PROFESSOR: OK, so good afternoon. Today, we will review probability theory. So I will mostly focus on-- I'll give you some distributions. So probabilistic distributions, that will be of interest to us throughout the course. And I will talk about moment-generating function a little bit. Afterwards, I will talk about law of large numbers and central limit theorem.

Who has heard of all of these topics before? OK. That's good. And I'll try to focus more on a little bit more of the advanced stuff. Then a big part of it will be review for you. So first of all, just to agree on terminology, let's review some definitions.

So a random variable x-- we will talk about discrete and continuous random variables. Just to set up the notation, I will write discrete at x and continuous random variable as y for now. So they are given by its probability distribution-- discrete random variable is given by its probability mass function. f sum x I will denote.

And continuous is given by probability distribution function. I will denote by x sub y. So pmf and pdf. Here, I just use a subscript because I wanted to distinguish f of x and x of y. But when it's clear which random variable we're talking about, I'll just say f.

So what is this? A probability mass function is a function from the sample space to non-negative reals such that the sum over all points in the domain equals 1. The probability distribution is very similar. The function from the sample space non-negative reals, but now the integration over the domain. So it's pretty much safe to consider our sample space to be the real numbers for continuous random variables. Later in the course, you will see some examples where it's not the real numbers.

But for now, just consider it as real Numbers.

For example, probability mass function. If X takes 1 with probability 1/3 minus 1 of probability 1/3 and 0 with probability 1/3. Then our probability mass function is fx 1 equals fx minus 1 1/3, just like that. An example of a continuous random variable is if-- let's say, for example, if x of y is equal to 1 for all y and 0,1, then this is pdf of uniform random variable where the space is 0.

So this random variable just picks one out of the three numbers with equal probability. This picks one out of this. All the real numbers are between 0 and 1 with equal probability. These are just some basic stuff. You should be familiar with this, but I wrote it down just so that we agree on the notation.

OK. Both of the boards don't slide. That's good.

A few more stuff. Expectation-- probability first. Probability of an event can be computed as probability of a is equal to either sum of all points in a-- this probability mass function-- or integral over a set a depending on what you're using. And expectation, our mean is expectation of x is equal to the sum over all x, x times that. And expectation of y is the integral over omega. Oh, sorry. Space. y times.

OK. And one more basic concept I'd like to review is two random variables x1 x2 are independent if probability that x1 is in A and x2 is in B equals the product of the probabilities for all events A and B. OK. All agreed?

So for independence, I will talk about independence of several random variables as well. There are two concepts of independence-- not two, but several. The two most popular are mutually independent events and pairwise independent events. Can somebody tell me the difference between these two for several variables? Yes?

AUDIENCE: So usually, independent means all the random variables are independent, like x1 is independent with every others. But pairwise means x1 and x2 are independent, but x1, x2, and x3, they may not be independent.

PROFESSOR: OK. Maybe-- yeah. So that's good. So let's see-- for the example of three random

variables, it might be the case that each pair are independent. x4 and x2, x1 is independent with x2, x1 is independent with 3x, x2 is with x3. But altogether, it's not independent. What that means is, this type of statement is not true. So there are that say 1, a2, a3, for which this does not hold.

But that's just some technical detail. We will mostly just consider mutually independent events. So when we say that several random variables are independent, it just means whatever collection you take, they're all independent.

OK. So a little bit more fun stuff [? in this ?] overview. So we defined random variables. And one of the most universal random variable, our distribution is a normal distribution. It's a continuous random variable.

Our continuous random variable has normal distribution, is said to have normal distribution if n mu sigma if the probability distribution function is given as 1 over sigma square root 2 pi e to the minus x minus mu squared. For all reals. OK? So mu mean over-- that's one of the most universal random variable distributions, the most important one as well.

OK. So this distribution, how it looks like-- I'm sure you saw this bell curve before. It looks like this if it's n 0 1, let's say. And that's your y. So it's centered around the origin, and it's symmetrical on the origin. So now let's look at our purpose. Let's think about our purpose. We want to model a financial product or a stock, the price of the stock using some random variable.

The first thing you can try is to use normal distribution. Normal distribution doesn't make sense, but we can say the price at day n minus the price at day n minus 1 is normal distribution. Is this a sensible definition? That's not really. So it's not a good choice. You can model it like this, but it's not a good choice. There may be several reasons, but one reason is that it doesn't take into account the order of magnitude of the price itself.

So the stock-- let's say you have a stock price that goes something like that. And say it was $10 here, and $50 here. Regardless of where your position is at, it says

that the increment, the absolute value of increment is identically distributed at this point and at this point. But if you observed how it works, usually that's not normally distributed. What's normally distributed is the percentage of how much it changes daily. So this is not a sensible model, not a good model.

But still, we can use normal distribution to come up with a pretty good model. So instead, what we want is a relative difference to be normally distributed. That is the percent. The question is, what is the distribution of price? What does the distribution of price? So it's not a very good explanation. Because I'm giving just discrete increments while these are continuous random variables and so on.

But what I'm trying to say here is that normal distribution is not good enough. Instead, we want the percentage change to be normally distributed. And if that is the case, what will be the distribution of the random variable? In this case, what will be the distribution of the price? One thing I should mention is, in this case, if each discriminant is normally distributed, then the price at day n will still be a normal random variable distributed like that.

So if there's no tendency-- if the average daily increment is 0, then no matter how far you go, your random variable will be normally distributed. But here, that will not be the case. So we want to see what the distribution of pn will be in this case.

OK. To do that-- let me formally write down what I want to say. What I want to say is this. I want to define a log normal distribution y or log over random variable y such that log of y is normally distributed.

So to derive the problem to distribution of this from the normal distribution, we can use the change of variable formula, which says the following-- suppose x and y are random variables such that probability of x minus x-- for all x. Then f of y of the first-- of x of x is equal to y. h of x.

So let's try to fit into this story. We want to have a random variable y such that log-wise normally distributed. Here-- so you can put log of x here. If y is normally distributed, x will be the distribution that we're interested in. So using this formula,

we can find probability distribution function of the log normal distribution using the probabilities distribution of normal. So let's do that.

AUDIENCE: [INAUDIBLE], right?

PROFESSOR: Yes. So it's not a good choice. Locally, it might be good choice. But if it's taken over a long time, it won't be a good choice. Because it will also take negative values, for example.

So if you just take this model, what's going to happen over a long period of time is it's going to hit this square root of n, negative square root of n line infinitely often. And then it can go up to infinity, or it can go down to infinity eventually. So it will take negative values and positive values. That's one reason, but there are several reasons why that's not a good choice.

If you look at a very small scale, it might be OK, because the base price doesn't change that much. So if you model in terms of ratio, our if you model it in an absolute way, it doesn't matter that much. But if you want to do it a little bit more like our scale, then that's not a very good choice. Other questions? Do you want me to add some explanation? OK.

So let me get this right. y. I want x to be-- yes. I want x to be the log normal distribution. And I want y to be normal distribution or a normal random variable. Then the probability that x is at most x equals the probability that y is at most-- sigma. y is at most log x. That's the definition of log over distribution.

Then by using this change of variable formula, probability density function of x is equal to probability density function of y at log x times the differentiation of log x of 1 over x. So it becomes 1 over x sigma square root 2 pi 8 to the minus log x minus mu squared. So log normal distribution can also be defined at the distribution which has probability mass function of this. You can use either definition.

Let me just make sure that I didn't mess up in the middle. Yes. And that only works for x greater than 0. Yes?

AUDIENCE: [INAUDIBLE]?

PROFESSOR: Yeah. So all logs are natural log. It should be log ln. Yeah. Thank you.

OK. So question-- what's the mean of this distribution here? Yeah?

AUDIENCE: 1?

PROFESSOR: Not 1. It might be mu. Is it mu? Oh, sorry. It might be e to the mu. Because log x as a normal distribution had mean mu. Log x equals mu might be the center. If that's the case, x is e to the mu will be the mean. Is that the case? Yes?

AUDIENCE: Can you get the mu minus [INAUDIBLE]?

PROFESSOR: Probably right. I don't remember what's there. There is a correcting factor. I don't remember exactly what that is, but I think you're right.

So one very important thing to remember is log normal distribution are referred to in terms of the parameters mu and sigma, because that's the mu and sigma up here and here coming from the normal distribution. But those are not the mean and variance anymore, because you skew the distribution. It's no longer centered at mu. Log x is centered at mu, but when it takes exponential, it becomes skewed. And we take the average, you'll see that the mean is no longer e to the new.

So that doesn't give the mean. That doesn't imply that the mean is e to the sigma. That doesn't imply that the variance is something like e to the sigma. That's just totally nonsense. Just remember-- these are just parameters, some parameters. It's no longer mean or variance. And in your homework, one exercise, we'll ask you to compute the mean and variance of the random variable.

But really, just try to have it stick in your mind that mu and sigma is no longer mean and variance. That's only the case for normal random variables. And the reason we are still using mu and sigma is because of this derivation. And it's easy to describe it in those. OK.

So the normal distribution and log normal distribution will probably be the

distributions that you'll see the most throughout the course. But there are some other distributions that you'll also see. I need this. I will not talk about it in detail. It will be some exercise questions. For example, you have Poisson distribution or exponential distributions. These are some other distributions that you'll see.

And all of these-- normal, log normal, Poisson, and exponential, and a lot more can be grouped into a family of distributions called exponential family. So a distribution is called to be in an exponential family.

A distribution belongs to exponential family if there exists a theta, a vector that parametrizes the distribution such that the probability density function for this choice of parameter theta can be written as h of x times c of theta times the exponent of sum from i equal 1 to k.

Yes. So here, when I write only x, x should only depend on x, not on theta. When I write some function of theta, it should only depend on theta, not on x. So hx tix depends only on x and c theta on my value theta depends only on theta. That's an abstract thing. It's not clear why this is so useful, at least from the definition.

But you're going to talk about some distribution for an exponential family, right? Yeah. So you will see something about this. But one good thing is, they exhibit some good statistical behavior, the things-- when you group them into-- all distributions in the exponential family have some nice statistical properties, which makes it good.

That's too abstract. Let's see how log normal distribution actually falls into the exponential family.

AUDIENCE: So, let me just comment.

PROFESSOR: Yeah, sure.

AUDIENCE: The notion of independent random variables, you went over how the-- well, the probability density functions of collections of random variables if they're mutually independent is the product of the probability densities of the individual variables. And so with this exponential family, if you have random variables from the same

exponential family, products of this density function factor out into a very simple form. It doesn't get more complicated as you look at the joint density of many variables, and in fact simplifies to the same exponential family. So that's where that becomes very useful.

PROFESSOR: So it's designed so that it factors out when it's multiplied. It factors out well.

OK. So-- sorry about that. Yeah, log normal distribution. So take h of x 1 over x. Before that, let's just rewrite that in a different way. So 1 over x sigma squared 2 pi e to the minus log x [INAUDIBLE] squared. Square.

Can be rewritten as 1 over x times 1 over sigma squared 2 pi e to the minus log x square over 2 sigma square plus mu log x over sigma square minus m square. Let's write it like that. Set up hx equals 1 over x c of theta-- sorry, theta equals mu sigma. c theta is equal to 1 over sigma square 2 pi e to the minus mu square.

So you will parametrize this family in terms of mu with sigma. Your h of x here will be 1 over x. Your c theta will be this term and the last term here, because this doesn't depend on x. And then you have to figure out what wnt is. You can let w1 of x be log x square t1-- no, t1 of x be log x square, w1 of theta be minus 1 over 2 sigma square. And similarly, you can let t2 equals log x and w2 equals mu over sigma.

It's just some technicality, but at least you can see it really fits in. OK. So that's all about distributions that I want to talk about. And then let's talk a little bit more about more interesting stuff, in my opinion. I like this stuff better.

There are two main things that we're interested in. When we have a random variable, at least for our purpose, what we want to study is given a random variable, first, we want to study statistics. So we want to study this statistics, whatever that means. And that will be represented by the k-th moments of the random variable. Our k-th moment is defined as expectation of x to the k.

And a good way to study all the moments together in one function is a moment-generating function. So this moment-generating function encodes all the k-th

moments of a random variable. So it contains all the statistical information of a random variable. That's why moment-generating function won't be interesting to us. Because when you want to study it, you don't have to consider each moment separately. It gets a unified way. It gives a very good feeling about your function. That will be our first topic.

Our second topic will be we want to study its long-term our large-scale behavior. So for example, assume that you have a normal distribution-- one random variable with normal distribution. If we just have a single random variable, you really have no control. It can be anywhere. The outcome can be anything according to that distribution.

But if you have several independent random variables with the exact same distribution, if the number is super large-- let's say 100 million-- and you plot how many random variables fall into each point into a graph, you'll know that it has to look very close to this curve. It will be more dense here, sparser there, and sparser there.

So you don't have individual control on each of the random variables. But when you look at large scale, you know, at least with very high probability, it has to look like this curve. Those kind of things are what we want to study. When we look at this long-term behavior or large scale of behavior, what can we say? What kind of events are guaranteed to happen with probability, let's say, 99.9%?

And actually, some interesting things are happening. As you might already know, two typical theorems of this type will be in this topic. It will be law of large numbers and central limit theory.

So let's start with our first topic-- the moment-generating function. The moment-generating function of a random variable is defined as-- I write it as m sub x. It's defined as expectation of e to the t times x where t is some parameter. t can be any real.

You have to be careful. It doesn't always converge. So remark does not necessarily

exist. So for example, one of the distributions you already saw, it does not have moment-generating function. The log normal distribution does not have any moment-generating function. And that's one thing you have to be careful.

It's not just some theoretical thing. The statement is not something theoretical. It actually happens for some random variables that you encounter in your life. So be careful. And that will actually show some very interesting thing I will later explain. Some very interesting facts arise from this fact.

Before going into that, first of all, why is it called moment-generating function? It's because if you take the k-th derivative of this function, then it actually gives the k-th moment of your random variable. That's where the name comes from. It's for all integers.

And that gives a different way of writing a moment-generating function. Because of that, we may write the moment-generating function as a sum from k equals 0 to infinity, t to the k, k factorial, times a k-th moment. That's like the Taylor expansion. Because all the derivatives, you know what the functions would be. Of course, only if it exists. This might not converge.

So if moment-generating function exists, they pretty much classify your random variables. So if two random variables, x y, have the same moment-generating function, then x and y have the same distribution. I will not prove this theorem. But it says that moment-generating function, if it exists, encodes really all the information about your random variables. You're not losing anything.

However, be very careful when you're applying this theorem. Because remark, it does not imply that all random variables with identical k-th moments for all k has the same distribution. Do you see it? If x and y have a moment-generating function, and they're the same, then they have the same distribution.

This looks a little bit controversial to this theorem. It says that it's not necessarily the k-th set. Two random variables, which have identical moments-- so all k-th moments are the same for two variables-- even if that's the case, they don't necessarily have

to have the same distribution.

Which seems like it doesn't make sense if you look at this theorem. Because moment-generating function is defined in terms of the moments. If two random variables have the same moment, we have the same moment-generating function. If they have the same moment-generating function, they have the same distribution. There is a hole in this argument. Even if they have the same moments, it doesn't necessarily imply that they have the same moment-generating function. They might both not have moment-generating functions. That's the glitch.

Be careful. So just remember that even if they have the same moments, they don't necessarily have the same distribution. And the reason is because-- one reason is because the moment-generating function might not exist. And if you look in to Wikipedia, you'll see an example of when it happens, of two random variables where this happens.

So that's one thing we will use later. Another thing that we will use later, it's a statement very similar to that, but it says something about a sequence of random variables. So if x1, x2 up to xn is a sequence of random variables such that the moment-generating function exists, and it goes to infinity. tends to the function of some random variable t x for some random variable x for all t. Here, we're assuming that all moment-generating function exists.

So again, the situation is, you have a sequence of random variables. Their moment-generating function exists. And in each point t, it converges to the value of the moment-generating function of some other random variable x. And what should happen? In light of this theorem, it should be the case that the distribution of this sequence gets closer and closer to the distribution of this random variable x.

And to make it formal, to make that information formal, what we can conclude is, for all x, the probability xi is less than or equal to x tends to the probability that at x. So in this sense, the distributions of these random variables converges to the distribution of that random variable.

So it's just a technical issue. You can just think of it as these random variables converge to that random variable. If you take some graduate probability course, you'll see that there's several possible ways to define convergence. But that's just some technicality. And the spirit here is just really the sequence converges if its moment-generating function converges.

So as you can see from these two theorems, moment-generating function, if it exists, is a really powerful tool that allows you to control the distribution. You'll see some applications later in central limit theorem. Any questions?

AUDIENCE: [INAUDIBLE]?

PROFESSOR: This one? Why?

AUDIENCE: Because it starts with t, and the right-hand side has nothing general.

PROFESSOR: Ah. Thank you. We evaluated that theorem. Other questions? Other corrections?

AUDIENCE: When you say the moment-generating function doesn't exist, do you mean that it isn't analytic or it doesn't converge?

PROFESSOR: It might not converge. So log normal distribution, it does not converge. So for all non-zero t, it does not converge for log normal distribution.

AUDIENCE: [INAUDIBLE]?

PROFESSOR: Here? Yes. Pointwise convergence implies pointwise convergence. No, no. Because pointwise, this conclusion is also rather weak. It's almost the weakest convergence in distributions.

OK. The law of large numbers. So now we're talking about large-scale behavior. Let x1 up to xn be independent random variables with identical distribution. We don't really know what the distribution is, but we know that they're all the same. In short, I'll just refer to this condition as iid random variables later. Independent Identically-distributed random variables.

And let mean be mu, variance be sigma square. Let's also define x as the average of n random variables. Then the probability that x for all. All positive [INAUDIBLE].

So whenever you have identical independent distributions, when you take their average, if you take a large enough number of samples, they will be very close to the mean, which makes sense. So what's an example of this? Before proving it, example of this theorem in practice can be seen in the Casino.

So for example, if you're playing blackjack in a casino, when you're playing against the casino, you have a very small disadvantage. If you're playing at the optimal strategy, you have-- does anybody know the probability? It's about 48%, 49%. About 48% chance of winning. That means if you bet $1 at the beginning of each round, the expected amount you'll win is $0.48. The expected amount that the casino will win is $0.52.

But it's designed so that the variance is so big that this expectation is hidden, the mean is hidden. From the player's point of view, you only have a very small sample. So it looks like the mean doesn't matter, because the variance takes over in a very short scale. But from the casino's point of view, they're taking a very large end there.

So for each round, let's say from the casino's point of view, it's like they are taking enormous value of n, n here. And that means as long and they have the slightest advantage, they'll be winning money, and a huge amount of money.

And most games played in the casinos are designed like this. It looks like the mean is really close to 50%, but it's hidden, because they designed it so the variance is big. But from the casino's point of view, they have enough players to play the game so that the law of large numbers just makes them money. The moral is, don't play blackjack. Play poker.

The reason that the rule of law of large numbers doesn't apply, at least in this sense, to poker-- can anybody explain why? It's because poker, you're playing against other players. If you have an advantage, if your skill-- if you believe that

there is skill in poker-- if your skill is better than the other player by, let's say, 5% chance, then you have an edge over that player. So you can win money. The only problem is that because-- poker, you're not playing against the casino. Don't play against casino.

But they still have to make money. So what they do instead is they take rake. So for each round that the players play, they pay some fee to the casino. And how the casino makes money at the poker table is by accumulating those fees. They're not taking chances there. But from the player's point of view, if you're better than the other player, and the amount of edge you have over the other player is larger than the fee that the casino charges to you, then now you can apply law of large numbers to yourself and win.

And if you take an example as poker, it looks like-- OK, I'm not going to play poker. But if it's a hedge fund, or if you're doing high-frequency trading, that's the moral behind it. So that's the belief you should have. You have to believe that you have an edge. Even if you have a tiny edge, if you can have enough number of trials, if you can trade enough of times using some strategy that you believe is winning over time, then law of large numbers will take it from there and will bring you money profit.

Of course, the problem is, when the variance is big, your belief starts to fall. At least, that was the case for me when I was playing poker. Because I believed that I had an edge, but when there is really swing, it looks like your expectation is negative. And that's when you have to believe in yourself. Yeah. That's when your faith in mathematics is being challenged. It really happened. I hope it doesn't happen to you.

Anyway, that's proof of there's numbers. How do you prove it? The proof is quite easy. First of all, one observation-- expectation of x is just expectation of 1 over n times sum of xi's. And that bi-linearity just becomes the sum of. And that's mu. OK. That's good.

And then the variance, what's the variance there? That's the expectation of x minus

mu square, which is the expectation sum over all i's minus mu square. I'll group them. That's the expectation of 1 over n sum of xi minus mu square. i is from [? one 10. ?]

What did I do wrong? 1 over n is inside the square. So I can take it out and square my square. And then you're summing n terms of sigma square. So that is equal to sigma square. That means the effect of averaging end terms does not affect your average, but it affects your variance. It divides your variance by n.

If you take larger and larger n, your variance gets smaller and smaller. And using that, we can prove this statement. There's only one thing you have to notice-- that the probability that x minus mu is greater than epsilon. And you multiply this epsilon square. This will be less than or equal to the variance of x.

The reason this inequality holds is because variances x is defined as the expectation of x minus mu square. For all the events when you have x minus mu at least epsilon, you're multiplying factor x square will be at least epsilon square. This term will be at least epsilon square when you fall into this event.

So your variance has to be at least x. And this is known to be sigma square over n. So probability that x minus mu is greater than epsilon is at most sigma square over ne squared. That means if you take n to go to infinity, that goes to zero. So the probability that you deviate from the mean by more than epsilon goes to 0.

You can actually read out a little bit more from the proof. It also tells a little bit about the speed of convergence. So let's say you have a random variable x. Your mean is 50. You epsilon is 0.1. So you want to know the probability that you deviate from your mean by more than 0.1. Let's say you want to be 99% sure. Want to be 99% sure that x minus mu is less than 0.1, or x minus 50 is less than 0.1.

In that case, what you can do is-- you want this to be 0.01. It has to be 0.01. So plug in that, plug-in your variance, plug in your epsilon. That will give you some bound on n. If you have more than that number of trials, you can be 99% sure that you don't deviate from your mean by more than epsilon.

So that does give some estimate, but I should mention that this is a very bad estimate. There are much more powerful estimates that can be done here. That will give the order of magnitude-- I didn't really calculate here, but it looks like it's close to millions. It has to be close to millions.

But in practice, if you use a lot more powerful tool of estimating it, it should only be hundreds or at most thousands. So the tool you'll use there is moment-generating functions, something similar to moment-generating functions. But I will not go into it.

Any questions? OK. For those who already saw large numbers before, the name suggests there's also something called strong law of large numbers. In that theorem, your conclusion is stronger. So the convergence is stronger than this type of convergence.

And also, the condition I gave here is a very strong condition. The same conclusion is true even if you weaken some of the conditions. So for example, the variance does not have to exist. I can be replaced by some other condition, and so on. But here, I just want it to be a simple form so that it's easy to prove. And you at least get the spirit of what's happening.

Now let's move on to the next topic-- central limit theorem. So weak law of large numbers says that if you have IID random variables, 1 over n times sum over x i's converges to mu, the mean in some weak sense. And the reason it happened was because this had mean mu and variance sigma square over n. We've exploited the fact that variance vanishes to get this.

So the question is, what happens if you replace 1 over n by 1 over square root n? What happens if for the random variable is 1 over square root n times i? The reason I'm making this choice of 1 over square root n is because if you make this choice, now the average has mean mu and variance sigma square just as in xi's. So this is the same as xi.

Then what should it look like? If the random variable is the same mean and same variance as your original random variable, the distribution of this, should it look like

the distribution of xi? If mean if mu. Thank you very much. The case when mean is 0. OK. For this special case, will it look like xi, or will it not look like xi? If it doesn't look like xi, can we say anything interesting about the distribution of this?

And central limit theorem answers this question. When I first saw it, I thought it was really interesting. Because normal distribution comes up here. And that's probably one of the reasons that normal distribution is so universal. Because when you take many independent events and take the average in this sense, their distribution converges to a normal distribution. Yes?

AUDIENCE: How did you get mean equals [INAUDIBLE]?

PROFESSOR: I didn't get it. I assumed it if x-- yeah. So theorem-- let x1 x2 to xn be iid random variables with mean, this time, mu and variance, sigma squared.

And let v-- or Yn. Yn be square root n times 1 over n of xi is mu. Then the distribution of Yn converges to that of normal distribution with mean 0 and variance sigma.

What this means-- I'll write it down again-- it means for all x, probability that Yn is less than or equal to x converges the probability that normal distribution is less than or equal to x. What's really interesting here is, no matter what distribution you had in the beginning, if we average it out in this sense, then you converge to the normal distribution. Any questions about this statement, or any corrections? Any mistakes that I made? OK.

Here's the proof. I will prove it when the moment-generating function exists. So assumed that the moment-generating functions exists. So proof assuming m of xi exists. So remember that theorem.

Try to recall that theorem where if you know that the moment-generating function of Yn's converges to the moment-generating function of the normal, then we have the statement. The distribution converges. So that's the statement we're going to use. That means our goal is to prove that the moment-generating function of these Yn's converge to the moment-generating function of the normal for all t pointwise

convergence.

And this part is well known. I'll just write it down. It's known to be e to the t square sigma square over 2. That just can be compute. So we want to somehow show that the moment-generating function of this Yn converges to that. The moment-generating function of Yn is equal to expectation of e to t Yn. E to the t 1 over square root n sum of xi times mu.

And then because each of the xi's are independent, this sum will split into products. Product of-- let me split it better. Meets the expectation-- we didn't use independents yet. Sum becomes products of e to the t 1 over square root n xi of x mu. And then because they're independent, this product can go out. Equal to the product from 1 to n expectation e to the t times square root n.

OK. Now they're identically distributed, so you just have to take the n-th power of that. That's equal to the expectation of e to the t over square root n xi minus mu to the n-th power. Now we'll do some estimation. So use the Taylor expansion of this. What we get is expectation of 1 plus that t over square root n xi minus mu plus 1 over 2 factorial, that squared, t over square root n, xi minus mu squared plus 1 over 3 factorial, that cubed plus so on. Then that's equal to 1 to the n-th power.

The linearity of expectation, 1 comes out. Second term ix 0, because xi has mean mu. So that disappears. This term-- we have 1 over 2 t squared over n xi minus mu square. xi minus mu square, when you take expectation, that will be sigma square.

And then the terms after that, because we're only interested in proving that for fixed t, this converges-- so we're only proving pointwise convergence. You may consider t as a fixed number. So as n goes to infinity-- if n is really, really large, all these terms will be smaller order of magnitude than n, 1 over n. Something like that happens.

And that's happening because we're fixed. For fixed t, we have to prove it. So if we're seeing something uniformly about t, that's no longer true. Now we go back to the exponential form. So this is pretty much just e to that term 1 over 2 t square sigma square over n plus little o of 1 over n to the n square.

Now, that n can be multiplied to cancel out. And we see that it's e to t-square sigma square over 2 plus the little o of 1. So if you take n to go to infinity, that term disappears, and we prove that it converges to that. And then by the theorem that I stated before, if we have this, we know that the distribution converges. Any questions?

OK. I'll make one final remark. So suppose there is a random variable x whose mean we do not know, whose mean is unknown. Our goal is to estimate the mean. And one way to do that is by taking many independent trials of this random variable.

So take independent trials x1, x2 to xn, and use 1 over x1 plus xn as our estimator. Then the law of large numbers says that this will be very close to the mean. So if you take n to be large enough, you will more than likely have some value which is very close to the mean.

And then the central limit theorem tells you how the distribution of this variable is around the mean. So we don't know what the real value is, but we know that the distribution of the value that we will obtain here is something like that around the mean. And because normal distribution have very small tails, the tail distributions is really small, we will get really close really fast.