PROFESSOR: OK, we have a busy day today, so let's get started. Want to go through Chernoff bounds and the Wald identity, which are closely related, as you'll see, and that involves coming back to the TG1Q a little bit and making use of what we did for that. It also means coming back to hypothesis testing and using that. Would probably have been better to start out with Wald's identity and the Chernoff bound and then do the applications when it was at the natural time for them. But anyway, this is the way it is this time, and next time we'll probably do it differently.

Suppose you have a random variable z. It has a moment generating function. Remember, not all random variables have moment generating functions. It's a pretty strong restriction.

You need a variance. You need moments of all orders. You need all sorts of things, s but we'll assume it exists in some region between r and r plus.

There's always a question, with moment generating functions, if they exist up to some maximum value of r because some of them exist at that value of r and then disappear immediately after that, and others just sort of peter away as r approaches r plus from below. I think in the homework this week, you have an example of both of those. I mean, it's a very simple issue. If you have an exponential distribution, then as r approaches, the rate of that exponential distribution, obviously, the moment generating function blows up because you're taking e to the minus lambda x, and you're multiplying it by a the r x. And when r is equal to lambda, bingo, you're integrating 1 over an infinite range, so you've got infinity. If you multiply that exponential by something which makes the integral finite when you set r equal to lambda, then of course, you have something which is finite at r star.

That is a big pain in the neck. It's usually not important. The notes deal with it very carefully, so we're not going to deal with it here. We will just assume here that we're talking about r less than r plus and not worry about that special case, which usually is not all that important. But sometimes you have to worry about it.

OK, the Chernoff bound says that the probability that random variable is greater than or equal to alpha is less than or equal to the moment generating function evaluated at some arbitrary value r times e to the minus r alpha. And if you put it in terms of the semi invariant moment generating function, the log of the moment generating function, then the bound is e to the gamma z of r minus alpha r. When you see something like that, you ought to look at it and say, gee, that looks funny because here, we're taking an arbitrary random variable and saying the tails of it have to go down exponentially.

That's exactly what this says. It says that a z takes on very large values. This is a fixed quantity here for a given value of r, and it's going down as e to the minus r times alpha. As you make alpha larger and larger, this goes down faster and faster.

So what's going on? How do you take an arbitrary random variable and say the tails of it is exponentially decreasing? That's why you have to insist that the moment generating function exists because when the moment generating function exists for some r, it means that the tail of that distribution is, in fact, going down at least that fast, so you get something that exists. So the question is what's the best bound of this sort of when you optimize o for r?

Then the next thing we did is we said that z is a sum of IID, then the semi invariant moment generating function for that sum is equal to n times the semi invariant moment generating function for the underlying random variable x. S of n is n of these IID random variable. So one thing you see immediately, and ought to be second nature to you now, is that if a random variable has a moment generating function over some range, the sum of a bunch of those IID random variables also has a moment generating function over that same range. You can just count on that because the semi invariant moment generating function is just n times this b.

OK, so then what we've said is the probability that sn is greater than or equal to na, where na is playing the role of alpha and sn is playing the role of z, is just a minimum over r of e to the n times gamma x of r minus ra, and the n is multiplying the ra as well as the n. OK, this is exponential n for a fixed a. In other words, what you do in this minimization, if you don't worry about the special cases or anything, how do you minimize something? Well, obviously, you want to minimize the exponent here, so you take the derivative of this gamma prime of r has to be equal to a. Then n can be whatever it wants to be when you find that optimum r, which is where gamma prime of r equals a, then this goes down exponentially with a.

Now, however, we're interested in something else. We're interested in threshold crossings. We're not interested in picking a particular value of a and asking, as n gets very, very big, what's the probability that the sum of random variable is greater than or equal to n times a. That is exponential in n, but what we're interested in is the probability that s of n is greater than or equal to just some constant alpha, and what we're doing, now, is instead of varying n and varying this with n also, we're holding the stick.

So we're asking as n gets very, very large, but you hold this alpha fixed, what happens on this bound over here? Well, when you minimize this, taking the same simple-minded view, now the n is not multiplied by the ra. It's just multiplied by the gamma x. You get n times gamma prime of r is equal to alpha s where the minimum is so that it says gamma prime of r is optimized when you pick gamma prime of r equal to alpha over and n. This quantity is minimized when you pick gamma prime of r equal to alpha over n.

So if you look at this bound as n changes, what's happening is, as n changes, r is changing also, so this is a harder thing to deal with for variable n. But graphically, it's quite easy to deal with. I'm not sure you all got the graphical argument last time when we went through it, so I want to go through it again. Let's look at this exponent r minus n over alpha times gamma of r, and see what it looks like.

We'll take r, pick any old r, there. What we want to do is show that this, if you take a slope of alpha over n, and take an arbitrary r, come down to gamma of x of r, draw a line in this slope, and look at where it hits the horizontal axis here, that point is r plus the length of this line here. The length of this line here is gamma of r, that's a negative value, times 1 over that slope of this line. And 1 over the slope of this line is n over alpha, so when I pick a particular value of r, the value of the experiment I have is this value here.

How do I optimize this over r? How do I get the largest exponent here? Well, I think of varying r, as I vary r from 0, and each time, I take this straight line here. And I start here, draw a straight line over there, start here, draw a straight line over, start at this tangent here, draw a straight line over.

And what happens when I come to larger values of r? Just because gamma of s of r is convex, what happens is I start taking these slope lines, slope alpha over n, and they intercept the horizontal axis at a smaller value. So this is optimized over r at the value of r0, which satisfies alpha over n equals gamma prime of r0.

That's the same answer we got before when we just used elementary calculus. Here, we're using a more sophisticated argument, which you learned about probably in 10th grade. I would argue that you learned mostly really sophisticated things when you're in high school, and then when you get to study engineering in college, somehow you always study these mundane things.

But anyway, aside from that, why is this geometric argument better? Well, when you look at these special cases of what happens when gamma of r comes around like this, and then suddenly it stops in midair and just doesn't exist anymore? So it comes around here, it's still convex, but then suddenly it goes off to infinity. How do you do that optimization then? Well, the graphical argument makes it clear how you do it, and makes it perfectly rigorous how to do it, whereas if you're doing it by calculus, you've got a really think it through, and it becomes fairly tricky.

OK, so anyway, now, the next question we want to ask-- I mean, at this point, we've seen how to minimize this quantity over r, so we know what this exponent is for a particular value of n. Now, what happens when we vary n? As you vary n, the thing that happens is we have this tangent line here, a slope alpha over n. When you start making n larger, alpha over n becomes smaller, so the slope becomes smaller. And as n approaches infinity, you wind up going way, way the heck out.

As n gets smaller, you come in again. You keep coming in until you get to this point here. And what happens then? We're talking about a line of-- maybe I ought to draw it on the board. It would be clearer, I think.

As n gets smaller, you get a point which is tangent here, this here. When you're here, the tangent gets right here, so we've moved all the way into this quantity we call r star, which is the root of the equation gamma of r equals 0. Gamma of r equals 0 typically has two roots, one here, and one at 0. It always has a root at 0 because moment generating function evaluated is 0 is always 1, so the log of it is always 0. There should be another root because this is convex, unless it drops off suddenly, and even if it drops off suddenly, you can visualize it as a straight line going off to infinity.

So when you get down to this point, what happens? Well, we just keep moving along. So as n increases, we start out very large. We come in. We hit this point, and then we start coming out again.

I mean, if you think about it, that makes perfect sense because what we're doing here is we're imagining experiment where this random variable has a negative expected value. That's what's indicated by this quantity there. We're asking what's the probability that the sum of a large number of IID random variables with a negative expected value ever rises above some positive threshold?

Well, the law of large numbers says it's not going to do that when n is very, very large, and this says that, too. It says the probability of it for n very large is extraordinarily small. It's e to the minus 10 times an exponent, which is very, very large. So as n gets very small, it's not going to happen either because it doesn't have time to get to the threshold.

So there's some intermediate value at which it's most likely to cross the threshold, if you're going to cross the threshold, and that intermediate value is just that value at which gamma of r star is equal to zero. So the probability this union of terms, namely the probability you ever cross alpha, is going to be, in some sense, approximately a to the minus alpha r star because that's where the dominant term is. The dominant term is where alpha over n is equal to gamma prime.

Blah, blah, blah, blah, blah, where'd I put that? r star satisfies gamma of r star equals 0. When you look at the line of slope, gamma prime of r plus of r star, that's where you get this critical value of n where it's most likely the cross the threshold.

OK, I put that somewhere. I thought it was on this slide, but it's the n, the critical n, let's call it n crit, is equal to gamma prime. Is at right? Alpha over m is the gamma prime. Alpha over n, 1 over n crit. n crit, this says, is alpha over gamma prime of r star.

OK, so that sort of nails down everything you want to know about the Chernoff bound accept for the fact that it is exponentially tight. The text proves that. I'm not going to go through that here. Exponentially tight means, if you take an exponent which is just a little bit larger than the one you found here, and look what happens as alpha gets very, very large, then you lose.

OK, let's go on, and at this point, we're ready to talk about Wald's identity. And we'll prove Wald's identity at the end of the lecture today. Turns out there's a very, very simple proof of it. There's hardly anything to it, but it seems more important to use it in several ways first so that you get a sense that it, in fact, is sort of important. OK, so we want to think about a random walk, s sub n and greater than or equal to n, so it's a sequence of sums of random variable, s sub n is equal to x1 plus up to x sub n. The x's are all IID.

This is the thing we've been talking about all term. We have a bunch of IID random variables. We look at the partial sums of them. We're interested in what happens to that sequence of partial sums.

The question we're asking here is does that sequence of partial sums ever cross a positive threshold? And now we're asking does it ever cross a positive threshold, or does it cross a negative threshold, and which does it cross first? So the probability that it crosses this threshold is the probability that it goes up, first. The probability that it crosses this threshold is the probability that it goes down, first.

Now, what Wald's identity says is the following thing. We're going to assume that x is not identically 0. If x is identically 0, then it's never going to go any place. We're going to assume that it has a semi invariant moment generating function in some region, r minus to r plus. That's the same as assuming that it has a generating function in that region, so it exists from some value less than zero to some value greater than zero.

And we picked two thresholds, one of them positive, one of them negative, and we let j be the smallest value of n. j is a random variable, now, because we've start to run this random walk. We run it until it crosses one of these thresholds, and if it crosses the positive threshold, j is the time at which it crosses the positive threshold. If it crosses the negative threshold, j is the time that it crosses the negative threshold. We're only looking at the first threshold that it crosses.

Now, notice that j is a stopping trial. In other words, what that means is you can determine whether you've crossed a threshold at time n solely in terms of s1 up to s sub n. If you see all these sums, then you know that you haven't crossed a threshold up until time n. You know you have crossed it at time n. Doesn't make any difference what happens at times greater than n.

OK, so it's a stopping trial in the same sense as the stopping trials we talked about before. You get the sense that Wald's identity, which we're talking about here, is sort of like Wald's equality, which we talked about before. Both of them have to do with these stopping trials. Both of them have everything to do with stopping trials.

Wald was a famous statistician, not all that much before your era. He didn't die too long ago. I forget when, but he was one of the good statisticians. See, he was a statistician who recognized that you wanted to look at lots of different models to understand the problem, rather than a statistician who only wanted to take data and think that he wasn't assuming anything. So Wald was a good guy.

And what his identity says is, and the trouble with his identity, is you look at it, and you blink. The expected value of e to the r s sub j. s sub j is the value of the random walk at the time when you cross a threshold minus the time at which you've crossed a threshold times gamma of r. So when you take the expected value of e to the this, you're averaging over j over the time that you crossed the threshold, and also at the value at which you crossed the threshold, so you're averaging over both of these things.

And Wald says this expectation not is less than or equal to 1, but it's exactly 1, and it's exactly 1 for every r between r minus and r plus. So it's a very surprising result. Yes?

AUDIENCE: Can you please explain why j cannot be defective? I don't really see it.

PROFESSOR: Oh, it's because we were looking at two thresholds. If we only had one threshold, then it could be defective. Since we're looking at two thresholds, you keep adding random variables in, and the sum starts to have a larger and larger variance. Now, even with a large variance, you're not sure that you crossed a threshold, but you see why you must cross a threshold. Yes?

AUDIENCE: If the MGF is defined at r minus, r plus, then is that also [INAUDIBLE] quality?

PROFESSOR: Yes. Oh, if it's defined at r plus. I don't know. I don't remember, and I would have to think about it hard. Funny things happen right at the ends of where these moment generating functions are defined, and you'll see why when we prove it.

I can give you a clue as to how we're going to prove it. What we're going to do is, for this random variable x, we're going to define another random variable which has the same distribution as x except it's tilted. For large values of x, you multiply it by e to the rx. For small values of x, you multiply it by e to the rx also.

But if r is positive, that means the positive values could shifted up, and the small values get shifted down. So you're taking some of the density that looks like this, and when you shift it to this tilted value, you're shifting the whole thing upward. When r is negative, you're shifting the whole thing downward. Now, what this says is that tilted random variable, when it crosses the threshold, the time of crossing the threshold is still a random variable.

You will see that this simply says that the expected value of that tilted random variable is equal to-- it says that tilted random variable is, in fact, the random variable. It's not defective. And it's the same argument as before, that has a finite variance, and therefore, since it has a finite variance, it keeps expanding. It will cross one of the thresholds eventually.

OK, so the other thing you can do here is to say, suppose instead of crossing a threshold, you just fix this stopping rule to say we'll stop at time 100. If you stop at time 100, then what this says is expected value of e to the r100 minus 100 times gamma of r is equal to 1. But that's obvious because the expected value of e to the r is j is, in fact-- it's j times the expected value of rx, so then you're subtracting off j times the log of the expected value of rx. So it's a trivial identity if x is fixed.

OK, so Wald's identity says this. Let's see what it means in terms of crossing a threshold. We'll assume both thresholds are there. Incidentally, Wald's identity is valid in a much broader range of circumstances than just where you have two thresholds, and you're looking at a threshold crossing. It's just that's a particularly valuable form of the Wald identity.

So that's the only thing we're going to use. But now, if we assume further that this random variable x has a negative expectation, when x has a negative expectation, gamma of r starts off going down. Usually, it comes back up again. We're going to assume that this quantity r star here, where it crosses 0 again, we're going to assume there is some value of r, for which gamma of r star equals 0.

Mainly, we're going to assume this typical case in which it comes back up and crosses the 0 point. And in that case, what it says is the probability that sj is greater than or equal to alpha is just less than or equal to e to the minus r star times alpha. Very, very simple bound at this point. You look at this, and you sort of see why we're looking, now, not at r in general, but just r star.

At r star, gamma of r star is equal to 0. So this term goes away, so we're only talking about the expected value of e to the r sj, e to the r star sj is equal to 1. So let's see what happens. We know that e to the r star sj is greater than or equal to 0 for all values of sj because e to the anything real is going to be positive.

OK, since e to the r star sj is greater than or equal to 0, what we can do is break this expected value here, this term is 0, now, remember, break it into two terms. Break it into the term where s sub j is bigger than alpha, and break it into the term where s sub j is less than beta. So I'm just going to ignore the case where it's less than or equal to beta.

I'm going to take this expected value. I'm going to write it as the probability that s sub j is greater than or equal to alpha times e times the expected value of either the r star s sub j given s sub j greater than or equal to alpha. There should be another term in here to make this equal, and that's the probability that s sub j is less than or equal to beta times e to the r star s sub j, given that s sub j is less than or equal to beta. We're going to ignore that, and that's why we get the less than or equal to 1 here.

Now, you can lower bound e to the r star sj under this condition. What's a lower bound to s sub j given that s sub j is greater than or equal to alpha? Alpha.

OK, we're looking at all cases where s sub j is greater than or equal to alpha, and we're going to stop this experiment at the point where it first exceeds alpha. So we're going to lower bound the point where it first exceeds alpha by alpha itself, so this quantity is lower bounded, again, by taking the probability that s sub j greater than or equal to alpha times e to the r star alpha, and that whole thing is less than or equal to 1. That says the probability that sj is greater than or equal to alpha is less than or equal to e to the minus r star alpha, which is what this inequality says here. OK, so this is not rocket science. This is a fairly simple result if you believe in Wald's identity, which we'll prove later.

OK, so it's valid for all choices of this lower threshold. And remember, this probability here, it doesn't look like it's a function of both alpha and beta, but it is because you're asking what's the probability that you cross the threshold alpha before you cross the threshold beta. And if you make beta very, very large, it makes it more likely that you're going to cross the threshold. If you make beta very close to 0, then you're probably going to cross beta first, so this inequality here, this quantity here, depends on beta also. But we know that this inequality is valid no matter what beta is, so we can let beta approach minus infinity, and we can still have this inequality.

There's a little bit tricky math involved in that. There's an exercise in the text which goes through that slightly tricky math, but what you find is that this bound is valid with only one threshold, as well as with two thresholds. But this proof here that we've given depends on a lower threshold, which is somewhere. We don't care where.

Valid for all choices of beta, so it's valid without a lower threshold. The probability that the union overall n of sn less than or equal to alpha. In other words, the probability that we ever crossed a threshold alpha--

AUDIENCE: It's not true equal.

PROFESSOR: What?

AUDIENCE: It's supposed to be sn larger [INAUDIBLE] as the last time?

PROFESSOR: It's less than or equal to e to the minus r star alpha, which is--

AUDIENCE: Oh, sn? sn?

PROFESSOR: n.

AUDIENCE: You just [INAUDIBLE] [? the quantity? ?]

PROFESSOR: Oh, it's a union overall n greater than or equal to 1. OK, in other words, this quantity we're dealing with here is the probability that sn--- oh, I see what you're saying. This quantity here should be greater than or equal to alpha. You're right.

Sorry about that. I think it's right most places. Yes, it's right. We have it right here. The probability of this union is really the same as the probability that the value of it, after it crosses the threshold, is greater than or equal to alpha.

OK, now, we saw before that the probability that s sub n is greater than or equal to alpha. Excuse me, that's the same. When you're writing things in LaTeX, the symbol for less than or equal to is so similar to that for greater than or equal to that's hard to keep them straight. That quantity there is a greater than or equal to sign, if you're going from right to left instead of right to left. So all we're doing here is simply using this, well, greater than or equal to.

OK, the corollary makes a stronger and cleaner statement that the probability that you ever cross alpha is less than or equal to-- my heavens, my evil twin got a hold of these slides. And let me rewrite that one. The probability that the union overall n of the event s sub n greater than or equal to alpha is less than or equal to e to the minus r star alpha.

OK, so we've seen from the Chernoff bound that for every n this bound has satisfied, this says that it's not only satisfied for each n, but it says it's satisfied overall n collectively. Otherwise, if we were using the Chernoff bound, what would we have to do to get a handle on this quantity? We'd have to use the union bound, and then when we use the union bound, we can show that for every n, the probability that sn is greater than or equal to alpha is less than or equal to this quantity.

But then we'd have to add all those terms, and we would have to somehow diddle around with them to show that there are only a few of them which are close to this value, and all the rest are negligible. And the number that are close to that value is only growing with n and goes through a lot of headache. Here, we don't have to do this anymore because the Wald identity has saved is from all that difficulty.

OK, we talked about the G/G/1 queue. We're going to apply this corollary to the

G/G/1 queue to the queueing time, namely to the time w sub i that the i's arrival spends in the queue before starting to be served. You remember, when we looked at that, we found that if we define u sub i to be equal to the ith interarrival time minus the i minus first service time, those are independent of each other, so this is the difference between those. So ui is the difference between the i's arrival time and the previous service time.

What we showed was that this sequence, u sub i, the sequence of the sums of u sub i as a modification of a random walk. In other words, the sums of the u sub i behave exactly like a random walk does, but every time it gets down to 0, if it crosses 0, it resets to 0 again. So it keeps bouncing up again.

If you look in the text, what it shows is that if you look at this sequence of u sub i's, and you look at the sum of them, and you look at them backward, if you look at the sum of u sub i plus u sub i minus 1 plus u sub i minus 2, and so forth, when you look at the sum that way, it actually becomes a random walk. Therefore, we can apply this bound to the random walk, and what we find is that the probability that the waiting time of n queue, of the nth customer, is probability that it's greater than or equal to an arbitrary number alpha is less than or equal to the probability that w sub infinity is greater than or equal to alpha, and it's less than e to the minus r star alpha. So again, all you have to do is you have this inner arrival time x, you have this service time y, you take the difference of the two, that's a random variable, you find a moment generating function of that random variable, you find the point of r star at which that moment generating function equals 1, and then the bound says that the probability that the queueing time that you're going to be dealing with is less than or equal to this quantity alpha here. Yes?

AUDIENCE: What do you work with when you have the gamma function go like this, and thus have infinity, and you cross it there. [INAUDIBLE] points that we're looking for?

PROFESSOR: For that, you have to read the text. I mean, effectively, you can think of it just as if gamma of r is a convex function like anything else. It just has a discontinuity in it, and bingo, it shoots off to infinity. So when you take these slope arguments, what happens is that for all slopes beyond that point, they just seesaw around at one point. But the same bound holds.

OK, so that's the Kingman bound. Then we talked about large deviations for hypothesis test. Well, actually we just talked about hypothesis test, but not large deviation for them. Let's review where we were on that.

Let's let the vector y be an n tuple of IID random variables, y1 up to y sub n. They're IID conditional on hypothesis 0. They're also IID conditional on hypothesis 1, so the game is nature chooses either hypothesis 0 or hypothesis 1. You take n samples of some IID random variable, and those n samples are IID conditional on either nature choosing 0 or nature choosing 1. At the end of choosing those n samples, you're supposed to guess whether h0 is the right hypothesis or 1 is a right hypothesis.

Invest in Apple stock 10 years ago, and one hypothesis is it's going to go broke. The other hypothesis is it's going to invent marvelous things, and your stock will go up by a factor of 50. You take some samples, you make your decision on that. Fortunately, with that, you can make a separate decision each year, but that's the kind of thing that we're talking about. We're just restricting it to this case where you have n sample values that you're taking one after the other, and they're all IID when the particular value of the hypothesis that happens to be there.

OK, so we said there is something called a likelihood ratio. The likelihood ratio for a particular sequence y is lambda of y is equal to the density of y given h1 divided by the density of y given h0. Why is it h1 on the top and h0 on the bottom? Purely convention, nothing else.

The only thing that distinguishes hypothesis 1 from hypothesis 0 is you choose one and call it 1, and you choose the other and call it 0. Doesn't make any difference how you do it. So after we make that choice, the likelihood ratio is that ratio.

Now, the reason for using semi invariant moment generating functions is that this density here is a product of densities. This density is a product of densities, and therefore when you take the log of this ratio of products, you get the sum from i equals 1 to n of this log likelihood ratio for just a single experiment. It's a single experiment that you're taking based on the fact that all n experiments are based on the same hypothesis, either h0 or h1. So the game that you're playing, and please remember what the game is if you forget everything else about this game, is the hypothesis gets chosen, and at the same time, you take n sample values. All n sample values correspond to the same value of the hypothesis.

OK, so when you do that, we're going to call z sub i, this logarithm here, this log likelihood ratio. And then we showed last time that a threshold test is-- well, we define the threshold test as comparing the sum with the logarithm of a threshold. And the threshold is equal to p0 over p sub 1, if in fact you're doing a maximum a posteriori probability test, and p0 and p1 are the probabilities of hypothesis.

Remember how we did that. It was a very simple thing. You just write out what the probability is of hypothesis 0 and a sequence of n values of y. You write out what the probability is of hypotheses 1 and that same sequence of values with the appropriate probability on that sequence for h equals 1 and h equals 0. And what you get out of that is that the threshold test sums up all the z sub i's, compares it with the threshold, and makes a choice, and that is the map choice.

OK, so conditional on h0, you're going to make an error if the sum of the z sub i's is greater than the logarithm of eta. And conditional on h1, you're going to make an error if the sum is less than or equal to log eta. I denote these as the random variable z sub i 0 to make sure that you recognize that this random variable here is conditional on h0 in this case, and it's conditional on h1 in the opposite case.

OK, so the exponential bound for z sub i sub 0-- OK, so what we're doing now is we're saying, OK, suppose that 0 is the actual value of this hypothesis. 0 is the value of the hypothesis. The experimenter doesn't know this. What the experimenter does is does what the experimenter has been told to do, namely the experimenter take these n values, y1 up to y sub n, finds the likelihood ratio, compares that likelihood ratio with the threshold, and if the threshold is larger than the threshold, it decides 1. If it's smaller than the threshold, that decides opposite thing. It decides 1 if it's above the threshold, 0 if it's below the threshold.

Well, first thing we want to do, then, is to find the log likelihood ratio under the assumption that 0 is the correct hypothesis, and something very remarkable happens here. Gamma sub 0 of r is now the logarithm because it's a semi invariant moment generating function of the expected value of this quantity of e to the r times z sub i. When we take the expected value, we integrate over f of y given h0 times e to the r times log of f of y given h1 over f of y given h0. You look at this, and what do you get? This quantity here is e to the r times log of f of y given h1.

That whole quantity in there is just f of y given h1 to the rth power. So what we have is, in this quantity here, is f of y given h0 to the minus r power. So this term combined with this term gives us f of 1 minus r of y given h0, and this quantity here is f to the r of y given h1 dy. So the semi invariant moment generating function is this quantity here. At r equals 1, this is just f of y given h1, so the log of it is equal to 0.

So what we're saying is that, for any old detection problem in the world, so long as this moment generating function exists, what happens is it starts at 0, it comes down, comes back up again, and r star is equal to 1. That's what we've just shown. When r is equal to 1, this whole thing is equal to 1, so the log of 1 is equal to 0. For every one of these problems, you know where this intercept is, you know where this intercept is, one is at 0, one is at 1.

What we're going to do now is try to find out what the probability of error is given that h is 0, h equals 0, is the correct hypothesis. So we're assuming that the probabilities are actually f of y given h0. We calculate this quantity that looks like this, and we ask what is the probability that this sum of random variables exceeds the threshold, exceeds the threshold eta. So the thing that we do is we draw a line, a slope, natural log of eta divided by eta. We draw that slope along here, and we find that the probability of error is upper bounded by gamma 0 of this quantity, defined by the slope, minus r0 times log of eta divided by eta.

That's all there is to it. Any questions about that? Seem obvious? Seem strange?

OK, so the probability of r conditional on h equals 0 is e to the n times gamma 0 of r0 minus r0, natural log of eta over eta. And ql of eta is the probability of error given that h is equal to l. OK, we can do the same thing for hypothesis 1. We're asking what's the probability of error given that h equals 1 is the correct hypothesis, and given that we choose a threshold, say we know the a priori probabilities, so we choose a threshold that way.

OK, we go through the same argument, z1 of s is the natural log of f of y given F1 times e to the s, we're using s in place of r here, times the natural log of f of y given h1 over f of y given h0. And this quantity, now, f of y given h1, the f of y given h1 is upstairs, so we have f of 1 plus s of y given h1. This quantity is down here, so we have f of minus s of y given h0. And we notice that when s is equal to minus 1, this is again equal to 0, and we notice also, if you compare this, gamma 1 of s is equal to gamma 0 of r minus 1. These two functions are the same, just shifts it by one.

OK, so this one of the very strange things about hypothesis testing, namely you are calculating these expected values, but you're calculating the expected value of a likelihood ratio. And the likelihood ratio involves the probabilities of the hypotheses also, so when you calculate that ratio, what you get this is funny quantity here, which is related to what you get when you calculate the semi invariant moment generating function given the other hypothesis. So that now, what we wind up with is a gamma 1 of the eta, is e to the n times gamma 0 of r0.

I'm using the fact that gamma 1 of s is equal to gamma 0 of r minus 1, s is just r shifted over by 1, so I can do the same optimization for each. So what I wind up with is the probability of error conditional on hypothesis 0, is this quantity down here. That's this one, and the probability of error conditional on the other hypothesis, the exponent is equal to this quantity here.

OK, so what that says is that as you shift the threshold-- in other words, suppose instead of using a map test, you say, well, I want the probability of error to be small when hypothesis 0 correct. I want it to be small when hypothesis 1 is correct. I have a trade off between those two. How do I choose my threshold in order to get the smallest value overall?

So you say, well, you're stuck. You have one exponent under hypothesis 0. You have another exponent under hypothesis 1. You have this curve here. You can take whatever value you want over here, and that sticks you with a value here.

You can rock things around this inverted seesaw, and you can make one probability of error bigger by making the other one smaller, or you make the other one bigger by making the other one smaller. Namely, what you're doing is changing the threshold, and as you change the threshold, as you make the threshold positive, what you're doing is making it harder to accept h1, h equals 1, and easier to accept h equals 0. When you move the threshold the other way, you're making it easier the other way.

This, in fact, gives you the choice between the two. You decide you're going to take n tests. You can make both of these smaller by making n bigger. But there's a trade off between the two, and the trade off is given by this tangent line to this curve here. And you're always stuck with r star equals 1 and all of these problems.

So the only question is what does this curve look like? Notice that the expected value of the likelihood ratio given h equals 0 is negative. The expected value given h equals 1 is positive, and that's just because of the form of the likelihood ratio.

OK, so this actually shows these two exponents. These are the exponents for the two kinds of errors. You can view this as a large deviation form of the Neyman Pearson test. In the Neyman Pearson test, you're doing things in a very detailed way, and you're taking a choice between choosing different thresholds to make the probability of error of one type bigger or less than the other one, just the other way.

Here, we're looking at the large deviation form of it that becomes an upper bound rather than an exact calculation, but it tells you much, much more because for most of these threshold tests, you're going to do enough experiments that your probability of error is going to be very small. So the only question is where do you really want the error probability to be small? You can make it very small one way by shifting the curve this way, and make it very small the other way by shifting the curve the other way. And you take your choice of which you want.

OK, the a priori probabilities are usually not the essential characteristic when you're dealing with this large deviation kind of result because, when you take a large number of tests, this threshold, log eta over eta over n, when n becomes very large, when you have a large number of experiments, log eta over n becomes relatively small. So that's not the thing you're usually concerned with. What you're concerned with is whether one test, the patient dies, and the other tests costs a lot of money; or one test, the nuclear plant blows up, and the other test, you waste a lot of money, which you wouldn't have had to pay otherwise.

OK, now, here's the important part of all of this. So far, it looked like there wasn't any way to get out of this trade off between choosing a threshold to make the error probability small one way, or making the error probability small the other way. And you think, well, yes, there is a way to get around it. What I should do is what I do in real life, namely if I'm trying to decide about something, what I'm normally going to do, I don't like to waste my time deciding about it, so as soon as the decision becomes relatively straightforward, I make up my mind. If the decision is not straightforward, if I don't have enough evidence, I keep doing more tests, so sequential tests are an obvious thing to try to do if you can do it.

What we have here, what we've shown, is we have two coupled random walks. Given hypothesis h equals 0, we have one random walk, and that random walk is typically going to go down. Given h equals 1, we have another random walk. That random walk is typically going to go up. And one is going to go down, one is going to go up, because we've defined the random variable involved is a log of f of y given h1 divided by f of y given h0, which is why the 1 walk goes up, and the 0 walk goes down.

Now, the thing we're going to do is do a sequential test. We're going to keep doing experiments until we cross a threshold. We're going to decide what threshold is going to give us a small enough probability of error under each condition, and then we choose that threshold. And we continue to test until we get there. So we want to find out whether we've gained anything by that, how much we've gained if we gain something by it, and so forth.

OK, when you use two thresholds, alpha's going to be bigger than 0. Beta's going to be less than 0. The expected value of z given h0 is less than 0, but the value of z given h1 is greater than 0. That's why the walks are coupled, so we can handle each of them separately until we can get the answers for one from the answers for the other. Crossing alpha is a rare event for the random walk with h0 because a random walk with h0, you're going to go down typically. You hardly ever go up. Yes?

AUDIENCE: Can you please explain again sign of expectations?

PROFESSOR: The sign of the expectations? Yes, z is the log, so that when we actually have h equals 1, the expected value of this is going to be lined up with this term on top. We have f of y. When we have h equals 0, this lined up with the term on the bottom.

I mean, actually, you have to go through and actually show that the integral of f of y given h1 of this quantity is greater than 0, and the other one is less than 0. We don't really have to do that because, if we calculate this moment generating function, we can pick it off of there. When we look at this moment generating function, that slope there is the expected value of z conditional on h equals 0, and because of the shifting property, this slope here is the expected value of z given h equals 1, just because the 1 curve is shifted from the other by one unit. It's really because of that ratio. If you defined it the other way, you just changed the sign, so nothing important would happen.

OK, so r start equals 1 for the h0 walk, so the probability of error, given h0, is less than or equal to e to the minus alpha. Well, that's a nice simple result, isn't it? In fact, that's really beautifully. You just calculate this moment generating function, you find the root of it, and you're done. You have a nice bound, and in fact, it's an exponentially tight bound.

And on the other hand, when you deal with the probability of error given h1 by symmetry, it's less than or equal to e to the beta. Beta is a negative number, remember, so this is exponentially going down as you choose beta, smaller and smaller. So the thing that we're getting is we can make each of these error probabilities as small as we want, this one, by making alpha big. We can make this one as small as we want by making beta big negative.

There must be a cost to this. OK, but what's the cost? What happens when you make alpha big?

When hypothesis 1 is the correct hypothesis, what normally happens is that this random walk is going to go up roughly at a slope of the expected value of z given h equals 0. So when you make alpha very, very large, you're forced to make a very large number of tests when h is equal to 1. When you make beta very, very large, you're forced to take a large number of tests when h is equal to 0.

So the trade off here is a little bit funny. You make your error probability for h equals 0 very, very small by costing more money when hypotheses 1 is the correct hypothesis because you don't make a decision until you've really climb way up on this random walk. And that means it takes a long time when you have h equals 1. Since when h is equal to 1, the probability of crossing this lower threshold is it is almost negligible, this expected time that it takes is really just a function of h equals 1. I'm going to show that in the next slide.

When you increase alpha, it lowers the probability of error given h equals 0. Excuse me, I should have h equals 0 instead of h sub 0. Exponentially, it increases the expected number of steps until you make a decision given h1. Expected value of j given h1 is effectively equal to alpha divided by expected value of z given h1.

Why is that? That's essentially Wald's equality. Not Wald's identity, but Wald's equality because-- Yes, it says from Wald's equality, since alpha is essentially equal to the expected value of s of j given h equals 1, the number of testing you have to take when h is equal to 1, when alpha is very, very large, is effectively the amount of time that it takes you to get up to the point alpha. That expected amount of time is typically pretty close to the mean value.

So alpha there is close to the expected value of s of j given h equals 1. So Wald's equality, given h equals 1, says the expected value of j given h1 is equal to the expected value of sj given h equals 1, that's alpha, divided by the expected value of z given h1, which is just the underlying likelihood ratio. So to get this result, we just substitute alpha for the expected value. And then the probability of error, given h equals 0, if we write it this way, we see the cost immediately. That's the expected value of j given h equal to 1.

In other words, the expected number of tests given h equals 1 times the expected value of the log likelihood ratio given h equals 1. When you decrease beta, that lowers the probability of error given h1 exponentially, but it increases the number of tests when h0 is the correct hypothesis. So in that case, you get the probability of error given h equals 1 is effectively equal to the expected value e to the expected value of j equals j equals 0. This is just the number of tests you have to do when h is equal to 0. This is the expected value of the log likelihood ratio when h is equal to 0.

This is very approximate, but this is how you would actually choose how big you make alpha, how big do you make beta if you want to do a test between these two hypotheses. Now, this shows what you're gaining by the sequential test over what you're gaining by the non-sequential test. You don't have this in your notes, so you might just jot it down quickly.

The expected value of z, conditional on h equals 0, is this slope here, the slope of the moment generating function is z equals 0. That's the slope of the underlying random variable. Since this point is r equal to 1, this point down here is the expected value of z given h equals 0. That's the exponents that you get when h equals 0 is, in fact, the correct exponent. When given the probability of error given that h is equal to 0, namely the probability that you choose hypothesis 1.

Same way over here. This slope here is the expected value of the log likelihood ratio given h equals 1. This hits down here at minus expected value of z given h equals 1. So you have this exponent going one way, you have this exponent going the other way when the thing multiplying the exponent is not an absolute value but is, in fact, the number of tests you have to do than the other test.

Now, if we do the fix test, what we're fixed with is a test where you take a line tangent to this curve, which goes from here across here to there. We can see-saw it around. When we see-saw it all the way in the limit, we can get this result here. But we get this result here at the cost of an error, which is almost one in the other case, so that's not a very good deal.

This says that sequential testing, well, it shows you how much you gain by doing a sequential test. I mean, it might not be intuitively obvious why this is happening. I mean, really the reason it's happening is that the times when you want to make the test very long are those times when if h is equal to 0, you normally go down.

The next most normal thing is you wobble around without doing anything for a long time, in which case you want to keep doing additional tests until finally it falls down, or finally it goes up. But by taking additional tests, you make it very unlikely that you're ever going to cross that threshold. So that's the thing you're gaining. You are gaining the fact that the error is small in those situations where the sum of these random variables stays close to 0 for a long time, and then you don't make errors in those cases.

We now have just a little bit of time to prove Wald's identity. I don't want to have a lot of time to prove it because proofs of theorems are things you really have to look at yourselves. This one, you almost don't have to look at it. This one is almost obvious as soon as you understand what a tilted probability is.

So let's suppose that x sub n is a sequence of IID discrete random variables. It has a moment generating function for some given r. We're going to assume that these random variables are discrete now to make this argument simple. If they're not discrete, this whole argument has to be replaced with all sorts of [INAUDIBLE] integrals and all of that stuff. It's exactly the same idea, but it just is messy mathematically.

So what we're going to do is we're going to define a tilted random variable. A tilted random variable is a random variable in a different probability space. OK, we start out with this probability space that we're interested in, and then we say, OK, suppose that we, just to satisfy our imaginations, we suppose the probabilities are different. We assume that the probabilities for a given r is the probability that the random variable X is equal to little x, namely this quantity here, is equal to the original probability that X is equal to little x.

All the sample values are the same, it's just the probability's changed, times e to the rx minus gamma of r. So we're taking these probabilities when X is large. We're magnifying them when x is small. We're knocking them down.

What's the purpose of this? It's just a normalization factor. e to the minus gamma of r is 1 over the moment generating function of r, so you take p of x, e to the rx, divide it by g of r. So this is a probability mass function, as well as this. This is the correct probability mass function for the model you're looking at.

This is an imaginary one, but you can always imagine. You can say let's suppose that we had this model instead of the other model. All the sample values are the same, but the probabilities are different. So we want to see what we can find out from these different probabilities in this different probability model.

If you sum over x here, this sum is equal to 1, as we just said. So we'll view q sub xr of x as the probability mass function on x in a new probability space. We can use all the laws of probability in this new space, and that's exactly what we're going to do. And we're going to say things about the new space, but then we can always come back to the old space from this formula here because whatever we find out in the new space will work in the old space.

One thing we'd like to do is to be able to find the expected value of the random variable x in this new probability space, so this isn't the expected value in the old space. It's a probability in the new space. It's the sum over x of x times q sub xr of x. That's what the expected value is.

X is the same in both spaces. That's just the probabilities that have changed. These are p of x times z to the rx minus gamma of r, so when you sum this, what you get is 1 over g of xr, which is that term, times the derivative of p sub x of x, e to the rx. When you take this derivative, then you get an x in front, which is that x there.

So you get g prime of xr over gx of r, which is gamma prime of r. OK, so in terms of that graph we've drawn, when you take these tilted probabilities, you move that slope, that r equals 0, and now you're looking at a slope at whatever r you're looking at. And that gives you the expected value there.

OK, if you have a joint tilted probability mass function-- and don't think it gets any more complicated. It doesn't. I mean, you've already gone through the major complication of this argument.

The joint tilted PMF is the probability of x1 to xn is the old probability of x1 to xn times all of these tilted factors here. If you let a of sn be the set of n tuples which have the same sum, then all these terms become r times s sub n. So what you get is that for each xn for which the sum is sn, this tilted probability becomes the old probability times e to the r sn minus n gamma of r, which says that when we look at the tilted probability of the sum, namely we said that when we tilt these probabilities, we can do everything in a new space that we could do in the old space. We can do everything that probability theory allows us to do, so we can look at the probability of s sub n in the new space also. The probability of sn in the old space, namely we're summing this quantity, overall xn in a of sn, so we sum up all of those as the probability sub s sub n at sn times this quantity, which is fixed.

So this is the key to a lot of large deviation theory. Any time you're dealing with a difficult problem, and you want to see what's happening way, way away from the mean, you want to see what these sums look like for these exceptional cases, what we do is we look at a new model where we tilt the probability so that the region of concern becomes the main region for that tilted model. So for r equals 0, we're tilting the probability towards large values, and you can use the law of large numbers, essential limit theorem, whatever you want to, in that new space, then.

Now, we can prove Wald's equality. What Wald's identity is is the statement that when you tilt these probabilities, a stopping rule in this tilted world is still the stopping time is still a random variable, namely you still stop with probability 1. Somebody questioned whether you stop with probability 1 in the old world. Like I said, you do because you have this positive variance, and the thing with two thresholds keeps growing and growing. Here, you have the same thing.

I mean, the mean doesn't make any difference at all. I mean, you're looking at trying to exceed one of two different thresholds, and eventually, you exceed one of them no matter where you set r. So what this is saying is the probability that j is equal to n in this tilted space is equal to the probability that j is equal to n in the old space times z to the r sn minus gamma of r. So this quantity is equal to the expected value of e to the r sn minus gamma of r. Given j equals n times the probability that j is equal to n, you sum this over n and, bingo, you're back at the Wald identity.

So that's all the Wald identity is, is just a statement that when you tilt a probability, and you have a stopping rule on the original probabilities, you then have a stopping rule on the new probabilities. And Wald's identity says-- well, Wald's identity holds whenever that tilted stopping rule is a random variable. OK, that's it for today. We will do martingales on Wednesday.