PROFESSOR: OK, let's get started. We'll start a minute early and then maybe we can finish a minute early if we're lucky.

Today we want to talk about the laws of large numbers. We want to talk about convergence a little bit. We will not really get into the strong law of large numbers, which we're going to do later, because that's a kind of a mysterious, difficult topic. And I wanted to put off really talking about that until we got to the point where we could make some use of it, which is not quite yet.

So first, I want to review what we've done just a little bit. We've said that probability models are very natural things for real-world situations, particularly those that are repeatable. And by repeatable, I mean they use trials, which have essentially the same initial conditions. They're essentially isolated from each other. When I say they're isolated from each other, I mean there isn't any apparent contact that they have with each other.

So for example, when you're flipping coins, there's not one very unusual coin, and that's the coin you use all the time. And then you try to use those results to be typical of all coins.

Have a fixed set of possible outcomes for these multiple trials. And they have an essentially random individual outcomes. Now, you'll see there's a real problem here when I use the word "random" there and "probability models" there because there is something inherently circular in this argument. It's something that always happens when you get into modeling where you're trying to take the messy real-world and turn it into a nice, clean, mathematical model. So that really, what we all do, and we do this instinctively, is after we start getting used to a particular model, we assume

that the real-world is like that model.

If you don't think you do that, think again. Because I think everyone does. So you really have that problem of trying to figure out what's wrong with models, how to go to better models, and we do this all the time.

OK, for any model, an extended model-- in other words, an extended mathematical model, for a sequence or an n-tuple of independent identically distributed repetitions is always well-defined mathematically. We haven't proven that, it's not trivial. But in fact, it's true.

Relative frequencies and sample averages. Relative frequencies apply to events, and you can represent events in terms of indicator functions and then use everything you know about random variables to deal with them. Therefore, you can use sample averages. In this extended model, essentially become deterministic. And that's what the laws of large numbers say in various different ways. And beyond knowing that they become deterministic, our problem today is to decide exactly what that means.

The laws of large numbers specify what "become deterministic" means. They only operates within the extended model. In other words, laws of large numbers don't apply to the real world. Well, we hope they apply to the real world, but they only apply to the real world when the model is good because you can only prove the laws of large numbers within this model domain.

Probability theory provides an awful lot of consistency checks and ways to avoid experimentation. In other words, I'm not claiming here that you have to do experimentation with a large number of so-called independent trials very often because you have so many ways of checking on things. But every once in a while, you have to do experimentation. And when you do, somehow or other the idea of this large number of trials, and either IID trials or trials which are somehow isolated from each other. And we will soon get to talk about Markov models. We will see that with Markov models, you don't have the IID property, but you have enough independence over time that you can still get these sorts of results.

So anyway, the determinism in this large number of trials really underlies much of the value of probability. OK, in other words, you don't need to use this experimentation very often. But when you do, you really need it because that's what you use to resolve conflicts, and to settle on things, and to have different people who are all trying to understand what's going on have some idea of something they can agree on.

OK, so that's enough for probability models. That's enough for philosophy. We will come back to this with little bits and pieces now and then. But at this point, we're really going into talking about the mathematical models themselves.

OK, so let's talk about the Markov bound, the Chebyshev bound, and the Chernoff bound. You should be reading the notes, so I hope you know what all these things are so I can go through them relatively quickly.

If you think that using these lectures slides that I'm passing out plus doing the problems is sufficient for understanding this course, you're really kidding yourself. I mean, the course is based on this text, which explains things much more fully. It still has errors in it. It still has typos in It. But a whole lot fewer than these lecture slides, so you should be reading them and then using them try to get a better idea of what these lectures mean, and using that to get the better idea of what the exercises you're doing mean.

Doing the exercises does not do you any good whatsoever. The only thing that does you some good is to do an exercise and then think about what it has to do with anything. And if you don't do that second part, then all you're doing is you're building a very, very second rate computer. Your abilities as a computer are about the same for the most part as the computer in a coffee maker. You are really not up to what a TV set does anymore. I mean, TV sets can do so much computation that they're way beyond your abilities at this point. So the only edge you have, the only thing you can do to try to make yourself worthwhile is to understand these things because computers cannot do any of that understanding at all. So you're way ahead of them there.

OK, so what is the Markov model? What it says is, if y is a non-negative random variable-- in other words, it's a random variable, which only takes one non-negative sample values. If it has an expectation, expectation of y. And for any real y greater than 0, the probability that Y is greater than or equal to little y is the expected value of Y divided by little y.

The proof of it is by picture. If you don't like proofs by pictures, you should get used to it because we will prove a great number of things by pictures here. And I claim that a proof by picture is better than a proof by algebra because if there's anything wrong with it, you can see from looking at it what it is.

So we know that the expected value of Y is the integral under the complimentary distribution function. This square in here, this area of Y times probability of Y greater or equal to Y. The probability of Y greater than or equal, capital Y, the random variable y greater than or equal to the number, little y, is just that point right up there. This is the point y. This is the point probability of capital Y greater than little y. It doesn't make any difference when you're integrating whether you use a greater than or equal to sign or a greater than sign. If you have a discontinuity, the integral is the same no matter which way you look at it.

So this area here is y times the probability that random variable y is greater than or equal to number y. And all that the Markov bound says is that this little rectangle here is less than or equal to the integral under that curve. That's a perfectly rigorous proof.

We don't really care about rigorous proofs here, anyway since we're trying to get at the issue of how you use probability, but we don't want to have proofs which mislead you about things. In other words, proofs which aren't right. So we try to be right, and I want you to learn to be right. But I don't want you to start to worry too much about looking like a mathematician when you prove things.

OK, the Chebyshev inequality. If Z has a mean-- and when you say Z has a mean, what you really mean is the expected value of the absolute value of Z is finite. And it

has a variance sigma squared of Z. That's saying a little more than just having a mean.

Then, for any epsilon greater than 0. In other words, this bound works for any epsilon. The probability that the absolute value of Z less than the mean. In other words, that it's further away from the mean by more than epsilon, is less than or equal to the variance of Z divided by epsilon squared.

Again, this is a very weak bound, but it's very general. And therefore, it's very useful. And the proof is simplicity itself. You define a new random variable y, which is Z minus the expected value of Z quantity squared. The expected value of Y then is the expected value of this, which is just the variance of Z.

So for any y greater than 0, we'll use the Markov bound, which says the probability that random variable Y is greater than or equal to number little y is less than or equal to sigma of Z squared. Namely, the expected value of random variable Y divided by number Y. That's just the Markov bound.

And then, random variable Y is greater than or equal to number y if and only if the positive square root of capital Y-- we're dealing only with positive non-negative things here-- is greater than or equal to the square root of number y. And that's less than or equal to sigma Z squared over y.

And square root of Y is just the magnitude of Z minus Z bar. We're setting epsilon equal to square root of y yields the Chebyshev bound.

Now, that's something which I don't believe in memorizing proofs. I think that's a terrible idea. But that's something so simple and so often used that you just ought to think in those terms. You ought to be able to see that diagram of the Markov inequality. You ought to be able to see why it is true. And you ought to understand it well enough that you can use it. In other words, there's a big difference in mathematics between knowing what a theorem says and knowing that the theorem is true, and really having a gut feeling for that theorem.

I mean, you know this when you deal with numbers. You know it when you deal with

integration or differentiation, or any of those things you've known about for a long time. There's a big difference between the things that you can really work with because you understand them and you see them and those things that you just know as something you don't really understand. This is something that you really ought to understand and be able to see it.

OK, the Chernoff bound is the last of these. We will use this a great deal. It's a generating function bound. And it says, for any number, positive number z, and any positive number r greater than 0, such that the moment generating function-- the moment generating function of a random variable z is a function given the random variable z. It's a function of a real number r. And that function is the expected value of e to the rZ. It's called the generating function because if you start taking derivatives of this and evaluate them, that r equals 0, what you get is the various moments of z. You've probably seen that at some point.

If you haven't seen it, it's not important here because we don't use that at all. What we really use is the fact that this is a function. It's a function, which is increasing as r increases because you put-- well, it just does. And what it says is the probability that this random variable is greater than or equal to the number z-- I should really use different letters for these things. It's hard to talk about them-- is less than or equal to the moment generating function times e to the minus rZ.

And the proof is exactly the same as the proof before. You might get the picture that you can prove many, many different things from the Markov inequality. And in fact, you can. You just put in whatever you want to and you get a new inequality. And you can call it after yourself if you want.

I mean, Chernoff. Chernoff is still alive. Chernoff is a faculty member at Harvard. And this is kind of curious because he sort of slipped this in in a paper that he wrote where he was trying to prove something difficult. And this is a relationship that many mathematicians have used over many, many years. And it's so simple that they didn't make any fuss about it. And he didn't make any fuss about it. And he was sort of embarrassed that many engineers, starting with Claude Shannon, found this to

be extraordinarily useful, and started calling it the Chernoff bound. He was slightly embarrassed of having this totally trivial thing suddenly be named after him. But anyway, that's the way it happened. And now it's a widely used tool that we use all the time. So it's the same proof that we had before for any y greater 0, Markov says this. And therefore, you get that.

This decreases exponentially with Z, and that's why it's useful. I mean, the Markov inequality only decays as 1 over little y. The Chebyshev inequality decays as 1 over y squared. This decays exponentially with y. And therefore, when you start dealing with large deviations, trying to talk about things that are very, very unlikely when you get very, very far from the mean, this is a very useful way to do it. And it's sort of the standard way of doing it at this point. We won't use it right now, but this is the right time to talk about it a little bit.

OK, next topic we want to take up is really these laws of large numbers, and something about convergence. We want to understand a little bit about that. And this picture that we've seen before, we take X1, up to Xn as n independent identically distributed random variables. They each have mean expected value of X. They each have variance sigma squared. You let Sn be the sum of all of them. What we want to understand is, how does S sub n behave? And more particularly, how does Sn over n-- namely, the relative-- not the relative frequency, but the sample average of x behave when you take n samples?

So this curve shows the distribution function of S4, of S20, and of S50 when you have a binary random variable with probability of 1 equal to a quarter, probability of 0 equal to 3/4. And what you see graphically is what you can see mathematically very easily, too. The mean value of S sub n is n times X bar. So the center point in these curves is moving out within.

You see the center point here. Center point somewhere around there. Actually, the center point is at-- yeah, just about what it looks like. And the center point here is out somewhere around there.

And you see the variance, you see the variance going up linearly with n. Not with n

squared, but with n, which means the standard deviation is going up with the square root of n. That's sort of why the law of large numbers works. It's because the standard deviation of these random-- of these sums only goes up with the square root of n. So these curves, along with moving out, become relatively more compressed relative to how far out they are.

This curve here is relatively more compressed for its mean than this one is here. And that's more compressed relative to this one.

We get this a lot more easily if we look at the sample average. Namely, S sub n over n. This is a random variable of mean X bar. That's a random variable of variance sigma squared over n. That's something that you ought to just recognize and have very close to the top of your consciousness because that again, is sort of why the sample average starts to converge.

So what happens then is for n equals 4, you get this very "blech" curve. For n equals 20, it starts looking a little more reasonable. For n equals 50, it's starting to scrunch in and start to look like a unit step. And what we'll find is that the intuitive way of looking at the law of large numbers, or one of the more intuitive ways of looking at it, is that the sample average starts to look-- starts to have a distribution function, which looks like a unit step. And that step occurs at the mean. So this curve here keeps scrunching in. This part down here is moving over that way. This part over here is moving over that way. And it all gets close to a unit step.

OK, the variance of Sn over n, as we've said, is equal to sigma squared over n. The limit of the variance as n goes to infinity takes the limit of that. Don't even have to know the definition of a limit, can see that when n gets large, this get small. And this goes to 0. So the limit of this goes to 0.

Now, here's the important thing. This equation says a whole lot more than this equation says. Because this equation says how quickly that approaches 0. All this says is it approaches 0. So we've thrown away a lot that we know, and now all we know is this. This 3 says that the convergence is as 1/n. This doesn't say that. This just says that it converges.

Why would anyone in their right mind want to replace an informative statement like this with a not informative statement like this? Any ideas of why you might want to do that? Any suggestions?

AUDIENCE: Convenience.

PROFESSOR: What?

AUDIENCE: Convenience. Convenience. Sometimes you don't need to--

PROFESSOR: Well, yes, convenience. But there's a much stronger reason.

This is a statement for IID random variables. This law of large numbers, we want it to apply to as many different situations as possible. To things that aren't quite IID. To things that don't have a variance. And this statement here is going to apply more generally than this. You can have situations where the variance goes to 0 more slowly than 1/n if these random variables are not independent. But you still have this statement. And this statement is what we really need, so this really says something, which is called convergence and mean square. Why mean square? Because this is the mean squared. So obvious terminology. Mathematicians aren't always very good at choosing terminology that makes sense when you look at it, but this one does.

Definition is a sequence of random variables Y1, Y2, Y3, and so forth. Converges in mean square to a random variable Y if this limit here is equal to 0.

So in this case, Y, this random variable Y, is really a deterministic random variable, which is just the deterministic value, expected value of X. This random variable here is this relative frequency here. And this is saying that the expected value of the relative frequency relative to the expected value of X-- this is going to 0. This isn't saying anything extra. This is just saying, if you're not interested in the law of large numbers, you might be interested in how a bunch of random variables approach some other random variable.

Now, if you look at a set of real numbers and you say, does that set of real numbers

approach something? I mean, you have sort of a complicated looking definition for that, which really says that the numbers approach this constant. But a set of numbers is so much more simple-minded than a set of random variables. I mean, a set of random variables is-- I mean, not even their distribution functions really explain what they are. There's also the relationship between the distribution functions. So you're not going to find anything very easy that says random variables converge. And you can expect to find the number of different kinds of statements about convergence.

And this is just going to be one of them. This is something called convergence and mean square-- yes?

AUDIENCE: Going from 3 to 4, we don't need IID anymore? So they can be just--

PROFESSOR: You can certainly find examples where it's not IID, and this doesn't hold and this does hold. The most interesting case where this doesn't hold and this does hold is where you-- no, you still need a variance for this to hold. Yeah, so I guess I can't really construct any nice examples of n where this holds and this doesn't hold. But there are some if you talk about random variables that are not IID.

I ought to have a problem that does that. But so far, I don't.

Now, the fact that this sample average converges in mean square doesn't tell us directly what might be more interesting. I mean, you look at that statement and it doesn't really tell you what this complementary distribution function looks like. I mean, to me the thing that is closest to what I would think of as convergence is that this sequence of random variables minus the random variable, the convergence of that difference, approaches a distribution function, which is the unit step. Which means that the probability that you're anywhere off of that center point is going to 0.

I mean, that's a very easy to interpret statement. The fact that the variance is going to 0, I don't quite know how do interpret it, except through Chebyshev's law, which gets me to the other statement. So what I'm saying here is if we apply Chebyshev to that statement before number 3-- this one-- which says what the variance is. If we

apply Chebyshev, then what we get is the probability that the relative frequency minus the mean, the absolute value of that is greater than or equal to epsilon. That probability is less than or equal to sigma squared over n times epsilon squared.

You'll notice this is a very peculiar statement in terms of epsilon. Because if you want to make epsilon very small, so you get something strong here, this term blows up. So the way you have to look at this is pick some epsilon you're happy with. I mean, you might want these two things to be within 1% of each other.

Then, epsilon squared here is 10,000. But by making n big enough, that gets submerged. So excuse me. Epsilon squared is 1/10,000 so 1 over epsilon squared is 10,000. So you need to make n very large. Yes?

AUDIENCE: So that's why at times when n is too small and epsilon is too small as well, you can get obvious things, like it's less than or equal to a number greater than 1?

PROFESSOR: Yes. And this inequality is not much good because there's a very obvious inequality that works. Yes.

But the other thing is this is a very weak inequality. So all this is doing is giving you a bound. All it's doing is saying that when n gets big enough, this number gets as small as you want it to be. So you can get an arbitrary accuracy of epsilon between sample average and mean. You can get that with a probability 1 minus this quantity. You can make that as close to 1 as you wish if you increase n. So that gives us the law of large numbers, and I haven't stated it formally, all the formal jazz as in the notes. But it says, if you have IID random variables with a finite variance, the limit of the probability that Sn over n minus x bar, the absolute value of that is greater than or equal to epsilon is equal to 0 in the limit, no matter how you choose epsilon.

Namely, this is one of those peculiar things in mathematics. It depends on who gets the first choice. If I get to choose epsilon and you get to choose n, then you win. You can make this go to 0.

If you choose n and then I choose epsilon, you lose. So it's only when you choose first that you win. But still, this statement works. For every epsilon greater than 0,

this limit here is equal to 0.

Now, let's go immediately a couple of pages beyond and look at this figure a little bit because I think this figure tells what's going on, I think better than anything else. You have the mean of x, which is right in the center of this distribution function. As n gets larger and larger, this distribution function here is going to be scrunching in, which we sort of know because the variance is going to 0. And we also sort of know it because of what this weak law of large numbers tells us. And we have these.

If we pick some given epsilon, then we have-- if we look at a range of two epsilon, epsilon on one side of the mean, epsilon on the other side of the mean, then we can ask the question, how well does this distribution function conform to a unit step?

Well, one easy way of looking at that is saying, if we draw a rectangle here of width 2 epsilon around X bar, when does this distribution function get inside that rectangle and when does it leave the rectangle? And what the weak law of large numbers says is that if you pick epsilon and hold it fixed, then delta 1 is going to 0 and delta 2 is going to 0. And eventually-- I think this is dying out. Well, no problem.

What this says is that as n gets larger and larger, this quantity shrinks down to 0. That quantity up there shrinks down to 0. And suddenly, you have something which is, for all practical purposes, a unit step.

Namely, if you think about it a little bit, how can you take an increasing curve, which increases from 0 to 1, and say that's close to a unit step? Isn't this a nice way of doing it? I mean, the function is increasing so it can't do anything after it crosses this point here. All it can do is increase, and eventually it leaves again.

Now, another thing. When you think about the weak law of large numbers and you don't state it formally, one of the important things is you can't make epsilon 0 here and you can't make delta 0. You need both an epsilon and a delta in this argument. And you can see that just by looking at a reasonable distribution function.

If you make epsilon equal to 0, then you're asking, what's the probability that this sample average is exactly equal to the mean? And in most cases, that's equal to 0.

Namely, you can't win on that argument.

And if you try to make delta equal to 0. In other words, you ask-- then suddenly you're stuck over here, and you're stuck way over there, and you can't make epsilon small. So trying to say that a curve looks like a step function, you really need two fudge factors to do that. So the weak law of large numbers. In terms of dealing with how close you are to a step function, the weak law of large numbers says about the only thing you can reasonably say.

OK, now let's go back to the slide before. The weak law of large numbers says that the limit as n goes to infinity of the probability that the sample average is greater than or equal to epsilon equals 0. And it says that for every epsilon greater than 0.

An equivalent statement is this statement here. The probability that Sn over n minus x bar is greater than or equal to epsilon is a complicated looking animal, but it's just a number. It's just a number between 0 and 1. It's a probability. So for every n, you get a number up there. And what this is saying is that that sequence of numbers is approaching 0.

Another way to say that a sequence of numbers approaches 0 is this way down here. It says that for every epsilon greater than 0 and every delta greater than 0, the probability that this quantity is less than or equal to delta is-- this probability is less than or equal to delta for all large enough n. In other words, these funny little things on the edge here for this next slide, the delta 1 and delta 2 are going to 0. So it's important to understand this both ways.

And now again, this quantity here looks very much like-- these two equations look very much alike. Except this one tells you something more about convergence than this one does. This says how this goes to 0. This only says that it goes to 0. So again, we have the same thing. The weak law of large numbers says this weaker thing, and it says this weaker thing because sometimes you need the weaker thing.

And in this case, there is a good example. The weak law of large numbers is true, even if you don't have a variance. It's true under the single condition that you have

a mean. There's a nice proof in the text about that. It's a proof that does something, which we're going to do many, many times.

You look at a random variable. You can't say what you want to say about it, so you truncate it. You say, let me-- I mean, if you think a problem is too hard, you look at a simpler problem. If you're drunk and you drop a coin, you look for it underneath a light. You don't look for it where it's dark, even though you dropped it where it's dark. So all of us do that. If we can't solve a problem, we try to pose a simpler, similar problem that we can solve. So you truncate this random variable.

When you truncate a random variable, I mean you just take its distribution function and you chop it off to a certain point. And what happens then?

Well. at that point you have a variance. You have a moment generating function. You have all the things you want. Nothing peculiar can happen because the thing is bounded. So then the trick in proving the weak law of large numbers under these more general circumstances is to first truncate the random variable. You then have the weak law of large numbers.

And then the thing that you do is in a very ticklish way, you start increasing n, and you increase the truncation parameter. And if you do this in just the right way, you wind up proving the theorem you want to prove.

Now, I'm not saying you ought to read that proof now. If you're sailing along with no problems at all, you ought to read that proof now. If you don't quite have the kind of mathematical background that I seem to be often assuming in this course, you ought to skip that. You will have many opportunities to understand the technique later. It's the technique which is important, it's not the-- I mean, it's not so much the actual proof.

Now, the thing we didn't talk about, about this picture here, is we say that a sequence of random variables-- Y1, Y2, et cetera-- converges in probability to a random variable y if for every epsilon greater than 0 and every delta greater than 0, the probability that the n-th random variable minus this funny random variable is

greater than or equal to epsilon, is less than or equal to delta. That's saying the same thing as this picture says.

In this picture here, you can draw each one of the Y sub n's minus Y. You think of Y sub n minus Y as a single random variable. And then you get this kind of curve here and the same interpretation works. So again, what you're saying with convergence and probability is that the distribution function of Yn minus Y is approaching a unit step as n gets big. So that's really the meaning of convergence and probability. I mean, you get this unit step as n gets bigger and bigger.

OK, so let's review all of what we've done in the last half hour. If a random variable, generic random variable x, has a standard deviation-- in other words, if it has a finite variance. And if X1, X2 are IID with that standard deviation, then the standard deviation of the relative frequency is equal to the standard deviation of x divided by the square root of n. So the standard deviation is the relative of the sample average. Excuse me, sample average, not relative frequency. Of the sample average is going to 0 as n gets big.

In the same way, if you have a sequence of arbitrary random variables, which are converging to Y in mean square, then Chebyshev shows that it converges in probability. OK so mean square convergence implies convergence in probability. Mean square convergence is a funny statement which says that this sequence of random variables has a standard deviation, which is going to 0. And it's hard to see exactly what that means because that standard deviation is a complicated integral. And I don't know what it means.

But if you use the Chebyshev inequality, then it means this very simple statement, which says that this sequence has to converge in probability to Y.

Mean square convergence then implies convergence in probability. Reverse isn't true because-- and I can't give you an example of it now, but I've already told you something about it. Because I've said that's the weak law of large numbers continues to hold if the generic random variable has a mean, but doesn't have a variance because of this truncation argument.

Well, I mean, what it says then is a variance is not required for the weak law of large numbers to hold. And if the variance doesn't hold, then you certainly don't have convergence in mean square. So we have an example even though you haven't proven that that example works. You have an example where the weak law of large numbers holds, but convergence in mean square does not hold.

OK, and the final thing is convergence in probability really means that the distribution of Yn minus Y approaches the unit step. Yes?

AUDIENCE: So in general, convergence in probability doesn't imply convergence in distribution. But it holds in this special case because--

PROFESSOR: It does imply convergence in distribution. We haven't talked about convergence in distribution yet. Except it does not imply convergence in mean square, which is a thing that requires a variance. So you can have convergence in probability without convergence in mean square, but not the other way. I mean, convergence in mean square, you just apply Chebyshev to it, and suddenly-- presto, you have convergence in probability.

And incidentally, I wish all of you would ask more questions. Because we're taking this video, which is going to be shown to many people in many different countries. And they ask themselves, would it be better if I came to MIT and then I could sit-in class and ask questions? And then they see these videos and they say, ah, it doesn't make any difference, nobody ask questions anyway. And because of that, MIT will simply wither away at some point. So it's very important for you to ask questions now and then.

Now, let's go on to the central limit theorem. This sum of n IID random variables minus n times the mean-- in other words, we just normalized it to 0 mean. S sub n minus n x bar is a 0 mean random variable. And it has variance n times sigma squared. It also has second moment n times sigma squared.

And what that means is that you take Sn minus n times the mean of x and divide it by the square root of n times sigma. What you get is something which is 0 mean

and unit variance. So as you keep increasing n, this random variable here, Sn minus n x bar over the square root of n sigma, just sits there rock solid with the same mean and the same variance, nothing ever happens to it. Except it has a distribution function, and the distribution function changes.

I mean, you see the distribution function changing here as you let n get larger and larger. In some sense, these steps are getting smaller and smaller. So it looks like you're approaching some particular curve and when we looked at the Bernoulli case-- I guess it was just last time. When we looked at the Bernoulli case, what we saw is that these steps here were going as e to the minus difference from the mean squared divided by 2 times sigma squared. Bunch of stuff, but what we saw was that these steps were proportional to the density of a Gaussian. In other words, this curve that we're converging to is proportional to the distribution function of the Gaussian random variable.

We didn't completely prove that because all we did was to show what happened to the PMF. We didn't really integrate these things. We didn't really deal with all of the small quantities. We said they weren't important. But you sort of got the picture of exactly why this convergence to a normal distribution function takes place. And the theorem says this more general thing that this convergence does, in fact, take place. And that is what the central limit theorem says. It says that that happens not only for the Bernoulli case, but it happens for all random variables, which have a variance. And the convergence is relatively good if the random variables have a certain moment that can be awful otherwise.

So this expression on top then is really the expression of the central limit theorem. It says not only does the normalized sample average-- I'll call this whole thing the normalized sample average because Sn minus n X bar has variance square root of n times sigma sub x. So this normalized sample average has mean 0 and standard deviation 1. Not only does it have mean 0 and variance 1, but it also becomes closer and closer to this Gaussian distribution. Why is that important?

Well, if you start studying noise and things like that, it's very important. Because it

says that if you have the sum of lots and lots of very, very small, unimportant things, then what those things add up to if they're relatively independent is something which is almost Gaussian. You pick up a book on noise theory or you pick up a book which is on communication, or which is on control, or something like that, and after you read a few chapters, you get the idea that all random variables are Gaussian.

This is particularly true if you look at books on statistics. Many, many books on statistics, particularly for undergraduates, the only random variable they ever talk about is the normal random variable. For some reason or other, you're led to believe that all random variables are Gaussian.

Well, of course, they aren't. But this says that a lot of random variables, which are sums of large numbers of little things, in fact are close to Gaussian. But we're interested in it here for another reason. And we'll come to that in a little bit. But let me make the comment that the proofs that I gave you about the central limit theorem for the Bernoulli case-- and if you fill-in those epsilons and deltas there, that really was a valid proof. That technique does not work at all when you have a non-Bernoulli situation. Because the situation is very, very complicated. You wind up-- I mean, if you have a Bernoulli case, you wind up with this nice, nice distribution, which says that every step in the Bernoulli distribution, you have terms that are increasing and then terms that are decreasing.

If you look at what happens for a discrete random variable, which is not binary, you have the most god awful distribution if you try to look at the probability mass function. It is just awful. And the only thing which looks nice is the distribution function. The distribution function looks relatively nice. And why is hard to tell.

And if you look at proofs of it, it goes through Fourier transforms. In probability theory, Fourier transforms are called characteristics functions, but it's really the same thing. And you go through this very complicated argument. I've been through it a number of times. And to me, it's all algebra. And I'm not a person that just accepts the fact that something is all algebra easily. I keep trying to find ways of making sense out of it. And I've never been able to make sense out of it, but I'm

convinced that it's true. So you just have to sort of live with that.

So anyway, the central limit theorem does apply to the distribution function. Namely, exactly what that says. The distribution function of this normalized sample average does go into the distribution function of the Gaussian. The PMFs do not converge at all. And nothing else converges, but just that one thing.

OK, a sequence of random variables converges in distribution-- this is what someone was asking about just a second ago, but I don't think you were really asking about that. But this is what convergence in distribution means. It means it's the limit of the distribution functions of a sequence of random variables. Turns into the distribution function of some other random variable. Then you say that these random variables converge in distribution to Z. And that's a nice, useful thing. And the CLT, the Central Limit Theorem, then says that this normalized sample average converges in distribution to the distribution of a normal random variable. Many people call that density city phi and call the normal distribution phi. I don't know why. I mean, you've got to call it something, so many people call it the same thing.

This convergence and distribution is really almost a misnomer. Because when random variables converge in distribution to another random variable, I mean, if you say something converges, usually you have the idea that the thing which is converging to something else is getting close to it in some sense. And the random variables aren't getting close at all, it's only the distribution functions that are getting close.

If I take a sequence of IID random variables, all of them have the same distribution function. And therefore, a sequence of IID random variables converges. And in fact, it's converged right from the beginning to the same random variable, to the same generic random variable. But they're not at all close to each other. But you still call this convergence in distribution.

Why do we make such a big fuss about convergence in distribution? Well, primarily because of the central limit theorem because you would like to see that a sequence of random variables, in fact, starts to look like something that is interesting, which is

the Gaussian random variable after a while. It says we can do these crazy things that statisticians do, and that-- well, fortunately, most communication theorists are a little more careful than statisticians. Somebody's going to hit me for saying that, but I think it's true.

But the central limit theorem, in fact, does say that many of these sums of random variables you look at can be reasonably approximated as being Gaussian.

So what we have now is convergence in probability implies convergence in distribution. And the proof, I will-- on the slides, I always abbreviate proof by Pf. And sometimes Pf is just what it sounds like it, it's "poof." It is not quite a proof, and you have to look at those to get the actual proof.

But this says the convergence is a sequence of Yn's in probability means that it converges to a unit step. That's exactly what convergence in probability mean. It converges to a unit step, and ti converges everywhere but at the step itself. If you look at the definition of convergence in distribution, and I might not have said it carefully enough when I defined it back here. Oh, yes, I did. Remarkable. Often I make up these slides when I'm half asleep, and they don't always say what I intended them to say. And my evil twin brother comes in and changes them later.

But here I said it right. A sequence of random variables converges in distribution to another random variable Z if the limit of the distribution function is equal to the limit-- if the limit of the distribution function of the Z sub n is equal to the distribution function of Z. But it only says for all Z where this distribution function is continuous.

You can't really expect much more than that because if you're looking at a distribution-- if you're looking at a limiting distribution function, which looks like this, especially for the law of large numbers, all we've been able to show is that these distributions come in down here very close, go up and get out very close up there. We haven't said anything about where they cross this actual line. And there's nothing in the argument about the weak law of large numbers, which says anything about what happens right exactly at the mean. But that's something that's the central limit theorem says-- Yes?

AUDIENCE: What's the Zn? That's not the sample mean?

PROFESSOR: When we use the Zn's for the central limit theorem, then what I mean by the Z sub n's here is those normalized random variables Sn minus n x bar over square root of n times sigma. And in all of these definitions of convergence, the random variables, which are converging to something, are always rather peculiar. Sometimes they're the sample averages. Sometimes they're the normalized sample averages. God knows what they are. But what mathematicians like to do-- and there's a good reason for what they like to do-- is they like to define different kinds of convergence in general terms, and then apply them to the specific thing that you're interested in.

OK, so the central limit theorem says that this normalized sum converges in distribution to phi, but it only has to converge where the distribution function is continuous. Yes?

AUDIENCE: So the theorem applies to the distribution. Why doesn't it apply to PMF?

PROFESSOR: Well, if you look at the example we have here, if you look at the PDF for this normalized random variable, you find something which is jumping up, jumping up. If we look at it for n equals 50, it's still jumping up. The jumps are smaller. But if you look at the PDF for this-- well, if you look at the distribution function for the normal, it has a density. This PDF function for the things which you want to approach a limit never have a density. All the time they have a PDF. The steps are getting smaller and smaller. And you can see that here as you're up to n equals 50. You can see these little tiny steps here. But you still have a PMF.

You'll want to look at it in terms of density. You have to look at in terms of impulses. And there's no way you can say an impulse is starting to approach a smooth curve.

OK, so we have this proof that converges in probability, implies convergence in distribution. And since convergence in mean square implies convergence in probability, and convergence in probability implies convergence in distribution, we suddenly have the convergence in mean square implies convergence in distribution also. And you have this nice picture in the book of all the things that converge in

distribution. Inside of that-- this is distribution. Inside of that is all the things that converge in probability. And inside of that is all the things that converge in mean square.

Now, there's a paradox here. And what the paradox is, is that the central limit theorem says something very, very strong about how Sn over n-- namely, the sample average-- converges to the mean. The convergence in distribution is a very weak form of convergence. So how is this weak form of convergence telling you something that says specific about how a sample average converges to the mean? It tells you much more than the weak law of large numbers does. Because it tells you if you at this thing, it's starting to approach a normal distribution function.

And the resolution of that paradox-- and this is important I think-- is that the random variables converge in distribution to the central limit theorem are these normalized random variables. The ones that converge in probability are the things which are normalizing in terms of the mean, but you're not normalizing them in terms of variance.

So when you look at one curve relative to the other curve, one curve is a squashed down version of the other curve. I mean, look at those pictures we have for that example.

If you look at a sequence of distribution functions for S sub n over n, what you find is things which are squashing down into a unit step. If you look at what you have for the normalized random variables, normalized to unit variance, what you have is something which is not squashing down at all. It gives the whole shape of the thing.

You can get from one curve to the other just by squashing or expanding on the x-axis. That's the only difference between them. So the central limit theorem says when you don't squash, you get this nice Gaussian distribution function. The weak law of large numbers says when you do squash, you get a unit step.

Now, which tells you more? Well, if you have the central limit theorem, it tells you a lot more because it says, if you look at this unit step here, and you expand it out by

a factor of square root of n, what you're going to get is something that goes like this. The central limit theorem tells you exactly what the distribution function is at x bar. It tells you that that's converging to what? What's the probability that the sum of a large number of random variables is greater than n times the mean? What is it approximately?

AUDIENCE: 1/2.

PROFESSOR: 1/2. That's what this says. This is a distribution function. It's converging to the distribution function of the normal. It hits that point, the normal is centered on this x bar here. And it hits that point exactly at 1/2. This says the probability of being on that side is 1/2. The probability of being on this side is 1/2. So you see, the central limit theorem is telling you a whole lot more about how this is converging than the weak law of large numbers is.

Now, I come back to the question I asked you a long time ago. Why is the weak law of large numbers-- why do you see it used more often than the central limit theorem since it's so much less powerful? Well, it's the same answer as before. It's less powerful, but it applies to a much larger number of cases. And in many situations, all you want is that weaker statement that tells you everything you want to know, but it tells you that weaker statement for this enormous variety of different situations.

Mean square convergence applies to fewer things. Well, of course, convergence in distribution applies for even more things. But we saw that when you're dealing with the central limit theorem, all bets are off on that because it's talking about a different sequence of random variables, which might or might not converge.

OK, so finally, convergence with probability 1. Many people call convergence with probability 1 convergence almost surely, or convergence almost everywhere. You will see this almost everywhere.

Now, why do I want to use convergence with probability, and why is that a dangerous thing to do? When you say things are converging with probability 1, it sounds very much like you're saying they converge in probability because you're

using the word "probability" in each. The two are very, very different concepts. And therefore, it would seem like you should avoid the word "probability" in this second one and say convergence almost surely or convergence almost everywhere. And the reason I don't like those is they don't make any sense, unless you understand measure theory. And we're not assuming that you understand measure theory here.

If you wanted to do that first problem in the last problem set, you had to understand measure theory. And I apologize for that, I didn't mean to do that to you. But this notion of convergence with probability 1, I think you can understand that. I think you can get a good sense of what it means without knowing any measure theory. And at least that's what we're trying to do.

OK, so let's go on. We've already said that a random variable is a lot more complicated thing than a number is. I think those of you who thought you understood probability theory pretty well, I probably managed to confuse you enough to get you to the point where you think you're not on totally safe ground talking about random variables. And you're certainly not on very safe ground talking about how random variables converge to each other. And that's good because to reach a greater understanding of something, you have to get to the point where you're a little bit confused first. So I've intentionally tried to-- well, I haven't tried to make this more confusing than necessary. But in fact, it's not as simple as what elementary courses would make you believe.

OK, this notion of convergence with probability 1, which we abbreviate WP1, is something that we're not going to talk about a great deal until we come to renewal processes. And the reason is we won't need it a great deal until we come to renewal processes. But you ought to know that there's something like that hanging out there. And you ought to have some idea of what it is.

So here's the definition of it. A sequence of random variables convergence with probability 1 to some other random variable Z all in the same sample space. If the probability of sample points in omega-- now remember, that a sample point implies a value for each one of these random variables. So in a sense, you can think of a

sample point as, more or less, equivalent to a sample path of this sequence of random variables here.

OK, so for omega and capital Omega, it says the limit as n goes to infinity of these random variables at the point omega is equal to what this extra random variable is at the point-- and it says that the probability of that whole thing is equal to 1.

Now, how many of you can look at that statement and see what it means? Well, I'm sure some of you can because you've seen it before. But understanding what that statement means, even though it's a very simple statement, is not very easy. So there's the statement up there. Let's try to parse it. In other words, break it down into what it's talking about.

For each sample point omega, that sample point is going to map into a sequence of-- so each sample point maps into this sample path of values for this sequence of random variables.

Some of those sequences-- OK, this now is a sequence of numbers. So each omega goes into some sequence of numbers. And that also is unquestionably close to this final generic random variable, capital Z evaluated at omega.

Now, some of these sequences here, sequences of real numbers, we all know what a limit of real numbers is. I hope you do. I know that many of you don't. And we'll talk about it later when we start talking about the strong law of large numbers. But this does, perhaps, have a limit. It perhaps doesn't have a limit.

If you look at a sequence 1, 2, 1, 2, 1, 2, 1, 2, forever, that doesn't have a limit because it doesn't start to get close to anything. It keeps wandering around forever.

If you look at a sequence which is 1 for 10 terms, then it's 2 the 11th term, and then it's 1 for 100 terms, 2 for 1 more term, 1 for 1,000 terms, then 2 for the next term, and so forth, that's a much more tricky case. Because in that sequence, pretty soon all you see is 1's. You look for an awful long way and you don't see any 2's. That does not converge just because the definition of convergence. And when you work with convergence for a long time, after a while you're very happy that that doesn't

converge because it would play all sorts of havoc with all of analysis.

So anyway, there is this idea that these numbers either converge or they don't converge. When these numbers converge, they might or might not converge to this. So for every omega in this sample space, you have this sequence here. That sequence might converge. If it does converge, it might converge to this or it might converge to something else. And what this is saying here is you take this entire set of sequences here. Namely, you take the entire set of omega. And for each one of those omegas, this might or might not converge. You look at the set of omega for which this sequence does converge. And for which it does converge to Z of omega.

And now you look at that set and what convergence with probability 1 means is that that set turns out to be an event, and that event turns out to have probability 1. Which says that for almost everything that happens, you look at this sample sequence and it has a limit, which is this. And that's true in probability. It's not true for most sequences.

Let me give you a very quick and simple example. Look at this Bernoulli case, and suppose the probability of a 1 is one quarter and the probability of a 0 is 3/4. Look at what happens when you take an extraordinarily large number of trials and you ask for those sample sequences that you take. What's going to happen to them?

Well, this says that if you look at the relative frequency of 1's-- well, if they converge in this sense, if the set of relative frequencies-- excuse me. If the set of sample averages converges in this sense, then it says with probability 1, that sample average is going to converge to one quarter. Now, that doesn't mean that most sequences are going to converge that way. Because most sequences are going to converge to 1/2. There are many more sequences with half 1's and half 0's than there are with three quarter 1's and one quarter 0's-- with three quarter 0's and one quarter 1's. So there are many more of one than the other.

But those particular sequences, which have probability-- those particular sequences which have relative frequency one quarter are much more likely than those which have relative frequency one half. Because the ones with relative frequency one half

are so unlikely. That's a complicated set of ideas. You sort of know that it has to be true because of these other laws of large numbers. And this is simply extending those laws of large numbers one more step to say not only does the distribution function of that sample average converge to what it should converge to, but also for these sequences with probability 1, they converge. If you look at the sequence for long enough, the sample average is going to converge to what it should for that one particular sample sequence.

OK, the strong law of large numbers then says that if X1, X2, and so forth are IID random variables, and they have an expected value, which is less than infinity, then the sample average converges to the actual average with probability 1. In other words, it says with probability 1, you look at this sequence forever.

I don't know how you look at a sequence forever. I've never figured that out. But if you could look at it forever, then with probability 1, it would come out with the right relative frequency. It'll take a lot of investment of time when we get to chapter 4 to try to sort that out. Wanted to tell you a little bit about it, read about in notes a little bit in chapter 1, and we will come back to it. And I think you will then understand it.

OK, with that, we are done with chapter 1. Next time we will go into Poisson processes. If you're upset by all of the abstraction in chapter 1, you will be very happy when we get into Poisson processes because there's nothing abstract there at all. Everything you could reasonably say about Poisson processes is either obviously true or obviously false. I mean, there's nothing that's strange there at all. After you understand it, everything works. So we'll do that next time.