PROFESSOR: OK, I guess we're all set for getting close to the end, coming now to a race about whether we could say anything meaningful about Martingales or not. But I think we can. I want to spend a little time reviewing the Wald identity today and also sequential tests.

It turns out that last time on the slides-- I didn't get the thresholds confused-- I got hypothesis 0 and hypothesis 1 interchanged from the way we usually do them. And it doesn't make any difference. There's no difference between hypothesis 0 and hypothesis 1. And you can do it either way you want to.

But it gets very confusing when you switch from one to the other when you're halfway through an argument. So I'm going to go through part of that again today. And we will get revised slides on the web, so that if you want to see them with the hypotheses done in a consistent way, you will see it there. That should be on there by this afternoon I hope.

OK, so let's go on and review what Mr. Wald said. He was talking about a random walk. Random walk consists of a bunch of a sequence of IID random variables. The random walk consists of the sequence of partial sums of those random variables.

And the question is if this random walk is taking place and you have two thresholds, one at alpha and one of beta-- and beta is below 0 and alpha is above 0-- and you start at 0, of course, and the question is when do you cross one of these two thresholds and which threshold do cross? What's the probability of crossing each other, and everything else you can say about this problem.

And it turns out this is a very major problem as far as stochastic processes are concerned, because it comes up almost everywhere. It certainly comes up as far as hypothesis testing is concerned. It's probably the major problem there when you get into sequential analysis. It's the major problem there. So it's a very important problem.

And what Wald said was if you let the random variable J be the stopping time of this random walk, namely, the time in which the walk first crosses either alpha or crosses the beta, and then he said no matter what r you choose and the range of points where the moment generating function g sub X of r exists. Y You can pick any r in that range, and then what you get is this strange looking equality here.

And I pointed out last time it just wasn't all that strange, because if instead of using the stopping time of when you cross a threshold, if instead you used as a stopping time just some particular end. You go for some number of steps, and then you stop. And at that point, you have the expected value of E to the rsn. The expected value of E to the rsn by definition is the moment generating function at r of S sub n, which is exactly equal to the minus J times gamma n times gamma of r.

So all we're doing here, all that Wald did-- as it turns out, it was quite a bit-- was to say that when you replace a fixed end with a stopping time, you still get the same result. We're stating it here just for the case of two thresholds. Wald stated it in much general terms. We'll use it a more general terms, when we say more about martingales.

X, now remember is the underlying random variable. S sub n is the sum of the X's. If X bar is less than 0, and if gamma r star equals 0, r star is the r at which gamma of r equals 0. It's the second root of gamma of r.

Gamma of r, if you remember, looks like this. This is r star here. This is the expected value of X as the slope here. And we're assuming that X bar is less than 0 for this. I don't know how that greater than 0 got in.

And then what it says is the probability that SJ is greater than or equal to alpha is less than or equal to the minus alpha r star. And last time, remember, we went through a long, messy bunch of equations for that.

I looked at it again, and this is just the simple old Markov inequality again. All you do to get this is you say, OK, think of the random variable, E to the r SJ. SJ is a very complicated random variable, but it's a random variable, nonetheless. So either the rSJ is a random variable and the expected value of that random variable is at r star, the expected value of it is just one.

I'll write down. It'll be easier. The expected value e to the r star S sub J is equal to 1. And therefore, the probability that E to the r star SJ is a greater than or equal to E to the r star alpha is just less than or equal to 1 over E to the r star alpha, OK? And that's what the inequality says. So that's all there is to it.

OK, what?

AUDIENCE: I don't really see why these two [INAUDIBLE]? They don't [INAUDIBLE].

PROFESSOR: You need x1 negative so that you get another root so that r star exists. If r star is positive, if the expected value of x is positive, then r star is down here at negative r. I mean, you're talking about the other threshold in a sense.

OK, this is valid for all lower thresholds. And it's also valid for no threshold. OK, in other words, this equation here does not have beta in it at all. So this equation is an upper bound on the probability, that you're going to cross that threshold of alpha. And that upper bound is valid, no matter where you put the lower bound at all.

So you can go to the limit as the lower bound goes to infinity. And this inequality should still be valid. You have a homework problem where you actually prove that.

Sometimes when things go to infinity, funny things happen. And that proves that nothing funny happens then. So what happens then is the probability that you ever cross a threshold at plus alpha, when you have a random variable, which has a negative mean, is this exponent here.

And we also sort of showed by looking at the turn off bound that this bound is pretty tight. So in other words, what this is saying is when you're looking at threshold crossing problems-- this quantity here, this quantity where the second root of gamma of r is-- that's sort of the crucial parameter that you want to know. Usually the first thing you want to know about a random variable is its mean, its variance, all sorts of things like that.

This is saying if you're interested in thresholds, forget about all those things, look at r star. If r star is positive that means it means is negative, so there's no problem there. But this one quantity here is sort of the most important parameter of all of these problems.

OK, so let's go back to look at a hypothesis testing again, where we're looking at the likelihood ratio of being the ratio of the density for hypothesis 0 divided by hypothesis 1. What you get then is you observe this sequence Y sub n. These are the observations that you're taking.

In other words, nature at the beginning of this whole experiment chooses either H equals 0 or H equals 1. At that point, you start to make measurements. Now whether nature chooses H equals 0 before or after or when doesn't make any difference. The point is the experiment consists of nature choosing one of these two hypotheses.

You know all the probabilities that exist in the world in this model. You go making these measurements. All you of observe is these measurements. You don't observe what the hypothesis is, so you define this likelihood ratio of the ratio of the densities of the vector Y for H equals 0 and the vector Y with H equals 1.

These quantities exist no matter what the a priori probabilities of the thresholds are or anything else. Even without all of that, so long as you have a model which tells you what the densities of these observations are, conditional on each hypothesis, you can define this. This doesn't depend on a priori probabilities at all.

OK, so now you look at the probability that H is equal to 0, given all these observations divided by the probability it's equal to 1. What you get here now, you have the a priori probabilities, p0 over p1. Here is the likelihood ratio here. So what you have this p0 over p1 times the likelihood ratio of this vector of however many observations you have observed. It's just a nice way of breaking up the problem into the likelihood ratio and the a priori probabilities.

Incidentally, we haven't talked about this at all, but there's an important idea and all of this hypothesis testing of a sufficient statistic, and what do you think a sufficient statistic is. It's anything from which you can calculate the likelihood ratio. In other words, what we're saying here, the point we're making, is that any intelligent choice of hypothesis is it based on a threshold test on the likelihood ratio. And therefore, the only thing you can really be interested in in all your observations is just what is the likelihood ratio?

If they make all these 1,000 observations complicated sort of thing, you calculate one number. And that's the only thing you're interested in. And anything from which that number could be calculated is a sufficient statistic. And anything from which it can't be calculated you've thrown away some of the information that you have.

If you study communication and you study detection, you study how to receive data that's being sent, what you find is that right at the beginning, even before you do any detection, even before you do any filtering, there's some idea of a sufficient statistic there. That's what you need in order to calculate everything else. And you want to make sure that you have that. So that's an important idea there.

OK, but anyway, the MAP rule, which comes right from this, says if you have these a priori probabilities, and you're trying to maximize the probability of choosing correctly, what do you do? Well, your probability of H equals 0 was the correct hypothesis, given all the observations you made, is in fact this. The probability that H equals 1 is the correct hypothesis is this.

What do you do if you want to maximize the probability of being correct? You choose the one which is biggest. In other words, what you do is you look at this number. And if this number is bigger than 1, you choose 0. If it's less than 1, you choose hypothesis 1.

And what it turns out to is threshold of rule. You take this likelihood ratio. You compare it with p1 over p0. And in this case, you select h equals 0. In this case you select H equals 1.

And the last time I just a 1 and 0 reversed, which is fine, but if you reverse them one place, you want a reverse them every place. And every other threshold test does something like this, except you replace p1 over p0 with some arbitrary threshold. You say whatever reason you want to find for that threshold, that's the only intelligent kind of test you can make.

OK, then we define the log-likelihood ratio of the logarithm of the likelihood ratio. And that was nice because it was a sum of this quantity related to the individual observations. For each observation you really want to know what f of Y given H of Y given 0, divided by Y given 1. You want to divide those two. You want to take the logarithm of it. And then you have those numbers, and the sufficient statistic that you're interested in is just a sum of those numbers.

So you're looking at a sum of IID, random variable. IID, why IID? Well, under the hypothesis that H is equal to 1, those Yi's are IID. And therefore under the hypothesis that H is equal to 1. Well, little Zi is just a sample value. If you look at the random variable which has these sampled values, Z sub i, under the probability measure, corresponding to H equals 1, those Z sub i's are IID.

So what that says is when you look at these sums of random variables, the sum of Zi from 1 to n, under hypothesis H equals 1, what do you get? You get a random walk. You get a sum of IID random variables.

If you take more observations, S sub n just changes. With n changing, then you have a larger number of observations. So the random walk goes a little further out, and you might get closer to a threshold or whatever. And that's what we're trying to do here.

OK, so the Z sub i's under the hypothesis H equals 1, or IID, and the moment generating function of the Z sub i's given H equals 1, is this. Let's be careful about this. The sampled values of the Z sub i do not depend on the hypotheses at all.

Namely, you make an observation. You make an observation of Y sub i. You calculate Z sub i from Y sub i. That has nothing to do with whether H equals 0 or H equals 1. You try to calculate this moment generating function, however. And you want to know what the probability density of the Y's are. And you get a different probability density for H equals 1, then you get on the other hypothesis.

If the observations behaved the same way under both hypotheses, it wouldn't make much sense to do the observation. Unless you have a government grant, and you're trying to get money out of the government instead of trying to do anything worthwhile. Under those circumstances, you keep on making observations. You now perfectly well that nothing is going to come from them. But otherwise, it's a little silly.

So this moment generating function under the hypothesis H equals 1 is given by this quantity here. And this density here is the same as this density here. So you get this density to the 1 minus r power, and you get this density to the r power. So you get the product of these two densities. You integrate it over Y, and that's what gamma 1 of r is.

Now I said that the really important thing in all of these threshold problems is what is our star? And for this problem, r star is trivial. It's always the same. r star is always equal to 1.

And the reason is when you set r equal to 1 here, this quantity becomes 1. This quantity becomes the density of Y conditional on H equals 0. When you integrate that, you get 1.

So for all of these hypothesis testing problems, r star is equal to 1. Gamma 1 of 1 is equal to 0. And this is what this curve says. OK, this is gamma 1 of r here. This curve starts out here, negative slope. It comes up here. r star is equal to 1 in this case. And that's sort of the end of the story for that.

Now if you are doing a test with a fixed value of n, you say I'm going to make n observations, it's all I have time for. The week is over. I'm going on vacation next week. I've got to stop this test. I've got to write my paper. Take the end test. You write your paper.

And what do you do? You go through the optimal tests the best you can. And what you find is given H equals 1, an error is going to occur if the sum of random variables, namely the log-likelihood likelihood ratio, exceeds the logarithm of your threshold.

OK, this is whatever threshold you decide to establish. And we showed before that the probability that S sub n is greater than or equal to log of the threshold is evaluated as E to the n times this quantity right here. The probability of error given H equals 1 is this quantity here.

Probability of error given H equals 1 is the probability that the data looks like H equals 0 was a right hypothesis. In other words, that you crossed the threshold at plus alpha, instead of crossing the threshold at beta. Excuse me. We have too many cases here we're looking at, so it gets confusing. What I'm looking at here is the probability that the log-likelihood ratio exceeds this threshold data, whatever we set beta to be.

Eta is set, depending on the cost of making errors of both types on our a priori beta, if we have any and all of those things. And the probability of error given H equals 1 is this quantity here, which has the threshold in it over there. We've looked at that a number of times in a lecture. We looked at it in chapter one. And then those, we looked at it in chapter seven.

And you calculate it by taking this moment generating function, drawing attention to it at the point where slope natural log of eta divided by n. And then you take where it comes in to this vertical axis here. And that's the exponent of the error of probability when hypothesis 1 is correct.

Now, if the hypothesis is H equals 0 instead, at that point with H equals 0, the expected value of this log-likelihood ratio is going to be positive. The situation is going to be a curve that comes over here, comes back at some point here. And what we've showed is that this curve is just a translation of this curve by 1.

OK, namely if you calculate the moment generating function for H equals 0, you get the same thing that we got before. I'm not going to go through all the details of this. Now you have 0 here instead of 1.

Over here you're going to have just minus r. And over here you're going to have 1 plus r. So this whole thing is translated by 1. The action happens here. But if you translate it by 1 over in this direction, what happens is the error of probability is determined by this.

It's this. The exponent is this point right here, gamma 1 of r0 plus 1 minus r0, log of eta over n. And the r0 was again determined by this point at which the slope is equal to log eta over n.

We did that before. Then I want to make clear you understood it, because to really understand it, you have to go through the arithmetic yourselves at least once. And you can do that easily by following the notes, because it does it in almost excruciating detail.

So that's the argument you get. We had this idea before, of the Neyman-Pearson principle, which says you don't assume a priori probabilities. You look at the probability of making an error as being a trade off between the error you make when H is equal to 1 and the error you make when H is equal to 0.

In terms of the Chernoff bound, this trade off is very clear. As you change the exponent that you want to get under H equals 1, this point moves. The tangent then moves. And the exponent over here moves.

So you have this inverted seesaw. And the exponent for one kind of error is over here. And the exponent for the other kind of error is over there. Then the next thing we said was this is really stupid, unless you're going on vacation this Friday. If you're not going on vacation this Friday, if you're really serious about making the right decision, then what you're going to do is keep on making observations until you're pretty sure you're right.

Now somebody at the end of the lecture last time pointed out something, which says that when you do experiments and you keep on making observations until you get the data that you want, there's something very unethical about that. Is this that kind of unethical behavior? Or is this really valid?

Well, I claim this is valid, because what we're doing when we're doing sequential testing is we're deciding what we're going to do ahead of time. Namely, we've decided what we're going to do is we're going to continue testing until we cross a threshold and threshold gives us a suitable probability of error. So we're not cooking the books at all. What we're doing is we're following this preset procedure we've set up. And the only question is can we get a very small error probability by using a smaller number of observations on the average than what we need otherwise?

Put it in terms of a communication system. One kind of communication system, you have to send some data from one point to another. You're not going to get any feedback on it. You've got to get the data through the first time. It's got to be right. What are you going to do?

You're going to send this data a very large number of times or use a very powerful coding technique on it. And by time it gets through, you're going to be very sure you're right. Now a much better procedure, and the thing which is used in almost all communication systems, and the thing which we use as human beings all the time, and the thing which control people use all the time, the thing which almost everybody uses, because most of us have common sense if we spend some time trying to do these things, is instead of trying to get it right the first time, we try little bit to get it right the first time. And we make sure that if we don't get it right the first time, we have some way of finding out about it and getting it right the second time.

And in the scientific way of looking at it, what we do is we decide ahead of time exactly what our procedure is going to be for making repetitions-- something called ARQ in communication systems, which means automatic repeat request. It's automatic, which means you don't try to make your decision depending on whether you'd like to receive this 0 or like to receive a 1. You make the decision ahead of time that if you have a clean enough answer, you're going to accept it. If it looks doubtful, you're going to send it over again. That's exactly the same sort of thing we're doing here.

OK, when we do that, given H equals 1, we again have this S sub n as a function of n as a random walk. It's a sum of IID random variables and conditional on H equals 1. You have a random walk.

Conditional on H equals 1, you have a negative slope on this random walk. The random walk starts out and on the average is going to go down, and it's going to continue going down forever. And if you're looking for across some positive threshold, if it doesn't cross it pretty soon, it's not going to cross it.

But anyway, we have a test which said we have some positive threshold. We have some negative threshold. If we ever cross the positive threshold, we say H is equal to 0. If we ever cross the negative threshold, we say H is equal to 1. And then we're done with it.

OK, now, let me give you another argument why that makes sense. I gave you one argument last time. I'll give you another argument this time. If S sub J is greater than or equal to 0, we're going to decide that 0 is the correct hypothesis.

If H equals 1 is the correct hypothesis, then we're going make an error when S sub J is greater than or equal to alpha. If S sub J is less than or equal to the beta, we're going to decide H equals 1. And conditional on H equals 1, an error is made if SJ is greater than or equal to alpha. Conditional on H equals 0, an error is made if SJ is less than or equal to beta.

OK, so the probability of the error conditional on H equals 1 is the probability that S sub J is greater than or equal to alpha, given H equals 1, which is less than or equal to E to the minus alpha or star. This is the thing that we said before. r star is the root of gamma of r. And gamma of r is equal to this.

OK, so, let's just make life a little easier for ourselves assume that our a priori probabilities are each 1/2. This is also called maximum likelihood decision. You take this likelihood ratio, and you just decide on the basis of the likelihood ratio.

OK, then at the end of trial end, the probability of H equals 0 given Sn divided by the probability of H equals 1 given Sn, the a prioris cancel out. It is just E to the S sub n. That's what it is.

It's the likelihood ratio. S sub n is the log-likelihood ratio. So this is what it is. If you now take probability of H equals 0 on probability that H equals 1 given S of in, this equation becomes this equation. And then the probability of H equals 1 given S sub n is just E to the minus Sn over 1 plus E to the minus Sn.

Now if Sn is a large number, E to the minus Sn is going to be totally trivial. And the probability that H equals 1 given Sn is essentially E to the minus Sn. It means when you can choose different values of n, this very directly gives you a control on what the probability of error is.

The probability of error is essentially E to the minus Sn. So if you choose a threshold alpha, what you're doing is you're guaranteeing that the probability of error cannot be less than E to the minus alpha. OK, so this is more than just talking about averages. This is saying if you use a threshold rule, then what you're doing is guaranteeing that the probability of error is never going to be less than this quantity of specified here.

OK, we saw last time the cost of choosing alpha to be large is that you have to make a very large number of trials, at least given H equals 0. Why don't I worry about the number of trials for H equals 1? I mean, it's nothing to be thought through here. If my thresholds are large, my probability of error is very small. The expected values of things for very large log-likelihood ratios are determined almost entirely by H equals 0.

H equals 1 sometimes. You sometimes make a mistake, because it's something very, very unusual. But that has very little influence on the expected number of tests you're making. So what happens then is the expected number of tests you make under the hypothesis that H is equal to 0-- now we're using Wald's equality rather than Wald's identity-- it's equal to the expected value of S sub J given H equals 0, divided by the expected value of Z, given H equals 0. Z is the log-likelihood ratio of one trial.

This is just Wald's equality with this condition thrown into it. Now what's the expected value of SJ given H equals 0? It's essentially alpha, and if you want to be more careful, it's alpha plus the expected overshoot given H equals 0. And that's divided by the expected value of Z, given H equals 0.

This is the answer we got last time. So the number of tests you have to make, if you set a positive threshold alpha, is essentially the number of tests you have to make when the hypothesis is equal 0. So the funny thing which is happening here is that as you change alpha, you're changing the probability of error for hypothesis H equals 1. And you're changing the number of tests you're going to have to do when H is equal to 0.

When you change beta, it's just the opposite. So that when you change beta, if you make beta a very large negative, you have to make an enormous number of tests under the circumstance that H is equal to 1. But you might make an error when H is equal to 0. So the trade off is between number of trials under one hypothesis, error of probability under the other hypothesis.

That's almost all we wanted to say about Wald's identity. There's one other huge thing that we want to talk about. If you take the first two derivatives of Wald's identity at r equals 0, you get some interesting things coming out. I mean, Wald's identity, you can use it any value of r you want to. And when you use it for a large value of r, you that an interesting result about large deviations. When you use it at a small value of r, you get something more about typical cases.

So looking at it at r equals 0, what you want to do is you want to take the derivative with respect to r of Wald's identity. This expected value in here we know is equal to 1. It's equal to 1 whatever value of r we choose. And therefore, when we take the derivative of this, we have to get 0. But we also want to take the derivative of it to see what we get.

So when you take the derivative of this quantity here and you don't worry about what exists and what doesn't exist-- you have to take the derivative here-- so you get an S sub J there. You take the derivative here, you get a gamma prime of r there. If you get SJ minus J times gamma prime of r, and this E to the what have you just sits there. You take the derivative of E to something, you never get rid of the E to something. You just get piled up stuff in front of it.

OK, so when we evaluate that at r equals 0, what happens? Well, what's the value of the gamma prime of 0? It's the expected value of X, yes. And this quantity here is all equal to 1, so we can forget about that. When r is equal to 0, this is equal to 0. When r is equal to 0, gamma of r is equal to 0.

So this whole thing in here is 0. So E to the 0 is 1. So we've got a 1 there. We got expected value of S sub J minus J times X bar is equal to 0.

What is that? That's Wald's equality. So Wald's equality falls out of the Wald's identity as what happens as the derivative of Wald's identity that r equals 0. Well, since we're so successful with that, let's go on and take another derivative. Yes?

AUDIENCE: I guess you want the final equal to 0 [INAUDIBLE].

PROFESSOR: Oh, the final equal to 0 comes from the fact that this quantity here that you're starting with is equal to 1 for all values of r. Therefore, I want to take the derivative with respect to r, I get 0. So that's one equation. The other thing is I just go through the mechanics of taking the derivative.

OK, so let's try to take the second derivative. Take the second derivative by taking the derivative of the first derivative. And what happens is then is this quantity in here I get an extra term of that sitting over there. And along with that, I get the derivative of this with respect to r.

I should probably have written that down there but since I didn't, let me see if I can do it. I get the expected value of SJ minus J gamma prime of r. And this quantity is squared now, because I have this there. I'm taking the derivative of this term with respect to r.

And also, I have to take the derivative of this with respect to r. So that gives me minus J times gamma double prime of r. And all of this times E to the r SJ minus J gamma of r.

Now I want to evaluate this at r equals 0. Evaluating this at r equals 0, this term goes away. So I wind up with the expected value of SJ minus J gamma prime of r where minus J gamma double prime of r is equal to 0.

Well, this doesn't look bad. But if you try to use it if you expand this term here, you get a term the expected value of S of J times J. And you can struggle with that. And it's ugly. That's very ugly.

But if you now say, if we have a mean we can use Wald's equality. It tells us what we want to know. If we don't have a mean, then Wald's equality doesn't tell us anything. But this is going to tell us something.

So we're going to make the assumption here that r is equal to 0 and X bar is equal to 0. And if X bar is equal to 0, gamma prime of 0 is equal to 0. And gamma double prime of 0 is equal to sigma squared of X.

So you do all of that. What you get is the expected value of S sub J squared minus sigma X squared of J is equal to 0. This is the same kind of thing that we got from Wald's equality.

From Wald's equality, it didn't tell us anything. It just gave us a relationship between the expected value of S sub J and expected value of J. This is doing the same thing. It's giving us a relationship between the expected value of S sub J squared and the expected value of J. So we get the same kind of quantity. It's doing the same thing for us.

Now you look at this for a 0 means simple random walk. Now you would have thought before you started to take this class that a simple random walk with mean 0 was the simplest thing in the world. And we've seen by looking at that it really isn't all that simple, that you come play these silly games like you can gamble forever, where with probability 1/2 you lose $1, with probability 1/2 you win $1-- perfectly fair game.

And with probability one, you to make $1 out of that, and quit and go home. And since you to make $1 out of it, and quit and go home, you can then quickly come back again and it again.

You can make $2. You can make $10. You can make $1,000. You can make $1 million with probability 1. So the simple random walk is no longer simple. It becomes puzzling.

But Wald's identity is dealing with two thresholds, one of alpha and one at beta. When we apply this and you observe it as a simple random walk, where you either go up by 1 or go down by 1, each with probability 1/2, the mean of X is 0 and the variance of X is 1. So this quantity here is 1.

You can then play games with what the probability is that you hit the upper threshold and the probability that you hit the lower threshold. I mean, it's done in the text. You don't have to take my word for it.

And when you do that, what you find is the expected value of J is equal to minus beta times alpha. Theta is a negative number, remember, so this is expected value of J is the magnitude of theta times the magnitude of alpha.

Now that's a little bizarre, but then you think about it a little bit. You think what happens. And this is really exact. I mean, this isn't an approximation or anything.

If alpha is very large, and beta is very large and negative, and you play this random walk game, you're going to fluctuate a long time. You're going to disperse slowly. You're going to disperse according to the square root of n, or the number of tests you take. So the amount of time it takes you until you get way out to these thresholds should be-- to the namely the value that n has to have-- roughly the square of alpha when beta and alpha are both the same.

This is something more general than that. It says that if Sn, the stop-when-you're-ahead game, we make alpha equals 1, the expected value of J depends on what the lower threshold is. And that suddenly makes sense, because what that's saying is if we have a lower threshold at 10, an upper threshold at one, then most of the time you win.

But when you lose, you lose $10. When you win, you win, $1. When you set a lower threshold at 100, when you lose, you lose $100. When you win, you win $1.

And suddenly, that stop-when-you're-ahead game does not look quite as attractive as it did before. What you're doing is taking a chance where you're probably going to win of winning $1, and you're risking your life's assets for it, which doesn't make too much sense anymore. OK this, I think, it gives you a better idea of what's going on on the simple random walk than anything else I've seen.

So it's time to start talking about martingales. A martingale, like most of the other things we've been talking about in the course, is a sequence of random variables. This is a more general kind of sequence than most of them.

Almost all of the processes we've talked about so far have been the kinds of things you can sort of get your hands on. And this is defined very abstractly in terms of a peculiar property that it has. And then the peculiar property it has is the expected value as the nth term, in this thing called a martingale, conditional on knowing the values of all the previous values, expected value of Z sub n given the value of Z and minus 1, Z and minus 2, all the way down to Z1 is equal to Z sub n minus 1. Namely, the expected value here is what you had there.

The word martingale comes from gambling, where gamblers used to spend a great deal of time trying to find gambling strategies when to stop, when to start betting bigger, when to start betting smaller, when to do all sorts of things, all sorts of strategies for how to lose less money. Let me put it that way, because you rarely find that opportunity where you can play a fair game. But if you play a fair game, martingales are what sort of rules on that.

And what that says is if you play this game for a long time, your capital is Z sub n minus 1. This says, figure expected capital after you play one more time. No matter what strategy you use, your expected capital is going to be the same as was as the actual capital the time before.

If this is too abstract for you, and it's too abstract for me half the time, because I look at this, and I say, gee, that's not much of a restriction, is it? What we're talking about is expected values here. But it's more than that, because it's saying for every choice of sample value for all of these things, none of them make any difference, except the last one.

And that's what happens in gambling. It doesn't make any difference how your capital has gotten to the point where it is at time n minus 1. You make a bet in a fair bet, and what you win is solely a function of what you've bet, if the game is fair. And that's what this is saying.

So when you write it out this way, the expected value of Zn, given that 1, the random variable Zn minus 1 has a particular value Zn minus 1. You started out with a particular value. It says that expected value is equal to what you said at the last time. And this is true for all sample values Zn minus 1 down to Z1, which is why it's a much stronger statement than it appears to be. now there's a lemma. I want to talk about that a little bit, because it's a good time to get you used to what these expected values mean. For martingale, the expected value of Zn given Zi, Zi minus 1, all the way down to Z1. This expected value is equal to Z sub i.

In other words, it's not only that your expected capital, given all of the past, is equal to what you had on the last time instant. If you're not given anything for 100 years back, and all you know is what your capital was 100 years ago, and if we think we're playing a fair game all of this time, which is of course always a question, the expected value of what we have now, conditional on everything from 100 years back through recorded history, is just that last term. In other words, it's the same kind of isolation of the past from the future as we had with Markov change.

With Markov change, remember, it's only what happens at one instant, given what happens a one instant, it makes the past independent of the future. Here it's not quite that way, because the past and the future are separated only in terms of the details of the past, and the expected value the future. It says expected value of Z sub n given all of the details of the past, no matter what the details of the past are, the effective value of Zn is equal to the actual value at times Z sub i.

So I want to improve this for you, and I warn you you're not going to follow this proof. And that's part of the reason for me to do it, because I want you to go back to Chapter 1 and think that through. Because in dealing with martingales, you have to think this through.

Because if you don't think it through, you're stuck with this notation all the way through. And if you try to use this notation with martingales, this is nice notation when you get confused, but you don't want to use it all the time. So you have to be able to go through arguments like this.

What I want to show is that if E to the Z3, given Z1 and Z2 is equal to Z2, then one special case of this lemma is that expected value of Z3 given Z1 is equal to Z1. And how do we show that?

Well, what we do is we use this law complete expectation. Well, first we remember the expected value of an arbitrary random variable X is the expected value of the expected value of X given Y. Now what does that mean?

The expected value of the random variable X, given Y, is a random variable. It's a random variable which depends on Y. That's a function of the sample value of Y. Namely, if you look at this quantity up here, expected value of X given Y equals 1. Expected value of X given Y equals 2. Expected value of X given Y equals 3.

We have all of these values here. We have a probability measure on it. This is a random variable, which is a function of Y. You've averaged that over X, but you're left move why because of the conditioning here.

So this quantity in here is now a function of Y. So when we take this equation and we add the conditioning on Z1, namely, this is being used for Z3 and Z2. Expected value of Z3 is equal to the expected value over Z2 of the expected value of Z3 given Z2, whole thing dependent on Z1.

OK, so what it says is this expected value is the expected value of the expected value of Z3 condition on Z2 and Z1. This quantity here as a function of what? That's a random variable. It's a function of what random variables?

AUDIENCE: Z2, Z1.

PROFESSOR: Z1 and Z2, yes. So this is a function of Z1 and Z2. What value is it as a function of Z1 and Z2? It's just equal to Z2.

So this quantity in here is Z2. so we're asking what's expected value of Z2 given Z1. And by definition of martingale, it is equal to Z1.

Now I imagine about half you could follow that, and half of you couldn't, and half of you sort of followed it. This is a kind of argument we'll be using all the way through on this stuff. So make sure you understand it. I mean, once you get it, it's easy. And you can apply it in all sorts of places. So it's worth doing it.

In the same way, you can follow the same kind of argument through the expected value Z sub i plus 2, using this total expectation based on Zi plus 1. And you go through the whole thing. When you go down to i equals 1, it says the expected value of z is equal to the expected value of Z1.

If you want to become wealthy, have a wealthy parent, who leads to a lot of money 20 years ago. That's the easiest way to make a million dollars is to start out with 2 million dollars is the way that some people put it.

OK, let's have some simple examples a martingales. One of them is a zero-mean random walk. Mainly, what I'm trying to do here is to show you this martingales are really pretty general things. And since there are many very general theorems that hold for all martingales, you can then apply them to all of these special cases, which is kind of neat.

To have a zero-mean random walk, let Z sub n be the sum of X1 plus Xn and the X sub i's are IID and zero mean. The fact that they're IID makes it a random walk. The fact that there's zero mean makes it a special zero mean random walk, and the expected value of Z sub n, given Zn minus 1.

All the way back, Zn now is Xn plus Zn minus 1. OK, Zn is the sum of all these random variables. So you get up n minus 1 of them, and then you add the last one in. So it's Zn minus 1 plus Xn, so its expected value of Xn plus Zn minus 1, given all the stuff before that.

The expected value of Xn, given all this stuff, is what? Xn is independent of all the other X's, therefore it's independent of all the earlier Z's. And therefore, that's just expected value of Xn. So we have the expected value of Zn minus 1, given Zn minus 1 back to Z1.

What's expected value of Zn minus 1, given Zn minus 1? Well, it's Zn minus 1. That's no problem there. So this is 0. So this is equal to Zn minus 1, as it's supposed to be.

All of these things you ought to go back and think them through yourself. Because the first time you look at martingales, all of this stuff, it's all pretty easy, but it all looks a little strange at first. The next one is sums of arbitrary dependent random variables. They're not quite arbitrary.

Suppose you have a sequence of random variables, X sub i, i greater or equal to 1. And they satisfy the expected value of Xi, given all the earlier X of i's is equal to 0. It's similar to what a martingale is. But here, we're just saying the Xi's all have expected value of 0.

And Zn, the sum of these, has to be a martingale. And I'm not going to write all a proof of that. I mean, this proof is really the same as this proof.

This is really a pretty important thing, because given any martingale, you can always look at the partial sums between the terms of the martingale-- namely, given Z1, Z2, up to Z sub n. You can always look at Z2 minus Z1. You can look at Z3 minus Z2. You could at Z4 minus Z3, and so forth.

And each of the Z's is just the sum of those other random variables. So given any martingale in the world, you can always define the set of arbitrary depending random variables would satisfy this rule here. So Zn, in this case, is a martingale. And if Zn is a martingale, you can always define a set of random variables, which satisfy this property.

I think it's almost easier to see what a random variable really has to do with gambling, which is where it started, by looking at this. This is not your capital at time n. This is how much you win or lose at time i. And what it's saying is your winnings or losings at time i has zero mean independent of everything in the past.

In other words, in a fair game, you can bet whatever you want to and depending on what you bet, that's the expected amount you get on that trial. And that's what this says. This says essentially, you're applying a fair game. So martingales really have to do is fair games.

If you can find fair games, why, that's great. But we always look for games where we have an edge. But what you want to avoid is games where Las Vegas has an edge.

OK so, that's a general one. Here's an interesting one, because I think this is an example which you can use. I mean, in any field you study, there are always generic examples, which can be used to generate counter examples to any simple thing you might want to think of. And this is to me the most interesting one of those for martingales.

Suppose that Xi is the product of two random variables-- one is either plus 1 or minus 1, each with probability 1/2. And the other one, Y sub i is anything it wants to be. I don't care what Y sub i is. Y sub i is non-negative, might as well make it non-negative. I don't care about it. I don't care how it's related to all the other Y sub i's. All I want is the that the U sub i's are all independent of all the Y sub i's.

And what happens then? I take the expected value of X sub i, give it anything in the past, and what do I get? U sub i is independent of Y sub i. And therefore, the expected value of U sub i times Y sub i is expected value of U sub i, which is what-- plus 1 or minus 1 of probability 1/2 each.

The expected value of U sub i is equal to 0. That makes expected value of X sub i of 0 whatever the past is. So you automatically have the-- I don't know what to call in them, the terms between the terms of a martingale-- the interarrival terms, so to speak.

I mean, it's like those for a renewal process. Those terms always have mean 0. And therefore, the sums of these turn out to be this simple kind of martingale. So that's a nice martingale to use as counter examples for almost anything.

The next one is product for martingales. Product for martingales are things we use quite a bit too, because now when we're using generating functions, we're in the habit of multiplying things together. And that's a useful thing to do.

So the expected value of Z to the n, given Z to the n minus 1, now to Z1, where Zn is this product of terms. OK, Z sub n then is equal to Xn times Z sub n minus 1, which is what we're doing here. Expected value of Zn conditional on the past is the expected value of Xn times Zn minus 1, conditional on the past, Xn and Zn minus 1.

Oh, the expected value of Xn for any given value of Zn minus 1, all the way back, is just the expected value of X sub n. So we have expected value of X sub n times the expected value of Z sub n minus 1, given Zn n minus 1 down to Z1. So that's just Zn minus 1.

Ah, the missing quantity, fortunately I wrote it here. The X sub i's are unit means random variables. And they're IIDs. They're independent of each other. And since the X sub i's are independent of each other, X sub n is independent of Xn minus 1, all the way back to X1.

Zn minus 1 back to Z1 is a function of Xn minus 1, down to X1. So Xn is independent of all those previous Z's also. That's why I could split this apart in this way. And suddenly I wind up with Zn n minus 1 again. So product form martingales work.

Special form of product form martingales-- this again is favored counter example for when you can and can't get around with going to limits and interchanging limits. And it's a simple one.

Suppose that X sub i's are IID, as in the previous example. And they're [INAUDIBLE] probably 2 or 0. I mean, this is a game you often play-- double or nothing.

You start out with dollar. You play the game. If you win, you get $2. If you lose, you're broke.

If you win, you play your $2. If you win again, you have $4. If you lose, you're broke.

You play again. If you win, you have $8. If you lose again, you're broke.

So the probability that Z sub n, which is your capital after n trials, is equals to 2 to the n, namely you've won all n times, is 2 to the minus n. And every other instance you've lost. So the probability that Zn is equal to 0 is 1 minus 2 to the minus n.

So for each n, if you calculate the expected value of Z sub n, it's equal to 1. Namely, with probability 2 to the minus n, your capital is 2 to the n, so that's 1. With all the other probability, you have nothing.

So your expected value of Z sub n is always equal to 1. That's what this product form martingale says. And this is a product form martingale.

However, the limit as n goes to infinity of Zn is equal to 0 with probability 1. If you play double or nothing, eventually you lose. And then you're wiped out.

In other words, there's no real purpose to playing the game, because eventually you lose. If you're playing with somebody else, and they're playing double or nothing, then of course you get their money eventually. Or you go broke and the bank that you bank at fails, and all that stuff. We won't worry about that.

OK, but the point of this is that the limit as n goes to infinity of Zn is equal to 0. And the limit of the expected value of Z sub n is equal to 1. And therefore, the limit of Zn and the expected value of the limit of Zn.

The expected value of the limit of Zn is 0. The limit as expected value of Zn is equal to one. So this is a case where you can't interchange limit and expectation. It's an easy one to keep in mind, because we all know about playing double or nothing.

Might as well define submartingales and supermartingales. Because the first thing to know about them is they're like martingales, except they're defined with inequalities. And for a submartingale, the expected value of Zn, given all the previous terms, is greater than or equal to Zn minus 1.

So submartingales go up. Supermartingales are the opposite. Supermartingales goes down.

What else could you expect from a mathematical theory? Things that should go up, go down. Things that should go down, go up. Only thing you have to remember about submartingales and supermartingales is you figure out what terminology should have been used, and you remember the terminology they use was the opposite of what they should use.

I don't know whether I've ever seen stupider terminology than this. And someone once explained the reasoning for it, and the reasoning was stupid too. So there's no excuse for that one.

We're only going to refer to submartingales in what we're doing, partly because that's where most of the neat results are. And the other thing is if you have to deal with a supermartingale, what you might as well do is instead of dealing with a sequence-- Z1, Z2-- deal with the sequence minus Z1, minus Z2, and so forth. And if you change the sign on all the terms, then you change supermartingales into submartingales and vice versa. You don't really have to deal with both of them.

Let me talk briefly about an inequality that I'm sure most of you heard of. How many people have heard of Jensen's Inequality? Maybe half of you, so not everyone.

Well, it's one of the main work horses of probability theory. Even though we haven't seen it yet this term, you will see it many times.

So what a convex function is. A convex function in simple minded terms is something, a convex function from r into r. A real value convex function is a function which has a positive second derivative everywhere. So it curves down and comes back up again. Since you also want to talk about functions which don't have second derivatives, you want something more general than that. So you go from derivatives.

You go back to your high school ideas, and you draw a picture. And function is convex. If all the tangents to the curve lie not strictly below, but all the tangents can curve lie beneath the curve, so wherever you draw a tangent, you get something which doesn't cross the curve.

Magnitude of X is a convex function of X. Magnitude X as a function of X looks like this. You go down or up. And all tangents to this, this goes off to infinity and this goes off to infinity. So there's no way to get something like that in this tangent.

So you have one tangent here. You have a bunch of tangents along here. And you have one tangent there. And they all lie below the curve. So X bar is a convex function too.

And Jensen's Inequality says if H is convex, and if Z is a random variable, it has finite expectation, then H of the expected value of Z is less than or equal to the expected value of H of Z. You can interchange expected value and function with inequality like this, if the function is convex.

Now, why is this true? You can see why it's true automatically, if you're dealing with a random variable that has only two values. If you have two values for the random variable-- Z is a random variable here.

You have one variable here, one value here, one sample value here. Look at what the expected value of H of Z is. The expected value of H of Z is the expected value of this and this with the appropriate probability put in on it, so at some point that lies on the straight line between here and there.

When you look at the H of the expect the value, then you find the expected value along here. You can think of finding it along the straight line here. And then it's that point there. So since the curve is convex, the H of the expected value of Z is there's always a bunch of points. Average them, which lie on a straight line beneath the curve.

And the expected value of H of Z is taking the average directly along the curve. So you get this boosting up everywhere. It's like saying that the absolute value of expected value of Z is less than or equal to the expected value of the absolute value of Z.

And I think I will stop there instead of going on, because we had a lot of new things today. And somehow sequential detection always wears one's mind out in a short period of time. That should be enough.