TheFourierTransformAndItsApplications-Lecture10

Instructor (Brad Osgood):First thing, a quick announcement – two quick announcements. The latest homework set is posted up on the Web, so you can get that at the usual course Web site.

Also, Thomas had to change his office hours just as – just for today; is that right? Okay. He'll be in Packard 107 today from 12:00 to 2:00 for those who wanna spend a little quality time with Thomas.

Any questions about anything? Anything on anybody's mind?

All right. Big day today. We're going to talk about – we're going to do our final application of convolution. I suppose I shouldn't say "final application of convolution" because it is the kind of operation that comes up repeatedly throughout the course. But sort of as the – as a last treatment of the kind of areas we've been talking about, I wanna talk about application convolution to the central limit theorem.

And this is one of my favorite topics because it just is so – it's such an important result, and it's such, in some sense, a surprising application of convolution. And the result itself is just so – I don't know. It has this air of mystery about it that it's a real – I think it's a real treat to see how it's – see how they – see how these ideas play out.

So I wanna talk about convolution and the central limit theorem. I will describe this, but it actually takes a little while to set up. We have to develop the appropriate language to get a precise statement of what the result is, and then understand how we're going to approach it. But this – the central limit theorem is, I say, one of the cornerstones of probability. And it's not only very important from a theoretical point of view from the development probability, but it's extremely important from the practical point of view. It has to do with the ubiquity of the bell-shaped curve, or why is it that so many things are distributed according to a Gaussian.

The central limit theorem, which, like all good things, has a three-letter acronym that goes along with it – the CLT somehow explains the universal – explains the bell-shaped curve – the universal appearance of the bell-shaped curve, that is to say the Gaussian in probability.

And there is a quote, actually, on this. There's a quote that's often repeated and connection with the central limit theorem. I put it in the notes, and I just wanted to read it to you. It's by G. Litmann, who's a French physicist.

He says something like, "Everyone believes in the normal approximation."

The normal approximation is another word for the Gaussian approximation for the bell-shaped curve.

Says, "Everyone believes in the normal approximation, the mathematicians because they think that with a," – excuse me – "the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact. It's got something for everyone."

Everybody buys it. Everybody believes it. And today, you will know why it's true.

Well, what does it say? Well, again, it's gonna take us a little while – a couple of iterations – before we get to a precise statement. What it says is, it says that most probabilities – I mean, this is sort of an intuitive or informal way of putting it. Most probabilities – some kind of average, really, is the way of thinking about it – in some kind of average sense are calculated or approximated, at least, as if they were distributed – as if they were determined by a bell-shaped curve – by a Gaussian.

The key word here, as it turns out in the precise understanding of this, is averaging. In an average sense – or averaging many outcomes or many outcomes contributing in an average sense to the final outcome are distributed according to a Gaussian.

Now, the picture of this, so – before I make that precise, the picture is something like this. If you have a Gaussian, we're going to be working with a number of standard Gaussians. I – we – in the past, we've used either the minus pi x squared. We're gonna use a slightly – we're gonna use a different – one that's scaled differently today. We've used either the minus pi x squared as a standard Gaussian. And that has the advantage that it's equal to its own Fourier transform.

For the central limit theorem, the normalization is to take – more standard normalization is to take – sorry, say p of x equals 1 over the square root of 2pi either to minus x squared over 2.

Now, that has the advantage – I'm going to be using a lot of terms today, by the way, that I'm not gonna completely define. I hope you will have read – you have read or will read the section on this so that you pick up some of the terminology that we're gonna use, like standard deviation, mean, variance, probability, distributions, etc.

If we went over all of those things, I wouldn't be able to get to the really core ideas in a short enough amount of time, sorry to say. So I'm counting on you to really pick up some of the background here on your own. And I wouldn't be surprised if you'd seen a lot of these things before. A matter of fact, I'd be surprised if you hadn't seen a lot of these things before.

So the reason why you take this as the standard Gaussian in this case is that it has mean 0 – mean or expected value 0, and it has standard deviation or variance 1.

And the way it comes up in calculating probability is, again, something that you have probably seen is the probability – the idea is that this is supposed to represent the distribution of all possible outcomes to a measurement or an experiment or whatever.

And the probability that a measurement lies within a certain range is given by the area under the curve.

So the probability that a measurement – that some sort of experiment or a measurement is between two numbers a and b is given by the area under the curve – the integral for a to b of either the minus – well, 1 over squared 2pi – either the minus x squared over 2, dx.

This has to be calculated numerically because you can't integrate directly to the minus x squared – either minus x squared over 2, or it does have – doesn't have a simple anti-derivative. That's the area under the Gaussian.

Now, there is a strong sense – actually, I haven't gotten to the precise name of the theorem, but even in the informal way we've been talking about it, there's a strong sense of universality of the central limit theorem. It says that most probabilities somehow are calculated or approximated by a Gaussian.

Every – in the average sense, about any – most probabilities that you're likely to run across – most measurements that you're likely to run across, if you average them all out, you find a – you find that they're distributed like a bell-shaped curve. Now, why should that be the case? Where's this element of universality coming from?

Well, let me give you some indication and actually show you why – not why yet, but at least give you some indication how convolution is coming into this by showing you a series of quite startling pictures, I think. So I'm gonna show some slides, if I may – some drawings, if I may. And I may because I'm the teacher. Sorry, I should've warned you beforehand. Hello, back there?

But first, a song.

So here's our friend, the screen, and there is the rectangle function – see 0 all the way up and all the way down and over. This is actually an appropriate lecture for Halloween time because of what you see – because of what you will see. Ah, the rectangle function. All of this buildup for the lousy rectangle function. Boy, that is impressive.

Now, you know what happens when you convolve the rectangle function with itself? You get the triangle function. A matter of fact, I even – we talked about that a number of times. I even showed this – I even showed that – I even asked you to do that in homework. And if you read this section, you've seen these pictures, but nevertheless, you have to see them in full glory.

So this rectangle function convolved with itself looks like a triangle function. Fine. And remember, that was already in the case – that was already interesting itself. That showed somehow that the convolution is some kind of smoothing or averaging operation – that the convolution of the discontinuous function of the rectangle with itself is continuous. It's smoother than the original function.

Now, let's look at the convolution of a rectangle with itself three times. It is smoother still, and it starts to look a little bell-shaped. You start to get a characteristic shape there that starts to resemble, in your imagination, the Gaussian. And if you take the rectangle convolved with a something – I wouldn't ask you to do that. This is all calculated with MATLAB, of course. I wouldn't ask you to actually carry out the integration and try to do this calculation. That's hard. I mean, it's actually a piece-wise defined curve, but it's getting smoother, and it's getting more bell-shaped.

And if you take the rectangle convolved with itself four times, you get something that is very decidedly bell-shaped.

Now, that's already, I think, kind of impressive. But after all, the rectangle is a pretty simple-looking function. There's nothing much to it. What's really spooky is if you start with something very much unlike a rectangle function – a random just assortment of – a random collection of measurement and just start to take in the convolution random function and start taking the convolution, you're gonna find similar results.

So for example, here is, again from the book – from the notes – here's a random signal. Now, what I mean by this is I just asked MATLAB to generate a bunch of random numbers in between – I think it's just in between 0 and 1, and then join them by straight lines. So just jumps around according to what the numbers are. So that's a well-defined – well-defined, so it's a random function. Somehow in the interval, it just jumps all the way around.

But okay. You can take the convolution of this thing with itself. And here's what you get. If you take the convolution of this with itself – I call it the function f, you get something that still jumps around, but it is now very much triangular, whereas the first one looked pretty random, pretty scattered all over the place.

This one, there's sort of a "it goes up; it goes back down again." Take the convolution three times, and you get something that looks like that. Now, it's still jaggedy. It's still – it has a lot of bumps into it, but look at it. I mean, it's getting kind of bell-shaped, and it's getting kind of smoother. And if you take the convolution four times with itself, you get something that looks like that – just four times.

Now, again, you can't see it quite so much on the resolution on the screen, but if I get close enough to it, I still see some – get some jaggedy edges, no doubt, no doubt. But they're getting smoothed out. And again, the overall shape of the curve is becoming more and more Gaussian.

Now, I don't know, but I really – it's pretty spooky. And I don't think there's any other word for it. I mean, something is really going on here that you just have to understand. I mean, it has to somehow be – it's just really spooky.

That's all good. Thank you very much. You can raise the screen.

Back to live action. So I wanna explain that. The content – the mathematical content of the central limit theorem is to explain that spookiness and to explain that universality that in the limit as something or other – as convolution is repeated or as we'll make the connection, as an averaging is taken – as an average is taken over more and more measurements, the result converges to something that is distributed according to a Gaussian. That's what we wanna get to.

So here is the setup. And again, it involves a certain – unavoidably, if you wanna get – if you wanna give a precise statement, that means you have to have precise language that goes into it. And I apologize ahead of time again for not defining all these terms. They're talked about more carefully in the notes, but I did wanna go under the assumption that you've seen these things so that we can get to the central point and not talk too much about background.

So there's the setup. The primitive notion here is a random variable and how it's distributed. Primitive notion – that is to say the first thing you define in order to talk about all other things is a random variable and how it's distributed.

Now, a random variable – let me break that a little bit more neatly. A random variable, and I owe this definition to Sam Savage, who's a friend of mine in management science and engineering – a random variable is a number you don't know yet. That means that it's an outcome of a certain measurement. You're gonna perform an experiment. You're gonna make a measurement, and then you're gonna get a number. But you don't know what that number is until you've run the experiment, so you somehow wanna keep the placeholder for doing the experiment.

So you call the random variable, generally speaking, by an uppercase letter and the actual measurement that you take by the corresponding lowercase letter. This is like Fourier transform but in reverse. So you call the random variable x and the measurement – that is, the value of the random variable, little x.

So again, intuitively, you think of X as the number you don't know yet. It's like measure the length of a pin. Measure the height of a person across an entire population. Or measure – or the random variable is height. The value of the random variable is the measurement of the particular height of the particular person. So it's a number that you have to compute by running an experiment – by doing a trial.

Now, what you are interested in is the distribution of measurements. And that's given by – or it's usually called the probability density function or the distribution associated with x. So you're interested in how the measurements x are distributed. The fraction of them within a certain range, the fraction of them between 1 and 2, the fraction of them between 2 and 3, and so on, and so on. This is given by a function p of x, which is non-negative. Or if I wanna indicate the [inaudible] on x, I may put a subscript in there – piece of x, capital X.

But it's sometimes called the probability density function or the distribution function or whatever. And it has a property that – not property, but by definition, the probability that the measurement lies in a range between a and b is the integral of p between a and b – the area under the curve.

So by definition, the probability that the measurements are gonna take a value between a and b is the integral of a to b of p of x, dx.

So whatever shape the curve is in – whatever shape the curve has, it's positive, I compute the area under the curve from a to b, and that gives me the probability.

Now, there's a special property of p, but only one, really two. One is that it's positive. The second is that the total of the area is 1. The measurement's gotta be somewhere, so the probability that it's between minus infinity and infinity is 1. So equals 1 – everybody gotta be someplace, right? Every measurement's gotta be somewhere. So that translates to a property of p, the integral for minus infinity to infinity of p of x; dx is equal to 1.

So again, in mathematical abstract treatments of this thing usually start by saying – by defining what you mean by random variable, by defining by what you mean by density function. And density function is just defined by the properties. It's greater or equal to 0. And its total integral is equal to 1. And then you define probability by this sort of integral.

And again, these things are probably familiar to you, and I apologize for not really talking about them – talking about the background too much. But I do wanna get to the punch line.

Now, here is the key result. Here is how averaging starts to come into this. Averaging can really help convolutions start to come into this. Here's how averaging and convolution, I should say, are joined in this study. So here's the thing that gets it – that launches the whole argument. The key result is suppose x1 and x2 are independent random variables – more on this in just a second – with distributions p, say 1 of x1, and p2 of x2.

So to say they're independent, that means that making a measurement of 1 doesn't affect the measurement of another. They're independent events. They don't have anything to do with each other. That's – I mean, that's sort of intuitive way of saying it, but again, it's the terminology that you have no doubt heard.

But the idea, again, is that they're numbers that you don’t know yet. And knowing one number doesn't affect knowing the other number. So making a measurement of x1 doesn't affect making a measurement of x2, and vice versa.

The question is how is the sum distributed? And the answer is by the convolution, it's distributed according to the convolution of the distributions of the separate random variables. That is, the distribution of x1 plus x2 is given by the convolution p1 convolved with p2 of – well, just p1 convolved with p2. I won't write the variable.

We'll bring averaging – averaging doesn't come into this quite at this stage. Just think of this as the distribution of the sum is given by the convolution of the distributions. Very impressive result. Very beautiful result.

Now, let me show you how this goes. I'll get a proof of this for you. The probability – here is the basic observation. The probability that x1 plus x2 for any – let me start – let me add one thing here. For any t – for any value of t, the probability that x1 plus x2 is less or equal to t is given by this integral – is given by a double integral over the region x1 plus x2 – I'll draw a picture in just a second. x1 plus x2 less than equal to t, p1 of x1 times p2 of x2, dx1, dx2.

Now, there's a certain amount to say here – certain amount I will say, and a certain amount that I won't say. Here is the region in the x1, x2 plane that I'm talking about. Here's x1; here's x2. The line x1 plus x2 equals t goes, say, like that. And we are interested in integrating under this region.

Now, two things are happening here. You're making a measurement of x1, and you're making a measurement of x2. You have a distribution for the random variable x1. You have a distribution for the random variable x2. Because they're independent, and this is where the assumption of independence comes in, to calculate the probability of both – something on both x1 and x2 – to say that x1 plus x2 is less than or equal to t means the values of the sum y underneath this line, you calculate the product. So it's like x1 is less than or equal to t.

This is not quite right, but there's an "and" coming in here, like x1 is less than or equal to t, and x2 is less than or equal to t. x1 plus x2 is less than or equal to t. I'm not saying this right.

You calculate the integral of the product of the distributions of the two random variables separately over the region that you're talking about. And the region lies below the line x1 plus x2 equals t.

There's more of an argument for this, again, in the notes. But I just wanna show you how the derivation goes once you accept this fact. And it's not so – if you think about this for a while, it's actually – it's not an unreasonable statement. A matter of fact, it's quite natural.

Now, what I'm gonna do is – this integral, because the region is a little bit complicated, I can't, as I have before, uncouple this integral into two iterated single integrals, at least so easily, not the way it is written now. The variables here are coupled, in some sense, because the region of integration is a little bit more complicated. It's not just a rectangle or something bounded by a horizontal or vertical line. But I can get to that picture if I make a change of variable in the integral. I want to work with this integral by making a change of variable – a relatively simple change of variable.

I wanna make a change of variable in the integral. Now, let me write down what it is, and then I'll explain what's going on here. The change of variable that I'm gonna make is it's –

the variables are coupled. I'm gonna let u equal x1, and I have to write this down because I wanna get it right – v equals x1 plus x2. That's – that defines u and v, my new variables in terms of the old variables x1 and x2 or the old variables in terms of the new variables are x1 is equal to u, and x2 is equal to v minus u.

So the line, x1 plus x2 is equal to t is described in the other variables by x1 plus x2 is u plus v minus u is just v is described by v equals t. In other words, geometrically, the picture is this. Here is the x1, x2 plane. Here's the line, x1 plus x2 equals t, and here's the region underneath it. Here is the uv plane over here. Here's the change of variable – sort of it takes you from the uv plane from the x1, x2 plane. Here is the line v equals t. This region corresponds to this region under that change of variables.

And I'm gonna make a corresponding – now, I have to carry that through and see what happens at the integral, and actually, there's a general theorem here that tells you how you make such a change of variables in a multiple integral, and I'm not gonna go through that again. But in this case, it's particularly simple. It is something you may not have seen or may not be as familiar with as the ordinary integration by substitution that you've done many times.

There's a little bit more involved in making a change of variable in a multiple integral – the so-called Jacobian transformation comes in. But as I like to say, if you're gonna design transistors, chips with millions and millions of transistors on them, you can God damn learn how to change variables in a double integral. I mean, it's not so hard. It's not beyond your abilities.

But nevertheless, I'm not gonna go – I won't go through the formula. How does it – how do the change of variable work? Change the variables in the integral. In this case, it's particularly simple. You just get the integral from over the region x1 plus x2 less than or equal to t, p1 of x1, p2 of x2, dx1, dx2 becomes the integral.

Oh, or I could describe the regions. Minus infinity – v is going from minus infinity up to t, u is going from minus infinity to plus infinity. And it is p1 of x1. x1 is u. p2 of x2, and x2 is v minus u, du, dd. That's what it turns out to be. Now, do I have my order of integration right? I have my order of integration wrong here, I guess. Sorry.

v is going from minus infinity up to t. u is going from minus infinity to infinity, so dv, du – yeah, to describe the region in the uv plane, d is going from minus infinity up to t. u is going from minus infinity to infinity. Now, I'm gonna swap the order of integration. Let me leave that down there. I don't wanna cut it off.

So that's equal to the integral from minus infinity up to t. These are constants. The limits of integration here are all constants, so it's easy just to swap the order of integration, something I could not have done so easily in the – when the integral was written in terms of – in the x1, x2 plane. But because all the limits of integration are constant, it becomes easy to switch the order of integration. So integral for minus t for minus infinity to infinity, p1 of u, p2 of v minus u, du, and then the results is integrated with respect to v.

Now there we are. We have eyes if only we but see, as you have heard me say before, what is inside the integral here? What is inside the inner integral? Is the integral p1 of u times p2 of v minus u integrative with respect to u for minus infinity to infinity. That's just our friend, the convolution. That's the integral for minus infinity up to t of p1 convolved with p2, v, our u, du. Evaluate it. Sorry, sorry, sorry, sorry, sorry. I'll get it; I'll get it. Evaluated v, and the results is integrated with respect to v.

The inner integral is the convolution of p1 and p2 at v. Integral for minus infinity of infinity, p1 of p2, v minus u, vu.

Where do we start? Where do we finish? We see that the probability that x1 plus x2 is less than or equal to t is given by the integral for minus infinity up to t of p1 convolved with p2, vdv. And that is enough to identify what you're integrating as the distribution for the probability of that random variable. And so this identifies this result, the fact that you get a probability by integrating, some function identifies this as the probability distribution. This identifies p1 convolved with p2 as the distribution of x1 plus x2. Two exclamation points – that's really quite a striking result. And it's an extremely important result.

Now, once again, this says that if you have two independent, random variables, x1 and x2, so again, the measurement of one does not affect the measurement of the other. You ask the question, how is their sum distributed?

The answer is their sum is distributed according to the convolution of each one. That's, in words, what we just showed here. And it's a very important, very fundamental result. It links the sum of random variables to convolution. It will link averages to convolutional – averages as come as in a little bit later. In a matter fact, it'll be just a few moments.

Now, also in the notes, you should take a look at – there's a discrete form of this. As a matter of fact, I think I – I think I started off by motivating the more general result by looking at the discrete form. That is, by rolling a couple of dice – rolling one dice, rolling two dice, looking at the average, looking at the sum of two dice and so on. And you see exactly this result come in in the discrete case.

So if you're looking for a little bit more intuition about why convolution should come in when you're considering probabilities and sums, then look at the discrete case the way it's described there because it also shows you – I think pretty convincingly – why it is that convolution should come in to describe the distribution of the sum.

We are, however, dealing with it – we're have to do everything in the continuous case – simply some things, actually. Some of the calculations are actually simpler in the continuous case than in the discrete case. But it's a good motivation.

Now, I should say, by the way, that the result for two random variables also holds for a sum of any finite number random variables. So you get a similar result for x1, x2. x1 plus x2 plus [inaudible] xn is distributed according to the convolution p1 star, p2 star, p3. If you know the convolution of each, if they're independent, then you – then the convolution – then the sum is distributed according to the convolution of all of them – same argument. We see – you just work inductively.

Now, with this, I can give the setup for the central limit theorem – the actual statement. We're not quite there. We need a little bit more – a few more assumptions, which are natural, actually, which are the kind of assumptions that come up often when you're applying this – these sorts of ideas in practice and actually figuring out the distributions for actual measurements, actual experiments.

So here's the setup for the CLT. So again, I have n random variables are n independent random variables just as before. And eventually, I'm gonna let n10 to infinity here. I'm also assume that they have the same distribution – that the distributions are the same for all the xis. That is, they're distributed in the same way.

Now, that may seem like a – quite a strong assumption. It is, in some sense. But it's also an assumption that is quite natural. That is, if you're performing an experiment – if the measurements may be uniformly distributed, they're different sorts of measurements, but there's no reason to think they're distributed any other way, that making a measurement of one aspect of something – one aspect of an experiment is different or distributed differently from making a different measurement of the experiment – that they all have the same distribution.

And like anything else in this business that has an acronym associated with it, you say the – that x1 up to xn are iid. That stands for independent identically distributed. I hate these things. Independent identically distributed – but you see this terminology a lot, so you should know it.

Now, let's call the distribution p. Call it a common distribution p. And some further normalization is also possible. Let's call the distribution p of x. So it's the same distribution for each one of the random variables. Then you can normalize further. And again, this is explained in the notes.

You can assume that the mean of all the random – each one of the random variables is 0. You can assume mean 0 – that is to – that means that the integral from minus infinity to infinity of x times p of x, dx is equal to 0. That's for all the random variables because they all have the same distribution. And you can assume – that's also called the expected value. And you can assume that the center of deviation, or variance – it's the square of the standard of deviation is 1.

I'm using the term "standard of deviation." I'm a little bit more used to that, actually. So I can assume standard of deviation is 1. And that amounts to the assumption that the integral for minus infinity to infinity of x squared, p of x. Once the mean is equal to 0, then the assumption that the standard deviation is equal to 1 is this assumption.

We always have – we already have another property of p that the integral for minus infinity to infinity to p of x, dx is equal to 1. The total area is equal to 1. p of x is positive, and the total area greater than or equal to 0. And the total area is equal to 1. We can further normalize to assume the mean of 0, which means that the integral of x, p of x, dx is equal to 0. And the standard deviation is equal to 1, which means this integral, x squared, p of x, is equal to 1.

Now, what is the distribution of the sum? That's what we wanna get to. So what is the distribution xn? Well, now, that's actually – let me call this sn, the sum. That's actually not quite what we wanna look at because it doesn't have quite the same qualities as the independent – as the separate and random variables.

The mean of sn is still 0. But the standard deviation scales, by the square root of n – but the standard deviation scales by the square root of n – that is to say the standard deviation of just the sum is the square root of n of sn is the square root of n.

We want the sum to be comparable to the separate random variables that are entering into it. So instead, we look at not just the sum, but this sum, sn over the square root of n is x1 plus xn divided by the square root of n. That will have mean 0 standard deviation 1.

Now, I can give you a statement of the central limit theorem. The statement of the central limit theorem says what happens to this average – this scaled average, sn divided by the square root of n, as n tenths to infinity. Say this – it says that the limit – there are different ways of stating it.

One way of stating it is this: The limit is n tenths to infinity of the probability that a is between – well, that the events that you're measuring – the average of events is between a and b, or should I say the scaled average of events is between a and b. The limit of that thing is n tenths to infinity as you take more and more measurements of average more and more or a greater number of measurements, it is the integral – whatever 2pi the integral from a to b of the corresponding Gaussian, which also has mean 0 and standard deviation 1.

So once again, this sum has mean 0 and standard deviation 1. You take greater and greater – you take more and more measurements. You take – you're averaging out a series of measurements over a whole bunch of experiments. How are your values distributed? What is the probability that your measurements are going to lie within a certain range is you take more and more measurements and average them all out. The answer is that in the limit, it's distributed according to a Gaussian. That probability is calculated as if you were calculating the probability based on a Gaussian.

So that means event is particularly large, for example, you're probably getting a pretty good approximation if you assume that the sum – that the scale sum over there is distributed approximately according to a Gaussian.

What I'm gonna show is – I'm gonna show – so this is one version of the central limit theorem. I'm gonna show an unintegrated form of this. Namely, if p of n of x is the distribution of that sum of x1 plus xn divided by the square root of n, then I'll show that piece of n itself tenth to the Gaussian, as n tenths to infinity. And piece to x tends to whatever the square root of 2pi, either to minus x squared over 2 as n tenths to infinity. If we know that, then the integrated form follows.

Now, looks like quite a complicated statement. Well, you know what piece of n is, actually. Piece of n is easy to write down.

The distribution – well, one thing at a time here. One thing at a time. So once again, p of x is the distribution for x1, x2, up to xn. They all have the same distribution. p convolved with itself n times – so that's p convolved with p convolved with p n times of x – that's the distribution for the sum. That's the key result. It relates to sum to the convolution is the distribution of x1 plus, plus xn. But I don't have that exactly. I have the scaled version of that.

And there's something a little extra that happens here. And again, I have to refer you to the notes for a derivation of this. Let me just write down what the result is. The distribution of x1 plus xn, if I scale it, is piece is the square root of n times the n full convolution evaluate the square root times x. So it actually both scales outside and inside the variable. That's a fact. That's sort of a mathematical fact that comes from change of variables in scaling and how probability distributions behave under change of variables.

So for that, I have to refer you to the notes. So just take that on faith right now, or look it up. So in other words, piece of n of x – that's the probability distribution of the scaled 1, x1 plus xn – [inaudible] square root of n – is given by this formula, the square root of n – the n full convolution of p with itself, evaluate the square root of n times x.

Now, how the hell are you gonna show that this n full convolution of the scaled version of this n full convolution tends to a Gaussian? I mean, it doesn't look like I made the problem any simpler. And convolution is given by a pretty complicated integral, and I have an n full. I don't just have a convolution of two functions. I have the convolution of n functions. And I wanna take the limit of that thing as n tenths to infinity.

The key to analyzing this is the Fourier transform because the Fourier transform turns convolution into multiplication. Multiplication is easier to analyze than convolution.

I wanna take the Fourier transform – do it down here. The board gets too covered up. Take the Fourier transform of that function, square root of n, p convolved with itself n times. Evaluate the square root of n times x. Now, what do you get? And the idea is that multiplication – that this turns convolution into multiplication – easier, we hope, to analyze.

Now, how do we do it? Well, I apply the stretch theorem. The Fourier transform of pn of x is the Fourier transform of the scaled convolution. So it is the square root of n – that comes out of the Fourier transform of the n full convolution evaluated at the square root of n times x. Now remember, this is not each individual function evaluated. I [inaudible] little careful here. This is not each individual function evaluated. This is where you have to be careful about your variables.

So let me try to get a sentence out here. It's not each individual function evaluated at the square root of n times x. It's the convolution evaluated at the square root of n times x. So it's this Fourier transform of some big, hairy function evaluate the square root of n times x. This is the square root of n divided – times one of the square root of n if I apply the shift theorem of the Fourier transform of p star n evaluated at s over the square root of n. Just applying the shift theorem.

Now the Fourier transform of the convolution is the product of the Fourier transform. It's an n full convolution, so I have the product of the Fourier transform of p with itself n times. So the square root of n here canceled one of the square root of n there. This is the Fourier transform of p raised to the nth power, evaluated it at s over the square root of n.

In other words, it's a Fourier – [inaudible] another way – it's a Fourier transform of p evaluated s over the square root of n raised to the nth power. Make sure you follow the statement there. I mean, the one thing that I didn't show you was why the distribution of the scale average looked like this. But once you know this, then you can find this Fourier transform just by applying the shift theorem.

And you have to – again, you have to be a little careful here. You have to realize it's the n full convolution evaluated at the square root of n times x, not each individual function, not each individual term in the convolution evaluated there, but the whole convolution there evaluated the square root of n times x.

Now, we are almost there, baby. Can you feel the excitement? I'm sure that you can. The Fourier transform of p and s over the square root of n is equal to the integral of minus infinity to infinity. I have to write down the formula for the Fourier transform, believe it or not.

e to the minus 2pi i, s over the square root of n, x, p of x, dx. Now, of all things, I'm gonna use the Taylor series – yes – for the complex exponential.

This is – this, I think, goes under the heading of dirty tricks. It's integral from minus infinity to infinity. What is the Taylor series of the complex exponential – any exponential? It's 1 minus 2pi i, s over the square root – let me write sx over the square root of n, 1 plus x plus x squared over 2 plus x cubed over 3 factorial and so on and so on.

So it's this term, 1, and there's a minus sign here because either the minus 2pi s and 2 x. And the next term is square, but there's an i there. And so it's become minus 2pi square – why don't I just write it out here – 2pi i – 2pi sx squared divided by square root of n, the whole thing squared, one half of this – minus one half. There we go – I got it, I got it, I got it – plus and so on and so on and so on – plus small terms. [Inaudible] terms – times p of x.

Check me here, check me here, check me here. So again, I just use e to the x or e to the anything. e to the x is 1 plus x plus x squared plus 2 factorial plus and so on and so on. That's all I'm using.

Now, watch this. This is the integral for minus infinity to infinity, p of x minus – let me do this – dx. Let me start to take out the terms here – minus integral for minus infinity to infinity of – there are constants out in front here, 2pi i s, s comes out over the square root of n, times the integral for minus infinity to infinity of x times p of x, dx, minus – what's the next term here?

2pi squared s squared, x squared. So 2pi squared – that's 4 pi squared, s squared divided by 2, gives me a 2pi squared, s squared divided by n times the integral for minus infinity to infinity of x squared, p of x, dx, plus other small terms, or terms that can be estimated.

I will say not much more about them now, but I can give you some more details about that later. The important thing is to just keep the first couple of terms. So all I did here was multiply p of x through by the first couple of terms in the Taylor series, and then there's another integral p of x times these high-order terms. Now, use the normalization. Integral p of x is 1. The integral of x times p of x – that's the mean. That's 0. And the integral of x squared, p of x is 1 because the variance – the standard deviation is 1. This integral is 1.

What results from this dirty trick? What's left? What's left is 1 minus 2pi squared, s squared over n plus – let me just say smaller in – plus error terms. I'll just call them small.

What has happened here? We gotta go, man. We have found that the Fourier transform of p at s over the square root of n is, let's say approximately 1 minus 2pi squared, s squared divided by n. That's the Fourier transform of the function p. We want the nth power of that, right? The Fourier transform of p that's over the square root of n to the n because that's the n full convolution that comes in is this raised to the nth power. 1 minus 2pi squared, s squared divided by n to the nth power.

Now you have to remember something from calculus. This – 1 minus 2pi squared, s squared over n to the n, 1 minus something over n to the n power, as n tends to infinity goes to an exponential. This quantity to the nth power – as n tends to infinity – n tends to an exponential. 1 minus 2pi squared, s squared over n to the n is approximately and actually tends to limit e to the minus 2pi squared over s squared, as n tenths to infinity. Well, it's approximately this. And it gets better and better as n tenths to infinity.

Now you take the [inaudible] Fourier transform. The Fourier transform is tending of the convolution of the distribution – the distribution is tending to this thing as n is tending to infinity. I think it's okay. Take the inverse Fourier transform, and you get the result because you know the Fourier transform. The Gaussian is the Gaussian. You know what to do with a scaled Gaussian by applying – by invoking the theorems. And you find – if I got it right – that the result of this is that the original function – fs of n – excuse me, ps of n – the distribution of the scaled average and random variables tends to a Gaussian as n tenths to infinity. It's unbelieveable.

And that wraps it up. We brought that baby home, only five minutes late. Next time, I may wrap this up a little –

[End of Audio]

Duration: 55 minutes