PROFESSOR: OK, I guess we might as well start a minute early since those of you who are here are here. We're coming to the end of course. We're deep in chapter 7 now talking about random walks and detection theory. We'll get into martingales sometime next week.

There are four more lectures after this one. The schedule was passed out at the beginning of the term. I don't know how I did it, but I somehow left off the last Wednesday of class.

The final is going to be on Wednesday morning at the ice rink. I don't know what the ice rink is like. It doesn't sound like an ideal place to take a final, but I assume they must have desks there and all that stuff. We will send out a notice about that.

This is the last homework set that you will have to turn in. We will probably have another set of practice problems and problems on-- but not things you should turn in. We will try to get solutions out on them fairly quickly, also. So you can do them, but also look at the answers right after you do them.

OK, so let's get back to random walks. And remember what we were doing last time. A random walk, by definition, you have a sequence of IID random variables. You have partial sums of those random variables. S sub n is a sum of the first n of those IID random variables. And the sequence of partial sums S1, S2, S3, and so forth, that sequence is called a random walk.

And if you graph the random walk, it's something which wanders up and down usually. And sometimes, if the mean of X is positive, it wanders off to infinity. If the mean of X is negative, it wanders off to minus infinity. If the mean of X is 0, it simply

diffuses somewhat as time goes on. And what we're trying to find that is exactly how do these things work. So our focus here is going to be on threshold-crossing problems. Namely, what's the probability that this random walk is going to cross some threshold by or at some particular value of n?

If you have two thresholds, one above and one below, what's the probability it's going to cross the one above? What's the probability it's going to cross the one below? And if it crosses one of these, when does it cross it? If it crosses it, how much of an overshoot is there?

All of those problems just come in naturally by looking at a sum of IID random variables. But here we're going to be trying to study them in some consistent manner looking at the thresholds particularly. We've talked a little bit about two particularly important applications. One is [? GG1Qs ?]. And even far more important than that is this question of detection, or making decisions, or hypothesis testing, all of which are the same thing.

You remember we did show that there was at least one threshold-crossing problem that was very, very easy. It's the threshold problem where the underlying random variable is binary. You either go up by 1 or you go down by 1 on each step. And the question is, what's the probability that you will cross some threshold at some k greater than 0?

And it turns out that since you can only go up 1 each time, the probability of getting up to some point k is the probability you ever got up to 1. Given that you got up to 1, it's the probability that you ever got up to 2. Given you got up to 2, it's the probability you ever got up to 3. That doesn't mean that you go directly from 2 to 3. After you go to 2, you wander all around, and eventually you make it up to 3. If you do, then the question is, do you ever get from 3 to 4, and so forth. And we found that the solution to that problem was p over 1 minus p to the k-th power of p is less than or equal to 1/2.

And we solved that problem, if you remember, back when we were talking about stop when you're ahead if you're playing coin tossing with somebody. And so let's

go further and look particularly at this problem of detection, and decisions, and hypothesis testing, which is really not a particularly hard problem. But it's made particularly hard by statisticians who have so many special rules, peculiar cases, and almost mythology about making decisions. And you can imagine why because as long as you talk about probability, everybody knows you're talking about an abstraction. As soon as you start talking about making a decision, it suddenly becomes real.

I mean, you look at a bunch of data and you have to do something. You look at a bunch of candidates for a job, you have to choose one. That's always very difficult because you might not choose the right one. You might choose a very poor one. But you have to do your best.

If you're investing in stocks, you look at all the statistics of everything. And finally you say, that's where I'm going to put my money. Or if you're looking for a job you say, that's where I'm going to work, and you hope that that's going to work out well. There are all these situations where you can evaluate probabilities until you're sick in the head. They don't mean anything. It's only when you make a decision and actually do something with it that it really means something. So it becomes important at this point.

The model we use for this, since we're studying probability theory-- well, actually, we're studying random processes. But we're really studying probability theory. You probably noticed that by now. Since we're studying probability, we study all these problems in terms of a probabilistic model. And in the probabilistic model, there's a discrete and, in most cases, binary random variable, H, which is called the hypothesis random variable. The sample values of H, you might as well call them 0 and 1. That's the easiest things to call binary things. They're called the alternative hypotheses. They have marginal probabilities because it's a probability model. You have a random variable. It can only take on the value 0 and 1, so it has to have probabilities of being 0 and 1.

Along with that, there are all sorts of other random variables. The situation might be

as complicated as you want. But since we're making decisions, we're making decisions on the basis of some set of alternatives. And here, since we're trying to talk about random walks, and martingales, and things like that, also we restrict our attention to particular kinds of observations. And the particular kind of observation that we restrict attention to here is a sequence of random variables, which we call the observation. You observe Y1. You observe Y2. You observe Y3, and so forth. In other words, you observe a sample value of each of those random variables. There are a whole sequence of them. And we assume, to make life simple for ourselves, that each of these are independent, conditional on the hypothesis. And they're identically distributed conditional on the hypothesis. That's what this says right here.

This makes one more assumption that assumes that these observations are continuous random variables. That doesn't make much difference, there are just a few peculiarities that come in if these are discrete random variables. There also a few peculiarities that come in when they're continuous. And there are a lot of peculiarities that come in when they're absolutely arbitrary. But for the time being, just imagine each of these are continuous random variables. So for each value of n, we look at n observations. We can calculate the probability density that those observations would occur conditional on hypothesis 0. We can find the conditional probability they could occur conditional on hypothesis 1. And since they're IID, that's equal to this product here.

Excuse me, they are not IID, they are conditionally ID. Conditional on the hypothesis. Namely, the idea is the world is one way or the world is another way. If the world is this way, then all of these hypotheses are IID. You're doing the same experiment again and again and again, but it's based on the same underlying hypothesis. Or, the underlying hypothesis is this over here. You make the number of observations all based on this same hypothesis, and you make as many of these IID observations conditional on that observation as you choose. And when you're all done, what do you do? You have to make your decision.

OK, so this is a very simple-minded model of this very complicated and very important problem. But it's close enough to the truth that we can get a lot of

observations from it.

Now, I spent a lot last time talking about this. Spend a lot of time this time talking about it because when we use a probability model for this, when we say that we're studying probability theory. And therefore, we're going to use probability, we have suddenly allied ourselves completely with people called Bayesian statisticians or Bayesian probabilists. And we have gone against, turned our back on people called Non-Bayesians, or sometimes classical. I hate using the word "classical" because I like the word "classics." I like the classics for such an unusual point of view.

And the unusual point of view is that we refuse to take a probability model. We accept the fact that on all the observations, all the observations are probabilistic. We assume we have a nice model for them, which makes sense. We can do whatever we want with that model. We can change the model. We can do whatever we want with a model. But if you once assume that these two hypotheses that you're trying to choose between, that they have a priori probabilities, then people get very upset about it because they say, well, if what the a priori probabilities are, why do you have to do a hypothesis test? You already understand everything there is to know about the problem. And they feel this is very strange.

It's not strange because you use probability models. You use models to try to understand certain things about reality. And you assume as many things as you want to assume about it. And when you get all done, you either use all the assumptions or you don't use them.

What we're going to find today is that when you use this assumption of a probability model, you can answer the questions that these classical statisticians go to great pains to answer. And you can ask them very, very simply. So that after we assume the a priori probabilities, we can calculate certain things which don't depend on those a priori probabilities. And therefore, we know two things. One, we know that if we did know the a priori probabilities, it wouldn't make any difference. And two, we know that if we can estimate the a priori probabilities, it makes a great deal of difference. And three-- and this is the most important point-- you make 100

observations of something. Somebody else says, I don't believe you, and comes in and makes another 100 observations. Somebody else makes another 100 observations.

Now, even if the second person doesn't believe what the first person has done, it doesn't make sense as a scientist to completely eliminate all of that from consideration. Namely, what you would like to do is say well, since this person has found such and such, the a priori probabilities have changed. And then I can go on and make my 100 observations. I can either make a hypothesis test based on my 100 observations or I can make a hypothesis test assuming that the other person did their work well. I can make it based on all of these observations.

If you try to do that those two things in a classical formulation, you run into a lot of trouble. If you try to do them in this probabilistic formulation, it's all perfectly straightforward. Because you can either start out with a model in which you're taking 200 observations or you can start out with a model in which you take 100 observations. And then suddenly, the world changes. This hypothesis takes on, perhaps a different value. You take another hundred observations. So you do whatever you want to within a probabilistic formulation.

But the other thing is, all of you that patiently have lived with this idea of studying probabilistic models all term long. You might as well keep on living with it.

The fact that we're now interested in making decisions should not make you think that everything you've learned up until this point is baloney. And to move from here to a classical statistical formulation of the world would really be saying, I don't believe in probability theory. It's that bad. So here we go.

I'm sorry, we did that. We were there.

Assume that on the basis of observing a sample value of this sequence of observations, we have to make a decision about H. We have to choose H equals 0 or H equals 1. We have to detect whether or not H is 1.

When you do this detection, you would think in the real world that you've detected

something. If you've made a decision about something, that you've tested a hypothesis and you found that which is correct. Not at all.

When you make decisions, you can make errors. And the question of what kinds of errors you're making is a major part of trying to make decisions. I mean, those people who make decisions and then can't believe that they might have made the wrong decision are the worst kind of fools. And you see them in politics. You see them in business. You see them in academia. You see them all over the place. When you make a decision and you've made a mistake, you get some more evidence. You see that it's a mistake and you change.

The whole 19th century was taken up with-- I mean, the scientific community was driven by physicists in those days. And the idea of Newton's laws was the most sacred thing they had. Everybody believed in Newtonian mechanics in those days.

When quantum mechanics came along, this wasn't just a minor perturbation in physics. This was a most crucial thing. This said, everything we've known goes out the window. We can't rely on anything anymore. But the physicists said, OK, I guess we made a mistake. We'll make new observations. We have new observations that can be made. We now see that Newtonian mechanics works over a certain range of things. It doesn't work in another ranges of things. And they go on and find new things.

That's the same thing we do here. We take these models. We evaluate our error probabilities. And evaluating them, we then say, well, we've got to go on and take some more measurements. Or we say we're going to live with it. But we face the fact that there are errors involved. And in doing that, you have to take a probabilistic model. If you don't take a probabilistic model, it's very hard for you to talk honestly about what error probabilities are. So both ways-- well, I'm preaching and I'm sorry. But I've lived for a long time with many statisticians, many of whom get into my own field and who cause a great deal of trouble. So the only thing I can do it urge you all to be cautious about this. And to think the matter through on your own. I'm not telling you to take my point of view on it. I'm telling you, don't take other people's

point of view without thinking it through.

The probability experiment here really-- I mean, every probability model we view in terms of the real world, as you have this set of probabilities, a set of possible events. You do the experiment. There's one sample point that comes out. And after the one sample point comes out, then you know what the result of the experiment is.

Here, the experiment consists both of what you normally view as the experiment. Namely, taking the observations. And it also involves a choice of hypotheses. Namely, there's not a correct hypothesis to start with. The experiment involves God throws his dice. Einstein didn't believe that God threw dice, but I do. And after throwing the dice, one or the other of these hypotheses turns out to be true. All of these observations point to that or they point to the other and you make a decision.

OK, so the experiment consists both on choosing the hypothesis and on taking a whole sequence of observations.

Now, the other thing to not forget in this-- because you really have to get this model in your mind or you're going to get very confused with all the things we do. The experiment consists on a whole sequence of observations, but only one choice of hypothesis. Namely, you do the experiment. There's a hypothesis that occurs, and there's a whole sequence of observations which are all IID conditional on that particular hypothesis. So that's the model we're going to be using.

And now life is quite simple once we've explained the model. We can talk about the probability that H is equal to either 0 or 1, conditional on the sample point we've observed. It's equal to the a priori probability of that hypothesis times the density of the observation conditional on the hypothesis divided by just a normalization factor. Namely, the overall probability of that observation period, which is the sum of probability that 0 is a correct hypothesis times this plus probability that 1 is a correct hypothesis times the density given 1.

This denominator here is a pain in the neck, as you can see. But you can avoid ever

dealing with a denominator if you take this for H equals 0, divide by this for H equals 1, and then you have this term divided by this term all divided by this term for l equals 1 divided by the same thing. So the ratio, the probability that H equals 0 given y over the probability that H is 1 equals y is just this ratio here.

Now, what's the probability of error if we make a decision at this point? If I've got in this particular sequence Y, this quantity here is, in fact, the probability that hypothesis 0 is correct in the model that we have chosen. So this is the probability that H is equal to 0 given Y. If we select 1 under these conditions, if we select hypothesis 1, if we make a decision and say, I'm going to guess that 1 is the right decision. That means that this is the probability you've made a mistake. Because this is the probability that H is actually 0 rather than 1. This quantity here is the probability that you've made a mistake given that 1 is the correct hypothesis.

So here we are sitting here with these probabilities of error. We don't have to do any calculations for them. Well, you might have to do a great deal of calculation to calculate this and to calculate this. But otherwise, the whole thing is just sitting there for you. So what do you do if you want to minimize the probability of error?

This was the probability that you're going to make an error if you choose 1. This is the probability of error if you choose 0. We want to minimize the probability of error and we see the observation Y, we want to pick the one of these which is largest. And that's all there is to it. This is the decision rule that minimizes the probability of an error. It's based on knowing what P0 and P1 is. But otherwise, probability that H equals l is the correct hypothesis given the observation is probability that H equals L given Y. We maximize the a posteriori probability of choosing correctly by choosing the maximum over l of probability that H equals l given Y.

This choosing directly, maximizing the a posteriori probability is called the MAP rule, Maximum A posteriori Probability. You can only solve the MAP problem if you assume that you know P0 P1.

We do know P0 and P1 if we've selected a probability model. So when we select this probability model, we've already assumed what these a priori probabilities are, so

we now make our observation. And after making our observation, we make a decision. And at that point, we have an a posteriori probability that each of the hypotheses is correct.

Anybody has any issues with this? I mean, it looks painfully simple when you look at this way. And if it doesn't look painfully simple, please ask now or forever hold your peace as they say. Yeah?

AUDIENCE: So can you explain how you get the equation? Can you explain how you get the equation on the first line?

PROFESSOR: On the first line right up here? Yes, I use Bayes' law.

AUDIENCE: So what is that? So that's P of A given B is equal to P of B given A?

PROFESSOR: Yes.

AUDIENCE: I don't quite see how to-- P of A given B is equal to P of B given A times P of A divided by P of A and B.

If you take this over there then it's-- am I stating Bayes' law in a funny way?

AUDIENCE: So the thing on the bottom is P of B? OK.

PROFESSOR: What?

AUDIENCE: OK, I get it.

PROFESSOR: I mean, I might not be explained it well.

AUDIENCE: [INAUDIBLE].

PROFESSOR: Except if you start out with P of A given B is equal to P of B given A times P of B divided by P of A. This quantity here is P of Y. So we have probability that H equals I times probability of Y given I divided by the probability of I to start with.

OK, so you maximize the a posteriori probability by choosing the maximum of these. It's called the MAP rule. And it doesn't require you to calculate this quantity, which is

sometimes a mess. All it requires you to do is to compare these two quantities, which means you have to compare these two quantities.

AUDIENCE: It's 10 o'clock.

PROFESSOR: Well, excuse me. Yes. Yes, I know.

These things become clearer if you state them in terms of what you call the likelihood ratio. Likelihood ratio only works when you have two hypotheses. When you have two hypotheses, you call the ratio of one of them to the other one the likelihood ratio.

Why do I put 0 up here and 1 down here? Absolutely no reason at all, it's just convention. And unfortunately, it's a convention that not everybody follows. So some people have one convention and some people have another convention. If you want to use the other convention, just imagine switching 1 and 1 in your mind. They're both just binary numbers.

Then, when you want to look at this MAP rule, the MAP rule is choosing the larger of these two things, which we had back here. That's choosing whether this is larger than this, or vice versa, which is choosing whether this ratio here is greater than the ratio of P1 to P0. So that's the same, that's the same thing.

So the MAP rule is to calculate the likelihood ratio for this given observation y. And if this is greater than P1 over P0, you select H equals 0. If it's less than or equal to P1 over P0, you select H1.

Why do I put the strict equality here and the strict inequality here? Again, no reason whatsoever. When you have equality, it doesn't make any difference which you choose. So you could flip a coin. It's a little easier if you just say, we're going to do this under this condition. So we state condition this way. We calculate the likelihood ratio. We compare it with a threshold. The threshold here is P1 over P0. And then we select something.

Why did I put a little hat over this?

AUDIENCE: Estimation.

PROFESSOR: What?

AUDIENCE: Because it's an estimation.

PROFESSOR: What?

AUDIENCE: It's an estimation?

PROFESSOR: Well, it's not really an estimation. It's a detection. I mean, estimation you usually view as being analog. Detection you usually view as being digital. And thanks for bringing that up because it's an important point. But in this model, H is either 0 or 1 in the result of this experiment. We don't know which it is. This is what we've chosen. So H hat is 0 does not mean that H itself is 0. So this is our choice. It might be wrong or it might be right.

Many decision rules, including the most common and the most sensible, are rules that compare lambda of y to a fixed threshold, say, eta, is P1 over P0, which is independent of y, which is just a fixed threshold.

The decision rules then vary only in the way that you choose the threshold.

Now, what happens as soon as I call this eta instead of P1 over P0? My test becomes independent of these a priori probabilities that statisticians have thought about for so long. Namely, after a couple of lines of fiddling around with these things, suddenly all of that has disappeared. We have a threshold test. The threshold test says, take this ratio-- everybody agrees that there's such a ratio that exists-- and compare it with something.

And if it's bigger than that something, you choose 0. If it's less than that thing, you choose 1. And that's the end of it.

OK, so we have two questions. One, do we always want to use a threshold test or are there cases where we should use things other than a threshold test? And the second question is, if we're going to use a threshold test, where should we set the

threshold? I mean, there's nothing that says that you really want to minimize the probability of error. I mean, suppose your test is to see whether-- I mean, something in the news today.

I mean, you'd like to take an experiment to see whether your nuclear plant is going to explode or not. So you come up with one decision, it's not going to explode. Or another decision, you decide it will explode. Presumably on the basis of that decision, you do all sorts of things. Do you really want to make it a maximum a posteriori probability decision? No. You recognize that if it's going to explode, and you choose that it's not going to explode and you don't do anything, there is a humongous cost associated with that.

If you decide the other way, there's a pretty large cost associated with that also. But there's not really much comparison between the two. But anyway, you want to do something which takes those costs into account.

One of the problems in the homework does that. It's really almost trivial to readjust this problem, so that you set the threshold to involve the costs also. So if you have arbitrary costs in making errors, then you change the threshold a little bit. But you still use a threshold test.

There's something called maximum likelihood that people like for making decisions. And maximum likelihood says you calculate the likelihood ratio. And if the likelihood ratio is bigger than 1, you go this way. If it's less than 1, you go this way. It's the MAP test if the two a priori probabilities are equal. But in many cases, you want to use it whether or not the a priori probabilities are equal. It's a standard test, and there are many reasons for using it. Aside from the fact that the a priori probabilities might be chosen that way. So anyway, that's one other choice.

When we go a little further day, we'll talk about a Neyman-Pearson test. The Neyman-Pearson test says, for some reason or other, I want to make sure that the probability that my nuclear plant doesn't blow up is less than, say, 10 to the minus fifth. Why 10 to the minus fifth? Pull it out of the air. Maybe 10 to the minus sixth,

that point our probabilities don't make much sense anymore. But however we choose it, we choose our test to say, we can't make the probability of error under one hypothesis bigger than some certain amount alpha than what test will minimize the probability of error under the other hypothesis.

Namely, if I have to get one thing right, or I have to get it right almost all the time, what's the best I can do on the other alternative? And that's the Neyman-Pearson test. That is a favorite test among the non-Bayesians because it doesn't involve the a priori probabilities anymore. So it's a nice one in that way. But we'll see, we get it anyway using a probability model.

OK, let's go back to random walks just a little bit to see why we're doing what we're doing.

The logarithm of the threshold ratio is logarithm of this lambda of y. I'm taking m observations. I'm putting that in explicitly, is the sum from N equals 1 to m of the log of the individual ratio.

In other words, when you-- under hypothesis 0, if I calculate the probability of vector y given H equals 0, I'm finding the probability of n things which are IID. So what I'm going to find this probability density is taking the product of the probabilities of each of the observations.

Most of you know now that any time you look at a probability, which is a product of observations, what you'd really like to do is to take the logarithm of it. So you're talking about a sum of things rather than a product of things because we all know how to add independent random variables. So the log of this likelihood ratio, which is called the log likelihood ratio as you might guess, is just a sum of these likelihood ratios.

If we look at this for each m greater than or equal to 1, then given H equals 0, it's a random walk. And given H equals 1, it's another random walk. It's the same sequence of sample values in both cases. Namely, as an experimentalist, we're taking these observations. We don't know whether H equals 0 or H equals 1 is what

the result of the experiment is going to be. But what we do know is we know what those values are. We can calculate this sum. And now, if we condition this on H equals 0, then this quantity, which is fixed, has a particular probability of occurring. So this is a random variable then under the hypothesis H equals 0. It's a random variable under the hypothesis H equals 1. And this sum of random variables behaves in a very different way under these two hypotheses.

What's going to happen is that under one hypothesis, the expected value of this log likelihood ratio is going to linearly increase with n. If we look at it under the other hypothesis, it's going to linearly decrease as we increase n.

And a nifty test at that point is to say, as soon as it crosses a threshold up here or a threshold down here, we're going to make a decision. And that's called a sequential test in that case because you haven't specified ahead of time, I'm going to take 100 tests and then make up my mind. You've specified that I'm going to take as many tests as I need to be relatively sure that I'm getting the right decision, which is what you do in real life.

I mean, there's nothing fancy about doing sequential tests. Those are the obvious things to do, except they're a little more tricky to talk about using probability theory.

But anyway, that's where we're headed. That's why we're talking about hypothesis testing. Because when you look at it in this formulation, we get a random walk. And it gives us a nice example of when you want to use random walks crossing a threshold as a way of making decisions. OK, so that's why we're doing what we're doing.

Now, let's go back and look at threshold tests again, and try to see how we're going to make threshold tests, what the error probabilities will be, and try to analyze them a little more than just saying, well, a MAP test does this. Because as soon as you see that a MAP test does this, you say, well, suppose I use some other test. And what am I going to suffer from that? What am I going to gain by it? So it's worthwhile to, instead of looking at even just threshold tests, to say, well, let's look at any old test at all.

Now, any test means the following. I have this probability model. I've already bludgeoned you into accepting the fact that's the probability model we're going to be looking at. And we have this-- well, we have the likelihood ratio, but we don't care about that for the moment. But we make this observation. We got to make a decision. And our decision is going to be either 1 or 0.

How do we characterize that mathematically? Or how do we calculate it if we want a computer to make that decision for us?

The systematic way to do it is for every possible sequence of y to say ahead of time to give a formula, which sequences get mapped into 1 and which sequences get mapped into 0. So we're going to call a set A the set of sample sequences that get mapped into hypothesis 1. That's the most general binary hypothesis test you can do. That includes all possible ways of choosing either 1 or 0. You're forced to hire somebody or not hire somebody. You can't get them to work for you for two weeks, and then make a decision at that point. Well, sometimes you can in this world. But if it's somebody you really want and other people want them, too, then you've got to decide, I'm going to go with this person or I'm not going to go with them.

So under all observations that you've made, you need some way to decide which ones make you go to decision 1. Which ones make you go to decision 0. So we will just say arbitrarily, there's a set A of sample sequences that map into hypothesis 1. And the error probability for each hypothesis using test A is given by-- and we'll just call Q sub 0 of A-- this is our name for the error probability.

Have I twisted this up? No. Q sub 0 of A is the probability that I actually choose-- it's the probability that I choose A given that the hypothesis is 0. Q sub 1 of A is the probability that I choose 1. Blah, let me start that over again.

Q0 of A is the probability that I'm going to choose hypothesis 1 given that hypothesis 0 was the correct hypothesis. It's the probability that Y is in A. That means that H hat is equal to 1 given that H is actually 0. So that's the probability we make an error given the hypothesis, the correct hypothesis is 0. Q1 of A is the

probability of making an error given that the correct hypothesis is 1.

If I have a priori probabilities, I'm going back to assuming a priori probabilities again. The probability of error is? It's P0 times the probability I make an error given that H equals zero. P1 a priori probability of 1 given that I make an error given 1. I add these two up. I can write it this way. Don't ask for the time being. I'll just take the P0 out, so it's Q0 of A plus P1 over P0 Q1 of A. So that's what I've called eta times Q1 of A.

For the threshold test based on eta, the probability of error is the same thing. But that A there is an eta. I hope you can imagine that quantity there is an eta. This is an eta. So it's P0 times Q0 of eta plus eta times Q1 of eta. So the eta probability, under this crazy test that you've designed, is P0 times this quantity. Under the MAP test, probability of error is this quantity here.

What do we know about the MAP test? It minimizes the error probability under those a priori probabilities. So what we know about it is that this quantity is less than or equal to this quantity. Take out the P0's and it says that this quantity is less than or equal to this quantity. Pretty simple.

Let's draw a picture that shows what that means. Here's a result that we have. We know because of maximum a posteriori probability for the threshold test that this is less than or equal to this. This is the minimum error probability. This is the error probability you get with whatever test you like. So let's draw a picture on a graph where the probability of error given H equals 1 is on the horizontal axis. The probability of error conditional on H equals 0 is on this axis. So I can list the probability of error for the threshold test, which sits here. I can list the probability of error for this arbitrary test, which sits here. And I know that this quantity is greater than or equal to this quantity. So the only thing I have to do now is to sort out using plain geometry, why these numbers are what they are.

This number here is Q0 of eta plus eta times Q1 of eta. Here's Q1 of eta. This distance here is Q1 of eta. We have a line of slope minus eta there that we've drawn. So this point here is, in fact, Q0 of eta plus eta times Q1 of eta . That's just

plain geometry. This point is Q0 of A plus eta times Q1 of A. Another line of slope minus et. What we've shown is that this is less than or equal to this. That's because of the MAP rule. This has to be less than or equal to that. So what have we shown here?

We've shown that for every test A you can imagine, when you draw that test on this two-dimensional plot of error probability given H equals 1 versus error probability given H equals 0. Every test in the world lies Northeast of this line here. Yeah?

AUDIENCE: Can you say again exactly what axis represents what?

PROFESSOR: This axis here represents the error probability given that H equals 1 is the correct hypothesis. This axis is the error probability given that 0 is the correct hypothesis.

So we've defined Q1 of eta and Q0 of eta as those two error probabilities. Using the threshold test, or using the MAP test where eta is equal to P0 over P1. And this point here is whatever it happens to be for any test that you happen to like.

You might have a supervisor who wants to hire somebody and you view that person is a threat to yourself, so you've taken all your observations and you then make a decision. If the person is any good, you say, don't hire him. If the person is good you say, hire them. So just the opposite of what you should do.

But whatever you do, this says this is less than or equal to this because of the MAP rule. And therefore, this point lies up in that direction of this line here.

You can do this for any eta that you want to do it for. So for every eta that we want to use, we get some value of Q0 of eta and Q1 of eta. These go along here in some way. You can do the same argument again. For every threshold test, every point lies Northeast of the line of slope minus eta through that threshold test. We get a whole family of curves when eta is very big, the curve of slope minus eta goes like this. When eta is very small, it goes like this. We just think of ourselves plotting all these curves, taking the upper envelope of them because every test has to lie Northeast of every one of those lines. So we take the upper envelope of all of these lines, and we get something that looks like this. We call this the error curve. And this

is the upper envelope of the straight lines of slope minus eta that go through the threshold tests at eta.

You get something else from that, too. This curve is convex. Why is the curve convex? Well, you might like to take the second derivative of it, but that's a pain in the neck. But the fundamental definition of convexity is that a one-dimensional curve is convex if all of its tangents lie underneath the curve. That's the way we've constructed this. It's the upper envelope of a bunch of straight lines. Yes?

AUDIENCE: Can you please explain, what is u of alpha?

PROFESSOR: U of alpha is just what I've called this upper envelope. This upper envelope is now a function.

AUDIENCE: What's the definition?

PROFESSOR: What?

AUDIENCE: What is the definition?

PROFESSOR: The definition is the upper envelope of all these straight lines.

AUDIENCE: For changing eta?

PROFESSOR: What?

AUDIENCE: For changing eta?

PROFESSOR: Yes. As eta changes, I get a whole bunch of these points. I got a whole bunch of these points. I take the upper envelope of all of these straight lines.

I mean, yes, you'd rather see an equation. But if you see an equation it's terribly ugly. I mean, you can program a computer to do this. as easily as you can program it to follow a bunch of equations.

But anyway, I'm not interested in actually solving for this curve in particular. I am particularly interested in the fact that this upper envelope is, in fact, a convex curve

and that the threshold tests lie on the curve. The other tests lie Northeast of the curve. And that's the reason you want to use threshold tests. And it has nothing to do with a priori probabilities at all.

So you see, the thing we've done is to start out assuming a priori probabilities. We've derived this neat result using a priori probabilities. But now we have this error curve. Well, to give you a better definition of what u of alpha is, u of alpha is the error probability under hypothesis 1 if the error probability under hypothesis 0 was alpha. You pick an error probability here. You go up to that point here. There's a threshold test there. You read over there. And at that point, you find the probability of error given H equals 1.

AUDIENCE: How do you know that the threshold tests lie on the curve?

PROFESSOR: Well, this threshold test here is Southwest of all tests. And therefore, it can't lie above this upper envelope.

Now, I've cheated you in one small way. If you have a discrete test, what you're going to wind up with is just a finite set of these possible points here. So you're going to wind up with the upper envelope of a finite set of straight lines. So the straight line is actually going to be-- it's still convex, but it's piecewise linear. And it's piecewise linear, and the threshold tests are at the points of that curve. And in between those points, you don't quite know what to do. So since you don't quite know what to do in between those points, as far as the maximum a posteriori probability test goes, you can reach any one of those points, sometimes using one test on one corner of-- I guess it's easier if I draw it.

And I didn't want to get into this particularly because it's a little messier. So you could have this kind of curve. And the notes talk about this in detail.

So the threshold test correspond to this point. This point says always decide one. Don't pay any attention to the tests at all, just say I think one is the right hypothesis. I mean, this is the testing philosophy of people who don't believe in experimentalism. They've already made up their mind. They look at the results.

They say, that's very interesting. And then they say, I'm going to choose this. These other points are our particular threshold tests.

If you want to get error probabilities in the middle here, what do you do? You use a randomized test. Sometimes you use this. Sometimes you use this. You flip a coin and choose whichever one of these you want to choose. So what this says is the Neyman-Pearson test, which is the test that says pick some alpha, which is the error probability under hypothesis 1 that you're willing to tolerate. So you pick alpha. And then it says, minimize the error probability of the other kind, so you read over there. And the Neyman-Pearson test, what it does is it minimizes the error probability under the other hypothesis.

Now, when this curve is piecewise linear, the Neyman-Pearson test is not a threshold test, but it's a randomized threshold test. Sometimes when you're at a point like this, you have to use this test and this test sometimes. For most of the tests that you deal with, Neyman-Pearson test is just the threshold test that's at that particular point.

Any questions about that? This is probably one of these things you have to think about a little bit. Yes?

AUDIENCE: When you say you have to use this test or this test, are you talking about threshold or are you talking about-- because this is always-- it's either H equals 0 or H equal 1, right? What do you mean when you say you have to randomize between the two tests?

I mean threshold tests-- if I have a finite set of alternatives, and I'm doing a threshold test on that finite set of alternatives, I only have a finite number of things I can do. As I increase the threshold, I suddenly get to the point where this ratio of likelihoods includes one more point. And then it gets to the point where it includes one other point and so forth. So that what happens is that this upper envelope is just the upper envelope of a finite number of points. And this upper envelope of a finite number of points, the threshold tests are just the corners there. So I sometimes have to randomize between them.

If you don't like that, ignore it. Because for most tests you deal with, almost all books on statistics that I've ever seen, it just says the Neyman-Pearson test looks at the threshold curve, at this error curve. And it chooses accordingly. Yes?

AUDIENCE: You can put the previous slide back? You told us that because of maximum a posteriori probability, if eta is equal to P0 divided by P1, then the probability of error is minimized. And so the errors of the test A are [INAUDIBLE]. But if we start changing eta from 0 to infinity, it doesn't have to be anymore. [INAUDIBLE], which means the error is not necessarily minimized. So the argument doesn't hold anymore.

PROFESSOR: As I change eta, I'm changing P1 and P0 also. In other words, now what I'm doing is I'm saying, let's look at this threshold test, and let's visualize what happens as I change the a priori probabilities. So I'm suddenly becoming a classical statistician instead of a Bayesian one. But I know what the answers are from looking at the Bayesian case.

OK, so let's move on. I mean, we now sort of see that these tests-- well, one thing we've seen is when you have to make a decision under this kind of probabilistic model we've been talking about-- namely, two hypotheses, IID random variable is conditional on each hypothesis. Those hypothesis testing problems turn into random walk problems.

We also saw that the [? GG1Q ?] when I started looking at when the system becomes empty, and how long it takes to start to fill up again, that problem is a random walk problem. So now I want to start to ask the question, what's the probability that a random walk will cross a threshold? I'm going to apply the Chernoff bound to it. You remember the Chernoff bound? We talked about it a little bit back on the second week of the term. We were talking about the Markov inequality and the Chebyshev inequality. And we said that the Chernoff inequality was the same sort of thing, except it was based on e to the rZ rather than x or x squared. And we talked a little bit about its properties.

The major thing one uses the Chernoff bound for is to get good estimates very, very far away from the mean. In other words, good estimates of probabilities that are very, very small.

I've grown up using these all my life because I've been concerned with error probabilities in communication systems. You typically want error probabilities that run between 10 to the minus fifth and 10 to the minus eighth. So you want to look at points which are quite far away. I mean, you take a large number of-- you take a sum of a large number of variables, which correspond to a code. And you look at error probabilities for this rather complicated thing. But you're looking very, very far away from the mean, and you're looking at very large numbers of observations. So instead of the kinds of things where we deal with things like the central limit theorem where you're trying to figure out what goes on close to the mean, here you're trying to figure out what goes on very far from the mean.

OK, so what the Chernoff bound says is that the probability that a random variable Z is greater than or equal to some constant b. We don't even need sums of random variables here, it's just a Chernoff bound is a bound on the tail of a distribution. Is less than or equal to the moment generating function of that random variable. g sub Z of r is the expected value of e to the rZ. These generating functions, you can calculate them if you want to. Times e to the minus rb. This is the Markov inequality for the random variable e to the rZ.

And go back and review chapter 1. I think it's section 1.43 or something. It's the section that deals with the Markov inequality, the Chebyshev inequality, and the Chernoff bound. And as I told you once when we talked about these things, Chernoff is still alive and well. He's a statistician at Harvard. He was somewhat embarrassed by this inequality becoming so famous because he did it as sort of a throw-off thing in a paper where he was trying to do something which was much more mathematically sophisticated. And now the poor guy is only known for this thing that he views as being trivial.

But what the bound says is the probability of Z is greater than or equal to b is this

inequality. Strangely enough, the probability that Z is less than or equal to b is bounded by the same inequality. But one of them, r is bigger than 0. And the other one, r is less than 0. And you have to go back and read that section to understand why.

Now, this is most useful when it's applied to a sum of random variables. I don't know of any applications for it otherwise. So if the moment-generating function-- oh, incidentally, also. When most people talk about moment-generating functions, and certainly when people talked about moment-generating functions before the 1950s or so, what they were always interested in is the fact that if you take derivatives of the moment-generating functions, you generate the moments of the random variable.

If you take the derivative of this with respect to r, evaluate it at r equals 0, you get the expected value of Z. If you take the second derivative evaluated at r equals 0, you get the expected value of Z squared, and so forth. You can see that by just taking the derivative of that.

Here, we're looking at something else. We're not looking at what goes on around r equals 0. We're trying to figure out what goes on way on the far tails of these distributions.

So if gX of r is e to the rX, then e to the e to the r Sn-- Sn is the sum of these random variables-- is the expected value of the product of e to the rXi. Namely, it's e to the r. Some of Xi. So that turns into a product. The expected value of a product of a finite number of terms is the product of the expected value. So it's gX or r to the n-th power.

So if I want to write this, now I'm applying the Chernoff bound to the random variable S sub n. What's the probability that S sub n is greater than or equal to n times a? It's gX to the n of r times e to the minus rna. That's what the Chernoff bound says. This is the Chernoff bound over on the other side of the distribution. This only makes sense and has interesting values when a is bigger than the mean or when a is less than the mean. And when r is greater than 0 for this one and less

than 0 for this one.

Now, this is easier to interpret and it's easier to work with. If you take that product of terms g to the r to the n-th power and you visualize the logarithm of g to the X. Visualize the logarithm of g to the X, then you get this quantity up here. You get the probability that Sn is greater than or equal to na is this e to the n times gamma x of r minus ra. Gamma is the logarithm of the moment-generating function. The logarithm of the moment-generating function is always called the semi-invariant moment-generating function. The name is, again, because people were originally interested in the moment-generating properties of these random variables.

If you sit down and take the derivatives, I can probably do it here. It's simple enough that I won't get confused. The derivative with respect to r of the logarithm of g of r is first derivative of r divided by g of r. And the second derivative is then the natural log of g of r. Taking the derivative of that is equal to g double prime of r over g of r squared.

Tell me if I'm making a mistake here because I usually do when I do this. Minus g of r and g prime of r. Probably divided by this squared. Let's see. Is this right? Who can take derivatives here?

AUDIENCE: First term doesn't have a square in it.

PROFESSOR: What?

AUDIENCE: First term doesn't have a square in the denominator.

PROFESSOR: First term? Yeah. Oh, the first thing doesn't have a square. No, you're right.

AUDIENCE: Second one doesn't have--

PROFESSOR: And the second one, let's see. We have-- we just have g prime of r squared divided by g of r squared. And we evaluate this at r equals 0. This term becomes 1. This term becomes 1. This term becomes the second moment x squared bar. And this term becomes x bar squared. And this whole thing becomes the variance of the moment of the random variable rather than the second moment.

All of these terms might be wrong, but this term is right. And I'm sure all of you can rewrite that and evaluate it at r equals 0. So that's why it's called the semi-invariant moment-generating function. It doesn't make any difference for what we're interested in. The thing that we're interested in is that this exponent here-- as you visualize doing this experiment and taking additional observations, what happens is the probability that you exceed na-- that the n-th sum exceeds n times some fixed quantity a is going down exponentially with [? the a. ?]

Now, is this bound any good? Well, if you optimize it over r, It's essentially exponentially tight. So, in fact, it is good.

What does it mean to be exponentially tight? That's what I don't want to define carefully. There's a theorem in the notes that says what exponentially tight means. And it takes you half an hour to read it because it's being stated very carefully. What it says essentially is that if I take this quantity here and I subtract-- I add an epsilon to it. Namely, e to the n times this quantity minus epsilon. So I have an e to the minus n epsilon, see it sitting in there? When I take this exponent and I reduce it just a little bit, I get a bound that isn't true. This is greater than or equal to the quantity with an epsilon.

In other words, you can't make an exponent that's any smaller than this. You can take coefficients and play with them, but you can't make the exponent any smaller.

OK, all of these things you can do them by pictures. I know many of you don't like doing things by pictures. I keep doing them by pictures because I keep trying to convince you that pictures are more rigorous than equations are. At least, many times.

If you want to show that something is convex, you try to show that the second derivative is positive. That works sometimes and it doesn't work sometimes. I mean, it works as a function is continuous and has a continuous first derivative. It doesn't work. otherwise.

When you start taking tangents of the curve, and you say the upper envelope of the tangents to the curve all lie below the function, then it works perfectly. That's what a convex function is by definition.

How do we derive all this stuff? What we're trying to do is to find-- I mean, this inequality here is true for all r, for all r greater than 0 so long as a is greater than the mean of X. It's true for all r for which this moment-generating function exists. Moment-generating functions can sometimes blow up, so they don't exist everywhere. So it's true wherever the moment-generating function exists. So we like to find the r for which this bound is tightest.

So what I'm going to do is draw a picture and show you where it's tightest in terms of the picture. What I've drawn here is the semi-invariant moment-generating function. Why didn't I put that down? This is gamma of r. Gamma of r at 0, it's the log of the moment-generating function at 0, which is 0. It's convex. You take its second derivative. Its second derivative at r equals 0 is pretty easy. Its second derivative of other values or r you have to struggle with it. But when you struggle a little bit, it is convex.

If you've got a curve that goes down like this, then it goes back up again. Sometimes goes off towards infinity. Might do whatever it wants to do. Sometimes at a certain value of r, it stops existing.

Suppose I take the simplest random variable you know about. You only know two simple random variables. One of them is a binary random variable. The other one's an exponential random variable. Suppose I take the exponential random variable with density alpha times e to the minus alpha X. Where does this moment-generating function exist? You take alpha and I multiply it by e to the rX when I integrate it. Where does this exist?

I mean, don't bother to integrate it. If r is bigger than alpha, this exponent is bigger than this exponent. And this thing takes off towards infinity. If r is less than a, the whole thing goes to 0. gX of r exists for r less than alpha in this case.

And in general, if you look at a moment-generating function, if the tail of that distribution function is going to 0 exponentially, you find the rate at which it's going to 0 exponentially. And that's where the moment-generating function cuts off. It has to cut off. You can't show a result like this, which says something is going to 0, faster than it could possibly be going to 0. So we have to have that kind of result. But anyway, we draw this curve. This is mu sub X of r. And then we say, how do we graphically minimize gamma of r minus r times a?

Well, what I do because I've done this before and I know how to do it-- I mean, it's not the kind of thing where if you sat down you would immediately settle on this. I look at some particular value of r. If I take a line of slope gamma prime of r, that's a tangent to this curve because this curve is convex. So if I take a line through here of this slope and I look at where this line hits here, where does it hit? It hits at gamma sub X of r, this point here, minus gamma X of r-- oh.

Well, what I've done is I've already optimized the problem. I'm trying to find the probability that Sn is greater than or equal to na. I'm trying to minimize this exponent here, gamma X of r minus ra. Unfortunately, I really start out by taking the derivative of that and setting it equal to 0, which is what you would all do, too.

When I set the derivative of this equal to 0, I get gamma prime of r minus a is equal to 0, which is what this says. So then we take a line of slope gamma x of r equals 0. It's tangent at this point here. You look at this point over here and you get the minimum value of the gamma X of r minus r0 a.

So what this says is when you vary a, you can go through this maximization tilting this curve around. I mean, a determines the slope of this line here.

If I use a smaller value of a, the slope is smaller. It hits in here. If I take a larger value of a, it comes in further down and the exponent gets bigger. That's not surprising. I want to find out the probability that S sub n is greater than or equal to a. As I increase a, I expect this exponent to keep going down as I make a bigger and bigger because it's harder and harder for it to be greater than or equal to a.

So anyway, when you optimize this, you get something exponentially tight. And this is what it's equal to. And I would recommend that you go back and read the section of chapter 1, which goes through all of this in a little more detail.

Let me go passed that. Don't want to talk about that.

Well, when I do this optimization, if what I'm looking at is the probability that S sub n is greater than or equal to some alpha rather than n times a when I'm do this optimization and I'm looking at what happens at different values of n, it turns out that when n is very big, you get something which is tangent there. As n gets smaller, you get these tangents that come down that comes in to there, and then it starts going back out again. This e to the r star is the tightest the bound ever gets. That's the n at which errors in the hypothesis testing usually occur. It's the point at which-- it's the n for which Sn greater than or equal to alpha is most likely to occur.

And if you evaluate that for our friendly binary case again, X equals 1 or X equals minus 1, what you find when you evaluate that point alpha r star is that r star is equal to log 1 minus P over P. And our bound of probability union of Sn is greater than or equal to alpha is approximately e to the minus alpha r star is 1 minus P over P to the minus alpha.

I mean, why do I torture you with this? Because we solved this problem at the beginning of the lecture, remember? The probability that the sum S sub n for this binary experiment is greater than or equal to k is equal to 1 minus P over P to the minus k. That's what it's equal to exactly.

When I go through all of this Chernoff bound stuff, I get the same answer. Now, this is a much harder way to do it, but this is a general way of doing it. And that's a very specialized way of doing it. So we'll talk more about this next time.