too tight. Newspaper stories will usually tell you, we have an accuracy of +/- 3 % points, instead of 1 % point. And because this number comes up as a square, by making it 3 % points instead of 1, saves you a factor of 10.

Then, the five percent confidence, I guess that's usually OK. If we use that factor of 10, then we make our sample that we gain from here, then we get a sample size of 10,000. And that's, again, a little too big. So where can we fix things?

Well, it turns out that this inequality that we're using here, Chebyshev's inequality, is just an inequality. It's not that tight. It's not very accurate. Maybe there's a better way of calculating or estimating this quantity, which is smaller than this. And using a more accurate inequality or a more accurate bound, then we can convince ourselves that we can settle with a smaller sample size.

This more accurate kind of inequality comes out of a difference limit theorem, which is the next limit theorem we're going to consider. We're going to start the discussion today, but we're going to continue with it next week.

Before I tell you exactly what that other limit theorem says, let me give you the big picture of what's involved here. We're dealing with sums of i.i.d random variables. Each X has a distribution of its own.

So suppose that X has a distribution which is something like this. This is the density of X. If I add lots of X's together, what kind of distribution do I expect? The mean is going to be n times the mean of an individual X. So if this is mu, I'm going to get a mean of n times mu.

But my variance will also increase. When I add the random variables, I'm adding the variances. So since the variance increases, we're going to get a distribution that's pretty wide. So this is the density of X1 plus all the way to Xn. So as n increases, my distribution shifts, because the mean is positive. So I keep adding things. And also, my distribution becomes wider and wider. The variance increases.

Well, we started a different scaling. We started a scaled version of this quantity when we looked at the weak law of large numbers. In the weak law of large numbers, we take this random variable and divide it by n. And what the weak law tells us is that we're going to get a distribution that's very highly concentrated around the true mean, which is mu.

So this here would be the density of X1 plus Xn divided by n. Because I've divided by n, the mean has become the original mean, which is mu. But the weak law of large numbers tells us that the distribution of this random variable is very concentrated around the mean. So we get a distribution that's very narrow in this kind. In the limit, this distribution becomes one that's just concentrated on top of mu. So it's sort of a degenerate distribution.

So these are two extremes, no scaling for the sum, a scaling where we divide by n. In this extreme, we get the trivial case of a distribution that flattens out completely. In this scaling, we get a distribution that gets concentrated around a single point.

Again, we look at some intermediate scaling that makes things more interesting. Things do become interesting if we scale by dividing the sum by square root of n instead of dividing by n. What effect does this have?

When we scale by dividing by square root of n, the variance of Sn over square root of n is going to be the variance of Sn over sum divided by n. That's how variances behave. The variance of Sn is n sigma-squared, divide by n, which is sigma squared, which means that when we scale in this particular way, as n changes, the variance doesn't change.

So the width of our distribution will be sort of constant. The distribution changes shape, but it doesn't become narrower as was the case here. It doesn't become wider, kind of keeps the same width. So perhaps in the limit, this distribution is going to take an interesting shape. And that's indeed the case.

So let's do what we did before. So we're looking at the sum, and we want to divide the sum by something that goes like square root of n. So the variance of Sn is n sigma squared. The variance of the sigma Sn is the square root of that. It's this number. So effectively, we're scaling by order of square root n.

Now, I'm doing another thing here. If my random variable has a positive mean, then this quantity is going to have a mean that's positive and growing. It's going to be shifting to the right.

Why is that? Sn has a mean that's proportional to n. When I divide by square root n, then it means that the mean scales like square root of n. So my distribution would still keep shifting after I do this division.

I want to keep my distribution in place, so I subtract out the mean of Sn. So what we're doing here is a standard technique or transformation where you take a random variable and you so-called standardize it. I remove the mean of that random variable and I divide by the standard deviation. This results in a random variable that has 0 mean and unit variance.

What Zn measures is the following, Zn tells me how many standard deviations am I away from the mean. Sn minus (n times expected value of X) tells me how much is Sn away from the mean value of Sn. And by dividing by the standard deviation of Sn -- this tells me how many standard deviations away from the mean am I.

So we're going to look at this random variable, which is just a transformation Zn. It's a linear transformation of Sn. S And we're going to compare this random variable to a standard normal random variable.

So a standard normal is the random variable that you are familiar with, given by the usual formula, and for which we have tables for it. This Zn has 0 mean and unit variance. So in that respect, it has the same statistics as the standard normal. The distribution of Zn could be anything -- can be pretty messy.

But there is this amazing theorem called the central limit theorem that tells us that the distribution of Zn approaches the distribution of the standard normal in the following sense, that probability is that you can calculate -- of this type -- that you can calculate for Zn -- is the limit becomes the same as the probabilities that you would get from the standard normal tables for Z.

It's a statement about the cumulative distribution functions. This quantity, as a function of c, is the cumulative distribution function of the random variable Zn. This is the cumulative distribution function of the standard normal. The central limit theorem tells us that the cumulative distribution function of the sum of a number of random variables, after they're appropriately standardized, approaches the cumulative distribution function over the standard normal distribution.

In particular, this tells us that we can calculate probabilities for Zn when n is large by calculating instead probabilities for Z. And that's going to be a good approximation. Probabilities for Z are easy to calculate because they're well tabulated. So we get a very nice shortcut for calculating probabilities for Zn.

Now, it's not Zn that you're interested in. What you're interested in is Sn. And Sn -- inverting this relation here -- Sn is square root n sigma Zn plus n expected value of X.

All right. Now, if you can calculate probabilities for Zn, even approximately, then you can certainly calculate probabilities for Sn, because one is a linear function of the other. And we're going to do a little bit of that next time. You're going to get, also, some practice in recitation. At a more vague level, you could describe the central limit theorem as saying the following, when n is large, you can pretend that Zn is a standard normal random variable and do the calculations as if Zn was standard normal.

Now, pretending that Zn is normal is the same as pretending that Sn is normal, because Sn is a linear function of Zn. And we know that linear functions of normal random variables are normal. So the central limit theorem essentially tells us that we can pretend that Sn is a normal random variable and do the calculations just as if it were a normal random variable.

Mathematically speaking though, the central limit theorem does not talk about the distribution of Sn, because the distribution of Sn becomes degenerate in the limit, just a very flat and long thing. So strictly speaking mathematically, it's a statement about cumulative distributions of Zn's. Practically, the way you use it is by just pretending that Sn is normal.

Very good. Enjoy the Thanksgiving Holiday.