PROFESSOR: OK, we're ready for the second lecture today. We will start to get into a little technical material, which doesn't mean necessarily that it's more important. It just means that it's easier because it deals with mathematics.

I'm going to spend a little bit more time reviewing probability, as you learned it before. I want to review it at a slightly more fundamental level than what you're used to seeing it as. You will understand why as we go on because when we get into stochastic processes, we will find that there are lots of very peculiar things that happen. And when peculiar things happen, the only thing you can do is go back to basics ideas. And if you don't understand what those basic ideas are, then you're in real trouble.

So we'll start out by talking about expectations just a little bit. Distribution function of a random variable often says more than you're really interested in. In other words, a distribution function is a function from the whole sample space into the real numbers. And that's a very complicated thing in general. And the expectation is just one simple number. And with that one simple number, you get an idea of what that random variable is all about, whether it's big or it's little or what have you.

They're a bunch of formulas that you're familiar with for finding the expectation. If you have a discrete random variable, the usual formula is you take all of the possible sample values, multiply each of them by the probability of that sample value, and you sum it up. This is what you learned right at the beginning of taking probability. If you've never taken probability, you learned it in statistic classes just as something you don't know where it comes from. But it's there.

If you have a continuous random variable, a continuous random variable is one that

has a density. You can find the expectation there. If you have an arbitrary random variable and it's not negative, then there's this peculiar formula here, which I will point to.

I think I'll point to it. Ah yes, I will point to it. This formula down here, you might or might not have seen. And I hope by the end of this course, you'll realize that it's one, more fundamental and two, probably more useful than either of these two. And then there's a final one, which this final formula, I'll tell you a little bit about that when we get to it.

OK, the formula for the expected value in terms of the integral of the complimentary distribution function. There's a picture here which shows you how it corresponds to the usual thing you're used to for a discrete random variable. Namely what you're doing is you're integrating this complimentary distribution function, which is the probability that the random variable x is greater than any particular x along the axis here.

So you integrate this function along here. And according to what I'm trying to convince you of, just integrating that function gives you the expected value. And the reason is that this top little square here is a1 times the probability that x is equal to a1. Next one is a2 times the probability it's equal to a2 and so forth down. And you can obviously generalize this to any discrete random variable, which is non-negative. And I'm just talking about non-negative random variables for the moment.

If x has a density, the same argument applies to any Riemann sum for that integral. You can take integrals. You can break them up into little slices. If you break them up into little slices, you can represent it in this way. And presto, again, you get that this integral is equal to the expectation. And if you have any other thing at all, you can always represent it in terms of this Riemann sum.

Now why is it even more powerful then that? Well, it's more powerful then that because if you took measure theory-- which most of you presumably have not taken yet, and many of you might never take it-- you will find out that this is really the fundamental definition after all. And integration, when you look and measure

theoretic terms, instead of taking little slices that go this way, you wind up taking little slices that go that way. So that any way this is the fundamental definition of expectation.

If you're worried about whether expectations exist or not, why is this much nicer? Because what you're integrating here is simply a function, which is monotonic decreasing with x. In other words, if you try to integrate it by integrating this function out to some largest value and then chopping if off there, what you get is some number. If you extend this chopping off point out, what you get is a number which keeps increasing.

What can happen? As you take a number which is increasing, you either get to infinity or you get to some finite limit. Nothing else can happen. So there aren't any limiting problems here.

And when you take expectations in other ways, there are always questions that you have to ask. And they're often serious. So this is just a much nicer way of doing it. Anyway, that's the way we're going to do it.

And so now we go on. Oh, I should mention where the other formula comes from. This formula back here.

You get that by representing x as both the positive part of x plus the negative part of x. And if you want to see how to do that exactly, it's in the notes where it talks about first this and then this. So you just put the two together. And then you get an expected value.

A word about notation here, and there's nothing I can do about this. It's an unfortunate thing. When somebody says that the expected value of a random variable exists, what do they mean? Any engineer would try to integrate it and would either get something which was undefined, because it was infinite going this way. It's minus infinity going that way. And there's no way to put the two together.

If you get infinity going this way, something finite going that way, like with a non-negative random variable, it's kind of silly to say the expectation doesn't exist.

Because really what's happening is the expectation is infinite. Now mathematicians and everybody who writes books, everybody who writes papers, everybody-- I think-- defines expected value as existing only if it's finite.

In other words, what you're doing is taking this integral over the set of real values. And you don't allow plus infinity or minus infinity. So you say that the expectation does not exist if in fact it's infinite or it's minus infinity or it is undefined completely. And you say it's undefined in all of those cases.

And that's just a convention that everybody lives by. So the other way of saying this is if the expected value of the magnitude of the random variable is infinite, then the expectation doesn't exist. So we will try to say it that way sometimes when it's really important.

OK, let's go on to indicator random variables. You're probably familiar with these. For every event you can think of, an event is something which is true, which occurs when some set of the sample points occur and is not true otherwise. So the definition of an indicator random variable is that the indicator for an event a-- as a function of the sample space-- is equal to 1, if omega is in the event a, and 0 otherwise.

So if you draw the distribution function of it, the distribution function of the indicator function is 0 up until the point 0. Then it jumps up to 1 minus the probability of a. At 1, it jumps all the way up to 1. So it's simply a binary random variable.

So every event has an indicator random variable. Every indicator random variable has a binary random variable. So indicator random variables are very simple. Events are very simple because you can map any event into an indicator random variable [INAUDIBLE]. And this also says that since we want to talk about events very often, binary random variables are particularly important in this field.

OK, but what this really says now is that any theorem about random variables can be applied to events. This is one of the few examples I know where it's much harder to find the expectation by taking the complimentary distribution function and

integrating it. It's not hard. But it's far easier to take the probability that the indicator random variable is 0, which is 1 minus probability of a. The probability is equal to 1, which is probability of a, and take the expectation, which is the probability of a, and the standard deviation, which is the square root the probability of a times 1 minus the probability of a. So random variables are sort of trivial things in a way.

OK, let's go on to multiple random variables. Now here's something that's a trick question in a way. But it's a very important trick question. Is a random variable specified by its distribution function? We've already seen that it's not really specified by its density or by its probability mass function. But we've said a distribution function is a more general thing so that every random variable has a distribution function. Does the distribution function specify the random variable?

No, that's the whole reason for what Kolmogorov did back in 1933. Or at least it was one of the main reasons for what he was doing. He wanted to straighten out this ambiguity which runs through the field about confusing random variables with their distribution function. Random variables are functions from the sample space to the real numbers. And they're not anything else.

So if you want to really define a random variable, you not only have to know what that random variable is but you also have to know what its relationships are. It's like if you're trying to understand the person. You can't understand the person without understanding something about who they know, how they know them, all those other things. All those relationships are important. And it's the same with random variables. You got to know about all the relationships.

Many problems you can solve just in terms of distribution function. But ultimately you have to-- or ultimately in many cases, you have to deal with these joint distribution functions. And random variables are independent. If the joint distribution function is equal to the product of the distribution functions for all x1 to xn, and that same form carries over for density functions and for probability mass functions.

OK, if you have discrete random variables, the idea of independence is a whole lot more intuitive if you express it in terms of conditional probabilities. The conditional

probability that the random variable x takes on some sample value x given that the random variable y takes on a sample value y.

Just as one side comment here, when you're doing problems, you will very often want to leave out the subscripts here saying what random variables you're dealing with. And you will use either capital or small letters here mixing up the argument and the function itself, which everybody does. And it's perfectly all right. I suggest that you try not to do it for a while because you get so confused doing this, not being able to sort out what's a random variable and what's a real number.

A lot of wags say random variables are neither random, because they're functions of the sample space, nor are they variables. And both of those are true. That's immaterial here. It's just that when you start getting confused about a problem, it's important to sort out which things are random variables and which things are arguments.

Now this conditional probability is something you're all familiar with. But x and y are independent then if the probability of x conditional on y is the same as the probability of x not conditional on y. In other words, if observing what y is doesn't tell you anything about what x is, that's really your intuitive definition of independence.

It's what you use if you're dealing with some real world situation. And you're asking what does this have to do with that? And if this has nothing to do with that, the random variables over here have nothing to do with the random variables over there, you would say in the real world that these things are independent of each other. When you have a probability model, you say they're statistically independent of each other.

OK, so that's the relationship between the real world and the models that we're dealing with all the time. We call it independence in both cases. But it means somewhat different things in the two situations.

OK, next about IID random variables. What are they? Well, the joint distribution function has to be equal to the product of the individual distribution functions.

You notice I've done something funny here, which is a convention I always use. A lot of people use it. If you have a bunch of independent random variables, they all have the same distribution function. If they all have the same distribution function, it gets confusing to refer to their distribution functions as a distribution function of the random variable x1, distribution function of the random variable x2. It's nicer to just take a generic random variable x, which has the same distribution as all of these, and express this numerically in this way. You have the same product form for probability mass functions and for density functions. So this works throughout.

OK next, think about a probability model in which r, the set of real numbers, is a sample space. And x is some random variable on that sample space. Namely x then is a function from the real numbers on to the real numbers.

The interesting thing here, and what I'm saying here is obvious. It's something that you all know. It's something that you've all been using even before you started to learn about probability theory. But at the same time, you probably never thought about it in a serious enough way that you would really make sense out of it.

You can always create an extended probability model in which the Cartesian space r to the n-- in other words, the space of n real numbers-- is the sample space. And x1 to xn are independent identity distributed random variables.

This is not obvious. And that's something you have to prove. But it's not hard to prove. And all you have to do is start out with a probability model for one random variable. And then just define all products to be what they're supposed to be and go from the products to all unions and all intersections.

We're just going to assume that that's true because we have to assume it's true if we don't want to use any measure theory here. This is one of the easier things to show using measure theory. But it's something you are always used to.

When you think of a random experiment, when you think of playing dice with somebody or playing cards with someone, you are-- from the very beginning when you started to talk about odds or anything of that sort, you have always had the idea

that this is a game which you can play repeatedly. And each time you play it, it's the same game. But the outcome is different. But all the probabilities are exactly the same.

What this says is you can always make a probability model that way. You can always make a probability model which corresponds to what you've always believed deep in your hearts all your lives. And fortunately, that's true. Otherwise, you wouldn't use probability.

OK, so let's move on from that. The page of philosophy, I will stop doing that pretty soon. But I have to get across what the relationship is between the real world and these models that we're dealing with. Because otherwise, you as engineers or business people or financial analysts or whatever the heck you're going to become will start believing in your probability models. And you will cause untold damage by losing track of the fact that these are supposedly models of something. And you better think of what they're supposed to be models of.

OK, in order to do that, we're going to study the sample average, namely the sum of n random variables divided by n. That's the way you take sample averages. You add them all up. You divide by n.

The law of large numbers, which we're going to talk about very soon, says that s sub n over n essentially becomes deterministic as n becomes very large. What we mean by that-- and most of you have seen that in various ways. We will review it later today. Well, we'll do it today and on Wednesday.

And there's a big question of about what becoming deterministic means. But there is an essential idea there. The extended model, namely when you have one random variable, you create a very large number of them. If it corresponds to repeated experiments in the real world, then s sub n over n corresponds to the arithmetic average in the real world.

In the real world, you do take arithmetic averages. Whenever you open up a newspaper, somebody is taking an arithmetic average of something and says, gee,

this is significant. This shows what's going on someplace.

Models can have two types of difficulties. This paragraph is a little different than what I wrote in the handout because I realized what I wrote in the handout didn't make a whole lot of sense. OK, the two types of difficulties you have with models, especially when you're trying to model things by IID random variables. In one, a sequence of real world experiments is not sufficiently similar and isolated to each other to correspond to the IID extended model.

In other words, you want to model things so that each time, each trial of this experiment, you do the same thing but get a potentially different answer. Sometimes you rig things without trying to do so in such a way that these experiments are not independent of each other and in fact are very, very heavily biased.

You find people taking risk models in the financial world where they take all sorts of these things. And they say, oh, all right, these things are all independent of each other. They look independent.

And then suddenly a scare comes along. And everybody sells simultaneously. And you find out that all these random variables were not independent at all. They were very closely related to each other but in a way you never saw before.

OK, the other way that these models are not true or not valid is that the IID extension is OK. But the basic model is not right.

OK, in other words, you model a coin. It's coming out heads with probability 1/2. And somebody has put a loaded coin in. And the probability that it comes up heads is 0.45. And the probability that it comes up tails is 0.55.

And you might guess that this person always bets on tails and tries to get you to bet on heads. So in that case, the basic model that you're using is not OK.

So you have both of these kinds of problems. You should try to keep them straight. But we'll learn about these problems through study of the models. Namely, we're

not going to go through an enormous amount of study on how you can bias a coin or things of this sort.

OK, science, symmetry, analogies, earlier models, all of these are used to model real world situations. Let me again talk about an example I talked about a little bit last time because the model was so trivial that you probably understood everything about the model in the situation. But you didn't understand what it was illustrating.

You have two dice. One of them is red. And one of them is white. You roll them. By symmetry, each one comes up to be 1, 2, 3, up to 6, each with equal probability.

If you roll them with two hands or something, they're going to be independent of each other. And therefore, the probability of each pair of outcomes-- red is equal to i. White is equal to j. Probability of each one of those is going to be 136 because that's the size of the sample space.

Now you take two white dice. And you roll them. What's the sample space?

Well, as far as the real world is concerned, you can't distinguish a red 1 from a white 1 and a white 2 from a red 2. In other words, those two possibilities can't be distinguished. So you might say I want to use a sample space which corresponds to the-- what's the word I used here?-- finest grain possible outcome that you can observe. And who would do that? You'd be crazy to do that.

I mean, you have a nice model of rolling dice where each outcome has probability 136. And you would replace that with something where the probability of a 1, 1 is 136. But the probability of a 1, 2 is 1/18 because you can get a 2 in two different ways. You get a 2 in two different ways, which says you're really thinking about a red die and a white die. Otherwise, you wouldn't be able to say that.

So the appropriate model here is certainly to think in terms of a red die and a white die. It's what everybody does. They just don't talk about it.

OK, so the point that I'm trying to make here is that what you call a finest grain model is not at all clear. And if it's not at all clear in the case of dice, it sure as hell is

not clear in most of the kinds of problems you want to deal with. So you need something considerably more than that.

OK, so neither the axioms nor experimentation motivate this model. In other words, you really have to use common sense. You have to use judgment.

And all of you have that. It's just that by learning all this mathematics, you eventually start to think that maybe you shouldn't use your common sense. So I have to keep saying that no, you keep on using your common sense.

You want to learn what these models are about. You want to use your common sense also. And you've got to go back and forth between the two of them.

OK, that's almost the end of our philosophy. I guess one more slide. I'm getting tired of this stuff.

Comparing models for similar situations and analyzing limited and effective models helps a lot in clarifying fuzziness. But ultimately, as in all of science, some experimentation is needed. This is like any other branch of science. You need experimentation sometimes. You don't want to do too much of it because you'd always be doing experiments.

But the important thing is that the outcome of an experiment is a sample point. It's not a probability. You do an experiment. You get an outcome. And all you find is one sample point, if you do the experiment once. And there's nothing that lets you draw a probability from that.

The only way you can get things that you would call probabilities is to use an extended model, hope the extended model corresponds to the physical situation, and deal with these law of large numbers kind of things. You don't necessarily need IID random variables. But you need something that you know about between a large number of random variables to get from an outcome to something you could reasonably call a probability.

OK, so that's enough. Let's go on to the law of large numbers. Let's do it in pictures

first. So you can lie back and relax for a minute or stop being bored by all this stuff.

What I've done here is to take the simplest random variable I can think of, which as you might guess is a binary random variable. It's either 0 or 1. Here it's 1 with probability 1/4 and 0 with probability 3/4.

I have actually calculated these things. The distribution function of x1 plus x2 plus x3 plus x4, this point down here, is the probability of all 0's, I guess. And then you get the probability of all 0's plus 1, 1 and so forth.

Here's where you take the sum of 20 random variables. And you're looking at the distribution function of the number of 1's that you get. And it comes out like this.

Here you're looking at s50. You're adding up 50 random variables. And what's happening as far as the gross picture is concerned here? Well, the mean value of s sub n is the mean of a sum of random variables. And that's equal to n times a mean of a single random variable when you have identically distributed random variables or random variables that have the same mean.

The variance is equal to n times sigma squared. Namely, when you take the expected value of this quantity squared, all these cross terms are going to balance out with the mean when you do. I mean, all of you know how to find the variance of s sub n. I hope you know how to find that.

And when you do that, it increases with n. And the mean increases with n. The standard deviation, which gives you a picture of how wide the distribution is, only goes up as the square root of n.

This is really the essence of the weak law of large numbers. I mean, everything else is mathematical detail. And then if you go on beyond this and you talk about the sample average, namely the sum of these n random variables-- assume them IID again. In fact, assume for this picture that they're the same binary random variables.

You look at the sample average. You find the mean of the sample average. And it's

the mean of a single random variable.

You find the variance of it. Because of this n here and the squaring that you're doing, the variance of the sum divided by n, the sigma squared divided by n, what happens as n gets large? This variance goes to 0. What happens when you have a random variable, a sequence of random variables, all of which have the same mean and whose standard deviation is going to 0? Well, you might play around with a lot of funny kinds of things that you might think of as happening.

But essentially what's going on here is the nice feature that when you add all these things up, the distribution function gets scrunched down into a unit step. In other words, since the standard deviation is going to 0, the sequence of random variables-- since they all have the same mean-- they all have smaller and smaller standard deviations. The only way you can do that is to scrunch them down into a limiting random variable, which is deterministic.

And you can see that happening here. Namely the largest value is the black thing, which is getting smaller and smaller. And the left side is going that way. On the right side, it's going that way.

So it looks like it's approaching a unit step. That has to be proven. And there's a simple proof of it. And we'll see that. And you've all seen that before. And you've all probably said, ho-hum. But that's the way it is.

Now the next thing to look at for this same set of random variables, the same sum, is you look at the normalized sum, namely sn minus n times the mean. And you divide that by the square root of n times sigma. And what do you get?

Well, every one of these random variables-- for every n has mean 0, has mean 0 because the mean of sn is n times x bar. So you're subtracting off of the mean essentially. And every one of them has variance 1. So you've got a whole sequence of random variables, which are just sticking there at the same mean, 0, and at the same variance.

What's extraordinary when you do that, and you can sort of see this happening a

little bit, this curve looks like it's going into a fixed curve, which starts out sticking to 0. And then it gradually comes up. And it looks fairly smooth. It goes off this way.

And if you read a lot about this or if you think that all respectable random variables are Gaussian random variables, and I hope at the end of this course you will realize that only most respectable random variables are Gaussian random variables. There are many very interesting random variables that aren't. But what the central limit theorem says is that as you add up more and more random variables and you look at this normalized sum here, what you get is in fact the normal distribution, which is this strange integral here, that e to the minus x squared over 2 times the x.

Now what I want to do with the rest of our time is to show you why in fact that happens. I've never seen this proof of the central limit theorem before. I'm sure that some people have done it. I'm only going to do it for the case of a binomial distribution, which is the only place where this works.

But I think in doing this you will see why in fact that strange e to the minus x squared over 2 comes up. It sure is not obvious by looking at this problem. OK, so that's what we're going to do. And I'm hoping that after you see this, you will in fact understand why the central limit theorem is true as well as knowing that it's true.

OK, so let's look at the Bernoulli process. You have a sequence of binary random variables, each of them is IID, each of them is 1 with probability p. And a 0 with probability q equals 1 minus p. You add them all up. They're IID. And the question is, what does the distribution of the sum look like?

Well, it has a nice formula to it. It's that formula down there. You've probably seen that formula before. Let's get some idea of where it comes from and what it means.

Each n tuple that starts with k1's and then ends with n minus k0's, each one of those has the same probability. And it's p to the k times q to the n minus k. In other words, the probability you get a 1 on the first toss is p. The probability you get a 1 on the second toss also, since those are independent, probability you get two 1's in a row is p squared. Probably you get three 1's in a row is p cubed and so forth up to

k.

Because we're looking at the probability that the first k outputs are 1, so the probability of that is p to the k. That's this term. And the probability that the rest of them are all 0's is q to the n minus k.

And this is sometimes confusing to you because you often think that this is going to be maximized when k is equal to p over n. You have some strange view of the law of large numbers. Well no, this quantity-- if p is less than 1/2, it's going to be largest at k equals zero.

The most probable single outcome from n tosses of a coin, and it's a biased coin, it comes out 1's more often than 0's. 0's are more probable than 1's. The most probable output is all 0's. Very improbable, but that's the most probable of all these improbable things.

But as you probably know already, there are n choose k different n tuples, all of which have k1's in them and n minus k0's. If you don't know that, I didn't even put that in the text. I put most things there. This is one of those basic combinatorial facts.

Look it up in Wikipedia. You'll probably get a cleaner explanation of it there than anywhere else. But look it up in any elementary probability book or in any elementary combinatorics book. I'm sure that all of you have seen this stuff.

So when you put this together, the probability that the sum of a n random variables, all of which are binary, the probability of getting k1's is n choose k times p to the k times q to the n minus k. Now you look at that. And if k is 1,000 and if n is 1,000, I mean, your eyes boggle because you can't imagine what that number looks like. So we want to find out what it looks like. And here's a tricky way of doing it.

What we want to do is to see how this varies with k. And in particular, we want to see how it varies with k when n is very large and when k is relatively close to p times n. So what we're going to do is take the ratio of the probability of k plus 1 1's to the ratio of k1's. And what is that?

I've written it out. n choose k, n choose k plus 1-- which is this term here-- is n factorial divided by k plus 1 factorial times n minus k minus 1 factorial. This term here-- you put the n factorial down on the bottom, k factorial times n minus k quantity factorial. And then you take the p's and the q's. For this term here you have p to the k plus 1 q to the n minus k minus 1. And for this one here you have p to the k times p to the n minus k. All that stuff cancels out, which is really cute.

When you look at this term you have p to the k plus 1 over p to the k. That's just p. And here you have q to the n minus k minus 1 over q to the n minus k. That's just q in the denominator. So this goes into p over q.

This quantity here is almost a simple n factorial over n factorial is 1. k plus 1 factorial divided by k factorial is k plus 1. That's this term here. And the n minus k over n minus k minus 1 is n minus k. So this ratio here is just that very simple expression there.

Now this ratio is strictly decreasing in k. How do I see that? Well, as k gets bigger and bigger, what happens?

As k gets bigger, the numerator gets larger. The denominator-- excuse me, as k gets larger, the numerator gets smaller. The denominator gets larger. So the ratio of the two gets smaller. So this whole quantity here, as k gets larger and larger for fixed n, is just decreasing and decreasing and decreasing.

Now let's look a little bit at where this crosses 1, if it does cross 1. And what I claim here is that when k plus 1 is less than or equal to pn, what happens here? Well if I can do this-- I usually get confused doing these things.

But if k plus 1 is less than or equal to pn, this is the last of these choices here, then k is also less than or equal to pn. And therefore, n minus k is greater than-- in fact, k is strictly less than pn. And n minus k is strictly greater than n minus pn. And since q is 1 minus p, this is n times q.

OK, so you take this divided by this. And you take this divided by this. And sure

enough, this ratio here is greater than 1 any time you have a k which is smaller than what you think k ought to be, which is p over n. OK, so you have these three quantities here. Let me go on to the next slide.

Since these ratios are less than 1 when k is large, approximately equal to 1 when k is close to pn, and greater than 1 when it's smaller than 1, if I plot these things, what I find is that as k is increasing, getting closer and closer to pn, it's getting larger and larger. As k is increasing further, getting larger than pn, this ratio says that these things have to be getting smaller and smaller. So just from looking at this, we know that these terms have to be increasing for terms less than pn and have to be decreasing for terms greater than pn.

So this is a bell-shaped curve. We've already seen that. It might not be quite clear that it's bell-shaped in the sense that it kind of tapers off as you get smaller. Because these ratios are getting bigger and bigger, as k gets bigger and bigger, the ratio of this term to this term gets bigger and bigger. So what's happening there?

As this ratio gets bigger and bigger, these terms get smaller and smaller. But as these terms get smaller and smaller, they're getting closer and closer to 0. So even though they're going to 0 like a bat out of hell, they still can't get any smaller than 0. So they just taper down and start to get close to 0. So that is roughly how this sum of binary random variables behave.

OK, so let's go on and show that the central limit theorem holds for the Bernoulli process. And that's just as easy really. There's nothing more difficult about it that we have to deal with.

This ratio, as we've said, is equal to n minus k over k plus 1 times p over k. What we're interested in here-- I mean, we've already seen from the last slide that the interesting thing here is the big terms. And the big terms are the terms which are close to pn. So what we'd like to do is look at values of k which are close to pn.

What I've done here is to plot this as k minus the integer value of pn. So we get integers. What I'm going to assume now, because this gets a little hairy if I don't do

that, I'm going to assume that pn is equal to an integer. It doesn't make a whole lot of difference to the argument. It just leaves out a lot of terms that you don't have to play with. So we'll assume that pn is an integer.

With this example, we're looking at where p is equal to 1/4. Pn is going to be an integer whenever n is a multiple of 4. So things are fine then. If I try to make p equal to 1 over pi, then that doesn't work so well. But after all, no reason to chose p in such a strange way.

OK, so I'm going to look at this for a fixed value of n. I'm going to look at it as k increases for k less than pn. I'm going to look at it as it decreases for k greater than pn. And I'm going to define k to be equal to the i plus pn. So I'm going to put the whole thing in terms of i instead of k.

OK, so when I substitute i equals k minus pn for k here, what I'm going to get is this term. It's going to be the probability of pn plus i plus 1 over p of-- fortunately when you're using textures, you can distinguish different kinds of p's. I have too many p's in this equation.

This is the probability mass function. This is just my probability of a 1. And p's are things that you like to use a lot in probability. So it's nice to have that separation there.

OK, when I take this and I substitute it into that with k equal to i, what I get is n minus pn minus i. That's n minus k over pn plus i plus 1 times p over q. Fair enough, OK, all I'm doing is replacing k with pn plus i because I want i to be very close to 0 in this argument. Because I've already seen that these terms are only significant when i is relatively close to 0. Because when I get away from 0, these terms are going down very, very fast.

So when I do that, what do I get? I get n minus pn is equal to qn. That's nice.

So I have an nq here. I have a q here. I have a pn here. I have a p here.

I'm going to multiply this p by n. I'm going to multiply this q by n. And I'm going to

take a ratio of this pair of things. So when I take this ratio, I'm going to get nq over nq, which is 1. And the other terms there become minus i over nq.

In the denominator, I'm going to divide pn plus i plus 1 by p, by pn, which gives me 1 plus i plus 1 divided by np. So I get two terms, ratio of two terms, which are both close to 1 at this point and which are getting closer and closer to 1 as n gets larger and larger. Now let's take the logarithm of this.

Let me justify taking the logarithm of it in two different ways. One of them is that what we're trying to prove-- and I'm playing the game that all of you always play in quizzes. When you're trying to prove something, what do you do? You start at the beginning. You work this way. You start at the end. You work back this way. And you hope, at some point, the two things come together.

If they don't come together, you get to this point. And you say, obviously. And then you go to that point which leads to-- yeah.

[LAUGHTER]

PROFESSOR: OK, so I'm doing the same thing here. This probability that we're trying to calculate-- well, I've listed it here in terms of-- I have put it here in terms of a distribution function. I will do just as well if I can do it in terms of an PMF. And what I'd like to show is that the PMF of sn minus nX bar over square root of n times sigma is somehow proportional to e to the minus x squared over 2.

Now if I want to do that, it will be all right if I can take the logarithm of this term and show that it's a square nX. And if I want to show that this logarithm is a square nX, and I'm looking at the differentials at each time, what are the differentials going to be if the sum of the differentials is quadratic? If the sum of these differentials is quadratic, then the individual terms have to be linear.

If I take a bunch of linear terms, if I add up 1 plus 2 plus 3 plus 4 plus 5, you've all done this I'm sure. And down here you write n plus n minus 1 plus, plus 1. And what do you get? You get n times n plus 1 over 2. You can also approximate that by integrating. Whenever you add up a sum of linear terms, you get a square term.

And I'm just curious. How many of you have seen that? Good, OK. Well, it's only about 1/2. So it's something you've probably seen in high school. Or your haven't seen it at all.

So let's go on with this argument. OK, so I'm going to take the logarithm of this expression here. I'm going to take the logarithm. I'm going to have the logarithm of 1 minus i over nq minus the logarithm of 1 plus i plus 1 over np. And I'm going to use what I think of as one of the most useful inequalities that you will ever see, which is the natural log of 1 plus x.

If we use a power expansion, we get x minus x squared over 2 plus x cubed over 3 minus-- it's an alternating series. If x is negative, this term is negative. This term is negative. This term is negative.

And all this makes sense because if I draw this function here, logarithm of 1 plus x at x equals 0. This is equal to 0. It comes up with a slope of 1. And it levels off. And here it's going down very fast.

So these terms, you get these negative terms. And on the positive side, you get these alternating terms. So this goes up slowly, down fast.

The slope here is x, which is this term here. The curvature here gives you the minus x squared over 2. And the approximation, which is very useful here, is that the logarithm of 1 plus x, when x is small, is equal to x plus what we call little l of x. Namely something which goes to 0 faster than x as x goes to 0.

OK, all of you know that, right? Well, if you don't know it, now you know it. It's useful. You will use it again and again.

OK, so what we're going to do is-- well, that's pretty good. Where did I get to that point? I skipped something.

What I have shown is that this increment in the probability, in the PMF for s sub n, namely the increment as you increase i by 1, is linear in i. And in fact, the logarithm

of this increment is linear in i. So therefore, by what I was saying before, the logarithm of the actual terms should be rather than linear in i, they should be quadratic in i. So that's what I'm trying to do here. I just missed this whole term here.

What I'm interested in now is getting a handle on pn plus some larger value, j, divided by the probability of sn for pn. What am I trying to do here? I should've said what I was trying to do.

This term is just one term. It's fixed. Be nice if we knew what it was, we don't at the moment.

But I'm trying to express everything else in terms of that one unknown term. And what I'm trying to do is to show that the logarithm of this everything else is going to be quadratic in j. And if I can do that, then I only have one undetermined factor in this whole sum. And I can use the fact that PMF summed to 1 to solve the whole problem.

So I'm going to express the probability that we get pn plus j1's divided by the probability that we get pn1's. It's the sum of the probability pn plus i plus 1 over pn plus i. And we increase i. We start out at i equals 0. And then the denominator is probability of pn plus 0, which is this term.

And each time we increase i by 1, this term cancels out with the previous or the next value of this term. And when I get all done, all I have is this expression here.

Everybody see that? OK, I see a lot of-- if you don't see it, just look at it. And you'll see that this-- I think you'll see that this works.

OK, so now I take this expression here. This logarithm is this linear term here. What do I want to do?

I want to sum i from 0 up to j minus 1. What do I get when I sum i from 0 to j minus 1? I get this expression here.

I get j times j minus 1 divided by 2n. Oh, I was-- I skipped something. Let's go back

a little bit. Because it'll look like it was a typo.

When I took this logarithm and I applied this approximation to it, I got minus i over nq minus i over np minus 1 over np plus square terms in n. When I take i over nq minus i over np, I can combine those two things together. I can take ip over npq minus iq times npq. And q plus p is equal to 1.

So the numerator all goes away. And these two terms combine to be minus i over n times p times p times q. And I just has this one last little term left here. Don't know what to do with that. But then I add up all these terms.

This one is the one that leads to j times j minus 1 over 2 npq. This one is the one that leads to j over np. And I just neglect this term, which is negligible compared to j squared. I get minus j squared over 2 npq.

Let me come back later to say why I'm so eager to neglect this term except that that's what I have to do if I want to get the right answer. OK, so we'll see why that has to be negligible in just a little bit. But now this logarithm is coming out to be exactly the term that I want it to be.

So finally, the logarithm of the sum of these random variables pn plus j, namely j off the mean, divided by that if pn is equal to minus j squared over 2 npq plus some negligible terms. And this says when I exponentiate things, that the probability that sn is j off the mean is approximately equal to this term, the probability that it's right at the mean, times e to the minus j squared over 2 npq. What that is saying is that this sum of terms that I was looking at before-- this term here, this term, this term, this term, this term, and so forth down-- these terms here are actually going as minus j squared over 2 npq, which is what they should be going as if you have a Gaussian curve here.

OK, now there's one other thing we have to do, which is figure out what this term is. And if you look at this as an undetermined coefficient on these Gaussian-type terms and you think of what happens if I sum this over all i, well, if I sum it over all i what I'm going to get is the sum of all of these terms here, which are negligible except

where j squared is proportional to n.

So I don't have to sum them beyond the point where this approximation makes sense. So I want to sum all these terms. In summing these terms, when n gets very, very large, these things are dropping off very, very slowly. The curve is getting very, very wide. If i scrunch the curve back in again, what I get is a Riemann approximation to a normal density curve. Therefore, I can integrate it. And believe me. If you don't believe me, I'll go through it. And you won't like that.

When you go through this, what you get is in fact this expression right here, which says that when n gets very, very large and j is the offset from the mean and is proportional to-- well, it's proportional to the square root of n. Then what I get is this PMF here, which is in fact what the central limit theorem says.

And now if you go back and try to think of exactly what we've done, what we've done is to show that the logarithm of these differences here is in fact linear in i. Therefore, when you sum them, you get something which is quadratic in j. And because of that, all you have to do is normalize with a center term. And you get this.

The central limit theorem, especially for the binary case, is almost always done by using a Stirling approximation. And a Stirling approximation is one of these things which is black magic. I don't know any place except in William Feller's book where anyone talks about where this formula comes from.

If you now go back and look very carefully at this derivation, this tells you what the Stirling approximation is. Because if you do this for p equals q, what you're doing is actually evaluating n choose k where k is very close to and over 2. And that will tell you exactly what Stirling's approximation has to be.

In other words, that's a way of deriving Stirling's approximation. The very backward way of doing things it seems. But often backward ways are the best ways of doing these things.

OK, so I told you I would stop at some point and ask for questions. Yes?

AUDIENCE: Can you please go back one slide before this slide where can you neglect a term, which [INAUDIBLE], minus j over np.

PROFESSOR: Why did I neglect the j over np? OK, that's a good question. If you look at this curve here, and I put the j in. I can put the j in by just making this expression here look at one smaller value of j or one larger value of j. And you get something different whether you're looking at the minus side or the plus side.

In fact, if p is equal to q, this term cancels out. If p is not equal to q, what happens is that the central limit theorem is approximately symmetric. But in this first ordered term, it's not quite symmetric.

It can't be symmetric because this is p times n. And you have all these terms out to 1. And you have many, many fewer terms back to 0. So it has to be slightly asymmetric.

But it's only asymmetric over at most a unit of value here, which is not significant. Because as n gets bigger, these terms-- well, as I've done it, the terms do not get close together. But if I want to think of it as a normalized Gaussian curve, I have to make the terms close together. So that extra term is not significant.

I wish I had a nicer way of taking care of all the approximations here. I haven't put this in the notes because I still haven't figured out how to do that. But I still think you get more insight from doing it this way than you do by going through Stirling's approximation, all those pages and pages of algebra.

Anything else? OK, well, see you Wednesday then.